P8130 Final Project Report

Yuechu Hu, Leyang Rui, Yifei Yu, Jinghan Zhao

Abstract

This study analyzes a breast cancer survival dataset of 4024 observations to develop predictive models for patient survival. Using logistic regression, we identified several significant predictors of survival times through model selection and validation. Survival analysis was applied to explore the changes in patients' survival status over time and its multiple risk factors. Results indicate that higher odds of death are associated with older age, greater regional node positivity, undifferentiated tumors, and advanced cancer stages, while marital status and race interact significantly. The findings highlight key risk factors while recognizing the study's observational limitations. Future research could focus on interventional research addressing these risk factors to optimize patient outcomes.

Introduction

The data we used for this analysis originates from a breast cancer survival dataset collected from a prospective study. The dataset contains 14 important predictor variables, including patients' age, race, marital status, tumor size, cancer stages (T Stage, N Stage, A stage, and 6th Stage), differentiated grade, estrogen and progesterone status, and regional node involvement. The dataset also records patients' survival times in months and their final survival status (dead or alive), which are the outcome variables of interest. The description of variables are shown in Table 1, 2, and 3. Our objective is to develop models that predict the risk of death among breast cancer patients using these features. Specifically, we aim to identify which variables significantly impact patients' survival outcomes, determine potential interactions among the variables, and assess the performance of the models. Additionally, we will investigate fairness in the model predictions to ensure equitable accuracy between the majority race group and the minorities.

Methods

The dataset comprises 4024 observations and 16 variables, with no missing values. Initially, we cleaned and tidied the data using data-wrangling techniques such as mutating. Then, graphical tools such as histograms and box plots helped us to figure out the distributions of the variables and check for potential outliers or influential points. We also examined the pairwise relationships between variables through scatter plots. We tried logarithmic transformation for variables with right-skewed distributions, but ultimately used the original data because the impact of the transformation was not obvious.

Since the response variable, status, is a binary categorical variable that indicates the survival result of the patients, we chose to build a logistic regression model to predict its estimated probability. Since breast cancer grade (i.e. variable *grade*) is exactly determined by the degrees of the variable *differentiate*, we removed *grade* from the model. In addition, the level IIIC of variable x6th_stage also correlates with other variables in the dataset. However, the AJCC system is a complex criterion based on multiple aspects, so we cannot simply remove this variable as part of its information may still be useful.

To find the best model for predicting the patients' survival status, we started by using all 3 automated procedures–forward selection, backward elimination, and stepwise regression to choose models with statistically significant predictors respectively. Next, the criterion-based procedure–AIC and BIC are applied to compare the values among the three automatically generated models and the full models to choose the best final model. For model diagnosis, the variance inflation factor (VIF) and the generalized variance inflation factor (GVIF) were used to detect multicollinearity among numerical and categorical variables. Since the residual versus fitted values plot always shows a pattern in logistic regression because of the binary response variable, we randomized the quantile residuals to examine the assumption regarding residuals. Additionally, a residual versus leverage plot is used to explore potential outliers. Finally, we employed a 10-fold cross-validation and evaluated the goodness-of-fit of the model by log loss and AUC. Similarly, to test the performance of the model among the majority white and other races, we did another stratified cross-validation and found potential space for improvement.

We also performed the survival analysis with patients' status and survival months as response variables. The Kaplan-Meier survival time curve represents the survival rate over time, and the log-rank test tells us whether there is a difference in survival times between patients whose cells have different degrees of differentiation. Since both methods are limited in that only one variable can be tested at a time, we computed Cox proportional hazard models to adjust for multiple risk factors simultaneously.

Results

As shown in Figure 1, most patients are between 40 and 70 years old, and the most frequent survival times are larger than 45 months. The number of examined regional nodes for most patients is smaller than 30, and the majority of patients have nearly 12 examined regional nodes. It is worth noting that the distributions of both variables *regional_node_positive* and *tumor_size* are significantly skewed to the right. However, since logistic regression does not require all predictors to be normally distributed and we reexamined their distributions after transformation and confirmed that skewness was not significantly reduced, we would not use the transformation in further analysis. Over 2500 subjects only have 1 or 2 positive regional nodes, the most frequent number of positive regional nodes. Most tumor sizes are smaller than 50 mm, and we found that the most frequent size is around 19 mm, followed by around 14 mm. Figure 2 distributed the survival time by the status, the "Dead" group is concentrated in the shorter survival months, while the alive group is predominant in longer survival months, particularly beyond 60 months. According to Figure 3, as the t_stage changes from stage 1 to stage 4, the size of the tumor also increases. We also noticed some potential outliers both in the T1 stage and T3 stage. Looking at Figure 4, the survival time is longer in the Regional stage, and the "Alive" group shows higher survival times across both stages. In Figure 5, the undifferentiated group has larger tumor sizes compared to the other categories, while the well, moderately, and poorly differentiated groups all display similar distributions with the majority of tumor sizes being small except for numerous high-value outliers. Figure 6 highlights the differences in tumor size distribution and trends with age between individuals who are alive and those who are deceased. While the "Alive" group shows no significant relationship between age and tumor size, the "Dead" group exhibits a pattern where larger tumors are associated with younger ages. Finally, according to Figure 7, as it changes from well-differentiated to undifferentiated, the negative correlation between the number of positive regional nodes and the patients' survival months strengthens.

After comparison, the stepwise regression model was chosen as the final model since it has the smallest AIC and BIC value, shown in Table 4, and the results of the final model are represented by Table 5. All but *marital status* are highly significant variables with very small p-values. For example, people with undifferentiated tumors have 6.46 times the odds of death compared to those with well-differentiated tumors. With respect to model diagnostic, as Table 6 reveals, all the adjusted GVIFs (a measure corrected for the degree of freedom and provides a scale similar to VIF for continuous variables) are less than or not much different from 2, implying the absence of multicollinearity. The randomized quantile residuals versus fitted values plot (Figure 8) displays a pattern of randomized residuals equally distributed around the 0.5 line, satisfying the assumptions of linearity and residual equal variance. Moreover, the residual versus leverage plot (Figure 9) indicates that observations 3527, 1561, and 3074 may be potential outliers, but they are not necessarily influential and we will keep them for future attention. The results of the 10-fold cross-validation, displayed in Table 7, show the goodness of fit by log loss and area under curve (AUC). The mean of log loss is 0.372, and the mean of AUC is 0.742. The prediction performance is better in the majority race group "White" than the minority "Black" and "Other" as shown in the results of stratified validation by levels of race (Table 8), where lower log loss and larger AUC indicate better test performance. Since the distribution of survival months is different between races with different marital statuses (Figure 10), we added an interaction term *marital_status*race* in the model, which further reduced the gap between race groups and improved the performance of our model (Table 9).

