STATISTICAL MODELLING FOR THE SURVIVAL OF HIV/AIDS INFECTED PATIENTS TAKING ANTIRETROVIRAL TREATMENT IN NTCHEU: A RETROSPECTIVE COHORT STUDY

MSc (Biostatistics)

RABSON JOSEPH MANGISON DZINZA

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

AUGUST, 2015

STATISTICAL MODELLING FOR THE SURVIVAL OF HIV/AIDS INFECTED PATIENTS TAKING ANTIRETROVIRAL TREATMENT IN NTCHEU: A RETROSPECTIVE COHORT STUDY

MSc (Biostatistics)

By

RABSON JOSEPH MANGISON DZINZA

BSc (Mathematics and Statistics)-University of Malawi

Thesis submitted to Faculty of Science in Partial Fulfillment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

AUGUST, 2015

Declaration

I the undersigned hereby declare that this thesis/dissertation is my own original work which has not been submitted to any other institution for the similar purpose. Where other people work has been used acknowledgement has been made.

RABSON JOSEPH MANGISON DZINZA

Name

_____

Signature

AUGUST, 2015

Date

Certificate of approval

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

Signature_____Date: _____

JIMMY J. NAMANGALE,PhD (ASSOCIATE PROFESSOR)

Supervisor

Signature _____ Date: _____

MAVUTO MUKAKA, PhD (SENIOR LECTURER)

Supervisor

Signature_____Date: _____

TSIRIZANI KAOMBE, MSc (LECTURER)

Course coordinator

Dedication

To Almighty God

and

To my wife and children

Acknowledgements

First and foremost, I would like to thank the Almighty God for His guidance and good health He provides me daily and this thesis could be in vain without His hand. May His name be glorified. May I thank associate Professor Jimmy J Namangale PhD, my supervisor, for continuously directing me on how to do this thesis. This thesis could not be successfully done without the tireless effort provided by him. He is a person of encouragement. His vast knowledge in research has enabled me to be independent in my dealings; hence, may Almighty God bless him abundantly. I'm deeply in debited to Dr Sarah White as well as Dr Mavuto Mukaka, who opened my eyes on how to do survival analysis. I'm sincerely grateful to Dr Bagrey Ngwira for providing me with some pieces of advice. I want to thank my family and friends for their support and encouragement. I would like also to thank District Health Officer (DHO) and entire management for Ntcheu district hospital for providing me with the data used in this study.

Abstract

The Cox proportional Hazard model has been broadly used and most popular technique in survival analysis. Under certain situations parametric models may offer advantages over Cox proportional hazard model. In this study five parametric models and the Cox proportional hazard model were fitted. The main objective in this study was to compare the performance of five models: exponential, loglogistic, Weibull, lognormal and Gompertz as well as the semi-parametric Cox model on the survival among HIV/AIDS infected patients taking ART in Ntcheu district. Five parametric models as well as Cox proportional model were fitted to 6670 HIV/AIDS patients who registered for ART from 2007-2012.The results of the analysis using Cox proportional hazard model and the Gompertz model were comparable. In both models, WHO clinical stage 4, body mass index and being male were found to be significantly associated with the survival of HIV/AIDS patients taking ART in Ntcheu district. In the multivariable analysis all the parametric models fit better than Cox model with respect to AIC and the Gompertz model was found to be the best model for modelling the survival among the HIV/AIDS infected patients taking ART in Ntcheu district.

Table of contents

List of Figures

List of Tables

List of abbreviations and acronyms

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
ARVs	Antiretrovirals
BMI	Body mass index
CART	Combination of antiretroviral Therapy
CI	Confidence Interval
EGPAF	Elizabeth Glaser Pediatric AIDS Foundation
HIV	Human immunodeficiency Virus
HR	Hazard Ratio
PEPFAR	President's Emergency Plan for AIDS Relief
TB	Tuberculosis
UNAIDS	Joint United Nations Programme on HIV and AIDS
UNC	University of North Carolina
UNICEF	United Nations Children Education Fund
USAID	United States Agency for International Development
WHO	World Health Organization

Chapter 1

1. Introduction

1.0 Background information

1.1 Global HIV/AIDS prevalence

The world as a whole has been greatly hit with the high prevalence of HIV/AIDS. The disease has claimed lives of most people. The world Health Organization estimates that by the end of 2011, 34 million people (31.4 to 35.9million) were living with HIV/AIDS globally (UNAIDS, 2012). The annual deaths worldwide have been estimated to be around 1.7 million (1.5 to 1.9million) in 2011 (UNAIDS, 2012) and this was 24% fewer deaths than in 2005. However statistical figures have shown that there is a decrease in the new HIV/AIDS infections by 35% since 2000 by the end of 2014 (UNAIDS, 2015). The decrease has been due to the antiretroviral therapy which reduces the viral load of a person with HIV to virtual undetectable levels and it also reduces the risk of transmitting the virus to uninfected partner (UNAIDS, 2015). A new report from Joint United Nations Programme on HIV/AIDS, the World Health Organization and the UNICEF on HIV treatment 2013, indicates a huge acceleration in the roll out and uptake of antiretroviral therapy since 2011.Arecord of 14.9 million people living with HIV were accessing treatment in 2014 compared to 13.0 million in 2013 (UNAIDS, 2015).

1.2 HIV/AIDS prevalence in sub-Saharan Africa

The sub-Saharan Africa remains the most heavily affected region in the global HIV epidemic with 1 death in every 20 people living with HIV (UNAIDS, 2012). In 2011 an estimated 23.5 million (22.1 to 24.8million) people living with HIV resided in sub-Saharan Africa, representing 69% of the global HIV burden (UNAIDS, 2012).The number of newly infected cases continue to fall and there has been a decline in the number from AIDS-related illness (UNAIDS, 2012). In 2011, there were an estimated 1.8 million (1.6 to 2 million) new HIV infections in Sub-Saharan Africa compared to 2.4 million (2.2 million-2.5 million) new infections in 2001, a 25% decline (UNAIDS, 2012).UNAIDS (2012) has indicated that between 2005 and 2011, the number of people dying from AIDS-related causes in sub-Saharan Africa declined by 32%, from 1.8million (1.6million-1.9million) to 1.2 million (1.1 million-1.3 million).

1.3 HIV/AIDS prevalence in Malawi

Malawi is one of the countries from the sub-Saharan region and is divided into three regions, South, Central and North. The 2012 Global AIDS Response Progress Report(2010- 2011) has revealed that the HIV prevalence in Malawi has decreased significantly from 17.6% in 2004 to 14.5% in 2010 in the southern region and from 8.1% in 2004 to 6.6% in 2010 in the northern region. The central region has experienced an increase in the prevalence of HIV as it has gone up from 6.5% in 2004 to 7.6% in 2010(Global AIDS Response progress Report, 2012). The USAID/Malawi has indicated that the overall HIV prevalence in urban is 17.4% and that of rural areas is 8.9% with an estimated of about 920000 people infected with HIV/AIDS in 2009.

The Malawi Demographics profile(2013) indicate that the HIV/AIDS adult prevalence rate in Malawi is 11% as estimated in 2009 of which 51000 people had died of HIV/AIDS in the same year. The Malawian government started giving free antiretroviral therapy in public sectors in 2004 (UNAIDS, 2012) and a total of 10761 were on ART. The Malawian government in 2011 received funding from the President's Emergency Plan for AIDS Relief (PEPFAR) through USAID for essential HIV/AIDS programme and services (USAID Malawi, 2012). The funding has helped to register significantly large number of HIV infected patients up to about 477022 who were initiated on ART by the end of 2011. It has been indicated that 347983(73%) HIV/AIDS infected patients were on ART by the 31st March, 2012(Ministry of Health Malawi, 2012). This is an improvement compared to 2004 HIV/AIDS infected people who were registered on ART.

1.4 HIV/AIDS prevalence in Ntcheu

Ntcheu district is located in the central region. It has the population size of 471589 (Malawi population census, 2008).The district hospital has 269 beds and there are 37 health centres of which 15 are private health centres.The hospital started providing free Antiretroviral (ARV) drugs in 2005.However out of these 37 health centres, 5 do not provide the ARVs due to the fact that that they have no trained staff like Medical assistants and nurses. In addition to that the district has not enough qualified personnel to administer HIV/AIDS test. Ntcheu had since registered 11340 HIV/AIDS infected patients (Ministry of Health Malawi, 2012) and all them are on ART.

A number of nongovernmental organizations are taking part in addressing the dangers of HIV/AIDS in the district so as to help in the reduction of the HIV/AIDS prevalence. These include the Concern Universal, the University of North Carolina (UNC) and Elizabeth Glaser Pediatric AIDS Foundation(EGPAF).The activities done by these nongovernmental organizations include prevention of mother to child transmission, community discussion with people living HIV/AIDS and training of health workers. In addition to what the nongovernmental organizations are doing, the hospital has also resources which are being used to address the HIV/AIDS prevalence like drugs for the prevention of mother to child transmission of HIV. It does also provide information about HIV/AIDS, civic education and communications through posters among others. Despite these efforts lack of some drugs like flaconazole, vinaristine as well as cotrimoxazole preventive therapy for treating opportunistic infections continue to be a major factor contributing to the mortality of HIV/AIDS infected patients and this poses challenge for the district hospital.

1.5 Problem statement

The most common approach to analysis of HIV/AIDS data is the “time to event” analysis. Often, the question of interest is either “time to death”; “time to be discharged alive from hospital” or “time to CD4 immuno recovery” just to mention a few. Researchers commonly apply the semi-parametric Cox regression model to address these questions. Much as the Cox regression is considered to be robust to model misspecifications in many cases, the choice of this model among medical researchers is often based on routine use of the method. Parametric models seem to play little or no role in these analysis despite the huge literature on these methods. In

fact many medical researchers may not be aware that parametric methods also exists that might perform better than the Cox in some isolated cases. Furthermore, many researchers do not even make an attempt to test the goodness of fit of the Cox model. The plausibility of the Proportional Hazards assumption is also often not tested. I feel that there is a gap among health researchers in the analysis approach for this data. I believe that in some cases, the parametric model may be more applicable than routinely used Cox model. I would, thus, advocate that when analyzing "time to event" data, it is important to consider both the semi-parametric (e.g. Cox regression) as well as the parametric methods. Under each data scenario, both the semi-parametric and the parametric methods should be fitted, compared and the best performing model should be chosen. Emphasis should also be on the testing of the goodness of fit of the model and the model assumptions. In order to provide an informed guidance on how to approach "time to event" analyses, I analysed data from cohort studies that were done at Ntcheu district hospital using both semi-parametric and parametric methods. The fitted models were compared and goodness of fit was assessed. The assumptions were also tested.

1.6 Main objective

The main objective in this study is to compare five statistical parametric models as well as Cox Proportional hazard model for modelling the survival among HIV/AIDS infected patients receiving ART therapy in Ntcheu district.

1.7 Specific objectives

a) To compare statistical parametric models on the survival among HIV/AIDS infected Patients receiving ART therapy in Ntcheu district.

b) To determine factors associated with mortality of HIV/AIDS infected patients.

1.8 Justification

Survival parametric models make it possible for clinicians to interpret the treatment benefit in terms of an effect on expected duration of illness. Hence the models have explanatory advantage in that covariates have a direct effect on survival rather than on hazard functions as in the proportional hazard model. The parametric models allow survival probabilities to be projected beyond observed follow–up. Therefore, the parametric models are more valuable, realistic and are known to be more accurate.

1.9 The structure of the thesis

The thesis has been structured as follows: In chapter two literature review has been cited for HIV/AIDS and related survival models. Chapter 3 contains statistical methodology which has been used in this study: Cox proportional hazard model, parametric models (Gompertz, Loglogistic, Lognormal, Weibull and Exponential), logrank test, Kaplan-Meier methods and Cox-Snell residuals. The results and discussion of the study are represented in chapter four. Finally, conclusion, limitation with associated recommendations are presented in chapter five.

Chapter 2

2. Literature review

The survival patterns following HIV infections in African population in the era before antiretroviral therapy form an important baseline for measuring future success of treatment programmes and the knowledge of the survival times of patients with AIDS and variables that influence survival is important both for increasing understanding of the patho-physiology of the disease, clinical decision making and planning health services intervention (Isingo et al, 2007; Jerene, 2007).The survival of patients with AIDS may depend on a variety of factors including hosts, the patterns of diseases present, access to health care, diagnostic routines and therapeutic interventions (Robert, Gilbert and Jean (1995).

The factors affecting the chance of survival/death status of HIV-positive people under the antiretroviral treatment programme were evaluated at a Dama hospital in Ethiopia, in a retrospective cohort study. The author used multiple logistic regressions for analysis and argued that the logistic is preferred from multiple regression and discriminant analysis as it results in a biologically meaningful interpretation

It is mathematically flexible and easily used distribution and it requires fewer assumptions. The model showed that condom use, alcohol, baseline weight, baseline CD4 cell count among others were found to be the factors associated with the death status of HIV infected patients (Nuredim,2007). In their study on CD4 cell count recovery among HIV infected patients with very advanced immunodeficiency commencing antiretroviral treatment in sub-Saharan Africa, separate multiple linear regression models were used to estimate the average change in CD4 cell count per month in South Africa at the Gugulethu community Health centre, Cape Town. In the analysis the separate multiple linear regression was used to examine factors associated with rates of CD4 cell counts increase per month and multiple logistic regression was used to assess factors associated with binary outcomes of a CD4 cell count increase and a study that was conducted on 122925 adult HIV-infected patients aged 15years or order from East Southern and West Africa, Asia pacific and Latin America so as to provide estimates of mortality among HIV-infected patients taking combination of antiretroviral therapy, the authors used piece wise exponential model fit through Poisson regression. The model reported that patients' mortality was high during the first 6 months after therapy for all patient groups. The following were the factors considered as predictors of mortality; age, gender, CD4 count at the initiation of therapy and at 6, 12, 24 and >24 months after the start of treatment (Lawn et al 2006; Constantin et al, 2012).

The factors associated with increased levels of self-reported quality of life among HIV-infected patients were identified using the logistic regression for the analysis of patients who were recruited at the two public health referral centres for AIDS, Belo

Horizonte, Brazil. The logistic regression indicated that >8years of education, non/mild symptoms of anxiety and depression, no antiretroviral switch, lower number of adverse reactions and better quality of life at baseline were independently associated with good/very good quality of life over four months of treatment.

A retrospective cohort study was conducted to examine the morbidity and mortality patterns of hospitalized adult HIV/AIDS at Ahmadu Bello University Teaching Hospital between June 2006 and December 2009 in northern Nigeria. In the study, the Mann-Whitney and Pearson chi-square tests were used for difference in variables by ART status. The unconditional binary logistic regression analysis was used to determine independent predictors of mortality. Out of 207 reviewed, 66(31.9%) patients died with higher mortality in males. Another examination on long-term survival in HIV-positive patients receiving cART was conducted in the Australia HIV observation Database (AHOD) to describe changes in mortality compared to the general population. The Cox proportional hazard and parametric survival models were used. The models revealed that improved survival was associated with increased recent CD4, reduced recent viral load, and younger patients. Parametric models showed a fairly constant mortality risk by year of cART up to 15years of treatment (Lorenza et al, 2009; Ogoina et al; 2012, McManus et al, 2012).

The effect of antiretroviral therapy on survival of HIV/TB infected patients in Ukraine was assessed in prospective cohort study. The Kaplan-Meier method was used to determine the survival of 80 patients and the effect of HAART on survival was evaluated using Cox proportional hazard models. From the results of the analysis, it

was found that patients with CD4 cell count<100μl had 5-fold higher risk of mortality and those with pulmonary tuberculosis with a 2 fold increased risk .Another assessment as to whether highly active ART was associated with improved survival in critically ill HIV-infected patients was conducted at the Saint–Louis teaching Hospital in Paris, France. In this study, multivariable logistic regression was used to identify risk factors for death. It was found that five factors were independently associated with increased intensive care unit mortality: delayed intensive care unit (ICU) admission, acute renal failure, hepatic cirrhosis (ICU), admission for coma and severe sepasis. (Mykhailo and Dmytro, 2013; Isaline et al, 2010).

The predictors of change in CD4 lymphocyte count and weight among HIV/AIDS infected patients taking ART were determined in Eastern Ethiopia, in a retrospective cohort study on 1540 patients from 2005 to 2010.The study used linear regression to characterize and screen data for problems of multi-collinearity and mixed model regression was used to examine changes in CD4 cell count and weight after the base baseline measurement. The results showed that median CD4 lymphocyte counts and weight improved in the follow up periods. Advanced WHO clinical stage, lower baseline CD4 cell count and baseline hemoglobin levels were found to be factors associated with decline in weight. The adherence to ART in Benin City and identification of the contributing factors in a prospective study conducted on 125 out patients were assessed in the University of Benin Teaching Hospital. The authors used logistic regression models to determine the predictors of adherence relative to socio-demographic and clinical treatment variables. The proportional data were compared using Chi-square test or fisher exact at statistical significance of 95% confidence

interval. The results from the findings showed that poor financial status, medication adverse effects, lack of confidentiality, occupational factors and stigmatization were the major reasons given for non-adherence (Ayalu et al, 2012; Era and Arute, 2008).

In a study conducted in north-eastern Vietnam on survival and causes of death among HIV/AIDS infected patients starting ART, the Kaplan-Meier analysis was used to describe survival trends and the Cox proportion hazard model was used to identify predictors. It was found that age>35years, clinical stage 3 or 4, body mass index (BMI) <18kg/hm^2, CD4 count<100µl, hemoglobin level<1100g/l and plasma viral load>100000 copies/ml were the predictive factors for AIDS-related death and tuberculosis (TB) was the most cause of death. In Malawi, a study was conducted on gender difference in retention and survival on antiretroviral therapy of HIV-infected adults. In the study the Kaplan-Meier estimates were used to analyze gender difference and rate ratios were derived from Poisson regression adjusting for confounding. It was reported that 4670 ART patients (49.8% female) were followed up for a median of 8.7 months after starting ART and the probability of death was significantly higher for men than women ($p<0.001$). Controlling for age WHO clinical stage and occupation, men experienced nearly 2 times the mortality of women, (Cuong et al, 2012; Katie et al, 2010).

The investigation on the relationship between tuberculosis infection and death in HIV/AIDS patients was conducted on 1575 subjects residing in both rural and urban areas of Yala province in Thailand and were followed between January 1992 and April 2010.Cox proportional hazard model was used to analyse the relationship and

the model reported statistically significant relationship in HIV/AIDS patients with tuberculosis and patients without tuberculosis. The HIV/AIDS patients with tuberculosis were more likely to live longer compared to those patients without tuberculosis after accounting for demographic factors. Another study to investigate the predictors of mortality in HIV associated hospitalizations in Portugal was done through a hierarchical survival model. In this study Kaplan-Meier plots were used to examine differences in survival curves. Cox proportional models with frailty were applied to identify independent predictors of hospital mortality to calculate hazard ratios .The results from the Cox proportional frailty showed that male gender, older patients, great number of diagnoses and pneumonia increased the hazard of HIV related hospital mortality. It was also found that tuberculosis was associated with a reduced risk of death. It was indicated that the frailty variance was so small but statistically significant showing hazard ratio heterogeneity among hospital that varied between 0.67 and 1.34 (Chaimay et al, 2011; Diass et al, 2009).

