Modeling Time to Death of HIV Infected Patients on Antiretroviral Therapy the case of Hossana Queen Elleni Mohammad Memorial Hospital, South Ethiopia

By

Adinew Handiso Arficho

A Thesis Submitted to the Department of Statistics, College of Natural Science, Jimma University as a Partial Fulfillment for the Requirements of Master of Science (MSc) Degree in Biostatistics

June, 2016

Jimma, Ethiopia

Modeling Time to Death of HIV Infected Patients on Antiretroviral Therapy the case of Hossana Queen Elleni Mohammad Memorial Hospital, South Ethiopia

BY

Adinew Handiso Arficho

Main Advisor

Dr. K.Sudhir (PhD.)

Co- Advisor

Yasin Negash (MSc.)

June, 2016

Jimma, Ethiopia

STATEMENT OF THE AUTHOR

First, 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 partial fulfillment for the Degree of Master of Science. The thesis can be deposited in the library of the university 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.

ADINEW HANDISO ARFICHO

Date: _____

Signature: _____

Jimma University, Jimma

DEPARTMENT OF STATISTICS, SCHOOL OF GRADUATE STUDIES

JIMMA UNIVERSITY

As thesis research advisors, we here by certify that we have read the thesis prepared by **ADINEW HANDISO ARFICHO** under our guidance, which is entitled "**Modeling Time to Death of HIV Infected Patients on Antiretroviral Therapy in case of Hossana Queen Elleni Mohammad Memorial Hospital, South Ethiopia**", 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. K.Sudhir (PhD)	_____	_____
Advisor	Signature	Date
Yasin Negash (MSc.)	_____	_____
Co-advisor	Signature	Date

As the members of the board of examiners of MSc 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

ACKNOWLEDGMENT

First and foremost, I am indebted to the Almighty God because of whose full mercy and grace I could complete my study.

I would like to express my deepest gratitude to my main advisor Dr. K.Sudhir and Co-advisor Mr. Yasin Negash for their excellent guidance, instructions and constructive comments beginning from the proposal development up to the final work of this thesis. This thesis would not be possible without their advice, suggestions and encouragement.

My grateful thanks to my parents, especially to my brother Amenuel Handiso and my sister Birhanesh Handiso who were always there to provide me with continuing motivation and encouragement, and constructive advice that helped me to complete this work**.** I also am so grateful to my parents Petros A., Bekele A., meheretu S. and Adimasu J. for their encouragement and supports.

I would like to thank Hadiya Zone Trade and Industry department for providing me the financial support for this study. I would also like to thank all my classmates for their encouragement and support that they gave me during my study period and this thesis work.

Lastly not from least, I would like to thank Wondimagegn Geranibo and Tagesse Erimias; data clerks of Hossana District Queen Elleni Mohamad Memorial Hospital ART clinic for helping me in collecting the data for this study.

ABSTRACT

Human Immunodeficiency Virus (HIV) is the virus that causes Acquired Immune Deficiency Syndrome (AIDS). HIV attacks and destroys certain types of white blood cells that are essential to body's immune system, the biological ability of the human body to fight infections. . The main aim of this study is modeling the factors that affect survival time of HIV infected patients by using Cox ph and parametric survival regression models. This study is a retrospective cohort study based on data from the ART clinical in Hossana Queen Elleni Mohamad Memorial Hospital , south Ethiopia. All HIV positive patients who are 15 years old and above placed under ART in between February 2011 to January 2016 were population in this study. The analytical methodologies were used the Kaplan-Meier and Log Rank Test to estimate Descriptive analysis , Cox's regression model was employed to identify the covariates that have a statistical significant effect on the survival time of HIV infected patients and exponential, weibull, log logistic and log-normal survival regression models were applied to compare efficiency of the models. The overall mean estimated survival time of patients was 51.5 months. The Cox Proportional Hazards regression Model result revealed that baseline weight, ART adherence, baseline CD4 count, WHO clinical stage, level of education, substance use and TB co-infection of patients are the major factors that affect significantly survival time of HIV infected patients. For future researchers on this area should apply Weibull survival regression model because Weibull distribution is unique that means only one that simultaneously both proportional and accelerated and also it predicts well the covariate that are associated with high risk of mortality.

***Key Words**: Survival analysis; Cox Proportional Hazard Regression model; Weibull Regression Model ; Hazard ratio*

Table of Contents

TITLES	PAGES
STATEMENT OF THE AUTHOR	I
ACKNOWLEDGMENT	III
ABSTRACT	IV
List of Tables	VII
List of Figures	IX
LIST OF ACRONYMS	X
CHAPTER ONE	1
1. INTRODUCTION	1
1.1 Background of the study	1
1.2.Statement of the Problem	5
1.3.Objectives of the study	6
1.3.1. General objective of the study	6
1.3.2. Specific objectives of the study	6
1.4. Significance of the study	6
1.5. Scope of the study	6
CHAPTER TWO	7
2. LITERATURE REVIEW	7
2.1. General overview on HIV/AIDS	7
2.2. Review of ART	9
2.3. Survival Modelling Approaches	10
CHAPTER THREE	15
3. DATA AND METHODOLOGY	15
3.1.Source of Data and Design	15
3.2. Study area and period	15
3.3. Study population	15
3.3.1 Eligibility Criteria	15
3.4. Data collection procedures	16
3.5. Variables in the study	17
3.5.1. Response Variable	17

3.5.2. Predictor (independent) Variables 17
3.6. Methods of Survival Analysis 18
3.6.1. Descriptive Methods for Survival Data 18
3.6.1.1. Survival Function 19
3.6.1.2. Hazard Function 19
3.6.1.3. Kaplan-Meier Estimator of the Survival Function 20
3.6.1.4. Comparing Survival Function 21
3.6.2. Modeling Survival Data 23
3.6.2.1. Sem-Parametric Survival Models 23
3.6.2.1.1. The Cox Proportional Hazards Regression Model 23
3.6.2.1.2. Fitting the Cox Proportional Hazard Regression Model 25
3.6.2.1.3. Assessment of Model Adequacy 28
3.6.3.1.4 Checking for the Proportional Hazards Assumption 32
3.6.2.2 Parametric Survival Regression Models 34
3.6.2.2.1 The Exponential Survival Regression Model 34
3.6.2.2.2 The Weibull Survival Regression Model 35
3.6.2.2.3 Log-logistic Survival Regression Model 36
3.6.2.2.4 Log-normal Survival Regression Model 37
3.6.2.2.5 Fitting parametric Survival Regression Models 38
3.6.2.3.6 Assessment Adequacy of parametric Survival Models 39
3.7. Model Development 40
3.7.1. Variable Selection 40
3.7.2. Model Comparison 41
3.8. Ethical Considerations 42
CHAPTER FOUR 43
4. RESULTS AND DISCUSSION 43
4.1. Descriptive Analysis of HIV patients 43
4.2. Comparison of Survival Experience 44
4.3. Results of the Cox proportional hazards Regression Model 48
4.3.1. Univariate Analysis of Cox ph Regression Model 48
4.3.2. Multivariate Analysis of Cox ph Regression Model 50
4.3.3. Assessment of Model Adequacy 52

4.3.3.1 Test the assumption of proportional hazards 52
4.3.3.2. Identification of Influential Observation 53
4.3.3.3. Overall Goodness of Fit 54
4.3.4. Interpretation and Presentation of the Final Cox PH Regression Model 55
4.4. Parametric Survival Regression Model Analysis 57
4.4.1. Model Comparison for Time to death of HIV infected Patients 57
4.4.2. Univariate Analysis of Weibull Regression Model 59
4.4.3. Multivariate Analysis of Weibull Regression Model 60
4.4.4. Assessment of adequacy and Interpretation of the Weibull Regression Model 62
4.5 .DISCUSSION 64
CHAPTER FIVE 66
5. CONCLUSION AND RECOMMENDATION 66
5.1. Conclusion 66
5.2. Recommendation 66
Limitation of the Study 67
REFERENCES 68
APPENDIX 73

List of Tables

Table 3.1: Covariates which were used for the analysis of data in this study..... 17
Table 4.2 Comparison of Survival Experience of HIV Patients Using socio-demographic and clinical variables. 47
Table 4.3: Single Covariate Analysis of Cox proportional hazards regression model of HIV Patients..... 49
Table 4.4: the Parameter Estimates, Standard Errors and the Hazard Ratios of the Final Cox Proportional Hazard Regression Model..... 51
Table 4.5: Results of the multivariable proportional hazards regression model containing the 53
variables in Table 4.4 and their interaction with log time (in months)..... 53
Table 4.6: The Likelihood Ratio, Score and Wald tests for overall measures of goodness of fit of the final Cox PH model in table 4.4..... 55
Table 4.7: Selection of parametric survival regression model by using Log likelihood and AIC .. 58
Table 4.8: *Comparison of* AIC *values* for Cox proportional model and weibull survival regression model…… 59
Table 4.9: The Result of Deviance, -2*LL and Pr (>Chi) of univariate weibull regression model Analysis..... 60
Table 4.10: Summary result of Parameter Estimates, Standard Errors and the 95% CI of the final multivariate Weibull regression model Analysis..... 61
Table 4. 11: The Likelihood Ratio Test of the Final Weibull Regression Model for Time to death of HIV Patients 62
Table 4.1.1: Summary of descriptive statistics for Socio-Demographic Variables 73
Table 4.1.2: Summary of descriptive statistics for clinical Variables 74
Table 4.12: Results of the multivariable proportional hazards Cox regression model containing the variables significant at 25% level in the single covariate proportional hazards Cox regression model..... 75
Table 4.13: The five highest differences in the parameter estimates of the variables included in the model in Table 4.6 when the data value for each patient is in turn deleted from the model 76
Table 4.14: The five highest likelihood displacement values when each observation is in turn deleted from the model in Table 4.6 77
Table 4.15: Univariate Parametric Survival Regression Model analysis for modeling time to..... 78
death of HIV infected patients. 78

List of Figures

Figure 4.1: the plot of the overall estimate of Kaplan-Meier survivor function of HIV Patients... 45
Figure 4.2: Plots of Kaplan-Meier survivor function estimates for the variable baseline weight and education level 45
Figure 4.3: plots of Kaplan-Meier survivor function estimates for the variable substance use and ART adherence 45
Figure 4.4: Cumulative hazard plot of the Cox-Snell residual for final Cox PH model 54
Figure 4.5: The Cox Snell plot after fitting Weibull regression model 57
Figure 4.6: The Cox Snell plot after fitting Exponential regression model 57
Figure 4.7: The Cox Snell plot after fitting log logistic regression model 58
Figure 4.8: Cox Snell plot after fitting lognormal regression model 58
Figure 4.9: plots of Kaplan-Meier survivor functions estimates for the variable WHO clinical stage and ART baseline CD4 79
Figure 4.10: plots of Kaplan-Meier survivor function estimates for the variable TB co-infection and Residence 79
Figures 4.13: (a – e) :Plot of log (-log (survival)) versus log survival time for categorical predictors in the fitted model, for TB co-infection, for baseline weight, for CD4 count, for substance, and for ART adherence of the patient seen respectively. 80
Figure 4.14: The plot of Scaled Schoenfeld residual for baseline weight and level of education respectively, to check the validity of the PH assumption 81
Figure 4.15: The plot of Scaled Schoenfeld residual for ART adherence and WHO clinical stage respectively, to check the validity of the PH assumption 81

LIST OF ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral therapy
ARV	Antiretroviral
AIC	Akaike Information Criteria
CI	Confidence Interval
D4T	Stavudine
EFV	Efavirenz
HAART	Highly Active Antiretroviral Therapy
HZHD	Hadiya Zone Health Department
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
HQEMMH	Hossana Queen Elleni Mohamad Memorial Hospital
EDHS	Ethiopian Demographic and Health Survey
STDs	Sexually transmitted diseases
MoH	Ministry of Health
FHAPCO	Federal HIV/AIDS Prevention and Control Office
NAIDSRC	National AIDS resource center
LR	Likelihood Ratio
MLE	Maximum Likelihood Estimate/Estimator
PLWHA	People Living with HIV/AIDS
PMTCT	Prevention of Mother to Child Transmission
SNNPR	Southern Nations Nationalities and People's Region
TB	Tuberculosis
UNAIDS	United Nationals Agency for AIDS
WHO	World Health Organization
ZDV	Zidovudine
3TC	Lamuvidine

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the study

A pattern of highly unusual infection in otherwise healthy young adults emerged in the early 1980s in the united States of America. This pattern or clusters of diseases that appeared in those whose immune system being attacked, came to be called Acquired Immune Deficiency Syndrome (AIDS). Between the 1983 and 1994 a new virus called Human Immunodeficiency Virus (HIV) has been identified as a cause of AIDS (UNAIDS, 2005).

Human Immunodeficiency Virus (HIV) is the virus that causes Acquired Immune Deficiency Syndrome (AIDS). People are said to be HIV positive when the HIV antibody is detected in their blood. HIV attacks and destroys certain types of white blood cells that are essential to body's immune system, the biological ability of the human body to fight infections. HIV infects primarily vital cells in the human immune system such as helper T cells (to be specific, CD4+ T cells), macrophages, and dendritic cells that are necessary to activate B-lymphocytes and induce the production of antibodies. The infected person becomes susceptible to a wide range of opportunistic infections, such as tuberculosis and Pneumocistic Carinii Pnemonia, and rare cancer such as Caposis Sarcoma (WHO, 2007).

There are two types of HIV: HIV-1 and HIV-2. Both types are transmitted by sexual contact, through blood, and from mother to child, and they appear to cause clinically indistinguishable AIDS. However, it seems that HIV-2 is less easily transmitted, and the period between initial infection and illness is longer in the case of HIV-2.Worldwide, the predominant virus is HIV-1, and generally when people refer to HIV without specifying the type of virus they will be referring to HIV-1. The relatively uncommon HIV-2 type is concentrated in West Africa and is rarely found elsewhere (Marlink*et,* 1994).

HIV/AIDS have caused the world most distressing tragedy and danger. More than 25 million people worldwide have died of AIDS since 1981. According to UNAIDS, around 33.4 million people are living with HIV throughout the world, including the approximately 2.7 million newly infected in 2008 and over 2 million have lost their lives to the disease leaving behind orphaned children and

ravaged communities in 2008 Joint (UNAIDS, 2009). There are 34.2 million people living with HIV, 2.5 million new HIV infections and 1.7 million deaths due to AIDS in 2011 worldwide (WHO, 2012).

The greatest burden of the disease is concentrated in developing countries with the least ability to cope. More than 66% of the 34.2 million people living with HIV/AIDS are in sub Saharan Africa, where AIDS is the leading cause of death. The countries of sub-Saharan Africa are home to approximately 22.6 million people living with HIV/AIDS. The HIV pandemic created unprecedented burden on the economies and health care systems of affected countries, particularly in Sub-Saharan Africa, where prevalence is highest. Some of the most explosive epidemics have been seen in Southern Africa. South Africa has the largest number of people living with HIV/AIDS in the world. Botswana and Swaziland have the highest prevalence levels, 38% and 33% respectively. West Africa has been relatively less affected by HIV infection than other regions of sub-Saharan Africa. Uganda and Senegal represent two success stories. Uganda has brought estimated prevalence rate down to 5% by the end of 2011 from an estimated peak of close to 14% in the early 1990s with strong prevention campaigns. HIV prevalence has stabilized in Senegal at a relatively low level (Stephen *et al.,* 2011).

In Ethiopia since the first two AIDS case reported in 1986, the prevalence rate has continuously increased until the year 2000 when it begun to show some decline (Merso, 2008). Adult HIV prevalence in 2009 was estimated to be between 1.4% and 2.8% in the country. Prevalence was 1.8% for males and 2.8% for females, and women accounted for 59% of the HIV-positive population. There were an estimated 131,145 new HIV infections and 44,751 AIDS-related deaths of which females accounted for 57% of the total infections and deaths. The total estimated number of HIV-positive pregnant women and annual HIV positive births in the same year were 84,189 and 14,140, respectively. There were an estimated 72,945 children less than 15 years old living with HIV, out of which 20,522 needed ART. Due to the combined effect of poverty and AIDS, more than 5.4 million children under the age of 18 years were orphaned out of which 855,720 lost at least one parent due to AIDS (FHAPCO, 2010).

From the total number of people who have died due to HIV/AIDS in 2006 alone was 88,997 and in 2007 it was estimated that 71,902 people would die (FMOH, 2007). In 2010, AIDS related death is expected to decline to 28,073 which might be as a result of ART. It is estimated that 398,717 of the HIV positive cases are in need of ART out of which 26,053(6.5%) are children under 15 years of age. It is also estimated that the all ages HIV prevalence in SNNPR in 2013 is 0.9% with 18,557 male and

27,221 female cases who live with the virus, an estimated 45,778 people are living with HIV/AIDS this may increase the number of HIV positive patients in the region (NAIDSR, 2014).

Hadiya zone is one of South Nations, Nationalities and Peoples Region (SNNPR), Ethiopia. SNNPR is one of the largest regions in Ethiopia, accounting for more than 10 percent of the country's land area and the current population is approximately 17 million with an average household size of 4.8 in 2007. More than 91 percent of the SNNPR population lives in rural areas. The mid-2012 population was estimated at nearly 17,745,000. The region is divided into 13 administrative zones including Hadiya zone. Hadiya Zone has 10 woradas and one town administration with an estimated total population of 1.5 million in 2013. It has one zonal hospital, 37 functional public health centers and 282 health posts among which ten health centers and one zonal hospital are provide a total of 2899 HIV infected patients have visited ART clinic, 2039 ever started ART of which 258 have died*(*HZHD, 2014).

The Antiretroviral Therapy (ART) drugs improve the quality of HIV infected persons by helping them to stay well much longer than they otherwise would. The drugs slow down the replication of HIV within the body. Although the treatments are not a cure and continue to present new challenges with respect to side-effects and drug resistance ART as disease modifying therapy for established HIV infection has produced dramatic effects on morbidity and mortality among HIV/AIDS patients. Recognizing the urgent need for antiretroviral treatment, the government of Ethiopia issued the first antiretroviral guideline in 2003, which is the same year as the antiretroviral treatment has, began on 900 HIV patients. The number of people who were able to access ART has substantially increased from 900 in 2003 to 180,447 in 2008 and in 2009 as part of the global issue the government (Seyoum *et al.,* 2009).

Survival analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event. 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. In medical science, time to event can be time until recurrence in a cancer study, time to death or time until infection. In the social sciences, interest can lie in analyzing time to events such as job changes, marriage, birth of children and so forth (Klembaum, 1996).

The method of survival analysis has experienced tremendous growth during the latter half of the 20th century. The methodological developments of survival analysis that had the most profound impact are the Kaplan-Meier method for estimating the survival function, the log-rank test for comparing the equality of two or more survival distributions, the Cox proportional hazards (PH) model for examining the covariate effects on the hazard function and the accelerated failure time (AFT) model for Analysis in Studies with time-varying covariates. The basic concepts of survival analysis, Descriptive Methods for Survival Data (the Kaplan-Meier method and the log-rank test), semi-parametric methods (the Cox PH model) and parametric methods such as exponential survival regression model, weibull survival regression model, lognormal survival regression model and log-logistic survival regression model for analyzing survival data (Lawless, 1982).

The Cox proportional hazard model is the most widely used for the analysis of survival data in the presence of covariates or prognostic factors. This is the most popular model for survival analysis because of its simplicity, and not being based on any assumptions about the survival distribution. 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 hazard function. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory. These developments have led to the introduction of several new extensions to the original model. However the Cox PH model may not be appropriate in many situations and other modifications such as stratifed Cox model or Cox model with time-dependent variables can be used for the analysis of survival data. The accelerated failure time (AFT) model is another alternative method for the analysis of survival data (Collett, 2003).

1.2. Statement of the Problem

Today, Ethiopia has made progress in reducing the number of HIV/AIDS death nationally, but the observed changes are not sufficient enough compared to the desired goals of the response against the epidemic. Numerous researches have also been conducted that tried to address many of the issues that arise in connection with the HIV epidemic. One thing that should, however, be noted is that many of these research works mainly focused on the assessment of the prevalence and the study of the numerous prevention measures that should be undertaken to stop or reduce the spread of the epidemic (NAIDSRC. 2010), but it seems that little attention has been given to study high risk factors that facilitate mortality of those people living with HIV/AIDS. Investigating the existence of significant associations between the different factors and HIV/ADIS mortality can provide evidence for informed protection mechanisms. The effectiveness of ART could vary from region to region (this variation is also generally a reflection of the variation that exists between and within countries and regions as regards HIV prevalence and its epidemiological patterns) because of the difference in background disease burden (such as tuberculosis or intestinal parasites), viral subtypes, and possible genetic differences in drug metabolism. However, such arguments are based on little data from the resource-limited settings **(**UNAIDS, 2009**,** Degu *et al*., (2008). Given this as a back drop, this thesis will focus on the consideration of some of the possible factors/variables that may possibly influence the survival status of people who are following ART in HQEMMH, South Ethiopia. Furthermore, modeling time to death of HIV infected patients on ART is helpful to identify covariates that facilitate mortality of those people living with HIV/AIDS (Leigh *et al.,* 2009).

In addition, a study conducted previously in HQEMMH used the Multilevel logistic regression model (Gezaghen, 2013), but the Multilevel logistic regression model is not well suited to survival data for several reasons. According to (Collett, 2003), the survival times are not normally distributed and the censored data are the result of missing values on the dependent variable, but in this study the survival analysis method used to identify the risk factors as well as to compare the efficiency of Cox ph and parametric survival regression models. Many covariates will collect to reduce possible modeling bias, when a large semi parametric/parametric model is built. An important and the first challenging task are to efficiently select a subset of significant variables upon which the hazard function depends (Hosmer and Lemeshow, 1999).

In general, the motivation behind this study is intended to address the following two major research questions:

- Which factors significantly affect survival time of HIV infected patients on ART?
- Which type of survival model, Cox ph or parametric regression model, predicts well the covariate that are associated with high risk of mortality?

1.3. Objectives of the study

1.3.1. General objective of the study

The main objective of this study is modeling the factors that affect survival time of HIV infected patients by using Cox ph and parametric survival regression models based on HQEMMH data.

1.3.2. Specific objectives of the study

1. To compare survival time among the different categories of covariates on HIV-infected patients
2. To estimate time-to-death of HIV infected patients treated with ART
3. To determine significant covariates that are high risk of mortality on HIV-infected patients
4. To compare the efficiency of Cox ph and parametric survival regression models in case of HQEMMH

1.4. Significance of the study

1. The result of this study provide information to the government program planners, decision makers, ART program implementers at different level and other stakeholder who work in the areas of giving care, support and treatment for HIV/AIDS patients.
2. The study useful to identify death risk of patients under significant factors at different time during and after HIV-treatment.
3. The study helps for both donors and government to understand factors that influence survival time of HIV patients.
4. The result of this study can provide better opportunity for further study in future.

1.5. Scope of the study

The study would have been covered HIV infected patients from 1st February 2011 to 1st January 2016, if it had been carried out in Hossana district level. However, it is limited to identify the risk factors and compare models in HQEMMH data.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. General overview on HIV/AIDS

According to UNAIDS (2004), the statistics about the impact of HIV/AIDS worldwide are overwhelming. Estimate of the United Nationals Agency for AIDS (UNAIDS) indicate that over 40 million people were living with HIV/AIDS in 2004, that nearly 25 million people have died of AIDS since the disease was first discovered in the early 1980's, and that more than 15.6 million children under 15 have lost either their mother or both parents as a direct result of AIDS.

