SURVIVAL ANALYSIS OF TREATMENT EFFECT FOR BRAIN CANCER: BASED ON THE SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS DATABASE

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

By

Madison Jane Mathiason

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

Survival Analysis of Treatment Effect For Brain Cancer: Based on The Surveillance, Epidemiology, and End Results Database

By

Madison Jane Mathiason

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. Bong-Jin Choi
Chair

Dr. Gang Shen

Dr. Rick Jansen

Approved:

11/20/2020
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Cancer is one of the leading causes of death in the United States. The Surveillance, Epidemiology, and End Results (SEER) data from the National Cancer Institute is a population based cancer registry, which geographically covers 34.6% of the US population. The SEER database was used to model surivial time for 21,524 patients with primary malignant brain tumors. The Kaplan-Meier surivial curves and the logrank test were used to compare the effect of treatment in each grade. The Cox Proportional Hazard Model was used to show the simultaneous effect of treatment, sex, and age on the risk of death for patients in each grade. Elderly patients had the lowest survival time, while adults had the highest. The risk of death for males was slightly higher than females. The results demonstrate that the survival curves of the three treatment groups only significantly differ among participants with grade 4 primary brain tumors.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Bong-Jin Choi for being my advisor during my graduate research and taking the time to help me learn and grow throughout my time at North Dakota State University. Also, I would like to say how much I appreciate my other committee members Dr. Gang Shen and Dr. Rick Jansen for their time and consideration on my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Survival analysis is used to analyze data to determine the time it takes for an event to occur. For survival analysis, the event of interest is usually death, but it can also be time till failure, time till remission, or end of life for a machine part. In engineering this is called reliability analysis. Survival analysis estimates the survival time based on one or more predictors which are commonly referred to as covariates. A fundamental aspect of survival analysis is censoring. Censoring occurs when the subject does not experience the event of interest in our study time frame and is therefore censored. Survival analysis includes subjects that are censored on the estimation of survival time by keeping them in the risk set until they are censored. The concept of censoring, or the idea that we measure someone that does not experience our event of interest, is the difference between survival analysis and regression.

There are many methods in survival analysis, including the Kaplan-Meier estimator, the logrank test, and the Cox Proportional Hazards Model. The Kaplan-Meier estimator takes into account a single categorical variable on survival time. The logrank test is the most popular way of comparing the survival of groups (Bland & Altman, 2004). The logrank test compares the survival curves from the Kaplan-Meier estimator to determine if there is a significant difference between the values of the covariate on survival time. The Cox Proportional Hazards Model takes into the account the effect of several variables simultaneously, and the model allows for continuous and categorical variables. This analysis will use the National Cancer Institutes Surveillance, Epidemiology, and End Results (SEER) data to model survival times of 21,524 patients with primary malignant brain tumors.

Survival curves are shown by the Kaplan-Meier estimator for grade and treatments in each grade. The logrank test will be used to the compare the Kaplan-Meier survival curves for

treatment in each grade. When there is a significant difference between the survival curves, the pairwise logrank test will be used to see which pairs of survival curves are different from each other by testing each pair of treatments for a significant difference. There are 3 comparisons total from the pairwise logrank test. A Cox Proportional Hazards Model is also used to model the effect of treatment, age, and sex on survival time. The goal of my research is show the effect of treatment in each grade on survival times for patients with primary malignant brain tumors, adjusting for age and sex.

2. SCIENTIFIC REVIEW

The scientific review will cover fundamental concepts in survival analysis including censoring, the survival function, the hazard function, the Kaplan-Meier estimator, the logrank test, and the Cox Proportional Hazards Model.

2.1. Censoring

There are three different types of censoring: left, right, and interval, with right being the most common. Right censoring is when the observed individual drops out or does not experience the event of interest during the study period. Left censoring is when the event has occurred before the study began, but it is unknown at what exact time. Interval censoring is when the event occurred in between two periods of time. In this paper, right censoring will be used because not all patients die from their cancer, so the event of interest does not happen during the study period. Illustratations of the 3 types of censoring are shown below. In figure 1, subjects 2 and 4 are right censored. In figure 2, subjects 2 and 3 are left censored since the study began at time 5. In figure 3, subjects 1, 3, and 5 are interval censored since the individuals were measured at time 5 and did not have an event, but had an event at time 10 when they were measured next.

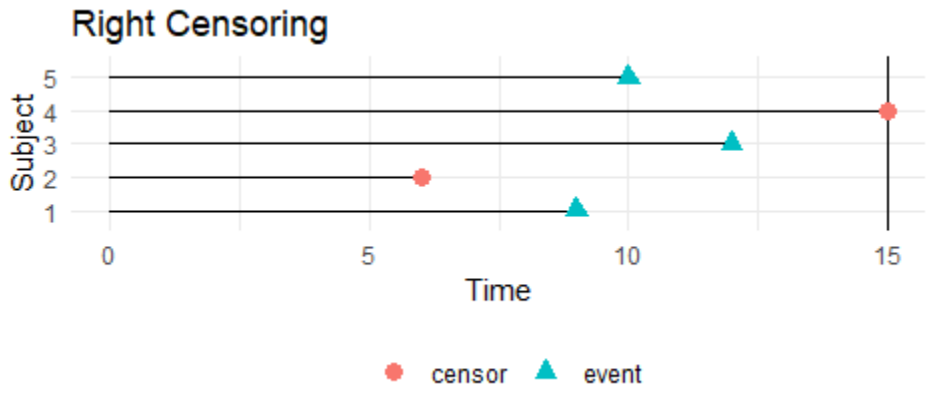

Figure 1. Right Censoring.

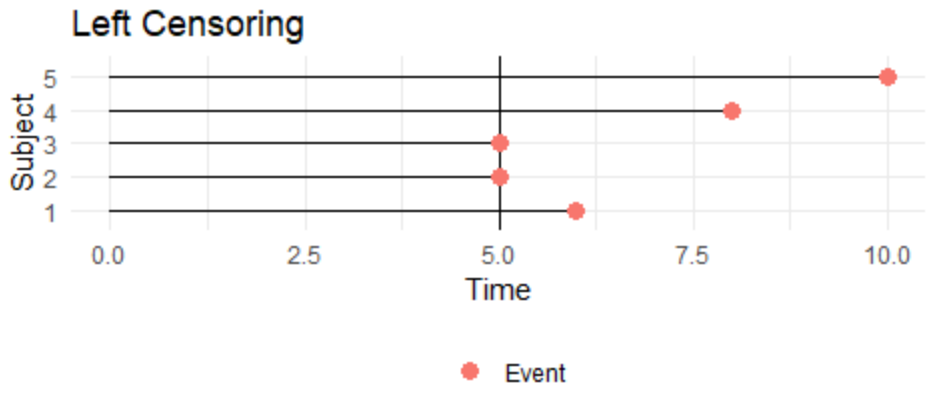

Figure 2. Left Censoring.

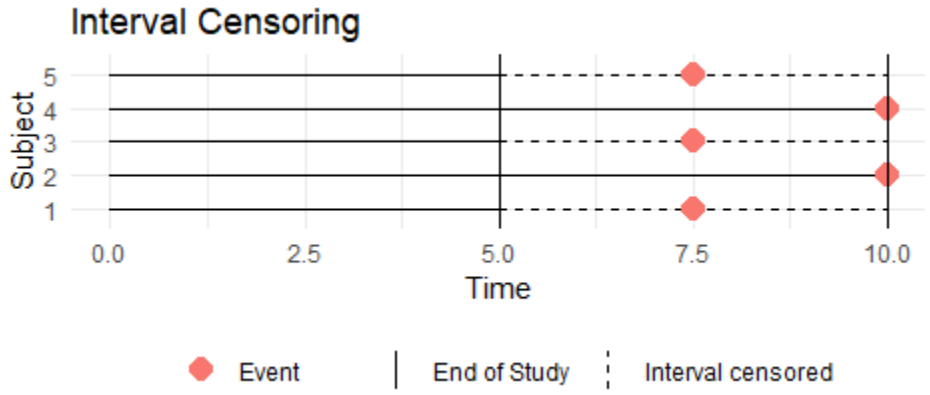

Figure 3. Interval Censoring.

2.2. Survival Function

Let T represent the survival time where $T \geq 0$ and $f(t)$ is the probability distribution function (PDF). The cumulative distribution function (CDF) is $F(t) = P(T \leq t)$ where $F(t)$ represents the probability of failure by some time t. The CDF gives the probability that the survival time is less than or equal to some time t. From the CDF, one can calculate the survival function. The survival function is:

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

This can also be written as $S(t) = P(T > t)$.

