Prognostic Modeling of Brain Tumor Survival

Aryan Mukherjee
University High School, Irvine, CA

Abstract

Brain tumors, including gliomas and meningiomas, vary in prognosis depending on tumor class, location, volume, and patient-specific factors such as health, gender, and treatment. Using an open-source dataset from the Masaryk Memorial Cancer Institute in Brno, we analyze survival outcomes of patients treated with radiation therapy between 2004 and 2011. Models considered include the Cox proportional hazards model, the exponential parametric model, the log-logistic model, and the random survival forest. Comparative results highlight differences in fit, predictive accuracy, and interpretability, offering guidance for biostatistical approaches to brain tumor survival analysis.

Keywords: Brain tumor, survival data, censoring, Cox proportional hazards model, hazard, log-rank test, exponential parametric model, log-logistic model, random survival forest

1 Introduction

1.1 Preliminary Information

Brain and other central nervous system (CNS) tumors are responsible for a significant health burden in the United States. Between 2016 and 2020, malignant CNS tumors caused approximately 17,206 deaths per year, corresponding to an annual mortality rate of 4.42 per 100,000 population. [1] Gliomas represent around 26% of all brain tumors, with glioblastoma being the most common malignant subtype; meningiomas are the most prevalent non-malignant tumor. [1] These figures underscore the urgent need to identify prognostic factors that influence survival.

Survival analysis is the statistical framework for modeling time-to-event data, in the presence of censored data. Censoring occurs when an individual who hasn't experienced an event year leaves the study, or the study concludes before an event is observed in this individual.

This paper is part two of a three-part series on predictive modeling for brain tumor survival outcomes. It focuses on four approaches: the Cox proportional hazards model, the exponential model, the log-logistic model, and the random survival forest model. We further employ log-rank tests for both two-group and four-group comparisons to assess the significance of predictors.

The structure of this paper includes data description, a concise literature overview, methodological details, model-specific outcomes, and implications for prognosis and model selection.

1.2 Data Description

The study analyzes 87 patients treated from 2004 to 2011 at the Masaryk Memorial Cancer Institute in Brno, all of whom received radiation therapy. [2] The original dataset had 88 patients, but we removed one due to a missing value. 52 of the remaining patients are censored. Here is the complete list of variables:

- **Gender:** Patient's self-reported gender (male/female).
- **Diagnosis:** Type of brain tumor (e.g., glioma, meningioma).
- **Location:** Tumor location in the brain, categorized as infratentorial (below the tentorium cerebelli, affecting cerebellum or brainstem) or supratentorial (above the tentorium, affecting cerebral hemispheres).
- **Stereotactic Method:** Type of radiation therapy applied—SRS (stereotactic radiosurgery, single high-dose) or SRT (stereotactic radiotherapy, multiple lower-dose fractions).
- **Karnofsky Index (KI):** A performance status score ranging from 0 to 100, with higher values indicating better ability to carry out daily activities.
- **Gross Tumor Volume (GTV):** The measured physical volume of the tumor (in cm^3), derived from imaging.
- **Duration:** Survival time, measured in months from study enrollment to death (or last follow-up if censored).
- **Censoring:** Patient status at last observation— $\delta = 0$ = censored (patient alive or lost

to follow-up at last contact), $\delta = 1$ = event occurred (death).

1.3 Overview of the Literature

Many cardinal works are associated with the field of survival analyses, several of which pertain to the study of brain tumors. Recent notable works are enumerated in this section.

Mahmoudi et. al employed a radiomic model to determine significant predictors of survival of glioblastoma patients from MRI data, demographics, and tumor biomarkers, and prognosis was characterized by a binary metric of over or under 18 months for predicted survival. Stepwise logistic regression was performed on significant variables, namely radiomics, age, and the methylation status of the MGMT gene. An AUC/Sensitivity/Specificity metric was used for model accuracy determination, and the radiogenomic model aggregating all significant predictors yielded values of 0.89/100/78.6 respectively. [3]

Awuah et al. further reviews the prospect of radiomic models coupled with non-imaging models, discussing the benefits of a combined AI model incorporating random survival forest or other nonparametric models with Artificial Neural Networks (ANNs), reducing the mean absolute error of survival prediction by 3.4 months. [4]

Rikan et al. developed five machine learning models to predict glioblastoma patient survival, obtaining the greatest regression accuracy from Deep Neural Networks (DNNs) and the second greatest from Random Forest (RF). They used the Ordinary Least Square (OLS) method to compute feature importance, finding all but gender and marital status to be significant. [5]

Selingerová et al. compared a kernel smoothing method with Cox regression for primary brain tumor patients. They found a significant p-value for gender, diagnosis, KI, and Stereotactic method, splitting KI at the 80% mark. The lack of evidence of a significant contribution by the covariate of age in fitting the model differed from the kernel estimate's conclusion, and rather than reflecting model comparison in measures of accuracy, the kernel estimate is deemed a better option due to imperfect Cox model assumptions. [6]

Noia et al. explored performance of various AI models for Overall Survival (OS) time prediction, focusing on gliomas. Of their models, Random Forest received a C-index of 0.91 on a testing dataset, which was used due to the large number of cases that could be reported for splits. This was the most accurate model, out-performing the Cox proportional hazards' model's C-index of 0.79. Both models were trained from a combination of radiomic and clinical patient data. [7]

