Survival Analysis of Prostate Cancer Cases in the Kingdom of Saudi Arabia

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

BY

ATHEER SALEH ALHUWAYSHIL

DR. MUNNI BEGUM-ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

DECEMBER 2016
Survival Analysis of Prostate Cancer Cases in the Kingdom of Saudi Arabia

A THESIS
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
BY
ATHEER SALEH ALHUWAYSHIL

Committee Approval:

_________________________  _________________________
Committee Chairperson     Date

_________________________  _________________________
Committee Member          Date

_________________________  _________________________
Committee Member          Date

Departmental Approval:

_________________________  _________________________
Departmental Chairperson  Date

_________________________  _________________________
Dean of Graduate School    Date

BALL STATE UNIVERSITY
MUNCIE, INDIANA
DECEMBER 2016
Dedication

To mom and dad

who always make me successful

and the special people in my life
ACKNOWLEDGEMENT

It is with immense pleasure that I express my gratitude to my advisor Dr. Munni Begum for her insightful guidance during the study period. Her cordial cooperation led me to the accomplishment of the thesis. In this occasion I would like to have an opportunity to thank the committee members Dr. Rahmatullah Imon and Dr. Yayuan Xiao for their help throughout the work. I would also like to thank the faculties and stuff of the Department of Mathematical Sciences for their continuous support during my attachment to the department.

My government has indeed been supportive especially with financial assistance just to make sure that I pursue my education dreams to a higher level. Without its generous support I wouldn’t be where I am today. My career looks bright thanks to the government’s support. It is for this reason that I will always endeavor to give back to my country. I look forward to offering my services in relation to my profession so that other people can also benefit.

I would like to specifically appreciate my siblings for having been patient through this past period. They had a choice of making my life difficult but they did not. They have been amazing whenever I needed anything from them irrespective of the time of the day. The fact that I could meet and call them anytime of the day enhanced my focus even more. I am glad to point out that my brother Yasir was always available to my aid whenever my parents were not around or were not in a position to assist me. He would even go an extra mile to ask me if I was okay or needed any kind of help. If I did need anything then I would not hesitate. My lovely sister Areej was amazing too. She understood how time was not on my side and...
would happily step in and even cover for me on various duties at home. This was very special because then I could have some extra time to concentrate.

Last but not least I would like to thank my best friends who also stood by my side. They helped me emotionally. With these special support team, I would nothing more than being a success, I owe it all to them. I am highly indebted to my friends and family members for their all-out support to accomplish my degree successfully, without their cooperation I could not think of it.
## Contents

1 **Introduction**  
   1.1 Background  
   1.2 Objectives of the study  
   1.3 Organization of the thesis  
2 **Methodology**  
   2.1 Introduction  
   2.2 Data source  
   2.3 Data and Variable Description  
   2.4 Data Cleaning  
   2.5 Statistical Analysis  
      2.5.1 Univariate Analysis  
      2.5.2 Bi-variate Analysis  
      2.5.3 Survival Analysis  
   2.6 Statistical Packages  
   2.7 Conclusion  
3 **Results: Univariate & Bivariate Analysis**  
   3.1 Introduction  
   3.2 Univariate Analysis  
   3.3 Bivariate Analysis  
   3.4 Conclusion  
4 **Result: Survival Analysis**  
   4.1 Introduction  
   4.2 Kaplan Meier (KM) Survival curves  
   4.3 Log-rank test  
   4.4 Assessing PH assumption using Graphical Approaches  
   4.5 Cox-PH Model  
      4.5.1 Univariate  
      4.5.2 Multivariate  
   4.6 Model comparison and selection  
   4.7 Interpretation of the coefficients from the selected model  
   4.8 Conclusion  
5 **Conclusion**  
   Bibliography
<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Region-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>15</td>
</tr>
<tr>
<td>3.2 Extent-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>17</td>
</tr>
<tr>
<td>3.3 Grade-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>18</td>
</tr>
<tr>
<td>3.4 Morphology-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>19</td>
</tr>
<tr>
<td>3.5 Status-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>20</td>
</tr>
<tr>
<td>3.6 Basis of diagnosis of Prostate cancer cases in Saudi Arabia</td>
<td>21</td>
</tr>
<tr>
<td>3.7 Summary statistics for age and survival time of Prostate cancer cases in Saudi Arabia</td>
<td>21</td>
</tr>
<tr>
<td>3.8 Results of Chi-square test</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Result from the Cox-PH model by using covariate region</td>
<td>35</td>
</tr>
<tr>
<td>4.2 Result from the Cox-PH model by using covariate extent</td>
<td>36</td>
</tr>
<tr>
<td>4.3 Result from the Cox-PH model by using covariate morphology</td>
<td>36</td>
</tr>
<tr>
<td>4.4 Result from the Cox-PH model by using covariate grade</td>
<td>36</td>
</tr>
<tr>
<td>4.5 Result from the Cox-PH model by using all covariates of interest</td>
<td>37</td>
</tr>
<tr>
<td>4.6 Model diagnostics by using correlation between Schoenfeld residuals and ranked survival time</td>
<td>39</td>
</tr>
<tr>
<td>4.7 Result from the Stratified Cox-Model</td>
<td>41</td>
</tr>
<tr>
<td>4.8 Table of AIC values for the competing Cox-PH models</td>
<td>42</td>
</tr>
</tbody>
</table>
## List of Figures

### Chapter 3

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Region-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>16</td>
</tr>
<tr>
<td>3.2</td>
<td>Extent-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>17</td>
</tr>
<tr>
<td>3.3</td>
<td>Grade-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>18</td>
</tr>
<tr>
<td>3.4</td>
<td>Morphology-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>19</td>
</tr>
<tr>
<td>3.5</td>
<td>Status-wise distribution of Prostate cancer cases in Saudi Arabia</td>
<td>20</td>
</tr>
<tr>
<td>3.6</td>
<td>Basis of diagnosis of Prostate cancer cases in Saudi Arabia</td>
<td>21</td>
</tr>
<tr>
<td>3.7</td>
<td>Histogram for age of Prostate cancer patients in Saudi Arabia</td>
<td>22</td>
</tr>
<tr>
<td>3.8</td>
<td>Histogram for survival time of Prostate cancer patients in Saudi Arabia</td>
<td>23</td>
</tr>
</tbody>
</table>

### Chapter 4

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Unadjusted KM curve for the prostate cancer cases in Saudi Arabia</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>Extent-wise KM curves for the prostate cancer cases in Saudi Arabia</td>
<td>28</td>
</tr>
<tr>
<td>4.3</td>
<td>Morphology-wise KM curves for the prostate cancer cases in Saudi Arabia</td>
<td>29</td>
</tr>
<tr>
<td>4.4</td>
<td>Grade-wise KM curves for the prostate cancer cases in Saudi Arabia</td>
<td>30</td>
</tr>
<tr>
<td>4.5</td>
<td>Log-log curves by region for the prostate cancer cases in Saudi Arabia</td>
<td>32</td>
</tr>
<tr>
<td>4.6</td>
<td>Log-log curves by extent for the prostate cancer cases in Saudi Arabia</td>
<td>33</td>
</tr>
<tr>
<td>4.7</td>
<td>Log-log curves by morphology for the prostate cancer cases in Saudi Arabia</td>
<td>33</td>
</tr>
<tr>
<td>4.8</td>
<td>Log-log curves by grade for the prostate cancer cases in Saudi Arabia</td>
<td>34</td>
</tr>
<tr>
<td>4.9</td>
<td>PH-assumption checking plot</td>
<td>40</td>
</tr>
</tbody>
</table>
Abstract

THESIS: Survival Analysis of Prostate Cancer Cases in the Kingdom of Saudi Arabia

STUDENT: Atheer Saleh Alhuwayshil

DEGREE: Master of Science

COLLEGE: Science and Humanities

DATE: December 2016

PAGES: 73

Prostate cancer is the third leading cancer cases among males in Saudi Arabia. To the best of our knowledge, no survival analysis of prostate cancer cases among Saudi population is found in literature. In order to have a better understanding on prostate cancer with respect to some selected variables such as region, extent, morphology, grade and on survival times of the patients prompted to the study.

In this study we analyzed 2795 prostate cancer cases (with complete information for all subjects) obtained from Saudi Cancer Registry covering the period January 1994 to March 2016. Frequency tables, graphs and some descriptive statistics for the study variables are presented as a part of exploratory analysis. Bi-variate association of status of the patients and prostate cancer grade, extent, region and morphology respectively are performed using Chi-square test. As a part of survival analysis, we applied non-parametric log rank test and parametric Cox proportional hazards (PH) model. Stratified Cox PH model has been considered as the final model based on lowest AIC value.
Almost seventy-five percent of the prostate cancer cases are reported from Riyadh, Makkah and Eastern regions. Distant metastasis and localized consist of more than eighty percent of the prostate cancer cases in Saudi Arabia. More than half of the patients are diagnosed as cancer patient while they stay at grade III. The portion of the patients diagnosed at early stage is very low. More than ninety-five percent of the patients had Adenocarcinoma. Approximately sixteen percent of the patients died of prostate cancer. Histology of primary cancel cell is the most widely used technique for diagnosis of the disease in Saudi Arabia. Age more than 60 years is the most vulnerable time for occurring prostate cancer in males. Grade, region, extent and morphology are significantly related with status (died or alive) of the patients. City Jazan is significantly different than other cities in Saudi Arabia with respect to hazard of survival. If patients are diagnosed at grades I and II, the probability of surviving can be increased significantly.
Chapter 1

