# SURVIVAL ANALYSIS OF UNPLANNED HOSPITAL READMISSION WITHIN 30 DAYS OF POST CANCER THERAPY TREATED PATIENTS

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

by

**MD AKTER HOSSAIN**

**DR. MUNNI BEGUM - ADVISOR**

BALL STATE UNIVERSITY

MUNCIE, INDIANA

DECEMBER, 2019

# Acknowledgements

In no particular order of significance, I would like to acknowledge the support and input of certain people who have been tremendous in the production of this work. My profound appreciation goes to my thesis supervisor, Dr. Munni Begum, for her enormous support and guidance throughout this work. The availability of her support, contribution and tolerance were always forthwith. I am immensely grateful to her.

I am also very grateful to Dr. Rebecca Pierce and Dr. Drew Lazar, who are members of my thesis committee, for their sincere, constructive and collective contributions, insightful comments and patience.

Finally, I owe gratitude to all our family members for their continuous support and motivation, special thanks to my wife-Sufia Begum, daughter-Nahsita Hossain and son-Taseen Hossain for the sacrifices they made in last three years.

# ABSTRACT

**THESIS:** SURVIVAL ANALYSIS OF UNPLANNED HOSPITAL READMISSION WITHIN 30 DAYS OF POST CANCER THERAPY TREATED PATIENTS

**STUDENT:** Md Akter Hossain

**DEGREE:** Master of Science

**COLLEGE:** Science and Humanities

**DATE:** December 2019

**PAGES:** 49

Cancer treatment has been declared a crisis in the United States because of the growing demand of services, increasing complexity of treatment and dramatically rising costs of treatment. The vast majority of adverse events occur within 30 days after receiving treatment. This suggests that this 30-day period is a very important time frame to observe side effects of treatment.

We analyzed unplanned hospital readmission data of cancer patients at Ball Memorial Hospital from June 2018 to May 2019 and explored how certain demographic and clinical characteristics affected patients' risk of unplanned readmission within 30 days. Our key explanatory/predictive variables were patients' age, gender, cancer stage and elapsed days and the response variable was the time to event which is patients' unplanned return to hospital within 30 days. We performed survival analysis to identify factors affecting hospital readmission and their significance. We applied the semi-parametric Cox Proportional Hazard model and the parametric Exponential and Weibull models, to determine if there were any significant differences in the results obtained from the two methods.

We considered models with gender, age and cancer stage and found that cancer stage was highly associated with risk of hospital readmission. Moreover, after adjusting for age, cancer stage became even more significant predictor for the risk of hospital readmission. The results from both semi-parametric and parametric models were consistent.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Introduction and Background

### Introduction

Cancer treatment has been declared a crisis in the United States because of the growing demand for services, increasing complexity of treatment and dramatically rising costs of cancer treatment [2]. Some 1.6 million individuals are diagnosed with cancer each year, and the number of cancer survivors is projected to increase dramatically because of the aging population and improvements in treatment [2]. By 2020, cancer care costs are expected to reach $173 billion, reflecting a considerable increase from $72 billion in 2004 [2]. According to [2], program and policies to reduce hospital readmission are increasingly viewed as promising avenues to reduce spending and improve healthcare quality and efficiency as well as patient experiences. For many cancer patients, readmission following hospitalization may be preventable and should be addressed to potentially lower costs and improve patient outcomes. Each year, approximately twenty-two percent of patients with cancer receive chemotherapy [7]. According to the existing literature, the vast majority of adverse events occur within 30 days after receiving treatment [7]. The thirty-day readmission rate for cancer patients discharged from medical services are as high as twenty seven percent [5]. However, an inadequate number of studies have been conducted to identify preventable readmission and few strategies exist to reduce readmission in cancer patients.

**Background**

Hospital admission and unplanned emergency department (ED) visits following outpatient care could be a reflection of poor quality of care. To measure quality and service, the Centers for Medicare and Medicaid Services (CMS) implemented the Hospital Outpatient Quality Reporting Program (HOQRP), which gives hospitals incentives to provide quality services. In 2006, the Tax Relief and Health Act mandated that hospitals submit quality measure data on their outpatient departments. Reimbursement to hospitals may be reduced by up to two percent if hospitals fail to meet requirements [6]. Historically, outpatient cancer therapy had been excluded from HOQRP but was added in 2017. The specific outcomes and quality measures that the CMS is looking at for outpatient chemotherapy are the occurrence of one or more hospital admissions or ED visits for any of the 10 quality measures: dehydration, diarrhea, emesis, fever, nausea, neutropenia, anemia, pain, pneumonia, or sepsis that occurred within 30 days of chemotherapy treatment.

## 1.2 Literature Review

According to Mathematica Policy Research [7] about twenty-two percent of patients with cancer receive chemotherapy, increasingly in Hospital Outpatient Departments (HOPDs). Outpatient hospital-based chemotherapy rose from 17 to 30 percent of all chemotherapy provided to Medicare patients from 2008 to 2012, and this trend is likely to continue. The study by Mathematica Policy Research [7] reveals that chemotherapy treatment can have severe, predictable side effects, which, if appropriately managed, can reduce patients' quality of life and increase healthcare utilization and costs. Approximately 40 percent of those admissions and 50 percent of those ED visits were caused by complications from chemotherapy. The study [7] used a two level hierarchical logistic regression model to estimate the risk-standardized outcome rates. The study determined a risk-adjusted model among all hospital types and used the model to calculate the risk-standardized inpatient admission rate and ED visits. Another study [2] reviewed the existing literature on readmission rates, predictors, and reasons for readmission among adults with cancer. The study found that the highest readmission rates were observed in patients with bladder, pancreatic, ovarian or liver cancer. Significant predictors of the

readmission included comorbidities, older age, advanced disease as measured by cancer stage, tumor size or lymph node involvement and length of hospital stay. This research also revealed that the common reasons for readmission included gastrointestinal and surgical complications, infection, and dehydration. Moreover, the study measured hospital level performance. However, a study entitled, "All-causes admissions and readmissions 2017", conducted by National Quality Forum [4] mentioned that although a wide variety of healthcare stakeholders support the goal of reducing readmissions, debates remain on the target rate for readmissions. Systematic reviews have found that less than a third of readmissions could be considered preventable [4]. Moreover, many factors related to readmission rates may be outside of a hospital's control, such as the resources available to the community it serves. Research also shows that readmissions and penalties have been significantly higher in hospitals that serve a larger proportion of low-income Medicare patients and in major teaching hospitals which tend to care for the sickest patients. At the same time, low readmission rates may be associated with higher rates of observation stays or ED use. According to [5], higher readmission rates among those discharged home with help suggests that services supplied are not sufficient to address their health needs. Franco, R. et al. in [5] assessed the prevalence of potentially preventable readmissions and associated factors in adult patients with metastatic cancer.

However, no research report was found to address patient survival probability and factors associated with this probability. Our study aims to model the longitudinal data using survival analysis methodology where the outcome variable is the time until an occurrence of event. In this type of analysis both components, (i) if an event (i.e. readmission in 30 days) occurs or not and (ii) when the event occurs can be incorporated simultaneously. Thus, the benefit of survival analysis over logistic regression or other statistical methods is the ability to add a time component to the model while also effectively handling censored data.

# Chapter 2

# Methodology

## 2.1 Introduction

This chapter provides an outline of the methodology used to answer the research questions. Data sources, data collection process and methods of data analysis are also described. A brief overview of survival or time to event analysis, basic notation and terminologies are also included. A time to event analysis usually refers to the time variable as survival time or time to event because it gives the time that an individual has 'survived' over some follow-up period. Survival time typically refers to the event as failure, because the event of interest typically is death, disease incidence or some other negative individual experience. In this study the event or failure was the hospital readmission within 30 days.

## 2.2 Data Source and Study Population

Our secondary data were collected under the direct supervision of a Resident Pharmacist, Department of Pharmacy, Indiana University of Health Ball Memorial Hospital, Muncie, Indiana. Cancer outpatients who received chemotherapy from June 2018 to May 2019 were the participants of the study. All together, 224 outpatients received cancer therapy during this time period. Sixty-four patients revisited the hospital and received post-cancer therapy treatment within 30 days of receiving outpatient chemotherapy. The rest, 160, patients did not visit the hospital to receive post-cancer therapy within 30 days. The data has three main parts: (1) the response variable, which is the time until

the occurrence of an event, (2) the time an observation is censored since for some subjects the event of interest did not occur and (3) predictor variables which may have an effect on the time to event.

## 2.3 Research Questions

This study sought to answer of the following research questions:

i) What was the probability that a cancer patient will be readmitted to the hospital within 30 days of receiving chemotherapy?

ii) How did certain demographic and clinical characteristics affect patient chance of hospital readmission?

To answer these questions, survival analysis was carried out to identify patients at risk and to predict the probability of as well as the pattern of hospital readmission within 30 days.

## 2.4 Exploratory Data Analysis

We used exploratory data analysis to summarize the patients main characteristics, primarily with visual methods. We present univariate and bivariate analysis of patients demographic characteristics. In particular, we consider the distribution of patients age, distribution of age according to gender, summary statistics of elapsed days as well as gender and the distribution of elapsed days according to gender. These analyses reveal the nature and pattern of the data, and allow covariates to be identified.

## 2.5 Notations and Terminologies

Survival analysis is a collection of statistical procedures for which the outcome variable of interest is time until an event occurs. By time, we mean the number of years, months, weeks, or days that occurred from the beginning until an event occurs. By event, we mean death, disease incidence, relapse from remission, recover or any designated experience of interest that may happen to an individual. Survival time gives the time that an individual has "survived" over some follow-up period. We denote, the random vari-

able for a person's survival time as capital T. Since T denotes time, its possible values include all non-negative numbers; i.e. ($T \geq 0$). A lower-case t denotes any specific value of interest of the random variable T.

The survival function denoted by S(t) and the hazard function denoted by h(t) are two functions used in survival analysis. The survival function S(t) gives the probability that a person survives longer than some specific time t. Thus, S(t) gives the probability that the random variable T exceeds the specific time t. The survival function is fundamental to survival analysis because obtaining survival probabilities for different values of t provides crucial summary information about survival data. On the other hand, the hazard function h(t) gives the probability that the subject will survive in a particular time interval, given that the subject has not failed up to that point in time. The hazard rate is interpreted as the rate at which failure occurs at that point in time or the rate at which risk is accumulated. This interpretation coincides with the fact that the hazard rate is the derivative of the cumulative hazard function H(t).

Another important characteristics of survival analysis is that data can be incomplete due to the inability to continuously track the subject. This characteristic is referred to as censoring. In essence, censoring occurs when we have some information about an individual survival time, but we do not know the survival time exactly. There are generally three reasons whey censoring may occur: (i) a person does not experience the event before the study ends, (ii) a person is lost to follow-up during the study period, or (iii) a person withdraws from the study because of death (if death is not the event of interest) or some other reason. There are different types of censoring, right censoring and left censoring. Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred, whereas left censoring is when the event of interest has already occurred before enrollment.

## 2.6 Estimating the Survival Function

In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Unlike ordinary regression models, survival models incorporate information from both uncensored and censored data to evaluate important features. One of the paramount aspects of survival analy-

sis is that covariate information that varies over time. This aspect of time-dependent covariates makes survival analysis very unique in a way that other standard statistical methods cannot handle. Therefore, the three main purposes of survival analysis are (1) to estimate and interpret the survival and/or hazard functions, (2) compare survival and hazard functions and (3) assess the relationship of explanatory variables to survival time using different step-by-step survival analysis techniques. The most popular approaches are Kaplan-Meier method, the log-rank test, Cox Proportional Hazard Model and the Parametric Method. Details about each of these techniques is discussed in the following sections. In addition, the proportional hazard assumptions are checked using Schoenfeld's test, which is similar to a goodness of fit test.