Moraes *et al*., (2008) conducted a national study of adults diagnosed with AIDS in Brazil to determine whether and to what extent survival time has increased with widespread access to antiretroviral drugs. Their results showed that predictors of longer survival included antiretroviral treatment, year of diagnosis, higher education, sexual exposure category, female sex and Pneumocystis carinii pneumonia prophylaxis. In further analysis, the predictive value of most of these was attenuated or disappeared, leaving antiretroviral treatment as the main predictor of survival.

Johannessen *et al*., (2009), in developing countries of Africa 39.4 million Peoples were living with HIV/AIDS. Adults contribute 37.2 million. About 5 million peoples were newly infected of which 4.3 million were adults from these more than 95% of new infections were in developing countries. Over 6 million infected need ART but 350,000-400,000 was treated in developing countries. By December 2006 two million people in low and middle income countries were receiving ART but this was still only 28% of those estimated to be in urgent need of it.

According to AIDS Resource Center (2008), as one of the countries in Africa, Ethiopia's GDP has increased in the last couple of years with double-digit economic growth. With the current effort of achieving the Millennium Development Goals, the country has made significant lead in several development sectors as in Health, Education and poverty reduction. Despite this dedicated effort to avert the burden of poverty, the hindrance caused by HIV/AIDS associated morbidity and mortality has posed an obstacle in the productive part of the society. The fact that the pandemic is predominately affecting part of individuals between the ages of 14-59, the productive age group, is a significant loss of labor supply. The protracted morbidity and eventual mortality resulting from

HIV/AIDS causes significant lost time to illness, reduced productivity, shortage of manpower, increased absenteeism in the workplace and rising medical costs.

According to UNAIDS (2010), HIV/AIDS has the greatest challenges to the Ethiopian health system, as elsewhere in sub-Saharan African countries. It has remained among the major causes of deaths over the past two decades. In 2010, more than one million people were estimated to be living with HIV in Ethiopia of whom nearly 397,818 need ART care and treatment. According to the FHAPCO single point estimate for prevalence of HIV/AIDS in Ethiopia, the adult (15-49) HIV prevalence for 2007 is estimated at 2.1% of which 7.7% is urban and 0.9% is rural (EMOH 2007). In 2010, the FHAPCO estimates of the overall adult (15-49) HIV prevalence is 2.4%. Urban and rural HIV prevalence rates were 7.7% and 0.9%, respectively. In 2010, an estimated 28,073 Ethiopians died of AIDS scaling the number of children who have lost one or both parents to AIDS to 804,184.

A retrospective cohort study was done by Reda *et al*., (2013) among HIV infected patients on ART in Hiwot Fana, Jugal and Dil Chora hospitals located in eastern Ethiopia with objective of examining mortality and its predictors among a cohort of HIV infected patients on antiretroviral treatment retrospectively followed for five years. It was found that in the multivariate analysis factors such as WHO stage, weight, CD4 cell counts and level of education were predictors of mortality of HIV infected patients on antiretroviral treatment. The results revealed that WHO stage IV patients were 3 times more likely to die compared to stage I and II patients (HR 3.19; 95% CI 1.51–6.76). Patients who reported to have lost more than 10% of their weight at baseline were 5 times more likely to die compared to those patients who did not (HR 4.93; 95% CI 1.20–20.41). Patients whose CD4 cell counts between 201–300 were 60% less likely to die compared to those whose CD4 counts less than 200 (HR 0.40; 95% CI 0.17–0.93). Those patients with primary education were almost 3 times more likely to die than illiterate counterpart (HR 2.79; 95% CI 1.26–6.16).

Asefa *et al*., (2005) a case-control study that was carried out in Addis Ababa provided evidence that substance abuse, particularly alcohol, was found to be a significant risk factor for HIV infection. The study suggested the need for health education to bring about behavioral changes and further study to identify the prevalence and role of substance in exposure to HIV infection in the community. HIV/AIDS has been and still is the greatest challenges to the Ethiopian health system, as elsewhere in sub-Saharan African countries. It has remained among the major causes of deaths over the past two decades.

FHAPCO (2006), In Southern Nation Nationalities and peoples Representative Region(SNNPR), the HIV prevalence was estimated to be 2.3% (2.6% for females and 2% for males) in 2005.The HIV prevalence among the urban population was estimated 10.2 % (11.5% in females and 8.9% in males). The corresponding estimate among the rural population was 1.5 % (1.7% for females and 1.3% for males).

2.2. Review of ART

Recsky *et al.* (2010), Pearson's X_2, the Cochran-Armitage and the Wilcoxon rank-sum tests have been used to determine the degree to which antiretroviral resistance may contribute to mortality among HIV-infected individuals enrolled in the centralized HIV/AIDS Drug Treatment Program in British Columbia, Canada, who had died between July 1997 and December 2001. In the investigation, of a total of 637 deaths, 83 (13.0 %) were attributed to accidental causes; and the remaining 554 deaths (87.0%) were attributed to non-accidental causes. The accidental causes were illicit-drug overdose (57.8 %), concussion (18.1%) and the remaining percentage accounted to suicide, traffic accidents, assaults, and other in-juries. The non-accidental causes were identified as 383 (69.1%) directly related to HIV infection (JCD-9 and 34 (6.1%) related to liver disease, 25 (4.5%) to various cardiac conditions, 20 (3.6 %) to viral and/ or bacterial infections, 18 (3.2%) to malignant neoplasms, 43 (7.8%) to other causes, and 31 (5.6%) to unknown causes. The study concluded that not only treatment failure due to antiretroviral resistance was a major factor influencing mortality in this cohort but co-morbidities, and other factors had got a lion share as well.

According to Isakidis, (2009). study which includes 285 HIV-positive children treated with first-line ART for at least 24 months to identify risk factors associated with treatment failure in two hospitals of Cambodia at Angkor Hospital for Children and Donkeo Referral Hospital, has shown that ART treatments have the desired effect to improve the survival of children, CD4 count progress, viral load suppression and weight for age Z-score increment, for children with viral load less than 1000 copies/ml. On the other hand for children associated with viral load greater than 1000 copies/ml only CD4 counts and CD4 percentages below the threshold is responsible for severe immunologic suppression at 24 month which predicts treatment failure at this time. However, orphan status and caregiver characteristics, including literacy, age and socioeconomic status, were not associated with treatment failure after 24 months of ART.

2.3. Survival Modelling Approaches

Some of the modeling approaches in survival data analysis particularly risk factors of mortality proposed by different authors were reviewed as follows:

Lavori *et al.,* (1985), study the origin of survival analysis goes back to the time when life tables were introduced. Life tables are one of the oldest statistical techniques and are extensively used by medical statisticians and by actuaries. Yet relatively little has been written about their formal statistical theory. Kaplan and Meier, 1958 gave a comprehensive review of earlier work and many new results. Cox, 1972 was largely concerned with the extension of the results of Kaplan and Meier to the comparison of life tables and more generally to the incorporation of regression like arguments into life table analysis. Survival models have the capability of handling censored data. Cox, 1972 and Cox and Oakes, 1984 used survival analysis in modeling human lifetimes. Fergusson, 1984 used hazard functions to study the time to marital breakdown after the birth of child. Hazard functions had been also used in studies of time to shift in attentions in classroom Females, 1983; in study of relapse of mental illness.

Sethi *et al.,* (2009) , biomedical researchers tend to choose semi parametric methods to model time-to-event data, in a study by, 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.

Katubulushi and Chanda (2009), study conducted in adults shows that the main Factors associated with the greatest reduction in risk of death from time of study entry were current use of HAART, HR 0.13 (0.06–0.30, $p<0.001$), and CD4 count 200 at entry, HR 0.16 (0.08–0.35, p 0.001). Current use of

HAART was the strongest predictor of survival from time of AIDS diagnosis, HR 0.11 (0.05–0.25, $p< 0.001$). The use of HAART therapy and CD4 count were primary predictors of survival (Under the hypothesis that the patients lost to follow-up were dead, study in Senegal shows the probabilities of dying were respectively 13.4% (95% CI, 10.4–17.1%) and 21.0% (95% CI, 17.4 25.4%) at 12 and 24 months of follow-up. Mortality rate decreased from 12.5 deaths/100 person-years (95% CI, 9.4–16.7) during the first year of treatment to 6.6/100 person-years (95% CI, 4.3–10.0) during the second year [hazard ratio (HR), 1.9; 95% CI, 1.1–3.1 P< 0.01] and kept decreasing thereafter (4.5, 2.3 and 2.2/100 person-years for years 3, 4 and 5, respectively) (20). For 0-3, 3-9 and >9 months from ART start mortality rates were 24, 13 and 6/100 per years respectively. 43% of the deaths were in the first 3 months of treatment. Adjusted hazard ratios (HRs) for participants with hemoglobin <8, 8.1-9.9, >11.9(f)/12.9 (m) g/mL were 4.99, 3.05 and 0.12 respectively comparing to 10-11.9 (f)/12.9 (m)g/mL in the first 3 months of ART. AHRs for CD4 counts were 0.40, 0.38 and 0.34 for 50-99, 100-200 and >200.

Ferradini *et al*., (2006), study conducted in Sub-Saharan Africa based on data from 18 published cohort studies containing 39,536 HIV/AIDS patients had employed the Kaplan-Meir method to assess the proportion of survival time and random-effects model to find hazard ratio of prognostic variables. Thus, a result of the study suggested advanced WHO clinical stage and low CD4 cell count as indicator of high mortality. Similarly, a study in Malawi based on 1308 patients employed Kaplan-Meier method to assess the probability of survival and the Cox proportional hazards model to assess the potential predictors of death. The study found low body-mass index, WHO clinical stage IV, male gender, and baseline CD4 count lower than 50cells/ml as independent determinants of death.

Jerene *et al*., (2006), the study was conducted in south Ethiopia between August 2003 and August 2005 on two cohorts of patients: the pre-HAART cohort (185 patients) and the HAART (180 patients) cohort. In the study, the Kaplan-Meier method was used to assess the event-free survival, the log-rank test was employed to test for the statistical significance and the Cox proportional hazards model was used to find out the effect of HAART on mortality and on tuberculosis incidence rates. Thus, the study indicated that the HAART improved survival and decreased tuberculosis incidence. Furthermore, the study recommended the importance of strengthening tuberculosis prevention efforts with the scale-up of treatment programmers.

Kaufmann *et al*., (2011), In Uganda a retrospective cohort of 427 HIV-1 patients who were initiated on ART was studied to establish the effect of AIDS defining complexes (ADCs) on immunological

recovery among patients initiated on ART. Kaplan-Meier survival curves were employed to estimate median times, log rank test to compare different categories and Cox proportional hazard models were used at multivariate analysis. The median time to gaining 50 CD4 cells/µl from the baseline value after ART initiation was longer in the ADC group (9.3 months) compared to the non-ADC group (6.9 months) (log rank test, p = 0.027). At multivariate analysis after adjusting for age, sex, baseline CD4 count, baseline HIV viral load, total lymphocyte count and adherence level, factors that shortened the median time to immunological recovery after ART initiation were belonging to the non-ADC group (HR = 1.31, p = 0.028), adherence to ART of ≥ 95% (HR = 2.22, p = 0.001) and a total lymphocyte count ≥ 1200 cells/mm3 (HR = 1.84, p =0.003). A low baseline CD4 count of ≤ 200 cells/µl (HR = 0.52, p = 0.001) was associated with a longer time to immunological recovery. There was no interaction between low CD4 counts and ADC group. Patients with ADCs take longer to regain their CD4 counts due to the defect in the immune system. This may prolong their risk of morbidity and mortality.

Sterensund (1989), PH modeling is the most frequently use type of the survival analysis modeling in many research areas, having been applied to topics such as smoking relapse and employee turnover (Morita, *et al.,* 1989), and in medical areas for identification of important covariates that have as significant impact on the response of the interested variables. Derbachew, 2012 used PH modeling and parametric `models to examine causes of Survival of Patients with Diabetes Mellitus. Tesfaye Getachew, 2013 used Kaplan-Meier estimation method, Cox PH model and parametric regression to model Survival Analysis of Time to Recovery from Obstetric Fistula.

Andinet and Sebastian (2010), study conducted on 272 HIV/AIDS patients in Shashemene and Assela Hospitals employed Kaplan Meier method to construct survival curves and the Cox proportional hazards model to determine predictors of mortality. The median survival time of the study was 104.4 weeks. The findings of the study showed WHO clinical stage IV, hemoglobin 510 g/dL, and cotrimoxazole prophylaxis therapy (CPT) initiation as the independent determinants of mortality. By the same token, Jerene*et al.* (2006) based on 162 patients, who were enrolled and treated between August 2003 and January 2005, ascertained that advanced disease stage (WHO clinical stage IV) and having total lymphocyte count (TLC) of up to 750/mcL were the major prognostic factors of mortality. The study also recommended identifying and treating patients early through improved counseling and testing strategies.

A study conducted by Gezahegn (2011), on Cox proportional hazard regression to calculate the bivariate and adjusted hazard rate and then determine independent predictors of time to death in CD4 cell counts data in Durame and Hosanna hospital. The estimated mortality was 7%, 8%, 11.3 %, 15.7% and 21% at 6, 12, 24, 36 and 48 months respectively. After adjustment, the independent significant predictors of death in patients living with HIV/AIDS after initiation of ART remain poor ART adherence(AHR=5.09[95% CI: 5.51-49.48]), Advanced WHO staging (AHR=1.5[95% CI: 1.18-2.16]), positive TB test (AHR=3.9[95% CI: 1.89-8.07]), not married or single (AHR=10.27[95% CI: 1.35-78.3), male gender (AHR=1.704[95% CI: 1.23-2.24]) and older age(AHR=1.45[95% CI: 1.1-1.96). This study demonstrated that simple laboratory and clinical data, available to health care providers prior to ART initiation, can predict which patients are at increased risk of death when they start therapy.

Another study that was done by Ketema (2011), in Armed Forces General Teaching Hospital (AFGTH) located in Addis Ababa, Ethiopia and applied Kaplan-Meier survival curves and Log-Rank test to compare the survival experience of different category of ART patients, and employed proportional hazards Cox model to identify independent predictors of mortality. 734 patients on ART were followed for a median of 38.5 months (IQR 10.75, 53). The independent predictors of mortality were low CD4 cell count at baseline, (HR = 0.995, 95% CI: 0.991 -0.999), ambulatory and bedridden functional status, (HR=2.011, 95% CI: 1.018 - 3.973) and (HR=3.358, 95% CI: 1.734 - 6.500), respectively, WHO clinical stages III and IV (HR=7.052,95% CI: 1.677- 29.658) and (HR=12.64, 95% CI: 3.003 - 53.199), respectively, TB co-infection,(HR=1.734, 95% CI: 1.039 - 2.893) and OIs (HR=8.985, 95% CI: 1.240 - 65.085).

Xueyan *et al.,* (2008), the world health organization reported in 1999 that of a total 53.9 million deaths, 1.5 million deaths was caused by TB. In addition, it was claimed as TB co-infection is the leading cause of mortality among those infected with HIV worldwide. A finding of a cross-sectional study based on 241 cases reported from nine domestic hospitals throughout mainland China was in agreement with the stated claim. The patients in the study were followed from January 2003 to December 2005. In spite of the fact that treatments for TB and HIV were provided to the patients, mortality attributable to co-infection was reported for 15.8% of the cases. As a result, the study concluded that HIV/TB co-infection was related to high mortality even when HAART or drug therapy for TB was provided.

Mohammed *et al.,*(2011) conducted a case control study in Jimma and Mettu Karl hospitals where the two hospitals serve as referral and treatment centers for HIV and TB in south-west Ethiopia from January to March, 2009. The study population consisted of 162 cases and 647 controls. Cases were adult people living with HIV/AIDS who developed active pulmonary tuberculosis and controls were people living with HIV/AIDS without active tuberculosis. The final multivariate model was obtained by a forward and backward variables selection procedure. Then, the result reveals that, after adjustment for potential confounders, an initial weight less than 18.5 kg (OR=4.1; 95% CI: 2.3, 7.4), a CD4 lymphocyte count less than 200 cells/mm3 (OR=9.8‘95% CI: 5.5, 17.5), a WHO clinical stage IV (OR=4.3; 95% CI: 2.6, 6.8) and not taking antiretroviral treatment (OR=3.1; 95%CI: 1.9,4.9), were independently associated with the development of active tuberculosis in people living with HIV/AIDS.

Yiannoutsos (2009), in certain circumstances parametric models may offer advantages over semi-parametric model. Using semi-parametric model would have required a much more complex modeling exercise, where factors associated with the change-points would have to be included among the model predictors. In such a case, the parametric models such as Lognormal, Weibull, Exponential, and Log logistic are the common options. These models provide the interpretation based on a specific distribution for duration times without need to proportional hazard assumptions. Nevertheless, the results of data analysis using parametric models are similar to the semi-parametric model. Although the hazard ratios in semi-parametric and parametric models are approximately similar but the Weibull and Exponential regression models are the most favorable for survival analysis of the data (Dehkordi *et al*., 2008). Furthermore, the Weibull regression model is a generalization of the common Exponential regression model (having shape 1). It is more flexible for many real-world situations as, in contrast to the Exponential regression model, it does not assume constant hazard of death.

CHAPTER THREE

3. DATA AND METHODOLOGY

3.1. Source of Data and Design

This study is a retrospective cohort study based on data from the ART clinical in Hossana Queen Elleni Mohamad Memorial Hospital (HQEMMH), Hadiya Zone, SNNP Region of Ethiopia. The survival data were extracted from the patient's chart which contains epidemiological, laboratory and clinical information of HIV patients under ART follow-up including a detailed antiretroviral therapy history.

3.2. Study area and period

The study was conducted in Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, and Ethiopia, from 1st February 2011 to 1st January 2016. Hadiya zone is one of 13 zones in SNNPR. There are 10 woredas and one town administration in the zone and Hosanna town its administrative center which is 235 km away from Addis Ababa. In the town there is one hospital and three health center which gives preventive, curative and rehabilitative service for the population. The hospital has a separate ART clinic and the clinic has one doctor, one nurse, one pharmacist and two data clerks.

3.3. Study population

In determining our sample, first we have to know our source population in this case all HIV positive patients who were 15 years old and above placed under ART in between 1st February 2011 to 1st January 2016 in Hossana Queen Elleni Mohamad Memorial Hospital. This study was based on a review of the patients' intake forms and follow-up cards of HIV patients. For uniformity use in the country so that those forms can be used to document almost all relevant clinical and laboratory variables. In this study were a total of 933 HIV infected patients were investigated who ever started ART.

3.3.1 Eligibility Criteria

Patients a eligible for ART on the basis of the 2003 World Health Organization (WHO) Guidelines Inclusion criteria:

- ✓ HIV infected patients aged 15 years or older who have started ART

Exclusion criteria:

- ✓ Diagnosis made outside of the hospital
- ✓ HIV infected infants and children

3.4. Data collection procedures

The data were extracted from the available standard national medical registers which have been adopted by Federal Ministry of Health (FMOH) to be uniformly used by clinicians to simply identify and document clinical and laboratory variables. The registers include pre-ART register and follow up form, ART intake form, patients' card and death certificate complemented registration by home visitors. Three days training was given for supervisors and data collectors. The overall activity was controlled by the researcher. Data quality was controlled by designing the proper data collection materials and through continuous supervision. The completed data collection forms were examined for completeness and consistency during data management, storage and analysis. The data were collected by data clerks working in the clinic and coded and analyzed using the statistical packages STATA and R.

3.5. Variables in the study

3.5.1. Response Variable

The response (dependant) variable is the survival time of HIV patients, the length of time from ART start date until the date of death or censor measured in months.

3.5.2. Predictor (independent) Variables

The following predictor (covariate) variables were considered for this study.

Table 3.1: Covariates which were used for the analysis of data in this study

No.	Name	Representation	Categories / codes
1	Age of Patients	Age	Less than 40=0 , 40 or above=1
2	Sex of patients	Sex	Male =1, Female =0
3	Marital status	Marital	Single =1,Married =2,Divorced =3, Widowed =4
4	Education level	Edu	No education=1, Primary=2, Secondary and above=3
5	Functional Status	Fun	Working =1,Ambulatory =2,Bedridden =3
6	Residence of the patients	Res	Rural=0 and urban=1
7	Substance use (Alcohol)	Subs	No=0, Yes=1
8	TB co-infection	TBco	No=0, Yes=1
9	ART Adherence	ART.ad	Good =0,Poor=1
10	Drug Regimen	Dur.R	D4T-3TC-NVP=1, AZT-3TC-NVP=2 , TDF-3TC-EFV=3
11	Base Line Weight	BLW	Less than 50kg=0 ,50kg or above=1
12	WHO clinical stage	Stage	Stage I=1, Stage II=2, Stage III=3, Stage IV=4
13	Base line CD4 cell counts	CD4	<200 cells=0, $\geq$ 200 cells=1

3.6. Methods of Survival Analysis

Survival analysis is an important statistical technique used to describe and model time to event data. The purpose of survival analysis is to model the underlying distribution of the failure time variable and to assess the dependence of the failure time variable on covariates. The term survival analysis suggests that the event is death, but that is not necessarily so. Events could also denote success, such as recovery from therapy. Survival time then describes the time from a certain origin to the occurrence of an event.

One of the most important differences between the outcome variables modeled via linear and logistic regression analyses and the time variable in the survival data is the fact that we may only observe the survival time partially. The variable time actually records two different things. For those subjects who experienced the event (most of the time death), it is the outcome variable of interest, the actual survival time. However, for subjects who were alive at the end of the study, for patient who were lost to follow-up, patient withdrawing from the study, competing event (e.g. death due to some cause other than the cause of interest) time indicates the length of follow-up(which is a partial or incomplete observation of survival time). These incomplete observations are referred to as being censored. Most survival analyses consider a key analytical problem called censoring. In essence, censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly (Collett, 2003). The most common form of incomplete data is right censoring. A survival time is said to be right censored if it is recorded from its beginning until a well-defined time before its end time. It means a subject's follow-up time terminates before the outcome of interest is observed. For instance, if an HIV-1 patient is followed until he has a viral load high than 1000 copies/µl and is followed without experiencing this scenario until the end of the observation period. In other words, a survival time is said to be right censored if it begins at time $t = 0$ and terminates before the outcome of interest is observed.

3.6.1. Descriptive Methods for Survival Data

In any applied setting, a statistical analysis should begin with description of the data. In particular, an initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times in a particular group. Routine applications of standard measures of central tendency and variability will not yield estimates of the desired parameters when the data include censored

observations. In summarizing survival data, the two common functions of applied are the survivor function and the hazard function (Hosmer and Lemeshow, 1999).