Introduction

1.1 Background

Prostate gland is an organ of male reproductive system. Development of cancer in prostate gland is referred to as the carcinoma of prostate or prostate cancer. Most of the prostate cancer types are of adenocarcinoma which takes place in cells that produce and secret mucus and other kinds of fluids. In most cases, prostate cancer patients do not show any early symptoms. Severe stage of the prostate cancer may cause frequent urination and interruptions the flow of urine. However, mostly these kinds of symptoms are resultant of benign prostate condition. The other most common symptoms are bone pain, often in the vertebrae (bones of the spine), pelvis, or ribs. In usual cases, the prostate cancer develops with a very slow pace. Most of the prostate cancer patients are diagnosed on or after 65 years old [1]. The cancer cell may spread to other parts of the body such as lymph nodes and bones [2]. Diagnosis of prostate cancer can be made in a number way, such as prostate imaging by Ultra Sound (US) or by Magnetic Resonance Imaging (MRI), biopsy of prostate (a way of examining prostate tissue by urologists or radiologists), Gleason Score a process of evaluation of microscopic feature of cancer cell, Tumor markers (a means of examining Prostate-Specific Antigen (PSA)), Staging that determines how far the cancer spreads [3]. Once a patient is diagnosed with prostate cancer, the patient is usually treated with one or a combination of multiple procedures: active surveillance, surgery, radiation therapy, chemotherapy, and biological therapy. Early detection of prostate cancer helps to prevent deaths from prostate cancer.
According to the world cancer fact sheet: world cancer burden (2012), cancer is one of the leading causes of disease worldwide. Approximately 14.1 million new cases occurred in 2012 globally. Among them 1.112 (7.87%) million is the prostate cancer cases. In the year 2012, the estimated prevalence of prostate cancer was 3.924 million (12.07% of all types of cancer) while prevalence of all types of cancers was 32.5 million. Approximately 8.2 million deaths occurred in 2012 due to cancer globally, among them 0.307(3.74%) million was due to prostate cancer. In 2008, it is recoded that approximately 169.3 million years of healthy life were lost worldwide due to cancer and the share of prostate cancer was 4.041 (2.38%) million years. The above noted cancer fact sheet also presented a projection to 2030 and depicts that if the current trend of major types of cancer is going on, an increase of 68% new cases would happen by 2030 as compared to 2012 [4].

World Health Organization – Cancer Country Profiles, 2014 [5] presents that 4,900 deaths among males and 4,300 deaths among females has been recorded due to all sorts of cancer in Saudi Arabia in 2014. In the year 2014, 703 new cases of prostate cancer cases have been recorded which is third leading cancer type among the males in Saudi Arabia. Cancer Incidence Report Saudi Arabia 2010, illustrates that among the male cancer patients of 60-74 years’ age group, prostate cancer is the fourth leading cancer cases while among the cancer patients of age group 75+ years it is the first leading cancer type. In 2010, among 4200 male cancer cases 278 cases were recoded as prostate cancer accounted 6.1% of all newly diagnosed cases. This cancer was ranked sixth in the year 2010. The age specific rate of prostate cancer was 5.5 per 100,000 males in Saudi Arabia. The five regions namely Eastern region, Riyadh, Tabuk, Jouf and Northern region were recorded as high incidence region for prostate cancer in 2010. The median age of diagnosis was 73 years [6].
It is evident from available researches that age, family history and race enhance the risk of prostate cancer. The higher age of male increases the risk for getting prostate cancer, particular genes passed from generation to generation may affect offspring’s prostate cancer risk. So far, it is not sure that a single gene is responsible for higher or lower the risk of prostate cancer. However, a man with first blood relative’s prostate cancer has two to three fold higher risks of prostate cancer [7]. The other risk factors may be included as dietary, medication exposure, infectious disease and sexual factors [3]

In the USA, among the African-American men prostate cancer is more prevalent than other races. There is research evidence that among the African-American men it starts at earlier ages and grow faster than in other racial or ethnic groups. In order to determine the root cause of prostate cancer and to find out a preventive measure, intensive research is going on all over the world but the researchers do not agree on the factors influencing the risk of developing the prostate cancer, either positively or negatively [7].

The National Cancer Institute (NCI), USA maintains survival statistics for stage wise prostate cancer and shows that 5-year survival for both local stage and regional stage is nearly 100% while for the distant stage it about 28% [8], which indicates early detection may increase the survival times of the prostate cancer patients. Literature depicts that survival analysis on the prostate cancer patients has been carried out in different countries. For example, a study on survival and mortality in prostatic cancer has been done based on Swedish cancer register [9]. However, as no survival analysis found in literature on Saudi prostate cancer patients, this research project address a survival analysis to identify the risk factors influencing survival time of the prostate cancer patients.
We have collected prostate cancer data from the Saudi Cancer Registry (SCR) for this study. SCR collects data on cancer patients from all over the country through designated office and hospital outlets. Our data includes information on patient’s age at diagnosis, date of birth, date of diagnosis, and last data of contact, city, extent, morphology, topography, laterality, and status of the patient at the time of last contact, causes of death. Subjects with complete information on all variables have been analysed in this study.

1.2 Objectives of the study

This research intends an in-depth analysis of prostate cancer scenario exists in Saudi Arabia with respect to the variables of interest.

The main objective of the research is to fit a survival model that best describes survival time of the prostate cancer patients in relation to the appropriate predictors in the data set.

The specific objectives are:

(1) To examine the distribution of prostate cancer cases with respect to the characteristics such as grade, extent, region, morphology, status (case dead or alive at the time of last contact), basis of diagnosis and age.

(2) To examine the relationship between status (case dead or alive at the time of last contact) with each of the variables of interests namely region, extent, morphology and grade of prostate cancer respectively.
1.3 Organization of the thesis

The current chapter describes the rationale of the study and objectives followed by organization of the thesis. The second chapter discusses the methodology of the analysis of data with brief sketch of statistical methods. Chapter 3 includes Exploratory Data Analysis (EDA) results. We present results and interpretations of findings from survival analysis along with Cox Proportional hazards model in chapter 4. Chapter 5 concludes the study followed by a reference. R- codes are presented in the appendix.
Chapter 2
Methodology

2.1 Introduction

The schematic approach followed in a research study known as methodology of the research, which includes formulation of research questions, collection of information or data in light of the research question, analyzing the data and finally presenting conclusion and implications. The current study explores facts and figures on prostate cancer in Saudi Arabia with a view to providing in-depth understanding of prevalence of prostate cancer as well as the survival pattern and the risk factors for the group of prostate cancer patients. In this chapter, we discuss about the data source, data and variables, data cleaning, univariate, bivariate and multivariate analysis as well.

2.2 Data source

Customized secondary data on prostate cancer have been collected from Ministry of Health (MOH), Saudi Arabia [10] by contacting open data library personnel. Saudi Cancer Registry (SCR) is responsible for following up the cancer patients in different parts of the country, collecting and keeping records accordingly. The Cancer Registry provides the support for early detection and screening of cancer cases. It also helps cancer research projects through providing accurate cancer data. For nationwide cancer data collection, five hospital based offices and five regional branches are employed. The SCR is supposed to supervise the regional offices to ensure wide coverage and authentic data. The supervision process entails verification of site, morphology, and staging information collection along with case linkage
(tumor and patient). Cancer is categorized as mandatory notifiable diseases in Saudi Arabia. Cancer data are accumulated from patient’s medical record, clinic, and/or histopathological diagnosis by SCR trained cancer registrars [11]. The prostate cancer data show that the data collection is continuous follow up of the patients since the diagnosis of the diseases starting from 1994 to March, 2016.

2.3 Data and Variable description

The data set provided by Cancer Registry consists of 4501 prostate cancer patients. Information was recorded for sex, age, address code (region), date of diagnosis of prostate cancer, topography, morphology, behavior of the cancer, grade of the cancer, extent, laterality, diagnosis basis, last contact date, status (death or alive) and causes of death (in case of death) at the time of last contact. Patient’s age at the time of first diagnosis is the quantitative variable in the data set and the others are categorical (either binary or multi-category). Descriptions of some of the variables are given below.

Region:

The variable region includes the big cities of Saudi Arabia. They are Asir, Baha, Hail, Jazan, Jouf, Madianh, Makka, Najran, Estern (a number of cities in eastern region), Northern (a number of cities in northern region), Qassim, Riyadh, Tabuk, some other unknown cities and patients from international community.

Date of diagnosis:

Date of diagnosis refers to the date on which the patient is diagnosed as prostate cancer cases for the first time.
**Morphology:**

Refers to the histology of the malignancies that are identified and coded in accordance with the Classification of Diseases for Oncology 3rd Edition (ICD-O-3) [11, 12]. The topography indicates the site of the cancer in the prostate. All of them are recoded as “prostate gland” in the study data set.

**Grade:**

The pathologists want to estimate how closely the cancer cells match with the healthy and mature cells whenever a cancer is detected. The cancer cells that do not match with their healthy counterparts are referred to as undifferentiated. Pathological grade assigned to a tumor according to the degree of aggressiveness determined by how the tissue looks under the microscope. The grades are defined as well differentiated (grade 1), moderately differentiated (grade 2), poorly differentiated (grade 3), or undifferentiated (grade 4). Grade I is treated as healthier condition than Grade 4 [13].

**Extent:**

Extent of disease is a detailed description of how far the tumor has spread from the organ or site of origin [13]. In this data extent for cases recorded as distant metastasis, localized, regional: not otherwise specified (NOS), direct extension and lymph node, direct extension and lymph node.
Laterality:

Laterality refers to the occurrence of cancer at which prostate. In this data set it is recorded as “not paired (unknown)” for all cases. Status reveals that whether a patient is alive or dead at time of last contact.

One of the primary objectives of this research is to study the survival pattern and to identify the risk factors for the group for prostate cancer patients in cancer registry of KSA. Thus to create a survival time variable, we subtracted the date of diagnosis from the date of last contact for each patient and converted them as survival time in years. For the status variable, death is considered as event and alive is considered as censored.

2.4 Data cleaning

The original dataset we obtained from Saudi Cancer Registry contains a lot of missing information on some variables for the patients. We have deleted those cases to create final analysis data set. For a number of variables some of the categories were unknown and some of the categories had negligible frequencies, we have deleted those from the data set. For example, for the region variable we have deleted all unknown items and the category ‘international’ as it had very negligible frequency. In particular, the variable Morphology contains a huge number of categories, we have kept only couples of categories with high frequencies. We have kept death cases only for due to prostate cancer cases. Therefore, we dealt with only complete cases. Finally, our analysis data set is consisting of 2795 complete cases.
2.5 Statistical Analysis

In order to explore our data, we conducted univariate analysis and bivariate analysis as the measures descriptive statistics and measures of pairwise association among the variables of interests respectively. For survival analysis we considered Kaplan-Meier non-parametric analysis and Cox proportional hazards regression model.