**Kaplan-Meier Method/Product Limit Estimator:**

The survival function S(t) gives the probability that a person survives longer than some specified time t. Notationally we write S(t)=$Pr(T > t)$, where T is the survival time [3]. A random variable indicating either censorship or failure is denoted as follows

$$\partial = \begin{cases} 1, & \text{if failure} \\ 0, & \text{if censored} \end{cases}$$

The estimated survival probabilities are computed using a product limit formula, given by the Kaplan-Meier (KM) method. The general formula for a KM survival probability at failure time $t_i$ is

$$\begin{aligned} \hat{S}(t_{(j)}) &= \prod_{i=1}^{j} \hat{P}_r[T > t_{(i)}/T \geq t_{(i)}] \\ &= \hat{S}(t_{(j-1)}) \times \hat{P}_r(T > t_{(i)}/T \geq t_{(i)}). \end{aligned}$$

The KM survival curve shows the difference between two groups, and we can test if the difference is statistically significant using the log-rank test. The log-rank test ignores other factors. However, there may be other factors at play which we want to control in our model.

**Log-Rank Test:**

The log-rank test is a method to compare survival curves to determine if the difference is due to chance or if there is a real difference in outcomes. When the two curves are

significantly different, the result can be generalized for the population at large [6]. The log-rank test is a statistical test of the null hypothesis specifying a common survival curve. For two groups, the log-rank statistic is calculated as the ratio of square of the observed minus expected scores for a given group divided by its variance estimate. The test statistic is approximately chi-square for large samples with G-1 degrees of freedom, where G denotes the number of groups being compared. Thus, the the null and alternative hypothesis are as follows:

$$H_{(0)} : S_1(t) = S_2(t)$$

$$H_{(1)} : S_1(t) \neq S_2(t)$$

and the log-rank test statistic:

$$log - rank = \frac{(O_i - E_i)^2}{var(O_i - E_i)}$$

where i=1, 2.

The variance formula involves the number in each group, $n_{ij}$, and the number of failures in each group, $mij$, at time j. For our case of two groups, we have

$$var(O_i - E_i) = \sum_j \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)}$$

where $m_{ij}$ are the observed counts and $e_{ij}$ are the expected counts calculated as $(\frac{n_{ij}}{n_{1j}+n_{2j}})\times$ $(m_{1j} + m_{2j})$. We use the log-rank test to check whether there is any difference in the survival curves for several groups. However, since the log-rank test does not consider other explanatory variables, it is unlikely to detect a difference between groups when survival curves cross.

**Cox Proportional Hazards Model:**

The Cox Proportional Hazard (Cox PH) Model, the most popular survival regression model, investigates the relationship of predictors and the time to event through the hazard function. The hazard function describes the number of events per unit time [6].The Cox PH model is similar to usual regression models as it fits the response variable, time

to event, through the hazard function. This model, unlike the log-rank test, allows us to consider other factors. An important feature is that the model is semi-parametic since it contains two components, namely a baseline hazard function of time and an exponential function involving predictors but not time. Thus, Cox PH model does not depend on a distribution assumption for T, time to event. The Cox PH model is defined by

$$h(t,x) = h_o(t)exp(\sum_{i=1}^{p} \beta_i X_i)$$

where $h_o(t)$ is the baseline hazard function and X represents the vector of predictors, $(X_1, X_2, ....X_p)$. The hazard ratio is the estimate of $h(t, X^*)$ divided by the estimate of h(t,X), where $X^*$ denotes the set of predictors for one individual and X denotes the set of predictors for the other individual.The equation below

$$\hat{HR} = \frac{\hat{h}(t,X^*)}{\hat{h}(t,X)} = exp\left[\sum_{i=1}^{p} \beta(X_i^* - X_i)\right]$$

is used to test the effect of other independent variables on the survival times of different groups of patients, similar to a multiple regression model. Hazard, the dependent variable, can be defined as the probability of dying at a given time, assuming that the patient has survived up to that given time. The hazard ratio is defined as the ratio of the risk of hazard occurring at any given time in one group compared with another group at that same time.

A log-log plot is used to check if survival curves cross. However, if the curves cross, it indicates the proportional hazard assumption is violated and it is not appropriate to use the Cox PH model. Another way to check the proportional hazard assumption is using Schoenfeld's test which is similar to a goodness of fit test. When the hazard assumption is violated, there are two approaches, one approach is to stratify the data and fit a separate Cox PH model for the different levels of covariates. While the other approach is to fit a Cox PH model with time dependent covariates or coefficients.

# Chapter 3

# Data and Variables

In this chapter we discuss an important aspect of my study, the data and the variables. Included are a discussion of the sources of data, the method of data collection, the inclusion and exclusion criteria, the data processing, the variables included for analysis, and the data limitations.

## 3.1 Data

Data was gathered under the direct supervision of the Resident Pharmacist, Department of Pharmacy, Indiana University of Health Ball Memorial Hospital, Muncie, Indiana. Data was recorded on cancer hospital admissions who received chemotherapy as an outpatient from June 2018 to May 2019. A total of 224 patients received cancer therapy during this study period, and 64 patients revisited the hospital to receive post-cancer therapy treatment within 30 days of receiving the outpatient chemotherapy. Thus, 160 patients did not visit the hospital to receive post-cancer therapy treatment within 30 days. These two data sets were categorized as “Readmit” and “Not-readmit”, respectively. As this research is performing ’time to event’ or ’survival analysis’, thirty days is a very important predictor. Both data sets have some common variables such as patient’s clinical and demographic information along with some other hospital record related general information.

## 3.2 Data Processing

The original data sets were in two different Excel files, which were exported to SAS and appended to a SAS data file for data processing and analysis purposes. The combined data set consisted of a total of 224 observations/patients. In the combined data set, a variable named 'event' was created. If the patients visited the hospital for post-cancer therapy treatment within 30 days, then the 'event' occurred and was coded as 'Yes'. If the patients did not visit within 30 days, then the 'event' did not occur and was coded as 'No'. In addition, the value for the variable elapsed days was given 30+ for those patients who did not visit for post-cancer therapy treatment within 30 days of receiving chemotherapy. They were treated as censored patients/observations. On the other hand, if 'elapsed days' is less then or equal to 30 days, then the actual number of days is written under the variable 'elapsed days' and those patients were considered as uncensored.

## 3.3 Explanatory and Response Variables

In a regression modeling setup, such as the Cox PH model, the explanatory variables, also known as features/independent variables/predictors, are used to predict or explain differences in the response or outcome variable. The Readmit dataset, for patients who received post-cancer therapy treatment, consisted of the variables age, gender, cancer stage, elapse days, event, reasons for readmission, types of treatment received, and cancer diagnosis. On the other hand, the variable for the not-readmitted dataset consisted of only the variables age, gender, cancer stages, elapsed days, and event. Thus, the common variables were patient's age, gender, cancer stages, elapsed days and the event. In the combined dataset, the variable 'event' is the response variable and all other variables were either demographic or clinical variables, which were considered as explanatory variables. In survival analysis, the response variable is the time until the occurrence of the event of interest. In this study, the 'event' of interest was if patients visited for post-cancer therapy treatment within 30 days of receiving chemotherapy. The number of patients who experienced the 'event' was 64.

## 3.4 Method of Data Collection

The Retrospective Chart Review (RCR), also known as a medical record review, was used to gather data from patients. Patients who met the following criteria were included in the study:

- Medicare-paid patient
- Age greater than 18 year-old
- Confirmed cancer diagnosis
- Receiving cancer therapy in an outpatient setting

Patients with the following criteria were excluded from the study:

- Confirmed leukemia diagnosis
- Bone marrow transplant recipient
- Organ transplant recipient
- Patients receiving planned inpatient chemotherapy setting
- Patients in rehabilitation care

## 3.5 Data Limitations

A major data limitation is the presence of heavy censoring due to the artificial consideration of majority of patients not readmitting as censored period. The majority, 71%, of the patients did not receive unplanned post-cancer therapy treatment within 30 days. Those patients were considered as censored for this study. In the combined data set, there are only three common explanatory variables from both data sets; therefore, an in-depth causal analysis and drawing inferences based on those explanatory variables may not be possible. Additionally, some patients visited the hospital multiple times in the study period. However, they were counted only once for their first visit to avoid duplication of record.

# Chapter 4

# Findings

## 4.1 Exploratory Data Analysis

This chapter discusses patient demographic and clinical characteristics such as sex, age and cancer stages. Data for this exploratory analysis was gathered from 224 admitted outpatients who received cancer therapy during the study period which was June 2018 to May 2019. Among them 64 patients were readmitted to the hospital within 30 days of receiving outpatient chemotherapy. If patients were readmitted in 30 days this is categorized as an 'event' and if not then we say the event has not occurred. The analysis will be performed both on all 224 patients and on the 64 readmitted patients separately.

Table 4.1 describes the distribution of gender according to event type. The proportion of females and males who had the event are 50.6% and 47.3%, respectively. A chi-square test was performed to see if there is a significant difference in the event occurring in female versus male patients. The test does not show a significant effect of gender on the event occurring with a chi-square value of 0.61.

Table 4.2 and graph 4.1 show tabular and graphical distributions of elapsed days to readmission by gender. The mean elapsed days for female is 26.5 and for males it is 25.4 with standard deviations of 8.5 and 8.7, respectively. A t-test was carried out to see if the difference of average elapsed days for female and male was significant. The test did not show a significant difference with a p-value of 0.84.

Table 4.2 and graph 4.2 show the tabular and graphical distributions of age by gender. The mean age of female and male patients are 70.8 and 73.6 with standard deviations of

9.6 and 7.9, respectively. A t-test was carried out to see if the difference of average ages for females and males was significant and this test did not show a significant difference.

