CHAPTER ONE

1.0 INTRODUCTION

1.1 BACKGROUND INFORMATION

Survival analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event (Klembaum,1996).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.

The engineering sciences have also contributed to the development of survival analysis which is called failure time analysis since the main focus is in modeling the lifetimes of machines or electronic components (Lawless,1982).The developments from these diverse fields have for the most part been consolidated into the field of survival analysis. Because these methods have been adapted by researchers in different fields, they also have several different names: event history analysis (sociology), failure time analysis (engineering), duration analysis or transition analysis (economics). These different names do not imply any real difference in techniques, although different disciplines may emphasize slightly different approaches. Survival analysis is the name that is most widely used and recognized (Lee and Wan, 2003).

1.2 Models for Survival Data

The complexities provided by the presence of censored observations led to the development of a new field of statistical methodology. The methodological developments in survival analysis were largely achieved in the latter half of the 20th century. Although Bayesian methods in survival analysis (Ibrahim and Sinha,2001) are well developed and are becoming quite common for survival data, my application will focus on frequentist methods. One of the oldest and most straight forward non-parametric methods for analyzing survival data is to compute the life table, which was proposed by Berkson and Gage (1950) for studying cancer survival. One important development in non-parametric analysis methods was obtained by Kaplan and Meier (1958). While non-parametric methods work well for homogeneous samples, they do not determine whether or not certain variables are related to the survival times. This need

leads to the application of regression methods for analyzing survival data. The standard multiple linear regression models are not well suited to survival data for several reasons. Firstly, survival times are rarely normally distributed. Secondly, censored data result in missing values for the dependent variable (Klembaum, 1996). The Cox proportional hazards (PH) model is now 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, which was discussed in the books of Fleming and Harrington (1991) and of Andersen et al (1993).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 stratified Cox model (Klembaum,1996) or Cox model with time-dependent variables (Collett,2003),can be used for the analysis of survival data. The accelerated failure time (AFT) (Collett,2003) model is another alternative method for the analysis of survival data. The purpose of this study is to compare the performance of the Cox models and the AFT models. This will be studied by means of real dataset from a cohort on treatment for HIV/AIDS.

1.3 Statement of the Problem

The proportional hazards model (Cox 1972) and the accelerated failure time models are the two major approaches to the regression analysis of censored data (Cox and Oakes 1984). Due to the availability of efficient inference procedures that are implemented in all statistical software packages, the proportional hazards model is used almost exclusively in practice. As reported by Reid (1994) the accelerated failure time models i.e. standard parametric models such as Weibull, Exponential and Lognormal are accelerated failure time models. These models are “in many ways more appealing because of its quite direct physical interpretation, “especially when the response variable does not pertain to failure time. This model may provide more accurate or more concise summarization of the data than the proportional hazards model in certain applications. Despite all these advantages Accelerated Failure Time Models are least used in medical research.

1.4 Objectives

1.4.1 General Objective

The main aim of this study is to compare Accelerated Failure Time and COX Proportional Hazards Models at HAART inception in a cohort of HIV infected adults in determining survival of patients.

1.4.2 The specific objectives are to:

1. Apply cox proportional hazard and accelerated failure time models on real data set.
2. Evaluate the models using Akaike Information Criterion.

1.5 Significance of the study

The outcome of this study would provide information about the risk factors or the most influential covariate that have significant impact on survival of HIV patients during treatment. Laboratory measurements, such as numbers of CD4 cells and levels of plasma HIV RNA, are helpful in determining the stage of infection and may serve as prognostic markers. Other factors may also influence outcome. This study will look at a number of factors including demographic and other host factors that may play a role in disease progression as well as describing the important impact antiretroviral therapy has had in disease progression. The study will try to identify death risk extent of patients under these significant factors at different time during their care. The study will also help in comparing the utility of COX and accelerated failure time (AFT) models, findings will help in making a decision as to which model to apply under specified conditions defined by predictor variables. Findings of the study will also give a deeper insight on how the concept of standardized measure of variability and Akaike information criterion can be applied in survival analysis.

1.6 Limitation of the Study

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

1.7 Research Questions

1. Between accelerated failure time model and cox proportional hazard model which is more efficient in analysis of time to event data?
2. How can Akaike information criterion be applied in survival analysis?

CHAPTER TWO

2.0 REVIEW OF LITERATURE

2.1 Introduction to the Chapter

The literature selected and discussed in this section are those that are more related and relevant to this study. Studies or research related to the survival of HIV/AIDS especially in Kenya are scarce. The literature review given below has three parts. The first part is a general overview of HIV/AIDS pandemic and the Kenya's situation. The second part concerns survival analysis in health research and the third is about Akaike Information System application in health research.

2.2 HIV/AIDS

Human Immunodeficiency Virus (HIV) is the virus that causes acquired immunodeficiency syndrome (AIDS). Being a member of a group of viruses called retroviruses. HIV infects human cells and uses the energy and nutrients provided by those cells to grow and reproduce. AIDS is a disease in which the body's immune system breaks down and is unable to fight off certain infections, known as "opportunistic infections", and other illnesses that take advantage of a weakened immune system. When a person is infected with HIV, the virus enters the body and lives and multiplies primarily in the white blood cells. These are the immune cells that normally protect us from disease. The hallmark of HIV infection is the progressive loss of a specific type of immune cell called T-helper or CD4 cells.

As the virus grows, it damages or kills these and other cells, weakening the immune system and leaving the individual vulnerable to various opportunistic infections and other illnesses, ranging from pneumonia to cancer (free encyclopedia Wikipedia. Facts about HIV/AIDS, 2001). The U.S. Centers for Disease Control and Prevention (2011) defines someone as having a clinical diagnosis of AIDS if they have tested positive for HIV and meet one or both of these conditions:

1. They have experienced one or more AIDS-related infections or illnesses
2. The number of CD4 cells has reached or fallen below 200 cells per cubic milliliter of blood (a measurement known as cd4-cell count)

In healthy individuals, the CD4 count normally ranges from 450 to 1200 cells/mm^3.For many years, there were no effective treatments for AIDS. Today, people in Kenya, other developing and developed countries can use a number of drugs to treat HIV infection and AIDS.

Some of these are designed to treat the opportunistic infections and illnesses that affect people with HIV/AIDS. In addition, several types of drugs seek to prevent HIV from reproducing and destroying the body's immune system. Many HIV patients are taking several of these drugs in combination a regimen known as highly active antiretroviral therapy (HAART). When successful, combination or "cocktail" therapy can reduce the level of HIV in the bloodstream to very low, even undetectable, levels and sometimes enable the body's CD4 immune cells to rebound to normal levels.

2.3 The Burden in HIV/AIDS

With around 2.6 million people becoming infected with Human Immunodeficiency Virus in 2009,there are now an estimated 33 million people around the world who are living with HIV, including millions who have developed AIDS,(AVERT, 2011). Since the beginning of the epidemic, AIDS has killed nearly 19 million people worldwide. AIDS has replaced malaria and tuberculosis as the world's deadliest infectious disease among adults and is the fourth leading cause of death worldwide.

In 2011 Kenya estimates that approximately 6.2% of the adult population is HIV-infected. HIV prevalence in Kenya is believed to have peaked in 1995–1996, at 10.5%, subsequently falling by approximately 40% and remaining relatively stable for the last several years. Historically a key marker for national progress in the AIDS response, HIV prevalence becomes more difficult to interpret as antiretroviral treatment is scaled up. Because treatment extends life and reduces rates of AIDS deaths, increases in HIV prevalence are likely even with incremental declines in the rates of new infections. Accordingly, performance indicators for Kenya’s most recent national AIDS strategy project a relatively modest decline in HIV prevalence between 2007 and 2013, with an actual uptick on overall HIV prevalence anticipated over time due to the health benefits of improved treatment access.

An estimated 1.6 million Kenyans were living with HIV in 2011. This equals the peak number of HIV-infected people that had previously been maintained annually between 1996 and 2002, and it represents a nearly four-fold increase over the 400,000 people estimated to be living with HIV in Kenya in 1990. Kenya has the third largest population of people living with HIV in sub-Saharan Africa and the highest national HIV prevalence of any country outside Southern Africa (UNAIDS, 2008). As people living with HIV are living longer as a result of improved

access to HIV treatment, it is anticipated that the total number of HIV-infected individuals in Kenya will continue to increase, approaching 1.8 million by 2015.

There is considerable geographic variability in the burden of HIV in Kenya. Provincial HIV prevalence ranges from a high of 13.9% in Nyanza Province to a low of 0.9% in North Eastern Province a more than 15 fold variation (Kenya National Bureau of Statistics, 2010). Nyanza Province alone accounts for one in four HIV-infected people in Kenya.

Kenya's epidemic disproportionately affects women who account for 59.1% of adults living with HIV. Among people aged between 15 and 49 years, HIV prevalence among women (8.0%) is nearly twice that among men which is 4.3% (Kenya National Bureau of Statistics, 2010). The odds of being infected increase as individuals' transition from adolescence to adulthood. Although HIV is most likely to affect young adults, a considerable number of older people are living with HIV. In 2008–2009, roughly one out of 11 (9.1%) Kenyan men ages 50–54 were HIV-positive (Kenya National Bureau of Statistics, 2010).

For Kenyans as a whole, urban residents have historically more likely to be HIV-infected than rural dwellers (Kenya National Bureau of Statistics, 2010). However, there is a notable distinction between men and women in this regard, with men in rural areas more likely to be HIV-infected than their urban counterparts (4.5% vs. 3.7%) (Kenya National Bureau of Statistics, 2010). Over time, HIV prevalence in urban and rural settings has converged, with HIV prevalence in urban areas only modestly higher than prevalence in rural settings. HIV affects Kenyans from all socioeconomic strata. Highest HIV prevalence (7.2%) is among the top wealth quintile, with the second highest HIV prevalence among the second lowest (6.8%). The poorest Kenyans (lowest wealth quintile) are least likely to be living with HIV, with a prevalence of 4.6%.For sub-Saharan Africa generally, educational attainment is inversely correlated with HIV risk for women, at least according to surveys conducted over the last 10–15 years (Hargreaves et al, 2008). In Kenya, this pattern is not so clearly established. Although women with secondary education or higher have lower HIV prevalence (6.9%) than women who completed only primary education (8.9%), lowest HIV prevalence is reported among women with no education (5.8%) (Kenya National Bureau of Statistics, 2010).

Muslim Kenyans have HIV prevalence roughly half the national average (3.3%), compared with 5.9% of Roman Catholics and 6.6% of people of Protestant or another Christian denomination (Kenya National Bureau of Statistics, 2010). Among Kenyan tribes, the Luo are

notably more likely to be living with HIV than other ethnicities, with more than one in five Luo (20.2%) testing HIV-positive in the 2008–2009 national household survey (Kenya National Bureau of Statistics, 2010). Somalis have the lowest HIV prevalence of any ethnicity (0.8%).

2.4 HIV/AIDS Mortality in Kenya.

Since the epidemic began, HIV has claimed the lives of at least 1.7 million people in Kenya. In 2011, an estimated 49,126 people in Kenya died of AIDS-related causes. The AIDS death toll in 2010 represents a nearly two-thirds drop from the peak in AIDS deaths in 2002–2004, when an estimated 130,000 people died each year. Peak mortality followed peak HIV incidence in Kenya by roughly a decade, which is expected given the roughly 10-year life expectancy of a newly infected individual in the pre-ART era. Were current trends to continue, Kenya would achieve its 2013 target for reducing the annual number of AIDS deaths to 61,000 or lower. Indeed, current projections indicate that 26,720 Kenyans are likely to die of AIDS-related causes in 2013.

2.5 The Impact of HIV in Kenya

The epidemic continues to have far-reaching social, economic, health and population effects. In addition to the harms directly inflicted on HIV-infected individuals and the households in which they live, AIDS has had indirect effects that are nevertheless real and substantial on communities and the whole of society.

In particular, HIV infection results in severe economic consequences for affected households (Bates et al., 2004). One out of nine households in Kenya has been affected by AIDS, with the head of household having HIV in more than three out of four AIDS-affected households (NASCOP, 2009).

The epidemic has resulted in a sharp deterioration of basic health indicators. Between 1998 and 2003 or roughly between the epidemic's peak in Kenya and the early introduction of antiretroviral therapy , the adult mortality rate (ages 15–49) rose by 40% for women and by 30% among men (Gelmond et al., 2009, citing findings from consecutive Demographic and Health Surveys). With a large number of newborns newly infected each year, the epidemic has also increased mortality among children under five (Gelmond et al., 2009).

The concentration of the epidemic's burden among young adults has visited particular hardships on Kenya's children, regardless of whether children themselves become HIV positive (K'Oyugi,2002). In 2011, an estimated 1.1 million children in Kenya had lost one or both parents

to AIDS. Kenyan children with one or more HIV infected parents are significantly less likely than other children to be in school, more likely to be underweight, and less likely to receive basic medical care (Mishra et al., 2005).

While children have experienced among the harshest effects of the epidemic, AIDS has burdened Kenyans from all age strata and all walks of life. Nearly one in five (18%) Nairobi residents over age 50 report having been personally affected by AIDS, such as becoming infected, caring for an AIDS patients or orphaned child, or losing a loved one (Kyobutungi et al., 2009).

AIDS appears to have affected fertility patterns. On average, HIV infected women have 40% fewer children than the norm (Akinyi et al, 2010). HIV-infected women are notably less likely to express a desire for a child within the next two years than women who had tested HIV-negative or who had not received HIV test results; women living with HIV are also significantly more likely than other women to report not desiring to have a child at any point in the future (NASCOP, 2009).

2.6 Treatment and Care for People Living With HIV: The Challenge of Sustaining Recent Gains.

Over the last several years, the HIV landscape in Kenya has been transformed by the rapid expansion of access to life-preserving antiretroviral therapy. The scaling-up of treatment programmes throughout Kenya has reduced HIV-related morbidity and mortality, prevented vulnerable households from falling deeper into poverty, rejuvenated entire communities, helped alleviate the stigma long associated with HIV infection, supported national efforts to improve maternal and child health, and contributed to gains in Kenya's fight against tuberculosis. These achievements are nothing short of historic. Kenya's most recent national HIV strategy – KNASP III – emphasizes the country's long-term commitment to HIV treatment access. During the four-year (2009–2013) period covered by KNASP III, it is projected that treatment and care will account for 57.9% of all HIV-related spending (NACC, 2009).

In this ongoing national undertaking to achieve universal treatment access, important challenges remain. With the number of people who will need antiretroviral treatment in the future outweighing those who are currently medically eligible, it is clear that sustaining treatment access will demand unflagging national commitment for decades to come. Uncertainties regarding future international funding for continued treatment scale-up, as well as the inevitable

growth over time in demand for costly second-line antiretroviral regimens, merely underscore the many challenges that await AIDS stakeholders in the coming years.