2.5.1 Univariate analysis

As a means of exploratory data analysis of the study variables, we constructed frequency tables for each of the categorical variables, which contain percentages for respective categories. In particular, frequency tables are created for region, grade of the prostate cancer, morphology, extent, status, basis of diagnosis and for causes of death. As there is no variation in behavior and in laterality, we have excluded them from analysis. Descriptive statistics, such as five number summary, are calculated for the continuous variables age and survival times. Histograms for the continuous variables and bar diagrams and pie diagrams for categorical variables are constructed as a means of graphical presentations.

2.5.2 Bi-variate analysis

One of our objectives was to examine the association between Status and Grade, Status and morphology, Status and extent and Status and region. For this purpose, we have performed Chi-square tests. Where status variable corresponds to either death or alive. The mathematical form of Chi-square test statistics can be written as:
\[ \chi^2_{\text{calculated}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \quad i = 1.2...r \text{ and } j = 1.2...c \]

With \( df = (r - 1)(c - 1) \), Where \( r \) indicates the number of rows and \( c \) the number of columns in an \( r \times c \) contingency table. If \( P(X > \chi^2_{\text{calculated}}) \) is less than 0.05, we would consider that there is association between two variables under consideration.

### 2.5.3 Survival Analysis

In survival analysis part we have considered survival time in year as well as censoring information status as our response variable. The variables region, grade, morphology and extent are considered as the potential covariates in according to our research hypothesis: is there any impact of those variables on survival time while other factors remain constant? We have excluded basis diagnosis from our analysis considering that it might not have any causal relationship with survival time.

For survival analysis we have followed the book: Survival Analysis-A Self-Learning Text, third edition [14] and the R code provided in the book. At first we have constructed Kaplan-Meier survival curve popularly known as KM curve with confidence interval without considering covariates. Secondly we have made KM curve for each of the variables of interest namely for region, grade, extent and morphology. Thirdly, we have checked whether survival curves differ with respect to the categories of the respective variables by using log-rank test. Before fitting Cox regression, we have checked Proportional hazard (PH) assumption by graphical analysis in the fourth stage. However, graphical approach did not work for some variables due to too many categories. In the fifth step, a univariate Cox model is fitted followed by a full Cox regression model. Model diagnostics are checked by using correlation between Schoenfeld residuals and ranked survival time. It is found that only Extent did not satisfy PH assumption. Finally, a stratified cox model is fitted
by considering Extent as the strata variable and model diagnostics are checked again. The stratified model with covariates region, grade and morphology is selected as the final model.

Some of the survival concepts and methods used in the study have been presented below:

(a) In order to calculate survival time, we have considered right censoring. The date of last contacts are treated as the end of the study for each of the patients.

(b) Hazard function formula: \( \hat{h}(t) = \lim_{\delta t \to \infty} \frac{P(t \leq S + \delta | S \geq t)}{\delta t} \)

(c) Survival function formula: \( S(t) = \exp\left(- \int_0^t h(u) du \right) \)

(d) KM product limit formula: \( \hat{S}(t) = \prod_{i=1}^{t-1} \Pr(T > t_i | T \geq t_i) \)

(e) The log-rank test for several groups is approximately \( \chi^2 \) test with degrees of freedom (G-1), where G is the number of groups to be compared. However, this is not required as computer program calculates the exact log rank statistic. The \( H_0 \) is considered as: no difference among the group-wise survival curves.

(f) The formula for Cox PH model is: \( h(t, x) = h_0(t) \exp(\sum \beta_i X_i) \); where \( h_0(t) \) is known as baseline hazard. The Cox PH model does not require any distributional assumption of the response variable (survival time). It is also known as semi-parametric model.

(g) Formula for Hazard ratio comparing two individuals:

\[
\frac{h(t, X^*)}{h(t, X)} = \exp(\sum \beta_i (X^*_i - X_i)) \], where \( X_1, X_2, ... X_p \) denotes the covariates.
(h) The meaning of PH assumption: Hazard ratio is independent of time and hence hazard of two X’s are proportional:

\[ \frac{h(t, x^*)}{h(t, x)} = \tilde{\theta} \Rightarrow h(t, X^*) = \tilde{\theta} h(t, X) \]

2.6 Statistical Packages used

The original data file comes in Excel file. Therefore, some initial data manipulation done by Excel. The rest of analysis is performed using R computational software. The survival R package “survival” has been used for survival analysis.

2.7 Conclusion

All the analysis is performed with R computational software. We have considered 5% level of significance for hypothesis testing. For model comparison in survival part we have used AIC value. The model with lowest AIC value is considered as the best model under comparison.
Chapter 3
Result: Univariate and Bivariate Analysis

3.1 Introduction

This study uses information on 2795 cases (complete cases) of prostate cancer patient from Saudi Arabia as mentioned in Chapter 2 (section 2.4). The current chapter includes summary results from univariate and bivariate analysis of the prostate cancer data collected from Saudi Arabia. We have performed some exploratory analysis in the univariate analysis part that basically includes frequency tables and graphs of the categorical variables and descriptive statistics for the continuous variables. The categorical variables we have included here are region, extent, grade, morphology, status and basis of diagnosis. The quantitative variable are age and survival time in years, both of them are continuous in nature. In particular, we have constructed frequency tables for all categorical variables that contain counts, percentages. As a part of graphical presentation we have constructed bar diagram for region, grade, morphology, pie chart for extent, status and histogram for age and survival time (years). Summary statistics for age and survival time are also presented.

In bi-variate analysis part we have checked the association between status (whether a patient is died or alive at the time of last contact) variable and each of the selected categorical variables respectively. One of our objective is to observe how the selected variables interact with survival time of the prostate cancer patients, that is why we are also interested in examining their association with survival status (status) respectively. We have excluded basis of diagnosis with a consideration that this variable might not have any impact on survival status of the patients. In order to check the association, Chi-square test has been performed.
3.2 Univariate Analysis

In the univariate analysis part, we are doing exploratory analysis for each of the variables individually without considering any association or causal relationship with other variables. The analysis provides us better insight of the study data with the respective variable’s distribution prevailing in Saudi Arabia.

Table 3.1: Region-wise distribution of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Region</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asir</td>
<td>244</td>
<td>8.73</td>
</tr>
<tr>
<td>Baha</td>
<td>37</td>
<td>1.32</td>
</tr>
<tr>
<td>Eastern</td>
<td>624</td>
<td>22.33</td>
</tr>
<tr>
<td>Hail</td>
<td>34</td>
<td>1.22</td>
</tr>
<tr>
<td>Jazan</td>
<td>85</td>
<td>3.04</td>
</tr>
<tr>
<td>Jouf</td>
<td>38</td>
<td>1.36</td>
</tr>
<tr>
<td>Madinah</td>
<td>111</td>
<td>3.97</td>
</tr>
<tr>
<td>Makkah</td>
<td>644</td>
<td>23.04</td>
</tr>
<tr>
<td>Najran</td>
<td>35</td>
<td>1.25</td>
</tr>
<tr>
<td>Northern</td>
<td>12</td>
<td>0.43</td>
</tr>
<tr>
<td>Qassim</td>
<td>74</td>
<td>2.65</td>
</tr>
<tr>
<td>Riyadh</td>
<td>824</td>
<td>29.48</td>
</tr>
<tr>
<td>Tabuk</td>
<td>33</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Both Table 3.1 and Figure 3.1 show that the highest number of prostate cancer cases reported in Riyadh city (824 cases, 29.48%) followed by Makkah (644 cases, 23.04%). The eastern region occupies the third position with respect reporting of cancer cases (624 cases, 22.33%). Basically, the eastern region is defined as group of cities located in eastern part of Saudi Arabia. Among the cities noted here, the northern cities reported the least number of cancer cases (12 cases, 0.43%). Approximately 9% of the prostate cancer reported from the city Asir. The cities Tabuk, Hail, Najran, Jouf and Baha look similar (ranging from 1.18% to 1.36%) reporting wise. The city Madinah experiences approximately 4% of the prostate cancer cases.
Figure 3.1: Region-wise distribution of Prostate cancer cases in Saudi Arabia

The Table 3.2 and Figure 3.2, illustrate that the highest frequent extent of prostate cancer is localized (1593 cases, 56.99%) followed by distant metastasis (894 cases, 31.99%). The regional: direct extension is reported as 8.12% of the total cases. The least frequent extent is Regional: Not otherwise specified

Only the Regional: Lymph node shares a portion of 1.68% (47 cases) of the total cases while Regional: Direct extension & Lymph node shares 1.11% (31 cases).
### Table 3.2: Extent-wise distribution of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Extent category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distant Metastasis</td>
<td>894</td>
<td>31.99</td>
</tr>
<tr>
<td>Localized</td>
<td>1593</td>
<td>56.99</td>
</tr>
<tr>
<td>Regional : Not otherwise specified</td>
<td>3</td>
<td>0.11</td>
</tr>
<tr>
<td>Regional : Direct extension &amp; Lymph node</td>
<td>31</td>
<td>1.11</td>
</tr>
<tr>
<td>Regional: Direct Extension</td>
<td>227</td>
<td>8.12</td>
</tr>
<tr>
<td>Regional: Lymph node</td>
<td>47</td>
<td>1.68</td>
</tr>
</tbody>
</table>

### Figure 3.2: Extent-wise distribution of Prostate cancer cases in Saudi Arabia

Grade refers to the status of prostate cancer cells while the patient first time diagnosed as cancer patient. The higher the grade the worse the patient’s condition. The Table 3.3 as well as Figure 3.3 depict that majority (1577 cases, 56.42%) of the prostate cancer cases are diagnosed as cancer patient while cancer cell status remains at grade III, which is quite alarming. Only 33.38% patients are diagnosed while at Grade II. There are also some cases (43 cases, 1.54%) who diagnosed at grade IV. At the primary stage of the disease, only 8.66% (242 cases) are diagnosed as prostate cancer cases in Saudi Arabia.
Table 3.3: Grade-wise distribution of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Grade Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade I (well diff)</td>
<td>242</td>
<td>8.66</td>
</tr>
<tr>
<td>Grade II (Mod diff)</td>
<td>933</td>
<td>33.38</td>
</tr>
<tr>
<td>Grade III (Poor diff)</td>
<td>1577</td>
<td>56.42</td>
</tr>
<tr>
<td>Grade IV (Undiff Anaplastic)</td>
<td>43</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Figure 3.3: Grade-wise distribution of Prostate cancer cases in Saudi Arabia

