University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Biometry

Modelling Survival Among Patients With Prostate Cancer in Kenya

Research Report in Biometry, Number 42, 2021

John Gaitho

July 2021

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Modelling Survival Among Patients With Prostate Cancer in Kenya

Research Report in Biometry, Number 42, 2021

John Gaitho

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis
Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Prostate cancer is among the most common types of cancers in men in Kenya and worldwide. The exact cause of the prostate cancer is unclear. However, previous research has indicated that there are risk factors likely to be associated with prostate cancer such as: old age, race where black people are at more risk, family history, obesity among other risk factors. In the year 2018 among men, it was the second most diagnosed cancer (1.3 million, 14.5%) after lung cancer (1.4 million, 15.5%) worldwide

This study aims to compare survival between clinical stages and treatment categories. The study further aims to investigate the impact age on survival among prostate cancer patients in Kenya. The study used secondary population-based prostate cancer data obtained from the KEMRI National Cancer Registry. Kaplan-Meier estimation method was used to compare survival function for clinical stage and treatment. Cox proportional hazards regression model was used to model age against survival time.

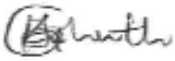
The Cox proportional hazards regression results checking for association between survival time and age were as follows: HR = 1.032 (CI: 1.004 – 1.061). Thus for each additional year an individual diagnosed with prostate cancer is 3.2% more likely to die per unit time (in days). However, the Kaplan-Meier results showed no significant differences in survival for clinical stage and treatment type categories.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.



23-Aug-2021

Signature Date

JOHN GAITHO
Reg No. I56/35921/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

August 23rd 2021

Signature Date

Dr Idah Orowe
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: orowe@uonbi.ac.ke

Dedication

This project has been dedicated to my family members for their encouragement and support.

Contents

Figures and Tables

Figures

Tables

Acknowledgments

To start with, I thank my supervisor Dr. Ida Orowe for her guidance, inspiration, and unending reassurance while working on this project. Her immense knowledge, supervision and corrections saw me through successful completion of this work. Secondly, I thank my classmates, for their encouragement and critique. Thirdly, I thank the School of Mathematics, University of Nairobi for the chance to study Masters of Science in Biometry. Not forgeting my family members for their patience, encouragement and understanding during the study period.

Again to you all, thank you.

John Gaitho

Nairobi, 2021.

1 Chapter 1: Introduction

1.1 Background

Cancer is characterized by an uncontrollable cell division and growth in an organ or a specific part of the body that results in destruction of the body tissue. It emerges from alteration of normal body cells into tumor cells through a multistage process starting from pre-tumor to a malignant tumor. In some instances, this abnormal growth of cells shows a tendency to proliferate and metastasize (spread from original organ to other body parts). Cancers are often named according the body part/organ where they first appear.

Prostate cancer is one of the most common types of cancer in men. It affects the prostate gland in men. The cause of the prostate cancer is unclear. However, previous research has indicated that there are risk factors that are associated with prostate cancer such as: old age, race where black people are at more risk, family history, obesity among other risk factors.

In the year 2018, the most diagnosed cancers across the world were lung cancer (2.09 million, 12.3%), breast cancer (2.09 million, 12.3%), colorectal cancer (1.8 million, 10.6%), and prostate cancer (1.3 million, 7.5%). Among men, prostate cancer was the second most diagnosed cancer (1.3 million, 14.5%) after lung cancer (1.4 million, 15.5%).

Cancer is the third leading cause of deaths after infectious diseases and cardiovascular diseases. Prostate cancer is one of the cancers with incidences that have been on the rise in most regions of the world. Even though prostate cancer cases remain high in developed countries, there are relatively more deaths in low- and middle-income countries as a result of late diagnosis and lack of treatment facilities.

For some undetermined reasons, prostate cancer has the highest incident rate among men of African descent. For instance, the number of prostate cancer cases diagnosed among men of African descent is close to 60% higher than Caucasian American men, with a mortality rate of between two and three times more.

1.2 Problem Statement

Worldwide, with exception of non-melanoma skin cancer, the most common cancer among males is lung cancer, accounting for about 15.5% of all cancer cases detected during 2018,

followed by prostate cancer which accounted for about 14.5%. Cancer incidence burden has been predicted to rise to over 85% by 2030 thereby resulting to an increase in cancer-related deaths. Of all cancer cases documented in the Nairobi Cancer Registry between 2004 and 2008, prostate cancer was the most diagnosed type of cancer among males with age standardized incidence rate of 40.6 per every 100,000. As a result there were more deaths due to prostate cancer among males than any other cancer type. Statistical modelling of cancer survival have not been explored adequately by previous research in the Kenyan setting. The present study will be an important source of information for the factors significantly associated with survival time among prostate cancer in Kenya.

1.3 Objectives

1.3.1 General Objective

The general objective of this study is to investigate factors associated with survival time among prostate cancer patients in Kenya.

1.3.2 Specific Objectives

The following are the specific objectives:

1. To compare survival among prostate cancer patients between clinical stages at diagnosis in Kenya.
2. To compare survival among prostate cancer patients treated with a single treatment and those treated with a combination of two or more treatment types in Kenya.
3. To find out the impact of age on survival among prostate cancer patients in Kenya.

1.4 Significance of the Study

This study intends to identify factors significantly associated with survival among prostate cancer patients for consumption by the Government of Kenya (GoK) and the general public. Through deeper understanding of these factors the government can put in place strategies and formulate policies directed towards thwarting prostate cancer mortality, for example, by setting up cancer centers, public awareness, and so on.

2 Chapter 2: Literature Review

A number of studies have been carried out to model association between various factors and survival among prostate cancer patients. This chapter provides a summary of some of research done before.

Bechis et al., 2011 studied the impact of age at diagnosis on treatment of prostate cancer and overall survival in the United States. The study used data of men in the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) database. The study used only data with complete information on risk, treatment and follow-up. Logistic regression was used to model the likelihood of receiving local treatment (that is, Radical Prostatectomy (RP), External Beam Radiation Therapy (EBRT), Brachytherapy, or Cryotherapy) versus Primary Androgen Deprivation Therapy (PADT) by age at diagnosis controlling for validated Cancer of the Prostate Risk Assessment (CAPRA) score, and the treatment year. They found that older men were more likely to be treated with PADT compared to younger men. Overall survival was modelled using proportional hazards model with survival as the response variable and age as the univariate predictor. The results showed decreased survival with increasing age ($P<.01$).

In another U.S study Hoffman et al., 2001 studied racial and ethnic differences in outcomes of advanced stage prostate cancer. 3,173 men diagnosed with prostate cancer between 1994 and 1995 were enrolled into a population-based cohort. Data on age, race, ethnicity, education level, income, employment status, comorbidity, urinary tract infection (UTI), PSA level, grade and stage was collected. Weighted logistic regression analysis was used to obtain odds ratios for advanced stage prostate cancer. Adjusting for all other covariates, there was statistically significant risk of advanced stage prostate cancer among African-American men compared to non-Hispanic white men (aOR = 2.26; CI = 1.43 to 3.58) as opposed to Hispanic men compared to non-Hispanic white men (aOR = 1.23; CI = 0.73 to 2.08).

Schröder et al., 2012 conducted an 11 years follow-up of 182,160 men from eight European countries to study mortality due to prostate cancer among men between 50 and 74 years of age randomized into either prostate-specific antigen (PSA) based screening or control group (without PSA-based screening). The primary outcome of interest was death from prostate cancer. Nelson–Aalen method was used for calculating the cumulative hazard. Poisson regression was used for calculating rate ratios adjusted per study site. The results showed 21% reduction in deaths due to prostate cancer in the PSA-based screening group compared to the control group. The overall rate ratio was 0.79 (P=0.001). After

further adjusting for noncompliance and selection bias, they obtained a rate ratio of 0.71 (P=0.001).

