Survival Analysis Approaches for Prostate Cancer

By

Eman Alhasawi

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science (MSc) in Computational Sciences

The Faculty of Graduate Studies

Laurentian University

Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	Survival Analysis Approaches for Prostate Cancer
Name of Candidate Nom du candidat	Alhasawi, Eman
Degree Diplôme	Master of Science
Department/Program Département/Programme	Computational Sciences
Date of Defence Date de la soutenance	April 15, 2015

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur de thèse)

Dr. Mazen Saleh
(Committee member/Membre du comité)

Dr. Hafida Boudjellaba
(Committee member/Membre du comité)

Dr. Chakresh Jain
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Dr. David Lesbarrères
M. David Lesbarrères
Acting Dean, Faculty of Graduate Studies
Doyen intérimaire, Faculté des études supérieures

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

Abstract

Survival time has become an essential outcome of clinical trial, which began to emerge among the latter half of the 20th century. A present study was carried out on the survival analysis for patients with prostate cancer. The data was obtained from Memorial Sloan Kettering where each sample was collected from the recipients of the treatment of radical prostatectomy. The Kaplan-Meier method was used to obtain and estimate the survival function and median time among the primary and metastatic tumor of prostate cancer. Results showed that the metastatic tumor has a poor survival rate compared to the primary tumor, which give us a hint that primary tumor has a higher probability of surviving. The log-rank test was used to test the differences in the survival curves. The results showed that the difference in survival rate between the patients of the two groups of tumor was significant with a p-value of 4.44e-15. The second approach was based on the efficiency of cox proportional hazards model and parametric model. Some criteria of residuals were used for judging the goodness of fit among the candidate models. The cox proportional hazard (PH) model provided an effective covariate on the hazard function. As a result of cox PH model, the influence of standard clinical prognostic factors is based on the hazard rate of prostate cancer patients. These prognostic factors are: prostate specific antigen (PSA) level at diagnosis, tumor size, Secondary Gleason grade, and Gleason score which is helpful to determine the treatment. The Gleason score [HR 4.835, 95% CI 2.7847- 8.3937, p=2.20E-08] has the most significant progression-associated prognosticators and reveal to be an effective criteria leading to death in prostate cancer. The Accelerated Failure Time (AFT) was applied to the data with four distortions. AFT with Weibull distortions was chosen to be the best model for our data by testing the AIC.

Acknowledgements

I would like to thank my God, who got me this far; who blessed me with the right people to help me during the different stages of my study.

It gives me great pleasure to express my deepest respect and sincere thanks to my advisor Professor Kalpdrum Passi for his encouragement, valuable suggestions, discussion and guidance throughout my graduate studies. He continually and convincingly conveyed a spirit of adventure in regard to research. He was patient with my writing style and taught me how to explain my thoughts and present them clearly in writing. Without his guidance and persistent help this thesis would not have been possible.

I am deeply indebted to my committee Dr. Hafida Boudjellaba who always found time to provide constructive feedback to my thoughts. She provided me with technical support and become more of a mentor friend, than a professor. She answered my detailed oriented questions and helped me progress. I am grateful for her tremendous help at the initial stages of developing my thesis project.

I would like to express my regards and thanks to Dr. Mazen Saleh, a member of my supervisory committee for reading my thesis and providing valuable feedback on my thesis.

I would like to send my appreciation and respects to Dr. Peter Adamic for his help and suggestions.

It is with immense gratitude to thank my family for their love, helps, and supports, especially my parents Ahmed Ali and Anisah Ahmed for being supportive and helping me get all the annoying little things done, my wonderful brother, Ali for supporting me in my pursuit of this degree. I would like to express my gratefulness towards my sisters Azhar and Asia who were always there for me and cheering me on all situations.

I am also grateful to all my friends here in Sudbury and my friends in Saudi Arabia for their encouragement and to help change my career path. I couldn’t have achieved this without their help.

I wish to express my deepest appreciation to the King Abdullah, for giving Saudi women the scholarship to complete studying. I recognize that thesis would not have been possible without the financial assistance of Saudi Cultural Bureau in Canada and the Saudi Ministry of Higher Education.

This thesis is dedicated to
My family and friends,
Without whose support and inspiration
I would never have had the courage to follow my dreams.
I love you and I miss you.

Table of Contents

Contents

Abstract ii

Acknowledgements iii

Table of Contents vi

List of figure viii

List of Table ix

List of APPENDIX x

Introduction ..... **Error! Bookmark not defined.**

1.1 Prostate Cancer ..... **Error! Bookmark not defined.**

1.1.1 Tumors ..... **Error! Bookmark not defined.**

1.1.2 Prognostic Factors in Prostate Cancer ..... **Error! Bookmark not defined.**

1.1.3 Treatment ..... **Error! Bookmark not defined.**

1.2 Survival Analysis ..... **Error! Bookmark not defined.**

1.2.1 Censored data ..... **Error! Bookmark not defined.**

1.2.2 Functions related to survival analysis ..... **Error! Bookmark not defined.**

1.3 Objectives ..... **Error! Bookmark not defined.**

Chapter 2 ..... **Error! Bookmark not defined.**

Literature Review ..... **Error! Bookmark not defined.**

2.1 Survival Analysis Study ..... **Error! Bookmark not defined.**

Vinh-Hung, V. et al. (2002), Post-surgery radiation in early breast cancer: survival analysis of registry data ..... **Error! Bookmark not defined.**

Ray, M.E. et al. (2009), Potential surrogate endpoints for prostate cancer survival: analysis of a phase III randomized trial ..... **Error! Bookmark not defined.**

Chan, Y.M. (2013), Statistical Analysis and Modeling of Prostate Cancer **Error! Bookmark not defined.**

Pulte, D. (2012), Changes in survival by ethnicity of patients with cancer between 1992–1996 and 2002–2006: is the discrepancy decreasing? ... **Error! Bookmark not defined.**

Chapter 3 ..... **Error! Bookmark not defined.**

Materials and Methodology ..... **Error! Bookmark not defined.**

3.1 The Data source ..... **Error! Bookmark not defined.**
3.2 Methodology ..... **Error! Bookmark not defined.**
3.2.1 Non-parametric Methods ..... **Error! Bookmark not defined.**
Kaplan Meier Estimates (K-M): ..... **Error! Bookmark not defined.**
Log Rank ..... **Error! Bookmark not defined.**
3.2.2 Semi-parametric Methods ..... **Error! Bookmark not defined.**
Cox proportional hazard: ..... **Error! Bookmark not defined.**
The Adequacy of a model: ..... **Error! Bookmark not defined.**
Testing the proportional hazards assumption ..... **Error! Bookmark not defined.**
3.2.3 Parametric Methods ..... **Error! Bookmark not defined.**
Accelerated Failure Time Model (AFT): ..... **Error! Bookmark not defined.**
Chapter 4 ..... **Error! Bookmark not defined.**
Results and Discussion ..... **Error! Bookmark not defined.**
4.1 Kaplan-Meier (K-M) Estimation ..... **Error! Bookmark not defined.**
4.2 Log-Rank Survival Estimates ..... **Error! Bookmark not defined.**
4.3 Cox Fit Model ..... **Error! Bookmark not defined.**
4.3.1 Testing the proportional hazards assumption using Schoenfeld's residuals .. **Error! Bookmark not defined.**
4.3.2 Evaluating overall model fitting ..... **Error! Bookmark not defined.**
4.3.3 Functional Form of Predictors ..... **Error! Bookmark not defined.**
4.3.4 Checking for Outliers ..... **Error! Bookmark not defined.**
4.4 Output of Accelerated Failure Time (AFT) ..... **Error! Bookmark not defined.**
4.5 Discussion ..... **Error! Bookmark not defined.**
Chapter 5 ..... **Error! Bookmark not defined.**
Conclusion ..... **Error! Bookmark not defined.**
Future work ..... **Error! Bookmark not defined.**
References ..... **Error! Bookmark not defined.**
Appendix ..... **Error! Bookmark not defined.**

List of figure

Figure **page**

1.1 Illustration of left, right and interval censoring (Aaserud,2011)…….…… 10

1.2 Generally used AFT in survival analysis(Sewalem, 2012).…….…….……. 13

1.3 The following steps were providing of analyzing the clinical trial for survival analysis in R……..14

3.1 Description illustrated of the clinical data for prostate cancer..…….…… 22

4.1 Survival curve of two tumor groups (primary and Met) for the prostate data in table 4.1… 43

4.2 Shows the lines for the prostate cancer data with two types of tumors…….…… 45

4.3 Survival times of patients with primary tumor according to Gleason grade.…...….... 51

4.4 The Cox proportional hazard PH with error bars show 95% confidence intervals...…….……...…….....……..... 53

4.5 Schoenfeld residuals for each explanatory variable versus transformed time in a model fit ...to the prostate cancer data. ...……...….……...… 55

4.6 Cumulative hazard plot of the Cox-Snell residual for Cox PH model to indicate the overall model……...…….…….…….…….…….. 56

4.7 Plot of martingale residuals vs. covariates. ……..…….……...…. 58

4.8 Deviance residuals consist of information about the influential and outlier data. ……...… 60

4.9 Cumulative hazard plot of the Cox-Snell residual for Weibull AFT model. ……...… 63

List of Table

Table		page
1.1	Four stages of Tumor (University of Maryland).	6
1.2	The survival time	8
1.3	The main model in survival analysis.	13
3.1	Clinical data of prostate cancer.	21
	Descriptive statistics for the distributions of the variables (Taylor, 2010)	24
4.1	Initial sorted table for Kaplan- Meier and Log- Rank analysis	41
4.2	Calculation for the K-M estimate of the survival function for primary type of tumor	46
4.3	Calculation for the K-M estimate of the survival function for Met type of tumor	47
4.4	Calculation for the log- rank test to compare tumor groups for the data in Table 4.1	48
4.5	The Cox's proportional hazards analysis for the prostate cancer patient	51
4.6	The hazard rate	51
4.7	Scaled Schoenfeld Residuals of Significant Covariates on the PH	54
4.8	Deviance residuals against the risk score	59
4.9	The log-likelihoods and Akaike Information Criterion (AIC) in the AFT models	61
4.10	Results from AFT models for time to progression with Weibull distribution.	62

List of APPENDIX

Appendix page

Appendix A ..... **Error! Bookmark not defined.**

Appendix B 78

Appendix C: 81

Appendix D: 83

Abbreviations

HR	hazard ratio
PH	Promotional hazard
CI	confidence interval
PSA	prostate specific antigen
RP	radical prostatectomy
Mets	metastasis
GS	Gleason score
BCR	biochemical recurrence
K-M	Kaplan–Meier
AIC	Akaike's information criterion
AFT	Accelerated failure time
MSKCC	Memorial Sloan Kettering Cancer Center
PathStage	Tumor stage
PathGGS	Combined Gleason score
PathGG1	Secondary Gleason grade
PreDxBxPSA	PSA level at diagnosis
BCR_FreeTime	Time until recurrence (months)
BCR_Event	Recurrence event (as defined by rise of PSA level)
Race	Patient race
DxAge	Age at diagnosis (years)

BxGGS	Biopsy combined Gleason score
BxGG1	Biopsy primary Gleason grade
BxGG2	Biopsy secondary Gleason grade
ClinT_Stage	Clinical Tumor stage
SMS	Surgical margin status
ECE	Extra-capsular extension
SVI	Seminal vesicle invasion
LNI	Lymph node involvement
Ng/ml	nanogram/millilitre
IQR	interquartile range

Chapter1

Introduction

This chapter starts with an essential issue in health, which is cancer. Specially, a review of prostate cancer with the related prognostic factors is presented. An overview of survival analysis is discussed along with important models that are relevant to the present study.

1.1 Prostate Cancer

Cancer is a term used for group of diseases where the cells have abnormal behavior of growth and division. There are more than 100 different types of cancer. The prostate cancer originates in prostate gland in the male reproductive system. Its function is producing fluid that protects and nourishes sperm cells in semen. The cancer cell can spread in different ways, such as through tissue, lymph system, and blood (National Cancer Institute). Prostate cancer is the second most common malignant cancer causing death in men, after lung cancer, and its incidence increases with age. Compared to other cancers, men with prostate cancer can live many years, since it grows slowly (Prostate Cancer Canada). Fortunately, prostate cancer in early stages of the disease, in half of the new cases, is still confined to the prostate. However, there are a significant number of cases of aggressive prostate cancers that can be very devastating. The cell of prostate cancer can spread to other parts of the body, which is called metastasis, such as the bones and lymph nodes.

The numbers of new cases of prostate cancer diagnosed each year in the US are approximately 220,000 and 30,000 of them die of the disease. In addition, the number of new cases of prostate cancer in 2013 was 238,590, and the number of deaths was 29,720. In 2014, the new cases of prostate cancer diagnosed were about 233,000 and about 29,480 of them died (American Cancer Society).

Research has identified the fundamental risk factors of prostate cancer; they are: age, race/ ethnicity, and family history. Among these, age has been found to be the most important factor, especially in older men over 60. It is found in the research that surgical radical prostatectomy (RP) had better results in young men than in older men. Literature has proved that African Americans have higher risk of prostate cancer, approximately 60%, than whites (Litwin et al.,

2000). If close family members, such as parents and grandparents, have had the disease, it is more likely that their children would have it (Prostate Cancer Canada). There are other possible factors that may increase the risk such as diet, body mass index (BMI), concomitant medical conditions, and hormone profiles,

1.1.1 Tumors

Primary Tumor

When the cancer has begun in any organ or tissue, the original tumor site is referred as the primary tumor or cancer.

Metastatic Tumor

Metastasis is a process that refers to the spread of cancer. This process can be understood as cancer cells breaking away from the primary tumor in the body and then entering the bloodstream or lymphatic system. A metastatic tumor or a metastasis is a tumor that is made via metastatic cancer cells. The cells can spread to the adrenal gland, bones, liver, and lungs. Metastatic cancer occurs exclusively in male patients, affecting the prostate. While the cancer can cause severe pain in patients, the depletion of testosterone or ingestion of medications, can improve the patient's urinary function and relieve some pain and discomfort.

Metastatic cancer is considered to be similar to the primary cancer. In many cases of the metastatic, if it is found as the first tumor, then the primary can also be found. However, some patients can have the metastatic without the primary tumor (National Cancer Institute). A pathologist examines to determine if a cancer is a primary or a metastatic tumor.

1.1.2 Prognostic Factors in Prostate Cancer

Clinical prognostic factors can be obtained through physical examination such as blood tests, radiological evaluation, and microscopy of biopsy material. The survival and prognosis of prostate cancer is affected by several clinical prognostic factors that give information about the cancer characteristics before planning a treatment decision. Some of the factors are the Gleason grades, PSA test, and size of tumor stage. They determine the survival rate after surgical radical prostatectomy.

Prostate-specific antigen (PSA)

The prognosis as well as diagnosis of prostate cancer has been achieved using the PSA (prostate specific antigen). Serum PSA, particularly free PSA is used widely as a marker for monitoring the performed surgery and treatment provided specifically for the prostate tissue. The PSA is a screen test commonly used for identifying early stages of prostate cancer. Prostate gland has cells, which are used for producing a protein called PSA. The human blood has PSA levels that are measured using a PSA test. The abnormality or normality of cells is indicated by the PSA results. After diagnosis of cancer PSA levels may be used for determining the extent of the disease. The levels of PSA that range from 4.0 (ng/ml) or lower were regarded by doctors as normal. In contrast, levels of PSA that are above 4 (ng/mL) are an indication that most parts of the body are affected by cancer (metastasis). However, from a general perspective, when the human body has high levels of PSA, then there is a high possibility that such a person could be suffering from prostate cancer. Therefore, it gives us a hint that the PSA test is not perfect. Additionally, prostate cancer may be indicated by a gradual increase in the PSA levels.

The levels of serum PSA are suitable determinants of prognosis outcomes after radiotherapy of prostate cancer and tend to increase the prognostic data that is free of tumor stages as well as grade (Buhmeida et al., 2006). Once a patient undergoes a radical prostatectomy, doctors will typically monitor the PSA levels, looking for any rise in levels which are typical indicators of a recurrence of clinical carcinoma (Penn State Hershey Medical Center). Physicians refer to the rate of increase as PSA velocity (PSAV). The PSAV is then used to determine the most applicable type of treatment along with the treatment's starting time.

Although PSA plays an important role in the prediction of long-term survival in patients, the follow-up period for monitoring PSA levels needs to be seriously considered (ibid). More research that focuses on the length or duration of the follow-up period is necessary before researchers can effectively determine the efficiency of this process.

Gleason grade

Notably, other factors that pose a high risk revolve around the Gleason score and represent the severity involving the prostate cancer tissue; it plays a critical role by helping the doctor in the identification of methods that are suitable for treatment of a specific case of cancer. The Gleason

technique is used for classifying the scores of cancer cells through analysis of the microscopic structure. An important characteristic of Gleason score revolves around its prognostic factor that is useful for determining the progression of cancer as well as death.

In addition, it is the most significant development, which influences the results. Univariate as well as multivariate prognosis analyses involving prostate cancer usually considers the Gleason grade as a key predictor of patient results (Buhmeida et al., 2006). The Gleason grade is an approach that was designed in 1960s by Dr. Donald Gleason (Humphrey, 2004). Two grades, which can be used by a pathologist to produce the score when a biopsy sample is being examined using a microscope for a specific pattern that ranges between 1-5, whereby 1 represents the normal prostate tissue and 5 represents the abnormal prostate tissue. The calculation of the Gleason score can be achieved after primary as well as secondary grades have been identified. The primary grade indicates the common tumors, which are over 50 percent, whereas the secondary grade is a representation of less frequent tumors that produce a score of below 5 percent. The Gleason score is created by combining the two grades that have a highest score of five. It contains a range of 2-10 (Humphrey, 2004). For instance, when the grade of primary tumor is three and that of secondary tumor is 4, the sum of the two grades will produce the Gleason score, that is 3+4=7 (Russell, et al., 2003). To have a clear understanding of the ways in which the Gleason score can indicate same biological behavior, they are classified into various groups. The lowest level of cancer is indicated by 6 on the Gleason score (significant differentiation of the tumor tissue), whereas 7 represents a mild grade of cancer (moderate differentiation of the tumor tissue). Additionally, a level ranging from 8-10 is an indication of higher level of cancer (poor differentiation of the tumor tissue). The highest level produces a severe cancer that has a rapid rate of separation compared to lower level cancer.

Tumors Stage

Notably, prostate cancer staging was achieved using two types of data namely clinical and histo-pathological data. Clinical information is usually obtained from external cancer symptoms, whereas histo-pathological data is obtained after surgically removing and examining the prostate tissues. Clinical information plays a critical role in enabling doctors to make decisions regarding the treatment. On the other hand, histo-pathological data is widely used for prognosis prediction. Because of this, prostate cancer staging considers clinical as well as histo-pathologic data.

Doctors particularly examine the size of the tumor (T), the involvement of the lymph node (N), visceral presence/metastasis (M) as well as tumor grade (G).

Additionally, the tumor size is regarded as a possible risk factor of prostate cancer. Studies have shown that an increase in tumor size led to an exponential increase in malignant tumors (Chan, 2012). Indeed, tumor cell has a relation with the risk of mortality amongst prostate cancer patients who are on observation (Andreas Josefsson, 2012). After the diagnosis of cancer, tumor stage checks the magnitude of severity as well as cancer spread. Table 1.1 describes four stages of tumor as T1 – T4, which represent the size as well as the spread of the tumor.

