Modeling Time-to-Death of HIV Positive Patients under Highly Active Antiretroviral Therapy : A Case of Jimma University Specialized Hospital

College of Natural Science

Department of Statistics

A Thesis to be submitted to the School of Graduate Studies Jimma University Department of Statistics in Partial Fulfillment of the Requirements for the Degree of Master of Science in Biostatistics

By:

Zeleke Baheda

October, 2015

Jimma, Ethiopia

Modeling Time-to-Death of HIV Positive Patients under Highly Active Antiretroviral Therapy: A Case of Jimma University Specialized Hospital

MSc Thesis

By:
Zeleke Baheda

Advisor: Dr. Yehenew Getachew(PhD)
Co-advisor: Mr. Tesfaye Getachew(Msc.)

October, 2015
Jimma, Ethiopia

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES, DEPARTMENT OF STATISTICS

As thesis research advisors, we hereby certify that we have read the thesis prepared by Zeleke Baheda under our guidance, which is entitled "**Modeling Time -to -Death of HIV Positive Patients under Highly Active Antiretroviral Therapy A Case of Jimma University Specialized Hospital**", in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill University and Department style requirements; (2) its illustrative materials including tables and figures are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the University library.

Dr. Yehenew Getachew	_____	_____
Advisor	Signature	Date
Mr. Tesfaye Getachew	_____	_____
Co-advisor	Signature	Date

As the members of the board of examiners of M.Sc. thesis open defense examination, we certify that we have read and evaluated the thesis and examined the candidate. Hence, we recommend that the thesis be accepted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

_____	_____	_____
Name of Chairman	Signature	Date
_____	_____	_____
Name of Major Advisor	Signature	Date
_____	_____	_____
Name of Co-advisor	Signature	Date
_____	_____	_____
Name of Internal Examiner	Signature	Date
_____	_____	_____
Name of External Examiner	Signature	Date

STATEMENT OF AUTHOR

I declare that this thesis is a result of my genuine work and all sources of materials used, for writing it, have been duly acknowledged. I have submitted this thesis to Jimma University in the partial fulfillment for the degree of Master of Science in Biostatistics. The thesis can be deposited in the University library to be made available to borrowers for reference. I solemnly declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Brief quotations from this thesis are allowed without requiring special permission provided that an accurate acknowledgement of the source is made. Requisites for extended quotations for the reproduction of the thesis in whole or in part may be granted by the head of the department of statistics when in his or her judgment the proposed use of the material is for a scholarly interest. In all other instances, however, permission must be obtained from the author.

Zeleke Baheda

Date: _____

Signature: _____

October, 2015

Jimma, Ethiopia

DEDICATION

This thesis is dedicated to my family, my wife Getie Bessir and all those who supported me to make my educational dream come true.

ACKNOWLEDGEMENT

First; I am very grateful to Dr. Yehenew Getachew, my thesis advisor and instructor for his invaluable comments and suggestions that contributed to the successful realization of the thesis. I also gratefully acknowledge my co-advisor Tesfaye Getachew (Msc.) for his continuous and passionate support by giving me constructive comments and suggestions which made me to concentrate on my thesis and my special thanks go to Belay Berlie (PhD scholar) who give me essential suggestions and future directions that really gives me a great motivations concerning the survival modeling.

My stay in the university was also a great pleasure and I am highly indebted to all of the staff members of department of statistics. I also extend my great full thank to Semera University for giving me the chance to attend the program as well as for all the financial support in two years of my stay in Jimma University.

I am very much indebted to my friends and family members, especially to my brother Hailu Baheda who highly contributed to my life by giving me motivations as well as for his financial support starting from my childhood to today. Finally, I would to thank Jimma University specialized hospital for allowing me the data.

ABSTRACT

Back ground: The survival time of HIV positive patients depends not only on the quality of care and highly active anti-retroviral therapy, but also on a host of other baseline demographic factors and clinical characteristics of the patients. The use of the appropriate model will certainly lead to identify more reliable and precise predictors of patients' survival time after initiation of ART and thereby help to have a more effective treatment program.

Objective: To model the time-to-death of HIV positive patients under highly active antiretroviral therapy using Cox PHs and AFT models.

Methods: The study consists of 440 HIV positive patients who were aged 18 years or above and who were placed under HAART any time in between 1st January 2010 to 30th June 2015 in JUSH. The data were analyzed using Cox PHs a Semi-Parametric model and AFT a parametric models. The performances of these models were compared using AIC criteria. For nested models using LR test and for non-nested models using AIC

Results: We fit a cox proportional model and check the proportionality assumption. For all variables except baseline CD4 count the assumption holds. To overcome the violation of proportional hazards we fit two alternative models; time-dependent cox PH and AFT model and compare their goodness of fit. The result shows the lognormal AFT model was the "best" fitting model for this dataset. Using the selected model Functional status, WHO clinical disease stage, pre-TB positive test, regimen type at ART initiation, body mass index and baseline CD4 count were found to be significantly associated with survival time of patients on HAART.

Conclusion: The AFT model is a more valuable and realistic alternative to the Cox PH model in situation when the PH assumption does not hold and therefore should be considered as an alternative to the Cox PH in modeling time-to-death of patients under HAART. Patients with low CD4 count, WHO stage III and IV, being underweight, pre-TB positive test and being ambulatory and bedridden are factors that accelerate survival time on HAART.

Table of Contents

List of Tables

LIST OF FIGURES

LIST OF ACRONYMS

AFT	Accelerated failure time model
AIC	Akaike information criterion
AIDS	Acquired Immune Deficiency Syndrome
ANC	Antenatal care
ART	Antiretroviral Therapy
CD4	Cluster of Differentiation 4 T-helper cells
HAART	Highly active antiretroviral therapy
HAPCO	HIV/AIDS Prevention and Control Office
HIV	Human Immune Deficiency Virus
HR	Hazard ratio
JUSH	Jimma University Specialized Hospital
LOWESS	Locally weighted least squares estimation
LRT	Likelihood Ratio Test
OARAC	Office of AIDS Research Advisory Council
PH	Proportional Hazard
PLWHA	People Living with HIV/AIDS
SV	Standardized Variability
TB	Tuberculosis
TR	Time ratio
UNAIDS	United Nations Program of HIV/AIDS
WHO	World Health Organization

1. INTRODUCTION

1.1. Back ground of the study

HIV/AIDS: HIV is the virus that causes AIDS. It attacks the immune system cell called the T-helper or CD4 cells. CD4 cells are crucial to the normal function of the immune system, which defends the body against infections. HIV causes many T-helper cells to be damaged or destroyed. As a result, a major reduction in the number of T-helper cells can have a serious effect on the immune system to fight illnesses. As HIV infection progresses and replicate within its human host, it eventually weakens the infection fighting cells called T-cell lymphocytes. This makes the host susceptible to opportunistic infections (e.g. the top list mycobacterium that cause TB) as well as other complications of HIV disease which cause severe or fatal health problems.

1.2. Burden of HIV/AIDS

The expansion of the epidemic has now become a burning issue globally and this is particularly so more in developing countries; specially, in Sub Saharan Africa (HAPCO, 2012). In fact, Sub-Saharan Africa accounts for 22.4 million infections, which is about 67% of the total HIV burden. The number of people estimated to acquire new infections is around 1.9 million accounting for 68 % of the total number of new infections (UNAIDS/WHO, 2009). This was noted as a significant reduction in the number of new cases since 2001.

However, it was also reported that HIV/AIDS has become the leading cause of death in the region (HAPCO, 2012). Ethiopia is one of the Sub-Saharan African countries most severely affected by the HIV/AIDS pandemic. Currently, the national adult prevalence rate is estimated at 2.3 percent and an estimated number of 1.2 million people are living with HIV/AIDS. An estimated 67,000 lost their lives due to AIDS at the end of 2007 (USAID, 2010). According to the Country Progress Report on HIV/AIDS, 2012 by the HIV/AIDS Prevention and Control Office (HAPCO), an estimated 790,000 persons were infected with HIV in 2011, with an adult prevalence in 15-49 year olds of 1.4%, and 950,000 orphans were of parents who had died of AIDS.

1.3. Highly active antiretroviral therapy (HAART)

HAART: refers to the use of combinations of various antiretroviral drugs with different mechanisms of action to treat for those infected with HIV. It usually involves a combination of at least three drugs from two or more classes of antiretroviral treatment (Boskey, 2010). The main effect of HAART is to suppress viral replication, allowing the individual's immune system to recover and protecting him/her from the development of AIDS and death.

During these years HIV infection has changed from a fatal condition to a manageable chronic illness mainly due to the development of antiretroviral therapy (ART). ART for the treatment of HIV infection has been shown to profoundly alter HIV disease progression, including incidence of opportunistic infections in PLWHA (Khamofu *et al*, 2012). The goal of this therapy is to improve survival; to reduce HIV associated morbidity and mortality, to increase the quality of life, to restore immune function and to achieve maximal and sustained suppression of viral replication (OARAC, 2008). As access to antiretroviral therapy (ART) improves, the population living with HIV increases as fewer people die from AIDS-related causes. In 2012, an estimated 35.3 million people were living with HIV – with 9.7 million people in low- and middle-income countries receiving ART. It has been estimated that during the period 1995–2012, ART cumulatively averted 5.5 million deaths in such countries .Globally, an estimated 1.6 million people died of HIV/AIDS in 2012; down from the peak of 2.3 million in 2005 (WHO,2014).

As a result of the rapid scaling-up of the antiretroviral treatment (ART) program, nearly 60% of PLWHA with CD4 counts below 350 received ART in 2011. Ethiopia is one of the few countries in Sub-Saharan Africa that achieved more than a 25% decline in new HIV infections, as revealed by a decline of new infections among ANC attendees from 5.6% in 2005 to 3.5% in 2007 and 2.6% in 2011 (HAPCO 2012, reference no. 26 in the Prevention Research section). But, the critical issue, to the success of ART, is retention to the treatment regime since ART is a lifelong commitment that requires patients to adhere diligently to daily medication, dosing schedules and make frequent clinic visits for care.

Survival models: Survival analysis is a collection of statistical method to analyze time-to-event data where the outcome variable of interest is the "time to the occurrence of an event" This variable is also often called "survival time" The "survival time" refers to a number of years, months, weeks or days from the beginning of the patient observance till the occurrence of an observed event in this particularly study case time to-death (Kleinbaum, 2005, Klein *et al*, 2005). Hence, survival analysis is also referred to as "time-to-event analysis", which is applied in a number of applied fields, such as medicine, public health, social science, and engineering.

Non- parametric, semi- parametric and parametric methods are the statistical methods used to analyze a survival data. Non-parametric methods work well for homogeneous samples; they do not determine whether or not certain variables are related to the survival times. This leads to the application of regression methods for analyzing survival data. The standard multiple linear regression models are not well suited to survival data for several reasons. First, survival times are non-normally distributed. Second, censored data result in missing values for the dependent variable (Klein *et al*, 2005).

In survival analysis studies, characterizing the different survival distributions that correspond to different subgroups within a heterogeneous population, a descriptive summary of such a comparison could consist of parametric or semi parametric methods (Kleinbaum, 2005, Klein *et al*, 2005). There are two major regression models used for survival data: proportional hazards model (Cox) as a semi parametric method (Cox, 1972), and accelerated failure time model or linear model representation in log time as a parametric model.

The model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the nature or shape of the baseline hazard function, .that is why Cox's model is referred to as a semi-parametric model for the hazard function (Klein *et al*, 2005, 1997, Kleinbaum *et al*, 2005). This model keeps the baseline hazard as an arbitrary, unspecified, and nonnegative function of time.

It is the most popular and commonly used model by researchers in medical sciences mainly because of its simplicity, and not being based on any assumptions about the survival distribution (Therneau *et al*, 2000). However, Cox PH model has the restriction that proportional hazards assumption holds with time-fixed covariates; and it may not be appropriate in many situations and other modifications such as stratified Cox model or Cox model with time-dependent variables are required for the analysis of survival data (Collett, 2003, Klein *et al*, 2005).

The Accelerated Failure Time (AFT) model is another alternative method for the analysis of survival data. Many of the standard parametric models such as Weibull, Exponential, Lognormal and log logistic are accelerated failure time models (Klein *et al*, 2005). Although the Cox regression model is the most favorable employed technique in survival analysis, parametric models (Andersen *et al*, 1993) do have a number of benefits. Parametric approach offers more in the way of predictions, and the AFT formulation allows the derivation of a time ratio, which is arguably more interpretable than a ratio of two hazards.

The natural history of HIV disease the time from HIV/AIDS diagnosis to death (survival time) has been modeled using a number of parametric and non-parametric survival models, Even if ART treatment has shown significant clinical importance by meeting the goal of therapy, we are still facing a number of deaths that can otherwise be avoided by appropriate interventions on socio demographic factors, CD4 cell counts, HIV viral load and opportunistic infections.

Generally, the purpose of this study is to support an argument for the consideration of the AFT model as an alternative to the Cox PH model in the analysis of time-to-event data in this particular study time-to-death; and to identify factors affecting time-to-death significantly.

1.4. Statement of the problem

Studies have shown that high initial death on HAART initiation from resources limited settings. However, there is dearth of evidence on treatment outcomes and associated significant factors.

The survival time of HIV positive patients depends not only on the quality of care and highly active anti-retroviral therapy (HAART), but also on a host of other baseline demographic factors and clinical characteristics of the patients. Identifying predictors of patients' survival time after initiation of HAART is of importance not only because it enables the physicians to detect the factors whose changes affect patients' survival time, but also helps them to make the best decision about patients' treatment. The use of appropriate model will certainly lead to identify more reliable and precise prognostic factors and thereby help to have a more effective treatment program.

The proportional hazards model (Cox, 1972) and the accelerated failure time models are the two major approaches to the regression analysis of censored data (Cox and Oakes, 1984). Review of literature shows the extensive use of nonparametric methods such as Kaplan-Meier and Cox Proportional hazards model in comparing and identifying factors that affect the survival time of HIV positive patients under HAART; the latter is used when the effect of covariates on the hazard ratio is desired.

The proportional hazards model is used almost exclusively in practice. This is probably due to the fact that this model allows us to estimate and make inference about the parameters without assuming any distribution for the survival time. However, it does have the requirement of proportional hazards. When the PH assumption is not met, it is improper to use standard Cox model as it may entail serious bias and loss of power in estimating and making inference about the effect a given predictors of death (Moran,2008). According to a review paper of survival analysis published in cancer journals, it was found that only five percent of all studies using the Cox PH model check PH assumption (Atman *et al*, 1985).

As reported by (Reid, 1994) the accelerated failure time models i.e. standard parametric models such as the exponential, Weibull, log-logistic, lognormal and the generalized gamma (GG) are accelerated failure time models.

These models are in many ways more appealing because of its quite direct physical interpretation. This model may provide more accurate or more concise summarization of the data than the proportional hazards model in certain applications.

Since recently AFT models have not been used very often and the few usage of these models are found in HIV/TB co infection Survival Data studies (Nawumbeni, 2014 & Qi, Jiezhi, 2009). More recently, an AFT model has been implemented to analysis of the time to AIDS onset in the Women's Interagency HIV Study (Komarek *et al*., 2004). Thus, requirements using models which can provide a closer examination of disease processes for the data have generated interest in assessing factors that affect the survival time of HIV positive patients by fitting a statistical model that can explain the data in most meaningful manner.

In this study, the performance of Cox proportional hazards and accelerated failure time models was compared using HAART data. Therefore, this study intended to address the following research questions:

Which covariates are statistically significantly associated with survival time of HIV positive patients under HAART?

Which parametric model predicts survival time of HIV positive patients under HAART?

Between AFT model and Cox PHs model which is more efficient in analysis of survival time of HIV positive patients under HAART?

1.5. Objectives of the study

General objective

The general objective of this study was to model the time-to-death of HIV positive patients under highly active antiretroviral therapy (HAART) using Cox PHs and AFT models and to identify factors affecting time-to-death significantly.

Specific objectives

This study addresses the following specific objectives:

- To estimate the survival time and compare the survival curves of HIV positive patients among groups.
- To identify factors that influence the Survival time of HIV positive Patients.

- To estimate the effects of covariates on acceleration or deceleration of the survival time of HIV positive Patients.
- To investigate the comparative performance of Cox PHs and AFT models in a survival analysis of patients on HAART.

1.6. Significance of the Study

The results of this study will provide information about the risk factors that have significant impact on the survival of HIV positive patients during treatment. It is useful in the development of an effective HIV care and antiretroviral therapy (ART) patient monitoring system. This study tries to identify death risk extent of patients under these significant factors at different time during their care. The study also helps in comparing the utility of Cox and accelerated failure time (AFT) models, findings will help in making a decision as to which model to apply under specified conditions defined by predictor variables.

1.7. Limitation of the Study

- The study was restricted to adults, and results might not be applicable to infants and children.
- The study presumed that all deaths are caused by HIV infection.
- Parts of information on individuals are missed because of censored observations.
- The study is based on baseline values of the variables of interest.

2. LITERATURE REVIEW

2.1.Non-Parametric and Parametric Models for Studying Time to Event

The semi parametric Cox proportional hazards model is more popular than parametric methods to analyze time-to-event data because no assumption is needed about the shape of the underlying hazard of the event over time. Examples of hazard distributions include exponential, Weibull, and log-logistic. Semi parametric and parametric methods both yield the relative hazard (RH) as the measure of association, allowing researchers to gain insight into the actual risk process from onset of exposure to an event of interest. Some distributions allow modeling of actual failure times. The accelerated failure time (AFT) models produce a "time ratio" (TR) as its measure of association, and the time when the nth percentile of subjects achieves the outcome of interest can be directly estimated. Using time-to-event data in which the underlying hazard is assumed to fit a Weibull distribution.

Biomedical researchers tend to choose semi parametric methods to model time-to-event data, in a study by (Sethi, *et al*, 2009) data was analyzed from a prospective cohort study of 195 adults receiving HIV/AIDS care and highly active antiretroviral therapy in Baltimore they were followed for 1188 visits between February 2000 and December 2001. Kaplan-Meier estimation and cox and Weibull regressions were performed. Results showed that illicit drug users experienced a greater hazard of clinically significant antiretroviral resistance as compared to non-users. Weibull regression demonstrated that a quarter and a half of illicit drug users developed resistance within 5 and 20 months of viral suppression, respectively, compared to 20 and 85 months, respectively, for non-users. Both semi parametric and parametric methods demonstrated an increased hazard of clinically significant resistance associated with illicit drug use. The parametric model facilitated the estimation of elapsed time to resistance associated with illicit drug use. From the study above the relative hazard produced in semi parametric and parametric proportional hazards modeling helped researchers identify risk factors for an outcome of interest.

Parametric models in the accelerated failure time metric are not commonly used despite the time ratio being a more easily interpretable measure of association than the relative hazard.

AFT models also facilitate the estimation of elapsed time between exposure and outcome, which has more clinical interpretability than a hazard ratio. In the analysis of the above study illicit drug use was associated with a doubling of the hazard of rebound with resistance even after adjustment by other factors.

According (Sethi, *et al*, 2009), one could even argue the analysis, as reported, would have little impact on HIV care. However, the finding that a quarter of illicit drug users were predicted to rebound with resistance within 5 months of achieving viral suppression has important implications. This reveals the imminence of rebound with resistance among illicit drug users despite achieving treatment success and emphasizes a need for physicians to ascertain substance use among patients and schedule more frequent follow-up visits for these patients. Researchers conducting survival analyses should consider the use of parametric models. When properly fitted to the data, these models produce inferences identical to those drawn from Cox regression. The estimation of time ratios and elapsed time are especially advantageous as they have interpretations that can directly translate to clinical and public health practice. Concerns about misspecification of the model, while valid, can be minimized by the use of broad classes of parametric models that encompass a wide variety of hazard shapes (Sethi, *et al*, 2009).

In a study by (Pierre De Beaudrap *et al*, 2008), 404 HIV-1- infected Senegalese adult patients were enrolled and data censored as of September 2005. Predictor effects on mortality were first examined over the whole follow-up period (median 46 months) using a Cox model and Shoenfeld residuals. Then, changes of these effects were examined separately over the early and late treatment periods; i.e., less and more than 6-month follow-up. They found out that during the early period, baseline body mass index and baseline total lymphocyte count were significant predictors of mortality (Hazard Ratios 0.82 [0.72-0.93] and 0.80 [0.69-0.92] per 200 cell/mm^3, respectively) while baseline viral load was not significantly associated with mortality. During the late period, viro-immunological markers (baseline CD4-cell count and 6-month viral load) had the highest impact. In addition, the viral load at 6-month was a significant predictor (HR = 1.42 [1.20-1.66]).They concluded that impaired clinical status could explain the high early mortality rate while viro-immunological markers were rather predictors of late mortality.

This study underlined changes over time in mortality predictors among HIV-1 infected patients. Disappearance of the predictive value of prognostic variables may often occur in medical studies.

The previous finding of an early peak in mortality rate prompted them to investigate more carefully the early period after HAART initiation. Whereas the effect of baseline CD4 cell count on the hazard of death remained roughly constant during the whole follow-up, the impacts of BMI and of total lymphocyte count were important immediately after HAART initiation before fading out after 6 months.

It is important to note that the total lymphocyte count has been proposed as a surrogate for CD4 cell count in low income country settings (WHO 2006).However, in their study, the total lymphocyte count was predictive only over very short time-to-event periods, which means that therapy initiation on basis of total lymphocyte count only may occur too late. They pointed out that additional studies are needed to assess the prognosis value of total lymphocyte count. Changes in the extent of predictor effects between the early and the late period were striking. Indeed, the effect of BMI and of total lymphocyte count at baseline disappeared. Therefore, these variables have a high prognostic value during the first months after HAART initiation, but lost it later. On the other hand, whereas they did not find significant association between viral load measured at baseline and subsequent risk of death, viral load measured at 6 months was noted to be important predictor of death for patients who survived until 6 months.

They pointed out that this was consistent with other studies that did not find a significant association between the viral load at baseline and the risk of death in advanced stage disease with low CD4 cell count. BMI within 3 months of HIV diagnosis was obtained from 1657 patients aged>=15 years, recruited in a seroprevalent clinical cohort in the Gambia since 1992 and followed up at least once. Baseline CD4+ counts and clinical assessment at time of diagnosis were done. The mortality hazard ratio (HR) of those with a baseline BMI <18 compared with those with a baseline BMI >=18 was 3.4 (95% CI, 3.0–3.9). The median survival time of those presenting with a BMI <16 was 0.8 years, in contrast to a median survival of 8.9 years for those with a baseline BMI >=22. Baseline BMI <18 remained a highly significant independent predictor of mortality after adjustment for age, sex, co-trimoxazole prophylaxis, tuberculosis, reported wasting at diagnosis, and baseline CD4+ cell count (adjusted HR = 2.5, 95% CI 2.0–3.0). Sensitivity and specificity of baseline BMI <18 was comparable to that of a CD4+ count <200 in predicting mortality within 6 months of diagnosis.

2.2.AFT Models for Survival Analysis in Studies with Time-Varying Treatments

As it is widely believed two useful models for survival analysis are the Cox proportional hazards model and the accelerated failure time (AFT) model. The widely used Cox model measures causal effect on the hazard (rate) ratio scale, whereas the less used AFT model, measures causal effect on the survival time ratio scale. Both the Cox model and semi parametric versions of the AFT model according to (Miguel *et al*, 2005) are models that leave the baseline hazard (or, equivalently, the baseline survival distribution) unspecified. However, even in the absence of unmeasured confounding and model misspecification, these standard models for survival analysis will provide estimates that fail to have a causal interpretation when: (i) there exists a measured time-dependent risk factor for survival that also predicts subsequent treatment, and (ii) past treatment history predicts subsequent risk factor level. Factors that meet condition (i) are known as time-dependent confounders. For example, when estimating the causal effect of highly active antiretroviral therapy (HAART) on the survival of individuals infected with the human immunodeficiency virus (HIV), condition (i) is met by the variable CD4 cell count because a low CD4 cell count is both a risk factor for survival and used by clinicians to decide whether to initiate HAART. Also, condition (ii) is met because prior HAART use increases CD4 cell count. Therefore, including the time-dependent confounder CD4 cell count in a standard Cox or AFT model may not appropriately adjust for confounding.

In contrast to standard Cox and AFT models, structural Cox and AFT models can be used to estimate causal effects when conditions (i) and (ii) hold. Marginal structural Cox model has previously been used to estimate the causal effect of HAART on the hazard of AIDS or death of HIV-infected individuals. The causal hazard ratio from the marginal structural model was 0.54 (95% confidence interval [CI]: 0.38, 0.78) when comparing continuous treatment with HAART versus no treatment with HAART. This hazard ratio was estimated by inverse probability weighting. The simultaneous presence of conditions (i) and (ii), and thus the problem of time-dependent confounding by factors affected by prior treatment, is ubiquitous in pharmaco epidemiology.