Farris et al., 2018 carried out case-control study on the effect of post-diagnosis of prostate cancer alcohol consumption on survival among prostate cancer patients in Alberta, Canada between years 1997 – 2000. 829 men diagnosed with prostate cancer were enrolled and followed for a period of up to 19 years to study survival outcomes. Cox proportional hazards regression model was used to model survival time. Survival time was the response while the covariate was post-diagnosis alcohol consumption, adjusting for other factors.The results showed a significant association between alcohol consumption and overall survival with adjusted hazard ratio of drinkers to those not drinking being 1.82 (CI: 1.07–3.10). However, the study recommended future research to confirm burden of disease.

Ardakani et al., 2017 conducted a retrospective study on survival outcomes in men with a prostate cancer diagnosis in Iran. The study used secondary data collected from 100 men in two hospitals in Iran from 2001 to 2012. Kaplan-Meier estimation method was employed to determine survival over time whereas Cox proportional hazards regression analysis was carried out to estimate the hazard ratios with survival time as the response and with age (in years), grade (gleason score), stage, treatment, and residence as the independent variables. The results showed a mean survival age of 70 ± 4.94 month. 1-year's survival rate was 97%, 3-year's survival rate was 73%, and 5-year's survival rate was 54%. Further results from Cox regression analysis showed a significant correlation between grade, stage, age, treatment and survival of men ($P < .05$). The worst survival was seen among men with Gleason score of between 8 and 10 and clinical stage IV. However, there was no significant association between area of residence and survival time ($P > .05$). The study recommended further research to keep track of survival of patients.

In another study conducted in Sweden, Epstein et al., 2011 investigated the impact dietary zinc on survival among men with a prostate cancer diagnosis. This study consisted of 525 men aged below 80 years and with a diagnosis from 1989 to 1994 in Orebro, Sweden. Data on food frequency, and dietary zinc was collected. Cox proportional hazards regression model analysis was used with time to mortality as the depend variable and dietary zinc as the independent variable, adjusting for disease clinical stage at the time of diagnosis (that is, localized or advanced), age, history of prostate cancer, smoking status, year diagnosed (1989–1991 or 1992–1994), grading, and body mass index (BMI). The results showed a reduced risk with high dietary zinc intake (HR was 0.64; with a CI: 0.44–0.94). The results further showed a stronger association among men with localized disease (HR: 0.24; CI: 0.09–0.66).

Bonn et al., 2015 researched on the association between physical exercise and prostate cancer. The study used data that included 4,623 men who had a localized prostate cancer

diagnosis between 1997 and 2002. They were followed up until the year 2012. Cox proportional hazards regression method was used to model the association between physical activity (recreational metabolic equivalent of task (MET hours/day), time spent cycling/walking, on household activities/exercising) and time to death due to prostate cancer and any cause. The model was adjusted for any possible confounder(s). In the follow-up period, there were 561 deaths from any cause whereas there were 194 deaths resulting from prostate cancer. The results further showed a reduced rate of overall mortality among men with ≥ 5 recreational MET-hours/day (HR was 0.63, with a CI: 0.52–0.77), $\geq$ 20 minutes/day cycling/walking (HR was 0.70, with a CI: 0.57–0.86), ≥ 1 hour/day engaging in household activities (HR was 0.71, with a CI: 0.59–0.86), or ≥ 1 hour/week exercising (HR was 0.74; with a CI: 0.61–0.90) in comparison with less active men for each category. They further observed a reduced death rate from prostate cancer among men with ≥ 20 minutes/day cycling/walking (HR was 0.61; with a CI: 0.43–0.87), or ≥ 1 hour/week exercising (HR was 0.68; with a CI: 0.48–0.94) in comparison with less active men for each category.

Pascale et al., 2013 studied the association between the human papillomavirus (HPV) and the risk of prostate cancer. They recruited 150 (out of whom 112 were HPV-positive) to find out the effect of HPV on survival among men with prostate cancer. Kaplan-Meier survival probability estimation method was used in estimating the survival time. The results showed that the prostate cancer patients positive to HPV had reduced overall survival (median was 4.59 years) in comparison with those negative to HPV (median was 8.24 years, with $P < 0.05$). However, they suggested that there is need for further research to clarify probable role of papillomavirus in prostate cancer.

In summary, most studies have explored association of various factors with prostate cancer using a number of methods such as Kaplan-Meier, Cox proportional hazards regression, logistic regression and so on. However, there has been inadequate statistical modelling of prostate cancer in the Kenyan setting. In the current work, log-rank tests, Kaplan-Meier survival probability estimation method, and Cox proportional hazards regression methods are used to investigate the impact of age, clinical cancer stage, and treatment method on mortality from prostate cancer.

3 Chapter 3: Methodology

3.1 Data Source

The study used population-based prostate cancer data obtained from from the Kenya Medical Research Institute (KEMRI) National Cancer Registry.

3.2 Data Management and Analysis

Data management and statistical analysis has been done with Microsoft Excel, SAS version 9.4 and R. Only complete cases were included in the analysis. The statistical inferences were done using the 95% confidence interval.

3.3 Survival Analysis

3.3.1 Introduction

Unlike Ordinary Least Squares regression that model factors associated with a normally distributed response, survival analysis comes in handy in modelling independent variables associated with survival time. Survival analysis is also able to handle **censoring**. Censoring is defined as failure to experience the death/event of interest within the study period. Censoring might occur due to any of the following three reasons:

1. A study participant fails to experience the death/event of interest within the study follow up period.
2. A study participant becomes lost to follow-up before the end of the study period.
3. Due to withdrawal from the study by the participant.

3.3.2 Probability Density Function, PDF

Let $T;\mid T \geq 0$ be a random variable that denotes survival time. The probability density function, also denoted as pdf or $f(t)$, refers to the probability of observing the random variable, $T;\ T \geq 0$, at time t corresponding to any other survival time. To get the probability of observing survival time in an interval a and b, the pdf is integrated.

$$P(a \leq T \leq b) = \int_a^b f(t)\ dt \tag{3.1}$$

Assuming an exponential distribution of survival time T, we have the following equation:

$$P(a \leq T \leq b) = \lambda \int_a^b e^{-\lambda t}\ dt \tag{3.2}$$

with λ being the rate parameter and is equivalent to the inverse of the mean of survival time.

3.3.3 Cumulative Distribution Function, CDF

The cumulative distribution function, also referred to cdf, gives the likelihood of finding $P(T \leq t)$. The following equations show how CDF can be expressed.

$$F(t) = P(T \leq t) \tag{3.3}$$

$$F(t) = \int_0^t f(u)\ du \tag{3.4}$$

This implies that for a given cdf, pdf is obtained by differentiating the cdf as follows:

$$f(t) = \frac{d}{dt}F(t) \tag{3.5}$$

3.3.4 The Survival Probability Function, $S(t)$

The survival probability function, also denoted as $S(t)$), gives the probability of living beyond a given time t (also denoted by $P(T > t)$). It is expressed as follows:

$$S(t) = P(T > t) \tag{3.6}$$

$$S(t) = 1 - F(t) \tag{3.7}$$

where $F(t)$ is the cumulative distribution function.