2.3. Hazard Function

The hazard function λ is the instantaneous risk of dying at time t, given the individual has survived till time t. It is written as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t) \big| (T \geq t))}{\Delta t} = \frac{f(t)}{S(t)} \quad (2)$$

2.4. Kaplan-Meier Estimate

The Kaplan-Meier estimator is one of the most common and simplest (Goel et. al., 2010) survival estimates; Kaplan and Meier's paper Nonparametric Estimation from Incomplete Observations has been cited over 57,000 times. The Kaplan-Meier estimator, also known as the product limit, is a non-parametric way to estimate the survival time of time to event data. The estimation is by a step function with discontinuities at time of death (or any event of interest). Time is split into many small intervals and the Kaplan-Meier estimate is the probability of surviving to a given length of time.

The Kaplan-Meier estimator was createdfrom the need to estimate the proportion P(t) of items in a population that exceed some time t from incomplete data, without making any assumptions about the form of P(t). The formulation of the Kaplan-Meier estimator is as follows. Suppose you have a random sample of n values (T_1, T_2, …. , T_N) of a random variable where each value has a 1/N probability. The F(t) value, or sample distribution, equals 1/N times the number of values in your random sample that are less than t. The estimate is a step function with discontinuities at the time of observed deaths even when observations are incomplete. Samples are observed lifetimes, where t_i=min(T_i, L_i) where T_i is the lifetime and L_i is the limit of the

observation, both T_i, $L_i \geq 0$. If $T_i \leq L_i$ then $t_i = T_i$ (a death), or $L_i < T_i$ and $t_i = L_i$ (a loss). When there are no losses, this reduced to the binomial estimate. The formula is written as:

$$P(t) = \prod_{i,t_i<t} 1 - \frac{d_i}{n_i} \tag{3}$$

Where d_i is the number of events at time t, and n_i is the number of subjects at risk prior to time t. The patients at risk can be defined as the number of patients that have not experienced an event yet. To find the survival probability at a specified time, multiply all previously calculated survival times for every time interval prior to the specified time (Kleinbaum & Klein, 2013).
It is a non-parametric function meaning there are no (possibly incorrect) assumptions made about the form of the distribution. The important property of non-parametric estimators is that if the age scale is transformed from t to t*=f(t), where f is strictly increasing. F_{hat}*(f(t)), the distribution function, is then equivalent to F_{hat}(t). Kaplan-Meier estimate is an example of univariate analysis. It works with a single categorical variable like gender or treatment. To compare the survival between groups, the logrank test is used.

2.5. The Logrank Test

The logrank test is a comparison between survival curves. It does not require any assumptions about the shape of the survival curve or distribution of survival times (Bland & Altman, 2004). The null hypothesis for the logrank test is that there is no significant difference between the survival curves, or there is no difference in probability of experiencing an event. At each time of event the logrank test compares the observed number of deaths to the expected number of deaths if there was really no difference in survival curves. If survival time is censored the individual is not considered at risk of dying in the subsequent calculations. The logrank test cannot determine the size of the difference, rather just an overall difference in probabilities of

events occurring (Bland & Altman, 2004). The modified Peto-Peto's weighted logrank test will be used when the survival curves cross. This is because the regular logrank test has a large bias when the survival curves cross (Bland & Altman, 2004).

2.5.1. The Pairwise Logrank Test

The pairwise logrank test is for pairwise comparisons between levels in a group. For example, if there are three groups, it can tell you which curves are different than eachother rather than an overall difference between the three. It also uses corrections for multiple testing. I will use the Benjamini Hochberg method for multiple comparisons.

2.6. Cox Proportional Hazards Model

The Cox Proportional Hazards Model proposed by Cox in 1972 is a semiparametric model used to calculate survival times based on the simultaneous effect of several covariates. The model is expressed by a hazard function, $h(t)$:

$$h(t) = h_0(t) * \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \quad (4)$$

Where t represents the survival time, $h(t)$ is the hazard function based on p covariates, h_0 is the baseline hazard which represents the value of the hazard if all the covaraties are equal to 0, and $(\beta_1, \beta_2, \ldots, \beta_p)$ are the coefficients for each of the p covariates. The hazard ratio is $\exp(\beta_1 + \beta_2 + \cdots + \beta_p)$. By maximizing the partial likelihood the parameter estimates are obtained.

The baseline hazard, $h_0(t)$ is the probability of experiencing the event of interest when all covariates are equal to 0. This is the only time-dependent part of the model, but an assumption of this test is that when comparing two indivduals, their hazard is proportional to each other regardless of time t. The baseline hazard can take any form, but the covariates enter the model linearly, making this a semiparametric model. The Cox Proportional Hazards Model cannot take into account any nonlinear effects, to do so one must use the extended cox model for

time changing covariates; this model is used because nonlinear effects violate the proportional hazard assumption. The baseline hazard is a function of time, but does not involve the covariate, while the exponential sum involves the covariates but not the time (Kleinbaum & Klein, 2013). Using the model above, if one of the covariates effects did change depending on time, that effect would not be taken into account in the model and rather the overall average effect would be used.

2.6.1. Hazard Ratio

The $exp(\beta_i)$ is called the hazard ratio. β_i, or the coefficient determines the impact size of the covariates. A value of β_i greater than zero, or in other words $exp(\beta_i)$ greater than one, decreases the survival time and is a poor diagnostic factor. The opposite is also true, a negative value of β_i , or an $exp(\beta_i)$ less than 1, increases the survival time and is considered good diagnostic factor. The hazard ratio (HR) is the ratio of the expected hazards corresponding to two levels of a covariate in a discrete case. In a continuous case a hazard ratio is the risk of event if the covariate in question increases by 1 unit. A way to compare two hazard functions:

$$HR(t) = \frac{h_2(t)}{h_1(t)} \tag{5}$$

Where $h_2(t)$ is the expected hazard of a patient in treatment A, and $h_1(t)$ is the expected hazard of a patient in a control group. The hazard ratio in this case represents the instantaneous risk at any point during the study period of a patient in treatment A compared to a patient in the control group. For example, say the hazard ratio was .5, then half of the patients in treatment A were experiencing an event compared to the control group. The proportional hazards assumption is that the hazard ratio, or effect size of the covariates, doesn't depend on time. This is equivalent to saying HR(t)=HR (Zahid, 2019).

2.6.2. Partial Likelihood Function

The coefficients are estimated through the partial likelihood function. “The partial likelihood is a product over the observed failure times of conditional probabilities, of seeing the observed failure, given the risk set at that time and that one failure is to happen.” (“The Proportional Hazards Regression Model,” n.d.). The partial likelihood uses (X_i, δ_i, Z_i) for each individual i. X_i is the failure time forpatient i,, either event or censored, δ_i is the censoring indicator, and Z_i is the value of the covariates. X_i can be any non-negative value and δ_i can be either 0 for censored or 1 for an event. Z_i can be any possible value for the covariate. The risk set, or the patients at risk for failure at time t is denoted as $(t) = \{\, i : X_i \geq t \,\}$. At each failure time X_j, the contribution to the partial likelihood is

$$L(B) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j \in R(X_i)} \lambda_j(X_i)} \right]^{\delta_i} \tag{6}$$

$$= \prod_{i=1}^{n} \left[\frac{\lambda_0(X_i)\exp(\beta' Z_i)}{\sum_{j \in R(X_i)} \lambda_0(X_i)\exp(\beta' Z_j)} \right]^{\delta_i}$$

$$= \prod_{i=1}^{n} \left[\frac{\exp(\beta' Z_i)}{\sum_{j \in R(X_i)} \exp(\beta' Z_j)} \right]^{\delta_i}$$

The partial likelihood is the value of exp (Z_i) at the observed failure time, divided by the sum of all other values of exp (Z_j), where the j individuals are in the risk set at the failure time. This value is multiplied over every failure time to get the value of the coefficient. The partial likelihood allows $\mathrm{h}_0(\mathrm{t})$, the baseline hazard function, to be excluded during the estimation of beta. The baseline hazard function can take any form, and is the nonparametric part of the model.

2.6.3. Assumptions of the Cox Proportional Hazards Model

Assumptions of the test:

1. Survival times are independent between individuals in the sample.
2. There is a multiplicative effect between the predictors and the hazard.
3. There is a constant hazard ratio over time, or in other words HR(t)=HR.