For survival analysis, the Kaplan-Meier curve (Figure 11) shows the overall survival rate over months. We found out significant difference in survival between patients whose cancer cells have different degrees of differentiation in the log-rank test (Figure 12). To further discuss the multiple risk factors to survival time, we performed the Cox proportional hazard model. The assumption of the Cox model was tested based on the scaled Schoenfeld residuals, that is, the survival curves of the two different strata of the risk factor must have a proportional hazard function that varies over time. Table 10 shows that *a_stage*, *estrogen_status,* and *progesterone_status* are not constant over time, so we removed these variables in further cox model analysis. The forest plot (Figure 13) shows the results of the cox model, that is, the relationship between multiple variables and the probability of death. A hazard ratio greater than 1 indicates an increased probability of death, and a hazard ratio less than 1 indicates a decrease. The smaller the p-value, the greater the weight of evidence that there is a difference between the groups.

Conclusion/Discussion

Through the visualization of the variables, survival months were significantly higher for the "Alive" group compared to the "Dead" group and were longer in the Regional stage compared to the Distant stage. Additionally, as the T stage progresses from stage 1 to stage 4, tumor size consistently increases. Meanwhile, the undifferentiated group has larger tumor sizes compared to the other categories. As tumor differentiation shifts from well-differentiated to undifferentiated, the negative correlation between the number of positive regional nodes and patients' survival months becomes stronger. While the "Alive" group shows no significant relationship between age and tumor size, the "Dead" group exhibits a pattern where larger tumors are associated with younger ages.

The overall survival rate in this dataset decreased over time and finally remained above 75%. According to the results of our final model (Table 5), the odds of death increase as age and number of positive regional nodes increase, and decrease as the number of examined regional nodes increases. The odds of death are higher for black people, separated couples, and widowed individuals, with higher T stages, and higher N stages. In addition, the more undifferentiated the tumor is, the higher the odds of death is. The results of survival analysis show that the hazard of death is significantly higher in black people, separated couples, patients with higher N stages, and lower differentiated cancer cells.

Although many variables are significant in predicting the odds of death, it is important to note that this dataset is from an observational study, so the conclusion is limited to correlations between these predictors and survival status and no causation can be ascertained. Further research could focus on prescribing drugs to patients with different stages and types of tumors and reassessing their survival status afterwards to effectively find the most appropriate drugs for patients with different conditions.

Group members' Contributions

Leyang and Jinghan focused on statistical methods, building models to extract meaningful insights. Yuechu and Yifei concentrated on data description and visualization, presenting information through clear visuals. All four of us contributed to the writing of the report, integrating our individual efforts into a cohesive final report that reflects both analytical depth and visual clarity.

Appendix

Descriptive Tables

Table 1: Data Dictionary

Variable	Description
age	The age of the patient (in years)
race	The race of the patient, categorized as Black, White or Other
marital_status	The marital status of the patient, categorized as Divorced, Married, Separated, Single, or Widowed
t_stage	Adjusted AJCC 6th T, categorized as T1, T2, T3, or T4
n_stage	Adjusted AJCC 6th N, categorized as N1, N2, or N3
x6th_stage	Breast Adjusted AJCC 6th Stage, categorized as IIA, IIB, IIIA, IIIB, or IIIC
differentiate	Tumor differentiation grade, categorized as Well differentiated, Moderately differentiated, Poorly differentiated, or Undifferentiated
grade	Tumor differentiation grade, categorized as 1, 2, 3, or anaplastic; Grade IV
a_stage	Categorized as Regional (a neoplasm that has extended) or Distant (a neoplasm that has spread to parts of the body remote from)
tumor_size	The size of tumor (in millimeters)
estrogen_status	The status of the patient's estrogen, categorized as Positive or Negative
progesterone_status	The status of the patient's progesterone, categorized as Positive or Negative
regional_node_examined	The number of examined regional nodes
regional_node_positive	The number of positive regional nodes
survival_month	The time of a patient with breast cancer is expected to live after their diagnosis (in months)
status	The status of the patient, categorized as Alive or Dead

Table 2: Summary Statistics for Numeric Variables

Variable Name	Mean	SD	Median	IQR
Age	53.972167	8.963134	54	14
Tumor Size	30.473658	21.119696	25	22
Regional Nodes Examined	14.357107	8.099675	14	10
Regional Nodes Positive	4.158052	5.109331	2	4
Survival Months	71.297962	22.921429	73	34

Table 3: Summary Statistics for Categorical Variables

Variable Name	Level	Count	Proportion
Race	Black	291	0.0723
Race	White	3413	0.8482
Race	Other	320	0.0795
Marital Status	Divorced	486	0.1208
Marital Status	Married	2643	0.6568
Marital Status	Separated	45	0.0112
Marital Status	Single	615	0.1528
Marital Status	Widowed	235	0.0584
T Stage	T1	1603	0.3984
T Stage	T2	1786	0.4438
T Stage	T3	533	0.1325
T Stage	T4	102	0.0253
N Stage	N1	2732	0.6789
N Stage	N2	820	0.2038
N Stage	N3	472	0.1173
6th Stage	IIA	1305	0.3243
6th Stage	IIB	1130	0.2808
6th Stage	IIIA	1050	0.2609
6th Stage	IIIB	67	0.0167

Variable Name	Level	Count	Proportion
6th Stage	IIIC	472	0.1173
Differentiate	Well	543	0.1349
Differentiate	Moderate	2351	0.5842
Differentiate	Poor	1111	0.2761
Differentiate	Undifferentiated	19	0.0047
Grade	1	543	0.1349
Grade	2	2351	0.5842
Grade	3	1111	0.2761
Grade	4	0	0.0000
A Stage	Distant	92	0.0229
A Stage	Regional	3932	0.9771
Estrogen Status	Positive	3755	0.9332
Estrogen Status	Negative	269	0.0668
Progesterone Status	Positive	3326	0.8265
Progesterone Status	Negative	698	0.1735
Status	Alive	3408	0.8469
Status	Dead	616	0.1531