Table 4.1: Patients According to Gender and Event

| Event | Gender | | |
|---|---|---|---|
| | Female | Male | Total |
| No | 86 (72.8) | 74 (69.8) | 160 (71.4) |
| Yes | 32 (27.1) | 32 (30.2) | 64 (28.6) |
| Total | 118 (52.6) | 106 (47.3) | 224 (100) |

Table 4.2: Demographic Information by Elapse Days and Ages

| Statistics | Elapse Days | | | Age | | |
|---|---|---|---|---|---|---|
| | Female | Male | Overall | Female | Male | Overall |
| N | 118 | 106 | 224 | 118 | 106 | 224 |
| Mean | 25.57 | 25.41 | 25.49 | 70.75 | 73.58 | 72.08 |
| Std | 8.50 | 8.67 | 8.57 | 9.62 | 7.89 | 8.94 |
| Min. | 01 | 01 | 01 | 40.00 | 47.00 | 40.00 |
| Max. | 30+ | 30+ | 30+ | 97.00 | 92.00 | 97.00 |

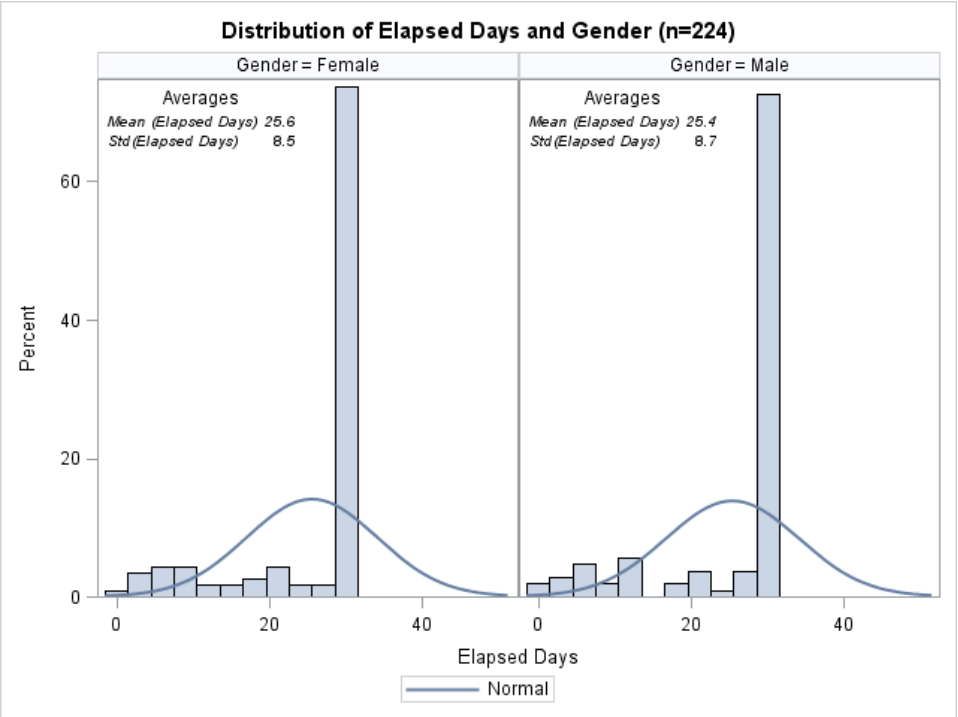


Figure 4.1: Elapsed Days and Gender

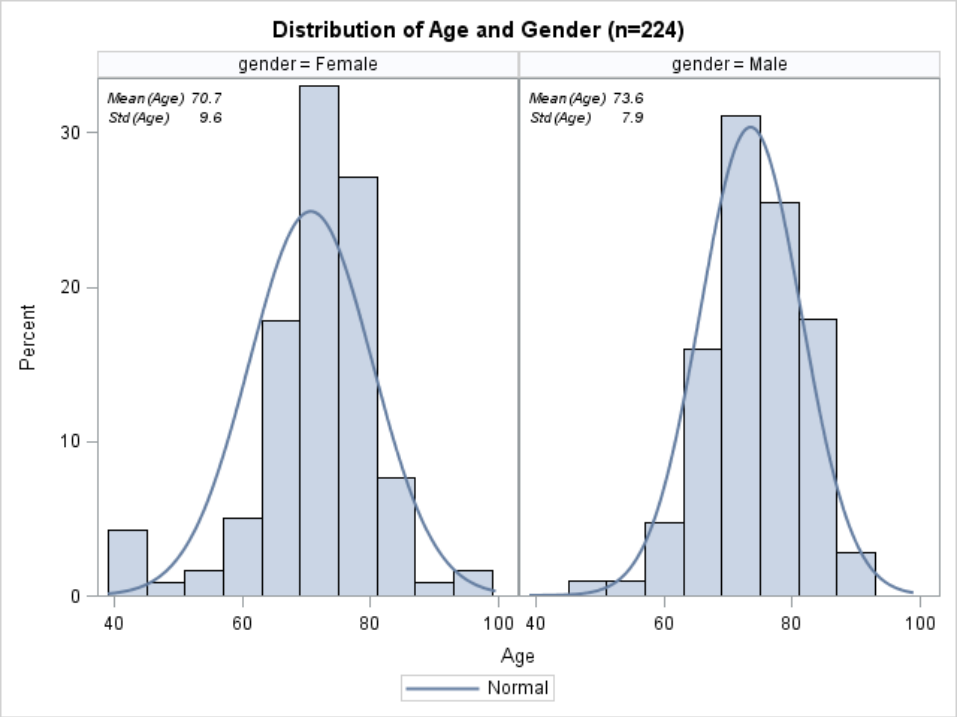


Figure 4.2: Age and Gender

Graph 4.3 shows the distribution of age according to event. The average age of patients for which the event occurred is 72.4 and the average age of patients for which the event did not occur is 70.4 with standard deviations of 8.8 and 5.8, respectively. A t-test for the differences in the mean age of patients who had the event and for patients who did not have the events was carried out. This test did not show any significant difference with a p-value 0.41.

Table 4.3 shows the distribution of event according to cancer stage. The table shows the higher the cancer stage the higher the chance of returning to the hospital within 30 following chemotherapy. A Chi-square test was performed to determine the significance of the association and the result was a statistically significant association between event and cancer stage with a p-value of 0.01. Graph 4.4 shows the distribution of cancer stages according to age. The average age of patients according to stage varies for each cancer stage. A t-test was performed to see if there are significant difference between the first three cancer stages and age. The test does not show any significant difference in cancer stages according to age with a p-value of 0.61. Graph 4.5 does not show any strong to moderate correlation among the variables. The highest sample correlation is between elapsed days and cancer stage with $r = -0.18$, followed by elapsed days and age with $r = -0.04$, where $r$ is the sample correlation coefficient. The lowest sample correlation between stage and age is $r = -0.02$.

Table 4.3: Patient Event Type According to Cancer Stage

| Event | Cancer | | | | |
|---|---|---|---|---|---|
| | Stage-1 | Stage-2 | Stage-3 | Stage-4 | Total |
| No | 10 (6.3) | 24 (15.0) | 33 (20.6.4) | 93 (58.1.4) | 160 (71.4) |
| Yes | 0 (0.0) | 3 (4.7) | 11 (17.2) | 50 (78.1) | 64 (28.6) |
| Total | 10 (4.5) | 27 (12.1) | 44 (19.6) | 143 (63.8) | 224 (100) |