Senders et al. compared numerous parametric and nonparametric models in their dis-

criminatory ability estimating overall survival of patients with glioblastoma multiforme. Employing an integrated concordance index on the 2005 to 2015 Surveillance Epidemiology and End Results (SEER) database, their log-logistic distribution fit on an accelerated failure time model yielded the highest C-index of 0.70, while the cox proportional hazards model and random survival forest models produced C-indices of 0.69 and 0.68 respectively. [8]

Marko et al. accessed data on glioblastoma patients from the MD Anderson Cancer Center Neuro-Oncology database, and performed log-logistic regression under the influence of the covariates of age, Karnofsky performance index, extent of tumor removal/resection (before therapy), and adjuvant chemoradiotherapy. The model was validated using 20% bootstrapping and cross-validation in which results from a single test set were aggregated, yielding an average pseudo-R^2 value of 0.305 and an average C-index of 0.69. The researchers claim that their log-logistic model had superior predictive capabilities compared to the Cox proportional hazards model, citing the presence of an explicit hazard function in the former as part of the reason. [9]

Weltman et al. analyzed the significance of factors such as age, Karnofsky Performance Status (KPS), and number of lesions, among others, on the survival of 65 patients with metastases (secondary brain tumor) who underwent stereotactic radiosurgery. They employed a univariate and multivariate (multiple groups, when applicable) log-rank test at $\alpha = 0.05$, and identified extracranial disease status and KPS as the important (significant) factors for prognosis. [10]

Alexopoulos et al. reviewed 1975-2018 data and conducted a population-based study of glioblastoma multiforme patients. Kaplan-Meier survival estimates encouraged the fitting of an exponential model, alongside other parametric models such as Weibull, Gompertz, and a generalized gamma. AIC and likelihood tests were employed, and the exponential model was found to be inferior to the lognormal model. The lognormal accelerated failure time model found age and tumor location, among miscellaneous demographic factors, to be significant patient survival predictors. [11]

Tunthanathip et al. performed a retrospective cohort study in which patients admitted to a tertiary center for glioblastoma in Southern Thailand between 2007 and 2021 were selected, with predictors identified backward and stepwise to be used in a multivariate Cox, Gompertz and Weibull regression. Random survival forest, among other nonparametric models, was investigated to complement the prognosis analyses with machine learning. Of the models tested, Cox and Weibull tied for a median C-index accuracy of 0.648, while

Random Forest tied for the lowest accuracy with a C-index of only 0.640. [12]
A comparison of the above literature with our results is presented in section 4.

2 Survival Analysis: Theory

2.1 Kaplan-Meier Estimator and Log-rank Test

Let T be a nonnegative random variable denoting the time to the event of interest (e.g., death, relapse). The *survival function* is defined as the probability of surviving beyond time t, that is,

$$S(t) = \mathbb{P}(T > t).$$

Survival analysis refers to the collection of statistical methods used to analyze and model survival data through the survival function [12]. A key feature that distinguishes survival analysis from standard statistical techniques is the presence of *censoring*, which arises when the event of interest is not observed for some individuals within the study period. In the most common case of right censoring, we only know that an individual has survived up to a certain time without experiencing the event, but the exact event time remains unknown.

The Kaplan–Meier (KM) estimator is a nonparametric estimator of the survival function in the presence of censoring. If the ordered event times are $t_1 < t_2 < \cdots < t_k$, with d_j events and n_j individuals at risk at time t_j, the estimator is

$$\widehat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right).$$

A graph of this estimated piecewise constant function is called the Kaplan–Meier curve.

Further, to compare survival between two groups, we test $H_0 : S_1(t) = S_2(t)$ for all $t \geq 0$ against $H_1 : S_1(t) \neq S_2(t)$ for some t. The log-rank test is used. Let d_{1j} be the observed number of events in group 1 at time t_j, with d_j total events, and n_{1j} and n_j the numbers at risk in group 1 and overall, respectively. The expected number of events in group 1 at time t_j under the null hypothesis is

$$E_{1j} = \frac{n_{1j} d_j}{n_j}.$$

The log-rank test statistic is

$$Z \;=\; \frac{\sum_{j=1}^{k}(d_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{k} \frac{n_{1j}(n_j - n_{1j})d_j(n_j - d_j)}{n_j^2(n_j - 1)}}},$$

which is approximately standard normal under the null hypothesis.

The log-rank test can be extended to incorporate more than two groups. Suppose there are $G > 2$ groups, and we would like to test $H_0 :\ S_1(t) = S_2(t) = \cdots = S_G(t)$ for all $t \geq 0$ against $H_1 :\ S_i(t) \neq S_j(t)$ for some i, j and $t \geq 0$. The test statistic has the form

$$\chi^2 \;=\; (O - E)' V^{-1} (O - E),$$

where $O = \sum_{i=1}^{k} (d_{1i}, d_{2i}, \ldots, d_{Gi})'$ is the sum of column-vectors of observed events in groups 1 through G over times $t_1 < t_2 < \cdots < t_k$, and $E = \sum_{i=1}^{k} (E_{1i}, E_{2i}, \ldots, E_{Gi})'$ is the sum of column-vectors of the corresponding expected values. The matrix V is a variance-covariance matrix with the diagonal entries

$$V_{jj} \;=\; \sum_{i=1}^{k} \frac{n_{1i}(n_i - n_{1i})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \;\; j \;=\; 1 \ldots, G,$$

and with off-diagonal entries

$$V_{jj'} \;=\; -\sum_{i=1}^{k} \frac{n_{ji} n_{j'i} d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \;\; j, j' \;=\; 1, \ldots, G, j \neq j'.$$

Under the null hypothesis, the test statistic has a chi-squared distribution with $G - 1$ degrees of freedom.