2.7 Approach to HIV Treatment and Care

Consistent with international recommendations, Kenya has adopted a public health approach to HIV treatment scale-up. National guidelines specify standardized first-line regimens, mandate routine patient monitoring, offer guidance on how and when to change regimens, and identify evidence-based approaches to clinical practice, including management of common treatment related toxicities and drug interactions. These guidelines have also formed the basis for extensive clinical training and capacity-building initiatives to ensure that diverse cadres of health care workers have the needed competence to play their respective roles in the administration of HIV treatment and care.

A national network of Comprehensive Care Clinics facilitates the ready access of people living with HIV to treatment, care and support services (IPPF et al., 2008). Roughly one in six health facilities (16%) were providing antiretroviral therapy in 2010. The number of facilities administering antiretroviral therapy increased from 731 in 2008 to 1,171 by early 2011 for adults and 1,105 for pediatric. As of December 2011, 1,405 facilities (including 1,242 public sector facilities) offered antiretroviral therapy.

In 2009, 213,521 HIV-positive patients were newly enrolled in HIV care, representing roughly one-third of cumulative enrolment (621,813) (NASCOP, 2010). The number of newly enrolled females in 2009 (140,639) was roughly twice the number of new male enrollees (72,882) (NASCOP, 2010). In 2009, 52% (111,744) of newly enrolled patients were started on antiretroviral therapy in 2009 (NASCOP, 2010).

Kenya has exempted people living with HIV from the usual cost-sharing requirements for antiretroviral therapy and treatment for tuberculosis. However, patients may remain liable for certain costs associated with nutritional support, laboratory investigations and treatment of opportunistic infections. The Government of Kenya has long recognized the value of engaging families and communities in treatment efforts. In 2002, the National AIDS/STD Control Programme in the Ministry of Health issued guidelines to aid families and community workers in undertaking home-based care for people living with HIV (NASCOP, 2002). KNASP III calls for initiatives to strengthen home and community-based care services (NACC, 2009).

2.8 Antiretroviral Therapy

Since the detection of first AIDS case in mid 80s, Kenya primarily provided symptomatic treatment and palliative care and later slowly introduced mono and dual therapy especially in the private sector. In the mid-1990s, the emergence of a new class of antiretroviral compounds ushered in the era of highly active antiretroviral therapy.

Antiretroviral therapy was first introduced through the private sector in the late 1990s but only became widely available through the private sector beginning in 2003–2004. Declines in the price of antiretroviral drugs, abetted in part by generic competition, have enabled the country to progressively increase coverage of antiretroviral treatment.

2.9 Initiation of Antiretroviral Therapy

Consistent with changes in international practice, as set forth in updated WHO recommendations, Kenya recommends the early initiation of antiretroviral therapy. Prior to 2007, Kenya provided for initiation of therapy when a patient's CD4 count fell below 200 cells/mm^3. In 2007, as evidence began to suggest that earlier initiation of therapy was advisable, Kenya began using a CD4 threshold of 250cells/mm^3 for starting therapy. As of June 2010, Kenya began calling for antiretroviral treatment to begin once a patient's CD4 count reached or fell below 350cells/mm^3. Kenya standard operating procedure requires definitive evidence of a positive HIV test result before antiretroviral therapy may be prescribed, although national guidelines permit clinicians to presume a diagnosis of HIV in symptomatic children when virologic confirmation is not possible. These changes have significantly increased the number of adults who are eligible for treatment, with further increases anticipated in the coming years.

Kenya's recommendations for starting treatment in HIV-infected children have also evolved to reflect expanded knowledge of state-of-the-art approaches. Drawing on emerging evidence of optimal approaches to the management of HIV infection in children, Kenya recommended, beginning in 2009, that all HIV-positive children under 18 months be started on antiretroviral therapy (NASCOP, 2009).

The changes in eligibility requirements for antiretroviral therapy have significantly increased the number of children in need of therapy. In 2011, more than 150,000 children in Kenya were eligible for antiretroviral therapy. As a result of scaled-up services to prevent mother-to-child transmission, it is anticipated that the number of children in need of therapy will decline in future years.

2.10 Predictors of HIV/AIDS survival

In a study by Moore et al., (2006), where they aimed to determine the prognostic value of baseline CD4 percentage in terms of patient survival in comparison to absolute CD4 cell counts for HIV-positive patients initiating highly active antiretroviral therapy (HAART). In the study a cohort study of 1623 antiretroviral therapy-naive HIV-positive individuals who initiated HAART between 1 August 1996 and 30 June 2002 was conducted. Cumulative mortality rates were estimated using Kaplan–Meier methods. Cox proportional hazards regression was used to model the effect of baseline CD4 strata and CD4 percentage strata and other prognostic variables on survival. A subgroup analysis was conducted on 417 AIDS-free subjects with baseline CD4 counts between 200 and 350cells/ml. In multivariate models, low CD4 percentages were associated with increased risk of death (CD4 %< 5, relative hazard (RH) = 4.46; CD4% 5–14, RH =2.43; P<0.01 for both) when compared with those subjects with an initial CD4 fraction of 15% or greater, but had less predictive value than absolute CD4 counts. In subgroup analyses where absolute CD4 strata were not associated with mortality, a baseline CD4 fraction below 15% (RH=2.71; 95% confidence interval (CI) 1.20–6.10), poor adherence to therapy and baseline viral load 4100 000 HIV-1 RNA copies/mL were associated with an increased risk of death. They concluded that CD4 percentages below 15% are independent predictors of mortality in AIDS-free patients starting HAART, including those with CD4 counts between 200 and 350 cells/mm^3. CD4 percentages should be considered for inclusion in guidelines used to determine when to start therapy.

2.11 Non-Parametric and Parametric Models for Studying Time to Event

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

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

From the study above the relative hazard produced in semi parametric and parametric proportional hazards modeling helped researchers identify risk factors for an outcome of interest. Parametric models in the accelerated failure time metric are not commonly used despite the time ratio being a more easily interpretable measure of association than the relative hazard. AFT models also facilitate the estimation of elapsed time between exposure and outcome, which has more clinical interpretability than a hazard ratio. In the analysis of the above study illicit drug use was associated with a doubling of the hazard of rebound with resistance even after adjustment by other factors.

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

parametric models that encompass a wide variety of hazard shapes (Sethi, et al, 2009). In a study by Pierre De Beaudrap and et al (2008), 404 HIV-1- infected Senegalese adult patients were enrolled and data censored as of September 2005. Predictor effects on mortality were first examined over the whole follow-up period (median 46 months) using a Cox model and Shoenfeld residuals. Then, changes of these effects were examined separately over the early and late treatment periods; i.e., less and more than 6-month follow-up. They found out that during the early period, baseline body mass index and baseline total lymphocyte count were significant predictors of mortality (Hazard Ratios 0.82 [0.72-0.93] and 0.80 [0.69-0.92] per 200 cell/mm^3, respectively) while baseline viral load was not significantly associated with mortality. During the late period, viro-immunological markers (baseline CD4-cell count and 6-month viral load) had the highest impact. In addition, the viral load at 6-month was a significant predictor (HR = 1.42 [1.20-1.66]). They concluded that impaired clinical status could explain the high early mortality rate while viro-immunological markers were rather predictors of late mortality.

This study underlined changes over time in mortality predictors among HIV-1 infected patients. Disappearance of the predictive value of prognostic variables may often occur in medical studies .The previous finding of an early peak in mortality rate prompted them to investigate more carefully the early period after HAART initiation. Whereas the effect of baseline CD4 cell count on the hazard of death remained roughly constant during the whole follow-up, the impacts of BMI and of total lymphocyte count were important immediately after HAART initiation before fading out after 6 months.

Clinical variables (BMI, CDC stage) were strongly associated with early death but not with death after 6 months. In the study cohort, the clinical status at enrolment may be, in average, more advanced than in other cohorts as shown by the low median CD4 cell count at baseline (128 in study cohort versus 168 in the Euro SIDA cohort, 192 in the Swiss cohort, and 250 in the panel of the ART-CC cohorts) and by the clinical stage (more than half of the cohort under study were in CDC stage C versus 25% in the Swiss cohort 34 and 21% in the ART-CC).

Anaemia was an independent and stable predictor of death as previously found in the Euro SIDA cohorts (Lundgren et al., 2003). The pathogenesis of anaemia in AIDS remains complex. According to the study although malaria could play a role in addition to inflammatory, deficiency and bleeding causes, it is unlikely that this effect was important in the highly urbanized setting of Dakar. It was interesting to note that total lymphocyte count at baseline was

a stronger predictor than CD4 cell count at baseline during the early period despite the correlation between the two counts. Similar results were found in children and this may reflect the better prognosis value of CD4 cell percentage than that of CD4 cell count. It is important to note that the total lymphocyte count has been proposed as a surrogate for CD4 cell count in low income country settings (WHO 2006).However, in their study, the total lymphocyte count was predictive only over very short time-to-event periods, which means that therapy initiation on basis of total lymphocyte count only may occur too late. They pointed out that additional studies are needed to assess the prognosis value of total lymphocyte count. Changes in the extent of predictor effects between the early and the late period were striking. Indeed, the effect of BMI and of total lymphocyte count at baseline disappeared.

Therefore, these variables have a high prognostic value during the first months after HAART initiation, but lost it later. On the other hand, whereas they did not find significant association between viral load measured at baseline and subsequent risk of death, viral load measured at 6 months was noted to be important predictor of death for patients who survived until 6 months. They pointed out that this was consistent with other studies that did not find a significant association between the viral load at baseline and the risk of death in advanced stage disease with low CD4 cell count. Also the initial response to HAART as assessed by the viral load at 6 months has a great prognostic value; this has been already demonstrated in developed countries (Egger M et al., 2002). This result emphasizes the importance of updating viral load in patient monitoring after HAART initiation in a low-income setting. The CD4 cell count at baseline remained a predictive factor for death after 6 months, with a stronger effect than the 6-month CD4 cell count. They did put forward several hypotheses to explain the disagreement between this result and those of other studies (Egger et al, 2002).Firstly; they argued that missing data may reduce the effect of the CD4 cell count at 6 months. In order to examine the effect of missing data, they used the same Cox model with and without that variable in the sub-population with known CD4 value at 6 months and found the same effects and significance for all the other covariates. Secondly, these results were applicable only to patients who survived at least 6 months after enrolment and therefore to a subset of the study population. At baseline, time since infection varied greatly between patients of the cohort, who may have produce a mixture of sub-populations with different prognoses and may explain their results. The more frailty patients would be characterized by clinical variables and their results argue for the use of immunological

markers instead of clinical stage or total lymphocyte count to decide initiation of HAART even in low-income countries. Conversely they recommended that, patients with an advanced disease, a low BMI or a functional dependence should be carefully monitored and more intensive care could be proposed during the first months.

In another study by Jon Michael Gran et al (2010), they argued that when estimating the effect of treatment on HIV using longitudinal data, standard methods may produce biased estimates due to the presence of time-dependent confounders. Such confounding can be present when a covariate, affected by past exposure, is both a predictor of the future exposure and the outcome, they gave an example of CD4 cell count, being a marker for disease progression for HIV patients, but also a marker for treatment initiation and influenced by treatment. Fitting a marginal structural model (MSM) using inverse probability weights is one way to give appropriate adjustment for this type of confounding. In their paper they studied a simple and intuitive approach to estimate similar treatment effects, using observational data to mimic several randomized controlled trials. Each trial was constructed based on individuals starting treatment in a certain time interval. An overall effect estimate for all such trials was found using composite likelihood inference. The method offered an alternative to the use of inverse probability of treatment weights, which is unstable in certain situations. The estimated parameter was not identical to the one of an MSM; it was conditioned on covariate values at the start of each mimicked trial. This allowed the study of questions that were not that easily addressed fitting an MSM. The analysis could be performed as a stratified weighted Cox analysis on the joint data set of all the constructed trials, where each trial was one stratum. The model was applied to data from the Swiss HIV cohort study. In their study an example of a time-dependent confounder was estimating treatment effects for HIV is the CD4 cell count, which, as an indicator of immune status, it is a predictor of treatment and outcome (AIDS or death), while at the same time influenced by treatment. To deal with this type of confounding, Robins *et al,* (2000) introduced a new type of model, called the marginal structural model (MSM). When fitting an MSM, time-dependent confounding is typically adjusted for using inverse probability of treatment (IPT) weighting. Each individual's probability of being treated is calculated conditioned on their observed covariates at each time point, which then are used to construct the IPT weights for that individual. The time-dependent confounding variables are no longer predictors of the exposure in the weighted analysis. The rest of the parameters in the MSM can therefore be estimated using a

weighted time-dependent Cox analysis, adjusting only for baseline covariates. Even though IPT weighting is an elegant way to adjust for time-dependent confounding, it has properties that make the weights unstable in certain situations.

The main problem lies in the instability of the estimated weights at the time where individuals go from being off treatment to on treatment. When the conditional probability of initiating treatment is small, the denominator in the expression for the weight can be close to zero, making the estimated weights unstable. In other words, individuals with unusual covariate histories when starting treatment can be given very large weights. The fact that the individuals keep this weight constant for their remaining event history after initiating treatment adds to the problem. In their paper they consider an alternative approach to time-dependent confounding, than the IPT weights used to fit an MSM. Their method was seeking to estimate a similar treatment effect as the MSM, but now by looking at the causal or counterfactual effect of treatment in many mimicked randomized controlled trials, each trial being distinguished by the time of treatment start. This approach also allowed them to investigate some questions that would not be that easy to answer with an MSM; such as estimating separate treatment effects for individuals with different CD4 counts at treatment start. Where in the MSM the time-dependent confounding is typically adjusted for using weighting, they considered a method of many successive Cox analyses, comparing the event histories of individuals starting treatment and the ones not yet on treatment in different time intervals separately. Individuals not on treatment by the start of the trial were considered to be artificially censored at the time of later treatment start.

One of the motivations behind the use of sequential Cox approach in the study was to look at alternatives to IPT weighting. In the sequential Cox method the IPT weights were avoided, partly by using artificial censoring to censor individuals at later treatment start. It was to be expected that individuals with certain covariate histories were more likely to get artificially censored due to later treatment start than others, which would make the artificial censoring dependent on disease history. In addition, ordinary censoring could also be dependent. To adjust for this bias, both types of dependent censoring were accounted for using IPC weighting. The problem of unstable IPT weights was based on the fact that the weights for individuals on treatment were calculated using the inverse of the probability of starting treatment. That way, an estimated small probability of starting treatment was to give a large weight. IPC weights were only calculated using the probability of not being censored. Considering this there were usually

no situations which would involve dividing by a number close to zero. Thus, the same problem of unstable weights was not present. In summary, Jon Michael Gran *et al*(2010) made five main assumptions for their estimate of the treatment effect to be a causal estimate; these are (i) the chosen covariates are sufficient to adjust for confounding, (ii) the model for estimating the hazard rate is correct, (iii) the model estimating the weights used to adjust for any dependent censoring is correct, (iv) the effect of treatment is the same in all mimicked trials, and (v) the effect is the same for all covariate histories before the start of the mimicked trials given covariates at the starting time.

