PROPENSITY SCORE AND SURVIVAL ANALYSIS FOR LUNG CANCER

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mohammad Gulam Mostofa

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

November 2020

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Propensity Score and Survival Analysis for Lung Cancer

By

Mohammad Gulam Mostofa

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr
Chair

Dr. Rhonda Magel

Dr. Changhui Yan

Approved:

November 20, 2020
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Propensity scores were used to assess covariate balance between black and white groups in each lung cancer stage of a large data set. Pairwise log rank tests were used to test the equality of survival distribution for treatment and race groups. Cox regression models were used to estimate the hazard ratios for each treatment in all stages. In stage one, radiation and surgery were found the best treatment. In stage two, treatment of chemotherapy was found as the best option. Radiation and chemo were found to be the best treatment combinations in stage three. Based on hazard ratios, the treatment chemo was the best for stage four. Statistically significant differences in survival curves were found between different gender and race combinations in stages one and three, but not in stages two or four.

ACKNOWLEDGEMENTS

I am grateful to Almighty Allah for the good health and well-being that were necessary to complete this research. I would like to thank the Department of Statistics at North Dakota State University for all the support I was given to excel in my studies. I would like to thank Professor Dr. Choi for his research idea and support to continue my research. I wish to express my profound gratitude to Dr. Orr for her advisory, patience, and suggestions to complete my research. She helped me a lot with the writing of this paper. My sincere appreciation goes to Professor Dr. Magel, for her continuous suggestions, financial support, and for making resources available during my study. I would like to thank my committee member, Dr. Changhui Yan, for his encouragement, insight, and comments.

I wish to thank my wife Dr. Fatima, my son Mohiuddin, my daughter Mubassira, my parents, my mother-in-law, and my father-in-law for their love, prayer, and support throughout my studies. It would have been impossible to finish my research and thesis without their support.

DEDICATION

I dedicate this thesis to my wife, my kids, and my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDC	Disease Control and Prevention
ACS	American Cancer Society
PSM	Propensity Score Matching
SEER	Surveillance, Epidemiology, and End Results
NSCLC	Non-small Cell Lung Cancer
TRT	Thoracic Radiotherapy
CT	Chemotherapy
SCLC	Small Cell Lung Cancer
HR	Hazard Ratio
RT	Radiotherapy
SPLC	Second Primary Lung Cancer
IPLC	Initial Primary Lung Cancer
NIH	National Cancer Institute
KM	Kaplan-Meier
SMD	Standardize Mean Difference

CHAPTER 1: INTRODUCTION

Cancer is a great public health concern around the world. It is the second leading cause of death, with the death rate for lung cancer higher than that of the breast, colorectal, pancreatic, and prostate cancer in the United States (“Centers for Disease Control and Prevention”). The death rate and incidence rate both are high for breast and lung cancer. In 2018, the projected new cases of lung and breast cancers are 234,030 and 268,670, respectively. Females are mostly affected by breast cancer due to a lack of breastfeeding and hormone levels. However, lung cancer is common for both males and females. The number of new cancer patients and death rates are recorded every year by the American Cancer Society through Surveillance, Epidemiology, and End Results (SEER) program (Siegel et al. 2020). There are four kinds of lung cancer: large cell carcinomas, small cell carcinomas, squamous cells, and adenocarcinomas. It is obvious smoking is the cause of most lung cancers. Due to smoking, the relative risk of lung cancer is higher than any other cancer. In the US, lung cancer incidence also varies due to racial disparities. The risk and epidemiologic factors are determined for all cancers according to age, gender, race, sex, socio-economic status, and other factors (e.g. smoking habit, family history, health conditions) (McCarthy et al. 2012). Lung cancer control and proper decision-making regarding treatment are always challenging for patients, doctors, and other personnel. There are lots of studies that address lung cancer survival, but there is no study on the estimation of a balanced score based on racial groups. First, the goal of this study is to get balanced sample data using propensity score methods and a statistical approach to predict the survival rate of lung cancer patients. Since SEER lung cancer data are for observational studies, the goal of propensity score matching in this study is to balance white and black lung cancer patients with multiple relevant covariates. The purpose of the study is to estimate the causal effect on lung cancer survival rate based on the race groups and reduce the bias due to group differences. The main objective of this study is to select the best treatment groups using matched data in all stages.

Moreover, we compared treatment groups with different covariates and designed a statistical model to predict the survival rate. Overall, this research focused on survival analysis of lung cancer and the factors which significantly contribute to the chance of survival of lung cancer patients. Our results will be helpful for doctors, policymakers, patients, and other personnel to make good decisions regarding lung cancer treatment in the future.

CHAPTER 2: LITERATURE REVIEW

Zhang et al. (2018) estimated the radiotherapy advantage for stage 4 lung cancer patients with non-small cell lung cancer (NSCLC). The researchers used the nearest neighbor (1:1) matching method for a balanced score. They elucidated that stage four lung cancer patients' survival rate significantly improved if patients received radiotherapy. For small cell lung cancer, Thoracic radiotherapy (TRT) in combination with chemotherapy (CT) improved patient's survival in comparison to only CT. They showed that 12,019 patients received CT, and 2348 patients received CT and TRT. The risk of death is higher for CT (HR=1.74) rather than CT and TRT (HR=1.70).

Govindan et al. (2006) showed within all lung cancer histology types that small cell lung cancer (SCLC) declined from 17.26 percent in 1986 to 12.95 percent in 2002. On the contrary, the proportion of women diagnosed with small cell lung cancer augmented from 28 percent in 1973 to 50 percent in 2002. The researchers further estimated the 2 and 5-year survival rates for both limited-stage small cell lung cancer patients and extensive-stage small cell lung cancer patients throughout the study period.

Lu et al. (2019) demonstrated that the average incidence of lung cancer was 59.0/100,000 people for the years from 1973 to 2015. Specifically, the incidence rate was at an apex in 1992 and then constantly declined. The male incidence rate was higher than female and black patients' incidence rate was higher than that of other racial groups patients. Furthermore, the lung cancer surgical rate was approximately 25 percent and the patient's chemotherapy burgeoned in 1985, while the radiation therapy was descending during that time. The surgical rate for small cell lung cancer was less than non-small cell lung cancer but CT was higher for small cell lung cancer.

Indeed, CT and RT were higher for an advanced stage than that of the early stage. The researcher viewed that the 5-year survival rate was raised with time but was less than 21%.

Lim et al. (2017) examined the propensity score matching and survival with postoperative radiotherapy in thymic carcinoma using lung cancer SEER data. Most importantly, the researchers showed that if the patients received postoperative radiotherapy their overall 5-year survival rates were better, both before and after matching. Besides, it was noticed that if the patient's age is greater than 63, treatment with debulking surgery and no received post-radiotherapy, resulted in a similar survival rate for stage III and Stage IV. On the other hand, if the patients received postoperative radiotherapy, then their survival was better for stage III and Stage IV. Furthermore, the researchers examined that the hazard ratio for the tumor is 0.31, node-negative (HR=0.58), and surgical extent of local excision (HR=0.44).

Che et al. (2018) ascertained from their study population that 21.3 percent of the lung cancer patient in stage I, 6.2 percent of the lung cancer patient in stage II, and 72.5 percent of the lung cancer patient in stage III. Furthermore, the hazard ratio was 0.33 in stage I, 0.51 in stage, 0.46 in stage III, accordingly. It concluded that for patients who did the surgery, their survival rate was significantly higher for overall survival compared to lung cancer-specific survival.

Lampaki et al. (2016) exposed the median survival time for small cell lung cancer patients was 5 years if the patient received treatment, but this medium time was less than 7 percent compared to the overall survival time. On the contrary, if the small cell cancer patients did not receive treatment, then the median survival time was 2 to 4 months. The researcher concluded CT with RT was the best treatment for limited disease and extensive disease patients for small cell lung cancer.

Caprario et al. (2013) showed the black female survival rate was higher with factors such as limited stage disease, receiving treatment, and lower comorbidity score. From the quantile regression, the researchers showed if patients received chemotherapy, then the median survival time improved by 6.5 months.

Earle et al. (2000) revealed that about 22 % of younger patients with comorbidities received CT during metastatic non-small cell lung cancer. The researcher explored the patients who received CT with nonmedical factors such as higher socioeconomic status, nonblack race, who were living in the Seattle/Puget Sound or Los Angeles SEER regions, etc.

Fu et al. (2005) evinced that the median age at diagnosis was 66 years for both men and women. The researcher used 228,572 patients from the SEER dataset where 35.8% were female. Among those females, 40.9% were less than 50 years and 35.4% were older patients. The authors showed that the survival rate was higher for females at all stages of the disease, but the male gender was a negative prognostic factor.

Hubbard et al. (2012) estimated the survival rate was 55.4%, 33.1%, and 24.3% for 10, 15, and 18 years, respectively. Hazard ratio (0.88) for squamous cell cancers was lower than that of adenocarcinoma (HR=1.08).