Table 1.1: Four stages of Tumor (University of Maryland).

Stage, T1 - T4	Description
T1	The tumor cannot be felt or seen using imaging techniques.
	T1a. Cancer cells are incidentally found in 5% or less of tissue samples from prostate surgery unrelated to cancer.
	T1b. Cancer cells found in more than 5% of samples.
	T1c. Cancer cells identified by needle biopsy, which is performed because of high PSA levels.
T2	The cancer is confined to the prostate but can be felt as a small well-defined module.
	T2a. Tumors are in half a lobe.
	T2b. Tumors are in more than half a lobe.
	T2c. Tumors in both lobes.
T3	The tumor extends through the prostate capsule.
	T3a: the tumor has spread through the capsule on one or both sides
	T3b: the tumor has invaded one or both seminal vesicles
T4	The tumor is fixed to or invades adjacent structures.

1.1.3 Treatment

Prostate cancer in men may be treated using a variety of techniques such as bisphosphonate therapy, chemotherapy, hormone therapy, radiation therapy as well as surgery. The aforementioned techniques are utilized separately; however, in certain instances they can be integrated.

The most commonly used technique for treating cancer that has not spread beyond the gland is called surgery. A radical prostatectomy (RP) refers to a surgical process, which entails the removal of prostate gland alongside the attached seminal vesicles. During this procedure, the lymph nodes that are adjacent to the prostrate may be removed simultaneously. Radical prostatectomy (RP) is a common alternative for treating prostate cancer that has not spread to other areas of the body. Specifically, it is preferred when the patients are younger than 70 years, otherwise radiation therapy is preferred (Stangelberger, 2008). The follow up after the RP can detect the PSA level, which identifies the patients with elevated risk of local treatment failure or metastatic disease. Though PSA level after surgery is high in some cases, patients are still free of symptoms for extended periods of time. Therefore, the PSA level may not be enough to initiate additional treatment (National Cancer Institute).

1.2 Survival Analysis

Since cancer ranks as the second leading cause of death in the world, survival analysis techniques have been used to measure the risk, hazards, and average survival time for cancer patients. The common research involving cancer is based on time called the survival time. The term ‘survival time’ is used in reference to the number of days, weeks and years from the time patient’s observance begins until death takes place. Since 1950s, survival analysis has proved to be an important technique (Langova, 2008). There are several areas where survival analysis is applied which include demography, economics, engineering, epidemiology, health, medicine and biology. Additionally, there has been an increase in the use of survival analysis in areas of biostatistics as well as pharmaceuticals. There are several objectives for survival analysis which include estimation as well as interpretation of survivor function using survival data, comparison of survivor functions coupled with assessment of the link involving defined variables and time of survival (Langova, 2008). Since the survival analysis was provided with cancer data, we need

special data that is called clinical trials. They are conducted to determine the effectiveness of new treatment (Singh, and Mukhopadhyay, 2011). Usually in survival studies, the patients are kept over a long period of time, so other factors are important to be still continual over the period.

The dependent variable within the survival analysis is composed of two attributes namely, time-to-event as well as event status. An endpoint occurs either when the event occurs or when the follow-up time has ended. There are several endpoints that can be the events such as death, relapse of disease, recurrence of a tumor, recovery or any designated experience of interest as shown in Table 1.2. It marks the indicator variable as 1 if the event of interest was observed or 0 if it was censored.

Table 1.2: The survival times

Starting Point	End Point
Surgery	Death/ Relapse/ Recurrence.
Diagnosis	Death/ Relapse/ Recurrence.
Treatment	Death/ Relapse/ Recurrence.

Standard statistical methods may not be used widely because the inherent distribution is abnormal and there is censoring of data (Bewick, 2004). Censoring of survival time occurs when the time of follow-up is available though it might have taken place unnoticed or has not taken place. Several techniques utilized in survival analysis include non-parametric, semi-parametric as well as parametric. Within the techniques two kinds of information involving clinical trial for survival analysis exist namely, censored as well as uncensored data. Exact data (uncensored data) refers to a situation whereby the participant is aware until the occurrence of event-of-interest.

1.2.1 Censored data

Censored data emerges as a critical issue for consideration in survival analysis, because it helps to indicate the kind of missing information. Censored data arises when a negative event takes place for instance, withdrawal of participant, difficulties in tracking the participant, participant has not encountered the suitable results or the relevant data is unavailable. Notably, an indicative variable with value 1 is used when the uncensored survival time has been identified and value of 0 is used for right-censored times (Zhao, 2008). Three separate scenarios involving censored data that rely on the follow-up times are in existence. However, this relies largely on the stage-level as well as risk-level for the patient.

1. *Right censored* refers to a patient who may not encounter a time failure for the event-of-interest until the follow-up duration elapses or withdraws from the study before it ends.

2. *Left censored* refers to a situation whereby the event-under-interest takes place prior to enrolment. This scenario is not common.

3. *Interval censored* this situation takes place when an event-under-interest has a tendency of occurring in a specific time (Zhao, 2008).

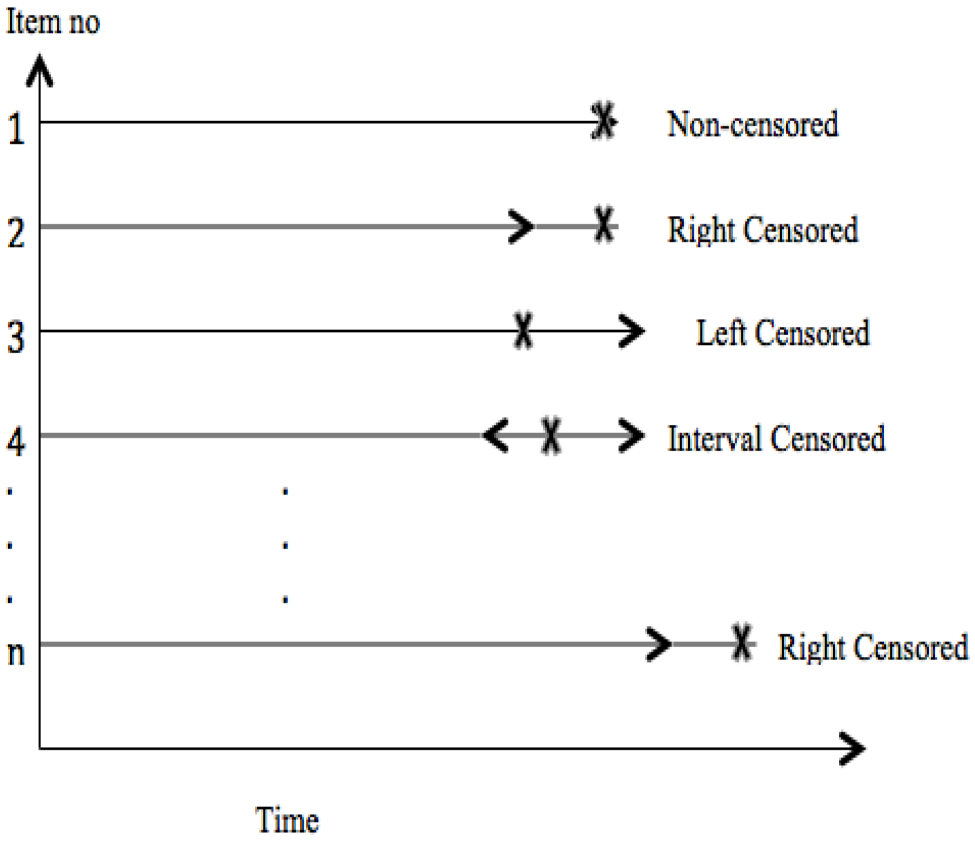


Figure1.1: Illustration of left, right and interval censoring

The crosses in Figure 1.1 indicate when the failure occurs, while the arrows perpendicular to the time axis show the actual times (Aaserud, 2011).

Overall, the feature of censoring implies that special techniques of analyzing are essential. Most widely used technique for analyzing is right censored.

1.2.2 Functions related to survival analysis

Before choosing a technique for use in survival analysis, it is imperative that two functions, which are time dependent, are considered. They include survival function and hazard function that may be explained using the survival information.

- The survival function S(t) produces the survival probability approximately to time t. The survival function is essential for survival analysis. The Kaplan-Meier curve provides the survival function.

$$S(t) = P(T > t) = 1 - F(t), t \geq 0. \qquad \text{(Fox, 2002)}$$

T represents a positive random variable that covers the time from commencement of the observation up to survival. $F(t)$ is the distribution function.

- The hazard function h(t) represents the probability condition of death at time t after survival time.

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}, t \geq 0.$$

A relationship between S(t) and h(t) is shown in the formula below:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d \log S(t)}{dt}$$

$f(t)$ is the density function which gives the fraction of the original group for whom the event occurs during the time interval at t adjusted for the width of the time interval.

If one of $S(t)$ or $h(t)$ is known, the other can be calculated.

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp\big(-H(t)\big), t \geq 0.$$

Where $H(t)$ is the cumulative hazard function. $H(t)$ is difficult to interpret, but there is an easier way to make a clear interpretation. The way is to think of $H(t)$ as the cumulative force of mortality or if the event were a repeatable process, the number of events expected for each individual by time t.

The probability density function of T can be defined as

$$f(t) = h(t)\ \exp\left[-\int_0^t h(u)du\right], t \geq 0.$$

Hazard ratio (HR)

It is expressed as the relative risk that is used to estimate the ratio of the hazard rate. In addition, it has been utilized for describing the outcome of the trials therapy in order to figure out the range the treatment can shorten the duration of the disease (Singh, and Mukhopadhyay, 2011).

The processing of statistical data should entail application of relevant techniques. The survival module is characterized by four essential techniques to fit survival models. These models are illustrated in Table 1.2. The last strategy in Table 1.2 that is more direct is the parametric technique (accelerated failure time), whereby there is assumption on the specified functional form of the baseline hazard (t). Within this technique several distributions which acquire a central point are in existence such as Weibull, generalized gamma, log-logistic and lognormal. The aforementioned approaches are explained in chapter 3 and can be integrated into our information to ascertain their suitability.

Different models can be classified as: proportional hazards model (exponential and Weibull) and proportional odds model (log-logistic). Figure 1.2 illustrates that the Weibull and exponential models can be both the accelerated failure time (AFT) model and proportional hazards model. In addition, it provides the commonly used parametric models.

Table 1.2: The main models for survival analysis

Technique	Goal
Kaplan-Meier	Estimate the probability of an individual surviving for a given time period
Log-rank Test	Compare survival of two different groups of individuals.
Cox regression	Detect clinical/ genomic/ epidemiologic variables, which contribute to the risk.
Accelerated failure time (AFT)	Used as an alternative model to the Cox model where the proportional hazard assumption is not held constant.

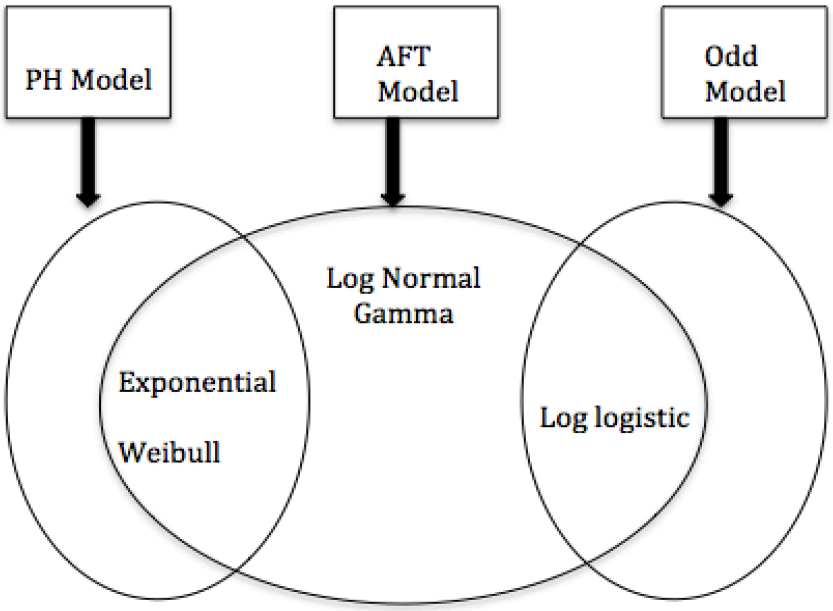


Figure 1.2: Parametric models in survival analysis (Sewalem, 2012).

The above methods have different properties and interpretations. However, each model can summarize survival data.

The steps of the survival analysis using R programming are shown in Figure 1.3 (Yang et al., 2011).

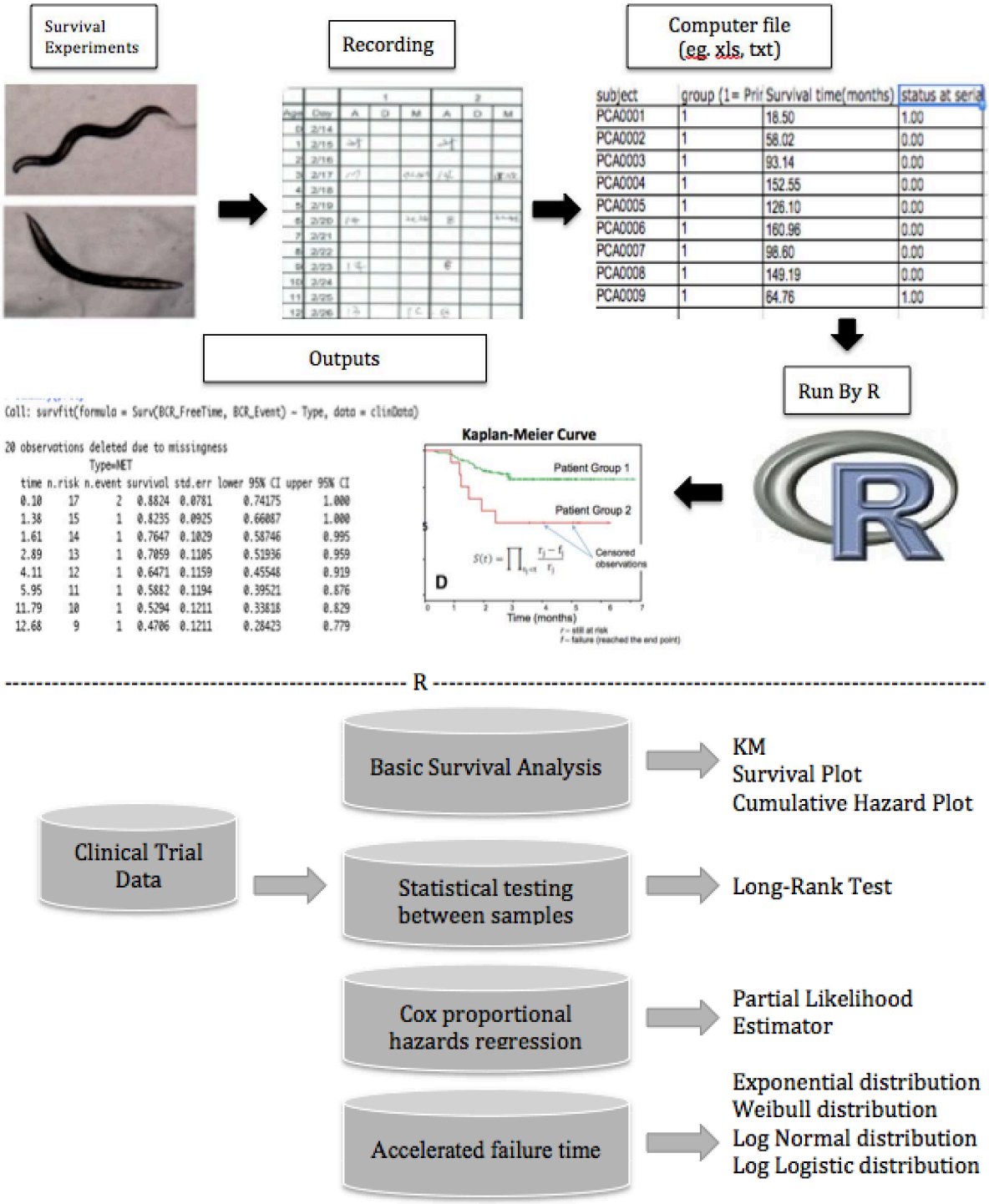


Figure 1.3: Steps for analyzing the clinical trial data for survival analysis in R.

1.3 Objectives

There are some essential goals that are presented in the current study in order to clearly understand the important meaning of the survival analysis and how it deals with prostate cancer data.

1. The objective of the present study is to address some important questions related to prostate cancer survivorship for patients with primary or metastatic tumor. It is commonly understood that the risk of developing prostate cancer is higher in metastatic tumor than the primary. The objective of this thesis is to examine the efficiency of several methods that are commonly used to estimate survival functions in the presence of censored data. Kaplan-Meier analysis was performed to estimate survival in univariate analysis. We compare different techniques to estimate survival functions.

2. This research investigates the influence of standard clinical prognostic features on the survival time of prostate cancer patients. Particularly it seeks independent variable patterns to determine the survival times and identify the correlations among the variables of interest. For this goal the Cox model performed well, which identified covariates associated with survival.

The rest of the thesis is organized as follows:

Chapter 2 provides a literature review on survival analysis for cancer dataset. Chapter 3 presents the data set and its description to clearly understand it. Additionally, it provides in detail the methodology that was performed for survival analysis suitable for the given data. Chapter 4 summarizes and discusses the results. Chapter 5 concludes the study. Finally, appendices summarize R code and a few outcomes.

Chapter 2

Literature Review

2.1 Survival Analysis Study

Litwin, et al., (2000) faced significant challenges that kept them from bringing data together from different studies in order to assess disparities in results of treatment in various institutions. Initially, they were presented with different endpoints from the studies. Following this, they noticed that the different studies showed varying disease severity. Finally, usefulness of the results was limited by the differences in the techniques used to measure patient-focused outcomes. There are several clinical trials where survival analysis model was used. We present here some of the techniques of survival analysis for cancer data especially prostate and breast.

Vinh-Hung, V. et al. (2002), Post-surgery radiation in early breast cancer: survival analysis of registry data

Vinh-Hung et al (2002) showed the survival advantage in patients diagnosed with early breast cancer, treated with post-surgery radiation. The current study tries to prolong these results through an organized population data analysis. This study made use of Epidemiology, Surveillance, and End Results (SEER) data on 83,776 women suffering from breast cancer diagnosed between 1988 and 1997, stage T1–T2, node positive or node negative. The proportional hazard models were used for the analysis.

Results showed that the best rates of survival were found with combined radiation and breast-conserving surgery in all cases. The available data indicate that post-surgery radiation provides a survival advantage irrespective of the type of surgery in node positive patients. Likewise, survival advantage was observed with post-surgery radiation and breast-conserving procedure in node negative patients.