Morphology refers to the histology of the malignancies are identified in the prostate cancer cell. In the study data set five categories of morphology have been reported, among them the category Adenocarcinoma, NOS reported as the highest frequent case (2672 cases, 95.6%). Transitional cell carcinoma, NOS is the least frequent case of morphology of prostate cancer (5 cases, 0.18%) in Saudi Arabia as shown in the Table 3.4 and Figure 3.4 as well. Acinar Cell carcinoma has been recorded as 1.75% (49 cases) of the total. Carcinoma (Not otherwise specified: NOS) has been recorded as the second highest frequent cases (2.25) of morphology.
Table 3.4: Morphology-wise distribution of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Morphology category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acinar Cell carcinoma</td>
<td>49</td>
<td>1.75</td>
</tr>
<tr>
<td>Adenocarcinoma,NOS</td>
<td>2672</td>
<td>95.6</td>
</tr>
<tr>
<td>Carcinoma,NOS</td>
<td>63</td>
<td>2.25</td>
</tr>
<tr>
<td>Neoplasm,malignant</td>
<td>6</td>
<td>0.21</td>
</tr>
<tr>
<td>Transitional cell carcinoma,NOS</td>
<td>5</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Figure 3.4: Morphology-wise distribution of Prostate cancer cases in Saudi Arabia

The study consists of 2795 prostate cancer cases, among them 2336 (85.58%) were alive at the date of last contact the rest are dead as illustrated in Table 3.5 and Figure 3.5.
Table 3.5: Status-wise distribution of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Status</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>2336</td>
<td>83.58</td>
</tr>
<tr>
<td>Dead</td>
<td>459</td>
<td>16.42</td>
</tr>
</tbody>
</table>

Figure 3.5: Status-wise distribution of Prostate cancer cases in Saudi Arabia

The Table 3.6 and Figure 3.6 show that histology of primary tumor cell is the mostly used means of diagnosis of prostate cancer in Saudi Arabia. This method is used in almost 99% cases. The other means such as cytology/hematological, history of metastases are rarely useful in Saudi Arabia.
Table 3.6: Basis of diagnosis of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Basis of Diagnosis</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytology/Hematological</td>
<td>12</td>
<td>0.43</td>
</tr>
<tr>
<td>Histology of metastases</td>
<td>15</td>
<td>0.54</td>
</tr>
<tr>
<td>Histology of primary tumor cell</td>
<td>2763</td>
<td>98.86</td>
</tr>
<tr>
<td>Unknown</td>
<td>5</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Figure 3.6: Basis of diagnosis of Prostate cancer cases in Saudi Arabia

Table 3.7: Summary statistics for age and survival time of Prostate cancer cases in Saudi Arabia

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>1st Qu</th>
<th>Median</th>
<th>3rd Qu</th>
<th>Max</th>
<th>Mean</th>
<th>St Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(year)</td>
<td>18</td>
<td>64</td>
<td>71</td>
<td>77</td>
<td>109</td>
<td>70.73</td>
<td>10.48</td>
</tr>
<tr>
<td>Survival time(year)</td>
<td>0</td>
<td>0.27</td>
<td>1.18</td>
<td>2.085</td>
<td>15.26</td>
<td>1.526</td>
<td>1.71</td>
</tr>
</tbody>
</table>
The Table 3.7 shows that 75% of prostate cancer patients are of age more than 64 years. 50% of the patients are from the age interval of 64-77 years. The median age is 71 years while the mean age is 70.73 years with a standard deviation 10.48 years. The earlier incidence recorded at 18 years old and later case reported at 109 years old. The summary statistics for survival time shows that median time of surviving 1.18 years and mean is 1.52 years with a standard deviation 1.52 years. Almost 75% patients die within 2.085 years of diagnosis. The maximum survival time recorded as 15.26 years.

Figure 3.7: Histogram for age of Prostate cancer patients in Saudi Arabia

The figure 3.7 illustrates that the distribution of age of the prostate cancer patients is negatively skewed. There are very few patients who are got infected with prostate cancer before age of 40 years. The histogram shows that ages 65-75 are riskier years for getting infected with prostate cancer.
The histogram of survival time (Figure 3.8) shows that survival time is highly positively skewed. Almost every patient dies within 10 years of diagnosis of the diseases. Approximately, 45% of the prostate cancer patients die within one year of diagnosis and more than 70% die within 2 years of diagnosis.

3.3 Bivariate Analysis

One of our interests was to examine if there is any association between status of the patients with region, grade, extent and Morphology of the prostate cancer respectively in Saudi Arabia. To answer those question, we did run Chi-square tests and reported to the following table (Table 3.8). Chi-square value with P-value<0.05 is considered as the basis of significance of the test.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable</th>
<th>Null Hypothesis</th>
<th>Chi-square Value</th>
<th>Degrees of Freedom (df)</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Grade</td>
<td>Status is independent of Grade</td>
<td>53.103</td>
<td>3</td>
<td>&lt;0.001</td>
<td>Significant</td>
</tr>
<tr>
<td>Status</td>
<td>Extent</td>
<td>Status is independent of Extent</td>
<td>267.95</td>
<td>5</td>
<td>&lt;0.001</td>
<td>Significant</td>
</tr>
<tr>
<td>Status</td>
<td>Morphology</td>
<td>Status is independent of morphology</td>
<td>11.298</td>
<td>4</td>
<td>0.0234</td>
<td>Significant</td>
</tr>
<tr>
<td>Status</td>
<td>Region</td>
<td>Status is independent of region</td>
<td>34.839</td>
<td>12</td>
<td>&lt;0.001</td>
<td>Significant</td>
</tr>
</tbody>
</table>

The Table 3.8, it is concluded that grade, extent, morphology and region have significant association with status of the prostate cancer patients at 5% level of significance since all the Chi-square values are too big with relatively very small p-values (<0.05).
3.4 Conclusion

It is observed that almost 75% of the prostate cancer cases reported from Riyadh, Makkah and Eastern cities. Distant metastasis and localized consist of 89% (approx.) of the prostate cancer cases in Saudi Arabia. More than 56% of the patients are diagnosed as cancer patient while they stay at grade III. The portion of the patients diagnosed at early stage is very low. More than 95% of the patients had Adenocarcinoma, NOS. Approximately 16% of the patients died of prostate cancer. Histology of primary cancel cell is the most widely used technique for diagnosis of the disease in Saudi Arabia. Age more than 60yeras is the most vulnerable time for occurring prostate cancer in males. Almost 75% of the patients die within 2.08 years of the diagnosis. Almost 45% die within 1 year of diagnosis. Grade, region, extent and morphology are significantly related with status (died or alive) of the patients.
Chapter 4
Result: Survival Analysis

4.1 Introduction

This chapter includes summary results of survival analysis of the prostate cancer data from Saudi Arabia. In this chapter we discuss about the Kaplan Meier (KM) curve without adjusting covariates first and then for each of the selected covariates individually. The covariates of interest in the survival analysis is region, grade, extent and morphology. The response variable survival time is in years. Status as dead is our event and alive is considered as censored cases. Log-rank test has been applied to test whether the survival curves for the categories of each of the covariates differ or not. The proportional Hazards (PH) assumptions are checked before fitting Cox Proportional Hazard (Cox-PH) model. We report results obtained from univariate PH-models as well as from full model with corresponding AICs. Model diagnostics has been performed by using Schoenfeld residuals (a method for checking PH assumptions) [5]. Finally, we have presented our suggested model with interpretation of the parameters in terms of hazard ratios.

4.2 Kaplan Meier (KM) Survival curves

The Kaplan–Meier estimator is a non-parametric approach of estimating survival probabilities and is one of the most frequently used methods of survival analysis. KM survival curve plots the estimated survival probabilities with respect to survival time. The nature of the curve is that it is downward slopping. At the beginning of the survival time the height of the survival curve is equal to 1 (as survival probability is 1 at the beginning) and as time goes up it goes down.
Figure 4.1 shows the KM survival curve of the prostate cancer patients with 95% confidence interval without considering any covariates. Up to 5 years of survival time the curve sharply goes down indicating that the number of deaths occurring in this interval is high. In the survival time interval (5 years-10 years) the curve is dropping slowly. There are a few cases who survived until 15 years.

**Figure 4.1: Unadjusted KM survival curve for the prostate cancer cases in Saudi Arabia**

We have constructed the region-wise KM curves for the prostate cancer cases in Saudi Arabia. It is observed that until 3 years (approx.) survival probability in Jazan was the highest, after that until 6/7 years’ survival probability in Baha was the highest, in 7-10 years range the survival probability in Jazan was the highest again, after that Riyadh occupies the highest position. However, this graph is too busy
because of having too many categories of region variable, therefore, it is not presented here.

The Figure 4.2 shows that survival for the Regional: NOS is the highest, it seems to be unjustified because data contain only three cases for this category. Other than this group, survival probability for the localized is higher than any other groups, while distant metastasis group experience the lowest survival.

Figure 4.2: Extent-wise KM curves for the prostate cancer cases in Saudi Arabia
Figure 4.3: Morphology-wise KM curves for the prostate cancer cases in Saudi Arabia

Figure 4.3 depicts that survival for adenocarcinoma group is higher than any other group after 3 years. Before 3-year survival for carcinoma, NOS is the lowest.
Figure 4.4: Grade-wise KM curves for the prostate cancer cases in Saudi Arabia

From the Figure 4.4, it observed that on an average survival of the grade I cases is the highest while for grade IV, it is the lowest. The result is quite natural. During first 1 years it seems that survival for the grade II patients are higher than any other group.