An important assumption of the Cox Model is the proportional hazards assumption. This states that for any two groups, the hazard ratio remains constant over time (independent of time) and proportional. There are a few ways to check this assumption. One way is if the hazard functions for any two groups cross when graphed then the proportional hazard assumption is not met. It is also possible for the proportional hazard assumption to not be met when the hazard functions do not cross. Another way to check is to plot the scaled Schoenfeld residuals. Any non-random pattern against time on the plot is an indication of non-proportional hazards (Kleinbaum & Klein, 2013). A third way is to add an interaction term with time and see whether the term is significant. Checking the proportional hazards assumption after fitting the proportional hazards model is the same as identifying time dependent coefficients (Zhang, 2018).

3. METHODOLOGY

3.1. SEER Data

The SEER database contains data from population based cancer registries across the United States. The benefit of population based is that there is diversity and a lack of bias presented by treatment and referral patterns (Davis et al., 1998). The SEER database has data from 18 different registries across the United States that covers 34.6% of the US population geographically ("Surveillance, Epidemiology, and End Results Program," n.d.). SEER database contains information on patient demographics, tumor site, tumor morphology, stage, treatment, and follow-up for vital status. The variables used in this analysis were as follows:

- Sex: Male or Female.
- Age: Patients at Diagnosis, split into 3 groups: Adolescent (<15), Adult (15-65), and Elderly (65+).
- Grade: Grade with values 1,2,3,4, and unknown.
- Survival Months: Patients survival time.
- Censory: Either alive or dead due to cause other than their cancer (1) or dead due to their cancer (2).
- Treatment: Either Surgery and Radiation, Surgery without Radiation, or No surgery or radiation.
- Year of Diagnosis: Year the patient was diagnosed.
- Number of Malignant tumors: How many malignant tumors a patient has.
- First Malignant Tumor: Either yes or no.
- Histology: Histologic ICD-O-3 for subsetting the data for Brain and CNS cancers.
- Surgery: Surgery performed.

- Radiation: Radiation performed.
- Number of Oberservations: Patient observation number.

3.2. Brain Cancer Background

Brain cancer is the growth of a tumor, either primary or secondary, in the brain. Primary brain tumors begin in the brain and secondary brain tumors are tumors that have metastasized to the brain from other parts of the body, most commonly lung or breast. Secondary brain tumors are more common than cancer that begins in the brain cells ("Learn About Brain Cancer: Information, Facts & Overview," 2020). Tumors in the brain can be benign, borderline, or malignant. Depending on the location and size of the tumor, benign brain tumors can be harmful. For this reason, in 2004 it became required for registries to start reporting benign tumors.

It is uncommon for primary brain tumors to move outside of the brain or Central Nervous System (CNS). Brain cancer is therefore graded instead of staged like most other cancers. Brain tumors are graded on how aggressive they appear under a microscope. Grades for primary tumors include 1-4 and secondary tumors are brain metastasis(TNM). The TNM staging is used only for secondary brain tumors. Primary brain tumors are classified by the cell or tissue the cancer effects, size and location, and resectability (likelihood it can be removed by surgery) ("Learn About Brain Cancer: Information, Facts & Overview," 2020). The focus of this paper is only primary malignant brain tumors.

To identify brain tumors, doctors use magnetic resonance imaging (MRI), Computed tomography (CT) scans, positron emission tomography (PET) scans, and biopsies. The main treatment for brain tumors is surgery. Surgery depends on the location, size, and patients overall health; surgery cannot always be performed. Radiation and chemotherapy are also treatments

after surgery. Brain cancer is a very active field of study and researchers are always looking at new technologies ("Brain Tumor," 2019).

3.3. Data Cleaning

From SEER data, analysis was conducted on 21,524 patients from 2004-2016 who were classified with primary Brain or CNS tumors. The presence of brain cancer was from histologic ICD-O-3 codes, which had a specific variable for brain groupings, HISTRECB. Patients with any previous or subsequent tumors were excluded from analysis because their survival was compromised by the additional cancers. Additionally patients diagnosed at death, patients with no age information, and patients with unknown surgery status were also excluded. Subsetting of the data was done in SEER*Stat. Demographic information was collected including age at diagnosis, gender, and year of diagnosis. Tumor information was collected including grade, histology, and treatment. Surgery was located in the SEER text files and radiation treatment was located in SEER*Stat. Patients were placed into 3 groups: patients that had surgery but no radiation, patients with surgery and radiation, and patients with neither surgery nor radiation. Primary brain tumors are graded on a scale of 1-4 by World Health Organization (WHO) Classification (Louis et. al., 2016). The grading scale was as follows:

Table 1. Grading Scale.

Grading Scale	Description
Grade 1	The tumor grows slowly and rarely spreads into nearby tissues. It may be possible to completely remove the tumor with surgery.
Grade 2	The tumor grows slowly but may spread into nearby tissues or recur.
Grade 3	The tumor grows quickly, is likely to spread into nearby tissues, and the tumor cells look very different from normal cells.
Grade 4	The tumor grows and spreads very quickly, and the tumor cells do not look like normal cells.

3.4. Analysis

After the data were subsetted, the distribution of grade was reviewed. The Kaplan-Meier estimator was used to plot survival curves for grade as well as treatment in each grade. Treatments for each grade (1-4) were patients that had surgery but no radiation, patients with surgery and radiation, and patients with neither surgery nor radiation. The logrank test was performed on treatments in each grade to see if there was a significant difference between the survival curves of the 3 treatment groups. The null hypothesis of the logrank test is that there are no difference in survival curves between the groups. The modified Peto-Peto's weighted logrank test was used when the survival curves crossed. The weighted logrank test is needed because the regular logrank test has a large bias when the survival curves cross. When the logrank test did conclude that there was a significant difference, the pairwise logrank test were used to test which pairs of survival curves differed from each other.

Finally, a Cox Proportional Hazards Model was fit to the data using grade, treatment, age, and sex as predictors. Age was split into 3 groups, adolescents (<15), adults (15-65), and elderly (65+) (Ritchie & Roser, 2019). The grade and treatment variables were the same as stated previously. Sex was either male or female. After testing the proportional hazards assumption, the model was stratified by grade. A plot of the hazard ratios and their respective 95% confidence intervals were shown and model interpretations were done.

4. RESULTS

First the distrubtion of grade was analyzed. Grade is how a brain tumor grows, and how doctors assess what treatment plan is best ("Understand how Brain Cancer is Staged and Graded," 2020). Approximately 6% of the patients were in grade 1, 13% of the patients were in grade 2, 7% of the patients were in grade 3, and 74% of the patients were in grade 4. The Kaplan-Meier survival curves were constructed for grade and treatment in each grade (1-4). Grade was split up tp show the survival times of treatment in each grade, and to test whether there is a significant difference between the effect on survival time of treatment in each grade. In Figures 6-9 the pink line corresponds to no surgery or radiation, the green line corresponds to both surgery and radiation and the blue line corresponds to surgery without radiation. The p-value on each Kaplan-Meier survival curve figure corresponds to the logrank test statistic. For the logrank test statistic:

Ho: $S_1(t) = S_2(t) = S_3(t)$

Ha: At least one of the $S_i(t)$'s is different for some time t

A significant p-value means the null hypothesis that all curves were the same could be rejected. Therefore it could be concluded that at least one of the curves survival times were significantly different. In order to determine which pairs of survival curves were significantly different from each other a pairwise logrank test was used.

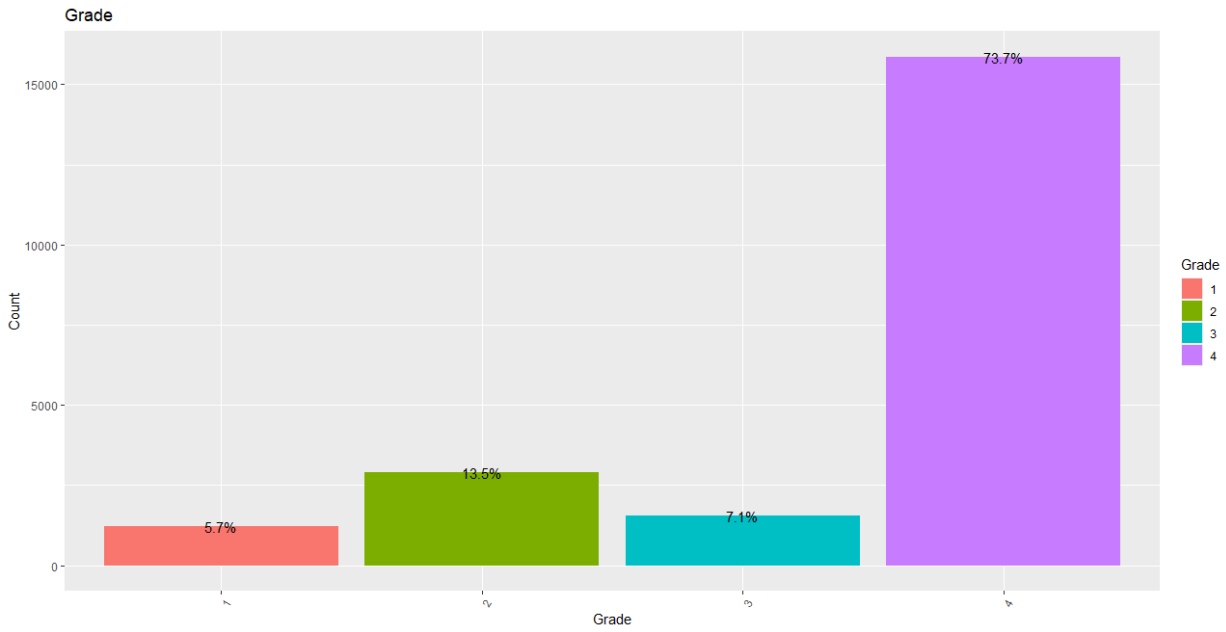

Figure 4. Graph of Grade.