2.11.1 Accelerated Failure Time Models for Survival Analysis in Studies with Time-Varying Treatments

As it is widely believed two useful models for survival analysis are the Cox proportional hazards model and the accelerated failure time (AFT) model. The widely used Cox model measures causal effect on the hazard (rate) ratio scale, whereas the less used AFT model, measures causal effect on the survival time ratio scale. Both the Cox model and semi parametric versions of the AFT model according to Miguel et al (2005) are models that leave the baseline hazard (or, equivalently, the baseline survival distribution) unspecified. However, even in the absence of unmeasured confounding and model misspecification, these standard models for survival analysis will provide estimates that fail to have a causal interpretation when: (i) there exists a measured time-dependent risk factor for survival that also predicts subsequent treatment, and (ii) past treatment history predicts subsequent risk factor level. Factors that meet condition (i) are known as time-dependent confounders. For example, when estimating the causal effect of highly active antiretroviral therapy (HAART) on the survival of individuals infected with the human immunodeficiency virus (HIV), condition (i) is met by the variable CD4 cell count because a low CD4 cell count is both a risk factor for survival and used by clinicians to decide whether to initiate HAART. Also, condition (ii) is met because prior HAART use increases CD4 cell count. Therefore, including the time-dependent confounder CD4 cell count in a standard Cox or AFT model may not appropriately adjust for confounding. In contrast to standard Cox and AFT models, structural Cox and AFT models can be used to estimate causal effects when conditions (i) and (ii) hold. Marginal structural Cox model has previously been used to estimate the causal effect of HAART on the hazard of AIDS or death of HIV-infected individuals. The causal hazard ratio from the marginal structural model was 0.54 (95% confidence interval [CI]:

0.38, 0.78) when comparing continuous treatment with HAART versus no treatment with HAART. This hazard ratio was estimated by inverse probability weighting. The simultaneous presence of conditions (i) and (ii), and thus the problem of time-dependent confounding by factors affected by prior treatment, is ubiquitous in pharmaco epidemiology. Other examples of time-dependent confounders that are affected by prior treatment are upper gastrointestinal bleeding when studying the effect of NSAIDs on gastric cancer, measures of disease severity when studying the effect of methotrexate on the mortality of patients with rheumatoid arthritis and hematocrit when studying the effect of erythropoietin on the mortality of dialyzed patients.

In their paper Miguel et al., (2005) reviewed the differences between structural models and standard regression models for survival analysis. They describe a structural AFT model, and illustrated the application of this model for estimating the effect of HAART on AIDS-free survival in two prospective cohort studies of HIV-infected individuals. They found out that Nested structural AFT models and marginal structural Cox models can be used to consistently estimate the effect of a time-dependent exposure on survival in the presence of time-dependent confounders affected by prior exposure. On the other hand, standard models for survival analysis may yield biased estimates of causal effect because they adjust for time-dependent confounding by including the confounders as covariates in the model. To avoid this problem, structural models adjust for time-dependent confounding by g-estimation or inverse probability weighting. Using a nested structural AFT model, they estimated that continuous HAART increased survival time by 2.5 fold in the MACS/WIHS.

Their causal effect estimates from a structural AFT model are consistent with those from a marginal structural Cox model. It is reassuring that these two very different methods for estimating causal effects yield similar results, and that both arrive at the same qualitative conclusion as a previously conducted randomized trial (Hammer et al.,1997). In contrast, a standard associational Cox model did not find a substantially lower mortality rate among those treated compared with those untreated with HAART.

According to Donglin and Lin (2007), the accelerated failure time model provides a natural formulation of the effects of covariates on potentially censored response variable. The existing semi parametric estimators are computationally intractable and statistically inefficient. In their article they proposed an approximate non-parametric maximum likelihood method for the accelerated failure time model with possibly time-dependent covariates. They estimated the

regression parameters by maximizing a kernel-smoothed profile likelihood function. The maximization was achieved through conventional gradient-based search algorithms. The resulting estimators were consistent and asymptotically normal. The limiting covariance matrix attained the semi-parametric efficiency bound and could be consistently estimated. They also provide a consistent estimator for the error distribution. Extensive simulation studies demonstrated that the asymptotic approximations were accurate in practical situations and the new estimators were considerably more efficient than the existing ones.

2.12 Akaike Information Criterion (AIC)

AIC is an asymptotically un-biased estimator of the expected relative Kullback-Leibler information quantity or distance (K-L) (Kullback and Leibler, 1951), which represents the amount of information lost when we use model A to approximate model B.upon computing AIC, the preferred model is the one with the smallest AIC value (Akaike, 1974).The AIC for a given model is a function of its maximized log-likelihood (l) and the number of estimable parameters (K):

$$\mathrm{AIC} = -2l + 2\mathrm{K}$$

In a study by Mohamad et al,(2007) where their main objective was to compare two survival regression methods; cox regression and parametric models in patients with gastric adenocarcinomas who registered at Taleghani hospital. They retrospectively studied 746 cases from February 2003 through January 2007. Gender, age at diagnosis, family history of cancer, tumor size and pathologic distant of metastasis were selected as potential prognostic factors and entered into the parametric and semi parametric models. Weibull, exponential and lognormal regression were performed as parametric models with the Akaike Information Criterion (AIC) and standardized of parameter estimates to compare the efficiency of models. For the aim of comparison among parametric and semi parametric models they used Akaike Information Criterion (AIC) and standardized of parameter estimates. The AIC proposed in Akaike (1974), is a measure of the goodness of fit of an estimated statistical model. It is grounded in the concept of entropy. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data.

The survival results from both Cox and Parametric models showed that patients who were older than 45 years at diagnosis had an increased risk for death, followed by greater tumor size and presence of pathologic distant metastasis.The evaluation criteria indicated Cox and

Exponential model are similarly the best models in multivariate analysis and same conclusions in univariate analysis. Although it seems that there may not be a single model that is substantially better than others, the data strongly supported the log normal regression among parametric models in univariate analysis and it can be lead to more precise results as an alternative for Cox.

CHAPTER THREE

3.0 METHODOLOGY

3.1 Introduction to the Chapter

This chapter describes the method that was used to meeting the study objectives. Cox Proportional Hazard and Accelerated Failure Time Models were used on historical data set and evaluated using Akaike information criterion. A description of the research setting, research tools used, research procedure used and the ethical issues relating to the study are also given.

3.2 Study setting

The study was carried out in Karuri Health Center. Karuri Health Centre was started in 1940's; the running of the health Central is by the Centre Government (Ministry of Health). It is classified as a level 3 facility. It offers various services including curative, preventive and promotive. It is located in Karuri Location which is a location of Kiambu East District in Kiambu County. The health center is headed by a Clinical Officer who is deputized by a Nursing Officer. It has a Comprehensive Care Centre (CCC) which serves HIV positive clients. Currently 3459 patients are enrolled in the clinic with 65% of them already on HAART

3.3 Empirical study

The data that used in this study was obtained from Karuri Health Centre in Kiambu County. Kiambu East District has several centers which provide ART care to PLWHIV where the units in the clinic have nurses, a pharmacy, data clerks, rooms, equipment and etc. The district started offering free ART service in 2005. Data for this study were extracted from the available standard national medical registers, which have been adopted by the Ministry of Health (MOH). The registers include the Pre ART register (register of patients at their first visit), the ART register (registration after ART initiation), and the follow-up patient form. Sampling was carried out using simple random sampling.The resulting sample from the sample frame comprised all cases of HIV-infected patients older than 18 years who have started HIV/AIDS treatment between January 1st 2008 to 31st December 2012 and followed for the outcomes of either the event (death) or censored (dropped out, lost to follow up, transferred out to other centers or on follow up at the end of study time). The end of the follow-up time was 31st December 2012.

3.4 Variables of the study

3.4.1 The Response Variable

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

3.4.2 Predictor Variables

The predictor variables in survival data analysis are called covariates. These are explanatory variables which are assumed to influence the survival of HIV infected patients and are given below.

1. Age in years... ... ... X_1
2. Sex (Male, Female) ... ... X_2
3. Baseline Weight (kg) ... ... X_3
4. Substance abuse (Smoking, Alcohol)(no, yes) ... ... X_4
5. Base line CD4 cell/mm^3 ... ... X_5
6. Regimen type... ... ... X_6
7. WHO Clinical staging on HAART initiation. ... ... X_7
8. BMI... ... ... X_8
9. Tb treatment status... ... ... X_9
10. Drug adherence... ... ... X_{10}

3.5 Study design

It was a retrospectively study where a total of 248 subjects were sampled .The period of study was from 1st January 2008 through 31st December 2012.Gender, age at initiation of HAART, baseline weight (kg) ,substance abuse (smoking, alcohol or any other drug), CD4 cell count (cells/mm^3)taken at the beginning of the study, regimen type (TDF +3TC+EFV or NVP and

AZT+3TC+NVP or EFV),history of drug adherence, whether the patient is on tuberculosis treatment or not and the world health clinical staging at the initiation of HAART Were the predictor variables.Cox Proportional Hazard model a semi-parametric model and Accelerated Failure Time a parametric model (Weibull, Exponential and lognormal form) were performed with the Akaike Information Criterion (AIC) been used to compare the efficiency of models.

3.6 Sampling Method

Sample size determination was addressed in the original study.

The following formula was used for sample size computation:

$$n = \frac{(z_{1-\frac{\alpha}{1}} + z_{1-\beta})^2 2\sigma^2}{(\mu_1 - \mu_2)^2}$$

where α = significant level (0.05)

1-β= the power of the study (90%)

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

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

$\mu_1 - \mu_2$ = the expected difference between the subjects on TDF +3TC+EFV or NVP and AZT+3TC+NVP or EFV.

This gave a total of 288 but only those whose data was complete were analyzed giving a total of 248 subjects.

3.7 Ethical Aspects

All facets of the relevant ethics were adequately addressed by the primary study; hence was not replicated here except for a formal application and subsequent acquisition of the original datasets from the Ministry of Health (MOH), Karuri Health Centre.

3.8 Validity and Reliability

Content validity is based on the adequacy with which the items in an instrument measure the attributes of the study (Nunnally, 1978). Content validity of the method was ensured through constructive criticism from colleagues in my class and my supervisors who have an extensive experience in research. The study will be revised and improved according to advice and suggestions made. Reliability is the extent to which any measuring procedure/method yields the same results on repeated trials (Carmines & Zeller, 1979). The reliability of the method will be ensured through fitting the model to hypothetical datasets. Furthermore, the reliability and validity of the results will be obtained through member checks to help indicate whether the findings appeared to match with perceived authenticity. This will be done in order to limit the distorting effects of random errors on the findings.

3.9 Method

3.9.1 Survival Analysis

3.9.2 Basic Concepts on Survival Analysis

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

There are three types of censoring: right censoring, left censoring, and interval censoring. Right censoring is said to occur if the event occurs after the observed survival time. Let C denote the censoring time, that is, the time beyond which the study subject cannot be observed. The

observed survival time is also referred to as follow up time. It starts at time *0* and continues until the event *X* or a censoring time *C*, whichever comes first.

The observed data are denoted by (T, δ), where T = min (X, C) is the follow-up time, and δ = $I_{(X \leq C)}$ is an indicator for status at the end of follow-up,

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

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

Censoring can also occur if we observe the presence of a condition but do not know where it began. In this case we call it left censoring, and the actual survival time is less than the observed censoring time. If an individual is known to have experienced an event within an interval of time but the actual survival time is not known, we say we have interval censoring. The actual occurrence time of event is known within an interval of time. Right censoring is very common in survival time data, but left censoring is fairly rare. An important assumption for methods presented in survival analysis studies for the analysis of censored survival data is that the individuals who are censored are at the same risk of subsequent failure as those who are still alive and uncensored. i.e. a subject whose survival time is censored at time *C* must be representative of all other individuals who have survived to that time. If this is the case, the censoring process is called non-informative. Statistically, if the censoring process is independent of the survival time. i.e.

$P(X \geq x; C \geq x) = P(X \geq x) \, P(C \geq x)$,

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

3.9.2.1 Survival time distribution

Let *T* be a random variable denoting the survival time. The distribution of survival times is characterized by any of three functions: the survival function, the probability density function or the hazard function. The survival function is defined for both discrete and continuous *T*, and the probability density and hazard functions are easily specified for discrete and continuous *T*.

The survival function is defined as the probability that the survival time is greater or equal to t.

S (t) = P (T ≥ t), t ≥ 0.

T discrete

For a discrete random variable T taking well-ordered values $0 \leq t_1 < t2 < \ldots t_n$, let the probability mass function be given by P (T = t_1) = f (t_i), i = 1; 2…, then the survival function is

$$S(t) = \sum_{j/tj \leq t} f(t_j)$$

$$= \sum f(t_j) I_{(tj \geq t),}$$

Where the indicator function $I_{(}t_j \geq_{t)} := \begin{cases} 0 \text{ if } t_j < t \\ 1 \text{ if } t_j \geq t \end{cases}$

In this case, the hazard function *h (t)* is defined as the conditional probability of failure at time t_j given that the individual has survived up to time t_j,

$$h_j = h(t_j) = P(T = t_j | T \geq t_j) = \frac{f(t_j)}{S(t_j)} = \frac{S(t_j) - S(t_j+1)}{S(t_j)} = 1 - \frac{S(tj+i)}{S(t_j)}$$

hence

1- $h(t_j) = \frac{S(t_j+1)}{S(t_j)}$, and

$$\prod_{j/t_j<t}(1 - h(tj)) \text{ co } = \frac{S(t_2)}{S(t_1)} * \frac{S(t_3)}{S(t_2)} * \ldots * \frac{S(t_j+1)}{S(t_j)} = S(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3.1)$$

Since $S(t_1) = 1$ and $S(t) = S(t_{j+1})$

Moreover,

$f(t_j) = h(t_j) * S(t_j)$

=

$$h(tj)\prod_{i=1}^{j-1}(1-h(t_i))$$

……..(3.2)

3.9.2.2 *T* is absolutely continuous

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

f (t)=F'(t)= -S'(t), t≥0

The hazard function gives the instantaneous failure rate at t given that the individual has survived up to time *t,* i.e.

$h(t) = lim_{\Delta t \downarrow 0} \frac{P(t \le T < +\Delta t \mid T \ge t}{\Delta t}$,*t≥0*

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

h(t) = f(t)/S(t) $= \frac{d}{dt} logs(t)$ …….(3.3)

then s(t)=exp $[-\int_0^t h(x)dx]$= exp (*-H(t)), t≥0,*……..……..…..(3.4)

where *H(t)*= $\int_0^t h(x)du$ is called the cumulative hazard function, which can be obtained from the survival function since H (t) = -log S (t).

The probability density function of T can be written

f(t) = h(t) exp$[-\int_0^t h(x)dx]$, *t≥0*

These three functions give mathematically equivalent specification of the distributions of the survival time *T*. If one of them is known, the other two are determined. One of these functions can be chosen as the basis of statistical analysis according to the particular situations.

The survival function is most useful for comparing the survival progress of two or more groups. The hazard function gives a more useful description of the risk of failure at any time point.