Exploratory Analysis

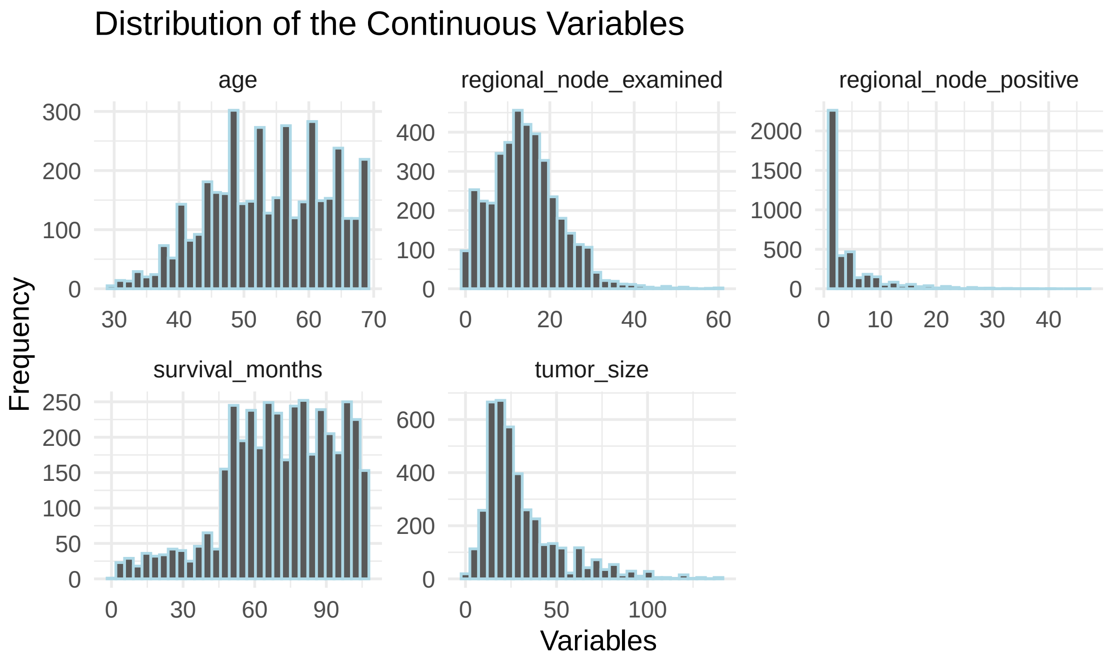

Figure 1: Distribution of the Continuous Variables

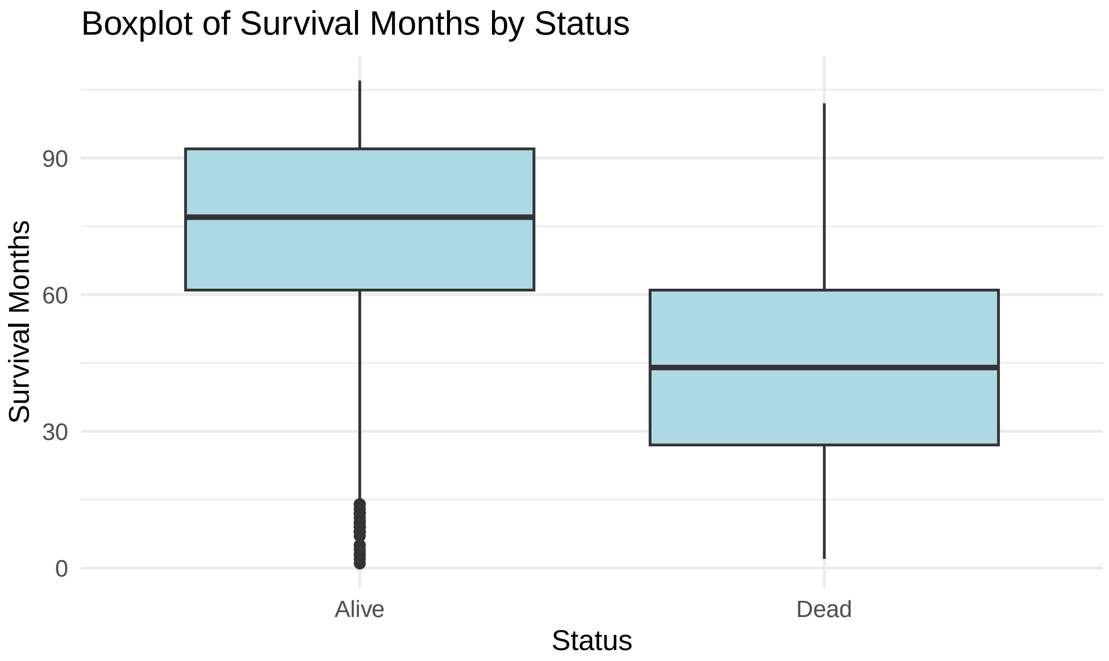

Figure 2: Survival Months by Status

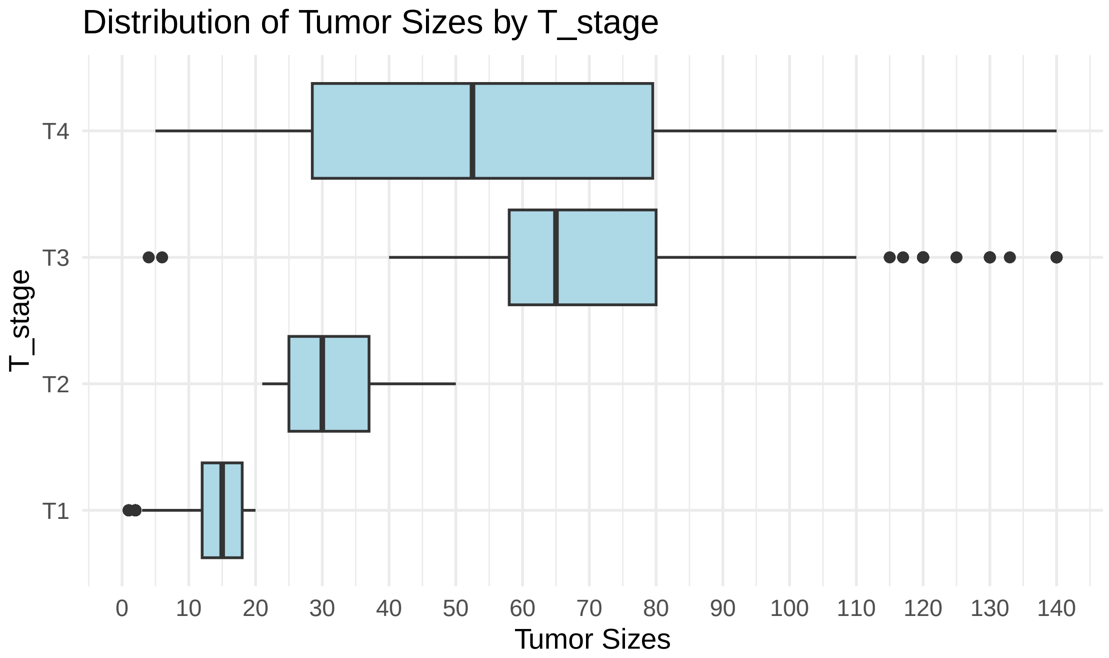

Figure 3: Tumor Sizes by T stage

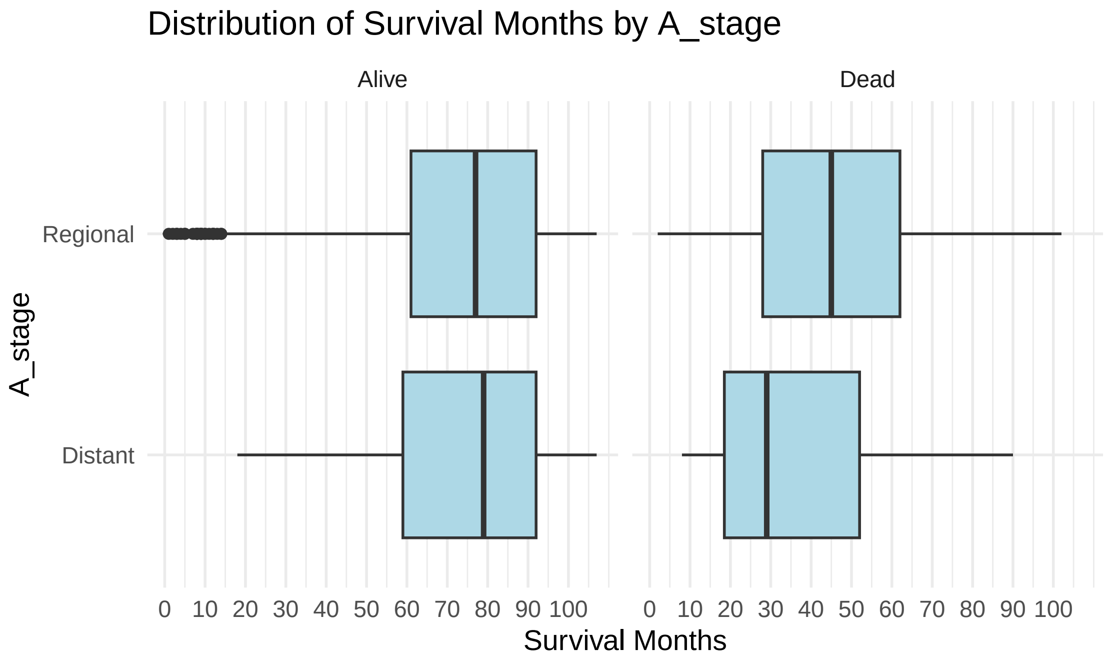

Figure 4: Survival Months by A stage Based on Status

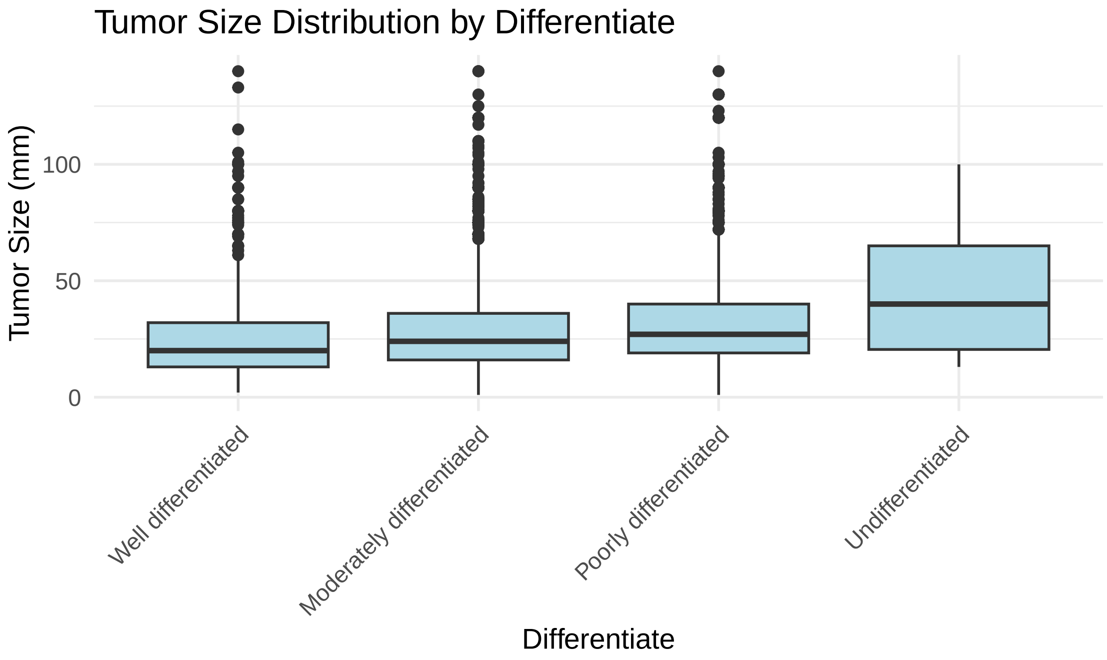

Figure 5: Tumor Size by Differentiate

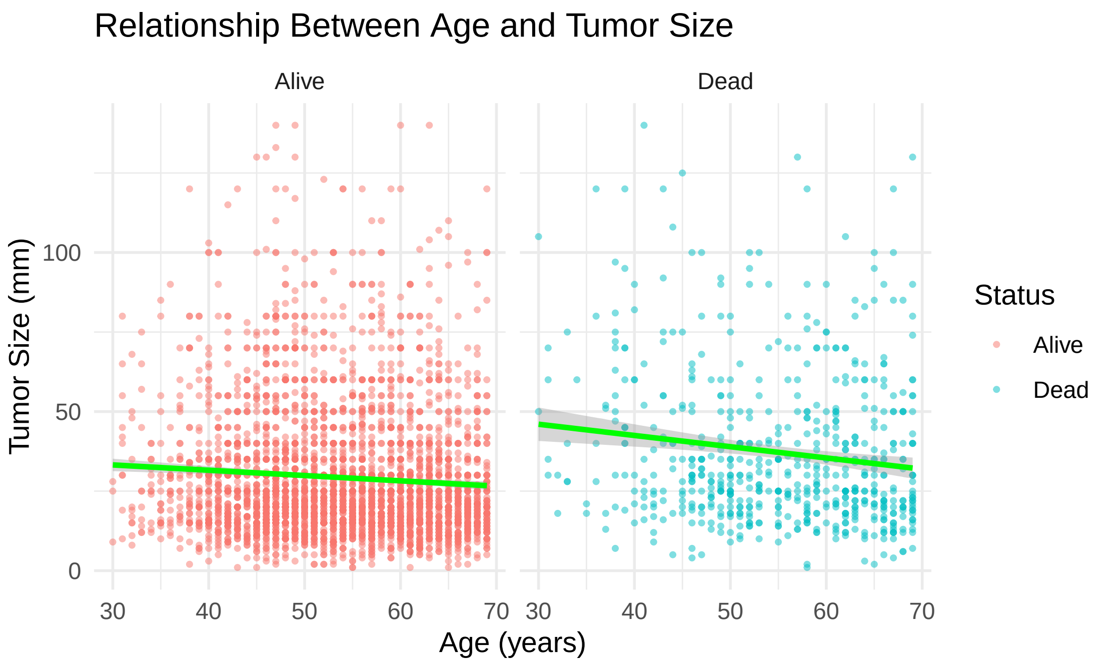

Figure 6: Relationship Between Age and Tumor Size across Status

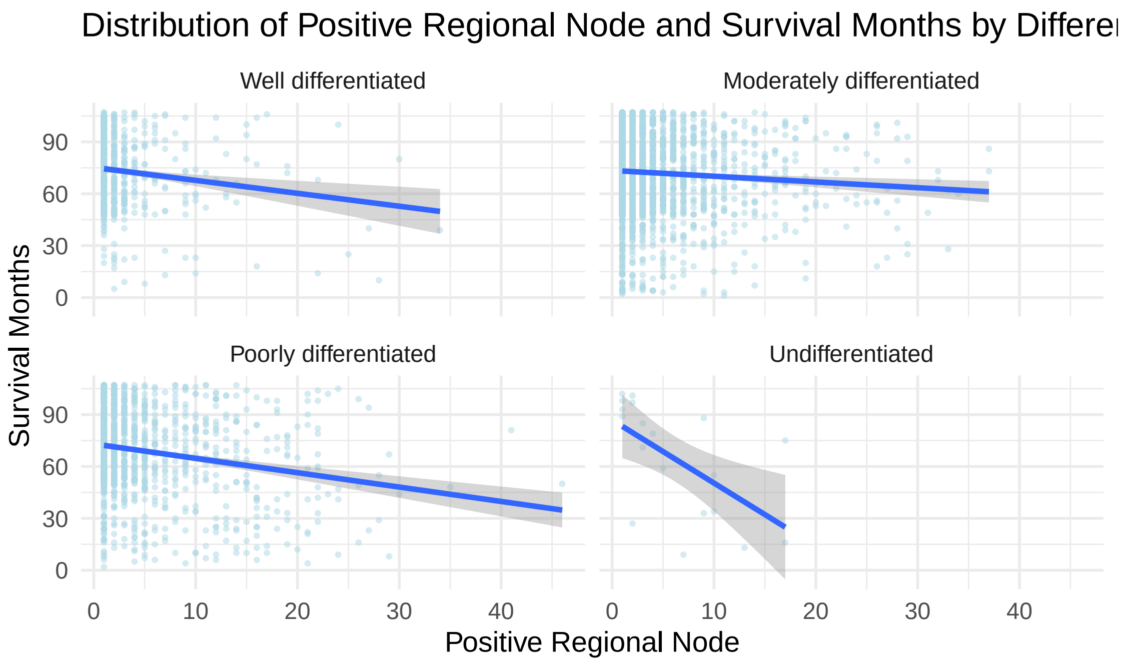

Figure 7: Positive Regional Node vs Survival Months Across Differentiate

Logistic Regression Model

Model Selection

Table 4: Model Selection

type	AIC	BIC
full	3002.000	3159.500
forward	3002.000	3159.500
backward	2993.771	3119.771
stepwise	2993.771	3119.771

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data, and models with lower AIC are generally preferred. Similarly, the Bayesian information criterion (BIC) is also a criterion for model selection among a finite set of models. They both resolve the overfitting problem by introducing a penalty term for the number of parameters in the model.