4.1. Kaplan-Meier Estimator for Treatment in Each Grade

The Kaplan-Meier estimator was applied to treatment in each grade. The Kaplan-Meier estimator is a nonparametric way to measure survival probabilities so there is no underlying assumptions about the distribution of the survival times. In Figure 5, there was a significant difference between the survival curves of grade shown by a p-value of <.0001 from the modified Peto-Peto's weighted logrank test. Next, the effect of treatment in each grade was shown.

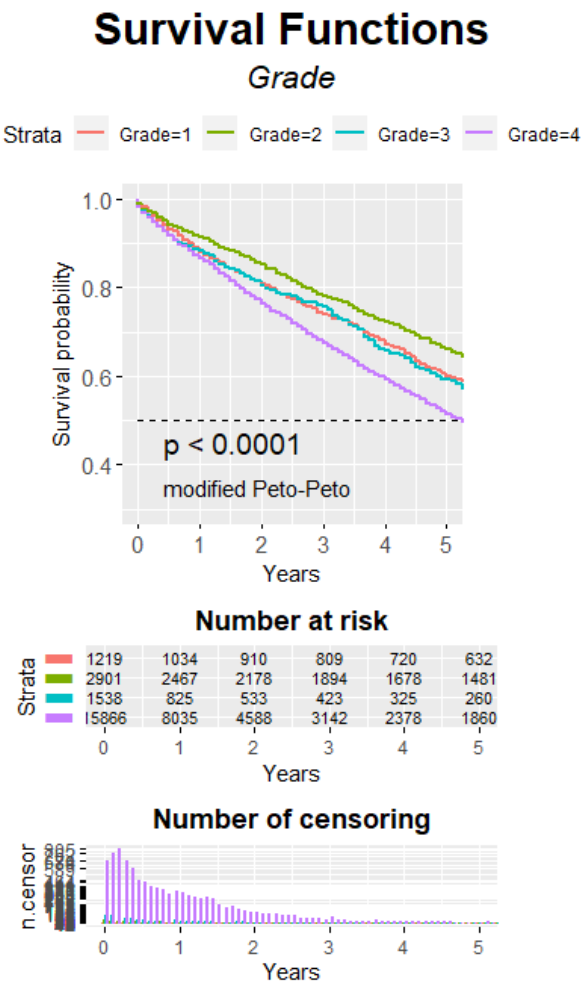

Figure 5. Graph of Kaplan-Meier Survival Curve for Grade.

From Figure 6, there was no significant difference between the survival times in the three treatments in grade 1 shown by a p-value of 0.15 from the modified Peto-Peto's weighted logrank test. The p-value is the probability of obtaining results at least as extreme as our data, given the null hypothesis is true. Any p-value greater than 0.05 means that the survival curves are likely the same and the null hypothesis cannot be rejected.

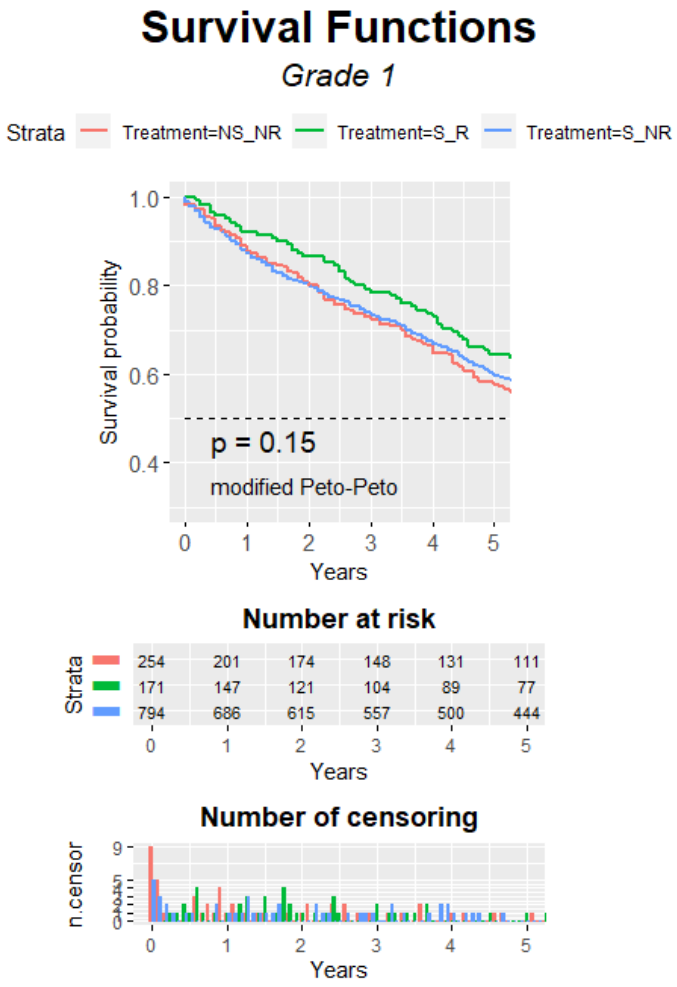

Figure 6. Graph of Kaplan-Meier Survival Curve for Treatment in Grade 1. (NS_NR : No Surgery or Radiation. S_R : Surgery and Radiation. S_NR : Surgery without Radiation.)

Next the survival cuves for treatment in grade 2 were analyzed in Figure 7. The modified Peto-Peto's weighted logrank test statistic p-value was equal to 0.20, meaning there was not a significant difference in the survival curves at any time point. The results were similar when looking at the survival curves for treatment in grade 3 in Figure 8. The modified Peto-Peto's weighted logrank test p-value was equal to 0.36 meaning there was not a significant difference between the survival curves at any time point.

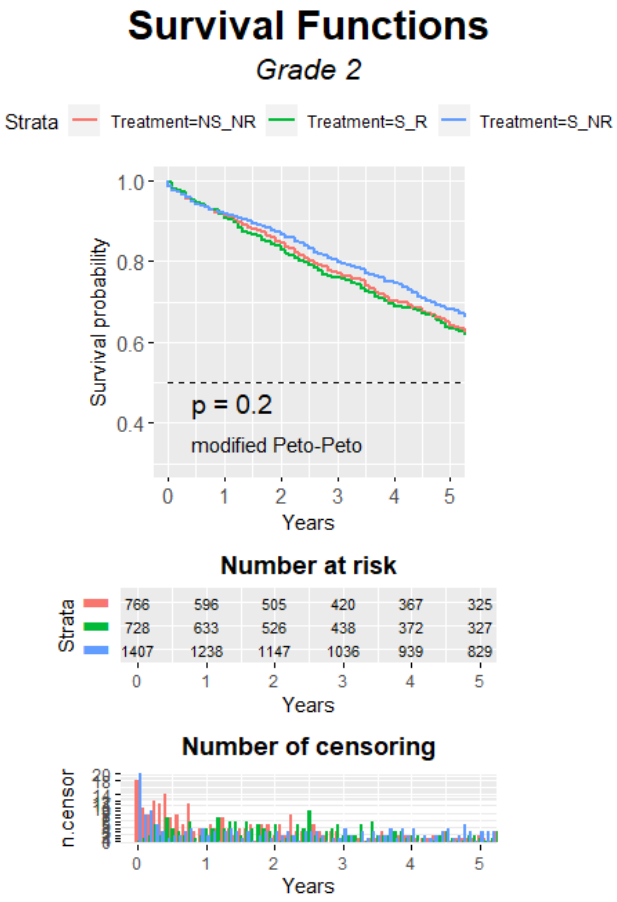

Figure 7. Graph of Kaplan-Meier Survival Curve for Treatment in Grade 2. (NS_NR : No Surgery or Radiation. S_R : Surgery and Radiation. S_NR : Surgery without Radiation.)