The prospective study aimed at estimating the short-term disease progression among HIV-infected patients was conducted in Asia and Pacific region. In this study the authors used Cox proportional hazard model to assess the predictors of disease progression and prognostic models were developed using Weibull models. It was found out in the analysis that the patients' not on treatment had higher rate of disease progression with 8.1 per 100 person-years against 17.6 per 100 person-years in the patients receiving antiretroviral treatment. The results showed that the baseline CD4 count was the strongest predictor of disease progression. The authors reported that prognostic models were successful at identifying patients at high risk of short-term

disease progression. The study was conducted on the outcome of antiretroviral treatment in rural public hospital in South Nations, Nationalities and people region in Ethiopia. The study used historical retrospective cohort study for patients visiting from January 1, 2005 to January 31, 2009.In the study the authors used Kaplan-Meier models to estimate mortality and Cox proportion hazard models to identify predictors of mortality. It was found that the hazard of death was higher in males Patients with WHO stage IV at baseline compared to WHO stage 1(Zhou and Kumarasamy, 2005;Tsegaye and Worku,2011).

The survival rate of HIV/AIDS patients after receiving free antiretroviral treatment was determined in Dehong Prefecture, Yunnan Province in a retrospective cohort analysis which was conducted on all the HIV/AIDS patients aged over 16 years who had started ART during January 2007 throughout December 29 in Dehong Prefecture. The authors used the Cox proportional hazard regression model analysis. The model indicated that after adjusting for age, sex and marital status, the baseline CD4 (+)T cells and transmission route could significantly predicate the rates of survival. It was found that those who were with baseline CD4 (+) T cells counts as 200-350/mm were less likely to die of AIDS than those with CD4 (+) T cell counts as <200/mm and HIV-infected through mother-to-child transmission or routes other than through heterosexual transmission were less likely to die of AIDS than through injecting drug use (Yao et al, 2010).

2.1 Theory for analysis of time to event data

2.1.1 Survival time distribution

Survival analysis generally refers to statistical analysis for time to event data. The outcome variable of interest is time to event, usually called failure time or survival time or life time (Kleinbaum, 1996). It is applied in a number of fields, such as medicine, public health, social science and engineering. The distribution of survival time, T, is normally described or characterized by three functions: survival function, hazard rate and probability density function of survival time, T.

2.1.2 Survival function

The survival function is the probability that the survival time is greater or equal to t (Kalbfleisch and Prentice, 2002),

$$S(t) = P(T > t), \quad \textbf{(1)}$$

$$S(t) = 1 - F(t),$$

$$F(t) = P(T \leq t) \quad \textbf{(2)}$$

The probability density function is the slope of the cdf (failure function),

$$f(t) = \lim \Delta t \to 0 \frac{p(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t} \quad \textbf{(3)}$$

Hence $\frac{\partial S(t)}{\partial t}$ is the probability of an individual dying in the interval $\frac{t, t + \Delta t}{\Delta t}$.

The survival function *S (t)* and the failure function *F (t)* are each a probability and hence have the properties of probabilities. It can be observed that, in particular, the

survivor function lies between zero and one and strictly decreasing function of t .The survivor function is equal to one at the start of the spell*(t=0)*and zero at infinity.

$$0 \le S(t) \le 1 \quad \textbf{(4)}$$

$$S(0) = 1$$

$$\lim_{t \to 0} S(t) = 0 \quad \textbf{(5)}$$

$\frac{\partial s}{\partial t} < 0$, and the density function is none-negative,

$$f(t) \ge 0 \quad \textbf{(6)}$$

2.1.3 The hazard rate

The hazard rate gives the instantaneous failure rate at time given that the individual has survived up to time *t* (Kalbfleisch and Prentice, 2002). The continuous time, hazard rate is defined as,

$$\theta(t) = \frac{f(t)}{1 - f(t)} = \frac{f(t)}{S(t)} \quad \textbf{(7)}$$

It can also be demonstrated that there is a clear relationship between hazard and survival functions given as follows:

$$\theta(t) = \frac{f(t)}{1 - F(t)} \quad \textbf{(8)}$$

$$= \frac{-\partial[1 - F(t)]}{\partial t} / [1 - F(t)] \quad \textbf{(9)}$$

$$= \frac{\partial\{-\ln[S(t)\}}{\partial t} \quad \textbf{(10)}$$

Integrating both sides gives the following:

$$\int_0^t \theta(u)du = -\ln[1-F(t)]\,|_0^t \quad \textbf{(11)}$$

Since $F(0) = 0$ and $ln(1) = 0$ then

$$\ln[1-F(t)] = \ln[S(t)] = -\int_0^t \theta(u)du$$

$$S(t) = \exp(-\int_0^t \theta(u)du$$

$$S(t) = \exp[-H(t)] \quad \textbf{(12)}$$

Then it follows that:

$$H(t) = \int_0^t \theta(u)du \quad \textbf{(13)}$$

which is cumulative hazard function,

$$=-\ln[S(t)$$

From this it can be observed that,

$$H(t) \geq 0$$

$$\frac{\partial H(t)}{\partial t} = \theta(t) \quad \textbf{(14)}$$

2.1.4 The probability density function of survival time, T

This function can be written as follows

$$f(t) = \theta(t)\exp[-\int_0^t \theta(u)du], t \geq 0 \quad \textbf{(15)}$$

The three functions outlined above are equivalent specifications of the distributions of the survival time. The survival function is useful for comparing survival progress of

two or more groups. Among the functions of the survival analysis, the hazard function provides useful description of the risk of failure at any time point.

2.2.0 Statistical methods

2.2.1 Kaplan-Meier method

The Kaplan-Meier estimator, also known as the product limit, is an estimator for estimating the survival function from lifetime data. It measures the fraction of patients living for a certain amount of time after treatment (Kaplan and Meier, 1958). A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant. The method is defined as: Let $x_i, x_{2,\ldots} x_n$ be independently identically distributed survival times having distribution function F(x) and let $G(c)$ be the distribution of independently identically distributed censoring times $c_1, c_{2,\ldots} c_n$ and c_i are assumed to be independent. Let $t_i = min\ \{x_i, c_i\}$ is the observed survival time and $\delta_i = I(x_i \leq c_i)$ indicate whether the survival time is censored or event. Let the number of individuals who are alive just before time t_i including those who are about to die at this time, be n_i and d_i denotes the number who die at this time. The Kaplan-Meier estimator is defined as:

$$\hat{S}km(t) = \frac{\prod_{i=1}^{n}((n_i - d_i))^{\delta i}}{n_i}. \qquad \textbf{(16)}$$

The assumption of the Kaplan-Meier survival function is that the distribution of censoring times is independent of exact survival times (interstat.statjournals.net, 2011).

2.2.2 Logrank method

Logrank test is used to find out whether the true survival curves differ from group to group:

H_0: No differences between survival times curves

H_1: There is a difference between survival times

It consists of observed verses expected events. For Example, letting $t_{(1)<\ldots}t_r$ be r distinct deaths times for each group. At time $_{(j)}$, let $d_{(1j)}$ and $d_{(2j)}$ be the number of deaths in group I and II respectively $_{(1j)}$ and $n_{(2j)}$ be the number of persons at risk prior to the time to time $t_{(j)}$:Then the logrank test statistic is:

$$\chi^2 = \frac{(\sum_{j=1}^{r}(d_{1j} - Ex_{ij}))^2}{\text{var}} \sim \chi^2(1) \qquad \textbf{(17)}$$

where $Ex_{ij} = \frac{n_1 d_j}{n_j}$ the mean of the hyper geometric random variable and the variance of d_{1j} $d_{1j}$are given as,

$$\text{var}(d_{ij}) = \frac{n_{1j} n_{2j} d_j (n_{ij} - d_j)}{n^2 j(n_j - 1)}$$

where,

$$\text{var} = \sum_{j=1}^{r} \text{var}(d_{1j})$$

Mantel, 1966).

2.2.3 Cox proportional hazard model

Cox proportion hazard model is a semi parametric model which is popular in survival data proposed by Cox (1972).The Cox model is defined as

$$h(t \mid x) = h_0(t)\exp\{\beta_1 x_1 + ..., \beta_k x_k\} \quad \textbf{(18),}$$

where $h(t|x)$ is the hazard function at time t for a subject with explanatory variables $X=(x_1,x_2 \ldots x_k)$, $h_0(t)$ is the baseline hazard function, that is the hazard function when all covariates equal to zero and β_1 is the regression coefficient for the *ith* covariate. The baseline can take any form (unspecified) but not negative. The Cox model assumes that the hazard functions for the two different levels of a covariate are proportional for all values of it and is given by

$$\frac{h(t_k \mid X_k)}{h(t_j \mid X_j)} = \frac{h_0(t)\exp\{\beta_1 X_k 1 + ..., \beta_p X_k p\}}{h_0(t)\exp(\beta_1 X_j + ..., \beta_p X_j p}$$

$$= \exp\{\beta_1(X_k 1 - X_j) + \beta_p(X_k p - X_j p) \quad \textbf{(19)}$$

hence $h_0(t)$ cancels out and this means that the ratio is the same at all-time points.

$x_1 = age, x_2 = bmi, x_3 = weight, x_4 = whostage, x_5 = occupation, x_6 = sex, x_7 = cd4 \le 250$, $x_8 = regimen$

2.2.4 Assumption of the Cox model

The Cox model makes the following assumptions:

a) The structure of the model is assumed correct. That is for example, model is multiplicative and all relevant covariates have been met.

b) The continuous covariates have a linear form

c) The proportional hazard assumptions are satisfied.

2.2.5 Cox model popularity

The Cox model has the following key properties:

a) It is robust hence a safe choice of model in many situations.

b) Estimated hazard are always non-negative

c) Even though $h_0(t)$ is unspecified, $\beta_i's$ can be estimated and thus compute the hazard ratio.

d) The $h_0(t|x)$ and $S(t|x)$ can be estimated for a Cox model using a minimum of assumption.

The Cox proportional hazard model can fit by maximizing the likelihood function and this procedure estimates the $h_0(t)$ and β.The popular approach is proposed by Cox (1975) in which a partial likelihood also called Cox likelihood function that doesn't rely on $h_0(t)$ is realized for β. The partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameter $h_0(t)$ in the Cox PH model. Assume K different failure times $t_{(1)}, t_{(2)}...t_{(k)}$ such that there is exactly one failure at each $t_{(i)}$, $i=1,...,k$ .Let $[i]$ denote the subject with an event time $t_{(i)}$ and $R(t)$ the risk set at time t, then the partial likelihood is given as

$$L(\beta) = \frac{\prod_{j=1}^{k} \exp\left(\sum_{i=1}^{p} \beta_i X[j]i\right)}{\sum_{i \in R(t|j)} \exp\left(\sum_{i=1}^{p} (\beta_i Xji)\right)} \qquad \textbf{(20)}$$

The likelihood considers probabilities for subjects who fail and doesn't consider probabilities for censored subject explicitly. The censored subjects are taken into account in the risk set. The estimates of βs is denoted by $\hat{\beta}_i s$ .The $\hat{\beta}_i s$ solves

$$\frac{\partial \log L}{\partial \beta_i} = 0, i = 1,...p. \quad \textbf{(21)}$$

Therefore $\hat{\beta}_i s$ maximize the Cox likelihood.

2.2.6 Proportional hazard assumption checking

The proportional hazard is the core assumption of the Cox model. There are a number of procedures for ensuring that a model satisfies the assumption of proportionality before the model results can be safely applied(Brant,2004).The proportional hazard means that the survival for two subjects have hazard functions that are proportional overtime(constant relative) (Brant,2004).

2.2.7 Graphical method

The Cox PH survival function can be obtained by the relationship between hazard function and survival function,

$$S(t \mid X) = S_0(t)\exp(\textstyle\sum_{i=1}^{k} \beta_k X_k) \quad \textbf{(22)}$$

Where X=(x_1, $x2…x_k$) is the value of the vector of predictor variables for a particular individual. Taking the logarithm twice, we have

$$\ln[-\ln s(t \mid X)] = \sum_{i=1}^{k} \beta_k X_k + \ln[-\ln S_0(t)$$

It can be noted that the difference in log-log curves corresponding to two different individuals with variables X_1=(x_{11}, $x_{12}…x_{1k)}$ and X_2=(x_{21},$x_{22}…x_{2k}$) is given by

$$\ln[-\ln S(t \mid X_1)] - \ln S(t \mid X_2)] = \sum_{i=1}^{k} \beta_i (X_{1i} - X_{2i}) \quad \textbf{(23)}$$

which doesn't rely on *t*. By plotting estimated log (-log (survival) versus survival time for the two groups parallel curves would be realized if the hazards are proportional. However, this method doesn't work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Moreover, the curves are sparse when there are few time points and it may be difficult to gauge how close to parallel is close enough.

2.2.8 Adding time-dependent covariates in the Cox model

This is done by creating interactions of the predictors and a function of survival time. Let X_i be the predictor of interest, and creating $X_{j(t)}$ as a time –dependent covariate, then

$$X_{j(t)=}\mathrm{X_i} * g_{(t)} \qquad \textbf{(24)}$$

where $g_{(t)}$ is a function of time. The model assessing PH assumption for X_j adjusted for other covariates is

$$h(t,(t)) = h_0(t)\exp[\beta_1 X_1 + \beta_2 X_2 + ...\beta_j X_j + \beta_p X_p + \delta X_j * g(t)] \qquad \textbf{(25),}$$

where $X(t) = (x_1 x_2 ..., x_{p,} x_j(t))$ is the value of predictor variables for a particular individual. The null hypothesis, to check proportionality is that δ=0.The test statistic can be done using either a Wald test or a likelihood ratio test. These statistics have chi-square distribution with one degree of freedom under the null hypothesis. If the time –dependent covariate is significant, then the predictor is not proportional.

2.2.9 Deviance residuals

The deviance residuals (Therneau, Grambsch and Fleming, 1990) is defined by

$$rD_i = sign(rx_i[-2\{rx_i + \delta_i \log(\delta_i - rx_i\}]^{\frac{1}{2}} \quad \textbf{(26)}$$

Where the function *sign*(.) is the sign function which takes the values 1 if rx_i is positive and -1 if rx_i if negative;

$$rx_i = \delta_i - rc_i$$

is the martingale residuals for the i^{th} individual; and δ_i=1 for uncensored observation, and δ_i=0 for censored observation. The deviance residuals are normalized transformation of the martingale residuals (Therneau, Grambsch and Fleming, 1990).They have a mean zero but are approximately symmetrically distributed about zero when fitted model is appropriate. Very large or small value can indicate potential outliers.

2.3.0 Schoenfeld Residuals

Schoenfeld residuals are computed with one per observation per covariate. It only defined at observed event times for the i^{th} subject and k^{th} covariate, the estimated Schoenfeld residual, r_{ik}, is given by

$$\hat{r}_{ik} = X_{ik} - \hat{\bar{X}}_{wik}$$

where X_{ik} is the value of the K^{th} covariate for individual *I* and $\hat{\bar{X}}_{wik}$ is the weighted mean of covariate values for those in the risk set at the given event time. Positive value of risk shows X value that is higher than expected at that death time. The Schoenfeld residuals sum to zero.

2.3.1 Parametric survival models

Parametric survival models are regression models in which the distribution of the response is chosen to be consistent with what one would see if the response is time-to- failure (Gutierrez**,** 2002).The parametric models are fitted to the survival data using maximum likelihood method, the procedure is described as follows:

Suggesting that the survival times $t_1, t_2 \ldots t_n$ are observed and q of the n individuals die at times $t_{(1)} < t_{(2)} \ldots t_{(q}$ and that the survival times of the remaining n-q(q<n) individuals are censored. If $f(t)$ denotes the probability density function of the survival time t and $S(t)$ be the survival function, then, the likelihood is given by,

$$\prod_{i=1}^{n} \{f(t_i)\}^{ci} \{S(t_i)\}^{1-ci} \qquad \textbf{(27)}$$

where, c is an indicator variable, taking value 0 when the survival time is censored and 1 for the uncensored survival time. Five parametric survival models: Weibull, Exponential, Gompertz, Log-logistic, Lognormal have been considered in the study.

2.3.2 Weibull model

The Weibull distribution is the generalized version of the exponential distribution.

It is preferred for performing survival data analysis in industrial engineering (Weibull, 1951).However, when implementations in the discipline of medicines are examined, one may see that it is an important distribution model. It is a flexible distribution that allows monotonous increasing and decreasing of mortality ratio in patients groups. In a study carried by Viscomi et al (2006) the distribution of the survival period of childhood leukemia patients was analyzed using the Weibull distribution. In a study conducted in Italy on the national wide estimations of the cancer patients, some

estimations were made for defining the parameters of Weibull distribution. The Weibull distribution has the following functions:

$$h(t \mid x_i) = \lambda p(t)^{p-1} \quad \textbf{(28)}$$

where $\lambda = \exp(x_i \beta)$ is the hazard function. The survival function is given by

$$S(t) = \exp(-\lambda t) p \quad \textbf{(29)}$$

and the Probability density function is defined as

$$f(t) = \lambda p(t)^{p-1} \exp(-\lambda t) p \quad \textbf{(30)}$$

In weibull model the hazard function for a person with explanatory variables (x_1,x_2, …,x_p) is given by:

$$h(t \mid x) = \lambda p(t)^{p-1} \exp(\beta_1 x_1 + \beta_2 x_2 + ... \beta_p x_p) = \lambda p(t)^{p-1} \exp(\beta' x) \quad \textbf{(31)}$$

The Weibull is a two parameter model (λ and p) where λ is the location parameter and p is the shape parameter since it determines whether the hazard is increasing p>1 or decreasing p<1 or constant p=1 overtime. When considering the applicability of Weibull distribution the validity of monotonic hazard must be followed.

