ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS

APPLICATION OF COX PROPORTIONAL HAZARDS MODEL IN CASE OF TUBERCULOSIS PATIENTS IN SELECTED ADDIS ABABA HEALTH CENTERS, ETHIOPIA

KABTAMU TOLOSIE

A THESIS SUBMITTED TO
THE DEPARTMENT OF STATISTICS

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

ADDIS ABABA UNIVERSITY
ADDIS ABABA, ETHIOPIA
JUNE, 2013

Addis Ababa University

School of Graduate Studies

Department of Statistics

This is to certify that the thesis prepared by Kabtamu Tolosie, entitled: Application of Cox Proportional Hazards Model in case of Tuberculosis Patients in Selected Addis Ababa Health Centers, Ethiopia and submitted in partial fulfillment of the requirements for the degree of master of science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Advisor Prof. M.K. Sharma Signature M.K Sha Date 24/06/13

Examiner Prof. Getachew Legesse Signature _____ Date 25/06/2013

Examiner Mekonnen Tadesse Signature _____ Date 25/06/2013

Chair of Department or Graduate Program Coordinator

Abstract

Application of Cox Proportional Hazards Model in case of Tuberculosis Patients in Selected Addis Ababa Health Centers, Ethiopia

Kabtamu Tolosie

Addis Ababa University, 2013

Tuberculosis (TB) is a chronic infectious disease and mainly caused by mycobacterium tuberculosis (MTB). It has been one of the major causes of mortality in Ethiopia. The objective of the study was to identify factors that affecting the survival of the patients with tuberculosis who started treatment for tuberculosis. This was a retrospective study in six randomly selected health centers in Addis Ababa, Ethiopia. The data were obtained from medical records of TB patients registered from September 2012 to August 2013 and treated under DOTS strategy. Out of the total 826 registered TB patients 105 (12.71%) died during the study period and 712(87.29%) were censored. Based on Kaplan Meier survival curves, log rank test and Wilcoxon test it was found that the patients had statistically significant differences in survival experience with respect to age, body weight at initiation of treatment, TB patient category and HIV status. Multivariable Cox hazards regression analysis revealed that the covariates age, TB patient category, HIV and age by HIV interaction were significant risk factors associated with death status in TB patients.

Acknowledgment

I would like to express my deepest gratitude to my advisor and instructor Prof. M.K. Sharma, for his commitment of encouragement, guidance, suggestions and constructive comments on my study.

I am grateful to the Addis Ababa Health Bureau for the permission to access data from Bole, Kolfe, Shiromeda, Kazanchis, Nifas Silk Lafto health centers and Yekatit 12 Hospital. My appreciation also goes to the Department of Statistics Addis Ababa University.

Finally I would like to express my warm appreciation to my family for financing and friends for their encouragement throughout my study.

Acronyms

AFB	Acid-fast Bacilli
AIDS	Acquired Immunodeficiency Syndrome
CDC	Centers for Disease Control and Prevention
COPD	Chronic Obstructive Pulmonary Disease
DOT	Directly Observed Therapy
DOTS	Directly Observed Treatment strategy
EPTB	Extra pulmonary Tuberculosis
HIV	Human Immunodeficiency Virus
HMIS	Health Management Information System
HR	Hazard Ratio
IQR	Inter Quartile Range
IUATLD	International Union against Tuberculosis and Lung Disease
KM	Kaplan-Meier
FMOH	Federal Ministry of Health
MPLE	Maximum Partial Likelihood Estimate
MTB	Mycobacterium tuberculosis
NLCP	National Tuberculosis and Leprosy Control Program
PH	proportional Hazards
PTB	Positive Pulmonary Tuberculosis
TB	Tuberculosis
WHO	World Health Organization

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the study

1.1.1. Tuberculosis Epidemiology

Tuberculosis (TB) is a chronic infectious disease mainly caused by mycobacterium tuberculosis (MTB). The main source of infection is untreated smear-positive pulmonary tuberculosis (PTB) patient discharging the bacilli. It mainly spreads by airborne route when the infectious patient expels droplets containing the bacilli. It is also transmitted by consumption of raw milk containing Mycobacterium bovis (Bates, 1980).

Tuberculosis (TB) is the leading cause of death from a single bacterial species among adults around the world. The World Health Organization (WHO) estimated that one third of the world population is infected with *Mycobacterium tuberculosis*, with 9.4 million new cases and 1.3 million deaths in 2009 (WHO, 2010).

Reducing death, eliminating disease, and preventing the development of drug-resistant TB are the major goals of TB control (WHO; 2009). TB-related death is often referred to as a TB control indicator (Kolappan *et al.*, 2006).

Although, the developed countries in Europe and North America have well-equipped treatment facilities and provide free and sufficient anti-TB drugs, treatment success rates there are still below WHO's goal of 85%, which may be because of the relatively high death rates. In those countries, TB cases are found in the relatively higher age group and are associated with comorbidities (Faustini *et al.*, 2005).

In previous studies, even in developed countries with good national TB control programs, more than 10% of cases died during the follow-up period. In high-resource settings, but not in most high TB-burden countries, most cases of death during TB treatment is often because of causes other than TB (Oursler *et al.*, 2002).

On March 24, 1882, Dr. Robert Koch announced the discovery of *Mycobacterium tuberculosis*, the bacteria that cause tuberculosis (TB). Dr. Koch's discovery was the most important step taken toward the control and elimination of this deadly disease. In 1982, a century after Dr. Koch's announcement, the first World TB Day was sponsored by the World Health Organization (WHO) and the International Union against Tuberculosis and Lung Disease (IUATLD). The event was intended to educate the public about the devastating health and economic consequences of TB, its effect on developing countries, and its continued tragic impact on global health (Gore, 2005).

1.1.2 Tuberculosis Transmission

When people with TB in their lungs or throat cough, laugh, sneeze, sing, or even talk the germs that cause TB may be spread into the air. If another person breathes in these germs there is a chance that the person will become infected with tuberculosis.

It is important to understand that there is a difference between being infected with TB and having TB disease. People who are infected with TB have the TB germs, or bacteria, in their body.

Someone with TB disease is sick and can spread the disease to other people. A person with TB disease needs to see a doctor as soon as possible.

It is not easy to become infected with tuberculosis. Usually a person has to be close to someone with TB disease for a long period of time. TB is usually spread between family members, close friends, and people who work or live together. TB is spread most easily in enclosed spaces over a long period of time. However, transmission in an airplane, although rare, has been documented.

Even if someone becomes infected with tuberculosis, that does not mean he/she will get disease. Most people who become infected do not develop TB disease because their body's immune systems protect them.

1.1.3 Tuberculosis Diagnosis

Tuberculosis is diagnosed by finding *Mycobacterium tuberculosis* bacteria in a clinical specimen taken from the patient. While other investigations may strongly suggest tuberculosis as the diagnosis, they cannot confirm it.

A complete medical evaluation for TB must include a medical history, a physical examination, a chest X-ray and microbiological examination (of sputum or some other appropriate sample). It may also include a tuberculin skin test, other scans and X-rays, surgical biopsy (Kumar *et al.*, 2007).

1.1.4 Tuberculosis Prevention

Three strategies are fundamental to the prevention and control of TB. The first priority is identifying and treating persons who have active TB; this priority entails identifying persons who have TB, ensuring that they complete appropriate therapy, and, in exceptional circumstances, using confinement measures. The second priority is finding and screening persons who have been in contact with TB patients to determine whether they have TB infection or disease and providing them with appropriate treatment. The third priority is screening high-risk populations to detect persons who are infected with *M. tuberculosis* and who could benefit from therapy to prevent the infection from progressing to TB disease (Snider and Roper, 1992).

1.1.5 Tuberculosis Treatment

TB bacteria become active (multiplying in the body) if the immune system can't stop them from growing. When TB bacteria are active, this is called TB disease. TB disease will make a person sick. People with TB disease may spread the bacteria to people with whom they spend many hours.

TB disease can be treated by taking several drugs for 6 to 12 months. It is very important that people who have TB disease should finish the medicine, taking the drugs exactly as prescribed. If they stop taking the drugs too soon, they can become sick again; if they do not take the drugs correctly, the germs that are still alive may become resistant to those

drugs. TB that is resistant to drugs is harder and more expensive to treat. In some situations, staff of the local health department meets regularly with patients who have TB and to watch them take their medications. This is called directly observed therapy (DOT). DOT helps the patient complete treatment in the least amount of time treat (CDC, 2011).

1.1.6 Risk Factor Related to TB

Age, gender, cavities, positive sputum smear, diabetes, HIV infection, kidney diseases, hepatitis, chronic obstructive pulmonary disease (COPD), cancer, poor nutrition, smoking, unemployment, alcoholism, delayed diagnosis, and delayed cure are factors associated with death in TB cases. Previous studies were performed mainly in Europe, North America, or HIV-prevalent countries of Africa. A similar study was done in Singapore; however, the study only explored comorbidities with diabetes, cancer, and renal failure, which were adopted from self-report forms rather than health insurance claim database. Another study conducted in Shanghai focused on sputum culture-positive cases. The study explored absent, present, and unknown comorbidities and did not collect actual disease condition data (Yuan *et al.*, 2009).

Here are the members of high-risk groups:

- **Seniors:** As people age, their immune systems aren't quite as able to fight off infections like tuberculosis, increasing seniors' risk for contracting TB.
- **Babies:** Infants who haven't yet developed strong enough immune systems to ward off tough bacterial infections like tuberculosis and are at a greater risk.
- **HIV-positive or AIDS patients:** HIV infection compromises the immune system, making even common illnesses a serious threat. People with HIV infections are much more susceptible to developing active tuberculosis.
- **People with diabetes:** Diabetes is another disease that weakens the immune system, leaving the body less able to defend itself against bacterial infections like tuberculosis.
- **Cancer patients:** Chemotherapy drugs also suppress the immune system, leaving the body more vulnerable to tuberculosis infection.

- **Organ transplant recipients:** The medications used to keep the body from rejecting a new organ suppresses the immune system, potentially allowing TB to develop.
- **People with kidney disease:** This illness can weaken the body, including the immune system, making a person more susceptible to illness.
- **People undergoing treatment for autoimmune diseases:** Some treatments for illnesses like Crohn's disease and rheumatoid arthritis can affect the immune system, creating an opportunity for tuberculosis infection to occur.
- **Malnourished individuals:** The body can't defend itself as well if somebody is underweight or malnourished due to poor diet or illness.
- **People surrounded by potential cases of TB:** This could be anyone who lives with or works closely with someone known to be infected; hospital workers, nursing home staffs, and prison staff, for example, are at greater risk.

Other groups of people who may also be more at risk for TB are drug users and low income individual who may not have access to medical care and testing. Anyone who has been around someone with tuberculosis is at a greater risk of contracting the disease and should get a skin test to diagnose the infection early (Rodriguez, 2009).

1.2 Statement of the problem

Tuberculosis (TB) is a potentially serious infectious disease. It has been one of the major causes of morbidity and mortality in Ethiopia for long. Accordingly, the Ethiopian Ministry of Health and its stakeholders have put their unreserved and integrated efforts in this health problem. Among these efforts is the well developed Health Management Information System (HMIS) for Tuberculosis programs.

However, the direction to where Tuberculosis in Ethiopia is heading hasn't been well analyzed and unpackaged by epidemiologically relevant factors. So, what are the factors that affect the survival of TB patients? It might be Gender, Age, TB patients Category, Type of TB, Smear result, HIV and body weight at initiation of treatments of patients. Therefore, to fill this gap there is a need to study the factors that are affecting the survival of patients with tuberculosis.

1.3 Objectives

1.3.1 General objective:

The general objective of this study is to identify the factors that are affecting the survival of the patients with tuberculosis.

1.3.2 Specific objectives

- To estimate survival time probabilities of the TB patients.
- To compare the survival probabilities of the TB patients with respect to different risk factors.
- To identify the factors influencing death status of patient by using the Cox proportional hazards model.

1.4 Significance of the study

The Ministry of Health and other bodies could benefit as the findings of the study can be used as a basis for further studies. In addition, it is expected that this study could provide information to government and other concerned bodies in setting policies, strategies and further investigation for reducing death to TB patient.

1.5 Limitation of the study

- The study is conducted based on secondary data which might have incomplete and biased information. Previous studies found that alcohol drinking, drug use, malnutrition, unemployment, and multi-drug resistance were risk factor of death in TB patients, but because of restrictions of the administration's data, we could not gather these variables in our study. Also information might have been missed in case of many censored observations.
- In many tuberculosis patients, multiple causes of death may act simultaneously, so the cause of death may not be determined accurately.
- Lack of literature on our country related to the subject under study.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Introduction

This chapter reviews relevant literature from previous studies on factors affecting the survival of the patients with tuberculosis. The scope of this literature review, therefore, is to synthesize evidence from textbooks, published research, scientific reports and other credible sources of scientific work done globally, mainly on TB factors issues.

2.2 Risk factors for tuberculosis

According to Horne *et al.* (2010) the study conducted in Washington State, during a median follow-up of 6 years in 3451 patients treated for tuberculosis, there were 417 deaths. Mortality was independently associated with increasing age, male gender, HIV co-infection, and U.S. birth.

A study of community-based cohort of patients with drug-susceptible pulmonary TB in Maryland showed that, a total of 29 (21%) of the 139 patients died during treatment; the median time to death among these patients was 39 days. The median follow-up for survivors was 202 days. Age was strongly associated with the risk of death. Patient's aged >60 years had a 5-fold increased risk of death compared with patients aged 18–60 years and HIV infection were associated with an increased risk of death during treatment (Oursler *et al.*,2002).

A study was carried out in Brazil aimed to analyze survival probability and identify risk factors for death from tuberculosis in a cohort of patients living in Recife who started treatment for tuberculosis. At the end of the follow-up period, the survival probability after beginning TB treatment was 95.9%. Older age, HIV positive and late initial treatments were statistically associated with death from TB in one year follow-up. When the analysis was done considering the total period of follow-up, older age, HIV positive, late initial treatment, weight loss, and history of previous treatments were statistically significant in the multivariable Cox regression model (Albuquerque *et al.*, 2009).

A study in Pernambuco State, Brazil showed that out of 5,451 individuals reported with tuberculosis, 320 (5.9%) died (incidence and mortality rates of 98.6 and 12.2/100 thousand inhabitants, respectively. TB/HIV co-infection is an important predictor of TB mortality. However, among individuals without HIV/AIDS, mortality is still highly associated with older age, mixed clinical forms, and treatment dropout. In both the univariate and multivariate analysis, the relative risk of dying from TB increased in a dose-response gradient with age, i.e., the older the patient, the greater the risk of dying. Patients with other forms combined with the pulmonary presentation showed half the median survival time and a threefold risk of death from TB as compared to cases with the isolated pulmonary form. Of the deaths in patients with mixed forms, 19.7% had HIV co-infection or AIDS. In fact, in the model for the group classified as TB/HIV/AIDS co-infection, the mixed pulmonary/extra pulmonary form was the only factor independently associated with death, while in the group without TB/HIV/AIDS diagnosis, age at diagnosis, mixed pulmonary/extra pulmonary form, and previous treatment dropout remained significantly associated with risk of death from TB (Domingos *et al.*, 2008).

A study by Manissero *et al.* (2010) in European Union and European Economic Area, reported older age as a risk factor in undesirable treatment of TB. It has also been shown that mortality rate in TB patients increased with age and the highest mortality rate is observed in people aged 65 and more.

A cohort study in Finland found that among 629 cases a favourable outcome was achieved in 441 (70.1%), 17.2% (108) died and other unfavourable outcome took place in 12.7% (80). Significant independent risk factors for death were male sex, high age, non-HIV-related immunosuppressant and any other than a pulmonary specialty being responsible for stopping treatment (Vasankari *et al.*, 2007).

A retrospective study was conducted to determine the risk factors and causes of death among TB patients in Russia; Tomsk, Siberia. Of 1,916 patients who initiated treatment, 183 (9.6%) died during treatment. Old age, history of previous treatment for TB, multi-drug resistance TB and alcoholism have significance effect for death during treatment period (Mathew *et al.*, 2006).

The study aimed to estimate excess mortality among tuberculosis patients and identify risk factors for tuberculosis-associated mortality was conducted in the Netherlands. The survival probabilities for tuberculosis patients were estimated as 95% after six months of treatment and 6% died within one year while on treatment. The study identified the following significant factors for mortality of tuberculosis patients: male sex, age (>65 years), presence of malignancy, human immune-deficiency virus (HIV) infection, addiction to alcohol or drugs, localization of tuberculosis (pulmonary and extra pulmonary tuberculosis) and type of medical officer having made the diagnosis (specialist internal medicine). However, the prognosis was very good for those aged less than 65 yrs and without human immunodeficiency virus infection or a malignancy (Borgdorff *et al.*, 1998).