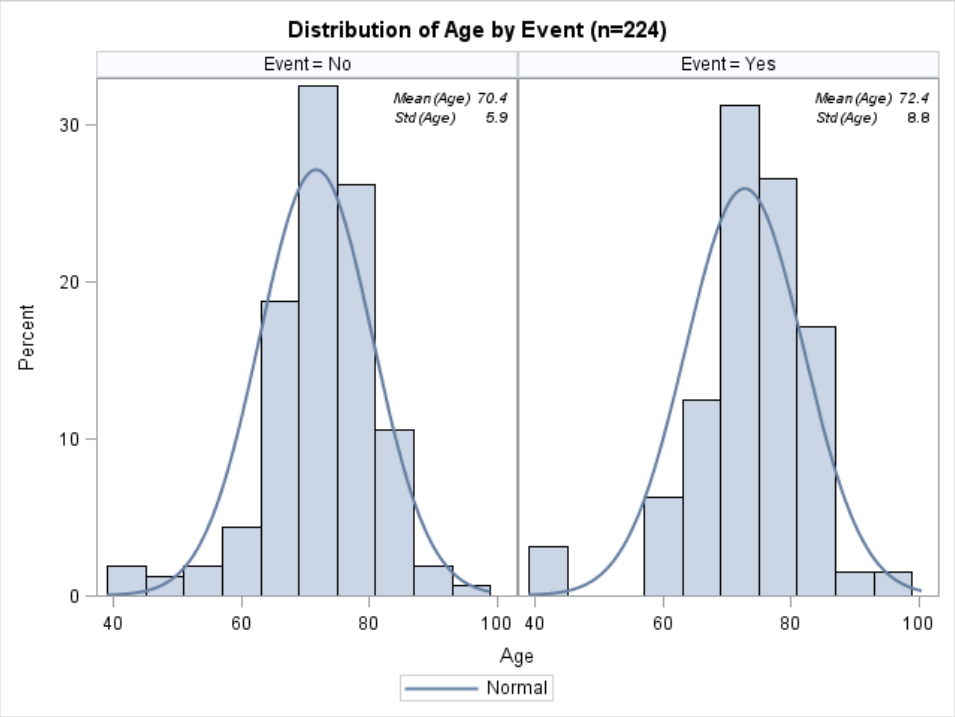


Figure 4.3: Age and Event

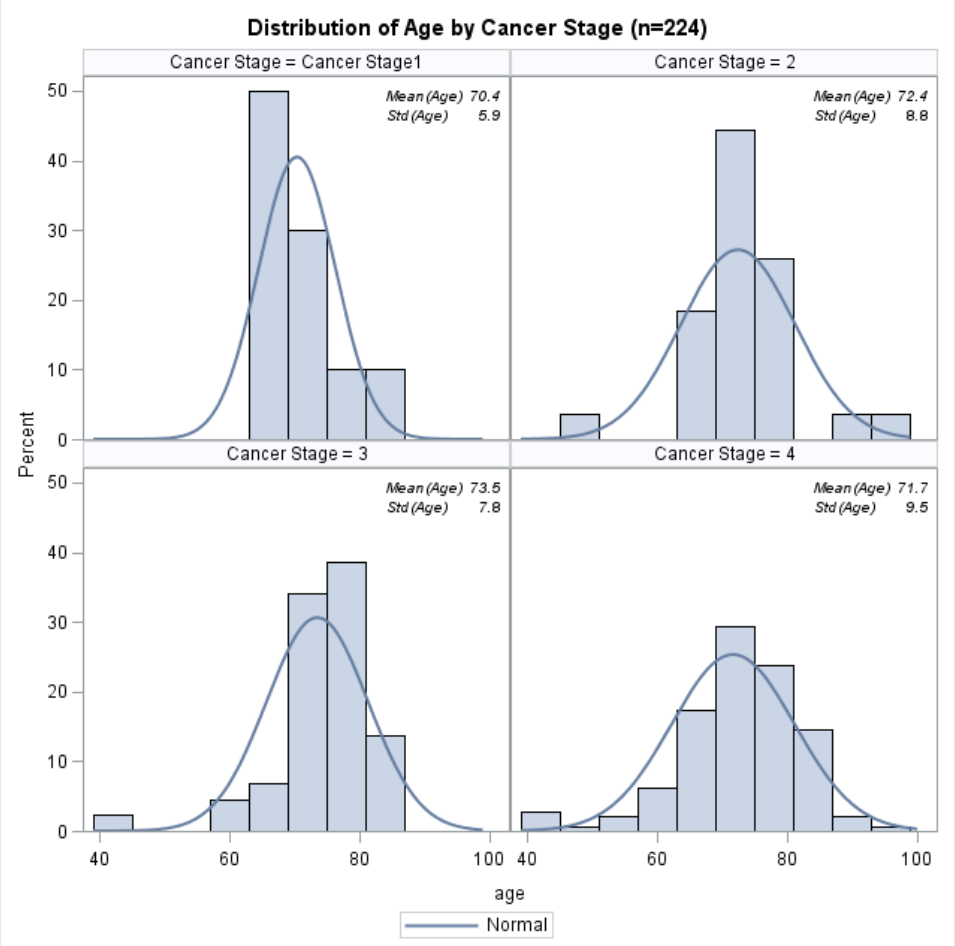


Figure 4.4: Age and Cancer Stages

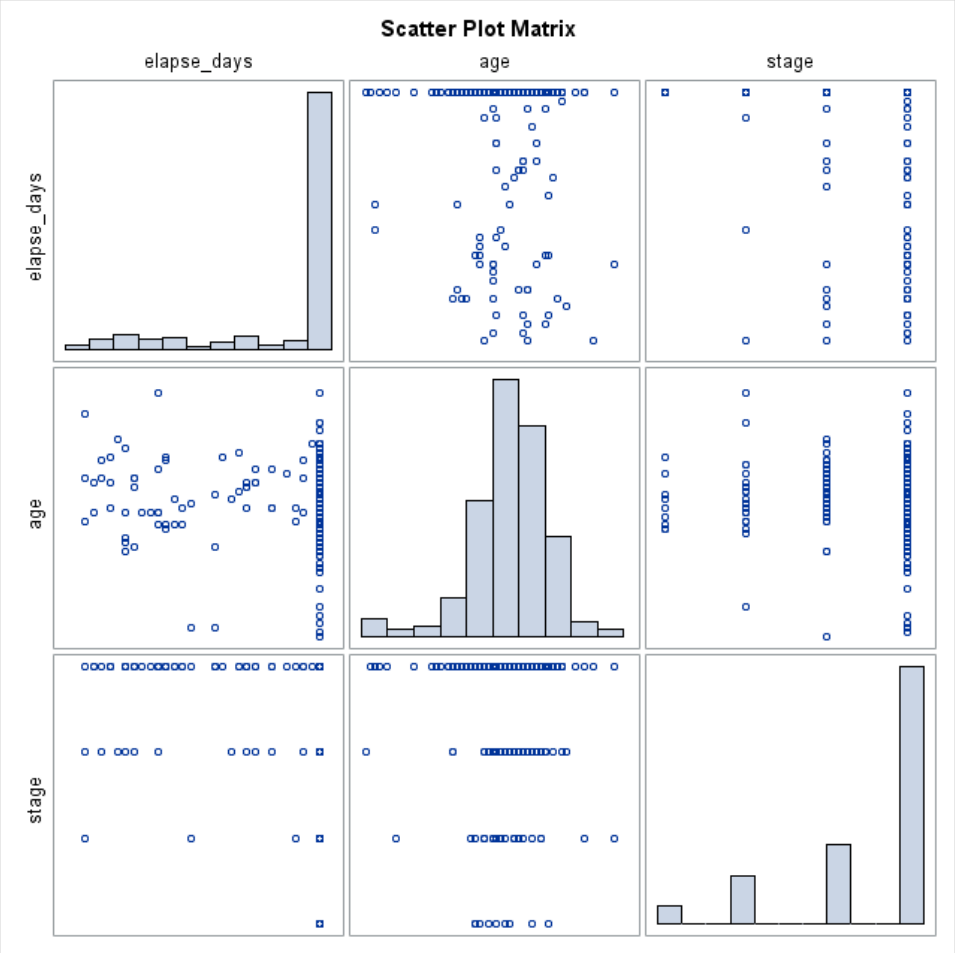


Figure 4.5: Correlation Among Exploratory Variables

## 4.2 Readmitted Patients

This section briefly discusses the exploratory analysis of readmitted patients (n=64) who visited the hospital within 30 days following chemotherapy. Graph 4.6 and Table 4.4 shows the graphical and tabular distribution of elapsed days and age according to gender. Average elapsed days for females and males are 13.7 and 14.8 with standard deviations of 8.5 and 9.4, respectively. A t-test was conducted to see if there is a difference of elapsed days between females and males. The test does not show any significant difference between females and males. Graph 4.7 and table 4.4 show the distributions of average age of females (71.2) and males (74.5) with standard deviations of 10.4 and 14.9 respectively. A t-test was conducted to see if there is any significant difference of patients' age between females and males. However, no significant difference was found with a p-value of 0.15. Graph 4.8 describes the distribution of ages according to cancer stages. However, according to a Chi-square test, there is no significant relationship between cancer stage and age group with a p-value of 0.61.

Table 4.4: Demographic Information by Elapsed Days and Ages

| Statistics | Elapsed Days | | | Age | | |
|---|---|---|---|---|---|---|
| | Female | Male | Overall | Female | Male | Overall |
| N | 32 | 32 | 64 | 32 | 32 | 64 |
| Mean | 13.7 | 14.5 | 14.2 | 71.2 | 74.5 | 72.9 |
| Std | 8.5 | 9.4 | 8.9 | 10.4 | 14.8 | 14.2 |
| Min. | 01 | 01 | 01 | 42.0 | 61.0 | 42.0 |
| Max. | 30 | 30 | 30 | 97.0 | 92.0 | 97.0 |

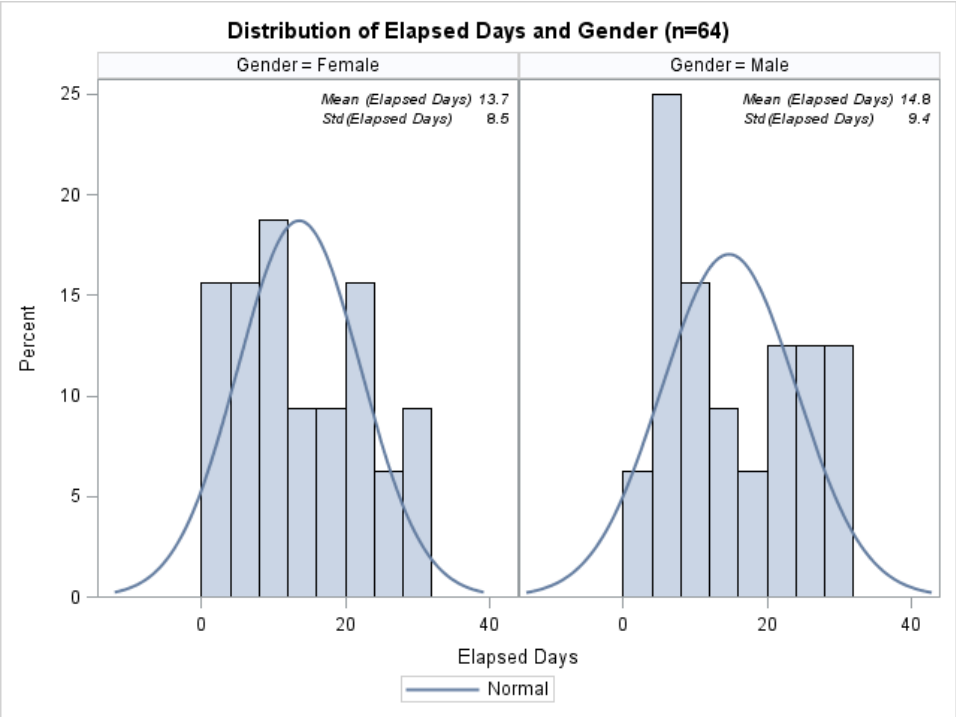


Figure 4.6: Elapsed Days and Gender

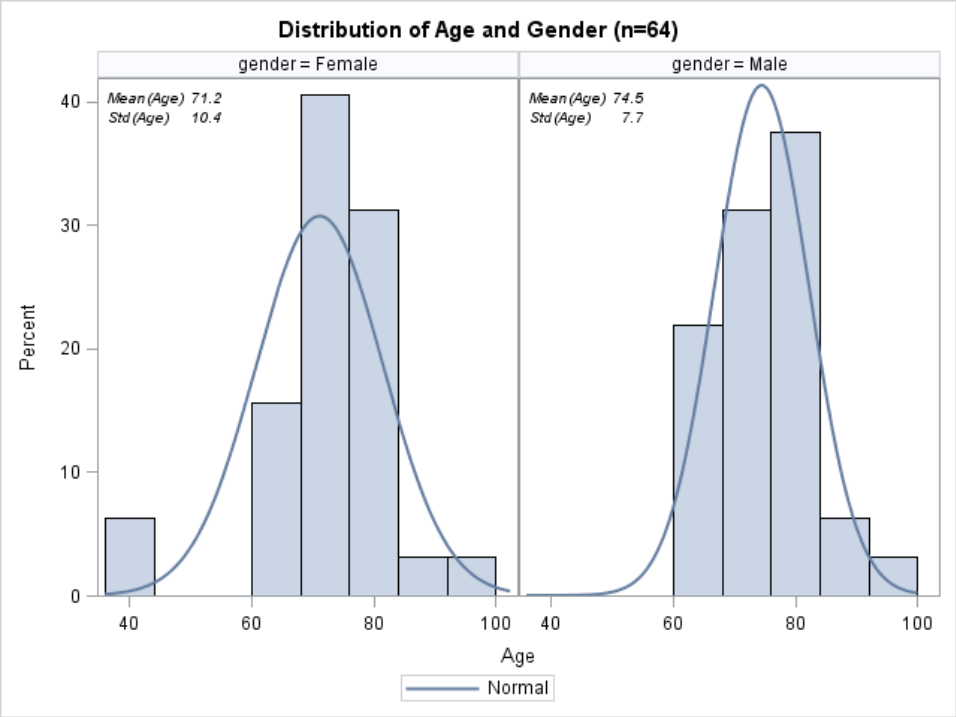


Figure 4.7: Age and Gender

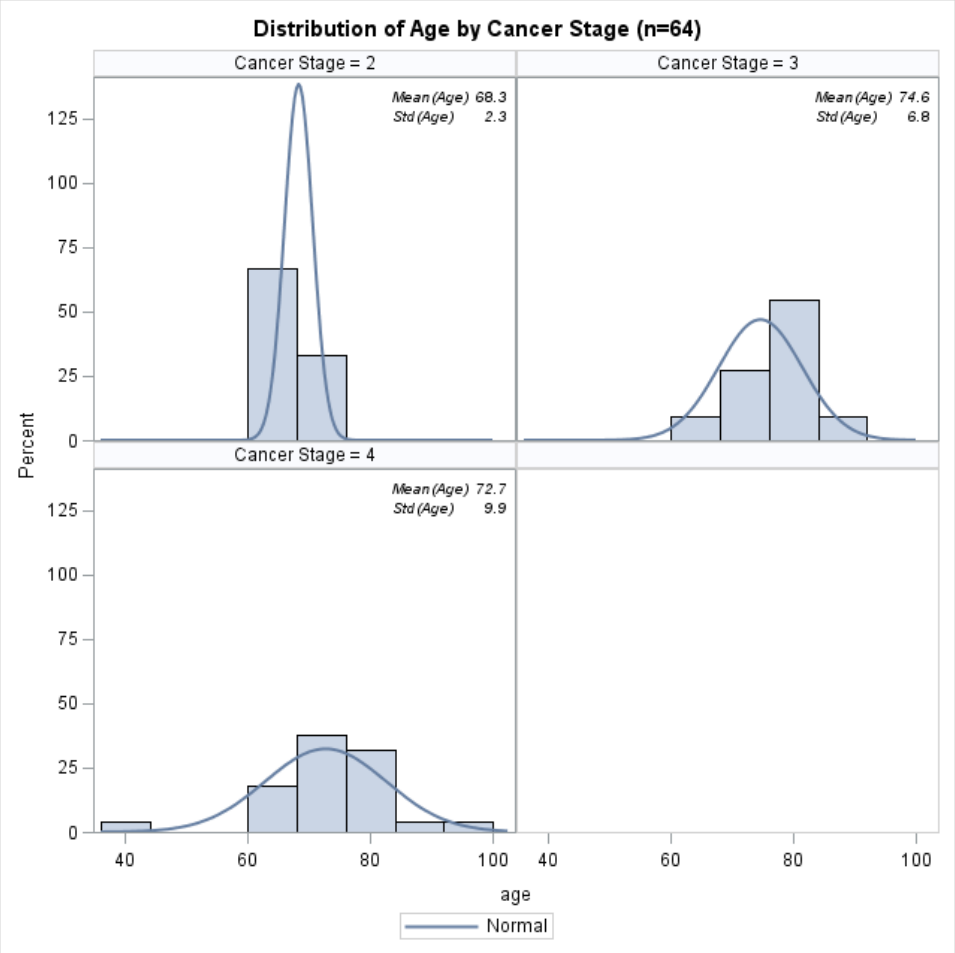


Figure 4.8: Age by Cancer Stages

From the above graphical and tabulated distributions of both data sets, admitted and readmitted, we determined the patients in the two data sets have similar demographic and clinical characteristics in terms of average age, the proportions that are male and female and cancer stages. In addition, we found very low correlation among the predictor variables.