By comparing AIC and BIC, we can see the model given by backward elimination or stepwise regression works slightly better than the full model or forward selection model. Therefore, we will choose the former to be our "best model".

Odds Ratios

Table 5: Final Model Results with Adjusted-Odds Ratio

	estimate	std_error	z_value	p_value	adjusted_odds_ratio
(Intercept)	-2.2838	0.4385	-5.2085	0.0000	0.1019
age	0.0238	0.0056	4.2426	0.0000	1.0241
raceOther	-0.9346	0.2485	-3.7616	0.0002	0.3928
raceWhite	-0.5148	0.1617	-3.1845	0.0014	0.5976
marital_statusMarried	-0.2110	0.1416	-1.4900	0.1362	0.8097
marital_statusSeparated	0.6691	0.3881	1.7240	0.0847	1.9526
marital_statusSingle	-0.0646	0.1748	-0.3696	0.7117	0.9374
marital_statusWidowed	0.0175	0.2211	0.0791	0.9369	1.0176

	estimate	std_error	z_value	p_value	adjusted_odds_ratio
t_stageT2	0.4111	0.1130	3.6372	0.0003	1.5085
t_stageT3	0.5516	0.1488	3.7077	0.0002	1.7360
t_stageT4	1.0988	0.2445	4.4934	0.0000	3.0005
n_stageN2	0.4363	0.1284	3.3987	0.0007	1.5470
n_stageN3	0.5872	0.2345	2.5034	0.0123	1.7989