3.6.1.1. Survival Functions

For most statistical application it is usual to describe models for probability distribution in terms of either the probability density function f(x) or the distribution function F(x). For survival analysis it is usually more appropriate to work with other functions which characterize the probability distribution. Let T be a positive random variable from a homogeneous population, representing the time until the relevant event occurs. In order to characterize the distribution of T one of the most often used functions is survivor function. The survivor function, S (t), is defined for both discrete and continuous distribution as the probability that an individual survives beyond time t i.e., for continuous random variable T, the density function, f (t), is given by

$$F(t) = \mathrm{P}\,(\mathrm{T} < t) \;\; = \int_0^t f(u)du\,, \mathrm{t} \geq 0 \qquad (3.1)$$

Which represents the probability that a subject selected at random will have a survival time less than some stated value t. Then, the survival function$S(t)$ is defined as:

$$S(t) = \mathrm{P}\left(\mathrm{T} \geq \mathrm{t}\right) = 1 - F(t) \qquad (3.2)$$

The survivor function can be used to represent the probability that an individual survives from the time origin to sometime beyond t and then relationship between the probability density function f(t) and S(t) will be:

$$f(t) = \frac{d(1-S(t))}{dt} = \frac{-\mathrm{dS(t)}}{\mathrm{dt}} \qquad (3.3)$$

3.6.1.2. Hazard Function

The hazard function is widely used to express the risk or hazard of experiencing the event (death) at some time t, and is obtained from the probability that an individual experiencing the event at time t, conditional on he or she has survived (censoring) to that time. That is, the function represents the instantaneous failure rate for an individual surviving to time t.

The hazard function $\mathrm{h(t)}$ is defined by: -

$$\mathrm{h(t)} = \lim_{\Delta \mathrm{t} \to 0} \frac{\mathrm{p\{an\ individual\ fails\ in\ the\ time\ interval(t, t + \Delta t)\backslash it\ survived\ until\ time\ t\}}}{\Delta \mathrm{t}}$$

$$\mathrm{h(t)} = \lim_{\Delta \mathrm{t} \to 0} \frac{\mathrm{P}\,[\mathrm{t} < T < t + \Delta \mathrm{t} \backslash \mathrm{T} \geq \mathrm{t}]}{\Delta \mathrm{t}} \qquad (3.4)$$

By applying the theory of conditional probability and the relationship in equation (3.4), the hazard function can be expressed in terms of the underlying probability density function and the survivor function as follows (Collett, 2003).

$$\mathrm{h(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\} \quad (3.5)$$

The corresponding cumulative hazard function $\mathrm{H}(t)$is defined by:

$$H(t) = \int_0^t h(u)du = -\log S(t) \quad (3.6)$$

Hence the survival function can be rewritten as

$$S(t) = \exp\{-H(t)\}, \quad (3.7)$$

The hazard rate is not a probability, it is a probability rate. Therefore it is possible that a hazard rate can exceed one in the same fashion as a density function f(t) may exceed one.

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 Kaplan-Meier Estimator of the Survival Function since they require no assumptions about the distribution of survival time. In order to compare the survival distribution of two or more groups, log-rank tests can be used (Collett, 2003).

3.6.1.3. Kaplan-Meier Estimator of the Survival Function

The Kaplan-Meier (KM) estimator proposed by Kaplan and Meier (1958) is the standard non parametric estimator of the survival function(Collett, 2003). Which is also called the Product-Limit estimator incorporates information from all observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. The Kaplan-Meier estimator is used to estimate the survival time (time of censoring) of a patient and construct survival curves to compare the survival experience of a patient between different categorical variables. The first step in the analysis of ungrouped censored survival data is normally to obtain the Kaplan-Meier estimate of the survivor function.

Suppose the data consist of n survival times t_1, t_2,…,t_n and some of these observations are right-censored times, i.e. for some of the t_j, it is only know that individual j was still censoring at time t_j. Let r be the number of distinct failure times, $r \leq n$, and $t_{(1)} < t_{(2)} < \ldots\ldots < t_{(r)}$ be the ordered failure times.

And assume that n_j is the number of patients at censored just before $t_{(j)}$ and d_j is the number of patients who was died at time $t_{(j)}$. Then the Kaplan-Meier estimator of the survival function at time t is given by:

$$\hat{S}(t) = \prod_{j=1}^{r} \left\{ \frac{n_j - d_j}{n_j} \right\} \quad (3.8)$$

for $t_{(k)} \leq t_{(k+1)}$, j=1,2,…,r, with $\hat{S}(t)=1$ for $t < t_{(1)}$.

Where, n_j is the number of individuals who are at risk of dying at time t_j and d_j is the number of individuals who failed (died) at timet_j. The variance of Kaplan-Meier survival estimator is estimated using Greenwood's formula (Collett, 2003) given as:

$$\text{var}\big(\hat{s}(t)\big) = (\hat{s}(t))^2 \sum_{j=1}^{r} \frac{d_j}{n_j(n_j - d_j)} \quad (3.9)$$

3.6.1.4. Comparing Survival Functions

The simplest way of comparing the survival times obtained from two or more groups is to plot the Kaplan-Meier curves for these groups on the same graph. However, this graph does not allow us to say whether or not there is a real difference between the groups. Assessing whether or not there is a real difference between groups can only be done by utilizing statistical tests. Thus, the Mantel-Hanzel (1959), currently called the "log-rank" test is used for comparison of two or more survival distributions in this thesis work.

Let $t_1 \leq t_2 \leq \cdots t_m$ be the m distinct ordered death times across two groups. Suppose that d_j failures occur at $t_{(j)}$ and that n_j subjects are at risk just prior to $t_{(j)}$ (j = 1, 2,..., m). Let d_{ij} and n_{ij}be the corresponding numbers in group i (i = 1, 2). Then the log-rank test compares the observed number of deaths with the expected number of deaths for group i. consider the null hypothesis: $S(1) = S(2)$; i.e. there is no difference between survival curves in two groups. Given $n_j$and d_j the random variable d_{1j} has the hypergeometric distribution

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}$$

Under the null hypothesis, the probability of experiencing an event at $t_{(j)}$ does not depend on the group, i.e. the probability of experiencing an event at t_j is $\frac{d_j}{n_j}$. So that the expected number of deaths in group one is

$$E(d_{1j}) = e_{1j} = \frac{n_{1j} d_j}{n_j}$$

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

$$U_L = \sum_{j=1}^{m} (d_{1j} - e_{1j})$$

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

$$\mathrm{v_{1j}} = \mathrm{Var}(\mathrm{d_{1j}}) = \frac{\mathrm{n_{1j} n_{2j} d_{j(n_j - d_j)}}}{\mathrm{n_j}^2(\mathrm{n_j} - 1)}$$

So that the variance of U_L is given by

$$Var(U_L) = \sum_{j=1}^{m} v_{1j} = V_L$$

Under the null hypothesis, statistic has an approximate normal distribution with zero mean and variance V_L. This then follows $\frac{{U_L}^2}{V_L} \sim {x^2}_1$

The general form of the test statistic to test the equality of survival curves which can also be used by several alternatives to the log-rank test, such as the Wilcoxon test, may be defined as follows:

$$Q = \frac{\sum_{j=1}^{m} w_j (d_{ij} - \hat{e}_{1j})}{\sum_{j=1}^{m} {w_j}^2 \hat{v}_{1j}} \qquad (3.10)$$

Where:w_j are weights whose values depend on the specific test

The Cochran-Mantel-Haenszel Log Rank Test

The log rank test, sometimes called the Cox-Mantel test, is the most well known and widely used test statistic. This test is based on weights equal to one, i.e. wi = 1. Therefore, the log rank test statistic becomes:

$$Q = \frac{\sum_{j=1}^{m} w_j (d_{ij} - \hat{e}_{1j})}{\sum_{j=1}^{m} \hat{v}_{1j}} \qquad (3.11)$$

Log rank test is based on weights equal to one, i.e. $w_j = 1$. And it is appropriate when hazard functions for two groups are proportional over time, i.e. $\boldsymbol{h_1(t) = \psi h_2(t)}$.

The Generalized Wilcoxon Test

Gehan (1965) and Breslow (1970) generalized the Wilcoxon rank sum test to allow for censored data. This test uses weights equal to the number of subjects at risk at each survival time, i.e. wi = ni and is

called Wilcoxon or Generalized Wilcoxon test in most software packages. Thus the Wilcoxon test can be defined as: $Q = \frac{\sum_{j=1}^{m} w_j(d_{ij} - \hat{e}_{1j})}{\sum_{j=1}^{m} n_j{}^2 \hat{v}_{1j}}$ (3.12)

3.6.2. Modeling Survival Data

Both the non-parametric methods defined earlier are examples of univariate analysis; they describe the survival with respect to the factor under investigation, but necessarily ignore the impact of any others. In clinical investigations it is more common to have a situation where covariates potentially affect patient forecast. When investigating survival in relation to any one factor, it is often desirable to adjust for the impact of others. Moreover, while the log-rank test provides a P-value for the differences between the groups, it offers no estimate of the actual effect size.

Through a modeling approach of survival analysis can explore how the survival experience of a group of individuals depends on the values of one or more explanatory variables, whose values have been recorded for each individual at the time origin. There are two broad reasons for modeling survival data. One objective of the modeling process is to determine which combination of potential explanatory variables affects the form of the hazard function.

Another reason for modeling the hazard function is to obtain an estimate of the hazard function itself for an individual. A variety of models and methods have been developed for doing sort of survival analysis (Collett, 2003).

3.6.2.1. Sem-Parametric Survival Models

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 baseline hazards unspecified for actual distribution of lifetimes. Due to the easy concept and accessibility of software the most popular semi-parametric model is the Cox proportional hazards regression model (Collett, 2003).

3.6.2.1.1. The Cox Proportional Hazards Regression Model

The basic model for survival data is the Cox proportional hazard model. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of covariates, while keeping the baseline hazards unspecified and can take only positive values. With this parameterization, the Cox hazard function is specified as a function of time and the covariates:

$$h(t, \boldsymbol{X}, \boldsymbol{\beta}) = h_0(t).\exp(\boldsymbol{\beta' X}) \quad (3.13)$$

where, $h_0(t)$ is the baseline hazard function that characterizes how the hazard function changes as a function of survival time, $h(t, \boldsymbol{X}, \boldsymbol{\beta})$ represents the hazard function at time t with covariates $\mathbf{X} = (X_1, X_2, \dots, X_p)$, $\boldsymbol{\beta'} = (\beta_1, \beta_2, \dots, \beta_p)'$ is a column vector of p regression parameters, $\exp(\boldsymbol{\beta' X})$ characterizes how the hazard function changes as a function of subject covariates. The model (3.13) is referred to as Cox model, or Cox proportional hazards model or simply the proportional hazards model. There are two assumptions of proportional hazards model, those are:-

1. The hazard of occurrence of an event at any given time for an individual in one group is proportional to the hazard at that time for an individual in the other group. When there are covariates in the analysis, which are times dependent, this assumption may not hold. This can be verified by considering the hazard ratios of different individuals (Collett, 2003).

For two different individuals with covariates $\mathbf{X_1}$= (x_{11}, $x_{12,\dots}$, x_{1m})' and $\mathbf{X_2}$= (x_{21},$x_{22,\dots}$, x_{2m})', the proportion

$$\frac{\mathrm{h}(\mathrm{t}, \mathbf{X1}, \boldsymbol{\beta})}{\mathrm{h}(\mathrm{t}, \mathbf{X2}, \boldsymbol{\beta})} = \frac{h_0(t)\exp(\mathbf{X}_1'\boldsymbol{\beta})}{h_0(t)\exp(\mathbf{X}_2'\boldsymbol{\beta})} = \frac{\exp(\mathbf{X}_1'\boldsymbol{\beta})}{\exp(\mathbf{X}_2'\boldsymbol{\beta})} = \exp\left((\mathbf{X}'_1 - \mathrm{X}'_2)\boldsymbol{\beta}\right) \quad (3.14)$$

Called the hazards ratio, and clearly this ratio is independent of time which means that the log hazard ratio is constant at any given time.

2. The relationship between log hazard or log cumulative hazard and a covariate is linear. The Cox proportional hazards model can equally be regarded as linear model, as a linear combination of the covariates for the logarithm transformation of the hazard ratio given by:

$$log\left\{\frac{h(t,\boldsymbol{X},\boldsymbol{\beta})}{h_0(t)}\right\} = log\{e^{\boldsymbol{\beta' X}}\} = \boldsymbol{\beta' X} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (3.15)$$

The quantity $\boldsymbol{\beta' X} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ is called the linear combination of the Cox proportional hazards model.

The hazard function in the Cox model is called semi-parametric function since it does not explicitly describe the baseline hazard function, h_o (t). The survival function of the proportional hazard model is estimated as:

$$S(t, \boldsymbol{X}, \boldsymbol{\beta}) = e^{-H(t,\boldsymbol{\beta' X})} \quad (3.16)$$

Where,$H(t, X, \beta)$ is the cumulative hazard function at time t for a subject with covariate x. Since we have assumed that survival time is absolutely continuous; the value of the cumulative hazard function is expressed as:

$$H(t, X, \beta) = H_0(t).\exp(\boldsymbol{\beta' X}) \quad (3.17)$$

Consequently, from the proportional hazards function, we obtained the survivor function given by:

$$S(t, \boldsymbol{X}, \boldsymbol{\beta}) = [So(t)]^{\exp(\boldsymbol{\beta' X})} \quad (3.18)$$

Where, $H_0(t)$ is the baseline cumulative hazard function and $So(t)$ is the baseline survival function

3.6.2.1.2. Fitting the Cox Proportional Hazard Regression Model

Fitting the Cox model to observed survival data requires estimating the unknown regression coefficients ($\boldsymbol{\beta}$). Also, the baseline hazard function must be estimated. It turns out that these two components of the model can be estimated separately. The coefficients should be estimated first and the estimates are then used to construct an estimate of the baseline hazard function. The regression coefficients in the proportional hazards Cox model, which are the unknown parameters in the model, can be estimated using the method of maximum likelihood (Collett, 2003).

In Cox proportional hazards model we can estimate the vector of parameters **β** without having any assumptions about the baseline hazard, $h_0(t)$. As a consequence, this model is more flexible and an estimate of the parameters can be obtained easily.

Maximum Likelihood Estimation

Suppose the survival data based on n independent observations are denoted by the triplet (t_i, δ_i, $\mathbf{X_i}$), i=1, 2...n.

Where

t_i - the survival time for the i^{th} individual.

δ_i - an indicator of censoring for the i^{th} individual. Given by i=0 for censored and i= 1 for event experience

$\mathbf{X_i} = (X_{i1,} X_{i2...} X_{im})'$ - column vector of m covariates for individual i.

The full likelihood function for right censored data can be constructed as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} h(ti, \mathbf{X}i, \boldsymbol{\beta})^{\delta i} S(ti, \boldsymbol{X}i, \boldsymbol{\beta}) \quad (3.19)$$

Where, $h(ti, \mathbf{X}i, \boldsymbol{\beta}) = h_0(ti)e^{\boldsymbol{\beta}'\mathbf{X}i}$ is the hazard function for the i[th] individual.

$S(ti, \mathbf{X}i, \boldsymbol{\beta}) = [So(ti)]^{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}$ is the survival function for the i[th] individual. It follows that,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[h_0(ti)e^{\boldsymbol{\beta}'\mathbf{X}i}\right]^{\delta i} [So(ti)]^{\exp(\boldsymbol{\beta}'\mathbf{X}_i)} \quad (3.20)$$

The full maximum likelihood estimator of **β** can be obtained by differentiating the right hand side of equation (3.20) with respect to the components of **β** and the base line hazard, $h_0(t)$.

This implies that unless we explicitly specify the base line hazard, $h_0(t)$, we cannot obtain the maximum likelihood estimators for the full likelihood. To avoid the specification of the base line hazard, Cox (1972) proposed a partial likelihood approach that treats the baseline hazard as a nuisance parameter remove it from the estimating equation.

Partial Likelihood Estimation

Instead of constructing a full likelihood, we consider the probability that an individual experiences an event at time $t_{(i)}$ given that an event occurred at that time. Suppose that data are available for n individuals, amongst them there are r distinct failure times and n - r right-censored survival times, and assume that only one individual was died at each ordered failure time, so that there are no ties. The r ordered failure times will be denoted by $t_{(1)}<t_{(2)}<\ldots.< t_{(r)}$, so that $t_{(i)}$ is the i[th] ordered failure time. The set of individuals who are at risk at time $t_{(i)}$ is the i[th] ordered failure (experiences an event) time, and denoted by R ($t_{(i)}$). And let $\mathbf{X}_{(i)}$ be the vector of explanatory variables for an individual who experiences an event at $t_{(i)}$.

The partial likelihood function is derived by taking the product of the conditional probability of a failure at time $t_{(i)}$, given the number of individuals who are at risk of experiencing the event at time $t_{(i)}$. Then,

P(j[th] individual will experience an event at time $t_{(i)}$)$= \frac{\exp(\boldsymbol{\beta}'\boldsymbol{X}(i))}{\sum_{j\in R\,(t(i))} \exp(\boldsymbol{\beta}'\boldsymbol{X}j)}$ (3.21)

Where, the summation in the denominator is over all individuals in the risk set. Thus the partial likelihood is the product over all event time t(*i*) for *i*= 1,2,...,n of the conditional probability (3.21) to give the partial likelihood function and can be expressed in the form:-

$$L_p(\boldsymbol{\beta}, \boldsymbol{X}(i)) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{X}(i))}{\sum_{j \in \mathrm{R}(\mathrm{t}(\mathrm{i}))} \exp(\boldsymbol{\beta}' \boldsymbol{X} j)} \right]^{\delta i} \tag{3.22}$$

The expression assumes that there are no tied times, and designed in such a way that it excluded terms when $\delta i = 0$, as a result the equation in (3.22) becomes. The product is over the r distinct ordered survival times. The corresponding log-partial likelihood function is given by:

$$log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i)) = \sum_{i=1}^{n} \{ \boldsymbol{\beta}' \boldsymbol{X}(i) - log[\sum_{j \in \mathrm{R}(\mathrm{t}(\mathrm{i}))} \exp(\boldsymbol{\beta}' \boldsymbol{X} j)] \} \tag{3.23}$$

The maximum likelihood estimates of the regression parameters in the proportional hazards model can be found by maximizing the log-likelihood function in equation (3.23) using numerical methods. This maximization is accomplished using the Newton-Raphson procedure (Collett, 2003). The Newton-Raphson procedure is used to maximize the partial likelihood function based on the following iterative procedure. An estimate of the vector of **β**-parameters at the $(s+1)^{th}$ cycle of iterative procedure, $\widehat{\boldsymbol{\beta}}_{s+1}$, is given by:

$$\widehat{\boldsymbol{\beta}}_{s+1} = \widehat{\boldsymbol{\beta}}_{s} + \mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}_{s}) \mathbf{U}(\widehat{\boldsymbol{\beta}}_{s}), \text{ for s = 0, 1, 2, ...}$$

$$\mathrm{U}(\widehat{\boldsymbol{\beta}}_{s}) = \left(\frac{\partial log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_1}, \ldots\ldots\ldots, \frac{\partial log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_p} \right) \tag{3.24}$$

Where $\mathbf{U}(\widehat{\boldsymbol{\beta}}_{s})$ is the $p \times 1$ vector of first derivatives of the log-likelihood function in equation (3.23) with respect to the **β**-parameters and this quantity known as the vector of efficient scores evaluated at $\widehat{\boldsymbol{\beta}}_{s}$. $\mathbf{I}(\widehat{\boldsymbol{\beta}}_{s}) = -\frac{\partial^2 \log \mathrm{L_p}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}$ is the $p \times p$ matrix and known as observed information matrix.

$$I(\widehat{\boldsymbol{\beta}}_{s})_{pxp} = -\begin{bmatrix} \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_1^2}, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1}, \ldots\ldots\ldots\ldots\ldots\ldots, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_p} \\ \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1}, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_2^2}, \ldots\ldots\ldots\ldots\ldots\ldots, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_p} \\ \vdots \\ \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_p \partial \boldsymbol{\beta}_1}, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_p \partial \boldsymbol{\beta}_2}, \ldots\ldots\ldots\ldots\ldots\ldots, \frac{\partial^2 log L_p(\boldsymbol{\beta}, \boldsymbol{X}(i))}{\partial \boldsymbol{\beta}_p^2} \end{bmatrix} \tag{3.25}$$

$\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}_{s})$ is the inverse of the observed information matrix evaluated at $\widehat{\boldsymbol{\beta}}_{s}$is the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$, $var(\widehat{\boldsymbol{\beta}})$, can be approximated by the inverse of the information matrix evaluated at $\widehat{\boldsymbol{\beta}}_{s}$ i.e. $\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}}_{s})$.

The partial likelihood derived above is valid when there are no ties in the data set. But in most real situations tied survival times are more likely to occur. In addition to the possibility of more than one experience an event at a time, there might also be more than one censored observations at a time of event. To handle this real-world fact, partial likelihood algorithms have been adopted to handle ties. There are three approaches commonly used to estimate regression parameters when there are ties. These are Breslow (1974), Efron (1977) and Cox (1972) approximations (Collett, 2003).The most popular and easy approach is Breslow's approximation. In many applied settings there will be little or no practical difference among the estimators obtained from the three approximations. Because of this, and since the Breslow approximation is more commonly available, otherwise, analysis presented in this study was based on it.

The Breslow Approximation

This approximation is proposed by Breslow and Peto by modifying the partial likelihood which takes the following form;-

$$L_B\big(\boldsymbol{\beta}, \boldsymbol{X}(i)\big) = \prod_{i=1}^{n} \frac{\prod_{\mathrm{i:yi=y(j),}\delta i=1,}^{\mathrm{n}} \exp(\boldsymbol{\beta}'\mathrm{Xi})}{\left[\sum_{j\in \mathrm{R\ (t(i))}} \exp(\boldsymbol{\beta}'\boldsymbol{X}j)\right]^{d_i}} \qquad (3.26)$$

Where xi is the sum of covariates over d_i subjects at time $t_{(i)}$, d_i is the number of experienced an event occurred at time $t_{(i)}$.

Now the partial log likelihood of (3.26) is given as .

$$\log \mathrm{L_B}\big(\boldsymbol{\beta}, \mathbf{X}(\mathrm{i})\big) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left[\boldsymbol{\beta}'\mathrm{Xi} - \mathrm{Log}\Big(\sum_{\mathrm{j}\in \mathrm{R\ (t(i))}} \exp(\boldsymbol{\beta}'\mathrm{Xj}) \Big) \right] \qquad (3.27)$$

We obtain the Breslow maximum partial likelihood estimator, adjusted for tied observation, by differentiating equation (3.27) with respect to the component of β and setting the derivative equal to zero and solving for the unknown parameters.

3.6.2.1.3. Assessment of Model Adequacy

The adequacy of the model needs to be assessed after the model has been fitted to observed survival data. Model-based inferences depend completely on the fitted statistical model. For these inferences to be valid, the fitted model must provide an adequate summary of the data upon which it is based.

Indeed, the use of diagnostic procedures for model checking is an essential part of the modeling process.

As model assumptions checking are based on residuals, we will first introduce the different types of residuals used in survival analysis, and more specifically in the semi-parametric proportional hazards model. Residuals are values that can be calculated for each observation and have the feature that their behavior is known, at least approximately, when the fitted model is satisfactory.