Studies in Kerman, Iran showed that treatment success was weak in patients whose sputum smear remained positive at the end of the second month of treatment. Unsuccessful treatment in those whose sputum smear was positive at the end of the second month of treatment was 16.9 times more than those whose sputum smear was negative at the end of the second month (Moosazadeh and Amiresmaili, 2011).

Talay *et al.* (2007) conducted a study in Turkey on factor associated with treatment success for tuberculosis patients. In multivariate regression analysis, risk factors for non-successful treatment outcome were determined by re-treatment patients, patients order than 46 years of age, and the presence of rifampicin resistance.

A study in Singapore aimed to identify risk factors associated with mortality among tuberculosis patients on treatment, older age and male sex were significant risk factors for death in TB patients (Low *et al.*, 2009).

Vasantha *et al.* (2008) conducted a study in South India. The results showed that 3818 TB patients who were initiated on treatment, 96%, 94% and 97% of category – I, II and III respectively, survived after completion of treatment. A multivariate analysis of Cox regression showed found that age (≥45), body weight at initiation of treatment (<35 kgs), previous history of treatment and alcoholism were risk factors associated with death during anti-tuberculosis treatment period.

A retrospective cohort study at District Tuberculosis Centre (DTC) in Yavatmal, India: A total of 716 patients were registered at the tuberculosis unit. The survival rates by the end of the intensive phase were 96%, 93% and 99% in categories I, II and III of DOTS, respectively. There was a significant difference in the survival curves amongst the three DOTS categories (log rank statistic= 7.26, d.f. = 2, P=0 0.02) and amongst the different age groups (log rank statistic= 8.78, d.f. = 3, P= 0.012). There was no difference in the survival curves of male and female patients (log rank statistic= 0.05, d.f. = 1, P= 0.80) and according to type of disease (log rank statistic= 5.63, d.f. = 2, P= 0.05). Cox proportional hazard analysis showed that age was important risk factor for death in tuberculosis patients. Age above 40 years was a significant risk factor for death in patients of tuberculosis (Pardeshi, 2009).

A study done by Lo *et al.* (2011) in Taiwan showed that, 50% of the deaths occurred within 2 months only 4% of the patients died of TB and 12.5% died of other diseases. Age and positive sputum bacteriology were the most significant predictors of death.

In Vietnam study, among 27 deceased patients the median time interval from treatment initiation to death was 8 months with inter quartile range (IQR 5 to 11 months). Also multidrug resistance may explain the high combined failure and death rate of 55% among patients with transfer-out (Vree *et al.*, 2007).

Increasing age has been noted as a risk factor for death and new disease in contrast to recurrent disease was also associated with high early mortality during the first 4 weeks of treatment for tuberculosis patients in Malawi (Harries *et al.*, 2001).

A study was conducted among patients registered at the Addis Ababa Tuberculosis Centre in Ethiopia to determine the rate of defaulting from treatment and factors associated with it. A high rate of defaulting, 82% was found. The rates of defaulting were higher in males, in the older age groups and in those living near to the TB centre. Social problems and feeling of improvement were the top two reasons for patients to default. Inadequate knowledge, low educational level, nearer distance and negative attitude toward the TB Centre were found to be statistically significant predictors for defaulting (Demissie and Kebede, 1994).

Getahun *et al.* (2011) conducted a study in Addis Ababa, Ethiopia. From a total of 6,450 registered TB patients 236(3.7%) died. More than 75% death occurred within eight month of treatment initiation. The mean and median times of survival starting from the date of treatment initiation were 7.2 and 7.9 months, respectively. Comparison of survival curves using Kaplan-Meier curves method with log-rank test showed that the survival status was significantly different between Ababa showed that the survival status was significantly different between Patient age, weight at initiation of anti-TB treatment, patient category, year of enrollment and treatment center ($P < 0.05$). The death rate of pulmonary positive, pulmonary negative and extra pulmonary TB patients were 2.7%, 3.6%, and 4.3%, respectively. In multivariable Cox's regression analysis patient age, weight at initiation of anti-TB treatment, patient category, year of enrolment and treatment center were significantly associated with time to death.

A study conducted in the Southern Region of Ethiopia showed that a total of 6,547 patients (55.6% male, 44.4% female) with mean age of 27.5 years were registered for treatment; 2,873 (43.9%) were smear-positive, 2,493 (30.1%) smear-negative and 1,157 (17.7%) had extra-pulmonary TB. Most (6,033, 92%) were new cases; 4,900 (74.8%) had a successful outcome and 1,095 (16.7%) a poor treatment outcome. Of those with poor outcome, 667 (60.9%) patients defaulted, 404 (36.9%) died and 24 (2.2%) failed treatment. Logistic regression analysis showed that age, sex, patient category, and TB classification were independent risk factors for poor outcome (Muñoz-Sellart *et al.*, 2010).

CHAPTER THREE

3 Data and Methodology

3.1. Study area and population

3.1.1. Study area

Ethiopia is Africa's oldest independent country. It is the tenth largest country in Africa, covering 1,104,300 square kilometres (with 1 million sq km land area and 104,300 sq km water) and is the major constituent of the landmass known as the Horn of Africa. It is bordered in the north-northeast by Eritrea, in the east by Djibouti and Somalia, in the south by Kenya, and on the west-southwest by Sudan and South Sudan.

Ethiopia is a country with great geographical diversity ranging from peaks up to 4,550m above sea level down to a depression of 110m below sea level. More than half of the country lies above 1,500 meters. The predominant climate type is tropical monsoon, with three broad climatic variations: the "Kolla", or hot lowlands, below approximately 1,500 meters, the "Wayna Dega" at 1,500- 2,400 meters and the "Dega" or cool temperate highlands above 2,400 meters (FMOH,2010).

Despite major strides to improve the health of the population in the last one and half decades, Ethiopia's population still face a high rate of morbidity and mortality and the health status remains relatively poor. The major health problems of the country are largely preventable communicable diseases and nutritional disorders (FMOH, 2010).

3.1.2 Population

The target population in this study is the full set of individuals who could be included in the retrospective study. It consists of all TB patients' male and female about whom follow-up time was available in the Addis Ababa Governmental health centers, Addis Ababa, Ethiopia.

3.1.3 The Data

The data was a retrospective cohort study based on TB patients that were registered in unit TB registers in the health facilities providing direct observed treatments strategy (DOTS) in six randomly selected Addis Ababa Governmental health centers, Addis Ababa, Ethiopia. In this study we used secondary data which was collected from patient follow-up records. Information for this study was extracted from documents of all TB cases registered from September 2012 to August 2013 in six DOTS clinic located in six randomly selected health centers.

3.2 Variables in the study

The explanatory (independent) variables of interest in this study include demographic factors, and disease and medicine related factors and characteristics of the disease. The response (dependent) variable is continuous; it is length of time of treatment for tuberculosis patients.

3.2.1 The Response Variable

The dependent variable or response is the waiting time until the occurrence of an event (dead=1, alive or censored=0). Observations are censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed.

3.2.2 Predictor Variables

Predictors or explanatory variables which are called covariates are those whose effect on the waiting time we wish to assess. The covariates can be categorical or continuous. In developing a model we should also establish if the covariate is time dependent or not. Such cases affect how the covariates would be modelled in Cox proportional hazard procedure. The predictor (covariate) variables which are assumed to influence the survival of TB patients included in the model are:

- ✓ Age
- ✓ Gender
- ✓ TB patients Category

- ✓ Type of TB
- ✓ Smear result
- ✓ Initial weight of patients
- ✓ HIV status

3.2.3 Definition of variables

Type of TB

According to the standard definitions of the Ethiopian National Tuberculosis and Leprosy control Program guideline (NLCP) (FMOH, 2008) type of TB was defined as:

i) **Pulmonary TB, smear-positive** (a patient with at least two sputum specimens which were positive for acid-fast bacilli (AFB) by microscopy, or a patient with only one sputum specimen which was positive for AFB by microscopy and chest radiographic abnormalities consistent with active pulmonary TB).

ii) **Pulmonary TB, smear negative** (a patient with symptoms suggestive of TB, with at least two sputum specimens which were negative for AFB by microscopy, and with chest radiographic abnormalities consistent with active pulmonary TB (including interstitial or miliary abnormal images), or a patient with two sets of at least two sputum specimens taken at least two weeks apart, and which were negative for AFB by microscopy, and radiographic abnormalities consistent with pulmonary TB and lack of clinical response to one week of broad spectrum antibiotic therapy) and.

iii) **Extra pulmonary TB** (EPTB) (this included tuberculosis of organs other than the lungs, such as lymph nodes, abdomen, genitourinary tract, skin, joints and bones, meninges, etc.).

Patient category

At entry to DOTS patients were categorized according to the Ethiopian NLCP national guideline as:

i) **New cases** (patient who has never had treatment for TB before or has been on anti- TB treatment for less than four weeks).

ii) **Relapse** (a patient who has been declared cure or treatment completed of any form of TB in the past but who reports back to the health service and is found to be AFB smear positive or culture positive).

iii) **Treatment failure** (a patient who, while on treatment remained smear-positive or become again smear-positive at the end of the five month or later after commencing treatment).

iv) **Returned after default** (a patient who had previously been recorded as defaulted from treatment and returns to the health facility and found to be smear positive sputum).

v) **Transfer in** (a patient who started treatment in one health facility and transferred to another health facility to continue treatment), and.

vi) **Other** (a patient who does not fit in any of the above mentioned categories).

Treatment Outcome

This variable was recorded as:

i) **Cured** (finished treatment with negative bacteriology result at the end of treatment).

ii) **Treatment completed** (patient finished treatment but no bacteriology result at the end of the treatment).

iii) **Treatment failure** (smear positive at five months despite correct intake of medication), defaulter (patients who interrupted their treatment for two consecutive months or more after registration), died (patients who died from any cause during the course of treatment).

iv) **Transferred out** (patients whose treatment results are unknown due to transfer to another health facility) and.

v) **Successfully treated** (the sum of patients who are declared "cured" and those who have "completed" treatment). Survival time was defined as the time in

days from the beginning of treatment to death from tuberculosis as the main or associated cause. Censoring occurred either at the end of the study or due to death from other causes and include defaulters and transfer outs.

3.2.4 Description of Variables in the study

Table 1: Response Variable

Variable	Category
Length of waiting Time	continuous

Table 2: Independent variables/covariates

Variable	Categories
Age	Continuous
Gender	0=Male 1=Female
TB patients Category	0= Non-new cases (Transfer in, Default, Relapse and others). 1= New cases
Type of TB	0= Category I(Pulmonary Positive) 1= Category II(Extra-Pulmonary) 2= Category III(Pulmonary Negative)
Smear result	0= Negative, 1= Positive
HIV status	0= HIV Negative, 1= HIV Positive
Initial weight of patients	continuous

3.3 Method of data analysis

3.3.1 Survival analysis

The study focused on time to event (time to death by tuberculosis), so the appropriate method of this particular study was survival analysis.

In many biomedical applications the primary endpoint of interest is time to a certain event. Examples are; time to death, time it takes for a patient to respond to a therapy, time from response until disease relapse (i.e., disease returns) etc.

We may be interested in characterizing the distribution of "time to event" for a given population as well as comparing this "time to event" among different groups (*e.g.*, treatment vs. control in a clinical trial or an observational study), or modelling the relationship of "time to event" to other covariates (sometimes called prognostic factors or predictors). Typically, in biomedical applications the data are collected over a finite period of time and consequently the "time to event" may not be observed for all the individuals in our study population (sample). This results in what is called censored data. That is, the "time to event" for those individuals who have not experienced the event under study is **censored** (by the end of study). It is also common that the amount of follow-up for the individuals in a sample vary from subject to subject. The combination of censoring and differential follow-up creates some unusual difficulties in the analysis of such data that cannot be handled properly by the standard statistical methods. Because of this, a new research area in statistics has emerged which is called Survival Analysis or Censored Survival Analysis (Klein and Moeschberger, 2003 and Kleinbaum and Klein, 2005).

Survival analysis is a branch of statistics which involves the modeling of time to event data.

To study, we must introduce some notation and concepts for describing the distribution of "Time to event" for a population of individuals. Let the random variable T denote time to the event of our interest. Of course, T is a positive random variable which has to be

unambiguously defined; that is, we must be very specific about the **start** and **end** with the length of the time period in-between corresponding to T.

Survival time (in general): measured from **birth** to **death** for an individual. This is the survival time we need to investigate in a life expectancy study.

Survival time of a treatment for a population with certain disease: measured from the time of **treatment initiation** until **death**.

Survival time due to heart disease: (the event is death from heart disease): measured from **birth** (or other time point such as treatment initiation for heart disease patients) to death caused by heart disease. (This may be a bit tricky if individuals die from other causes. This is competing risk problem. That is, other risks are competing with heart disease to produce an event (death).

Describing the Distribution of Time to an Event

The distribution of the random variable T can be described in a number of equivalent ways. There is of course the usual (cumulative) distribution function

$$F(t) = P(T \leq t),\ t \geq 0 \qquad (1)$$

If T is a continuous random variable, then it has a density function $f(t)$, which is related to $F(t)$ through following equations

$$F(t) = \frac{dF(t)}{dt},\quad F(t) = \int_0^t f(u)du \qquad (2)$$

The **survival function:**

$$S(t) = P(T \geq t) = 1 - F(t) \qquad (3)$$

Where, $S(t)$ is the probability that a randomly selected individual will survive beyond time t.

$S(t)$ is a non-increasing function with $S(0) = 1$. Typically, $S(\infty) = 0$.

$S(\infty) > 0$ is possible for survival data. Ex: Let T be the time from response to relapse of a curable disease.

$S(\infty)$ is the proportion of cured individuals.

If T is continuous, then

$$S(t) = \int_t^{\infty} f(u)du \quad \text{and} f(t) = -\frac{dS(t)}{dt} \tag{4}$$

- Mean survival: $\mu = E(T)$

Due to censoring, the sample mean of observed survival times will not be an unbiased estimator of μ. If we can estimate $S(t)$, then we can estimate μ taking advantage of

$$E(T) = \int_0^t S(t)dt \tag{5}$$

- Median Survival Time

The median survival time m is such that

$$S(m) = 0.5 \text{ iff. } \quad m = S^{-1}(0.5) \tag{6}$$

- Hazard Rate

The hazard rate $\lambda(t)$ is the instantaneous rate of failure; i.e. death at time t given that an individual is alive at time t. This is the limit of the mortality rate as the time interval goes to 0.

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \le T < t+h | T \ge t)}{h} \tag{7},$$

$$\text{Thus } \lambda(t) = \frac{\lim_{h \to 0} \frac{P(t \le T < h | T \ge t)}{h}}{P(T \ge t)} = \frac{f(t)}{S(t)} \tag{8}$$

$$\text{This gives } \lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{\log S(t)}{dt} \tag{9}.$$

Integrate both sides

$$\wedge(t) = \int_0^t \lambda(u)du = -[\log S(t) - \log S(0)] = -\log S(t) \tag{10}$$

$\wedge(t)$: is known as the cumulative hazard rate

Thus

$$S(t) = exp\{-\wedge(t)\} = exp\left\{-\int_0^t \lambda(u)du\right\} \tag{11}$$

3.3.2 Kaplan-Meier Estimator and comparison of survival function

The **Kaplan-Meier** or **product limit** estimator is the limit of the life-table estimator when intervals are taken so small that only at most one distinct observation occurs within an interval. Kaplan and Meier (1958) demonstrated in a paper that this estimator is "maximum likelihood estimate."

The **Kaplan-Meier** or **product limit** estimator is the extension of the empirical distribution function for censoring. The KM estimator is a nonparametric survival model that was first introduced by Edward Kaplan and Paul Meier in their only collaboration published in 1958 named Nonparametric Estimation from incomplete data (Kaplan and Meier, 1958).

The KM estimator estimates the probability of surviving to or past time t by only using information of the people at risk at time t. Formula of the KM survival estimator is

Let $d(x)$ be the number of deaths at time x. Generally, $d(x) = 0, 1$. But we allow the possibility of tied survival times. So, $d(x) > 1$ is possible.

Let $n(x)$ be the number of at-risk individuals at time x; i.e., the number of individuals who are neither dead nor censored at time x.