Since t can take any value within the interval $[0, \infty)$, the survival probability function $S(t)$ has the following characteristics:

1. $S(t)$ has non-increasing property.
2. $S(t) = 1$ at time $t = 0$. Thus the probability of living beyond time $t = 0$ is 1.
3. $S(t) = 0$ at time $t = \infty$, that is, $S(\infty) = 0$ at time $t = \infty$. Thus the survival probability goes to 0 as time t tends to infinity.

3.3.5 The Hazard Function, $h(t)$

The main aim of survival analysis is mainly to model the hazard function. The hazard rate, also denoted by $h(t)$, is defined as the instantaneous rate of occurrence of events at time t given that the individual has survived until time t. The hazard rate is obtained as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \tag{3.8}$$

Hazard rate $h(t)$ has the following relationship with the probability density function $f(t)$ and survival probability function $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} \tag{3.9}$$

3.3.6 The Cumulative Hazard Function, $H(t)$

The cumulative hazard function, also denoted by $H(t)$, is defined as the cumulative hazards over time. Cumulative hazard $H(t)$ is obtained by integrating the hazard function, $h(t)$, over a given time interval.

$$H(t) = \int_0^t h(u) \ du \tag{3.10}$$

This implies that

$$\frac{d}{dt} H(t) = h(t) \tag{3.11}$$

Thus the cumulative hazard function provides a description of accumulation of hazard rates over time.

With the help of the relationships highlighted below

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ f(t) &= \frac{d}{dt} F(t) \\ f(t) &= \frac{d}{dt} (-S(t)) \end{aligned}$$

the equations below can be obtained to show relationships between other survival functions with cumulative hazard function $H(t)$:

(i) The following equation shows the relationship between survival probability function $S(t)$ and cumulative hazard function $H(t)$:

$$S(t) = e^{-H(t)} \tag{3.12}$$

(ii) The following equation shows the relationship between cumulative distribution function $F(t)$ and cumulative hazard function $H(t)$:

$$F(t) = 1 - e^{-H(t)} \tag{3.13}$$

(iii) The following equation shows the relationship between probability density function $f(t)$, cumulative hazard function $H(t)$ and hazard function $h(t)$:

$$f(t) = h(t)e^{-H(t)} \tag{3.14}$$

From the above relationships, the survival functions $S(t)$ and the cumulative hazard function $H(t)$ exhibit a relationship that is monotonic. This means that at the instance where the survival function $S(t)$ is at its highest at the beginning of the observation time, the cumulative hazard $H(t)$ is at its lowest, and as time moves the survival function tends to its lowest whilst the cumulative hazard function tends to its highest.

From the above relationships, it is also clear that the probability density function $f(t)$ will be high at the instance where hazard rate $h(t)$ is high. This happens in most cases at the start of the study. When the cumulative hazard rate $H(t)$ is low, at the start of the study. In short, it is expected to find many failures in a specific time interval if:

(i) There are many individuals at risk.

(ii) There is high hazard rate.

3.3.7 Kaplan-Meier Estimator

The Kaplan-Meier (KP) Estimator is a non-parametric survival probability estimator also referred to as product-limit estimator. It is given by the following equation:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \tag{3.15}$$

The estimate for the variance for the Kaplan-Meier Estimator is given by the following equation:

$$\hat{Var}\left(\hat{S}(t)\right) = \left(\hat{S}(t_i)\right)^2 \sum_{t_i \leq t} \frac{d_i}{n_i\left(n_i - d_i\right)} \tag{3.16}$$

The $100(1-\alpha)\%$ confidence interval for the Kaplan-Meier estimator at time t_i is obtained as follows:

$$CI = \hat{S}(t_i) \pm Z_{\frac{\alpha}{2}} \sqrt{\left(\hat{S}(t_i)\right)^2 \sum_{t_i \leq t} \frac{d_i}{n_i \left(n_i - d_i\right)}} \tag{3.17}$$

where

n_i denotes the number of individuals at risk at time t_i

d_i denotes the number of individuals experiencing the event of interest at time t_i

Therefore each term for the Kaplan-Meier Estimator is obtained by multiplication of the conditional probability of surviving past time t_i given that the individual has lived up to time t_i.

3.3.8 Tests of Hypothesis on Survival Function

Log-Rank and Wilcoxon tests are the common non-parametric tests used for testing of equality of survival probability functions between groups, for example between males and females. These tests are most appropriate for survival data that is skewed to the right and has censoring. The test statistic is obtained as follows :

$$Q = \frac{\left\{\sum_i w_j \left(d_{ij} - \hat{e}_{ij}\right)\right\}^2}{\sum_i w_j^2 \hat{v}_{ij}} \tag{3.18}$$

where

d_{ij} is the actual number of deaths in group i at time j.

$\hat{e}_{ij}$ is the expected number of deaths in group i at time j.

$\hat{v}_{ij}$ is the estimated variance of actual deaths d_{ij}

w_j is weighted difference at time j

The above statistics add the differences (weighted) between actual number deaths and the expected number of deaths for each group at each time, with the assumption that survival probability function is the same for each group. That is, if the survival probability function is the same for all groups, then proportion of deaths is expected to be the same in each time interval. In case the proportions are significantly different among groups at

different times, then Q will have a big value leading to rejection of the null hypothesis of no difference among groups.

The only difference between the Log-Rank and the Wilcoxon tests is the weights used. For the Log-Rank test the weight used is $w_j = 1$ thus equal weighting is done at all intervals whereas for Wilcoxon test the weight used is $w_j = n_j$ thus weighting is done by the number of individuals at risk at all intervals hence more weight is given to differences occurring earlier in the observation time.

3.3.9 Non-parametric Estimation of the Hazard Function

Normal non-parametric methods most of time fail to directly estimate the hazard function $h(t)$. However, a curve of kernal-smoothed estimate may be used to give an idea of hazard rate. Since the hazard function $h(t)$ can be obtained by differentiating the cumulative hazard function $H(t)$, it is possible to obtain a rough estimate of change in cumulative hazard via getting differences within adjacent points as follows:

$$\Delta\hat{H}(t) = \hat{H}(t_i) - \hat{H}(t_{i-1}) \tag{3.19}$$

3.3.10 Cox Proportional Hazards Regression Model

The Cox proportional hazards regression model (Cox, 1972) is a semi-parametric method used for determining the association between the survival (time to event) with a set of predictor variables or a single predictor. The predictor variables can either be continuous or categorical. The model can be used to model the effect of several predictor variables simultaneously on the survival time.

The model is semi-parametric in that no assumptions about the baseline hazard function shape is made. However, the model makes some assumptions as we shall see in the next subsection.

It is possible to express the cox proportional hazards regression model in terms of hazard function as follows:

$$h(t) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} \tag{3.20}$$

Equation (3.20) can as well be expressed as follows:

$$log\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.21}$$

where

- t denotes the survival time.
- $h_0(t)$ denotes the baseline hazard.
- $h(t)$ denotes the hazard due the the effects of predictor variables.
- x_i $i = 1, 2, \ldots, p$ are the predictor variables.
- β_i $i = 1, 2, \ldots, p$ are the coefficients measuring the effect size of the predictor variables.

The values e^{β_i} are the hazard ratios comparing hazard $h(t)$ due to a given predictor variable at a given time point with baseline hazard $h_0(t)$.

If β_i is greater than zero then this means that the hazard ratio is greater than one. This then means that the hazard increases as i^{th} predictor variable increases implying that the length of the survival time reduces.

If β_i is less than zero then this means that the hazard ratio is less than one. This then means that the hazard decreases as i^{th} predictor variable decreases implying that the length of the survival time increases.

If value of β_i is equal to zero. This means that the hazard ratio is equal to one. This the means that there is no effect on hazard for the i^{th} predictor.