Different types of residuals are typical for survival analysis due to the fact that censoring has to be taken into account. Ordinary residuals from linear or generalized linear models are therefore often not applicable.

i. Cox-Snell Residuals

The residual that is widely used in the analysis of survival data is the Cox-Snell residual, it is a particular example of the general definition of residuals given by Cox and Snell (1968). The Cox-Snell residual for the i^{th} individual, i=1, 2, ..., n, is given by:

$$rc_i = \exp(\widehat{\boldsymbol{\beta}}'\mathbf{x}_i)\widehat{\mathrm{H}}_o(t)) \qquad (3.28)$$

Where $\widehat{\mathrm{H}}_o(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of that individual. Note that from equation (3.28), the Cox-Snell residual, rc_i, is the value of $\widehat{\mathrm{H}}_i(t_i) = -log\hat{S}_i(t_i)$, where $\widehat{\mathrm{H}}_i(t_i)$ and $\hat{S}_i(t_i)$ are the estimated values of the cumulative hazard and survivor function of the i^{th} individual at t_i.

ii. Schoenfeld Residuals

Two disadvantages of Cox–Snell residuals depend heavily on the observed survival time and require an estimate of the cumulative hazard function. These disadvantages are overcome in a residual proposed by Schoenfeld (1982). These residuals were originally termed partial residuals, but are now commonly known as Schoenfeld residuals.

Schoenfeld residual differs from those considered previously in one other important respect. This is that there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model.

The i^{th} partial or Schoenfeld residual for X_j, the j^{th} explanatory variable in the model, is given by:

$$rSik = \delta_i\{\mathrm{xji} - \frac{\sum_{l\epsilon R(t_i)} X_{ji}\exp(\widehat{\boldsymbol{\beta}}\boldsymbol{X}_1)}{\sum_{l\epsilon R(t_i)}\exp(\widehat{\boldsymbol{\beta}}\boldsymbol{X}_1)}\}, \qquad (3.29)$$

Where, X_{ji} is the value of the j^{th} explanatory variable, j=1,2,..,p, for the i^{th} individual in the study, and if individuals in the risk set are indexed by l and $R(t_i)$ is the set of all individuals at risk at time of t_i.

Schoenfeld residuals are also used to check the proportionality of the covariates over time that is to check the validity of the proportional hazards assumption. If the model fits well then the residuals are randomly distributed without any systematic pattern around the zero line, reference line. For greater diagnostic power the scaled schoenfeld residual is preferred. The scaling can be done on the variance of the i^{th} subject Schoenfeld residuals. If the plot of scaled Schoenfeld residuals versus the logarithm of time is a random, smooth, straight line about zero the proportional hazards assumption will be satisfied.

iii. 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, $j = 1, 2, \ldots, p$ 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 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 (Klein and Moeschberger 1997), which are modifications of Schoenfeld residuals and are defined for all the observations, and $Var(\hat{\beta})$ is the 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. $\Delta_i\hat{\beta}_j = \hat{\beta}_j - \hat{\beta}_{j(-1)}$ 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.

iv. Overall Goodness of Fit

One method of checking goodness of fit of the model is to use R^2. In proportional hazards regression model as in all regression analyses there is no single, simple method of calculating and interpreting

R^2, a measure analogous to R^2 in linear regression, as a measure of model performance of a Cox model is also used. Hosmer and Lemeshow (1999) recommended a statistic based on the log-likelihood of the model, which is defined as follows;

$$R_p{}^2 = 1 - \left\{exp\left[\frac{2}{n}\left(L_0 - L_p\right)\right]\right\} \tag{3.30}$$

where L_0 the log-likelihood for null model i.e. the model with no covariates, L_p the log-likelihood for the fitted final Cox model with p covariates and n number of observations included in the study. For reason of censored lower the value of R^2 indicate that the better fit of the model.

Under the assumption of proportional hazards, there are three different tests for model assessment (the significance of the coefficients): the partial likelihood ratio test, the Wald test and the score test

a. Partial Likelihood Ratio Test

Partial likelihood ratio test is the easiest test to compute and the best of the three tests for assessing the significance of the fitted model for testing the significance of a subset of q explanatory variables from p explanatory variables.

The partial likelihood ratio test statistic,U_{LR}, is given by:

$$U_{LR} = 2\{logL_p(\hat{\beta}) - logL_p(\beta_0)\} \approx \text{H0 } \chi^2_q \tag{3.31}$$

Where, $logL_p(\widehat{\boldsymbol{\beta}})$ - the log-partial likelihood evaluated at $\widehat{\boldsymbol{\beta}}$

$logL_p(\boldsymbol{\beta}_0)$ - the log-partial likelihood evaluated at $\boldsymbol{\beta}_0 = \mathbf{0}$.

Under the null hypothesis, H_o: $\boldsymbol{\beta}_{q\times 1} = \boldsymbol{\beta}_0 = \mathbf{0}_{q\times 1}$,that all q coefficients are simultaneously equal to zero, and under mathematical regularities and large sample size conditions U_{LR} follows a chi-square distribution with q degree of freedom, χ^2_q.

b. Wald test

The Wald test is used to check the overall goodness of fit as well as checking the significance of each parameter of the model.

Under the hypothesis, H_o: $\boldsymbol{\beta}_0 = \mathbf{0}_{q\times 1} = (0, 0, 0, \dots, 0)'$vs H_1: at least one $\beta_i \neq 0$, $\widehat{\boldsymbol{\beta}}$ will be asymptotically normally distributed with mean $\mathbf{0}$ covariance matrix estimated by $\widehat{\text{V}}\text{ar}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{I}\left(\widehat{\boldsymbol{\beta}}\right)^{-1}$. Then, the Wald test statistic, $U^2{}_w$, given by:

$$U^2 = \left(\hat{\beta} - \beta_0\right)' I\left(\hat{\beta}\right)^{-1}\left(\hat{\beta} - \beta_0\right) = \hat{\beta}' I\left(\hat{\beta}\right)^{-1}\hat{\beta} \approx \text{H0 } \chi^2_q. \tag{3.32}$$

Follows a chi-square distribution with q degree of freedom, $X^2(q)$.

c. Score test

The score test statistic, to test H_0: $\boldsymbol{\beta}_0 = \mathbf{0} = (0, 0, 0\ldots.0)'_{q \times 1}$ is defined as

$$U^2{}_{SC} = U(\beta_0)'I(\beta_0)^{-1}U(\beta_0) \tag{3.33}$$

Where,$U(\boldsymbol{\beta}_0)$ and $\boldsymbol{I(\beta}_0)^{-1}$ are the score vector and inverse of the observed information matrix evaluated at $\boldsymbol{\beta}_0$. Under null hypothesis and for large sample $U^2{}_{SC}$ is asymptotically distributed as chi-squared with q degree of freedom, X^2_q.

3.6.3.1.4 Checking for the Proportional Hazards Assumption

The basic assumption of the Cox model is the assumption of proportional hazards. There are several methods for verifying that a model satisfies proportionality assumption.

Graphical Methods

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

$$S(t, X) = \big(S_0(t)\big)^{exp(\sum_{i=1}^{p}\boldsymbol{\beta}_i x_i)} \tag{3.34}$$

Where, $X = (x_1, x_2, \ldots, x_p)$ are the values of the vector of explanatory variables for a particular individual. When taking the logarithm twice, we can easily get

$$\log\big(-\log\big(S(t,X)\big)\big) = \sum_{i=1}^{p}\beta_i x_i + log\left(-log\left(\big(S_0(t)\big)\right)\right) \tag{3.35}$$

Then the difference in log-log curves corresponding to two different individuals with variables $x_1 = \big(x_{11}, x_{12}, \ldots, x_{1p}\big)$ and $x_2 = \big(x_{21}, x_{22}, \ldots, x_{2p}\big)$ is given by:-

$$\log\big(-\log\big(S(t,X_1)\big)\big) - \log\big(-\log\big(S(t,X_2)\big)\big) = \sum_{i=1}^{p}\beta_i(x_{1i} - x_{2i}) \tag{3.36}$$

Which does not depend on t. This relationship is very helpful to identify situations where we may have proportional hazards. By plotting estimated log (-log (survival probability)) 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.

Looking at the K-M curves and log(-log(survival probability)) is not enough to be certain of proportionality since they are univariate analysis and do not show whether hazards will still be proportional when a model includes many other predictors. But they support our argument for proportionality. There are some other statistical methods for checking the proportionality.

Adding Time-Dependent Covariates in the Cox Model

We create time-dependent covariates by creating interactions of the predictors and a function of survival time and including them in the model. For example, if the predictor of interest is x_j, then we create a time-dependent covariate $x_j(t)$, $x_j(t) = x_j * g_i(t)$; where $g_i(t)$ is some specified function of time, usually $g_i(t) = \ln(t)$ is used by most software's by default . Then the model assessing PH assumption for x_j adjusted for other covariates is:

$$h\big(t, x(t)\big) = h_0(t) exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_p x_p + \delta x_j * g_i(t)\right) \quad (3.37)$$

Where,$x(t) = \big(x_1, x_2 \ldots, x_j \ldots, x_p\big)$; x_j 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

$$W = \left(\frac{\hat{\delta}}{se\big(\hat{\delta}\big)}\right)^2$$

The likelihood ratio test calculates the likelihood under null hypothesis, L_0 and the likelihood under the alternative hypothesis, L_a. The LR statistic is then

$$LR = -2\log\left(\frac{L_0}{L_a}\right) = -2(l_0 - l_a)$$

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

3.6.2.2 Parametric Survival Regression Models

In the analysis of survival data, survival models can also be used in addition to hazards model. One advantage of such models is that the proportionality assumption of the hazards is not required. The parametric survival regression models work analogous to the multiple linear regression of logarithm of survival time on explanatory variables. Such survival models are termed as parametric accelerated failure time models or simply AFT models. Because these models work on survival, the complementary concept of hazard, the sign of the regression coefficients in an AFT model will be opposite to those in PH models (Klein and Moeschberger 1997).

Most commonly used parametric Survival Regression models are Exponential, Weibull, Log-Logistic and Log-normal. Exponential and Weibull parametric models can work both in PH and in AFT models. These models are equally appropriate viewed in either model. And one can transform regression coefficients computed in PH model into the regression coefficient in AFT model or vice versa for Exponential and Weibull parametric survival models. That means:-

- For exponential $\beta_i = -\alpha_i$, the exponential PH and AFT are in fact the same model, except that the parameterization is different, hence HR=exp $(-\alpha_i)$ is the hazard ratio of the ith group with the reference groups.
- For weibull, $\beta_i = -\rho\alpha_i$, where ρ is the shape parameter and hence, HR=exp $(-\rho\alpha_i)$ is the hazard ratio of the ith group with the reference groups.

Other parametric survival models such as Log-Logistic and Log-normal work only in AFT model as these models do not fit into the proportional hazards frame work.

3.6.2.2.1. The Exponential Survival Regression Model

The simplest model for the hazard function is to assume that it is constant over time. The hazard of death at any time after the time origin the study is then the same, irrespective of the time elapsed (Collett, 2003).

From the constant baseline hazard function; $h_{0i=}\lambda$, the corresponding survivor function is:

$$S_0(t) = \left\{-\int_0^t \lambda du\right\} = exp(-\lambda t) \quad (3.38)$$

And so the implied probability density function of the survival times is

$$f_0(t) = \lambda\, exp(-\lambda t) \quad (3.39)$$

For the time data and skewed to the right and with distribution of the time is exponentially, the time of survival for covariates matrix *X*, which is called, accelerated failure time, expressed as:

T = exp ($\beta' x_i$+ ε)

This model can be transformed by taking the natural log of each side of the equation as:

lnT = $\beta' x_i$+ ε*

Where ε is the error component and $\beta' = (\beta 1, \beta 2, \ldots \beta k)'$

The exponential model (t ~exp($\boldsymbol{\alpha}$)) is the simplest parametric model and assumes a constant risk or hazard over time, which reflects the property of the distribution appropriately called "lack of memory". Reparameterizing the exponential distribution using $\beta_i = -\alpha_i$, the exponential regression model for the k covariates and i [th] individual is expressed as:

$h_i(t, \beta, \mathbf{X}i) = h_{0i} \exp(\beta_0 + \beta_1 X_{11} + \cdots + \beta_1 X_{1k})$ (3.40)

For the exponential regression survival models the hazard ratio, with one unit increase in covariate Xi while other covariates being held fixed, at a time t is HR (Xi) = exp(-β_1) .

Generally exponential hazard model can be presented as

$h_i(t) = \lambda \exp(\beta' x_i)$

$S_i(t) = (\exp -\lambda \exp(\beta' x_i) t)$

$f_i(t) = \lambda \exp(\beta' x_i)(\exp -\lambda \exp(\beta' x_i) t)$

3.6.2.2.2 The Weibull Survival Regression Model

The Weibull distribution is a generalization of the exponential distribution. However, unlike the exponential distribution, it does not assume a constant hazard rate and therefore has broader application. The distribution was proposed by Weibull (1939) and its applicability to various failure situations discussed again by Weibull (1951). Suppose that survival times are assumed to have a Weibull distribution with scale parameter and shape parameter, the Weibull density function can be expressed as:

$$f(t, \mu, \alpha) = \frac{\alpha}{\mu} \left(\frac{t}{\mu}\right)^{\alpha - 1} \exp\left(\left(-\frac{t}{\mu}\right)\right)^{\alpha} \quad (3.41)$$

The baseline hazard model for Weibull distributed event times is given by:

$$ho(ti, \mathrm{X}) = \frac{a}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1}$$

Reparameterizing the Weibull distribution using $\rho = \boldsymbol{\sigma}^{-1}$, $\lambda = \mu^{-\alpha}$ then ho= $\lambda\rho t^{\rho-1}$ would be the baseline hazard function. Now incorporate covariates matrix X in the hazard function, the Weibull regression model becomes:

$$h_i(t, \beta, \mathbf{X}\mathrm{i}) = \lambda\rho t^{\rho-1}\exp\left(\beta 0 + \beta_1 X_{11} + \cdots + \beta_1 X_{1k}\right) \quad (3.42)$$

Generally Weibull survival regression model can be presented as

$$h_i(t) = \lambda\rho t^{\rho-1}\exp(\beta' x_i)$$

$$S_i(t) = (\exp -\lambda\exp(\beta' x_i)t^{\rho}) \quad (3.43)$$

$$f_i(t) = \lambda\rho t^{\rho-1}\exp(\beta' x_i)(\exp -\lambda\exp(\beta' x_i)t^{\rho}) \quad (3.44)$$

The event time of the i^{th} subject is then characterized by the Weibull distribution with scale parameter λ and shape parameter ρ. The shape of the hazard function critically depends up on the values of ρ that means:-

If $\rho < 1$: hazard decreases monotonically with time

If $\rho > 1$: hazard increases monotonically with time

If $\rho = 1$: constant hazard (equivalent to exponential distribution)

Thus, all subjects share the shape parameter but differ with respect to their scale parameter. The model assumes that individual i and j with covariates X_i and X_j have proportional hazard function of the forms. A different parameterization is used with intercept v and covariate effects γi having relationship with original parameterization as $\beta = \frac{-\gamma i}{\boldsymbol{\sigma}}$ and $\mu = \exp(v)$.

$\frac{h(t;X_i)}{h(t;X_j)} = \frac{exp(\beta' X_i)}{exp(\beta' X_j)} = exp\left(\beta'\left(X_i - X_j\right)\right)$, the quantities $\exp(\beta)$ can be interpreted as hazard ratios

3.6.2.2.3. Log-logistic Survival Regression Model

The log-logistic model assumes that the disturbance term, in an accelerated failure time, has a standard logistic distribution. Covariate incorporated log logistic accelerated failure time may be expressed as:

$$\mathrm{T} = \exp((\mu + \beta' x_i) + \varepsilon\boldsymbol{\sigma}) \quad (3.45)$$

This model can be transformed by taking the natural log of each side of the equation as:

$\log T_i = \mu + \boldsymbol{\beta}'\boldsymbol{x}_i + \delta\varepsilon_i$

Where, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$, μ is intercept, δ is scale parameter and ε_i is a random variable used to model the deviation of values of $\log T_i$ from the linear part of the model. ε_i is assumed to have a particular probability distribution supposed to be followed by the survival time under study.

Suppose a random variable T, representing survival time, follows Log-Logistic distribution with shape parameter k and scale parameter θ with probability distribution

$$f(t) = \frac{e^{\theta} k t^{k-1}}{\left(1+e^{\theta} t^{k}\right)^{2}} \tag{3.46}$$

The baseline hazard function is given by

$$h_0(t) = \frac{e^{\theta} k t^{k-1}}{1+e^{\theta} t^{k}} \text{ for } k > 0 \tag{3.47}$$

If $k \leq 1$ the hazard function decreases monotonically and if $k > 1$, the hazard function has single mode (Collett 2003).

The corresponding baseline survival function for the equation as

$$S(t, z) = [1 + \exp(z)]^{-1} \tag{3.48}$$

Where, z is the standardized log-time outcome variable, that is $y - \frac{\boldsymbol{\beta}' \boldsymbol{x}_i}{\delta}$ and y =ln(t).

Since (t, z) is the probability of surviving to time *t* for any given time *t*, the ratio is often called the *odds* of surviving to time *t*. The ratio $\frac{S(t,X,\delta)}{1-S(t,X,\delta)}$ =exp(-z) is often called the odds of surviving time.

Therefore, with one unit increase in covariate while other covariates being held fixed, the survival odds ratio (SOR) at a time t is given by

$$\text{SOR} = \frac{\frac{\exp\left(-y-\boldsymbol{\beta i}'(\boldsymbol{x}_i+1)\right)}{\delta}}{\frac{\exp\left(-y-\boldsymbol{\beta i}'(\boldsymbol{x}_i)\right)}{\delta}} = \exp\left(\frac{\boldsymbol{\beta i}}{\delta}\right) \tag{3.49}$$

This is independent of time

3.6.2.2.4 Log-normal Survival Regression Model

When a random variable T is said to have a Log-normal distribution with parameters μ and σ the probability density function is given as follows:

$$f(\mathrm{t}, \mu, \sigma_2) = \frac{1}{\sigma\sqrt{(2\pi)}}\, t^{-1} exp\left\{-\left(\frac{(\log t-\mu)^2}{2\sigma^2}\right)\right\}, \text{For } 0 \le t < \infty, \sigma > 0 \quad (3.50)$$

From which the survivor and hazard functions can be derived. The survivor function is given by

$$s(\mathrm{t}, \mu, \sigma_2) = 1 - \Phi\left(\frac{log(t)-\mu}{\sigma}\right) \quad (3.51)$$

Where $\Phi(.)$ is the standard normal distribution function. The hazard function can be found from the relation $h(t) = f(t)/S(t)$. This function is zero when $t = 0$, increases to a maximum and then decrease to zero as t tends to infinity (Collett, 2003).

3.6.2.2.5 Fitting parametric Survival Regression Models

The survival likelihood for Weibull distributed survival data with event times and right censored data is generally given by

$$L = \prod_{i=1}^{n}\left\{\left(t_i\lambda\rho x_i{}^{\rho-1}\exp(-\lambda x_i{}^{\rho})\right)^{\delta_i}(\exp(-\lambda x_i{}^{\rho}))^{1-\delta_i}\right\} \quad (3.52)$$

Resulting in the log likelihood function

$$l = d\log(\lambda\rho) + (\rho-1)\sum_{i=1}^{n}\delta_i log x_i - \lambda\sum_{i=1}^{n}x_i^{\rho} \quad (3.53)$$

with $d$the total number of events. Maximum likelihood estimators can be obtained by equating the first derivatives of l with respect to λ and ρ to zero and we get.

$$\hat{\lambda} = \frac{d}{\sum_{i=1}^{n} x_i^{\hat{\rho}}} \text{ and } \quad \frac{d}{\hat{\rho}} + \sum_{i=1}^{n}\delta_i log x_i - \frac{d}{\sum_{i=1}^{n} x_i^{\hat{\rho}}}\sum_{i=1}^{n} x_i^{\rho}\, log x_i = 0$$

which is nonlinear in $\hat{\rho}$ and can only be solved by a numerical procedure such as the Newton Raphson algorithm.

The likelihood function is derived from the log-linear function of the model defined in equation (3.51). The likelihood function of n observed survival times, $t_1, t_2, \ldots, t_n$ for the log-linear form of the parametric Survival Regression model is given by

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{n}[f_i(t_i)]^{\delta_i}\,[S_i(t_i)]^{(1-\delta_i)} \quad (3.54)$$

Where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions for the i^{th} individual at time t_i and δ_i is the event indicator for the observation and has value zero for censored and one for uncensored

individuals. If $f_{\varepsilon_i}(z_i)$ and $S_{\varepsilon_i}(z_i)$ are probability density function and survival function respectively of the random variable $\boldsymbol{\varepsilon_i}$ in equation (3.54) in such a way that

$S_i(t_i) = S_{\varepsilon_i}(z_i)$ and $f_i(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}(z_i)$

Where, $z_i = \left(\frac{\log t_i - (\mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_P x_{Pi})}{\sigma}\right)$

The resulting likelihood function using survival function and density function of assumed probability distribution represented by random variable ε_i is as follows:

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} \left[f_{\varepsilon_i}(z_i)\right]^{\delta_i} \left[S_{\varepsilon_i}(z_i)\right]^{(1-\delta_i)}$$

The log-likelihood function is:

$$\log\big(L(\beta, \mu, \sigma)\big) = \sum_{i=1}^{n} \big\{-\delta_i \log(\sigma) + \delta_i \log f_{\varepsilon_i}(z_i) + (1 - \delta_i) \log S_{\varepsilon_i}(z_i)\big\} - \sum_{i=1}^{n} \delta_i \log t_i \qquad (3.55)$$

The term $(-\sum_{i=1}^{n} \delta_i \log t_i)$ is omitted as it does not involve any unknown parameters. Hence the full log-likelihood function is given by

$$\log\big(L(\beta, \mu, \sigma)\big) = \sum_{i=1}^{n} \big\{-\delta_i \log(\sigma) + \delta_i \log f_{\varepsilon_i}(z_i) + (1 - \delta_i) \log S_{\varepsilon_i}(z_i)\big\} \qquad (3.56)$$

The maximum likelihood estimates of the parameters are estimated by using iterative Newton-Raphson procedure.

3.6.2.3.6 Assessment Adequacy of parametric Survival Regression Models

Once the model has been finalized, it is necessary to test the overall fit of it. For assessing the goodness of fit of a parametric Survival Regression model different methods can be applied. In this study Cox-Snell residuals plot and maximum likelihood ratio Test statistics are applied. The overall fit of the parametric Survival Regression model is evaluated by using the diagnostic plot of Cox-Snell residuals as described in Cox regression model. However, the calculation of Cox-Snell residuals in parametric Survival Regression is different from that of Cox regression model because of the difference in formulation between these two families. The Cox-Snell residuals for parametric survival regression model are calculated by using standardized residuals which are defined as

$$rS_i = \frac{\{\log t_i - (\hat{\mu} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi})\}}{\hat{\sigma}} \qquad (3.57)$$

Where $\hat{\mu},\ \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ and $\hat{\sigma}$ are the maximum likelihood estimates of $\mu, \beta_1, \beta_2, \ldots, \beta_p$ and σ respectively. In the Cox-Snell residuals plot, if the plotted points lie on a line that has an intercept zero and slope unity, then it indicates that the fit is good.

Maximum Likelihood Ratio Test

Maximum likelihood ratio test is the easiest test to compute and the best tests for assessing the significance of the fitted model for testing the significance of a subset of q explanatory variables from p explanatory variables.