Ray, M.E. et al. (2009), Potential surrogate endpoints for prostate cancer survival: analysis of a phase III randomized trial

In their study, Ray, M. E. et al. (2009) determined that surrogate endpoints for prostate cancer specific survival may reduce the length of the clinical trials for a patient's prostate cancer. This

study assessed distant metastasis and failure of general medical treatment as possible substitutes for prostate cancer-specific survival using data from the Radiation Therapy and Oncology Group 92-02 randomized experiment.

They use data where patients (n = 1554 assigned randomly and 1521 evaluable for the study) having locally advanced prostate cancer had undergone 4 months with neoadjuvant and simultaneous androgen deprivation therapy with external treatment of beam radiation. These patients were then randomly assigned to either no additional treatment (control arm) or 24 extra months of androgen deprivation therapy (experimental arm). Statistics coming from the point of origin of examinations at three and five years for failure of general clinical treatment (meaning the recorded progression of local disease, distant or regional metastasis, induction of androgen deprivation therapy, or a prostate-specific antigen level of 25 ng/mL or more following radiation therapy) and/or distant metastasis were examined as substitute final stages for prostate cancer-specific survival at 10 years using Prentice's four criteria. The Cox proportional hazard (PH) models were utilized to provide the hazard ratio (HR) between the treatments. All statistical investigations were two-sided.

At 3 years, 1364 patients were surviving and provided data for evaluation. Both general clinical treatment failure and distant metastasis at 3 years indicated consistency with all four of Prentice's criteria for being a substitute for final stages for prostate cancer-specific survival at 10 years. At 5 years, 1178 patients were surviving and offered more data for evaluation. Even if prostate cancer-specific survival did not statistically vary substantially between arms of treatment at 5 years (P = .08), both final stages showed consistency with the remaining Prentice's criteria. They concluded that general clinical treatment failure and distant metastasis at 3 years may be possible with alternative endpoints for prostate cancer-specific survival at 10 years. However, these final stages must be verified with other collections of data.

Chan, Y.M. (2013), Statistical Analysis and Modeling of Prostate Cancer

The study performed by Chan (2013) provided a comparison of survival between African American and White men at the four distinct stages of prostate cancer under the same treatment. Moreover, the study made it possible to estimate the average difference in survival between

White and African American males diagnosed with prostate cancer and addressed some of the critical issues related to treatment.

There is a common perception that African American men have a greater risk of developing prostate cancer than men of other races. Nevertheless, with the use of parametric analysis, Chan's study showed that the perception is more of a myth than reality. The study further recognized the presence of racial/ethnic differences by relating the average size of tumors, the median time of survival time, and the survival function between African American and White men. These outcomes emphasize the need for acknowledging the role that racial background plays in improving clinical targeting, and in so doing, advancing clinical results. Moreover, parametric survival analysis was conducted to approximate the survival rate of white men going through various treatments at every stage of prostate cancer. In addition, to comprehend the risk factors (tumor size, age, age and tumor size interaction) linked with survival time, a framework of accelerated failure time was created. The model could precisely forecast the survival rates of white men at each stage of prostate cancer according to the received treatment.

Lastly, the outcomes of parametric survival analysis and the framework of accelerated failure time model are compared among white men going through the same treatment at each disease stage.

Pulte, D. (2012), Changes in survival by ethnicity of patients with cancer between 1992–1996 and 2002–2006: is the discrepancy decreasing?

The study organized by Pulte et al (2012) emphasized that those patients of marginal race or ethnicity report lower rates of survival in most of cancer diagnosis cases. Over time, there exists few data regarding changes in the differences. Here, they evaluated changes in survival rates of patients with common cancers in two latest periods of time by ethnicity or race.

In their methods they utilized a structured period analysis to define relative survival (RS) for African–American (AA), non-Hispanic white (nHw), and Hispanic patients in the Epidemiology, Surveillance, and End Results database detected with usual solid and hematological distortions.

The results show the RS for five years became better for nHw for every tumor detected, between +2% points (pancreatic cancer) to +16.4% points [non-Hodgkin's lymphoma, (NHL)]. Greater

progress was seen for Hispanics and AA than nHw in NHL, and prostate and breast cancer. There was less improvement for Hispanics and AA than for nHw for pancreatic and lung cancer. Statistically, for Hispanics and AA with acute leukemia or myeloma, no substantial improvement was observed. Disparities in survival remained between AA and Hispanics from 0.5% points for myeloma to 13.1% points for breast cancer

They conclude that there has been advancement in reducing the differences in survival between minorities and nHw in NHL, prostate and breast cancer. Minimal improvement has been possible in decreasing the differences for their cancers.

Chapter 3

Materials and Methodology

3.1 The Data source

Prostate survival data was obtained and generated from the Memorial Sloan Kettering Cancer Center (MSKCC). At MSKCC, each sample had been collected from the recipients of radical prostatectomy treatment. Out of the sample group, 181 patients were diagnosed with primary tumors along with one patient who has both metastases and primary tumor and 37 with metastases including the aforementioned patient. The association between survival, the tumor type, and the raw data contains 230 cases (rows) and 164 variables (columns). Additionally, for this recent analysis 40 variables have been chosen to improve the results. Detailed description of the data set can be found in the article written by Taylor et al. (2010). The format of the dataset is shown in Table 3.1, where they defined the data such as the type of tumor (primary or metastatic), age (years), race (White Non-Hispanic, Black Non-Hispanic, White Hispanic, and Black Hispanic) and so on. The data was converted from a factor into numeric, in order to apply it in R.

The data collected from MSKCC has some primary criteria to select patients after the radical prostatectomy (RP) treatment:

- Serum PSA testing every 3 months for the first year, 6 months for the second year, and annually thereafter. For all analyses described here, biochemical recurrence (BCR) was defined as PSA $\geq$ 0.2 ng/ml on two occasions.
- Following radical prostatectomy, patients were followed-up with history, and physical exam.
- The size of the primary tumor was known
- Initial biopsy Gleason score is between 6 to 9.
- At the time of data analysis, patient follow-up was completed through December 2008. The age was between 37.3 and 83.

Table 3.1: Clinical trial data of prostate cancer.

Sample ID	Type	MetSite	Race	PreDxBxPSA	DxAge	BxGG1	….
PCA0171	PRIMARY	NA	White Non-Hispanic	8.20	60.87	3.00	….
PCA0172	PRIMARY	NA	White Non-Hispanic	17.20	52.00	4.00	….
PCA0173	PRIMARY	NA	White Non-Hispanic	5.24	66.16	3.00	….
PCA0174	PRIMARY	NA	White Non-Hispanic	4.60	54.32	4.00	….
PCA0175	PRIMARY	NA	Black Non-Hispanic	5.60	51.00	3.00	….
PCA0176	PRIMARY	NA	Black Non-Hispanic	8.10	53.55	3.00	….
PCA0177	PRIMARY	NA	White Non-Hispanic	5.63	59.78	4.00	….
PCA0178	PRIMARY	NA	White Non-Hispanic	4.60	61.83	3.00	….
PCA0179	PRIMARY	NA	White Non-Hispanic	40.24	64.89	4.00	….
PCA0180	PRIMARY	NA	White Non-Hispanic	8.03	67.17	4.00	….
PCA0181	PRIMARY	NA	White Non-Hispanic	27.00	69.01	4.00	….
PCA0182	MET	Node	White Hispanic	30.00	58.00	4.00	….
PCA0183	MET	Bone	White Non-Hispanic	182.10	82.00	5.00	….
PCA0184	MET	Node	White Non-Hispanic	27.00	69.00	3.00	….
PCA0185	MET	Node	White Non-Hispanic	27.00	69.00	3.00	….
….	….	….	….	….	….	….	….

The headings found in Table 3.1 are described in detail using proper medical language in Figure 3.1.

Variables	Description
Sample ID	Sample ID
Type	Sample type (primary / metastasis / cell line)
MetSite	Site of the metastasis
Race	Patient race
PreDxBxPSA	PSA level at diagnosis
DxAge	Age at diagnosis (years)
BxGG1	Biopsy primary Gleason grade
BxGG2	Biopsy secondary Gleason grade
BxGGS	Biopsy combined Gleason score
PreTxPSA	PSA level prior to radical prostatectomy
ClinT_Stage	Clinical Tumor stage
NeoAdjRadTx	Neoadjuvant therapy
ChemoTx	Chemotherapy
HormTx	Hormonal therapy
RadTxType	Radiation
RP_Type	Type of radical prostatectomy
SMS	Surgical margin status
ECE	Extra-capsular extension
SVI	Seminal vesicle invasion
LNI	Lymph node involvement
Num_Nodes_Removed	Number of lymph nodes removed for examination
Num_Nodes_Positive	Number of lymph nodes involved by tumor
PathStage	Tumor stage based on pathologic examination of the radical prostatectomy
PathGG1	Primary Gleason grade in the radical prostatectomy specimen
PathGG2	Secondary Gleason grade in the radical prostatectomy specimen
PathGGS	Combined Gleason score in the radical prostatectomy specimen
BCR_FreeTime	Time until recurrence (months)
BCR_Event	Recurrence event (as defined by rise of PSA level)
MetsEvent	Metastasis resulting from the primary tumor
SurvTime	Overall survival (months)
Event	Death
Nomogram PFP_PostRP	5 -year probability of freedom from biochemical recurrence after radical prostatectomy
Nomogram NomoPred_ECE	Probability of having extra-capsular extension
Nomogram NomoPred_LNI	Probability of having lymph node metastases
Nomogram NomoPred_OCD	Probability of having organ confined disease
Nomogram NomoPred_SVI	Probability of having seminal vesicle invasion
Copy-number Cluster	Copy-number cluster assignment
ERG-fusion aCGH	ERG fusion status determined by copy-number (intragenic deletion)

Figure 3.1: Description of the clinical data for prostate cancer.

Descriptive statistics can provide some information about the distributions of the variables such as the average, minimum and maximum time in days for the censored and failed observations. In our data the descriptive statistics provided information for the two types of tumors: primary and metastatic as shown in Table 3.2. Patients in our study ranged in age from 37.3 to 83 years.

An example of these statistics can be found in the following table. As the table shows, the PSA value was tested at less than 4, between 4 and 10, and greater than 10 ng/ml. There are 31 (17.2%) patients with primary tumors whose PSA was less than 4, and 4 patients of metastatic (12.5%). There are 105 (58.3%) patients with primary tumors whose PSA was between 4 and 10, and 6 (18.75%) patients of metastatic. Finally, there are 44 (24.5%) patients with primary tumors whose PSA was greater than 10, and 22 (68.75%) patients with metastatic.

Table 3.2: Descriptive statistics for the distributions of the variables (Taylor et al., 2010).

Characteristic	Primary tumor	Metastatic
Age		
Median	58.3	60
Mean	58.3	60
Standard deviation	7	8.6
Min-max	37.3–83	41–82
PSA at diagnosis (ng/ml)		
Median	6 (IQR 4.4, 9)	17 (IQR 8.6, 46.6)
<4	31 (17.2%)	4 (12.5%)
4–10	105 (58.3%)	6 (18.75%)
>10	44 (24.5%)	22 (68.75%)
Initial biopsy Gleason score		
5	2 (1%)	--
6	101 (56%)	2 (6%)
7	61 (34%)	16 (46%)
8	11 (6%)	8 (23%)
9	6 (3%)	9 (25%)
Initial clinical stage		
cT1c	95 (52.4%)	8 (22%)
cT2	76 (42%)	12 (33%)
cT3	9 (5%)	3 (8%)
cT4	--	1 (3%)
Not available	--	9 (25%)

3.2 Methodology

We examine non-parametric, semi-parametric and parametric methods to estimate the survival function for the given data.

3.2.1 Non-parametric Methods

Kaplan Meier Estimates (K-M)

The first point to consider is how censoring can be adjusted in the K-M method in order to estimate the survival function. As the K-M method makes no assumption about the shape of the underlying survival curve, it is categorized as a non-parametric method for estimating a survival function. However, using a non-parametric analysis typically generated much wider confidence bounds than those calculated via parametric analysis. Parametric analysis shows how predictions outside the range of observations are not possible with non-parametric analysis.

As we have defined earlier in chapter 1, the survival function is the mathematical equation that describes a smooth curve, depicting the lifetime of the population of interest. When the survival time differs especially when some subjects are excluded in the study, K-M curve is one of the most common curves to tackle the likewise circumstances.

The characterization of all the subjects of the survival analysis by K-M method can use only three variables (Rich, et al., 2010). The first variable is the serial time which begins with the commencement of the treatment and gets censored from the study when it reaches the end point. At the end of the serial time, the second variable consists of the patient's status. The third variable is the study groups the patients belong to.

The idea of this method is based on the probability of the surviving in k or more periods in the study and is a product of k probabilities when each period is observed under it. It is written by the following expression:

$$S(k) = p_1 \times p_2 \times p_3 \times ... \times p_k$$ (Bewick, et al., 2004)

In the above equation p_1 constitutes surviving proportion in the first period, p_2 is the proportion survived over the second period, and so on. The equation below gives the proportion of surviving for period i where they survived up to period i:

$$p_i = \frac{r_i - d_i}{r_i}$$

Where,

r_i is the number of patients living at start of the period i, and

d_i is the number of deaths.

Considerations in the K-M curves:

Identifications have to be made in the analysis of the K-M curve regarding the units of measurement according to the axes and the events of interest. After that, the evaluation of the curve, the number of the censored subjects and their distribution are very important. The number of participating subjects is much more if the curve consists of many small steps. However, limited number of participating subjects results in a curve with large steps. The time period of being alive of those patients who have been treated is measured in many medical studies.

In current study survival package of R program was used for survival analysis. In our data the first variable to characterize each subject of K-M survival analysis is the serial time in the column BCR_FreeTime, which is time until death (in months). The second variable status is given by column BCR_Event, its event (after RP as defined by rise of PSA level). The time to event (death) was estimated for a group of individuals. The third variable study group is the type of prostate cancer (primary or met). The instruction in R to estimate the survival function is given by:

```
pFit <- survfit(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData)
```

The results sorted by ascending serial times beginning with the shortest times for each group are shown in Chapter 4. A subset of the data is used to look at primary tumors only according to the tumor Gleason score to show the estimate the survival time and to show which tumor grade has a higher rate.

The median is chosen for a summary measure as the distribution of survival time is positively skewed. Therefore, the median survival time is defined as 50% of the individuals under study expected to survive in their time.

Log Rank

The K-M survival curves can provide us an idea about the difference between survival functions among two or more groups. However, it cannot give us whether this observed difference is statistically significant. Hence, many methods can be utilized to test the equality of the survival functions for different groups. The log-rank is one commonly used non-parametric test for comparing two or more survival distributions of the patients; it is also called Mantel log-rank. Additionally, this method is useful when the risk of an event is always greater for one group than another in order to detect a difference between groups.

The steps to complete this method start with arranging the survival time for both censored and observed times. The log rank test is a form of Chi-square test distribution with one degree of freedom (Singh, and Mukhopadhyay, 2011) that calculates a test statistic used for testing a null hypothesis. In our study the log rank test was used to test the null hypothesis that the survival curves for all groups are the same. To clarify, it was used to test whether or not there is a difference between the populations in the probability of an event (here death) at any point in time. For each point of time the observed number of deaths in each group and the number of expected deaths are calculated to determine if there was a difference. The number of expected deaths is determined by multiplying the total number of events at a given point in time with the proportion of subjects who are at risk at that point (Dakhil, et al., 2012).

The calculation of the test is:

$$X^2(long\ rank) = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$$ (Bewick, et al., 2004)

Here, O_1 and O_2stand for the number of total events that have been observed within the groups of 1 and 2 respectively.

The expected number of events is represented by E_1 andE_2.

The *survdiff* function in R implements the log rank test. In our study, this method is useful to detect the difference between two groups of tumor - primary and metastasis.

```
logDiff <-  survdiff(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData)
```

3.2.2 Semi-parametric Methods

Cox proportional hazard

The non-parametric methods are not useful for controlling the covariates and it requires categorical predictors. Therefore, the multivariate approaches are used when we have several prognostic variables. The most widely applicable and broadly implemented multivariate method in the survival analysis is the regression model of the Cox proportional hazards. In the year 1972 Cox showed the first light to the Cox model (Fox, 2002). The explanatory variables (determining the features of the patients with the estimation of the number of covariates and risk of death) and the response variables are combined. As any form can be adopted by the disturbance of the baseline, the nature of the model is semi-parametric (Fox, 2002). Disturbances are defined as the hurdles of death and the moments risked by death, which have been outlasted by the patients within a given period of time. There can be a number of homogenous regression models of Cox but hazard function is the variable to depend upon. However, the time factor does not affect the hazard function as it does in the survival function.

The Cox model, a regression method for survival data, provides an estimate of the hazard ratio which is always non-negative and its confidence interval. The hazard ratio is an estimate of the ratio of the hazard rate based on comparison of event rates. The hazard rate is the probability that if the event in question has not already occurred, it will occur in the next time interval, divided by

the length of that interval. The time interval is made very short, so that in effect the hazard rate represents an instantaneous rate. An assumption of proportional hazards regression is that the hazard ratio is constant over time.

The mathematical equation of the Cox model is:

$$h(t) = \exp\{ h_0 (t) + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \}$$

Or $$\log h(t) = h_0 (t) + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$ (Fox John, 2002)

Here,

h(t) represents the hazard function within the limited time period of t.

The covariates of x_1, x_2, …., x_p are also the explanatory variables.

If each explanatory variable x_i is zero ($\exp h_0 (t) = 1$) then the hazard of the baseline is represented by h_0(t). If covariates (risk factors) is characterized by dichotomy and is coded 1 if present and 0 if absent, respectively.

According to the implication of the proportionality, the quantities $\exp(b_i)$ are known as the ratios of hazards. The interpretation of the quantity $\exp(b_i)$ is an event when relative risks are evolved immediately irrespective of time . The individuals with the risk factors and individuals without the risk factors work equally with all the covariates. If the covariates are presupposed to be continuous then the interpretation of the quantity $\exp(b_i)$ is an event where the risk factors are evolved immediately irrespective of time where the value of the covariate increased by 1 was compared with another individual provided the effects of all the individuals on the covariates are same. Binary and continuous are the two types of covariates. The h (t)/ $h_0 (t)$ is known the hazard ratio.

The statistic package that triggers the changes in the hazards that are anticipated, helps to estimate the coefficients $b_1, b_2 \ldots b_p$. Since this method focuses on the relation, it is necessary for us to interpret the coefficients for each explanatory variable. The hazards will increase if the occurrence of the regression coefficient in explanatory variables is positive, and the hazards

decrease as an improved prognostic is yielded by the regression coefficient only when it is negative.

A Cox model was fitted by using an appropriate computer program in R to find the equation for the hazard as a function of several explanatory variables. This model was used to investigate the influence of standard clinical prognostic features on the survival time of Prostate cancer patients:

Before we fit the Cox proportional hazard, the variables should be converted to numeric as follows:

```
clinData $PathStage <- as.numeric(clinData $PathStage)
clinData $PathGGS <- as.numeric(clinData $ PathGGS)
clinData $ PathGG2 <- as.numeric(clinData $ PathGG2)
```

The best fit model with four explanatory variables is given by:

```
coxFit2 <- coxph(formula = Surv(BCR_FreeTime, BCR_Event) ~ PreDxBxPSA + PathGGS +
PathGG2 + PathStage, data = clinData)
```

The Adequacy of a model:

When clinical trials are performed, the primary analysis of the biomedical and biological applications has a number of available variables. However, the number of predictable errors can increase if the covariates are invalid (Liang and Guohua, 2008). The task of determining covariates for the statistical model has always been very sensitively critical in the process of data analysis. The selection of a proper model is broadly depended on the Akaike's information criterion (AIC).