2.2 Random Censoring Model

Many models are used to describe the distribution of event times T. Parameter estimation is typically carried out via maximum likelihood. However, specifying the likelihood is not straightforward in the presence of censoring. A standard approach is to adopt the random censoring model. Here we describe how it works. We observe n pairs (t_i, δ_i), where $t_i = \min(T_i, C_i)$ is the recorded time, T_i the survival time with density $f(t)$ and cumulative distribution function (cdf) $F(t)$, C_i the censoring time with density $g(t)$ and cdf $G(t)$, and

$\delta_i = 1$ if $T_i \leq C_i$ (event observed), $\delta_i = 0$ otherwise. Assuming T_i and C_i are independent, the joint density is

$$f(t, \delta) = \big[f(t)(1 - G(t))\big]^{\delta} \big[(1 - F(t))g(t)\big]^{1-\delta}.$$

Hence, the partial likelihood function that depends only on $f(t)$ and $F(t)$ takes the form

$$L_p = \prod_{i=1}^{n} f(t_i)^{\delta_i} (1 - F(t_i))^{1-\delta_i}.$$

Maximum likelihood estimators are obtained by solving the score equations, and the survival function is estimated by $\hat{S}(t) = 1 - \hat{F}(t)$ with fitted parameters.

2.3 Cox Proportional Hazards Model

The Cox proportional hazards model, formulated in terms of the survival function, is expressed as

$$S(t \mid \mathbf{X}) = \big[S_0(t)\big]^{\exp(\boldsymbol{\beta}'\mathbf{X})},$$

where $\mathbf{X}$ is a vector of predictors, $\boldsymbol{\beta}$ is a vector of corresponding regression coefficients, and $S_0(t)$ is the baseline survival function, defined as the survival function of an often hypothetical "baseline" individual for whom all predictors have zero values.
The parameters $\boldsymbol{\beta}$ are estimated via the *partial likelihood* in the random censoring model, which is further reduced to avoid specifying the baseline survival function $S_0(t)$. This partial likelihood function is

$$L_p(\boldsymbol{\beta}) =, \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{X}_j)} \right]^{\delta_i}$$

with $R(t_i)$ denoting the risk set at time t_i. Maximizing this function gives the estimator $\hat{\boldsymbol{\beta}}$.

The baseline survival function $\hat{S}_0(t)$ is obtained via the maximum likelihood approach. Denote by $\pi_i = \mathbb{P}(T > t_i | \, T > t_{i-1})$, the conditional survival probability at time t_i for a baseline subject. The conditional survival probability of a subject with predictors $\mathbf{X}$ can be obtained by raising π_i to the power $\exp(\boldsymbol{\beta}'\mathbf{X})$. Then the likelihood function takes the form:

$$L(\pi_1, \ldots, \pi_k) = \prod_{i=1}^{k} \prod_{j \in D(t_i)} (1 - \pi_i^{\exp(\boldsymbol{\beta}'\mathbf{X}_j)}) \prod_{j \in R(t_i) \setminus D(t_i)} \pi_i^{\exp(\boldsymbol{\beta}'\mathbf{X}_j)}.$$

Here $R(t_i)$ and $D(t_i)$ denote the risk set and event set at time t_i, respectively. This function is maximized with respect to π's after the partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is plugged in. Finally, the survival curve for a subject with covariates $\mathbf{X}$ is estimated as

$$\hat{S}(t \mid \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{X})}.$$

2.4 Exponential Survival Model

In the exponential model, the survival time T has an exponential distribution with density $f(t) = \lambda \exp\{-\lambda t\}$ and cdf $F(t) = 1 - \exp\{-\lambda t\}$, $t \geq 0$, $\lambda > 0$, where $\lambda = \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\}$. In the random censoring model, the partial likelihood function is

$$L_p(\beta_0, \ldots, \beta_k) \;=\; \prod_{i=1}^{n} \left[\lambda_i \exp\{-\lambda t_i\}\right]^{\delta_i} \left[\exp\{-\lambda_i t_i\}\right]^{1-\delta_i}$$

with $\lambda_i = \exp\{-(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})\}$. The model parameters $\beta_0, \ldots, \beta_k$ are estimated by maximizing this function. The estimator of the survival function is

$$\hat{S}(t) \;=\; \exp\{-\hat{\lambda} t\},\; t \geq 0,$$

where $\hat{\lambda} = \exp\{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)\}$.