4.3 Log-rank test

Log-rank test is basically equivalent to Chi-square test. It is used to test whether significant difference exists among the survival curves of interest. In this section we present the log-rank test results for testing differences of survival curves among the categories for each of the variables, such as region, grade, extent and morphology. For this test the null hypothesis is: There is no difference among the
survival curves with respect to categories of region or grade or extent or morphology. Log-rank test for Region- wise KM curves for the prostate cancer cases in Saudi Arabia depicts that Chi-square value=16.4 with a p-value 0.0175 (<0.05) which means that the survival curves are statistically different than each other city at 5% level of significance. Log-rank test for Extent- wise KM curves for the prostate cancer cases in Saudi Arabia shows that Chi-square value of 354 with a p-value 0.000 (<0.05) which indicates that the survival curves are statistically different for the extent categories at 5% level of significance. Log-rank test for morphology- wise KM curves for the prostate cancer cases in Saudi Arabia illustrates that Chi-square value of 20.4 with a p-value 0.0004(<0.05) which means that the survival curves are significantly different for the categories of morphology at 5% level of significance. Log-rank test for grade- wise KM curves for the prostate cancer cases in Saudi Arabia shows that Chi-square value of 72.1 with a p-value <0.001(<0.05) which indicates that the survival curves for different grades are statistically different at 5% level of significance.

4.4 Assessing PH assumption using Graphical Approaches

The graphical approach presented here for checking the PH assumption are comparing log-log survival curves. The log-log survival curve is a transformation of an estimated survival curve obtained from taking the natural log of an estimated survival probability twice [5]. PH-assumption refers to the fact that hazard ratio is independent of time meaning that hazard ratio for two individuals is proportional. However, we have used Survival R package to produce the graphs. If the graphs look parallel then we conclude that PH assumption satisfied, otherwise not satisfied.
Figure 4.5: Log-log survival curves by region for the prostate cancer cases in Saudi Arabia

In Figure 4.5, we observe that log-log survival curves for the prostate cancer patients in administrative regions (different categories in of region) are not parallel to each other, some of the curves are crossing each other, that’s why we cannot have a clear idea about whether the variable region satisfies the PH assumption or not.
The Figure 4.6 depicts that the log-log survival curves for the prostate cancer patients in Saudi Arabia for the categories of extent are not parallel to each other. Therefore, it seems that extent does not satisfy the PH assumption.

Figure 4.7: Log-log curves by morphology for the prostate cancer cases in Saudi Arabia
In the Figure 4.7, it is observed that the log-log survival curves for the prostate cancer patients in Saudi Arabia for the categories of morphology are not parallel to each other. Therefore, it is hard to conclude that morphology does satisfy the PH assumption.

Figure 4.8: Log-log curves by grade for the prostate cancer cases in Saudi Arabia

In the Figure 4.8, the grade wise log-log survival curves for the prostate cancer patients in Saudi Arabia are not parallel to each other. Therefore, from the figures (4.5-4.8), it appears that the variables do not satisfy PH assumption. However, the graphical approach is merely an approximation, not exact. Hence, we should check PH-assumption by an analytical approach. Therefore, we will check PH-assumption further by using Schoenfeld residuals.
4.5 Cox-PH Model:

4.5.1 Univariate

In this section we present the results obtained by fitting Cox-PH model considering each of the covariates separately. Table 4.1, 4.2, 4.3 and 4.4 show that each of the covariates such as region, extent, morphology and grade are statistically significant at 5% level of significance. We will carry them forward for multivariate model.

Table 4.1: Result from the Cox-PH model by using covariate Region

<table>
<thead>
<tr>
<th>Region categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baha</td>
<td>-0.2994</td>
<td>0.7412</td>
<td>0.530</td>
<td></td>
</tr>
<tr>
<td>Eastern</td>
<td>-0.2729</td>
<td>0.7611</td>
<td>0.165</td>
<td></td>
</tr>
<tr>
<td>Hail</td>
<td>-0.3313</td>
<td>0.7179</td>
<td>0.451</td>
<td></td>
</tr>
<tr>
<td>Jazan</td>
<td>-1.1743</td>
<td>0.3090</td>
<td>0.007</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Jouf</td>
<td>-0.1737</td>
<td>0.8405</td>
<td>0.692</td>
<td></td>
</tr>
<tr>
<td>Madinah</td>
<td>-0.1737</td>
<td>1.0227</td>
<td>0.935</td>
<td></td>
</tr>
<tr>
<td>Makkah</td>
<td>0.0225</td>
<td>0.7193</td>
<td>0.088</td>
<td></td>
</tr>
<tr>
<td>Najran</td>
<td>-0.3294</td>
<td>0.6850</td>
<td>0.471</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>-1.0868</td>
<td>0.3373</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td>Qassim</td>
<td>-0.1386</td>
<td>0.8705</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>Riyadh</td>
<td>-0.0924</td>
<td>0.9117</td>
<td>0.606</td>
<td></td>
</tr>
<tr>
<td>Tabuk</td>
<td>0.3042</td>
<td>1.3556</td>
<td>0.392</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2: Result from the Cox-PH model by using covariate Extent

<table>
<thead>
<tr>
<th>Extent categories</th>
<th>Coefficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized</td>
<td>-1.827</td>
<td>0.0161</td>
<td>&lt;0.001</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Regional (NOS)</td>
<td>-15.61</td>
<td>1.668 e(+07)</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>Regional: Direct ext.&amp; lymph node</td>
<td>-1.365</td>
<td>0.2555</td>
<td>0.002</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Regional: Direct extension</td>
<td>-1.078</td>
<td>0.0344</td>
<td>&lt;0.001</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Regional: lymph node</td>
<td>-1.154</td>
<td>0.0312</td>
<td>&lt;0.001</td>
<td>Significant at 5%</td>
</tr>
</tbody>
</table>

Table 4.3: Result from the Cox-PH model by using covariate Morphology

<table>
<thead>
<tr>
<th>Morphology categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenocarcinoma (NOS)</td>
<td>0.2447</td>
<td>1.2773</td>
<td>0.626</td>
<td></td>
</tr>
<tr>
<td>Carcinoma (NOS)</td>
<td>1.1991</td>
<td>3.3171</td>
<td>0.030</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Neoplasm, Malignant</td>
<td>1.3464</td>
<td>3.8437</td>
<td>0.228</td>
<td></td>
</tr>
<tr>
<td>Transitional cell Carcinoma NOS</td>
<td>1.2188</td>
<td>3.3830</td>
<td>0.159</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Result from the Cox-PH model by using covariate Grade

<table>
<thead>
<tr>
<th>Grade categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade II</td>
<td>0.3061</td>
<td>1.3582</td>
<td>0.2412</td>
<td></td>
</tr>
<tr>
<td>Grade III</td>
<td>1.1245</td>
<td>3.0786</td>
<td>&lt;0.001</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Grade IV</td>
<td>1.4871</td>
<td>4.4244</td>
<td>&lt;0.001</td>
<td>Significant at 5%</td>
</tr>
</tbody>
</table>

4.5.2 Multivariate

In this part we have fitted Cox-PH model with the covariates found significant in Cox-PH univariate models. The corresponding model results have been reported. We have performed model diagnosis by checking PH assumption for the variables when they are adjusted in the model.
<table>
<thead>
<tr>
<th>Region categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baha</td>
<td>-0.1341</td>
<td>0.8745</td>
<td>0.781</td>
<td></td>
</tr>
<tr>
<td>Eastern</td>
<td>-0.2066</td>
<td>0.8133</td>
<td>0.300</td>
<td></td>
</tr>
<tr>
<td>Hail</td>
<td>-0.8256</td>
<td>0.4380</td>
<td>0.064</td>
<td>Significant at 10%</td>
</tr>
<tr>
<td>Jazan</td>
<td>-1.253</td>
<td>0.2858</td>
<td>0.004</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Jof</td>
<td>-0.6496</td>
<td>0.5223</td>
<td>0.144</td>
<td></td>
</tr>
<tr>
<td>Madinah</td>
<td>-0.1866</td>
<td>0.8297</td>
<td>0.506</td>
<td></td>
</tr>
<tr>
<td>Makkah</td>
<td>-0.3268</td>
<td>0.7212</td>
<td>0.093</td>
<td>Significant at 10%</td>
</tr>
<tr>
<td>Najran</td>
<td>-0.4284</td>
<td>0.6515</td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>-0.7413</td>
<td>0.4765</td>
<td>0.465</td>
<td></td>
</tr>
<tr>
<td>Qassim</td>
<td>-0.4619</td>
<td>0.6301</td>
<td>0.154</td>
<td></td>
</tr>
<tr>
<td>Riyadh</td>
<td>-0.0278</td>
<td>0.9725</td>
<td>0.877</td>
<td></td>
</tr>
<tr>
<td>Tabuk</td>
<td>0.2265</td>
<td>1.25</td>
<td>0.532</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extent categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized</td>
</tr>
<tr>
<td>Regional (NOS)</td>
</tr>
<tr>
<td>Regional: Direct ext.&amp; lymph node</td>
</tr>
<tr>
<td>Regional: Direct extension</td>
</tr>
<tr>
<td>Regional: lymph node</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morphology categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenocarcinoma (NOS)</td>
</tr>
<tr>
<td>Carcinoma (NOS)</td>
</tr>
<tr>
<td>Neoplasms, Malignant</td>
</tr>
<tr>
<td>Transitional cell Carcinoma NOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade II</td>
</tr>
<tr>
<td>Grade III</td>
</tr>
<tr>
<td>Grade IV</td>
</tr>
</tbody>
</table>
Extent categories

Table 4.5 shows that all the covariates have significant impact on survival time except morphology at 5% level of significance. However, before selecting the final model we are checking whether the covariates included in the model satisfy the PH-assumption. Table 4.6 illustrates the correlation between Schoenfeld residuals and ranked survival time, we observe that only the correlation (rho=0.1127, p-value=0.0151 <0.05) for the category of extent named regional: direct extension is statistically significant. It is supported by the Figure 4.9, if the plotted points (upper and lower) are parallel then it satisfies the PH –assumption, otherwise not satisfied. In our plot the points are not parallel. Therefore, we conclude that only extent does not satisfy the PH assumption.