In their paper (Miguel *et al*., 2005) reviewed the differences between structural models and standard regression models for survival analysis. They describe a structural AFT model, and illustrated the application of this model for estimating the effect of HAART on AIDS-free survival in two prospective cohort studies of HIV-infected individuals. They found out that Nested structural AFT models and marginal structural Cox models can be used to consistently estimate the effect of a time-dependent exposure on survival in the presence of time-dependent confounders affected by prior exposure. On the other hand, standard models for survival analysis may yield biased estimates of causal effect because they adjust for time-dependent confounding by including the confounders as covariates in the model.

To avoid this problem, structural models adjust for time-dependent confounding by g-estimation or inverse probability weighting. Using a nested structural AFT model, they estimated that continuous HAART increased survival time by 2.5 fold in the MACS/WIHS. Their causal effect estimates from a structural AFT model are consistent with those from a marginal structural Cox model. It is reassuring that these two very different methods for estimating causal effects yield similar results, and that both arrive at the same qualitative conclusion as a previously conducted randomized trial (Hammer *et al*.,1997). In contrast, a standard associational Cox model did not find a substantially lower mortality rate among those treated compared with those untreated with HAART.

According to (Donglin and Lin, 2007), the accelerated failure time model provides a natural formulation of the effects of covariates on potentially censored response variable. The existing semi parametric estimators are computationally intractable and statistically inefficient. In their article they proposed an approximate non-parametric maximum likelihood method for the accelerated failure time model with possibly time-dependent covariates. They estimated the regression parameters by maximizing a kernel-smoothed profile likelihood function. The maximization was achieved through conventional gradient-based search algorithms. The resulting estimators were consistent and asymptotically normal. The limiting covariance matrix attained the semi-parametric efficiency bound and could be consistently estimated. They also provide a consistent estimator for the error distribution. Extensive simulation studies demonstrated that the asymptotic approximations were accurate in practical situations and the new estimators were considerably more efficient than the existing ones.

(Ponnuraja &Venkatesan, 2010) in their study, the proportional hazard model and its extension were used comprehensively to assess the effect of an intervention in the presence of covariates. They observed that in situation where the effect of the intervention is to accelerate the PH assumptions may not hold hence the AFT model is also appropriate. Their study was aimed to formulate a model that yields biological plausible and interpretable estimates of the effect of important covariates on survival time. It was revealed that the AFT model gives better prediction than the Cox PH model.

3. MATERIALS AND METHODS

3.1.Data source

The data used for this study was obtained from Jimma University Specialized Hospital HIV Outpatient Clinic, South West of Ethiopia. The survival data was extracted from the patient's chart which contains epidemiological, laboratory and clinical information of all HIV patients under ART follow-up including a detailed antiretroviral therapy history.

3.2. **Study population**

The study population consisted of all HIV positive patients who were aged 18 years or above placed under HAART any time in between 1st January 2010 to 30th June 2015 in Jimma University Specialized Hospital was included in the study. Therefore, among the total of 2048 HIV positive patients registered from 2010 to 2015, only HIV+ patients who were 18 years and older having at least one CD4 count measurement after first January 2010 and before 30th June 2015 were considered for the study.

Sample Size Based on the (Schoenfeld, 1981, 1983)

The total number of patients required by given by:

$$n = \frac{Events}{Pr\{Events\}} = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{(log\Delta)^2 \pi_1 \pi_2 \ \boldsymbol{\psi}}$$

Where α=significance level (0.05)

1-β=the power of the study (90%)

$Z_{1-\alpha/2}$= Z-value attributed to $\alpha/2$(1.96)

$Z_{1-\beta}$ =Z-value attributed to 1-β (1.28)

π_1 and π_2 are the proportions patients to be allocated to regimen type 1 and 2, for equal allocation: $\pi_1 = \pi_2$=0.5

Δ= the expected change in hazard ratio between the subjects on regimen type 1 and regimen type 2. In many clinical trials hazard ratio is chosen to be 1.5 since a 50% increase in survival is regarded as being clinically important and biologically feasible.

ψ= is the overall probability of death in the entire follow-up. (20%)

For the designed significance level and power to detect a hazard ratio of 1.5, this would require 88 deaths. But, how many patients should we plan to enroll to get 88 deaths? For 20% death rate for entire patients the sample size is 220 for n1 (being on regimen type 1) and 220 for n2, finally resulting the total sample size of 440.

3.3. **Variables of the study**

3.3.1. The Response Variable

The response or outcome variable for this study was the survival time (time-to-death) measured in months from the date of ART treatment's start until the date of the patient's death or censor.

The date of death was recorded for all HIV-infected patients, those died during the study period while on ART. Censor –Patients missing their follow-up visits for more than 3 months were counted as loss to follow-up and the date of the last registered follow-up visit was recorded as date of loss to follow-up. ART using HIV-infected patient, who were transferred to another Hospital before experience the event, were recorded as transferred-out cases and their dates of transferred-out were also recorded and the patients who are still on ART follow up to 30th June 2015.

3.3.2. Predictor Variables

Patients' baseline demographic and clinical characteristics were used as the predictor variables; in survival data analysis predictor variables are called covariates. These are explanatory variables which are assumed to influence the survival of HIV infected patients and are given below.

Table 3.1: Description of covariates together with their values/codes

Covariates	**Values/ Codes**
Age	in years
Sex	0=Female and 1= male
Marital status	0=Never married, 1= Married, 2= Others
Educational level	0 = Not Educated , 1 = Primary, 2 = Secondary, 3 = Tertiary
Residence	0=Rural and 1=urban
Weight	Kilogram
CD4 count	cells/mm^3
Pre-TB positive test	0=No and 1=Yes
WHO clinical stage	1 = Stage I, 2 = Stage II, 3 = Stage III, 4 = Stage IV
Functional status	0 = Working, 1 = Ambulatory, 2 = Bedridden
Type of initial regimen	0=AZT-3TC-NVP and 1= TDF-3TC-EFV
BMI	0 = Underweight , 1 =Normal, 2 = Overweight

There are twelve covariates; of these covariates three are continuous while nine of them are categorical covariates.

Notice that WHO Clinical disease Stage which is classified into four; I, II, III and IV; where Stage I indicates asymptomatic disease, Stage II indicates mild disease, Stage III indicates advanced disease and Stage IV indicates severe disease. Hence disease severity increases from Stage I to Stage IV. Functional Status of the patients is also categorical covariate with three categories: Working, Ambulatory and Bedridden. Working patients are those patients who can able to work day to day while ambulatory patients are those patients who can able to work some time but bedridden patients cannot able to work due to the disease and body mass index is also categorical (underweight, normal and overweight) .

3.4.Basic topics in survival analysis

3.4.1. Introduction to Survival analysis

The primary concept in survival analysis is ***survival time*** which is also called failure time. Survival time is a length of time that is measured from time origin to the time the event of interest occurred. To determine survival time precisely, there are three requirements: A time origin must be unambiguously defined, a scale for measuring the passage of time must be agreed upon and finally the definition of event (often called failure) must be entirely clear. The specific difficulties in survival analysis arise largely from the fact that only some individuals have experienced the event and other individuals have not had the event in the end of study and thus their actual survival times are unknown. This leads to the concept of ***censoring***. Censoring occurred when we have some information about individual survival time, but we do not know the survival time exactly. There are three types of censoring: right censoring, left censoring, and interval censoring.

Right censoring is said to occur if the event occurs after the observed survival time. Let *C* denote the censoring time, that is, the time beyond which the study subject cannot be observed. The observed survival time is also referred to as follow up time. It starts at time *0* and continues until the event *X* or a censoring time *C,* whichever comes first. The observed data are denoted by (T, δ), where T = min (X, C) is the follow-up time, and $\delta = \boldsymbol{I}_{x \leq c}$ is an indicator for status at the end of follow-up,

$$\delta = I_{x \leq c} = \begin{cases} 0, if\ X > C(\ observed\ censoring) \\ 1, if\ X \leq C(observed\ failure) \end{cases}$$

There are some reasons why right censoring may occur, for example, no event before the study ends, loss to follow-up during study period, or withdrawal from the study because of some reasons. The last reason may be caused by competing risks. The right censored survival time is then less than the actual survival time.

Censoring can also occur if we observe the presence of a condition but do not know where it began. In this case we call it left censoring, and the actual survival time is less than the observed censoring time.

If an individual is known to have experienced an event within an interval of time but the actual survival time is not known, we say we have interval censoring. The actual occurrence time of event is known within an interval of time. Right censoring is very common in survival time data, but left censoring is fairly rare.

An important assumption for methods presented in survival analysis studies for the analysis of censored survival data is that the individuals who are censored are at the same risk of subsequent failure as those who are still alive and uncensored. i.e. a subject whose survival time is censored at time *C* must be representative of all other individuals who have survived to that time. If this is the case, the censoring process is called non-informative. Statistically, if the censoring process is independent of the survival time i.e. $P(X \geq x; C \geq x) = P(X \geq x)\, P(C \geq x)$,

Then we will have non-informative censoring. Independence censoring is a special case of non - informative censoring. In this study, we assumed that the censoring is non-informative right censoring.

3.4.2. Analysis of Survival data

Let *T* be a continuous random variable denoting the survival time. The distribution of survival times is characterized by any of three functions: the survival function, the probability density function or the hazard function. The following definitions are based on textbook (Kleinbaum D, 2005).

The survival function is defined as the probability that an individual survives longer than t.

$$S(t) = P(T > t), t \geq 0$$

The range of $S(t)$ is 0 and 1 i.e. $0 \leq S(t) \leq 1$ .The graph of survival function is a step function and is called survival curve. At time zero, S(t) reaches to its maximum value 1 and if the last observed time is event time S(t) achieves the minimum value zero.

For an absolutely continuous variable T, The probability density function of T is

$$f(t) = \lim_{\Delta t \to 0} P(t \leq t < t + \Delta t)/\Delta t = \frac{d}{dt} F(t), t \geq 0$$

Where F(t)=1-S(t)

The probability density function is also known as the unconditional failure rate

The hazard function is a measure of the probability of failure during a very small interval, assuming that the individual has survived at the beginning of the interval. It is defined as

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

The function is also known as instantaneous failure rate, force of mortality, conditional rate and age specific failure rate. The hazard function is not a probability as it does not lie between 0 and 1. The function is commonly used for identifying the models. Such as exponential, Weibull or gamma curve that fits one's data.

There is a clearly defined relationship between *S(t)* and *h(t),* which is given by the formula

h(t)=f(t)/s(t) =d/dt logs(t)…… (3.1)

The cumulative hazard (cumulative risk) function is denoted by

$$\Lambda(t) = \int_0^t h(x)dx$$

Whereby Λ(t) is defined as the sum of the risks one may face going from duration 0 to t.

Then $s(t) = \exp\left(-\int_0^t h(x)\,dx\right) = \exp(-\Lambda(t)\,, t \geq 0$ …… (3.2)

$$\Rightarrow \Lambda(t) = -\log S(t)$$

The probability density function of T can be written

$f(t) = h(t) \exp\left(-\int_0^t h(x)dx\right)$ t≥0

These three functions give mathematically equivalent specification of the distributions of the survival time *T*. If one of them is known, the other two are determined. One of these functions can be chosen as the basis of statistical analysis according to the particular situations. The survival function is most useful for comparing the survival progress of two or more groups. The hazard function gives a more useful description of the risk of failure at any time point.

3.4.3. Non-parametric methods

In survival analysis, it is always a good idea to present numerical or graphical summaries of the survival times for the individuals. In general, survival data are conveniently summarized through estimates of the survival function and hazard function. The estimation of the survival distribution provides estimates of descriptive statistics such as the median survival time.

These methods are said to be non-parametric methods since they require no assumptions about the distribution of survival time. In order to compare the survival distribution of two or more groups, we will use log-rank tests (Kleinbaum D, 2005, Klein *et al*, 1997).

Preliminary analysis of the data using non-parametric methods provides insight into the shape of the survival function for each group and get an idea of whether or not the groups are proportional, i.e., if the estimated survival functions for two groups are approximately parallel (do not cross).

3.4.3.1.Kaplan-Meier estimator of the survival function

The Kaplan-Meier Estimator is a nonparametric estimator of the survival function (Kaplan & Meier, 1958) which is not based on the actual observed event and censoring times, but rather on the ordered in which events occur. This principle of nonparametric estimation of the survival function is to assign probability to and only to uncensored failure times. Suppose there are n observations,$t_1, \dots, t_n$, with corresponding censoring indicators,$\delta_1, \dots, \delta_n$. Let the number of distinct event times be r (r ≤ n), with the ordered event times given by$t_{(1)} < \cdots < t_{(r)}$ and corresponding number of events $d_{(1)}, \dots, d_{(r)}$. And also let $R(t_{(j)})$ denote the risk set at the event time $t_{(j)}$, i.e., the set of subjects that did not yet experience the event and were not yet censored before time $t_{(j)}$ and thus still at risk for the event at that time. Therefore, the Kaplan-Meier estimate of the survival function at time t is given by:

$$\hat{S}_{KM}(t) = \prod_{j=1}^{k} \left(\frac{R(t_{(j)}) - d_{(j)}}{R(t_{(j)})} \right), \text{for } t_{(j)} < t < t_{(j+1)}, k=1,2,\dots,r. \quad \dots\dots(3.3)$$

with $\hat{S}(t)=1$ for $t < t_{(1)}$ and the assuption that $t_{(r+1)} = \infty$

This estimator is a step function that changes values only at the time of each death. The size of the jump at a certain event time t(j) depends on the number of events observed at t(j), as well as on the pattern of the censored event times before t(j) . The variance of the Product-Limit estimator is estimated by Greenwood's formula (13), which is given by;

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j:t_{j \le t}} \frac{d_j}{R(t_{(j)})(R(t_{(j)}) - d_j)} \text{ ; } j=1,2,\dots,r \dots\dots(3.4)$$

3.4.3.2. Estimator of median and percentile of survival time

Since the distribution of survival time tends to be positively skewed, the median is preferred for a summary measure. The median survival time is the time, beyond which 50% of the individuals under study are expected to survive, i.e., the value of t(50) at $\hat{S}t(50) = 0.5$

The estimated median survival time is given by$\hat{t}(50) = min\{t|\hat{S}(t) < 0.5\}$, where ti is the observed survival time for the ith individual, i = 1, 2... n. In general, the estimate of the p^{th} percentile is $\hat{t}(P) = min\{t|\hat{S}(t) < 1 - \frac{p}{100}\}$

A confidence interval for the percentiles using delta method was discussed in the text- books (32, 44). The variance of the estimator of the pth percentile is

$$Var[\hat{S}(t(p))] = \left(\frac{d\hat{S}(t(p))}{dt(p)}\right)^2 var(t(p))$$

$$= \left(-f(t(p))\right)^2 var(t(p))$$

The standard error of $\hat{t}(p)$ is therefore given by

$$SE[\hat{t}(p)\ ] = \frac{1}{\hat{f}(\hat{t}(p))}\ SE[\hat{S}(t(p))]$$

The standard error of $\hat{S}(t(p))$ can be obtained using Greenwood's formula, given in equation (3.4). An estimate of the probability density function at the pth percentile b(p) is used by many software packages.

$$\hat{f}(\hat{t}(p)) = \frac{\hat{S}(\hat{u}(p)) - \hat{S}\left(\hat{l}(p)\right)}{\hat{l}(p) - \hat{u}(p)}$$

Where

$$\hat{u}(p) = max\{S(t_j) \geq 1 - \frac{p}{100} + \delta\}$$

$$\hat{t}(p) = min \leq \{S(t_j)1 - \frac{p}{100} - \delta\}$$

$t_j$is jth ordered death time, j=1,…,r; ε = 0.05 is typically used by a number of statistical packages.

Therefore, for median survival time, $\hat{u}(50)$ is the largest observed survival time from the K-M curve for which $\hat{S}(t) \geq 0.55$ and $\hat{t}(50)$ is the smallest observed survival time from the K-M curve for which $\hat{S}(t) \leq 0.45$. The 95% confidence interval for the p^{th} percentile $\hat{t}(p)$ has limits of $\hat{t}(p) \pm 1.96 * SE[\hat{t}(p)]$

3.4.3.3.Nonparametric Comparison of Survival Distributions

The estimated Kaplan- Meier survival curves shows the pattern of one survivorship function lying above another, this means the group defined by the upper estimated curve lived longer, or had a more favorable survival experience than the group defined by the lower estimated curve.
However, this is an exploratory method which may not be concluded upon, and hence, the need for formal statistical method to assess whether the two groups are different with respect to their survival time. In the comparison of two groups of survival data, there are a number of methods which can be used to quantify the extent of between-group differences. Two non-parametric approaches that are commonly used in practice are *log rank test* and *Wilcoxon test.*

There are a number of methods that can be used to test equality of the survival functions in different groups. One commonly used non-parametric test for comparison of two or more survival distributions is the log-rank test which we used in this study.
For two groups, let *t (1)<t (2)<…<t (k)* be the ordered death times across the two groups. Suppose that *d*j failures occur at *t* (j) and that rj subjects are at risk just prior to $t_{(j)}$ (*j* = 1, 2,…,*k*). Let d_{ij} and r_{ij} be the corresponding numbers in group *i* (*i* = 1, 2).The log-rank test will compare the observed number of deaths with the expected number of deaths for group *i.* Consider the null hypothesis S1 (*t*) = S2 (t), i.e. there is no difference between survival curves in two groups. Given rj and dj, the random variable d1j has the hyper geometric distribution.
Under the null hypothesis, the probability of death at t(j) does not depend on the group, i.e. the probability of death at t(j) is $\frac{d_j}{r_j}$. So that the expected number of deaths in group one is

$$E(d_{1j}) = e_{1j} = r_{1j} d_j {r_j}^{-1}$$

The test statistic is given by the difference between the total observed and expected number of deaths in group one

$U_L = \sum_{j-1}^{r} (d_{1j} - e_{1j})$ …… (3.5)

Since d_{1j} has the hyper geometric distribution, the variance of d_{1j} is given by

$$V(d_{1j}) = \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j{}^2 (r_j - 1)}$$

$$V(U_L) = \sum_{j=1}^{r} V_{ij} = V_L$$

Under the null hypothesis, statistic (3.5) has an approximate normal distribution with zero mean and unit variance i.e. $Z = \frac{U_L}{\sqrt{V_L}} \sim N(0,1)$ Assuming that the contingency tables at different death times are independent, the log rank test is given by. $Z^2 = W_L = \frac{{U_L}^2}{V_L} \sim \chi^2{}_1$ (Under the null hypothesis), Where W_L implies Wald test.

A large value of W_L would lead to the conclusion that two groups do not have equal survival function. The test is more appropriate, powerful and reliable as compared to the other tests in a situation where two or more survival curves do not cross i.e. whose hazard functions are proportional. (Peto and Peto, 1972) and (Prentice, 1978), they suggested using a weight function that depends more explicitly on the observed survival experience of the combined sample. The weight is then given as $W_j = S(t_{j-1}) \times \frac{r_j}{r_{j+1}}$

3.4.4. Regression Models for Survival Data

In most medical studies which give rise to survival data, supplementary information is collected on each individual so that the relationship between the survival experience of individuals and various explanatory variables may be investigated. A variety of models and methods have been developed for doing this sort of survival analysis – parametric and semi-parametric.

3.4.4.1.Cox Regression Model

Semi parametric models are models that parametrically specify the functional relationship between the lifetime of an individual and his characteristics (demographic, socio-economic, etc.) but leave the actual distribution of lifetimes arbitrary.

The most popular of the semi-parametric models is the *proportional hazards model,* which has the property that ratio of the hazards of two individuals at time *t* can depend on the values of their explanatory variables, say x1, x2…but does not depend on time *t*.

One very popular model in survival data is the Cox proportional hazards model, which is proposed by (Cox, 1972).

The Cox Proportional Hazards model is given by:

For fixed covariates

$$h(t|X_i) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t)\exp(\beta' X)\text{,i=1, 2,…, p} \quad \ldots\ldots\ldots\ldots(3.6)$$

Where

- $h_0(t)$ is the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero.
- X = (x1,x2, … , xp)' is the values of the vector of baseline explanatory variables for a particular individual.
- β'= (β1, β2, …, βp) is a vector of regression coefficients.

The corresponding survival functions are related as follows:

$$S(t|x) = P(T > t/x) = S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i X_i)}$$

If all of the covariates are zero the model (3.6) above becomes $h_i(t) = h_0(t)$ because of this we call the term the baseline hazard function. In this model, no distributional assumption is made for the survival time; the only assumption is that the hazards ratio $\psi = \frac{h_i(t)}{h_j(t)}$ does not change over time (i.e., proportional hazards). This model is semi-parametric in the sense that a parametric form is assumed only for the covariate effect, with the baseline hazard function treated non-parametrically.

The measure of effect is called hazard ratio. The hazard ratio of two individuals with different covariates x and x^*will be given by:

$$\widehat{HR} = \frac{h_0(t)exp\left(\hat{\beta}' x\right)}{h_0(t)exp\left(\hat{\beta}' x^*\right)} = exp\left\{\sum \hat{\beta}'(x - x^*)\right\}$$

This hazard ratio is time-independent, which is why this is called the proportional hazards model. The parameter of the Cox proportional hazard model refers to the hazard ratio of one group in comparison to the other groups for categorical covariates and change in hazard ratio with a unit change of the covariate for the continuous variables when other covariates are fixed.

The change in hazard ratio for the continuous covariate is given by:

$$\frac{h(t|x_k+1)}{h(x_k)} = \frac{\exp(\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k(x_k+1)+\cdots+\beta_p x_p)}{\exp(\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\cdots+\beta_p x_p)} = \exp(\beta_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

which represents change (equivalently, $\exp(\beta_k)$*100% percentage change) hazard function with unit change in covariate provided that other covariates remains fixed.

For a categorical covariate **X** with l levels, the model contains (l−1) dummy variables defined as **Z**i = 1 if **x**= i , and 0 otherwise for i = 1, 2,..., l −1. Let $\beta_1, \beta_2 \ldots \beta_{l-1}$denote the coefficients in front of the appropriate dummy variables. Then the ratio of the hazard of two subjects, one with **X** at level j and the other with **X** at level k (j, k = 1,2,..., l −1), provided the values of all other covariates for these subjects are the same, the hazard ratio between these two categories is given by:

$$\frac{h(t|z_j)}{h(t|z_k)} = \frac{\exp(\beta_j)}{\exp(\beta_k)} = \exp(\beta_j - \beta_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

The quantity $\exp(\beta_j - \beta_k)$ 100% signifies the ratio (expressed as a percentage) of hazard functions for subjects at level j and at level k of the covariate (j, k = 1,2,..., l -1), provided the other covariates have equal values.

3.4.4.1.1. Partial likelihood estimate for Cox PH model

Fitting the Cox proportional hazards model, we estimated $h_0(t)$ and β. One approach was to attempt to maximize the likelihood function for the observed data simultaneously with respect to $h_0(t)$ and β. A more popular approach is proposed by (Cox & Oakes,1984) in which a partial likelihood function that does not depend on $h_0(t)$ is obtained for β. Partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameters ($h_0(t)$ in the Cox PH model. In this part, we constructed the partial likelihood function based on the proportional hazards model.