Thakur et al. (2018) elucidated that Second Primary Lung Cancer (SPLC) was the highest for both males and females irrespective of age. On the contrary, Initial Primary Lung Cancer (IPLC) was the highest for a young age patient. The researchers considered the young age group (20—49) years and estimated the Standardized Incidence Ratio (SIR) was the highest for females irrespective of race. They got SPLC was 1.10 % per patient per year and cumulative risk factors increased over time. Their median survival time 59 to 62 months. The surgery rate was higher percent (76.7%) of white patients compared to black patients (64 %).

Austin (2011) compared baseline characteristics between treated and untreated patients and estimated the effects of treatment outcomes using the propensity score. He concluded that exposure of interest was a receipt of smoking cessation before hospital discharge.

Austin (2010) showed that the propensity score reduced the effects of confounding factors. The author compared the regression-based method and the propensity score method for the analysis of observational studies. The researcher estimated the four different propensity score methods such as covariate adjustment, stratification approach, matching approach, and inverse probability treatment weighting. The author showed that risk differences, variance estimation, and MSE using propensity score matching.

Lonardo et al. (2014) estimated matched patients between propofol-lorazepam (2,250) and propofol -midazolam (1,054) groups. The researchers used propensity score matching to achieve baseline patient characteristics and assessed the balance of measure between two groups.

CHAPTER 3: METHODOLOGY

3.1. Propensity Score Matching

Propensity score matching is a statistical matching method which was introduced by Rosenbaum and Robin in 1983. The propensity score is the predicted probability of being a particular race given the set of covariates. Baseline covariates are expected to be balanced between white and black groups based on the propensity score. The main goal for propensity score is to achieve balanced scores which display similar propensity score distributions in the matched groups. For statistical analyses, the propensity score is widely used in observational designs, which are more feasible but can be more challenging to analyze due to a lack of balance in the covariates between treatment groups (race groups). Thus, propensity score matching is used to achieve this balance between treatment groups (Littnerova et al. 2013). It is one of the best tools to investigate the treatment effects where potential bias might exist. If we use the imbalanced observations for survival analysis, it would give us biased results. Confounding variables affect lung cancer survival. For example, socio-economic factors such as income, education, and employment can have a hidden effect on lung cancer patient's survival. To control potential baseline confounding factors across the groups we used the propensity score method to give us balanced scores. It plays a significant role in improving the accuracy of statistical inference. Since SEER lung cancer data is observational, the utility of propensity score matching is to balance white versus black groups with multiple relevant covariates.

In our study, race groups represent the treatment indicator where Z=1 if the race is white and Z=0 if the race is black. Let's assume that X_i= (X_1, X_2, X_3, X_4, X_5, $X_{6,}$ X_7) observed covariates, where X_1= Age, X_2= Gender, X_3=Marital status, X_4=Radiation, X_5=Surgery, X_6=Chemo, X_7=Stages

The formal definition of a propensity score (PS) is as follows,

$$PS = P(Z_i = 1|X_i)$$

The probability for each patient in both groups ranges from 0 to 1 (0<PS<1) (Rosenbaum and Rubin 1983).

Propensity score matching analysis was done in R with the "MatchIt" package. MatchIt plays a significant role to reduce the dependence of causal inferences and generated balanced data. We explored individual observations with matching covariates between white and black groups. Sometimes, it is very difficult to find exact matches. Thus, MatchIt implemented closeness or distance that generated matched data. We used Nearest Neighbor Matching which was the default method in MatchIt packages (Daniel Ho, Kosuke Imai, Gary King 2011). MatchIt () function worked under MatchIt package for Nearest Neighbor Matching. The race was on the left side of the tilde which is our treatment variable. On the right side of the tilde were the variables we wanted to match. Assessing the balance of covariates is done by MatchIt after matching including mean difference. Furthermore, MatchIt enables selected well-matched subsets of original race groups. The "tableone" package was used to compare baseline characteristics between two groups.

Standardized Mean Difference (SMD) is a very popular statistical method for balance diagnostic. We used SMD to check the balance of covariates after propensity score between black and white groups. If SMD is greater than 0.1, it indicates imbalance. In our study, there was no imbalance (SMD<0.1) of covariates between two groups after propensity score matching. Therefore, we do not need to check the misspecification for propensity score. Based on the above reasons, it is justified to use the balance score for further analysis since the balance condition was fulfilled. The SMD formula as follows:

$$\text{SMD}=\frac{\hat{P}_1-\hat{P}_2}{\sqrt{\frac{[\hat{P}_1(1-\hat{P}_1)+\hat{P}_2(1-\hat{P}_2)]}{2}}}$$

here, $\hat{P}_1$ and $\hat{P}_2$ are dichotomous variables for white and black groups.

Since age and stages are categorical baseline variables with k levels, we used the multivariate Mahalanobis distance method to generalize the standardized difference metric for the multinomial sample.

$$\text{Suppose } T=\left(\widehat{P_{12}},\widehat{P_{13}},\dots\dots\dots.\widehat{P_{1k}}\right)$$

$$C=(\widehat{P_{22,}}\widehat{P_{23}},\dots\dots\dots\dots.\widehat{P_{2k}})$$

here, T= White group, C= Black group, $P_{JK}=\Pr(\text{category } k|\text{Race group } j)\,, j\in\{1,2\},$ and $k\in$

$$\{2,3,\dots..,k\}$$

The standardized difference is defined as follows:

$$d=\sqrt{(T-C)'S^{-1}(T-C)}\quad,$$

where S is a $(k-1)\times(k-1)$ covariance matrix. The covariance matrix S is defining as follows (Yang and Dalton 2012):

$$S=[S_{kl}]=\begin{cases}\dfrac{[\widehat{P_{1k}}\left(1-\widehat{P_{1k}}\right)+\widehat{P_{2k}}(1-\widehat{P_{2k}})]}{2}, & k=l\\ \dfrac{[\hat{P}_{1k\,\widehat{P_{1l}}+\widehat{P_{2k}}\,\widehat{P_{2l}}}]}{2} & ,k\ddagger l\end{cases}$$

3.2. Survival Analysis

Survival analysis is consistent with a set of statistical approaches used to study the time for an event of interest to occur. There are three different types of basic concepts of survival analysis: (i) survival time and event, (ii) censoring, and (iii) survival function and hazard function. Cancer studies include two important measures that correspond to the time between response to a recurrence of the disease and treatment such as the time to death and the relapse-free survival tim*e*; this is known as event-free survival time. For survival analysis, time is considered as days, weeks, months, and years. In this study, months were considered as survival time.

3.2.1. Survival analysis related to two functions

The survival probability and hazard probability are two related probabilities that are used to describe survival data. The probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t is known as the survival probability which is also known as the survivor function $S(t)$, The probability that an individual who is under observation at a time t has an event at that time is defined as the hazard and denoted by $\lambda(t)$

Distribution of time to an event: First we define survival time and specific value for survival time. We denoted, $T =$ survival time, where $T \geq 0$ and random variable.

$$t = \text{specific value for } T\ ,$$

where $T > t,$ If T is a continuous random variable, so density function is $f(t)$.

$$f(t) = (dF(t)))/dt, F(t) = \int_0^t f(\mu)du,$$

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

The survival function $S(t)$ is a decreasing function over time taking on the value. If T is a continuous random variable, then

$$\text{S(t)}= \int_t^\alpha f(\mu)du, f(t) = \text{-(d s(t)))/dt.}$$

The hazard function denoted by $\lambda(t)$ represents the probability condition of death at time t after survival time.

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

If dt is very small,

$$\lambda(t). dt \approx \frac{P(t \leq T < t + dt)}{T > t}$$

$$\lambda(t) = lim_{dt\rightarrow 0} \frac{P(t \leq T < t + dt)}{P(T \geq t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} = \int \frac{dlog\big(S(t)\big)}{dt}$$

$$\lambda(t) = \int_0^t \lambda(\mu)du = -\log\big(S(t)\big)$$

Here, $\lambda(t)$ is the cumulative hazard function (DG Kleinbaum 2006).

3.2.2. Kaplan Meier method

Kaplan Meier method is a nonparametric method (Kaplan and Meier 1958) which compares and estimates survival function *S(t).* It is a good method for survival curve estimation because there is no assumption of an underlying probability distribution. It also deals with censor information in a very easy way. Kaplan Meier estimate is a step function because it depends on the total number of observed failures at the time t_j and the number of risks at the time t_j. The Kaplan Meier estimator of survival function is as follows:

$$\hat{S}(t) = (T > t) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j})$$

Where $t_1,t_2,\ldots\ldots t_n$ are observed event times

d_j = number of events at the time t_j

$n_{j=}$number of patients at risk at the time t_j

$\frac{d_j}{n_j}$=Proportion of failed at the event time at t_j

$1 - \frac{dj}{n_j}$ =Proportion of surviving the event time at t_j

The main advantage of the Kaplan Meier method is to estimate the survival probability of an event which is demonstrated by the y-axis. In our study months were considered survival time, which is represented by the x-axis. The survival probability is one if time is zero. Another benefit is that it is considered right censoring, which showed the vertical tick marks in the survival curves. The survival analysis was executed using a "survival package" in R to estimate the survival function for different groups.