2.5 Log-Logistic Survival Model

The log-logistic survival model assumes that survival time T follows a log-logistic distribution. The density of this distribution is

$$f(t) \;=\; \frac{(\gamma/\lambda)\,(t/\lambda)^{\gamma-1}}{\left(1 + (t/\lambda)^{\gamma}\right)^2}, \;\; t \geq 0,$$

and the cumulative distribution function is given by

$$F(t) \;=\; \frac{(t/\lambda)^{\gamma}}{1 + (t/\lambda)^{\gamma}}, \;\; t \geq 0,$$

where the scale parameter λ depends on predictors $x_1, \ldots, x_k$ as follows:

$$\lambda = \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\},$$

and $\gamma > 0$ is a shape parameter.

The parameters of the model $\beta_0, \ldots, \beta_k$ and γ are estimated by maximizing the likelihood function. In the presence of random censoring, the partial likelihood function is of the form

$$L(\beta_0, \ldots, \beta_k, \gamma) = \prod_{i=1}^{n} \Big[\frac{(\gamma/\lambda_i)\,(t_i/\lambda_i)^{\gamma-1}}{\Big(1 + (t_i/\lambda_i)^{\gamma}\Big)^2} \Big]^{\delta_i} \Big[1 \; - \; \frac{(t_i/\lambda_i)^{\gamma}}{1 + (t_i/\lambda_i)^{\gamma}} \Big]^{1-\delta_i},$$

with $\lambda_i = \exp\{-(\beta_0 + \beta_1\, x_{1i} + \cdots + \beta_k\, x_{ki})\}$. The fitted survival function is

$$\hat{S}(t) \; = \; \frac{1}{1 + \big(t\, \exp(\hat{\beta}_0 + \hat{\beta}_1\, x_1 \; + \; \ldots \; + \; \hat{\beta}_k\, x_k)\big)^{\hat{\gamma}}}.$$

2.6 Random Survival Forest

The random survival forest is a tree-based ensemble method for survival analysis. Each tree is grown using a bootstrap sample of the data, while a random subset of predictors is considered at each split. Splits are chosen to maximize survival differences between child nodes, using the log-rank statistic. Terminal nodes contain subsets of individuals with similar survival patterns, and the survival function for a node is typically estimated by the Kaplan–Meier estimator. The survival estimate for an individual is obtained by averaging the node-specific survival functions across all trees in the forest.

2.7 Goodness of Model Fit Assessment

Model assessment in survival analysis is often based on information criteria and discrimination measures. For Cox proportional hazards and parametric survival models, the Akaike Information Criterion (AIC) is widely used, defined as

$$\text{AIC} \; = \; -2 \ln L(\hat{\theta}) \; + \; 2p$$

where $\ln L(\hat{\theta})$ is the maximized log-likelihood function and p is the number of parameters. Smaller AIC values indicate better trade-off between model fit and complexity; thus, a model with the smallest AIC value has the best fit.

In addition, predictive performance is commonly evaluated using the concordance index (C-index), which measures the proportion of correctly ordered pairs of survival times based on the predicted risk scores. A C-index of 0.5 corresponds to random prediction, while a value close to 1 indicates strong discriminative ability.

3 Data Analysis Results

The Kaplan–Meier estimator was used to obtain a nonparametric estimate of the survival function for our data set, with the resulting curve shown in Figure 1 [13].

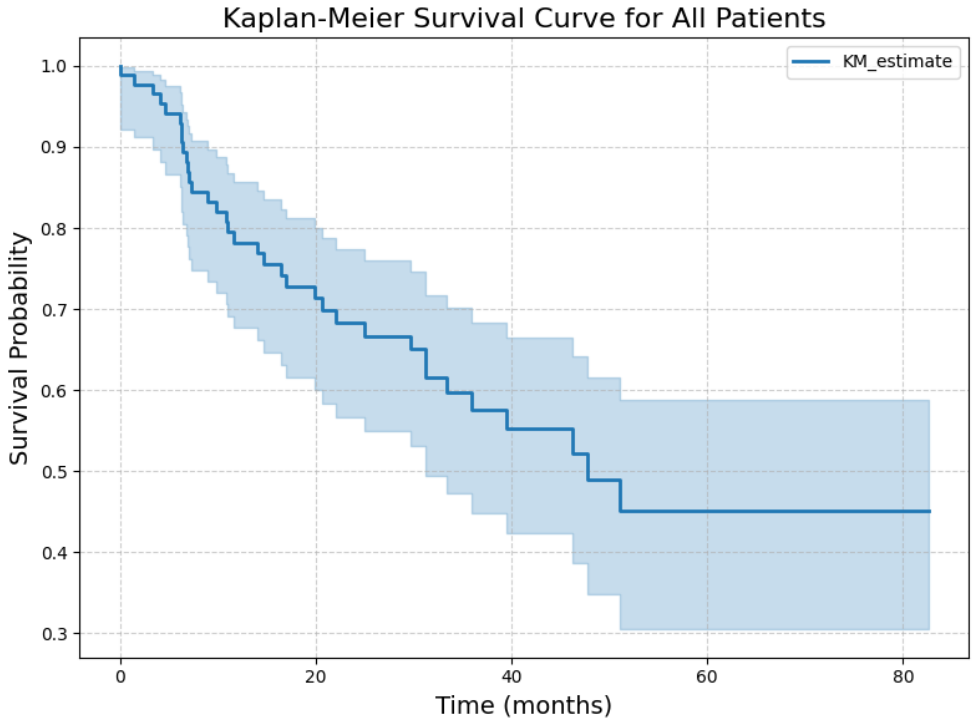

Figure 1. Kaplan-Meier curve for survival function for brain tumor data.