Let $t_1, t_2, .., t_n$ be the observed survival time for n individuals. Let the ordered death time of r individuals be $t_{(1)} < t_{(2)} <, \dots, < t_{(n)}$ and let$R(t_{(j)})$ be the risk set just before $t_{(j)}$and r_j for its size. So that $R(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The conditional probability that the *i*th individual dies at $t_{(j)}$ given that one individual from the risk set on $R(t_{(j)})$ dies at $t_{(j)}$ is

P(individual *i* dies at $t_{(j)}$ | one death from the risk set $R(t_{(j)})$ at $t_{(j)}$).

$$= \frac{P\left(idividual\ i\ dies\ at\ t_{(j)}\right)}{P\left(one\ death\ at\ t_{(j)}\right)}$$

$$= \frac{P\left(individual\ i\ dies\ at\ t_{(j)}\right)}{\sum_{k \in R\left(t_{(j)}\right)} P\left(individual\ k\ dies\ at\ t_{(j)}\right)}$$

$$= \frac{\frac{P\left\{individual\ i\ dies\ at\ (t_{(j)}, t_{(j)}+\Delta t)\right\}}{\Delta t}}{\sum_{k \in R\left(t_{(j)}\right)} \frac{P\left\{individual\ k\ dies\ at\left(\ t_{(j)}+\Delta t\right)\right\}}{\Delta t}}$$

$$= \frac{\lim_{\Delta t \to 0} \frac{P\left\{individual\ i\ dies\ at\ (t_{(j)}, t_{(j)}+\Delta t)\right\}}{\Delta t}}{\lim_{\Delta t \to 0} \sum_{k \in R\left(t_{(j)}\right)} \frac{P\left\{individual\ k\ dies\ at\left(\ t_{(j)}+\Delta t\right)\right\}}{\Delta t}}$$

$$= \frac{h_i\left(t_{(j)}\right)}{\sum_{k \in R\left(t_{(j)}\right)} h_k\left(t_{(j)}\right)}$$

$$= \frac{h\left(t_{(j)}\right) \exp\left(\beta' x_i\left(t_{(j)}\right)\right)}{\sum_{k \in R\left(t_{(j)}\right)} h_0\left(t_{(j)}\right) \exp\left(\beta' x_k\left(t_{(j)}\right)\right)}$$

$$= \frac{\exp\left(\beta' x_i\left(t_{(j)}\right)\right)}{\sum_{k \in R\left(t_{(j)}\right)} \exp\left(\beta' x_k\left(t_{(j)}\right)\right)}$$

Then the partial likelihood function for the Cox PH model is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\left(\beta' x_i\left(t_{(j)}\right)\right)}{\sum_{k\in R\left(t_{(j)}\right)} \exp(\beta' x_k(t_{(j)})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

in which xi(t(j)) is the vector of covariate values for individual i who dies at t(j): The general method of partial likelihood was discussed by (Cox ,1973).Note that this likelihood function is only for the uncensored individuals. Let $t_1, t_2, \ldots, t_n$ be the observed survival time for n individuals and δ_i be the event indicator, which is zero if the i^{th} survival time is censored, and unity otherwise. The likelihood function in equation (3.9) can be expressed by

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\beta' x_i\left(t_{(j)}\right)\right)}{\sum_{k\in R\left(t_{(j)}\right)} \exp(\beta' x_k(t_{(j)})}\right]^{\delta_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

The partial likelihood is valid when there are no ties in the dataset. That means there is no two subjects who have the same event time.

The partial likelihood given by equation (3.10), although it describes only part of the data, could be regarded as a likelihood function allowing the estimation of β with standard procedures. In general, large sample properties like normality and consistency of maximum likelihood estimators of β based on partial likelihood have been shown to be the same as those of any estimator from complete likelihood (Kleinbaum D, 2005, Klein, J. P. &Moeschberger, 1997).

3.4.4.1.2. Checking of Proportional hazard assumption

The main assumption of the Cox proportional hazards model is proportional hazards. PHs means that the hazard function of one individual is proportional to the hazard function of the other individual, i.e., the hazard ratio is constant over time. There are several methods for verifying that a model satisfies the assumption of proportionality.

a. Graphical methods

We can obtain Cox PH survival function by the relationship between hazard function and survival function

$$S(t|x_1) = S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i x_i)}$$

Where X = (x1,x2, …xp)' is the values of the vector of explanatory variables for a particular individual. When taking the logarithm twice, we get

$$\log[-logS(t|x_1)] = \exp\left(\sum_{i=1}^{p}\beta_i x_i\right) + \log[-logS_0(t)]$$

Then the difference in log-log curves corresponding to two different individuals with variables x1= (x11; x12 ,…, x1p) and x2 = (x21 ,x22 ,…, x2p) is given by

$$\log[-logS(t|x_1)] - \log[-logS(t|x_2)] = \sum_{i=1}^{p}\beta_i\,(x_{1i} - x_{2i})$$

does not depend on *t*. This relationship is very helpful to help us identify situations where we may or may not have proportional hazards. By plotting estimated log (-log (survival)) versus survival time for two groups we would see parallel curves if the hazards are proportional. This method does not work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Furthermore, the curves are sparse when there are few time points and it may be difficult to tell how close to parallel is close enough.

b. Adding time-dependent covariates in the Cox model

Time-dependent covariates are created by creation of interactions of the predictors and a function of survival time and including them in the model. For example, if the predictor of interest is Xj , then we create a time-dependent covariate Xj(*t*), Xj(*t*) = Xj x *g*(*t*) , where *g* (*t*) is a function of time, e.g., *t*, log *t* or Heaviside function of *t*. The model assessing PH assumption for X*j* adjusted for other covariates is

$$h\big(t\big|x(t)\big) = h(t)exp\big[\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_j x_j + \cdots + \beta_p x_p + \delta x_j * g(t)\big]$$

Where x (t) = (x1; x2 ,…, xp ,xj (t))' is the values of the vector of explanatory variables for a particular individual. The null hypothesis to check proportionality is that δ = 0. The test statistic can be carried out using either a Wald test or a likelihood ratio test.

In the Wald test, the test statistic is $W = \left(\frac{\delta_i}{se(\delta_i)}\right)^2$.

The likelihood ratio test will calculates the likelihood under null hypothesis, L0 and the likelihood under the alternative hypothesis, La. The LR statistic is then,

$LR = -2\log\left({L_a}/{L_0}\right) = -2(l_a - l_0)$, where l0 and la are log likelihood under two hypothesis respectively

Both statistics have a chi-square distribution with one degree of freedom under the null hypothesis. If the time-dependent covariate is significant, i.e. the null hypothesis is rejected, then the predictor is not proportional. In the same way, we assessed the PH assumption for several predictors simultaneously.

c. Tests Based on the Schoenfeld Residuals

The other statistical test of the proportional hazards assumption is based on the Schoenfeld residual (Schoenfeld, D, 1982) .The Schoenfeld residuals are defined for each subject who is observed to fail. If the PH assumption holds for a particular covariate then the Schoenfeld residual for that covariate will not be related to survival time. So this test will be accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual survival times. The null hypothesis is that the correlation between the Schoenfeld residuals and the ranked survival time is zero. Rejection of null hypothesis concludes that PH assumption is violated.

3.4.4.1.3. Model diagnostics: Cox PH model

After a model has been fitted, the adequacy of the fitted model needs to be assessed. The model checking procedures below are based on residuals. In linear regression methods, residuals are defined as the difference between the observed and predicted values of the dependent variable. However, when censored observations are present and partial likelihood function is used in the Cox PH model, the usual concept of residual is not applicable. A number of residuals have been proposed for use in connection with the Cox PH model. We will use three major residuals in the Cox model: the Cox-Snell residual, the deviance residual, and the Schoenfeld residual.

a. Cox-Snell residuals

The Cox-Snell residual is given by Cox and Snell (Cox, D. R. & Snell, E. J, 1968) The Cox-Snell residual for the i^{th} individual with observed survival time t_i is defined as

$$r_{ci} = \widehat{H}_0(t_i)\exp\left(\sum_{i=1}^{p} x_{ik}\hat{\beta}_k\right) = \widehat{H}_i(t) = -\log\left(\hat{S}(t_i)\right)$$

Where , $\widehat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i which was derived by (Kalbfleisch, J. D., & Prentice, R., 1973) .This residual is motivated by the following result:

Let *T* have continuous survival distribution *S(t)* with the cumulative hazard *H(t)* = -log(*S(t)*). Thus, *S(t)* = exp (-H(t)). Let *Y = H(T)* be the transformation of *T* based on the cumulative hazard function. Then the survival function for Y is

$$\begin{aligned} S_Y(y) = P(Y < y) &= P(H(T) > y) \\ &= P\left({H_T}^{-1}(y)\right) = S_T\left({H_T}^{-1}(y)\right) \\ &= \exp\left(-H_T\left({H_T}^{-1}(y)\right)\right) = \exp(-y) \end{aligned}$$

Thus, regardless of the distribution of *T*, the new variable *Y = H(T)* has an exponential distribution with unit mean. If the model was well fitted, the value would $\hat{S}(t_i)$ have similar properties to those of Si (ti) . So *r*ci = -log$\hat{S}(t_i)$ will have a unit exponential distribution with *f*R*(r)* = exp (-r). Let SR*(r)* denote the survival function of Cox-Snell residual *rci* . Then

$$S_R(r) = \int_r^{\infty} f_R(x)dx = \int_r^{\infty} \exp(-x)\,dx = \exp(-r)\,,$$

$$H_R(r) = -logS_R(r) = -log(\exp(-r)) = r$$

Therefore, we will use a plot of *H(rci)* versus *rci* to check the fit of the model. This gives a straight line with unit slope and zero intercept if the fitted model is correct. The Cox-Snell residuals will not be symmetrically distributed about zero and cannot be negative.

b. Martingale Residuals

As far as one – event models are concerned, martingale residual for the i^{th} subject at the moment t is defined as follows $M_i(t) = \delta_i(t) - H(t|x_i)$ and is interpreted as a difference between (observed) and expected (resulting from the model) number of event occurrence till the moment t. It is calculated for the given subject, at the given time point t. With, $\delta_i(t)$ being the dummy variable to indicate that if $\delta_i = 1$ for uncensored observation and $\delta_i = 0$ for censored observation. Usually martingale residuals are subject - specific and are calculated as of at the end of the study. As residuals of this type do not have symmetric distribution, they can be transformed into deviance residuals that are supposed to have a symmetric distribution with the mean equal to zero, assuming proper specification of the model. Martingale residuals are useful while examining assumption of linear effect of covariates on logarithm of hazard.

c. Deviance residuals

The deviance residuals help in identifying poorly fitted subjects, which is defined in (16) as

$$Di = sign(\widehat{M})\sqrt{-2(\widehat{M}_i + d_i) + log(d_i - \widehat{M}_i)}.$$

It is known that the deviance residuals are symmetrically distributed about zero when the fitted model is adequate, and individuals with large positive or negative deviance residuals are poorly predicted by the model.

The function sign (.) is the sign function which takes the value 1 if M*i* is positive and -1 if M*i* is negative.

The martingale residuals take values between negative infinity and unity. They have a skewed distribution with mean zero (Hosmer DW, Lemeshow S, 1999). The deviance residuals are, however, a normalized transform of the martingale residuals. They also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Deviance residual can also be used like residuals from linear regression.

The plot of the deviance residuals against the covariates can be obtained. Any unusual patterns may suggest features of the data that have not been adequately fitted for the model.

Very large or very small values suggest that the observation may be an outlier in need of special attention. In a fitted Cox PH model, the hazard of experiencing the event for the i^{th} individual at any time depends on the value of $\exp(\beta' x_i)$ whichis called the risk score. A plot of the deviance residuals versus the risk score is a helpful diagnostic to assess a given individual on the model. Potential outliers will have deviance residuals whose absolute values are very large. This plot will give the information about the characteristic of observations that are not well fitted by the model.

d. Schoenfeld residuals

All the above three residuals are residuals for each individual. We will describe covariate-wise residuals: Schoenfeld residuals (Kleinbaum D, 2005, Klein, J. P., & Moeschberger, 1997). The Schoenfeld residuals were originally called partial residuals because the Schoenfeld residuals for *i*th individual on the *j*th explanatory variable Xj is an estimate of the *i*th component of the first derivative of the logarithm of the partial likelihood function with respect to βj. Hence, the Schoenfeld residuals are interpreted as 'input' of a given subject in the derivative of logarithm of partial likelihood function with respect to the given covariate (or: a difference between actual value of the given covariate for the given subject and expected value of particular covariate in a risk set)

From equation (3.10), this logarithm of the partial likelihood function is given by $\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{p} \delta_i \{x_{ij} - \alpha_{ij}\}$; Where xij is the value of the jth explanatory variable j = 1; 2,… ,p for the i^{th} individual and $\alpha_{ij} = \frac{\sum_{i \in R(t_i)} x_{ij} \exp(\beta' X_{ij})}{\sum_{i \in R(t_i)} \exp(\beta' X_{ij})}$

The Schoenfeld residual for i^{th} individual on zj is given by $r_{pij} = \delta_i \{x_{ij} - \alpha_{ij}\}$. The Schoenfeld residuals sum to zero.

e. Diagnostics for influential observations

Observations that have an undue effect on model-based inference are said to be influential.

In the assessment of model adequacy, it is important to determine whether there are any influential observations. The most direct measure of influence is $\hat{\beta}_j - \hat{\beta}_{j(i)}$where $\hat{\beta}_j$ is the j^{th} parameter, j = 1, 2, ..., p in a fitted Cox PH model and $\hat{\beta}_{j(i)}$ is obtained by fitting the model after omitting observation i. In this way, we have to fit the n + 1 Cox models, one with the complete data and n with each observation eliminated. This procedure involves a significant amount of computation if the sample size is large. We would like to use an alternative approximate value that does not involve an iterative refitting of the model. To check the influence of observations on a parameter estimate, Cain and Lange (Cain, K. C., & Lange, N. T,1984) showed that an approximation to $\hat{\beta}_j - \hat{\beta}_{j(i)}$ is the j^{th} component of the vector

$$r_{si}'Var(\hat{\beta})$$

Where, r_{si} is the p ×1 vector of score residuals for the i^{th} observation (Collett,2003), which are modifications of Schoenfeld residuals and are defined for all the observations, and is the $Var(\hat{\beta})$ variance-covariance matrix of the vector of parameter estimates in the fitted Cox PH model. The j^{th} element of this vector is called delta-beta statistic for the j^{th} explanatory variable, i.e., $\Delta_i\hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$ which tells us how much each coefficient will change by removal of a single observation. Therefore, we can check whether there are influential observations for any particular explanatory variable.

3.4.4.3.4. **Strategies for analyzing of non-proportional data**

If the statistic tests or other diagnostic techniques will give strong evidence of non-proportionality for one or more covariates To deal with this we use two popular methods: stratified Cox model and Cox regression model with time-dependent variables which are particularly simple and can be done using available software in my analysis. Another alternative to consider is to use a different model. A parametric model such as an AFT model, which we will describe in next section, is more appropriate.

a. Stratified Cox model

One method that we can use is the stratified Cox model, which stratifies on the predictors not satisfying the PH assumption. The data are stratified into subgroups and the model is applied for each stratum. The model is given by

$$h_{ig}(t) = h_{0g}(t)\exp(\beta' X_{ig})$$

Where g represents the stratum

The hazards are non-proportional because the baseline hazards may be different between strata. The coefficients β are assumed to be the same for each stratum *g*. The partial likelihood function is simply the product of the partial likelihoods in each stratum. A drawback of this approach is that we cannot identify the effect of this stratified predictor.

b. Cox regression model with time-dependent variable

Until now we have assumed that the values of all covariates did not change over the period of observation. However, the values of covariates may change over time *t*. Such a covariate is called a time-dependent covariate. The second method that we will consider is to model non proportionality by time-dependent covariates. The violations of PH assumption are equivalent to interactions between covariates and time. That is, the PH model assumes that the effect of each covariate is the same at all points in time. If the effect of a variable varies with time, the PH assumption is violated for that variable. To model a time-dependent effect, we will create a time dependent covariate *X (t), then β X (t) = βX* x *g (t). g (t) is* a function of *t* such as *t*; log *t* or Heaviside functions, etc. The choice of time-dependent covariates may be based on theoretical considerations and strong clinical evidence. The Cox regression with both time independent predictors *Xi* and time-dependent covariates *Xj(t)* can be written

$$h\big(t|X(t)\big) = h_0(t)\exp\Big(\sum_{i=1}^{p1}\beta_i x_i + \sum_{j=1}^{p2}\alpha_j x_j\,(t)\Big)$$

The hazard ratio at time t for the two individuals with different covariates Z and Z* is given by

$$exp\hat{\beta}_i(x_i{}^* - x_i) + \sum_{j=1}^{p2}\hat{\alpha}_j\left(x_j{}^*(t) - x_j(t)\right)$$

Note that, in this hazard ratio formula, the coefficient $\hat{\alpha}_j$ is not time-dependent. $\hat{\alpha}_j$, represents overall effect of Xj*(t)* considering all times at which this variable was measured in this study. But the hazard ratio depends on time *t*. This means that the hazards of event at time *t* is no longer proportional, and the model is no longer a PH model.

In addition to considering time-dependent variable for analyzing a time-independent variable not satisfying the PH assumption, there are variables that are inherently defined as time-dependent variables. One of the earliest applications of the use of time-dependent covariates is in the report by (Crowder, J., & Hu, M.1977), on the Stanford Heart Transplant study. Time-dependent variables are usually classified to be internal or external. An internal time-dependent variable is one that the change of covariate over time is related to the characteristics or behavior of the individual.

For example, blood pressure, disease complications, etc. The external time-dependent variable is one whose value at a particular time does not require subjects to be under direct observations,

i.e., values changes because of external characteristics to the individuals. For example, level of air pollution.

3.4.5. Parametric model

Cox regression models are used for survival data which do not require any specific distributional assumptions about the shape of the survival function (time) but they make an assumption of the proportional hazards. However, when the proportional hazards assumption is questionable, these models will not be suitable. When these parametric models provide a good fit to data, they tend to give more precise estimates of the quantities of interest because these estimates are based on fewer parameters. Of course, if the parametric model is chosen incorrectly, it may lead to consistent estimators of the wrong quantity. In this section, we will introduce parametric model, in which specific probability distribution is assumed for the survival times. Most often used parametric models can be represented as; Parametric PH models; AFT models or log linear models.

3.4.5.1.Parametric PH models

The parametric proportional hazards model is the parametric versions of the Cox proportional hazards model. It is given with the similar form to the Cox PH models. The hazard function at time t for the particular patient with a set of p covariates (x1, x2... xp) is given as follows:

$$h(t|x) = h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t)\exp(\boldsymbol{\beta}' x)$$

The key difference between the two kinds of models is that the baseline hazard function is assumed to follow a specific distribution when a fully parametric PH model is fitted to the data, whereas the Cox model has no such constraint. The coefficients are estimated by partial likelihood in Cox model but maximum likelihood in parametric PH model. Other than this, the two types of models are equivalent. Hazard ratios have the same interpretation and proportionality of hazards is still assumed. A number of different parametric PH models may be derived by choosing different hazard functions. The commonly applied models are exponential, Weibull models.

a. Exponential PH model

The exponential PH model is a special case of the Weibull model when γ = 1. The hazard function under this model is to assume that it is constant over time. The survival and hazard function are written as

$$S(t) = \exp(-\lambda t)\ ; h(t) = \lambda$$

Under the exponential PH model, the hazard function of a particular patient is given by

$$h(t|x) = \lambda exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = \lambda\exp(\boldsymbol{\beta}' x)$$

b. Weibull PH model

Suppose that survival times are assumed to have a Weibull distribution with scale parameter λ and shape parameter γ, so the survival and hazard function of a W (λ, γ) distribution are given by:

$S(t) = \exp(-\lambda t^{\gamma})$;$h(t) = \lambda\gamma(t)^{\gamma-1}$ with $\lambda, \gamma > 0$……(3.11)

The hazard rate increases when, $\gamma > 1$ and decreases when $\gamma < 1$ as time goes on. When = 1, the hazard rate remains constant, which is the special exponential case.

Under the Weibull PH model, the hazard function of a particular individual with covariates (x1; x2, . . ., xp) is given by

$$h(t|x) = \lambda\gamma(t)^{\gamma-1}exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = \lambda\gamma(t)^{\gamma-1}\exp(\boldsymbol{\beta}'x) \quad \ldots\ldots\ldots\ldots\ldots (3.12)$$

We can see that the survival time of this patient has the Weibull distribution with scale parameter λ exp(β'x) and shape parameter γ: Therefore the Weibull family with fixed γ possesses PH property. This shows that the effects of the explanatory variables in the model alter the scale parameter of the distribution, while the shape parameter remains constant. From equation (1.2), the corresponding survival function is given by

$S(t|x) = \exp(-exp(\beta'x)\lambda t^{\gamma})$ …… (3.13)

After a transformation of the survival function for a Weibull distribution, we can obtain

$$log\{-logS(t)\} = log\lambda + \gamma logt$$

The log{-logS(t)} versus log(t) should give approximately a straight line if the Weibull distribution assumption is reasonable. The intercept and slope of the line will be rough estimate of logλ and γ respectively. If the two lines for two groups in this plot are essentially parallel, this means that the proportional hazards model is valid. Furthermore, if the straight line has a slope nearly one, the simpler exponential distribution is reasonable.

In the other way, for a exponential distribution, there is log S(t) = -λt. Thus we can consider the graph of log S(t) versus t. This should be a line that goes through the origin if exponential distribution is appropriate.

Another approach to assess the suitability of a parametric model is to estimate the hazard function using the non-parametric method. If the hazard function were reasonably constant over time, this would indicate that the exponential distribution might be appropriate. If the hazard function increased or decreased monotonically with increasing survival time, a Weibull distribution or Gompertz distribution might be considered.

3.4.5.2.Accelerated failure time model

Although parametric models are very applicable to analyze survival data, there are relatively few probability distributions for the survival time that can be used with these models. In these situations, the accelerated failure time model (AFT) is an alternative to the PH model for the analysis of survival time data. Under AFT models we measured the direct effect of the explanatory variables on the survival time instead of hazard. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean survival time Similar to the PH model, the AFT model describes the relationship between survival probabilities and a set of covariates.

The accelerated failure-time model states that the survival function of an individual with covariate X at time t is the same as the survival function of an individual with a baseline survival function at a time t*exp($\alpha'\mathbf{X}$) here $\alpha' = (\alpha_1\alpha_2, \dots, \alpha_p)$ is a vector of regression coefficients. In other words, the accelerated failure-time model is defined by the relationship

$$S(t|X) = S_0[t * \exp(\alpha' X)] \quad ; for\ all\ x.$$

3.4.5.3.Linear Model Representation in Log Time

The corresponding log-linear form of the AFT model with respect to time is given analogous to the classical linear regression approach. In this approach, the natural logarithm of the survival time Y =log (T) is modeled. This is the natural transformation made in linear models to convert positive variables to observations on the entire real line. A linear model is assumed for Y, namely,

$$Y = \log T = \boldsymbol{\mu} + \alpha' x + \boldsymbol{\sigma}\varepsilon$$

Where

$\alpha' = (\alpha_1, \alpha_2, \dots, \alpha_p)$ is a vector of regression coefficients

μ is intercept

σ is scale parameter and

ε is the error distribution assumed to have a particular distribution.

For each distribution of ε there is a corresponding distribution for T. The members of the AFT model class include the exponential AFT model, Weibull AFT model, log- logistic AFT model, log-normal AFT model, and gamma AFT model. The AFT models are named for the distribution of *T* rather than the distribution of ε or log *T*.

Table 3.2: Summary of parametric AFT models

Distribution of ε	Distribution of T
Extreme value(1 parameters)	Exponential
Extreme value (2 parameters)	Weibull
Logistic	Log-logistic
Normal	Log-normal
Log-gamma	Gamma

This model can be related to the accelerated failure-time model representation (15) as in.
The survival function of *Ti* can be expressed by the survival function of ε_i.

$$
\begin{aligned}
S_i(t) &= P(T_i \geq t) \\
&= P(logT_I) \geq logt \\
&= P(Y_I \geq logt) \\
&= P(\mu + \alpha' x + \sigma\varepsilon \geq logt) \\
&= P\left(\varepsilon_i \geq \frac{logy-\mu-\alpha\prime x}{\sigma}\right) \\
&= S\varepsilon_I\left(\frac{logt-(\mu+\alpha\prime x)}{\sigma}\right) \quad \text{.................................. (3.14)}
\end{aligned}
$$

The effect size for the AFT model is the time ratio. The time ratio comparing two levels of covariate *xi (xi = 1 vs. xi = 0)*, after controlling all the other covariates is $\exp(\alpha_i)$, which is interpreted as the estimated ratio of the expected survival times for two groups.

A TR>1 for the covariate will implies that this covariate prolongs the time to event, while a TR <1 indicates that an earlier event is more likely. Therefore, the AFT models will be interpreted in terms of the speed of progression of a disease.

The effect of the covariates in an accelerated failure time model is to change the scale, and not the location of a baseline distribution of survival times.