3.2.3. Log rank test

The Log rank test is a hypothesis test that was introduced by Mantel (1966) and Peto and Peto (1972). It was used to test the equality of survival distribution for two or more groups. We compared survival distribution stratified by two race groups using this test. The test statistic is approximately a chi-square with (n-1) degrees of freedom.

The test statistic was computed by chi-square test which was

$$\chi 2 = \sum_{j=1}^{n} \frac{\left(O_j - E_j\right)^2}{E_j} \sim N(0,1) \text{ ; under } H_0 \quad \text{with } (n-1)df$$

Where, O_j = observed number of failures in group j,

E_j = expected number of failures in group j.

The expected number of failures for one group is the sum of the expected number of failures at the time of each failure It can be calculated as the risk of failure at that time multiplied by the number alive in a group (Bewick et al. 2004).

We also compared survival distribution for seven treatment groups by pairwise log rank test. We used “survminer package” in R with the “pairwise_survdiff” function to get the P value for a pair of groups. If the P value is less than 0.05 for each group, it concludes that the survival distribution for the pair of treatment groups is different, but if the P value is larger than 0.05, it indicates the pair of treatment groups have similar survival distribution.

3.3. Semi-parametric Methods

3.3.1. Cox proportional hazard model

This model was first introduced by D.R Cox in 1972 (Cox 1972). This model is called a semi-parametric model because the proportional assumption exists, but hazard distribution is not specified here. It estimates which variables contribute to the risk. It also depends on the proportional hazard assumptions. The main assumption of this model is that hazard is constant over time. It allows investigating the influence of factors on the rate of a specific event happening at a specific point in time. Moreover, we computed the hazard ratio for treatment groups, gender, and race groups. We denoted the hazard function by $\lambda(\mathrm{t})$. The main feature of the proportional hazard is that the baseline hazard does not include the covariates, but it is a function of it. On the contrary, the exponential part does not include the time but includes the covariates. The covariates express the multiplicative effects of a hazard. The proportional hazard equation is as follows:

$$\lambda(\mathrm{t}) = \lambda_0(\mathrm{t}) \exp(\beta X) \quad \text{, for all } t \geq 0$$

Here, $\lambda_0(\mathrm{t})$ is the baseline function. β is the vector of the regression coefficient. A set of covariates is expressed by X. We consider $\exp(\beta)$ for a hazard ratio of covariates. X indicates the explanatory variables. We used age, gender, marital status, and treatments groups.

3.3.2. Cox regression

The cox model we considered is a multiple regression with the logarithm of hazard based on covariates X (i.e. Treatment groups, age, race, gender, and marital status). In this study, the selected covariates describe the multiplicative effect of lung cancer survival. For example, if the value of covariates increases, the hazard increases. Therefore, the length of survival would be decreased. We computed the hazard ratio for six treatment groups and other covariates.

Suppose $\lambda_2(\mathrm{t}) = \text{black group patients},$ $\quad \lambda_3(\mathrm{t}) = \text{White group patients},$

the hazard ratio for race group= $\frac{\lambda 2(t)}{\lambda_3(t)} = \frac{\exp(\beta)}{\exp(\beta')},$ for all $t \geq 0$

Hazard ratio can be greater than 1 because it is not a probability. If a covariate with a hazard ratio is greater than one (i.e. β is greater than zero) is called bad survival. On the other hand, if one covariate with a hazard ratio is less than one (i.e. β is less than zero) is called good survival.

3.4. Data Description

We used surveillance, epidemiology, and end results (SEER) data for lung cancer. This data is very reliable and dependable. National Institute of Health (NIH) organized and gave support to collect cancer incidence from the population-based cancer registries covering about 34.6 percent of the U.S. population. They follow up with patients for vital status as observed in the NIH.. We used lung cancer incidence data from 2005 to 2016. This SEER data is very a good source for various cancer statistics for the US. It appeared that cancer mortality statistics, survival, and prevalence are also available on this website. The SEER Cancer Statistics Review (CSR) demonstrated about 6.3 percent of men and women were diagnosed with lung and bronchus cancer based on incidence and mortality. The American Joint Committee on Cancer (AJCC) first classified lung cancer stages by T (the primary tumor), N (regional lymph nodes), and M (distant metastasis). Subsequently, AJCC used a numeric system to describe lung cancer stages which started from Stage0 to stage IV. Stage0 indicates in situ which describes diseases to the surface. Stage I describes evidence of cancer growth and tissue of origin. Stage II describes cancer in the limited local spread. Stage III indicates the extension of local and regional spread. Stage IV describes distant metastasis. In our study, we used lung cancer stages from Stage I to Stage IV. One important point we need to discuss is that stage2 total observations were lower than that of other stages observations as shown in Table 2 because data collection from the source registries was incomplete and inaccurate. SEER data has some limitations which include incompleteness of individual-level data collection for specific cancer risk and treatment. According to SEER reports, tumor recurrence data were presently not collected. Consequently, the effectiveness of salvage therapy, correlates of local, regional, and distant control, and progression-free survival cannot be assessed. Information gaps in treatment and follow-up about a cancer behavior can occur in non-SEER regions, out of SEER regions, and bias conclusions. Therefore, inaccuracies in the source data can occur either due to the available data to the regional registrars for coding incorrect or miscoding the data transmitted to NCI by the regional registries. Therefore, miscoding, inaccuracies, and incompleteness were causes of lower observations in stage2 (Duggan et al. 2016).

3.5. Data Preprocessing

Data pre-processing is important to prepare the original data. For improving the data quality, pre-processing is a good step to transform the original data into real data which is used to apply for analysis. Data pre-processing is considered to handle unknown and missing values. For instance, survival months, and other variables contained unknown and missing values. Lung cancer stages were classified based on the code which was defined by the SEER.

Table 1. Each year per stage censor information for matching dataset

Year	Stage1	Stage2	Stage3	Stage4	Total censor
2005	136 (2.77)	36 (0.73)	61 (1.24)	72(1.47)	305 (6.22)
2006	145 (2.95)	30(0.61)	74(1.51)	69(1.41)	318 (6.48)
2007	97 (1.98)	31(0.63)	93(1.90)	75(1.52)	296 (6.03)
2008	147(2.99)	25 (0.50)	67 (1.36)	79(1.61)	318 (6.48)
2009	151(3.08)	23 (0.47)	108(2.20)	77(1.56)	359(7.31)
2010	174 (3.55)	32(0.65)	77(1.56)	95(1.94)	378(7.71)
2011	173(3.52)	40(0.83)	122(2.49)	90 (1.84)	425 (8.68)
2012	205 (4.18)	25 (0.51)	140 (2.85)	106 (2.16)	476(9.70)
2013	203 (4.14)	51(1.04)	146 (2.98)	135(2.75)	535(10.91)
2014	237 (4.83)	57(1.16)	164(3.34)	201(4.09)	659(13.44)
2015	310(6.32)	55(1.12)	227(4.63)	244 (4.97)	836(17.04)
Grand total					4905

*Note. The percentage of censor information inside the parenthesis

Year based censor information on four stages of lung cancer from 2005 to 2015 are shown in Table 2. In stage 1, the number of censors was 136 (2.77%) in 2005, which increased every year except 2007. In 2015, the number of censors was twice compared to 2005. Over the years 2005-2015, the number of censors was fluctuated in stage 2 and stage 3, while it was increased in stage 4. Overall, the total number of rate of censors was increased.

CHAPTER 4: RESULTS

4.1. Results for Propensity Score Method

We conducted a matched cohort analysis using the propensity score method based on race groups. After matching, we had 15122 patients including 7561 white and 7561 black patients for our subsequent analysis. The baseline characteristics were well-balanced in the matched cohort since absolute standardized mean difference <0.10 across the covariate groups as shown in Table 2.

Table 2. Patient baseline characteristics before and after PSM: White versus Black races

Characteristics		Unmatched cohort (%)			Matched cohort (%)		
		White (n=52891)	Black (n=7561)	SMD	White (n=7561)	Black (n=7561)	SMD
Marital Status	Single	9414 (17.8)	3836 (50.7)	0.740	3819 (50.5)	3836 (50.7)	0.004
	Married	43477 (82.2)	3725 (49.3)		3742 (49.5)	3725 (49.3)	
AGE	Young	588 (1.1)	139 (1.8)	3.050	144 (1.9)	139 (1.8)	0.003
	Middle	23358 (44.2)	4701 (62.2)		4682 (61.9)	4701 (62.2)	
	Adult	28945 (54.7)	2721 (36.0)		2735 (36.2)	2721 (36.0)	
SEX	Male	32665 (61.8)	4655 (61.6)	0.004	4705 (62.2)	4655 (61.6)	0.002
	Female	20226 (38.2)	2906 (38.4)		2856 (37.8)	2906 (38.4)	
Stages	StageI	12555 (23.7)	1351 (17.9)	0.114	1353 (17.9)	1351 ((17.9)	0.004
	StageII	3148 (6.0)	388 (5.1)		358 (4.7)	388 (5.1)	
	StageIII	12420 (23.5)	1918 (25.4)		1941 (25.7)	1918 (25.4)	
	StageIV	24768 (46.8)	3904 (51.6)		3909 (51.7)	3904 (51.6)	
Radiation	Radiation (yes)	26028 (49.2)	3913 (51.8)	0.051	3893 (51.7)	3913 (51.8)	0.005
	Radiation (No)	26863 (50.8)	3648 (48.2)		3650 (48.3)	3648 (48.2)	
Chemo	Chemo (yes)	29507 (55.8)	4169 (55.1)	0.013	4179 (55.3)	4169 (55.1)	0.003
	Chemo (no)	23384 (44.2)	3392 (44.9)		3392 (44.7)	3392 (44.9)	
Surgery	Surgery (yes)	15973 (30.2)	1785 (23.6)	0.149	1771 (23.4)	1785 (23.6)	0.004
	Surgery (no)	36918 (69.8)	5776 (76.4)		5776 (76.4)	5776 (76.4)	

*Note: Percentages of patients characteristics in the parenthesis MAR_STAT=Marital status at diagnosis, AGE_DX=Age at diagnosis, SEX= male, female; D_AJCC_S = American Joint Committee on Cancer Staging, RADIATNR= Radiation therapy, no radiation; CHEMO_RX_REC= chemotherapy, no chemo; NO_SURG= Performed surgery, no surgery; SMD=Standardized Mean Difference (SMD).