The KM estimate is then

$$\hat{S}^{KM}(t) = \prod_{x \leq t} \left\{1 - \frac{d(x)}{n(x)}\right\} \qquad (12)$$

Note: The product changes only at times $d(x) \geq 1$ (death).

The KM estimator produces survival estimates on the basis of independent censoring. The advantage of the KM estimator is that it does not hold any assumption about the distribution of the survival function. The KM plot, which is a step function, gives some indications about the shape of the survival distribution for more in depth analyses.

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. The figure in general shows if the pattern of one

survivorship function lying above another which means the group defined by the upper curve lived longer, or had a more favourable survival experience, than the group defined by the lower curve.

In biomedical applications, it has become common practice to use nonparametric tests; that is, using test statistics whose distribution under the null hypothesis does not depend on specific parametric assumptions on the shape of the probability distribution. With censored survival data, we use the following tests;

Log-rank Test

The log-rank test (also known as the Mantel log-rank test, the Cox Mantel log-rank test, and the Mantel- Haenszel test) is the most commonly used test for comparing survival distributions. It is applicable to data where there is progressive censoring and gives equal weight to early and late failures. It assumes that hazard functions for the two groups are parallel. The test takes each time point when a failure event occurs and a 2 x 2 table showing the number of deaths and the total number of subjects under follow up is created. For each table the observed deaths in each group, the expected deaths and the variance of the expected number are calculated. These quantities are summed over all tables to yield a chi-square statistic with 1 degree of freedom (known as the Mantel-Haenszel or log-rank test statistic). The log-rank test calculations also produce for each group the observed to expect ratio which relates the number of deaths observed during the follow up with the expected number under the null hypothesis that the survival curve for that group would be the same as that for the combined data. In order to measure the strength of the evidence against the null hypothesis, we must be able to evaluate the distribution of the test statistic (at least approximately) under the null hypothesis. Specifically, the weighted log-rank test statistic is given by

The log-rank test statistic for testing H_0: $S_0(t) = S_1(t)$ is

$$T(w) = \frac{\sum_x w(x)\left[dN_1(x) - \frac{dN(x) \times Y_1(x)}{Y(x)}\right]}{\left\{\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x)-dN(x)]}{Y^2(x)[Y(x)-1]}\right]\right\}^{1/2}} \qquad (13)$$

Where, $w(x)$ is weight for censor adjustment at failure time x.

$dN_0(x)$ and $dN_1(x)$ be the number of deaths observed at time x from group 0 and 1, respectively

$Y_0(x)$ and $Y_1(x)$ be the number of individuals at risk at time x from group 0 and 1, respectively

$Y(x) = Y_0(x) + Y_1(x)$ be the total number at risk at time x.

$dN(x)$ be total number of death observed at time x.

Under the null hypothesis, this test statistic is approximately distributed as a standard normal

$$T(w) \sim N(0, 1)$$

Therefore, a level α test (two-sided) will reject H_0: $S_0(t) = S_1(t)$ whenever

$$|T(w)| \geq z_{\alpha/2}$$

Breslow's test (also known as Gehan's generalised Wilcoxon test) is applicable to data where there is progressive censoring. It is more powerful than the log-rank test when the hazard functions are not parallel and where there is little censoring. It has low power when censoring is high. It gives more weight to early failures. The Cox Mantel test is similar to the log-rank test. It is applicable to data where there is progressive censoring. More powerful than Gehan's generalised Wilcoxon test. The Peto and Peto modification of the Gehan- Wilcoxon test is similar to Breslow's test and is used where the hazard ratio between groups is not constant. Cox's F test is more powerful than Breslow's test if sample sizes are small.

Generally,

The log-rank test uses $w(x) = 1$ for all x.

Gehan's generalization of the Wilcoxon test uses $w(x) = Y(x)$.

Peto-Prentice's generalization of the Wilcoxon test uses $w(x)$ = Kaplan-Meier estimator using the combined sample (Kalbfleisch and Prentice, 2002).

$$\prod_{u<x}\left[1 - \frac{dN(x)}{Y(x)}\right] \qquad (14)$$

3.3.3 The Proportional Hazards Model

The **Cox proportional hazards (PH) model,** a popular mathematical model used for analyzing survival data. A proportional hazards model proposed by Cox (1972) assumes that

$$\lambda(t|z) = \lambda_0(t)e^{Z^T\beta} \qquad (15)$$

Where, $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$

$\boldsymbol{Z}$ is a $p \times 1$ vector of covariates such as treatment indicators, prognostic factors, etc., and β is a $p \times 1$ vector of regression coefficient.

Obviously,

$$\lambda(t \mid Z = 0) = \lambda_0(t) \qquad (16)$$

So $\lambda_o(t)$ is often called the baseline hazard function. It can be interpreted as the hazard function for the population of subjects with $z = 0$. The only requirement is that $\lambda_o(t) > 0$. $\lambda_o(t)$ is the nonparametric part of the model and $z^T\beta$ is the parametric part of the model. So Cox's proportional hazards model is a semi parametric model.

Assumptions:

- The baseline hazard , $\lambda_o(t)$ depends on t, but not on covariates Z
- The hazard ratio, that is, $exp(z^T\beta)$, depends on the covariates Z, but not on time t.
- The covariates Z do not depend on time t.
- Time is measured on a continuous scale.
- Censoring occurs randomly.

Interpretation

The survival function of the subpopulation with covariate z is

$$S(t|Z) = [S_0(t)]^{\exp(Z^T\beta)} \qquad (17)$$

Where, $$S_0(t) = e^{-\int_0^t \lambda_0(t)dt} \qquad (18)$$

For any two sets of covariates $\mathbf{z}_0$ and $\mathbf{z}_1$

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_0)} = e^{(Z_1 - Z_0)^T \beta} \quad (19)$$

This is constant. Hence, a proportional hazards model with one unit increase in z_k while other covariates being held fixed, we get

$$\frac{\lambda(t|Z_k+1)}{\lambda(t|Z_k)} = e^{\beta_k} \quad (20)$$

It is clear that β characterizes the "effect" of z. So it is the focus of our inference while $\lambda_0(t)$ is regarded as a nuisance parameter.

A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox PH model is a "robust" model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

3.3.4 Parameter Estimation

The formula for the Cox model likelihood function is actually called a "partial" likelihood function rather than a (complete) likelihood function. The term "partial" likelihood is used because the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. Thus the likelihood for the Cox model does not consider probabilities for all subjects, and so it is called a "partial" likelihood. Since the baseline hazard $\lambda_0(t)$ is completely unspecified (infinite dimensional), ordinary likelihood methods cannot be used to estimateβ. Cox proposed the method of partial likelihood to remove the nuisance parameter $\lambda_0(t)$ from the proposed estimating.

Let $Y_i(x)$ denote the indicator of whether the i^{th} individual is at risk at time x. That is, $Y_i(x) = I(X_i \geq x)$.

Let us break the time axis (patient time) into a grid of points. Assume the survival time is continuous. We hence can take the grid points dense enough so that at most one death can occur within any interval.

Let $dN(x) = \sum_{i=1}^{n} dNi(x)$ be the total number of deaths in $[x, x + \Delta x)$. Since Δx is small, we can assume that $dN(x) = 0, 1$.

Let $Y(x) = \sum_{i=1}^{n} Y_i(x)$ be the total number at risk at time x.

The probability that an individual dies at time x_i with covariates $\boldsymbol{Z}(i)$, given one of the individuals in $Y(x_i)$ dies at this time, is given by

$$P[\text{individual dies at } x_i \mid \text{one death at } x_i] = \frac{P(\text{individual dies at } x_i | \text{survival to } x_i)}{P(\text{one death at } x_i | \text{survival to } x_i)}$$

$$= \frac{\lambda_0(t_i)exp[\boldsymbol{\beta}^T \boldsymbol{Z}(i)]}{\sum_{J \epsilon Y(x_i)} \lambda_0(t_i)exp[\boldsymbol{\beta}^T \boldsymbol{Z}(i)]}$$

$$= \frac{exp[\beta Z(i)]}{\sum_{J \varepsilon Y(x_i)} exp[\beta Z(i)]} \qquad (21)$$

The partial likelihood is formed by multiplying these conditional probabilities over all deaths, so we have the likelihood function

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(Z^T{}_i \beta)}{\sum_{j=1}^{n} \exp(Z^T{}_j \beta) Y_j(x_i)} \right\}^{\delta_i} \qquad (22)$$

Where, δ_i an indicator of censoring for the i^{th} individual given by 0 for censored and 1 for event/death

This is treated as a usual likelihood, and inference is carried out by usual means. It is of interest to note that the numerator of the likelihood depends only on information from the individual who experiences the event, whereas the denominator utilizes information about all individuals who have not yet experienced the event (including some individuals who will be censored later).

Partial Likelihoods When Ties Are Present

A central assumption in survival analysis is that time is continuous. Sometimes (particularly in veterinary epidemiological research) the outcome of interest is not measured on a continuous scale and outcome events may occur simultaneously (e.g. service number when conception occurred). When the number of tied events is large, approximate methods yield regression coefficients that are biased towards zero. There are

three common methods for dealing with ties: 1. Breslow approximation, 2. Efron approximation and 3. Exact partial likelihood

The Breslow estimator is:

$$\widehat{\lambda}_o(t) = \sum_{x<t} \left[\frac{dN(x)}{\sum_{i=1}^{n} e^{\hat{\beta}^T Z_i Y_i(x)}} \right] \qquad (23)$$

The importance of using the partial likelihood is that this function depends only on the parameter of interest, and is free of the baseline hazard λ_0 (t), which is infinite dimensional nuisance function. Cox suggested treating *PL* as a regular likelihood function and making inference on β accordingly. For example, we maximize the partial likelihood to get the estimate of β, often called MPLE (maximum partial likelihood estimate), and use the minus of the second derivative of the log partial likelihood as the information matrix, etc.

3.3.5 Tests to assess the significance of the coefficient

There are three different tests to assess the significance of the coefficients: the partial likelihood ratio test, the Wald test and the score test.

i) The partial likelihood ratio test

The partial likelihood ratio test, G, is computed as twice the difference between the log partial likelihood of the model containing covariates and the log partial likelihood of the model not containing the covariates. Specifically,

$$G = 2\{L_P(\hat{\beta}) - L_P(0)\} \qquad (24)$$

Where, $Lp\ (0) = -\sum_{i=1}^{m} ln(n_i)$ and n_i denotes the number of subjects in the risk set at observed survival time $t(i)$.

Under the null hypothesis, that the coefficients are equal to zero, that is H_0: $\beta = (0, 0...$ 0), the statistic, G, follows a chi-square distribution with 1 degree-of-freedom for "sufficiently" large sample size. This distribution can be used to obtain p-values to test the significance of the coefficient.

ii) The Wald test

The Wald statistic is defined as the ratio of the estimated coefficient to its estimated standard error.

$$Z = \frac{\hat{\beta}}{\hat{S}E(\hat{\beta})} \qquad (25)$$

Under the null that is a single parameter H_0: $\beta_i=0$, the Wald statistic follows a standard normal distribution (i.e., $z \sim N(0, 1)$). Obviously, the square of Wald statistic follows a chi square distribution with 1 degree-of-freedom.

iii) The score test

The score test which is obtained by computing the ratio of the derivative of the log partial likelihood to the square root of the observed information, all evaluated at β=0.
The partial log-likelihood is

$$L(\beta) = \sum_x dN(x)\left[Z^T{}_i(x)\beta - \log\left(\sum_{j=1}^{n} \exp\left(Z^T{}_j\beta\right) Y_j(x)\right)\right] \qquad (26)$$

The equation for the score test is

$$Z^* = \frac{\partial L_P/\partial\beta}{\sqrt{I(\beta)}}\Big|_{\beta=0} \qquad (27)$$

Under the null hypothesis that each parameter is equal to zero, H_0: $\beta_i=0$, this statistic follows a standard normal distribution. The score test, when computed by a software package, may be reported as the square of the value of z^* , which will follow a chi-square distribution with 1 degree-of-freedom under the null hypothesis.

In practice, the numeric values of the three tests ($\sqrt{G}$, Z and z^*) should be quite similar and thus lead to draw the same conclusion about the significance of the coefficient. In situations where there is a disagreement, making it necessary to choose one test, the partial likelihood ratio test is the preferred choice.

In this study we used SAS and STATA statistical software for model fitting process.

3.4 Model building

Selection of covariates

An initial step in model building process is to identify a set of explanatory variables that have the potential for being included in the linear component of a proportional hazards model. This set will contain those variates and factors that have been recorded for each individual, but additional terms corresponding to interactions between factors or between variates and factors may also require.

Model building in proportional hazards regression analysis requires critical decisions in selecting subsets of covariates as it is likely that more covariates are present in real life problems; selection of interaction terms to be included in the model and checking the linearity of continuous covariates and choosing the appropriate transformation for non-linear covariates. The methods of selecting a subset of covariates in a proportional hazards regression model are essentially similar to those used in any other regression models. The most common methods are purposeful selection, step-wise (forward selection and backward elimination) and best sub-set selections (Hosmer and Lemeshow, 1999).

Recommendable procedure in selecting variables in the study

Hosmer and Lemeshow (1999) and Collett (2003) recommended the following procedure in variable selection.

1. Include all variables that are significant in the univariable analysis at the 20 to 25 percent level and also any other variables which are presumed to be clinically important to fit the initial multivariable model.

2. The variables that appear to be important from step 1 are then fitted together in a multivariable model. In the presence of certain variables others may cease to be important. Consequently, backward elimination is used to omit non-significant variables (i.e. those variables that do not significant increase the value of $-2log\hat{L}$

when they are omitted) from the model. We therefore compute the change in the value of $-2log\hat{L}$ when each variable on its own is omitted from the set. Only those that lead to a significant increase in the value of $-2log\hat{L}$ are retained in the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn should be examined.

3. Variables, that were not important on their own, and so were not under consideration in step 2, may become important in the presence of others. These variables are therefore added to the model from step 2, with forward selection method (i.e. any that reduce $-2log\hat{L}$ significantly are retained in the model). This process may result in terms in the model determined at step 2 ceasing to be significant.

4. A final check is made to ensure that no term in the model can be omitted without significant increasing the value of $-2log\hat{L}$ and that no term not included significantly reduces $-2log\hat{L}$. At this stage the interactions between any of the main effects currently in the model can be considered for inclusion if the inclusion significantly modifies the model.

In searching for possible confounders, it is useful to examine the simple relationship between the other explanatory factors and survival, adjusted for the factor of interest because, if there is obviously no relationship between a factor and survival, then, it is not likely to be a confounder. Thus, the next step is to consider the relationship between each of the other explanatory factors and survival, given that the factor stated in the basic hypothesis is already in the model. This process is continued by exploring the relationships between each of the remaining explanatory variables and survival, given that the factor stated in the basic hypothesis and the one next most related to survival (assuming that the basic variable is in the model) are in the model. If no significant confounders are found at any step in this process, then we stop and base our inference about the primary hypothesis on the last model.

3.5 Checking for linearity of continuous covariates

To check the linearity of continuous covariates we plot hazards against the midpoints of the class and also using plots of martingale residuals.

There are two residual-based plots:

1. A plot of the covariate values versus the Martingale residuals (and their smooth) from a model that excludes the covariate of interest, and
2. A plot of the covariate values versus the log of the ratio of smoothed censor to smoothed cumulative hazard.

To construct the second plot:

1. Fit the preliminary main effects model, including the covariate of interest (let x is included covariate).

2. Save the Martingale residuals (M_i) from this model,

3. Calculate $H_i = c_i - M_i$, where c_i is the censoring variable,

4. Plot the values of c_i versus the covariate of interest and calculate a lowess smooth (called cLSM),

5. Plot the values of H_i versus the covariate of interest and calculate a lowess smooth (called HLSM),

6. The smoothed values from these plots are used to calculate:

$$y_i = ln\left(\frac{cLSM}{HLSM}\right) + \beta_x * X_i \qquad (28),$$

and the pairs (y_i, X_i) are plotted and connected by straight lines. There should be a linear relationship between the covariate values and each of the described parameters.

3.6 Testing the proportional hazards assumption

Once a suitable set of covariates has been identified, it is wise to check each covariate to ensure that the proportional hazards assumption is valid. To assess the proportional hazards assumption we examine the extent to which the estimated hazard curves for each level of strata of a covariate are equidistant over time.