3.4.5.4.Distributions used in AFT models

To be used in an AFT model, a distribution must have a parameterization that includes a scale parameter. The logarithm of the scale parameter is then modeled as a linear function of the covariates

a. The Weibull AFT Distribution:

The Weibull distribution (including the exponential distribution as a special case) as shown above can also be parameterized as an AFT model, and they are the only family of distributions to have this property. The results of fitting a Weibull model can therefore be interpreted in either framework (32, 44). Then the Weibull distribution is very flexible model for time-to-event data. It has a hazard rate which is monotone increasing, decreasing, or constant. It is stated above that If T has a Weibull distribution, then ε has an extreme value distribution (Gumbel distribution). The survival function of Gumbel distribution is given by$S_\varepsilon\varepsilon = \exp\big(-exp(\varepsilon)\big)$

From equation (3.14), the AFT representation of the survival function of the Weibull model is given by

$$S_{\varepsilon i}(t) = exp\left(-exp\left(\frac{logt-[\mu+\alpha' x]}{\sigma}\right)\right) \Leftrightarrow exp\left(-exp\left(\frac{-[\mu+\alpha' x]}{\sigma}\right)t^{\frac{1}{\sigma}}\right) \quad \text{(3.15)}$$

From equation (3.11), the PH representation of the survival function of the Weibull model is given by:

$$S_i(t|x) = exp\{-exp(\beta'^{x_i})\lambda t^{\gamma}\} \quad \text{(3.16)}$$

Comparing the above two formulas (3.15) and (3.16), we can easily see that the parameter λ, γ, βj in the PH model can be expressed by the parameters μ, α, σ in the AFT model:

$$\lambda = exp\left(\frac{-\mu}{\sigma}\right) \; ; \; \gamma = {}^{1}/_{\sigma} \; ; \; \beta_j = {}^{-\alpha_j}/_{\sigma} \quad \text{(3.17)}$$

Using equation (3.14), the AFT representation of hazard function of the Weibull model is given By

$h_i(t) = \frac{1}{\sigma} t^{\frac{1}{\sigma}-1} exp\left(\frac{-\mu-\alpha_i x_i}{\sigma}\right)$ …… (3.18)

Suppose the pth percentile of the survival distribution for the ith individual is ti (p) which is the value such that

$$S_i\big(t_i(p)\big) = \frac{100-p}{100}$$

From equation (4.5), we can easily get

$$t_i(t) = exp\left[\sigma log\left\{-log\left(\frac{100-p}{p}\right)\right\} + \mu + \alpha' x_i\right]$$

The median survival time is

$t_i(50) = exp[\sigma \log(\log 2) + \mu + \alpha'^{x_i}]$ …… (3.19)

b. Log-logistic AFT model

One limitation of the Weibull hazard function is that it is a monotonic function of time.

However, the hazard function can change direction in some situations. The log-logistic distribution provides the most commonly used AFT model. Unlike the Weibull distribution, it can exhibit a non-monotonic hazard function which increases at early times and decreases at later times. It is similar in shape to the log-normal distribution but its cumulative distribution function has a simple closed form, which becomes important computationally when fitting data with censoring (Hosmer DW, & Lemeshow S., 1997). The log-logistic survival and hazard function for log linear model with no covariates (logT=μ+δε) are.

$S(t) = \frac{1}{1+e^{\theta} t^{\gamma}}$ …… (3.20) $h(t) = \frac{e^{\theta}\gamma t^{\gamma-1}}{1+e^{\theta} t^{\gamma}}$ …… (3.21)

Where $\theta = {}^{-\mu}/_{\sigma}$ and $\gamma = {}^{1}/_{\sigma}$ are unknown parameters and γ > 0; When γ ≤1, the hazard rate decreases monotonically and when γ > 1, it increases from zero to a maximum and then decreases to zero.

Suppose that the survival times have a log-logistic distribution with parameter θ and γ, and then from equation (3.4), under the AFT model, the hazard function for the ith individual is

$$h_i(t|x) = h_0\big(texp(-\alpha' x)\big)\exp(-\alpha' x)$$

$= \frac{\gamma \exp((\theta) t exp(-\alpha' x)}{1+\exp((\theta)\{t exp(-\alpha' x)\}^{\gamma}}$ …….. (3.22)

The log-logistic accelerated failure time model with a covariate x may be expressed as

$Y = \log T = \mu + \alpha' x + \sigma\varepsilon$ where $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_p)$; ε has the standard logistic distribution. Here, the survival function for the time to the event with covariate x is given by:

$$S_T(t|x) = \frac{1}{1 + \lambda exp(\beta' x) t^{\gamma}} = \frac{1}{1 + exp\{\log \lambda + (\beta' x)\}}$$

$$h_T(t|x) = \frac{\gamma t^{\gamma-1} \lambda exp(\beta' x)}{1 + \lambda exp(\beta' x) t^{\gamma}} = \frac{\gamma t^{\gamma-1} \lambda exp(\beta' x)}{1 + exp\{\log \lambda + (\beta' x)\}}$$

As for the Weibull distribution, the parameters are related by

$$\lambda = exp\left(\frac{-\mu}{\sigma}\right) = \exp(\theta) \; ; \; \gamma = {}^{1}/_{\sigma} \; ; \; \beta_j = {}^{-\alpha_j}/_{\sigma}$$

Therefore, the survival time for the ith individual has a log-logistic distribution with parameter $\log \lambda + (\beta' x)$ and γ; log-logistic distribution has AFT property.

If the baseline survival function is $S_0(t) = \frac{1}{1+\lambda t^{\gamma}}$ where θ and γ are unknown parameters, then the baseline odds of surviving beyond time t are given by $\frac{S_0(t)}{1-S_0(t)} = \frac{1}{1+\lambda t^{\gamma}}$

To interpret the factor $\exp(\beta' x)$ for the log logistic model, one can notice that the odds of survival beyond time t for the logistic model is given by $\frac{S_T(t)}{1-S_T(t)}$

We can see that the log-logistic distribution has the proportional odds (PO) property. So this model is also a proportional odds model, in which the odds of an individual surviving beyond time t are expressed as

$$\frac{S_T(t)}{1 - S_T(t)} = \exp(-\alpha' x) \frac{S_0(t)}{1 - S_0(t)}$$

So, the factor $\exp(-\alpha' x)$ is an estimate of how much the baseline odds of survival at any time changes when an individual has covariate x.

Note that $\exp(\alpha' x)$ is the relative odds of experiencing the event for an individual with covariate x compared to an individual with the baseline characteristics.

As this representation of a log logistic regression is as an accelerated failure-time model with a log logistic baseline survival function, then the log logistic model is the only parametric model with both a proportional odds and an accelerated failure-time representation

If Ti has a log-logistic distribution, then ε_i has a logistic distribution. The survival function of logistic distribution is given by

$$S_{\varepsilon i}(\varepsilon) = \frac{1}{1 + exp(\varepsilon)}$$

Using equation (4.14), the AFT representation of survival function of the log-logistic model is given by

$$S_i(t) = \left[1 + t^{\frac{1}{\sigma}} \exp\left(\frac{-\mu - \alpha' x}{\sigma}\right)\right]^{-1} \quad \ldots\ldots\ldots\ldots (3.23)$$

According to the relationship of survival and hazard function, the hazard function for the i^{th}individual is given by

$$h_i(t) = \frac{1}{\sigma t}\left\{1 + t^{-\frac{1}{\sigma}} exp\left(\frac{-\mu - \alpha' x}{\sigma}\right)\right\}^{-1} \quad \ldots\ldots\ldots\ldots (3.24)$$

The p[th] percentile of the survival distribution for the i[th] individual is

ti (p), from equation (3.24),is

$$t_i(p) = exp\left[\sigma log\left(\frac{100 - p}{100}\right) + \mu + \alpha' x_i\right]$$

The median survival time is

$$t_i(50) = exp(\mu + \alpha' x_i) \quad \ldots\ldots\ldots\ldots (3.25)$$

plot of $\log\left[\frac{1-S(t)}{S(t)}\right]$ versus log t should be linear if log-logistic distribution is appropriate.

c. Log-normal AFT model

If the survival times are assumed to have a log-normal distribution, the baseline survival function and hazard function are given by

$$S_0(t) = 1 - \Phi\left(\frac{logt - \mu}{\sigma}\right) \quad , \qquad h_0(t) = \frac{\Phi\left(\frac{logt}{\sigma}\right)}{\left[1 - \Phi\left(\frac{logt}{\sigma}\right)\right]\sigma t}$$

Where, μ and $\sigma$are parameters, ϕ(x) is the probability density function and Φ(x) is the cumulative density function of the standard normal distribution. The survival function for the ith individual is

$$S_i(t) = S_0\left(t * exp(\mu + \alpha' x)\right)$$
$$= 1 - \Phi\left(\frac{logt - \mu - \alpha' x}{\sigma}\right)$$

Therefore, the log survival time for the i[th] individual has normal$(\mu + \alpha' x_i, \sigma)$. The log-normal distribution has the AFT property.

In a two group study, we can easily get $\Phi^{-1}\left(1 - S(t)\right) = \frac{logt - \mu - \alpha' x}{\sigma}$ where xi is the value of a categorical variable which takes the value one in one group and zero in the other group. This implies that a plot of $\Phi^{-1}\left(1 - S(t)\right)$ versus logt will be linear if the log-normal distribution is appropriate.

d. Gamma AFT model

There are two different gamma models discussed in survival analysis literature. The standard (2-parameter) and the generalized (3-parameter) gamma models The gamma model means the generalized gamma model in this thesis. The probability density function of the generalized gamma distribution with three parameters λ ,α and γ is defined by

$$f(t)=\frac{\alpha\lambda^{\alpha\gamma}}{\Gamma(\gamma)} t^{\alpha\gamma-1} \exp\left[-(\lambda t)^{\alpha}\right] \qquad t>0, \lambda>0, \alpha > 0, \gamma > 0$$

Where, γ is the shape parameter of the distribution. The survival function and the hazard function do not have a closed form for the generalized gamma distribution. The exponential, Weibull and log-normal models are all special cases of the generalized gamma model. It is easily seen that this generalized gamma distribution becomes the exponential distribution if $\gamma = \alpha = 1$; the Weibull distribution if $\gamma = 1$; and the log-normal distribution if $\gamma \to \infty$. The generalized gamma model can take on a wide variety of shapes except for any of the special cases. For example, it can have a hazard function with U or bathtub shapes in which the hazard declines reaches a minimum and then increases.

3.4.5.5.Parameter estimation for parametric model

AFT models are fitted using the maximum likelihood method. The likelihood of the n observed survival times, $t_1, t_2, \dots, t_n$ is given by the likelihood function for right-censored data, following the construction the likelihood function for right-censored data, is given by

$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} f_j\left(t_j\right)^{\delta_j} S_j\left(t_j\right)^{1-\delta_j}$ …… (3.26)

Where $f_j\left(t_J\right)\ and\ S_j\left(t_j\right)$ where fj(tj) and Sj(tj) are the density and survival functions for the jth individual at tj and $\delta_j$is the event indicator for the jth observation and has different functional form depending on the type of survival distribution we used.

$\log L(\alpha, \mu, \sigma) = \sum_{i=1}^{n}\{-\delta \log\left(\delta t_j + \delta_j \log f_i\left(x_i\right) + (1-\delta_i) \log s_i(W_i)\right)\}$…….. (3.27)

Where $W_i = \left\{\log t_j - \frac{(\mu+\alpha_{1j}x_i+\cdots\alpha_{pj}x_{pj})}{\delta}\right\}$ and $Z = \{Z_{ij}\}$ is a vector of covariates

for jth individual. The maximum likelihood estimates of the p+2 unknown parameters, $\mu, \delta, \alpha_1, \alpha_2, \dots, \alpha_p$ are found by maximizing this function using the Newton-Raphson procedure

3.4.5.6.Checking the Adequacy of parametric Model

The graphical methods can be used to check if a parametric distribution fits the observed data. Specifically, if the survival time follows an exponential distribution, a plot of $log[-logS(t)]$ versus $logt$ should yield a straight line with slope of 1. If the plots are parallel but not straight, then PH assumption holds but not the Weibull. If the lines for two groups are straight but not parallel, the Weibull assumption is supported but the PH and AFT assumptions are violated. The log-logistic assumption can be graphically evaluated by plotting $log\left(\frac{1-S(t)}{S(t)}\right)$ versus$logt$.

If the distribution of survival function is log-logistic, then the resulting plot should be a straight line. For the log-normal distribution, a plot of $\Phi^{-1}\left(1-S(t)\right)$ versus logt should be linear. All these plots are based on the assumption that the sample is drawn from a homogeneous population, implying that no covariates are taken into account. So this graphical method is not very reliable in practice.

There are other methods to check the fitness of the model.

a. Using quantile - quantile plot

An initial method for assessing the potential for an AFT model is to produce a quantile-quantile plot. For any value of p in the interval (0; 100), the p^{th} percentile is

$$t(p) = S^{-1}\left(\frac{100-p}{100}\right)$$

Let $t_0(p)$ and $t_0(p)$ be the p^{th}percentiles estimated from the survival functions of the two groups of survival data. The percentiles for the two groups may be expressed as

$$t_0(p) = {S_0}^{-1}\left(\frac{100-p}{100}\right) \; ; \; t_1(p) = {S_1}^{-1}\left(\frac{100-p}{100}\right)$$

Where $S_0(t)$ and $S_0(1)$ are the survival functions for the two groups. So we can get

S_1 [t1 (p)) = S_0[t0(p)):

Under the AFT model , $S_1(t) = S_0\big(t * exp(-\alpha' x)\big)$, and so $S_1[t_1(p)] = S_0[t_1(p) * exp(-\alpha' x)]$

Therefore, we get $t_0(p) = t_1(p) * exp(-\alpha' x)$

The percentiles of the survival distributions for two groups can be estimated by the K-M estimates of the respective survival functions. A plot of percentiles of the K-M estimated survival function from one group against another should give an approximate straight line through the origin if the accelerated failure time model is appropriate. The slope of this line will be an estimate of the acceleration factor $\exp(-\alpha' x)$

b. Using statistical criteria

The need to select a model is of great importance in statistics in order to ensure goodness of fit and adjust or penalize for model complexity. The observed data is usually from an unknown probability distribution. As a result, several models are fitted in order to find the best. The models that are not very close to the actual distribution have to be discarded. Below are different statistics that are commonly used in the model selection.

We can use statistical tests or statistical criteria to compare models. For the aim of comparison among parametric and semi parametric models Akaike Information Criterion (AIC) and standardized of parameter estimates can be used. Nested models can be compared using the likelihood ratio or Wald test.

A likelihood ratio test is a statistical test used to compare the fit of two models. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favor of the alternative model.

LRT is based on the joint probability density function of observable random variables. At the null model, subsequent updated model are estimated and it is viewed as the function of parameters given the realized random variables. The LRT then uses the Chi-square test to assess if any updated model offers an improvement in goodness of fit against the null model. The exponential model, the Weibull model and log-normal model are nested within gamma model.

For comparing models that are not nested, the Akaike information criterion (AIC) can be used instead, which is defined as

$$\mathrm{AIC} = -2(\mathrm{loglikelihood}) + 2(\mathrm{k} + \mathrm{c})$$

Where, k is the number of covariates in the model and c is the number of model-specific ancillary parameters. The addition of 2(k +c) can be thought of as a penalty if non predictive parameters are added to the model. Although the best-fitting model is the one with the largest log likelihood, the preferred model is the one with the smallest AIC value.

But there is a difficulty in using the AIC in that there are no formal statistical tests to compare different AIC values. When two models have very similar AIC values, the choice of model may be hard and external model checking or previous results may be required to judge the relative plausibility of the models rather than relying on AIC values alone.

The efficiencies can also be compared using Standardized measure of variability which is analogous to the coefficient of variation and defined by the ratio of standard error to the corresponding parameter estimate $\left(sv = \frac{\sigma_\beta}{|\beta|}\right)$ in which the smaller is preferred.

c. Using residual plots

Residual plots can be used to check the goodness of fit of the model. Procedures based on residuals in the AFT model are particularly relevant with the Cox PH model.

One of the most useful plots is based on comparing the distribution of the Cox-Snell residuals with the unit exponential distribution. The Cox-Snell residual for the i[th] individual with observed time ti is defined as

$r_{ci} = \widehat{H}_0(t_i)\exp\left(\sum_{i=1}^{p} x_{ik}\hat{\beta}_k\right) = \widehat{H}_i(t) = -\log\left(\hat{S}(t_i)\right)$; Where ti is the observed survival time for individual i, xi is the vector of covariate values for individual i, and Si (ti) is the estimated survival function on the fitted model. From equation (a), the estimated survival function for the i^{th} individual is given by

$\hat{S}_{\varepsilon i}\left(\frac{\log t_i - \hat{\mu} - x'_i\hat{\alpha}}{\hat{\sigma}}\right)$; where $\hat{\mu}, \hat{\alpha}\ and\ \hat{\sigma}$ are the maximum likelihood estimator of μ, α and σ respectively. $\hat{S}_{\varepsilon i}(\varepsilon)$ is the survival function of εi in the AFT model, and $\frac{\log t_i - \hat{\mu} - x'_i\hat{\alpha}}{\hat{\sigma}} = r_{si}$ is referred to as standardized residual.

The Cox-Snell residual can be applied to any parametric model. The corresponding form of residual based particular AFT model can be obtained. For example, under the Weibull AFT model, since $\hat{S}_{\varepsilon i}(\varepsilon) = exp\left(-exp(\varepsilon)\right)$, the Cox-Snell residual is th

$$r_{ci} = -\log\left(\hat{S}(t_i)\right) = \log S_{\varepsilon i}\,(r_{si}) = exp(r_{si})$$

Under the log-logistic AFT model, since $S_{\varepsilon i}(\varepsilon) = \left(1 + exp(\varepsilon)\right)^{-1}$, the Cox-Snell residual is then $r_{ci} = \log(1 + \exp(r_{si}))$

If the fitted model is appropriate, the plot of log (-log S(rci)) versus logrci is a straight line with unit slope through the origin.

These residuals lead to the deviance residuals for the particular AFT model. A plot of deviance residuals against the survival time or explanatory variables can be used to check whether there are particular times, or particular values of explanatory variables, for which the model is not a good fit.

4. STATISTICAL ANALAYSIS AND RESULTS

4.1.Descriptions of the data

In total, 440 individuals contributed to the data used in our analysis. The total observation time for one individual varied from 1 to 66 months, of which 88 (20%) were death and the remaining 352 (80%) were censored observations. Among 88 patients who died, 46 (52.27%) died within 3 months after start of treatment, this indicates that more of the patients were died early death or within 3 months after treatment.

From table 4.1 below the overall median follow-up time of all the patients was 34 months (Interquartile range: 9.0-54 .0 months). The median follow-up time of the death was 3.0 months (Interquartile range: 1.0-17.25 months) and censored patients was 46.0 months (Interquartile range: 16.75 -56 .0 months). The median age of the patients at start of ART was 31 years (Interquartile range 26.75–37 years), The median baseline bodyweight was 50 kg (Interquartile range 44–59). The median baseline CD4 count was 162.5 cells/mm^3 (Interquartile range was 89.75–233.25).

The estimated average follow up time, age, baseline weight and baseline CD4 count of died patients with their corresponding standard error were at least 32.42727(22.88227) in months,32.99318(8.934459) in years, 51.37614(11.00828) in kg and 176.6432(125.3515) cells/mm^3 respectively.

Table 4.1: Summary statistics for continuous variables included in the study of HIV positive patients under HAART in JUSH

Continuous variables	mean	St. err		maximum	median	Q1	Q3
Time	32.42727	22.88227	1	66	34	9.0	54.0
Event time	12.46591	17.19727	1	66	3.0	1.0	17.25
Censored time	37.42	21.3704	1	66	46.0	16.75	56.0
Age	32.99318	8.934459	18	67	31	26.75	37.0
Bweight	51.37614	11.00828	20.5	93	50	44.0	59.0
BCD4count	176.6432	125.3515	7	398	162.5	89.75	233.25

The demographic information and some basic base line covariate from the HIV positive patients were also reported on table 4 below. As observed from the table 4 below by the categorical group of the covariates out of total of 440 HIV positive patients 278(63.2%) of them were females and 49(55.7%) death were also occurred in female groups in comparison with male HIV positive patient groups.

Among the total of marital status category the larger number 215(48.5%) of the HIV positive patients were married and 138(31.4%) of HIV positive were belongs to others group(including widowed, divorced ,..) while smaller number 87(19.3%) of the HIV positive patients belong to never marital status groups. of the total deaths occurred in these groups 47(53.4 %) , 22(25.0%) and 19(21.6%) of the deaths occurred in married ,others and never married marital status groups respectively, in comparison married marital status groups represents the larger died percentages according to the marital status category.

When we look at the educational level category of the HIV positive patient's larger number 153(34.8%) were attended their primary education, 136(30.9%) were attended their secondary education, 97(22.0%) not educated patients and only 54(12.3%) attended their tertiary educations.

Among the total deaths occurred in educational level category 34(38.60%) deaths occurred in primary education level category groups while smaller number of deaths 11(12.5%) occurred in tertiary educational level group. There were 290(65.9%) patients who were able to work, 131(29.8%) were ambulatory and 19(4.3 %) were bedridden in the functional status categories, of the total patients deaths in these categories 41(46.6 %) of the deaths were occurred in patient group who was ambulatory at the base line.

More of these HIV positive patients 272(61.8%) came from the urban and large number 59(67.0%) of deaths also occurred in this group in comparison with rural areas of the town.

Of the total HIV positive patients by their baseline clinical stages 90(20.5 %) were at clinical stage I, 131(29.8 %) were at clinical stage II, 164(37.3 %) were at clinical stage III and the rest 55(12.5%) were at clinical Stage IV whereas of the total deaths occurred in clinical stage categories 72(81.9%) deaths were occurred in both clinical stage III and IV at base line time in comparison with remaining baseline clinical stages.

The larger number of HIV positive patients 296(67.3%) were at the regimen type TDF-3TC-EFV and 144(32.7) were at the regimen type AZT-3TC-NVP of the total death the larger number 76(86.4%) of death happens at the regimen type TDF-3TC-EFV. More of those HIV positive patients 343(78.0%) have no pre-TB positive test and a large number of 45(51.1%) of deaths occurred in no pre –TB positive test group in comparison with pre-TB status group.

There were 193(43.9%) patients whose BMI was less than 18.5, 217(49.3%) were BMI between 18.5 and less than 24.9 and 30(6.8%) were BMI greater than or equal to 25 in the body mass index categories, of the total patients deaths in these categories 50(56.8%) of the deaths were occurred in patient group whose BMI was less than 18.5 at the base line.

Table 4.2: Frequencies and percentages for the baseline categorical covariates together with the status of HIV positive patients

Variables	Categories	Total n (%)	Status of the observation	
			Censored observation n(%)	Observed events n(%)
Sex	Female	278 (63.2%)	229 (65.1%)	49 (55.7%)
	Male	162 (36.8%)	123 (34.9%)	39 (44.3%)
Marital status	Never married	87 (19.8%)	68 (19.3%)	19 (21.6%)
	Married	215 (48.9%)	168 (47.7%)	47 (53.4%)
	Others	138 (31.4%)	116 (33.0%)	138 (25.0%)
Education level	Not educated	97 (22.0%)	81 (23.0%)	16 (18.2%)
	Primary	153 (34.8%)	119 (33.8%)	34 (38.6%)
	Secondary	136 (30.9%)	109 (31.0%)	27 (30.7%)
	Tertiary	54 (12.3%)	43 (12.2%)	11 (12.5%)
Residence	Rural	168(38.2%)	139(39.5%)	29 (33.0%)
	Urban	272(61.8%)	213(60.5%)	59 (67.0%)
WHO stage	WHO stage I	90(20.5%)	86(24.4%)	4 (4.5%)
	WHO stage II	131(29.8%)	119(33.8%)	12 (13.6%)
	WHO stage III	164(37.3%)	121(34.4%)	43 (48.9%)
	WHO stage IV	55(12.5) %	26(7.4%)	29 (33.0%)
Regimen type	ATZ-3TC-NVP	144 (32.7%)	132 (37.5%)	12 (13.6%)
	TDF-3TC-EFV	296 (67.3%)	220 (62.5%)	76 (86.4%)
TB	No pre-TB status	343 (78.0%)	298 (84.7%)	45 (51.1%)
	Yes pre-TB status	97 (22.0%)	54 (15.3%)	43 (48.9%)
Body mass index	BMI less than 18.5	193 (43.9%)	143 (40.6%)	50 (56.8%)
	18.5<=BMI<=24.9	217 (49.3%)	180 (51.1%)	37 (42.0%)
	BMI>=25	30 (6.8%)	29 (8.2%)	1 (1.1%)

4.2. Descriptive and Non-parametric analysis

Before proceeding to more complicated models, we make a descriptive analysis that will use as initiation to our subsequent findings. Here we start with the Kaplan-Meier estimator is applied to estimate the survival curves for categorical covariates. The overall survival probability among HIV positive patients declined over follow-up time (see appendix Figure 1).Plot of the Kaplan-Meier estimates for only two selected categorical covariates; Regimen type and Functional Status were displayed below in Figure 2.