The total number of unmatched observations was 60,452. There were more married patients (82.2 % for white and 49.3 % for black) compared to single patients (17.8 % for white and 50.7 % for black). The Adult age group was considered above 65 years old. In this group, 54.7 % of patients were white, and 36.0 % of the patients were black. In the middle age group, 44.2 % of patients were white, and 62.2 % of patients were black. On the contrary, only 1.1 % of young patients were white, and 1.8 % of young patients were black. Considering gender groups, the male white patients were 61.8 % and male black patients were 61.6 %. On the other hand, 38.2 % of female patients were white but 38.4 % of female patients black. Radiation was received by 49.2 % of white and 51.8 % of black patients, whereas 55.8 of % white patients and 55.1 % of black patients received chemo. The percentages of white and black patients' surgery were 30.2 % and 23.6%, respectively. The total number of matched observations was 15,122. The total number of matched observations were stratified by covariates. The percentage of each covariate's observations was almost the same in the matched cohort (Table 2).

We explored the relevant reasons for using the matched sample. It plays a significant role in improving the accuracy of statistical inference. Based on observable covariates, the propensity score model assigned a probability for each individual, corresponding to a race group. The propensity score method was used to create a balanced score across the race groups. Standardized Mean Difference (SMD) was used to assess pre-matching and post-matching balance of covariates distribution between groups as shown in Table 3. SMD was expressed as a percentage of the pooled standard deviation (SD) of the covariates. After matching, the standardized mean difference was very small (<0.1) for all covariates between groups.

SMD ranged from 0.004 to 0.740 from unbalanced data and ranged from 0.003 to 0.019 from balanced data. It is observed that from table2, SMD was higher than 0.1 for marital status, age, stages, and surgery. It indicates that they were unbalanced. Conversely, SMD was less than 0.1 for all explanatory variables for balanced data. It is controlled confounding variables that could affect the estimation of balance covariates distribution. The above reasons support the use of propensity score techniques for this study. The following histogram shows the propensity score distribution.

4.1.1. Propensity score distribution through histogram

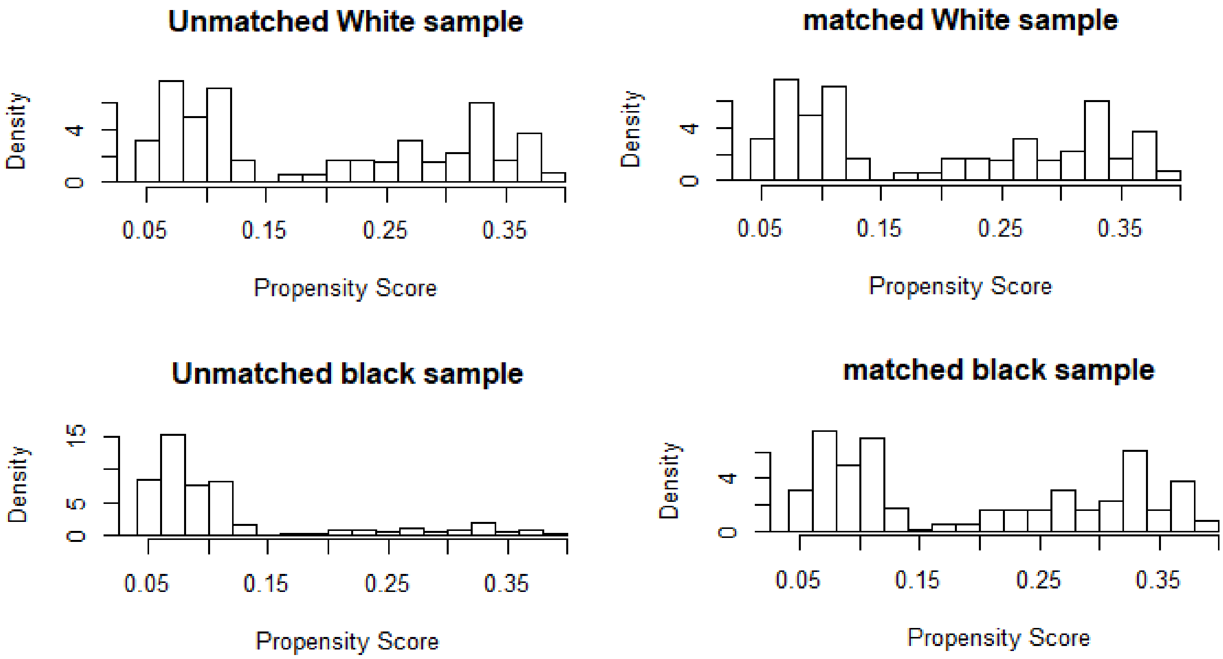

Figure 1. Final Matching Propensity Score distribution with histograms.

Histograms demonstrated the density of propensity score distribution between the white and black groups before and after matching. Most of the observations in the white patient group (raw control group) were left-skewed and the propensity score is lower than the black patient group (raw treated group). After matching, the density of propensity score distributions of the black patient group and white patient group are identical. Everybody in the black group is matched to somebody in the white group (Figure 1).

4.2. Results for Survival Analysis

We compared lung cancer patients' survival for seven treatment groups as well as two race groups by the Kaplan Meier method. A Corresponding log-rank test was used to detect the survival differences among the treatment groups and race groups in all stages. The log-rank test statistic was statistically significant if the p-value was less than 0.05, which can be interpreted as a difference in survival between race groups. Furthermore, we used a pairwise log-rank test for seven treatment groups to select which groups were different from each other. Subsequently, we selected the best treatment groups in all stages based on the hazard ratios.

4.2.1. Survival curves for treatment groups in each stage

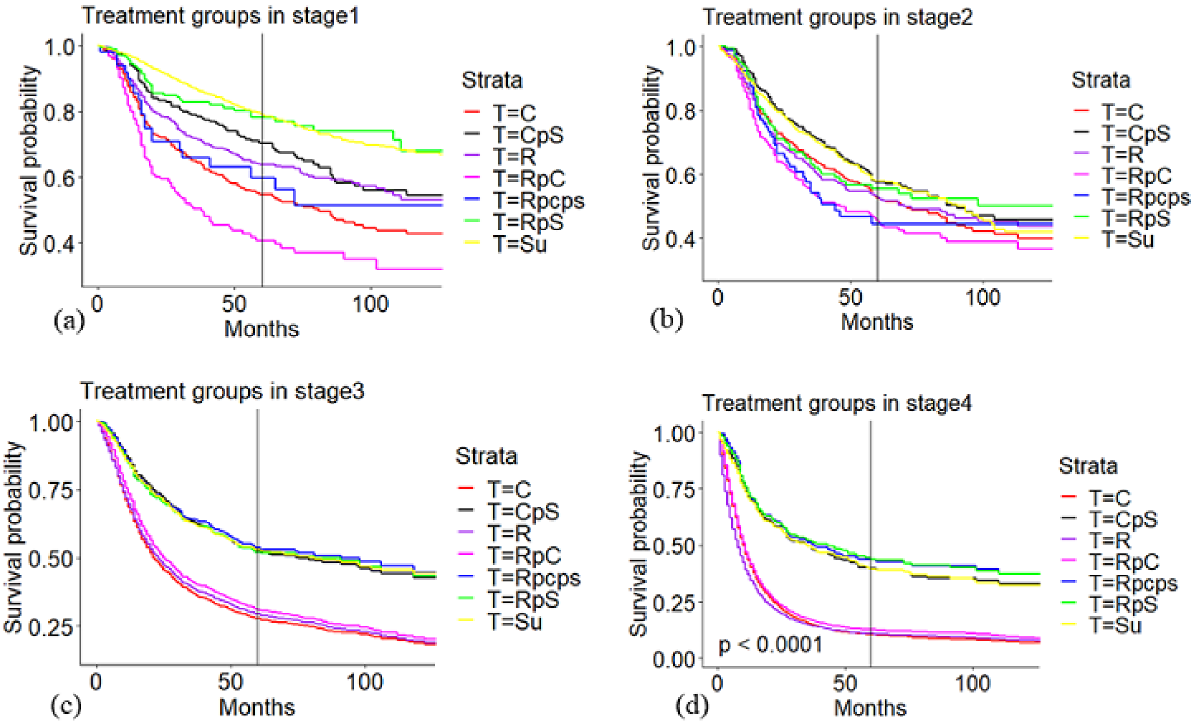

Figure 2. Kaplan-Meier survival curve of lung cancer patient’s survival stratified by six treatment groups. Here, T stands for treatment. Red, black, purple, magenta, blue, green, yellow lines represent Chemo (C), Chemo and Surgery (CPS), Radiation (R), Radiation and chemo (RpC), Radiation and chemo and surgery (Rpcps), Radiation and surgery (RpS), Surgery (Su) treatment groups respectively.