Table 4.6 Model diagnostics by using correlation between Schoenfeld residuals and ranked survival time

<table>
<thead>
<tr>
<th>Region categories</th>
<th>Rho</th>
<th>Chi-square</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baha</td>
<td>-0.0328</td>
<td>0.527</td>
<td>0.469</td>
<td></td>
</tr>
<tr>
<td>Eastern</td>
<td>0.0518</td>
<td>1.26</td>
<td>0.261</td>
<td></td>
</tr>
<tr>
<td>Hail</td>
<td>0.0288</td>
<td>0.409</td>
<td>0.522</td>
<td></td>
</tr>
<tr>
<td>Jazan</td>
<td>0.010</td>
<td>0.047</td>
<td>0.827</td>
<td></td>
</tr>
<tr>
<td>Jouf</td>
<td>0.0719</td>
<td>0.717</td>
<td>0.380</td>
<td></td>
</tr>
<tr>
<td>Madinah</td>
<td>-0.021</td>
<td>0.208</td>
<td>0.648</td>
<td></td>
</tr>
<tr>
<td>Makkah</td>
<td>0.033</td>
<td>0.531</td>
<td>0.466</td>
<td></td>
</tr>
<tr>
<td>Najran</td>
<td>0.0498</td>
<td>1.28</td>
<td>0.257</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>-0.0613</td>
<td>1.74</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td>Qassim</td>
<td>0.025</td>
<td>0.306</td>
<td>0.581</td>
<td></td>
</tr>
<tr>
<td>Riyadh</td>
<td>0.021</td>
<td>0.219</td>
<td>0.640</td>
<td></td>
</tr>
<tr>
<td>Tabuk</td>
<td>0.004</td>
<td>0.0116</td>
<td>0.9143</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.6 Model diagnostics by using correlation between Schoenfeld residuals and ranked survival time (continued)

<table>
<thead>
<tr>
<th>Extent categories</th>
<th>Rho</th>
<th>Chi-square</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized</td>
<td>0.0544</td>
<td>1.43</td>
<td>0.2323</td>
<td></td>
</tr>
<tr>
<td>Regional (NOS)</td>
<td>-0.2314</td>
<td>0.000</td>
<td>0.990</td>
<td></td>
</tr>
<tr>
<td>Regional: Direct ext.&amp; lymph node</td>
<td>-0.0164</td>
<td>0.124</td>
<td>0.7244</td>
<td></td>
</tr>
<tr>
<td>Regional: Direct extension</td>
<td>0.1127</td>
<td>5.90</td>
<td>0.0151</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Regional: lymph node</td>
<td>0.0754</td>
<td>2.77</td>
<td>0.096</td>
<td></td>
</tr>
</tbody>
</table>

**Morphology categories**

| Adenocarcinoma (NOS)                  | -0.0460 | 0.994     | 0.3187  |                 |
| Carcinoma (NOS)                       | -0.0853 | 3.52      | 0.0607  |                 |
| Neoplasm, Malignant                   | -0.0567 | 1.69      | 0.1931  |                 |
| Transitional cell Carcinoma NOS       | -0.006  | 0.019     | 0.889   |                 |

**Grade categories**

| Grade II                              | 0.038  | 0.725     | 0.394   |                 |
| Grade III                             | 0.055  | 1.53      | 0.215   |                 |
| Grate IV                              | 0.061  | 1.89      | 0.168   |                 |
When any of the covariates does not satisfy PH assumption, one of the alternatives is to fit a stratified model. Therefore, we fit a stratified Cox model considering extent as the stratified variable and checked again the PH assumption. We found each of the covariates in the stratified model satisfy PH-assumption.
### Region categories

<table>
<thead>
<tr>
<th>Region categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baha</td>
<td>-0.1475</td>
<td>0.862</td>
<td>0.7601</td>
<td></td>
</tr>
<tr>
<td>Eastern</td>
<td>-0.2679</td>
<td>0.764</td>
<td>0.1827</td>
<td></td>
</tr>
<tr>
<td>Hail</td>
<td>-0.8310</td>
<td>0.435</td>
<td>0.0692</td>
<td>Significant at 10%</td>
</tr>
<tr>
<td>Jazan</td>
<td>-1.236</td>
<td>0.290</td>
<td>0.0051</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Jouf</td>
<td>-0.6602</td>
<td>0.516</td>
<td>0.1448</td>
<td></td>
</tr>
<tr>
<td>Madinah</td>
<td>-0.2011</td>
<td>0.817</td>
<td>0.4754</td>
<td></td>
</tr>
<tr>
<td>Makkah</td>
<td>-0.3334</td>
<td>0.716</td>
<td>0.0879</td>
<td>Significant at 10%</td>
</tr>
<tr>
<td>Najran</td>
<td>-0.4818</td>
<td>0.617</td>
<td>0.3768</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>-0.7123</td>
<td>0.490</td>
<td>0.4829</td>
<td></td>
</tr>
<tr>
<td>Qassim</td>
<td>-0.4618</td>
<td>0.630</td>
<td>0.1554</td>
<td></td>
</tr>
<tr>
<td>Riyadh</td>
<td>-0.0520</td>
<td>0.949</td>
<td>0.7740</td>
<td></td>
</tr>
<tr>
<td>Tabuk</td>
<td>0.2319</td>
<td>1.261</td>
<td>0.5244</td>
<td></td>
</tr>
</tbody>
</table>

### Morphology categories

<table>
<thead>
<tr>
<th>Morphology categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenocarcinoma (NOS)</td>
<td>0.0651</td>
<td>1.067</td>
<td>0.8984</td>
<td></td>
</tr>
<tr>
<td>Carcinoma (NOS)</td>
<td>0.5335</td>
<td>1.704</td>
<td>0.3460</td>
<td></td>
</tr>
<tr>
<td>Neoplasm, Malignant</td>
<td>0.8255</td>
<td>2.283</td>
<td>0.4805</td>
<td></td>
</tr>
<tr>
<td>Transitional cell Carcinoma NOS</td>
<td>0.1327</td>
<td>1.141</td>
<td>0.8815</td>
<td></td>
</tr>
</tbody>
</table>

### Grade categories

<table>
<thead>
<tr>
<th>Grade categories</th>
<th>Co-efficient</th>
<th>Hazard ratio</th>
<th>P-value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade II</td>
<td>0.3836</td>
<td>0.383</td>
<td>0.1532</td>
<td></td>
</tr>
<tr>
<td>Grade III</td>
<td>0.7600</td>
<td>0.760</td>
<td>0.0036</td>
<td>Significant at 5%</td>
</tr>
<tr>
<td>Grate IV</td>
<td>1.0395</td>
<td>1.039</td>
<td>0.0113</td>
<td>Significant at 5%</td>
</tr>
</tbody>
</table>

### 4.6 Model comparison and selection
We have fitted six models and comparing AIC values for model selection purpose. We have chosen the model that has the lowest AIC value. From the table 4.8, we see that Model-6 has the lowest AIC value, therefore, we conclude that stratified model (model-6) is our selected model and we interpret the parameters of that model accordingly.

Table 4.8: Table of AIC values for the competing Cox-PH models

<table>
<thead>
<tr>
<th>Model</th>
<th>Response</th>
<th>Covariates</th>
<th>AIC value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model1</td>
<td>Survival time and Status</td>
<td>Region</td>
<td>6165.417</td>
<td></td>
</tr>
<tr>
<td>Model2</td>
<td>Survival time and Status</td>
<td>Extent</td>
<td>5857.603</td>
<td></td>
</tr>
<tr>
<td>Model3</td>
<td>Survival time and Status</td>
<td>Morphology</td>
<td>6152.99</td>
<td></td>
</tr>
<tr>
<td>Model4</td>
<td>Survival time and Status</td>
<td>Grade</td>
<td>6089.613</td>
<td>Highest</td>
</tr>
<tr>
<td>Model5</td>
<td>Survival time and Status</td>
<td>Region, Extent, Morphology, Grade</td>
<td>5847.36</td>
<td></td>
</tr>
<tr>
<td>Model6</td>
<td>Survival time and Status</td>
<td>Region, Morphology, Grade (Stratified by Extent)</td>
<td>5842.297</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

4.7 Interpretation of the coefficients from the selected model

From the Table 4.7, we see that city Asir is the reference for region, grade I is reference for grade and acinar cell carcinoma is the reference for morphology. The variable morphology is not statistically significant at 5% level, which means that morphology does not have any significant impact on explaining hazard of surviving of the prostate cancer in Saudi Arabia. The other covariates such as region and grade are statistically significant. The column Hazard Ratio represent the hazard ratios (HR) with respect to the reference category of the respective variables. The HR for city Jazan is 0.29031 with a p-value 0.00515 is significant. It means that hazard of surviving for prostate cancer patients in Jazan is [(1-0.29031) x100 %] 71% less in
comparing to the prostate cancer patients in Asir. The other cities in Saudi Arabia are seem be similar with respect hazard of surviving. The HR for grade 2.13831 with a p-value 0.00364 indicates that it is significant and the hazard of survival of the patients diagnosed at grade III is 2.13831 times higher than the patients diagnosed at grade I. The hazard ratio for grade IV is 2.82801 with a p-value 0.01138 indicates that the hazard of survival of the patients diagnosed at grade IV is 2.82801 times higher than the patients diagnosed at grade I. The hazards are seeming to be similar if the patients diagnosed at grade I and grade II.

4.8 Conclusion

City Jazan is significantly different than other cities in Saudi Arabia with respect to hazard of survival. If patients can be diagnosed at grade I and grade II, the probability of surviving as well as survival time can be significantly improved.
Chapter 5

Conclusion

Saudi Arabia is a country of 2,149,700 square kilometers that occupies four-fifths of Arabian Peninsula. The whole country is divided into 13 administrative units. Those are: Jouf, Northern, Tabuk, Hail, Qassim, Madinah, Makkah, Baha, Riyadh, Eastern, Asir, Najran and Jazan. In 2012, the Saudi national population was 19.84 million, among them 50.2% were male. Among the male population 1.15 million were older than 50 years. This population group is the most vulnerable to prostate cancer. Saudi cancer registry collects and monitors the cancer data from the whole country through offices and hospital outlets located in the above listed 13 administrative regions. Cancer is a mandatory notifiable disease in Saudi Arabia, once the cancer registry come across with any sort of cancer patients they do follow up until death of the patients [6].