differentiateModerately differentiated	0.5328	0.1838	2.8990	0.0037	1.7036
differentiatePoorly differentiated	0.9190	0.1924	4.7772	0.0000	2.5069
differentiateUndifferentiated	1.8649	0.5538	3.3672	0.0008	6.4551
estrogen_statusPositive	-0.7480	0.1775	-4.2140	0.0000	0.4733
progesterone_statusPositive	-0.5842	0.1275	-4.5811	0.0000	0.5576
regional_node_examined	-0.0359	0.0072	-5.0110	0.0000	0.9647
regional_node_positive	0.0797	0.0153	5.2076	0.0000	1.0829

Model Diagnostics

Table 6: Examination for Multicolinearity

	GVIF	Df	GVIF^(1/(2*Df))
age	1.1072	1	1.0522
race	1.0629	2	1.0154
marital_status	1.1291	4	1.0153
t_stage	1.1019	3	1.0163
n_stage	3.8068	2	1.3968
differentiate	1.1171	3	1.0186
estrogen_status	1.4754	1	1.2147
progesterone_status	1.4275	1	1.1948
regional_node_examined	1.4778	1	1.2157
regional_node_positive	4.2484	1	2.0612

Variance Inflation Factor is a commonly used method for detecting multicollinearity in regression models. VIF is generally calculated for the continuous variables, and Generalized Variance Inflation Factor (GVIF) is used for evaluating the multicollinearity for categorical variables.