The analysis of survival function stratified by seven treatment groups in each stage is shown in Figure 2. We observed that a horizontal gap indicated longer for one treatment to experience a certain fraction of deaths, while a vertical gap showed a specific period for one treatment group which had a greater probability of patients surviving. As displayed in Figure 2(a), with no adjustment of other covariates, patients’ five-year survival probability increased if they received surgery in stage1. Moreover, after five years, the survival probability increased if patients received treatment radiation and surgery. Conversely, patients’ survival decreased if they received treatment combinations of radiation and chemo. Figure 2(b) shows that the patients’ survival probability was higher in stage 2 if they received treatment chemo and surgery compared to other treatments. The patient's five -year survival probability increased as demonstrated by Figure 2(c) if they received treatment combinations of radiation and chemo and surgery compared to other treatments. Furthermore, patients' survival probability was higher if they took a treatment combination of radiation and surgery, as shown in Figure 2(d). Each stage showed lung cancer survival without a censoring indicator. If the censoring indicator was included in the graphs, then the graphs were not clear. Therefore, we excluded the censoring indicator from the survival curve. Since we had seven treatment groups, we decided which groups were different from each other by pairwise log-rank test, as shown in Table 4.

Table 3. Survived patients by race, gender, age, marital status in stage 3 and 4

Survival patient's information	Year 5		Year 6		Year 7		Year 8	
	Stage3	Stage4	Stage3	Stage4	Stage3	Stage4	Stage3	Stage4
Race (%)								
Black	74(10)	54 (4)	55(6)	36(3)	35(5)	28(2)	27(3)	17(1)
White	64 (8)	45 (3)	53(6)	32(3)	39(4)	25(2)	28(3)	19(1)
Age groups (%)								
Young (<35)	6 (7)	0	4 (4)	0	2 (2)	0	2 (2)	0
Middle (35< &<65	101(9)	62 (3)	79(7)	48(3)	37(6)	54(2)	28(4)	42(1)
Adult (65<)	37(8)	31(3)	25(4)	20(2)	18(4)	16(1)	11(2)	8(1)
Marital status (%)								
Single	69(8)	47(3)	57(6)	33(3)	38(4)	26(2)	33(4)	21(2)
Married	69(9)	52(3)	51(6)	35(3)	36(5)	27(2)	22(3)	15(1)
Gender (%)								
Male	55(11)	46(3)	41(8)	31(3)	23(6)	29(2)	21(5)	20(1)
Female	83(7)	53(3)	37(5)	30(3)	67(4)	45(2)	34(2)	16(1)

Note: The percentage of survival patient's in the parenthesis for the year of 5,6,7, and 8.

Some patients survive more than 5 years in stage 3 and stage 4, which is shown in Table 3. In the case of year 5, the survival rate of black patients was 10 % and 4% for stage 3 and stage

4, respectively; while the survival rate of white patients was 8% and 3% for stage 3 and stage 4, respectively. The survival rate of black and white patients declined year to year. Moreover, the survival rate of age groups, marital status groups, and gender groups were also declined year to year. We explored the reasons for more than 5 years of survival in stage3 and stage 4. There are many reasons, such as overall health condition, patient's age at early diagnosis, gender, each patient's socio-economic conditions, and the gene changes in the cancer cells period. A significant percentage of lung cancer could be prevented by reducing tobacco use and other unhealthy behaviors. Currently, lung cancer mortality is declined day by day due to upgrading treatment such as targeted therapy, chemotherapy, radiation therapy, and surgery. Furthermore, proper biomarker testing and the betterment of genetics can help to enhance the survival rate for more than 5 years in stages 3 and 4 (Zappa and Mousa 2016).

Table 4. P values for pairwise comparison with the log-rank test for different treatment groups

Stages	Treatment groups	C	C+S	R	R+C	R+C+S	R+S
Stage1	C+S	0.0005	--	--	--	--	--
	R	0.0028	0.209	--	--	--	--
	R+C	0.0032	<0.0001	<0.0001	--	--	--
	R+C+S	0.2096	0.0032	0.0185	0.988	--	--
	R+S	<0.0001	0.0543	0.003	<0.0001	0.0004	--
	S	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	0.776
Stage2	C+S	0.205	--	--	--	--	--
	R	0.011	0.205	--	--	--	--
	R+C	0.205	0.664	0.664	--	--	--
	R+C+S	0.205	0.904	0.217	0.699	--	--
	R+S	0.664	0.648	0.205	0.514	0.648	--
	S	0.726	0.237	0.011	0.237	0.237	0.837
Stage3	C+S	<0.0001	--	--	--	--	--
	R	0.3595	<0.0001	--	--	--	--
	R+C	0.0018	<0.0001	0.0532	--	--	--
	R+C+S	<0.0001	0.889	<0.0001	<0.0001	--	--
	R+S	<0.0001	0.969	<0.0001	<0.0001	0.871	--
	S	<0.0001	0.969	<0.0001	<0.0001	0.871	0.969
Stage4	C+S	<0.0001	--	--	--	--	--
	R	<0.0001	<0.0001	--	--	--	--
	R+C	0.047	<0.0001	<0.0001	--	--	--
	R+C+S	<0.0001	0.472	<0.0001	<0.0001	--	--
	R+S	<0.0001	0.386	<0.0001	<0.0001	0.921	--
	S	<0.0001	0.896	<0.0001	<0.0001	0.383	0.277

*Note: Inside the table, all values indicate the p-value. In each stage, p < 0.05 was considered as significant value. C= Chemo, C+S=Chemo and Surgery, R= Radiation, R+C= Radiation and Chemo, R+S=Radiation and Surgery, S= Surgery, R+C+S=Radiation and chemo, and surgery

We used pairwise log-rank tests to check the equality of seven treatment groups in each stage as shown in Table 4. It appeared that the following pairs of groups were not significantly different (p>0.05): R+C+S and C, R+S and C+S, R+C+S and R+C, S, and R+S. Conversely, other pairs were significantly different (p < 0.05) in stage1. In stage 2, the pairs R and C, S and R, were significantly different (p < 0.05). The other pairs were not statistically different (p>0.05). In stage 3, the following pair of groups were not significantly different (p>0.05): R and C, R+C+S and C+S, R+S and C+S, S, and C+S, R+C and R, R+S and R+C+S, S and R+C+S, S and R+S. In stage 4, the following pair of treatment groups were not significantly different (p>0.05): R+C and C, R+C+S and C+S, R+S, and C+S, S, and C+S, R+C and R, R+S and R+C+S, S and R+C+S, S and R+S. The pairwise log-rank test describes the equality of the survival distribution among the treatment groups. The best treatment was determined based on the hazard ratios which are shown in table 6.

4.2.2. Testing proportional hazards assumption

The cox model estimated which variables contributed to the risk. It also depends on the proportional hazard assumptions. The cox model assumes that the multiplicative effect of the covariate in the hazard function is constant over time. It is essential to compute which covariate is associated with a lower or higher risk of lung cancer death overtime. Before doing that, we checked the assumption using scaled Schoenfeld residuals as shown in Table 5.

Table 5. Scaled Schoenfeld Residuals of significant covariates on the PH at 4 different stages for different treatment groups.

Stages	Treatment groups	rho	P value
Stage1	C	0.0127	0.0684
	C+S	0.0621	0.0567
	R	0.0126	0.0686
	R+S	0.0101	0.0747
	S	0.123	0.0902
	R+C+S	-0.005	0.0870
	Global	NA	.10300
Stage2	C	-0.0616	0.0699
	R+C	-0.0936	0.0596
	R	-0.0959	0.0585
	R+S	-0.0693	0.0514
	S	-0.0215	0.5270
	R+C+S	-0.0603	0.0758
	Global	NA	0.1700
Stage3	C	0.0043	0.7468
	R+C	0.0336	0.1336
	C+S	0.0361	0.0783
	R+S	0.0153	0.2578
	S	0.0184	0.1742
	R+C+S	0.0170	0.2111
	Global	NA	0.0560
Stage4	C	0.1089	0.0985
	R+C	0.091	0.0590
	C+S	0.0492	0.330
	R+S	0.0442	0.0.531
	S	0.0441	0.0564
	R+C+S	0.0426	0.0940
	Global	NA	0.0584

The test for the Cox PH model is Schoenfeld, which indicates whether the model assumption is violated or not. Schoenfeld first introduced chi-square goodness of fit in 1980 to investigate the proportional hazard assumption using lagged residuals. The first column indicates the Pearson product-moment correlation (rho), which describes the Schoenfeld residuals and lagged residuals for each independent variable. The row Global indicates the global test of proportionality for each covariate. From the above output, it is apparent that the covariates C, R, S, C+S, R+S, R+C, R+C+S were not statistically significant (p>0.05) in all stages. In stage1, the p value for global test was 0.10 (>0.05). It means that this is not statistically significant. In stage2, stage3, and stage4, the p value was greater than 5 %, it is concluded that the global test was not statistically significant. All groups were likely to hold proportional hazard assumption true in all stages. There was no violation for all stages. Therefore, Cox model can be used for further analysis.