AIC is applied in order to choose one model from two competing ones having the possibilities of various ranges of parameters. The models mostly carry a great number of covariates that are feasible and due to this each parameter should be evaluated separately with dynamic scale in search of the most befitted model. AIC of lower amount in a model is the finest trigger in the betterment of a model (Symonds, et al., 2011).

The implementation of the proposed method is quite feasible in the concurrent software named R/Splus and others. The command used in R for this method is extractAIC.

```
coxAIC <- extractAIC(coxFit2)
```

The AIC will be:

$$AIC = -2LL + 2(c + a)$$ (Bradburn, et al., 2003)

Here,

The logarithm of the similarities of the models is specified in LL.

c and a stand for the number of covariates and ancillary parameters respectively.

Selection of covariates by step-wise selection using p-values:

For the creation of a preliminary model this is the best-suited approach and it is subject to change if its ability of befitting is good. To perform step wise p-value is measured to keep the variable or remove the variable from the model. The variable can be kept in the model when its associated significance level is less than this p-value. In contrast, if its associated significance level is greater than this p-value then the variable will be remove from the model.

Testing the proportional hazards assumption

The assumption of proportional hazards (PH) function is the finest techniques in the Cox model. This model help clarify the idea that multiplicative effect of each covariate in the hazards function is constant over time (Xue, et al., 2013). Quite often the assumption of PH is substantially important. The standard cohort studies offer a number of ways by which the assessment of the assumption of PH can be approached in addition to the statistical tests and graphical methods. “Conversely, graphical methods involve a moderate degree of subjectivity in interpretation. Statistical tests typically screen for the lack of fit of a Cox model.” (Xue, et al., 2013).

Using Schoenfeld’s residuals

To monitor the capability of befitting of a statistical model is best judged by this useful method of residuals. The method to be presented here is the scaled residual method. The testing of time dependent covariates is the same as the test for a non-zero slope (both are equivalent) on the functioning time of the scaled Schoenfeld residuals in a linear regression. If the assumption of proportional hazard is violated then it is indicated by a non-zero slope.

In a Cox model each predictor variable defines the Schoenfeld residuals. Therefore, both the number of predictor variables and Schoenfeld residual variables are same. They are constructed on the assistance of each of the predictor variable to the log partial likelihood. The diagnosis of Cox regression models is greatly benefitted from the scaled Schoenfeld residuals, particularly in the assessment of the assumption of the proportional hazards (Grambsch and Therneau, 1994). Theoretically, the adjustments in the Schoenfeld residuals are incorporated according to the inverse of the covariance matrix of the Schoenfeld residuals. By using the slope of scaled Schoenfeld residuals against a function of time as our basis for the null hypothesis of the test on proportional hazards, it can apply the results to the generalized linear regression approach (Grambsch and Therneau, 1994).

The basis of the null hypothesis of the experiments in the proportional hazards is the scaled Schoenfeld residuals, which is the slope of the Schoenfeld residuals. This can be against a function of time which is zero for each predictor variables.

According to Schoenfeld the ith residual value can be plotted against t_i to verify the assumption that residuals are not affected by time. Partial residual is defined as the difference between the observed and expected value of X_i in the risk set R_i. More precisely, the test statistic for an individual predictor variable is:

$$r_{ik} = X_{ik} - E(X_{ik}|R_i), \qquad \text{(Fitrianto and Jiin, 2013)}$$

In R program, the function of cox:zph calculates the assumption of the proportional-hazards with respect to covariates as it correlates the transforming time with the respective set of scaled Schoenfeld residuals. As a consequence, the proportional hazard assumption (PHA) does not need to be rejected provided the larger of p- value (>0.05) and its smaller p-value (<0.05) leads to the rejection of PHA.

In R the command is given as follows:

```
Resplot <- cox.zph(coxFit2)
```

Cox- Snell

Cox- Snell residuals are used as their optimum distribution is achieved in the exponential form. As a result when different models with varied distribution are being worked upon, a common method can be applied to all of them. The possible drawback of this method lies in the nonparametric estimation of the baseline hazard function, which might violate the approximation exponentially of the Cox- Snell residuals.

It must be noted that use of Cox-Snell residuals for assessing the accuracy of survival models has not gained wide acceptance particularly when semi parametric Cox- models are used.

The following equation defines the model as:

$$r_i = \widehat{H_0}(T_j) \exp(\widehat{\beta}\grave{X}_\iota) , i = 1, 2, \dots, n$$

where

r_i denotes a censored sample from a unit exponential distribution presuming that the applied Cox model will hold true.

$\widehat{H_0}$ stands for values that lie near the actual value of H_0.

$\hat{\hat{\beta}}$ denotes values that lie near the actual value of β.

Plotting the log cumulative hazard plot of Cox-Snell residual with its best fitted straight line in R:

```
Htilde <- cumsum(coxph.res2$n.event / coxph.res2$n.risk)
plot((coxph.res2$time), (Htilde), type = 's', col = 'blue')
abline(0, 1, col = 'red', lty = 2)
```

Martingale residual

This residual is mainly used to assess how well a Cox model fits a series of observations. The martingale residual can actually quantify and produce a variable that will denote the interrelationship between a continuous predictor and the survival expectation for individuals. This model is examined as the best functional form for the given covariates.

The graph can be interpreted easily by superimposing a smoothed curve. There are many of smoother curves that can fit to scatterplot. One of the common algorithm used is LOESS or LOWESS smooth which is implemented in R package. To calculate the correct functional form of the variable, a Cox model is chosen with excluding the variable and a graph is plotted between the LOESS smooth of the martingale residuals against some change in the parameters of the variable. If the change in the parameters is correct then the graph will show a linear distribution. The equation depicting this is given below:

$$\widehat{M_i}(t) = N_i\,(t) - \int_0^t Y_i\,(s) e^{\hat{\beta}\,\grave{X_i}\,(s)} d\widehat{H_0}(s)$$

Where:

$\widehat{M_i}$is an acronym for $\widehat{M_i}(\infty)$. The residual is defined at each t, as the

$N_i\,(t)$ stands for the number of observed events over a time period t and includes the ith subject.

Y_i is a 0-1 process determining if the ith subject is a risk at time t.

$\hat{\beta}$ which estimate by maximum partial likelihood estimator β

$e^{\hat{\beta}\,\grave{X_i}}$ is expected value for events for each ith subject

$\widehat{H_0}$ denotes cumulative baseline hazard.

At times it is difficult to understand the graphs of martingale residuals as they have a skewed distribution and can have values ranging from (-∞, 1). This is the reason why deviance residuals are considered a better option for assessing model accuracy and pinpointing outliers.

The plot of martingale residual can be done for each variable in R as follows:

```
plot(clinData2$PreDxBxPSA, rr,xlab="PreDxBxPSA",ylab="Residual")
lines(lowess(clinData2$PreDxBxPSA, rr,iter=0),lty=2)
```

Deviance residual

It has been used to work on improperly projected observations. It was introduced by Therneau et al (1990) as a solution to the disadvantages of the martingale residual.

The deviance residual has a more symmetrical distribution about zero and can be defined as:

$$d_i = sign(M_i)[-2\{M_i + \delta_i \log(\delta_i - M_i)\}]^{\frac{1}{2}}$$ (Fitrianto and Jiin, 2013)

Where:

M_i denotes martingale residual, function sign (.) stands for the sign function.

δ_i observed number of events for ith observation with 1 or 0.

Observations that lie on the extremes of the deviance residual are those that are not in accordance with the model parameters and are designated as outliers. In case of minimum censoring approximately < 25% deviance residuals achieve a more symmetrical form as compared to martingale residual and the overall distribution is in a pattern that closely resembles normal distribution.

This is how we can get deviance residuals from R:

```
plot(coxFit2$linear.predictor, dev.res, xlab = 'Risk Score', ylab = 'Deviance residuals')
abline(0,0,lty=2,col='red')
```

3.2.3 Parametric Methods

Accelerated Failure Time Model (AFT):

The type of this model is completely different and the survival time data is evaluated by it. It has been presupposed that the linear function of the covariates serves as the time logarithm. It has been assumed further that the scale of time would affect the change in the survival of the covariates. It has been represented as:

$$\log(T_i) = b_0 + b_1 x_1 + b_{\ 2} x_2 + \cdots + b_{\ p} x_p + \sigma \varepsilon_i$$

Here,

b_0 refers to the intercept and the unknown coefficients for the values of explanatory variables of p are $b_0, b_1, b_2 \ldots b_p$.

σ refers to the scale parameter and the deviation of values of $\log(T_i)$ is modeled implying a random variable- the quantity ε_i.

In the AFT structure, the survival times are normally supposed to maintain a particular pattern of distribution. For presenting survival data of this form, distributions like, log-logistic, log-normal, and Weibull can be applied. The AFT supposition is considered in case of comparison between survival times. On the other hand, PH supposition is used in case of hazard comparisons. According to the AFT process, the impact of the covariates is assumed to work on the log time scale and therefore multiplicatively on the time scale itself (Ponnuraja and Venkatesan, 2010). To make it simple and easy to express, the time ratio or TR, which is actually the exponentiated regression coefficients (i exp (β)), is suggested to express in the same way hazard ratio (HR) is interpreted in the models of proportional hazards. When TR is greater than 1 in case of a particular co-variant, it infers slowing down or extends the time for the event; on the other hand, when TR is less than 1 in case of a particular covariant it depicts higher probability of manifestation of the previous event (Khanal, et al., 2012). In order to offer an in-depth elucidation citing practical examples using the AFT model, AIC, the process of evaluating the rightness of fit of the AFT model is calculated in the following way. The detection plot of the Cox-Snell residuals is employed for evaluation of the overall fit of the AFT model.

Exponential AFT Model:

In survival studies, exponential distribution is considered as the most simple and the most vital form of distribution. Just in the way, normal distribution plays a vital role in different statistical cases, in lifetime studies, exponential distribution takes a major role (Hashemian, 2013). The exponential model can be a special case of Weibull model when $\frac{1}{\sigma} = 1$ and λ is a scale parameter.

The following equation gives the survival function of exponential distribution:

$$S_i(t) = \exp(-\lambda_i t),$$

Weibull AFT Model:

Weibull Distribution is nothing more than a generalized form of exponential distribution, which is extensively used in meteorology for weather prediction modeling and also in radar modeling for predicting the distribution of wind. Weibull Distribution is the only type of distribution that can be presented as accelerated failure time model as well as proportional hazard model. For the purpose of studying survival data, the Weibull distribution is actually borrowed from the field of engineering. At the time of study of the applicability of the distribution in the medical science, it was noted that the distribution model was vital because there is the probability of uniform increase or decrease in the patient mortality rate. Weibull distribution of a survival time T, means the same as of the Gumbell distribution of the survival time; and hence the AFT presentation of the survival function of Weibull model is represented as:

$$S_i(t) = \exp\left(-\lambda_i t^{\frac{1}{\sigma}}\right),$$

Where, $\lambda_i = \exp\left(\frac{-\mu - \acute{x}_i \beta}{\sigma}\right)$, which is a Weibull distribution with shape parameter λ_i and scale parameter $\frac{1}{\sigma}$.

The Weibull AFT model is implemented by the "survreg" function from survfit package in R as follows:

```
weibul12 <- survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData3, dist="weibull")
```

The Cox-Snell residual will take the form

$$\hat{r}_{c,i} = \exp\left(\frac{log(t_i) - \hat{r} - \acute{x}_i \hat{\beta}}{\hat{\sigma}}\right) = \exp(\hat{r}_{c,i})$$

In case of a perfect model they will have unit exponential distribution.

The plot of Cox- Snell was given in R with the command below:

plot(cs.fit$time, -log(cs.fit$surv),type = 's',xlab="Cox-Snell residual",ylab="Cumulative hazard of residual",main="Cox-snell plot for weibull model")

Lognormal AFT Model:

Low value mean, skew distributions, non-negative values and high variance, such as the variety of species are usually based on lognormal distribution. It is often applied for retorting to intoxicating biological elements, hardware repair time distribution, most patterns of survival data, load price study and financial studies. In case the survival times are presumed to be in a lognormal distribution, the baseline survival is presented by,

$$S_i = 1 - \Phi\left(\frac{\log(t) - \mu - \grave{\beta} x_i}{\sigma}\right),$$

$\mu\ and\ \sigma$ are parameters,

$\Phi(x)$ is the cumulative density function of the standard normal distribution.

The log survival time for ith individual has normal ($\mu + \grave{\beta} x_i$, σ)

Logistic AFT Model:

In case the rate of death for studying the survival reduces gradually after achieving the highest within a restricted span, it might be ideal to apply failure rate distribution to find the lifetime. Continuous distribution of random variables with probability and non-negative variable is a log-logistic. This type of distribution is utilized in analysis model of parametric survivals, where at the beginning the rate goes radically up and then starts to decrease.

The survival time even for the ith individual can have a log logistic distribution presented as:

$$S_i = \frac{1}{1+(\lambda t)^{\sigma}}\ ,\ \lambda > 0$$

λ is parameterized in term of predictor variables and regression parameters (shape parameter). σ is scale parameter.

Chapter 4

Results and Discussion

In our study, the subject of analysis was the data from 218 patients with prostate cancer that had the event (death) after radical prostatectomy (RP), as defined by a rise in PSA level.

The result of each method was performed by statistical package in R, which was used to analyze the data. After applying the Kaplan–Meier (K-M) method to the RP data, the results are tabulated in the tables in the following sections. There are some criteria and methods used to validate the model and check the over fitting of the model. Finally, the discussion will clarify the meaning of the results.

4.1 Kaplan-Meier (K-M) Estimation

The construction of a table is a necessary first step in order to analyze the K-M estimate, which requires three elements to function. These elements are serial time (survival time by month), status at serial time (1= death, 0=censored), and group (1 = primary and 2= Met). An Excel spreadsheet was used to build the table, beginning with the shortest times for each group and sorted by ascending serial time, which is shown in Table 4.1 below. The initial table is preparation for K-M analysis to be used by statistical program R.

Table 4.1: Initial sorted table for Kaplan- Meier and Log- Rank analysis

Subject	Group (1= Primary, 2=Met)	Survival time (months)	Status at serial time (1 =event, 0 = censored)
PCA0001	1	18.50	1
PCA0002	1	58.02	0
PCA0009	1	64.76	1
PCA0003	1	93.14	0
PCA0007	1	98.60	0
PCA0005	1	126.10	0
PCA0008	1	149.19	0
PCA0004	1	152.55	0
PCA0006	1	160.96	0
.	.	.	.
.	.	.	.
.	.	.	.
PCA0214	2	0.10	1
PCA0207	2	1.38	1
PCA0206	2	1.61	1
PCA0208	2	11.79	1
PCA0210	2	20.04	1
PCA0213	2	64.66	1
PCA0205	2	NA	0
PCA0209	2	NA	0
PCA0211	2	NA	0
PCA0212	2	NA	0

K-M curve

The plot of survival curves is an important part of survival analysis for each group of interest. However, the comparison between two groups is represented by log rank test. In our study, there are two types of prostate cancer tumors (Primary, and Met) that were compared to get the survival time. From Figure 4.1, the plot of K-M estimate of the survival function plays the role of a step function rather a smooth curve, which is between two times (times at adjacent deaths and the interval only decrease at each death). In the curve in Figure 4.1 the survival duration for the interval is represented by the lengths of the horizontal lines along the X-axis of serial times. Moreover, the cumulative probability of surviving a given time is seen on the Y-axis. In addition, the vertical distances between horizontals are important because they illustrate the change in cumulative probability. When the event of interest occurs, the interval is terminated.

Some subjects are censored (patients did not die during the follow up) and they are shown as vertical bar marks in the graph; these do not terminate the interval. The graph shows the median of survival time and the survival rate.

Presently, we will look at the censored subject as shown in the curve of Figure 4.1. The line of the group 1 curve ends with censored subject as seen in the plot. That provides us with a warning in terms of interpreting anything beyond this point, because the subjects might have the event (death) a few hours later. In contrast, the line of group 2 has no subjects left and the curve drops to zero after the seventeen intervals.

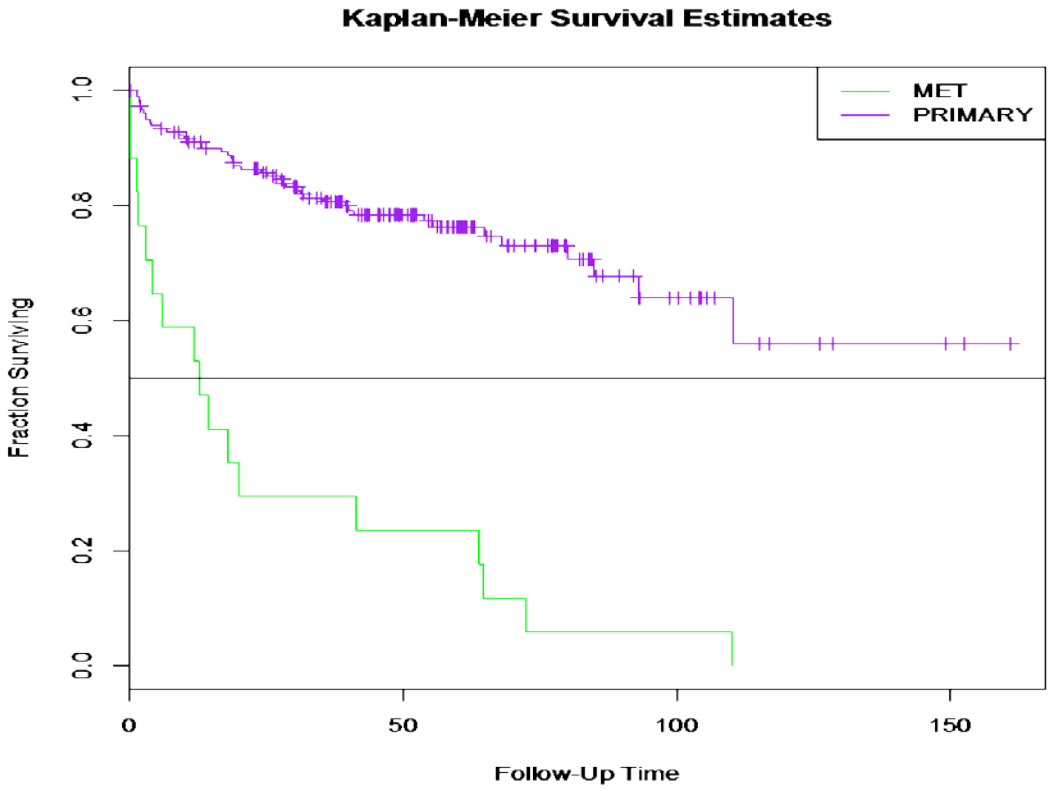


Figure 4.1: Survival curve for two tumor groups for the data in Table 4.1.

In Figure 4.2 *the fun = "cumhaz"* argument in R is used to generate the cumulative hazard curve rather than the survival curve for the participant enrolled in the study described above.