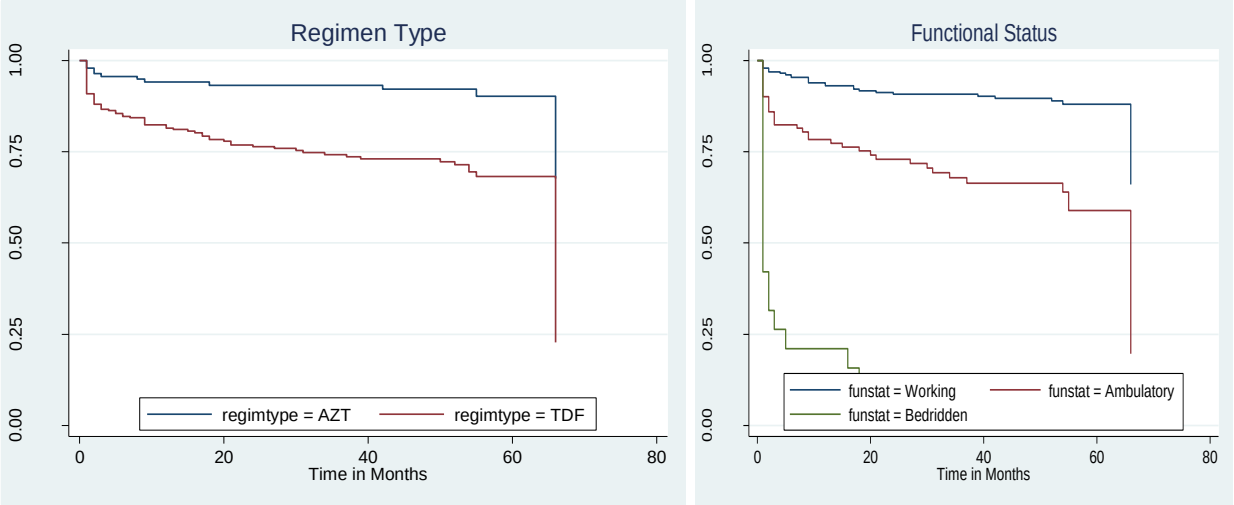


Figure 2: Plot of Kaplan-Meier Estimates of survivor function of HIV positive patients under HAART for Regimen type and Functional Status

Obviously, the left hand side plot of the above figure shows that Regimen type 1 patients have higher probability of survival throughout the five years HAART treatment period than regimen type 2 patients. Similarly, the right hand side plot of the figure shows that those patients who were able to work day to day when they started ART have higher probability of survival than those patients who were either ambulatory or bedridden. Also, it shows that those patients who were bedridden have less probability of survival or fail more quickly than those patients who were ambulatory. The survivorship estimate curve for gender, marital status, education level, pre-TB treatment status and body mass index (Figure 3, in the appendix) shows that there was difference in the survival times among the levels of covariates. For comparing the survival experiences between groups; the log-rank test was applied to all categorical variables.

The null hypothesis to be tested is that there is no difference between the probabilities of an event occurring at any time point for each population. The R and STATA results have been summarized in Table 5 below.

Table 4.3: Log-Rank and Peto & Peto test for equality of survival experience among different groups of covariates

Covariates	Log –Rank Test			Peto & Peto Test		
	Test Statistic	df	p-value	Test Statistic	df	p-value
Sex	3.39	1	0.0657	3.22	1	0.0729
Marital status	1.93	2	0.3809	2.01	2	0.3669
Education level	1.20	3	0.7540	1.24	3	0.7428
Functional status	178.37	2	0.0000	174.86	2	0.0000
Residence	0.89	1	0.3456	0.90	1	0.3423
WHO clinical stage	63.76	3	0.0000	61.50	3	0.0000
Regimen type	21.12	1	0.0000	20.67	1	0.0000
Pre-TB test	43.47	1	0.0000	41.47	1	0.0000
Body mass index	13.20	2	0.0014	13.99	2	0.0009

Note: * Test was carried out at $p<0.05$

From the table above both the log –rank and Peto & Peto test results suggests that functional status, WHO clinical stage, regimen type, pre-TB status and body mass index were covariates whose different levels have a significant difference in survival experience (i.e. time to death dependency on those covariates) while sex, marital status, education level and residence does not have a different survival experience.

4.3.Cox Proportional Hazard Regression Model

After making a comparison of the survivorship experience among groups of covariates, the next important step is model development. An initial step in the model building process is to identify sets of explanatory variables that have the potential for being included in the linear components of a multivariable proportional hazards model. We started with fitting univariable Cox proportional hazards regression model.

Table 4.4: Uni-variate analysis of Cox PHs on the survival time of HIV positive patients under HAART (at JUSH, during 2010-2015)

	Univar iate analysis			
Covariates	β	HR	95%CI	P-value
Gender	0.3993	1.491	(0.9763,2.277)	0.0645
Age	-0.0034	0.9966	(0.973, 1.021)	**0.78**
Marital status				
Married	0.02742	1.0278	0.6026 ,1.753)	**0.920**
Others	-0.31522	0.7296	(0.3940 ,1.351)	**0.316**
Education level				
Primary	0.29957	1.349	(0.7446,2.445)	**0.323**
Secondary	0.18309	1.201	(0.6470,2.229)	**0.562**
Tertiary	0.03031	1.031	(0.4730,2.246)	**0.939**
Functional status				
Ambulatory	1.3554	3.878	(2.406 ,6.251)	2.61e-08
Bedridden	3.3740	29.195	(15.823,53.868)	< 2e-16
residence	0.2055	1.228	(0.7873, 1.916)	0.365
Base line weight	-0.03770	0.963	(0.9429,0.9835)	0.00042
Base line CD4	-0.005094	0.9949	(0.9924,0.9974)	5.87e-05
WHO clinical stage				
Stage II	0.7842	2.191	(0.7065, 6.793)	0.17439
Stage III	1.9557	7.069	(2.5355,19.709)	0.00018
Stage IV	2.6978	14.846	(5.2163,42.255)	4.3e-07
Regimen type	1.3355	3.802	(2.061, 7.012)	1.9e-05
Pre-TB positive test	1.3171	3.733	(2.448, 5.692)	9.39e-10
Body mass index				
BMI >=18.5 &<24.9	-0.6047	0.5462	(0.35587,0.8385)	0.00568
BMI>= 25	-2.1860	0.1124	(0.01552 ,0.8136)	0.03045

- Others include Divorced, Widowed

The univariate Cox PHs results shows that functional status, base line weight, base line CD4 count, WHO clinical stage III and IV, regimen type , presence of pre-TB positive test and body mass index are statistically significant covariates that influence the survival time of HIV positive patients under HAART at 10 % level of significance. But remaining variables which were used in the single covariate analysis (such as gender, age, marital status, education level, place of residence) were found to be not statistically significant with survival.

We considered including the predictor in the multivariable model if the test for the univariate model has a p-value < 0.1 in the univariate analysis. If the predictor has a p-value greater than 0.2 in a univariate analysis it is highly unlikely that it will contribute anything to multivariable model which includes other predictors. We then fit the full multivariate Cox PH model including all the potential covariates which are significant at 10% at the univariate level. Then we selected among variables significant at 5% in the multivariable analysis. Accordingly, functional status, WHO clinical stage, CD4 count, regimen type, pre-TB positive test and body mass index were the best combination of variables to yield the minimum possible AIC of all the combinations (table 7).

Table 4.5: Multivariable analysis of Cox PHs on the survival time of HIV positive patients under HAART (at JUSH, during 2012-2015)

Covariates	Multivariable analysis						
	β	HR	Se(β)	z	95%CI		P-value
Functional status							
Ambulatory	0.7078	2.0296	0.2686	2.635	(1.1988	3.4360)	0.00841
Bedridden	2.7736	16.0169	0.3524	7.872	(8.0290	31.9516)	3.44e-15
Base line CD4	-0.0019	0.9981	0.0008	-2.239	(0.9964	0.9998)	0.02514
WHO clinical stage							
Stage III	1.4846	4.4131	0.5396	2.751	(1.5328	12.7061)	0.00593
Stage IV	1.6421	5.1663	0.5640	2.912	(1.7105	15.6039)	0.00359
Regimen type	0.99908	2.7158	0.3154	3.168	(1.4637	5.0389)	0.00153
Pre-TB status	1.2662	3.5472	0.2158	5.867	(2.32382	5.4146)	4.43e-09
Body mass index							
BMI>=18.5&<24.9	-0.5517	0.5759	0.2183	-2.527	(0.37544	0.8835)	0.0115
BMI>= 25	-2.0211	0.1325	1.0107	-2.000	(0.01828	0.9607)	0.0455

4.3.1. Assessment of Cox Proportional Hazards Model Adequacy

Checking of Proportional hazard assumption

The PH assumption checking with graphical method and two statistical test methods (adding time-dependent covariates in the Cox model and tests based on the Schoenfeld residuals were considered. We used the – (log (-log (survival))) plot versus survival time to check the PH assumption for all the categorical variables in included in the model (appendix1). The graphs for each of the categorical variable display lines that appeared to be parallel implying that the proportional-hazards assumption among categorical variable such as functional status, WHO clinical stage, regimen type, pre-TB positive test and body mass index has not been violated.

We also create the time-dependent covariate by creating interactions of the predictors and survival time and include them in the model. The result indicates that the PH assumption for baseline CD4 count is violated (p-value for baseline CD4 count*log time is less than 0.05). However the coefficient for the interaction term obtained to be statistically insignificant for the covariates functional status, WHO clinical stage, regimen type, pre-TB status and body mass index indicating that this covariate are time independent variables.

Table 4.6: Statistical test for proportional hazards assumption (PH) of the covariates and their interaction with log of time (time to death)

Covariates	HR(95%CI)	Se(β)	z	P-value	95%CI
Functional status					
Ambulatory	1.6691	.4468154	1.92	0.055	(0.988322,2.82121)
Bedridden	9.1348	3.387438	5.97	0.059	(0.416109,14.8949)
BaselineCD4	.9937161	.0019986	-3.13	0.002	(.9898067,.997641)
WHO clinical stage					
Stage II	1.653238	.8810801	0.94	0.346	(0.581693,4.69868)
Stage III	2.534833	1.292656	1.82	0.068	(0.932982,6.88692)
Stage IV	3.281873	1.768937	2.20	0.086	(0.141096,8.43890)
Regimen type	2.117805	.6169562	2.58	0.087	(0.141096,8.43890)
Pre-TB status	1.368304	.3380429	1.27	0.204	(0.843123,2.22061)
Body mass index					
BMI>=18.5&<24.9	.7122486	.1599486	-1.51	0.131	(0.458646,1.10607)
BMI>= 25	.3160411	.3229584	-1.13	0.260	(0.042649,2.34193)
BaselineCD4count*logtime	1.002226	.0007283	3.06	**0.002**	(1.000799,1.00365)
-2loglikelihood	885.79				

The Schoenfeld residuals are also used to check the PH assumption. We check the p-value for testing whether the correlation between Schoenfeld residual for these covariates and ranked survival time is zero. The p-value= 0.0260 for baseline CD4 count and GLOBAL test also not validate general acceptability of the proportional hazard assumption (p-value =0.0186).This provides the evidence that the covariate baseline CD4 count violates the PH assumption at the level of = 0.05 and the PH assumption is not violated for the remaining covariates (table 9).

Furthermore, plotting the scaled Schoenfeld residuals of each covariate against log time will be used to check whether the assumption of proportional hazards is violated or not. Clearly, a close look of this plot indicates that the residuals are random and LOESS curve are smooth and horizontal with zero slope for the covariates functional status, WHO clinical stage, regimen type, pre-TB positive test and body mass index. But, the impact of the baseline CD4 count clearly changes with time and this time-varying effect can be seen in (appendix Figure 5). When the assumption of proportionality does not hold, alternative models have to be considered We also assess goodness of fit by residual plots. A plot of the Cox-Snell residuals against the cumulative hazard of Cox-Snell residuals is presented in below figure 6. This plot reveals that there is some evidence of a systematic deviation from the straight line, which gives us some concern about the adequacy of the fitted model.

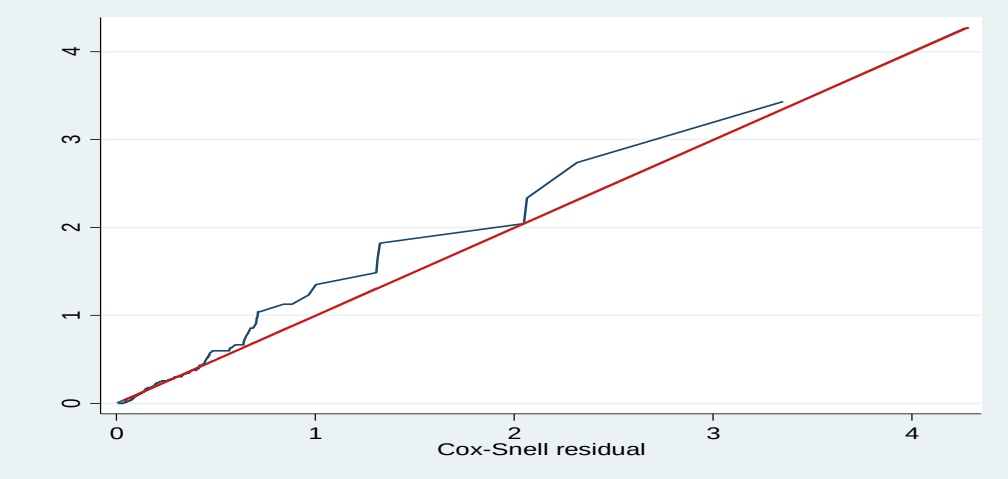


Figure 6: cumulative hazard plot of the Cox-Snell residual for Cox PH model

Checking for Linearity Assumption:

In order to assess the linearity assumption on the part of the covariates, we plotted martingale residuals compared by excluding the covariate to be checked for linearity against the values of the covariate. The scatter plots or the smoothed curve using Lowess can be used to check the linearity assumption. In our model "baseline CD4 count" is the only continuous variable among the six significant covariates. So we tried to search for an appropriate functional form of the effect of baseline CD4count. Accordingly, Martingale residuals for covariate baseline CD4count are plotted against the baseline CD4count values along with a Lowess smooth curve in the annex of figure 7 revealed that the plots are linear and showed a correct functional form. The result does not show any trend and the resulting smoothed plots (LOESS) can be described as horizontal straight lines. Thus, there are no signs of nonlinearity in baseline CD4count indicating that the assumption of linearity is fulfilled by baseline CD4count.

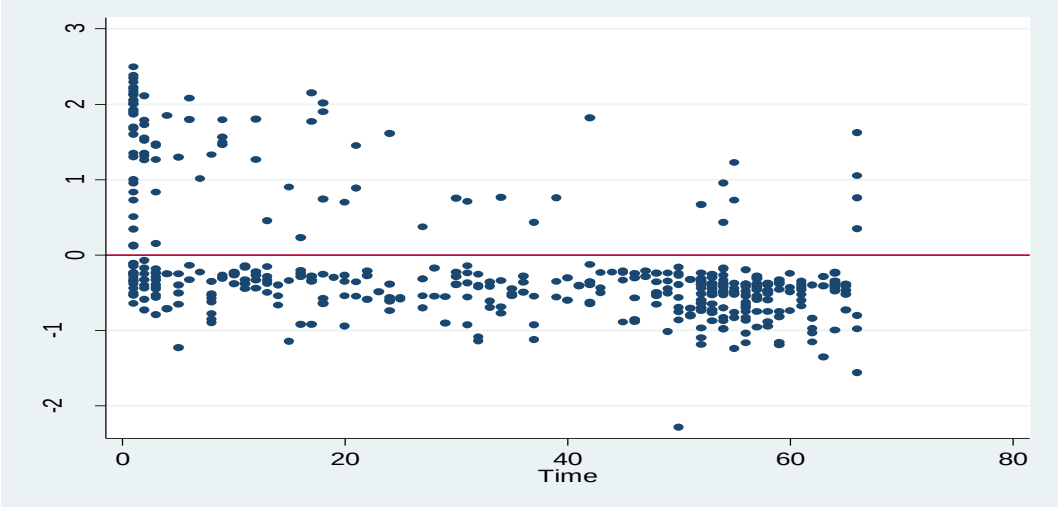


Figure 8: Deviance residuals versus the risk score for Cox PH model

The plot of deviance residual against the risk score shows that the deviance residuals seem not to be symmetrically distributed about zero. There are very high or very low deviance residuals which suggest that these observations may be outliers (Figure 8). Therefore, we have some concern about the adequacy of the fitted Cox PH model.

Checking for Influential Observations: We also use delta-beta statistic to measure the influential observations on the model as a whole (appendix figure 9). According to figure 9, the coefficient does not change too much when the observations corresponding to the largest delta-beta statistics are removed. That is, comparing the magnitudes of the largest dfbeta values to the regression coefficients suggests that none of the observations is terribly influential individually. Therefore, we do not remove them from the dataset and conclude that there are no influential observations.

Model Interpretation

$h(t|Z_i)=h_0(t)\exp(0.7078\text{FuncStat}_1+2.7736\text{FuncStat}_2-0.0019\text{BCD4count}+1.4846\text{WHOstage}_3$
$+1.6421\ \text{WHOstage}_4\ +0.99908\text{Regimen type}+1.2662\text{Pre-TB status}\ -0.5517\text{BMI}_1\ -2.0211\text{BMI}_2)$

The multivariable Cox PH model fitted to HIV HAART data shows that functional status, base line CD4 count, WHO clinical stage III &IV, regimen type, Pre-TB status and Body mass index are covariates that significantly influences the survival time of HIV positive patients under HAART at 5% significance levels as shown in Table 7.

The estimated hazard ratio and 95% confidence interval for Ambulatory and bedridden functional status group were 2.0296 (1.1988, 3.4360) and 16.0169(8.0290, 31.9516) respectively. This implies that Patients with baseline functional status group of ambulatory were two times more likely to die compared to patients with working functional status group. However, the risk of death increased 16.0169 times when the patients had baseline functional status of bedridden compared to patients with working functional status group holding other factors constant.

The estimated hazard ratio and 95% confidence interval for baseline CD4 count was 0.9981(0.9964, 0.9998). This implies that a unit increase in the CD4 count of a patient will decrease the estimated hazard by 0.19% assuming that all covariates are constant.

HIV-positive patients with WHO clinical stage III had increased risk of death compared to patients with stage I (HR 4.4131; 95% CI: 1.5328-12.7061), and the risk of death among WHO clinical stage IV patients was even higher- compared to stage I patients (HR 5.1663; 95% CI: 1.7105-15.6039) assuming that all other covariates are held constant.

The regimen type 2 is also significant at 5 % significance level with an estimated hazard ratio of 2.7158(1.4637, 5.0389) compared with regimen type 1. This implies that patient in regimen type 2 has higher estimated hazard compared patient in regimen type 1 holding other factors constant. Patient those have pre-TB positive test is also significant at 5% significance level with higher estimated hazard ratio of 3.5472 (2.32382, 5.4146) compared with patients with no pre-TB positive test holding other factors constant.

Regarding body mass index of HIV positive patients under HAART, for patients who had BMI greater or equal to 18.5 and less than 24.9 , there was a decrease in an estimated hazard ratio (HR= 0.5759 , 95% CI: 0.37544 ,0.8835), than those with BMI less than 18.5. Comparing patients with BMI greater than 25(HR= 0.1325, 95% CI: 0.01828, 0.9607) with those with BMI greater or equal to 18.5 and less than 24.9, it was observed that there was a decrease in hazard with 45 units.

4.3.2. Strategies for analyzing of non-proportional data

4.3.2.1. Cox model with time-dependent variable

We have shown that the Cox model displayed non proportionality for variable baseline CD4 count. We observed that there is an interaction between baseline CD4 count and time. It is not appropriate to use stratified Cox model because baseline CD4 count is a continuous variable. We then incorporate time-dependent covariate into the model.

Table 4.8: Time-dependent covariates represent different time periods

	Year				
Variable	0-0.6 year	0.6-1.2 years	1-2 years	2-3 years	>3 years
$g_1(t)$	0	1	0	0	0
$g_2(t)$	0	0	1	0	0
$g_3(t)$	0	0	0	1	0
$g_4(t)$	0	0	0	0	1

We define $X(t) = \text{baseline CD4 count} \times \text{t}$ and formulate a model (model A)

$h(t, \mathbf{x}(t)) = h_0(t)\{\exp(\beta_b BCD4count + \beta_1(BCD4count \times t) + \beta' x)\}$, \where x is the vector of all the fixed covariates (functional status, WHO clinical stage, regimen type, pre-TB positive test and body mass index) and β is the corresponding vector of the regression coefficient for the .fixed covariates.

The effect of CD4 count is $\exp(\beta_b + \beta_1 t)$. β_b can be interpreted as the effect of CD4 count at study enrollment. This model states that the effect of CD4 count is assumed to increase or decrease linearly in relation to time. Alternatively, the proportional hazards assumption may hold at least approximately over short time periods instead of the entire time period. In this situation, hazard ratio of CD4 count may change at discrete intervals. We partitioned the time period into 5 sub-periods and created four binary, time-dependent covariates to represent them (Table 10).

The first interval goes from 0 to half of a year; the second time interval goes from half to one year; the third time interval goes from 1 year to 2 years; the fourth interval goes from 2 to 3 years; and the last interval goes from 3 years onward. This model assumes that there are five different hazard ratios estimates in five intervals.

The Cox non-PH model is fitted as follows:

$$h(t, x(t)) = h_0(t) exp\{\delta_b CD4count + \delta_1(CD4count \times g_1(t)) + \delta_2(CD4count \times g_2(t)) + \delta_3(CD4count \times g_3(t)) + \delta_4(CD4count \times g_4(t)) + \boldsymbol{\delta}'\mathbf{x}\}$$

Table 4.9: Time-dependent effect of baseline CD4 count in four time Intervals

Period	$\boldsymbol{\delta}$	HR	P-value	95%CI
<12 months	-0.0113	0.989	0.000	0.98, 0.99
1-2 years	-0.0182	0.982	0.002	0.97, 0.99
2-3 years	-0.0115	0.988	0.110	0.97 1.01
>3 years	-0.0001	0.999	0.940	0.73, 1.09

Where x is the vector of all the .fixed covariates (functional status, WHO clinical stage, regimen type, pre-TB status, body mass index) and $\boldsymbol{\delta}$ is the corresponding vector of the regression coefficient for the fixed covariates.

The non-PH model allows the effect of CD4 count varies with time periods. The coefficients $\boldsymbol{\delta}_1$ to $\boldsymbol{\delta}_4$ denote the interaction effect between CD4 count and time. The effect of CD4 count in the first half year is given by$\exp(\delta_b)$. The effect of CD4 count in the subsequent period are estimated by$\exp(\beta_1 + \delta_i)$; i = 1; 2; 3; 4.

The results are presented in Table 11.The result shows a significant effect of CD4 count below 2 years and a non-significant effect of CD4 count after 2 years. Hazard ratios in the first two time intervals are similar and hazard ratios for the last two time intervals are similar, we therefore separate the data into two time intervals. We use one Heaviside function g (t), where

$$g(t) = \begin{cases} 0 & if\ t \leq 2 \\ 1 & if\ t > 2 \end{cases}$$

Table 4.10: Cox models with time-dependent covariate

Covariates	Piecewise Cox non-PH model			
	β	HR	95%CI	P-value
Functional status				
Ambulatory	0.573	1.774	(1.053,2.988)	0.031
Bedridden	2.276	9.740	(4.799,1.001)	0.000
Who clinical stage				
Stage III	1.208	3.348	(1.279,8.764)	0.014
Stage IV	1.424	4.152	(1.502,11.473)	0.006
Regimen type	0.782	2.184	(1.238,3.850)	0.007
Pre-TB status	1.029	2.798	(1.827,4.287)	0.000
Body mass index				
18.5 <=BMI<24.9	-0.472	0.624	(0.407,0.958)	0.031
BMI>25	-1.995	0.136	(0.018,0.986)	0.048
Baseline CD4 count	-0.0129	0.987	(0.982,0.992)	0.000
Baseline CD4 count*g(t)	0.0122	1.012	(1.006,1.018)	0.000
-2loglikelihood	893.94			

Note: g(t) is the Heaviside function, which is zero when time is less than or equal to 2 years and 1 when time is greater than 2 years.

Table 4.11: Time-dependent effect of CD4 count in two time intervals

period	β	HR	95%CI	P-value
0-2 years	-0.0129	0.988	0.983,0.994	0.000
>2 years	-0.0007	0.999	0.996,1.002	0.576

The fitted model with time-dependent variable was :

$$h(t, x(t) = \exp(-0.0129CD4\ count + 0.0122CD4\ count \times g(t) + \alpha' x))$$

In this model, we define one cut point (2 years) on the time axis. The two hazard ratios are given by separately exponetiating each of the two estimated coefficients. When time is less than 2 years, $HR = \exp(-0.0129) = 0.988$.When time is greater than 2 years,$HR = \exp(-0.0129 + 0.0122) = \exp(-0.0007) = 0.999$.We can obtain that the 95% confidence interval for the two hazard ratios are (0.983-0.994) and (0.996-1.002) based on the fitted model. The 95% confidence interval in the first two years does not include one, which means the effect of CD4 count is statistically significant. The 95% confidence interval includes one after two years, which means that the effect of CD4 count is not statistically significant any more after two years. The estimated CD4 count effects are presented in Table 13.