3.9.3 Non-parametric methods

In survival analysis, it is always a good idea to present numerical or graphical summaries of the survival times for the individuals. In general, survival data are conveniently summarized through estimates of the survival function and hazard function. The estimation of the survival distribution provides estimates of descriptive statistics such as the median survival time. These methods are said to be non-parametric methods since they require no assumptions about the distribution of survival time. In order to compare the survival distribution of two or more groups, we will use log-rank tests.

3.9.3.1 The Kaplan-Meier estimate of the survival function

The life table is the earliest statistical method to study human mortality meticulously, but its standing has been reduced by the modern methods, like the Kaplan-Meier (K-M) method. In clinical studies, individual data is usually available on time to death or time to last seen alive. The K-M estimator for the survival curves is usually used to analyze individual data, whereas the life table method applies to grouped data. Since the life table method is a grouped data statistic, it is not as precise as the K-M estimate, which uses the individual values. Suppose that r individuals have failures in a group of individuals. Let $0 \geq t(1) < \ldots < t(r) < \infty$ be the observed ordered death times. Let r_j be the size of the risk set at $t_{(j)}$, where risk set denotes the collection of individuals alive and uncensored just before $t_{(j)}$. Let d_j be the number of observed deaths at $t_{(j)}$, $j = 1; \ldots, r$. Then the K-M estimator of $S(t)$ is defined by

$$\hat{S}(t) = \prod_{j:t(j)<t} \left(1 - \frac{d_j}{r_j}\right)$$

This estimator is a step function that changes values only at the time of each death. Suppose that the distribution is discrete, with atoms h_j at finitely many specified points $0 \leq T_1 < T_2 < \ldots < T_j$: the survival function $S(t)$ may be expressed in terms of the discrete hazard function h_j as

$$\hat{S}(t) = \prod_{j|\tau_j<t} (1 - h_j)$$

To derive the full likelihood from a sample of *n* observations, we will first collect all the terms corresponding to the atom τ_j . Let $b_i = j$ if the *i*th individual dies at τ_j: Using (3.2), the contribution to the total log likelihood is

$log h b_i +$

$$\sum_{k<b_i} log(1 - h_k)$$

Let e_i= j if the *i*th individual is censored at τ_j ; using the equation (3.1), the log likelihood contribution to the total likelihood is

$$\sum_{K \le ei} \log(1 - h_k)$$

Then the total log likelihood is given by

$l = \sum_{death\ i} \log h_{bi} + \sum_{death\ i}[\sum_{k<bi} \log(1 - h_k)] + \sum_{censor\ i}[\sum_{k \le ei} \log(1 - hk)]$

$= \sum_j d_j\ log\ h_j + \sum_k[\sum_{j>k} d_j] \log (1\text{-}h_k) + \sum_k[\sum_{j \ge k} c_j] \log (1\text{-}h_k)$

$= \sum_j [d_j \log h_j + (r_j - d_j) \log (1 - h_j)]$,

Where d_j is the number of observed death at T_j , c_j is the number censored at $[\tau_j, \tau_j+1)$,and r_j is the number of living and uncensored at τ_j If h_j is the solution of

$$\frac{\partial l}{\partial hj} = \frac{d_j}{h_j} - \frac{r_j - d_j}{1 - h_j} = 0$$

then,

$$\hat{h_j} = d_j/r_j.$$

This maximizes the likelihood since the total log likelihood function is concave down. So that the K-M estimator of the survival function is

$$\hat{S}(t) = \prod_{j|Tj<t}(1 - \hat{h}_j)$$

=

$$\prod_{j|Tj<t}(1 - \frac{d_j}{r_j})$$

The K-M estimator gives a discrete distribution. If the observations are modeled to come from unknown continuous distribution, the maximum likelihood estimator does not exist.

3.9.3.2 Greenwood's formula

Confidence interval for the survival probability is calculated by Greenwood's formula; first, we will need the variances of the $\hat{h_j}s$. Let the number of individual at risk at $t_{(j)}$ be r_j and the number of deaths at $t_{(j)}$ be d_j. Given r_j, the number of individuals surviving through the interval $[t_{(j)}, t_{(j+1)})$, $r_j \_ d_j$, can be assumed to have binomial distribution with parameters r_j and $1\text{-}h_j$: The conditional variance of $r_j - d_j$ is given by

$$V(r_j - d_j / r_j) = r_j h_j(1 - h_j).$$

The variance of $\hat{h}_j$ is

$$V(\hat{h_j}/r_j) = V(1 - \hat{h_j}) = V(1 - \frac{d_j}{r_j}) = \frac{h_j(1-h_j)}{r_j}.$$

Since $\hat{h_j}$ is a conditional independent of $\hat{h_1}$,…, $\hat{h}_{j-1}$, given $r_1, \ldots r_j\text{-}1$, the delta method can be used to obtain.

$V\ (\text{In}\ \hat{S}(t)\ |_{r_j}: t_{(j)}<t) = V\ [\sum_{j:t(j)<t}(ln(1-\hat{h}_j))\ |\ r_j]$

$= \sum_{j:t(j)<t} V[\text{ln}\ (1-\hat{h}_j)\ |_{r_j}]$

$\approx \sum_{j:t(j)<t}(\frac{d}{dx} ln\ (1-x\ ))^2\ _{x\ =\ \hat{h}_j} V\ (\hat{h_j}/r_j$

$= \sum_{j:t(j)<t}\{-\frac{1}{1-\hat{h}_j}\}^2 \frac{h_j(1-h_j)}{r_j}$, $j=1, \ldots, r$

we estimate this by simply replacing h_j with $\hat{h}_j = d_j/r_j$, which gives

$V\square(ln\hat{S}(t)) = \sum_{j:t(t)<t} \frac{d_j}{r_j(r_j-d_j)}$, $j=1,\ldots, r$

Let $Y = ln\ \hat{S}\ (t)$, again using the delta method, we get,

$V\square\ (\hat{S}\ (t)) \approx [\hat{S}\ (t)]^2 \sum_{j:t(j)<t} \frac{d_j}{r_j-d_j}$. …….. (3.5)

This is known as Greenwood's formula. The K-M estimator and functions of it have been proved to be asymptotically normal distributed (Andersen, P. K., 1993).Thus the confidence intervals will be constructed by the normal approximation based on *S (t).*

3.9.3.3 Estimating the median and percentile of survival time

Since the distribution of survival time tends to be positively skewed, the median is preferred for a summary measure. The median survival time is the time, beyond which 50% of the individuals under study are expected to survive, i.e., the value of *t* (50) at $\hat{S}$ (*t* (50)) = 0:5. The estimated median survival time is given by

$t'(50)= \min\ \{t_i\ /\hat{S}(t_i)<0.5\}$,

Where t_i is the observed survival time for the *i*th individual, *i* = 1; 2; …; *n*. In general, the estimate of the *p*th percentile is

$t\square\ (p)$ = min $\{t_i | \hat{S}(t_i) < 1 - \frac{p}{100}\}$.

The variance of the estimator of the *p*th percentile is

$V[\hat{S}\ (t\ (p))] = (\frac{d\hat{S}(t(p))}{dt(p)})^2\ V\{t(p)\}$

$= (-f\square\ (t\ (p)))^2\ V\{t\ (p)\}.$

The standard error of *t' (p)* is therefore will be given by

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

The standard error of $\hat{S}\ (t'\ (p))$will be obtained using Greenwood's formula, given in equation (3.5) An estimate of the probability density function at the *p*th percentile t' (p) is used by many software packages

$f\square[t\square(p)] = \frac{\hat{S}[\hat{u}(p)] - \hat{S}[\hat{l}(p)]}{\hat{l}(p) - \hat{u}(p)}$,

where,

$\hat{u}(p)$ = max $\{t_{(j)} | \hat{S}(t_{(j)}) \geq 1 - \frac{p}{100} + \varepsilon\}$,

$l\square(p)$ = min $\{t_{(j)} | \hat{S}(t_{(j)}) \leq 1 - \frac{p}{100} - \varepsilon\}$,

$t_{(j)}$ is *j*th ordered death time, $j = 1, 2, \ldots r$. $\varepsilon = 0.05$ is typically used by a number of statistical packages. Therefore, for median survival time, û (50) is the largest observed survival time from the K-M curve for which $\hat{S}\ (t) \geq 0.55$, and *l'* (50) is the smallest observed survival time from the K-M curve for which $\hat{S}\ (t) \leq 0.45$

The 95% confidence interval for the *p*th percentile t' (p) has limits of

t'(p) ± 1.96SE{t'(p)}

3.9.3.4 Non-parametric comparison of survival distributions

The K-M survival curves can give us an insight about the difference of survival functions in two or more groups, but whether this observed difference is statistically significant requires a formal

statistical test. There are a number of methods that can be used to test equality of the survival functions in different groups. One commonly used non-parametric tests for comparison of two or more survival distributions is the log-rank test which we used in this study.

For two groups, let $t_{(1)}<t_{(2)}<\ldots<t_{(k)}$ be the ordered death times across the two groups. Suppose that d_j failures occur at $t_{(j)}$ and that r_j subjects are at risk just prior to $t_{(j)}$ ($j = 1, 2,\ldots,k$). Let d_{ij} and r_{ij} be the corresponding numbers in group i ($i = 1, 2$).The log-rank test will compare the observed number of deaths with the expected number of deaths for group i. Consider the null hypothesis $S_1(t) = S_2(t)$, i.e. there is no difference between survival curves in two groups. Given r_j and d_j, the random variable d_{1j} has the hyper geometric distribution

$$\frac{\binom{d_j}{d_{1j}}\binom{r_j-d_j}{r_{1j}-d_{1j}}}{\binom{r_j}{r_{1j}}}.$$

Under the null hypothesis, the probability of death at $t_{(j)}$ does not depend on the group, i.e. the probability of death at $t_{(j)}$ is $\frac{dj}{rj}$. So that the expected number of deaths in group one is

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

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

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}) \quad \ldots\ldots(3.6)$$

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

$$V'_{1j} = V(d_{1j}) = \frac{r1j\ r2j\ dj\ \big|\ (rj-dj)}{r2/j(rj-1)}, \quad \ldots\ldots(3.7)$$

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

Under the null hypothesis, statistic (3.6) has an approximate normal distribution with zero mean and variance V_L. This then follows

$$\frac{U_L^2}{V_L} \sim \varkappa_1^2$$

There are several alternatives to the log-rank test to test the equality of survival curves, for example, the Wilcoxon test .These tests may be defined in general as follows:

$$\frac{\sum_j^r w_j(d_{1j} - e_{1j})}{\sum_j^r w_j^2 V_{1j}}$$

where w_j are weights whose values depend on the specific test.

The Wilcoxon test uses weights equal to risk size at $t_{(j)}$, $w_j = r_j$. This gives less weight to longest survival times. Early failures receive more weight than later failures. The Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large. This type of weighting may be used to assess whether the effect of treatment on survival is strongest in the earlier phases of administration and tends to be less effective over time. Whereas the log-rank test uses weights equal to one at $t_{(j)}$, $w_j = 1$. This gives the same weight to each survival time. Therefore, Wilcoxon statistic is less sensitive than the log-rank statistic to difference of d_{1j} from e_{1j} in the tail of the distribution of survival times.

The log-rank test is appropriate when hazard functions for two groups are proportional over time, i.e., h_1 (t) = $Øh_2$(t) So it is the most likely to detect a difference between groups when the risk of a failure is consistently greater for one group than another and it was the one that was used in this study.

3.9.4 Cox Regression Model

The Cox Proportional Hazards model is given by:

$h\ (t \mid X) = h_0\ (t) \exp\ (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) = h_0(t) \exp\ (\mathbf{\beta' x})$,

where $h_0(t)$ is called the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero., $X = (x_1, x_2, \ldots, x_p)'$ is the values of the vector of explanatory variables for a particular individual, and $\boldsymbol{\beta'} = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients.

The corresponding survival functions are related as follows:

$$S(t/X) = s_0(t)^{exp^{(\sum_{i=1}^{p}\beta_i x_i)}}$$

This model, also known as the Cox regression model, makes no assumptions about the form of $h_0(t)$ (non-parametric part of model) but assumes parametric form for the effect of the predictors on the hazard (parametric part of model). The model is therefore referred to as a semi-parametric model. The beauty of the Cox approach is that this vagueness creates no problems for estimation. Even though the baseline hazard is not specified, we can still get a good estimate for regression coefficients β hazard ratio, and adjusted hazard curves. The measure of effect is called hazard ratio. The hazard ratio of two individuals with different covariates x and x^* will be given by:

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

This hazard ratio is time-independent, which is why this is called the proportional hazards model.

3.9.4.1 Partial likelihood estimate for Cox proportional hazards model

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

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

$$= \frac{P(\text{individual } i \text{ dies at } t_{(j)})}{P(\text{one death at } t_{(j)})}$$

$$= \frac{P(\text{individual } i \text{ dies at } t_{(j)})}{\sum_{k\epsilon R(t_{(j)})} P(\text{individual } k \text{ dies at } t_{(j)})}$$

$$\approx' \frac{P\{\text{individual } i \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t}{\sum_{k\epsilon R(t_{(j)}} P\{\text{individual } i \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t}$$

$$= \frac{\lim_{\Delta t \downarrow 0} P\{\text{individual } i \text{ } dies\ at\ (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t}{\lim_{\Delta t \downarrow 0} \sum_{k\epsilon R(t_{(j)}} P\{\text{individual } k \text{ dies at } t_{(j)}, t_{(j)} + \Delta t)/\Delta t}$$

$$= \frac{h_i(t_{(j)}}{\sum_{k\epsilon R(t_{(j)}} h_k(t_{(j)}}$$

$$= \frac{h_0(t_{(j)}) \exp\,(\beta'^{x}{}_i(t_{(j)})}{\sum_{k\epsilon R(t_{(j)}} h_0(t_{(j)} \exp\,(\beta'^{x}{}_k(t_{(j)}}$$

$$= \frac{\exp\,(\beta'^{x}{}_i(t_{(j)}}{\sum_{k\epsilon R(t_{(j)}} \beta'^{x}{}_k(t_{(j)}}\,.$$

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

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\,(\beta'^{x}{}_i(t_{(j)}}{\sum_{k\epsilon R(t_{(j)}} \beta'^{x}{}_k(t_{(j)}}\,, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.8)$$

in which $x_i(t_{(j)})$ is the vector of covariate values for individual i who dies at $t_{(j)}$: This likelihood function is only for the uncensored individuals. Let t_1, t_2 ,…, t_n be the observed survival time for n individuals and δ_i be the event indicator, which is zero if the ith survival time is censored, and unity otherwise. The likelihood function in equation can be expressed as

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\,(\beta'^{x}{}_i(t_{(j)}}{\sum_{k\epsilon R(t_{(i)}} \beta'^{x}{}_k(t_{(i)}}\right]^{\delta i}, \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

where $R(t_i)$ is the risk set at time t_i

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

3.9.4.2 Proportional hazard assumption checking

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

3.9.4.2.1 Graphical method

COX PH survival function is got by the relationship between hazard function and survival function

$S(t, x_1) = S_0(t)^{\exp(\sum_i^p \beta_i x_i)}$

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

$\ln[-\ln S(t,x)] = \sum_{i=1}^p \beta_i x_i + \ln[-\ln S_0(t)]$.