The maximum likelihood ratio test statistic, U_{ML}, is given by:

$U_{ML} = 2\{logL_p(\widehat{\boldsymbol{\beta}}) - logL_p(\boldsymbol{\beta}_0)\} \approx$ H0 X_q^2

Where,

$logL_p(\widehat{\boldsymbol{\beta}})$ - Log-maximum likelihood evaluated at $\widehat{\boldsymbol{\beta}}$

$logL_p(\boldsymbol{\beta}_0)$ - Log-maximum likelihood evaluated at $\boldsymbol{\beta}_0 = \mathbf{0}$.

Under the null hypothesis, H_o: $\boldsymbol{\beta}_{q\times 1} = \boldsymbol{\beta}_0 = \mathbf{0}_{q\times 1}$,that all q coefficients are simultaneously equal to zero, and under mathematical regularities and large sample size conditions U_{LR} follows a chi-square distribution with q degree of freedom, X_q^2.

3.7. Model Development

The methods of selecting a subset of Variable and Comparison of Models in a PHs regression model are essentially similar to those used in any other regression models. The most common methods are purposeful selection, step-wise (forward selection and backward elimination) and best sub-set selections. Survival analysis using Cox regression method begins with a thorough univariable analysis of the association between survival time and all important covariates (Hosmer and Lemeshow, 1999).

3.7.1. Variable Selection

When the number of variables is relatively large, it can be computationally expensive to fit all possible models. In this situation, automatic routines for variable selection that are available in many software packages might seem an attractive prospect. These routines are based on forward selection, backward elimination or a combination of the two known as the stepwise procedure. Thus, instead of

using automatic variable selection procedures, the following general strategy for model selection is recommended by Collet (2003).

1. The first step is to fit models that contain each of the variables one at a time. The values of $-2log\hat{L}$ for these models are then compared with that for the null model. The null model is a model in which there are no explanatory variables in the linear component of the hazard model and used to determine which variables on their own significantly reduce the value of $-2log\hat{L}$.
2. The variables that appear to be important from step 1 are then fitted together in a multivariable model. In the presence of certain variables others may cease to be important. Consequently, those variables that do not significantly increase the value of $-2log\hat{L}$ when they are omitted from the model can now be discarded. We therefore compute the change in the value of $-2log\hat{L}$ when each variable on its own is omitted from the set. Only those that lead to a significant increase in the value of $-2log\hat{L}$ are retained in the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn should be examined.
3. Variables that were not important on their own, and so were not under consideration in step 2, may become important in the presence of others. These variables are therefore added to the model from step 2, one at a time, and any that reduce $-2log\hat{L}$ significantly are retained in the model. This process may result in terms in the model determined at step 2 ceasing to be significant.
4. A final check is made to ensure that no term in the model can be omitted without significantly increasing the value of$-2log\hat{L}$, and that no term not included significantly reduces$-2log\hat{L}$.

When using this selection procedure, rigid application of a particular significance level should be avoided. In order to guide decisions on whether to include or omit a term, the significance level should not be too small. A level of around 20% - 25% is recommended.

3.7.2. Model Comparison

Different models can be compared on the basis of the variables selected and their coefficients in each model Cox-Snell residuals plot and AIC. If the models being compared have a similar set of covariates that have entered in the respective final models, it can be interpreted as all models are equally good or bad as far as the identification of important covariates associated with the outcome.

However, it is difficult to interpret either way if the selected variables in the models being compared are different, as there is no way of knowing the truth. The precision of the regression coefficients is another criterion that can be used to compare different models. The smaller the standard error, the more precise an estimate is expected to be a model with more precise coefficients can be considered as a more precise model. In this study Cox-Snell residuals plot and AIC are applied to compare different models in similar set of covariates.

i. Comparison Based on Cox-Snell Residuals Plots

The construction of the Cox-Snell residuals plot is explained in the respective sections above. Broadly, all models require the plot to be a straight line, passing through the origin to qualify for a good fit. So the plots under each model can be visually assessed as to which one of them is close to the requirements of a good fit.

ii. Comparison based on AIC

Akaike's Information Criterion (AIC) proposed by (Akaike, 1974) may also be used when comparing can be used to compare models that are not nested. The AIC of a model may be defined as

$$AIC = -2Log(L) + 2(p + k + 1) \quad (3.58)$$

Where L is the log-likelihood, p is the number of covariates and k is the number of model-specific ancillary parameters. A lower value of the AIC suggests a better model.

3.8. Ethical Considerations

Ethical clearance was obtained from Jimma University, College of Natural science Department of Statistics. And, the official ethical clearance also was obtained from Hossana Queen Elleni Mohamad Memorial Hospital medical director. To keep the confidentiality, the data collectors extracted the necessary data from the patient baseline and follow up card. Moreover, no personal identifier was used on data collection form. The recorded data was not accessed by a third person except the researcher, and was kept confidentially. Thus, the data obtained by checklist was organized by the researcher.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1. Descriptive Analysis of HIV patients

The study included 933 HIV patients, who started ART in Hossana Queen Elleni Mohamad Memorial Hospital between 1st February 2011 to 1st January 2016. Among those patients 15.3% were dead cases and the rest 84.7% were censored. The baseline socio-demographic variables of the cohort are summarized in Table 4.1.1 of the appendix. Out of these patients 501(53.7%) were females, death proportion were 14.2%. In case of age 530(57%) of the patients were less than 40 years old, death proportion were 13.2%. The settlement of the patients were 481(51.5%) lived in rural out of Hossana town, death proportion were 15.8% .When we see base line weight, 570(61%) were greater than or equal to 50kgms . Of the total patients, 226(24.2%) were single, 242(26%) were ambulatory and 234(25%) were not educated. Furthermore, the clinical variables of the patients are summarized in table 4.1.2 0f the appendix. Among the patients, the regimen TDF-3TC-EFV was frequently prescribed 398(43%) patients and 295(31.6%) were staged clinically as IV. In case of ART Adherence were 174(18.6%) not followed the drug combinations properly. From the total of HIV-infected patients 190(20.4%) were TB co- infected, 172(18.4%) were abuse Substance and 426(45.7%) were CD4 count less than 200 cells/μl.

The overall mean estimated survival time of patients under the study was 51.50 (95% CI: 50.30, 52.73) months. The minimum follow up time was 1month and the maximum was 60 months. Females and males have almost the same survival times i.e., 52.18 months with 95% CI, (50.55, 53.81) for females and 50.9 months with 95% CI, (49.13, 52.72) for males. Patients with age less than 40 had survived for about 52.19 months with 95% CI,(50.47, 53.91) while the mean survival time for older patients was 49.3 months with 95% CI, (47.70 ,51.99). The mean survival time of HIV-infected patients from rural, 52.35 months with 95% CI, (50.15 , 53.55) is greater than that of from urban, 51.14 months with its 95% CI (49.41 , 52.87). The mean survival time of patients based on different socio-demographic and clinical variables are summarized in Table 4.1(1 and 2).

4.2. Comparison of Survival Experience

The Kaplan-Meier survivor estimator is used to investigate the significance differences between the survival probabilities of different categories. In this study overall graph of the Kaplan-Meier survivor function showed that relatively small number of the deaths occurred in the earlier months of ART treatment which given in Figure 4.1. Separate graphs of the estimates of the Kaplan-Meier survivor functions for the covariates such as place of residence, baseline weight, functional status, marital status ,drug regimen, education level, ART adherence, WHO clinical stage, baseline CD4, substance use and TB co-infection are also presented in figures 4.(2 to 3) and figures 4.(9 to12) of the Appendix . In order to see whether there is difference in survival experience between different categories of individuals. Most of the graphs show differences between different categories; such as baseline weight, education level, functional status, WHO clinical stage, ART adherence, baseline CD4, substance use and TB co-infection. However, residence, marital status and drug regimen graphs did not show clear differences between the intended categories. In general, the pattern of one survivorship function lying above another means the group defined by the upper curve had a better survival than the group defined by the lower curve.

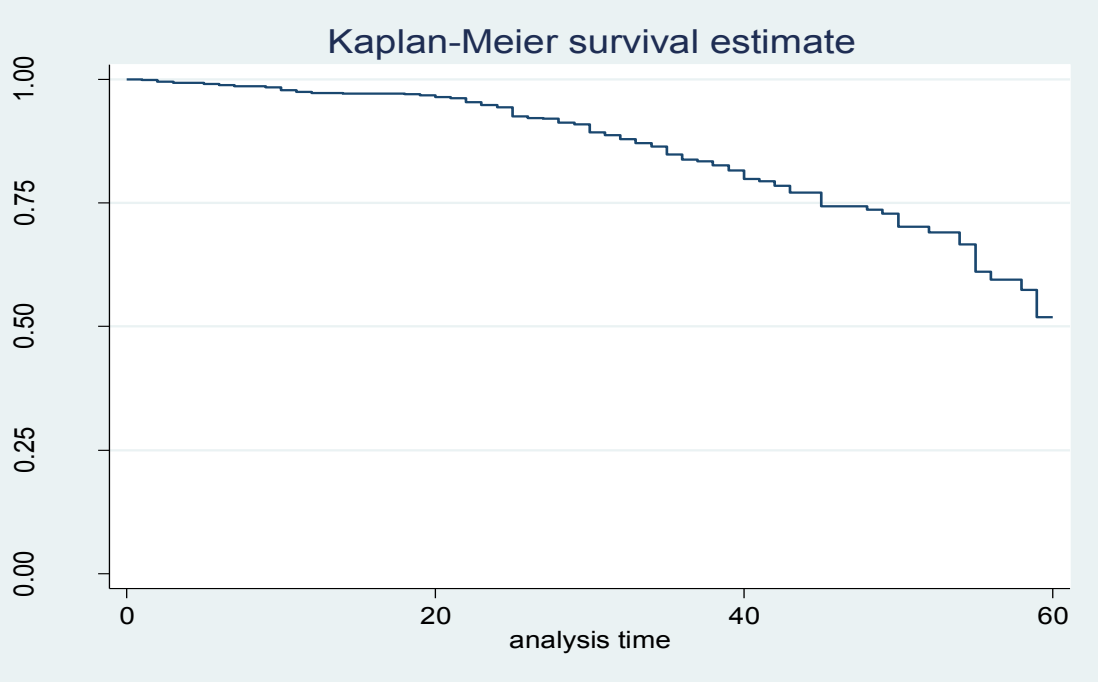


Figure 4.1: the plot of the overall estimate of Kaplan-Meier survivor function of HIV Patients

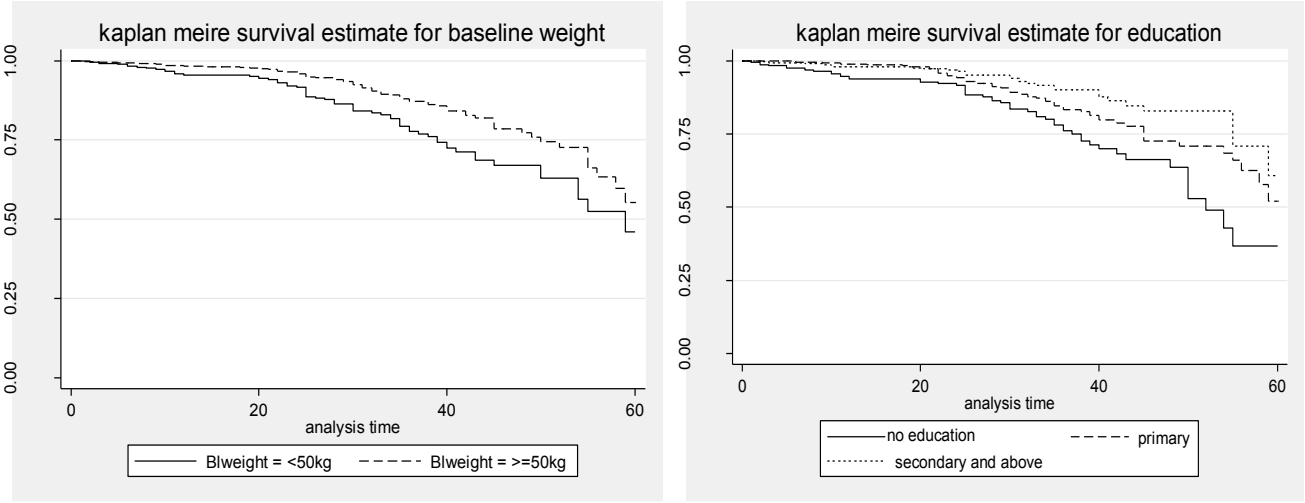


Figure 4.2: Plots of Kaplan-Meier survivor function estimates for the variable baseline weight and education level

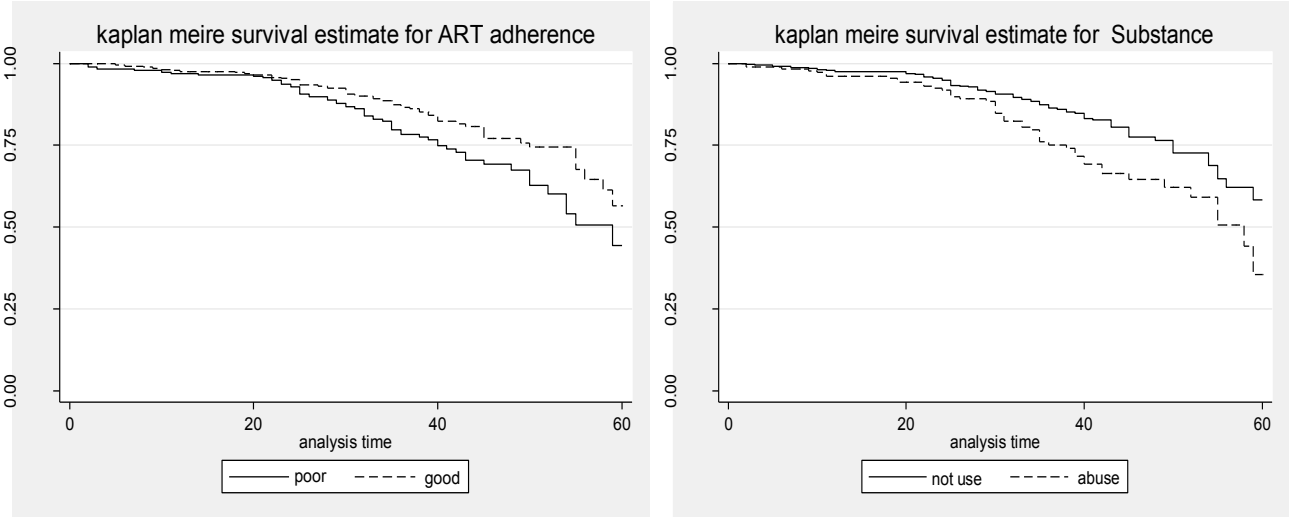


Figure 4.3: plots of Kaplan-Meier survivor function estimates for the variable substance use and ART adherence

To check for significance differences among categories of factors that are shown using the Kaplan-Meier estimates of the survivor functions, we employ a log-rank statistical test. Based on the Log-Rank test, there was no significant difference in survival experience between the various categories of gender, age, residence, marital status and drug regimen. However, the log-rank test shows that the survival experience of patients in different categories such as baseline weight, functional status, level of education, ART adherence, baseline CD4, substance use and TB co-infection are differs significantly which displayed in Table 4.2 below. A close examination of figures 4.(2 ,3) , figures 4.(9 to12) and Table 4.2 reveal that patients who had: baseline weight 50kgms or above, working functional status, secondary and above education level, good ART adherence , ≥200 line CD4 count, not abuse substance and no TB co-infected had better survival time compared with other categories.

Table 4.2 Comparison of Survival Experience of HIV Patients Using socio-demographic and clinical variables.

Covariates/Factors	DF	Chi-square		P-value	
		Log Rank (Mantel-Cox)	Breslow (Genera-lized Wilcoxon)	Log Rank (Mantel-Cox)	Breslow (Genera-lized Wilcoxon)
Gender	1	0.47	0.017	0.491	0.896
Age	1	0.0	1.741	0.992	0.187
Residence	1	0.779	3.326	0.377	0.068
Base line weight	1	11.594	13.595	0.010	0.000
Marital status	4	2.318	4.038	0.128	0.440
Functional status	2	6.007	4.175	0.034	0.041
Drug regimen	2	4.048	3.538	0.440	0.600
Education level	2	13.289	9.791	0.000	0.002
ART Adherence	1	6.789	5.114	0.009	.0024
WHO clinical stage	3	3.975	1.320	0.046	0.251
Base line CD4	1	4.016	4.199	0.045	0.040
Substance use	1	9.274	7.106	0.020	0.008
TB co-infection	1	17.667	19.804	0.000	0.000

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016 indicates statistical significance at 0.05 level of significance.*

4.3. Results of the Cox proportional hazards Regression Model

In order to study the relationship between survival time and covariates, a regression modeling approach to survival analysis using the Cox proportional hazards model can be employed for estimating the regression coefficients, making interpretation based on the hazard function, conducting statistical tests, constructing confidence intervals, checking the adequacy of model and its development precede interpretation of results obtained from the fitted model.

4.3.1. Univariate Analysis of Cox ph Regression Model

The univariate Cox proportional hazard model analysis is an appropriate method that is used to show potentially important variables before directly included in the multivariate model. The relationship between each covariates and survival probability of HIV patients are presented in Table 4.3. As can be seen from this Table, survival probability of the patients was significantly associated with base line weight, education level, ART adherence, baseline CD4, substance use and TB co-infection. But the covariates like gender, age, residence, drug regimen, marital status, WHO clinical stage and functional status were not statistically significant at 5% significance level. Furthermore, using a modest level of significance 25% to include in the multivariate model for further investigation were base line weight, marital status, WHO clinical stage, functional status ,drug regimen, education level , ART adherence, baseline CD4, substance use and TB co-infection.

Table 4.3: Single Covariate Analysis of Cox proportional hazards regression model of HIV Patients

Covariates	B	SE	Wald	DF	P-Value	HR	95.0% CI for HR
							Lower Upper
Gender	-0.115	0.168	0.469	1	0.493	0.891	[0.641 , 1.239]
Age	0.002	0.169	0.000	1	0.992	1.002	[0.719 , 1.396]
Residece	-0.147	0.168	0.769	1	0.380	0.863	[0.621 , 1.199]
Baseline weight	-0.559	0.167	11.175	1	0.001	0.5714	[0.411 , 0.793]
Marital status			10.351	4	0.055		
Married	0.609	0.442	1.893	1	0.169	1.838	[0.772 , 4.375]
Divorced	0.006	0.424	0.100	1	0.990	1.006	[0.438 , 2.309]
Windowed	0.506	0.476	1.132	1	0.087	1.659	[0.653 , 4.213]
Functional status			7.037	2	0.166		
Ambulatory	-.438	0.268	2.675	1	0.102	.645	[0.382 , 1.091]
Bedridden	0.012	0.283	0.002	1	0.967	1.012	[0.581 , 1.763]
Drug regimen			3.968	2	0.138		
AZT-3TC-NVP	0.395	0.199	3.922	1	0.048	1.484	[1.004 , 2.194]
TDF-3TC-EFV	0.227	0.207	1.204	1	0.272	1.255	[0.837 , 1.882]
Education level			18.518	2	0.000		
Primary	-0.503	0.188	7.145	1	0.008	0.604	[0.418 , 0.874]
secondary and above	-0.971	0.231	17.639	1	0.000	0.379	[0.241 , 0.596]
ART Adherence	-0.734	0.169	6.610	1	0.010	0.427	[0.244 , 0.721]
WHO clinical stage			16.986	3	0.053		
stage II	0.087	0.351	0.061	1	0.805	1.091	[0.548 , 2.171]
stage III	0.5007	0.349	0.025	1	0.035	1.650	[1.578 , 2.290]
stage IV	0.750	0.325	5.318	1	0.021	2.116	[1.119 , 4.002]
Base line CD4	-0.430	0.170	6.419	1	0.011	0.65	[0.465 , 0.907]
Substance use (alcohol)	0.558	0.177	9.982	1	0.002	1.748	[1.236 , 2.471]
TB co-infection	0.714	0.174	16.752	1	0.000	2.042	[1.708 , 2.383]

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016, SE= Standard Error, B=Parameter Estimate, HR= Hazard Ratio, CI = Confidence Interval, DF= Degrees of Freedom.

4.3.2. Multivariate Analysis of Cox ph Regression Model

One problem with any univariate analysis is that it ignores the possibility that a collection of covariates, each of which is weakly associated with the outcome may have a significant effect when used together with other covariates in the model. If this is thought to be a possibility, then we should choose a significance level large enough to allow the suspected variables to become candidates for inclusion in the multivariate model. That is why we used 20-25% significant for selection of variables that are candidates for the multivariate analysis from single covariate findings (Table 4.12 in Appendix). Consequently, the most important subset of these predictors to be included in the multivariable model will be selected by stepwise procedure, which based on their contribution to the maximized log partial likelihood of the model (-2*LL)*. The summary result indicate that the highest reduction in - 2*LL*(bˆ) is observed for drug regimen that reduced the value for the null/empty model, from 1707.449 to 1655.784, the difference is 51.66 and the next highest change is obtained for functional status of (48.789) followed by marital status (42.021). Therefore, all the covariates will be included in the multivariate study. The next step is to check the significance of the covariates in the multivariable model. The covariates which are not significant at 5% significance level, then those covariates eliminated from the model. Lastly, the final Cox ph regression model is fitted in Table 4.4 using the remaining significant covariates.

Table 4.4: the Parameter Estimates, Standard Errors and the Hazard Ratios of the Final Cox Proportional Hazard Regression Model

Covariates	DF	Parameter Estimate	SE	Wald	P-Value	HR	95.0% CI for the HR	
							Lower	Upper
Baseline weight								
<50kgms (Ref.)								
≥50kgms	1	-0.438	0.173	6.332	0.012	0.6455	[0.459	, 0.906]
Education level	2			18.518	0.000			
no education (Ref.)								
Primary	1	-0.504	0.188	7.145	0.008	0.604	[0.417	, 0.875]
secondary and above	1	-0.972	0.231	17.639	0.001	0.379	[0.242	, 0.596]
ART Adherence								
Poor (Ref.)	1	-0.7881	0.172	5.068	0.024	0.454	[0.284	, 0.749]
Good								
WHO clinical stage	3			13.923	0.003			
stage I (Ref.)								
stage II	1	0.325	0.354	0.845	0.358	1.384	[0.692	, 2.769]
stage III	1	0.507	0.351	0.159	0.022	1.650	[1.578	, 2.290]
stage IV	1	0.823	0.327	6.340	0.012	2.278	[1.700	, 4.323]
Base line CD4								
< 200 cells/μl (Ref.)								
≥ 200 cells/μl	1	-0.4033	0.1734	5.379	0.020	0.685	[0.495	, 0.907]
Substance use								
No (Ref.)	1	0.6034	0.184	10.739	0.001	1.828	[1.275	, 2.621]
Yes								
TB co-infection								
Not infected (Ref.)	1	0.3775	0.188	4.021	0.045	1.458	[1.008	, 2.109]
Co-infected								

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016 indicates statistical significance at 0.05 level of significance. SE= Standard Error, HR= Hazard Ratio, CI = Confidence Interval, Ref. = Reference, DF= Degrees of Freedom, AIC value=* 1698.571

4.3.3. Assessment of Model Adequacy

The adequacy of the model needs to be assessed after the model has been fitted to the observed survival data. At this point we have a preliminary fitted model and the next step is assessing the adequacy of the fitted model should be done in order to evaluate how well the fitted regression describes the data set. In Cox ph survival regression analysis assessment of model adequacy the study must test the assumption of proportional hazards, check influence observation and overall goodness of fit.