A number of different tests for assessing the PH assumption have been proposed in the literature. We present the test of Harrel and Lee (1986), a variation of a test originally proposed by Schoenfeld (1982) and based on the residuals defined by Schoenfeld, now called the Schoenfeld residuals. For each predictor in the model, Schoenfeld residuals are defined for every subject who has an event.

A plot of the scaled Schoenfeld residuals and a loess smoother as a function of time may be used to test proportionality of hazards. In a 'well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0. The idea behind this test is that if the proportional hazards assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time.

The implementation of the test can be thought of as a three-step process:

1. Run a Cox proportional hazards model and obtain the Schoenfeld residuals for each predictor,

2. Create a variable that ranks the order of failures (the subject who has the first (earliest) event gets a value of 1, the next gets a value of 2, and so on,

3. Test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is zero.

Suppose we have a time-independent covariate Z and we have the Cox PH model

$$\lambda(t|Z) = \lambda_0(t)e^{\beta Z} \qquad (29)$$

And hence

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_0)} = e^{\beta(Z_1 - Z_0)} \qquad (30)$$

Suppose to test whether the hazard ratio changed over time. Consider the following model:

$$\lambda(t|Z) = \lambda_0(t)e^{\beta Z + \gamma g(t)} \qquad (31)$$

Where $g(t)$ is some specified function of time (eg. $g(t) = log(t)$) chosen by the data analyst

For such a model, we have

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_0)} = e^{(Z_1 - Z_0)(\beta - \gamma g(t))} \qquad (32)$$

This model allows the hazard ratio to change over time giving us greater flexibility than proportional hazards assumption. In addition, testing whether or not γ is significantly from zero allows us the opportunity to evaluate the proportional hazards assumption. The model (31), we can be viewed as a Cox PH model with two components:

i) A time-independent covariate Z

ii) A time-dependent covariate $g(t)\,Z$.

Let further we use an indicator functions over intervals of time $0<\tau_1< \ldots <\tau_{k-1}<\infty$:

$I_1(t) = I[t \in [0,\tau_1)]$, $I_2(t) = I[t \in [\tau_1,\tau_2)] \ldots I_{K-1}(t) = I[t \in [\tau_{K-2},\tau_{K-1})]$ and $I_K(t) = I[t \in [\tau_{K-1},\infty)]$. We then define the model:

$$\lambda(t|Z) = \lambda_0(t)\, exp\{\beta Z + \gamma_1 Z I_1(t) + \gamma_2 Z I_2(t) + \cdots + \gamma_{K-1} Z I_{K-1}(t)\} \quad (33)$$

For such a model, we have:

$$\log\left(\frac{\lambda(t|Z_1)}{\lambda(t|Z_0)}\right) = (Z_1 - Z_0)(\beta + \gamma_1 I_1(t) + \cdots + \gamma_{K-1} I_{K-1}(t)) \qquad (34)$$

An omnibus test for the proportional hazards assumption can be obtained by testing $H_0: \gamma_1 = \cdots = \gamma_{K-1} = 0$, which would yield a chi-square test with (K - 1) degrees of freedom. We can use likelihood-based tests like Wald test. If $\gamma = 0$ is not rejected, β_i's are not time varying coefficients and hence the proportional hazards assumption is

satisfied. If $\gamma = 0$ is rejected then the proportional hazards assumption is not satisfied and we have to look for another model.

An important point about this approach is that the null hypothesis is never proven with a statistical test (the most that can be said is that there is not enough evidence to reject the null) and that p-values are driven by sample size. A gross violation of the null assumption may not be statistically significant if the sample is very small. Conversely, a slight violation of the null assumption may be highly significant if the sample is very large.
For categorical covariates the proportional hazards assumption can be visually tested by plotting –log [-log S (t)] vs. time for strata of each covariate. If the proportionality assumption holds the two (or more) curves should be approximately parallel and should not cross. Alternatively, run a model with each covariate (individually).

3.7 Identification of influential and poorly fit data

Another important aspect of model evaluation is a thorough examination of the regression diagnostic statistics to identify which, if any, subjects (1) have an unusual configuration of covariates, (2) exert an illegitimate influence on the estimates of the parameters and/or (3) have an undue influence on the fit of the model. Statistics similar to those used in linear and logistic regression are available to perform these tasks with a fitted proportional hazards model. There are some differences in the types of statistics used in linear and logistic regression and proportional hazards regression, but the essential ideas are the same in all the three settings (Hosmer and Lemeshow, 1999).

Leverage is a diagnostic statistic that measures how "unusual" the values of the covariates are for an individual. In linear and logistic regression leverage is the distance of the value of the covariates for a subject to the overall mean of the covariates. Leverage is not easily defined nor does it have the same nice properties in proportional hazards regression. This is due to the fact that subjects may appear in multiple risk sets and thus may be present in multiple terms in the partial likelihood.

The score process residual for the i^{th} subject on the k^{th} covariate may be expressed as

$$L_{ik} = \sum_{j=1}^{n} (x_{ik} - \bar{x}_{wjk})\, dM_i(t_j) \qquad (35)$$

is a weighted average of the distance of the value, x_{ik}, to the risk set means, x_{wjk} , where the weights are the change in martingale residual $dM_i(t_j)$ defined as.

$$dM_i(t_j) = dN_i(t_j) - Y_i(t_j)\, exp(x_j^T \beta)\lambda_0(t_j) \qquad (36)$$

where $dN_i(t_j)$ is the change in the count function for the this subject at time t_j , always equal to zero for censored subjects and one for non-censored subjects, at actual observed survival time. The function $Y_i(t_j)$ is called the risk process and is defined as follows:

$$Y_i(t_j) = \begin{cases} 1 \; if \; t_i \geq t_j \\ 0 \; if \; t_i < t_j \end{cases} \qquad (37)$$

and $\lambda_0(t_j)$ is the value of $\frac{\delta^i}{\sum_{j \varepsilon R(t)} \exp\,(x_j^T \beta)}$ evaluated at t_j. The net effect is that, for continuous covariates, the score residuals have the linear regression leverage property that the further the value is from the mean the larger the score residual is, but "large" may be either positive or negative. Thus, the score residuals are sometimes referred to as the leverage or partial leverage residuals. We plot score residuals against each continuous covariates to observe if there is individuals far away from the mean. Hence, continuous covariates are most amenable to having their score residuals examined graphically. The graph for the dichotomous are less interesting in that all of the values fall on two vertical bands at zero and one, the two covariate values.

In many occasions, the influence that each observation has on the estimated hazard function will be of interest, and it will then be important to identify observations that influence the complete set of parameter estimates in the model. In other words, it may happen that the structure of the fitted model is particularly sensitive to one or more observations in the data set. Such observations can be analyzed through diagnostics that are designed to highlight observations that influence the complete set of parameter estimates in the linear predictor. This could be done by fitting the model to all n observations in the data set, and then fitting the same model to the sets of n-1 observations obtained by omitting each of the n observations in turn (Collett, 2003).

To achieve this purpose, to examine influence in the proportional hazards setting, we need to use statistics analogous to Cook's distance in linear regression. This is denoted as

$$\Delta\hat{\beta}_{ik} = \hat{\beta}_k - \hat{\beta}_{k(-i)} \qquad (38)$$

Where $\hat{\beta}_k$ denotes the partial likelihood estimator of the coefficient computed using the entire sample of size n and $\hat{\beta}_{k(-i)}$ denotes the value of the estimator if the i^{th} subject is removed (Hosmer and Lemeshow, 1999 and Collett, 2003). Thes shows that an approximate estimator of (38) is the k^{th} element of the vector of coefficient changes

$$\Delta\hat{\beta}_i = (\hat{\beta} - \hat{\beta}_{(-i)}) = \widehat{var}(\hat{\beta})\hat{L}_i \qquad (39)$$

Where $\hat{L}_i$ is the vector of score residuals and $\widehat{var}(\hat{\beta})$ is the estimator of the covariance matrix of the estimated coefficients.

3.8 Residuals

Residuals analysis provides information for evaluating a fitted proportional hazards model. They identify leverage and influence measures and can be used to assess the proportional hazards assumption. By definition, residuals for censored observations are negative and residual plots are useful to get a feeling for the amount of censoring in the data set large amounts of censoring will result in 'banding' of the residual points. There are three types of residuals:

1. Martingale residuals. Martingale residuals are the difference between the observed number of events for an individual and the conditionally expected number given the fitted model, follow up time, and the observed course of any time-varying covariates. Martingale residuals may be plotted against covariates to detect non-linearity (that is, an incorrectly specified functional form in the parametric part of the model). Martingale residuals are sometimes referred to as Cox-Snell or modified Cox-Snell residuals.

2. Score residuals. Score residuals should be thought of as a three-way array with dimensions of subject, covariate and time. Score residuals are useful for assessing individual influence and for robust variance estimation. Consider the model (33) and the

hypothesis of proportional hazardsH_0: $\gamma_1 = \cdots = \gamma_{K-1} = 0$. The partial likelihood of $\theta = (\beta, \gamma_1 \ldots \gamma_{K-1})'$ is

$$PL\,(\theta) = \prod_{i=1}^{n} \left[\frac{e^{\beta Z_i + \sum_{j-1}^{K-1} \gamma_j Z_i I_j(X_i)}}{\sum_{K=1}^{n} e^{\beta Z_i + \sum_{j-1}^{K-1} \gamma_j Z_i I_j(X_i) Y_j(X_i)}} \right]^{\Delta_i} \tag{40}$$

The log partial likelihood is

$$L(\theta) = \sum_{i=1}^{n} \Delta_i \left[\beta Z_i + \sum_{j-1}^{K-1} \gamma_j Z_i I_j(X_i) \right] - \sum_{i=1}^{n} \log \left[\sum_{K=1}^{n} e^{\beta Z_i + \sum_{j-1}^{K-1} \gamma_j Z_i I_j(X_i) Y_j(X_i)} \right] \tag{41}$$

Now the score function is

$$\frac{\partial L}{\partial \gamma_j} \Big|_{\hat{\beta}(H_0)} = \sum_{i=1}^{n} \Delta_i I_j(X_i) \left[Z_i - \frac{\sum_{K=1}^{n} Z_K e^{\hat{\beta} Z_K Y_K(X_i)}}{\sum_{K=1}^{n} e^{\hat{\beta} Z_K Y_K(X_i)}} \right]$$

$$= \sum_{i=1}^{n} \Delta_i \left[Z_i - \bar{Z}\left(X_i, \hat{\beta}\right) \right] I_j(X_i) \tag{42}$$

Where, $\bar{Z}\left(X_i, \hat{\beta}\right)$ is the weighted average of the covariate for individuals at risk at time X_i.

$SCH_i = \Delta_i \left[Z_i - \bar{Z}\left(X_i, \hat{\beta}\right) \right]$ is known as the Schoenfeld residual. We can write the score as:

$$\frac{\partial L}{\partial \gamma_j} = \sum SCH_i \times I\{X_i \epsilon [\tau_{j-1}, \tau_j]\} \tag{43}$$

If PH, then on average these residuals should be 0. A noticeable trend away from 0 may be indicative of the lack of PH.

3. Schoenfeld residuals. Schoenfeld residuals are useful for assessing proportional hazards. Schoenfeld residuals provide greater diagnostic power than unscaled residuals. Sometimes it referred to as score residuals.

If we have a model with several covariates,

$$\lambda(t|Z) = \lambda_0(t) e^{\beta_1 Z_1 + \cdots + \beta_q Z_q} \tag{44}$$

For each covariate $Z_j, j = 1 \ldots q$, we can compute

$$SCH_{ij} = \Delta_i \left[Z_{ij} - \frac{\sum_{K=1}^{n} Z_{Kj} I_j(X_i) e^{\hat{\beta} Z_K Y_K(X_i)}}{\sum_{K=1}^{n} e^{\hat{\beta} Z_K Y_K(X_i)}} \right] \tag{45}$$

and produce q plots of SCH_{ij} versus X_i.

3.9 Overall goodness-of-fit

To assess the overall goodness-of-fit of a Cox proportional hazards regression model Arjas (1988) suggests plotting the cumulative observed versus the cumulative expected number of events for subjects with observed (not censored) survival times. If the model fit is adequate, then the points should follow a 45 degree line beginning at the origin. The methodology is as follows: (1) create groups based on covariate values (e.g. treated yes, treated no) and sort on survival time within each group, (2) compute the cumulative sum of the zero-one censoring variable and the cumulative sum of the cumulative hazard function within each group, (3) plot the pairs of cumulative sums within each group only for subjects with an observed survival time.

CHAPTER FOUR

4. Statistical Data Analysis and Discussion

4.1 Introduction

In this section summary statistics of our data, data analysis, interpretation and discussion are presented. The censoring indicator (status) is 0 for censored observations and 1 for event/death. The study, begins with present's summary statistics of factors considered in the study, then comparison of the survival time of risk factors associated with death in TB patients, fitting the model , check the adequacy of the model and interpretation and discussion of the results. The statistical packages SAS and STATA have been used to analyze the data.

4.2 Summary Statistics

Out of the total 826 registered TB patients 105(53 male and 52 female) or (12.71%) died during the study period and 712(87.29%) were censored. The age group (≥ 45 years) showed the highest percentage (18.3%) with respect to death proportions than the other two age groups. In TB patient category non-new case had higher percentage (21.93%) of death. The percentage of death 10.4%, 13.99% and 12.99% occurred in the patients with pulmonary positive, pulmonary negative and extra pulmonary types of TB, respectively, and the patients with positive smear result had lower death percentage. HIV positive TB patients are highest risk group for death i.e. (22.18%). Patients with body weight at initiation of treatment (≥ 35kg's) had lower risk group for death (Table 4.1).

Table 4.1: Characteristics of Tuberculosis patient data under DOTS from Six randomly selected governmental health centers in Addis Ababa, Ethiopia September, 2011- August 2012.

Patients characteristics		Summary of the Number of death and censored Values				
		Total	Death	Percent death	Censored	Percent censored
Gender	Male	414	53	12.80	361	87.20
	Female	412	52	12.62	360	87.38
Age categories	0-24	284	22	7.75	262	92.25

	25-44	389	55	14.14	334	85.86
	≥45	153	28	18.30	125	81.70
TB patients cat	New	712	80	11.24	632	88.76
	Non-new	114	25	21.93	89	78.07
Type of TB	Pul.positive	202	21	10.40	181	89.60
	Pul.negative	293	41	13.99	252	86.01
	Extra pulmonary	331	43	12.99	288	87.01
Smear result	Positive	202	21	10.4	181	89.60
	Negative	624	84	13.46	540	86.54
HIV status	Positive	239	53	22.18	186	77.82
	Negative	587	52	8.86	535	91.14
Initial weight	<35	62	13	20.97	49	79.03
	≥35	764	92	12.04	672	87.96

4.3 Descriptive survival analysis

Table 4.2 exhibits that out of 826 TB patients, 721 patients were censored (87.29%) and 105 patients died (12.71%). The median follow-up time was 168 days for patients that are censored (range, 15 to 284 days), 25% of the patients had 176 days of follow-up (upper quartile). The median time of death was 52 days (ranges, 1 to 190 days).

Table 4.2: Summary statistics of status of TB patients and days of follow up time

Status of TB patients		
status	Frequency	Percent
Censored	721	87.29
Dead	105	12.71

Status of patients	N Obs	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Min.	Max.
Censored	721	170.791	18.291	168.000	168.000	176.000	15.000	284.000
Dead	105	63.867	47.012	52.000	29.000	88.000	1.000	190.000

Next, we see whether the observed differences in data summary among different factors are statistically significant or not with the help of Kaplan-Meier survival estimates and the log-rank test. The following graph (Fig 4.1) of the estimate of overall Kaplan-Meier survival curve shows the cumulative proportion of TB patients over time and the plot shows that most of the event/deaths occurred in the earlier months of TB treatment initiation.