In summary,

- HR $>$ 1: Increased hazard.
- HR $<$ 1: Decreased hazard.
- HR = 1: No effect on hazard.

The parameters for the Cox proportional hazards regression model are estimated using the partial likelihood estimate method.
The equation for the partial likelihood estimation method is expressed as follows:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{x_i \beta}}{\sum_{j \in R_i} e^{x_j \beta}} \right\} \tag{3.22}$$

where R_i is the number of individuals still at risk at time t_i. Equation (3.22) gives the β values that give the the biggest combined probability.

Tests of Hypothesis

1. **Wald Test**

   For each predictor, the test of hypothesis can be stated as

$$H_0 : \beta_j = 0 \iff HR_j = 1$$
$$H_1 : \beta_j \neq 0 \iff HR_j \neq 1$$

   The test statistic for for the hypothesis stated above is

$$Z = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \tag{3.23}$$

   or

$$\chi^2 = \left(\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \right)^2 \tag{3.24}$$

   For a factor C with c levels then χ^2 is constructed with the following test statistic

$$\chi^2_{(c-1)} = \hat{\beta}'_C \text{Var} \left(\hat{\beta}_C \right)^{-1} \hat{\beta}_C \tag{3.25}$$

   where $\hat{\beta}_C = (\beta_1, \beta_1, \ldots, \beta_{c-1})'$ are the $(c-1)$ parameter estimates from the $Z_1, Z_2, \ldots, Z_{c-1}$ binary variables from factor C.
   We reject H_0 if the absolute value of the test statistic is greater than the critical value, or when p-value is less than the level of significance α.

2. The Likelihood Ratio Test

   This test involves comparing the full model with the reduced model. Low p-value indicates that the model is significant.

   Let there be $(p+q)$ covariates

$$x_1, x_2, \ldots, x_p, x_{p+1}, x_{p+2}, \ldots, x_{p+q} \tag{3.26}$$

   Consider the two models provided below:

   - **M1**: Model containing p predictors

$$h(t) = h_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} \tag{3.27}$$

- **M2**: Model containing $p+q$ predictors

$$h(t) = h_0(t)e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\beta_{p+1}x_{p+1}+\beta_{p+2}x_{p+2}+\cdots++\beta_{p+q}x_{p+q}} \tag{3.28}$$

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q}$$

The test statistic used has an approximately χ^2 distribution with q degrees of freedom.

$$\chi^2 = -2\{\log L(M1) - \log L(M2)\} \tag{3.29}$$

where $L(M.)$ is maximum partial likelihood for each model.

Cox Proportional Hazards Regression Model Assumptions

The Cox proportional hazards regression models makes some assumptions as noted below:

1. One of the main assumptions of the Cox proportional hazards regression model is that the hazard plots for different categories (or individuals) need to be proportional and should never cross.

   Considering two patients i and i' with different x values. We can express the corresponding hazard functions for the two patients as noted below:

   For the patient i, the hazard function can be written as follows:

$$h_i(t) = h_0 e^{\sum_{j=1}^{n}\beta_j x_j} \tag{3.30}$$

   For the patient i', the hazard function can be written as follows:

$$h_{i'}(t) = h_0 e^{\sum_{j=1}^{n}\beta_j x'_j} \tag{3.31}$$

   We notice that the hazard ratio for these two patients does not depend on time t

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0 e^{\sum_{j=1}^{n}\beta_j x_j}}{h_0 e^{\sum_{j=1}^{n}\beta_j x'_j}} = \frac{e^{\sum_{j=1}^{n}\beta_j x_j}}{e^{\sum_{j=1}^{n}\beta_j x'_j}} \tag{3.32}$$

   As a result, Cox model is clearly a proportional hazards model.

2. Another assumption for the Cox model, as seen above, is that the hazard function and the predictor variables has a multiplicative relationship (unlike a linear relationship as is the case with the multiple linear regression model).

$$HR = \frac{e^{\sum_{j=1}^{n} \beta_j x_j}}{e^{\sum_{j=1}^{n} \beta_j x'_j}} = e^{\sum_{j=1}^{n} \{\beta_j (x_j - x'_j)\}} \tag{3.33}$$

Diagnostics for the Cox proportional hazards regression model

After fitting a Cox proportional hazards regression model, it is important to check for validity of the of the model since it makes several assumptions. The following methods are used to check for the model assumptions:

1. **Proportional hazards assumption** is checked with the help of Schoenfeld residuals.
   Since the Schoenfeld residuals are not dependent on time, a plot showing a random pattern against time means that the proportional hazard assumption has been met. Additionally, a non-significant residuals association with time supports the proportional hazards assumption.

2. **Non-linearity assumption** is usually checked with the help of Martingale residuals. This is done by plotting Martingale residuals against the continuous predictors. Patterns in the plot may indicate a continuous predictor is not fit properly.
   Martingale residuals can contain values between $-\infty$ and $+1$. Values close to +1 denotes subjects that experienced the event very soon. Big negative values denoted subjects that stayed for too long without experiencing the event.
   A plot of continuous predictor against null Cox model's Martingale residuals may be useful in choosing the functional form of the continuous predictor. To meet the proportional hazards regression model assumption, the fitted lines with loess fuction need to be linear.

3. To examine **influential observations/outliers**, we use the Deviance residual. These are normalized transformations of the Martingale residuals. They need to be almost symmetric with a standard normal distribution. Non-negative values denote subjects that experienced the event very soon. Negative values denote subjects that stayed for too long without experiencing the event. Too big values or too small values are influential observations/outliers that have been poorly predicted.

   $dfbeta$ values are also used to test for influential observations/outliers. $dfbeta$ graphs the approximate change in regression parameter estimates after deleting each observation at a time. Influential observations can be identified by checking the $dfbeta$ values compared with the regression coefficients.

4 Chapter 4: Data Analysis and Results

This chapter provides a detailed description of data analysis and the statistical results.

4.1 Descriptive Statistics

The study used data for 413 men who had an average age of 69.0 $\pm$ 10.00 (SD) years, a minimum age of 20.0 years and a maximum age of 96.0 years. Most of men (353; 85%) had unknown clinical cancer stage at diagnosis, 17 (4%) had stage II cancer at diagnosis, 12 (3%) had stage III cancer at diagnosis and 31 (8%) had stage IV cancer at diagnosis. Close to half of men (177; 43%) had unknown treatment, 148 (36%) were treated with only one treatment method (from either surgery, chemotherapy, radiotherapy, hormonetherapy, or immunotherapy), while 88 (21%) were treated with combination of two or more treatment methods.

Table 4.1.1. Counts of participants by clinical stage at diagnosis

Clinical Stage	n	%
Stage II	17	4.1
Stage III	12	2.9
Stage IV	31	7.5
Unknown	353	85.5
Total	413	100.0

n = Number of the study participants. % = Percentages obtained as $\frac{n}{413} \times 100$.

Table 4.1.2. Counts of participants by treatment type categories

Treatment	n	%
Single Treatment Type	148	35.8
2 or more Treatment Types	88	21.3
Unknown	177	42.9
Total	413	100.0

n = Number of the study participants. % = Percentages obtained as $\frac{n}{413} \times 100$.

The least follow up time was 1 day whereas the maximum follow up time was 1820 days. There were 58 (14.0%) men with the outcome of death while 355 (86.0%) men did not experience the event of interest (death) and therefore were censored. Among men with clinical stage II prostate cancer at diagnosis, only 1 (0.2%) had death outcome, while 16 (3.6%)

Table 4.1.3. Summary statistics for age by clinical stage at diagnosis

Clinical Stage	n	Median	Mean	Minimum	Maximum	SD
Stage II	17	67.0	66.5	47	80	10.09
Stage III	12	61.0	60.3	20	74	15.61
Stage IV	31	66.0	67.2	46	90	8.72
Unknown	353	70.0	69.6	46	96	9.74
Total	413	69.0	69.0	20	96	10.00

n = Number of study participants. SD = Standard deviation.