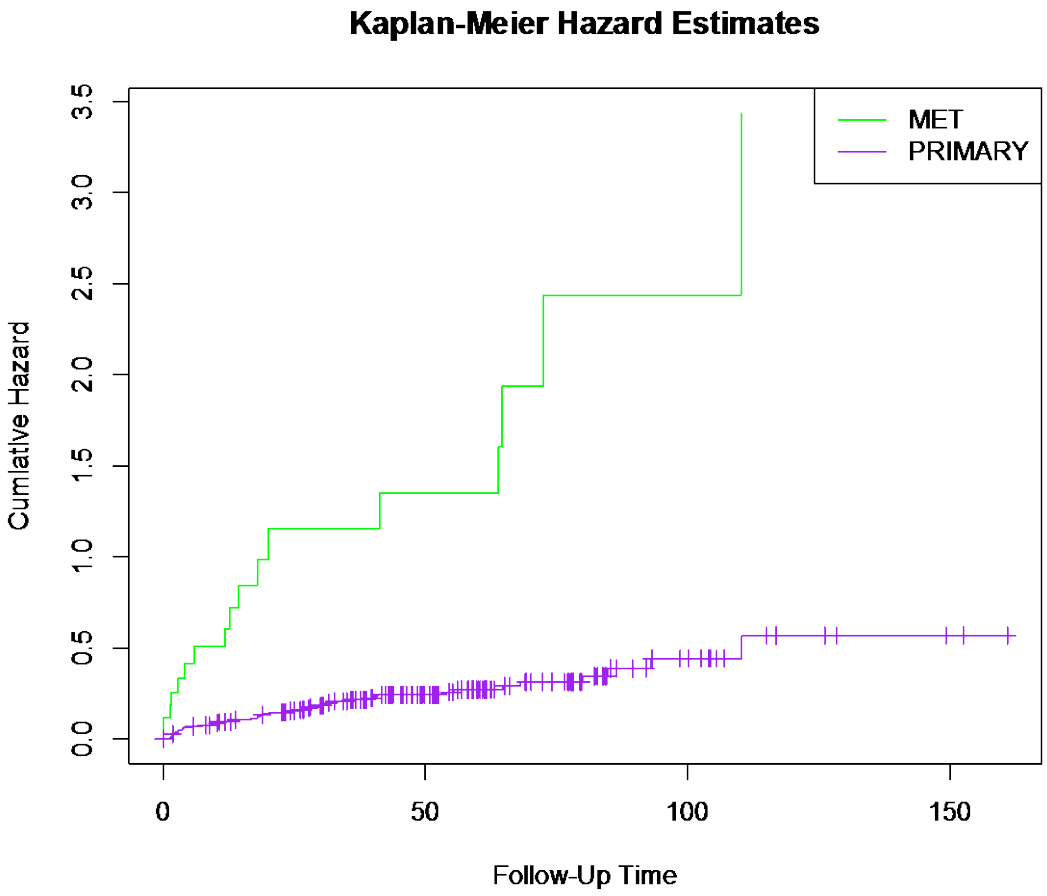


Figure 4.2: Cumulative hazard curve for the prostate cancer data with two types of tumors.

Table 4.2 and Table 4.3 can help explain the curve and the way the curve ends. The number of patients who are alive in group 1 before the serial time of 1.38 is 180 (column 2). Since one patients die at 1.38 (column 3), the probability of dying by a time of 1.38 is 179/180=0.994. The number of patients who are alive in group 2 just before the serial time of 0.10 is 17 (column 2). Since two patients die in 0.10 (column 3), the probability of dying by 0.10 is 15/17=0.8824. The probability for group 1 after 110.33 month is 0.56 (95% CI values from 0.405 - 0.772), while the probability for group 2 after the same time is 0. Moreover, the estimate probability for group 2 after 72 months is 0.0588.

Table 4.2 and Table 4.3 provide the confidence interval of medians for the survival time. The patients with primary tumor have probability of 0.5 to survive longer than 110 months, while the patients with metastatic tumor have probability of 0.5 to survive longer than 12 months. Those

patients with the primary tumor have a better chance of survival.

Table 4.2: Calculation for the K-M estimate of the survival function for primary tumor.

Time	n.risk	n.event	Survival	std.err	Lower 95% CI	Upper 95% CI
1.38	180	1	0.994	0.00554	0.984	1
1.41	179	1	0.989	0.00781	0.974	1
1.64	178	1	0.983	0.00954	0.965	1
1.81	177	1	0.978	0.01099	0.956	1
1.87	176	1	0.972	0.01225	0.949	0.997
2.1	174	1	0.967	0.01339	0.941	0.993
2.56	173	1	0.961	0.01443	0.933	0.99
2.92	172	2	0.95	0.01629	0.918	0.982
3.71	170	1	0.944	0.01712	0.911	0.978
3.94	169	1	0.939	0.01791	0.904	0.974
5.72	168	1	0.933	0.01865	0.897	0.97
6.7	166	1	0.927	0.01937	0.89	0.966
8.97	163	1	0.922	0.02007	0.883	0.962
9.86	162	1	0.916	0.02074	0.876	0.958
10.61	160	1	0.91	0.02138	0.869	0.953
13.04	156	1	0.905	0.02203	0.862	0.949
13.21	155	1	0.899	0.02265	0.855	0.944
16.82	153	1	0.893	0.02325	0.848	0.94
18	152	1	0.887	0.02382	0.841	0.935

18.5	151	1	0.881	0.02438	0.835	0.93
18.83	150	1	0.875	0.02491	0.828	0.925
19.02	148	1	0.869	0.02544	0.821	0.921
20.27	147	1	0.863	0.02594	0.814	0.916
23.92	143	1	0.857	0.02646	0.807	0.911
25.13	140	1	0.851	0.02697	0.8	0.906
27.6	136	1	0.845	0.02748	0.793	0.901
27.86	134	1	0.839	0.02799	0.786	0.895
28.65	132	1	0.832	0.02849	0.778	0.89
30.56	128	1	0.826	0.029	0.771	0.885
31.21	127	1	0.819	0.02949	0.763	0.879
31.8	125	1	0.813	0.02998	0.756	0.874
35.35	121	1	0.806	0.03047	0.748	0.868
39.49	110	1	0.799	0.03107	0.74	0.862
39.95	107	1	0.791	0.03166	0.732	0.856
40.9	106	1	0.784	0.03223	0.723	0.85
53.82	75	1	0.773	0.03345	0.71	0.842
55.39	72	1	0.763	0.03467	0.698	0.834
64.76	48	1	0.747	0.03741	0.677	0.824
68.04	45	1	0.73	0.04009	0.656	0.813
80.03	31	1	0.707	0.04519	0.623	0.801
84.83	24	1	0.677	0.05202	0.582	0.787
92.98	18	1	0.639	0.06124	0.53	0.772
110.33	8	1	0.56	0.09199	0.405	0.772

Table 4.3: Calculation for the K-M estimate of the survival function for Metastatic tumor.

Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
0.1	17	2	0.8824	0.0781	0.74175	1
1.38	15	1	0.8235	0.0925	0.66087	1
1.61	14	1	0.7647	0.1029	0.58746	0.995
2.89	13	1	0.7059	0.1105	0.51936	0.959
4.11	12	1	0.6471	0.1159	0.45548	0.919
5.95	11	1	0.5882	0.1194	0.39521	0.876
11.79	10	1	0.5294	0.1211	0.33818	0.829
12.68	9	1	0.4706	0.1211	0.28423	0.779
14.36	8	1	0.4118	0.1194	0.23329	0.727
18	7	1	0.3529	0.1159	0.18543	0.672
20.04	6	1	0.2941	0.1105	0.14083	0.614
41.4	5	1	0.2353	0.1029	0.09987	0.554
63.9	4	1	0.1765	0.0925	0.0632	0.493
64.66	3	1	0.1176	0.0781	0.032	0.432
72.41	2	1	0.0588	0.0571	0.00879	0.394
110.16	1	1	0	NaN	NA	NA

4.2 Log-Rank Survival Estimates

Table 4.4: Calculation for the log- rank test to compare tumor groups for the data in Table 4.1.

Type	N	O	E	(O-E)^2/E	(O-E)^2/V
MET	17	17	3.26	57.81	61.5
PRIMARY	181	44	57.74	3.27	61.5

The calculation determined that in order to document a significant difference in survival times for patients with primary and/or Met tumors, the p-value must be less than 0.05. From the table of chi-squared we get the p-value is 4.44e-15 (less than 0.05), then we reject the null hypothesis, which is H_0: $S_1(t) = S_2(t)$ because there is difference between the populations in the probability of an event (death) at any time point. In Table 4.4, the total number of expected (E) death for group 1(primary) is calculated as 57.74 and the total number of observed death is 44. In contrast, the total number of expected (E) death for group 2 (Met) is calculated as 3.26 and the total number of observed (O) death is 17. Therefore, the value of statistic (chi-squared) is calculated as follows: 61.5. The degrees of freedom are the number of groups minus one, 2 - 1 = 1. Based on our calculations, we can conclude that there is significant evidence of a difference in survival times for primary and Met.

K-M estimation for subset of primary with the Gleason score

According to the diagnostic factor in prostate cancer we consider the survival time among the most affected factors when using the Gleason score with only the group 1 (primary) subset. Through our analyses, we want to show if there is a significant difference in survival times within primary tumor patients.

The results show a lot of patients have a Gleason score of 6 and 7, and a lot of the subjects are censored as shown in Figure 4.3. Therefore, it gives us a hint that the patients with low Gleason score have better survival. In contrast, the grades of 8 and 9 have the highest rate of experiencing the event (death), because the Gleason score classified to be a high Gleason score, which indicates more aggressive tumors. The table is illustrated in Appendix D which explains the curve and how the curve ends for each grade.

As shown in the figure, there is a big difference between grade 6 and 9. The graph illustrates the p- value of 0, which means there is a significant difference for survival times of primary tumor patients. This gives us a hint that determining survival time using the Gleason score is dependent on the classification within the group.

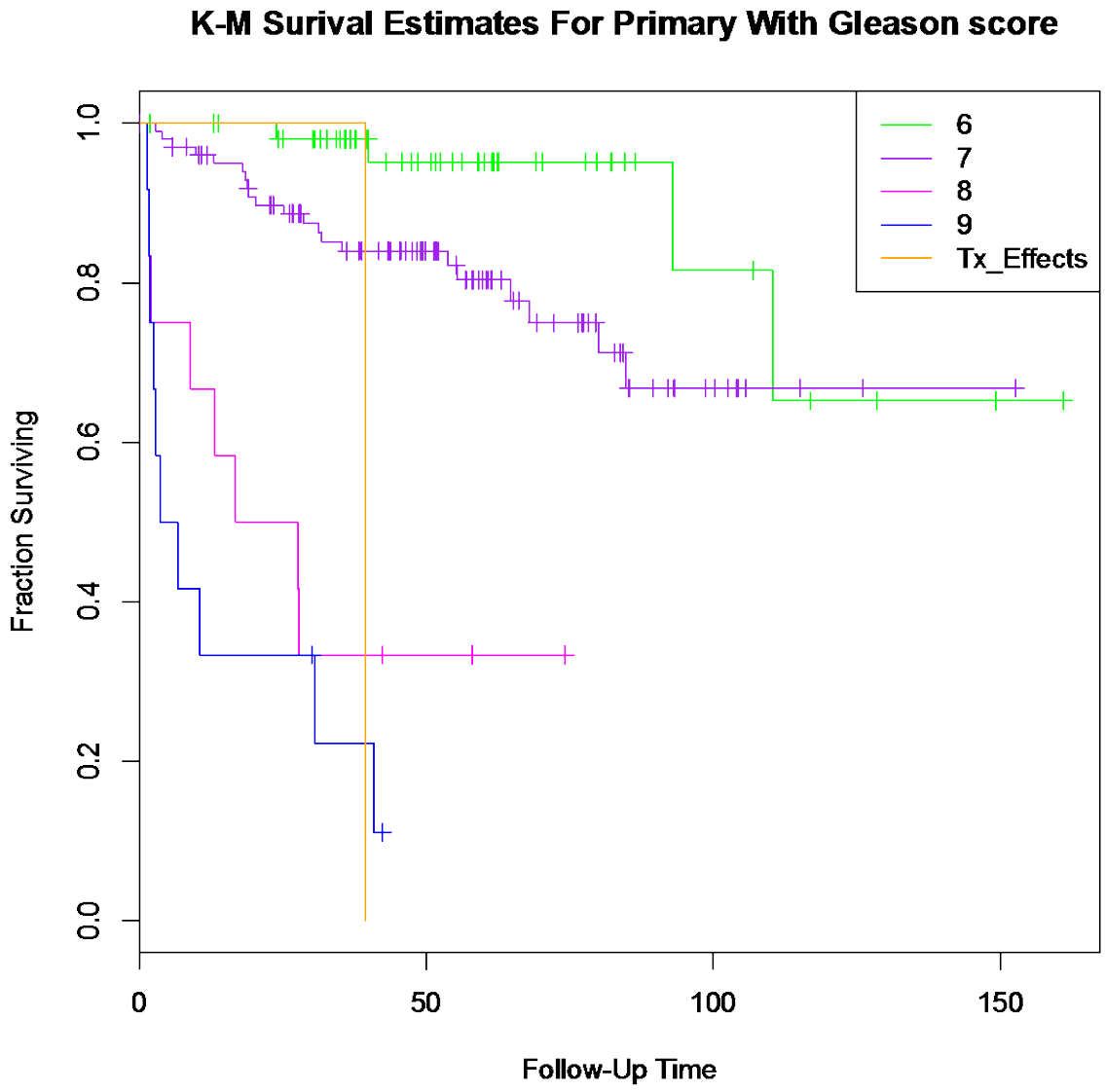


Figure 4.3: Survival times of patients with primary tumor according to Gleason grade.

4.3 Cox Fit Model

As we mentioned before the Cox model yields an equation for the hazard with the explanatory variable. The aim from this model is to show the prognostic factor impact on survival. In our study, Cox model was applied to our data using PSA levels, tumor stage, secondary Gleason grade, and Combined Gleason score as explanatory variables. They are chosen by using the AIC and p-value criteria to fit the model. The output is shown in Table 4.5 and Table 4.6.

The number of value of the cox model was represented as 192 and the number of events was 57. There are 26 observations that were deleted due to missing values.

Table 4.5: Multivariate analysis of prognostic factors for the prostate cancer patient using the Cox PH model.

charactristic	Coef	exp(coef)	se(coef)	z	P
PreDxBxPSA	0.00427	1.004	0.00152	2.81	4.90E-03
PathGGS	1.57581	4.835	0.28147	5.6	2.20E-08
PathGG2	-0.87512	0.417	0.31551	-2.77	5.50E-03
PathStage	0.22182	1.248	0.07861	2.82	4.80E-03

Table 4.6: The hazard rate

charactristic	exp(coef)	exp(-coef)	lower .95	upper .95
PreDxBxPSA	1.004	0.9957	1.0013	1.0073
PathGGS	4.835	0.2068	2.7847	8.3937
PathGG2	0.417	2.3992	0.2246	0.7736
PathStage	1.248	0.8011	1.0701	1.4563

Concordance	= 0.838 (se = 0.041)				
Rsquare	=0.361 (max possible= 0.94)				
Likelihood ratio test	= 86.07	on	4	df,	p=0
Wald test	= 88.1	on	4	df,	p=0
Score (logrank) test	= 121.2	on	4	df,	p=0

In the result, there are two tables: Table 4.5 for the coefficients and Table 4.6 for the hazard rate. In Table 4.6 the second column presents the regression coefficient. The sign of the coefficients is an important issue to consider since a positive sign means the hazard ratio for this variable is higher, while the negative sign will decrease the hazard risk (risk of death). For example in PathGG2 the coefficient is negative so the risk of death will decrease. Column three in Table 4.5 presents the estimate of hazard for instance, exp (-0.87512) = 0. 417, which is a 41% decrease in the risk of the death for patient with PathGG2.The estimation of hazard increases by exp (1. 57581) = 4.835for each grade of PathGGS. The fourth column in Table 4.5 is an approximate test of significance for each variable, and is obtained by dividing the regression estimate coefficient by its standard error SE(coef). The column z in Table 4.5 records the ratio of each regression coefficient to its standard error; a wald statistic is asymptotically standard normal under the hypothesis that the corresponding coefficient is zero. Finally, p-value shows the significance of the explanatory variable. The asymptotically equivalent tests of the omnibus null hypothesis that all of the coefficients are zero are likelihood ratio, wald score, chi-square statistic at bottom of the output. We can conclude that in cox model, if the coefficient is negative the hazard will decrease, but if the coefficient is positive the hazard will increase.

The plot of survival curves based on the cox model and Kaplan-Meier Estimates for the model is presented in the Figure 4.4. The estimated distribution of survival times for cox model is illustrated below by using *survefit* function graph (function to calculate survival time). It is illustrated the estimate survival function

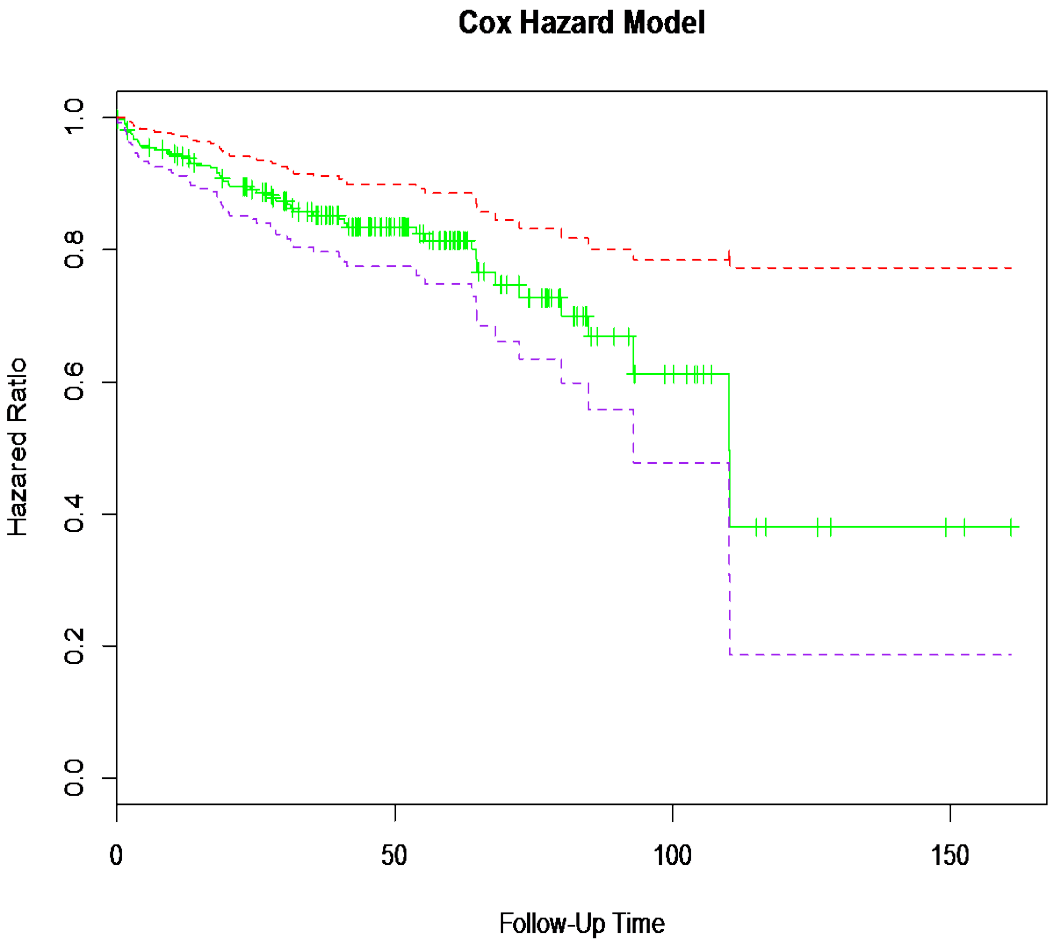


Figure 4.4: The Cox proportional hazard (PH) with error bars show 95% confidence intervals.