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

$$ln[-lnS(t,x)] - ln[S(t,x_2)] = \sum_{i=1}^{p} \beta_i (x_{1i} - x_{2i})$$

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

3.9.4.2.2 Adding time-dependent covariates in the Cox model

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

$$h(t, x(t)) = h_0(t) \exp[\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \ldots + \beta_p x_p + \delta x_j x\, g(t)].$$

Where $x(t) = (x_1; x_2, \ldots, x_p, x_j(t))'$ is the values of the vector of explanatory variables for a particular individual. The null hypothesis to check proportionality is that $\delta = 0$. 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 $W = (\delta\square/se(\delta\square))^2$. The likelihood ratio test will calculates the likelihood under null hypothesis, L_0 and the likelihood under the alternative hypothesis, L_a. The LR statistic is then $LR = -2 \ln (L_0/L_a) = -2 (l_0 - l_a)$, where l_0, l_a are log likelihood under two hypothesis respectively. Both statistics have a chi-square distribution with one degree of freedom under the null hypothesis. If the time-dependent covariate is significant, i.e., the null hypothesis is rejected, then the predictor is not proportional. In the same way, we assessed the PH assumption for several predictors simultaneously.

3.9.4.2.3 Tests based on the Schoenfeld residuals

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

3.9.4.2.4 Cox proportional hazards model diagnostics

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

3.9.4.2.5 Cox-Snell residuals and deviance residuals

The Cox-Snell residual is given by Cox and Snell (1968) The Cox-Snell residual for the *i*th individual with observed survival time t_i is defined as

$r_{ci} = \exp(\beta'x_i)\, H_0(t_i) = -\log S_i(t_i)$,

Where $H_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, which was derived by Kalbfleisch and Prentice (2002).This residual is motivated by the following result: Let *T* have continuous survival distribution *S(t)* with the cumulative hazard *H(t)* = -log(*S(t)*).

Thus, $S_T(t) = \exp(-H(t))$. Let $Y = H(T)$ be the transformation of *T* based on the cumulative hazard function. Then the survival function for Y is

$S_Y(y) = P(Y>y) = P(H(t)>y)$

$= P(T>H_T^{-1}(y)) = S_T(H_T^{-1}(y))$

$= \exp(-H_T(H_T^{-1}(y)) = \exp(-y)$.

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

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

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

and

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

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

$rD_i = sign(r_{mi})[-2\{r_{mi} + \delta_i \log(\delta_i - r_{mi}\}]^{1/2}$

Where the function *sign* (.) is the sign function which takes the value 1 if r_{mi} is positive and -1 if r_{mi} is negative; $r_{mi} = \delta_i - r_{ci}$ is the martingale residuals for the *i*th individual; and $\delta_i = 1$ for

uncensored observation, $\delta_i = 0$ for censored observation. The martingale residuals take values between negative infinity and unity. They have a skewed distribution with mean zero. The deviance residuals are a normalized transform of the martingale residuals . They also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Deviance residual can also be used like residuals from linear regression. The plot of the deviance residuals against the covariates can be obtained. Any unusual patterns will suggest features of the data that have not been adequately fitted for the model. Very large or very small values will suggest that the observation may be an outlier in need of special attention. In a fitted Cox PH model, the hazard of death for the *i*th individual at any time depends on the value of exp (β'0xi) which is called the risk score. A plot of the deviance residuals versus the risk score will be a helpful diagnostic to assess a given individual on the model. Potential outliers will have deviance residuals whose absolute values are very large. This plot will give the information about the characteristic of observations that are not well fitted by the model.

3.9.4.2.6 Schoenfeld residuals

All the above three residuals are residuals for each individual. I will describe covariate-wise residuals: Schoenfeld residuals The Schoenfeld residuals were originally called partial residuals because the Schoenfeld residuals for *i*th individual on the *j*th explanatory variable X_j is an estimate of the *i*th component of the first derivative of the logarithm of the partial likelihood function with respect to β_j : From equation (3.2), this logarithm of 20 the partial likelihood function is given by

$$\frac{\partial log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{p} \delta_i \{x_{ij} - a_{ij}\}$$

Where x_{ij} is the value of the *j*th explanatory variable $j = 1, 2, \ldots, p$ for the *i*th individual and

$$a_{ji} = \frac{\sum_{l \epsilon R(t_i)} x_{ji} \exp(\beta' x_l)}{\sum_{l \epsilon R(t_i)} \exp(\beta' x_i}$$

The Schoenfeld residual for *i*th individual on X_j is given by $r_{pji} = \delta_i\{x_{ji} - a_{ji}\}$. The Schoenfeld residuals sum to zero.

3.9.4.2.7 Strategies for analysis of non-proportional data

If the statistic tests or other diagnostic techniques will give strong evidence of non-proportionality for one or more covariates.To deal with this we use two popular methods: stratified Cox model and Cox regression model with time-dependent variables which are particularly simple and can be done using available software in my analysis.

3.9.4.2.8 Stratified Cox model

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

$h_{ig}(t)=h_{og}(t)exp(\beta' x_{ig})$ Where g represents the stratum.

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

3.9.4.2.9 Cox regression model with time-dependent variables

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

$$h(t \mid x(t)) = h_0(t)\exp\left[\sum_{i=1}^{P_1} \beta_1 x_1 + \sum_{j=1}^{P_2} \alpha_j x_j(t)\right],$$

The hazard ratio at time t for the two individuals with different covariates *x* and x* will be given by

H□R□' (*t*) = exp $[\sum_{i=1}^{P_1} \hat{\beta}_i(x_i^* - x_i) + \sum_{j=1}^{P_2} \alpha_j x_j^*(t)] - x_j(t))]$.

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

In addition to considering time-dependent variable for analyzing a time-independent variable not satisfying the PH assumption, there are variables that are inherently defined as time-dependent variables. One of the earliest applications of the use of time-dependent covariates is in the report by Crowley and Hu (1977), on the Stanford Heart Transplant study. Time-dependent variables are usually classified to be internal or external. An internal time-dependent variable is one that the change of covariate over time is related to the characteristics or behavior of the individual. For example, blood pressure, disease complications, etc. The external time-dependent variable is one whose value at a particular time does not require subjects to be under direct observations, i.e., values changes because of external characteristics to the individuals. For example, level of air pollution.

3.9.4.3 Accelerated Failure Time Model

Although parametric models are very applicable to analyze survival data, there are relatively few probability distributions for the survival time that can be used with these models. In these situations, the accelerated failure time model (AFT) is an alternative to the PH model for the analysis of survival time data. Under AFT models we measured the direct effect of the explanatory variables on the survival time instead of hazard. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean survival time. Currently, the AFT model is not commonly used for the analysis of clinical trial data, although it is fairly common in the field of manufacturing. Similar to the PH model, the AFT model describes the relationship between survival probabilities and a set of covariates.

For a group of patients with covariate $(X_1,X_2,\ldots,X_p)$, the model is written mathematically as $S(t|x) = S_0(t|n(x))$, where $S_0(t)$ is the baseline survival function and η is an 'acceleration factor that is a ratio of survival times corresponding to any fixed value of S(t). The acceleration factor is given according to the formula $\eta(x) = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p)$

Under an accelerated failure time model, the covariate effects will be assumed to be constant and multiplicative on the time scale, that is, the covariate impacts on survival by a constant factor (acceleration factor).

According to the relationship of survival function and hazard function, the hazard function for an individual with covariate $X_1;X_2,\ldots,X_p$ is given by

$h(t|x) = [1/\eta(x)]h_0[t/\eta(x)]$…….. (3.10)

The corresponding log-linear form of the AFT model with respect to time is given by

$log\ T_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_p X_{pi} + \sigma\epsilon_i$

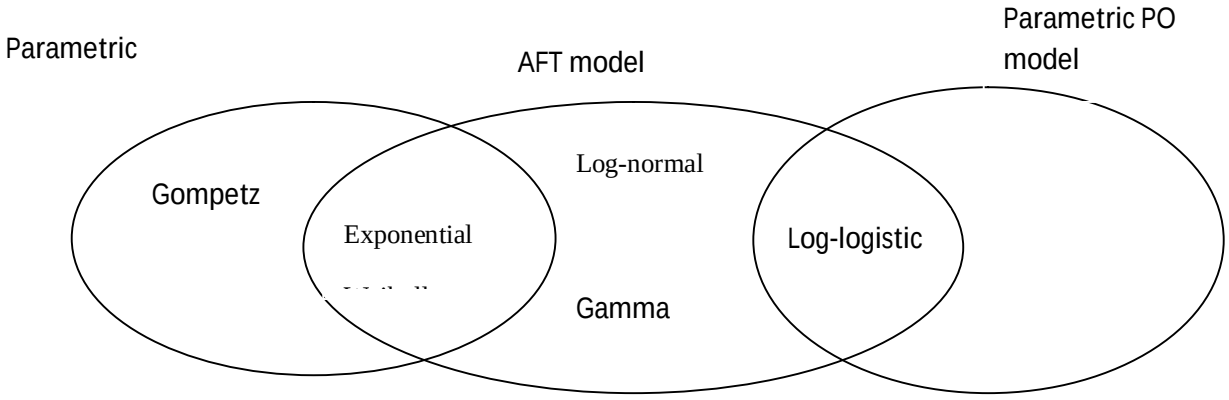
where μ is intercept, σ is scale parameter and ϵ_i is a random variable, assumed to have a particular distribution.

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

Table 3.1 Summary of parametric AFT models

Distribution of ϵ	Distribution of *T*
Extreme value 1	Exponential
Extreme 2	Weibull
Logistic	Log-logistic
Normal	Log- normal

Figure 3.1 Summary of parametric models



The survival function of T_i can be expressed by the survival function of ϵ_i

$S_i(t) = P\ (T_i \geq t)$

$= P\ (\log T_i \geq \log t)$

$= P\ (\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\epsilon_i \geq \log t)$ ……(3.11)

$= P(\varepsilon_i \geq \frac{\log t - \mu - \alpha x}{\sigma})$

$= S\varepsilon_i(\frac{\log t - \mu - \alpha x}{\sigma}$

The distributions of ε_i and the corresponding distributions of T_i are summarized in Table 3.1, and the summary of the commonly used parametric models are described in Figure 3.1.

The effect size for the AFT model is the time ratio. The time ratio comparing two levels of covariate x_i ($x_i = 1$ *vs.* $x_i = 0$), after controlling all the other covariates is $\exp(\alpha_i)$, which is interpreted as the estimated ratio of the expected survival times for two groups. A time ratio above 1 for the covariate will implies that this covariate prolongs the time to event, while a time ratio below 1 indicates that an earlier event is more likely. Therefore, the AFT models will be interpreted in terms of the speed of progression of a disease. The effect of the covariates in an

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

3.9.4.3.1 Estimation of AFT model

AFT models are fitted using the maximum likelihood method. The likelihood of the *n* observed survival times, $t_1, t_2, \ldots, t_n$ will be given by

$$L(\alpha,\mu,\sigma) = \prod_{i=1}^{n}\{f_i(t_i)\}\,\delta_i\{S_i(t_i)\}^{1-\delta_i}$$

where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions for the *i*th individual at t_i and δ_i is the event indicator for the *i*th observation. Using equation (3.11), the log-likelihood function will then given by

$$log\ L\ (\alpha,\mu,\sigma) = \sum_{i=1}^{n}\{-\delta_i log\ (\sigma t_i + \delta_i log f_{\varepsilon_i}(Z_i) + (1-\delta_i)\log S\varepsilon_i(Z_i))\},$$

$$where Z_i = log\ t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi}) \ / \sigma.$$

3.9.4.3.2 Weibull AFT model

Suppose the survival time *T* has $W(\lambda,\gamma)$ distribution with scale parameter λ and shape parameter γ From equation (3.10), under AFT model, the hazard function for the *i*th individual is

$$H_i(t) = [1/\eta_i(X)]h_0[t/\eta_i(X)]$$

$$=[1/\eta_i(X)]\lambda\gamma(t/\eta_i(X))^{\gamma-1}$$

$$=1/[\eta_i(X)]^{\gamma}\lambda\gamma(t)^{\gamma-1}$$

Where $\eta_i = \exp(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi})$ for individual *i* with *p* explanatory variables.

So the survival time for the *i*th patient is $W(1/[\eta_i(X)]^{\gamma}\lambda,\gamma)$. The Weibull distribution has the AFT property.

If T_i has a Weibull distribution, then ε_i has an extreme value distribution (Gumbel distribution). The survival function of Gumbel distribution is given by

$$S\varepsilon_i(\varepsilon) = \exp(-\exp(\varepsilon)).$$

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

$S_i(t)$ = exp [- exp ($\frac{\log t - \mu - \alpha_1 x_{1i} - \cdots - \alpha_p x_{pi})}{\sigma}$)]

= exp [- exp ($\frac{-\mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{p1})}{\sigma}$)$t^{1/\sigma}$]……..……(3.12)

The PH representation of the survival function of the Weibull model is given by

$S_i(t)$ = exp { - exp ($\beta_1 x_{1i} + \ldots + \beta_p x_{pi})\lambda t^{\gamma}$}……(3.13)

Comparing the above two formulas (3.12) and (3.13), we can easily see that the parameter λ,γ,β_j in the PH model will be expressed by the parameters μ,σ,α_j in the AFT model:

$\lambda = exp\ (-\mu/\sigma), \gamma = 1/\sigma, \beta_j = -\alpha_j/\sigma$……(3.14)

Using equation (3.3), the AFT representation of hazard function of the Weibull model will be given by

$h_i(t) = \frac{1}{\sigma} t \frac{1}{\sigma} - 1$ exp ($\frac{-\mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi})}{\sigma}$……(3.15)

Suppose the *p*th percentile of the survival distribution for the *i*th individual is $t_i(p)$, which is the value such that $S_i(t_i(p)) = \frac{100-p}{100}$. From equation (4.4), we can easily get

$t_i(p)$ = exp [σ log { - log ($\frac{100-p}{100}$)} + μ + α'x_i]

The median survival time is

$t_i(50)$ = exp [σlog(log2) + μ + α'x_i]……(3.16)

To calculate the standard error of $\beta_j$we can use the appropriate covariance of a function of two parameter estimate $\theta_1\theta_2$ which is given by

($\frac{\partial g}{\partial \theta_1}$)$^2 V(\theta_1)$ + ($\frac{\partial g}{\partial \theta_2}$)$V$ (□□$_2$) + 2 ($\frac{\partial g}{\partial \theta_1}\frac{\partial g}{\partial \theta_2}$) Cov ($\theta_1\theta_2$).

The approximate variance of β_j is expressed as

$V(\beta_1)= (\frac{-1}{\hat{\sigma}})^2 V(\alpha\square_j) + (\frac{\hat{\alpha}j}{\hat{\sigma}^2})^2 V(\sigma\square) + 2(\frac{-1}{\hat{\sigma}})(\frac{\hat{\sigma}_j}{\hat{\sigma}^2})\text{Cov}(\acute{\alpha}_j\hat{\sigma}$

The square root of this is the standard error of $\beta\square_j$ Then the 95% confidence interval can be calculated.