Table 4.1.4. Summary statistics for age by treatment type categories

Treatment type	n	Median	Mean	Minimum	Maximum	SD
Single Treatment Type	148	68.0	68.4	46	92	9.19
2 or more Treatment Types	88	66.0	66.2	20	87	10.36
Unknown	177	71.0	71.0	47	96	10.13
Total	413	69.0	69.0	20	96	10.00

n = Number of study participants. SD = Standard deviation.

were censored. Among those with clinical stage III disease at diagnosis, none had death outcome, while 12 (3.6%) were censored. Among those with clinical stage IV disease at diagnosis, 5 (1.2%) had death outcome, while 26 (6.3%) were censored. Among those with unknown clinical stage, 52 (12.6%) had death outcome, while 301 (72.9%) were censored. Among men treated with single treatment method, 11 (2.7%) had death outcome, while 137 (33.2%) were censored. Among those who received two or more treatment methods, 6 (1.5%) had death outcome, while 82 (19.9%) were censored. Among those with unknown treatment method, 41 (9.9%) had death outcome, while 136 (32.9%) were censored.

Table 4.1.5. Summary statistics for age by treatment type

Treatment sub-type	n	Median	Mean	Minimum	Maximum	SD
Surgery	48	67.0	68.3	51	90	8.74
Radiotherapy	51	68.0	68.0	46	92	8.80
Chemotherapy	36	69.5	69.6	49	90	10.45
Hormonetherapy	13	67.0	66.5	54	78	9.18
Surgery + Radiotherapy	2	54.5	54.5	54	55	0.71
Surgery + Chemotherapy	2	66.0	66.0	66	66	0.00
Surgery + Hormonetherapy	9	64.0	65.8	48	87	12.23
Radiotherapy + Chemotherapy	29	66.0	64.7	46	81	8.56
Radiotherapy + Hormonetherapy	18	63.5	62.1	20	78	13.41
Chemotherapy + Hormonetherapy	13	74.0	73.3	62	86	7.11
Surgery + Radiotherapy + Chemotherapy	3	71.0	73.3	65	84	9.71
Surgery + Radiotherapy + Hormonetherapy	3	72.0	64.0	46	74	15.62
Surgery + Chemotherapy + Hormonetherapy	1	70.0	70.0	70	70	NC
Radiotherapy + Chemotherapy + Hormonetherapy	7	69.0	70.0	64	80	4.93
Surgery + Radiotherapy + Chemotherapy + Hormonetherapy	1	71.0	71.0	71	71	NC
Unknown	177	71.0	71.0	47	96	10.13

n = Number of participants. SD = Standard deviation. NC = Not Calculated.

Table 4.1.6. Summary of outcome status by clinical stage at diagnosis

	n	%
Stage II		
Alive	16	3.9
Dead	1	0.2
Stage III		
Alive	12	2.9
Dead	0	0
Stage IV		
Alive	26	6.3
Dead	5	1.2
Unknown		
Alive	301	72.9
Dead	52	12.6

n = Number of study participants. % = Percentages obtained as $\frac{n}{413} \times 100$.

Table 4.1.7. Summary of outcome status by treatment type category

	n	%
Single Treatment Type		
Alive	137	33.2
Dead	11	2.7
2 or more Treatment Types		
Alive	82	19.9
Dead	6	1.5
Unknown		
Alive	136	32.9
Dead	41	9.9

n = Number of study participants. % = Percentages obtained as $\frac{n}{413} \times 100$.

Table 4.1.8. Summary of outcome status by specific treatment

	n	%
Chemotherapy		
Alive	43	10.4
Dead	5	1.2
Radiotherapy		
Alive	49	11.9
Dead	2	0.5
Chemotherapy		
Alive	33	8.0
Dead	3	0.7
Hormonetherapy		
Alive	12	2.9
Dead	1	0.2
Surgery + Radiotherapy		
Alive	2	0.5
Dead	0	0.0
Surgery + Chemotherapy		
Alive	2	0.5
Dead	0	0.0
Surgery + Hormonetherapy		
Alive	8	1.9
Dead	1	0.2
Radiotherapy + Chemotherapy		
Alive	25	6.1
Dead	4	1.0
Radiotherapy + Hormonetherapy		
Alive	18	4.4
Dead	0	0.0
Chemotherapy + Hormonetherapy		
Alive	12	2.9
Dead	1	0.2

Surgery + Radiotherapy + Chemotherapy		
Alive	3	0.7
Dead	0	0.0
Surgery + Radiotherapy + Hormonetherapy		
Alive	3	0.7
Dead	0	0.0
Surgery + Chemotherapy + Hormonetherapy		
Alive	1	0.2
Dead	0	0.0
Radiotherapy + Chemotherapy + Hormonetherapy		
Alive	7	1.7
Dead	0	0.0
Surgery + Radiotherapy + Chemotherapy + Hormonetherapy		
Alive	1	0.2
Dead	0	0.0
Unknown		
Alive	136	32.9
Dead	41	9.9

4.2 Comparing survival between clinical stages

Clinical stage III had no events of interest (death) and hence was excluded from the analysis.The “Unknown" category was also excluded when calculating the Kaplan-Meir survival probability estimates. Comparison of the survival fuction was therefore made between clinical stage II and IV.

Table 4.2.1. Kaplan-Meier estimates for survival function alongside the 95% CI for stage II and stage IV

t_i	n_i	d_i	$\hat{S}(t)$	**s.e** $\left(\hat{S}(t)\right)$	**Lower 95% CI**	**Upper 95% CI**
Stage II						
0	17	0	1	0	1	1
248	6	1	0.833	0.152	0.583	1
Stage II						
0	31	0	1	0	1	1
8	30	1	0.967	0.0328	0.905	1
39	26	1	0.929	0.0482	0.840	1
174	20	1	0.883	0.0644	0.765	1
497	10	1	0.795	0.1019	0.618	1
692	6	1	0.662	0.1477	0.428	1

t_i = survival time in days. n_i = number of patients at risk. d_i = number of events/deaths. **s.e** = standard error. $\hat{S}(t)$ = survival probability. **CI** = confidence interval.

Figure 4.2.1 shows the Kaplan-Meier curves comparing survival probabilities between prostate cancer patients with clinical stage II with those with stage IV disease. Both the Log-rank test (P value = 0.7) and the Wilcoxon test (P value = 0.7) showed that survival functions between patients with stage II and those with stage IV disease were not significantly different.