The Kaplan-Meier curve displays an exponentially decaying function, which suggests that the exponential parametric model provides a reasonable fit to the data. We fit both the Cox proportional hazards model and the exponential model and compare the results. The fitted Cox model is of the form:

$$\hat{S}(t) = [\hat{S}_0(t)]^{\mathrm{HR}}$$

where

$$\mathrm{HR} = \exp\{0.26(\mathrm{S}) + 0.58(\mathrm{D}) + 1.25(\mathrm{L}) - 0.05(\mathrm{K}) + 0.01(\mathrm{G}) + 0.33(\mathrm{Sm})\}$$

and each $\exp\{B_n\}$ in the product denotes the hazard ratio for the nth covariate. Note that S, D, L, K, G, and Sm stand for Gender, Diagnosis, Location, Karnofsky Index,

Gross Tumor Volume, and Stereotactic methods respectively. The estimate of the baseline survival function is given in tabular form in Table 1 below.

Time (months)	**Baseline Survival Probability**
0.065	0.994
1.180	0.994
1.410	0.987
1.541	0.987
2.033	0.987
3.377	0.980
⋮	⋮
57.639	0.456
65.016	0.456
67.377	0.456
73.738	0.456
78.754	0.456
82.557	0.456

Table 1. Estimate of baseline survival function in the Cox model for brain tumor data.

The fitted exponential model is given as

$$\hat{S}(t) = \exp\left\{ -\exp\Big(-\big(2.64 - 0.16S - 0.57D - 1.18L + 0.04K - 0.01G - 0.27Sm\big)\Big)\, t \right\}.$$

Additionally, the fitted log-logistic model is given as

$$\hat{S}(t) \;=\; \frac{1}{1 + \Big(t\, \exp\big(-5.89 + 0.35\, S + 0.58\, D + 1.34\, L + 0.12\, Sm - 0.39\, K + 0.24\, G\big)\Big)^{1.35}}.$$

The AIC for the Cox model is 257.35, whereas that for the exponential model is 351.35 and that for the log-logistic model is 349.96; thus, the Cox model has the best fit.

To complement these analyses, a random survival forest model was fitted, and its predictive performance was evaluated. Table 2 below reports concordance indices (C-indices).

Model	**C-index**
Cox Proportional Hazards	0.777
Exponential	0.222
Log-logistic	0.784
Random Survival Forest	0.866

Table 2. C-indices for the Cox model, the exponential model, and random survival forest for brain tumor data.

To further explore the results, Kaplan–Meier curves were plotted for significant categorical predictors, stratified by their levels, and compared using the log-rank tests. A p-value below 0.05 indicates a statistically significant difference between survival curves across levels of a predictor, which should also be evident from their visual separation. The corresponding p-values are summarized in Table 3 below. The graphs are presented in Figures 2 and 3 that follow.

Covariate	p**-value**
Gender	0.2301
Location	0.1103
Stereotactic methods	0.0325
Diagnosis	1.61e-06

Table 3. Results of the log-rank test with 2 groups (Gender, Location, and Stereotactic methods) and 4 groups (Diagnosis).

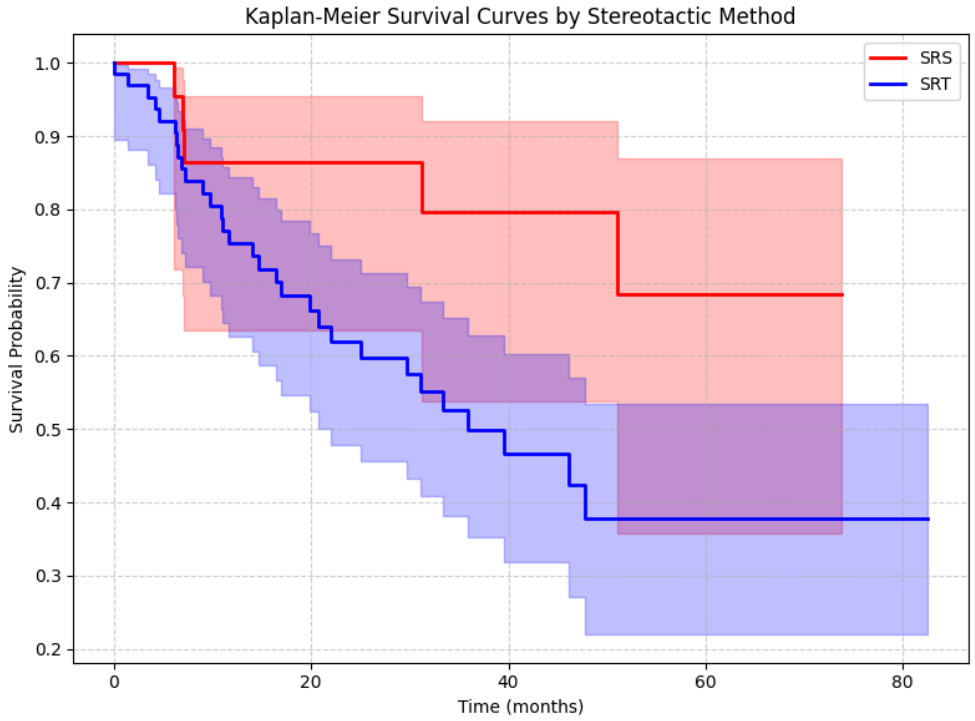

Figure 2: Kaplan-Meier Curves (stratified by treatment).