The important step after fitting the model for Cox is to evaluate the adequacy of the fitted model. As we mention in Chapter 3 the model that checks the analysis is based on residuals. In the analysis for the cox model four major criteria of residuals have been described, they are the Cox-Snell residual, the deviance residual, martingale residual and the Schoenfeld residual.

4.3.1 Testing the proportional hazards assumption using Schoenfeld's residuals

The approach of Schoenfeld is the global goodness of the fit test for Cox PH models, which are used to detect the insufficiency of covariates in describing the relative risks and the assumption of PH.

From function cox.zph in R (function will test proportionality of all the predictors) we got three columns of computation a test for each covariate as shown in Table 4.7. The column rho indicates the Pearson product-moment correlation between the scaled Schoenfeld's residuals and lagged residuals for each covariate; the column chisq gives the test statistics and the last row GLOBAL gives the global test of proportionality for all the interactions at once as illustrated in Table 4.7. The column p gives the p-values. According to these p-values, there is strong evidence of proportionality as shown by small global test statistics (large p-value). We compare the result of proportional hazards assumption checking by using the graphical and numerical methods.

The Schoenfeld residuals are computed and plotted against the time for each covariate "PreDxBxPSA", "PathGGS", "PathGG2', and "PathStage". The list of the residuals is ordered the same as predictor in the cox model.

Table 4.7: Scaled Schoenfeld Residuals of Significant Covariates on the PH.

	Rho	Chisq	P
PreDxBxPSA	0.06080	0.230375	0.631
PathGGS	-0.00391	0.000644	0.980
PathGG2	-0.07240	0.243562	0.622
PathStage	-0.05141	0.103435	0.748
GLOBAL	NA	1.750937	0.781

As illustrated from the output, it appears that PreDxBxPSA, PathGGS, PathGG2, and PathStage satisfy the PHA. According to Table 4.7, they have a slope which is not significantly different from zero since the p-values are $> \alpha = 0.10$ in which we failed to reject the null hypotheses. It means there is no correlation between each covariates of Cox model and time (not time

dependent), which implies that the proportional hazards assumption is fulfilled. Four Smoothed scaled Schoenfeld residual plots for these predictors provide an interpretation of the proportionality of the model in Figure 4.5.

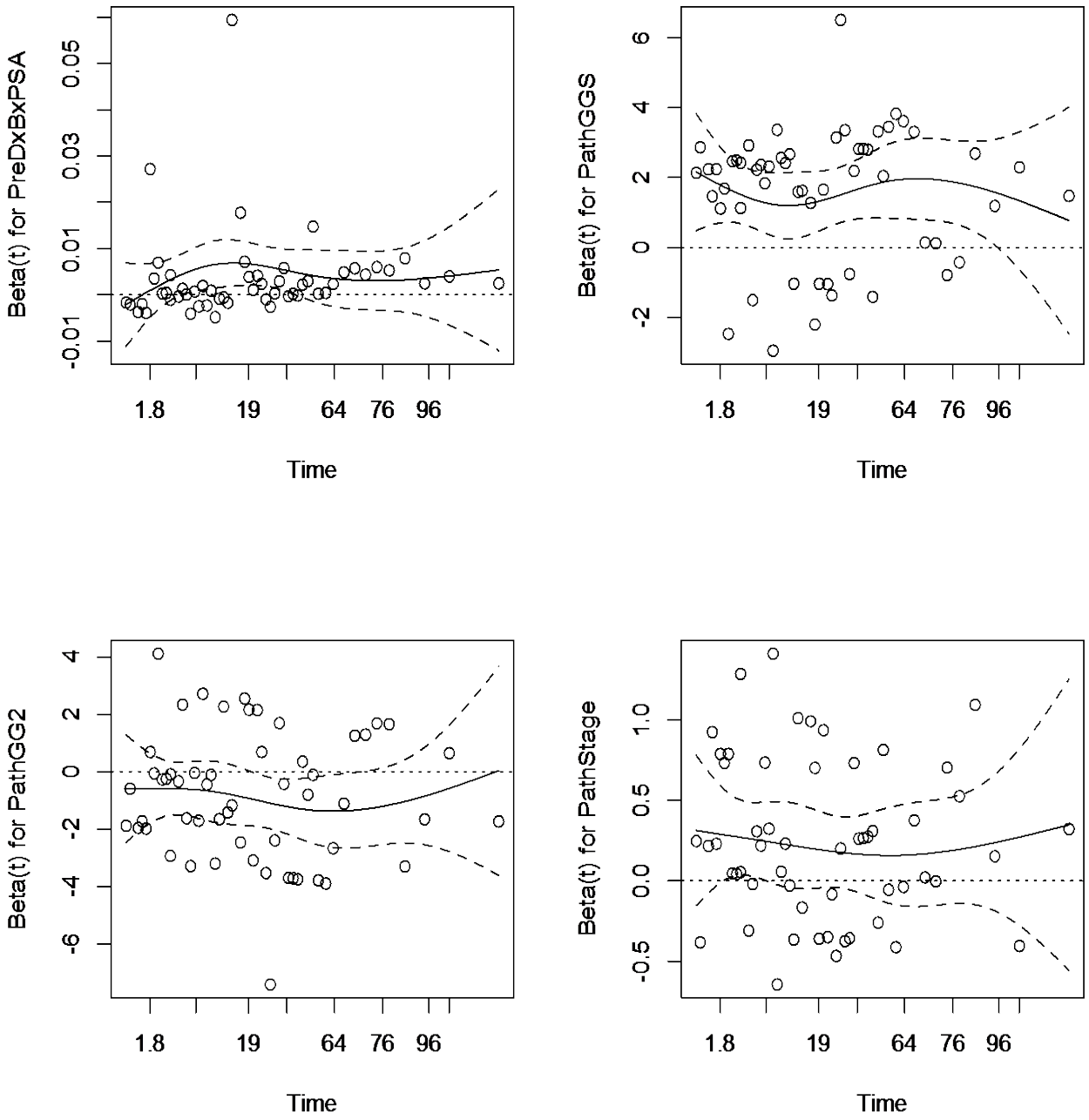


Figure 4.5: Schoenfeld residuals for each explanatory variable versus transformed time in a model fit to the prostate cancer data.

4.3.2 Evaluating overall model fitting

Cox-Snell residual is helpful to evaluate the overall model fitting. When the model does not fit the data well, cox-Snell does not show any indication of the reason. It only shows whether the model is fitted or not. When the step function coincides with the straight line, we say the model fits well.

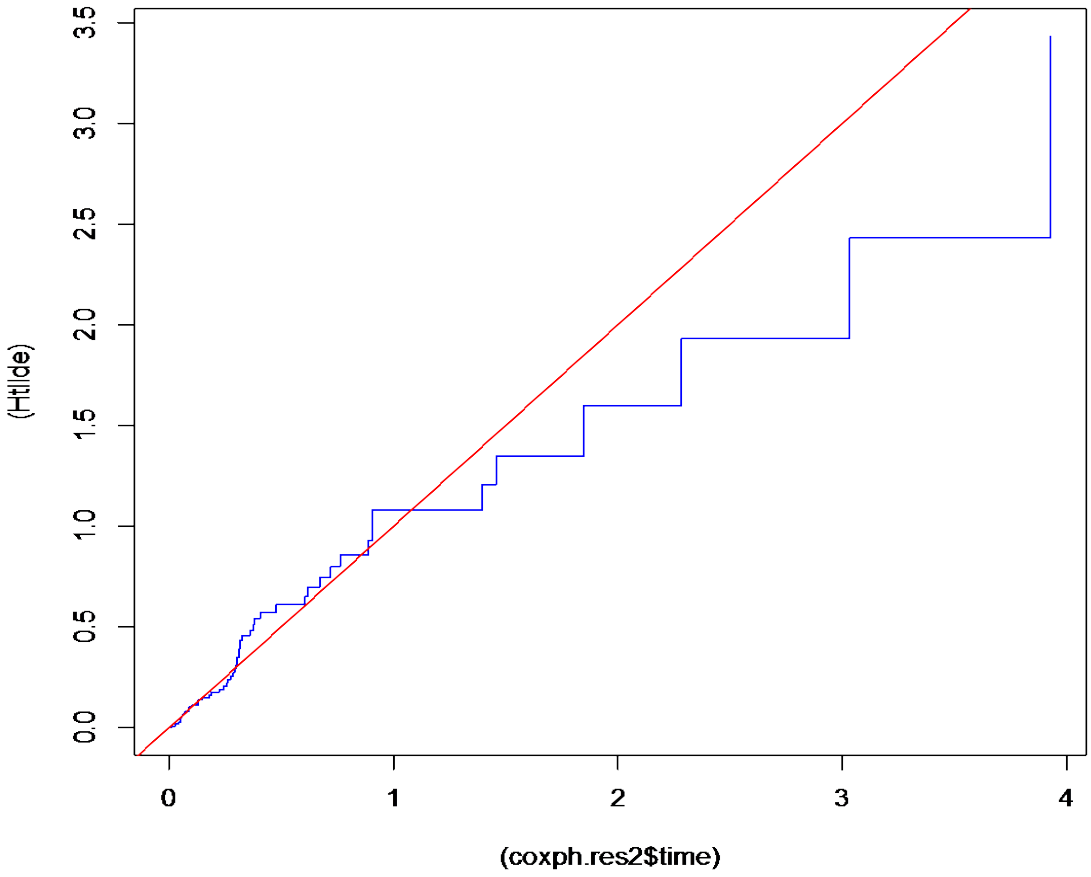


Figure 4.6: Cumulative hazard plot of the Cox-Snell residual for Cox PH model to indicate the overall model.

4.3.3 Functional Form of Predictors

Here we investigate the functional form of the covariates. Martingale residual and covariates were plotted to observe the covariates functional form. When the functional form is linear or near to linear then it would be satisfactory. If the functional form is different from linear, then we have to apply some transformation (eg. log, square, square root etc.) of the covariates.

According to Figure 4.7 below, the martingale residual illustrates a functional form for the covariates "PathStage", "PathGGS" and "PathGG2" that seems close to linear. For "PreDxBxPSA" the functional form seems not linear. Therefore, we have to apply transformation for "PreDxBxPSA". From this functional form of "PreDxBxPSA", log transformation would be more logical to use.

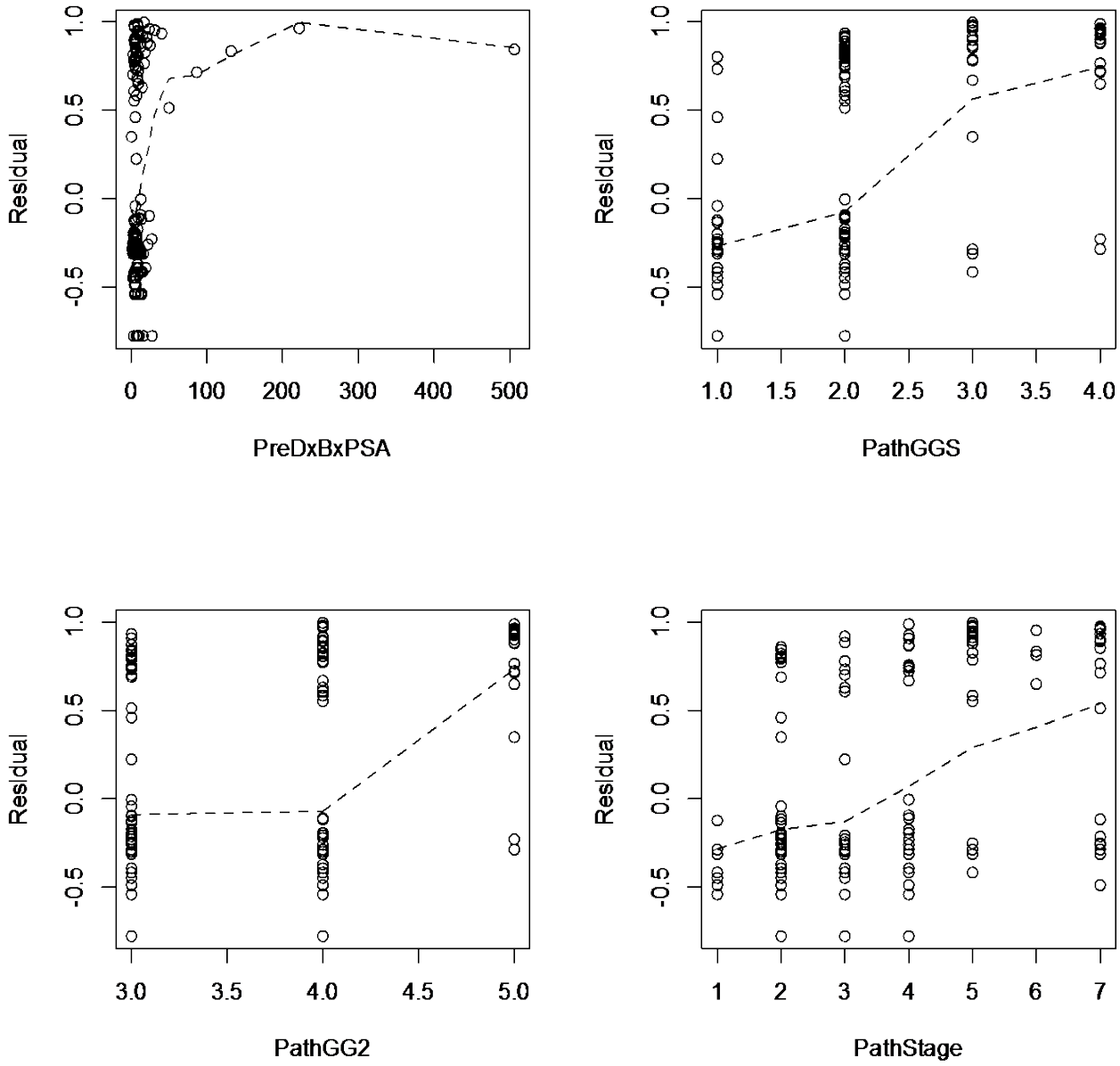


Figure 4.7: Plot of martingale residuals and lowess smoothed vs. covariates.

4.3.4 Checking for Outliers

We have defined before two types of residuals, which are cox-snell and martingale and it helps to obtain the deviance residual.

Deviance residual was used to detect poorly predicted observations, so we can find outliers. The residual deviance shows how well the response is predicted by the model when the predictors are included. Figure 4.8 shows that the pattern plot of deviance residual against the risk score seems to be symmetrically distributed about zero. This plot is a powerful diagnostic to detect individuals whose survival times are out of line.

Table 4.8 is illustrated that deviance residuals can be used to identify outliers.

Table 4.8: Deviance residuals against the risk score

dev	dev.res	BCR_FreeTime	BCR_Event	PreDxBxPSA	PathGGS	PathGG2	PathStage
2	2.643373	0.1	1	17	3	4	5
6	2.120942	14.357596	1	12.6	2	4	2
76	2.301396	3.942589	1	14.9	2	4	3
89	2.252719	23.91838	1	7	1	3	2
159	2.085014	1.412761	1	4	3	4	5

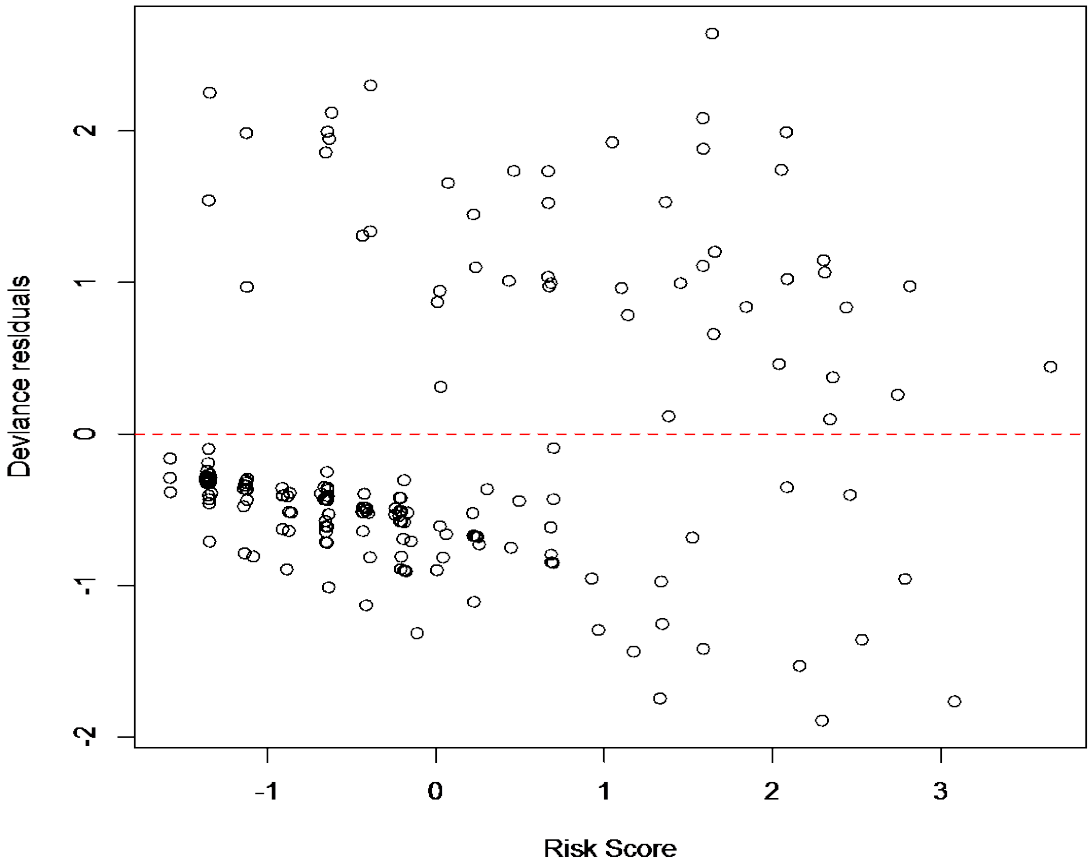


Figure 4.8: Deviance residuals consist of information about the influential and outlier data.

4.4 Output of Accelerated Failure Time (AFT)

These data sets were analyzed using the different AFTs such as exponential, Weibull, log-normal and log-logistic models.

The results from different AFT models applied to prostate cancer progression are presented in Table 4.9.There is no significant difference for the estimations in different models.