3.9.4.3.3 Log-normal AFT model

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

$S_0(t) = 1 - \Phi(\frac{logt-\mu}{\sigma})$, $h_0(t) = \frac{\phi(\frac{logt}{\sigma})}{1-\Phi(\frac{logt}{\sigma})]\sigma t}$,

Where μ and σ are parameters, $\boldsymbol{\phi}(x)$ is the probability density function and $\Phi(x)$ is the cumulative density function of the standard normal distribution. The survival function for the *i*th individual is

$S_i(t) = S_0(t/\eta_i)$

$=1-\Phi(\frac{logt-\alpha' x_i-\mu}{\sigma})$,

Where $\eta_i = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p)$: Therefore the log survival time for the *i*th individual has normal $(\mu + \alpha' x_i\ \sigma)$. The log-normal distribution has the AFT property. In a two group study, we can easily get

$\Phi^{-1}[1-S(t)] = 1/\sigma\ (\log t - \alpha' x_i - \mu)$,

where x_i is the value of a categorical variable which takes the value one in one group and zero in the other group. This implies that a plot of $\Phi^{-1}[1-S(t)]$ versus log *t* will be linear if the log-normal distribution is appropriate.

CHAPTER FOUR

4.0 RESULTS

4.1 Introduction

In this chapter, the results of the study are described and the analysis of the data presented. Analysis was done using Stata statistical soft wares. The results describe information on the subjects under study and how different predictors influence the outcome of interest, which is death by performing chi-square and regression analysis.

Table 4.1.1 summary of the survival data

```
. stset months, failure(status==1)

     failure event:  status == 1
obs. time interval:  (0, months]
 exit on or before:  failure

------------------------------------------------------------------------------
      248  total obs.
        0  exclusions
------------------------------------------------------------------------------
      248  obs. remaining, representing
       23  failures in single record/single failure data
     6715  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        57
```

The tables indicate 248 total observations. There were no exclusions, 23 failures and total of 6,715 total analysis times at risk.

Table 4.1.2 Gender of the Subjects

```
. stsum, by(sex)

        failure _d:  status == 1
  analysis time _t:  months
```

sex	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
1	2092	.0043021	84	.	.	.
2	4623	.0030283	164	.	.	.
total	6715	.0034252	248	.	.	.

There were a total of 84 males and 164 females. The total number of hours male were at risk is 2092 and the total number of hours female were at risk is 4623. The incident rate for male was 0.0043021 and for female was 0.0030283.

Table 4.1.3 Adherence to Prescribed Medication

```
. stsum, by(drugadh)

        failure _d:  status == 1
  analysis time _t:  months
```

drugadh	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
0	5642	.0010635	206	.	.	.
1	1073	.0158434	42	23	43	46
total	6715	.0034252	248	.	.	.

206 subjects took their medication as prescribed while 42 had poor adherence to prescribed drugs. The incident rate for those who had good adherence was 0.0010635and for those who had poor adherence was 0.158434. Median survival time for those with poor adherence was 43 hours.

Table 4.1.4 Patient who were on Tuberculosis Treatment

```
. stsum, by(tbtrtmnt)

        failure _d:  status == 1
  analysis time _t:  months
```

tbtrtmnt	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
0	5838	.0023981	214	.	.	.
1	877	.0102623	34	34	45	.
total	6715	.0034252	248	.	.	.

A total of 34 subjects were on tuberculosis treatment. Their median survival time was 45 months. The incident rate for those who had tuberculosis was 0.102623. Their total time at risk was 877months. A total of 214 subjects were not on tuberculosis treatment. Their incident rate was 0.0023981. Their time at risk was 5838 months.

Table 4.1.5 History of Drug Abuse

```
. stsum, by(drudabs)

        failure _d:  status == 1
  analysis time _t:  months
```

drudabs	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
0	5107	.0013707	190	.	.	.
1	1589	.0094399	57	39	46	.
total	6696	.0032855	247	.	.	.

A total of 57 clients were candidates of substance abuse. Their median survival time was 46 months and their incident rate was 0.0094399. The total number of months they were at risk was 1589. A total of 190 subjects were not candidates of drug abuse. Their incident rate was 0.0013707 and the total number of months they were at risk was 5107.

Table 4.1.6 World Health Organization Clinical Staging

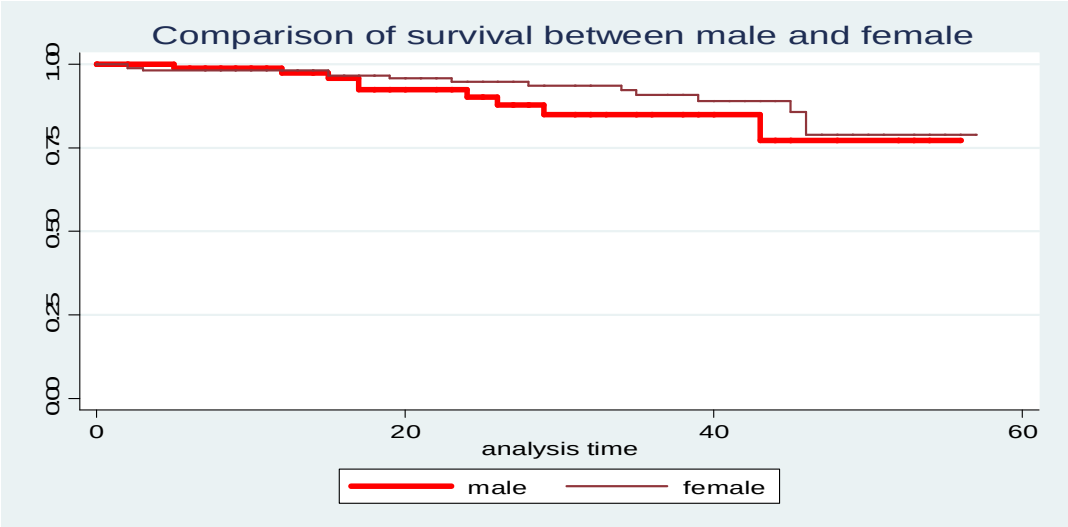
```
. stsum, by(whostgng)

        failure _d:  status == 1
  analysis time _t:  months
```

whostgng	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
1	2042	.0004897	68	.	.	.
2	1149	.0017406	42	.	.	.
3	2338	.0038494	91	46	.	.
4	1186	.0092749	47	34	.	.
total	6715	.0034252	248	.	.	.

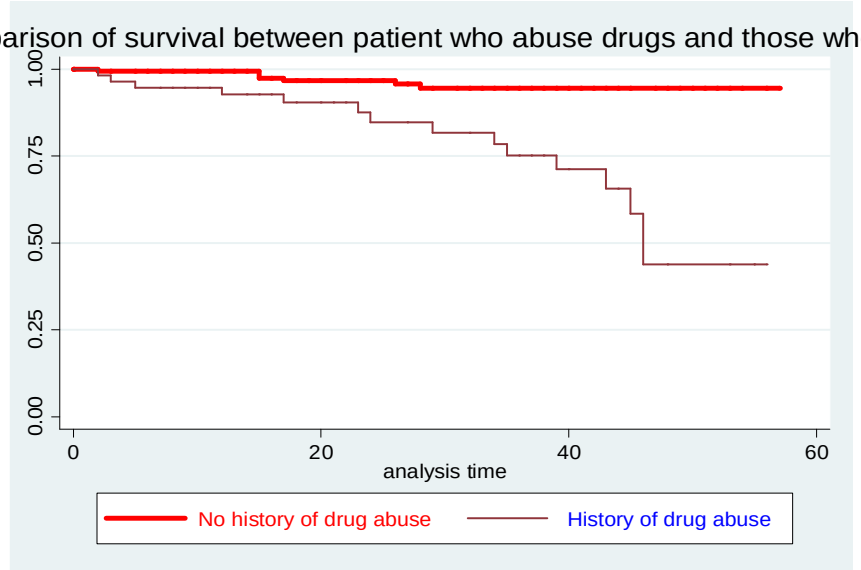
A total of 68 subjects were in WHO clinical stage 1. Their incidence rate was 0.0004897 and the total number of months they were at risk was 2042. A total of 42 subjects were in WHO clinical stage 2 and their incidence rate was 0.0017406 and the total number of hours they were at risk was 1149. A total of 91 subjects were in WHO clinical stage 3. Their incidence rate was 0.0038494 and the total number of hours they were at risk was 2338 months. A total of 47 subjects were in WHO clinical stage 4. Their incidence rate was 0.0092749. The total number of months they were at risk was 1186.

Figure 4.1.1 Kaplan Meier graph Male and Female.



The Kaplan Meier graph above indicates that gender is not significant in predicting survival.

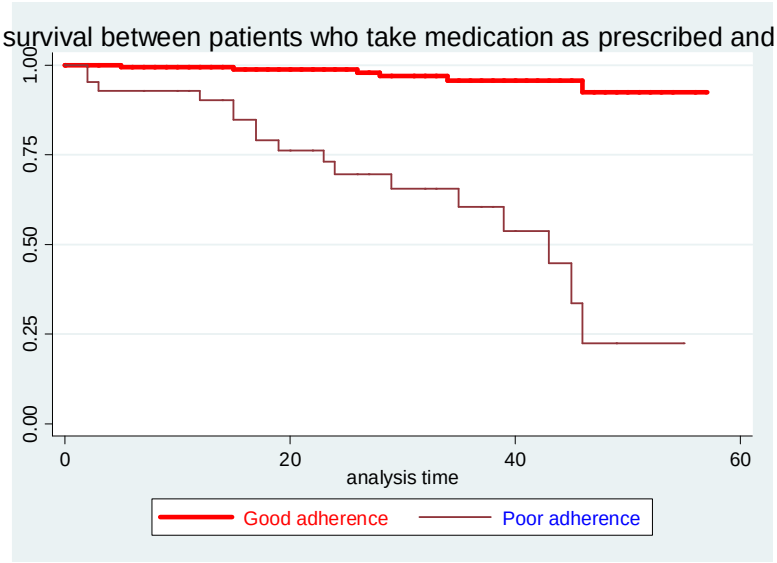
Figure 4.1.2 Kaplan Meier Graph for Drug Abuse



From the Kaplan Meier graph above, it is clear that there is a significance difference between those who abuse drugs and those who do not abuse drugs. Those who abuse drugs die more than those who do not.

Figure 4.1.3 Kaplan Meier Graph for Drug Adherence

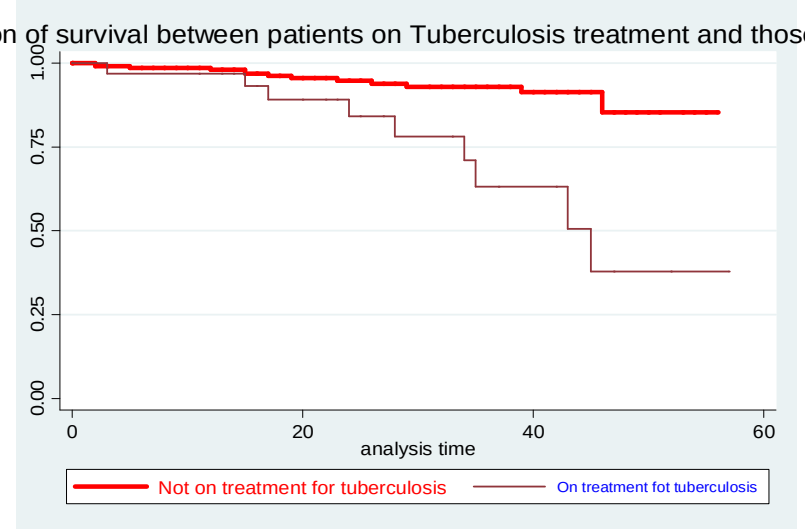
Comparison of survival between patients who take medication as prescribed and those who do not.



From the Kaplan Meier graph above, survival is dependent on drug adherence with those who have poor adherence dying more than those who do not.

Figure 4.1.4 Kaplan Meier Graph for Treatment for Tuberculosis

Comparison of survival between patients on Tuberculosis treatment and those who are not.



From the Kaplan Meier graph above, subjects on tuberculosis treatment have low survival than those who are free from tuberculosis.

Figure 4.1.5 Kaplan Meier Survival Estimates for Cd4 Cell Count

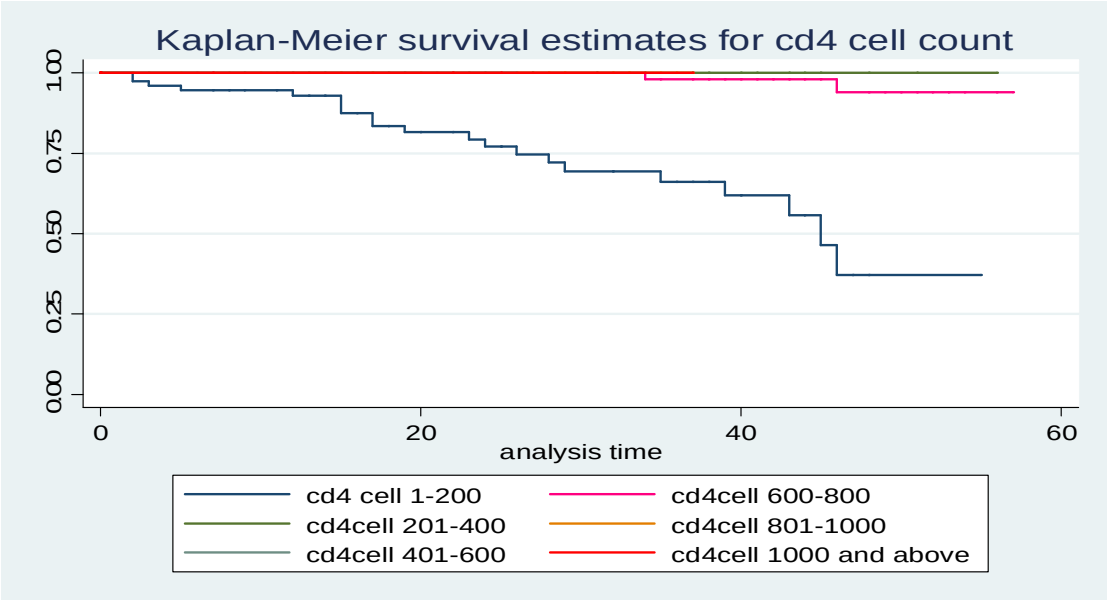
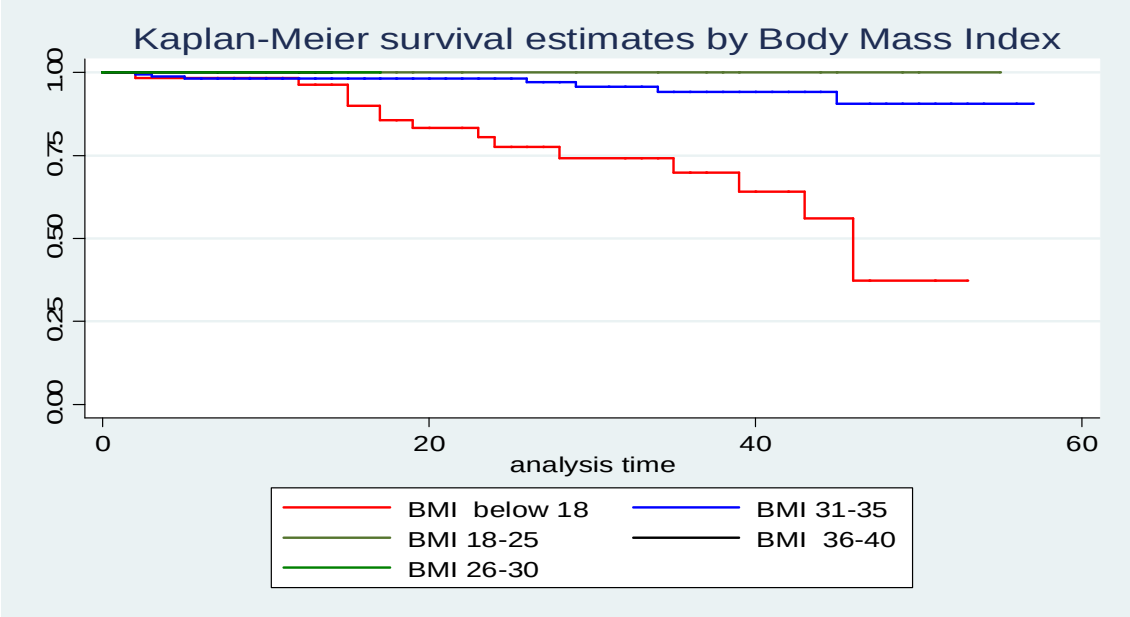


Figure 4.1.6 Kaplan-Meier Survival Estimates by Body Mass Index



From the figure above it shows that survival is affected by low Body Mass Index, with those who have low BMI at the beginning of treatment having a high mortality rate.