4.3.3.1 Test the assumption of proportional hazards

A proportional hazard is one of the very important assumptions in the Cox model. To check the PH assumption for all the categorical variables included in the fitted model, we used the log (-log (survival probability)) plot versus log survival time, which is called as a log-cumulative hazard plots are presented in Figure 4.13**:** (a – e) of the appendix. The graphs for each of the categorical variable display lines that appeared to be parallel implying that the proportional hazards assumption among groups of the categorical variables such as baseline weight, ART Adherence ,baseline CD4 count, substance use and TB co-infection has not been violated. The assumption also checked for each covariate in the final Cox regression model by adding an interaction term with log of time. The results after adding the interaction term with log time are presented in table 4.5. The coefficient for interaction effect of each covariate with log time is found not significant which indicates that the proportional hazards assumption is not violated. Moreover, the plot of the scaled Schoenfeld in Figure 4.14 and 4.15 of the Appendix shows that the residuals are random without any systematic pattern and the smoothed plot looks straight line without any departure from the horizontal line. This above interaction effect and figures indicate that the PH assumption is satisfied for all the covariates in the model.

Table 4.5: Results of the multivariable proportional hazards regression model containing the variables in Table 4.4 and their interaction with log time (in months)

Covariates Interacted with Log time	Parameter Estimate	SE	Wald	P-Value	Hazard Ratio	95.0% CI for the Hazard Ratio
Baseline weight	-0.336	0.806	0.418	0.676	0.714	(0.147, 3.469)
Education level	-0.351	0.517	0.680	0.497	0.703	(0.394, 3.221)
ART Adherence	-0.492	0.964	0.510	0.610	0.611	(0.255, 1.939)
WHO clinical stage	-0.314	0.497	0.063	0.950	0.969	(0.092, 4.051)
Base line CD4	-1.247	0.936	1.332	0.183	0.287	(0.165, 2.568)
Substance use	0.788	0.912	0.864	0.387	1.200	(0.586, 1.800)
TB co-infection	1.167	0.864	1.349	0.177	3.212	(1.315, 3.681)
Baseline weight *ln(Time)	-0.646	0.243	0.027	0.979	0.199	(0.058, 1.750)
Education *ln(Time)	-0.012	0.155	0.073	0.942	0.988	(0.617, 1.600)
ART Adherence*ln(Time)	-0.001	0.287	0.035	0.972	0.990	(0.728, 1.342)
WHO clinical stage*ln(Time)	0.007	0.147	0.310	0.757	1.047	(0.563, 1.739)
Base line CD4*ln(Time)	0.268	0.279	0.957	0.339	1.307	(0.784, 1.398)
Substance*ln(Time)	0.0157	0.271	0.058	0.954	1.016	(0.755, 2.261)
TB co-infection*ln(Time)	-0.193	0.260	0.740	0.460	0.824	(0.596, 1.729)

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016.
SE= Standard Error

4.3.3.2. Identification of Influential Observation

The next step this study follows in evaluation of regression diagnostic is to see if there are any observations that have undue influence on the estimates of the Cox regression parameters, or have an unexpected influence on the fit of the model. The DFBETA statistic for measuring the influence of the i^{th} *observation* is defined as the one-step approximation to the difference in the MLE of the regression parameter vector with i^{th} and the MLE of the regression parameter vector without the i^{th} observation. As a result, the first five largest changes in parameter estimates are shown in Tables 4.13 and 4.14 of the Appendix. From the tables deleting observation decreases the relative hazard of death,

but again the change is not big. Therefore, this study conclude that there is not much influential outlier observation in the HIV infected patient data from HQEMMH.

4.3.3.3. Overall Goodness of Fit

The final step in the model assessment is to measure the overall goodness of fit. For this objective the study use the Cox-Snell residuals, R^2 and Likelihood Ratio, Score and Wald tests. Plot of the Cox-Snell residuals was applied to test the overall fit of the model. The plot of the Nelson-Aalen estimate of the cumulative hazard function against the Cox-Snell residuals is presented in Figure 4.4 below. It can be seen that the plot of the residuals in Figure is almost close to the 45^0 straight line through the origin. Thus, the plot is evidence that the model fitted to the data is satisfactory. However, there is little evidence of a systematic deviation from the straight line at the left, this can be expected even if we have a well-fitting Cox model because of the reduced effective sample size caused by prior failures and censoring (Khanal 2009).

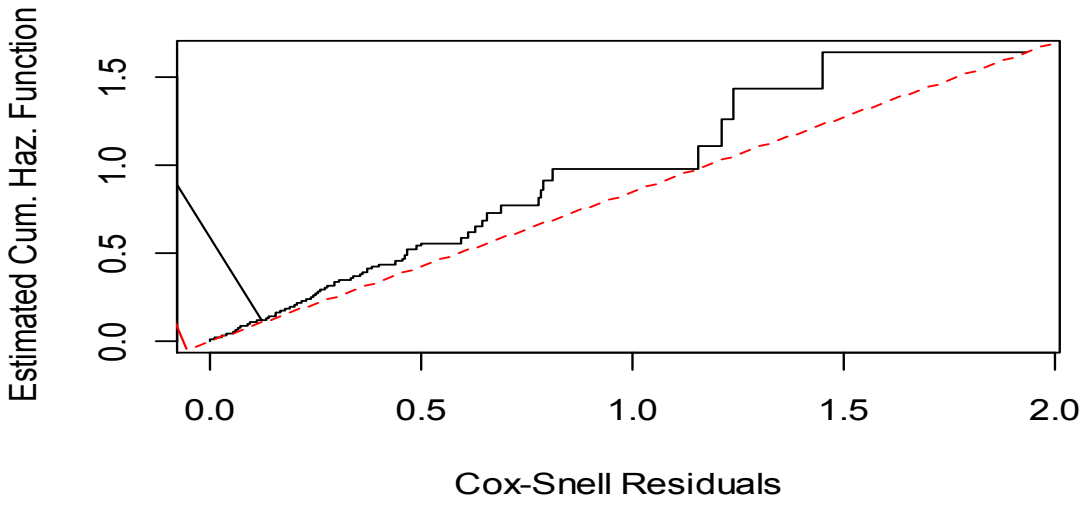


Figure 4.4: Cumulative hazard plot of the Cox-Snell residual for final Cox PH model

An adequate model is a model with low R^2 due to high percent of censored data. The value of the -2Log-Likelihood of the model with covariates in table 4.6 which is equal to 1649.303 and the -2Log-Likelihood for the null or empty model equals 1707.449. The measure of goodness of fit R^2_p is calculated as: R^2_p=1- exp[$\frac{2}{n}$(L_O - L_p)] = 1-exp[$\frac{2}{933}$((-853.7245-(-824.65))] = 0.0604. which is small, indicating that the model fit the data well. Furthermore, the results of the Likelihood ratio, Score and

Wald tests for model goodness of fit displayed in Table 4.6 which suggest that the model is good fit (i.e. significant at 5% level of significance). Therefore, the model with estimates as given in Table 4.4 is the final Cox PH Regression model.

Table 4.6: The Likelihood Ratio, Score and Wald tests for overall measures of goodness of fit of the final Cox PH model in table 4.4

Test	Chi-Square	DF	Pr>Chisq
Likelihood Ratio	72.34	10	<.0001
Score	76.23	10	<.0001
Wald	72.94	10	<.0001

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016

4.3.4. Interpretation and Presentation of the Final Cox PH Regression Model

The coefficient of the categorical covariates is interpreted as the logarithm of the hazard ratio of death to the baseline (reference group) hazard. That is, they are interpreted by comparing the reference group with others. Similarly, the coefficient for a continuous explanatory variable indicates the estimated change in the logarithm of the hazard ratio for a unit increase in the value of the respective covariate when the remaining covariates in the model are under control. Accordingly, the interpretation of the covariates included in the final Cox proportional hazard model of HIV infected patients in the case of HQEMMH is as follows.

The estimated hazard ratio of death for patients whose baseline weight is ≥50 kgms is $\widehat{HR}$= 0.6455 [95% CI: 0.4595-0.9068, p=0.012]. This means that the hazard rate of death of patients whose baseline weight ≥50 kgms reduced by 35.45% compared to patients whose baseline weight<50 kg controlling for other variables in the model. Similarly, the covariate baseline CD4 count is statistically significant influence on the survival time of the patients. The estimated hazard rate of death of patients whose CD4 count ≥200cells/μl is 0.687 [95% CI: 0.4957-0.9071, p=0.02]. This indicates that the estimated hazard rate of death of patients whose CD4 count ≥200cells/μl reduced by 31.3% compared to patients whose CD4 count < 200 cells/μl controlling for other variables in the model.

The estimated relative risk of death for a patient whose level of education are primary and secondary or above are 0.604 (95% CI: 0.417- 0.875, p=.008) and 0.379 (95% CI: 0.242-0.596, p<0.001),

respectively. This means that the hazard rate of death of patients whose level of education are primary and secondary or above are reduced by 39.6% and 62.1% respectively compared to patients with no education controlling for other variables in the model. Moreover, the estimated hazard ratio of primary level of education compared to secondary or above level of education is exp (-.972+.504) =0.626 (95% CI: 0.543 - 0.7083). Since the confidence interval does not contain 1, it indicate that an individual whose level of education is secondary or above has a significantly reduce hazard rate at any given time than patients with primary level of education. Similarly, the estimated relative risk of death for patients whose ART adherence good as compared to those who are ART adherence poor is,$\widehat{HR}$ =0.454 [95% CI: 0.2845 - 0.7498, p=0.024]. This means estimated hazard rate of death of patients whose good ART adherence was reduced by 54.6% times than patients whose poor ART adherence controlling for other variables in the model.

The estimated hazard ratio of death for patients in WHO stages III and IV are $\widehat{HR}$ = 1.650 [95% CI: 1.578-2.290, p=0.022] and $\widehat{HR}$=2.278 [95% CI: 1.700-4.323, p=0.012] respectively in comparison to that of stage I controlling for other variables in the model. This indicates that the hazard rate of death was 1.65 times higher in stage III and 2.3 times higher in stage IV as compared with stage I. On the other hand, the estimated hazard ratio of stage IV compared to stage III is 1.38 =exp (0.823-0.5007). This suggests, patients in stage IV are 38% more likely to die than patients in stage III.

The estimated hazard ratio of death for patients who were abuse substance (tobacco, alcohol, soft drugs) was 1.828 times higher than those who didn't uses substance[95% CI: 1.275- 2.621, p=0.001]. This indicates patients who were abuse substance was 82.8% higher risk of death than patients who did not use substance controlling for other variables in the model. Similarly, the estimated relative risk of death for patients who were TB co-infected was 1.458 times higher risk of death than patients not TB co-infected [95% CI: 1.008-2.109, p=0.045] controlling for other variables in the model.

4.4. Parametric Survival Regression Model Analysis

4.4.1. Model Comparison for Time to death of HIV infected Patients

From this Time to death of HIV infected patients the parametric regression models were fitted in Table 4.15 of the Appendix. This study consider model Comparison after adjusting for the effect of covariates and also compare models by using graphical method based on the Cox-Snell residual plots and Akaikie information criterion (AIC). In case of Cox-Snell residual plot, if the model is good, the plot of Cox-Snell residuals versus cumulative hazard estimates line should passes through the origin. Here this study presents the Cox-Snell residual plots for model comparison in Figures 4.5 to 4.8. From those figures Cox-Snell residuals plot for Weibull regression model shows deviation from the straight line passing through origin, it indicates that the Weibull regression model fit the data better, otherwise that the exponential, log normal and log logistic regression models fit the data poorly.

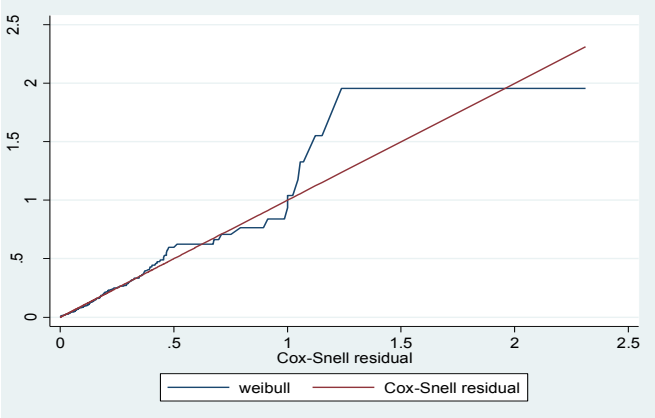


Figure 4.5 The Cox Snell plot after fitting Weibull regression model

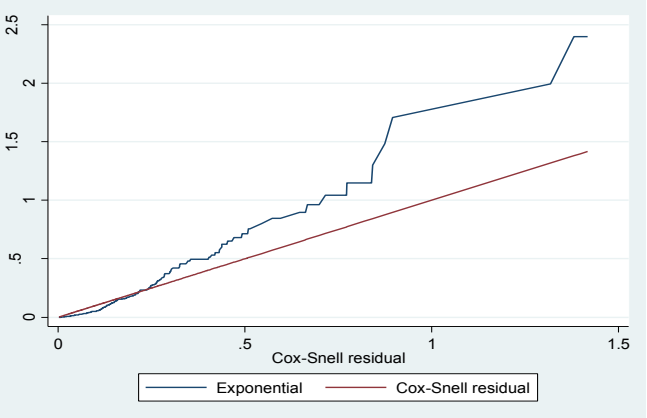


Figure 4.6 The Cox Snell plot after fitting Exponential regression model

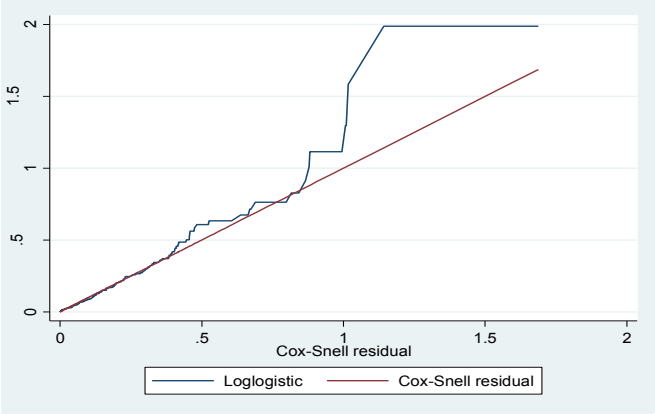


Figure 4.7 The Cox Snell plot after fitting log logistic regression model

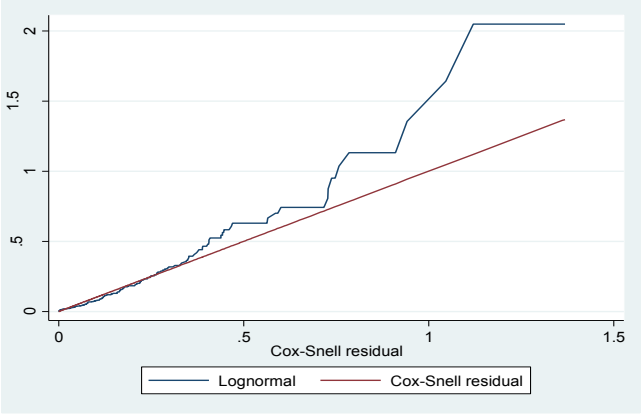


Figure 4.8 The Cox Snell plot after fitting lognormal regression model

But graphical methods may not assure the result. In order to select the appropriate parametric survival regression mode1, the most common applicable criterion called Akaikie information criterion (AIC). Nevertheless, the results of cox-snell were consistent with the results based on Akaikie‟s information criterion. Here, the models are not nested; it is not possible to compare the models using logliklihood values. When the models were compared using AIC in Table 4.7, among the parametric models, the result of table reveal that the Weibull regression model has the smallest AIC, which shows that weibull model is the appropriate parametric survival regression model for HIV infected patients from Hossana Queen Elleni Mohamad Memorial Hospital.

Table 4.7: Selection of parametric survival regression model by using Log likelihood and AIC

Model	log-likelihood	AIC
Exponential	-865.9	1759.879
Weibull	-827.1	1684.139
Log logistic	-828.6	1687.193
Lognormal	-840.8	1711.652

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016

AIC=Akaike's information criteria

And also using the same type of variables model comparison made between semi-parametric and parametric is carried out using AIC. The result given on Table 4.8 blow shows that AIC for Cox proportional hazard regression model is 1698.57 while the AIC value for Weibull regression model is 1684.139. Hence the Weibull regression model is better model than Cox proportional hazard regression model in fitting to the data on survival time of HIV infected patients considering in this study. Finally, the Weibull distribution is unique in that it is the only one that is simultaneously both proportional and accelerated so that both relative event rates and relative extension in survival time can be estimated.

Table 4.8: *Comparison of* AIC *values* for Cox proportional model and weibull survival regression model

Model	**AIC**
Weibull regression model	1684.139
Cox proportional hazard regression model	1698.571

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016, AIC=Akaike's information criteria

4.4.2. Univariate Analysis of Weibull Regression Model

According to the Weibull analysis of single covariate, the selected covariates for further analysis and interpretation are made here below. To have an idea about the individual effects of the different explanatory variables on survival of HIV infected patients fitted Weibull regression model separately for each explanatory covariate results displayed in Table 4.9 bellow. Which illustrate that statistically significant factors for the survival probability of HIV infected patients, thus are baseline weight, education level, ART adherence, WHO clinical stage, baseline CD4, substance and TB co-infection. Whereas the factors that were not statistically significant are gender, age group, residence of patients, marital status, functional status and drug regimen at 5% level of significance. The factors those were statistically significant included in the final Weibull regression model for the prediction of survival probability of HIV infected patients.

Table 4.9: The Result of Deviance, -2*LL and Pr (>Chi) of univariate weibull regression model Analysis.

Covariate	Df	Deviance	Resid. Df	-2*LL	Pr(>Chi)
NULL	NA	NA	931	1720.972	NA
Gender	1	4.42518e-01	930	1720.529	5.059107e-01
Age	1	4.068705e-04	930	1720.529	9.839069e-01
Residence	1	1.117299e+00	930	1719.952	2.905007e-01
Baseline weight	1	1.215160e+01	930	1719.877	4.904580e-04
Marital status	4	8.581178e+00	927	1711.708	3.930489e-01
Functional status	2	4.747349e+00	929	1713.989	6.724579e-02
Drug regimen	2	3.674616e+00	929	1717.146	5.893753e-02
Education level	2	1.747705e+01	929	1702.713	3.150739e-05
ART adherence	1	3.576980e+00	930	1714.454	4.383535e-02
WHO clinical stage	3	1.411636e+01	928	1704.820	2.750965 e-02
Baseline CD4	1	3.634150e+00	930	1714.177	3.653883 e-02
Substance use	1	1.212300e+01	930	1711.324	3.927276e-04
TB-co-infection	1	3.153572e+00	930	1704.493	8.930405e-03

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016, NA=Not applicable, Df= Degrees of freedom, LL= Loglikelihood.

4.4.3. Multivariate Analysis of Weibull Regression Model

The result of relationship between covariates and survival probability of HIV infected patients modeled by Weibull regression model are presented in Table 4.10. It indicate the parameter estimates of coefficients for the covariates in the final Weibull regression model along with the associated significance level, hazard ratio with corresponding standard error and 95% confidence interval for the hazard ratio. Survival time of HIV infected patients were significantly associated with baseline weight, WHO clinical stage, education level, ART adherence, baseline CD4, substance use and TB co-infection as can be seen from the Table 4.10.

Table 4.10: Summary result of Parameter Estimates, Standard Errors and the 95% CI of the final multivariate Weibull regression model Analysis

Covariates	Parameter Estimate	SE	Z	P-Value	Hazard Ratio	95.0% CI for the H R
Baseline weight						
≥50kgms (Ref.)						
< 50kgms	0.215	0.084	2.511	1.10e-02*	1.24	[1.048, 1.461]
Education level						
no education (Ref.)						
Primary	-0.230	0.084	-2.357	1.63e-02*	0.792	[0.619, 0.981]
secondary and above	-0.463	0.117	-3.948	8.62e-05*	0.629	[0.402, 0.845]
ART Adherence						
Poor (Ref.)						
Good	-0.589	0.084	-2.240	3.06e-02*	0.554	[0.389, 0.716]
WHO clinical stage						
Stage I (Ref.)						
stage II	0.086	0.122	0.712	5.04e-01	1.090	[0.844, 1.366]
stage III	0.237	0.107	2.207	2.90e-02*	1.267	[1.064, 1.476]
stage IV	0.648	0.170	0.866	3.86e-03*	1.711	[1.517, 2.044]
Base line CD4						
≥ 200 cells/μl (Ref.)						
<200 cells/μl	0.289	0.084	2.234	2.19e-02*	1.335	[1.170, 1.499]
Substance use						
No (Ref.)	0.492	0.091	3.196	1.29e-03*	1.636	[1.427, 1.734]
Yes						
TB co-infection						
Not infected (Ref.)						
Co-infected	0.388	0.092	2.049	3.46e-02*	1.473	[1.298, 1.654]

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016 indicates statistical significance at 0.05 level of significance. SE= Standard Error, HR= Hazard Ratio, CI = Confidence Interval, Ref. = Reference, AIC value=* 1684.139

In this study the baseline hazard for final weibull regression model obtained from equation (3.42) and with the parameters found in Table 4.10, the survival time of HIV patients with Weibull distribution can be expressed as t ~ Weibull(ρ, λ), with parameters $\lambda = \exp(\frac{-\mu}{\sigma})$= 2.98e^{-4} and $\rho = \frac{1}{\sigma}$ = 2.056 this shows hazard increases monotonically with time , time ~ Weibull (2.056, 2.98e^{-4}). By substituting the parameters in the final Weibull model with substitution of $\lambda = 2.98e^{-4}$ and $\rho = 2.056$, the Weibull hazard regression model that predicts the hazard rate of patients with identical data settings is:

$$h_i(t, X, B) = h_0(t)\exp(\beta' x_i) = \lambda\rho t^{\rho-1} \exp(\beta' X) = 2.98e^{-4} * 2.056 * t^{1.056} \exp(\beta' X) \quad (4.1)$$

Form the final Weibull regression model the baseline hazard vary with $\lambda\rho t^{\rho-1}$; so the base line hazard function of HIV infected patients for HQEMMH is given with formula of (4.2) in every increase in time.

$$h_o(t) = \lambda\rho t^{\rho-1} = 2.98e^{-4} * 2.056 * t^{1.056} \quad (4.2)$$

4.4.4. Assessment of adequacy and Interpretation of the Weibull Regression Model

The likelihood ratio test presented in Table 4.11, it illustrate that the model was significantly fit the data of HIV patients. And in using the log likelihood, the model has a significant improvement after the covariates is incorporated in the model.

Table 4. 11: The Likelihood Ratio Test of the Final Weibull Regression Model for Time to death of HIV Patients

Log likelihood (intercept only)	Log likelihood (Model)	LR chi-square	DF	P-values	Intercept	Scale
-860.5	-804.2	72.48	10	0.000	4.0712	0.482

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016 indicates statistical significance at 0.05 level of significance. DF= Degrees of Freedom.*

The importance of this interpretation is that for those data where it was considered reasonable to apply Cox regression to estimate the underlying hazard ratio, it should also be reasonable to apply a Weibull analysis to estimate the hazard ratio and using the estimated scale parameter. In this study Weibull regression model was considered as better fit to the data, and also both hazard ratio and survival probabilities can be still interpreted as the hazard rate of death or survival probabilities increase/decrease in survival time on the reference group relative to others.