2.3. 3 Exponential model

The exponential model is the special case of the Gamma distribution and is used in cancer survival model. It was once used by Dewals and Bouckaert (1985) on carrier bacteria. The exponential model is characterized by the constant hazard rate .Thus it has one parameter denoted by λ. This implies that the conditional probability of an event is constant overtime. If the exponential distribution is to be used it is important to consider whether the hazard is likely to remain constant over an entire lifetime. The exponential functions are,

$$h(t \mid x) = \lambda \quad \textbf{(32)}$$

is the hazard rate, and the hazard function is given by,

$$h(t) = \lambda t, \lambda > 0, t > 0 \quad \textbf{(33)}$$

The survival function is defined as

$$S(t) = \exp(-\lambda t) \quad \textbf{(34)}$$

The probability density function for the exponential regression model is given by

$$f(t) = \lambda \exp(-\lambda t) \quad \textbf{(35)}$$

The exponential model assumes that the baseline hazard is constant (Lawless, 2003). Therefore the hazard is given by,

$$h(t \mid x_i) = \exp((\beta_0 + \beta_1 x))t \quad \textbf{(36)}$$

The survival function is given by,

$$S(t \mid x_i) = \exp\{-\exp(\beta_0 + x_i \beta x)t \quad \textbf{(37)}$$

Then the hazard function for a particular person with the explanatory variables ($x_1 x_{2\ldots} x_p$) is given by:

$$h(t \mid x) = \lambda \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p) = \lambda \exp(\beta' x) \quad \textbf{(38)}$$

2.3.4 Gompertz model

Gompertz model is used frequently by medical researchers and biologists in modeling mortality ratio data. The model was formulated by Gompertz (1825).It has these functions:

Hazard function:

$$h(t) = \lambda \exp(\gamma t) \quad \textbf{(39)}$$

for $0 \le t < \infty$, where λ is positive value and is the scale parameter and Θ is the shape parameter. When $\gamma = 0$ survival times have an exponential distribution, where γ>0 the hazard increases monotonically with time and when γ<0 the hazard decreases with time.

$$h_0(t) = \exp(\gamma t)\exp(\beta_0) \quad \textbf{(40)}$$

The model now becomes:

$$h(t \mid x_i) = h_0(t)\exp(x_i \beta x)$$

$$= \exp(\gamma t)\exp((\beta_0 + x_i \beta x)) \quad \textbf{(41)}$$

The survival function is given by

$$S(t) = \exp(\frac{\lambda}{\gamma}(1 - \exp(\gamma t)) \quad \textbf{(42)}$$

The distribution is characterized by the fact that the log of hazard is linear in it. The hazard function for a particular person using the Gompertz distribution is given by:

$$h(t \mid x) = \lambda \exp(\gamma t)\exp(\beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p) = \lambda \exp(\beta' x)\exp(\gamma t) \quad \textbf{(43)}$$

2.3.5 Log-logistic model

The log logistic distribution is continuous for the random variable which is not negative in probability and statistics. The mortality ratio in a life analysis slowly decreases after it reaches to the maximum point over a finite period and it is suitable to use anon monotonic failure rate distribution model on the life and lost (Viscomi et

al, 2006). In case of censored data, using log logistics distribution is mathematically more advantages than other distributions (Hayati, 2010).

A study conducted by Byers et al(1988) on the spreading ratio of HIV virus in San Francisco between 1978 and 1986 indicated that log logistic was the most suitable model among many distributions to use with half censored data. Zhou et al (2007) conducted study in which he emphasized that the maximum likelihood estimation was the most suitable method in estimating the parameters when performing analyses using loglogistic distribution on grouped data such as half censored data.

These are the distributions functions for the log logistic:

$$h(t) = \frac{\lambda p t^{p-1}}{1+\lambda t^{p}} \qquad \textbf{(44)}$$

the hazard function, *h(t*), increases then decreases if *p*>1, monotonically, decreases when *p*=1, λ gives information on the covariate**,**

$\lambda_i = \exp(x_i \beta)$

and the following function,

$$S\,(t) = \frac{\lambda t^{p}}{1+\lambda t^{p}} \qquad \textbf{(45)}$$

is the survival function.The Accelerated Failure time for the loglogistic regression:

$$\lambda t_i = \exp(-x_i \beta x) t_i \qquad \textbf{(46)}$$

$t_i \sim$Loglogistic (β_0, γ)

This has the cumulative distribution function:

$$F(t) = 1 - [1 + \{\exp(-\beta_0) t\}^{\frac{1}{\gamma}}]^{-1} \qquad \textbf{(47)}$$

$$\ln(t_i) = x_i\beta x + \ln(\tau_i)$$

$$= \beta_0 + x_i\beta x + \mu_i \quad \textbf{(48)}$$

μ_i follows a logistic distribution with mean 0 and the standard deviation, $\frac{\pi\gamma}{\sqrt{3}}$.This follows**:**

$$E\{\ln(t_i) \mid x_i\} = [\beta_0 + x_i\beta x] \quad \textbf{(49)}$$

The base line survivor functions of t_i is given by,

$$S_0(t) = [1 + [1 + \{\exp(-\beta_0)t_i\}^{\frac{1}{\gamma}}]^{-1} \quad \textbf{(50)}$$

Hence the effect of the covariates is to accelerate time by the factor of $\exp(-x_i\beta x)$.

Then the Accelerated Failure time model is given by

$$S(t_i \mid x_i) = S_0\{\exp(-x_i\beta x)t_i \quad \textbf{(51)}$$

$$= [1 + \{\exp(-\beta_0)\exp(-x_i\beta x)t_i\}^{\frac{1}{\gamma}}]^{-1}$$

$$= [1 + \{\exp(-\beta_0 - x_i\beta x)t_i\}^{\frac{1}{\gamma}}]^{-1} \quad \textbf{(52)}$$

2.3.6 Lognormal

The lognormal is a skewed distribution where the average values are low, variances are high and the values are not negative. The theory of the lognormal distribution was characterized by McAlister (1897) and there is accordance to the lognormal distribution in many examples in the area of medicines. The history of lognormal distribution, its features, estimation problems and its use in economics were examined in detail in 1957. The survival distributions of Hodgkin's disease, chronic leukemia

were analyzed via lognormal distribution, which is positively skewed and with survival period distributed normally (Lee and Wang, 2003). Feinleib and McMahon (1974), in a study conducted on chronic lymphocyte and myelocythic leukemia patients, applied lognormal distribution. In the lognormal, the hazard function increases from 0 to reach maximum and then decreases monotonically approaching 0 as $t \to \infty$.The survival function is given by:

$$h(t) = 1 - \Phi\{\{\frac{\ln(t) - \mu}{\sigma}\}\} \quad \textbf{(53)}$$

where Φ is the standard normal cumulative density function and μ=*x*β.The hazard rate is given by:

$$\frac{1}{t\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}\{\ln(t) - \mu^2\}/(1 - \Phi\{\frac{\ln(t) - \mu}{\sigma}) \quad \textbf{(54)}$$

the hazard rate rises first then falls. The density function is given by:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}\{\ln(t) - \mu^2\}] \quad \textbf{(55)}$$

However the lognormal has no proportional hazard interpretation. Hence, its interpretation is in the AFT metric (Cleves, 2010).It assumes that t_i ~lognormal $(\beta_{0,\sigma})$ and it has the cumulative distribution as given by:

$$F(t) = \Phi(\frac{\ln - \beta_0}{\sigma}) \quad \textbf{(56)}$$

is the cumulative distribution function for the standard Gaussian(normal) distribution, hence

$$\ln(t_i) = x_i \beta x + \ln(\tau_i) \quad \textbf{(57)}$$

The lognormal model transforms time into ln(*time*) and converts the problem into simple linear regression:

$$E\{\ln(t_i) \mid x_i\} = \beta_0 + x_i \beta x. \quad \textbf{(58)}$$

The baseline survivor function is realized as:

$$S_0(t) = 1 - \Phi\{\frac{\ln(t) - \beta_0}{\sigma}\} \quad \textbf{(59)}$$

2.3.7 Accelerate failure time model

The Accelerated Failure Time model is a linear regression model in which the response variable is the logarithm or known monotone transformation of a failure time (Lee and Wang, 2003).The accelerated failure time model describes a relationship between the survivor function of any two individuals. Taking T_i to be a random variable denoting the failure time for the *ith* subject, and let $X_{i1}, X_{12}\ldots, X_ip$ be the values of p covariates of the subject. The model is then given by

$$\log T = \{\beta_0 + \beta_1 X_i 1 + \ldots + \beta_p X_i p + \delta\varepsilon_i\} \quad \textbf{(60)}$$

Where $\varepsilon_i \sim h_0(t), \varepsilon_i$ is a random disturbance term, $\beta_0, \ldots, \beta_p$ and σ are parameters to be estimated, $h_0(t)$ is a known baseline survival, T_i is actual survival times sometimes observed is a scale parameter and X_i is a fixed P*1 vector of covariates and X_i is assumed to affect *logT* linearly and no interactions. Moreover, is σ assumed to be constant and independent of X_i. The parametric accelerated failure time distribution is also assumed to be correctly specified. The survival function for the parametric baseline accelerated failure model is given as,

$$S(t_i) = S_0\{\exp(-X^t i\beta)t_i \quad \textbf{(61)}$$

where $S_0(t)$ is the baseline. These accelerated failure time models are named for the distribution of T rather than the distribution of ε or *logT*. This is so because different distributions have different implications for the shapes or hazard function (Cox and

Oakes D, 1996). It has been noted that the proportional hazard model is used exclusively in practice. However; the accelerated failure time model in many ways is more appealing because it is quite direct physical interpretation especially when the response variable does not pertain to failure time (Reid, 1994).

2.3.8 The fitness of the model

When the model has been fitted the adequacy of it needs to be assessed. There are a number of ways to check the adequacy, like using cox-Snell, deviance among others.

2.3.9 Cox –Snell residuals

The Cox-Snell residuals is given by Cox and Snell (Cox and Oakes,1984).The residuals for the ith individual with the observed survival time t_i is given as follows,

$$r_{ci} = \exp(\beta X_i) H_0(t_i) = H_t(t) = -\log \hat{S}_i(t_i) \quad \textbf{(62)}$$

Given that $H_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i and it was derived by Kalbfleish and Pretence (1973).

Letting T be the continuous survival distribution S (t) with the cumulative hazard,

$$H(t) = -\log(S(t)) \quad \textbf{(63)}$$

Then it follows thus,

$$S_T(t) = \exp(-H(t))$$

Taking $Y = H(t)$ be the transformation of T based on cumulative hazard function. It follows that the survival function for Y is now given as

$$S_Y = p(Y > y) = p\{H(t) > y\} \quad \textbf{(64)}$$

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

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

The new $Y = H(t)$ has an exponential distribution with unit one. If the model was well fitted, the actual value $\hat{S}_i(t_i)$ would have similar properties to those of $S_i(t_i)$ Therefore $r_{ci} = \log\hat{S}(t_t)$ will have a unit exponential distribution with $f_R(r) = \exp(-r)$. Letting $S_R(r)$ denote the survival function of Cox-Snell residuals r_{ci}, then

$$S_R = \int_r^{\infty} f_R \exp(-x)dx = \exp(-r) \quad \textbf{(65)}$$

and it follows that, $H_R(r) = -\log S_R(r) = -\log(\exp)(-r) = r$ **(66)**

Hence a plot of $H(r_{ci})$ versus r_{ci} is used to check the fit of the model. This gives a straight line with a unit slope and zero intercept if the fitted model is correct.

2.4.0 Model checking

There are a number of methods which are employed to check if a parametric distribution fits the observed data. In this study the Akaike information Criterion (AIC), a statistical criteria, used for comparing models and residuals plots used to check the goodness of fit of the model have been used.

2.4.1 Akaike information criterion (AIC)

The Akaike information criterion (AIC) proposed in AKaike (1974) is a measure of the goodness of fit of an estimated statistical model which compares the models which have been fitted. The AIC is an operation way of trading off the complex of an estimated model against how well the model fits the data. The AIC is defined by

$$AIC = -2l + 2(k + c) \quad \textbf{(67)}$$

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 $2(k+c)$ is thought as penalty if non-predictive parameters are added to the model. Lower values of the AIC suggest a better model. However, there is a difficulty in using AIC in the sense that there is no formal test statistically to compare different AIC values when two or more models have similar AIC values. Hence, the choice of the model may be hard to determine.

2.4.2 Residual plots

The residual plots can be used to check the goodness of fit of the model. Among the useful plots is based on comparing the distribution of the cox-Snell residuals with the unit exponential distribution. The cox-Snell residual for ith individual with observed time, $t_i$is defined as:

$$r_{ci} = \hat{H}(t_i \mid x_i) = -\log[\hat{S}(t_i)x_i] \qquad \textbf{(68)}$$

where t_i is the observed survival time for individual i, x_i is the vector covariate values for individual i, and $\hat{S}_i(t_i)$ is the estimated survival function on the fitted model. The estimated survival function for the ith individual is given by

$$\hat{S}_i(t_i) = S_{\varepsilon i}(\frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}}) \qquad \textbf{(69)}$$

where $\hat{\alpha}, \hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimator of α , μ and σ respectively, $S_{\varepsilon i}(\varepsilon)$ is the survival function of εi in the AFT model and

$$(\frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}}) = r_{\varepsilon i} \qquad \textbf{(70)}$$

is referred to as standard residuals.

Chapter 3

3 Research design & methodology

3.1 Research design

The data used in this project came from cohort studies that were done at Ntcheu District hospital. All the records with a total of 6670 patients were studied. The decision to analyse all the records was arrived at after seeing that there was a lot of missing information in most of the variables considered. Hence, the study would just left with fewer subjects had the study just used sample and the results would not be true.

3.2 Data collection

A total of 6670 patients were followed from 2007 to 2012. This data was collected from the medical records (master cards) of patients who were registered for ART from January 1, 2007-December31; 2012 at Ntcheu district hospital which was electronically stored into touch screen computers by Baobab health. The data was collected from the patients' records by two clerical statisticians, one surveillance, two data entry clerics and the HIV/AIDS counselor together with the ART clinical nurse. These included patients aged 15 years and above. The clinical nurse assisted in tracking down the quarterly data from ART supervision monitoring checklist from 1

January, 2007 to 31 December, 2012 to ensure that the data taken reflects what had been stored electronically. A permission was obtained from the Ministry of Health through the District Health Officer to have an access to the data required. The collected data was cleaned and validated to make sure that the data was of reasonable quality before conducting the analysis.

3.3 Response variable (dependent variable)

The response variable in this study is the survival time from the initiating ART till outcome. The outcome of interest in this study is the time it takes a patient on ART to die keeping all other variables constant.

3.4 Predictor variables (independent variables)

The following variables were considered as covariates in the statistical models: age in years, body mass index (bmi), occupation, World Health immunological staging (Who stage), CD4<=250 and sex. Age in this study has been considered looking at the role it plays in relation to HIV/AIDS patients. For instance age determines the type of regimen to be given to the patients. The regime is typical to adults or children depending on the conditions. It also helps to know when one did start ART. Furthermore, age helps clinicians to provide proper and constructive advice to the guardians of the HIV/AIDS infected patients on ART on how they can take care of the patients. Body mass index has been considered in the study as it helps to determine as to whether patient is having nutritious diet or not. Body mass index is associated with immunity of the body. The Cd4 has been included in the study due to the fact that for the HIV/AIDS patients to be initiated on ART there is need to check

cd4 cell count. In addition to that it helps the clinicians to check if there are any increase in cd4 cell count after the patient has been initiated on ART, thus to find out if there exist any positive impact of the treatment received by the HIV/AIDS patients. The WHO clinical staging (3&4) has been used in the study as predictor variable since it helps to determine the time an HIV/AIDS patient has to be initiated on ART. For example those patients who might be in WHO clinical stages 1 and 2 are not initiated on ART. However, at other times patient can be initiated on ART while in WHO clinical stage 1 or if the cd4 cell count is less than the threshold. Another predictor which is occupation helps the clinician to suggest the risk behavior depending on the type of occupation. It helps to have a clue to what led them to contrite the HIV. And this helps the patient to be given the appropriate advice. At other situations the type of work could be that of heavy duty and this can shorten the life of the patient, so there is need to be given advice basing on the type of occupation. Another example on occupation is that of a farmer in that if the type of crop grown is tobacco can make the patient to have higher risk of being infected with tuberculosis which can lead to higher mortality. HIV-related diseases are grouped into four WHO clinical stages that correlate with disease progression and prognosis of survival:

a. Stage 1: Asymptomatic
b. Stage 2: Mild (moderate weight loss<10% unexplained, seborrhoeic dermatitis)
c. Stage 3: Advanced(severe weight loss>10%) and /or BMI<18.5kg/m^2 unexplained, diarrhea, chronic(>1month) unexplained, oral candidiasis, severe,bacterial,infections(pneumonia,empyema,pyomyositis,bone/joint,m

eningitis,bacteraemia) acute necrotizing ulcerative stomatitis, gingivitis or periodontitis, hepatitis B or C infection.

d. Stage 4:Severe(HIV wasting syndrome(severe weight loss+persistent fever or severe weight loss+chronic diarrhea),bacterial pneumonia, recurrent severe, chronic herpes simplex infection, cytomegalovirus infection, toxoplasmosis of the brain, non-typhoidal salmonella bacteraemia, recurrent, invasive cancer of cervix, leishmaniasis (Ministry of Health,Malawi,2011).

Table 1 shows the variable description and categorization.

Table 1 : Variable description

VARIABLE	**CODING**	**DESCRIPTION**
Gender	1=female	categorical
	2=male	
Age (in years of the patient)		continuous
WHO	Stage 3=1	categorical
	Stage 4=2	
Cd4<=250 cells per μl	Cd4=1	1=initiated at <=250
	Cd4=0	0=otherwise
occupation	0=non worker	categorical
	2=worker	
Body mass index(BMI) in kg/cm^2		continuous
Weight of the patient in kilogram		continuous

3.5 Statistical methods

3.5.1 The Kaplan-Meier method

The Kaplan-Meier estimator, also known as the product limit, is an estimator for estimating the survival function from lifetime data. In this study it has been used to assess the survival probability overtime after initiation of ART comparing the survival among the patients, since this method measures the fraction of patients living for a certain amount of time after treatment (Kaplan and Meier, 1958). A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant. The method is defined as: Let x_i, x_2,... x_n be independently identically distributed survival times having distribution function F(x) and let G(c) be the distribution of independently identically distributed censoring times c_1 $c_{2,\ldots}c_n$ and c_i are assumed to be independent times $c_{1,}c_{2,\ldots}c_n$ and c_i.Let $t_i = min\ \{x_i, c_i\}$ is the observed survival time and $\delta_i = I(x_i \leq c_i)$ indicate whether the survival time is censored or event. Let the number of individuals who are alive just before time t_i including those who are about to die at this time, be n_i and d_i denotes the number who die at this time (equation:16). The assumption of the Kaplan-Meier survival function is that the distribution of censoring times is independent of exact survival times (interstat.statjournals.net, 2011).

3.5.2 Logrank method

Logrank test has been used to find out whether the true survival curves differ from group to group:

H_0: No differences between survival times curves

H_1: There is a difference between survival times curves

It consists of observed verses expected events. For Example, letting $t_{(1)<\ldots}t_r$ be r distinct deaths times for each group. At time$_{(j)}$,let $d_{(1j)}$ and $d_{(2j)}$ be the number of deaths in group I and II respectively, $n_{(1j)}$ and $n_{(2j)}$ be the number of persons at risk prior to the time to time $t_{(j)}$ (equation:17).

3.5.3 Cox proportional hazard model

Cox proportion hazard model is a semi parametric model which is popular in survival data proposed by Cox (1972).It has been used to determine the difference of survival time (in years) in relation to sex, WHO clinical stage 4, age, cd4, occupation and regimen. The Cox model assumes that the hazard functions for the two different levels of a covariate are proportional for all values of it (equation: 19). In this study,

$$x_1 = age, x_2 = bmi, x_3 = whostage, x_4 = occupation, x_5 = sex, x_6 = cd4 \leq 250, x_7 = regimen$$

3.5.4 Assumption of the Cox model

The Cox model makes the following assumptions:

a) The structure of the model is assumed correct. That is for example, model is multiplicative and all relevant covariates have been met.
b) The continuous covariates have a linear form
c) The proportional hazard assumptions are satisfied.