4.4.AFT model

The AFT model which is another alternative of the Cox PH model when the PH assumption is violated can be used to express the magnitude of effect in a more accessible way in terms of difference between covariates in survival time. We fitted the dataset using exponential, Weibull, log-logistic, log-normal and gamma AFT model. The Q-Q plot is used to check the AFT assumption. The q-q plot appear to be approximately linear for both covariates regimen type and pre-TB positive test indicating that the AFT model may provide an appropriate fit for time-to-death data.

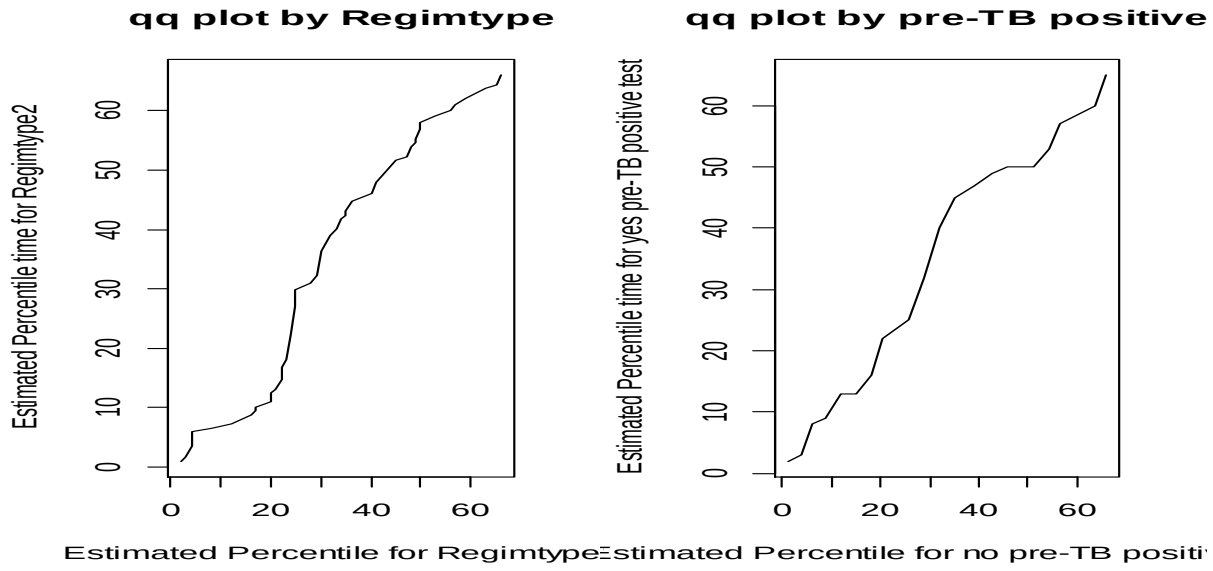


Figure 10: Q-Q plot to check the adequacy of the accelerated failure time model

Table 4.12: univariate analysis using Exponential Weibull , Lognormal and Log-Logistic AFT model.

	Exponential			Weibull			Log-normal			Log-logistic			Gamma		
	Coef.	TR	Pval	Coef.	TR	Pval	Coef.	TR	Pval	Coef.	TR	Pval	Coef.	TR	Pval
Covariates															
Functional status															
Ambulatory	-1.42	0.24	0.000	-2.18	0.12	0.000	-2.27	0.11	0.000	-2.18	0.112	0.000	-2.23	0.108	0.000
Bedridden	-3.89	0.020	0.000	-5.50	0.004	0.000	-5.74	0.003	0.000	-5.66	0.0034	0.000	-5.73	0.003	0.000
WHO stage															
stage II	-0.67	0.52	0.205	-1.18	0.30	0.216	-1.17	0.31	0.131	-1.14	0.317	0.202	-1.18	0.305	0.099
stage II	-1.89	0.150	0.000	-3.36	0.034	0.000	-3.34	0.035	0.000	-3.38	0.034	0.000	-3.31	0.036	0.000
stage IV	-2.62	0.073	0.000	-4.59	0.01	0.000	-4.57	0.010	0.000	-4.66	0.0094	0.000	-4.47	0.011	0.000
BaselineCD4	0.0062	1.006	0.000	0.010	1.011	0.000	0.0082	1.0082	0.000	0.008	1.008	0.000	0.005	1.005	0.000
Regimen type	-1.20	0.29	0.000	-2.05	0.13	0.000	-2.17	0.113	0.000	-2.16	0.1149	0.000	-1.06	0.344	0.000
Pre-TB status	-1.45	0.24	0.000	-2.56	0.078	0.000	-2.55	0.078	0.000	-2.60	0.0739	0.000	-2.49	-2.49	0.000
Body mass index															
18.5<=BMI<24.9	0.59	1.80	0.005	0.98	2.68	0.013	1.233	3.431	0.004	1.150	3.159	0.006	-0.04	0.958	0.693
BMI>=25	2.27	9.72	0.024	4.18	65.78	0.029	4.013	55.329	0.005	4.22	68.57	0.019	2.14	8.508	0.000

Significant variables in univariate analysis were considered in the multivariabe Lognormal and multivariable AFT model. Same set of six variables (functional status, WHO clinical stage, baseline CD4 count regimen type, pre-TB positive test and body mass index) came out at least marginally significant ($p \leq 0.20$) in univariate analysis for the model. Stepwise forward selection procedure with entry probability 0.05 and removal probability 0.051 was implemented in model. The coefficients, time Ratio (TR) of these variables and their p-values are given in Table 15.

Table 4.13: Multivariable analysis using Exponential, Weibull, and Lognormal and Log-Logistic AFT model

	Exponential			**Weibull**			**Log-normal**			**Log-logistic**			**gamma**		
Covariates	Coef.	TR	Pval	Coef	TR	Pval	Coef.	TR	Pval	Coef.	TR	Pval	Coef.	TR	pval
Functional status															
Ambulatory	-0.79	0.45	0.002	-1.10	0.33	0.006	-0.98	0.37	0.013	-1.023	0.36	0.009	-0.96	0.38	0.01
Bedridden	-3.25	0.038	0.000	-4.18	0.01	0.000	-4.29	0.01	0.000	-4.37	0.02	0.000	-4.27	0.02	0.00
WHO stage															
stage II	-0.45	0.636	0.400	-0.72	0.48	0.385	-0.53	0.590	0.414	-0.595	0.55	0.048	-0.515	0.59	0.42
stage III	-0.99	0.369	0.052	-1.56	0.21	0.051	-1.37	0.252	0.029	-1.36	0.26	0.031	-1.37	0.25	0.03
stage IV	-1.008	0.365	0.064	-1.74	0.17	0.042	-1.68	0.186	0.018	-1.64	0.19	0.003	-1.68	0.18	0.02
BaselineCD4	0.003	1.002	0.011	0.003	1.003	0.021	0.004	1.005	0.001	0.004	1.004	0.001	.0046	1.004	0.00
Regimen type	-0.74	0.477	0.011	-1.10	0.332	0.015	-1.35	0.258	0.001	-1.23	0.29	0.003	-1.36	0.25	0.00
Pre-TB status	-0.43	0.655	0.028	-0.66	0.517	0.0260	-0.63	0.531	0.024	-0.60	0.55	0.027	-0.63	0.53	0.02
Body mass index															
18.5<=BMI<24.9	0.59	1.817	0.011	0.65	1.925	0.048	0.60	1.84	0.036	0.56	1.74	0.038	0.614	1.84	0.02
BMI>=25	1.10	3.024	0.024	1.78	5.939	0.021	1.76	5.83	0.018	1.52	4.60	0.040	1.786	5.96	0.04
Constant	6.26			7.79			6.97			6.90			6.90		0.00
Logashape				-0.44			0.82			0.22			0.85	0.84	0.00
kappa													-0.080	-0.08	0.85
shape				0.65			2.275			1.25			2.32	2.32	

A covariate with a positive coefficient (i.e. time ratio above 1) implies that these variables prolong the time-to-death as they increase. Based on the multivariable AFT model, the acceleration factor for baseline CD4 count is greater than 1 indicating that high baseline CD4 count has the tendency to prolong the survival time. The time ratios for HIV positive patients under HAART whose body mass index is between 18.5 and less 24.9 and greater than 25 based on multivariate log-normal regression were greater than 1 ,indicating that HIV positive patients under HAART have prolonged the expected survival time as body mass index increase by one unit. For the covariates functional status, WHO stage, regimen type at ART initiation and pre-TB positive test the acceleration factor is less than 1, implies that an earlier death is more likely.

Table 4.14: The log-likelihoods and likelihood ratio (LR) tests, for comparing alternative AFT models for the survival time of HIV positive patients under HAART

	No of parameters	Loglikelihood	Testing against the Gamma distribution		
Distribution	m	L	LR	df	Prob > chi2
Exponential	1	-317.64516	166.22	2	0.0000
Weibull	2	-302.13173	140.45	1	0.0113
LogNormal	2	-297.08045	141.81	1	**0.8499**
Gamma	3	-297.06183			
Loglogistic	2	-299.74467	**Not nested**		

Table 4.15: Akaike Information Criterion (AIC) in the AFT models

Distribution	**ll(nul)**	**ll(model)**	**df**	**k**	**c**	**AIC= -2l+2(k+c)**	**BIC**
Exponential	-400.756	-317.6452	11	6	1	657.2903	702.2448
Weibull	-372.3569	-302.1317	12	6	2	628.2635	677.3048
Log Normal	-367.9857	-297.0804	12	6	2	**618.1609**	**667.2022**
Gamma	-353.2621	-297.0618	13	6	3	620.1237	673.2517
Log logistic	-371.77	-299.7447	12	6	2	623.4893	672.5306

We compared all these AFT models using statistical criteria (likelihood ratio test and AIC). The nested AFT models can be compared using the likelihood ratio (LR) test. The exponential, the Weibull and the log-normal models are nested within the model (Table 16).According to the LR test; the log-normal model fits better. However, the LR test is not valid for comparing models that are not nested. In this case, we use AIC to compare the models (Table 17). A model with smaller value of *AIC* can be considered as a better model compared to other models under consideration. The computed value of AIC for Lognormal AFT model is 618.1609. On the basis of AIC, Lognormal AFT model appears to be an appropriate AFT model according to AIC compared with other AFT models.

Table 4.16: The fitted multivariable lognormal AFT model with its coefficients and time ratio

Covariates	**Coef.**	**TR**	**Pval**	**95%CI**	
Functional status					
Ambulatory	-0.98	0.37	0.013	(-1.747753,	-.2098955)
Bedridden	-4.29	0.01	0.000	(-5.527208,	-3.070496)
WHO stage					
stage II	-0.53	0.590	0.414	(-1.790285,	0.7374195)
stage III	-1.37	0.252	0.029	(-2.611083,	-0.144734)
stage IV	-1.68	0.186	0.018	(-3.077793,	-0.2864098)
BaselineCD4	0.004	1.004	0.001	(0.0018,	0.0073078)
Regimen type	-1.35	0.258	0.001	(-2.156517,	-0.5489147)
Pre-TB status	-0.63	0.531	0.024	(-1.37324,	-0.1076674)
Body mass index					
18.5<=BMI<24.9	0.60	1.84	0.036	(0.347705,	8709628)
BMI>=25	1.76	5.83	0.012	(1..022293,	4.149807)
Constant	6.970		0.000	(5.426825,	8.514444)
Log scale	0.822		0.000	(0.662972,	.9816984)
scale	2.275			(1.940552,	2.668985)

Furthermore, we check the goodness of fit of the model using residual plots. The cumulative hazard plot of the Cox-Snell residuals in log normal model is presented in Figure 7. The plotted points lie on a line that has a unit slope and zero intercept. The graph could be considered roughly linear through the origin so there is no reason to doubt the suitability of this fitted lognormal model. At last, we conclude that the lognormal AFT model fitted the data better than other distribution in multivariablee analysis based on AIC criteria and residuals plot.

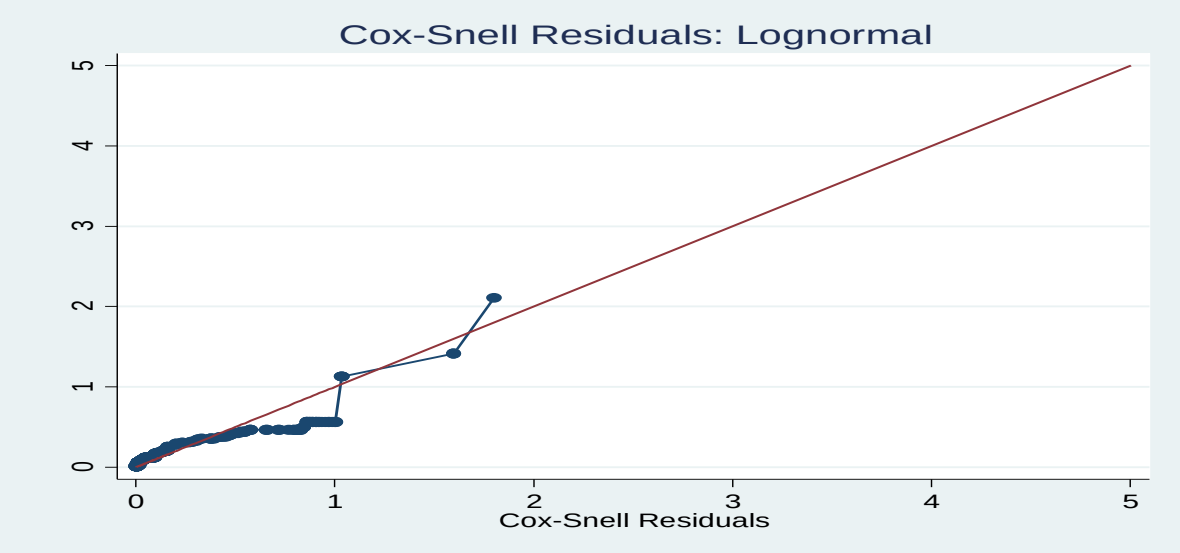


Figure 11: cumulative hazard plot of the Cox-Snell for the fitted log normal AFT model

Model Interpretation

According to the log linear form of the model, the Lognormal AFT model fitted can be represented by the following equation where T_i is the random variable associated with the survival time of the i^{th} .patient.

$$\log T_i = 6.970 - 0.98\ \text{Ambulatory} - 4.2\ 9\ \text{Bedridden} - 1.37\ \text{stageIII} - 1.68\ \text{stageIV} + 0.004\ \text{BCD4count} - 1.35\ \text{Regimtype} - 0.63\ \text{PreTB} + 0.6\ BMI_2 + 1.7\ 6\ BMI_3 + 2.75\ \varepsilon_i$$

Under the fitted multivariable lognormal AFT model, the estimated acceleration factor for an individual in ambulatory group and bedridden functional status group relative to an individual in working functional status group is 0.37 and 0.01 respectively. This indicates that the effect of ambulatory and bedridden functional status groups accelerates the time to death of patients under HAART (that is minimum survival time).

Similarly the estimated acceleration factor for an individual in stage III and stage IV of WHO clinical stage group relative to an individual in stage I WHO clinical stage group is 0.252 and 0.186 respectively. This indicates that the effect of stage III and stage IV of WHO clinical stage accelerates the time to death of patients under HAART as the stage of the disease increase.

The estimated time ratio for Baseline CD4 count is above 1 [TR $= e^{0.004} = 1.004$] implies that these variables improve the estimated survival time by 0.4% as one unit increase in CD4 count. Regimen type at HAART initiation TDF-3TC-EFV was associated with shorter survival time (TR $= e^{-1.35} = 0.258$); patients who initiated AZT-3TC-NVP have longer survival time compared to TDF-3TC-EFV.

The acceleration factor for pre-TB positive test is (TR $= e^{-0.63} = 0.581$) which indicates that shorter survival time are more likely for patients who have pre-TB positive test. That means it was associated with accelerates time- to –death.

The estimated acceleration factor results also show that for HIV positive patients who have normal BMI, the survival time is accelerated 1.84 times (TR $= e^{0.60}$) as compared to patients that are underweight, similarly patients that were overweight their survival time is accelerated 5.83 times (TR $= e^{1.76}$) as compared to patients that are underweight. That means the patient with BMI less than 18.5 has shorter Survival time than patients with normal BMI, but the patients with BMI above 25 has longer survival time than patients with BMI with normal BMI.

Table 4.18: Standardized variability and AIC for Cox and Parametric models of time to death in multivariable Analysis

Covariates	Cox		Exponential		Weibull		Log-normal		Log-logistic		Gamma	
	Coef.	SV	Coef.	SV	Coef.	SV	Coef.	SV	Coef.	SV	Coef.	SV
Functional												
Ambulatory	0.63	0.28	0.83	0.32	0.75	0.35	-1.04	0.394	-1.12	0.372	-1.02	0.408
Bedridden	2.42	0.15	3.34	0.11	2.75	0.13	-4.49	0.146	-4.57	0.129	-4.43	0.145
WHO stage							-					
stage III	1.137	0.49	1.16	0.48	1.17	0.47	1.64	0.424	-1.64	0.476	-1.62	0.412
stage IV	1.40	0.42	1.22	0.48	1.35	0.43	-2.00	0.388	-1.97	0.430	-1.98	0.382
BaselineCD4	-0.001	0.50	-0.002	0.39	-0.002	0.44	0.004	0.311	0.004	0.320	0.004	0.309
Regimen type	0.917	0.34	0.95	0.34	0.920	0.34	-1.69	0.269	-1.58	0.299	-1.72	0.261
Pre-TB status	0.315	0.79	0.34	0.75	0.34	0.583	-0.58	0.680	-0.53	0.744	-0.58	0.672
Body mass index												
18.5<=BMI<24.9	-0.414	0.55	-0.67	0.36	-0.48	0.48	0.73	0.498	0.68	0.526	0.74	0.490
BMI>=25	-1.028	0.99	-0.94	1.01	-1.01	1.02	1.68	0.758	1.41	0.97	1.74	0.716
AIC	**892.32**		657.2903		628.2635		**618.1609**		623.4893		620.1237	

As the scales of the parameters in Cox's model and in the parametric models differ, neither parameter estimates nor their estimated variances are suitable for comparisons. The efficiency of parameter estimates can be better compared using standardized measure of variability of the parameters in the model. Hence, the SV-values of the covariate can be compared across different models. In the univariate analysis of covariates, (table 19) the efficiencies of parametric and semi-parametric were more or less similar. According to table 20, the lognormal model seemed to be relatively efficient model for estimating the covariate over Cox PH model as it yielded the smallest SV values over the other models for these variables. However, a close investigations of tables 10 and 20 show that the difference in SV value is not that much large to conclude that one model is more efficient over the others. It seems that there may not be a single model that is substantially better than others in terms of efficiency. Oakes conferred that asymptotically parametric models are well fitted if parameter values deviate far from zero (Oakes, 1977). Parameter estimates from our study also significantly far from zero while considering all six covariates. But, a model which has smaller AIC value is preferred. Hence, the lognormal AFT model fits the data better than cox PH.

5. DISCUSION

Researchers in the field of medical sciences often tend to prefer Cox proportional hazard model more than parametric models (Efron B, 1977; Oakes D, 1977). This is probably due to the fact that this model allows us to estimate and make inference about the parameters without assuming any distribution for the lifetime, whose distribution is often unknown. However, it does have the requirement of proportional hazards, which is not always satisfied by the data (Atman *et al*, 1982). If this assumption does not hold there are various solutions to consider. One solution is to include the time-dependent variable for the predictors with non-proportional hazards. When this approach is used to account for a variable with non-proportionality, different results may be obtained from different choices of time-dependent variables. It is hard to choose between models. Alternatively we can use a model where we stratify on the non-proportional predictors. The stratified Cox model is not appropriate when the covariate with non-proportionality is continuous or of direct interests. And both ways are still based on comparison of hazards.

In these situations, AFT models such as lognormal, log logistic, Weibull and Exponential provide an alternative method to fit survival data even when hazards are not proportional.
Moreover, under these models we measure the direct effect of the explanatory variables on the survival time and not on a conditional probability, as we do in the Cox regression model. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean lifetime. Parametric models are, therefore, based on a specific distribution for duration times without need to proportional hazard assumptions (Bradburn MJ *et al*, 2003; Pourhoseingholi MA, *et al.*2008).When a suitable distribution can be found the parametric model is more informative than the Cox model. It is straightforward to derive the hazard function and to obtain predicted survival times when using a parametric model, which is not the case in the Cox proportional hazard model.

In addition, the appropriate use of these models should lead to more efficient parameter estimates than the cox model (Bradburn MJ *et al*, 2003).The Cox model expresses the multiplicative effect of covariates on the hazard whereas the AFT model expresses the multiplicative effect of covariates on survival times. The results from an AFT model are easier to interpret, more relevant to clinicians and provide a more appropriate description of survival data in many situations (Bradburn *et al*, 2003; Pourhoseingholi, *et al.*2008).

Hence researcher may use the parametric survival model as an alternative to the semi parametric counterparts (Cox PH model) when the proportional hazard assumption does not hold.

The main aim of the study was modeling time-to-death of HIV positive patients under highly active antiretroviral therapy (HAART) using Cox PHs and AFT models. The comparison of distributions of the models was done using the AIC criteria, where a model with minimum AIC is accepted to be the best (Mohamad *et al*, 2007), the data strongly supported the log normal regression among parametric models in univariate analysis and it can be lead to more precise results as an alternative for Cox. (Vallinayagam *et al.,* 2014) compared the performance of the common parametric models including the, Exponential, Weibull, Gompertz, Lognormal and Log-logistic using Breast Cancer data. Their study revealed that Log-normal model is better than other models.

The baseline covariates explored in the thesis were sex, age, marital status, education level, functional status, place of residence, weight, CD4 count, WHO clinical stage, regimen type, pre-TB positive test and body mass index. Both multivariate Cox PH and lognormal AFT model fitted to HIV HAART data shows that functional status, base line CD4 count, WHO clinical stage III &IV, regimen type, Pre-TB positive test and Body mass index were covariates that significantly influences the survival time of HIV positive patients under HAART at 5% significance levels. For multivariate analyses sex, age, marital status, education level, place of residence and weight was not an important predictor of survival. After fitting the multivariate Cox PH model, PH assumptions and the goodness of fit of the model was assessed through residual plots, the results shows that the PH assumption does not hold for covariate CD4 count and the residual plots seems to display a systematic deviation of fit of the model. A strategy for analyzing of non-proportional hazards was considered and the time dependent effects of covariate CD4 count shows that a varying effect at a different time intervals.

In contrast, based on AIC criteria and residuals plot Lognormal AFT model (Table 18) among parametric models in univariate and multivariate analysis it can be lead to more precise results as an alternative for Cox and provides an adequate description of the data.

In this study Functional status during ART initiation was significant predictor of death. This finding is consistent with many studies done in Ethiopia (Amuron *et al*., 2011) which showed that the risk of death among working patients is lowered by 55% than bedridden patients during ART initiation, similar study by (Mageda *et al*., 2012) revealed that ambulatory and bedridden functional status is 2.87 and 6.90 times at risk of death than the working status. Consistent findings by (Abebe *et al*, 2014) showed that Patients in ambulatory functional status and bedridden were at increased hazard rate of death by 2.72 and 2.38 times than patients in working functional status, respectively. This might be due to the fact that bedridden patients have been long in bed so that there clinical characteristics are deteriorating and the disease progression has implicated here with low CD4 count.

This study showed that that the two advanced WHO clinical stages III and IV were found to be significant predictors of death among HIV-patients on ART. This finding is consistent with the findings obtained from most studies carried out on patients with HIV positive patients under HAART. Findings by (Ashenif *et al*, 2010) showed that the significant predictors of mortality were advanced WHO stage. Similar study findings by (Worku and San Sebastian, 2009; David Jr., 1999; May *et al*, 2010).showed that Patients with advanced clinical diseases (WHO stage III or IV) had higher mortality compared to patients with WHO stage I or II. Therefore, the significant effects of these conditions of patients on mortality indicate that patients died mostly because of their late initiation of ART when they had the worst health conditions.

This study confirms that low baseline CD4 count in ART patients is a predictor for the progression to death. Similar results have been demonstrated with other studies, (Rafera,2012) where he asserts that the rate of dying among patients with higher CD4 cell count in Ethiopia is proportionally lower compared to patients with lower CD4 count). This finding was similar to the result of studies by (Pierre De Beaudrap *et al*, 2008) which revealed that during the early period, baseline CD4 count was significant predictors of mortality 0.80 [0.69-0.92] per 200 cell/mm^3.many authors show that the survival depends on the level of CD4 cell count at the beginning of treatment (Ashenif *et al*, 2010, Moore *et al*., 2006).