## 4.3 Survival Probability

In the clinical trial and biomedical fields, the KM estimator is the most widely used method for estimating survivor functions. It also know as the product-limit estimator [1]. Graph 4.9 shows the survival probabilities across elapsed days. It shows the median survival time is 25.5 days with standard error of 0.58. Graph 4.10 compares the survival distributions of females and males where the survival probabilities of females are slightly higher than that of males. However, the log-rank test does not confirm that the difference is significant with a p-value of 0.64. It is to be noted that the mean survival time and its standard error were under estimated due to a large number of right-censored observations. Graph 4.11 shows patient survival probabilities according to cancer stages. It is observed in Figure 4.11 that patients in cancer stages 1-3 have higher survival probabilities compared to patients in cancer stage 4. This indicates that the patients who are in earlier stages have a better survival prognosis than patients who are in later stages. The log-rank test shows that the difference is highly statistically significant with a p-value of $\approx 0.00$. Moreover, as the number of days increases, the two curves appear to get further apart, suggesting that being in cancer stage 4, increases the risk of unplanned hospital readmission. Graph 4.12 displays the patients survival probabilities according to age groups. The graph shows that the survival probabilities of patients in the seventies age group have higher survival probabilities compared to patients in the eighties age group. However, the log-rank test did not confirm that the difference is statistically significant with a p-value of 0.51.

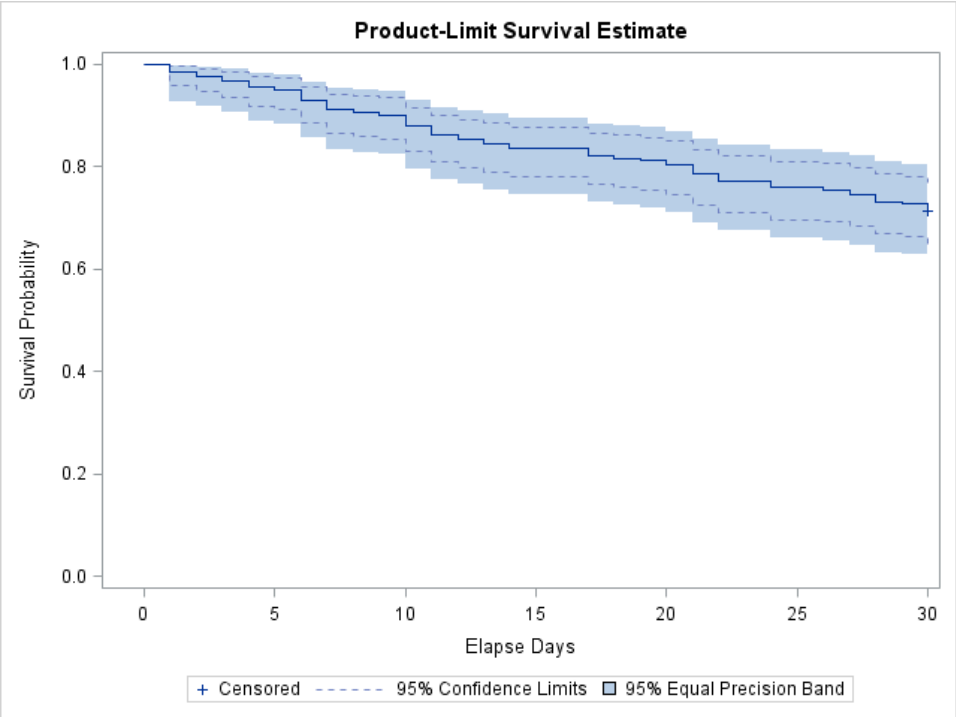


Figure 4.9: Product-Limit Survival Estimate

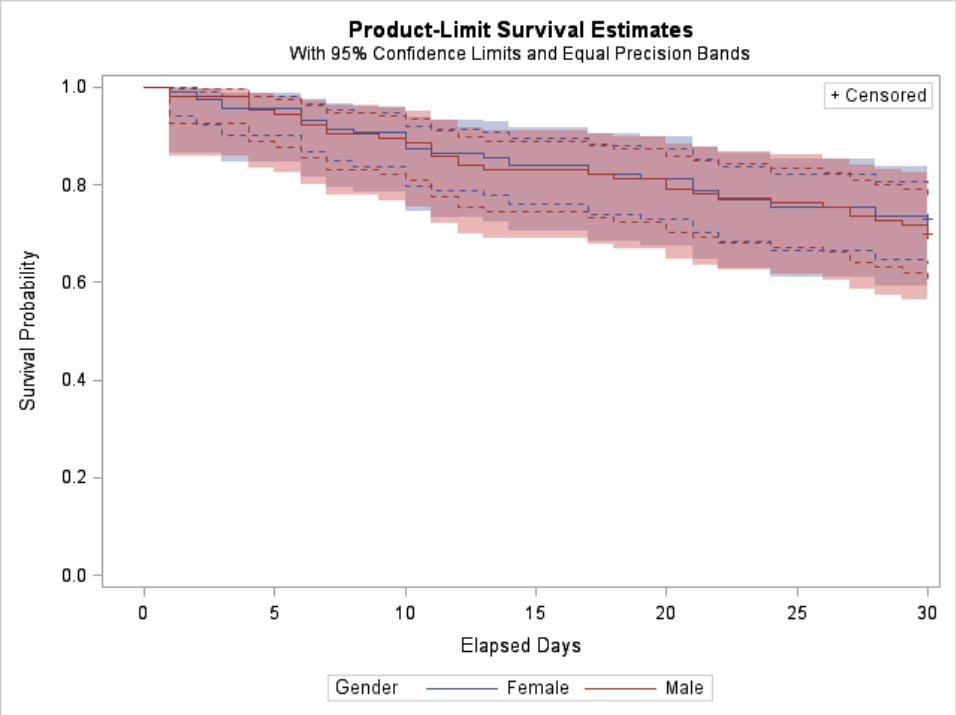


Figure 4.10: Product Limit Survival Estimate by Gender

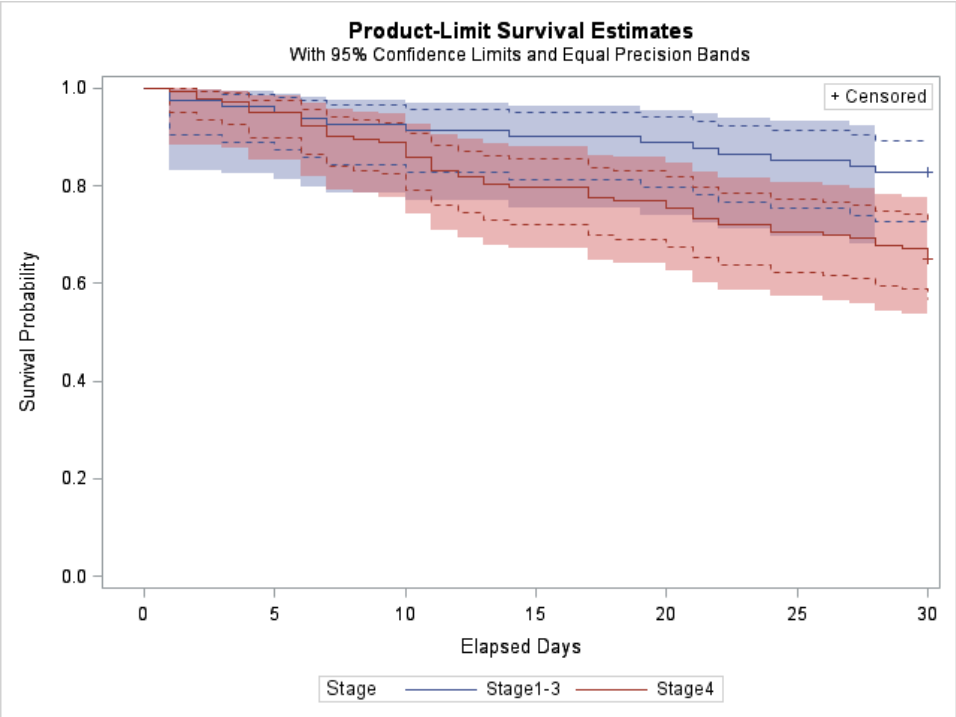


Figure 4.11: Product-Limit Survival Estimate by Stage

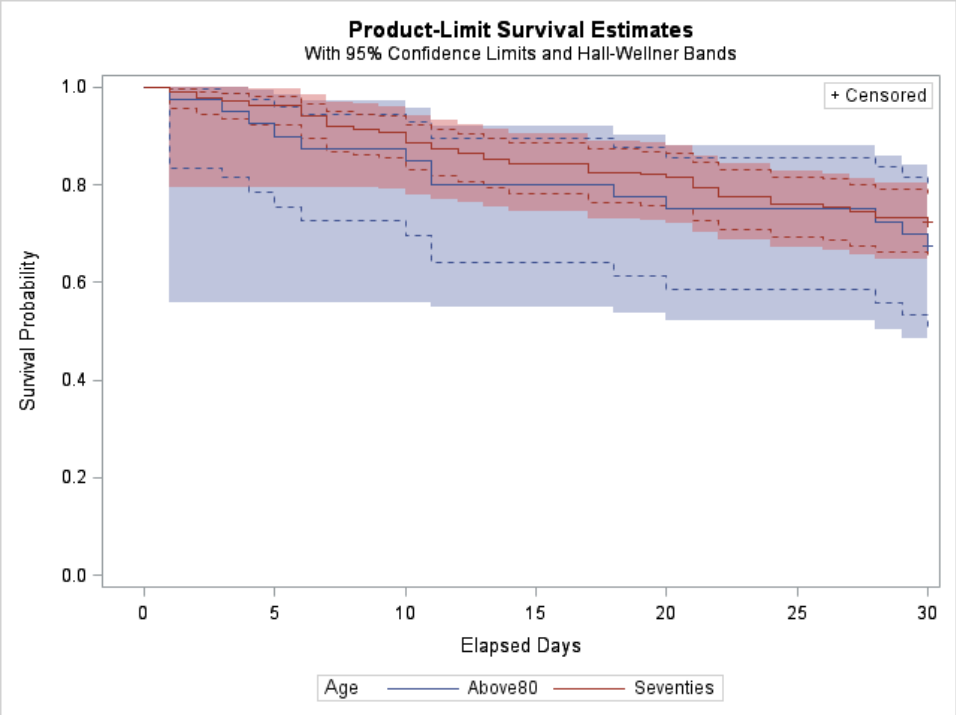


Figure 4.12: Product-Limit Estimate by Age

## 4.4 Proportionality Assumptions: Cox Proportional Hazard Model

Time-dependent covariates change at different rates for different individuals, so the ratios of their hazards cannot remain constant. The violation of the PH assumptions are equivalent to interactions between one or more covariates and time. That is, the Cox PH model assumes that the effect of each covariate is the same at all points in time. If the effect of a variable varies over time, the PH assumption is violated for that variable. To check the proportionality assumptions, a variety of residuals from the Cox regression models are considered. Martingale residuals are used to test for non-proportionality, using the method proposed by Lin, Wei, and Ying (1993) [1]. In the SAS ASSESS statement, the PH option requests an assessment of the proportional hazards assumption. For each covariate, the statement produces a graphical display of the empirical scores process, which is based on the martingale residuals.