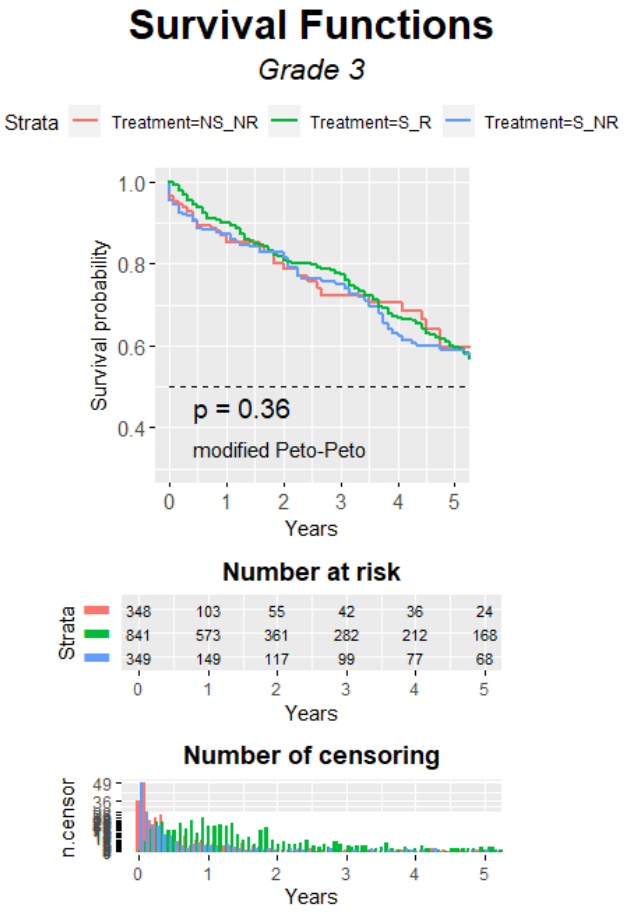

Figure 8. Graph of Kaplan-Meier Survival Curve for Treatment in Grade 3. (NS_NR : No Surgery or Radiation. S_R : Surgery and Radiation. S_NR : Surgery without Radiation.)

Lastly the Kaplan-Meier survival curves for treatment in grade 4 were analyzed in Figure 9. The results of the modified Peto-Peto's weighted logrank test were significant with a p-value of <0.0001. The null hypothesis that the survival curves were equal was rejected, and the pairwise logrank test was used to determine which pairs of treatments were significantly different from each other.

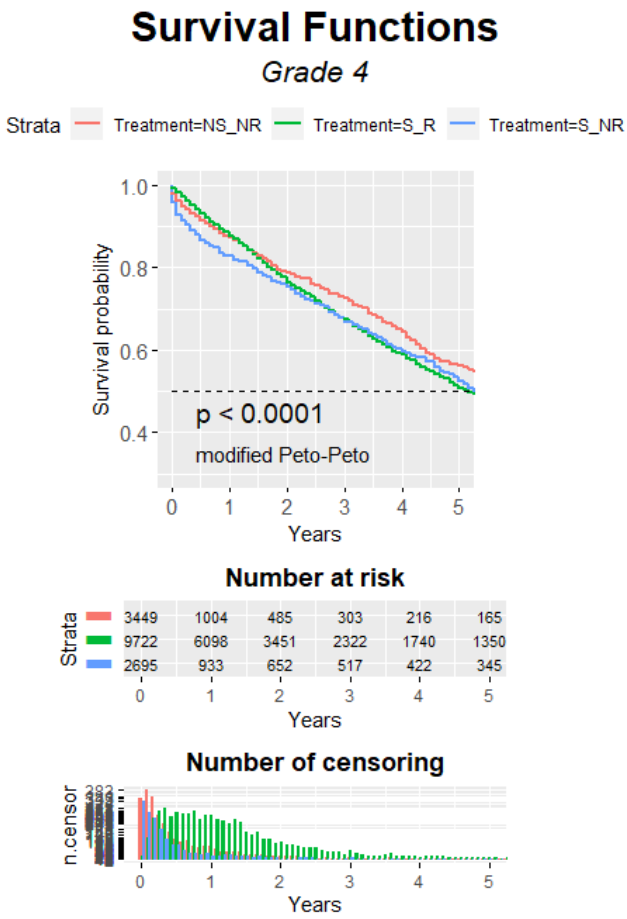

Figure 9. Graph of Kaplan-Meier Survival Curve for Treatment in Grade 4. (NS_NR : No Surgery or Radiation. S_R : Surgery and Radiation. S_NR : Surgery without Radiation.)

Results from the pairwise logrank test are below in Table 2. The Benjamini Hochberg method was used to control the type 1 error rate for false discoveries. For the pairwise logrank test, let: A= No Surgery or Radiation, B=Surgery and Radiation, and C=Surgery, No Radiation.

For Grade 4:

1. A vs B p=0.4630
2. A vs C p=6.2e-05
3. B vs C p=0.0040

Table 2. Results of Pairwise Logrank Test from Grade 4.

	No Surgery or Radiation	Surgery and Radiation
Surgery and Radiation	0.4630	N/A
Surgery, No Radiation	6.2e-05	0.0040

The pairwise logrank test shows that there was not a significant difference between patients that had surgery and radiation and patients that had neither surgery nor radiation. The pairwise logrank test shows that there was a significant difference between surgery without radiation and patients that had neither surgery nor radiation. There was also a significant difference between patients who had surgery and radiation and surgery without radiation.

4.2. The Cox Proportional Hazards Model

The Cox Proportional Hazards Model was fit to the data using grade, treatment, age, and sex as predictors. The Cox Proportional Hazards Model allows the simulatanous effect of several variables on survival time. The covarites used in this analysis were chosed based on clinically relevant factors from literature. The first model was fit with grade, treatment, age, and sex as predictors. When testing for proportional hazards, the first model fails this assumption with a global p-value of 8.3e-07. In Figure 10, when looking at the Schoenfeld residual plot for grade, grade violates the proportional hazards assumption. It was concluded that the effect of grade dimished over time as shown by a decreasing value for beta over time.

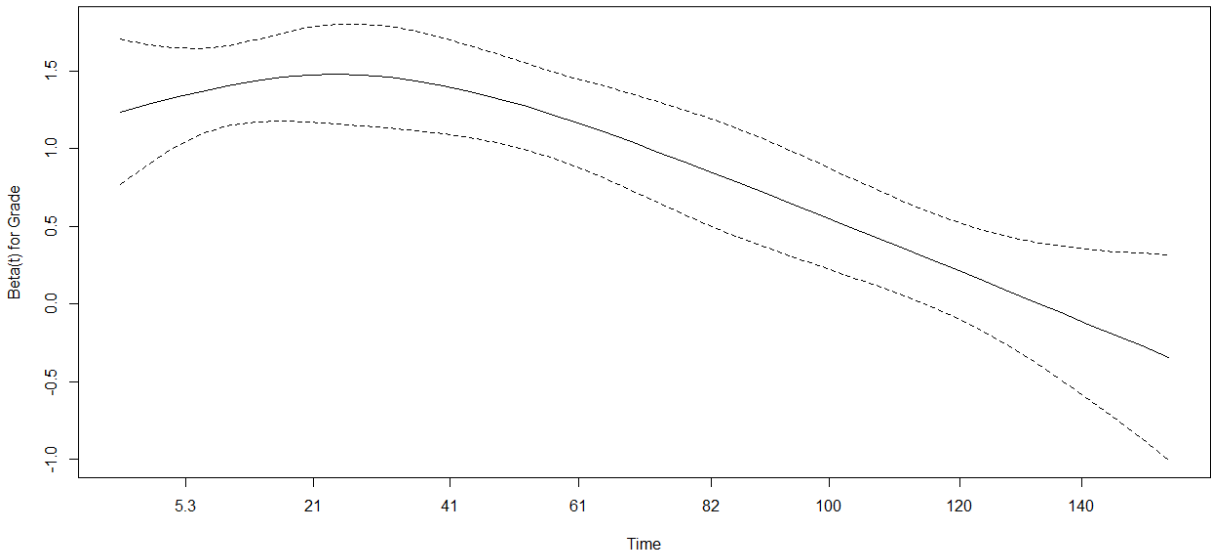

Figure 10. Schoenfeld Residual Plot of Beta(t) for Grade.