The current research study is conducted with a view to (i) examining the distribution of prostate cancer cases with respect to the characteristics such as grade, extent, region, morphology, status (case dead or alive at the time of last contact), basis of diagnosis and age (ii) examining the relationship between status (case dead or alive at the time of last contact) with each of the variables of interests namely region, extent, morphology and grade of prostate cancer respectively. The main focus of the research is to fit a survival model that best describes survival time of the prostate cancer patients in Saudi Arabia in relation to the appropriate predictors in the data set.

We have collected a customized version of secondary data on 4501 prostate cancer patients from Saudi Cancer Registry via personal contacts which includes
follow up cases from 1994 to 1st quarter of 2016. In accordance with our objectives we have analyzed data on 2795 complete cases. Almost 38% percent of the data were either misclassified, miscoded or missing. Exploratory analysis such as frequency tables, descriptive statistics and graphical analysis and bivariate association have been performed. Kaplan Meier (KM) survival curves, and Cox PH model have been fitted as a part of survival analysis of the prostate cancer. Finally based on lowest AIC values, stratified Cox-PH model with covariates region, grade and morphology is selected as the best among the competing models.

The region wise distribution of the patients shows that Riyadh shares 29.48%, Makkah shares 23.04% and Eastern shares 22.33% of the prostate cancer cases while the least share happens with northern with a percentage of 0.43%. In the survival analysis part, the administrative unit Asir is our reference for the region which shares 9% of the total prostate cancer cases. Extent wise distribution shows that almost 57% cases are localized followed by 32% in distant metastasis. Among the patients 56.42% are diagnosed at advance stages (grade III) levels followed by 33.38% in grade II levels. Only 8% patients are diagnosed at primary level. Approximately 96% prostate cancer cases are Adenocarcinoma (not otherwise specified) while 0.21% patients experience malignant type of carcinoma. Death is the event in survival analysis which shares 16.4 % of the total prostate cancer cases. Basis of diagnosis distribution shows that almost 99% of the patients are diagnosed by histology of primary cells. We observed that 71 years is the median age of prostate cancer and median survival time is 1.18 years which is quite alarming in case of prostate cancer in Saudi Arabia. About 75% patients die within 2.085 years of diagnosis. Almost 100% patients die within 10 years of diagnosis of the diseases. The in depth analysis shows that 45% of the prostate cancer patients die within one year of diagnosis and more than 70% die within 2 years of diagnosis. The maximum survival
time is recorded as 15.26 years. Bi-variate analysis depicts that grade, region, extent and morphology are significantly associated with the status (death or alive) of the patients during the last contact.

Unadjusted (without considering covariates) KM survival curve of the prostate cancer patients shows that until 5 years of survival time the curve drops sharply indicating that the number of deaths cases in this interval is enormously high. In the survival time interval (5 years-10 years) the curve drops at a slower pace. It is observed that until 3 years (approx.) survival probability in Jazan was the highest, after that until 6/7 years’ survival probability in Baha was the highest. Survival probability for the localized prostate cancer cases is the highest while for distant metastasis is the lowest. The survival for morphology type adenocarcinoma (NOS) is the highest while for carcinoma (NOS) is the lowest. Grade IV patient’s survival is the lowest which is quite natural while for grade I it is the highest. Thus it indicates that early detection can improve the survival. The log-rank test depicts that survival for the respective categories of the region, extent, grade, and morphology are significantly different than each other. PH assumption checking by graphical approach does not provide any clear indication while Schoenfeld residuals methods shows only the variable extent does not satisfy PH assumption. That is why Stratified Cox PH model with respect to extent is selected as final model.

The final model output depicts that morphology does not have any significant impact on explaining hazard of surviving of the prostate cancer in Saudi Arabia. The other covariates such as region and grade are statistically significant. Hazard of death for prostate cancer patients in Jazan is 71% less in comparing to the prostate cancer patients in Asir. The other cities in Saudi Arabia seem be similar with respect to hazard of death. The hazard of death of the patients diagnosed at grade III is 2.13831
times higher than the patients diagnosed at grade I. The hazard of death of the patients diagnosed at grade IV is 2.82801 times higher than the patients diagnosed at grade I. The hazards seem to be similar if the patients diagnosed at grade I and II.

This study result might be unveiling an insight to the researchers, policy makers for better understanding of prostate cancer cases in Saudi Arabia. Since the age distribution of prostate cancer patients shows that 75% of the patients are in the age group 64 and older male population, mandatory screening should be imposed to this population. Since the study shows that hazard of death for grade I and grade II are not significantly different, early detection may enhance the survival time of the patients. In addition, since data contain 38% missing or faulty classification, attention should be given to data quality improvement.

This research has several shortfalls. One of the limitations of the study is that it did not include too many relevant variables that may have significant impact on survival of prostate cancer patients. Missing data is another problem that causes potential loss of information. Stratified Cox-PH model is considered as the best model in this study. One of the pitfalls of stratified model is that it does not provide HR for stratification variable. Thus further scope of work would be to include more covariates, handle missing values by imputation or other viable methods and perhaps by considering parametric survival modeling approach.
Bibliography


Appendix

# Setting working directory and reading original csv datafile ( name:p_cancer_ksa.csv )

rm(list=ls())
getwd()
setwd("C:/Users/al_at/Desktop/atheer thesiss")

rdata <- read.csv(file="C:/Users/al_at/Desktop/atheer thesiss/P_cancer_ksa.csv", header=TRUE, sep="",)

dim(rdata)
colnames(rdata)
head(rdata)
View(rdata)

# Making my working data file: we need the listed variables (dataw)

myvars=c("ID","DOB","Age","Region","DOD","Morphology","Behavior","Grade","Extent","Laterality","B_diag","DoLC","Status","CoD")

dataw=rdata[myvars]
colnames(dataw)
View(dataw)

# To convert Date of Diagnosis (DOD) and Date of Last Contact (DoLC) as date format for calculating survival time

d1=as.Date(dataw[,5],"%m/%d/%Y") # To convert Date of Diagnosis (DOD) as date format

d2=as.Date(dataw[,12],"%m/%d/%Y")# To convert Date of Last Contact (DoLC) as date format

# for creating status (Status1) variable for survival analysis
dataw$Status1[dataw$Status=="Alive"] <- 0
dataw$Status1[dataw$Status=="Dead"] <- 1
dataw$Status1[dataw$Status=="Unknown"] <- NA

# Survival time in year calculation#

dataw$survtime <- (d2 - d1) # survival time in days subtracting DOD from DoLC
colnames(dataw)
st=as.numeric(dataw$survtime)# defining survival time as numeric data
dataw$st_yr=round(st/365,2)# survival time in years with two decimal places
dataw$st_yr
dataw$st_yr[dataw$st_yr==6.3] <- 0
colnames(dataw)

summary(dataw$st_yr)

View(dataw)

# df=data.frame(dataw, st_yr)

# defining the varaibles as factors factors

Region=as.factor(dataw$Region)
Grade=as.factor(dataw$Grade)
Morphology=as.factor(dataw$Morphology)
Behavior=as.factor(dataw$Behavior)
Extent=as.factor(dataw$Extent)
Laterality=as.factor(dataw$Laterality)
Status=as.factor(dataw$Status)
Status1=as.factor(dataw$Status1)
B_diag = as.factor(dataw$B_diag)
CoD = as.factor(dataw$CoD)

# frequency tables for checking ambiguous data to make data clean: These tables based on original data

tb1 = table(dataw$Region)
a = cbind(tb1, round(prop.table(tb1) * 100, 2))
colnames(a) <- c('Count', 'Percentage')
a

tb2 = table(dataw$Grade)
b = cbind(tb2, round(prop.table(tb2) * 100, 2))
colnames(b) <- c('Count', 'Percentage')
b

tb3 = table(dataw$Morphology)
c = cbind(tb3, round(prop.table(tb3) * 100, 2))
colnames(c) <- c('Count', 'Percentage')
c

tb4 = table(dataw$Behavior)
d = cbind(tb4, round(prop.table(tb4) * 100, 2))
colnames(d) <- c('Count', 'Percentage')
d

tb5 = table(dataw$Extent)
e = cbind(tb5, round(prop.table(tb5) * 100, 2))
colnames(e) <- c('Count', 'Percentage')
e
tb6 = table(dataw$Laterality)
f = cbind(tb6, round(prop.table(tb6) * 100, 2))
colnames(f) <- c('Count', 'Percentage')
f

tb7 = table(dataw$Status)
g = cbind(tb7, round(prop.table(tb7) * 100, 2))
colnames(g) <- c('Count', 'Percentage')
g

tb8 = table(dataw$B_diag)
h = cbind(tb8, round(prop.table(tb8) * 100, 2))
colnames(h) <- c('Count', 'Percentage')
h

tb9 = table(dataw$CoD)
i = cbind(tb9, round(prop.table(tb9) * 100, 2))
colnames(i) <- c('Count', 'Percentage')
i

tb10 = table(dataw$Status)
j = cbind(tb10, round(prop.table(tb10) * 100, 2))
colnames(j) <- c('Count', 'Percentage')
j

tb11 = table(dataw$Status1)
k = cbind(tb11, round(prop.table(tb11) * 100, 2))
colnames(k) <- c('Count', 'Percentage')
k
# data cleaning: Deleting missing data, unknown categories for the selected variables and rare categories.