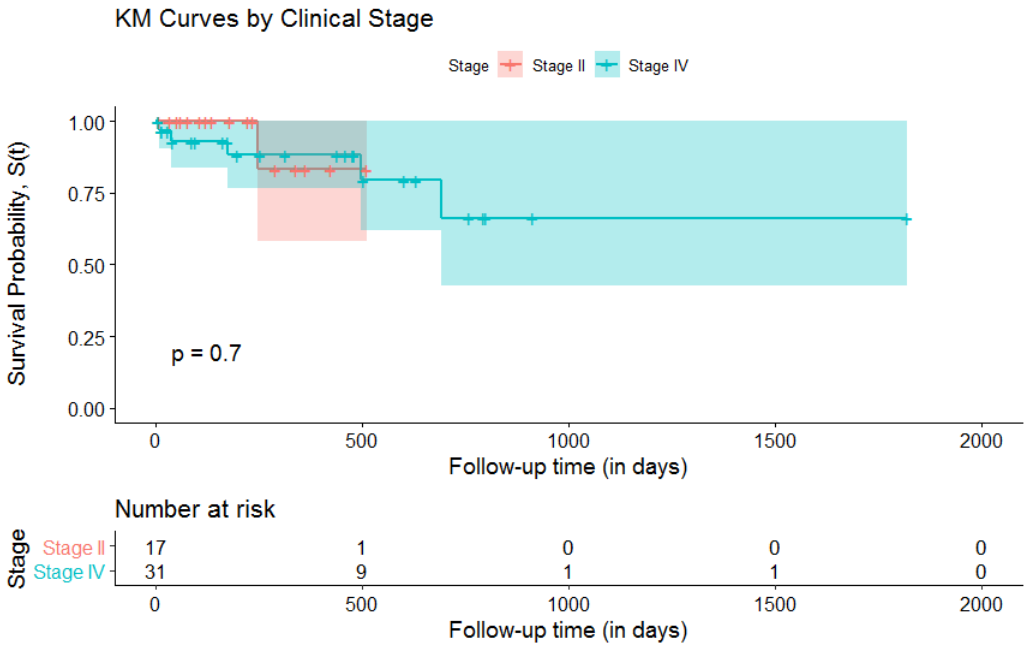


Figure 4.2.1. Kaplan-Meier survival estimates curves by clinical stage

4.3 Comparing Survival Between Patients Treated With Single Treatment Type and Those Treated With 2 or More Treatment Types

Patients were treated with either of the following treatment types: surgery, chemotherapy, radiotherapy, hormone-therapy,or immunotherapy; or with a combination of two or more treatment types listed as listed. This study categorized these treatments into either “single treatment type" or “2 or more treatment types" categories. The “Unknown" category was excluded when calculating the Kaplan-Meier survival functions. Comparison was therefore made between prostate cancer patients treated with single treatment type and those treated with two or more treatment types.

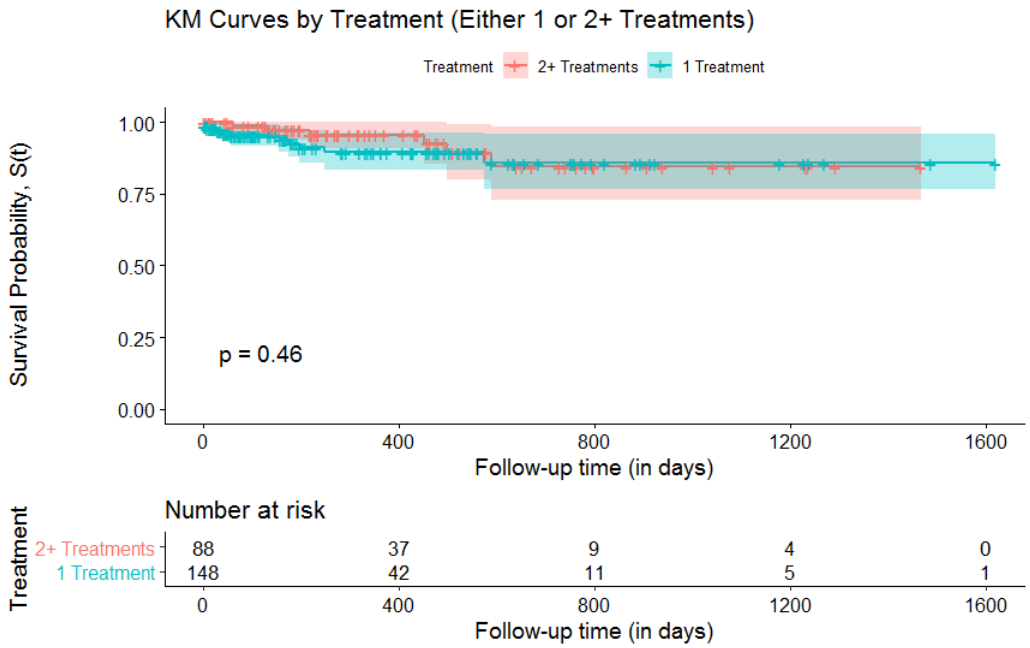


Figure 4.3.1. Kaplan-Meier survival estimates curves by treatment type category

Figure 4.3.1 shows the Kaplan-Meier curves comparing survival probabilities between prostate cancer patients treated with a single treatment with those treated with two or more treatment types. Both the Log-rank test (P value = 0.5) and the Wilcoxon test (P value = 0.4) showed that survival functions between patients treated with one treatment type and those treated with a combination of two or more treatment types were not significantly different.

Table 4.3.1. Kaplan-Meier estimates for survival function alongside the 95% CI for 1 treatment and 2+ treatments

t_i	n_i	d_i	$\hat{S}(t)$	**s.e** $\left(\hat{S}(t)\right)$	**Lower 95% CI**	**Upper 95% CI**
1 Treatment						
0	148	0	1	0	1	1
1	148	2	0.986	0.009	0.968	1
7	136	1	0.979	0.012	0.956	1
26	122	1	0.971	0.014	0.944	1
39	116	1	0.963	0.016	0.931	0.995
57	104	1	0.954	0.019	0.918	0.991
154	75	1	0.941	0.022	0.898	0.986
174	69	1	0.927	0.026	0.878	0.979
195	63	1	0.913	0.029	0.857	0.972
248	56	1	0.896	0.033	0.834	0.963
574	23	1	0.857	0.049	0.766	0.960
2+ Treatments						
0	88	0	1	0	1	1
60	77	1	0.987	0.013	0.962	1
130	70	1	0.973	0.019	0.937	1
218	57	1	0.956	0.025	0.908	1
451	33	1	0.927	0.038	0.856	1
497	27	1	0.893	0.049	0.801	0.995
588	19	1	0.846	0.065	0.727	0.984

t_i = survival time in days. n_i = number of patients at risk. d_i = number of events/deaths. **s.e** = standard error. $\hat{S}(t)$ = survival probability. **CI** = confidence interval.

4.4 Effect of Age on Survival

Cox proportional hazards regression model with time to death (in days) as the response and age as a single covariate was used. . A significant model with response survival time and age as the predictor was found (P value = 0.02).

Likelihood ratio test = 5.19 on 1 df, P value = 0.02

Table 4.4.1. Results of the Cox proportional hazards regression model

Variable	$\hat{\beta}$	s.e $\hat{\beta}$	HR (95% CI)	P value
Age	0.03156	0.01392	1.032 (1.004 − 1.061)	0.02

$\hat{\beta}$ = coefficient. **s.e** = standard error. **HR** = Hazards ratio. **CI** = confidence interval.

Table 4.4.1 shows the results of the Cox proportions hazards regression model model with survival among prostate cancer patients as the response and age as a single covariate. The results show that age is a significant predictor of survival time among prostate cancer patients (P value = 0.02).

4.4.1 Cox Proportional Hazards Model Diagnostics

1. **Proportional hazards assumption**

Table 4.4.2. Testing for proportional hazards assumption

	χ^2	d.f	P value
Age	0.0858	1	0.77
Global	0.0858	1	0.77

The results from table 4.4.2 show that there is no significance for age. The global test is also not significant, hence proportional hazards assumption is supported.

From figure 4.4.1, a plot of Schoenfeld residuals against the time (in days), there is no pattern with time futher supporting proportional hazards assumption.

2. **Checking for outliers**

A $dfbeta$ plot was used to check for influential observations.

From figure 4.4.2 there are no outlying observations after comparing the regression coefficients with magnitudes of the biggest $dfbeta$ values.