Next, the model was stratified by grade using the predictors treatment, age, and sex. The model was stratified by grade because grade violates the proportional hazard assumption. The global p-value for the stratified model is 0.0165. At the 1% significance level, the proportional hazards assumption is met. The overall model was significant with a p-value of <2.2e-16 from the likelihood ratio test. The results of the Cox Proportional Hazards Model is in Table 3.

Table 3. Cox Proportional Hazards Model Output.

Variable	Coefficient	Exp(Coefficient)	SE(Coefficient)	P-Value
Treatment: Surgery and Radiation	0.0206	1.0208	0.0331	0.5340
Treatment: Surgery without radiation	0.0552	1.0568	0.0342	0.1060
Sex: Male	0.0250	1.0253	0.0216	0.2480
Age: Adult	-0.0472	0.9538	0.0306	0.1230
Age: Elderly	0.3262	1.3857	0.0441	1.52e-13

The coefficient is the value for beta. A positive value for beta increases survival time, while a negative value decreases survival time. For example, being in the elderly age group increases your chance of death, while being in the adult age group decreases it. The exp(coefficient) is known as the hazard ratio. The hazard ratio is the effect size of the covarite. A hazard ratio greater than 1 increases chances of a death occurring in the group, while a hazard ratio less than one decreases the chance of death in the group. A hazard ratio of 1 has no effect on survival time. The hazard ratio and their 95% confidence intervals are shown in Figure 11. The hazard ratio of male was 1.02, this means that men had a 2% increased risk of death compared to women. For the treatment covariates, each variable was being compared to the treatment group of patients with neither surgery nor radiation. Patients in the surgery and radiation treatment group had a 2% increased risk of death compared to patients who had neither surgery nor radiation. Patients that had surgery without radiation had a 5% increased risk of death compared to patient with neither surgery nor radiation. For the age covarites, the groups adult and elderly were compared to adolescence. The adult age group had a 5% decreased risk of death compared to adolescents. The elderly age group had a 38% increased risk of death compared to adolescents.

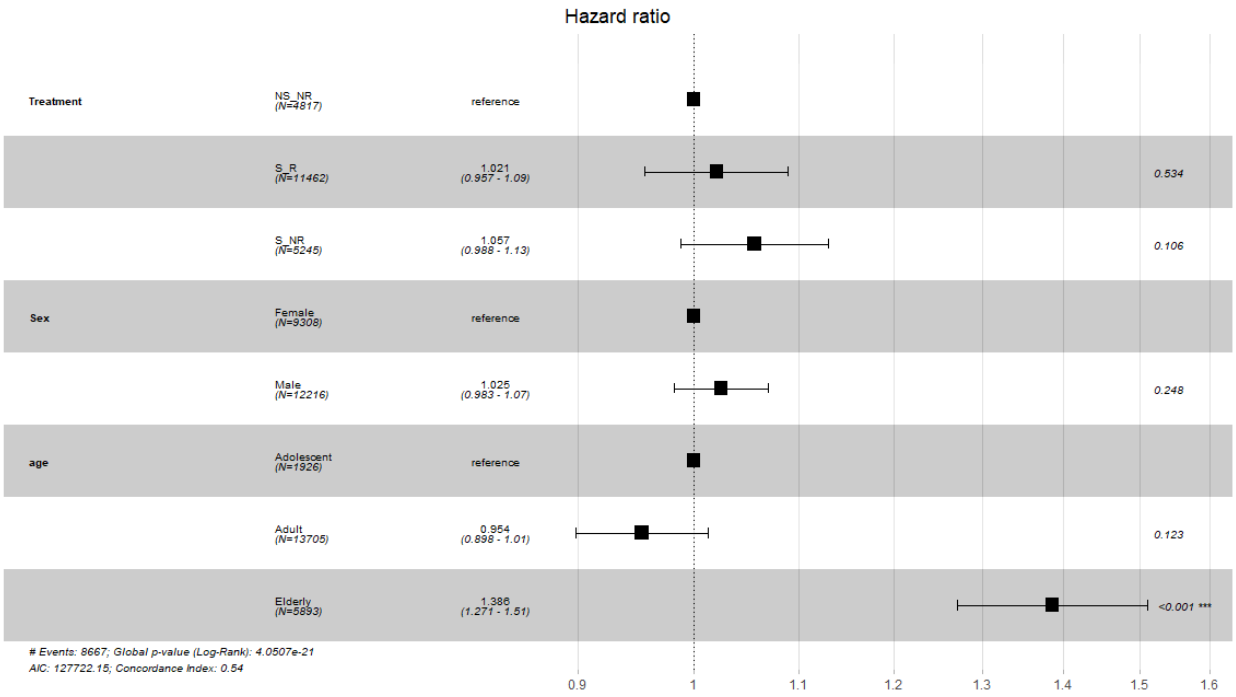

Figure 11. Hazard Ratios in Cox Proportional Hazards Model.

5. CONCLUSION

The effects of treatment in each grade were shown for patients with primary malignant brain tumors who had no previous or subsequtent cancers. Data from the National Cancer Instituates Surveillance, Epidemiology, and End Results (SEER) program was used. After comparing treatments using the Kaplan-Meier estimator and the logrank test, there was not a significant difference in the survival curves of the three treatments until grade 4. In grade 4 there was a significant difference between patients who had surgery without radiation and patients who had no surgery or radiation and patients that had both surgery and radiation. To measure the simultaneous effect on the risk of death for treatment, age, and sex a Cox Proportional Hazards Model was used. The results from the Cox Proportional Hazards Model showed that patients in the adult age group had an increased survival time compared to adolsecents, and patients in the elderly age group had a decreased survival time compared to adolescents. The model also showed the females have a slightly lower risk when compared to males. Patients who had surgery and radiation had a 2% increased risk of death compared to patients with neither surgery nor radiation, while patients who had surgery without radiation had a 5% increased risk of death compared to patients with neither surgery nor radiation.

This study was useful to show survival time for patients with primary malignant brain tumors. Knowing patient survival times and the probability of death in each treatment is valuable for patients who are on one of the treatment paths. Limitations of the study include additional factors that affect survival rates and treatment like histology.

6. FUTURE RESEARCH

Future work includes including additional variables in the Cox Proportional Hazards Model. The survival rates by region is of interest to see if patients survival times differ significantly based on their geographic location in the United States. There is also another treatment variable in SEER*Stat corresponding to chemotherapy. Extending treatment to chemotherapy would be of interest to show the effect of chemotherapy on survival times. Chemotherapy only has two reporting options, Yes and No/Unknown. For this reason it was excluded in this analysis.

REFERENCES

Bland, J. M., & Altman, D. G. (2004). The logrank test. BMJ (Clinical research ed.), 328(7447), 1073. https://doi.org/10.1136/bmj.328.7447.1073.

Brain tumor. (2019). http://www.mayoclinic.org/diseases-conditions/brain-tumor/diagnosis-treatment/drc-20350088.

Davis, F. G., Freels, S., Grutsch, J., Barlas, S., & Brem, S. (1998). Survival rates in patients with primary malignant brain tumors stratified by patient age and tumor histological type: an analysis based on Surveillance, Epidemiology, and End Results (SEER) data, 1973–1991, *Journal of Neurosurgery*, *88*(1), 1-10. https://thejns.org/view/journals/j-neurosurg/88/1/article-p1.xml.

Davis, F.G., McCarthy, B.J., Freels, S., Kupelian, V. and Bondy, M.L. (1999), The conditional probability of survival of patients with primary malignant brain tumors. Cancer, 85: 485-491. doi:10.1002/(SICI)1097-0142(19990115)85:2<485::AID-CNCR29>3.0.CO;2-L.

Deorah, S., Lynch, C. F., Sibenaller, Z. A., & Ryken, T. C. (2006). Trends in brain cancer incidence and survival in the United States: Surveillance, Epidemiology, and End Results Program, 1973 to 2001, *Neurosurgical Focus FOC*, *20*(4), E1. https://thejns.org/focus/view/journals/neurosurg-focus/20/4/foc.2006.20.4.e1.xml.

Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, *1*(4), 274–278. https://doi.org/10.4103/0974-7788.76794.

Hankinson, T. C., Dudley, R. R., Torok, M. R., Patibandla, M., Dorris, K., Poonia, S., Wilkinson, C., Bruny, J. L., Handler, M. H., & Liu, A. K. (2016). Short-term mortality following surgical procedures for the diagnosis of pediatric brain tumors: outcome

analysis in 5533 children from SEER, 2004–2011, *Journal of Neurosurgery: Pediatrics PED, 17*(3), 289-297. https://thejns.org/pediatrics/view/journals/j-neurosurg-pediatr/17/3/article-p289.xml.