From the Weibull regression model, Baseline CD4 count is statistically significant influence on the survival time of the patients. After adjusting the other covariates, the hazard rate of patients who were CD4 count < 200 cells/μl was 1.335 times higher than those patients who were CD4 count ≥ 200 cells/μl (adjusted HR=1.335, CI=1.170-1.499), this means the survival probability of HIV patients whose CD4 count < 200 cells/μl was declined by 33.5%. Similarly, keeping other covariates constant, the hazard rate of patients who were baseline weight < 50 kg was 1.237 times higher than those patients who were baseline weight ≥50 kg (adjusted HR=1.237, CI=1.04829-1.461633), which means the survival probability of HIV patients who were baseline weight less than 50 kg was reduced by 23.7%.

For WHO clinical stage, after fixing other covariates constant, The hazard rate of a patients in WHO stages III were 1.267 times that of patient in stages I (adjusted HR=1.267, 95% CI: 1.064 -1.476). And the hazard rate of a patients in WHO stages IV were 1.711 times that of patient in stages I (adjusted HR=1.711, 95% CI: 1.517 -2.044). This indicates that the survival probability of HIV patients in stage III was reduced by 26.7% and in stage IV was reduced by 71.1% as compared with stage I. For education level after fixing other covariates , the hazard rate of patients whose education level are primary and secondary or above are 0.792 (95% CI: 0.619-0.981) and 0.629 (95% CI: 0.402 -0.845), respectively. This means that the survival probability of HIV patients whose level of education are primary and secondary or above are increased by 20.8% and 37.1% respectively compared to patients with no education. Similarly, the After adjusting other covariates, the hazard rate of patients who were good ART adherence was reduced by 0.554 times than those patients who were poor ART adherence (adjusted HR=0.554, CI=0.389-0.716), which means that the survival probability of HIV patients who good adherence was increased by 44.6%.

The abuse substance (alcohol) had also a significant effect on the survival probability of HIV patients. After adjusting other covariates, the hazard rate of patient who were abuse substance was 1.636 times higher than those patient who didn't use substance (adjusted HR=1.636, CI=1.427-1.734), this pointed out that the survival probability of patients who use substance was reduced by 63.6%. Similarly, After adjusting other covariates, the hazard rate of patients who were co-infected with TB was 1.473 times higher than patients who had not co-infected (adjusted HR=1.473, 95% CI: 1.298-1.654). This means that the survival probability of HIV patients who TB co-infected was declined by 47.3%.

4.5 .DISCUSSION

The study used stepwise selection technique, enter and remove, and backward technique the most non-significant covariates are removed and the rest in the model are refitted. At the last step the procedure ended with the most likely selected covariates: baseline weight, ART adherence, baseline CD4 count, WHO clinical stage, education level, substance use and TB co-infection of the patients on ART. In both Cox proportional and Weibull regression models, the baseline body weight <50kgs was significantly associated with reducing the survival probability of HIV infected patients. This result confirms the finding observed in West Africa, in which body weight at initiation of ART treatment <50 kgs was significant risk factors of death during ART treatment period (Moha *et al*., 2007 and Ferradini *et al*., 2006). Similarly, our study revealed that the weight of patients is significantly associated with mortality and those patients having lower weight were at higher risk of mortality.

This finding showed that the majority of HIV patients had started antiretroviral treatment with more advanced immunodeficiency status. Since the majority of HIV patients had AIDS as defined by their CD4 cell counts < 200 cells/μl, indicated advanced immune suppression at initiation of ART. These findings indicated that the HIV patients with lower the CD4 cells count which greater chances of getting very serious diseases, which lead to high risk of mortality. The result of this study and findings of Lawn *et al*. (2008) are similar in low baseline CD4 cell count strong risk factor association to early mortality.

A study by Zubairu and Musa (2009) suggested that WHO clinical Stages III and IV have significant impact on reducing the survival probability and also it indicated that WHO clinical stage is a risk factor of HIV infected patients. According to previous studies by Mohammed *et al* (2011) showed that WHO clinical stage III and IV are significantly associated with high risk of mortality on HIV infected patients. Similarly, this study result revealed that the patients in advanced WHO clinical stage III and IV were significantly reduced the survival probability compared to those patients in Stages I. The study also revealed that the HIV infected patients whose level of education was secondary or above were more likely to survive to compare those patients with no education. This result is consistent with studies done previously by DeSilva *et al*.,(2009). Similarly, this study revealed that the patients whose education level was secondary or above were more likely to survive compared to those patients with no education which is also consistent result with previous study by Reda *et al*., (2013). In the same manner, this study revealed that the patients with primary education were more likely to survive than illiterate patients (patients with no education).

The study showed that substance abuse had a significant impact on the survival probability of HIV infected patients. This result confirmed the finding obtained from previous studies by Asefa (2005), Liu *et al.,* (2006) and Moattia *et al.,* (2000) alcohol and other substances abuse were associated with mortality, non-adherence to medication and lower quality of life of HIV infected patients. This study showed similar result that HIV infected patients who were substance abused more likely died compared to those patients who do not use. This might be due to ART non-adherence in addition to the complications that alcohol brings in to one's health. A study conducted in Uganda showed non adherent patients had a mortality of 42.5 deaths per 1000 person-years and after adjusting for age, sex and educational level were two times as likely to die as adherent participants (Kaufmann *et al*., 2011). In addition study conducted in Ethiopia revealed, the risk of death in non- adhered patients is 4 times higher compared to adhered patients (Bedru, 2009). Above studies were in agreement with this study results which indicated that there was strong association between mortality and ART adherence.

The result of this study showed that patients co-infected with TB had highly increased the rate of death than uninfected patients. Similar study conducted in China and Ethiopia showed that TB co-infection was associated with high risk of mortality on HIV infected patients (Xueyan *et al*., 2008, and Gezahegn, 2011). This study confirmed that HIV infected patients who were TB positives at ART initiation were more likely to die compared to those HIV infected patients who were TB negatives. This might be due to the fact that TB is the leading cause of death worldwide in HIV-infection and mycobacterium tuberculosis is a virulent organism that can produce disease in HIV-infected persons at any stage of disease even when the immuno suppression is minimal.

In this study comparison among four models were made, exponential, weibull, lognormal and log logistic regression models, the models were compared based on their Cox-Snell plots and Akaikies information criterion (Akaike, 1974). In this study, the Weibull regression model fit the data better than the other models. And also model comparison made between semi-parametric and parametric was carried out using AIC(Yiannoutsos ,2009 and Dehkordi et al., 2008). Hence the Weibull regression model was better model than Cox proportional hazard model in fitting to the data on survival time of HIV infected patients considering in this study. This result was similar to previous study by Dehkordi et al., (2008) and Tesfaye (2013). Finally, the Weibull regression model was appropriate for HQEMMH HIV infected patient data set.

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The results of Kaplan-Meier and log-rank test showed that patients who had: baseline weight 50kgms or above, working functional status, secondary and above education level, good ART adherence , ≥200 line CD4 count, not abuse substance and no TB co-infected had better survival time compared with reference groups. Univariate Cox Proportional Hazards regression models were developed to assess the relation between each covariate survival status and their selected variables. The result of multivariate Cox proportional hazards regression model showed that baseline weight, ART adherence, baseline CD4 count, WHO clinical stage, education level, substance and TB co-infection of patient were the major factors that affect the survival probability of HIV infected patients.

For modeling time to death of HIV patients Exponential, Weibull, lognormal and log logistic parametric regression models were applied. Among these using Cox-Snell residuals plot and AIC for model comparison, the Weibull survival regression model was better fitted model for time to death of HIV infected patients in case of Hossana Queen Elleni Mohamad Memorial Hospital than the other remaining parametric models. The Weibull regression model results revealed that baseline weight<50 kg, low CD4 count at baseline, no education, WHO stages III and IV, poor ART adherence, co-infection with TB and substance abuse are the categories that reduce the survival probability of HIV infected patients. Finally, The Weibull survival regression model provides better predictions to the survival probability of HIV patients.

5.2. Recommendation

Based on this study finding, the following recommendations can be forwarded for government program planners, decision makers, ART program implementers at different level and other stakeholder who work in the areas of giving care, support and treatment for HIV/AIDS patients.

- Health workers should be cautious when a patient has lower baseline CD4 and lower baseline weight.

- Health workers need to support those patients with no or little education by continuous awareness creation of taking care of themselves and knowing what factors facilitate death. Hence, education level of the patients has an important role in increasing their quality of life.
- Prompt initiation of TB treatment in order to reduce patient mortality and Patients who drink alcohol need to be given advice to reduce excessive drinking. And also Careful follow up for poorly adhered patients and giving them drug counseling is crucial to improve survival
- For future researchers on this area should apply Weibull survival regression model because Weibull distribution is unique that means only one that simultaneously both proportional and accelerated so that both relative event rates and relative extension in survival time can be estimated and it predict the survival probability of HIV patients well

Limitation of the Study

- The study presumed that all deaths are caused by HIV infection.
- The study is based on baseline values of the variables of interest i.e. CD4 cell count stability or improvement, weight loss or gain, which are associated to mortality of AIDS patients, are not included in the study because they were not consistently recorded.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification", IEEE Transactions on Automatic Control, **19** (6): 716–723, doi:10.1109/TAC.1974.1100705, MR 0423716.

Andinet and Sebastian. (2010).Determinants of survival in adult HIV patients on antiretroviral therapy in Oromiyaa, Ethiopia.

Assefa,T. (2005).The association between substance abuse and HIV infection among people visiting HIV counseling and testing centers in Addis Ababa, Ethiopia.

Bijaria, S. (2008).Dynamics of Naïve and Memory CD4+ T Lymphocytes in HIV-1 Disease Progression. Journal of Acquired Immune Deficiency Syndrome, **30**:41-58.

Bedru, M.(2009) Assessment of predictors of survival in patients living with HIV/AIDS after the advent of Highly active antiretroviral therapy in Addis Ababa Ethiopia .

Bhaskaran, K., Hamouda, O. and Sannes M. (2008). Changes in the Risk of Death after HIV Serocon version Compared with Mortality in the General Population. ; Malawi: 300(1):51-59.

Breslow, N. (1974). Covariance analysis of censored survival data. Biometric 30 (1974), 89.100

Cain, K., and Lange, N. (2003). Approximate case impudence for the proportional hazards Regression. Approximate case influence for the proportional hazards Regression model with censored data. Biometrika 40 , 493-499.

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

Cox, C.(1998). Delta method in Encyclopedia of Biostatistics P. Armitage and T. Colton, Eds., vol.2, pp. 1125.1127.74.

Cox, D. and Snell, E. (1968).A general definition of residuals with discussion. Journal of the Royal Statistical Society.Series B.,429.436.

Degu J. and Lindtjørn B. (2008).Disease Progression among Untreated HIV-Infected Patients in South Ethiopia: Implications for Patient Care. Medscape General Medicine;7(3):66.

Dehkordi, B., Safaee, M, Pourhoeingholi, R., Fatemi, Z., Tabeie,M.and Zali ,R. (2008). Statistical Comparison of Survival Models for Analysis of Cancer Data. *Asian pacific J. cancer Prev,*9: 417-420.

DeSilva, M., Merry, S., Fischer, P., Rohrer, J., Isichei, C., & Cha, S.(2009). Youth, unemployment, and male gender predict mortality in AIDS patients started on HAART in Nigeria. AIDS care, 21(1), 70-77.

Efron ,B. (1977). The efficiency of Cox‟s likelihood function for censored data.; 72: 557-65.*Ethiop. J. Health Dev*; **22**(1).

FerradiniL., Jeannin A., Pinoges L., Izopet J., Odhiambo D., Mankhambo L., Karung G. ,Szumilin .E., (2006). Scaling up of highly active antiretroviral therapy in a rural district of Malawi :an effectiveness assessment. Lancet; 367: 1335–42

Fleming,T., and Harrington.D, (1991).Counting Processes and Survival Analy-sis. W. NewYork.

Gehan, E. (1965). Generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika52 , 203.223.

Gezahegn, A. (2011). Survival Status among patient living with HIV AIDS who are on ART treatment in Durame and Hossana Hospitals. DOI: 10.1371/journal.pone.0087392 .doi:10.1371/journal.pone.0087392.t002.

Holford, T. (1976). Life tables with concomitant information. Biometrika 32, 587.597.

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

HZHD, (2014*).*Hadiya zone Health Department: Health Sector Development program IV: annual Health performance report.

Isakidis.,J.(2009). Determinants of survival in AIDS patients on antiretroviral therapy in a rural centre in the Far-North Province, Cameroon. *Tropical Medicine*

Jerene,G.(2006). Mortality and causes of death in adults receiving highly active antiretroviral therapy in Senegal: a 7-year cohort study *AIDS.*

Johannessen .K.(2008). Predictors of mortality in HIV-infected patients starting antiretroviral Therapy in a rural hospital in Tanzania.

Johansen, S.(1978). The product limit estimate as a maximum likelihood estimate. Scan-dinavian Journal of Statistics 5, 195.199.

Kaufmann G, Furrer H, Ledergerber B (2011) Characteristics, determinants, and clinical relevance of CD4 T cell recovery to <500 cells/microL in HIV type 1-infected individuals receiving Potent antiretroviral therapy.64 , 256. 387

Kaplan, E., and Meier, P.(1958).Nonparametric estimation from incomplete observations. Journal of American Statistical Association 53 , 457.481.

Katubulushi, M. and Chanda, W. (2009). Slow decline in CD4 cell counts in a cohort of HIV infected adults living in Lusaka, Zambia. Department of Medicine.

Ketema, K. (2011). Determining factors that affect the survival rate of HIV-infected patients on ART: The case of Armed Forces General Teaching Hospital, Addis Ababa, Ethiopia.

Klein, J., Moeschberger, M.(1997). Survival analysis: Techniques for censored and truncated data; New

York, Springer-Verlag.

Kleinbaum ,D.(1996). Survival analysis: A self- learning text, New York, Springer-Verlag.

Lavori,l,D., Moorgar, G.,. (2005). Declining morbidity and mortality among patients with advanced human immune deficiency virus infection. The New England Journal of Medicine 338, 13, 853.860.

Lawless ,J.(1998). Parametric models in survival analysis. In encyclopedia of Biostatistics. Wiley: New York, 3254-64.

Lawn, S., Little, F., Bekker, L., Kaplan, R., Campbel, E., Orrell, C. and Wood, R. (2008).Changing mortality risk associated with CD4 cell response to antiretroviral therapy in South Africa. AIDS; **23**:335–342.

Lee, E., and Wang, J.(2003). Statistical methods for survival data analysis, New Jersey, Wiley.

Leigh ,F. and Rob, E. (2009). Modeling the demographic impact of HIV/AIDS in South Africa and the likely impact of interventions. *Demographic research,14(22):541-574.*

Liu,C., Johnson, L., Ostrow, D., Silvestre, A., Visscher, B., Jacobson, L. (2006).Predictors for Lower Quality of Life in the HAART Era Among HIV-Infected Men. J Acq; **42**:470

Marlink , R.(1994). Reduced rate of disease development after HIV-2 infection as compared to HIV-1 Science 265, 1587–1590.

Moha R., Danela C., Messoua E., Ouassaa T., Gabillarda D., Anziana A., Aboa Y., (2007). Incidence and determinants of mortality and morbidity following early antiretroviral therapy initiation in HIV-infected adults in West Africa. AIDS, 21:2483–2491.

Mohammed T, *(2011)*. Risk factors of active tuberculosis in people living with HIV/AIDS in South west Ethiopia: a case control study. Ethiopian J Health Sci. ; 21(2): 131-9

MoH,(2012). An intentionally recognized, effective and important strategy for both preventive and care studies

Moattia, J., Carrieria, M., Spirea, B., Gastauta, J., Cassutod, J., Moreaue, J.(2000). Adherence to HAART in French HIV-infected injecting drug users: the contribution of buprenorphine drug maintenance treatment. AIDS; **14**:151-155.

Moraes, c. Ricardo, A and Alencar, X (2008). Survival of AIDS Patients and Characteristics of those Who Died Over Eight Years of Highly Active Antiretroviral Therapy, at a Referral Center in Northeast Brazil. *The Brazilian Journal of Infectious Diseases; 12 (4):269-277.*

NAIDSRC, (2010). 'Towards universal access: Scaling up priority HIV/AIDS interventions in the

health sector' and An intentionally recognized, effective and important strategy for both preventive and care studies

NAIDSR,(2014). Antiretroviral therapy for HIV infection in adults and adolescents in resource limited settings: towards universal access. Recommendations for a public health

Nelson,W. (1972). Theory and application of hazard plotting for censored failure data. Technometrics, *14: 945 – 965.*

Pourhoseingholi, M., Pourhoseingholi, A., Vahedi ,M., et al.(2011). Alternative for the Cox regression model: using parametric models to analyze the Survival of Cancer.

Reda, A., Biadgilign, S., Deribew, A., Gebre, B., Deribe, K. (2013). Predictors of Change in CD4 Lymphocyte Count and Weight among HIV Infected Patients on Anti-Retroviral Treatment in Ethiopia: A Retrospective Longitudinal Study. *PLoS ONE;* **8**(4):e58595.

Recskyet,H. (2010). "Journal of Antimicrobial chemotherapy "HIV infection in older patients In the HAART era". Volume 57, Oxford university press.

Seyoum*et ,L. (*2009).HIV infection in older patients in the HAART era. Journalof Antimicrobial chemotherapy Volume 57, Oxford university press.

Schoenfeld, D. (1982). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika; 7: 145-153.*

Stephen, D., Lawn, A., Harries, X., Landon, M. and Robin, W. (2011). Early mortality among Adults accessing antiretroviral treatment programmes in sub-Saharan Africa. *AIDS; 22:1897–1908.*

Sethi, E. (2009). Factors affecting CD4+T-lymphocyte count response to HAART in HIV patients. HIV medicine journal.Volume 8, Blackwell publishing, uk.pp135-141.

Tesfaye.G (2013). Survival Analysis of Time to Recovery from Obstetric Fistula: A Case Study at Yirgalem Hamlin Fistula Hospital, SNNPR, Ethiopia

UNAIDS,(2004). Report on Global AIDS Epidemic''

UNAIDPS, (2005).Human Development Report." International cooperation at a crossroads aid, trade and security in unequal world.

UNAIDS,(2006). Report on Global AIDS Epidemic'', http:// data.unaids.org/ pub/ Global Report executivesummary_en.pdf

UNAIDS, (2007). Global AIDS epidemic update.

UNAIDS, (2008).Report on Global AIDS Epidemic-Google Book: ebsite:http://www.unaids.

UNAIDS,(2009).AIDS epidemic update.UNAIDS, Geneva.http://hdr.undp.org/en/media

UNAIDS, (2010). Global report: UNAIDS report on the global AIDS epidemics.

UNAIDS, (2012). Global Report: UNAIDS Report on the Global AIDS Epidemic 2012.

UNAIDS, (2013).Global Report:UNAIDS Report on the Global AIDS Epidemic 2013.

Vernonet, (2013). Explaining urban-rural variations in health: A review of interactions between individual and environment. Social Science and Medicine 42: 923-935.

Weibull,W. (1939).A Statistical Theory of the Strength of Materials.(1951), A Statistical Distribution of Wide Applicability. Journal of Applied Mathematics, 18, 29 3-297.

WHO,(2002).HIV/AIDS treatment and care clinical protocols for the WHO European Region. DK-2100 Copenhagen, Denmark.

WHO, (2003).Scaling-up antiretroviral therapy in resource-limited settings. Treatment guideline for a Public health approach ,2003 revision.

WHO,(2007). Report on the global AIDS epidemic, treatment and care.

WHO, (2012).*Core Epidemiology Slides*.

Xueyan J., Hongzhou L., Yuexin Z., Zengquan Z., Hanhui Y., Qingxia Z., Hui W.,Fuxiang W., Boping Z. and Cao Y. (2008). A Cross-Sectional Study of HIV and Tuberculosis Co infection Cases in Mainland China.Southern Medic

Yiannoutsos C. (2009). Modelling AIDS Survival after Initiation of Antiretroviral Treatment by Weibull Models with Change-points. *Journal of the International AIDS Society,* 12:

Zubairu. L and Musa .B,(2009). Prevalence and Predictors of Tuberculosis Co-infection mong HIV-Seropositive Patients Attending the Aminu Kano Teaching Hospital, Northern Nigeria. J Epidemiol ; 19(2):81-87.al Association; 101(9):914-917.