Table 6. Hazard ratio calculation for various treatment groups for four different stages.

	Treatment groups	HR	95 % CI	P value
Stage1	R+C	Reference		
	C	0.71	0.55---0.90	0.004
	R	0.81	0.65---1.01	0.06
	C+S	1.27	0.96—1.67	0.09
	R+S	0.55	0.43---0.71	<0.0001
	Su	0.47	0.37---0.60	<0.0001
	R+C+S	1.17	0.86---1.57	0.30
	C	Reference		
	R	0.67	0.55---0.82	<0.0001
	C+S	1.79	1.42---2.28	<0.0001
	R+S	0.78	0.64—0.96	0.020
	Su	1.15	0.96—1.37	0.12
	R+C+S	1.65	1.22---2.15	0.0007
	C+S	Reference		
	R	0.37	0.29—0.47	<0.0001
	R+S	0.44	0.34—0.56	<0.0001
	Su	0.64	0.51---0.79	<0.0001
	R+C+S	0.92	0.68---1.23	0.579
	R	Reference		
	R+S	1.17	0.95—1.42	0.12
	Su	1.71	1.44—2.03	<0.0001
	R+C+S	2.46	1.90—3.19	0.0000
	R+S	Reference		
	Su	1.46	1.22---1.75	<0.0001
	R+C+S	2.10	1.62—2.74	0.0000
	Su	Reference		
	R+C+S	1.43	1.12---1.84	0.003

Table 6. Hazard ratio calculation for various treatment groups in four different stages (Continued).

	Treatment groups	HR	95 % CI	P value
Stage2	C+S	Reference		
	R	1.0	0.86---1.31	0.54
	R+C	1.3	1.09---1.73	0.005
	C	1.5	1.22---2.03	0.000
	R+S	1.1	0.90---1.44	0.27
	Su	1.5	1.25—1.88	<0.0004
	R+C+S	1.4	1.11---1.88	0.006
	R	Reference		
	C	0.93	0.76—1.15	0.05
	R+S	1.06	0.84—1.35	0.59
	R+C	1.29	1.02—1.63	0.03
	Su	1.43	1.16—1.77	0.0007
	R+C+S	1.35	1.03---1.77	0.026
	R+S	Reference		
	C	0.87	0.69---1.11	0.04
	Su	1.34	1.06---1.71	0.01
	R+C	1.21	0.94—1.56	0.15
	R+C+S	1.27	0.94—1.70	0.11
	R+C	Reference		
	C	0.72	0.57---0.90	0.24
	Su	1.11	0.88—1.40	0.03
	R+C +S	1.05	0.78—1.39	0.54
	Su	Reference		
	C	0.65	0.53---0.80	<0.0001
	R+C+S	0.94	0.72—1.2	0.67
	C	Reference		
	R+C+S	0.92	0.67—1.25	0.59

Table 6. Hazard ratio calculation for various treatment groups in four different stages (Continued).

	Treatment groups	HR	95 % CI	P value
Stage3	R	Reference	95 % CI	P value
	C	1.05	0.98---1.13	0.10
	R+C+S	0.96	0.88---1.02	0.05
	C+S	1.23	1.15---1.32	0.000
	R+S	1.14	1.06---1.23	0.000
	Su	1.47	1.38—1.57	0.000
	R+C	0.92	0.86—0.99	0.22
	R+C+S	Reference		
	C	1.0	0.93—1.08	0.82
	R+S	1.08	1.00---1.18	0.03
	C+S	1.17	1.09---1.27	<0.0001
	R+C	0.88	0.81—0.95	0.003
	Su	1.41	1.31---1.51	<0.0001
	Su	Reference		
	C	0.72	0.67---0.76	<0.0001
	C+S	0.84	0.78---0.89	<0.0001
	R+S	0.77	0.72---0.82	<0.0001
	R+C	0.63	0.58----0.67	<0.0001
	C+S	Reference		
	C	0.86	0.79---0.92	<0.0001
	R+C	0.75	0.69---0.81	<0.0001
	R+S	0.92	0.86—0.99	0.03
	R+C	Reference		
	C	1.41	1.06—1.23	0.0003
	R+S	1.23	1.14---1.33	<0.0001
	C	Reference		
	R+S	1.07	1.00—1.15	0.04

Table 6. Hazard ratio calculation for various treatment groups in four different stages (Continued)

	Treatment groups	HR	95 % CI	P value
Stage4	R	Reference		
	C	0.87	0.84---0.91	<0.0001
	Su	1.21	1.16---1.26	<0.0001
	C+S	0.94	0.90---0.98	0.016
	R+C	1.12	1.07---1.17	<0.0001
	R+S	0.82	0.78---0.87	<0.0001
	R+C+S	0.91	0.86---0.96	0.0003
	C+S	Reference		
	C	0.92	0.88—0.97	0.005
	R+C	0.87	0.83---0.92	<0.0001
	R+S	1.18	1.13—1.24	<0.0001
	Su	1.28	1.23---1.33	<0.0001
	R+C +S	0.96	0.91---1.01	0.12
	R+C	Reference		
	C	1.06	1.00—1.11	0.03
	Su	1.46	1.40---1.53	<0.0001
	R+S	1.36	1.29---1.43	<0.0001
	R+C+S	1.09	1.03---1.16	0.0012
	C	Reference		
	R+S	1.28	1.22---1.34	<0.0001
	Su	1.38	1.33---1.44	<0.0001
	R+C+S	1.04	0.99----1.09	0.14
	R+S	Reference		
	R+C+S	0.81	0.77---0.85	<0.0001
	Su	1.07	1.03---1.12	0.0002

First, the lower survival probability treatment group was selected as a reference group for all stages. Based on the hazard ratios, radiation and surgery were the best treatment for stage1. Treatment options depend on the patients' condition and stage of the disease. From the literature review, Chemo was the best treatment in stage2 due to the hazard ratio if we considered lower survival probability as a reference group. If cancer is spread out to other parts of the body and patients have both small cell lung cancer and non-small cell lung cancer, chemo and surgery were offered in stage2. In our study, the risk of death increased by 5% (HR=1.5,95% CI: 1.22 -2.03) in stage2 if patients received C. On the other hand, due to the hazard ratio, R+C was the best treatment group for stage 3. When lung cancer spread out to the whole body, more treatment combinations were needed to control lung cancer. In this case, radiation, chemo, and surgery were offered to patients' for stage3. In our study, if patients received radiation and chemo, the risk of death decreased by 8 % (HR=0.92, 95 % CI 0.86—0.99) compared to the reference group. The true value was lying between 1% and 14%. According to the hazard ratio, Chemo was the best treatment for stage4. The risk of death decreased by 13% (HR=0.87, 95 % CI 0.84—0.91) if patients received C in stage4. However, if we considered a different reference treatment group instead of the lower survival probability treatment group in all stages, the best treatment combinations are different.