3.5.5 Cox model popularity

The Cox model has the following key properties:

a) It is robust hence a safe choice of model in many situations.

b) Estimated hazard are always non-negative

c) Even though h_0 (t) is unspecified, $\beta_i's$ can be estimated and thus compute the hazard ratio.

d) The $h_0(t|x)$ and $S(t|x)$ can be estimated for a Cox model using a minimum of assumption.

The Cox proportional hazard model is fit by maximizing the likelihood function (equation: 20) and this procedure estimates the $h_0(t)$ and β. This popular approach was proposed by Cox (1975) in which a partial likelihood also called Cox likelihood function that doesn't rely on $h_0(t)$ is realized for β.The partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameter $h_0(t)$ in the Cox PH model. Assume K different failure times $t_{(1)}$, $t_{(2)}$...$t_{(k)}$ such that there is exactly one failure at each $t_{(i)}$, i=1,...,k.(section:2.2.5). The likelihood considers probabilities for subjects who fail and doesn't consider probabilities for censored subject explicitly. The censored subjects are taken into account in the risk set. Therefore $\hat{\beta}_i s$ maximize the Cox likelihood (equation: 21).

3.5.6 Proportional hazard assumption checking

The proportional hazard is the core assumption of the Cox model. There are a number of procedures for ensuring that a model satisfies the assumption of proportionality before the model results can be safely applied (Brant, 2004).The proportional hazard means that the survival for two subjects have hazard functions that are proportional overtime (constant relative) (Brant, 2004).

3.5.7 Graphical method

The Cox PH survival function can be obtained by the relationship between hazard function and survival function (equation: 22), where $X=(x_1,x_2\ldots x_k)$ is the value of the vector of predictor variables for a particular individual. The logarithm is taken twice, and it can be noted that the difference in log-log curves corresponding to two different individuals with variables $X_1=(x_{11},x_{12}\ldots x_{1k})$ and $X_2=(x_{21},x_{22},\ldots,x_{2k})$ is given as in equation 23. By plotting estimated log (-log (survival) versus survival time for the two groups parallel curves would be realized if the hazards are proportional. However, this method doesn't work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Moreover, the curves are sparse when there are few time points and it may be difficult to gauge how close to parallel is close enough.

3.5.8 Adding time-dependent covariates in the Cox model

The study has used this approach which is done by creating interactions of the predictors and a function of survival time (equation: 24). The model assesses PH assumption adjusted for other covariates (equation: 25). The null hypothesis to check proportionality is that $\delta=0$.The test statistic can be done using either a Wald test or a likelihood ratio test. These statistics have chi-square distribution with one degree of freedom under the null hypothesis. If the time –dependent covariate is significant, then the predictor is not proportional.

3.5.9 Schoenfeld Residuals

Schoenfeld Residuals are computed with one per observation per covariate. It detects if some form of time dependency in a particular covariate. The residual plot has two bands, one above zero for X=1 and one below 0 for X=0.It is only defined at observed event times for the i^{th} subject and k^{th} covariate, the estimated Schoenfeld residual, r_{ik}, is given as in section 2.3.0. A horizontal line shows the coefficient is constant and the proportion assumption is valid.

3.6.0 Parametric survival models

Parametric survival models are regression models in which the distribution of the response is chosen to be consistent with what one would see if the response is time-to- failure (Gutierrez, 2002). The present study has used parametric survival models as they measure follow-up time from a defined starting point to the occurrence of a given event, for example the time from diagnosis of a disease to death. In this dataset for the survival HIV/AIDS infected patients; the standard statistical techniques cannot usually be applied because the underlying distribution is rarely Normal and the data are often censored (Bewick et al, 2004).Parametric regression models assume particular families of probability distributions such as exponential, weibull, Gompertz, lognormal, log-logistic or gamma hence provide a complete probability specification for the data validation for the model (Allison, 2012),and they yield more precise estimates which are being slightly more efficient (Bradbum et al, 2003).The parametric survival analysis relates to portability and ease of manipulating a mathematical equation to understanding underlying phenomena, assess the impact of a risk factor and facilitate strategic medical decision making, particularly in

examining the hazard function, one can discover easily what is difficult non parametrically (Blackson, 2012). The parametric models explicitly model the function form of the event times using various statistical distribution. Generally parametric survival models involve two parameters, scale parameter and shape parameter. The shape parameter generally assumes constant across individuals and the scale relate to determinants via regression which quantifies the effect of predictors, particularly treatment. Parametric survival models can readily predict survival in future groups of similar patients as each patient can be given personalized survival curve as a solution of the mathematical equation for the patient characteristics (Blackson, 2012).However, this study has only selected five parametric models: Exponential, Weibull, Gompertz, lognormal and loglogistic. The study has used Akaike information criterion, statistical criteria, to check if these parametric distribution fits the observed data. The residual plots have also been used to check for the goodness of fit of the model used. The parametric models are fitted to the survival data using maximum likelihood method, the procedure is described as in section 2.3.1.

3.6.1 Weibull model

The Weibull distribution is the generalized version of the exponential distribution. It is preferred for performing survival data analysis in industrial engineering (Weibull, 1951). However, when implementations in the discipline of medicines are examined, one may see that it is an important distribution model. It is a flexible distribution that allows monotonous increasing and decreasing of mortality ratio in patients groups. The convenience of the Weibull model for empirical work stems on the one hand from this flexibility and on the other hand on the other form, the simplicity of the

hazard and survival function. In a study carried by Viscomi et al (2006) the distribution of the survival period of childhood leukemia patients was analyzed using the Weibull distribution. In a study conducted in Italy on the national wide estimations of the cancer patients, some estimations were made for defining the parameters of Weibull distribution. The Weibull distribution has been given as in equations 28-31.

3.6.2 Exponential model

The exponential model is the special case of the Gamma distribution and is used in cancer survival model. It was once used by Dewals and Bouckaert (1985) on carrier bacteria. The exponential model is characterized by the constant hazard rate .Thus it has one parameter denoted by λ. This implies that the conditional probability of an event is constant overtime. If the exponential distribution is to be used it is important to consider whether the hazard is likely to remain constant (flat) over an entire lifetime (equation: 32) (flat, which implies that the conditional probability of the event is the same, no matter when the observation is observed), The event occurs according to a Poisson process ("memory less").The lack of flat hazards means covariates are missing and there is need to use other models.

3.6.3 Gompertz model

Gompertz model is used frequently by medical researchers and biologists in modeling mortality ratio data. The Gompertz model has been considered in this study because it turns into the straight hump-shaped line within logarithmic coordinate. The model

was formulated by Gompertz (1825). It has functions as indicated and explained in section 2.3.4(equations: 39-43).

3.6.4 Log-logistic model

The log logistic distribution is continuous for the random variable which is not negative in probability and statistics. The log-logistic distribution has a fairly flexible form, it is one of the parametric survival time models in which the hazard may be decreasing, and increasing as well as hump-shaped. It is strictly accelerated failure time models in that it begins with the log –linear model. The mortality ratio in a life analysis slowly decreases after it reaches to the maximum point over a finite period and it is suitable to use anon monotonic failure rate distribution model on the life and lost (Viscomi et al, 2006). In case of censored data, using log logistics distribution is mathematically more advantages than other distributions (Hayati, 2010).

A study conducted by Byers et al (1988) on the spreading ratio of HIV virus in San Francisco between 1978 and 1986 indicated that log logistic was the most suitable model among many distributions to use with half censored data. Zhou et al (2007) conducted study in which he emphasized that the maximum likelihood estimation was the most suitable method in estimating the parameters when performing analyses using log logistic distribution on grouped data such as half censored data. The distributions functions for the log logistic has been given in section 2.3.5.

3.6.5 Lognormal

The lognormal is a skewed distribution where the average values are low, variances are high and the values are not negative. The theory of the lognormal distribution was characterized by McAlister (1897) and there is accordance to the lognormal distribution in many examples in the area of medicine. The history of lognormal distribution, its features, estimation problems and its use in economics were examined in detail in 1957. The survival distributions of Hodgkin's disease, chronic leukemia were analyzed via lognormal distribution, which is positively skewed and with survival period distributed normally (Lee and Wang, 2003). Feinleib and McMahon (1974), in a study conducted on chronic lymphocyte and myelocythic leukemia patients, applied lognormal distribution. In the lognormal, the hazard function increases from 0 to reach maximum and then decreases monotonically approaching 0 as $t \to \infty$,therefore it has been considered in this study because of the very reason that its base line hazard has a value of 0 at t=0 and becomes large when approaching 0. The survival function and other functions have given and explained in section 2.3.6.

3.6.6 Accelerate failure time model

The Accelerated Failure Time model is a linear regression model in which the response variable is the logarithm or known monotone transformation of a failure time (Lee and Wang, 2003). The accelerated failure time model describes a relationship between the survivor function of any two individuals. Taking T_i to be a random variable denoting the failure time for the *ith* subject, and let $X_{i1}, X_{12} \ldots, X_ip$ be the values of p covariates of the subject. From section 2.3.7 and the equations 60-61, the model is then given by

$$\log T = \{\beta_0 + \beta_1 X_i 1 + ... + \beta_p X_i p + \delta\varepsilon_i\}$$

where $\varepsilon_i \sim h_0(t), \varepsilon_i$ is a random disturbance term, $\beta_0,...,\beta_p$ and σ are parameters to be estimated, $h_0(t)$ is a known baseline survival, T_i is actual survival times sometimes observed is a scale parameter and X_i is a fixed P*1 vector of covariates and X_i is assumed to affect $logT$ linearly and no interactions. Moreover, is σ assumed to be constant and independent of X_i. The parametric accelerated failure time distribution is also assumed to be correctly specified. The survival function for the parametric baseline accelerated failure model is given as,

$$S(t_i) = S_0\{\exp(-X^t i\beta)t_i$$

where $S_0(t)$ is the baseline. These accelerated failure time models are named for the distribution of T rather than the distribution of ε or *logT*. This is so because different distributions have different implications for the shapes or hazard function (Cox and Oakes D, 1996). It has been noted that the proportional hazard model is used exclusively in practice. However; the accelerated failure time model in many ways is more appealing because it is quite direct physical interpretation especially when the response variable does not pertain to failure time (Reid, 1994).

3.6.7 The fitness of the model

When the model has been fitted the adequacy of it needs to be assessed. There are a number of ways to check the adequacy, like using cox-Snell, deviance among others. This study has used the cox-Snell residuals to check the fitness of the model.

3.6.8 Cox –Snell residuals

The Cox-Snell residuals is given by Cox and Snell (Cox and Oakes, 1984).From section 2.3.9 and equations 62-66, the residuals for the ith individual with the observed survival time t_i is given as follows,

$$r_{ci} = \exp(\beta x_i) H_0(t_i) = H_t(t) = -\log \hat{S}_i(t_i)$$

Given that $H_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i and it was derived by Kalbfleish and Pretence (1973).

Letting T be the continuous survival distribution S (t) with the cumulative hazard,

$$H(t) = -\log(S(t))$$

Then it follows thus,

$$S_T(t) = \exp(-H(t))$$

Taking $Y = H(t)$ be the transformation of T based on cumulative hazard function. It follows that the survival function for Y is now given as

$$S_Y = p(Y > y) = p\{H(t) > y\}$$

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

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

The new $Y = H(t)$ has an exponential distribution with unit one. If the model was well fitted, the actual value $\hat{S}_i(t_i)$ would have similar properties to those of $S_i(t_i)$

Therefore $r_{ci} = -\log \hat{S}(t_i)$ will have a unit exponential distribution with $f_R(r) = \exp(-r)$. Letting $S_R(r)$ denotes the survival function of Cox-Snell residuals r_{ci}, then $S_R = \int_r^{\infty} f_R \exp(-x)dx = \exp(-r)$ and it follows that

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

Hence a plot of $H(r_{ci})$ versus r_{ci} is used to check the fit of the model. This gives a straight line with a unit slope and zero intercept if the fitted model is correct.

3.6.9 Model checking

There are a number of methods which are employed to check if a parametric distribution fits the observed data. In this study the Akaike information Criterion (AIC), a statistical criteria, used for comparing models and residuals plots used to check the goodness of fit of the model have been used.

3.7.0 Akaike information criterion (AIC)

The Akaike information criterion (AIC) proposed in AKaike (1974) is a measure of the goodness of fit of an estimated statistical model which compares the models which have been fitted. The AIC is an operation way of trading off the complex of an estimated model against how well the model fits the data. From equation 67, the AIC is defined by

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

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 $2(k+c)$ is thought as penalty if non-predictive parameters are added to the model. Lower values of the AIC suggest a better model. However, there is a difficulty in using AIC in the sense that there is no formal test statistically to compare different AIC values when two or more

models have similar AIC values. Hence, the choice of the model may be hard to determine.

3.7.1 Residual plots

The residual plots have been used to check for the goodness of fit of the model. Among the useful plots is based on comparing the distribution of the cox-Snell residuals with the unit exponential distribution. From section 2.4.2 and equations 68-70, the cox-Snell residual for ith individual with observed time, t_i is defined as:

$$r_{ci} = \hat{H}(t_i \mid x_i) = -\log[\hat{S}(t_i)x_i]$$

where t_i is the observed survival time for individual i, x_i is the vector covariate values for individual ii, and $\hat{S}_i(t_i)$ is the estimated survival function on the fitted model. The estimated survival function for the ith individual is given by

$$\hat{S}_i(t_i) = S_{\varepsilon i}(\frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}})$$

All the statistical analyses have been performed using STATA version: 12.

Conclusion drawn from the aforementioned analyses has been limited by the retrospective study.

Chapter 4

4 Results

4.1 Descriptive

Table 2 gives a descriptive summary of the categorical predictor variables considered for the HIV/AIDS taking ART in Ntcheu district from 2007-2012.Theresults show that 482(7.2%) had died, 1558(23.4%) had transferred out, 1466(22%) defaulted and 3161(47.4%) were found to be alive and on ART. The results from the table further shows that more females, 4527(67.9%) were registered during the period as compared to their male counterparts, 2143(32.1%).On World health clinical staging (WHO), the results show that more patients were starting taking ART at WHO clinical stage3,4010(88.3) as compared to WHO clinical stage 4,531(11.7). The majority of these patients had no occupation,5182(77.7%)and 1485(22.3%) were the patients who reported to have an occupation by the time they were initiating ART. The cd4<=250 was another issue making patient to be initiated on ART. The results from the table indicates that 1462(21.9%) patients were initiated when cd4<=250. The table further reports the mean for the age, body mass index, weight and height as 37.9 years,20.5kg/m^2, 52.4kg and 158.9cm respectively. The median for age, body mass index weight and height has been reported to be 37years, 20.3kg/m^2, 51.6kg and 159cm respectively. Furthermore, the interquartile range for age, body mass index, weight and height has been reported to be 13years, 4.2kg/m^2, 11.2kg and 11cm respectively.

Table 2: Descriptive summary of baseline characteristics for HIV/AIDS patients taking ART in Ntcheu :2007-2012

VARIABLE	FREQUENCY	MEAN	MEDIAN	IQR
death	482 (7.23)			
Transferred out: censored	1558 (23.37)			
Defaulted: censored	1466 (21.99)			
Alive on ART	3161 (47.41)			
Total	6667			
Sex:				
male	2143 (32.1)			
female	4527 (67.9)			
Total	6670			
WHO clinical stage:				
3	4010 (88.3)			
4	531 (11.7)			
Total	4541			
Occupation:				
Working	1485 (22.3)			
non-working	5182 (77.7)			
Total	6667			
Cd4<=250:				
0	5204 (78.1)			
1	1462(21.9)			
Total	6666			
Age		37.9	37	13
BMI		20.5	20.3	4.2
Weight		52.4	51.6	11.2
Height		158.9	159	11

4.2 Univariate analyses

The univariate analysis and the tests of equality to explore whether or not to include the predictor in the final model and find out whether there exist significant difference in the groups for the survival of HIV/AIDS patients taking ART in Ntcheu district have been considered. In this study logrank test which is a non- parametric test has been employed to test equality of survival curves between categorical levels of variables which include: sex, occupation, cd4_250,WHO clinical stage and regimen to find out whether there exist significant difference for the effects of these categorical variables. Table 3 gives the results of the logrank test

Table 3 : Logrank test for equality of survival function

VARIABLE	CHISQUARE(1)	P-VALUE
Sex	56.61	0.00
occupation	1.32	0.254
WHO clinical stage 4	21.42	0.00
Cd4_250	10.31	0.0013
regimen	30.27	0.00

The results of the logrank-test of equality of survival function presented in table 3, shows a significant difference in the survival of patients for the sex group with a *p*-value ≤ 0.01.AWHOclinical stage 4, regimen as well as cd4_250 groups have also indicated the significant difference in the survival of patients taking ART with a *p*-value of $p \leq 0.01$. However, occupation has been reported to be statistically

insignificant as its *p*-value has been reported to be 0.25 ($p>0.05$.Therefore there is sufficient evidence of no difference in the survival of patients between working and nonworking patients taking ART in Ntcheu district. The following figures (1-4) are the Kaplan-Meier curves showing the probability of the survival of HIV/AIDS patients taking ART overtime in Ntcheu district after initiation of ART comparing the survival among subgroups taking into account the categorical variables at a time.