APPENDIX

Table 4.1.1: Summary of descriptive statistics for Socio-Demographic Variables

Covariates	categories	Status of patients				
		Total	Number of censored	Number Of death	**Mean**	95% CI
Gender	Female	501(53.7%)	430(85.8%)	71(14.2%)	52.18	(50.55 , 53.81)
	Male	432(46.3%)	360(83.3%)	72(16.6%)	50.92	(49.13 , 52.72)
Age group	below 40	530(57%)	460(86.8%)	70(13.2%)	52.19	(50.47, 53.91)
	40 and above	403(43%)	330(82%)	73(18%)	49.34	(47.70 ,51.99)
Residence	Rural	481(51.5%)	405(84.2%)	76(15.8%)	52.35	(50.15 , 53.55)
	Urban	452(48.5%)	385(85.2%)	67(14.8%)	51.14	(49.41, 52.87)
Base line weight	less than 50kg	363(39%)	293(80.7%)	70(19.3%)	48.99	(46.84 , 51.16)
	50kg or above	570(61%)	497(87.2%)	73(12.8%)	52.98	(51.55 , 54.42)
Marital status	Single	226(24.2%)	187(82.7%)	39(17.3%)	47.84	(44.58 , 51.10)
	Married	526(56.4%)	453(86%)	73(14%)	52.78	(51.32 , 54.24)
	Divorced	122(13%)	100(82%)	22(18%)	48.57	(44.23 , 52.93)
Functional status	Working	608(65%)	529(87%)	79(13%)	52.71	(51.28 , 54.15)
	Ambulatory	242(26%)	195(80.6%)	47(19.4%)	49.28	(46.74 , 51.83)
	Bedridden	83(9%)	66(80%)	17(20%)	49.96	(45.96 , 53.96)
Education level	no education	234(25%)	183(78%)	51(22%)	47.35	(44.68 , 50.06)
	Primary	383(41%)	324(84.6%)	59(15.4%)	52.09	(50.32 , 53.86)
	secondary and above	316(34%)	283(89.6%)	33(10.4%)	53.86	(51.93 , 55.80)

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016; mean: mean survival time, CI: Confidence Interval for mean

Table 4.1.2: Summary of descriptive statistics for clinical Variables

Covariates	categories	Status of patients			Mean	95% CI
		Total	Number of censored	Number Of death		
Drug regimen	D4T-3TC-NVP	255(27%)	207(81.2%)	48(18.8%)	49.80	(47.46 , 52.14)
	AZT-3TC-NVP	280(30%)	238(85%)	42(15%)	51.23	(48.91 , 53.57)
	TDF-3TC-EFV	398(43%)	345(86.7%)	53(13.3%)	52.68	(50.92 , 54.45)
Education level	no education	234(25%)	183(78%)	51(22%)	47.35	(44.68 , 50.06)
	Primary	383(41%)	324(84.6%)	59(15.4%)	52.09	(50.32 , 53.86)
	secondary and above	316(34%)	283(89.6%)	33(10.4%)	53.86	(51.93 , 55.80)
ART Adherence	Poor	174(18.6%)	132(75.8%)	42(24.2%)	49.38	(47.22 , 51.54)
	Good	759(81.4%)	658(86.7%)	101(13.3)	52.71	(51.27 , 54.15)
WHO clinical stage	stage I	263(28%)	233(88.6%)	30(11.4%)	53.20	(51.06 , 55.36)
	stage II	279(30%)	246(88.2%)	33(11.8%)	53.70	(51.72 , 55.68)
	stage III	295(31.6%)	236(80%)	59(20%)	47.76	(45.46 , 50.07)
	stage IV	96(10.4%)	75(78%)	21(22%)	46.40	(42.59 , 48.22)
Base line CD4 cell counts	less than 200	426(45.7%)	346(81%)	80(19%)	50.32	(48.52 , 52.13)
	200 or above	507(54.4%)	444(87.6%)	63(12.4%)	52.59	(50.96 , 54.23)
Substance use (alcohol)	No	761(81.6%)	663(87%)	98(13%)	52.41	(51.06 , 53.77)
	Yes	172(18.4%)	127(73.8%)	45(26.2%)	48.47	(45.76 , 51.18)
TB co-infection	No	743(79.6%)	642(86.4%)	101(13.6%)	52.86	(51.54 , 54.18)
	Yes	190(20.4%)	148(78%)	42(22%)	47.13	(44.32 , 49.94)
Over All			84.7%	15.3%	51.50	(50.30 , 52.73)

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016; mean: mean survival time, CI: Confidence Interval for mean

Table 4.12: Results of the multivariable proportional hazards Cox regression model containing the variables significant at 25% level in the single covariate proportional hazards Cox regression model

Covariates	B	SE	Wald	DF	Sig.	HR	95.0% CI for HR
Baseline weight	-0.414	0.175	5.594	1	0.018	0.661	[0.474 , .915]
Marital status			7.780	4	0.100		
Married	0.743	0.457	2.643	1	0.104	2.101	[0.858 , 5.145]
Divorced	0.236	0.437	0.293	1	0.588	1.266	[0.538 , 2.980]
Windowed	0.657	0.493	1.772	1	0.183	1.928	[0.733 , 5.072]
Functional status			2.983	2	0.225		
Ambulatory	0.025	0.281	0.008	1	0.929	1.025	[0.591 , 1.780]
Bedridden	0.341	0.294	1.342	1	0.247	1.406	[0.790 , 2.502]
Drug regimen			2.691	2	0.260		
AZT-3TC-NVP	0.335	0.205	2.666	1	0.103	1.397	[0.935 , 2.088]
TDF-3TC-EFV	0.125	0.216	0.338	1	0.561	1.134	[0.743 , 1.730]
Education level			17.618	2	0.000		
Primary	-0.504	0.188	7.145	1	0.008	0.605	[0.417 , 0.873]
secondary and above	-0.972	0.231	17.639	1	0.000	0.378	[0.242 , 0.591]
ART Adherence	-0.788	0.174	5.068	1	0.040	0.461	[0.294 , 0.816]
WHO clinical stage			12.216	3	0.007		
stage II	0.201	0.359	0.313	1	0.576	1.222	[0.605 , 2.469]
stage III	0.508	0.355	0.092	1	0.052	1.678	[1.568 , 2.332]
stage IV	0.739	0.328	5.067	1	0.024	2.093	[1.100 , 3.981]
Base line CD4	-0.364	0.175	4.329	1	0.037	0.695	[.4957 , 0.926]
Substance use (alcohol, soft drugs)	0.571	0.186	9.367	1	0.002	1.546	[1.674 , 1.965]
TB co-infection	0.336	0.193	3.030	1	0.048	1.498	[1.106 , 2.124]

Source: *Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016* indicates statistical significance at 0.05 level of significance. SE= Standard Error, DF= Degrees of Freedom, HR= Hazard Ratio, CI = Confidence Interval, B=Parameter Estimate.*

Table 4.13: The five highest differences in the parameter estimates of the variables included in the model in Table 4.6 when the data value for each patient is in turn deleted from the model

Covariates	**Deleted Observation** (i)	$\Delta j(-i) = \widehat{\boldsymbol{\beta}_j}$-$\widehat{\boldsymbol{\beta}_j(\text{-i})}$	$\lvert\Delta j(-i) = \widehat{\boldsymbol{\beta}_j}$-$\widehat{\boldsymbol{\beta}_j(\text{-i})}\rvert$
Baseline weight	315	0.0801931	0.0801931
	309	0.0802415	0.0802415
	419	0.0804372	0.0804372
	40	0.0806846	0.0806846
	624	0.0810465	0.0810465
Education level	876	0.0148935	0.0148935
	834	0.0149413	0.0149413
	440	0.0149916	0.0149916
	214	0.0151039	0.0151039
	117	0.0151124	0.0151124
ART Adherence	607	0.0760346	0.0760346
	591	0.0760713	0.0760713
	453	0.0760852	0.0760852
	293	0.0760861	0.0760861
	72	0.0760885	0.0760885
WHO clinical Stage IV	557	-0.0234476	0.0234476
	264	-0.0244906	0.0244906
	172	-0.0249060	0.0249060
	69	-0.0240998	0.0240998
	30	-0.0252037	0.0252037
Base line CD4	391	0.0952784	0.0952784
	440	0.0953783	0.0953783
	214	0.0954956	0.0954956
	643	0.0962211	0.0962211
	181	0.0965172	0.0965172
Substance(yes)	315	0.0118264	0.0118264
	866	0.0119088	0.0119088
	155	0.0120706	0.0120706
	117	0.0122306	0.0122306
	849	0.0122318	0.0122318
	682	0.0112145	0.0112145

TB (co-infected)	67	0.011345	0.0113453
	214	0.011354	0.0113547
	866	0.011377	0.0113774
	442	0.0114403	0.0114403

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016

Table 4.14: The five highest likelihood displacement values when each observation is in turn deleted from the model in Table 4.6

Deleted observation (*i*)	$\mathbf{LD_i = 2[LL(\beta) - LL(\beta - i)]}$
117	0.00630594
643	0.10305253
866	0.0148392
591	0.10102374
214	0.03259605

Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016

Table 4.15: Univariate Parametric Survival Regression Model analysis for modeling time to death of HIV infected patients.

Covariate	Exponential		Weibull		Log- logistic		Log-normal	
	β (95% CI coef)	p-value	**β (95% CI coef)**	P-value	**β (95% CI coef)**	P-value	**β (95% CI coef)**	**P-value**
Gender	-.11[-.460 , .252]	0.554	-.06 [-.23 , .12]	0.497	-.05[-.24 ,.13]	0.575	-.09[-.31 ,.13]	0.419
Age	-.14[-.48 3 , .197]	0.414	-.01 [-.16 ,.16]	0.981	.02[-.15 ,.19]	0.788	.01[-.21, .22]	0.960
Residence	-.11[-.44 , .224]	0.519	-.05 [-.21 , .11]	0.535	-.08[-.25 ,.09]	0.358	-.13[-.33 , .08]	0.229
Baseline weight	.41[.062 , .762]	0.021	.20 [.03, 37]	0.021	.23[.06 , .42]	0.010	.30[.08, .52]	0.007
Marital status	.11[-.066 , .280]	0.226	.06 [-.02 , .15]	0.140	.09[-.01, .18]	0.060	.12[.02 , .24]	0.27
Functional.s.	-.07[-.307 , .173]	0.584	-.04 [-.15 , .07]	0.498	-.03[-.16 ,.09]	0.593	-.05[-.20 , .11]	0.526
Drug regimen	.14[-.054 , .340]	0.157	.08 [-.010 ,.18]	0.080	.08[-.02 , .18]	0.121	.09[-.03 , .22]	0.126
Education level	.45[.223 , .669]	0.000	.22 [.11 , .33]	0.000	.23[.12 , .36]	0.000	.29[.15 , .44]	0.000
ART adherence	.36[.014 , .698]	0.042	.16 [01 , .32]	0.037	.15[-.03 , .33]	0.097	.22[.01 , .44]	0.045
WHO clinical stage	-.18[-.262 , -.093]	0.044	-.07 [-.12 , .01]	0.035	-.08[-.13,-03]	0.048	-.08[-.13, -.03]	0.073
Baseline CD4	.35[.008 , .692]	0.044	.16 [.01 , .32]	0.043	.18[.01 , .36]	0.038	.23[.02 , .44]	0.033
Substance use	-.31[-.46 , - .25]	0.001	-.27 [-.44,-.09]	0.001	-.29[-.49,-.10]	0.003	-.36[-.59, -.12]	0.004
TB-co-infection	-.24[-.48 , -.197]	0.004	-.24 [-.42,-.06]	0.004	-.28[-.48,-.09]	0.003	-.42[-.65, -.18]	0.001

*Source: Hossana Queen Elleni Mohamad Memorial Hospital, SNNPR, Ethiopia; from 1st February 2011 to 1st January 2016**

indicates statistical significance at 0.1 level of significance, CI = Confidence Interval, β =Parameter Estimate.

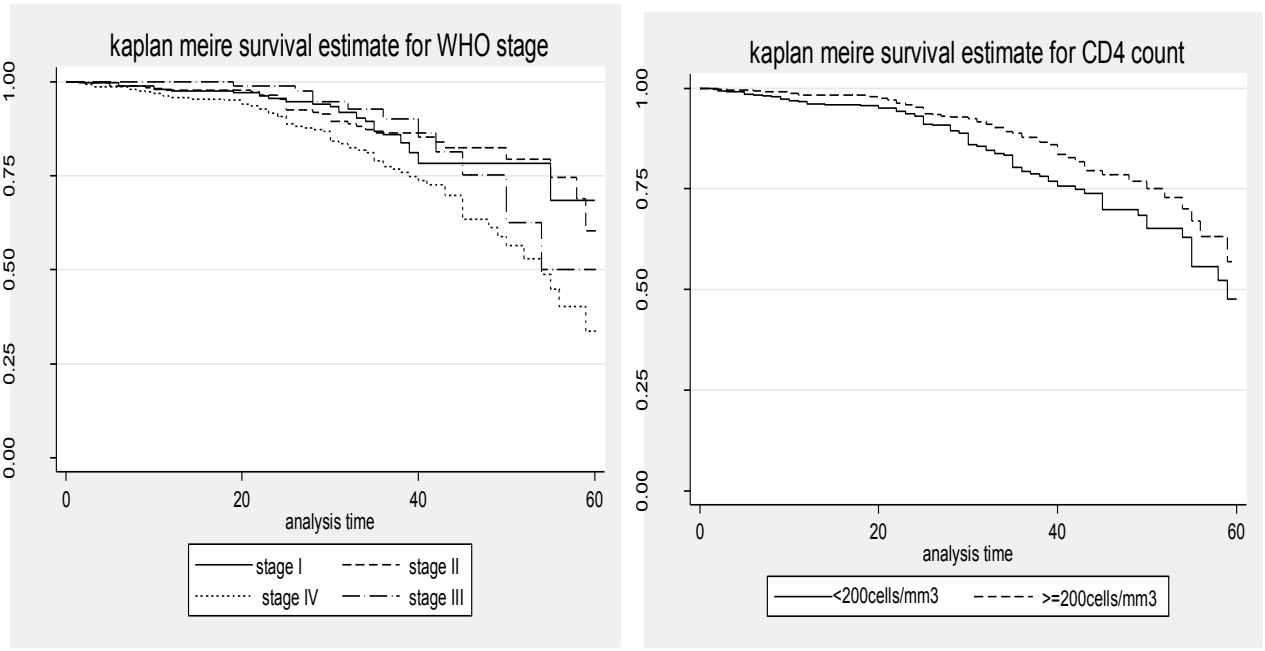


Figure 4.9: plots of Kaplan-Meier survivor functions estimates for the variable WHO clinical Stage and ART baseline CD4

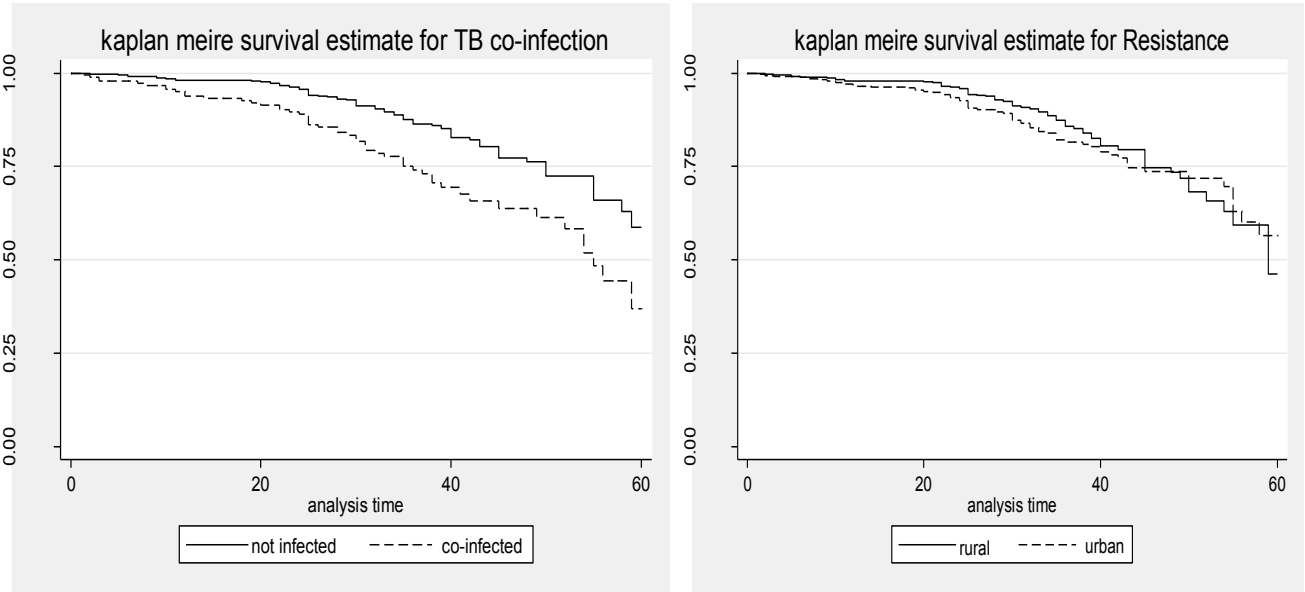
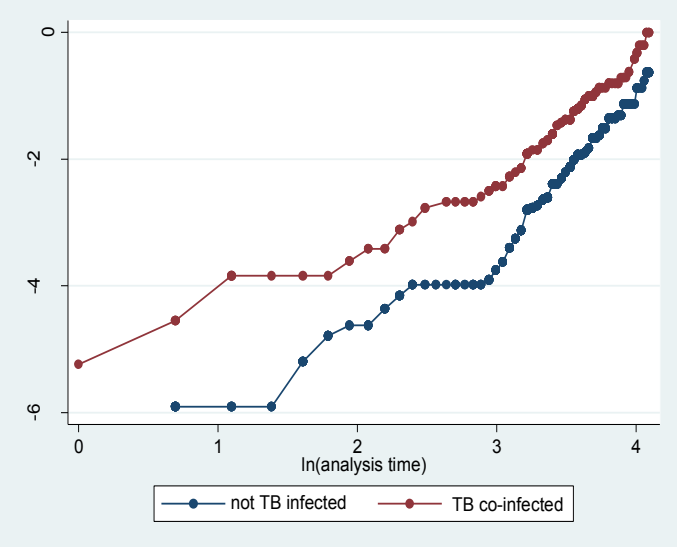
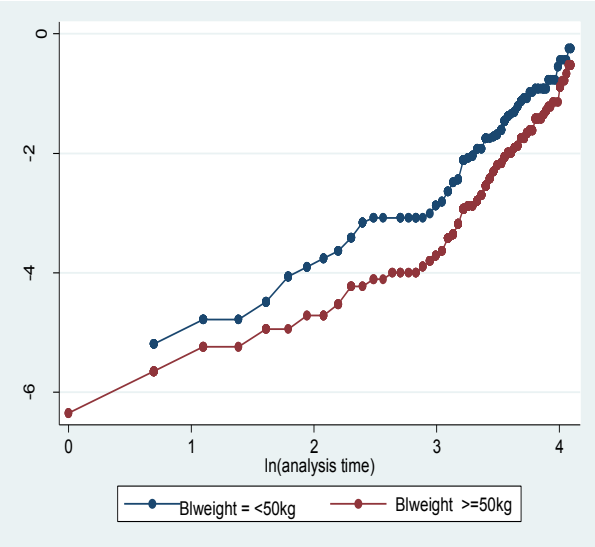


Figure 4.10: plots of Kaplan-Meier survivor function estimates for the variable TB co-infection and Residence

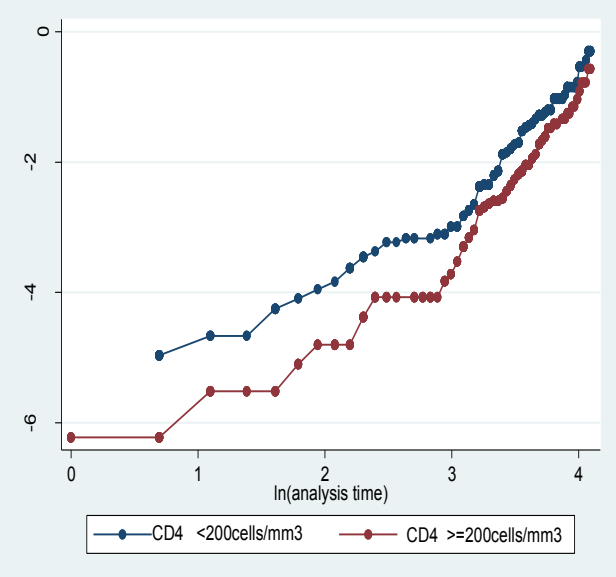
Figures 4.13: (a – e) Plot of log (-log (survival)) versus log survival time for categorical predictors in the fitted model, for TB co-infection, for baseline weight, for CD4 count, for substance use and for ART adherence of the patient seen respectively.



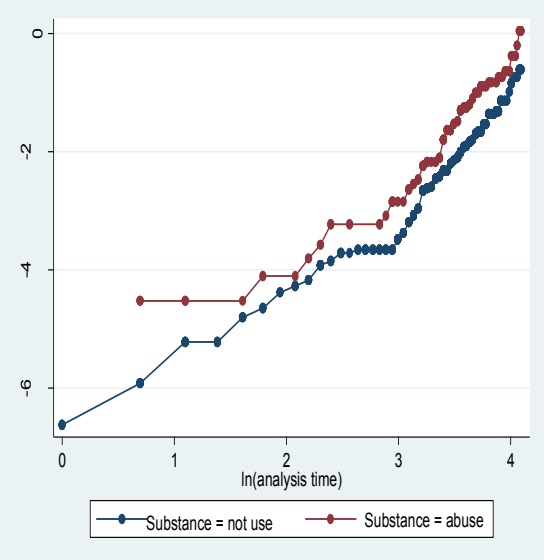
a. For TB co-infection



b. For baseline weight



d. For CD4 count



c. For ART adherence

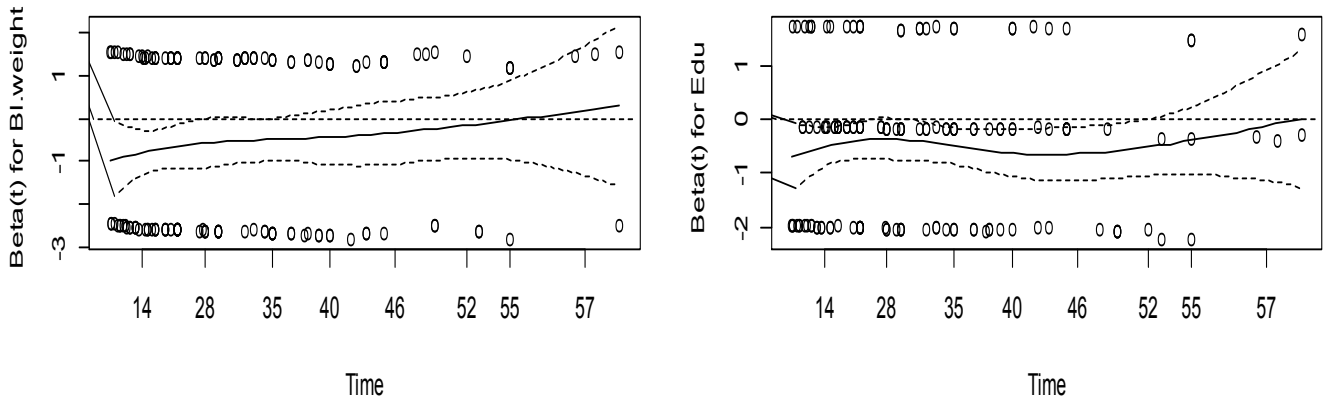


Figure 4.14: The plot of Scaled Schoenfeld residual for baseline weight and level of education respectively, to check the validity of the PH assumption

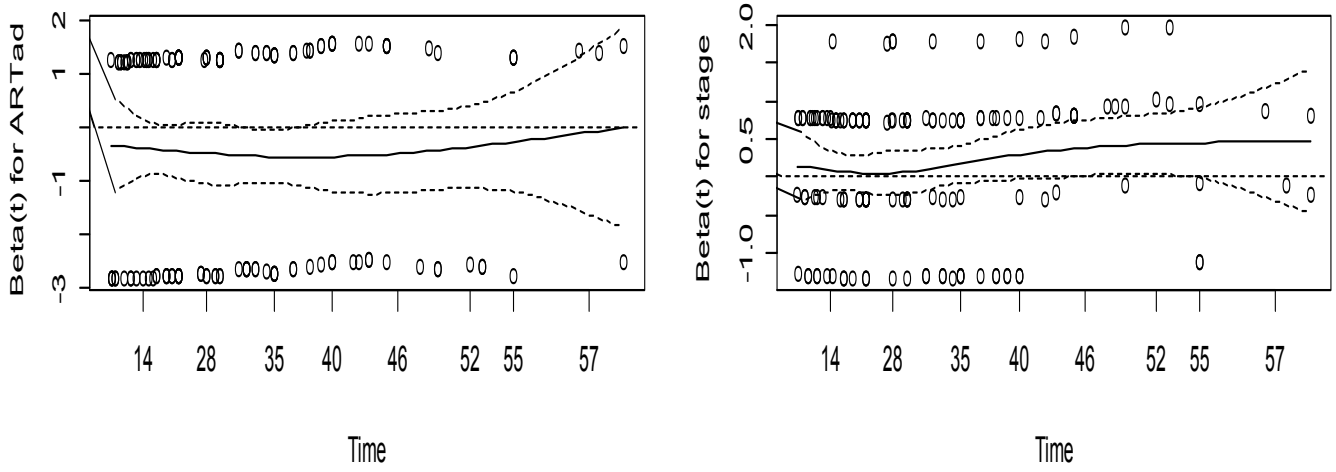


Figure 4.15: The plot of Scaled Schoenfeld residual for ART adherence and WHO clinical stage respectively, to check the validity of the PH assumption