Table 4.9: The log-likelihoods and Akaike Information Criterion (AIC) in the AFT models.

Distribution	Loglikelihood	K	C	AIC
Exponential	-256.5	14	1	589.0159
Weibull	-253.5	14	2	585.0454
LogNormal	-256.3	14	2	590.5098
LogLogistic	-254	14	2	586.0415

AFT models were compared by using statistical criteria (Maximum likelihood (ML) test and AIC). According to these criteria, with the AIC (the smaller AIC is better) and higher log-likelihood value. The computed value of AIC for Weibull AFT model is 585.0454. It appears to be an appropriate AFT model compared to the other AFT models as shown in Table 4.10. AIC's are only directly comparable if the number of parameters are the same. AFT model as there are two parameters involved. The AIC for exponential, if it had more parameters, it could be higher than 589.0159.

Table 4.10: Results from AFT models for time to progression with Weibull distribution.

	Value	TR	Std. Error	z	p
(Intercept)	6.19138		2.71474	2.2807	2.26E-02
Type	0.6827	1.979214	0.45513	1.5	1.34E-01
Race	0.09762	1.102544	0.11617	0.8404	4.01E-01
PreDxBxPSA	-0.00442	0.9955898	0.00179	-2.4685	1.36E-02
DxAge	-0.00189	0.99881118	0.02307	-0.0818	9.35E-01
BxGG1	-0.33915	0.7123756	0.3929	-0.8632	3.88E-01
BxGGS	0.24948	1.283358	0.26477	0.9423	2.28E-01
ClinT_Stage	0.08895	1.093026	0.07374	1.2062	2.28E-01
SMS	0.31704	1.373057	0.35166	0.9015	3.67E-01
ECE	0.005	1.005013	0.17329	0.0289	9.77E-01
SVI	-0.18032	0.835003	0.47891	-0.3765	7.07E-01
LNI	0.81869	2.267527	0.37071	2.2085	2.72E-02
PathStage	-0.23156	0.7932951	0.1316	-1.7595	7.85E-02
PathGG1	-0.69585	0.4986504	0.35703	-1.949	5.13E-02
PathGGS	-0.91593	0.4001443	0.22176	-4.1302	3.62E-05
Log(scale)	0.03463	1.035237	0.11168	0.3101	7.56E-01

Scale= 1.04

Weibull distribution

Loglik(model)= -253.5

Loglik(intercept only)= -331.7

Chisq= 103.27 on 14 degrees of freedom, p= 1.1e-15

Number of Newton-Raphson Iterations: 7, n= 190

After selecting the best-fitted parametric model from the AFT family, the performance of the parametric model was compared with the Cox model based on the Cox-Snell residual method. Furthermore, we check the goodness of fit of the model using residual plots. The cumulative hazard plot of the Cox-Snell residuals in Weibull model is presented in Figure 4.9; the plotted points lie on a line that has a unit slope and zero intercept.

According to the plot, there is no reason to doubt the suitability of this fitted Weibull model. We conclude that the Weibull model is the best fitting the AFT model based on AIC criteria and residuals plot.

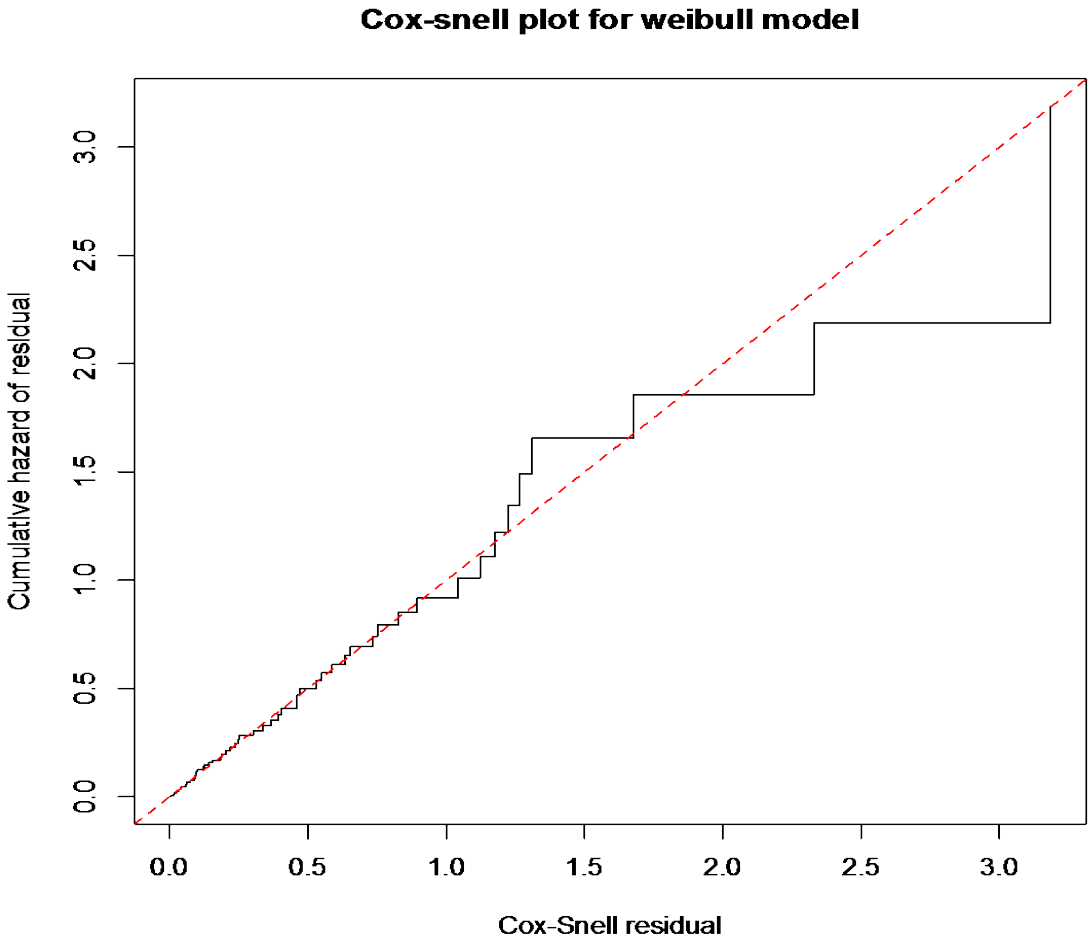


Figure4.9: Cumulative hazard plot of the Cox-Snell residual for Weibull AFT model

We can calculate the acceleration factors and the corresponding confidence interval for every pair of groups manually.

4.5 Discussion

Several statistical models have been suggested for analyzing special type of data, which is referred to as censored data in the survival analysis literature. Non-parametric, semi-parametric, and parametric survival models are mainly used in many clinical trials. These models direct the form of the conditional hazard function for a given set of variables of the survival time.

The Kaplan-Meier method gives very good estimations of survival probabilities. The pattern of this method in assuming on censoring is independent of the survival time as shown in Figure 4.2. Each group of tumors has a pattern independent of the survival time (Langova, 2008). The present study has demonstrated that the patients with a primary tumor have a lesser risk than those with metastatic considering the latter have the spread of cancer cells in the body. Therefore, their survival time will be decreased. The results have provided curve of K-M and the table of the survival time which may be useful in comparing the survival time of each group and checking the censored data. The survival time for almost 10 years is 0 for metastatic tumors while the survival time for those with primary tumors is 0.56 which is evidence of survival. The K-M graph displays the cumulative survival function on a linear scale by tumor (Figure 4.1). The survival curve of primary tumor patients was lower than that of metastatic tumor patients, which means that primary have a higher probability of surviving (not experiencing an event).

Table 4.3 presented the calculations from the log-rank test to show that there is a significant evidence of difference in survival times for groups (primary, and metastatic) since the p-value is less than 0.05. That means there is no significant relation between the survival times of each group of tumors.

The most popular method of examining the effect of explanatory variables on survival is the Cox PH model. This model requires the assumption of proportional hazards between strata formed by the combinations of levels of the different explanatory variables (Mostafa, 2013). Hence, we found the model that only includes the four significant variables, which was chosen with p-value and AIC criteria. Additionally, we can conclude that the Cox model was performed to evaluate

the joint prognostic significance factors. First, we can say the factor and percentage changes in the hazard ratio can be calculated for every significant factor: PathStage, PreDxBxPSA, PathGGS, and PathGG2. These variables are the prognostic factors of prostate cancer after surgical radical prostatectomy, which predict the hazard rate and show the effectiveness in the progress of the disease. Among these prognostic factors are PSA level, secondary grade, Gleason scores, and tumor stage. Moreover, these factors could be used to help determine the treatment strategy.

The results in Table 4.5 showed that "Gleason score: PathGGS" has the highly significant (p-value = 2.20E-08) progression-associated prognosticator. In addition, it gives the hazard ratio of 4.835 which means 83%. Gleason score remains the most powerful prognostic factors for prostate cancer. We can conclude that the important covariates in our study affect the risk is Gleason score as proven through medical research (Buhmeida, et al., 2006). Fijikawa et al. (1997) also achieved this result in their study, explaining that the Gleason score was an essential prognosis factor. In their study, Epstein et al. (2005) revealed that the best prognostic factor in prostate cancer was the Gleason score. Similarly, Yigitbasi et al. (2011) detected that the Gleason score was an essential prognostic factor.

The results of the Gleason score with primary tumors are illustrated in Appendix D. In the group with the lowest Gleason score (6), it was significant that the time was 110.3 months and survival time 0.652. In the group with the highest Gleason score (9), it was 40.9 months and the survival time 0.11, which was also significant. Consequently, the Gleason score at the diagnosis of the primary tumor indicated to be an independent prognostic factor.

"PreDxBxPSA: PSA's level" research has proved that the PSA level is not a very important prognostic factor variable. In our study, it has been found that PSA level had the lowest risk of death, since exp(0.00427) = 1.004. PSA can be an important factor in diagnosing, following and staging prostate cancer. However, in the literature there are different options to determine the PSA as evaluating it as prognostic factor (Yigitbasi et al., 2011). Schubert et al. (1994) mention that PSA can be helpful for monitoring a patient's response when the patients have been given the treatment. Additionally, it is also useful for the follow-up stage after surgical process relapse

and residual tumor. According to Zagars et al. (1995), the PSA is a separate prognostic factor regardless of stage and grade.

For the categorical variable “PathStage: Tumor size stage based on pathologic examination of the radical prostatectomy”, it had an affect on the hazard rate by 1.248, which meant a 24% increase in the hazard rate. While the last variable” PathGG2: Secondary Gleason grade in the radical prostatectomy specimen” decreases the hazard rate with 41%, because it is not one of prognostic factors.

After we chose the best model for Cox, the assumption of PHs were checked with different criteria of residuals. These criteria illustrated the usefulness of goodness-of-fit test and offered a number of established approaches in determining the validity of a fitted Cox PH model. They are the Cox-Snell residual, the deviance residual, martingale residual and the Schoenfeld residual. The scaled Schoenfeld residuals were used to check the PH assumption. Martingale residuals and deviance residuals were considered for influential observations in models and checking outliers.

Afterwards the AIC criterion was applied to determine the best model of AFT with four distributions. After consideration, the Weibull was determined to be the most appropriate model. Table 4.10 presents the Weibull for AFT model and has in column 3 the TR, which is important to interpret the results as HR reported in proportional hazards models. As we discussed in chapter 3, TR> 1 for the covariates reveals that this slows down the time to the event such as Type, Race, and BxGGS, while the TR<1 indicates that an earlier event is more likely such as PreDxBxPSA, and PathStage.

Chapter 5

Conclusion

The literature of survival analysis includes postulations of bunch of statistical models in order to analyze the censored data in presence of covariates. Different types of models like K-M, Cox PH model, and AFT model which are categorized under Non-parametric, semi-parametric, or parametric survival models respectively are generally used for agricultural, clinical, or biomedical purposes. A study on survival analysis has been done on the prostate cancer patients and presented in this thesis. The data for our analysis was taken from Memorial Sloan Kettering Cancer Center (MSKCC), especially taking the samples from the patients under the treatment of radical prostatectomy.

Purpose of this study is to find and approximate the survival function and median time of the primary as well as metastatic tumors of the prostate cancer. For achieving this goal, K-M method has been applied. In comparison with the growth of primary tumors, metastatic tumors are found to grow with double chances of spreading the cancer. In order to differentiate the survival curves, and log- rank test was applied. Results show the difference in survival rates between the patients having either group of tumor growth and these different survival rates were found as significant with the p-value of 4.44e-15.

Depending upon the diagnostic factors of prostate cancer, we are supposed to consider the life span among Gleason score along with the primary subset group 1. Additionally, we have shown that significant difference exists between the survival times of the patients diagnosed with primary tumor growth and having Gleason score. A large numbers of patients have a Gleason score 7, and lots of subjects are in censored condition. We conclude that they have lesser risk because the survival rate is better than other grade. Those who are identified with score 8 and 9 as well, are prone to have higher risk of developing the tumor in comparison with the others. According to the results, significant survival difference has been found for the patients identified with primary tumors and Gleason score between 6 and 9 and the p value is 0.

Function of conditional hazard of the survival time for a certain source of covariates is indicated by the Cox PH models. After justifying the explanatory variables, survival curves could be determined. Several literatures have been investigated for the prognostic factors, which is used to determine the treatment. Additionally, the influences of the standard clinical and pathological prognostic factors over the incidence of hazardous prostate cancer patients have been illustrated through the results. PSA levels, secondary Gleason grade, Gleason score, tumor stage are the considered factors in this case. The Gleason score is identified with higher significant progression-related prognosticators revealing the effective mean towards the cases of demise in the prostate cancer [HR 4.835, 95% CI 2.7847- 8.3937, p=2.20E-08]. In order to judge the goodness of the fit among all of the models of candidate, some specific features of residuals were applied.

AFT model may have a chance of providing a substitute method for fitting few survival data. The effective factor to the clinicians is time ratio because it is quite easier for them to track, interpret, and more importantly this is very significant indeed. In order to fit the data, we applied four diverse models to the dataset, such as Weibull AFT model, log-normal AFT model, log-logistic AFT model, and exponential AFT model. Among all the models, Weibull AFT model fits better and describes the data best. Moreover using the residual plots we also check with the goodness of fit. We mainly used Cox-Snell residuals' cumulative hazard plot in case of the Weibull model. Thus we conclude with the fact that the best fitting model is Weibull model from the aspect of AIC criteria in addition with residuals plot.

Future work

To improve the Cox PH model and the accelerated failure models, we could increase the number of attributed variables that are significant predictors of survival time such as some relevant risk factor, family history, smoking status or race of prostate cancer. These would help to understand the characteristics of health behaviors associated with survivorship for prostate cancer patients.

The other future work could include Markov analysis to examine the progression of prostate patients who took different treatments.

References

[1] Aaserud, S. (2011). Residuals and Functional Form in Accelerated Life Regression Models.

[2] Amaro, A., Esposito, A. I., Gallina, A., Nees, M., Angelini, G., Albini, A., & Pfeffer, U. (2014). Validation of proposed prostate cancer biomarkers with gene expression data: a long road to travel. *Cancer and Metastasis Reviews*, 1-15.

[3] Balogun, O. S., Role, M. R., & Dawodu, O. O. (2014). Survival Analysis of Prostate Cancer In Ilorin, Kwara State. *Survival*, *4*(06).

[4] Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *CRITICAL CARE-LONDON-*, *8*, 389-394.

[5] Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, *89*(3), 431.

[6] Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis Part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. *British journal of cancer*, *89*(4), 605.

[7] Buhmeida, A., Pyrhonen, S., Laato, M., & Collan, Y. (2006). Prognostic factors in prostate cancer. *Diagn Pathol*, *1*(4), 124.

[8] Chan, Y. M. (2013). *Statistical Analysis and Modeling of Prostate Cancer* (Doctoral dissertation, University of South Florida).

[9] Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, *89*(2), 232.

[10] Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part IV: further concepts and methods in survival analysis. *British journal of cancer*, *89*(5), 781.

[11] Dakhil, N. K., Al–mayali, Y. M., & Al–A'bidy, M. A. (2012). Analysis of Breast Cancer Data using Kaplan–Meier Survival Analysis. */Journal of Kufa for Mathematics and Computer*, *1*(6).

[12] Fitrianto, A., & Jiin, R. L. T. (2013). Several Types of Residuals in Cox Regression Model: An Empirical Study. *International Journal of Mathematical Analysis*, *7*(53), 2645-2654.

[13] Fox, J. (2002). Cox proportional-hazards regression for survival data. *See Also*.

[14] Hashemian, A. H., Beiranvand, B., Rezaei, M., & Reissi, D. (2013). A Comparison Between Cox Regression and Parametric Methods in Analyzing Kidney Transplant Survival. *World Applied Sciences Journal*, *26*(4), 502-507.

[15] Humphrey, P. A. (2004). Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology*, *17*(3), 292-306.

[16] Khanal, S. P., Sreenivas, V., & Acharya, S. K. (2012). Accelerated Failure Time Models: An Application in the Survival of Acute Liver Failure Patients in India.

[17] Langova, K. (2008). Survival analysis for clinical studies. *Biomedical Papers*, *152*(2), 303-307.

[18] Liang, H., & Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Computational statistics & data analysis*, *52*(5), 2538-2548.

[19] Litwin, M. S., Steinberg, M., Malin, J., Naitoh, J., & McGuigan, K. A. (2000). *Prostate cancer patient outcomes and choice of providers: development of an infrastructure for quality assessment* (No. MR-1227-BF). RAND CORP SANTA MONICA CA.

[20] Mazerolle, M. J. (2006). Improving data analysis in herpetology: using Akaike's Information Criterion (AIC) to assess the strength of biological hypotheses. *Amphibia Reptilia*, *27*(2), 169-180.

[21] Mostafa, A. A. (2013). Kaplan-Meier and Cox Proportional Hazards Survival Regression Analysis Illustrated with Immune Deficiency Cohort of Patients after Allogeneic Bone Marrow Transplantation. *International Journal of Advanced Computing, ISSN*, *2051*(0845), 1348-1365.

[22] *National Cancer Institute*; <http://www.cancer.gov/cancertopics/factsheet/detection/tumor-grade>.

[23] Penn State Hershey Medical Center. <http://pennstatehershey.adam.com/content.aspx?productId=10&pid=10&gid=000033>.

[24] Ponnuraja, C., & Venkatesan, P. (2010). Survival models for exploring tuberculosis clinical trial data-an empirical comparison. *Indian Journal of Science and Technology*, *3*(7), 755-758.

[25] Pocock, S. J., Clayton, T. C., & Altman, D. G. (2002). Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet*, *359*(9318), 1686-1689.

[26] Pulte, D., Redaniel, M. T., Brenner, H., & Jeffreys, M. (2012). Changes in survival by ethnicity of patients with cancer between 1992–1996 and 2002–2006: is the discrepancy decreasing?. *Annals of oncology*, mds023.

[27] Prostate Cancer Canada; <*http://www.prostatecancer.ca/Prostate-Cancer/Testing-and-Diagnosis/Grading#.VMhNrsYzDV0* >.

[28] Qi, J. (2009). *Comparison of proportional hazards and accelerated failure time models* (Doctoral dissertation, University of Saskatchewan).