4.2.3. Best treatment effect of race groups for all stages

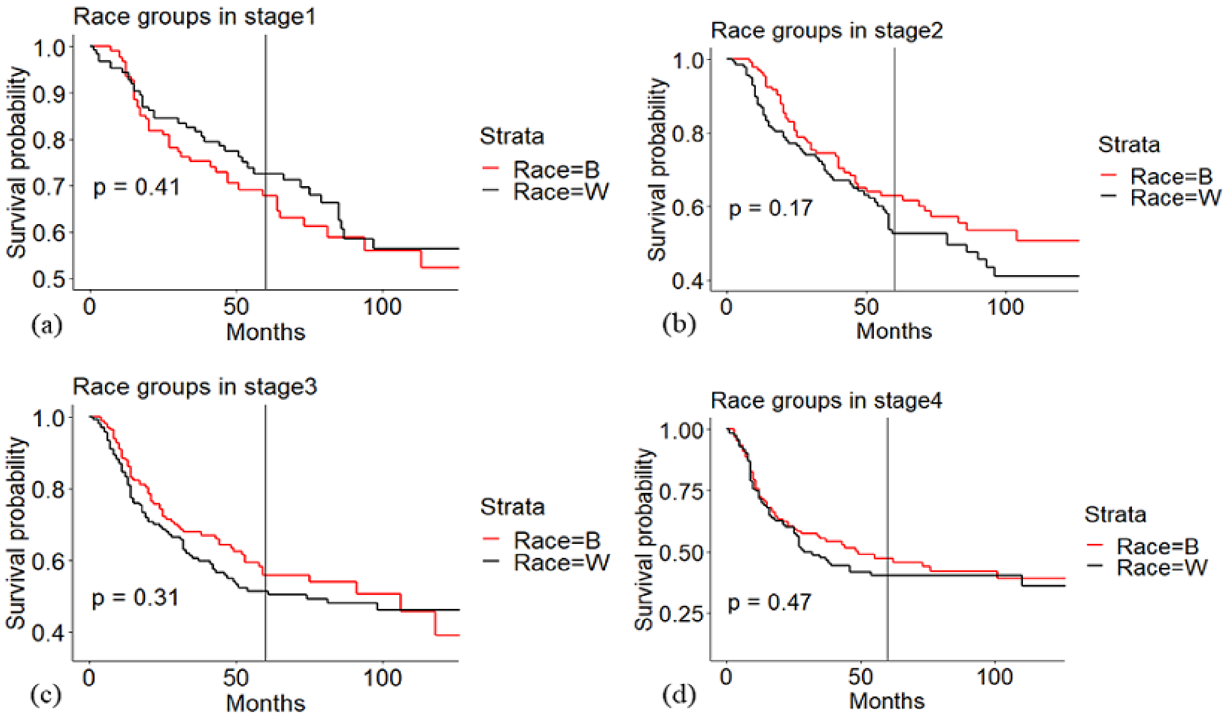

Figure 3. Kaplan-Meier survival curve of lung cancer patient's survival stratified by race groups using the best treatment groups. Here, B and W stands for black and white respectively. Red, and black lines corresponding to black and white race groups

The analysis of survival function stratified by race groups for all stages is shown in Figure 3. Our purpose is to test the hypothesis of a survival difference between black and white patients' groups if they received the best treatment for all stages. It appeared the best treatment effect of survival time was the same in all stages of our study. Nonetheless, a few studies illustrated that there was a difference between black and white patients' survival time. Yan et al. (2013) showed that black patients had a lower survival than white patients. They could not measure the exact reasons behind it, but they considered some reasons for that, such as socioeconomic conditions, racial discrimination, poor access to health facilities, smoking history, and family health conditions. However, we did not consider other covariates, which could influence the survival time of white and black patients in all stages. Controlling other factors, if white and black patients underwent radiation and surgery in stage1, chemo in stage 2, radiation and chemo in stage 3, and chemo in stage 4, there was no difference (p > 0.05) between their survival time for all stages.

Table 7. The computation for the log-rank test to check the equality between black and white patients using the best treatment groups

Stages	Race groups	Chi-square value	P value
Stage1	B	0.67	0.41
	W	0.67	
Stage2	B	1.84	0.17
	W	1.84	
Stage3	B	1.05	0.31
	W	1.05	
Stage4	B	0.52	0.47
	W	0.52	

Note: B=Black, W=White

A log-rank test was used to test equality between white and black patient groups. There was no significant difference (p > 0.05) in survival time between race groups in all stages. Holding other factors, our results demonstrated that if we used different treatment groups in all stages, there was no survival increase or decrease (p>0.05) between white and black patients (Table 7).

Table 8. Cox proportional hazard model for race groups in all stages

Stages	Race groups	HR	95 % CI	P value
Stage1	Black	Reference		0.42
	White	0.85	0.58---1.25	
Stage2	Black	Reference		0.87
	White	0.99	0.88---1.11	
Stage3	Black	Reference		0.06
	White	1.12	1.03—1.22	
Stage4	Black	Reference		0.42
	White	1.19	0.77—1.83	

Note: HR=Hazard Ratio; CI=Confidence Interval

Each hazard ratio indicates a relative risk of death that compares one group to another group. If the hazard ratio is greater than one, it indicates the risk of death increased. On the other hand, if the hazard ratio is less than one, it illustrates the risk of death decreased. In the white patient group, 15 % less died compared to the black patient group (HR=0.85, 95 % CI 0.58 to 1.25) as shown in stage 1. Furthermore, the risk of death decreased by 1 % for white patients (HR=0.99, 95% CI 0.88 to 1.11) compared to black patients for stage 2. Conversely, in stage 3, the white patients' group displayed a 12 % increased risk of lung cancer death compared to the black patients' group (HR=1.12, 95 % CI 1.03 to 1.22). The white patients' group also showed a 19 % increased risk of death (HR=1.19, 95 % CI 0.78 to 1.83) compared to the black patients' group for stage 4 (Table 8).

4.2.4. Survival function for race and gender groups in each stage using best treatment group

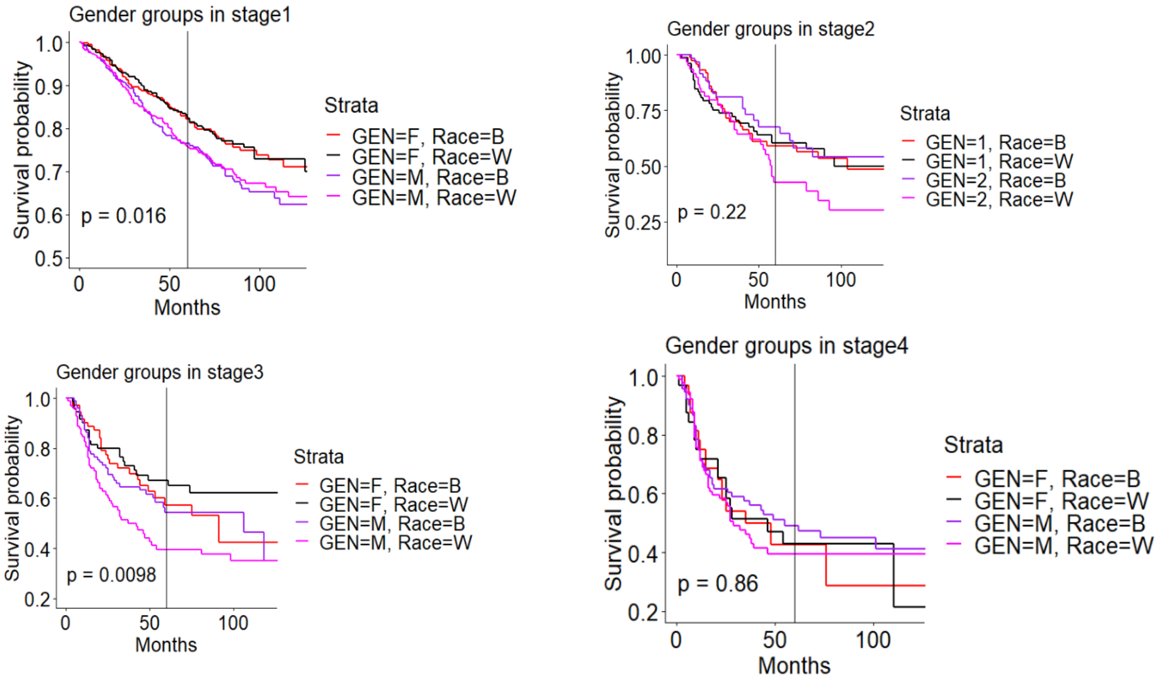

Figure 4. Kaplan-Meier survival curve of lung cancer patient's survival stratified by gender groups using the best treatment groups. Here, GEN stands for gender, F, and M stands for female and male respectively.

The survival function stratified by gender and race groups using the best treatment in each stage is shown in Figure 4. The survival experience of white and black female patients was higher than that of white and black male patients in stage1. White male and female patient's survival experiences were higher in stage 2. In contrast, the survival experience of black male and female patients was higher in stage 3 and stage4 until five years compared to white female and male patients.

Table 9. P values for pairwise comparison with the log-rank test for race and gender groups

Stages	Gender groups	B+F	F+W	M+B
Stage1	F+W	0.891	--	--
	M+B	0.045	0.642	--
	M+W	0.050	0.044	0.891
Stage2	F+W	0.520	--	--
	M+B	0.642	0.282	--
	M+W	0.067	0.241	0.113
Stage3	F+W	0.922	--	--
	M+B	0.684	0.031	--
	M+W	0.041	0.872	0.312
Stage4	F+W	0.093	--	--
	M+B	0.093	0.091	--
	M+W	0.242	0.091	0.059

The pair of gender groups white male and black female, black male and black female, white male and white female were significantly different (p<0.05) by pairwise log-rank test in stage1. In stage two and stage four, there was no significant difference between gender groups by the pairwise log-rank test. In stage three, the pair of gender group white male and black female, black male, and white female were significantly different (p<0.05) by pairwise log-rank test (Table 9).

Table 10. Cox proportional hazard model for gender groups in all stages

Stages	Gender groups	HR	95 % CI	P value
Stage1	Female	Reference		0.23
	Male	0.77	0.51 - 1.18	
Stage2	Female	Reference		0.008
	Male	1.9	1.19 - 3.3	
Stage3	Female	Reference		0.112
	Male	1.10	0.97 - 1.24	
Stage4	Female	Reference		0.012
	Male	1.11	1.02 - 1.21	

The male patients had a decreased risk of death of 77 % (HR=0.77) in stage 1 compared to female patients. Due to the lower hazard ratio, white and black male patients had improved survival in stage1. Conversely, black and white male patients had an increased risk of death at 9 %, 10%, and 11 % in stage 2, stage3, stage4, respectively. Since the hazard ratio was greater than one, it was associated with poor survival (Table 10).