The adjusted GVIF (i.e. GVIF^(1/(2*Df))) values are corrected for the degree of freedom and provide a scale similar to VIF. The high adjusted GVIF values (GVIF > 2) indicate the presence of moderate to strong multicollinearity.

The table shows that most variables do not show multicollinearity, with the exception of `regional_node_positive`. Since its adjusted GVIF is not much different from 2, we will keep this variable for now.

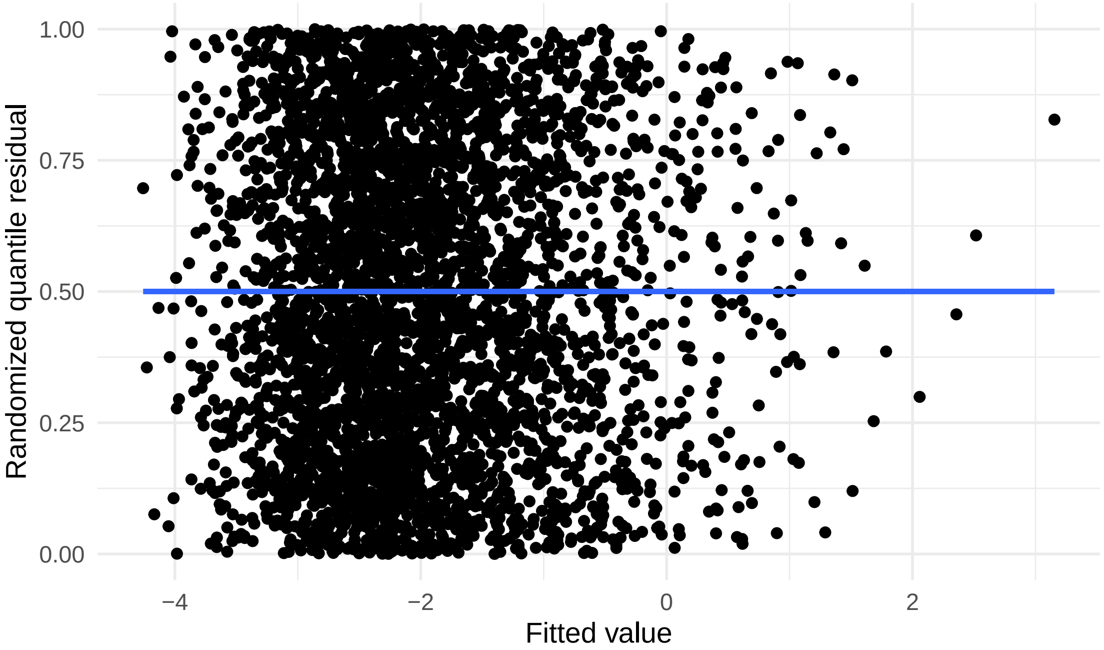

Figure 8: Random Quantile Residual versus Fitted Values Plot

By randomizing the quantile residuals, we resolve the problem that the RVF plot always shows a pattern in logistic regression because of the binary response variable. Since in the randomized quantile residual vs. fitted value plot, the residuals distribute randomly around the 0.5 horizontal line, the residual assumption is met and the model is a good fit.

The residual vs. leverage plot indicates that observations 3527, 1561, and 3074 may be potential outliers, but they are not necessarily influential.

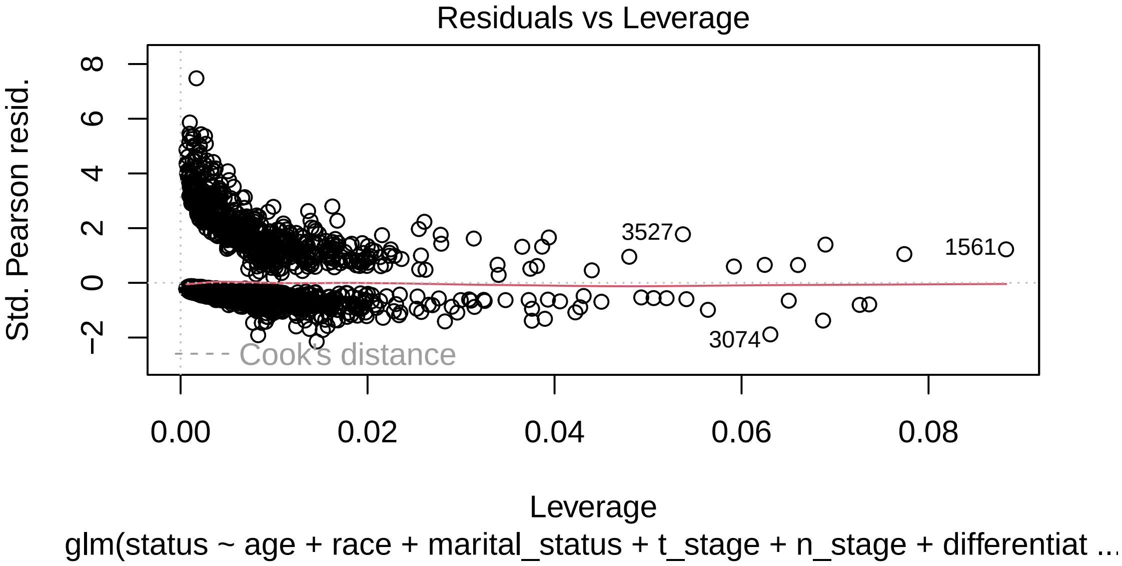

Figure 9: Residual versus Leverage Plot

Cross Validation

Table 7: Results of 10-Fold Cross Validation

log_loss	AUC
0.3681	0.6999
0.3639	0.7463
0.4069	0.7498
0.3773	0.7552
0.3575	0.7792
0.3780	0.7317
0.3679	0.6604
0.3636	0.7951
0.3681	0.7644
0.3718	0.7405

After applying 10-fold cross-validation, we evaluate the goodness of fit by log loss and AUC. The mean of log loss is 0.3723182, and the mean of AUC is 0.7422428.