Kleinbaum, D. G., & Klein, M. (2013). *Survival Analysis: A Self-Learning Text, Third Edition*. New York, NY: Springer Science+Business Media.

Learn About Brain Cancer: Information, Facts & Overview. (2020). http://www.cancercenter.com/cancer-types/brain-cancer/about.

Louis, D.N., Perry, A., Reifenberger, G. *et al.* (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131,** 803–820. https://doi.org/10.1007/s00401-016-1545-1.

Perkins, S.M., Mitra, N., Fei, W. *et al.* (2012). Patterns of care and outcomes of patients with pleomorphic xanthoastrocytoma: a SEER analysis. *J Neurooncol* 110, 99–10. https://doi.org/10.1007/s11060-012-0939-8.

Ritchie, H., & Roser, M. (2019). *Age Structure*. https://ourworldindata.org/age-structure.

Surveillance, Epidemiology, and End Results Program. (n.d.). https://seer.cancer.gov/.

The Proportional Hazards Regression Model. (n.d.). University of California at San Diego. https://www.math.ucsd.edu/~rxu/math284/slect5.pdf.

Therneau, T., Crowson, C., & Atkinson, E. (2020). Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. Retrieved from https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf.

Understand how Brain Cancer is Staged and Graded. (2020). Retrieved from https://www.cancercenter.com/cancer-types/brain-cancer/grades.

Zahid, T. (2019). Cox Proportional-Hazards Model. http://www.sthda.com/english/wiki/cox-proportional-hazards-model.

Zhang, Z., Reinikainen, J., Adeleke, K., Pieterse, M., & Groothuis-Oudshoorn, C. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals Of Translational Medicine, 6*(7), 11. doi:10.21037/atm.2018.02.12.