Table 11. Cox proportional hazard model for gender groups in all stages using unmatched data

Stages	Gender groups	HR	95 % CI	P value
Stage1	Female	Reference		
	Male	1.06	1.02 - 1.42	0.003
Stage2	Female	Reference		
	Male	1.14	0.96 - 1.36	0.136
Stage3	Female	Reference		
	Male	1.19	1.01 - 1.40	0.029
Stage4	Female	Reference		
	Male	1.18	0.91 - 1.21	0.889

The male patients had an increased risk of death of 6 % (HR=1.06) in stage 1 compared to female patients. Furthermore, the black and white male patients had an increased risk of death 14 %, (HR=1.14), 19% (HR=1.19), and 18 % (HR=1.18) in stage 2, stage3, stage4, respectively (Table 11). In terms of comparing the results between table 10 and table 11, we think that the results of the matched data are more reliable because the two samples are more comparable in terms of the covariates compared to the unmatched sample. Since the hazard ratios were greater than one for all stages (Table 11), it concludes that they were associated with poor survival. Therefore, the results obtained using the matched data are more reliable than unmatched data.

CHAPTER 5: CONCLUSION

First and foremost, we computed the propensity score and got balanced samples based on the race groups. Subsequently, the balanced samples were used to determine the best treatment in all stages. A pairwise log-rank test was used to check the equality of seven treatment groups in all stages. Furthermore, the log-rank test was used to test the hypothesis of a survival difference between black and white patients' groups if they received the best treatment in all stages. Based on hazard ratios, radiation and surgery were the best treatment in stage1. Moreover, chemotherapy was the best treatment in stage 2. Furthermore, the treatment combination of radiation and chemo was the best treatment in stage 3. In stage 4, chemotherapy was the best treatment according to the hazard ratio. White male patient survival was better in stage1 compared to female patients based on hazard ratio. Conversely, the white and black female patients were higher survival compared to white and black male patients in stage 2, stage 3, and stage 4, based on the hazard ratios.

5.1. Future Work

Future work should include the use of Weibull distribution, Gamma distribution, Exponential distribution, Log-normal, and Log-Logistic for further analysis to see the relationship between the survival time in months of patients diagnosed with lung cancer and the covariates. Nowadays, machine learning has a great significance for analysis in the health sector due to its high computational capacity for the early prediction of diseases. Therefore, machine learning techniques can be used for lung cancer survival prediction.

REFERENCES

"Centers for Disease Control and Prevention", https://www.cdc.gov/nchs/fastats/deaths.htm.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020), "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, American Cancer Society, 70, 7–30.

McCarthy, W. J., Meza, R., Jeon, J., and Moolgavkar, S. H. (2012), "Chapter 6: Lung Cancer in Never Smokers: Epidemiology and Risk Prediction Models," *Risk Analysis*, John Wiley & Sons, Ltd, 32, S69–S84.

Zhang, R., Li, P., Li, Q., Qiao, Y., Xu, T., Ruan, P., Song, Q., and Fu, Z. (2018), "Radiotherapy improves the survival of patients with stage IV NSCLC: A propensity score matched analysis of the SEER database," *Cancer Medicine*, John Wiley & Sons, Ltd, 7, 5015–5026.

Govindan, R., Page, N., Morgensztern, D., Read, W., Tierney, R., Vlahiotis, A., Spitznagel, E. L., and Piccirillo, J. (2006), "Changing Epidemiology of Small-Cell Lung Cancer in the United States Over the Last 30 Years: Analysis of the Surveillance, Epidemiologic, and End Results Database," *Journal of Clinical Oncology*, American Society of Clinical Oncology, 24, 4539–4544.

Lu, T., Yang, X., Huang, Y., Zhao, M., Li, M., Ma, K., Yin, J., Zhan, C., and Wang, Q. (2019), "Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades," *Cancer management and research*, Dove Medical Press, 11, 943–953.

Lim, Y. J., Song, C., and Kim, J.-S. (2017), "Improved survival with postoperative radiotherapy in thymic carcinoma: A propensity-matched analysis of Surveillance, Epidemiology, and End Results (SEER) database," *Lung Cancer*, 108, 161–167.

Che, K., Shen, H., Qu, X., Pang, Z., Jiang, Y., Liu, S., Yang, X., and Du, J. (2018), "Survival Outcomes for Patients with Surgical and Non-Surgical Treatments in Stages I-III Small-Cell Lung Cancer," *Journal of Cancer*, Ivyspring International Publisher, 9, 1421–1429.

Lampaki, S., Zarogoulidis, P., Lagoudi, K., Tsavlis, D., Kioumis, I., Papakala, E., Lazaridis, G., Huang, H., Hohenforst-Schmidt, W., Pavlidis, P., Darwiche, K., Barbetakis, N., Karapantzos, I., Karapantzou, C., Rapti, A., Karavasilis, V., and Zarogoulidis, K. (2016), "Small Cell Lung Cancer: Current and Future Strategies," *Oncomedicine*, 1, 4–13.

Caprario, L. C., Kent, D. M., and Strauss, G. M. (2013), "Effects of Chemotherapy on Survival of Elderly Patients with Small-Cell Lung Cancer: Analysis of the SEER-Medicare Database, " *Journal of Thoracic Oncology*, 8, 1272–1281.

Earle, C. C., Venditti, L. N., Neumann, P. J., Gelber, R. D., Weinstein, M. C., Potosky, A. L., and Weeks, J. C. (2000), "Who Gets Chemotherapy for Metastatic Lung Cancer?," *Chest*, 117, 1239–1246.

Fu, J. B., Kau, T. Y., Severson, R. K., and Kalemkerian, G. P. (2005), "Lung Cancer in Women: Analysis of the National Surveillance, Epidemiology, and End Results Database," *Chest*, 127, 768–777.

Hubbard, M. O., Fu, P., Margevicius, S., Dowlati, A., and Linden, P. A. (2012), "Five-year survival does not equal cure in non–small cell lung cancer: A Surveillance, Epidemiology, and End Results–based analysis of variables affecting 10- to 18-year survival," *The Journal of Thoracic and Cardiovascular Surgery*, 143, 1307–1313.

Thakur, M. K., Ruterbusch, J. J., Schwartz, A. G., Gadgeel, S. M., Beebe-Dimmer, J. L., and Wozniak, A. J. (2018), "Risk of Second Lung Cancer in Patients with Previously Treated Lung Cancer: Analysis of Surveillance, Epidemiology, and End Results (SEER) Data," *Journal of Thoracic Oncology*, 13, 46–53.

Austin, P. C. (2011), "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research*, Routledge, 46, 399–424.

Austin, P. C. (2010), "The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies," *Statistics in Medicine*, John Wiley & Sons, Ltd, 29, 2137–2148.

Lonardo, N. W., Mone, M. C., Nirula, R., Kimball, E. J., Ludwig, K., Zhou, X., Sauer, B. C., Nechodom, K., Teng, C., and Barton, R. G. (2014), "Propofol is associated with favorable outcomes compared with benzodiazepines in ventilated intensive care unit patients," *American Journal of Respiratory and Critical Care Medicine*, 189, 1383–1394.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Yang, D., and Dalton, J. (2012), "A unified approach to measuring the effect size between two groups using SAS ® Dongsheng Yang and Jarrod E . Dalton Departments of Quantitative Health Sciences and Outcomes Research Cleveland Clinic SAS Global Forum 2012 Statistics and Data Analysis," *SAS Gloabl Forum 2012*, Paper 335.

Daniel Ho, Kosuke Imai, Gary King, E. A. S. (2011), "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference," *Journal of Statistical Software*, 42.

D.G Kleinbaum, M. K. (2006), "Survival analysis: a self-learning text, Springer Science & Business Media."

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, Taylor & Francis, 53, 457–481.

Mantel, N. (1966), "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemother. Rep.*

Peto, R., and Peto, J. (1972), “Asymptotically Efficient Rank Invariant Test Procedures,” *Journal of the Royal Statistical Society: Series A (General)*, John Wiley & Sons, Ltd, 135, 185–198.

Bewick, V., Cheek, L., and Ball, J. (2004), “Statistics review 12: Survival analysis,” *Critical Care*, 8, 389.

Cox, D. R. (1972), “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, John Wiley & Sons, Ltd, 34, 187–202.

Duggan, M. A., Anderson, W. F., Altekruse, S., Penberthy, L., and Sherman, M. E. (2016), “The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship,” *The American journal of surgical pathology*, 40, e94–e102.

Zappa, C., and Mousa, S. A. (2016), “Non-small cell lung cancer: current treatment and future advances,” *Translational lung cancer research*, AME Publishing Company, 5, 288–300.

Yan, G., Norris, K. C., Yu, A. J., Ma, J. Z., Greene, T., Yu, W., and Cheung, A. K. (2013), “The relationship of age, race, and ethnicity with survival in dialysis patients,” *Clinical journal of the American Society of Nephrology : CJASN*, American Society of Nephrology, 8, 953–961.