The proportional hazards assumption for gender is presented in graph 4.14. The solid line shows the observed empirical score process and the dashed lines are the empirical score processes based on 20 random simulations that embody the proportional hazards assumptions. As in graph 4.14, the observed processes does not deviate from the simulated processes, which is evidence in favor of the proportional hazards assumption for gender. The lower right corner of the output gives a quantitative assessment in the form of a p-value for 1000 simulations. Over 82% of cases followed the simulated process. Thus, the quantitative values also provide evidence in favor of the proportional hazards assumption.

The proportional hazards assumption for patients' age is presented in graph 4.15. The solid line shows the observed empirical score process and the dashed lines are empirical score processes based on 20 random simulations that embody the proportional hazards assumptions. As the observed processes deviates distinctly from the simulated processes, there is evidence against the proportional hazards assumption for age. The lower right corner of the output gives a quantitative assessment in the form of a p-value for 1000

simulations. Only 32% of cases followed the simulated process. Thus, this is evidence against the proportional hazards assumption. The proportional hazard assumption for patients cancer stages is presented in graph 4.13. The solid line shows the observed empirical score process and the dashed lines are empirical score processes based on 20 random simulations that embody the proportional hazards assumptions. As the observed processes does not deviate from the simulated processes, it is evidence in favor of the proportional hazards assumption for cancer stages. The lower right corner of the output gives a quantitative assessment in the form of p-values for 1000 simulations. Only 76% of the cases followed the simulated process. This is evidence in favor of the proportional hazards assumption.

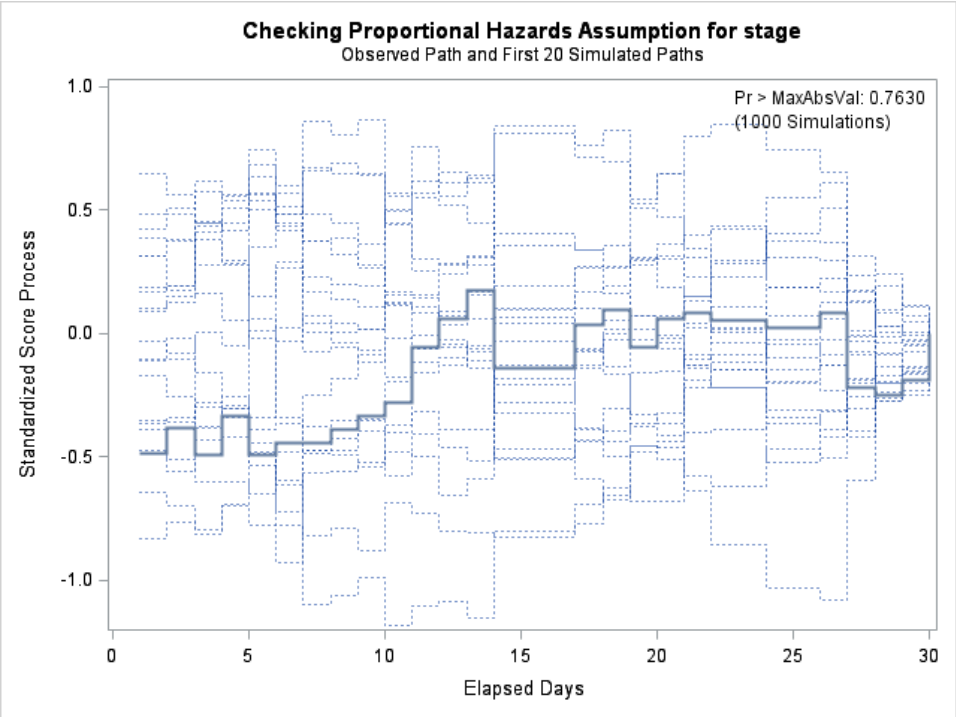


Figure 4.13: Checking Proportional Hazards Assumption for Cancer Stage

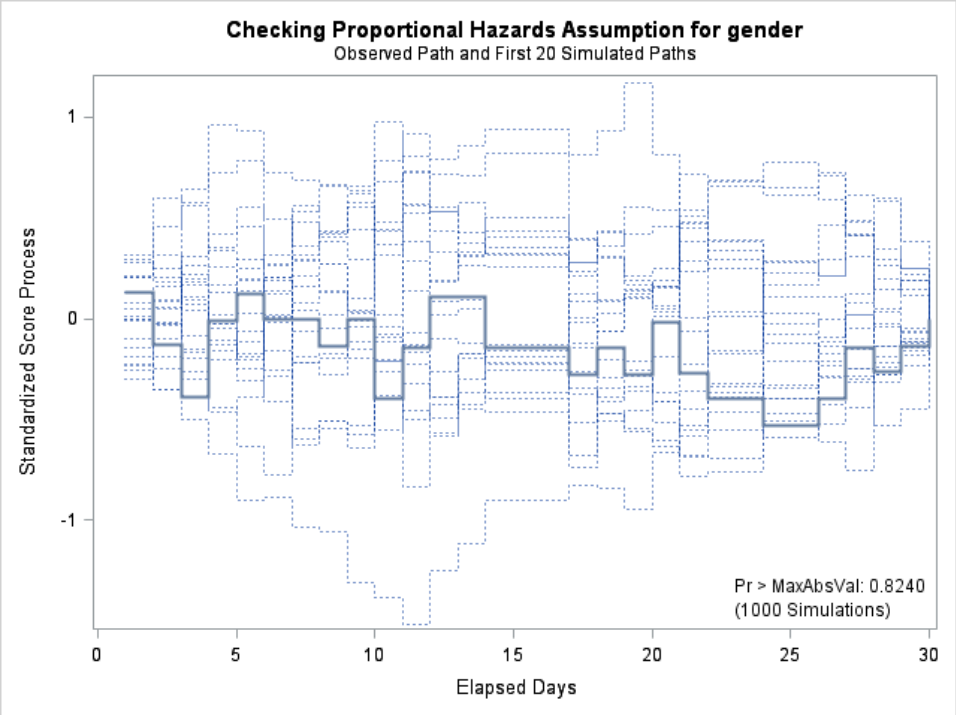


Figure 4.14: Checking Proportional Hazards Assumption for Gender

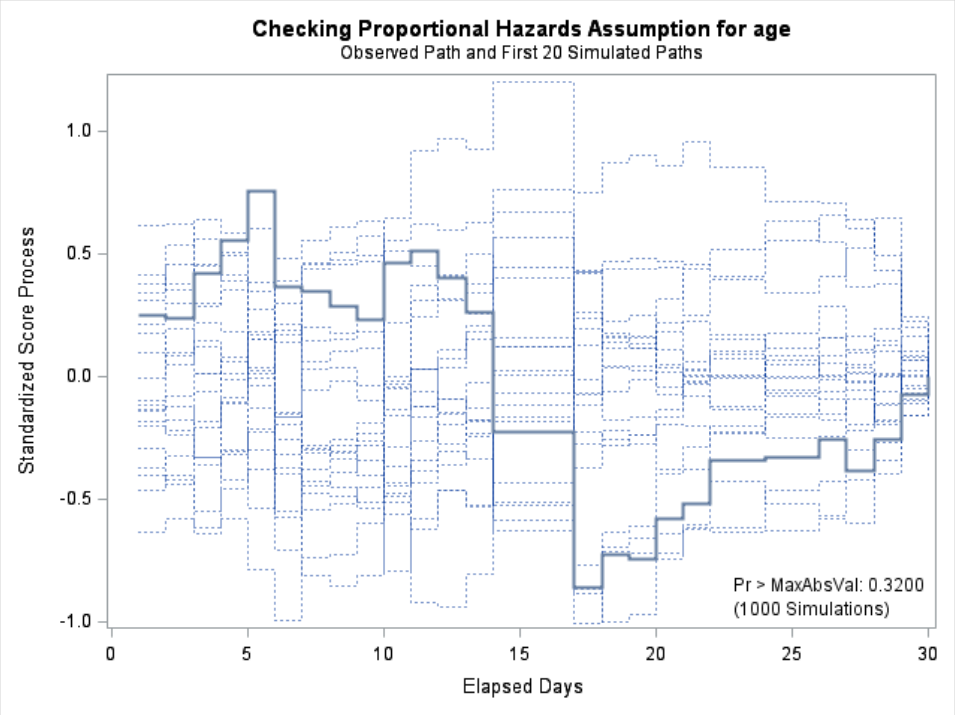


Figure 4.15: Checking Proportional Hazards Assumption for Age

## 4.5 Semi-parametric: Cox Proportional Hazard Model

In table 4.5, we see coefficient estimates and associated statistics. Here, the intercept is $\alpha(t)$, an arbitrary function of time, which drops out of the estimating equations. The Wald Chi-square test is for the null hypothesis that each coefficient is equal to 0. Statistics are calculated by squaring the ratio of each coefficient to its estimated standard error. The last column, labeled Hazard Ratio, is calculated with $\exp(\beta)$. For indicator variables, the hazard ratio can be interpreted as values of 1 and 0. We can interpret the hazard ratio as the ratio of the estimated hazard for those with a value of 1 to the estimated hazard for those with a value of 0 (controlling for other covariates). For quantitative covariates, a more helpful statistic is obtained by subtracting 1.0 from the hazard ratio and multiplying by 100. This gives the estimated percent change in the hazard for each unit increase in the covariate.

Table 4.5 displays the semi-parametric model with predictor variables gender, age and cancer stages. Table 4.5 shows cancer stage has the highest HR with the lowest p-value, compared to age and gender with relatively higher p-values. The HR for cancer stages after adjusting for age and gender is 1.92 which indicates that the hazard of readmission for the patients in stage 4 is about twice for the patients in stage 1-3. This result is

statistically significant with a p-value of 0.002 and a 95% confidence interval of (1.258, 2.915). Then we fit the model with cancer stage only and found that the HR is 1.88 with a p-value of 0.002. Further, the confidence interval was the same which indicates that the hazard for readmission is still about 2 times for patients in stage 4 compared to those in stage 1-3 while not considering age or gender. Then we developed the model separately modeling cancer stage and gender as well as cancer stage and age. We found as shown in table 4.7 that the HR of cancer stage is slightly higher with a value of 1.90 when considering age with a p-value 0.002 and a 95% confidence interval of (1.25, 2.87). This is compared to cancer stage and gender in table 4.8 with a HR of 1.87, a p-value of 0.003 and a 95% confidence interval of (1.24, 2.85).