# New pcdata file is the data file of the cleaned data

pcdata<-droplevels(dataw[!dataw$Region %in% c('International', 'Unknown', 'NaN') &
                       !dataw$Grade %in% c('Unknown', 'NaN') &
                       !dataw$Morphology %in% c('Acinar cell cystadenocarcinoma',
                                                  'Adenosquamous carcinoma',
                                                  'Carcinoma, anaplastic, NOS',
                                                  'Carcinoma, diffuse type',
                                                  'Carcinoma, undifferentiated, NOS',
                                                  'Cribriform carcinoma, NOS',
                                                  'Embryonal rhabdomyosarcoma, NOS',
                                                  'Granular cell carcinoma',
                                                  'Infiltrating duct carcinoma, NOS',
                                                  'Large cell carcinoma, NOS',
                                                  'Leiomyosarcoma, NOS',
                                                  'Malignant tumor, giant cell type',
                                                  'Malignant tumor, spindle cell type',
                                                  'Mucin-producing adenocarcinoma',
                                                  'Mucinous adenocarcinoma',
                                                  'Neuroendocrine carcinoma, NOS',
                                                  'Papillary adenocarcinoma, NOS',
                                                  'Papillary carcinoma, NOS',
                                                  'Papillary transitional cell carcinoma',
                                                  'Signet ring cell carcinoma',
                                                  'Small cell carcinoma, NOS')]}
& dataw$Morphology!=\'Small cell sarcoma\'
& dataw$Morphology!=\'Squamous cell carcinoma, NOS\'
& dataw$Morphology!=\'Tubular adenocarcinoma\'
& dataw$Morphology!=\'Tumor cells, malignant\'
& dataw$Morphology!=\'NaN\'
& dataw$Extent!=\'Unknown\'
& dataw$Extent!=\'NaN\'
& dataw$Status1!=\'NA\'
& dataw$Status1!=\'NaN\'
& dataw$CoD!=\'Other\'
& dataw$CoD!=\'Unknown\'
& dataw$CoD!=\'NaN\'
& dataw$Status!=\'NaN\'
& dataw$Status!=\'Unknown\',\})

dim(pcdata)
View(pcdata)
attach(pcdata)

write.csv(pcdata, file = "pcdata.csv")

#Univariate analysis based on clean data file: pcdata

ttb1=table(pcdata$Region)
a1=cbind(ttb1,round(prop.table(ttb1)*100,2))
colnames(a1) <- c('Count','Percentage')
a1

ttb2=table(pcdata$Grade)
b1=cbind(ttb2,round(prop.table(ttb2)*100,2))
colnames(b1) <- c('Count','Percentage')
b1

ttb3=table(pcdata$Morphology)
c1=cbind(ttb3,round(prop.table(ttb3)*100,2))
colnames(c1) <- c('Count','Percentage')
c1

ttb4=table(pcdata$Behavior)
d1=cbind(ttb4,round(prop.table(ttb4)*100,2))
colnames(d1) <- c('Count','Percentage')
d1

ttb5=table(pcdata$Extent)
e=cbind(ttb5,round(prop.table(ttb5)*100,2))
colnames(e) <- c('Count','Percentage')
e

ttb6=table(pcdata$Laterality)
f1=cbind(ttb6,round(prop.table(ttb6)*100,2))
colnames(f1) <- c('Count','Percentage')
f1

ttb7=table(pcdata$Status)
g1=cbind(ttb7,round(prop.table(ttb7)*100,2))
colnames(g1) <- c('Count','Percentage')
g1

ttb8=table(pcdata$B_diag)
h1=cbind(ttb8,round(prop.table(ttb8)*100,2))
colnames(h1) <- c('Count','Percentage')

h1

ttb9=table(pcdata$CoD)
i1=cbind(ttb9,round(prop.table(ttb9)*100,2))
colnames(i1) <- c('Count','Percentage')
i1

ttb10=table(pcdata$Status)
j1=cbind(ttb10,round(prop.table(ttb10)*100,2))
colnames(j1) <- c('Count','Percentage')
j1

ttb11=table(pcdata$Status1)
k1=cbind(ttb11,round(prop.table(ttb11)*100,2))
colnames(k1) <- c('Count','Percentage')
k1

# Descriptive statistics for Age
summary(pcdata$Age)
sd(pcdata$Age)

length(st_yr)
summary(pcdata$st_yr)
sd(pcdata$st_yr)

#Graphical Analysis based on clean data file: pcdata

hist(pcdata$Age,freq=FALSE, col="purple", main="Histogram of Age of prostate cancer patients", xlab="Age",ylab="Density")
lines(density(pcdata$Age), na.rm = TRUE)

hist(pcdata$st_yr, freq = FALSE, col = "green", main = "Histogram of survival time of prostate cancer patients after diagnosis", xlab = "Age", ylab = "Density")

lines(density(pcdata$st_yr), na.rm = TRUE)

# Bar diagram for Regionwise distribution
count1 = table(pcdata$Region)
count1
percentage1 = count1 * 100 / 2795


# Bar diagram for Extent
count2 = table(pcdata$Extent)
count2
percentage2 = count2 * 100 / 2795

barplot(percentage2, main = "Extent wise distribution of Prostate cancer in Saudi Arabia", xlab = "Extent category", ylab = "percentage", col = "purple", las = 2)

# Bar diagram for Morphology
count3 = table(pcdata$Morphology)
count3
percentage3 = count2 * 100 / 2795

barplot(percentage3, main = "Morphology wise distribution of Prostate cancer in Saudi Arabia", xlab = "Extent category", ylab = "percentage", col = "purple", las = 2)
# Pie chart Grade
mytable2 <- table(pcdata$Grade)
mytable2
gr <- paste(names(mytable2))
gr
pie(mytable2, labels = gr,
    main="Gradewise distribution of prostate cancer for the study data in Saudi Arabia")

### Bivariate table###
tgrade=table(pcdata$Grade,pcdata$Status)
tgrade
chisq.test(tgrade, correct=T)

tmorp=table(pcdata$Morphology,pcdata$Status)
tmorp
chisq.test(tmorp, correct=T)

textent=table(pcdata$Extent,pcdata$Status)
textent
chisq.test(textent, correct=T)

tregion=table(pcdata$Region,pcdata$Status)
tregion
chisq.test(tregion, correct=T)

#Survival Analysis#
library (survival)
y=Surv(pcdata$st_yr,pcdata$Status1==1)
# Kaplan-Mier(KM) Survival curve for over all data: Unadjusted
kmfit1=survfit(y~1)
summary(kmfit1)
plot(kmfit1,main="Unadjusted Survival Curve with 95% Confidence Interval",xlab="Survival times in years",ylab="Survival probability",col="red")
survdiff()

# Kaplan-Mier(KM) Survival curve for over cities
kmfit2=survfit(y~pcdata$Region)
summary(kmfit2)
plot(kmfit2, main="Region wise Survival Curve",xlab="Survival times in years",ylab="Survival probability",col="red")
# Log rank to compare region wise survival curves
survdiff(y~pcdata$Region)

# Kaplan-Mier(KM) Survival curve for Extent
kmfit3=survfit(y~pcdata$Extent)
summary(kmfit3)
plot(kmfit3,lty=c("solid","dashed","dotted","dotdash","longdash","twodash"),
main="Extent wise Survival Curve",xlab="Survival times in years",ylab="Survival probability",col=c("black","green","blue","pink","grey","purple"))
legend("topright",c("D_Metas", "Localised","R: NOS","R: De_Ly_No","R: D_Ext","R: L_No"),lty=c("solid","dashed","dotted","dotdash","longdash","twodash"),col=c("black","green","blue","pink","grey","purple"))

# Log rank to compare Extent wise survival curves
survdiff(y~pcdata$Extent)

# Kaplan-Mier(KM) Survival curve for Morphology
kmfit4 = survfit(y~pcdata$Morphology)
summary(kmfit4)
plot(kmfit4, lty=c("solid","dashed","dotted","dotdash","longdash"),
main="Morphology wise Survival Curve", xlab="Survival times in years",
ylab="Survival probability", col=c("black","green","blue","grey","purple"))
legend("topright", c("Acinar cell carcinoma", "Adenocarcinoma, NOS", "Carcinoma, NOS", "Neoplasm, malignant", "Transcell carcinoma, NOS"), lty=c("solid","dashed","dotted","dotdash","longdash"), col=c("black","green","blue","grey","purple"))

# Log rank to compare Morphology wise survival curves
survdiff(y~pcdata$Morphology)

# Kaplan-Mier(KM) Survival curve for Grade
kmfit5 = survfit(y~pcdata$Grade)
summary(kmfit5)
plot(kmfit5, lty=c("solid","dashed","dotted","dotdash"),
main="Grade wise Survival Curve", xlab="Survival times in years",
ylab="Survival probability", col=c("black","green","blue","purple"))
legend("topright", c("Grade I", "Grade II", "Grade III", "Grade IV"), lty=c("solid","dashed","dotted","dotdash"), col=c("black","green","blue","purple"))

# Log rank to compare Grade wise survival curves
survdiff(y~pcdata$Grade)

############ Assessing PH assumption using Graphical Approaches

plot(kmfit2, fun="cloglog", main ="Log-log curves by Region", xlab="Time in years using logarithmic scale", ylab="log-log survival")
plot(kmfit3, fun="cloglog", main ="Log-log curves by Extent", xlab="Time in years using logarithmic scale", ylab="log-log survival")
plot(kmfit4, fun="cloglog", main ="Log-log curves by Morphology", xlab="Time in years using logarithmic scale", ylab="log-log survival")
plot(kmfit5,fun="cloglog",main ="Log-log curves by grade",xlab="Time in yeras using logarithmic scale",ylab="log-log survival")

# FITTING cOX-pH Model: Univariate

m1=coxph(y~pcdata$Region)
summary(m1)
extractAIC(m1)

m2=coxph(y~pcdata$Extent)
summary(m2)
extractAIC(m2)

m3=coxph(y~pcdata$Morphology)
summary(m3)
extractAIC(m3)

m4=coxph(y~pcdata$Grade)
summary(m4)
extractAIC(m4)

##### Full Model

m5=coxph(y~pcdata$Region+pcdata$Extent+pcdata$Morphology+pcdata$Grade)
summary(m5)
extractAIC(m5)

##### Model diagnostics by using correlation between Schoenfeld residuals and ranked survival time
m5=coxph(y~pcdata$Region+pcdata$Extent+pcdata$Morphology+pcdata$Grade)
m5
cox.zph(m5, transform=rank)
plot(cox.zph(m5, transform=rank), se=F, var='pcdata$ExtentRegional: Direct Ext')

#### Stratified COX Model as Extent does not satisfy Proportional Hazard (PH) Assumption
m6=coxph(y~pcdata$Region+pcdata$Grade +pcdata$Morphology+strata(pcdata$Extent))
summary(m6)
extraAIC(m6)

m7=coxph(y~pcdata$Region+pcdata$Grade+strata(pcdata$Extent))
summary(m7)
extraAIC(m7)
cox.zph(m7, transform=rank)

#####
m7=coxph(y~pcdata$Region+pcdata$Extent+pcdata$Grade)
summary(m7)
extraAIC(m7)