In our study, treatment regimen has been associated with mortality. In sub-Saharan African study, there was little evidence for an association between excess mortality and treatment regimen (Brinkhof, 2009) compared with EFV, the adjusted HR for all-cause mortality for NVP was 2.28 during the first 6 months, which decreased to 1.31 after 6 month. Inconsistent findings by (Mzileni, 2008) showed that the mortality rates did not vary (p >0.05) between the 3 ART regimens: 8.8% (n = 184/2083) in NRTI (1a regimen), 8.9% (n = 7/78), and 8% (n = 2/26) in regimen 2(3rd NRTI)

This study arrived at pre-TB positive test had a significant impact on the survival of HIV positive patients on ART. According to (Qingxia Z, *et al*, 2008), TB co-infection was found to be a cause of high mortality of HIV patients. A consistent findings has been reported by (Kalyango J *et al*, 2009) which concluded that TB is the leading cause of death worldwide, and the virulence of the mycobacterium increases in HIV infected patients, where the host's immune system is suppressed, enabling it to establish infection very easily, similarly in this study patients who were experienced pre-TB positive test were at advanced disease stage. Moreover patients in advanced clinical stages are prone for TB infection (Mageda *et al*., 2012; Anglaret X *et al*, 2012; Tsehaineh, 2010; David Jr., 1999)

In this study body mass index was a significant predictor of death for patients on ART. This finding was similar to the result of studies by (Pierre De Beaudrap *et al*, 2008) the study found out that during the early period, baseline body mass index was significant predictors of mortality (Hazard Ratios 0.82 [0.72-0.93].Consistent findings by (Van der Sande *et al*, 2004) showed that the mortality hazard ratio (HR) of those with a baseline BMI <18 compared with those with a baseline BMI >=18 was 3.4 (95% CI, 3.0–3.9). The median survival time of those presenting with a BMI <16 was 0.8 years, in contrast to a median survival of 8.9 years for those with a baseline BMI >=22.Baseline BMI<18 remained a highly significant independent predictor of mortality after adjustment for other covariates in the study. A similar finding by (Damtew *et al*, 2014) which revealed that the risk of death in patients with a BMI<18.5kg/m^2 was more than two times higher compared to those with a BMI≥18.5kg/m^2 . BMI is an indicator of patient nutritional status but may also be influenced by late-stage AIDS conditions, such as wasting syndrome and opportunistic infections, or by progression of the HIV itself (Togun T *et al*, 2004).

6. CONCLUSION AND RECOMMENDATIONS

6.1.CONCLUSION

This study was based on a data set on time-to-death of HIV positive patients who were aged 18 years or above placed under HAART in JUSH with an aim of investigating the comparative performance of Cox and parametric models in a survival analysis of time-to death data. We used AIC and standardized variability of the coefficients for covariates in the models to evaluate the performance among models.

Using the selected model Functional status, WHO clinical disease stage, pre-TB positive test, regimen type at ART initiation, body mass index and baseline CD4 count were found to be significantly associated with survival time of patients on HAART. Patents' baseline clinical characteristics; ambulatory and bedridden functional status, WHO stage III and IV and the presence of pre-TB positive test were covariates that accelerate the time to death of patients on HAART. The time to death was shorter for regimen type I patients at ART initiation. Underweight patients are prone to have shorter survival time.

According to the Cox model with time-dependent variable, the predictive effect of baseline CD4 count clearly changes at about 2 years. Before 2 years, the risk death decreases as baseline CD4 count increases. After 2 years, the risk of death increases as baseline CD4 count increases. According to the lognormal AFT model, baseline CD4 count prolongs the time to death as it increase.

In univariate and multivariable analysis of covariates the efficiencies of parametric and semi-parametric were more or less similar. It seems that there may not be a single model that is substantially better than others in terms of efficiency

The result revealed that the lognormal AFT model provided a better fit to the studied data than the Cox proportional hazards model. Hence, it is better for researchers of HIV patients on HAART to consider AFT model when proportionality assumption of the Cox model is not satisfied.

6.2.Recommendation

- Based on the finding of the study it is necessary that a careful monitoring of patients with low baseline CD4 cell count, advanced WHO stage III and IV, presence of pre-TB positive test, ambulatory and bedridden functional status and being underweight must be undertaken in order to improve the survival of HIV positive patients under HAART initiation.
- There is a need of further study on the effect of TB treatment on AIDS progression and survival of HIV –infected adults and considering type of regimen at ART initiation for patients in impaired clinical status.
- The PH model is routinely applied to the analysis of survival data. The study considered here provides a situation where PH model is not hold. The result revealed that the lognormal AFT model provided a better fit to the studied data than the Cox proportional hazards model. To this content the AFT model may have explanatory advantage in that covariates have a direct effect on survival times rather on hazard functions as in the PH model. Hence, it is better for researchers of HIV patients on ART to consider AFT model as an alternative when proportionality assumption of the Cox model is not satisfied.

7. REFERENCES

Abebe, Yigeremu, Ab Schaap, Girmatchew Mamo, Asheber Negussie, Birke Darimo, and Dawit Wolday. "HIV prevalence in 72 000 urban and rural male army recruits, Ethiopia." *Aids* **17**, no. 12 (2003): 1835-1840.

Akaike, Hirotugu. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on* **19**, no. 6 (1974): 716-723.

Alemu, Andinet Worku, and Miguel San Sebasti. "Determinants of survival in adult HIV patients on antiretroviral therapy in Oromiyaa, Ethiopia." *Global Health Action* **3** (2010).

Amuha, Monica G., Paul Kutyabami, Freddy E. Kitutu, Richard Odoi-Adome, and Joan N. Kalyango. "Non-adherence to anti-TB drugs among TB/HIV co-infected patients in Mbarara Hospital Uganda: Prevalence and associated factors." *African health sciences* **9**, no. 2 (2009).

Amuron, Barbara, Jonathan Levin, Josephine Birunghi, Geoffrey Namara, Alex Coutinho, Heiner Grosskurth, and Shabbar Jaffar. "Mortality in an antiretroviral therapy programme in Jinja, south-east Uganda: a prospective cohort study."*AIDS Res Ther* **8**, no. 1 (2011): 39.

Andersen, P. K., Borgan, O., Gaill, R. D., and Keiding, N. Statistical Methods Based on Counting Processes. Springer, New York, 1993.

Anglaret, Xavier, Albert Minga, Delphine Gabillard, Timothée Ouassa, Eugene Messou, Brandon Morris,Moussa Traore et al. "AIDS and non-AIDS morbidity and mortality across the spectrum of CD4 cell counts in HIV-infected adults before starting antiretroviral therapy in Cote d'Ivoire." *Clinical infectious diseases* **54**, no. 5 (2012): 714-723.

Atman DG, De Stavola BL, Love SB, et al. Review of survival analysis published in cancer journals. British Journal of Cancer **72** (1985), 511.518.

Bradburn MJ, Clark TG, Love SB, et al. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit.British Journal of Cancer (2003) **89**, 605 – 611

Brinkhof M W G, Boulle A, Weigel R, Messou E, Mathers C,Orrell C, Dabis F, Pascoe M, Egger M and for the International epidemiological Databases to Evaluate AIDS (IeDEA). Mortality of HIV-Infected Patients Starting Antiretroviral Therapy in Sub- Saharan Africa: Comparison with HIV-Unrelated Mortality. PLoS Med. 2009; 6 (4): e1000066.

Cain, K. C., and Lange, N. T. Approximate case influence for the proportional hazards regression model with censored data. Biometrika **40** (1984), 493-499.

Carpenter, Charles CJ, Margaret A. Fischl, Scott M. Hammer, Martin S. Hirsch, Donna M. Jacobsen, David A. Katzenstein, Julio SG Montaner et al. "Antiretroviral therapy for HIV infection in 1997:

updated recommendations of the International AIDS Society—USA Panel." *Jama* **277**, no. 24 (1997): 1962-1969.

Collett, D. Modelling Survival Data in Medical Research. Chapman and Hall, London, 2003.

Cox DR. Regression models and life-table. *Journal of Royal Statistical Society B.* 1972; 34: 187-220

Cox, D. R., and Oakes, D. Analysis of Survival Data. Chapman and Hall, London, 1984.

Cox, D. R., and Snell, E. J. A general definition of residuals with discussion. Journal of the Royal Statistical Society. Series B **30** (1968), 248.275

Crowder, J., and Hu, M. Covariance analysis of heart transplant survival data. Journal of American Statistical Association **78** (1977), 27.36.

Damtew, Bereket, Bezatu Mengistie, and Tadesse Alemayehu. "Survival and Determinants of Mortality in Adult HIV/AIDS Patients Initiating Antiretroviral Therapy in Somali Region, Eastern Ethiopia." *J AIDS Clin Res* **5**, no. 327 (2014): 2.

De Beaudrap, Pierre, Jean-François Etard, René Ecochard, Assane Diouf, Allé Baba Dieng, Vannina Cilote, Ibrahima Ndiaye et al. "Change over time of mortality predictors after HAART initiation in a Senegalese cohort." *European journal of epidemiology* **23**, no. 3 (2008): 227-234.

Efron B. The efficiency of Cox's likelihood function for censored data.1977; 72: 557-65.

Federal HIV/AIDS Prevention and Control Office (HAPCO). Country Progress Report on HIV/AIDS Response, Federal Democratic Republic of Ethiopia. 2012.

Hernán, Miguel A., Stephen R. Cole, Joseph Margolick, Mardge Cohen, and James M. Robins. "Structural accelerated failure time models for survival analysis in studies with time-varying treatments." *Pharmacoepidemiology and drug safety* **14**, no. 7 (2005): 477-491.

Hosmer DW, Lemeshow S. Regression modeling of time-to-event data. Wiley, NewYork. 1999..

Kalbfleisch, J. D., and Prentice, R. L. Marginal likelihoods based on Cox.s regression and life model. Biometrika **60** (1973), 267.278.

Klein, J. P., and Moeschberger, M. L. Survival Analysis: Techniques for Cen-sored and Truncated Data. Springer, New York, 1997.

Kleinbaum D, Klein M. Survival analysis: A self- learning text, New York, Springer-Verlag. 2005.

Mageda, Kihulya, Germana Henry Leyna, and Elia John Mmbaga. "High Initial HIV/AIDS-related mortality and-its predictors among patients on antiretroviral therapy in the Kagera Region of Tanzania: a five-year retrospective cohort study." *AIDS research and treatment* 2012 (2012).

Modi, Manoj, Boskey Pancholi, Shweta Kulshrestha, Ajay Kumar S. Rawat, Swadesh Malhotra, and Satish K. Gupta. "Anti-HIV-1 activity, protease inhibition and safety profile of extracts prepared from Rhus parviflora." *BMC complementary and alternative medicine* **13**, no. 1 (2010): 158

Pourhoseingholi MA, Moghimi-Dehkordi1 B, Safaee A, et *al.* Statistical comparison of survival models for analysis of cancer data. Asian Pacific J Cancer Prev. 2008; **9**: 417-420.

Moran JL, Bersten AD, Solomon PJ, et al. Modeling survival in acute severe illness: Cox versus accelerated failure time models. Journal of Evaluation in Clinical Practice. 2008;14: 83-89

Mzileni M O, Longo-Mbenza B and Chephe T J. Mortality and causes of death in HIV-positive patients receiving antiretroviral therapy at Tshepang Clinic in Doctor George Mukhari Hospital. Pol Arch Med Wewnetrznej. 2008; 118 (10): 548-554

Nawumbeni, Derek Ngbandor, Albert Luguterah, and Timothy Adampah. "Performance of Cox Proportional Hazard and Accelerated Failure Time Models in the Analysis of HIV/TB Co-infection Survival Data." *Research on Humanities and Social Sciences* **4**, no. 21 (2014): 94-102.

Oakes D. The asymptotic information in censored survival data. *Biometrika.* 1977, 64, 441-8

Odafe S., Idoko O., Badru T., Aiyenigba B., Suzuki C., Khamofu H., Onyekwena O., Okechukwu E., Torpey K. and Chabikuli1 O. N. Patients' demographic and clinical characteristics and level of care associated with lost to follow-up and mortality in adult patients on first-line ART in Nigerian hospitals. 2012; 15: 17424.

Office of AIDS Research Advisory Council (OARAC). Guidelines for the use of antiretroviral agents in HIV-1 Infected Adults and Adolescents.2008.

Peto, Richard, and Julian Peto. "Asymptotically efficient rank invariant test procedures." *Journal of the Royal Statistical Society. Series A (General)*(1972): 185-207.

Ponnuraja C, Venkatesan P. Survival models for exploring tuberculosis clinical trial data an empirical comparison. Indian Journal of Science and Technology. 2010; 3: 755-58.

Pourhoseingholi, Mohamad Amin, Ebrahim Hajizadeh, Bijan Moghimi Dehkordi, Azadeh Safaee, Alireza Abadi, and Mohammad Reza Zali. "Comparing Cox regression and parametric models for survival of patients with gastric carcinoma." *Asian Pacific Journal of Cancer Prevention* 8, no. 3 (2007): 412.

Qi, Jiezhi. "Comparison of proportional hazards and accelerated failure time models." PhD diss., University of Saskatchewan Saskatoon, 2009.

Refera, H., Survival and risk factors of HIV/TB Co-infected Patients under Antiretroviral Therapy in Ambo Hospital, Ethiopia. Addis Ababa University, School of Graduate Studies Department of Statistics, 2012.

Schoenfeld, D. Partial residuals for the proportional hazards regression model. Biometrika **69** (1982), 239.241.

Schoenfeld, David A. "Sample-size formula for the proportional-hazards regression model." *Biometrics* (1983): 499-503.

Schoenfeld, David. "The asymptotic properties of nonparametric tests for comparing survival distributions." *Biometrika* **68**, no. 1 (1981): 316-319.

Sethi, Ajay K., and Stephen J. Gange. "Parametric Models for Studying time to antiretroviral resistance associated with illicit Drug use." *Wisconsin Medical Journal* **108**, no. 5 (2009): 266.

Tadele, Ashenif, Ashenafi Shumey, and Neway Hiruy. "Survival and Predictors of Mortality among Adult Patients on Highly Active Antiretroviral Therapy at Debre-Markos Referral Hospital, North West Ethiopia; A Retrospective Cohort Study." *Journal of AIDS & Clinical Research* (2014).

Therneau T, Grambsch P. Modeling survival data: Extending the Cox Model, New York, Springer-Verlag. 2000.

Tsehaineh, B. "Assessment of factors associated with high risk of mortality of hiv patients treated with highly active antiretrovairal therapy in Jimma Zone, South western Ethiopia: application of survival analysis methods." *Addis Ababa University* (2010).

UNAIDS/WHO. AIDS epidemic update UNAIDS 20 avenues Appia CH-1211 Geneva 2 Switzerland, 2009.

USAID. HIV/AIDS Health Profile for Ethiopia. AA. Ethiopia Wikipedia, the free encyclopedia (2008). List of countries by HIV/AIDS Adult prevalence rate. 2010; Accessed on 30/05/10

Vallinayagam, V., Prathap, S., and Venkatesan, P. Parametric Regression Models in the Analysis of Breast Cancer Survival Data. *International Journal of Science and Technology*, 3(3) 2049-7318.,2014

Van der Sande, Marianne AB, Maarten F. Schim van der Loeff, Akum A. Aveika, Saihou Sabally, Toyin

Togun, Ramu Sarge-Njie, Abraham S. Alabi, Assan Jaye, Tumani Corrah, and Hilton C. Whittle. "Body mass index at time of HIV diagnosis: a strong and independent predictor of survival." *JAIDS Journal of Acquired Immune Deficiency Syndromes* **37**, no. 2 (2004): 1288-1294.

Wei, L. J. "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis." *Statistics in medicine* **11**, no. 14-15 (1992): 1871-1879.

World Health Organization Statistics, 2014

Worku, Andinet. *Pattern and determinants of survival in adult HIV patients on antiretroviral therapy, Ethiopia*. Umeå International School of Public Health, 2009.

Xueyan, Jiang, Lu Hongzhou, Zhang Yuexin, Zhou Zengquan, Ye Hanhui, Zhao Qingxia, Wang Hui et al. "A cross-sectional study of HIV and tuberculosis coinfection cases in mainland China." *Southern medical journal* **101**, no. 9 (2008): 914.

Zeng, Donglin, and D. Y. Lin. "Efficient estimation for the accelerated failure time model." *Journal of the American Statistical Association* **102**, no. 480 (2007): 1387-1396.

8. Appendices

ANNEX 1: THE KAPLAN-MEIER SURVIVAL FUNCTION ESTIMATES

Figure 1 Kaplan-Meier survival curve of 440 HIV-positive patients on ART.

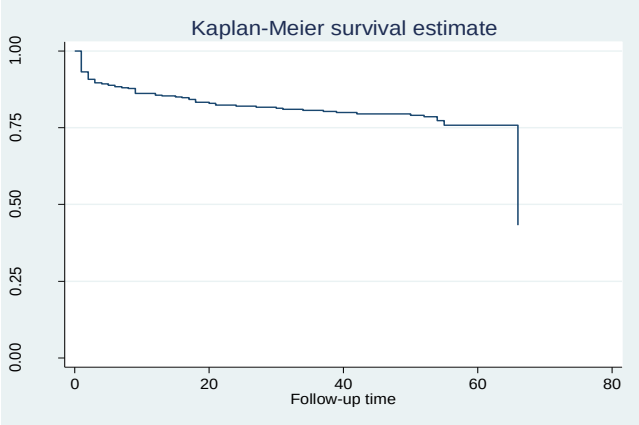
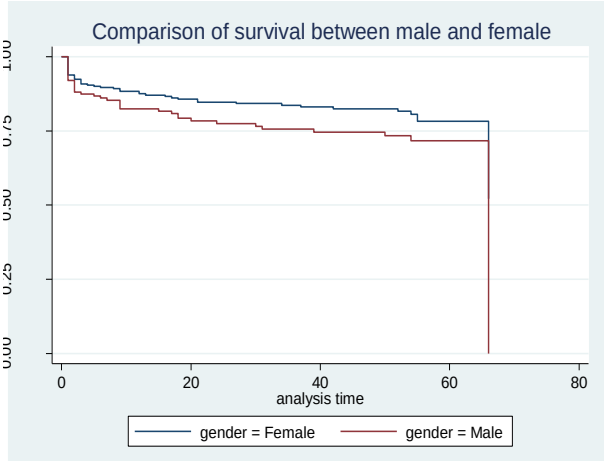
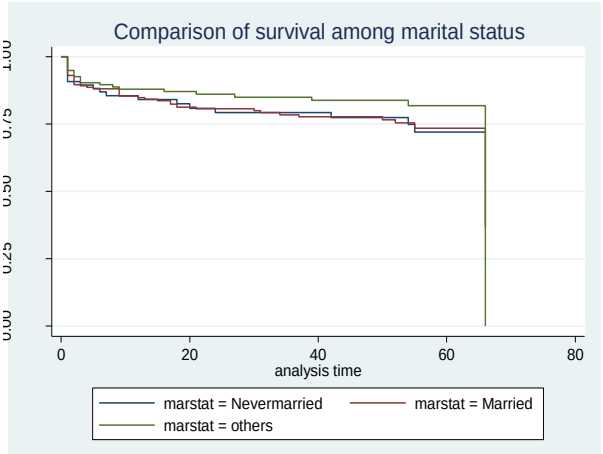
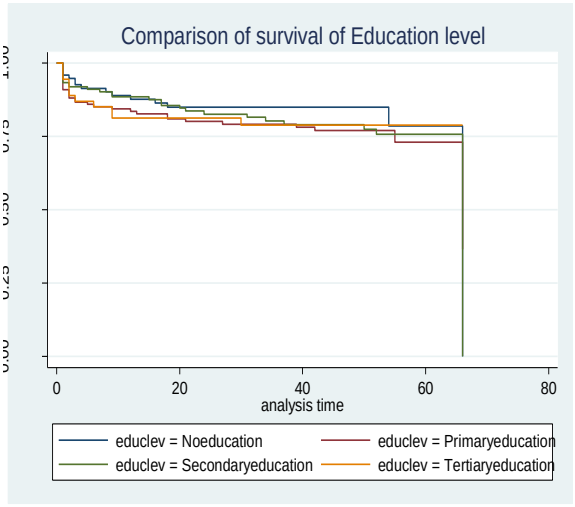


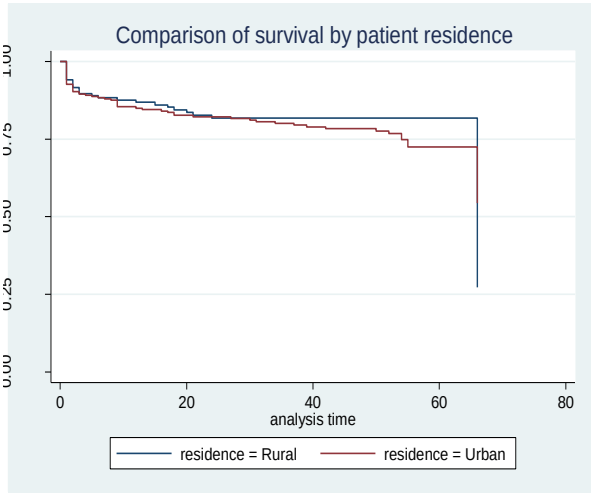
Figure 3: Kaplan-Meier survival curve of HIV-positive patients on ART for categorical covariate

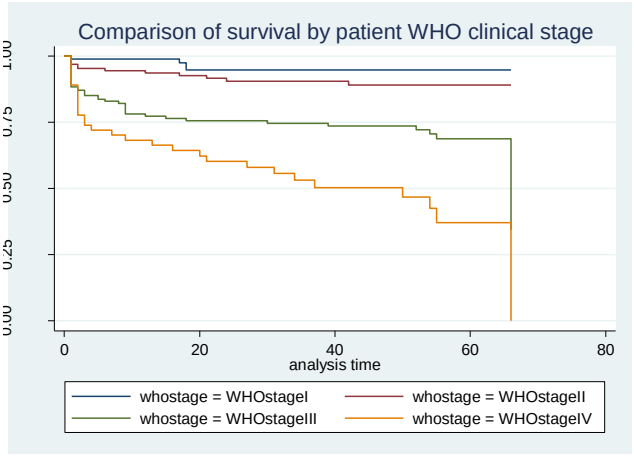


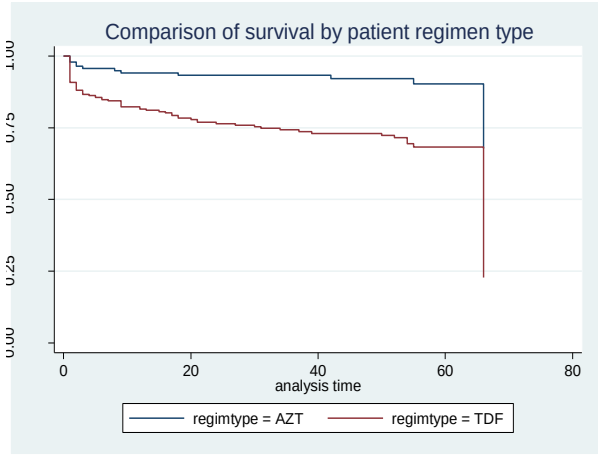


.