Table 4.5: Maximum Likelihood Estimates of Regression Coefficients

| Parameter | Parameter Estimate | Standard Error | Chi-Square | Pr>ChiSq | Hazard Ratio | 95% Confidence Limit | |
|---|---|---|---|---|---|---|---|
| Gender | -0.061 | 0.026 | 0.059 | 0.813 | 0.941 | 0.568 | 1.559 |
| Age | 0.015 | 0.015 | 0.986 | 0.321 | 1.015 | 0.986 | 1.045 |
| Stage | 0.650 | 0.214 | 9.194 | 0.002 | 1.915 | 1.258 | 2.915 |

Table 4.6: Maximum Likelihood Estimates of Regression Coefficients of Cancer Stage

| Parameter | Parameter Estimate | Standard Error | Chi-Square | Pr>ChiSq | Hazard Ratio | 95% Confidence Limit | |
|---|---|---|---|---|---|---|---|
| Stage | 0.635 | 0.210 | 9.137 | 0.003 | 1.88 | 1.250 | 2.850 |

Table 4.7: Maximum Likelihood Estimates of Regression Coefficients of Cancer Stage and Age

| Parameter | Parameter Estimate | Standard Error | Chi-Square | Pr>ChiSq | Hazard Ratio | 95% Confidence Limit | |
|---|---|---|---|---|---|---|---|
| Age | 0.014 | 0.0146 | 0.939 | 0.3349 | 1.014 | 0.986 | 1.044 |
| Stage | 0.6422 | 0.2113 | 9.2344 | 0.0024 | 1.901 | 1.256 | 2.876 |

Table 4.8: Maximum Likelihood Estimates of Regression Coefficients of Cancer Stage and Gender

| Parameter | Parameter Estimate | Standard Error | Chi-Square | Pr>ChiSq | Hazard Ratio | 95% Confidence Limit | |
|---|---|---|---|---|---|---|---|
| Gender | -0.006 | 0.252 | 0.001 | 0.982 | 0.994 | 0.607 | 1.630 |
| Stage | 0.633 | 0.212 | 8.933 | 0.003 | 1.874 | 1.244 | 2.854 |

## 4.6 Parametric Model

A parametric survival model is one in which survival time is assumed to follow a known parametric distribution. Examples of distributions that are commonly used for survival time are the Weibull, the gamma and the exponential which is a special case of the Weibull. We first consider the exponential model which is the simplest parametric survival model in that the hazard is constant over time (i.e. $h(t)=\lambda$). The purpose of this parametric model analysis is to compare findings with Cox PH model. Table 4.9 shows that the estimates of parameters for cancer stage is the highest at -0.654 and is statistically significant with a p-value of 0.002. Similarly, table 4.10 shows the most widely used parametric model with estimated parameter values. We see from the table that the parameter estimate for cancer stage is the highest at -0.653 and this is statistically significant. The findings from the exponential and Weibull models are consistent.

Table 4.9: Analysis of Maximum Likelihood Estimates: Exponential Distribution

| Parameter | Estimate | Standard Error | 95% Confidence Limit | | Chi-Square | Pr>Chi-Sq |
|---|---|---|---|---|---|---|
| Intercept | 7.865 | 1.378 | 5.164 | 10.565 | 32.573 | <0.000 |
| Gender | 0.063 | 0.258 | -0.443 | 0.568 | 0.062 | 0.807 |
| Age | -0.015 | 0.015 | -0.044 | 0.015 | 0.960 | 0.326 |
| Stage | -0.654 | 0.214 | -1.074 | -0.234 | 9.301 | 0.002 |
| Scale | 1.000 | 0.000 | 1.000 | 1.000 | - | - |
| Weibull | 1.000 | 0.000 | 1.000 | 1.000 | - | - |

Table 4.10: Analysis of Maximum Likelihood Estimate: Weibull Distribution

| Parameter | Estimate | Standard Error | 95% Confidence Limit | | Chi-Square | Pr>Chi-Sq |
|---|---|---|---|---|---|---|
| Intercept | 7.864 | 1.378 | 5.164 | 10.565 | 32.573 | <0.000 |
| Gender | 0.063 | 0.2579 | -0.4426 | 0.568 | 0.062 | 0.807 |
| Age | -0.015 | 0.0150 | -0.0440 | 0.0146 | 0.964 | 0.326 |
| Stage | -0.654 | 0.214 | 1.074 | -0.233 | 9.304 | 0.002 |
| Scale | 1.000 | 0.000 | 1.000 | 1.000 | - | - |
| Weilbull | 1.000 | 0.000 | 1.000 | 1.000 | - | - |

## 4.7 Conclusions and Limitations

We considered models with gender, age and cancer stage variables and found that the cancer stage variable is highly associated with the risk of hospital readmission within 30 days following chemotherapy. Moreover, after adjusting for age, cancer stage is an even more significant predictor for risk of hospital readmission. The results from both semi-parametric and parametric models are consistent.

A limitation of our study is that we had a relatively small number of patients in cancer stages 1-3 compared to cancer stage 4. We had 10 patients in cancer stage 1, 27 in

cancer stage 2, 44 in cancer stage 3 and 143 in cancer stage 4. We thus combined the first three stages into a single category of cancer stages, with 81 patients, for the purpose of analysis. It would be useful to have a sample with more patients in each of the first three cancer stages, to explore and identify effects on the likelihood of readmission of the different stages of cancer.

Another limitation is the generalizability of the results of this study. A limited number of patients with multiple types of cancer during one-year period of time from a single health care facility has been considered for the study. Type of cancer may have acted as a mediator or intervening factor in the relationship between hospital readmission and cancer stage. Therefore, results may not be generalized to a similar cohort of patients in other facilities.

# Chapter 5

# Bibliography


[1] P.D. Allison. Survival analysis using SAS. In *Practical Guide*, pages 30–31. SAS, 2010.

[2] J.F. Bell, R.L. Whitney, Reed S.C., H. Poghosyan, R.S. Lash, K. Kim, A. Davis, R.J Bold, and J.G. Joseph. Systematic review of hospital readmissions among patients with cancer in the united states. In *Oncol Nurs Forum*, volume 44, pages 176–91, 2017.

[3] R. Farhin. Survival analysis of events on prostate cancer. In *Facts from Cancer Genome*. Unpublished, 2017.

[4] National Quality Forum. All-causes admissions and readmissions 2017. In *Technical Report*. Centers for Medicare and Medicaid Services, 2017.

[5] R. Solomon, N. Egorova, R. Franco, and N. Bickell. Thirty-day readmissions in metastatic cancer patients: Room for improvement? In *Mathematica policy research*. Journal of Clinical Oncology, 2017.

[6] S. Subramaniam. Understanding survival analyses. In *Clinical Epidemiology of Chronic Liver Diseases*, pages 33–39. Springer, 2019.

[7] Mathematica Policy Research under subcontract to Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation. Admissions and emergency department visits for patients receiving outpatient chemotherapy. In *Mathematica policy research*. Centers for Medicare and Medicaid Services, 2016.