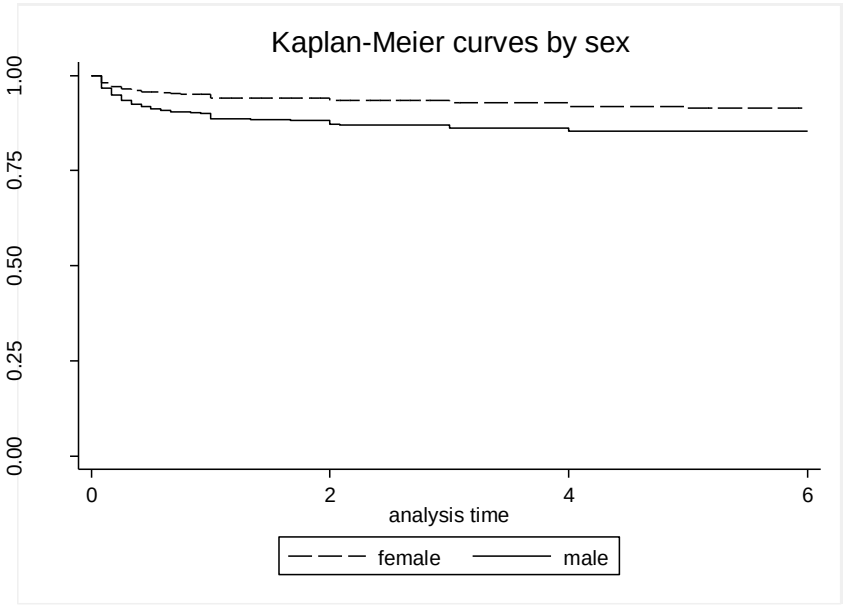

Figure 1: Kaplan-Meier curves comparing survival of male and female HIV/AIDS patients in Ntcheu district

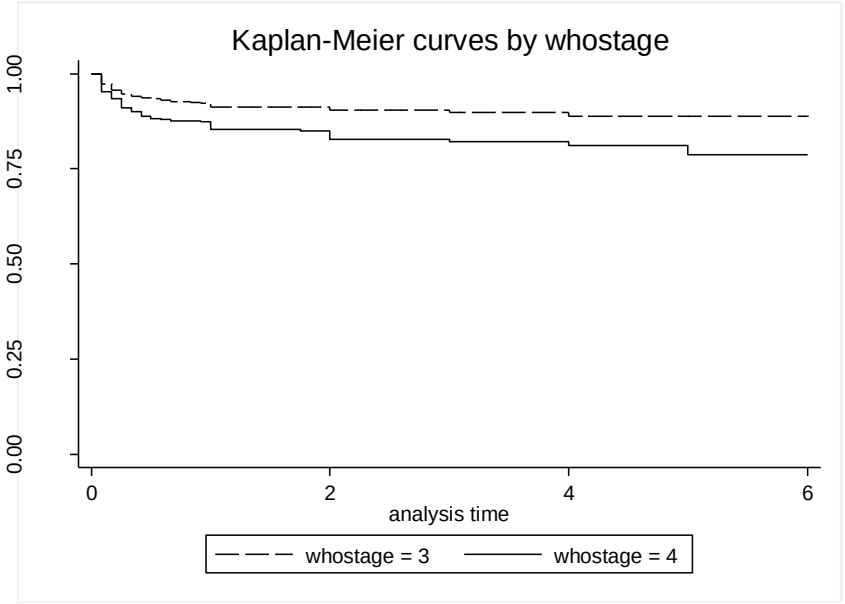

Figure 2: Kaplan-Meier curves comparing survival of HIV/AIDS patients in WHO clinical stages 3 and 4 in Ntcheu district

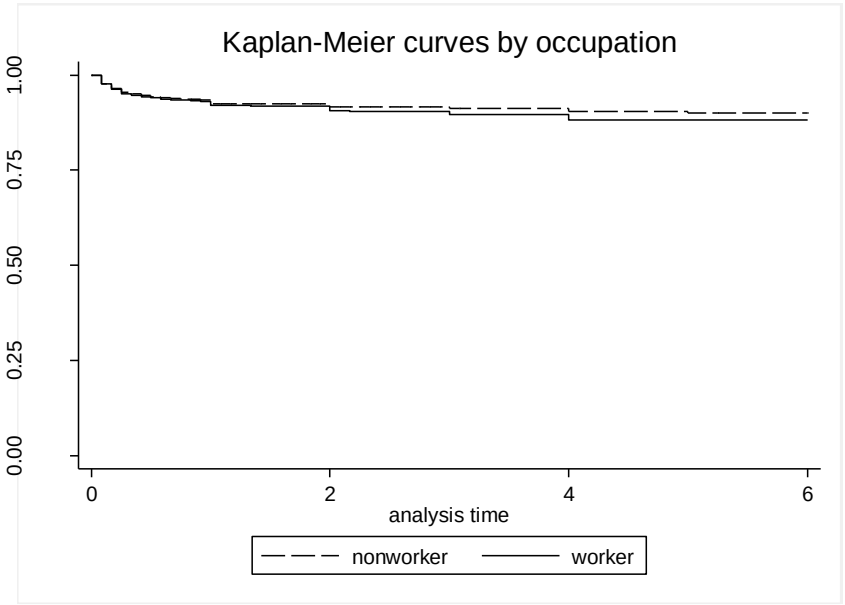

Figure 3: Kaplan-Meier curves comparing survival between working and nonworking HIV/AIDS patients in Ntcheu district

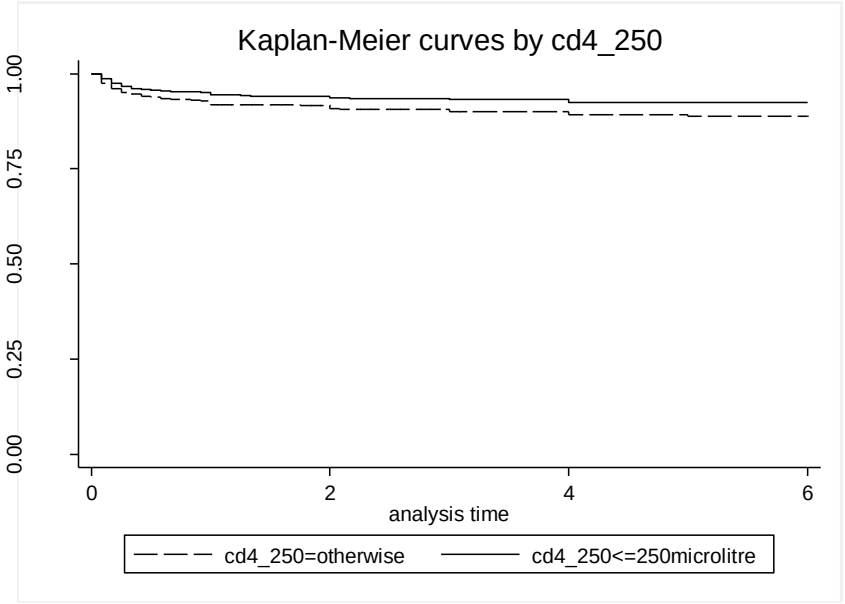

Figure 4: Kaplan-Meier curves comparing survival of HIV/AIDS patients initiating ART at CD4_250 microlitre and otherwise in Ntcheu district

The results from Kaplan-Meier survival estimates, figure 1, have shown that females have a better survival as compared to males, as evidenced by the survival for the females being positioned above of that of the males. It has also been reported that the patients who initiated ART at WHO clinical stage 3 showed better survival as compared to WHO clinical 4, figure 2. Although it has shown that HIV/AIDS patients who were non-working had better survival as compared to working in the occupation category, figure 3, there is no significant difference in the survival of patients statistically based on the logrank test table: 3.The HIV/AIDS infected patients who were initiated ART at cd4<=250microlitre showed better survival as compared to others, figure 4.

4.3 Cox proportion hazard model

Univariate analysis is helpful in determining which predictor is significant before being a candidate for complicated model. This is a kind of exploratory analysis. The results from table 4 indicates that body mass index (bmi) is significant in both univariate analysis (HR=0.84; $p \leq 0.01$) and multivariable analysis(HR=0.84; $p \leq 0.01$) Moreover, the results have shown that the bmi is associated with the decreased mortality of HIV/AIDS infected patients taking ART.The year increase in age has been insignificantly associated with hazard of death of HIV/AIDS patients taking ART in both univariate analysis (HR=1.02; $p \leq 0.01$) and multivariable analysis(HR=1.00; p=0.076).The male patients have been significantly associated with higher hazard of mortality of almost 2 times the hazard of the female patients in both univariate analysis(HR=1.95; $p \leq 0.01$) and multivariable analysis (HR: 1.55; $p \leq 0.01$). The occupation of the patients has been found not to be significant influence on their survival in both univariate analysis (HR=1.13; p=0.25) and multivariable analysis (HR=1.00; p=1.00).Therefore, it has not been considered in the final model. The patients in WHO clinical stage 4 have been significantly associated with 1.78 times higher risk of mortality as compared to the patients in WHO clinical stage 3 in univariate analysis (HR=1.78; $p \leq 0.01$) and 1.64 times higher mortality as compared to WHO clinical stage 3 in multivariable analysis(HR=1.70; $p \leq 0.01$).The cd4_250 _1 has been shown to have a significant impact in decreasing the hazard of the patients of HIV/AIDS infected patients taking ART (HR=0.68; $p \leq 0.01$) univariate analysis. However, it has been indicated to be insignificant in multivariable analysis (HR=1.29; p=0.352). Hence the cd4_250 has been excluded in the final model. Regimen 5 has been shown to have a significant impact in decreasing the hazard of

the patients of HIV/AIDS infected patients taking ART (HR=0.16; p≤0.01).However, it has been indicated to be insignificant in multivariable analysis (HR=1.4; p=0.563).Hence the regimen 5 has been excluded in the final model.

Table 4 : Cox proportional hazard model of HIV/AIDS infected patients taking ART in Ntcheu district

	UNIVARIATE ANALYSIS			MULTIVARIABLE ANALYSIS		
COVARIATES	HR	95%CI	P-VALUE	HR	95%CI	P-VALUE
Body mass index	0.84	0.81-0.86	0.00	0.84	0.81-0.87	≤0.01
age	1.02	1.00-1.02	0.00	1.00	1.00-1.02	0.076
Regimen 5	0.16	0.08-0.34	0.00	1.40	0.45-4.38	0.563
male	1.95	1.63-2.34	0.00	1.55	1.24-1.94	≤0.01
working	1.13	0.92-1.39	0.25	1.00	0.77-1.29	1.00
Cd4_250_1	0.68	0.53-0.86	0.00	1.29	0.75-2.21	0.352
WHO clinical	1.78	1.39-2.28	0.00	1.64	1.24-2.17	≤0.01

WHO clinical =WHO clinical stage 4

A multivariable Cox-proportion hazard model was finally fitted taking into consideration of only those predictors which were found significant in both univariate and multivariable analysis. Table 5 shows the coefficients of the Cox-proportion hazard model among HIV/AIDS infected patients taking ART in Ntcheu district.

Table 5 : Coefficients of proportional hazard model of HIV/AIDS infected patients taking ART in Ntcheu district

COVARIATE	COEFFICIENT	95%CI	P-VALUE
Body mass index	-0.17	-0.20, -0.14	$p \leq 0.01$
male	0.44	0.21, 0.66	$p \leq 0.01$
WHO clinicalstage4	0.53	0.25, 0.81	$p \leq 0.01$

From equation 18, the multivariable Cox Proportion hazard model is given by

$$h(t \mid x) = h_0(t)\exp(-0.17bmi + 0.44male + 0.53WHOclinicalstage4)$$

The model above entails that body mass index decreases the risk of death while being male and WHO clinical stage 4 increase the risk of death.

4.4 Checking the proportional hazard model assumption

When the proportional hazard model has been finally fitted, there is need to verify the assumptions. In the proportional hazard model, the proportion hazard is the core assumption. In this study, graphical approach (stcoxkm), time dependent covariates in Cox model and tests and graphs based on the Schoenfeld Residuals have been used.

4.4.1 Graphical approach: Cox and Kaplan-Meier curves (stcoxkm)

In this approach, Kaplan-Meier plots observed survival curves and compares them to the Cox predicted curves for the same variable. The closer the observed curves are to the predicted curves the less likely it is that the proportion hazard assumption has been violated. The figures(5- 6) show the stcoxkm plots.

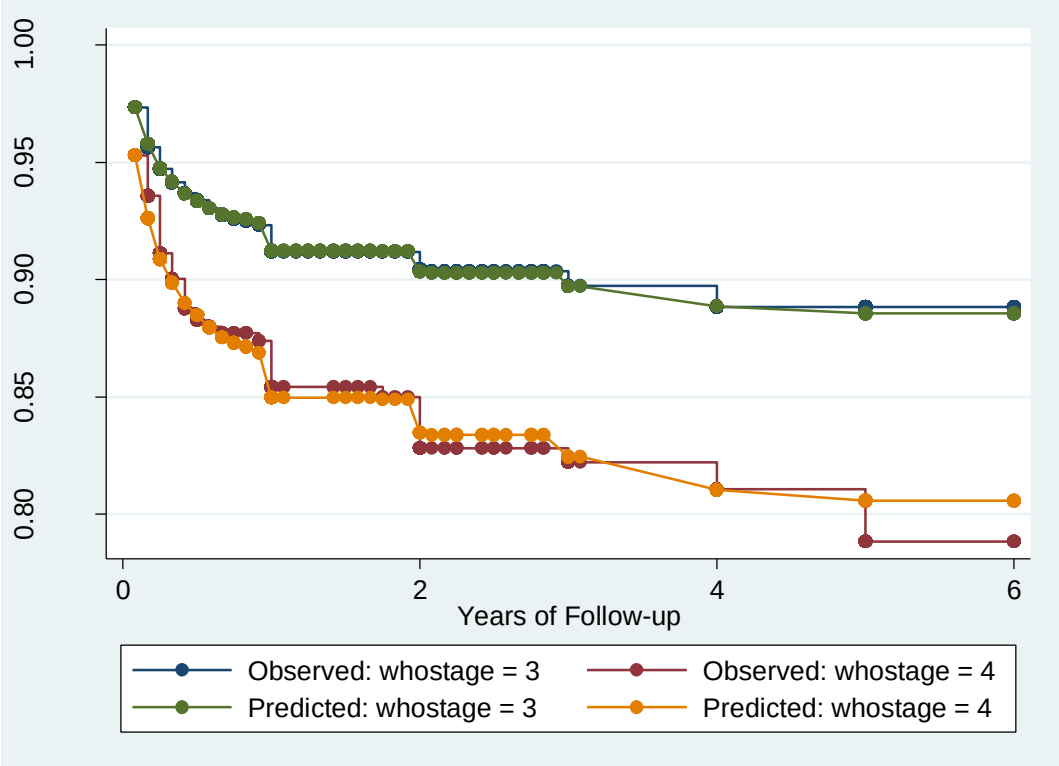

Figure 5 : Cox and Kaplan-Meier plots for categorical variables (WHO clinical stages) for checking proportional hazard model assumption among HIV/AIDS infected patients in Ntcheu district

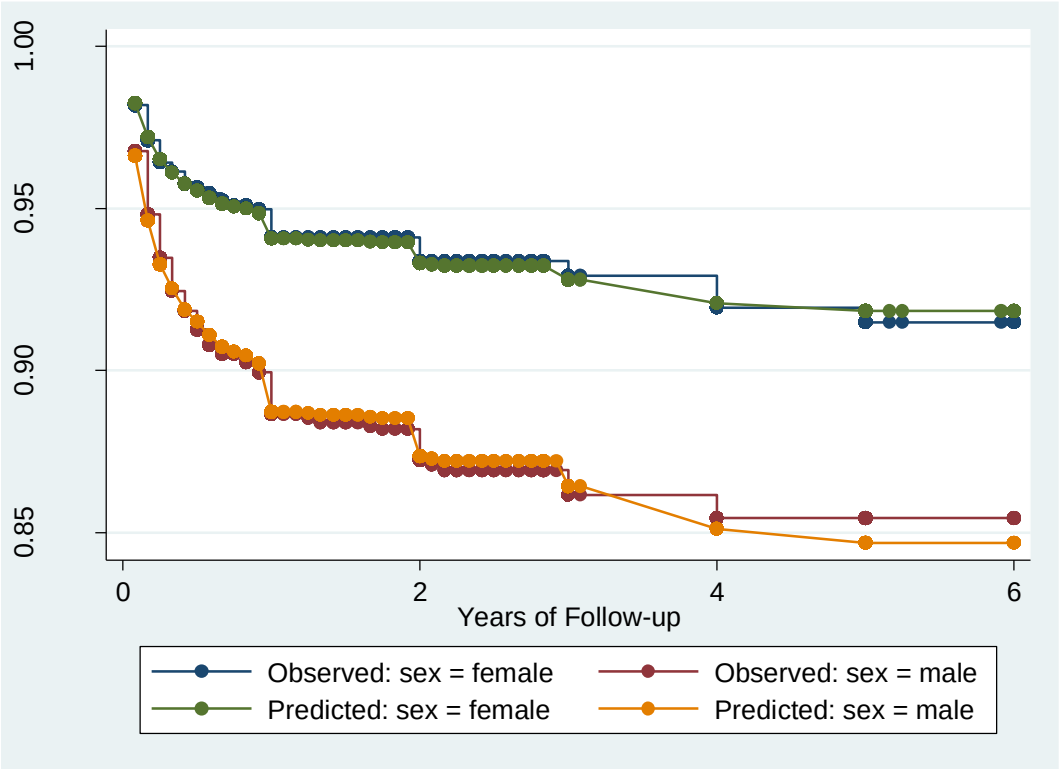

Figure 6 : Cox and Kaplan-Meier plots for categorical variable (sex) for checking proportional hazard model assumption among HIV/AIDS infected patients in Ntcheu district

The results from figures (5-6) show that the proportion hazard assumption for the two categorical variables hold. The observed values of the Kaplan-Meier are closer to the predicted values of the Cox.

4.4.2 Time dependent covariates in the Cox model

In the Cox approach, time dependent covariates are generated by creating interactions of the predictors and a function of survival time and include in the model. If any of the dependent covariates are significant then those predictors are not proportional. Table 6 shows the time dependent covariates in the Cox model.

`

Table 6 : The time dependent covariates in the Cox model

Function of time,t	Coefficient	95%	P-value
Main			
Body mass index	-0.13	-0.19, -0.7	≤0.01
Male	0.57	0.21, 0.93	0.002
WHO clinical stage 4	0.69	0.25, 1.13	0.002
Time varying coefficient			
Body mass index	0.03	-0.003,-0.06	0.08
Male	0.07	-0.14, -0.28	0.52
WHO clinical stage 4	0.12	-0.13, -0.37	0.36

The results from time dependent covariates, table 6, in the Cox model have shown that the time dependents covariates are not significant suggesting no violation of proportionality. Hence the results can be safely applied.

4.4.3 Test and Graph based on the Schoenfeld Residuals

Testing the time dependent covariates is equivalent to testing for non-zero slope in generalized linear regression of the scaled Schoenfeld residuals on function time. Anon-zero slopes is an indication of a violation of the proportional hazard assumption. The table 7 shows the results of the Schoenfeld based on global test.

Table 7 : Test of proportional hazard assumption on Scaled Schoenfeld residuals based on global test among HIV/AIDS infected patients taking ART in Ntcheu district

	rho	Chi2	df	p-value
Body mass index	0.05	0.77	1	0.38
Male	-0.01	0.03	1	0.85
WHOclinicalstage4	0.05	0.93	1	0.34
Global test		3.33	3	0.65

The results in table 7 indicate that the proportional hazards assumption is justified, since the global test is not significant.