Table 4.2.7 Log-Rank Test of Significance by Sex

Log-rank test for equality of survivor functions

sex	Events observed	Events expected
1	9	6.86
2	14	16.14
Total	23	23.00

chi2(1) = 0.97
Pr>chi2 = 0.3257

From the Chi-Square test above with 1 d.f (α=0.05) the significance level is 0.3257 which is more than 0.05 indicating that survival is independent of sex.

Table 4.2.8 Log- Rank Test of Significance by Drug Abuse

```
. sts test drudabs

         failure _d:  status == 1
   analysis time _t:  months
```

Log-rank test for equality of survivor functions

drudabs	Events observed	Events expected
0	7	16.65
1	15	5.35
Total	22	22.00

chi2(1) = 23.17
Pr>chi2 = 0.0000

From the Chi-Square test above with 1 d.f (α=0.05) the significance level is below 0.05. Therefore, survival is dependent on drug abuse.

Table 4.2.9 Log- Rank Test of Significance by drug adherence

```
. sts test drugadh
        failure _d:  status == 1
  analysis time _t:  months

Log-rank test for equality of survivor functions
```

drugadh	Events observed	Events expected
0	6	19.49
1	17	3.51
Total	23	23.00

chi2(1) = 61.75
Pr>chi2 = 0.0000

From the Chi-Square test above with 1d.f (α=0.05) the significance level is below 0.05 indicating that survival is dependent on drug adherence.

Table 4.2.10 Log Rank Test of Significance by Tuberculosis Status

Log-rank test for equality of survivor functions

tbtrtmnt	Events observed	Events expected
0	14	20.16
1	9	2.84
Total	23	23.00

chi2(1) = 15.39
Pr>chi2 = 0.0001

From the Chi-Square test above with 1 d.f (α=0.05) the significance level is below 0.05 and indication that survival is dependent on tuberculosis status.

Table 4.2.11 Log-Rank Test of Significance by WHO Staging

Log-rank test for equality of survivor functions

whostgng	Events observed	Events expected
1	1	7.21
2	2	4.01
3	9	7.65
4	11	4.14
Total	23	23.00

chi2(3) = 18.14
Pr>chi2 = 0.0004

The Chi-Square test above with 3 d.f (α=0.05) indicates that survival depends on WHO clinical stage.

Table 4.1.12 Cox Proportional Hazard Regression Stata Output

```
        failure _d:  statusattheendofthestudy == 1
  analysis time _t:  timeinmonths

Iteration 0:   log likelihood = -105.02045
Iteration 1:   log likelihood = -79.818325
Iteration 2:   log likelihood = -67.038923
Iteration 3:   log likelihood = -63.698112
Iteration 4:   log likelihood =  -63.54586
Iteration 5:   log likelihood = -63.544497
Iteration 6:   log likelihood = -63.544497
Refining estimates:
Iteration 0:   log likelihood = -63.544497

Cox regression -- Breslow method for ties

No. of subjects =          247                     Number of obs   =       247
No. of failures =           22
Time at risk    =         6696
                                                   LR chi2(10)     =     82.95
Log likelihood  =   -63.544497                     Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageinyears	1.025696	.0329788	0.79	0.430	.9630532	1.092413
sex	1.697321	1.22641	0.73	0.464	.4118348	6.995278
historyofd~e	2.55818	1.593383	1.51	0.132	.7546632	8.671795
drugadhere~e	2.987397	1.891913	1.73	0.084	.8634329	10.33611
tbtreatmen~s	.6513177	.3721881	-0.75	0.453	.212512	1.996192
weightatth~t	1.029458	.0655568	0.46	0.648	.9086646	1.16631
regimen	.8206991	.4880643	-0.33	0.740	.2558468	2.632618
cd4cellcou~a	.9825466	.0050197	-3.45	0.001	.9727572	.9924345
bmiatthebe~t	.7868494	.1489496	-1.27	0.205	.5429506	1.14031
whoclinica~g	1.035342	.3308147	0.11	0.913	.5534856	1.936697

In the proportional hazards model of Cox independent failure times $T_1, T_2, \ldots\ldots, T_n$ are studied, here the distribution is described by a hazard function $\lambda(t)$ given by:

$$\lambda(t) = \lambda_0(t) \times \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})$$

Therefore hazard rate from Table 4.1.12 will be given by:

$$\lambda(t)=\lambda_t \exp(\beta_0+(1.025696X_1)+(1.697321X_2)+(1.029458X_3)+(2.55818X_4)+(0.9825466X_5)+$$

$$(0.8206991X_6)+(1.035342X_7)+(0.7868494X_8)+(0.6513177X_9)+(2.987397X_{10})$$

From the output above, those with poor adherence have high mortality (they are 3 more times likely to die than those with good adherence) followed by those who abuse drugs. Tuberculosis treatment, the drug regime, CD4 cell count and BMI are not significant predictors of mortality.

Table 4.1.13 Weibull Regression – Accelerated failure-time form

```
        failure _d:  statusattheendofthestudy == 1
  analysis time _t:  timeinmonths

Fitting constant-only model:

Iteration 0:   log likelihood =  -85.18952
Iteration 1:   log likelihood = -83.588505
Iteration 2:   log likelihood = -83.561868
Iteration 3:   log likelihood = -83.561862

Fitting full model:

Iteration 0:   log likelihood = -83.561862
Iteration 1:   log likelihood = -62.148872
Iteration 2:   log likelihood = -46.105587
Iteration 3:   log likelihood = -42.991148
Iteration 4:   log likelihood = -42.366948
Iteration 5:   log likelihood = -42.361299
Iteration 6:   log likelihood = -42.361298

Weibull regression -- accelerated failure-time form

No. of subjects =          247                 Number of obs   =       247
No. of failures =           22
Time at risk    =         6696
                                               LR chi2(10)     =     82.40
Log likelihood  =   -42.361298                 Prob > chi2     =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageinyears	-.0226651	.0188084	-1.21	0.228	-.0595289	.0141987
sex	-.4876807	.4335439	-1.12	0.261	-1.337411	.3620496
historyofd~e	-.6353027	.3909335	-1.63	0.104	-1.401518	.1309129
drugadhere~e	-.4202942	.3530063	-1.19	0.234	-1.112174	.2715855
tbtreatmen~s	.2401341	.3332363	0.72	0.471	-.412997	.8932652
weightatth~t	-.01958	.0374351	-0.52	0.601	-.0929514	.0537913
regimen	.0480747	.3407168	0.14	0.888	-.619718	.7158674
cd4cellcou~a	.0101007	.0032049	3.15	0.002	.0038191	.0163823
bmiatthebe~t	.1592278	.1150586	1.38	0.166	-.066283	.3847386
whoclinica~g	-.0631921	.184681	-0.34	0.732	-.4251603	.298776
_cons	3.396271	1.818765	1.87	0.062	-.1684428	6.960985
/ln_p	.5271729	.1829523	2.88	0.004	.1685929	.8857529
p	1.694136	.3099461			1.183638	2.424809
1/p	.5902714	.1079915			.4124036	.8448527

In the Weibull model, the hazard rate is characterized as:

$h\,(t;X) = \lambda p(\lambda t)^{p-1}$ $\qquad P = \begin{cases} 1 & \text{for those with two parameters to be estimated} \\ 0 & otherwise \end{cases}$

$\lambda_i = e^{X_i\beta}$

Therefore $h\,(t;X)$ from Table 4.1.13 will be given by;

$h\,(t;x) = \lambda_2\{exp((3.230099)+(-.0226651X_1)+(-.4876807X_2)+$

$(-.01958X_3)+(-.6353027X_4)+(.0101007X_5)+(.0480747X_6)+$

$(-.0631921X_7)+(.1592278X_8)+(.2401341X_9)+(-.4202942X_{10})\}\}^{2-1}$

Table 4.2.14 Exponential Regression – Accelerated failure-time form

```
        failure _d:  statusattheendofthestudy == 1
  analysis time _t:  timeinmonths

Iteration 0:   log likelihood =  -85.18952
Iteration 1:   log likelihood = -63.567746
Iteration 2:   log likelihood = -49.136637
Iteration 3:   log likelihood = -46.263069
Iteration 4:   log likelihood = -45.767336
Iteration 5:   log likelihood = -45.762195
Iteration 6:   log likelihood = -45.762194

Exponential regression -- accelerated failure-time form

No. of subjects =          247                     Number of obs   =       247
No. of failures =           22
Time at risk    =         6696
                                                   LR chi2(10)     =     78.85
Log likelihood  =   -45.762194                     Prob > chi2     =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageinyears	-.0385914	.0297631	-1.30	0.195	-.096926	.0197432
sex	-.8851462	.7036777	-1.26	0.208	-2.264329	.4940366
historyofd~e	-1.127786	.5875255	-1.92	0.055	-2.279314	.0237432
drugadhere~e	-.6069931	.5763563	-1.05	0.292	-1.736631	.5226445
tbtreatmen~s	.2369674	.5471968	0.43	0.665	-.8355185	1.309453
weightatth~t	-.0346983	.0617055	-0.56	0.574	-.1556388	.0862422
regimen	.3091138	.5592424	0.55	0.580	-.7869812	1.405209
cd4cellcou~a	.0153152	.004402	3.48	0.001	.0066874	.0239429
bmiatthebe~t	.2822042	.1851358	1.52	0.127	-.0806553	.6450637
whoclinica~g	-.0954622	.300294	-0.32	0.751	-.6840277	.4931032
_cons	3.230099	2.980043	1.08	0.278	-2.610678	9.070876

.

In the exponential model, the hazard rate is characterized as: $h\ (t;\ X) = \lambda$

This implies that the conditional 'probability' of an event is constant over time (and that events occur according to a Poisson process). In other words, the risk of an event occurring is flat with respect to time. Modeling the dependency of the hazard rate on covariates entails constructing a model that ensures a non-negative hazard rate .the way to do this is by exponentiating the covariates such that:

$h\ (t;X) = \lambda_i = e^{X_i\beta}$

Therefore the from Table 4.2.14 $h\ (t;X)$ is given by;

$h\ (t;X)$ = exp ((3.230099)+(-.0385914 X_1)+(-.8851462X_2)+(-.0346983X_3)+(-1.127786X_4)+

(.0153152X_5)+(.3091138X_6)+(-.0954622X_7)+(.2822042X_8)+(.2369674X_9)+(-.6069931X_{10})

If h (t) < 1, then the hazard is monotonically decreasing with time.

If h (t) > 1, then the hazard is monotonically increasing with time.

If h (t) = 1, then the hazard is flat.

4.2.15 Lognormal regression – accelerated failure-time form

```
        failure _d:  statusattheendofthestudy == 1
  analysis time _t:  timeinmonths

Fitting constant-only model:

Iteration 0:   log likelihood = -110.87441
Iteration 1:   log likelihood = -92.904646
Iteration 2:   log likelihood = -85.780505
Iteration 3:   log likelihood = -85.002024
Iteration 4:   log likelihood = -84.966454
Iteration 5:   log likelihood = -84.966309
Iteration 6:   log likelihood = -84.966309

Fitting full model:

Iteration 0:   log likelihood = -84.966309  (not concave)
Iteration 1:   log likelihood = -61.228004
Iteration 2:   log likelihood = -57.070323
Iteration 3:   log likelihood = -46.037048
Iteration 4:   log likelihood = -44.629809
Iteration 5:   log likelihood = -44.605655
Iteration 6:   log likelihood = -44.605613
Iteration 7:   log likelihood = -44.605613

Lognormal regression -- accelerated failure-time form

No. of subjects =          247                     Number of obs   =       247
No. of failures =           22
Time at risk    =         6696
                                                   LR chi2(10)     =     80.72
Log likelihood  =   -44.605613                     Prob > chi2     =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageinyears	-.0173423	.0207457	-0.84	0.403	-.0580031	.0233185
sex	-.6466395	.4924402	-1.31	0.189	-1.611804	.3185255
historyofd~e	-.7591358	.3975502	-1.91	0.056	-1.53832	.0200482
drugadhere~e	-.4690429	.3967813	-1.18	0.237	-1.24672	.3086341
tbtreatmen~s	-.0329909	.398103	-0.08	0.934	-.8132584	.7472766
weightatth~t	-.0402286	.0437548	-0.92	0.358	-.1259865	.0455294
regimen	.2420188	.4262894	0.57	0.570	-.5934931	1.077531
cd4cellcou~a	.0104868	.002721	3.85	0.000	.0051538	.0158199
bmiatthebe~t	.2009889	.1330154	1.51	0.131	-.0597164	.4616942
whoclinica~g	-.0390171	.1895374	-0.21	0.837	-.4105036	.3324694
_cons	3.286084	1.920954	1.71	0.087	-.4789163	7.051084
/ln_sig	-.0025935	.1560936	-0.02	0.987	-.3085314	.3033444
sigma	.9974098	.1556893			.7345249	1.354381

CHAPTER FIVE

5.0 DISCUSSION OF THE RESULTS

5.1 Discussion

Health researchers are often interested in semi parametric models in analysis of time to event data more than the parametric models but, in a recent review of survival analyses in cancer journals (Altman et al., 1985), it was found that only 5 per cent of all studies using the Cox PH approach model with respect to checking the underlying assumptions. If this assumption does not hold, the Cox model can lead to the unreliable conclusions so Parametric models such as Lognormal, Weibull and Exponential are the common choices. These models provide the interpretation based on a specific distribution for duration times without need to proportional hazard assumptions.