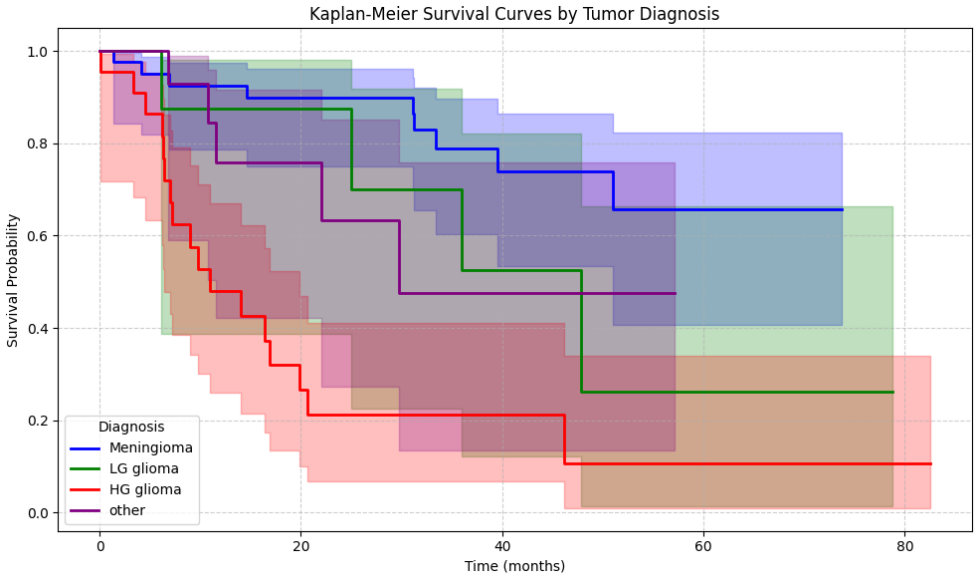

Figure 3: Kaplan-Meier Curves (stratified by diagnosis).

4 Discussion of Results

Our findings largely agreed with prior literature, with minor differences.

The literature generally agrees with our conclusion that nonparametric, machine learning alternatives similar to the random survival forest out-perform parametric and semi-parametric alternatives such as the Cox, exponential, and log-logistic models. Within that hierarchy, the literature also supports the superior discrimination capabilities of a log-logistic model compared to a Cox model, and is sparse on the exponential model due to its restrictive features that lower regression accuracy compared to other models.

Many papers disagreed with each other on the relative feature importance of several variables that were present in our own study. Specifically, gender received varying results, with papers both supporting and opposing our lack of evidence for its significance. Most papers, meanwhile, agreed that the tumor diagnosis type is a significant covariate.

The C-indices we found deviated from most literature. Our exponential model was worse than a random model, while our random survival forest outputted a value higher than expected. This is largely explained by the shortage of a publicly available data to support our research, resulting in a relatively low sample size compared to the number of censored patients and predictors involved. To fortify our findings, a larger sample size from multiple institutions is required, narrowing the C-index values and improving model reliability.

We discovered that both a random survival forest and a Cox proportional hazards model are moderately strong predictors of patient survival for our dataset. From the former, we discovered that the diagnosis of a tumor in a patient is most critical for determining their survival function, an idea that is both intuitively and theoretically well-supported. From the latter, we also found a hazard ratio that could confidently be placed above 1.00000, indicating a significant difference in relative risk of a patient depending on the type of tumor inflicted.

Aside from computing hazard ratios, we investigated the hypothesis that survival function differs within each group of our covariates with the log-rank test. From this, we determined that there is significant evidence at a 5% significance level that the type of radiation therapy performed on a patient influences their survival function. Specifically, Stereotactic

RadioSurgery, or SRS, which delivers a singular large dose of radiation to the tumor location, proved superior in its risk mitigation capabilities. Lastly, as previously signaled by both the ensemble and cox models, the log-rank test for 4 groups confirmed that at least one tumor type differs significantly from the others in its survival rate.

These results are significant in the fields of radiation oncology and radiology, as they indicate greater potential for SRS technology in advancing cancer treatment. They also demonstrate how, while important, other variables influencing a patient's health are secondary to their exact condition, helping regularize models predicting patient timelines.

Overall, we hope the results provide insight into variables influencing patient survival for brain tumor victims, and wish to extend and strengthen the models to a greater number of factors, machine learning techniques, and datasets.

Supplemental Materials

Supplemental materials, including the dataset and Python code used for analysis, are available at the following GitHub repository. To access the data file directly on Kaggle, use this link. To run the code directly, use this Google Colab link. Readers are encouraged to reproduce the results and explore the data further.