3. **Testing for non-linearity**

Figure 4.4.3 is a plot of Martingale residuals of the null model against age. It appears fitted lines are almost linear thus supporting the Cox proportional hazards regression model assumptions.

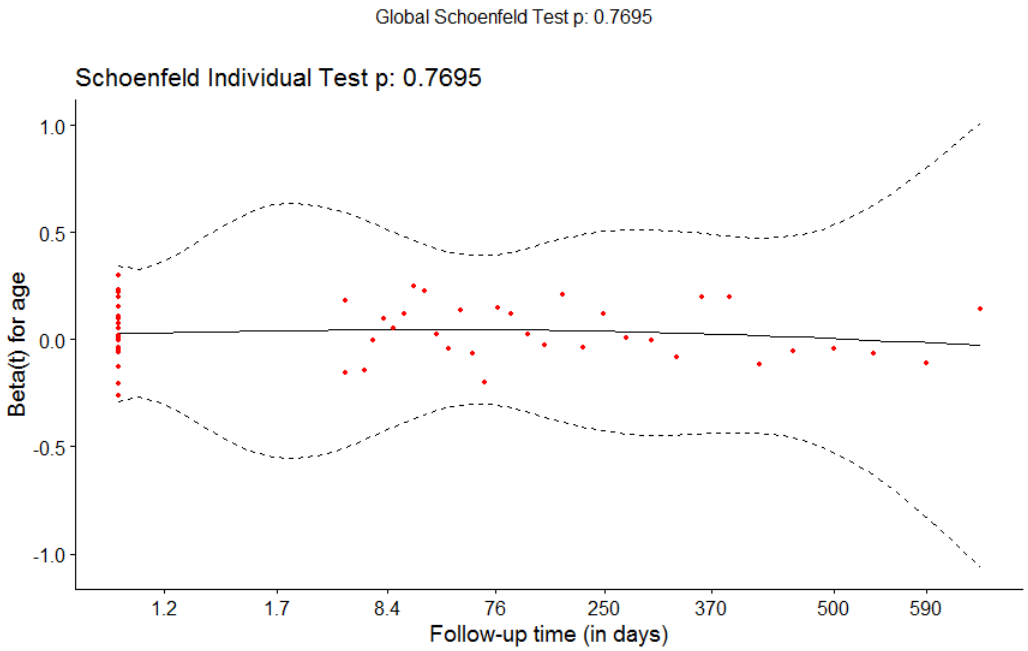


Figure 4.4.1. Schoenfeld residuals against the time (in days)

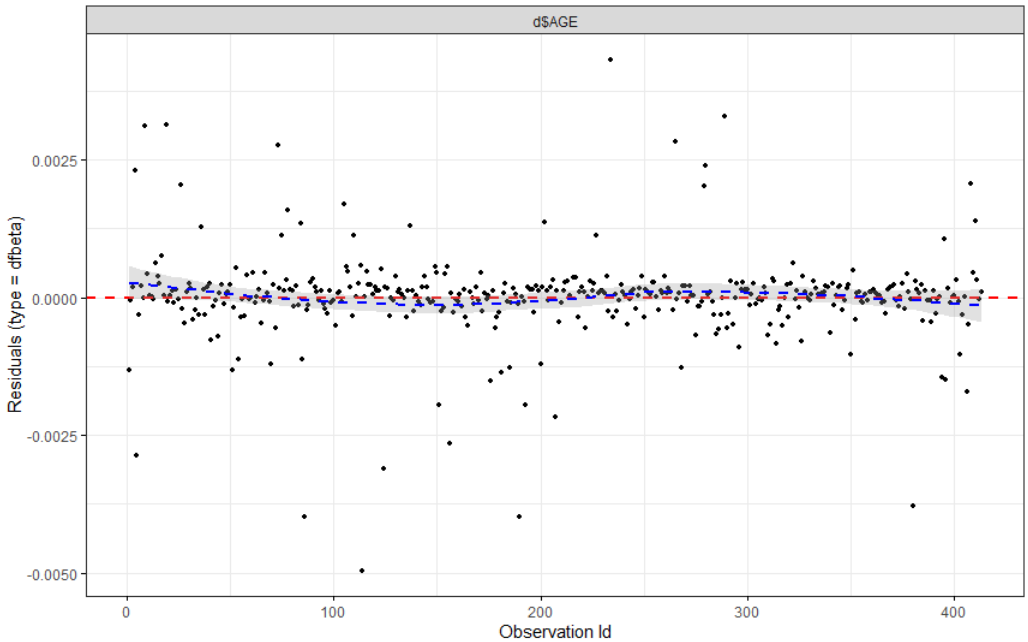


Figure 4.4.2. $dfbeta$ **residuals**

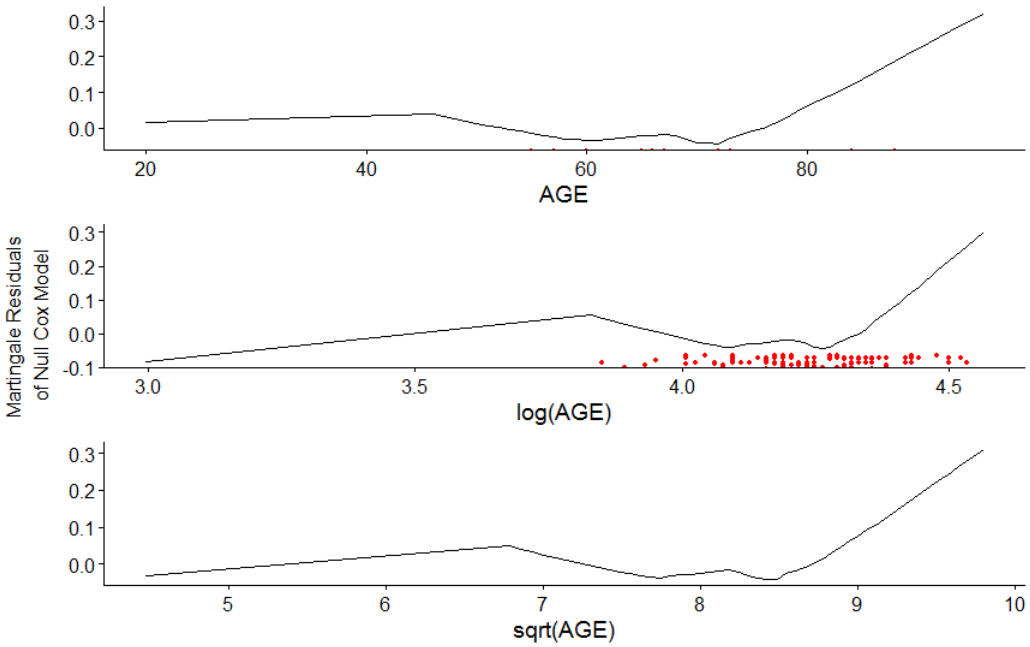


Figure 4.4.3. Martingale residuals for testing for non-linearity

4.4.2 Interpretation of the Results

For each additional year, an individual diagnosed with prostate cancer is 3.1% likely to die per unit time (in days).

5 Chapter 5: Conclusion & Recommendations

5.1 Conclusion

The study used used the Kaplan-Meier survival function estimation method to compare survival between different clinical stages. Kaplan-Meir survival function was further used to compare survival between patients treated with a single treatment and those treated with two or more treatments. Both the Log-rank and Wilcoxon tests were used to test for checking for significant differences in survival between either the clinical stage and treatment categories.