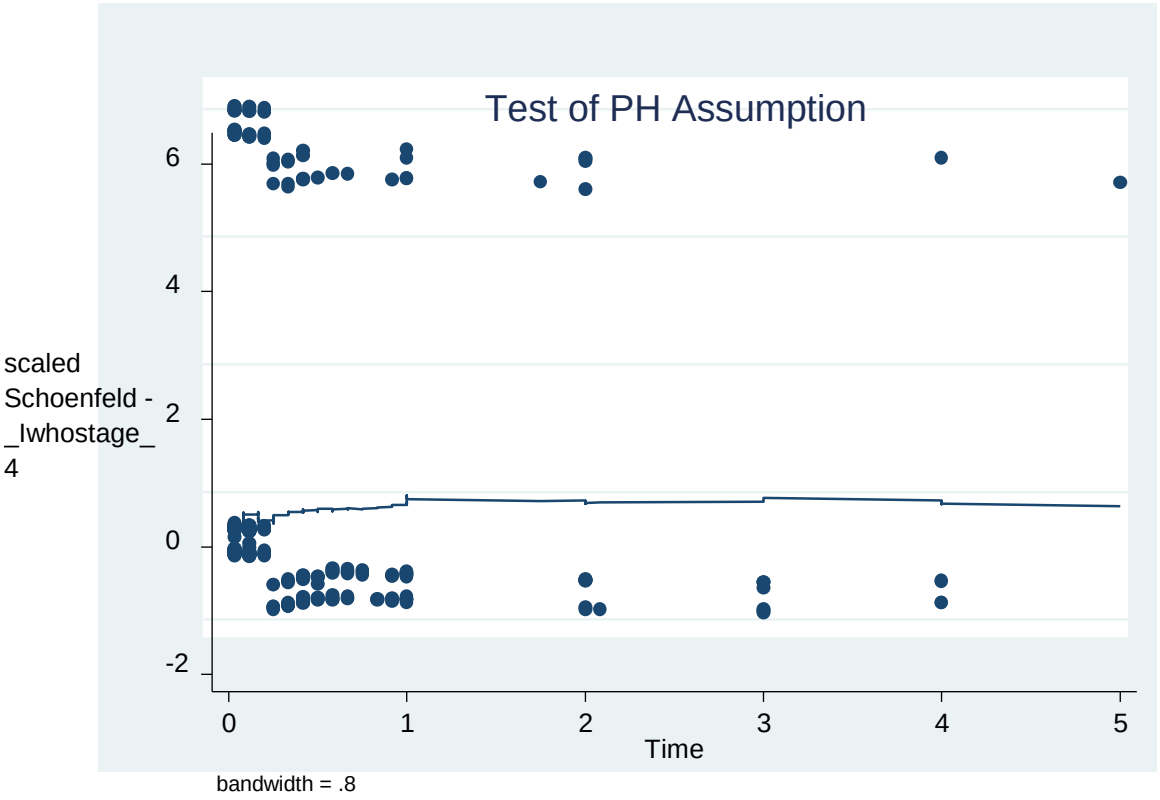

Figure 7 : Graph of Scaled Schoenfeld residuals among HIV/AIDS infected patients taking ART in Ntcheu district based on WHO clinical stage 4

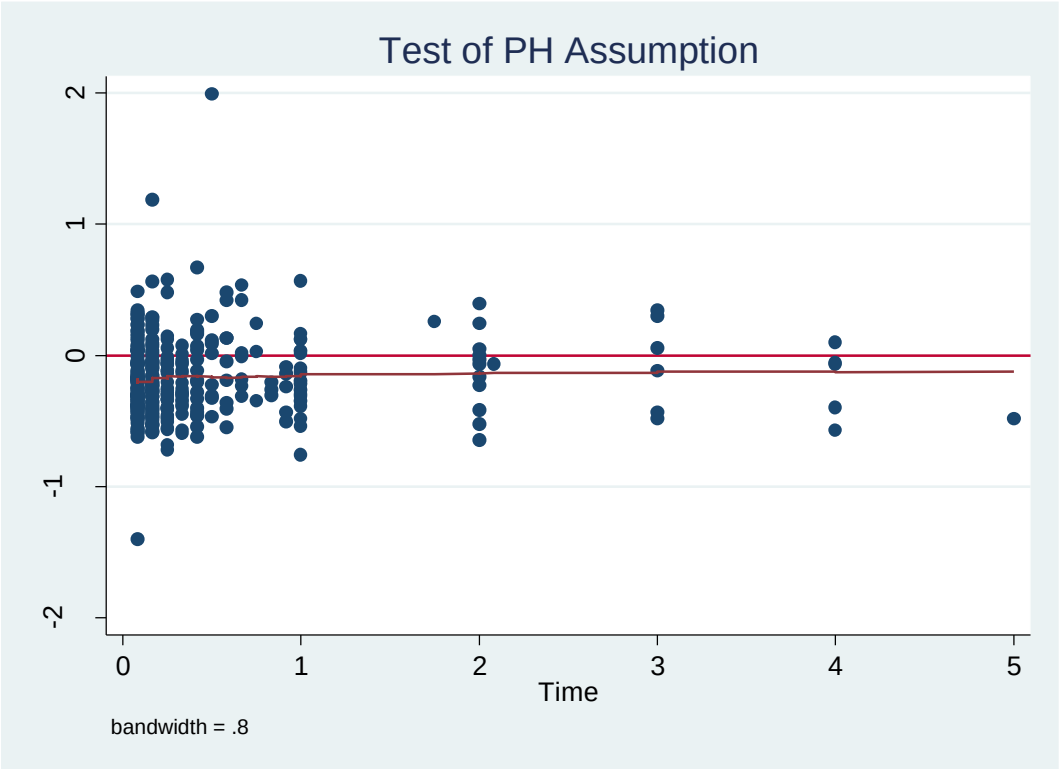

Figure 8 : Graph of Scaled Schoenfeld residuals among HIV/AIDS infected patients taking ART in Ntcheu district based on BMI

In all the plots of Schoenfeld residuals in figures 7 and 8, there is almost a flat line being realized. This is an indication that there is no proportionality problem.

4.4.4 The goodness of fit for the Cox-proportion hazard model

The fit of the Cox-proportion hazard model has been evaluated using Cox –Snell residuals. The Cox-Snell residuals are useful in assessing overall model fit. The model is said to fit the data well when the true cumulative hazard function conditional on the covariate vector has an exponential distribution with a hazard rate of one. Therefore, if the hazard rate follows the 45 degree line it suggests that its approximation has an exponential distribution with a hazard rate of one then the model fits the data well.

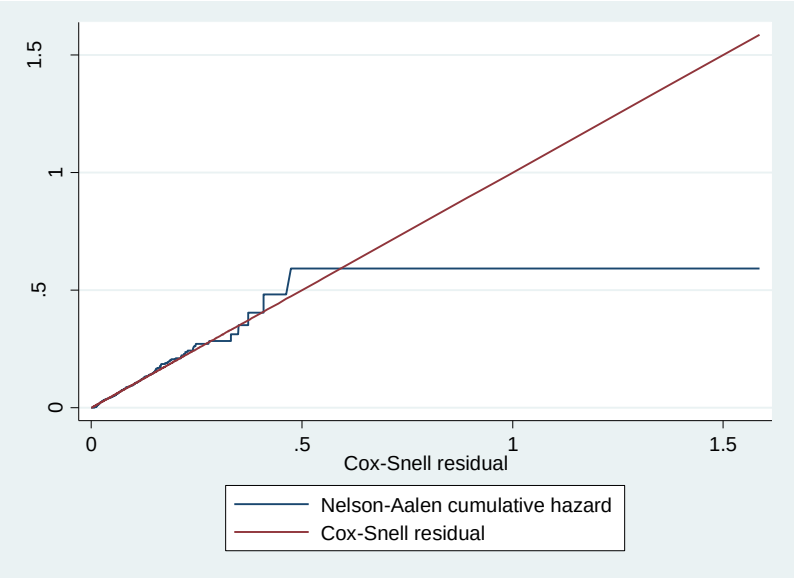

Figure 9 : The goodness of model fit for Cox model among HIV/AIDS infected patients taking ART in Ntcheu district

The result from figure 4 shows that the hazard function at least follows the 45 degree line very closely expect for very large values of time.

Table 8: Univariate parametric regression models with hazard ratio estimates

	EXPONENTIAL			WEIBULL			GOMPERTZ			LOGLOGISTIC			LOGNORM		
Covariate	HR	95% CI	P.V	HR	95% CI	P.V	HR	95% CI	P.V	COEF	95% CI	P. V	COEF	95% CI	P.V
male	2.01	1.68,2.41	0.00	1.97	1.65,2.36	0.00	1.98	1.66,2.37	0.00	-1.27	-1.61,-0.92	0.00	-1.24	-1.58,-0.90	0.00
age	1.01	1, 1.02	0.01	1.01	1, 1.02	0.00	1.02	1,1.02	0.00	-0.03	-0.04,-0.01	0.00	-0.03	-0.04,-0.01	0.00
Regimen5	0.35	0.17,0.74	0.01	0.21	0.10,0.45	0.00	0.18	0.08,0.37	0.00	2.87	1.48, 0.13	0.00	2.49	1.42, 3.57	0.00
Cd4_250	0.67	0.53,0.85	0.00	0.67	0.53,0.85	0.00	0.66	0.52,0.83	0.00	0.75	0.31, 1.19	0.00	0.74	0.31, 1.17	0.00
WHO clin	1.82	1.42,2.34	0.00	1.80	1.4,2.31	0.00	1.82	1.42,2.33	0.00	-1.16	-1.66,-0.66	0.00	-1.12	-1.64,-0.60	0.00
working	1.06	0.86,1.30	0.60	1.09	0.89,1.35	0.38	1.11	0.90,1.37	0.32	-0.17	-0.55,0.22	0.39	-0.15	-0.54,0.23	0.43
bmi	0.83	0.81,0.86	0.00	0.83	0.81,0.86	0.00	0.83	0.81,0.86	0.00	0.36	0.29,0.42	0.00	0.29	0.23, 0.34	0.00

P.V is p-value where 0.00=p-value≤0.01, WHO clin=WHO clinical stage 4

Table 9: Multivariable parametric regression models with coefficient estimates

	EXPONENTIAL			WEIBULL			GOMPERTZ			LOGLOGISTIC			LOGNORM		
Covariate	COEF	95% CI	P.V	COEF	95% CI	P.V.	COEF	95% CI	P.V	COEF	95% CI	P. V	COEF	95% CI	P.V.
male	0.55	0.33,0.78	0.00	0.49	0.26,0.71	0.00	0.47	0.24,0.69	0.00	-0.97	-1.41,-0.52	0.00	-0.94	-1.38,-0.49	0.00
Bmi	-0.20	-0.23,-0.16	0.00	-0.18	-0.22,-0.15	0.00	-0.18	-0.21,-0.14	0.00	0.38	0.30,0.46	0.00	0.34	0.27,0.41	0.00
Age	0.01	0.00,0.02	0.05	0.01	-0.001,0.02	0.1	0.01	-0.001,0.02	0.1	-0.02	-0.04,0.001	0.1	-0.02	-0.04,0.003	0.1
Regimen5	1.34	0.19,2.48	0.02	0.67	-0.48, 1.81	0.25	0.49	-0.65,1.63	0.40	-1.26	-3.45,0.94	0.26	-1.10	-3.24,1.05	0.32
WHOclin 4	0.53	0.25,0.81	0.00	0.53	0.26,0.81	0.00	0.55	0.27,0.83	0.00	-1.08	-1.65,-0.52	0.00	-1.14	-1.72,-0.57	0.00
Cd4_250	0.44	-0.11, 0.97	0.11	0.30	-0.23,0.84	0.27	0.27	-0.26,0.81	0.32	-0.51	-1.56,0.55	0.35	-0.32	-1.42,0.78	0.56
working	-0.06	-0.32, 0.19	0.63	-0.03	-0.28, 0.23	0.85	-0.02	-0.27,0.24	0.89	0.03	-0.47,0.54	0.90	0.02	-0.49,0.52	0.57
Cons	0.11	-0.66,-0.87	0.79	0.38	-0.37,1.13	0.32	1.34	0.59,2.09	0.00	-1.40	-2.92,0.12	0.07	-0.50	-1.93,0.93	0.5

P.V is p-value where 0.00=p -value≤0.01, WHO clin 4=WHO clinical stage 4

Table 9: Continued

	EXPONENTIAL			WEIBULL			GOMPERTZ			LOGLOGISTIC			LOGNORM		
Lnp				-0.66	-0.75,-0.56	0.00									
P				0.52	0.47,0.57										
1/p				1.93	1.76,2.12										
Lngamma										0.60	0.50,0.69	0.00			
Lnsigma													1.28	1.19,1.37	0.00
Sigma													3.55	3.24,3.89	

Table 10: Multivariable parametric regression models with hazard ratio estimates

	EXPONENTIAL			WEIBULL			GOMPERTZ			LOGLOGISTIC			LOGNORM		
Covariate	HR	95% CI	P.V	HR	95% CI	P.V	HR	95% CI	P.V	COEF	95% CI	P. V	COEF	95% CI	P.V.
Bmi	0.82	0.79,0.85	0.00	0.83	0.86,0.86	0.00	0.84	0.81,0.86	0.00	0.38	0.30,0.46	0.00	0.34	0.27,0.41	0.00
Age	1.01	1.00,1.02	0.05	1.01	1.0,1.02	0.07	1.01	1.0,1.02	0.09	-0.02	-0.04,-0.001	0.06	-0.02	-0.04,-0.003	0.10
Regimen 5	3.80	1.21,11.92	0.02	1.95	0.62,6.1	0.25	1.63	0.52,5.11	0.40	-1.26	-3.45,0.94	0.26	-1.10	-3.24, 1.05	0.32
male	1.73	1.38,2.17	0.00	1.62	1.30,2.03	0.00	1.60	1.27,2.00	0.00	-0.97	-1.47,-0.52	0.00	-1.94	-1.38,-0.49	0.00
WHO clin 4	1.70	1.29,2.24	0.00	1.71	1.29,2.25	0.00	1.63	1.32,2.29	0.00	-1.08	-1.65,-0.52	0.00	-1.14	-1.72,-0.57	0.00
Cd4_250	1.55	0.90,2.65	0.11	1.35	0.79,2.32	0.27	1.31	0.77,2.24	0.32	-0.51	-1.56,0.55	0.35	-0.32	-1.42,0.78	0.56
working	0.94	0.73,1.21	0.63	0.97	0.75,1.26	0.84	0.98	0.76,1.27	0.89	-0.03	-0.47,0.54	0.90	-0.02	-0.49,0.52	0.95
Cons	1.08	0.52,2.20	0.84	1.26	0.62,2.54	0.53	2.20	1.48,6.02	0.002	-1.27	-2.75,0.22	0.10	-0.19	-1.61,1.24	0.80

P.V is p-value where 0.00=p≤0.01, WHO clin 4=WHO clinical stage 4

Table 10: continued

	EXPONENTIAL			WEIBULL			GOMPERTZ			LOGLOGISTIC			LOGNORMAL		
Lnp				-0.66	-0.76,-0.56	0.00									
p				0.52	0.47, 0.57										
1⁄p				1.93	1.76, 2.13										
Lngamma										0.59	0.50, 0.69	0.00			
gamma							-1.47	-1.68,1.26	0.00	1.81	1.65, 1.99				
Lnsigma													1.27	1.18, 1.36	0.00
Sigma													3.55	3.25, 3.89	

P.V is p-value where 0.00=p≤0.01

The results from the fitted univariate regression models table: 8 have indicated that the continuous variable, age, is only significant (*p-value*<0.005) in exponential model. Hence it has been excluded in the final model as it has no significant effects on the survival of HIV/AIDS patients receiving ART.The multivariable regression models, table: 8c have reported that cd4_250, working, regimen 5 and age are insignificant and are not to be included in the final model as they don't have any significant effects on the survival of HIV/AIDS patients receiving ART in Ntcheu district.

4.5 The exponential model

The exponential model fitted has indicated that occupation and cd4_250 has no influence in the survival of HIV/AIDS patients taking ART in Ntcheu district as in both univariate and multivariable analysis, tables 8 and 9.On the hand, both the univariate and multivariable analysis have reported that sex (male), age of the patient, body mass index (bmi) and WHO clinical stage4are significantly influencing the survival among the HIV/AIDS patient taking ART in Ntcheu district. However, regimen 5 has been indicated to have significant effect on the survival of HIV/AIDS in univariate but not in multivariate analysis, hence has been excluded in the final model. Using equation (38), the fitted hazard function for the exponential model is:

$$h(surv_t \mid x) = 0.07\exp(-0.20bmi + 0.55male + 0.53WHOclinicalstage4)$$

4.5.1 The Weibull model

The results from fitting the weibull model, tables 8 and 9 have indicated that, sex (male), WHO clinical stage 4,body mass index have significant effect on the survival

among HIV/AIDS patients receiving ART in Ntcheu district as in both univariate and multivariable analysis. However cd4_250 and regimen 5 have been shown to be significantly associated with the survival among HIV/AIDS patient taking ART in the univariate analysis only. The variable working has been found not to have any influence for the survival of HVI/AIDS patients taking ART either in univariate or multivariable analysis of the Weibull model. The results have also revealed that the hazard is significantly decreasing as its shape parameter (p=0.52) is reported less than 1. From equation (31), the estimated hazard function for the weibull is

$$h(surv_t \mid x) = 0.38 * 0.52(t)^{0.52-1} \exp(-0.18bmi + 0.49male + 0.53WHOclinicalstage4)$$

4.5.2 The Gompertz model

The results of the Gompertz model have indicated that body mass index (bmi), WHO clinical stage 4 and sex (male) significantly influence the survival among HIV/AIDS patients taking ART from both univariate and multivariable analysis. On the other hand, cd4_250, regimen 5, age have only been reported to be significant in the univariate. The occupation_2(working) has been reported to be insignificant in both univariate and multivariable analysis, tables 8-10.The function and fitted model for the Gompertz, using equation (43), is

$$h(surv_t \mid x) = 1.34\exp(-1.47t)\exp(-0.18bmi + 0.55male + 0.53WHOclinicalstage4)$$