Evaluation Across Races

Table 8: Race Comparison Before Adding Interaction Terms

race	avg_log_loss	avg_AUC
Black or other	0.4231	0.6997
White	0.3651	0.7502

Low log loss and high AUC indicate better test performance.

To reduce the gap of prediction performance between the majority and minority, we focused on whether there were interactions between the variables. We extracted each variable from the best model and examined how it differed in survival months of survival by race. Most variables did not show significant differences by race, suggesting that there may not be an interaction between these variables and race. However, the variable marital status showed a different pattern.

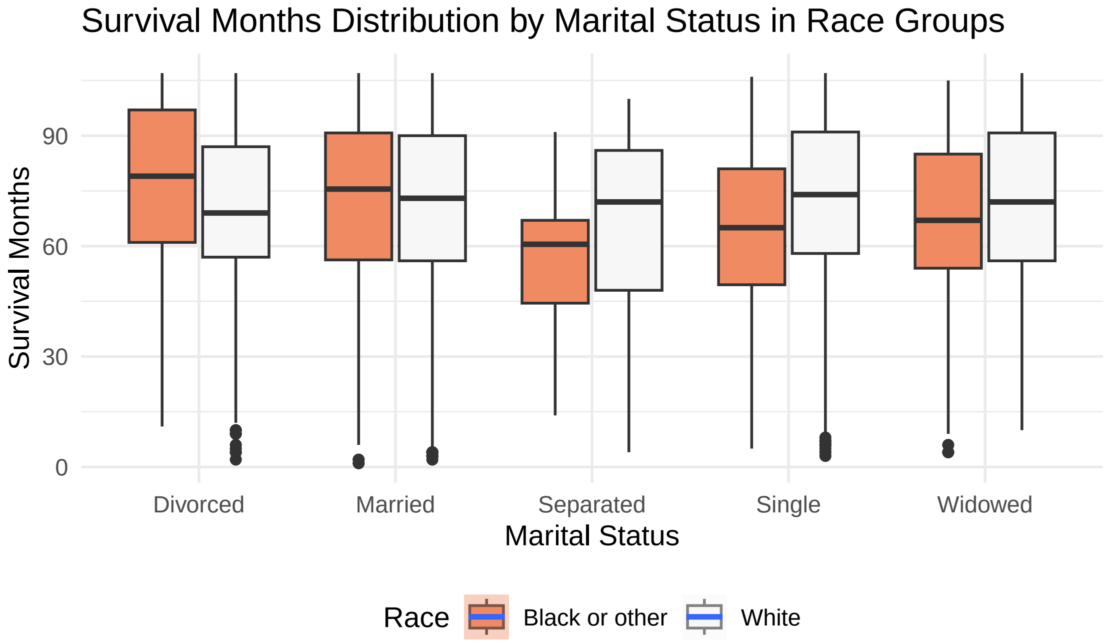

Figure 10: Survival Months Distribution by Marital Status in Race Groups

From the figure we can see that the distribution of survival months is different between races with different marital status. This indicates the potential interaction between race and marital status, and the interaction term can be added in the model to improve the fairness of the model.

Table 9: Race Comparison After Adding Interaction Terms

race	avg_log_loss	avg_AUC
Black or other	0.4169	0.7251
White	0.3647	0.7501

By adding interaction term `marital_status * race`, we can observe a decrease in log loss and an increase in AUC, which means an improve in the fairness between group “White” and the minority “Black” + “Other”.

Survival Analysis

Kaplan Meier Curve

The Kaplan Meier curve graphically represent the survival rate. Time is plotted on the x-axis and the survival rate is plotted on the y-axis.

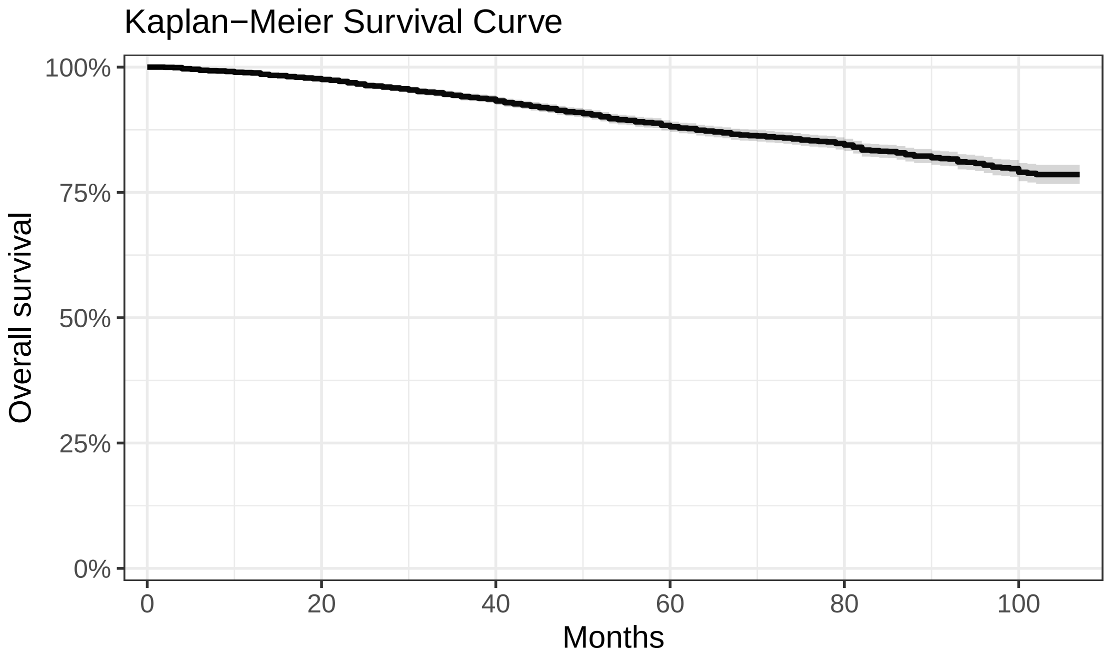

Figure 11: Kaplan-Meier Survival Curve

Log Rank Test

The log rank test lets us test whether there is a difference in survival times between groups of patients. For example, we want to find out whether there is a significant difference in survival between patients whose cells have different degrees of differentiation.