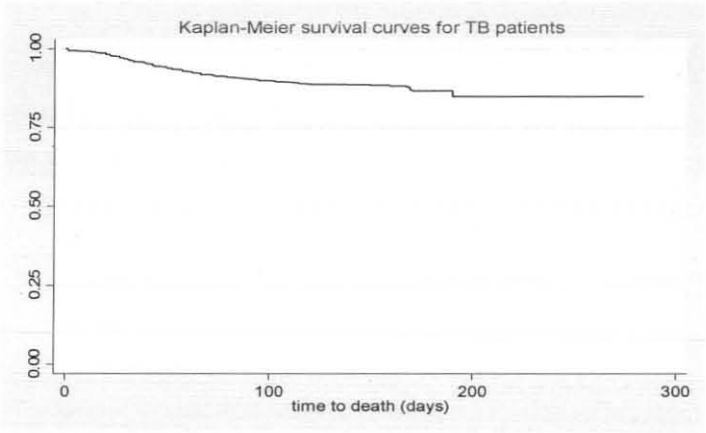
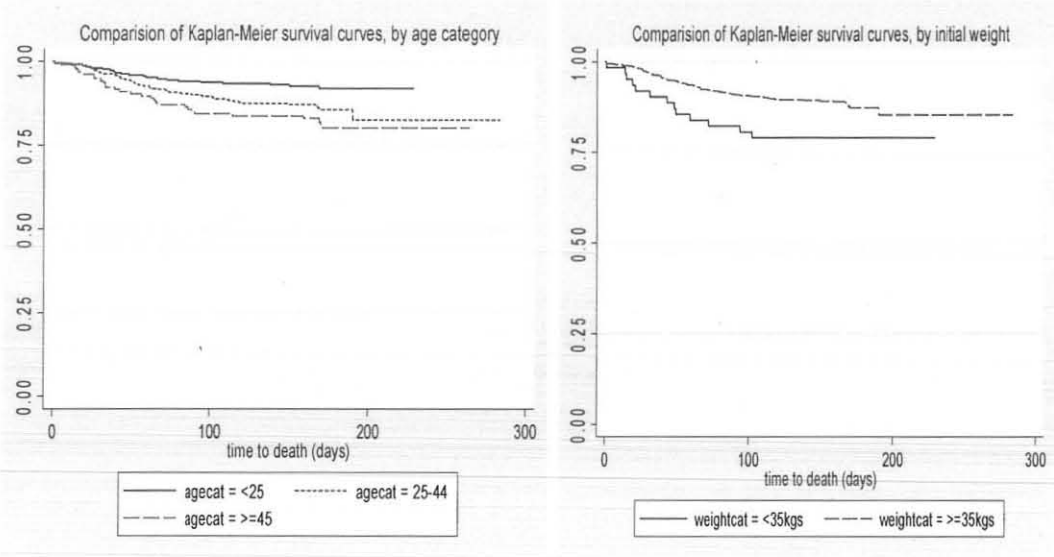


Figure 4.1: The plot of the overall estimate of Kaplan-Meier survival curves of TB under DOTS in AA health centers.

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. The figure in general shows the pattern of one survivorship function lying above another which means the group defined by the upper curve lived longer, or had a more favourable survival experience, than the group defined by the lower curve. From (Fig 4.2), there were differences among survivor curves of age category, initial weight, TB patient category and HIV status for TB patients. However, there are not clear differences among survivor curves of gender, smear result and type of TB. For detail see Appendix D (Fig a-g).

The estimated survival functions for the three age groups are shown in Figure 4.2(a). We see that the survival function for age group less than 25 years lies above those of the other two groups. This means that patients with age group less than 25 years had better survival probability than both age groups; the age groups 25-44 years survived longer than the age 45 years and above. In particular, those patients with age group age 45 years and above to had a smallest survival probability than both age groups. The plots of the survival functions in Figure 4.2(b) indicate a larger difference in the survival probability between the two groups over time. Patients whose initial weight was less than 35kgs had lived short experiences. Figure 4.2 (c) shows that patients new case TB groups had live longer than non-new case. Figure 4.2(d) exhibits that patients with HIV positive had worse survival experiences.

a) b)



c) d)

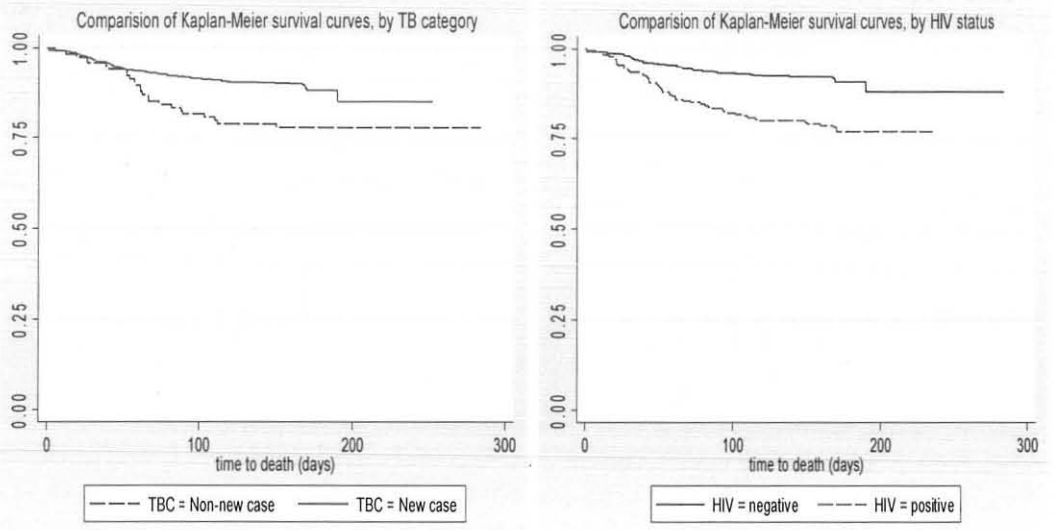


Figure 4.2: The plot of the estimate of Kaplan-Meier survivor curves of TB patients under DOTS in AA health centers a) age category, b) initial weight of patients, c) TB patient category and d) HIV status.

After comparing groups of subjects by survival curves, the next step is to conduct a homogeneity test across strata. The null hypothesis is that there is no difference among the survival curves. The statistical tests often employed are the log-rank test and the

Wilcoxon test. Under the null hypothesis the log-rank test statistic is approximately distributed as chi-square in large samples with K-1 degrees of freedom, where K denotes the number of groups being compared. The Wilcoxon test allows early failures to receive more weights than later failures whereas the log-rank test gives equal weight to failures at all failure time.

Testing Homogeneity of Survival Curves for survival time over Strata

Table 4.3: Results of the Log-rank test, Wilcoxon test and -2log (LR) for the categorical variables of TB patients under DOTS in six randomly selected AA health centers

Test of no difference in survival over Strata							
Covariates	DF	Log-Rank		Wilcoxon		-2Log(LR)	
		Chi-Square	Pr> Chi-Square	Chi-Square	Pr> Chi-Square	Chi-Square	Pr> Chi-Square
Gender	1	0.0020	0.9641	0.0167	0.8973	0.0011	0.9740
Age	2	11.0320	0.0040	10.4640	0.0053	11.5554	0.0031
Weight	1	4.4974	0.0339	5.6019	0.0179	3.9437	0.0470
Smear res.	1	1.6355	0.2009	1.7743	0.1828	1.7244	0.1891
TBC	1	9.7252	0.0018	11.3382	0.0008	8.4141	0.0037
TTB	2	1.8631	0.3939	2.0941	0.3510	1.9208	0.3827
HIV	1	27.6614	<.0001	28.9458	<.0001	25.6999	<.0001

Based on the above table, we find that log-rank test and Wilcoxon test are not significant in survival experience between the various categories of gender, smear result and type of TB. But, they are significant in survival experience of the patients in different categories of age, body weight at initiation of treatment, TBC and HIV status (at α= 5%).

4.4 Results of the Cox proportional hazards model

In order to explore the relationship between the survival experience of a patient and covariate variables, we applied a Cox proportional hazard modelling approach for the analysis of survival data. By this approach it can be explored how the survival experience of a group of patients depends on the value of one or more covariate variables, whose values have been recorded for each patient at the time origin. In this study, our aim was to determine which of covariate variable had an impact on the survival time of the TB patients. The Cox PH model could be employed for estimating the regression

coefficients, making interpretation based on the hazard function and conducting statistical tests. Checking the adequacy of model and its development precede interpretation of results obtained from the fitted model.

We begin with a multivariable model that contains all variables which were significant in the univariate Cox proportional hazard model at the 20-25 percent level.

Table 4.4: Results of the Univariable Proportional Hazards Cox regression model of TB patients DOTS in six randomly selected AA health centers.

Analysis of Maximum Likelihood Estimates								
covariates	DF	Parameter Estimate	Standard Error	-2LL	Lik. ratio p-value	Score p-value	Wald p-value	Hazard Ratio
gender	1	-0.00878	0.19526	1382.970	0.9641	0.9641	0.9641	0.991
age	1	0.01626	0.00608	1376.244	0.0095	0.0072	0.0074	1.016
weight	1	-0.00859	0.00772	1381.761	0.2711	0.2665	0.2659	0.991
Sresult	1	-0.31097	0.24441	1381.251	0.1896	0.2012	0.2033	0.733
TBC	1	-0.70235	0.22993	1374.782	0.0042	0.0018	0.0023	0.495
TTB cat II	1	0.26331	0.26657	1381.039	0.3805	0.3944	0.3233	1.301
TTB cat III	1	0.36374	0.26882				0.1760	1.439
HIV	1	0.98580	0.19524	1358.373	<.0001	<.0001	<.0001	2.680
The value of -2LL for the null model is 1382.972								

Table 4.4, exhibits the summary of seven covariate variables in the univariate analysis. As result, the most appropriate subset of these predictors to be included in the multivariable model will be selected based on their contribution to the maximized log partial likelihood of the model (*-2LL*). The highest reduction in -2*LL* ($\hat{\beta}$) is observed for HIV status that reduced the value for the null or empty model with no variables in it, from 1382.972 to 1358.373. This difference is 24.599 which is statistically significant (p-value <0.0001) when compared with percentage points of the chi-square distribution on 1 degree of freedom and suggests that an improvement over the null model would be achieved by including HIV status. The reduction in -2*LL* ($\hat{\beta}$) on adding TBC to the null model is 8.19, which is significant. The next reduction in -2*LL* ($\hat{\beta}$) on adding age of patients to the null model is 6.728, which also significant.

By using the Wald chi -square test, the variable age, smear result, TB category and HIV status are significant at the 25 percent level and therefore they are candidates for inclusion in multivariable model. Age, TB Category and HIV status have relatively strong associations to the death of TB patients. Omitting the predictors or covariates gender, initial weight and type of TB from the model does not bring significant changes in the value of $-2LL\ (\hat{\beta})$. Therefore, these predictors become the first to be removed from the multivariable models.

The next step is to fit the multivariable Cox proportional model that contains age, smear result, TBC and HIV status. The predictor smear result (sresult) is clearly not significant at 5% level of significance and we will drop it from the model. So, at this stage we have a multivariable model which includes the three main effect covariates age, TB category (TBC) and HIV status (HIV). These covariates are significant at 5% level of significance [see appendix A Tables: 1-3].

The other important step is considering variables that are non-significant at univariable analysis for possibility of confounders. This can proceed by adding (one at a time) to the model containing the three significant variables at 5% level of significance. The effect of adding each of the three variables gender, weight and types of TB (TTB) in the model is shown in Table 4.5. In particular, when gender is added, the increase in $-2LL\ (\hat{\beta})$ is 0.228, when initial weight is added the increase is 2.142 and when TTB is added it is 0.736. Each of these variables is not significant and because they do not improve the $-2LL\ (\hat{\beta})$. Hence, these variables are not significant in the univariate analysis and in the multivariable analysis as well as are not confounders of the main factors in the preliminary model [see appendix B Tables: 1-3].

Table 4.5: Result of Partial likelihood ratio test for those variables not significant in the univariable analysis fitted by included in the model containing those variable significant in multivariable analysis one at a time.

variables	$-2LL\ (\hat{\beta})$	$-2LL\ (\hat{\beta})$ difference	DF	Pr>chiSq
age+TBC+HIV	1349.058			
age+TBC+HIV+gender	1348.830	0.228	1	0.633

age+TBC+HIV+weight	1346.916	2.142	1	0.143
age+TBC+HIV+TTB	1348.302	0.736	2	0.692

The final step in the model building process is the consideration of interaction terms. This step begins with the creation of a list of possible interactions formed from the main effect as given in Table 4.6. Thus, to see if the interaction effect can increase or decrease the survival time of TB patients. The Wald test was used to assess the significance of reasonable and possible interactions. The null hypothesis to be tested is that the model with only main effect fits the model equally well as the model having the main effects and their interactions as predictors. The decision for rejection of the null hypothesis is $-2\ L\ L_2$- ($-2\ L\ L_1$)> χ^2 (α=0.05) =3.84, otherwise the null hypothesis was not be rejected. Table 4.6 shows that only the interaction between HIV status and age of patient (HIV, age) was significant. And this is an indication that the interaction of HIV and the age of the patient affects the survival time of the patient [see Appendix C Tables: 1-6].

Table 4.6: Result of Partial likelihood ratio test for the contribution of the interaction effect.

Model Fit Statistics				
Interactions	$-2\ L\ L_2$ With main effects only	$-2\ L\ L_1$ With main and interaction.effects	$-2\ L\ L_2$- ($-2\ L\ L_1$)	Sig.
Age, TBC	1368.802	1368.189	0.613	Don't reject
Age, HIV	1353.644	1345.312	8.332	Reject
TBC, HIV	1353.368	1350.085	3.283	Don't reject

The results above ensure that the preliminary model of the study will contain three main effects and one interaction effect. Now, all covariates are significant at 5% level of significance (Table 4.7). The next step is to check the linearity of continuous covariates in the preliminary multivariable model.

Table 4.7: Results of the multivariable proportional hazards Cox regression model containing the main and interaction effect.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	42.6252	4	<.0001
Score	45.1009	4	<.0001
Wald	40.3081	4	<.0001

Analysis of Maximum Likelihood Estimates								
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% HR Conf. Limit	
age	1	0.02745	0.00760	13.0603	0.0003	1.028	1.013	1.043
TBC	1	-0.54298	0.23231	5.4629	0.0194	0.581	0.369	0.916
HIV	1	2.29131	0.50215	20.8209	<.0001	9.888	3.695	26.457
agHIV	1	-0.04028	0.01354	8.8468	0.0029	0.961	0.935	0.986
The value of -2LL for the model is 1340.347								

4.5 Checking for linearity of continuous covariates in the model

The next step is to examine the scale of continuous covariates in the preliminary main effects model by using graphs. The plots of the martingale residuals are useful in assessing the functional form of a covariate to be entered into a Cox model. Sometimes the covariate may need to be transformed so that the transformed variable will satisfy the assumptions of the linearity. To find the appropriate functional form of a variable, we fit a Cox model excluding the variable and then plot a lowess smooth of the martingale residuals against some transformation of the variable in question. If the transformation is appropriate, then the lowess smooth should be approximately linear. Thus, the resulting plot, for the only continuous variable in the model age of patient and age by HIV interaction, together with a LOESS smoothed curve superimposed to ease interpretation is presented in Figure 4.3 below. It can be seen that the plots of martingale residuals are random showing no systematic patterns or trends, and the LOESS smoothed curve appears to lie about a horizontal line through zero, supporting the inclusion of the

untransformed covariates age and age and HIV interaction have a linear relationship with the survival time in Cox model.

a) b)

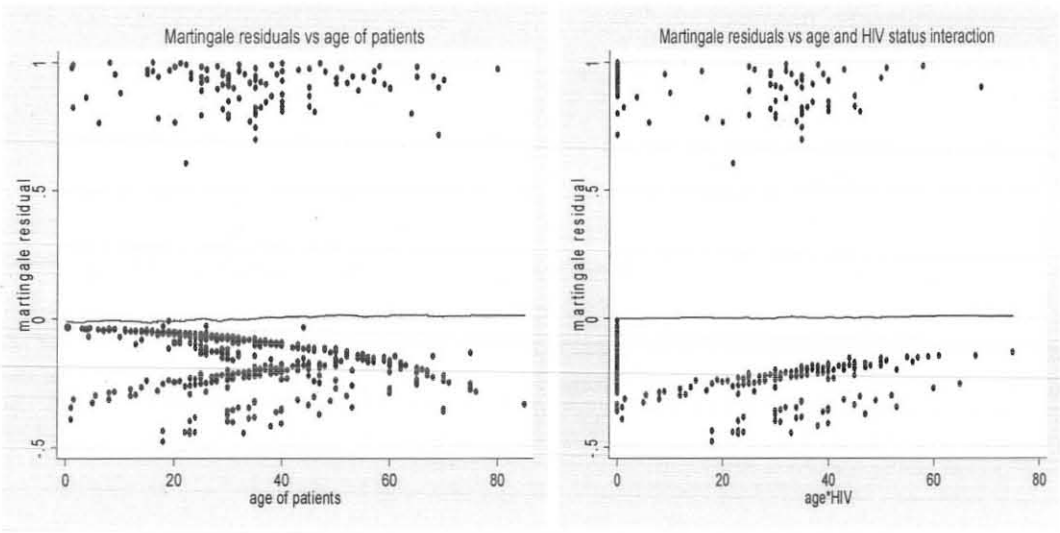


Figure 4.3: Plots of Martingale residuals computed for continuous covariates of TB patients under DOTS in AA health centers a) age and b) age and HIV status interaction

After a preliminary model has been fitted to an obsereved set of survival data, our next step would be to asses the adequacy of the fitted model. The model will not be identified as the final model until its fit and adherence to model assumptions.