The aim of this study was to investigate the comparative performance of Cox and parametric models in a survival analysis of patients on HAART. I used Akaike Information Criterion (AIC) to evaluate the performance of models in analysis. In my example the proportional hazard assumptions were hold and the all parametric model residual indicated a perfect fit. I explored the impact of gender, age in years, sex, history of drug abuse, history of drug adherence,TB treatment status,Weight,regimen,CD4 cell count, BMI at the beginning and WHO Clinical staging on survival time and all parametric and semi parametric models. Multivariate analysis showed an increased risk of death for patients who had poor drug adherence [HR=2.987397 SE= 1.891913 (P=0.05) CI (0.8634329 10.33611)], who where substance abusers [HR=2.55818 SE= 1.593383 (P=0.05) CI (.7546632 8.671795)] for multivariate analysis in cox PH model. Age, sex and WHO clinical staging was not an important

predictor of survival. From the parametric multivariate analysis the output produced exponentiated coefficients (hazard ratios were reported). From the Weibull output (Table 4.1.13) a one-unit change in age causes acceleration(risk of death) by 2.27% , whereas a one-unit change in weight cause acceleration by approximately 2% but a unit change in BMI improves survival, this is by causing deceleration by 1% .Interpreting results in the PH metric is easier, though regression coefficients are not difficult to interpret in the AFT metric. A positive coefficient means that time is decelerated by a unit increase in the covariate in question. This may seem awkward, but think of this instead as a unit increase in the covariate causing a delay in failure and thus increasing the expected time until failure. The difficulty that arises with the AFT metric is merely that it places an emphasis on log (time-to failure) rather than risk (hazard) of failure.

Drug abuse was a strong and independent prognostic factor for death as a result of HIV/AIDS, and these findings in multivariate analysis is in conformity with previous reports (Arveux et al., 2002) indicated poor survival for these patients. Lack of drug adherence is another important prognostic factor of survival (Haugstvedt et al., 2003) many authors show that the survival depends on the level of cd4 cell count at the beginning of treatment(Moore et al., 2006). Sex and treatment regime are not significant factor that affected the survival probability of patients in both univariate and multivariate analysis.

The evaluation criteria indicated accelerated failure-time form as the best models in multivariate analysis. Although it seems that there may not be a single model that is substantially better than others, the data strongly supported the Weibull (Table 5.1) AFT form regression among parametric models in multivariate analysis and it can be lead to more precise results as an alternative for Cox.

5.2 Akaike information criterion

Akaike information criterion (AIC) was used to compare all these AFT models. Nested models can be compared using the likelihood ratio test. The exponential model, the Weibull model and log-normal model are nested within gamma model. For comparing models that are not nested, the Akaike information criterion (AIC) can be used instead, which is defined as:

$$AIC = -2l + 2(k + c)$$

where l is the log-likelihood, k is the number of covariates in the model and c is the number of model specific ancillary parameters. The addition of 2(k +c) can be thought of as a penalty if non-predictive parameters are added to the model. Lower values of the AIC suggest a better model.

Table 5.1 AIC For Parametric Models in HIV/AIDS Survival - Multivariate Analysis

A F T MODEL	A I C VALUES	LOG LIKELIHOOD
WEIBULL	-2(-42.361298)+2(10+2)=**108.722596**	-42.361298
EXPONENTIAL	-2(-45.752194)+2(10+1)=**113.504388**	-45.762194
LOG-NORMAL	-2(-44.605613)+2(10+2)=**113.211226**	-44.605613
COX PH MODEL	-2(-63.544497)+2(10+2)=**151.088994**	-63.544497

Based on AIC, there is no major variability between the three parametric models as witnessed with Cox Proportional Hazard model (Table 5.1). Among the parametric models Wiebull is the best model in multivariate analysis. The Weibull AFT model (Table 5.1) appears to be an appropriate AFT model according to AIC compared with other AFT models, although it is only slightly better than exponential or log-normal model. Cox PH model fair poorly compared to other parametric models. The largest log likelihood for parametric models was obtained for the exponential model; which is also not preferred by the AIC. Although the best-fitting model is the

one with the largest log likelihood, the preferred model is the one with the smallest AIC value (Akaike, 1974).

CHAPTER SIX

STUDY LIMITATIONS, CONCLUSIONS AND RECOMMENDATIONS

6.1 INTRODUCTION

This chapter explains the study limitations, conclusions that are drawn from the study and the recommendations of what should be done to improve use of statistical models in health research.

6.2 STUDY LIMITATIONS

A limitation of this data is the percent of censoring. A good discrimination among parametric models requires the censoring percentage not to exceed 40-50 per cent (Nardi et al., 2003) although in my data the censoring was about 90 per cent, the parametric results were not performed and were significance.

6.3 CONCLUSION

Survival analysis can be used to analyze data on the length of time it takes for a specific event to occur. A characteristic of “time to event” data is that we did not know the actual time to event for every person in our data set. We knew this only for some individuals. In this regard, our data are incomplete (i.e. censored).The study was seeking to; (1) describe the pattern (distribution) of event times of the cohort under study; this was done using Kaplan-Meier”, (2) to compare patterns of time to event across groups which was done using Log Rank Test and to explore the influences of possibly several factors on “time to event” which was achieved by Cox Proportional Hazard regression and accelerated failure time model (considering weibull,exponential and lognormal forms).Evaluation was also carried out on the models to establish which model is more appropriate than the other in the analysis.

In this study, the survival/ death status of HIV/AIDS infected subjects who were on HAART treatment at Karuri Health Centre between January 1st 2008 to 31st December 2012 was studied. The Kaplan-Meier method was used to estimate the survival time of patients after commencement of HAART. The mortality rate was high in subjects who abused drug, those who did not adhere to treatment as prescribed, those who had tuberculosis and those who had low cd4 cell count and low BMI at the beginning of treatment (Figures 4.1.2, 4.1.3, 4.1.4, 4.1.5, 4.1.6).This claim was further authenticated by performing log lank test which produced similar results (tables 4.2.11 , 4.2.9, 4.2.10, 4.2.8) .Using Cox proportional hazard model covariates that significantly influence the survival of HIV/AIDS infected patients were identified. Three covariates that are identified to affect the survival of the patients at 0.05 level of significant were gender, history of drug abuse and poor treatment adherence (Tables 4.2.15, 4.1.13, 4.2.14, 4.1.12).

Akaike Information Criterion was used to evaluate the performance of the models in analyzing the data. There was no major variability between the three parametric models as witnessed with Cox Proportional Hazard model (Table 5.1). Among the parametric models Wiebull was the best model in multivariate analysis based on the value of the AIC. The Weibull AFT model (Table 5.1) appears to be an appropriate AFT model according to AIC compared to other AFT models,Cox PH model fair poorly compared to other parametric models. The largest log likelihood for parametric models was obtained for the exponential model; this model was also not preferred by the AIC.

6.4 RECOMMENDATIONS

One main disadvantage of using the AFT model is that the specific distribution of survival time is not known in many cases. As proposed by Wei (1992), further studies of this data could attempt using a non-parametric version of the AFT model which does not require the specification of the distribution can be applied in this dataset. The results from this model could then be compared with the standard AFT models and Cox PH models. In addition, further study can be carried out to appraise the effects of practical cases such as enormous censoring.

In spite of advantage in using these models for survival analysis I recommend that further studies should be carried out to evaluate the effects of practical cases such as small sample size, large censoring and changing in proportional hazard assumption or duration time's distribution.

REFERENCES

Abdelmonem Afifi, Virginia A. Clark & Susanne May, (2004), Computer Aided Multivariate Analysis, (4th Ed.) Boca Raton, Chapman & Hall/CRC.

Adachi Y, Oshiro T, Mori M, et al (1996). Prediction of early and late recurrence after curative resection for gastric carcinoma.

Ajay K. Sethi, Stephen J. Gange (2009) Wisconsin medical journal, Parametric Models for Studying Time to Antiretroviral Resistance Associated With Illicit Drug Use,

Akaike H (1974). A new look at the statistical model identification. IEEE Trans Automatic Control.

Altman DG, De Stavola BL, Love SB, et al (1985). Review of survival analyses published in cancer journals. Bri J Cancer,

Arveux P, Faivre J, Boutron M-C, et al (1992). Prognosis of gastric carcinoma after curative surgery. A population-based study using multivariate crude and relative survival analysis.

Adachi Y, Oshiro T, Mori M, et al (1996). Prediction of early and late recurrence after curative resection for gastric carcinoma.

Basavarajaiah D. M, Narasimahamurthy B. & Leelavathy B. (2013) International Journal of General Medicine and Pharmacy (IJGMP), Comparative Analysis Of Survival And Growth Model Approach With Relation To Cd4 Count In Hiv- Positive Pregnant Women ISSN.

Dianne M. Finkelstein, Dirk F. Moore & David A. SchoenfeldSource, (1993), Biometrics, A Proportional Hazards Model for Truncated AIDS Data.

D. M. Moore, RS. Hogg, B. Yip, K. Craib, E. Wood & JSG Montaner, (2006), HIV Medicine, CD4 Percentage is an Independent Predictor of Survival in Patients Starting Antiretroviral Therapy with Absolute CD4 Cell Counts between 200 And 350 Cells/µl.

Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents by The HHS Panel on Antiretroviral Guidelines for Adults and Adolescents A Working Group of the Office of AIDS Research Advisory Council (OARAC).

Hung Ju Li, (2007), Cox Proportional Hazard Model With Cross Sectional HIV Prevalence Data, Institute of Statistics, National University of Khaohsiung, Taiwan.

John Fox,(2002), Cox Proportional-Hazards Regression for Survival Data, Cox Proportional-Hazards Regression for Survival Data.

Loeys T & E Goetghebeur, (2003), Biometrics, A Casual Proportional Hazard Estimator for the Effect of Treatment Actually Received in a Randomized Trail with All or Nothing Compliance.

M. A. Hernan et al. (2005), Pharmaco-epidemiology and Drug Safety, Structural Accelerated Failure Time Models for Survival analysis in Studies with Time-Varying Treatments, Wiley InterScience (www.interscience.wiley.com).

Ministry of Health, Government of Kenya,(2001), Guidelines for Antiretroviral Drug Therapy in Kenya, National Aids and STI Control Program (NASCOP), 3rd Ed.

Miguel Angel Hernan, Babette Brumback, & James M. Robins, (2000), Epidemiology, Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men.

Mohamad Amin Pourhoseingholi et al, (2007), Asian Pacific Journal of Cancer Prevention, Comparing Cox Regression and Parametric Models for Survival of Patients with Gastric Carcinoma.

Pierre De Beaudrap et al, (2008) European Journal of Epidemiology, Change over Time of Mortality Predictors after HAART Initiation in a Senegalese Cohort.

Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer,J., Satten, G. A., Aschman, D. J., Holmberg, S. D., and The HIV Outpatient Study Investigators(1998). Declining morbidity and mortality among patients with advanced human immunode.ciency virus infection. The New England Journal of Medicine.

Raj S. Chhikara, Johnny Conkin and Laura A. Thompson, (1985), The NASA STI Program, Cox Proportional Hazards Models for Modeling the Time To Onset of Decompression Sickness in Hypobaric Environments.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R.(1990)Martingale-based residuals for survival models. Biometrika.

Tuner BJ, Markson LE, Mckee L, et al(1991). The aids-defining diagnosis and sub-sequent complications: a survival-based severity index. Journal of Acquired Immune Deficiency Syndromes .

Wei, L. J.(1992) The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine 11.

William B. Goggins & Dianne M. FinkelsteinSource, (2000), Biometrics, A Proportional Hazards Model for Multivariate Interval Censored Failure Time Data.

Zhiguo Li, Shiyu Zhou, Suresh Choubey & Crispian Sievenpiper, (2007), IIE Transactions, Failure Event Prediction Using The Cox Proportional Hazard Model Driven by Frequent Failure Signatures.

Yang, S. (1997), “Extended Weighted Log-Rank Estimating Functions in Censored Regression,” Journal of the American Statistical Association.

William B. Goggins, Dianne M. Finkelstein & Alan M. Zaslavsky, (1999), Biometrics, Applying of Cox Proportional Hazards Model When the Change Time of a Binary Time-Varying Covariate is Interval Censored.

APPENDIX I TIME FRAME

Task to be completed	DATE						
	August 2013	August 2013	August 2013	September 2013	September 2013	October 2013	October/ November 2013
Identification of research topic							
Collection of background information							
Drafting of proposal and presentation							
Drafting of proposal and presentation							
Data cleaning							
Data analysis							
Report writing, submission and presentation of final draft.							

APPENDIX II STUDY BUDGET

Items	Quantity	Unit Price (Kshs)	Total (Kshs)
Stationery & Equipment			
Printing Papers	5 reams	400.00	2,000.00
Black Cartridges (for HP 845C)	2	2,000.00	4,000.00
Coloured Cartridge (for HP 845C)	1	2,500.00	2,500.00
Writing Pens	1 packet	500.00	500.00
Flash Discs	2	1,000.00	2,000.00
Note Books	10	30.00	300.00
Full Scups (Compilation Sheets)	2 reams	250.00	500.00
Box Files	5	100.00	500.00
Document Wallets	6	50.00	300.00
Tape Recorder	2	3,500.00	7,000.00
Sub total			19,600.00
Research Proposal Development			
Printing drafts & final proposal	10 copies	500.00	5,000.00
Photocopies of final proposal	6 copies	100.00	600.00
Binding of copies of Proposal	5 copies	40.00	200.00
Sub total			5800.00
Personnel			
Research Assistants	2 for 2 months	5,000.00	20,000.00
Sub total			20,000.00
Thesis Development			
Printing of drafts and final thesis	10 copies	800.00	8,000.00
Photocopy of final thesis	6 copies	200.00	1,200.00
Binding of thesis at Main Campus Library	6 copies	250.00	1,500.00
Dissemination cost			15,000.00
Sub total			25,700.00
Miscellaneous			
Grand Total			72,100.00

APPENDIX III REQUEST FOR DATA

AUGUSTINE GATIMU NJUGUNA,

P.O BOX 640-00900,

KIAMBU,

06-10-2013.

TO:

CLINICAL OFFICER INCHARGE,

KARURI HEALTH CENTRE,

P.O BOX 39-00900,

KIAMBU.

RE: REQUESTING FOR DATA FROM COMPREHENSIVE CARE CENTER.

I am Augustine Gatimu Njuguna a student at university of Nairobi taking Masters Of Science In Medical Statistics. I am in the process of carrying out my project and I would like to make use of your health facility Comprehensive Care Centre data. Attached find a copy of my concept paper for a brief description of what I intend to do with the data. Your assistance will be highly appreciated.

Yours faithfully

Augustine .G. Njuguna