Acknowledgments

I would like to sincerely thank Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach, for her invaluable guidance, support, and encouragement throughout this research. I am also grateful to Mr. Eric Shulman, mathematics teacher at University High School, for his instruction in calculus and matrix theory, which laid a foundational knowledge essential to this work. Finally, I deeply appreciate my family for their support and patience, which made this journey possible.

References

[1] Ostrom, Q. T., Price, M., Neff, C., Cioffi, G., Waite, K. A., Kruchko, C., and J. S. Barnholtz-Sloan. (2024). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2017-2021. *Neuro-Oncology*, 26 (Supplement 6): vi1-vi85. doi: 10.1093/neuonc/noae145.

[2] Mahmoudi, K., Kihira, S., Nael, K., Kim, D., Tavakkol, E., Bauer, A., Tsankova, N., Khan, F., Hormigo, A., and V. Yedavalli. (2024). Multiparametric Radiogenomic Model to Predict Survival in Patients with Glioblastoma. *Cancers*, 16(3): 589. doi:10.3390/cancers16030589.

[3] Awuah, W. A., Ben-Jaafar, A., Roy, S., Nkrumah-Boateng P. A., Tan, J.K., Abdul-Rahman, T., and O. Atallah. (2025). Predicting survival in malignant glioma using artificial intelligence. *European Journal of Medical Research*. 30(1). doi:10.1186/s40001-025-02339-3.

[4] Babaei Rikan, S., Sorayaie Azar, A., Naemi, A., Bagherzadeh Mohasefi, J., Pirnejad, H., and U. K. Wiil. (2024). Survival prediction of glioblastoma patients using modern deep learning and machine learning techniques. *Scientific reports*. 14(1). doi:10.1038/s41598-024-53006-2.

[5] Selingerová, I., Horová, I., Doleželová, H., Zelinka, J., and S. Katina. (2016). Survival of Patients with Primary Brain Tumors: Comparison of Two Statistical Approaches. *PLOS ONE*. 11(2):e0148733. doi:10.1371/journal.pone.0148733.

[6] Di Noia, C., Grist, J. T., Riemer, F., Lyasheva, M., Fabozzi, M., Castelli, M., Lodi, R., Tonon, C., Rundo, L., and F. Zaccagna. (2022). Predicting Survival in Patients with Brain Tumors: Current State-of-the-Art of AI Methods Applied to MRI. *Diagnostics*. 12(9): 2125. doi:10.3390/diagnostics12092125.

[7] Senders, J. T., Taphoorn, M. J. B., Arnaout, O., Staples, P., Mehrtash, A., Cote, D. J., Reardon, D. A., Gormley, W. B., Smith, T. R., Broekman, M. L., and O. Arnaout. (2020). An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery*. 86(2): E184-E192. doi:10.1093/neuros/nyz403.

[8] Marko, N. F., Suki, D., Sawaya, R. E., Schroeder, J. L., Lang, F. F., and R. J. Weil. (2014). Extent of Resection of Glioblastoma Revisited: Personalized Survival Modeling Facilitates More Accurate Survival Prediction and Supports a Maximum-Safe-Resection Approach to Surgery. *Journal of Clinical Oncology*. 32(8): 774-782. doi:10.1200/jco.2013.51.8886.

[9] Weltman, E., Salvajoli, J. V., Brandt, R. A., de Morais Hanriot, R., Prisco, F. E., Cruz, J. C., de Oliveira Borges, S. R., and D. Ballas Wajsbrot. (2000). Radiosurgery for brain metastases: a score index for predicting prognosis. *International Journal of Radiation Oncology*, 46(5): 1155–1161. https://doi.org/10.1016/s0360-3016(99)00549-0.

[10] Alexopoulos, G., Zhang, J., Karampelas, I., Patel, M., Kemp, J., Coppens, J., Mattei, T. A., and P. Mercier. (2022). Long-Term Time Series Forecasting and Updates on Survival Analysis of Glioblastoma Multiforme: A 1975–2018 Population-Based Study. *Neuroepidemiology*, 56(2): 75–89. https://doi.org/10.1159/000522611.

[11] Tunthanathip, T., and T. Oearsakul. (2023). Comparison of predicted survival curves and personalized prognosis among Cox regression and machine learning approaches in glioblastoma. *Journal of Medical Artificial Intelligence*, 6: 10. https://doi.org/10.21037/jmai-22-98.

[12] Hosmer, D., Lemeshow, S., and S. May. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, 2nd edition.

[13] Nag, A. (2022). *Survival Analysis with Python*, Routledge, 1st edition.