4.6. Diagnosis of the model

4.6.1 Assessment of the proportional hazards assumption

The proportionality hazard assumption is vital to interpretation and use of a fitted proportional hazard model. To assess the proportional hazards assumption we examine the extent to which the estimated hazard curves for each level of strata of a covariate are equidistant over time. There are a number of different tests for assessing the PH assumption. In this study, we used residuals defined by Schoenfeld, called the Schoenfeld residuals. For each predictor in the model, Schoenfeld residuals are defined for every subject who has an event. A plot of the scaled Schoenfeld residuals and a loess smoother

as a function of time has used to test proportionality of hazards. In a 'well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0. In order to test the assumption, the extended Cox model is employed and graphical display was used to substantiate the same. For that reason, all interactions of covariates with the logarithm of survival times are modeled together with the main effects; and Wald statistic was used to test the significance of the interaction terms at 5% level of significance.

The results of tests of all the time-dependent variables in Table 4.8 were not significant either individually or collectively, so we do not have enough evidence to the reject proportionality assumption of all covariates at 5% level of significance. The that assumption of proportional hazards was met by all the covariates and the test proportionality (Wald Chi-Square = 2.1579, d.f=4, p-value= 0.7067).

Table 4.8: Result of test of proportionality assumption containing the variables in Table 4.7 and their interaction with log time.

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.01919	0.02212	0.7526	0.3857	1.019
TBC	1	0.20274	0.95433	0.0451	0.8318	1.225
HIV	1	1.09680	1.13372	0.9359	0.3333	2.995
agHIV	1	-0.02914	0.01722	2.8638	0.0906	0.971
age*log(time)	1	0.00223	0.00561	0.6904	0.6904	1.002
TBC*log(time)	1	-0.19525	0.23829	0.6714	0.4126	0.823
HIV*log(time)	1	0.31882	0.27050	1.3891	0.2385	1.375
agHIV*log(time)	1	-0.0001785	0.0001725	1.0712	0.3007	1.000

Linear Hypotheses Testing Results			
Label	Wald Chi-Square	DF	Pr > ChiSq
test proportionality	2.1579	4	0.7067

Now, we see that a plot of the scaled Schoenfeld residuals and a loess smoother as a function of time used to test proportionality of hazards. The plots of the scaled

Schoenfeld residuals and the lowness smooth are shown in Figure 4.4 (a-d) support the assumption of proportional hazards for each of the four covariates. That is, each subplot in the figure is random, smooth and approximates a horizontal through zero or slope approximately equal to zero. This indicates that none of the four covariates had interaction with log of time also the plots support the proportional hazards assumption.

a) b)

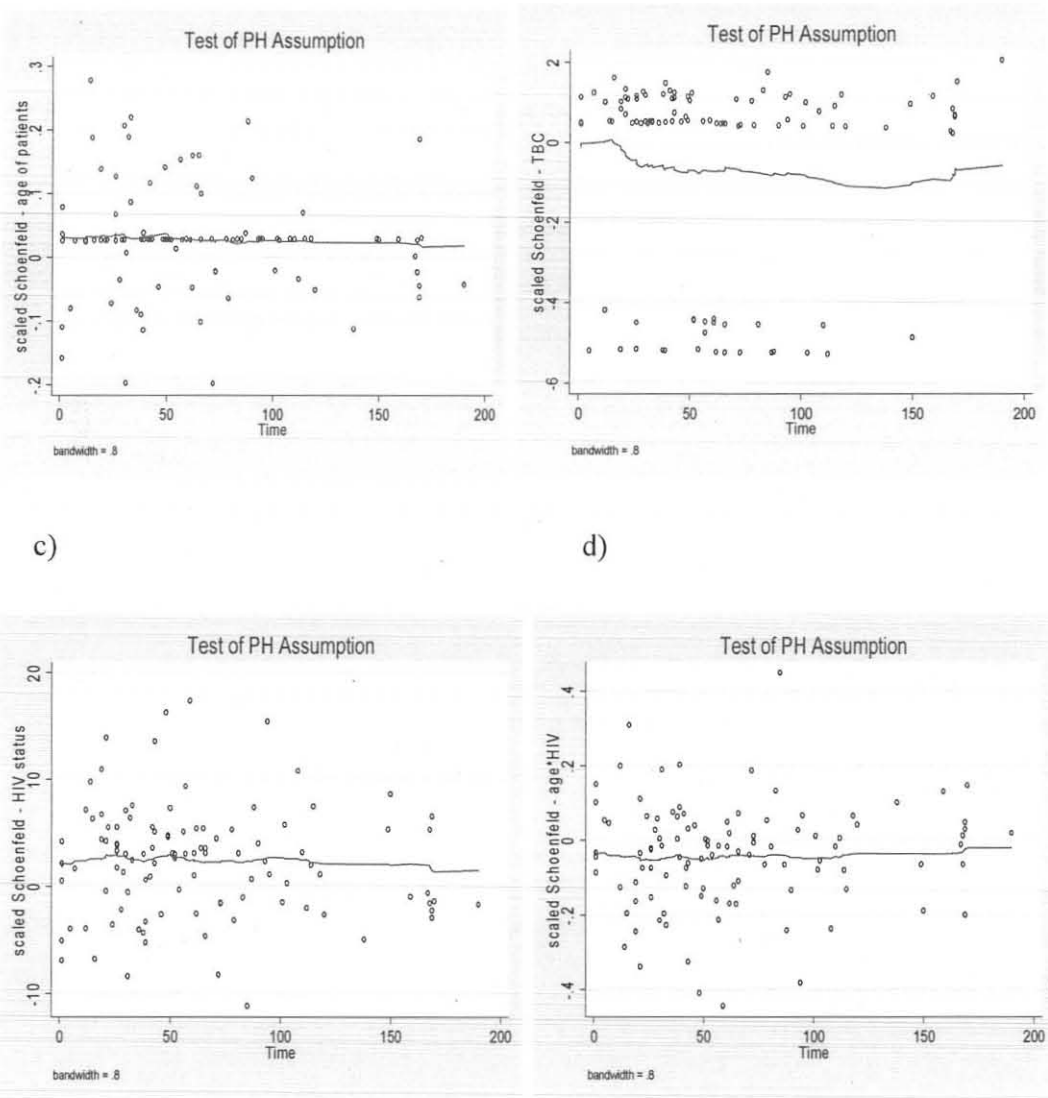


Figure 4.4: Graphs of the scaled Schoenfeld residuals and their LOESS smooth curves for the covariates a) age of patients, b) TB patient category, c) HIV status and d) age and HIV status interaction.

4.6.2 Identification of influential and poorly fit data

The graphs of the score residuals for the covariates age and the age and HIV interaction obtained from fitted model in Table 4.7 are shown in Figure 4.5. These two terms were chosen because they are the continuous variables in the fitted model. Hence, continuous covariates are most amenable to having their score residuals examined graphically. The graph for the dichotomous variables are less interesting in that all of the values fall on two vertical bands at zero and one, the two covariate values.

The score residuals for age in Figure 4.5 (a) display fan shape with smallest distance near the mean age of 35 and increasing in absolute value for ages increasingly order or younger than 35. The purpose of the plot is to see whether there are subjects whose ages yield unexpectedly large values. This would be seen in the graph as a point lying well away from the others in the plots. In the figure there is one point in the top right at age 80 that falls a bit away from the rest of the points. However, the distance between this point and the others is not striking. The oldest subject, age 85, has score residuals that are well within range of values. Thus, we conclude that there are no high leverage values for age of TB patients.

The score residuals for age × HIV interaction are ploted in Figure 4.5 (b). The plots has one point in the top left corner that falls a bit away from the rest of the others points. But, the distance between these point and the others is not striking. The two subjets, one in the top right and one in the bottom right, have score residuals that are well within range of values. Thus, we conclude that there are no high leverage values for age by HIV interaction. In general, the plots in Figure 4.5 have shown that, there are no strikingly large score residuals.

a) b)

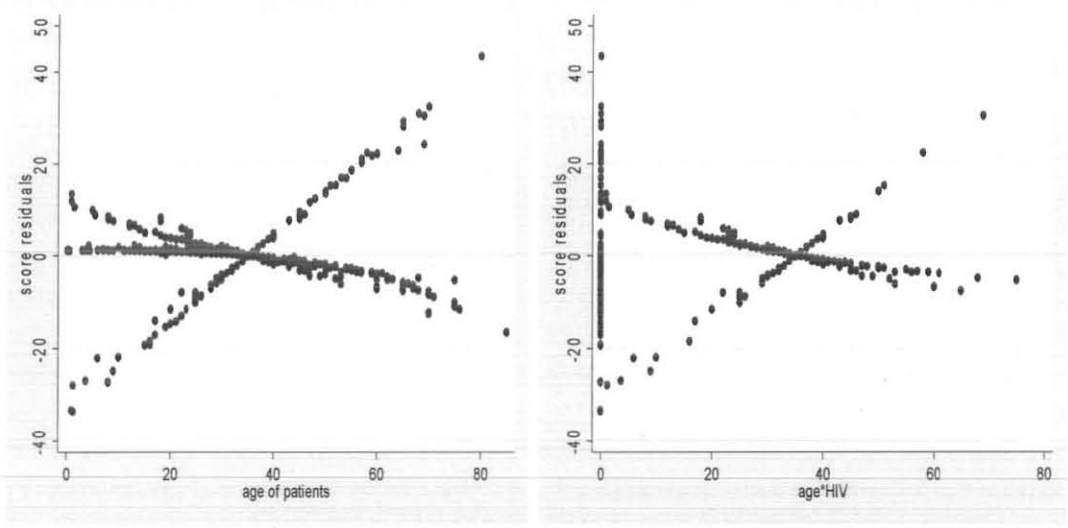


Figure 4.5 graphs of the score residuals computed from the model in Table 4.7 for (a) age of TB patients and (b) age by HIV status interaction.

The first six largest changes in parameter estimates are shown in Table 4.9. To begin with the largest difference for covariate age is observed for patient numbered 730. The result exhibits that the change in the parameter estimate if the data for this patient is discarded is 0.0023234. The standard error of the parameter estimate for age in the full data set is 0.0076. That is the percentage change in parameter estimate if the observation is removed is about 30.57% of the standard error i.e., less than one standard error. Thus, removing this observation would not bring a significant change on age of patients. The largest difference for covariates TBC, HIV and age by HIV interaction are observed for patients numbered 413, 133 and 548 with percentage change in parameter estimate if the observation were removed are 18.94%, 23.61% and 31.07% of the standard error respectively. These were less than one standard error and deleting the observations of the highest difference in parameter estimates and hence had no significant impact on the parameters of the covariates and on the fit of the model. Therefore, it can be concluded that there was no aberrant observation in the data set that illegitimately inflated the estimates of the parameters of the covariates in the final model.

Table 4.9: The six highest differences in the parameter estimates of the variables included in the model in Table 4.7 when the data value for each patient is in turn deleted from the model, approximate delta-betas for age($\widehat{\beta}_1$), TBC($\widehat{\beta}_2$), HIV($\widehat{\beta}_3$) and age by HIV($\widehat{\beta}_4$).

Deleted Obs. (i)	$\Delta_i\hat{\beta}_1$	%change parameter of the std. error	Deleted Obs. (i)	$\Delta_i\hat{\beta}_2$	%change parameter of the std. error
730	0.0023234	30.57	413	-0.0439987	18.94
162	-0.0021205	27.90	585	-0.0428435	18.44
262	-0.0020957	27.58	343	-0.0427599	18.41
421	-0.0017624	23.20	368	-0.0423473	18.23
542	0.0017002	22.37	544	-0.04214	18.14
580	0.001611	21.20	447	-0.041601	17.91
Sd.error($\hat{\beta}_1$) of full model= 0.00760			Sd.error($\hat{\beta}_2$) of full model= 0.23231		

Deleted Obs. (i)	$\Delta_i\hat{\beta}_3$	%change parameter of the std. error	Deleted Obs. (i)	$\Delta_i\hat{\beta}_4$	%change parameter of the std. error
133	0.1185484	23.61	548	0.0042068	31.07
548	-0.1155908	23.02	63	0.0032745	24.18
204	0.1146464	22.83	133	-0.0031662	23.38
244	-0.1056003	21.03	204	-0.0030298	22.38
162	-0.1008921	20.10	244	-0.0027001	19.94
262	-0.098443	19.60	282	-0.0024873	18.37
Sd.error($\hat{\beta}_3$) of full model= 0.50215			Sd.error($\hat{\beta}_4$) of full model= 0.01354		

4.6.3 Checking for overall goodness of fit

The last step in the process of assessment of model adequacy is an overall goodness fit of the model. We can evaluate the fit of the model by using the Cox-Snell residuals. The Cox-Snell residuals are most useful for examining overall fit of a model. If the model fits the data well then the true cumulative hazard function conditional on the covariate vector has an exponential distribution with a hazard rate of one.

The cumulative hazard plot of the Cox-Snell residuals is shown in Figure 4.6. We see that the hazard function is reasonably straight line that has a unit slope and zero intercept.

It is approximates the 45 degree line very closely except for very large values of time. It is very common for models with censored data to have some wiggling at large values of time and it is not something which should cause much concern. Overall we would conclude that the final model fits the data very well. Therefore, the model with estimates as given in Table 4.7 is the final model.

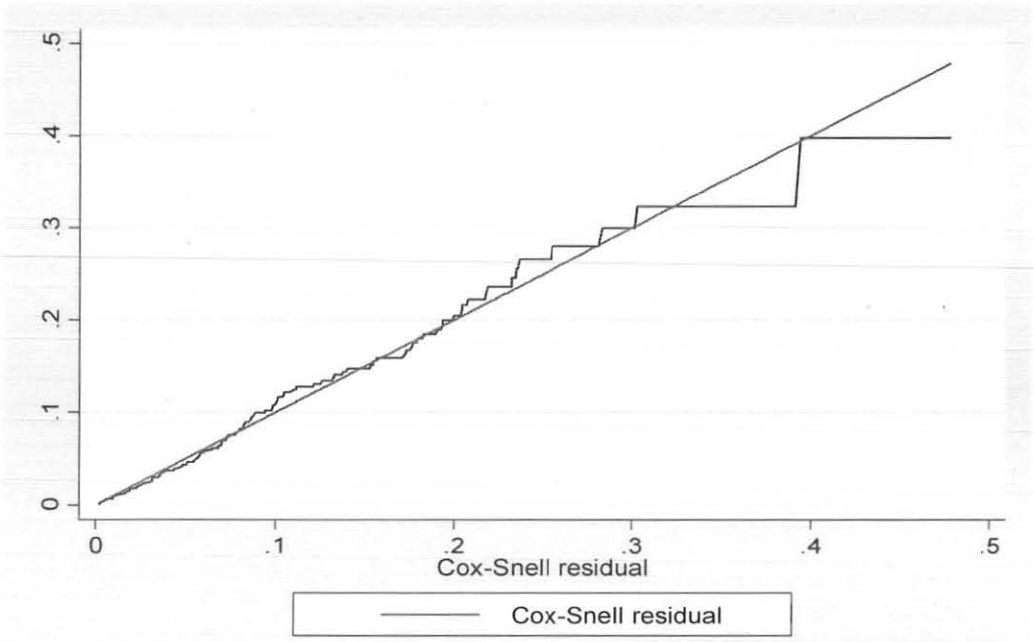


Figure 4.6: Cumulative hazard plot of the Cox-Snell residuals of the proportional hazards Cox regression model in table 4.7. The 45 degree-straight line through the origin is drawn for reference.

4.7 Interpretation and discussion of the results

This section presents an interpretation and discussion of the results. The discussion will focus on the research objectives of the study and compares the research findings with existing literature.

4.7.1 Interpretation of the results

The final model fit to the TB data, shown in Table 4.7, found three main effects and one interaction effect that are statistically significantly associated with the hazard of TB

patients. The model contains two dichotomous covariates (TB patient category (TBC) and HIV), one linear continuous covariate (age) and one interaction of a continuous and a dichotomous covariate (age and HIV).