The study used the Cox proportional hazards regression model to determine the association between age and survival among prostate cancer patients in Kenya. Including either the clinical stage and treatment variables resulted to a non-significant model, hence they were excluded from the model. However, age was a significant predictor.

The results showed that there were no significant differences in survival function between patients with stage II and stage IV disease. Stage I and II alongside the unknown categories were excluded from analysis since there were no deaths in stage I and III categories, and to avoid biased results for the unknown clinical stage. Similarly the results showed there no significant differences in survival function between single treatment category and two or more treatment categories. Unknown treatment category was removed from the analysis to avoid biased results.

The Cox proportional regression hazards model showed that age was a significantly associated with survival. The results further showed that the likelihood of experiencing the event of interest, that is, death increased with increase in age.

5.2 Recommendations

As indicated in the results section of the study, the likelihood of death increases with increase in age. Previous studies have shown similar results which has been confirmed by the present study in the Kenyan setting. The government is therefore encouraged to enhance prostate cancer screening among older men. It is also recommended that the government set up more cancer screening and treatment centers.

The KEMRI National Cancer Registry is encouraged to enhance data collection and documentation to ensure data completeness.

5.3 Future Research

This study has mainly focused on age, clinical stage and treatments since they were only the available variables from the the secondary data. Further research utilizing a bigger sample is recommended to confirm the findings in the current study. Effect of more factors on survival among prostate cancer patients in Kenya need to be explored.

References

Ardakani, H. A. V., Moghimi, M., Shayestehpour, M., Doosti, M., & Shahrbabaki, F. S. (2017). Survival outcome in men with prostate cancer in yazd province, central iran, from 2001 to 2012. *Asian Pacific Journal of Cancer Biology*, *2*(2), 23–26.

Balasubramaniam, G., Talole, S., Mahantshetty, U., Saoba, S., & Shrivastava, S. (2013). Prostate cancer: A hospital-based survival study from mumbai, india. *Asian Pacific Journal of Cancer Prevention*, *14*(4), 2595–2598.

Bechis, S. K., Carroll, P. R., & Cooperberg, M. R. (2011). Impact of age at diagnosis on prostate cancer treatment and survival. *Journal of Clinical Oncology*, *29*(2), 235.

Bonn, S. E., Sjölander, A., Lagerros, Y. T., Wiklund, F., Stattin, P., Holmberg, E., Grönberg, H., & Bälter, K. (2015). Physical activity and survival among men diagnosed with prostate cancer. *Cancer Epidemiology and Prevention Biomarkers*, *24*(1), 57–64.

Braun, D. P., Gupta, D., & Staren, E. D. (2012). Predicting survival in prostate cancer: The role of quality of life assessment. *Supportive Care in Cancer*, *20*(6), 1267–1274.

Cassell, A., Yunusa, B., Jalloh, M., Ndoye, M., Mbodji, M. M., Diallo, A., Kouka, S. C., Labou, I., Niang, L., & Gueye, S. M. (2019). Management of advanced and metastatic prostate cancer: A need for a sub-saharan guideline. *Journal of oncology*, *2019*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Epstein, M. M., Kasperzyk, J. L., Andrén, O., Giovannucci, E. L., Wolk, A., Håkansson, N., Andersson, S.-O., Johansson, J.-E., Fall, K., & Mucci, L. A. (2011). Dietary zinc and prostate cancer survival in a swedish cohort. *The American journal of clinical nutrition*, *93*(3), 586–593.

Farris, M. S., Courneya, K. S., Kopciuk, K. A., McGregor, S. E., & Friedenreich, C. M. (2018). Post-diagnosis alcohol intake and prostate cancer survival: A population-based cohort study. *International journal of cancer*, *143*(2), 253–262.

Hoffman, R. M., Gilliland, F. D., Eley, J. W., Harlan, L. C., Stephenson, R. A., Stanford, J. L., Albertson, P. C., Hamilton, A. S., Hunt, W. C., & Potosky, A. L. (2001). Racial and ethnic differences in advanced-stage prostate cancer: The prostate cancer outcomes study. *Journal of the National Cancer Institute*, *93*(5), 388–395.

Hosmer, D. (2008). *Applied survival analysis : Regression modeling of time-to-event data*. Wiley-Interscience.

Kenfield, S. A., Stampfer, M. J., Giovannucci, E., & Chan, J. M. (2011). Physical activity and survival after prostate cancer diagnosis in the health professionals follow-up study. *Journal of Clinical Oncology*, *29*(6), 726.

Kinyao, M., & Kishoyian, G. (2018). Attitude, perceived risk and intention to screen for prostate cancer by adult men in kasikeu sub location, makueni county, kenya. *Annals of Medical and Health Sciences Research*, *8*(3).

Korir, A., Okerosi, N., Ronoh, V., Mutuma, G., & Parkin, M. (2015). Incidence of cancer in n airobi, k enya (2004–2008). *International journal of cancer*, *137*(9), 2053–2059.

Mbugua, R. G., Oluchina, S., & Karanja, S. (2021). Prostate cancer awareness and screening among men in a rural community in kenya: A cross-sectional study. *African Journal of Urology*, *27*(1), 1–10.

Møller, H., Roswall, N., Van Hemelrijck, M., Larsen, S. B., Cuzick, J., Holmberg, L., Overvad, K., & Tjønneland, A. (2015). Prostate cancer incidence, clinical stage and survival in relation to obesity: A prospective cohort study in denmark. *International journal of cancer*, *136*(8), 1940–1947.

Mutua, K., Pertet, A. M., & Otieno, C. (2017). Cultural factors associated with the intent to be screened for prostate cancer among adult men in a rural kenyan community. *BMC Public Health*, *17*(1), 1–8.

Newton, R. U., Kenfield, S. A., Hart, N. H., Chan, J. M., Courneya, K. S., Catto, J., Finn, S. P., Greenwood, R., Hughes, D. C., Mucci, L., et al. (2018). Intense exercise for survival among men with metastatic castrate-resistant prostate cancer (interval-gap4): A multicentre, randomised, controlled phase iii study protocol. *BMJ open*, *8*(5).

Pascale, M., Pracella, D., Barbazza, R., Marongiu, B., Roggero, E., Bonin, S., & Stanta, G. (2013). Is human papillomavirus associated with prostate cancer survival? *Disease markers*, *35*.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., et al. (2012). Prostate-cancer mortality at 11 years of follow-up. *New England Journal of Medicine*, *366*(11), 981–990.

Sonn, G. A., Aronson, W., & Litwin, M. (2005). Impact of diet on prostate cancer: A review. *Prostate Cancer and Prostatic Diseases*, *8*(4), 304–310.

Steele, C. B., Li, J., Huang, B., & Weir, H. K. (2017). Prostate cancer survival in the united states by race and stage (2001-2009): Findings from the concord-2 study. *Cancer*, *123*, 5160–5177.

Tableman, M. (2004). *Survival analysis using s : Analysis of time-to-event data*. Chapman & Hall/CRC.

Wambalaba, F. W., Son, B., Wambalaba, A. E., Nyong'o, D., & Nyong'o, A. (2019). Prevalence and capacity of cancer diagnostics and treatment: A demand and supply survey of health-care facilities in kenya. *Cancer Control*, *26*(1).

Woodward, M. (2014). *Epidemiology : Study design and data analysis*. CRC Press, Taylor & Francis Group.

Yedjou, C. G., Mbemi, A. T., Noubissi, F., Tchounwou, S. S., Tsabang, N., Payton, M., Miele, L., & Tchounwou, P. B. (2019). Prostate cancer disparity, chemoprevention, and treatment by specific medicinal plants. *Nutrients*, *11*(2), 336.