Comparison of survival of Education level
0.00
0.25
0.50
0.75
1.00
0
20
40
60
80
analysis time
educlev = Noeducation
educlev = Primaryeducation
educlev = Secondaryeducation
educlev = Tertiaryeducation


Comparison of survival by patient residence
0.00
0.25
0.50
0.75
1.00
0
20
40
60
80
analysis time
residence = Rural
residence = Urban


Comparison of survival by patient WHO clinical stage
0.00
0.25
0.50
0.75
1.00
0
20
40
60
80
analysis time
whostage = WHOstageI
whostage = WHOstageII
whostage = WHOstageIII
whostage = WHOstageIV


Comparison of survival by patient regimen type
0.00
0.25
0.50
0.75
1.00
0
20
40
60
80
analysis time
regimtype = AZT
regimtype = TDF

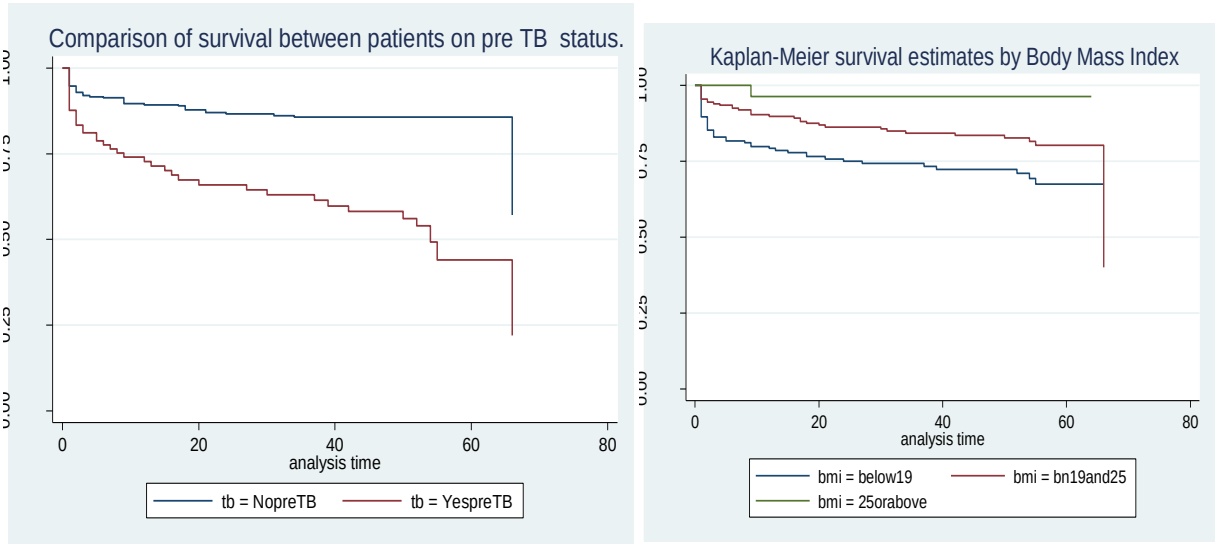
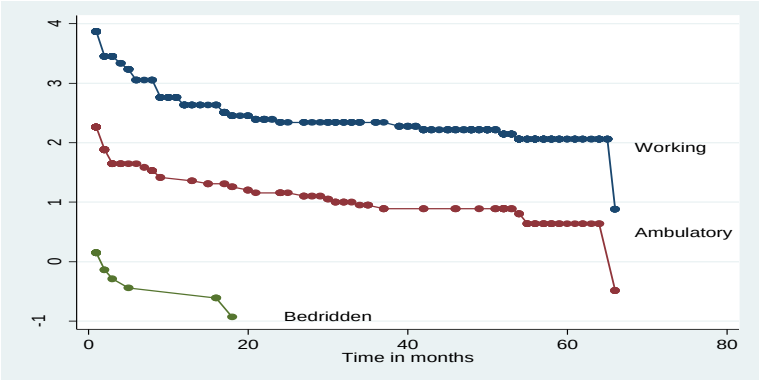
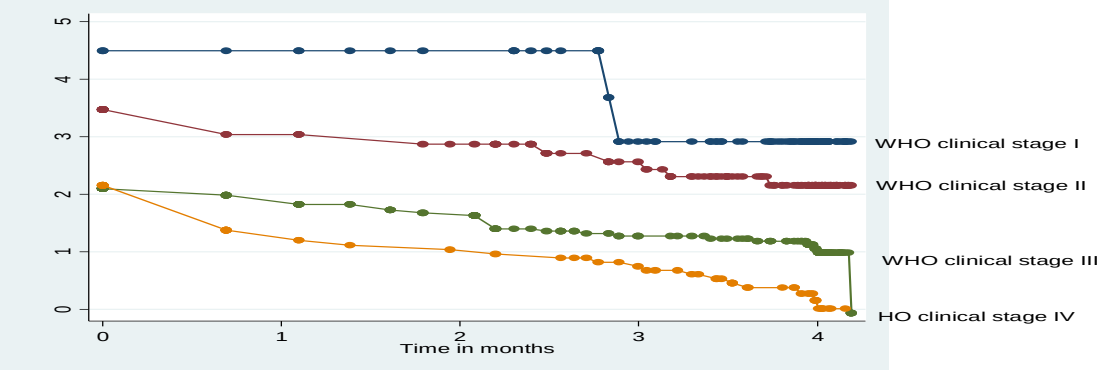


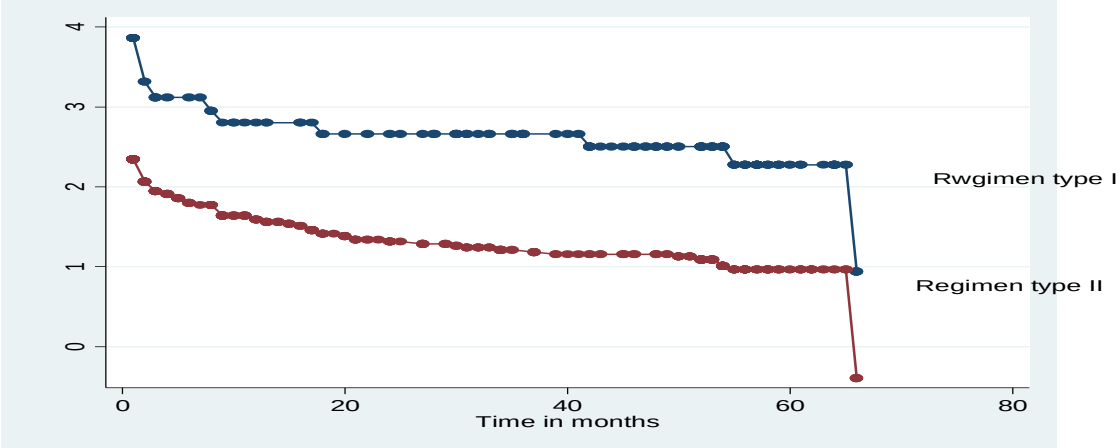
Figure 4: plot of log (-log (survival)) versus survival time for categorical variables to check the PH assumption for all the categorical variables



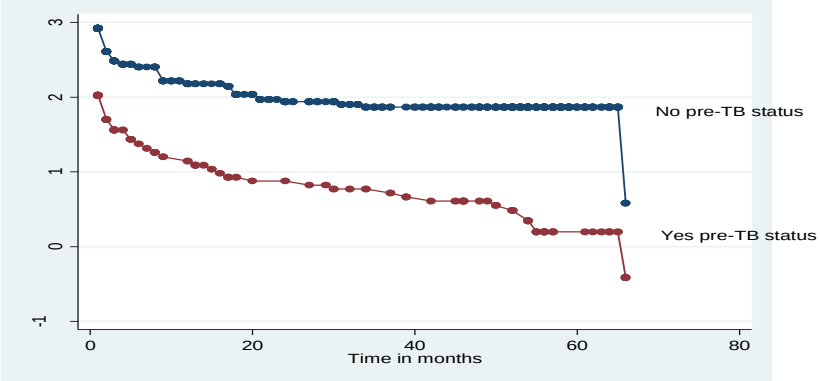
Plot of –log(-log(survival)) versus follow up time to assess the proportionality assumption for functional status.



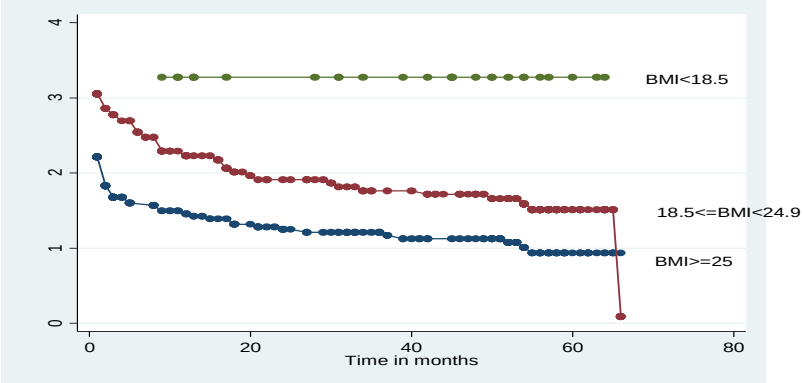
Plot of –log(-log(survival)) versus follow up time to assess the proportionality assumption for WHO clinical stage



Plot of –log(-log(survival)) versus follow up time to assess the proportionality assumption for regimen type.



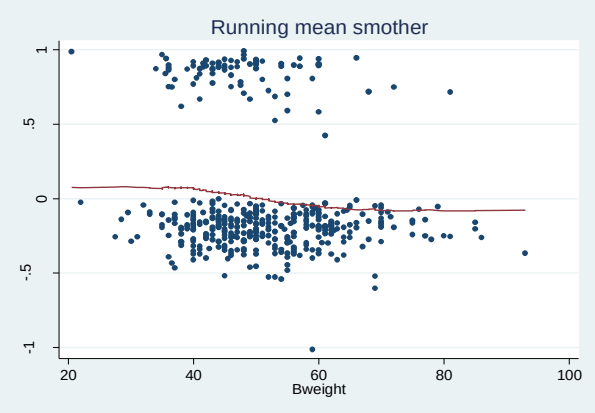
Plot of –log(-log(survival)) versus follow up time to assess the proportionality assumption for pre-TB status



.

Plot of –log(-log(survival)) versus follow up time to assess the proportionality assumption for body mass index.

Figure 7: Plots of Martingale residuals for the continuous covariates (a) Baseline weight and (b) Baseline CD4 count using Cox PH model



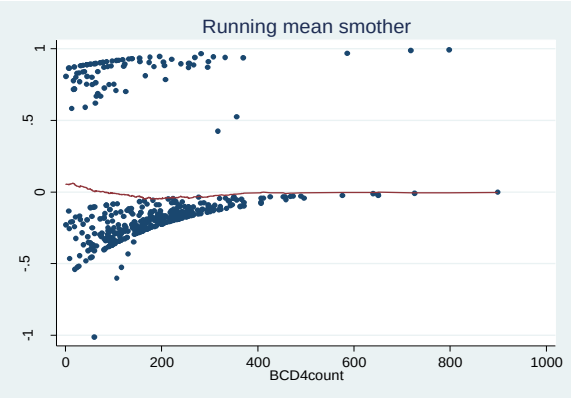
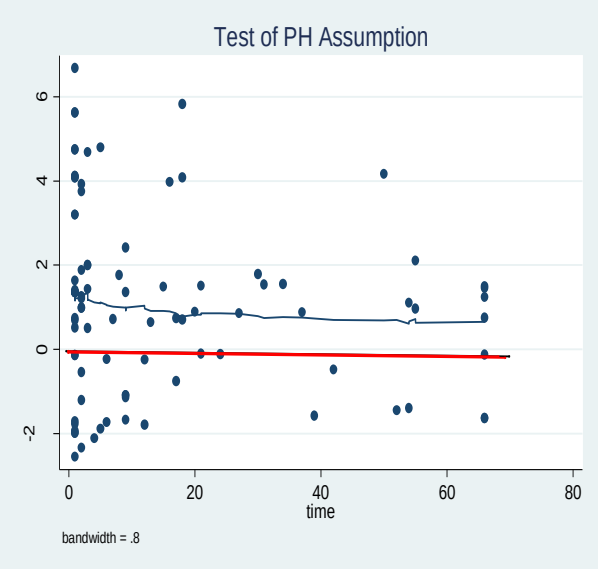
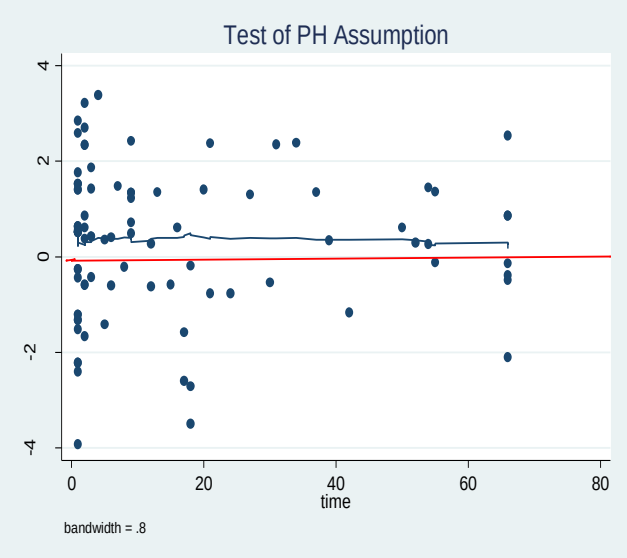


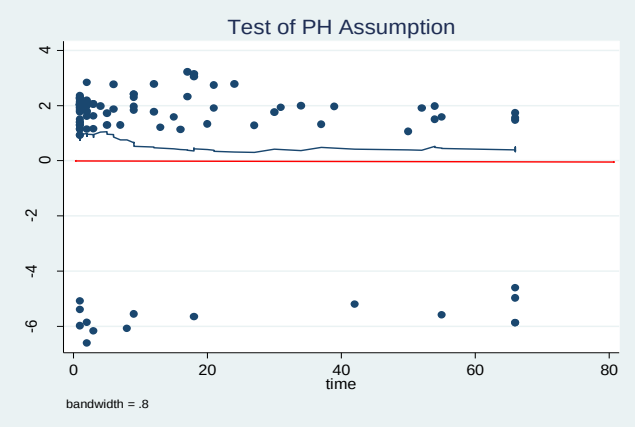
Table 4.7: Test for proportional hazard assumption based on Schoenfeld residuals

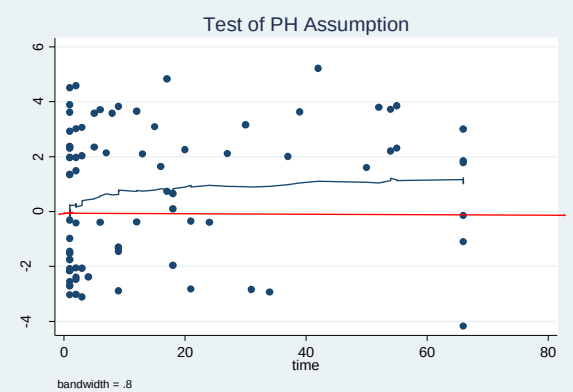
Covariates	rho	Chi-square	df	prob> Chi-square
Functional status				
Ambulatory	-0.09601	0.85	1	0.3559
Bedridden	-0.09408	0.96	1	0.3283
WHO clinical stage				
Stage II	-0.06294	0.37	1	0.5454
stage III	-0.11089	1.20	1	0.2732
stage IV	-0.05388	0.28	1	0.5981
Base line CD4 count	0.20038	4.96	1	**0.0260**
Regimen type	-0.22042	4.53	1	0.1813
Pre=TB status	0.25672	8.12	1	0.2126
Body mass index				
BMI1	0.09868	0.88	1	0.3479
BMI2	-0.02506	0.06	1	0.8083
Global test		21.37	10	**0.0186**

Figure 5: Graphs of the scaled Schoenfeld residuals and their LOESS smooth curves obtained from the model in Table 7 For the covariates functional status, WHO clinical stage,CD4 cell count regimen type, pre-TB status and body mass index.


Test of PH Assumption
-2
0
2
4
6
0
20
40
60
80
time
bandwidth = .8


Test of PH Assumption
-4
-2
0
2
4
0
20
40
60
80
time
bandwidth = .8


Test of PH Assumption
-6
-4
-2
0
2
4
0
20
40
60
80
time
bandwidth = .8


Test of PH Assumption
-4
-2
0
2
4
6
0
20
40
60
80
time
bandwidth = .8

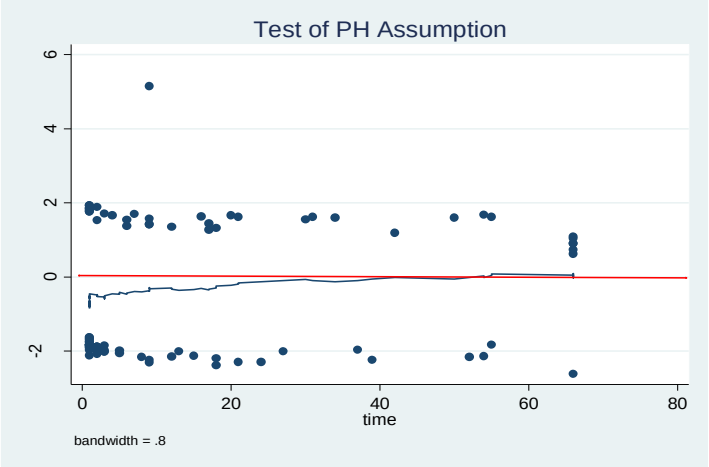
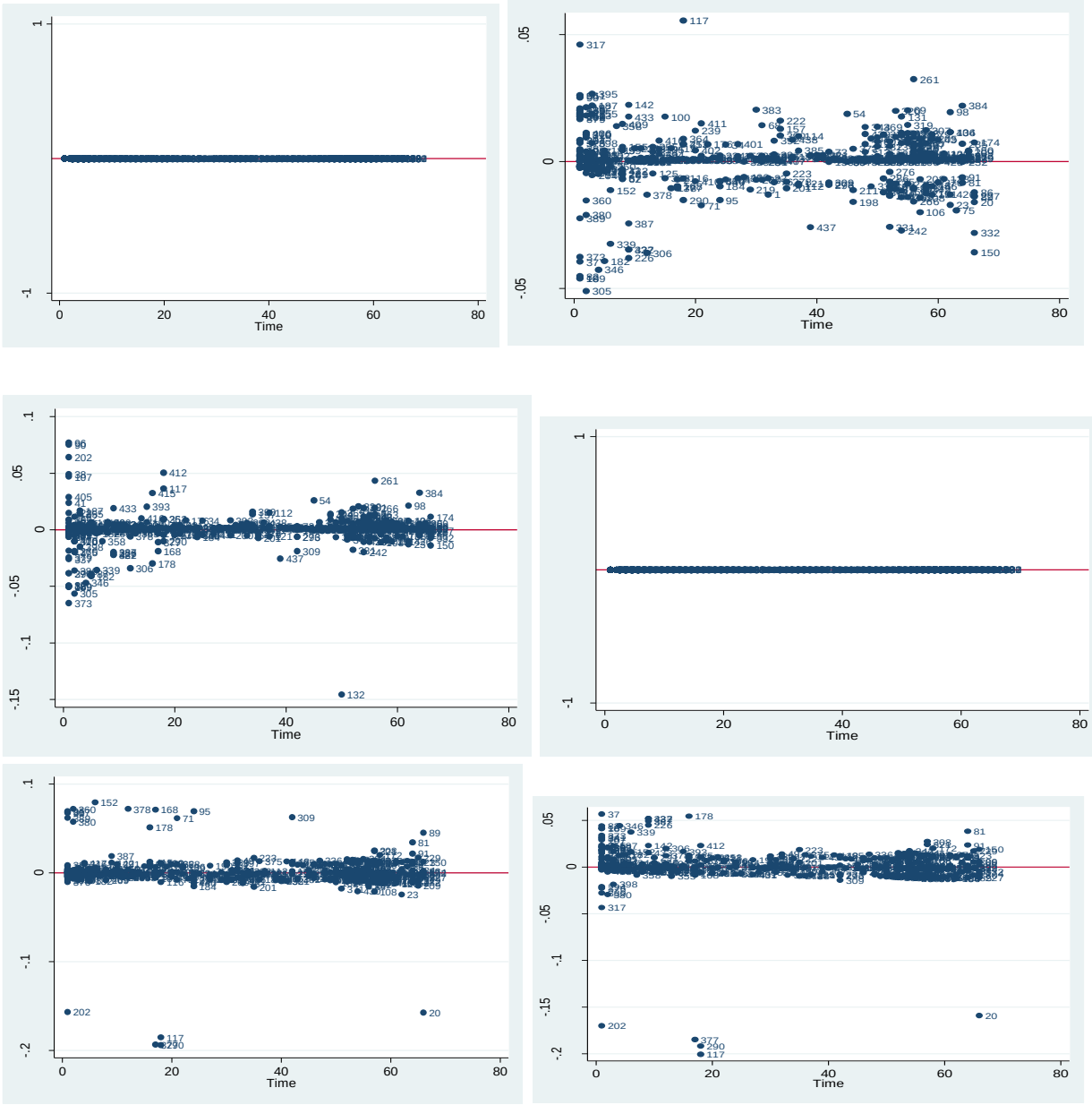
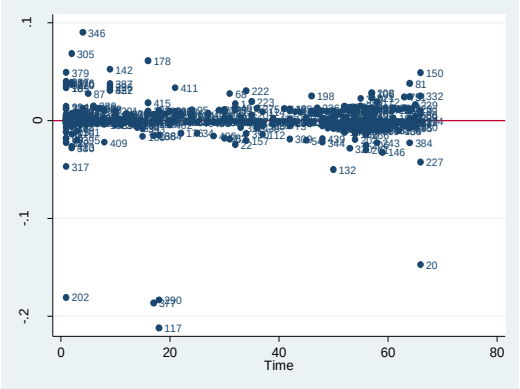
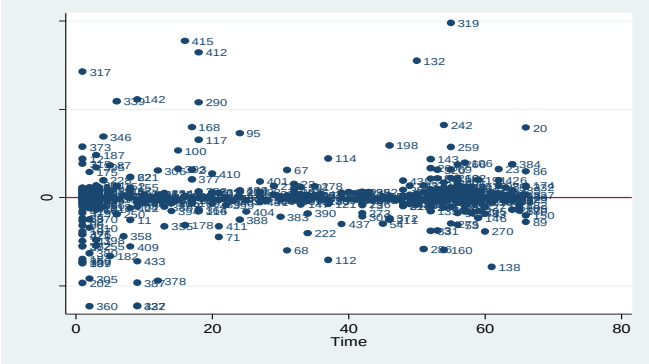
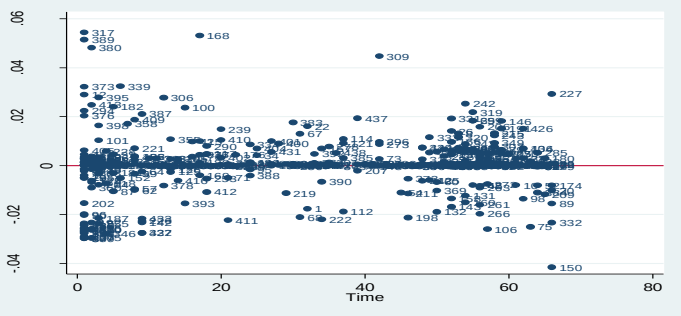

Test of PH Assumption
-2
0
2
4
6
0
20
40
60
80
time
bandwidth = .8

Figure 9: Index plots of dfbeta for the multivariate Cox regression of time-to-death




Time


Time


Time

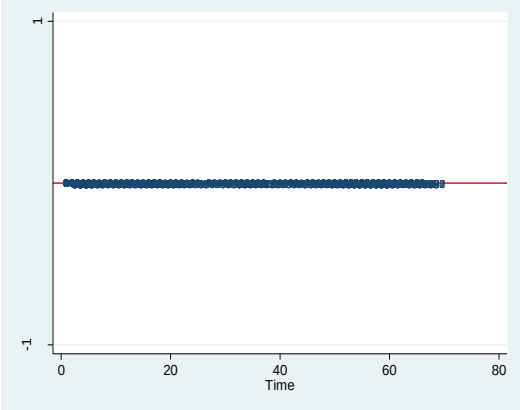

Time

Table 4.17: Standardized variability and AIC for Cox and Parametric Models of time to death in Univariate Analysis

Covariates	**Cox**			**Exponential**			**Weibull**		
	Coef.	SV	AIC	Coef.	SV	AIC	Coef.	SV	AIC
Functional status									
Ambulatory									
Bedridden	1.33	0.182	919.25	1.492	0.162	679.66	1.35	0.173	656.6445
	3.17	0.009		3.992	0.075		3.42	0.089	
WHO stage									
stage III	1.93	0.270	948.31	2.074	0.252	735.03	1.85	0.256	694.9958
stage IV	2.66	0.200		2.846	0.187		2.53	0.192	
BaselineCD4	-0.005	0.249	982.05	-0.006	0.213	772.37	-0.005	0.232	728.4351
Regimen type	1.31	0.237	977.85	1.416	0.219	770.05	1.12	0.252	730.9508
Pre-TB status	1.29	0.165	967.70	1.440	0.148	756.73	1.39	0.149	709.7455
Body mass index							-0.53	0.397	738.1923
18.5<=BMI<24.9	-0.597	0.367	988.94	-0.647	0.334	783.94	-2.25	0.448	
BMI>=25	-2.167	0.466		-2.25	0.448				

….Continued

Lognormal			**Log logistic**			**Gamma**		
-2.27	0.175	656.77	-2.18	0.182	656.94	-2.39	0.180	640.9497
-5.74	0.123		-5.66	0.106		-6.007	0.112	
-3.63	0.049	671.3064	-3.82	0.255	676.7417	-3.533	0.210	672.6414
-4.96	0.183		-5.20	0.201		-4.768	0.189	
0.0084	0.225	704.777	0.0104	0.238	712.0124	0.0032	0.273	669.1043
-2.49	0.225	702.635	-2.60	0.237	709.8545	-1.176	0.233	691.8896
-2.62	0.195	696.0736	-2.70	0.180	699.168	-2.634	0.203	698.0686
1.34	0.334	711.0979	1.28	0.349	720.3347	0.5714	0.569	688.2748
4.12	0.365		4.37	0.433		2.913	0.242	