The result of the fitted final model in Table 4.7, are interpreted in terms of hazard ratios (HR), sometimes called risk ratios. The coefficient of the categorical covariates is interpreted as the logarithm of the ratio of the hazard of death to the baseline (reference group) hazard. That is, they are interpreted by comparing the reference group with others. Similarly, the coefficient for a continuous explanatory variable indicates the estimated change in the logarithm of the hazard ratio for a unit increase in the value of the respective covariate when the remaining covariates in the model are controlled.

Only the covariate TB patient category (TBC) has hazard ratio that is estimated by exponentiating its estimated coefficient. This is because the covariates age and HIV are involved in interaction. The estimated coefficient for a new case TB was -0.54298, which decreases the hazards of experiencing death by a factor $exp(-0.54289) = 0.581$ (95% CI: 0.369-0.916) that is, the patients with a new case TB have about 41.9% lower mortality rate than patients with non-new case TB. The 95 percent confidence interval suggested that the rate could be as much as 63.1 percent lower to only 8.4 percent lower. The estimated hazard ratio pointed to a significant benefit for the new case of the two TB patient categories, controlling for all other model covariates.

Age and HIV status are present in the model, with both main effects and their interaction. Since HIV status is at two levels and is fixed by design of the study, we present hazard ratios for age at each HIV status rather than for HIV status at each age. The first step is to write down the equation for the hazard ratio as a function of the variables of interest, holding TBC fixed in the model.

$$Log\ \hat{H}(age, HIV, Z) = \hat{\beta}_1 age + \hat{\beta}_2 HIV + \hat{\beta}_3 age \times HIV + \hat{\boldsymbol{\beta}}' z \qquad (1)$$

The second step is to write down the expression for the difference of interest, in this case an increase say 15 years of age, holding HIV fixed:

$$Log\ \widehat{H}(age+15,\ HIV,\ Z) - Log\ \widehat{H}(age,\ HIV,\ Z) = \{\hat{\beta}_1(age+15) + \hat{\beta}_2 HIV + \hat{\beta}_3(age+15) \times HIV + \widehat{\boldsymbol{\beta}}'\boldsymbol{Z}\} - \{\hat{\beta}_1 age + \hat{\beta}_2 HIV + \hat{\beta}_3(age) \times HIV + \widehat{\boldsymbol{\beta}}'\boldsymbol{Z}\}$$

$$= 15\hat{\beta}_1 + 15\hat{\beta}_3 \times HIV \qquad (2)$$

The last step is to estimate hazard ratio for an increase of 15 years of age at HIV=0 and HIV=1, by using the estimated coefficient of age from Table 4.7, $\hat{\beta}_1$= 0.02745, and the estimated coefficient of the interaction of age and HIV $\hat{\beta}_3$= -0.04028.

$$\widehat{HR}(\text{age+15, age, HIV=0}) = exp\ (15 \times 0.02745) = 1.51$$

$$\text{and } \widehat{HR}(\text{age+15, age, HIV=1}) = exp(15 \times 0.0275 + 15 \times (-0.04028)) = 0.82$$

The endpoints of the 100(1-α) percent confidence interval estimator of the hazard ratio are computed by exponentiating the endpoints of the confidence interval of the estimator of (2), which are

$$15\hat{\beta}_1 + 15\hat{\beta}_3 \times HIV \pm Z_{1-\alpha/2}\, \widehat{SE}(15\hat{\beta}_1 + 15\hat{\beta}_3 \times HIV) \qquad (3)$$

where

$$\widehat{SE}(15\hat{\beta}_1 + 15\hat{\beta}_3 \times HIV) = \left[15^2 \times \widehat{var}(\hat{\beta}_1) + 15^2 HIV^2 \times \widehat{var}(\hat{\beta}_3) + 2 \times 15^2 \times HIV \times \widehat{cov}(\hat{\beta}_1, \hat{\beta}_3)\right]^{1/2} \qquad (4),$$

and since HIV is coded zero and one, HIV^2= HIV. Again the values of the variance and covariance needed to compute the standard error are available from software packages. These are, $\widehat{var}(\hat{\beta}_1)$= 0.0000576901, $\widehat{var}(\hat{\beta}_3)$= 0.0001833672 and $\widehat{cov}(\hat{\beta}_1, \hat{\beta}_3)$= 0.0022105405.

Using these values, the 95 percent confidence interval at HIV=1(positive) is (-2.200, 1.815).

This means that being older by 15 years at HIV negative increased the rate of death by about 51 percent and HIV positive reduces the rate of death by about 18 percent.

4.7.2 Discussion of the results

The study by Oursler *et al* (2002) showed that the median time to death among TB patient was 39 days and follow-up for survivors was 202 days. Lo *et al* (2011) showed

that, 50% of deaths occurred within 2 months. In this study we found that the median follow-up time is 168 days for patients that are censored (range from, 15 to 284 days), 25% of the patients had 176 days of follow-up (upper quartile). The median time to death among those patients who died was 52 days (ranges from, 1 to 190 days).

The survival curves of TB patient among age groups are significantly different. There was no difference in the survival curves of male and female patients and according to type of disease (Pardeshi, 2009). A study by Getahun *et al* (2011) conducted in Addis Ababa showed that the survival status was significantly different between Patient age, weight at initiation of anti-TB treatment, patient category, year of enrollment and treatment center. This study we found a similar result, no differences among survivor curves of gender and type of disease. However, there was a difference among survivor curves of age category, body weight at initiation of treatment, TB patient category and HIV status for TB patients.

Age has been identified as an important risk factor for death in tuberculosis patients. Different studies showed that, age was a factor that is affecting the survival of TB patients. According to Horne *et al.* (2010) in Washington State, mortality was independently associated with increasing age. A study in Maryland, community-based cohort of patients with drug-susceptible pulmonary TB, age was strongly associated with the risk of death (Oursler *et al.*, 2002). A study in Brazil showed that age was statistically significant in the multivariable Cox regression model (Albuquerque *et al.*, 2009 and Domingos *et al.*, 2008). Another study also showed that age has been identified as an important risk factor for death in tuberculosis patients (Vasankari *et al.*, 2007, Pardeshi, 2009, Lo *et al.*, 2011, Harries *et al.*, 2001, Getahun *et al.*, 2011 and Muñoz-Sellart *et al.*, 2010). Our study also found that age was statistically a significant risk factor for death.

HIV co-infection was statistically associated with an increased risk of death in TB patients during treatment (Horne *et al.*, 2010, Oursler *et al.*, 2002, Albuquerque *et al.*, 2009, Domingos *et al.*, 2008 and Borgdorff *et al.* 1998). We have also found a similar result, i.e. HIV co-infection was a statistically significant risk factor for death in TB patient.

Similar to our finding, other studies have also shown that TB patient category (TBC) was statistically associated with death of patients with tuberculosis (Albuquerque *et al.*, 2009, Domingos *et al.*, 2008, Mathew *et al.*, 2006 and Vasantha *et al.*, 2008).

The present study identified that body weight at initiation of treatment was not the risk factor for death in tuberculosis patients in multivariable Cox proportional regression model. However, studies by (Albuquerque *et al.*, 2009, Vasantha *et al.*, 2008 and Getahun *et al.*, 2011) reported that body weight at initiation of treatment was a risk factors associated with death during anti-tuberculosis treatment period.

In this study, the covariates gender and type of TB (TTB) were not found to be factors that affect the survival of patients with TB. But, different studies reported that male were at higher risk factor for death in TB patients (Horne *et al.*, 2010, Vasankari *et al.*, 2007, Borgdorff *et al.* 1998, Low *et al.*, 2009 and Demissie and Kebede, 1994) and types of tuberculosis (positive pulmonary, negative pulmonary and extra pulmonary tuberculosis) were identified the significant factors for mortality of tuberculosis patients (Domingos *et al.*, 2008, Borgdorff *et al.* 1998 and Muñoz-Sellart *et al.*, 2010).

CHAPTER FIVE

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This study is an attempt to identify the factors that affect the survival of the patients with tuberculosis. Based on our study findings we make the following conclusions.

- ✓ The study has shown the median follow-up time is 168 days for patients that are censored (range from, 15 to 284 days), 25% of the patients had 176 days of follow-up (upper quartile). The median time to death among dead patients was 52 days (ranges from, 1 to 190 days).The majority of the death occurs during the of anti-TB treatment initiation.

- ✓ From the Kaplan-Meier survival estimates there was a significant difference in survival by the age, body weight at initiation of treatment, TB patient category, and HIV status. However, there were no differences among survival curves of gender, smear result and type of TB. The log-rank test also showed that, there was no significant difference in survival experience between the various categories of gender, smear result and type of TB. However, the test showed that the survival experience of patients in different categories of age, initial weight, TBC and HIV status differ significantly.

- ✓ The multivariable Cox proportional hazards regression results analysis indicated that the three covariates age, TB patient category (TBC), HIV and the interaction age by HIV were significantly associated with death among TB patients. Gender, body weight at initiation of treatment, smear result and type of TB were not significantly associated with factors that are affecting the survival of patients with tuberculosis.

5.2 Recommendations

We recommend the following:

- ✓ Strengthen to follow-up of patients with TB treatment from the day of anti-TB treatment initiation to completion days.

- ✓ Paying close attention to individuals with disease especially HIV co-infected and non-new TB case can prevent and decrease the death rate among TB patients.

- ✓ All concerned bodies should develop recording systems for completeness of the registry about variables like alcohol drinking, drug use, malnutrition, unemployment, education level and multi-drug resistance for TB patients. Hence, these variables are important for identifying the factors that affect the survival/death in TB patients and conducting further studies.

REFERENCES

1. Albuquerque, M.F.P.M., Batista, J.A.L., Ximenes, R.A.A., Carvalho, M.S., Diniz, G.T.N. and Rodrigues, L.C. (2009). Risk factors associated with death in patients who initiate treatment for tuberculosis after two different follow-up periods. *Rev Bras Epidemiol,* 12(**4**), 513-22.
2. Arjas, E. (1988). A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model. *Journal of the American Statistical Association*, **83**, 204—212.
3. Bates, JH. (1980). Transmission and pathogenesis of tuberculosis. *Clin Chest Med*, 1(**2**), 167-174.
4. Borgdorff, M.W., Veen, J., Kalisvaart, N.A. and Nagelkerke, N. (1998). Mortality among tuberculosis patients in the Netherlands in the period 1993-1995. *Eur Respir J*, **11**, 816-820.
5. Centers for Disease Control and Prevention (CDC). (2011). Treatment for Tuberculosis disease: National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Division of Tuberculosis Elimination. http://www.cdc.gov/tb/topic/treatment/tbdisease.htm. Accessed Dec. 2012.
6. Collett, D. (2003). *Modelling survival data in medical research*, Second Edition, Chapman and Hall/ CRC, London.
7. Cox, D.R. (1972). "Regression models and life Tables (with Discussion)." *Journal of the Royal statistical society*, 34: 187-220.
8. Demissie, M.and Kebede, D. (1994). Defaulting from tuberculosis treatment at the Addis Abeba Tuberculosis Centre and factors associated with it. *Ethiop Med J*, 32(**2**), 97-106.
9. Domingos, M.P., Caiaffa, W.T. and Colosim, E.A. (2008). Mortality, TB/HIV co-infection, and treatment dropout: predictors of tuberculosis prognosis in Recife, Pernambuco State, Brazil. *Cad. Saúde Pública, Rio de Janeiro*, 24(**4**), 887-896
10. Faustini, A., Hall, AJ. And Perucci, CA. (2005). Tuberculosis treatment outcomes in Europe: a systematic review. *Eur Respir J.*, **26**, 503-510.

11. Federal Ministry of Health (2008). Tuberculosis, Leprosy and TB/HIV Prevention and Control Programme Manual, Addis Ababa, Ethiopia.
12. Federal Ministry of Health (2010). Health Sector Development Programme 5, Addis Ababa, Ethiopia.
13. Getahun, B., Ameni, G., Biadgilign, S. and Medhin, G. (2011). Mortality and associated risk factors in a cohort of tuberculosis patients treated under DOTS programme in Addis Ababa, Ethiopia. *BMC Infectious*, 11, 127.
14. Gore, R.S. (2005). Kentucky Epidemiologic Notes and Reports: TB Control Program Condensed from CDC National Center for HIV, STD and TB Prevention, Division of TB Elimination. Presented on World Tuberculosis Day—March 24, Kentucky, Vol. **40** No.2.
15. Harrel, F. and Lee, K. (1986). *Proceedings of the Eleventh Annual SAS User's Group International*. 823–828.
16. Harries, AD., Hargreaves, NJ., Gausi, F., Kwanjana, JH. And Salaniponi, FML. (2001). High early mortality in tuberculosis patients in Malawi. *Int J Tuberc Lung Disease*, **5**, 1000-1005.
17. Horne, D.J., Hubbard, R., Narita, M., Exarchos, A., Park, D.R. and Goss, CH.H. (2010). Factors associated with mortality in patients with tuberculosis. *BMC Infectious Diseases* 10,258.
18. Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis Regression Modeling of Time to Event Data*. John Wiley and Sons, Inc., New York.
19. Kalbfleisch, J.D. and Prentice R.L. (2002). *The Statistical Analysis of Failure Time Data*, Second Edition, John Wiley and Sons, Inc., New Jersey.
20. Kaplan, E.L., Meier, P. (1958). "Nonparametric estimation from Incomplete Data." *Journal of American Statistical Association*, 53(**282**), 457-481.
21. Klein, J.P. and Moeschberger, M.L. (2003). *SurvivalAnalysis techniques for censured and truncated data,* Second Edition. Springer-Verlag, New York.
22. Kleinbaum, D.G. and Klein, W. (2005). *Survival Analysis a self learning text*, Second Edition. Springer Science+Business Media, Inc., New York.

23. Kolappan, C., Subramani, R., Karunakaran, K. and Narayanana, PR. (2006). Mortality of tuberculosis patients in Chennai,India. *Bull World Health Organ*, **84**, 555-560.

24. Kumar, V., Abbas, A.K., Fausto, N. and Mitchell, R.N. (2007). Robbins Basic Pathology, Eight Editions. *Saunders Elsevier*, 516-522 ISBN 978-1-4160-2973-1.

25. Lo, H.Y., Suo, J. Chang, HJ., Yang, SL. and Chou, P. (2011). Risk Factors Associated With Death in a 12-Month Cohort Analysis of Tuberculosis Patients: 12-Month Follow-up After Registration. *Asia Pac J Public Health* published

26. Low, S., Ang, LW., Cutter, J., James, L., Chee, CB. and Wang, YT., et al. (2009). Mortality among tuberculosis patients on treatment in Singapore. *Int J Tuberc Lung Dis*, **13**, 328-34.

27. Manissero, D., Hollo, V., Huitric, E., Kdmn, C. and Amato-Gauci, A. (2010). Analysis of tuberculosis treatment outcomes in the European Union and European Economic Area: efforts needed towards optimal case management and control. *Euro Surveill*, 15(**11**)

28. Mathew, T.A., T. N. Ovsyanikova, T.N., Shin, S.S., Gelmanova, I., Balbuena, D.a., Atwood, S., Peremitin, G.G., A. K. Strelis, A.K. and Murray, M.B. (2006). Causes of death during tuberculosis treatment in Tomsk Oblast, Russia. *Int J Tuberc Lung Dis* 10(**8**):857–863.

29. Moosazadeh, M. and Amiresmaili, M. (2011). Risk Factors Associated with Unsuccessful Treatment outcomes among Patients with Smear Positive Ppulmonary Tuberculosis in Iran: A Logistic Regression Model, Kerman University of medical sciences, Kerman-Iran.

30. Muñoz-Sellart, M., Cuevas, L.E., Tumato, M., Merid,Y. and Yassin, M.A. (2010). Factors associated with poor tuberculosis treatment outcome in the Southern Region of Ethiopia. *Int J Tuberc Lung Dis* 14(**8**), 973–979.

31. Oursler, K.K., Moore, R.D., Bishai, W.R., Harrington, S.M., Pope, D.S. and Chaisson, R.E. (2002). Survival of patients with pulmonary tuberculosis: clinical and molecular epidemiologic factors. *Clin Infect Dis.*, **34**, 752-759.

32. Pardeshi, G. (2009). Survival Analysis and Risk Factors for Death in Tuberculosis Patients on directly observed Treatment-Short Course. *Indian J Med Sci*, Vol. **63**, No. 5.