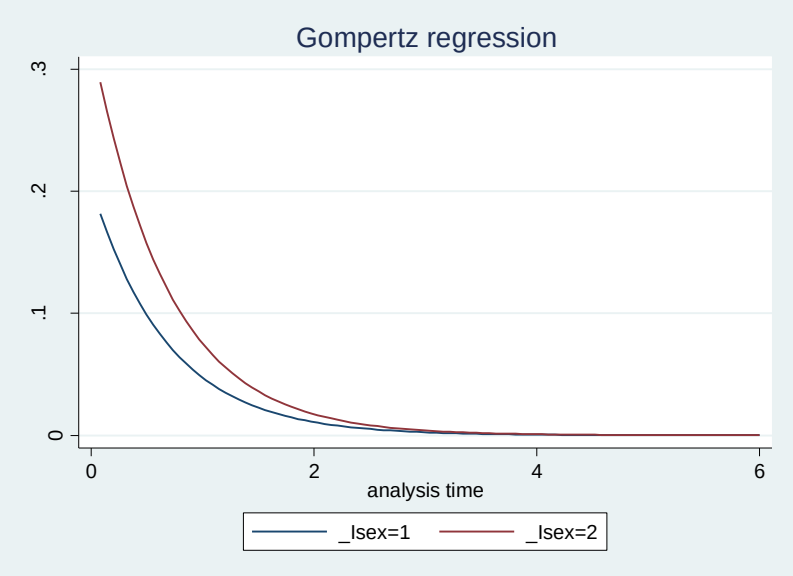

Figure 10: Hazard function for Gompertz model on categorical variable: Sex

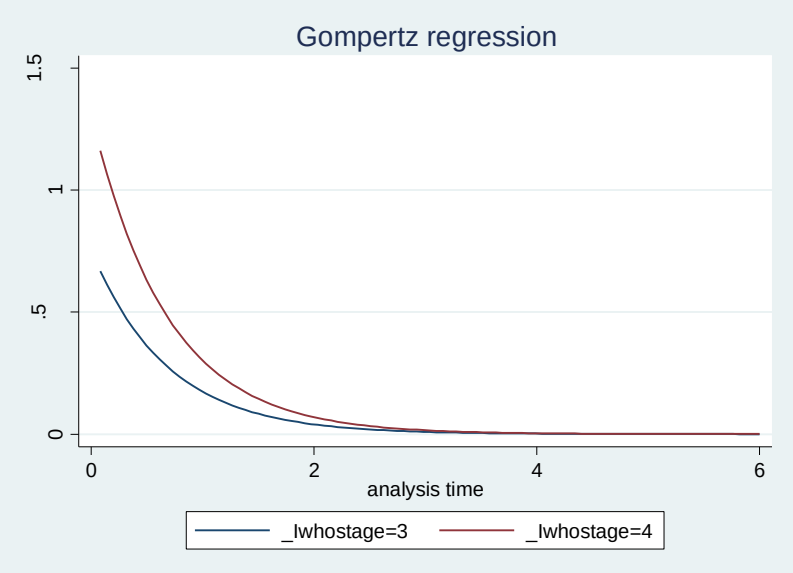

Figure 11: Hazard function for Gompertz model on categorical variable: WHO clinical stage

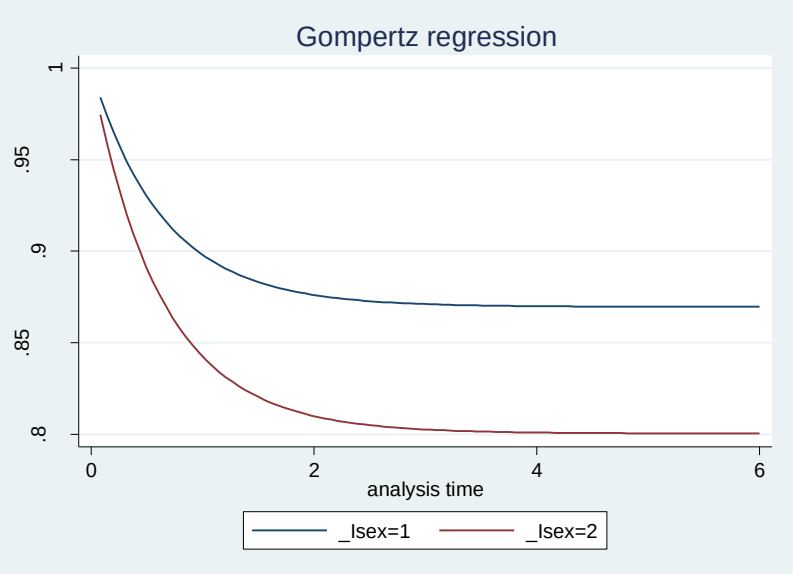

Figure 12: Survival function for Gompertz model on categorical variable: Sex

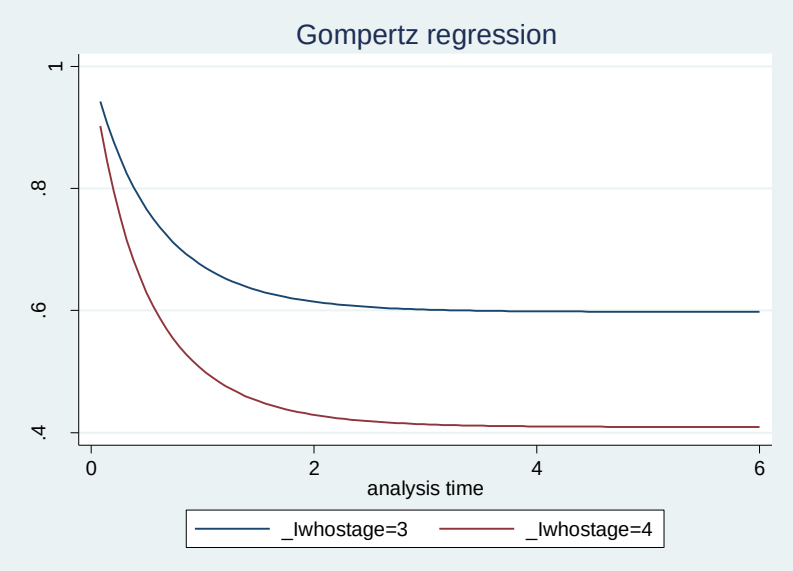

Figure 13: Survival function for Gompertz model on categorical variable: WHO clinical stage

4.5.3 The log logistic model

Similar to what has been found in the exponential, Weibull and Gompertz model, log logistic model has also reported the same results with sex (male), body mass index

(bmi) and WHO clinical stage 4 being having significant effects on the survival among HIV/AIDS infected patients receiving ART in both univariate and multivariable analysis, cd4_250, regimen5 and age has shown the significant influence in the univariate analysis. The working is reported to be insignificant in both univariate and multivariable. The log logistic regression model is only given in accelerated failure, times only. Therefore, from equation (52), table 10, the model for the log logistic is:

$$S(surv_t \mid x_i) = [1 + \{-1.40\exp(-0.38bmi + 0.97male + 1.08WHOclinicalstage4)surv_t\}^{\frac{1}{1.81}}]^{-1}$$

4.5.4 The lognormal regression model

The lognormal regression model fitted has produced the same results as of other models. The cd4_250, age and regimen 5 being significant in univariate analysis. The working has still been reported to be insignificant in both univariate and multivariable analysis. From equation (59), table 10, the estimated survival function for the lognormal in accelerated failure times is:

$$S(surv_t \mid x_i) = 1 - \Phi\{\frac{\ln(t) - 0.5 - 0.34bmi + 0.94male + 1.14WHOclinicalstage4}{3.55}\}$$

4.5.5 Selection of the best-fitted model for the patient receiving ART

There are a number of methods which are employed to check if a parametric distribution fits the observed data. In this study the Akaike information Criterion (AIC), a statistical criteria, used for comparing models and residuals plots used to check the goodness of fit of the model has been used.

4.5.6 Akaike information criterion (AIC)

The study has used the Akaike information criterion so as to select the best parametric model among the ones which have been fitted with the predictors reported to be significant in both univariate and multivariable analysis. The following are the three predictors which have been finally used in fitting the parametric models: male, body mass index (bmi) and WHO clinical stage 4.

Table 11: Akaike information Criterion (AIC) statistics for the parametric models

MODEL	OBSERVATION	LL(NULL)	LL(MODEL)	DF	AIC
EXPONENTIAL	3690	-1761.3	-1681.8	4	3371.5
WEIBULL	3694	-1632.0	-1560.0	5	3161.1
LOGLOGISTIC	3694	-1628.1	-1533.4	5	3116.7
LOGNORMAL	3694	-1605.5	-1535.6	5	3081.3
GOMPERTZ	3694	-1574.2	-1502.4	5	3045.0
COX	3694	-2621.2	-2554.3	3	5114.6

Table 11 shows the AIC statistics for the five parametric models and Cox model applied to the survival data for the selection of the best fitting parametric model. The criterion has demonstrated that Gompertz has indicated the achievement of the lowest AIC value and hence the suggested model for predicting the survival among

HIV/AIDS infected patients receiving ART in Ntcheu district and is followed by the lognormal model. The Cox model fits poorly.

4.5.7 Checking goodness of fit of the fitted parametric models

When the model has been fitted the adequacy of it needs to be assessed. There are a number of ways to check the adequacy, like using cox-Snell, deviance among others. This study has used the cox-Snell residuals to check the fitness of the model.

4.5.8 Cox-Snell residual plots

In this study the goodness of fit has been checked by Cox-Snell residual plots. In figures (14-18) show the Cox-Snell residuals for the five parametric models and from the results it suggests that Gompertz model has fitted the data well as compared to other parametric models. The Gompertz has at least followed the 45 degree line. Therefore, the Gompertz model has emerged the suitable model for modelling HIV/AIDS infected patients taking ART in Ntcheu district

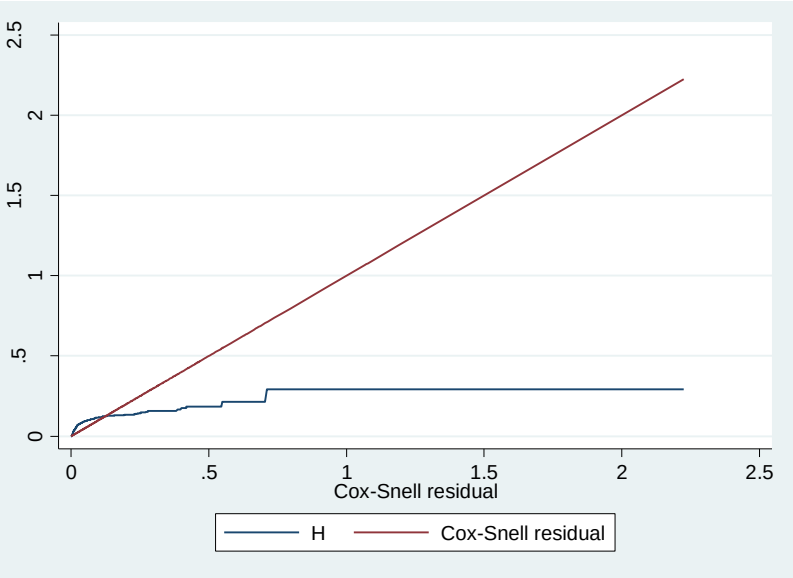

Figure 14: Exponential

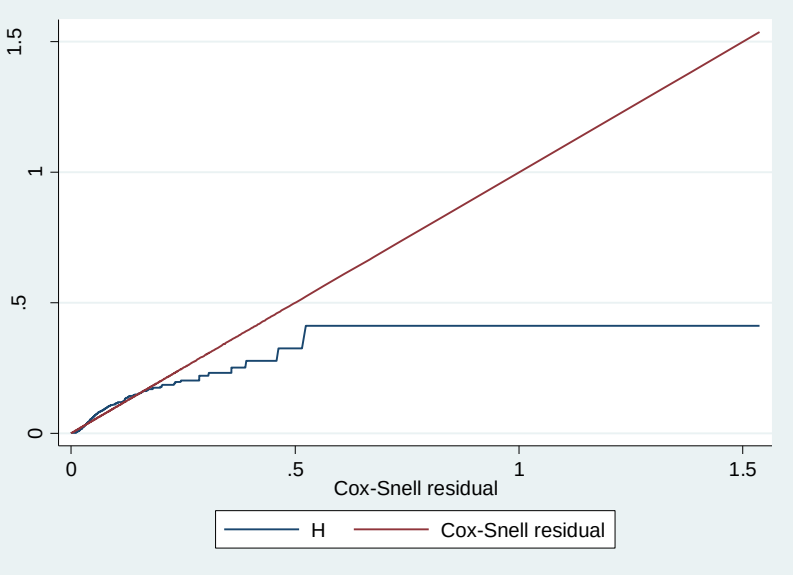

Figure 15: Weibull

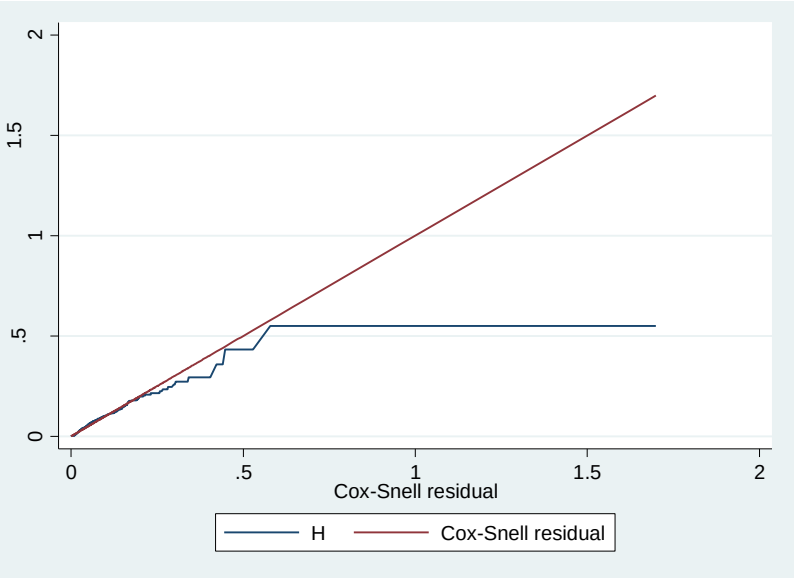

Figure 16: Gompertz

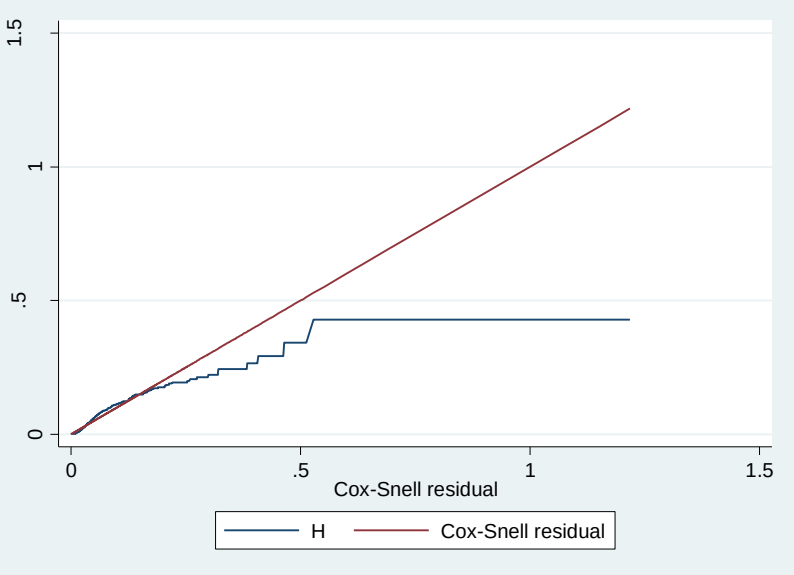

Figure 17: Loglogistic

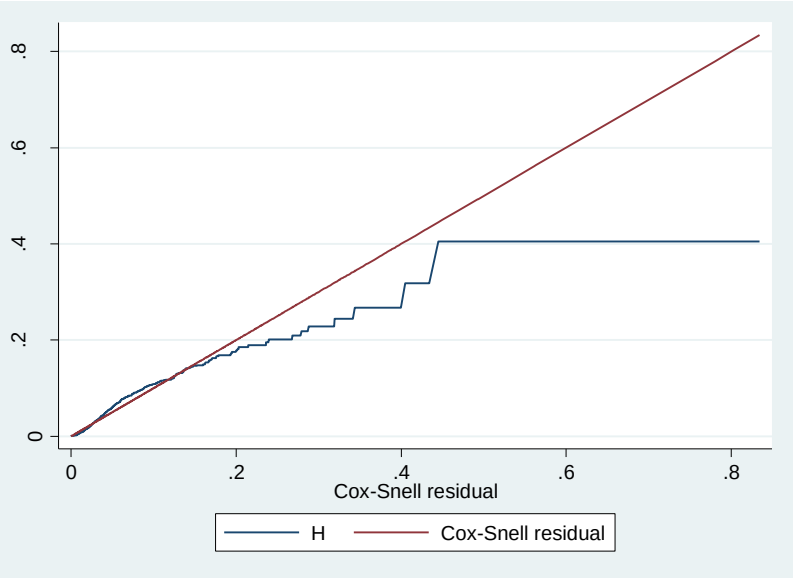

Figure 18: Lognormal

4.5.9: Survival cases of HIV/AIDS infected patients receiving ART in Ntcheu district.

The hazard ratios of HIV/AIDS infected patients receiving ART in Ntcheu estimated from the Gompertz are shown in table 12.

Table 12: Covariates of survival among HIV/AIDS patients taking ART in Ntcheu district, applying Gompertz model: 2007-2012

COVARIATE	HR	95% CI	P-VALUE
male	1.60	1.27, 2.00	≤0.01
Body mass index	0.84	0.81, 0.86	≤0.01
WHO clinical stage4	1.63	1.32, 2.29	≤0.01

From the table 12, it has been shown that the estimated hazard ratio for the body mass (bmi) had been reported to be 0.84(p-value≤0.01) and shows that hazard of death

decreased within increase in the body mass. On the other hand the WHO clinical stage 4 with reference to WHO clinical stage 3 had been noted to have the hazard ratio of1.63(p-value≤0.01).The findings for male had shown that men had 1.6 times higher rate of death (p-value≤0.01) as compared to women.

4. 6.0 Discussion

The study aimed at determining the statistical parametric model on the survival among HIV/AIDS infected patients in Ntcheu district. Five parametric models were fitted on to the data and the estimates of the coefficients were compared to the Cox Proportional hazard model. The results of the estimated coefficients from both the parametric and the Cox proportion hazard models were found to be comparable. To determine the best parametric model, AIC was used and the Gompertz model emerged the champion among the models. It has also been noted that the hazard rate decreases with the passage of time as denoted by gamma from the estimated Gompertz model, table 8b. In this study the Cox proportional hazard assumption was checked using the graphical approach (Cox and Kaplan-Meier survival estimates), time dependent covariates in Cox model, test and graphs based on the Schoenfeld residuals. These methods have indicated that the proportional hazards assumption has not been violated. However, when compared to the parametric models using AIC, the Cox model is the worst model among all the parametric models as it achieved the highest AIC, table 11. The adequacy of the Gompertz model was assessed using Cox- Snell residuals. The results showed that the hazard function followed the 45 degree line very closely. This suggest that the Gompertz model is appropriate and worth to be

used for modelling the survival among the HIV/AIDS infected patients in Ntcheu district.

In review of survival analysis, it was found that only 5% of studies used Cox model with respect to checking the underlying assumptions (Atman DG, e tal, 1985).Hence, when proportion hazard assumption fails to hold researchers should use parametric models as they provide the interpretation based on a specific distribution for duration times without need to proportion hazard assumption. This study has demonstrated that parametric models are best models compared to the Cox parametric model and as it has been already indicated the data has strongly supported the Gompertz which can provide precise results compared to Cox semi- parametric model.

In this study 6670 HIV/AIDS infected patients were followed from 2007 to 2012.More females, 4527(67.9%) were found to have registered for ART as compared to males, 2143(32.1%).The number of females is high due to the fact that they seek medical intervention more than males. Females also get tested for HIV/AIDS when they come for antenatal services. From this study, 482 patients died (7.2%), 1558(23.4%) were transferred out, 1466(22%) defaulted and a total 3161(47.4%) were still on ART. It has been found that the median for age, body mass index, weight and height is 37 years, 20.3kg/m^2, 51.4kg and 159cm respectively. The mean for the age, body mass index, weight and height is 37.9 years, 20.5kg/m^2, 52.4kg and 158.9cm respectively. The study has also indicated that the interquartile range (IQR) for age, body mass index, weight and height is 13 years, 4.2kg/m^2, 11.2kg and 11cm respectively.