Appendix

Data Frequency Charts

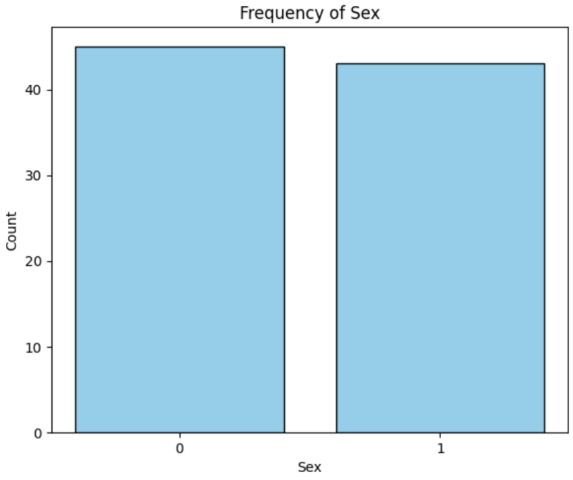

Figure 1: Gender

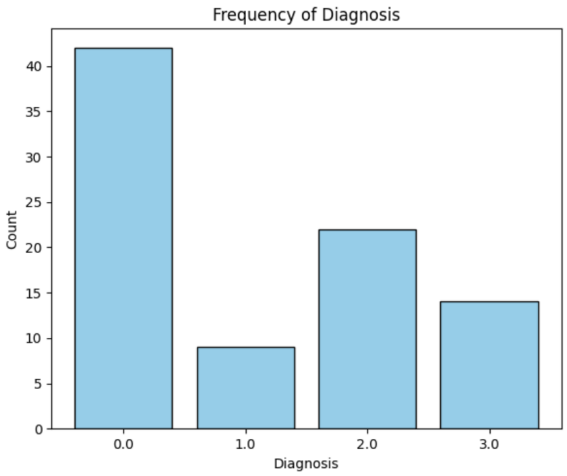

Figure 2: Diagnosis

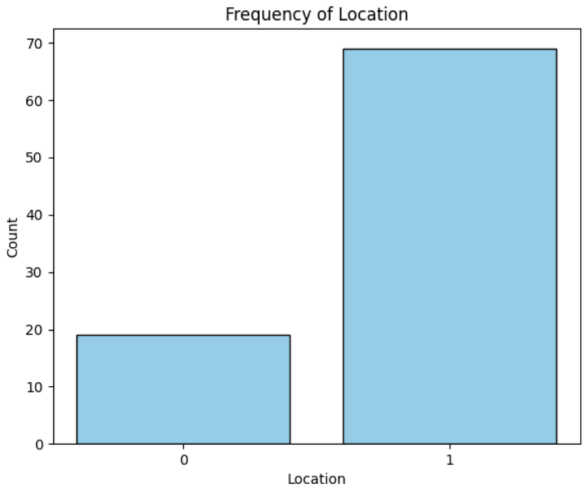

Figure 3: Location

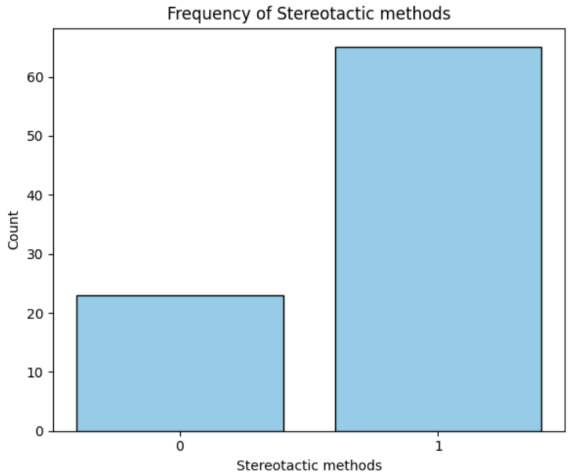

Figure 4: Stereotactic Methods

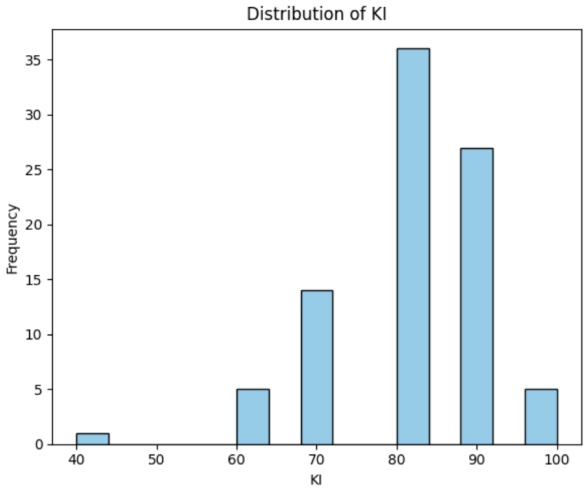

Figure 5: Karnofsky Index

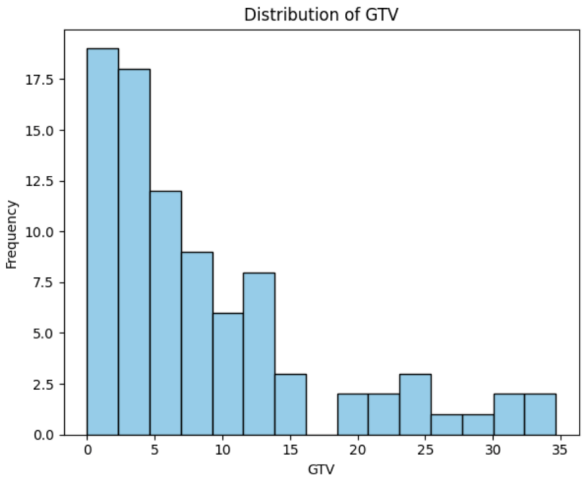

Figure 6: Gross Tumor Volume