APPENDIX. R CODE

```
#install libraries
library(ggplot2)
library(survival)
library(survminer)
library(plyr)
library(tidyverse)

##input data from SEER*STAT - file contains only malignant brain or CNS tumor
s
allData=read.csv(file="C:/Users/Madison/Downloads/SeerDataBrain1_4-29-2020.cs
v", header=TRUE)

#clean data
data=allData %>%
  filter(Year_of_diagnosis>=2004 & Totalnumberofinsitumalignanttu==1 &
SEERcausespecificdeathclassifi!="N/A not first tumor") %>%
  #clean data for total number of malignant tumors, first tumors only, and ye
ars 2004-2016.
   mutate(censory=factor(SEERcausespecificdeathclassifi,
                         levels=c("Dead (attributable to this cancer dx)",
                                  "Alive or dead of other cause",
                                  "Dead (missing/unknown COD)"),
                         labels=c(2,1,2)),
          surg=factor(Reasonnocancerdirected_surgery,
                      levels=c("Surgery performed",
                               "Not recommended",
                               "Not performed, patient died prior to recommend
ed surgery",
                               "Not recommended, contraindicated due to other
cond; autopsy only   (1973-2002)",
                               "Recommended but not performed, patient refused
",
                               "Recommended but not performed, unknown reason"
,
                               "Recommended, unknown if performed",
                               "Unknown; death certificate; or autopsy only (2
003+)"),
                      labels=c("Surgery", "No", "No", "No", "No", "No", "Unkno
wn", "Unknown")),
          rad=factor(Radiationsequence_with_surgery,
                     levels=c("Intraoperative rad with other rad before/after
surgery",
                              "Intraoperative radiation",
                              "Radiation after surgery",
                              "Radiation before and after surgery",
                              "Radiation prior to surgery",
                              "Sequence unknown, but both were given",
```

```
                              "Surgery both before and after radiation",
                              "No radiation and/or cancer-directed surgery"),
                    labels=c("Radiation & Surgery", "Radiation","Radiation &
Surgery", "Radiation & Surgery", "Radiation & Surgery", "Radiation & Surgery"
,"Radiation & Surgery","No Radiation")),
         trt=paste(surg, rad, sep=","),
         Treatment=factor(trt,
                          levels=c("No,No Radiation",
                                   "No,Radiation & Surgery",
                                   "Surgery,No Radiation",
                                   "Surgery,Radiation",
                                   "Surgery,Radiation & Surgery",
                                   "Unknown,No Radiation",
                                   "Unknown,Radiation & Surgery"),
                          labels=c("No Surgery or Radiation","Surgery and Radiat
ion","Surgery, No Radiation","Surgery and Radiation","Surgery and Radiation",
"Unknown Surgery and No radiation","Surgery and Radiation")))


##view Treatment & filter out unknown surgery values
table(data$Treatment)
data$trtUse=as.vector.factor(data$Treatment)
data=data %>%
  filter(Treatment!="Unknown Surgery and No radiation") %>%
  droplevels()

#filter out unknown grade value
table(data$Grade)
#data$Grade=as.vector.factor(data$Grade)
data=data %>%
  filter(Grade!="Unknown") %>%
mutate(Grade=factor(Grade,
             levels=c("Well differentiated; Grade I",
                      "Moderately differentiated; Grade II",
                      "Poorly differentiated; Grade III",
                      "Undifferentiated; anaplastic; Grade IV"),
             labels=c("1","2","3","4")))

#ad id statement
data$ID=1:nrow(data)
#change censory to numeric
data$censory=as.numeric(data$censory)
table(data$censory)
#condense data to only variables used for analysis
data=data %>%
  select(ID,Treatment, Grade, Survival_months, censory, Age_at_diagnosis, His
tologyrecodeBrain_groupings, Sex, Race_recode_White_Black_Other )
```

```r
#plot grade
p=ggplot(data.frame(data$Grade), aes(x=data$Grade, fill=data$Grade)) + geom_b
ar(aes(y = (..count..))) + geom_text(aes(y = (..count..),label = scales::perc
ent((..count..)/sum(..count..))), stat="count") + labs(title ="Grade", x = "G
rade", y = "Count")+theme(axis.text.x = element_text(angle = 60, hjust = 1))
p+ labs(fill = "Grade")

#grade 1 analysis

data=data %>%
mutate(Treatment=factor(Treatment,
             levels=c("No Surgery or Radiation",
                      "Surgery and Radiation",
                      "Surgery, No Radiation"),
             labels=c("NS_NR","S_R","S_NR")))


grade1=data[which(data$Grade==1),]

km.grade1=survfit(Surv(Survival_months,censory) ~ Treatment, data=grade1)
summary(km.grade1,c(12, 24, 36, 48, 60))

ggsurvplot(km.grade1, pval=TRUE, conf.int=FALSE,
           title="Survival Functions",log.rank.weights = "S2",pval.method = T
RUE,
            pval.method.coord = c(5, .35), # coordinates for the name
           pval.method.size = 4, pval.coord=c(5,.45)  ,
           subtitle="Grade 1",font.title = c(22, "bold", "black"),
           #change theme
           ggtheme = theme_grey() + theme(plot.title = element_text(hjust = 0
.5, face = "bold"))+
             theme(plot.subtitle = element_text(hjust = 0.5, size = 16, face
= "italic"))+theme(aspect.ratio = 1),  tables.theme =  theme(aspect.ratio = 0
.2),
           #change censor
           censor.shape = NULL, censor.size = 0,
           #x scale
           xlab="Years",break.x.by=12,xlim=c(0,60),xscale="m_y", ylim=c(.3,1)
,
           # changes the tick label on x axis
           font.xtickslab=c(11,"plain"),
           font.ytickslab=c(11,"plain"),
           risk.table = TRUE,
           risk.table.height = 0.2,
           risk.table.fontsize = 3.0 ,
           risk.table.y.text = FALSE,
           ncensor.plot = TRUE,
           ncensor.plot.height = 0.2,
           surv.median.line = "hv") # add the median survival pointer.
```

```r
#Grade 2 Analysis
grade2=data[which(data$Grade==2),]

km.grade2=survfit(Surv(Survival_months,censory) ~ Treatment, data=grade2)
summary(km.grade2,c(12, 24, 36, 48, 60))

ggsurvplot(km.grade2, pval=TRUE, conf.int=FALSE,
           title="Survival Functions",log.rank.weights = "S2",pval.method = T
RUE,
           pval.method.coord = c(5, .35), # coordinates for the name
           pval.method.size = 4, pval.coord=c(5,.45)  ,
           subtitle="Grade 2",font.title = c(22, "bold", "black"),
           #change theme
           ggtheme = theme_grey() + theme(plot.title = element_text(hjust = 0
.5, face = "bold"))+
             theme(plot.subtitle = element_text(hjust = 0.5, size = 16, face
= "italic"))+theme(aspect.ratio = 1),  tables.theme =  theme(aspect.ratio = 0
.2),
           #change censor
           censor.shape = NULL, censor.size = 0,
           #x scale
           xlab="Years",break.x.by=12,xlim=c(0,60),xscale="m_y", ylim=c(.3,1)
,
           # changes the tick label on x axis
           font.xtickslab=c(11,"plain"),
           font.ytickslab=c(11,"plain"),
           risk.table = TRUE,
           risk.table.height = 0.2,
           risk.table.fontsize = 3.0 ,
           risk.table.y.text = FALSE,
          ncensor.plot = TRUE,
          ncensor.plot.height = 0.2,
          surv.median.line = "hv") # add the median survival pointer.

grade3=data[which(data$Grade==3),]

km.grade3=survfit(Surv(Survival_months,censory) ~ Treatment, data=grade3)
summary(km.grade3,c(12, 24, 36, 48, 60))

ggsurvplot(km.grade3, pval=TRUE, conf.int=FALSE,
           title="Survival Functions",log.rank.weights = "S2",pval.method = T
RUE,
           pval.method.coord = c(5, .35), # coordinates for the name
           pval.method.size = 4, pval.coord=c(5,.45)  ,
           subtitle="Grade 3",font.title = c(22, "bold", "black"),
           #change theme
           ggtheme = theme_grey() + theme(plot.title = element_text(hjust = 0
.5, face = "bold"))+
             theme(plot.subtitle = element_text(hjust = 0.5, size = 16, face
```

```
= "italic"))+theme(aspect.ratio = 1),  tables.theme =  theme(aspect.ratio = 0
.2),
           #change censor
           censor.shape = NULL, censor.size = 0,
           #x scale
           xlab="Years",break.x.by=12,xlim=c(0,60),xscale="m_y", ylim=c(.3,1)
,
           # changes the tick label on x axis
           font.xtickslab=c(11,"plain"),
           font.ytickslab=c(11,"plain"),
            risk.table = TRUE,
           risk.table.height = 0.2,
           risk.table.fontsize = 3.0 ,
           risk.table.y.text = FALSE,
          ncensor.plot = TRUE,
          ncensor.plot.height = 0.2,
          surv.median.line = "hv") # add the median survival pointer.

grade4=data[which(data$Grade==4),]

km.grade4=survfit(Surv(Survival_months,censory) ~ Treatment, data=grade4)
summary(km.grade4,c(12, 24, 36, 48, 60))

ggsurvplot(km.grade4, pval=TRUE, conf.int=FALSE,
           title="Survival Functions",log.rank.weights = "S2",pval.method = T
RUE,
          pval.method.coord = c(5, .35), # coordinates for the name
          pval.method.size = 4, pval.coord=c(5,.45)  ,
           subtitle="Grade 4",font.title = c(22, "bold", "black"),
           #change theme
           ggtheme = theme_grey() + theme(plot.title = element_text(hjust = 0
.5, face = "bold"))+
             theme(plot.subtitle = element_text(hjust = 0.5, size = 16, face
= "italic"))+theme(aspect.ratio = 1),  tables.theme =  theme(aspect.ratio = 0
.2),
           #change censor
           censor.shape = NULL, censor.size = 0,
           #x scale
           xlab="Years",break.x.by=12,xlim=c(0,60),xscale="m_y", ylim=c(.3,1)
,
           # changes the tick label on x axis
           font.xtickslab=c(11,"plain"),
           font.ytickslab=c(11,"plain"),
            risk.table = TRUE,
           risk.table.height = 0.2,
           risk.table.fontsize = 3.0 ,
          risk.table.y.text = FALSE,
           ncensor.plot = TRUE,
          ncensor.plot.height = 0.2,
```

```
          surv.median.line = "hv") # add the median survival pointer.



##Best treatment on each grade using pairwise log rank test
res4=pairwise_survdiff(Surv(Survival_months,censory) ~ Treatment, data =grade
4)
res4

#no difference between surgery and radiation and no surgery or radiation.
grade4$Treatment=revalue(grade4$Treatment, c("No Surgery or Radiation"="Combi
ned",
                                             "Surgery and Radiation"="Combined",
                                             "Surgery, No Radiation"="Surgery, No
Radiation"))

km.grade4.2=survfit(Surv(Survival_months,censory) ~ Treatment, data=grade4)
ggsurvplot(km.grade4.2, pval=TRUE, conf.int=FALSE,
           title="Survival Functions",log.rank.weights = "S2",pval.method = T
RUE,
          pval.method.coord = c(5, .35), # coordinates for the name
          pval.method.size = 4, pval.coord=c(5,.45)  ,
           subtitle="Grade 4 Combined",font.title = c(22, "bold", "black"),
           #change theme
           ggtheme = theme_grey() + theme(plot.title = element_text(hjust = 0
.5, face = "bold"))+
             theme(plot.subtitle = element_text(hjust = 0.5, size = 16, face
= "italic"))+theme(aspect.ratio = 1),  tables.theme =  theme(aspect.ratio = 0
.2),
           #change censor
           censor.shape = NULL, censor.size = 0,
           #x scale
           xlab="Years",break.x.by=12,xlim=c(0,60),xscale="m_y", ylim=c(.3,1)
,
           # changes the tick label on x axis
           font.xtickslab=c(11,"plain"),
           font.ytickslab=c(11,"plain"),
            risk.table = TRUE,
           risk.table.height = 0.2,
           risk.table.fontsize = 3.0 ,
          risk.table.y.text = FALSE,
           ncensor.plot = TRUE,
          ncensor.plot.height = 0.2,
          surv.median.line = "hv")

#start cox ph

##split age at diagnosis
table(data$Age_at_diagnosis)
```

```r
data$age=cut(data$Age_at_diagnosis, breaks=c(0,15,65,Inf),labels = c("Adolesc
ent", "Adult", "Elderly"),  right = FALSE)


#fit coxph function
coxp=coxph(Surv(Survival_months,censory) ~ strata(Grade)+Treatment+Sex+age, d
ata=data)
coxp
test.ph=cox.zph(coxp)
test.ph
trace(ggforest, edit = TRUE)
ggforest(coxp, data=data)

#create censoring
#right
right=tibble(Subject = as.factor(1:5),
                   Years = c(9,6,12,15,10),
                   censor = c("event", "censor", "event", "censor", "event" )
)

ggplot(right, aes(x=Subject)) +
  geom_linerange(aes(ymin = 0, ymax = Years))+
  geom_hline(yintercept=c(15))+
    geom_point(data = right,
               aes(Subject, Years, color = censor, shape = censor),
               size = 3)+ ylab("Time")+
coord_flip()+
theme_minimal() +
    theme(legend.title = element_blank(),
          legend.position = "bottom") +ggtitle("Right Censoring")

#left
left=tibble(Subject = as.factor(1:5),
                   Years = c(6,5,5,8,10),
                   censor = c("Event", "Event", "Event", "Event", "Event"))

ggplot(left, aes(x=Subject)) +
  geom_linerange(aes(ymin = 0, ymax = Years))+
  geom_hline(yintercept=c(5))+
    geom_point(data = left,
               aes(Subject, Years, color = censor, shape = censor),
               size = 3)+ ylab("Time")+
coord_flip()+
theme_minimal() +
    theme(legend.title = element_blank(),
          legend.position = "bottom") +ggtitle("Left Censoring")


#interval
```

```r
interval=tibble(Subject = as.factor(1:5),
                left=c(5,5,5,5,5),
                  right = c(10,10,10,10,10),
                interval=left+(right-left)/2,
                   censor = c(1,0, 1, 0, 1 ),
                censored=c("Event", "Event", "Event", "Event", "Event"))

ggplot(interval, aes(x=Subject)) +
  geom_linerange(aes(ymin = 0, ymax = left)) +
    geom_linerange(aes(ymin = left, ymax = right, linetype = as.factor(censor
))) +
    geom_point(aes(y = ifelse(censor, interval, right),color=censored, shape=
censored),
        size = 4)+
  geom_hline(yintercept=c(5,10))+
  scale_linetype_manual(name = "Censoring", values = c(1, 2),
        labels = c("End of Study", "Interval censored"))+
  xlab("Subject") +  ylab("Time")+
theme_minimal() +
coord_flip()+
    theme(legend.title = element_blank(),
          legend.position = "bottom") +ggtitle("Interval Censoring")
```