# Appendix: SAS Code

```
*Read Excel Data with Return (Failure) Cases;
libname Ball "C:\Users\Md Akter Hossain\Desktop\THESIS\Ball Data\Combin";
PROC IMPORT OUT= Ball.main_import1 replace
DATAFILE= "C:\Users\Md Akter Hossain\Desktop\THESIS\Ball Data\03222019.xlsx"
     DBMS=EXCEL REPLACE;
     RANGE="maindata";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES; RUN;
data ball.main_import;
set ball.main_import1;
if elapse_days= 36 then elapse_days= 30; *elapsed days one 36 found in the data
set, I replaced it with 30;*
        run;
Data Ball.data_main;
set Ball.main_import
(rename =(cancer_diagno=cancer cancer_diagno_coded=cancer_code) drop=note);
        run;
*Read Excel Data with Right Censored Cases;
PROC IMPORT OUT= Ball.censor_import
DATAFILE= "C:\Users\Md Akter Hossain\Desktop\THESIS\Ball Data\03222019.xlsx"
     DBMS=EXCEL REPLACE;
     RANGE="Denominator$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES; RUN;
```

```
Data Ball.censor_data (rename=(patient_number=pid cancer_diagnosis=cancer_code));
        set Ball.censor_import; run;
*Combinded main and censor datasets;
Data Ball.Combined_1st;
        length chemo $150 cancer $117;
        set Ball.data_main Ball.censor_data; Run;
* Dataset without missing values* MAIN DATASETS;
data Ball.Combined_wtmis (keep=pid gender age stage elapse_days);
        set Ball.Combined_1st;
        where stage ne .; run;
* created EVENT variable (if returned within 30 days)for combined dataset;
data Ball.combinedw;
set Ball.combined_wtmis;
if elapse_days =. then do; elapse_days=30; event=0; end; /*elapse_days for
censored data with if statement*/ else event=1; run;
/* Exploratory data analysis*/
/*Distribution according to Gender and Event: Table-1*/
        proc freq data=ball.combinedw;
        table event * stage/chisq;
        run;
/*Distribution according to Elapsed days and Age: Table-2*/
proc means data=ball.combinedw nway;
        var age elapse_days;
        class event;
        output out=stats mean(age)=Mean_age std(age)=Std_age ; run;
/* Distribution of Mean Elapsed days by Gender: Fig 1/5;
/*For inset (Step-1)*/
ods noproctitle;
ods graphics off;
options nodate nonumber;
proc means data=ball.data_main nway;
        var elapse_days;
```

```
        class gender;
        output out=stats mean(elapse_days)=Mean_elapse_days
        std(elapse_days)=Std_elapse_days ; run;
/*Sort the data (Step-2)*/
proc sort data=ball.data_main out=gender_srt;
        by gender;
        run;
/*Match-merge the inset data with original data and label (Step-3)*/
data merged;
        merge gender_srt stats;
        by gender;
        label Mean_elapse_days="Mean (Elapsed Days)";
        label Std_elapse_days="Std(Elapsed Days)";
        run;
/*Create panel graph using Inset statement (Step-4)*/
proc format;
        value sex
        0=Female
        1=Male;
        run;
proc sgpanel data=merged;
        Title Distribution of patients Elapsed Days by Gender (n=64);
        panelby gender;
        histogram elapse_days;
        Density elapse_days;
        Inset Mean_elapse_days Std_elapse_days /
        position=topright Textattrs =(style=italic) /*title ="Averages"*/;
        format _numeric_ 8.1;
        format gender sex.;
        label elapse_days=Elapse Days;
        label  gender=Gender;
        run;
```

```
*Significance test;
proc ttest data=ball.data_main;
        class gender;
        var elapse_days;
        run;
/* Distribution of Mean Age by Gender: Fig 2/6;
/*For inset (Step-1)*/
ods noproctitle;
ods graphics off;
options nodate nonumber;
proc means data=ball.data_main nway;
        var age;
        class gender;
        output out=stats mean(age)=Mean_age std(age)=Std_age ;
        run;
/*Sort the data (Step-2)*/
proc sort data=ball.data_main out=gender_srt;
        by gender;
        run;
/*Match-merge the inset data with original data and label (Step-3)*/
data merged;
        merge gender_srt stats;
        by gender;
        label Mean_age="Mean(Age)";
        Label Std_age="Std(Age)";
        run;
/*Create panel graph using Inset statement (Step-4)*/
proc format;
        value sex
        0=Female
        1=Male;
        run;
```

```
proc sgpanel data=merged;
        Title Distribution of patients Age by Gender (n=64);
        panelby gender;
        histogram age;
        Density age;
        Inset Mean_age Std_age/
        position=topright Textattrs =(style=italic) /*title ="Averages"*/;
        format _numeric_ 8.1;
        format gender sex.;
        label  gender=Gender;
        run;
*Significance test;
proc ttest data=ball.data_main;
        class gender;
        var age;
        run;
/* Distribution of Mean Age by Event:Fig 3;
/*For inset (Step-1)*/
ods noproctitle;
ods graphics off;
options nodate nonumber;
proc means data=ball.combinedw nway;
        var age;
        class event;
        output out=stats mean(age)=Mean_age std(age)=Std_age ;
        run;
/*Sort the data (Step-2)*/
proc sort data=ball.combinedw out=event_srt;
        by event;
        run;
/*Match-merge the inset data with original data and label (Step-3)*/
data merged;
```

```
        merge event_srt stats;
        by event;
        label Mean_age="Mean(Age)";
        Label Std_age="Std(Age)";
        run;
/*Create panel graph using Inset statement (Step-4)*/
proc format;
        value eventcat
        0="No"
        1="Yes";
        run;
proc sgpanel data=merged;
        Title Distribution of patients Age by Event (n=64);
        panelby event;
        histogram age;
        Density age;
        Inset Mean_age Std_age/
        position=topright Textattrs =(style=italic) /*title ="Averages"*/;
        format _numeric_ 8.1;
        format event eventcat.;
        label  event=Event;
        run;
*Significance test;
proc ttest data=ball.data_main;
        class gender;
        var age;
        run;
* Distribution of Event by Cancer Stage: Table-3;
proc freq data=ball.combinedw;
        table event*stage/chisq;
        run;
/* Distribution of Mean Age by Stage:Fig 4/8;
```

```
/*For inset (Step-1)*/
ods noproctitle;
ods graphics off;
options nodate nonumber;
proc means data=ball.data_main nway;
        var age;
        class stage;
        output out=stats mean(age)=Mean_age std(age)=Std_age ;
        run;
/*Sort the data (Step-2)*/
proc sort data=ball.data_main out=stage_srt;
        by stage;
        run;
/*Match-merge the inset data with original data and label (Step-3)*/
data merged;
        merge stage_srt stats;
        by stage;
        label Mean_age="Mean(Age)";
        Label Std_age="Std(Age)";
        run;
/*Create panel graph using Inset statement (Step-4)*/
proc format;
        value stagecat
        1="Cancer Stage1";
        2="Cancer Stage2";
        3="Cancer Stage3";
        4="Cancer Stage4";
        run;
proc sgpanel data=merged;
        Title Distribution of patients Age by Stage (n=64);
        panelby stage;
        histogram age;
```

```
        Density age;
        Inset Mean_age Std_age/
        position=topright Textattrs =(style=italic) /*title ="Averages"*/;
        format _numeric_ 8.1;
        format stage stagecat.;
        label  stage=Cancer Stage;
        run;
*Significance test;
proc glm data=ball.combinedw;
        class stage;
        model age=stage;
        run;
        quit;
* Correlation among exploratory variables;
Proc corr data=ball.combinedw plots (maxpoints=10000)=Matrix(histogram);
var elapse_days age stage;
run;
/*Graphical Distribution of Elaspsed_Days*/
proc sgplot data=ball.data_main;
        Title Univariate Distribution of Elapse days;
        histogram elapse_days/fillattrs=(color=pink);
        density elapse_days/lineattrs= (color=cyan);
        density elapse_days/type=kernel lineattrs=(color=red);
        run;
/*Distribution of Age of Patients */;
proc sgplot data=ball.data_main;
        Title Univariate Distribution of Elapse days;
        histogram age/fillattrs=(color=pink);
        density age/lineattrs= (color=cyan);
        density age/type=kernel lineattrs=(color=red);
        run;
/*Goodness of FIT test for Elapsed_days by MALE & FEMALE*/
```

```
ods noproctitle;
proc univariate data=ball.data_main noprint;
        class gender;
        var elapse_days;
        format numeric 8.1;
        histogram elapse_days/normal;
        inset n median;
        format gender sex.;
        title "Distribution of Elapse Days";
        run;
/***SURVIVAL FUNCTION, KAPLAN-MEIER ESTIMATE: NONPARAMETIC MODEL***/
libname Ball "C:\Users\Md Akter Hossain\Desktop\THESIS\Ball Data\Combin";
*For Product Limit Survival Estimate WITHOUT Confidence Level (CL)
& Equal Precision (EP) bands;
proc lifetest data=Ball.combinedw plots=(s);
        time elapse_days*event(0);
        label elapse_days=Elapse Days;
        run;
*For Product Limit Survival Estimate WITH CL & EP bands;
proc lifetest data=Ball.data_main plots=s(CL CB=EP); *notable;
        time elapse_days*event(0);
        label elapse_days=Elapse Days;
        run;
*survival probability according to Gender;
proc format;
        value gendercat
        0=Female
        1=Male;
        run;
proc lifetest data=Ball.combinedw plots=s(CL);
        time elapse_days*event(0);
        format gender gendercat.;
```

```
        Label elapse_days=Elapsed Days gender=Gender;
        strata gender;
        run;
*P_L Estimate without CL EP;
        proc format;
        value stagecat
        1,2,3="Stage1-3"
        4="Stage4"; run;
proc lifetest data=Ball.combinedw plots=(s);
        time elapse_days*event(0);
        format stage stagecat.;
        strata stage;
        label elapse_days=Elapse Days stage=Stage;
        run;
*PL Estimate with CL EP FIG: FigSurPlot3;
proc lifetest data=Ball.combinedw plots=s(CL CB=EP);
        time elapse_days*event(0);
        format stage stagecat.;
        strata stage;
        label elapse_days=Elapsed Days stage=Stage;
        run;
proc format;
        value agecat
        low-79 = Seventies
        80-high = Above80;
        run;
proc lifetest data=Ball.combinedw plots=s(CL CB)notable;
        time elapse_days*event(0);
        format age agecat.;
        strata age;
        label elapse_days=Elapse Days age=Age;
        run;
```

```
*Cox Proportional Hazard Model Assumption for Stage, Age, Gender and Event;
ods graphics on;
proc phreg data=Ball.combinedw;
        Title Checking Proportional Hazard Assumption for Cancer Stage;
        model elapse_days*event(0)=stage gender age/rl;
        label elapse_days= "Elapsed Days";
        label stage="Cancer Stage";
        label gender="Gender";
        label age="Age";
        assess PH / RESAMPLE;
        run;
        title;
ods graphics off;
/*Semi-Parametric Model;*/
*Cox Proportional Hazard (CPH) Model for "Stage";
        proc format;
                value stagecat
                1,2,3="Stage 1-3"
                4="Stage 4";
        run;
*ODS graphics on;
proc phreg data=Ball.combinedw;
                model  elapse_days*event(0)=stage
                /ties= exact rl;
                strata stage;
                format stage stagecat.;
                label elapse_days="Elapsed Days";
run;
*ODS graphics off;
*Cox Proportional Hazard (CPH) Model for "Gender";
proc phreg data=Ball.combinedw plots(overlay=row)=s;
model  elapse_days*event(0)=gender
```

```
                /ties= exact rl;
                label elapse_days="Elapsed Days";
run;
*Cox Proportional Hazard (CPH) Model for "Age";
        proc format;
                value stagecat
                1,2,3="Stage 1-3"
                4="Stage 4";
        run;
proc phreg data=Ball.combinedw plots(overlay=row)=s;
model  elapse_days*event(0)=stage
                /ties= exact rl;
                label elapse_days="Elapsed Days";
                format stage stagecat.;
run;
*Cox proportional hazard model For ALL Covariates ;
proc phreg data=Ball.combinedw;
        *class gender (ref="1");
        model elapse_days*event(0)=gender age stage/rl;
        Label elapse_days="Elapsed Days" age="Age" stage="Cancer Stage"
        gender="Gender";
        run;
*Cox proportional hazard model For Age and Sage;
proc phreg data=Ball.combinedw;
        *class gender (ref="1");
        model elapse_days*event(0)=age stage/rl;
        Label elapse_days="Elapsed Days" age="Age" stage="Cancer Stage";
        run;
/*Cox Proportional Hazard (CPH) Model Age*/
ODS graphics on;
proc phreg data=Ball.combinedw plots(overlay=row)=s;
model  elapse_days*event(0)=age
```

```
                /ties= exact rl;
                label elapse_days="Elapsed Days"; run;
ODS graphics off;
/* Hazard Ratio for age*/
proc phreg data=Ball.combinedw;
        model elapse_days*event(0)=age
         /ties=exact rl;
         Label elapse_days="Elapsed Days" age="Age" stage="Cancer Stage";
         hazardratio age/units=5 10;
         run;
*Cox proportional hazard model with Stage and Gender ;
proc phreg data=Ball.combinedw;
        *class gender (ref="1");
        model elapse_days*event(0)=stage gender/rl;
        Label elapse_days="Elapsed Days" gender="Gender" stage="Cancer Stage";

*PARAMETRIC MODEL;
*Exponential Regression coefficient;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=age gender stage/d=exponential;
label elapse_days= Elapsed Days age=Age gender=Gender; run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=stage/d=exponential;
label elapse_days= Elapsed Days stage=Stage; run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=gender/d=exponential;
label elapse_days= Elapsed Days gender=Gender; run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=age/d=exponential;
label elapse_days= Elapsed Days age=Age; run;
*Weiball Regression Coefficient;
proc lifereg data=Ball.combinedw plots=(probplot);
```

```
model elapse_days*event(0)=gender age stage/d=weibull;
label elapse_days=Elapsed Days;
run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=stage/d=weibull;
label elapse_days= Elapsed Days stage=Stage; run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=gender/d=weibull;
label elapse_days= Elapsed Days gender=Gender; run;
proc lifereg data=Ball.combinedw plots=(probplot);
model elapse_days*event(0)=age/d=weibull;
label elapse_days= Elapsed Days age=Age; run;
proc phreg data=Ball.combinedw;
        model elapse_days*event(0)=gender stage;
        assess PH / RESAMPLE;
        title1 font="Times new roman/bold" Checking proportional hazard
        assumption for Gender;
        title2 font="Times new roman"Observed path & first 20 simulated path;
        run;
/*Checking PH assumptions using log[log] survival probability:
-Source Alan B. Cantor page#136;
 *Note: if we make Age one category (below 69 years)* results looks better;
 *The following test was suggested by Dr. Begum*/
proc lifetest plots= (loglogs) nocensplot;
        time elapse_days*event(0);
        label age=Age elapse_days=Elapse Days;
        strata age;
        format age agecat.; run;
proc lifetest plots= (loglogs) nocensplot;
        time elapse_days*event(0);
        label gender=Gender;
        strata gender;
```

```
        format gender gendercat.; run;
/*Schoenfeld residual*/
proc phreg data=Ball.combinedw;
        model elapse_days*event (0)=gender stage age/
        ties=efron;
        output out= Ball.combinedw_a ressch=schgender schstage schage;
        run;
*ods powerpoint close;
data Ball.combinedw_b;
        set Ball.combinedw_a;
        elapse_days1=log(elapse_days);
        elapse_days2=elapse_days**2;
proc corr;
        var elapse_days elapse_days1 elapse_days2 schgender schstage schage;
        run;
proc means data=Ball.combinedw maxdec=2;
        class event gender;
        format gender gendercat. age agecat. event eventcat.;
        var age;
        run;
proc freq data=ball.combinedw (rename=(event=Event gender=Gender));
        tables event*gender;
        format event eventcat. gender gendercat.;
        run;
proc freq data=Ball.combinedw;
        table event*stage;
        format gender sex.;
        run;
Proc means data=Ball.combinedw maxdec=2;
        var age;
        class gender;
        format gender gendercat.;
```

```
        run;
proc print data=combinedw;
        run;
* Survival and hazard functions;
proc lifetest data=Ball.combinedw method=act plots=(s(name=Actsurv),
   h(name=Acthz)) notable;
   time elapse_days*event(0);
   run;
* Survival and hazard functions by sex;
proc lifetest data=Ball.combinedw method=act plots=(s(name=Actsurv),
    h(name=Acthz)) notable;
        time elapse_days*event(0);
        format gender sex.;
        strata gender;
        run;
*Paremetic and semi-parametric model;
*exponential regression coefficient;
proc lifereg data=Ball.combinedw;
        model elapse_days*event(0)=gender age stage/d=exponential;
        run;
*Weibull regression coefficient;
proc lifereg data=Ball.combinedw;
        model elapse_days*event(0)=gender age stage/d=weibull;
        run;
```