33. Rodriguez, D. (2009). Risk Factors for Tuberculosis Infection. http://www.everdayhealth.com/tuberculosis/risk-factors.aspx.AccessedDec. 2012.

34. Schoenfeld, D. (1982). Partial residuals for the proportional hazards model. *Biometrika*, **69**, 51–55.

35. Snider, D.E. and Roper, W.L. (1992). The new tuberculosis. *N Engl J Med,* **326**, 703–5.

36. Talay, F., Kunbetli, S. and Altin, S. (2007). Factors Associated with Treatment Success for Tuberculosis Patients .Research hospital for chest disease and thoracic surgery, Stanbul, Turkey.

37. Vasankari, T., Holmström, P., Ollgren, J., Liippo, K., Kokki, M. and Ruutu, P. (2007). Risk factors for poor tuberculosis treatment outcome in Finland: a cohort study. *BMC Public Health*, **7**, 291.

38. Vasantha, M., Gopi P. G. and Subramani, R. (2008) .Survival of Tuberculosis patients treated under DOTS in rural Tuberculosis Unit (TU), South India. *Indian J Tuberc,* **55,** 64-69.

39. Vree, M., Huong, N.T., Duong, BD., Sy, DN., Van le, N., Co, NV., Cobelens, FG. and Borgdorff, MW. (2007). Mortality and failure among tuberculosis patients who did not complete treatment in Vietnam: a cohort study. *BMC Public Health*, **7,** 134.

40. WHO report (2009). Treatment of Tuberculosis: Guidelines for National Programmes. World Health Organization, Geneva, Switzerland.

41. WHO report (2010). Global Tuberculosis Control: Surveillance, Planning, Financing. World Health Organization, Geneva, Switzerland.

42. Yuan, Z., Shen, X. and Deriemer, K. (2009). Death among tuberculosis cases in Shanghai, China: who is at risk?. *BMC Infect Dis.*, **9**, 95.

APPENDIX

APPENDIX A: Results of the multivariable proportional hazards Cox regression model

Table 1: Results of the univariable proportional hazards Cox regression model of TB patients DOTS in six randomly selected AA health centres.

Testing Global Null Hypothesis: BETA=0							
Covariates	DF	Likelihood Ratio		Score		Wald	
		Chi-Square	Pr> ChiSq	Chi-Square	Pr> ChiSq	Chi-Square	Pr> ChiSq
Gender	1	0.0020	0.9641	0.0020	0.9641	0.0020	0.9641
Age	1	6.7277	0.0095	7.2104	0.0072	7.1647	0.0074
Weight	1	1.2110	0.2711	1.2346	0.2665	1.2379	0.2659
Smear result	1	1.7208	0.1896	1.6334	0.2012	1.6188	0.2033
TBC	1	8.1897	0.0042	9.7121	0.0018	9.3306	0.0023
TTB	2	1.9324	0.3805	1.8607	0.3944	1.8437	0.3978
HIV	1	24.5990	<.0001	27.6246	<.0001	25.4948	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
gender	1	-0.00878	0.19526	0.0020	0.9641	0.991
age	1	0.01626	0.00608	7.1647	0.0074	1.016
weight	1	-0.00859	0.00772	1.2379	0.2659	0.991
Smear result	1	-0.31097	0.24441	1.6188	0.2033	0.733
TBC	1	-0.70235	0.22993	9.3306	0.0023	0.495
TTB cat II	1	0.26331	0.26657	0.9757	0.3233	1.301
TTB cat III	1	0.36374	0.26882	1.8308	0.1760	1.439
HIV	1	0.98580	0.19524	25.4948	<.0001	2.680
The value of -2LL for the null model is 1382.972						

Table 2: Results of the multivariable proportional hazards Cox regression model containing the variables significant at 20 - 25% level in the univariable proportional hazards Cox regression model of TB patients under DOTS

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.5184	4	<.0001

Score	38.5787	4	<.0001
Wald	35.9138	4	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.01357	0.00661	4.2183	0.0400	1.014
Smear result	1	-0.18871	0.24703	0.5836	0.4449	0.828
TBC	1	-0.53159	0.23288	5.2105	0.0225	0.588
HIV	1	0.87529	0.19865	19.4144	<.0001	2.400

Table 3: Results of the multivariable proportional hazards Cox regression model after eliminating the variable smear result from the multivariable proportional hazards Cox regression model in Table 2.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.9139	3	<.0001
Score	37.9981	3	<.0001
Wald	35.3213	3	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.01407	0.00662	4.5230	0.0334	1.014
TBC	1	-0.52104	0.23244	5.0247	0.0250	0.594
HIV	1	0.89072	0.19770	20.2984	<.0001	2.437
The value of -2LL for the model is 1349.058						

APPENDIX B: The result of multivariable Cox hazard model those variables not significant in the univariable analysis fitted by included in the model containing those variable significant in multivariable analysis one at a time.

Table 1: Result of multivariable Cox proportional hazard model containing variables in Table 3(Appendix A) including the variable gender which is not significant in univariable analysis.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.1421	4	<.0001
Score	38.1607	4	<.0001
Wald	35.4751	4	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
gender	1	0.09417	0.19707	0.2284	0.6327	1.099
age	1	0.01443	0.00664	4.7204	0.0298	1.015
TBC	1	-0.52584	0.23262	5.1100	0.0238	0.591
HIV	1	0.89448	0.19782	20.4452	<.0001	2.446
The value of -2LL for the model is 1348.830						

Table 2: Result of multivariable Cox proportional hazard model containing variables in Table 3(Appendix A) including the variable initial weight of patients which is not significant in univariable analysis.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	36.0554	4	<.0001
Score	40.1081	4	<.0001
Wald	37.3936	4	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
weight	1	-0.01261	0.00860	2.1489	0.1427	0.987
age	1	0.01703	0.00667	6.5241	0.0106	1.017

TBC	1	-0.50480	0.23260	4.7098	0.0300	0.604
HIV	1	0.86526	0.19822	19.0546	<.0001	2.376
The value of -2LL for the model is 1346.916						

Table 3: Result of multivariable Cox proportional hazard model containing variables in Table 3(Appendix A) including the variable initial type of TB which is not significant in univariable analysis.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.6693	5	<.0001
Score	38.7888	5	<.0001
Wald	36.1260	5	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TTB catII	1	0.22970	0.26768	0.7363	0.3908	1.258
TTB catIII	1	0.14388	0.27358	0.2766	0.5990	1.155
age	1	0.01389	0.00666	4.3480	0.0371	1.014
TBC	1	-0.53427	0.23312	5.2523	0.0219	0.586
HIV	1	0.88122	0.19931	19.5490	<.0001	2.414
The value of -2LL for the model is 1348.302						

APPENDIX C: Results of the multivariable proportional hazards Cox regression model containing two covariates and their interaction.

Table 1: Results of the multivariable proportional hazards Cox regression model containing age, TB patients' category and their interaction.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.7831	3	0.0020
Score	16.0814	3	0.0011
Wald	15.4054	3	0.0015

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.00474	0.01552	0.0935	0.7598	1.005
TBC	1	-1.13072	0.62975	3.2238	0.0726	0.323
agTBC	1	0.01308	0.01692	0.5977	0.4395	1.013

Table 2: Results of the multivariable proportional hazards Cox regression model containing only age and TB patients' category.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.1695	2	0.0008
Score	15.9143	2	0.0004
Wald	15.4384	2	0.0004

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.01563	0.00620	6.3531	0.0117	1.016
TBC	1	-0.66782	0.23018	8.4177	0.0037	0.513

Table 3: Results of the multivariable proportional hazards Cox regression model containing age, HIV status and their interaction.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.6597	3	<.0001
Score	39.1291	3	<.0001
Wald	34.8034	3	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.02752	0.00751	13.4205	0.0002	1.028
HIV	1	2.29752	0.49592	21.4632	<.0001	9.949
agHIV	1	-0.03882	0.01332	8.4954	0.0036	0.962

Table 4: Results of the multivariable proportional hazards Cox regression model containing only age and HIV status.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.3278	2	<.0001
Score	32.1899	2	<.0001
Wald	29.8431	2	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.01456	0.00654	4.9660	0.0258	1.015
HIV	1	0.94509	0.19565	23.3345	<.0001	2.573

Table 5: Results of the multivariable proportional hazards Cox regression model containing TBC, HIV status and their interaction.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.8865	3	<.0001
Score	35.2507	3	<.0001
Wald	31.6797	3	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TBC	1	-0.99846	0.32124	9.6605	0.0019	0.368
HIV	1	0.28893	0.40059	0.5202	0.4708	1.335
TBHIV	1	0.83393	0.45885	3.3031	0.0691	2.302

Table 6: Results of the multivariable proportional hazards Cox regression model containing only TBC and HIV status.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.6039	2	<.0001
Score	33.9676	2	<.0001

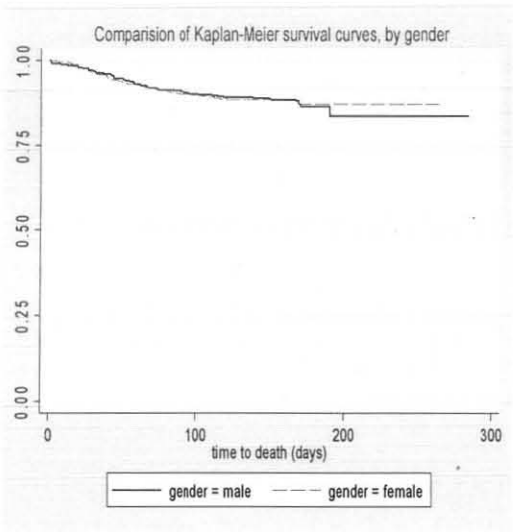
Wald	31.5157	2	<.0001

Analysis of Maximum Likelihood Estimates						
Covariates	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TBC	1	-0.54520	0.23228	5.5091	0.0189	0.580
HIV	1	0.92757	0.19738	22.0843	<.0001	2.528

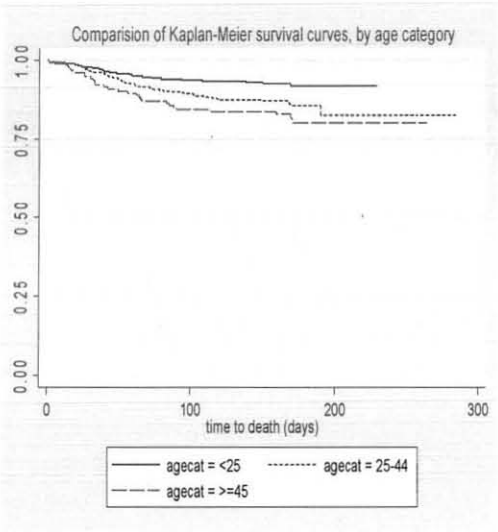
APPENDIX D: Figures

Figure (a – g): Plots of Kaplan-Meier survivor functions for different categories or groups

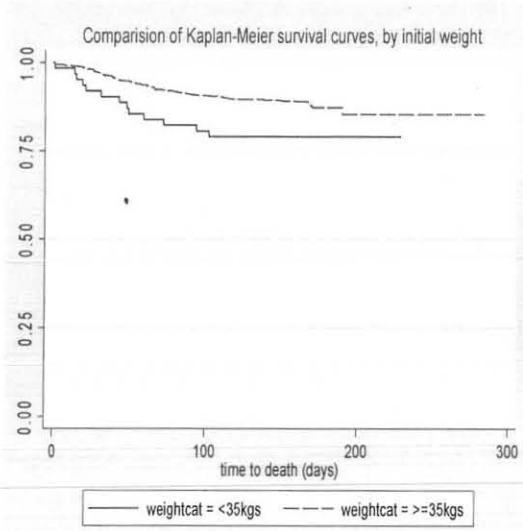
a)



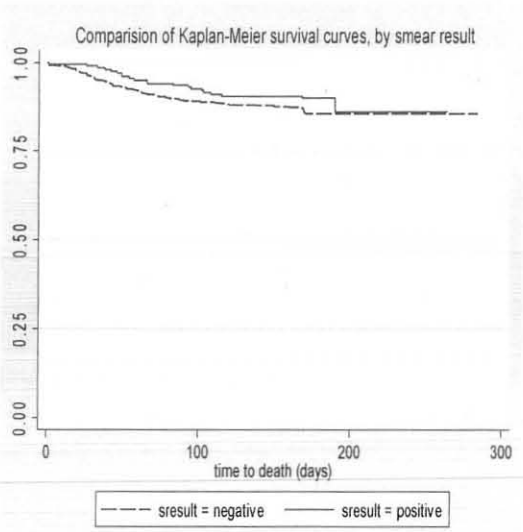
b)



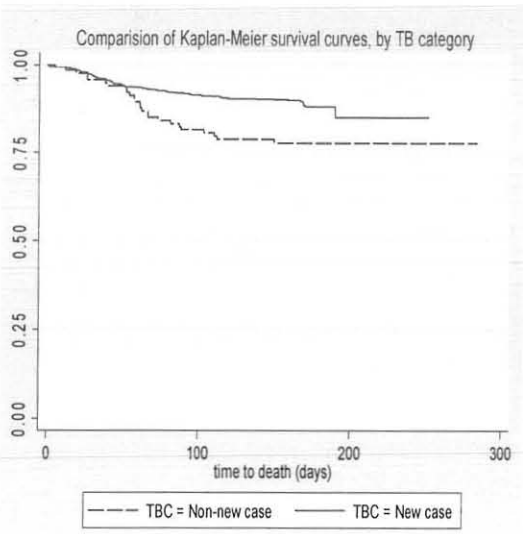
c)


Comparision of Kaplan-Meier survival curves, by initial weight
0.00
0.25
0.50
0.75
1.00
0
100
200
300
time to death (days)
weightcat = <35kgs
weightcat = >=35kgs

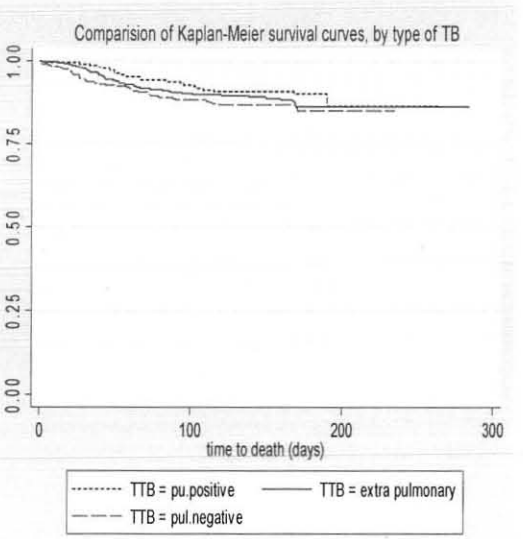
d)


Comparision of Kaplan-Meier survival curves, by smear result
0.00
0.25
0.50
0.75
1.00
0
100
200
300
time to death (days)
sresult = negative
sresult = positive

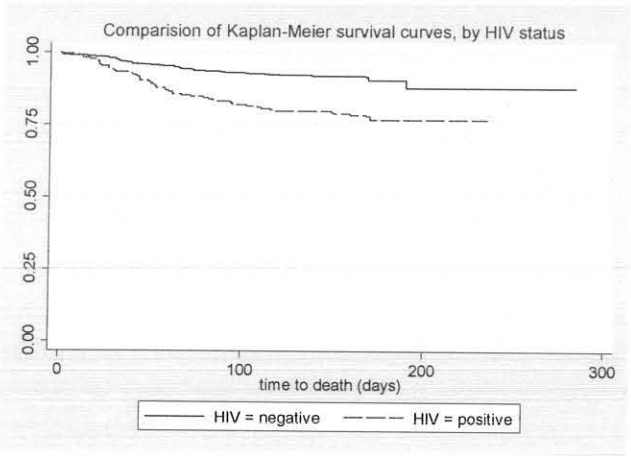
e)


Comparision of Kaplan-Meier survival curves, by TB category
0.00
0.25
0.50
0.75
1.00
0
100
200
300
time to death (days)
TBC = Non-new case
TBC = New case

f)


Comparision of Kaplan-Meier survival curves, by type of TB
0.00
0.25
0.50
0.75
1.00
0
100
200
300
time to death (days)
TTB = pu.positive
TTB = extra pulmonary
TTB = pul.negative

g)


Comparision of Kaplan-Meier survival curves, by HIV status
1.00
0.75
0.50
0.25
0.00
0
100
200
300
time to death (days)
HIV = negative
HIV = positive