[29] Ray, M. E., Bae, K., Hussain, M. H., Hanks, G. E., Shipley, W. U., & Sandler, H. M. (2009). Potential surrogate endpoints for prostate cancer survival: analysis of a phase III randomized trial. *Journal of the National Cancer institute*.

[30] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology-Head and Neck Surgery*, *143*(3), 331-336.

[31] Russell, P. J., Jackson, P., & Kingsley, E. A. (Eds.). (2003). *Prostate cancer methods and protocols* (Vol. 81). Humana Press.

[32] Sewalem, A. (2012). *Semiparametric Analysis of Survival Data with Applications in Agricultural Science* (Doctoral dissertation).

[33] singh, R., & Mukhopadhyay, K. (2011). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, *2*(4), 145.

[34] Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13-21.

[35] Stangelberger, A., Waldert, M., & Djavan, B. (2008). Prostate cancer in elderly men. *Reviews in urology*, *10*(2), 111.

[36] Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., ... & Gerald, W. L. (2010). Integrative genomic profiling of human prostate cancer. *Cancer cell*, *18*(1), 11-22.

[37] Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147-160.

[38] University of Mariland medical center. <http://umm.edu/health/medical/reports/articles/prostate-cancer#ixzz3KHgF4STn>.

[39] Vinh-Hung, V., Burzykowski, T., Van de Steene, J., Storme, G., & Soete, G. (2002). Post-surgery radiation in early breast cancer: survival analysis of registry data. *Radiotherapy and oncology*, *64*(3), 281-290.

[40] Walters, S. J. (1999). *What is a Cox model?*. Hayward Medical Communications.

[41] Xue, X., Xie, X., Gunter, M., Rohan, T., Wassertheil-Smoller, S., Ho, G. Y., ... & Strickler, H. D. (2013). Testing the proportional hazards assumption in case-cohort analysis. *BMC medical research methodology*, *13*(1), 88.

[42]Yang, J. S., Nam, H. J., Seo, M., Han, S. K., Choi, Y., Nam, H. G., ... & Kim, S. (2011). OASIS: online application for the survival analysis of lifespan assays performed in aging research. *PloS one*, *6*(8), e23525.

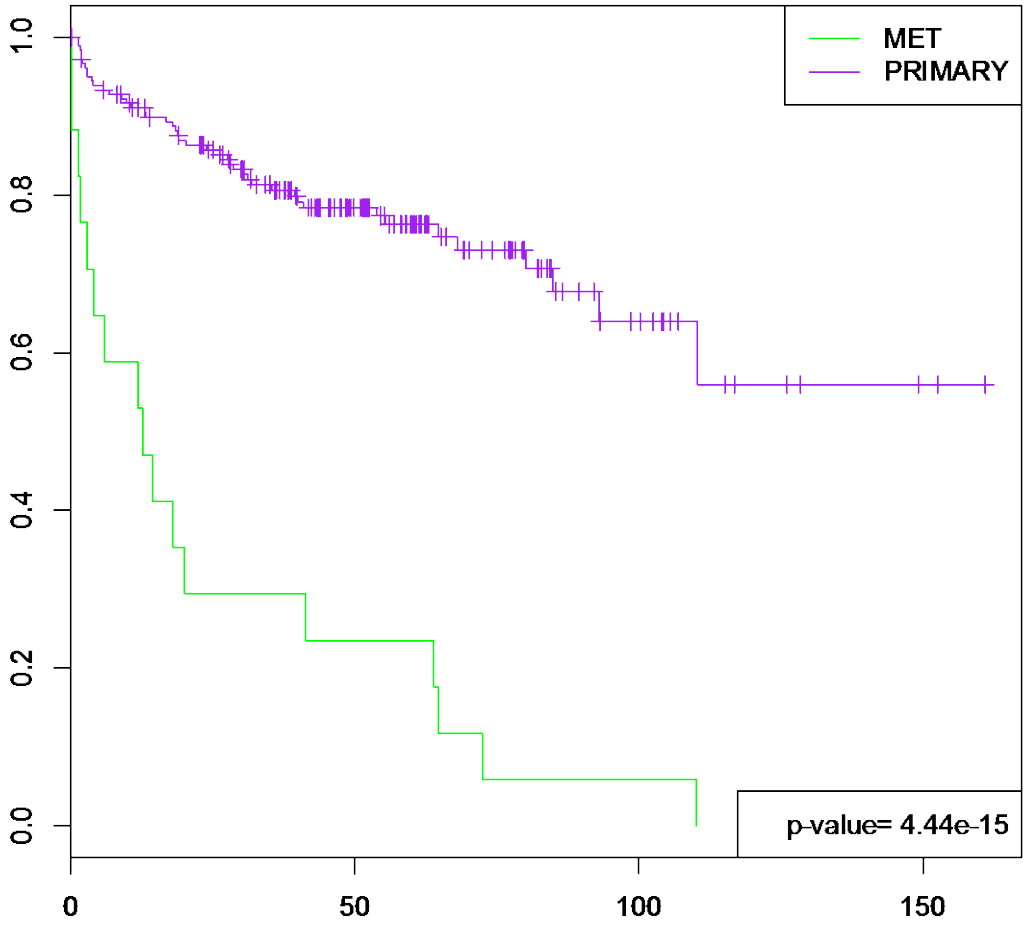
[43] Yigitbasi, O., Ozturk, U., Goktug, H. G., Gucuk, A., & Bakirtas, H. (2011, April). Prognostic factors in metastatic prostate cancer. In *Urologic oncology: seminars and original investigations* (Vol. 29, No. 2, pp. 162-165). Elsevier.

[43] Zhao, G. (2008). *Nonparametric and parametric survival analysis of censored data with possible violation of method assumptions*. ProQuest.

Appendix A

```
# The aim of this code is survival analysis
# The goal of this is to learn to run Survival analysis using high resolution data from prostate
 cancer study (Taylor et al, Cell 2010).
# The goal is to understand how to apply the clinical data and learn basic R functions for survival
 analysis.

# Change to my documents where the file is stored
setwd("/Users/emanalhasawi/Documents/survivalAnalysis")


######### Load the package         #############

# The R function require (survival) accomplishes the same

library(survival)
#read the functions descriptions:
#Read about the syntax of the Kaplan-Meyer estimator function:
?survfit
#Read about the syntax of the log-rank test
?survdiff

######### read the clinical data in#############

clinData<-read.table("TaylorClinicalData_for_CNA_data.txt", header = TRUE, sep = "\t",
 na.strings = "NA", quote = "", comment.char = "")

######### Run Kaplan-Meyer analysis #############

#fit and plot Kaplan-Meyer curves for Primary vs Mets:
pFit <- survfit(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData)

plot(pFit, xlab="Follow-Up Time", ylab="Fraction Surviving",
+ main="Kaplan-Meier Survival Estimates",col = c("green", "purple"))
> legend("topright", levels(clinData$Type), lty = 1, col = c("green", "purple"))
> abline(a=.5, b=0)
pFit <- survfit(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData, type='fleming'
plot(pFit, fun="cumhaz", xlab="Follow-Up Time", ylab="Cumlative Hazard",
+ main="Kaplan-Meier Hazard Estimates",col = c("green", "purple"))
> legend("topright", levels(clinData$Type), lty = 1, col = c("green", "purple"))
```

```
######### Run Log-rank test #############
# Find and plot log- ranks test for the two groups of the tumors

survdiff(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData)

#define a custom function to pull the p-value out of the Log-rank test
getPval <- function(x) {
   if( is.matrix(x$obs))
      etmp <- apply(x$exp, 1, sum)
   else
      etmp <- x$exp
   df<- (sum(1 * (etmp > 0))) - 1
   pv <- 1 - pchisq(x$chisq, df)
   format(pv, digits = 3)
}
pValue<-getPval(survdiff(Surv(BCR_FreeTime, BCR_Event) ~ Type, data = clinData))
> pValue
[1] "4.44e-15"
legend("bottomright", paste("p-value=", pValue), col = "black")
```



######### Subset the data to look at primary tumors only. Plot survival according to the tumour grade ##############

```
# now we want to look at  subset of primary tumor with only Gleason score in order to get the survival time and the difference between the groups of them
#the factor has to be rebuilt since the column still knows about the #"NA" values;
#Use columns "Type" and "PathGGS" - Pathological Gleason Grade Score.
subsetData<-subset(clinData, Type=="PRIMARY")
subsetData$Type<-factor(subsetData$Type)
```

```
pFit <- survfit(Surv(BCR_FreeTime, BCR_Event) ~ PathGGS, data = subsetData)
plot(pFit, , xlab= "Follow-Up Time", ylab= "Fraction Surviving", main=" K-M Surival Estimates
 For Primary With Gleason score", col = c("green", "purple", "magenta", "blue", "orange"))
legend("topright", levels(subsetData$PathGGS), lty = 1, col = c("green", "purple", "magenta",
 "blue", "orange"))
pValue<-getPval(survdiff(Surv(BCR_FreeTime, BCR_Event) ~ PathGGS, data = subsetData))
legend("bottomright", paste("p-value=", pValue), col = "black")
```

Appendix B

```
######### Run Cox proportional hazard #############
# Before fit the Cox proportional hazard, the variables should convert to numeric:
clinData $PathStage <- as.numeric(clinData $PathStage)
clinData $PathGGS <- as.numeric(clinData $ PathGGS)
clinData $ PathGG2 <- as.numeric(clinData $ PathGG2)

# Calculate AIC value for each model with the add new variable, and the “best” model is the one
with minimum AIC value
coxAIC <- extractAIC(coxFit)
##### Fit Cox proportional hazard and plot curves for the covariates:
# combines vector by columns
clinData2<-
 cbind(clinData$BCR_FreeTime,clinData$BCR_Event,clinData$PreDxBxPSA,clinData$PathG
 GS,clinData$PathGG2,clinData$PathStage)
# This function creates data frames, which used for storing data table
clinData2<-data.frame(clinData2 )

colnames(clinData2)<-c("BCR_FreeTime","BCR_Event"
 ,"PreDxBxPSA","PathGGS","PathGG2","PathStage")
##remove the cases with missing clinical variables
clinData2<-na.omit(clinData2)

#dim(clinData2)

#head(clinData2)
# Run coxph to fits a Cox proportional hazards regression model
coxFit2<-coxph(formula = Surv(BCR_FreeTime, BCR_Event) ~ PreDxBxPSA +

+     PathGGS + PathGG2+PathStage , data = clinData2, ties = 'breslow')
```

```
plot(survfit(coxFit2, xlab ="Follow-Up Time", ylab="Hazared Ratio", col = c("green", "purple",
 "red"),main= "Cox PH model") , col = c("green", "purple", "red"),main= "Cox PH model")

summary(coxFit2)


####Evaluating proportionality assumption using Schoenfeld residuals for lactation against
 transformed time for each covariate in a model fit

resplot<-cox.zph(coxFit2)

resplot

par(mfrow=c(2,2))

#plot(cox.zph(coxFit2))

plot(resplot[1])
abline(h=0, lty=3)
plot(resplot[2])
abline(h=0, lty=3)
plot(resplot[3])
abline(h=0, lty=3)
plot(resplot[4])
abline(h=0, lty=3)

###Evaluating overall model fitting
/*plotting the log cumulative hazard plot of Cox-Snell
residual with it's best fitted straight line*/
#The default residuals of coxph in R are the martingale residuals.

cox.snell <- clinData2$BCR_Event - resid(coxFit2,type = "martingale")
coxph.res2 <- survfit(Surv(cox.snell, clinData2$BCR_Event) ~ 1)
#summary(coxph.res2)

Htilde <- cumsum(coxph.res2$n.event / coxph.res2$n.risk)
plot((coxph.res2$time), (Htilde), type = 's', col = 'blue')
abline(0, 1, col = 'red', lty = 2)
par(mfrow=c(1,1))
```

```
#####Functional Form of Predictors

## This could be used to determine the functional form of a covariate

clinData2$PreDxBxPSA<-as.numeric(clinData2$PreDxBxPSA)
coxFit3<-coxph(formula = Surv(BCR_FreeTime, BCR_Event) ~1,data=clinData2)
rr<-resid(coxFit3)



#martingle residual
par(mfrow=c(2,2))
plot(clinData2$PreDxBxPSA, rr,xlab="PreDxBxPSA",ylab="Residual")
lines(lowess(clinData2$PreDxBxPSA, rr,iter=0),lty=2)
plot(clinData2$PathGGS, rr, xlab=" PathGGS",ylab="Residual")

lines(lowess(clinData2$ PathGGS, rr,iter=0),lty=2)

plot(clinData2$PathGG2, rr, xlab="PathGG2",ylab="Residual")

lines(lowess(clinData2$PathGG2, rr,iter=0),lty=2)

plot(clinData2$PathStage, rr, xlab="PathStage",ylab="Residual")

lines(lowess(clinData2$PathStage, rr,iter=0),lty=2)


#Checking for Outliers by 'Deviance residual

dev.res <- resid(coxFit2, type = "deviance")

#length(dev.res)

#length(clinData2$BCR_FreeTime)

plot(coxFit2$linear.predictor, dev.res, xlab = 'Risk Score', ylab = 'Deviance residuals')

abline(0,0,lty=2,col='red')

cbind(dev.res, clinData2)[abs(dev.res) > 2, ]
```

Appendix C

```
######### Run AFT   model with specific distribution #############

# Apply AFT with four distributions
clinData<-read.table("TaylorClinicalData_for_CNA_data.txt", header = TRUE, sep = "\t", quote
= "", comment.char ="")

clinData3<-
cbind(clinData$BCR_FreeTime,clinData$BCR_Event,clinData$Type,clinData$Race,clinData$Pr
eDxBxPSA,clinData$DxAge,clinData$BxGG1,clinData$BxGGS,clinData$ClinT_Stage,clinData
$SMS,clinData$ECE,clinData$SVI,clinData$LNI,clinData$PathStage,clinData$PathGG1,clinDa
ta$PathGGS)
clinData3<-data.frame(clinData3)

colnames(clinData3)<
c("BCR_FreeTime","BCR_Event","Type","Race","PreDxBxPSA","DxAge","BxGG1","BxGGS"
,"ClinT_Stage","SMS","ECE", "SVI","LNI","PathStage","PathGG1","PathGGS")
#str(clinData3)
##remove the cases with missing clinical variables
clinData3<-na.omit(clinData3)

exponenl12= survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData, dist="exponential")
> aicexpo= extractAIC(exponenl12)

> logaic12= survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData, dist="loglogistic")
> aicloglog= extractAIC(logaic12)

> lognorm12= survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData, dist="lognormal")
> aiclognorm= extractAIC(lognorm12)

weibul12= survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData, dist="weibull")
```

- aicweible12= extractAIC(weibul12)

```
weibul12= survreg(formula = Surv(BCR_FreeTime, BCR_Event) ~ Type+ Race+
PreDxBxPSA+ DxAge+BxGG1+ BxGGS+ ClinT_Stage+ SMS+ ECE+ SVI+ LNI+ PathStage+
PathGG1+ PathGGS, data = clinData3, dist="weibull")
summary(weibul12)

hat.sig = weibul12$scale
hat.alpha = 1/hat.sig
reg.linear = weibul12$linear.predictor
reg.linear.mdf = -reg.linear/hat.sig
tt=cbind(Surv(clinData3$BCR_FreeTime, clinData3$BCR_Event))[,1]
cs.resid = exp(reg.linear.mdf)*tt^(hat.alpha)

cs.fit = survfit(Surv(cs.resid,clinData3$BCR_Event)~1,type="fleming-harrington")
#summary(cs.fit)
par(mfrow=c(1,1))
plot(cs.fit$time, -log(cs.fit$surv),type = 's',xlab="Cox-Snell residual",ylab="Cumulative hazard
of residual",main="Cox-snell plot for weibull model")
abline(0, 1, col = 'red', lty = 2)
```

Appendix D

PathGGS	records	n.max	n.start	Events	median	0.95LCL	0.95UCL
PathGGS=6	53	53	53	4	NA	110.33	NA
PathGGS=7	102	102	102	21	NA	NA	NA
PathGGS=8	12	12	12	8	22.21	8.97	NA
PathGGS=9	12	12	12	10	5.21	2.56	NA
PathGGS=Tx		1	1	1	39.49	NA	NA

PathGGS=6

time	n.risk	n.event	survival	std.err	0.95LCL	0.95UCL
23.9	50	1	0.98	0.0198	0.942	1
40	34	1	0.951	0.0343	0.886	1
93	7	1	0.815	0.1292	0.598	1
110.3	5	1	0.652	0.1788	0.381	1

PathGGS=7

time	n.risk	n.event	survival	std.err	0.95LCL	0.95UCL
2.92	101	1	0.99	0.00985	0.971	1
3.94	100	1	0.98	0.01386	0.953	1
5.72	99	1	0.97	0.01689	0.938	1
9.86	96	1	0.96	0.01951	0.923	0.999
13.04	92	1	0.95	0.02191	0.908	0.994
18	91	1	0.939	0.02403	0.893	0.988
18.5	90	1	0.929	0.02593	0.879	0.981
18.83	89	1	0.918	0.02766	0.866	0.974
19.02	87	1	0.908	0.02928	0.852	0.967
20.27	86	1	0.897	0.03079	0.839	0.96
25.13	82	1	0.886	0.0323	0.825	0.952
28.65	76	1	0.875	0.03391	0.811	0.944
31.21	75	1	0.863	0.03541	0.796	0.953

31.8	74	1	0.851	0.0368	0.782	0.927
35.35	73	1	0.84	0.0381	0.768	0.918
53.82	47	1	0.822	0.04127	0.745	0.907
55.39	45	1	0.804	0.04421	0.721	0.895
64.76	31	1	0.778	0.04981	0.686	0.882
68.04	28	1	0.75	0.05523	0.649	0.866
80.03	20	1	0.712	0.06394	0.597	0.849
84.83	16	1	0.668	0.07384	0.538	0.829

PathGGS=8

time	n.risk	n.event	survival	std.err	0.95LCL	0.95UCL
1.41	12	1	0.917	0.0798	0.773	1
1.64	11	1	0.833	0.1076	0.647	1
2.1	10	1	0.75	0.125	0.541	1
8.97	9	1	0.667	0.1361	0.447	0.995
13.21	8	1	0.583	0.1423	0.362	0.941
16.82	7	1	0.5	0.1443	0.284	0.88
27.6	6	1	0.417	0.1423	0.213	0.814
27.86	5	1	0.333	0.1361	0.15	0.742

PathGGS=9

time	n.risk	n.event	survival	std.err	0.95LCL	0.95UCL
1.38	12	1	0.917	0.0798	0.7729	1
1.81	11	1	0.833	0.1076	0.647	1
1.87	10	1	0.75	1250	0.541	1
2.56	9	1	0.667	0.1361	0.4468	0.995
2.92	8	1	0.583	0.1423	0.3616	0.941
3.71	7	1	0.5	0.1443	0.284	0.88
6.7	6	1	0.417	0.1423	0.2133	0.814
10.61	5	1	0.333	0.1361	0.1498	0.742
30.56	3	1	0.222	0.1283	0.0717	0.689

40.9	2	1	0.111	0.1014	0.0186	0.665

PathGGS=Tx_Effects

time	n.risk	n.event	survival	std.err	0.95LCL	0.95UCL
39.5	1	1	0	NAN	NA	NA