More patients initiated ART at WHO clinical stage 3, 4010(88.3%) and 581(11.7%) patients initiated ART at WHO clinical stage 4.In this population of study, 5182(77.7%) were nonworking while 1485(22.3%) were found to be working. A total of 1462(21.9%) initiated on ART at less or equal to 250 Cd4 cell count threshold. The survival among HIV/AIDS infected patients receiving ART in Ntcheu district was modeled using parametric models. Using AIC for selecting the model, it was found that Gompertz parametric model was emerged to be the champion as the best fitting parametric model for modelling the survival among HIV/AIDS infected patients receiving ART in Ntcheu district.

Applying the Gompertz parametric model it was indicated that men had higher risk of death as compared to the females. It has been found out that most men in Ntcheu report late for medical attention after they are already critically ill. It has also been noted that they don't accept to go for HIV/AIDS testing when called to do so. Other findings have also reported the same that late reporting of men to care and treatment clinics causes men to experience more risk of death than women (Egger et al, 2002;Alibhai et al, 2010).Most men would go to HIV testing after experiencing HIV/AIDS related symptoms. Other studies showed that males start ART with advanced disease as compared to females, (Geng et al, 2011). It has been noted that most men in Ntcheu are not compliant to ART treatment, as once they start getting better they stop taking ART. This is increasing the risk of higher mortality in them. Moreover, once they stop complying with ART, they tend to go and continue with immoral behavior like sexual immorality which expose them to a lot of sexually transmitted diseases as well as more HIV/AIDS viral load. The finding on non-

compliant to treatment is consisted with Nakhaeeetal (2011).It has been observed that most men in the district do indulge in other risky behaviours like smoking and beer drinking and these do weaken the immunity of the body leading them to increased risk of mortality. Another factor which has been observed in the study which is making men with HIV/AIDS to experience higher rate of mortality is that they do miss food, say when they go for business and even when they go for social activities like beer drinking. The habit derails the immune system of the body and become easily affected with various opportunistic diseases. On the part of women in the district, they have reduced risk of mortality due to that they freely go for HIV/AIDS testing even before the development of any signs or symptoms. Hence, if tested positive they are enrolled on ART on time making them to have a reduced risk of mortality. Another point is that when they are pregnant they start ART regardless of WHO clinical staging. Hence they stand a better chance of being initiated on WHO clinical stages1 or 2 while men wait till WHO clinical stage 3.

The findings from this study showed that WHO clinical stage 4 was found to be a strong predictor of mortality. It has been noted that most of HIV/AIDS patients in Ntcheu who are dying fast are the ones who developed diseases like Kaposi sarcoma (skin cancer), bacterial pneumonia and had wasting syndrome (severe malnutrition) among others. These complications are associated with WHO clinical stage 4. The finding from this study is consistent with other studies (Coetteet al, 2004; Brailtstein et al,2006;Jerene et al, 2006;Amuron et al,2011). The study has found these HIV/AIDS infected patients delay in seeking medical attention as well as unwillingness to go for HIV testing which could otherwise help them to be initiated

on ART on time, for example at WHO clinical stage 3. As a result they go for medical help while they are already weak hence making them unlikely to survive longer though initiated on ART since their cd4 cell counts are already very low. However, other studies had shown that WHO clinical stage 4 was not associated with mortality (Hambisa, Ahmed and Yadeta, 2013) and it was assumed that this was due to the fact that the majority of the patients (95.9%) have good ARV adherence and large proportions (33.2%) started ART early at WHO stages 1 and 2.

The study has also indicated that HIV/AIDS patients with body mass index > 16kg/m^2 are associated with decreased hazard of death. The other studies have documented that patients with low BMI<16kg/m^2 at ART initiation had a mortality rate double that of patients with BMI ≥ 18.5 kg/m^2(Johannessenet al,2008).In other studies it was shown that Protein-Calorie Malnutrition(PCM) is associated with suppression of antigen-specific arms of immune system and several generalized host defence mechanisms(Mcmurray,1981).It had been noted that persons with Protein-Calorie Malnutrition are more susceptible to opportunistic infections and suffer greater morbidity(Schneider et al, 2004; Schaible and Kaufmann, 2007)and it was reported further that Protein-Calorie Malnutrition is associated with reactivation of viral infections and decreased T-cell primary antibody response and memory response (Najera et al,2007).

Chapter 5

5 Conclusion and Recommendations

5.1 Conclusion

The study aimed at determining the statistical parametric survival model among HIV/AIDS infected patients receiving ART in Ntcheu district. Five parametric models were fitted on to the data and these are exponential, weibull, gompertz, loglogistic and lognormal. All the models fitted produced the same coefficient estimates for the covariates. In this study WHO clinical stage4and sex was found to be significantly associated with increased hazard of mortality rate whereby body mass index was found to be significantly associated with decreased hazard mortality rate.Morever, cd4_250 occupation and regimen 5 and age had no effect in the Gompertz multivariate model which deemed to be the best model. The study used the AIC to compare models and Gompertz was reported to have the lowest AIC hence suggesting it to be the best parametric model for modelling among HIV/AIDS infected patients in Ntcheu district. The goodness- of – fit for the Gompertz model was assessed using Cox-Snell residuals.

The Cox proportional hazard model was fitted in this study so as to compare with the determined parametric survival model, Gompertz, since Cox proportional hazard

model is robust. The study has shown that the results found in Cox proportional hazard model are comparable to what has been reported from the Gompertz model. However, when Cox model is compared to the Gompertz model using AIC it proves to fit the data poorly with its highest value of AIC, table 9. Therefore, researchers can use Gompertz model for modelling the survival among the HIV/AIDS infected patients receiving ART in Ntcheu district. The study has demonstrated that parametric models are the best as compared to Cox model. Hence there is need to fit both parametric and semi-parametric model to survival data and the proportion hazard assumption of the Cox model should assessed .If the proportion hazard assumption of the Cox doesn't hold , researchers should use parametric models which do not depend on proportional hazard assumption. Moreover, the parametric models can be easily conducted by maximum likelihood estimates and allow the researchers to explore the data through the different relationship consisting of linear trend, nonlinear ones or interactions and when the proportions hazard assumption fails to hold these methods lead to valid estimates.

5.2 Recommendation

Adherence is very important for the effectiveness of ART, so there is need to make a follow up to identify reasons for their poor adherence mostly in male patients.

5.3 Study Limitations

a. The major limitation in this study is that data had a lot of missing information. For example, most of the patients had their cd4 cell count not recorded.

b. Furthermore, it was very difficult to trace the patients who were transferred out as to whether they were alive or not or were still on ART or had defaulted. Their outcomes after the study were not known.

c. The study has only looked at HIV/AIDS patients of 15 years of age and above and there is no any information of HIV/AIDS patients below 15years of age on their survival.

REFERENCES

Ahibhai,W., Kippo,L.,&Saunders,D.etal.(2010). Gender- related mortality for HIV-infected patients on highly active antiretroviral therapy (HAART) in rural Uganda. *International journals of women`s health* 2, 45-52.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic Control,* 19(6), 716 -23.

Ata, N., Karasoy, D., &Sozer,MT.(2008). Stratified Cox regression model for non-proportional hazards and an application on breast Cancer patients Turkiyeklinikles. *J mediscie*, 28, 327-32

Bewick, V., Check, L., & Ball J,. (2004).Statistical review IL: *Survival Analysis.Crit Care*. 8(5): 389-394.

Bhunyabhad, C. (2011*).*The survival time of HIV/AIDS with and without Tuberculosis in Yala province in Thailand. *Asia journal of public health2, accessed from:www.as.ap.org* (www.as.ap.org)

Blossfeld, H.P., &Hamerle,A.(1989). *Event History Analysis*, New Jersey: Lawrence Erlbaum Associates

Bradbum, M .J.,Clark, T. G., Love,S.B., &Altman. D. G.(2003).Survival analysis Part II: Multivariate data analysis-an introduction to concepts and methods. *British Journal of Cancer* 89, 431-436.

Brailtstein P., Brinkhof, M.W. (2006*). Mortality of HIV -1- infected patients in the first year of antiretroviral therapy:*Comparison between low-income and high -income countries. *Lancet* 367, 817-824.

Byers,R.H.J.R., Morgan, W.M., Darrow, W.W.W., Doll, L., Jaffe, H.W., &Rutherford, G.(1988). Estimating Aids infection rates in the Sanfransisco Cohort. *AIDS*, 2(3), 207-10.

Chaison, R.E., Keruly, J.C., &Moore, R.D., (1995). Race, sex, drugs use and progress of immunodefiency Virus disease. *New England Journal of medicine*, 12, 751-6.

Christensen, K. B., Anderson, P. K., Nlelsen, M. L., &Christensen,T. S.(2007).Analysing sickness absence with Statistical models for survival data.*Scand work enviro health.* 33(3), 233-9.*Pubmed* PMID: 17572833.

Cleves, M.A.,Gould,W.W., &Gutierrez,R.G.(2004).*An introduction to survival analysis using STATA*. Place of publication .A press.

Constantin, T.Y., &Leish, F.J. (2012). Estimated mortality of adult HIV-infected patients starting treatment with combination antiretroviral therapy. *Sex transform infect*, 88(suppl2), 133-i43.

Cox D.R, Oakes, D. (1984). *Analysis of survivor data.* London: Chapman and hall.

Cox, D. R. (1975).Partial likelihood.*Biometrika,* 62, 269-276.

Cox,D. R., & Snell, E. J. (1968). A general definition of residuals with discussion. *Journal of the Royal Statistical Society,*30, 248-275.

Dias, S.S., Andreozz, V., Martins, M.O., &Torgal, J.(2009).*Higher institute of statistics and information management*. Lisbon: New University of Lisbon.

Dimie,O., &Reginald,O.(2012).Morbidity and mortality patterns of hospitalized adult HIV/AIDS patients in the era of highly active antiretroviral therapy. A 4 year retrospective review from Zaria, Northern Nigeria. *AIDS Research and Treatment volume* article ID 940580, 10 pages doi: 10.1155/2012/940580.

Egger, M.(2002).prognosis of HIV -1- infected patients starting HAART collaborative analysis of prospective studies. *The lancet, l* 360, (9327), 119-129.

Feinleib, M &McMahon B. (1960). Variation in the duration of survivor of patients with the chronic leukemia. *Blood*, 15(3), 332-49.

Geng,E.H., Haunt, P.W.,&Diero,L.O.(2011). Trends in the clinical characteristics of HIV – infected patients initiating antiretroviral therapy in Kenya, Uganda and Tanzania between 2002 and 2009, *Journal of international AIDS Society*, 14, 46.

Geskus, R., Meyer, L., Hubert, J.B.,Schuitemaker,H., Berkhont, B.,& Rouzioux, C.(2003). Causal pathways of effects of age and the CCR5 -Delta 32 CCR2-64 and SDF-1 3' Alleles AIDS development.*Journal of acquired immune deficiency syndrome*,39 (3), 321-6 .

Gupta RC, Akman O, Lvin S. (1995). A study of log-logistic model in survival analysis.*Biom J*, 41 (4), 431-43.

Hosmer,D.W., &Lemeshow,S. (1999).*Applied Survival Analysis: Regression Modeling of Time to event Data*. New York: John Willey & Sons.

Isingo, R., Zaba, B., Marston, M.,Ndenge, M.,Wringe, A., Beckles, D.,Changalucha, J., Urassa, M. (2007). *Survival after HIV infection in the pre antiretroviral therapy era in rural Tanzania cohort. AIDS,* 21(Suppl.6), 55-513.

Jerene, D. E., Ndale, A., Hailu, Y., Lindtjorn, B. (2006).Predictors of early death in a cohort of Ethiopian patients treated with HAART. *BMC infect Dis*; 6: 136.doi 10. 1186/1471-2334-6-136.

Jerene, G.H.V. (2007). Antiretroviral therapy in Ethiopia*: Overcoming implementation challenges*. Unpublished doctoral thesis, University of Bergen, Norway.

Kalbfleisch, J. D & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life models. *Biometrika*, 60267-278.

Kalbfleisch, J. D., & Prentice R.L. (2002). *The statistical Analysis of Failure Data*,(2nded.). New York: Wiley.

Kalbfleish,J.D.(1980). Prentice. *The Statistical Analysis of Failure Time Data*.New York: John Willey and sons.

Kaplan, E.L.&Meier, P.L.(1958). Non parametric estimation from incomplete observation: *Journal of the American statistical association*, 53, 457-481.

Katies, T.S., & Hannock,T .(2010). Gender –differences on antiretroviral therapy of HIV-1 infected adults in Malawi. *Malawi Medical Journal*, 22(2), 49-56.

Keiding, N., Anderson, P.K., & Klein, J.P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistical in medicines*. 16, 215-24.

Kleibunn, D.G., &Klein, M. (2003).*Survival analysis: a Self –learning text* (second edition*).* New Jersey: John Willey & Sons.

Klein,J.P.(1997). Moeschberger.*SurvivalAnalysis.Techniques for censored data and Truncated data*, New York: Springer.

Kleibum, D.G. (1996).*Survival analysis a self-learning text*. Springer, New Yolk.

Lawn, S.D., Little, F., Bekker, L.G., Kaplan, R., Campel, E., Orrell, C.,& Wood, R.(2009). Changing mortality risk associated with CD4 cell response to antiretroviral therapy in South Africa. *AIDS*, 23; 335-342 doi 10. 1097/QAD.ob 013e32832. 1823 f.

Lee, E.T., &Wang, J.W. (2003).*Statistical methods for survival analysis* (3rd ed.). New Jersey: John Willy & sons.

Lee, E.T.,&Wang, J.W.(2003). Some well-known parametric survival distribution and their applicability. Statistical methods for survival data analysis (3rd ed). Hoboken, New Jersey: *John Willey.*

Lee, E.T., & Wang, J.W. (2003).*Statistical methods for survival data analysis*. New Jersey: John Willey.

Lorenza, N.C. (2009). Quality of life among HIV infected patients in Brazil after initiation of treatment .*Clinics* (*Sao Paulo*), 64(9), 867-875.

Madec,Y. S., Zumillin, E., &Genevier,C.(2009).Weight gain at 3 months of antiretroviral therapy is strongly associated with survival: Evidence from the developing countries. *AIDS* 27, 853~ 61.

Manfred, R. (2002). *Drugs and aging*, 19(9), 647-69.

Mantel, N. (1966).Evaluation of survival data and two new rank statistics arising its consideration. *Cancer – thermotherapy Report,* 50,163-170.

Maria, F. (2002). Survival of adult AIDS patients in reference hospital of a metropolitan in Brazil. *Http: lldx.doi.org*/10.1590/50034-89102002000300004.

Marrazi, M.C.,Liotta, G., & Germano,P. (2008). Excessive mortality in the first year of treatment in HIV type -1- infected patients initiating antiretroviral therapy in resource limited region settings. *AID HIV Retrovirases,* 555-60.

Mcgil Christ, C.A., Alsbett, C.W. (1991).Regression with frailty in survival analysis .*Biometrics* 47, 461-466.

Mcmanus, H., O'Connor, C.C., Boyd, M., Broom, J.,& Russell D.(2012). *Long term survival in HIV positive patients with up to 15 years of antiretroviral therapy*. Plos one 7(11). E48839. Doi: 10.1371/ journal. Pone.0048839.

Ministry of Health Malawi. (2011).*Clinical Management of HIV in Children and Adults. Lilongwe:MoH.*.

Muron, A &Levin, J.(2011). Mortality in an antiretroviral therapy programme in Jinja, South East Uganda .*Prospective cohort study*. AIDS Res Ther; 8;39. Doi: 10:1186/1742-6405-8-39.

Mykhailo, A., Dmytro Z.(1213). Effect of antiretroviral therapy on survival of HIV/tb – infected patients in Ukraine. *Our Dermatol online*, 4(2).

Najera, O., Gonzaleze, C., Cortes, E., Toledo, G.,& Ortiz, R. (2007). Effect T-Lymphocytes in well –nourished and malnourished infected children. *Clinexpimmunol,* 148,501-6.

Nakhaee, F., &Law, M.(2011).*Parametric modeling of survival following HIV and AIDS in the era of highly active antiretroviral therapy*.EM/J VOL 17 NO 3.Accessed from:www.emro.who.int/emhj/v17/03/17-3-2011-0231-0232.

Patrick,O., & John, E. A.(2008).*Adherence of HIV/AIDS patients to antiretroviral therapy in a tertially health facility in Benin City. African journal of pharmacy and pharmacology*.vol 2 (7), P.145-152.Accessed .from:www.academijournals.org/ajpp ISSN 1996-0816.

Phillips, A.N. (1991). More rapid progression to AIDS in older HIV-infected people: The role of cd4 and T-cell counts *Journal of acquired immune deficiency syndrome*, 1991, 4970-975.

Robert, Z., Gilbert, R., & Jean, P. (2012). Survival differences in Australian Patients with acquired Immunodeficiency Syndrome.*European Journal of Epidemiology* 1995, 11:519-526.*AScand J infect Dis*

mar: 44(3), 201-8. Do10.3109/00365548.2011.631937. Epub 2011 nov 28.

Schaible,U.E., Kaufmann, S.H. (2007).Malnutrition and infection *Plos med*., 4 , 806 -12.

Schoenfeld, Partial, D. (1982).Residuals for the proportional hazards regression model.*Biometrika,* 69239-241.

Shikuma, C.M., Zackin, R., & Sattler F. (2004). Changes in weight and lean body mass during highly active antiretroviral therapy. *Clin infect Di*, 39, 1223-30.

Skiest, D.J., Rubinstien, E., Carley, N., Goiella, L.,&Lyons, R.(1996).The importance of Co morbidity in HIV- infected patients over 55 . A retrospective case –control study. *The American Journal of medicine*, 10, (6), 605-11.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990) .Martingale-based residuals for survival models. *Biometrika*, 77,147-160.

Vaupel, J.W., Manton, K.G., & Stallard, E.(1997). The impact of Heterogeneity in individuals frailty on the dynamics of mortality. *Demography,* 6, 439 – 454.

Viscomi, S., Pastorore, G., Damma, E., Zuccolo, L., &Peace, N. (2006). Life expectancy as an indicator in Follow up of population-based Cancer registries. *The example of childhood leukemia*. *Ann Oncol*, 17 (1), 167-71.

Zachariah, R., &Massaqwi, M. (2006).Risk factors for high early mortality in patients on antiretroviral treatment in rural district of Malawi*: AIDS* 20, 2355-60.

Zhou,J., &Nkumarasamyi.(2005).*Predicting short-term disease progression among HIV-infected patients in Asia and Pacific region:* Preliminary results from the treat Asia HIV observation Database CTAHOOD.

Zhou,Y.Y.(2007).Estimating of parameters in logistic and log-logistic distribution with grouped data. *Lifetime Data Anal,* 13(3), 421-9.