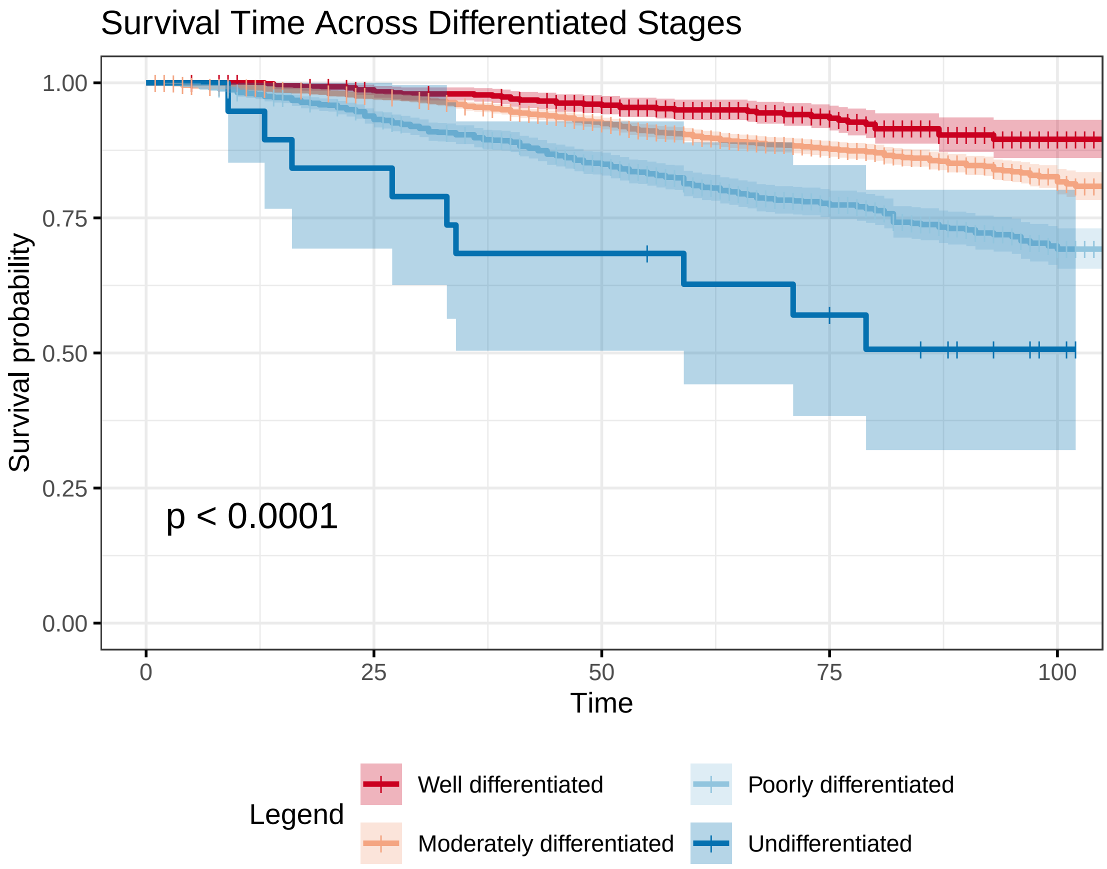

Figure 12: Survival Time Across Differentiated Stages

Cox Model

The limitation of KM curves and log-rank tests is that we can only test one variable at a time. To further discuss the risk factors to survival time, we will compute the cox proportional hazard model to adjusts for multiple risk factors simultaneously.

The cox proportional hazard model has a assumption: the survival curves for two different strata of a risk factor must have hazard functions that are proportional over time. This assumption is satisfied when the change in hazard from one category to the next does not depend on time. That is, a person in one stratum has the same instantaneous relative risk compared to a person in a different stratum, irrespective of how much time has passed.

We will test this assumption based on the scaled Schoenfeld residuals. Here is an interpretation of the results: When p-val < 0.05, there is evidence against the proportional hazards assumption, meaning that the HR is not constant over time. Similarly, the larger the chi-square value, the greater the violation of the assumption.

Table 10: Results of Cox Proportional Hazard Model

	chisq	df	p
age	0.1328	1	0.7156
race	0.9335	2	0.6270
marital_status	2.6670	4	0.6150
t_stage	0.2144	3	0.9752
n_stage	1.7178	2	0.4236
x6th_stage	3.8545	3	0.2776
differentiate	1.8899	3	0.5956
a_stage	5.2218	1	0.0223
tumor_size	0.9310	1	0.3346
estrogen_status	28.9294	1	0.0000
progesterone_status	32.1281	1	0.0000
regional_node_examined	0.0187	1	0.8912
regional_node_positive	0.0324	1	0.8571
GLOBAL	57.2155	24	0.0002

We can see from the table that variable `a_stage`, `estrogen_status`, `progesterone_status` are not constant over time, which means it's not proper to contain these covariates in cox regression. To reduce bias of the model, we can remove these variables and take a closer look at the result.

The hazard ratio is similar to relative risk, but differs in that the HR is the instantaneous risk rather than the cumulative risk over the entire study.

The x-axis of this forest plot represents hazard ratios. Hazard ratio = 1 means no significant difference compared to the reference, and a HR higher than 1 means it increases the hazard ratio of the event, death, and a HR lower than 1 decreases it. The smaller the p-value is the stronger the weight of evidence that the two groups are different.

We can conclude from the plot that for the variable race, blacks have the highest hazard of death, followed by whites, while the lowest mortality rate is for other ethnic groups. In the variable marital status, the hazard of death is significantly higher for separated people, but this may be due to information bias caused by fewer observations. The confidence intervals for the other categories of marital status all contain the null hypothesis, meaning that there is no significant difference.

The hazard of death is highest for patients with N stage N3, followed by N2, and finally N1. Differently, although T stage also shows a similar trend, the confidence intervals of each stage level contain the null hypothesis, meaning that there is no significant difference between levels. For the 6th stage, IIIB has the highest hazard of death, followed by IIB, and then IIA, but there is no significant difference. For stage IIIC, since it contains the same information as N3 of N stage, no comparison is made in this variable.

In the variable differentiated, the hazard of death is significantly highest for undifferentiated, and then decreases in the order of poorly differentiated, moderately differentiated, and well differentiated.

For the variables tumor size, regional node examined, and regional node postive, we did not observe significant differences in the hazard of death.

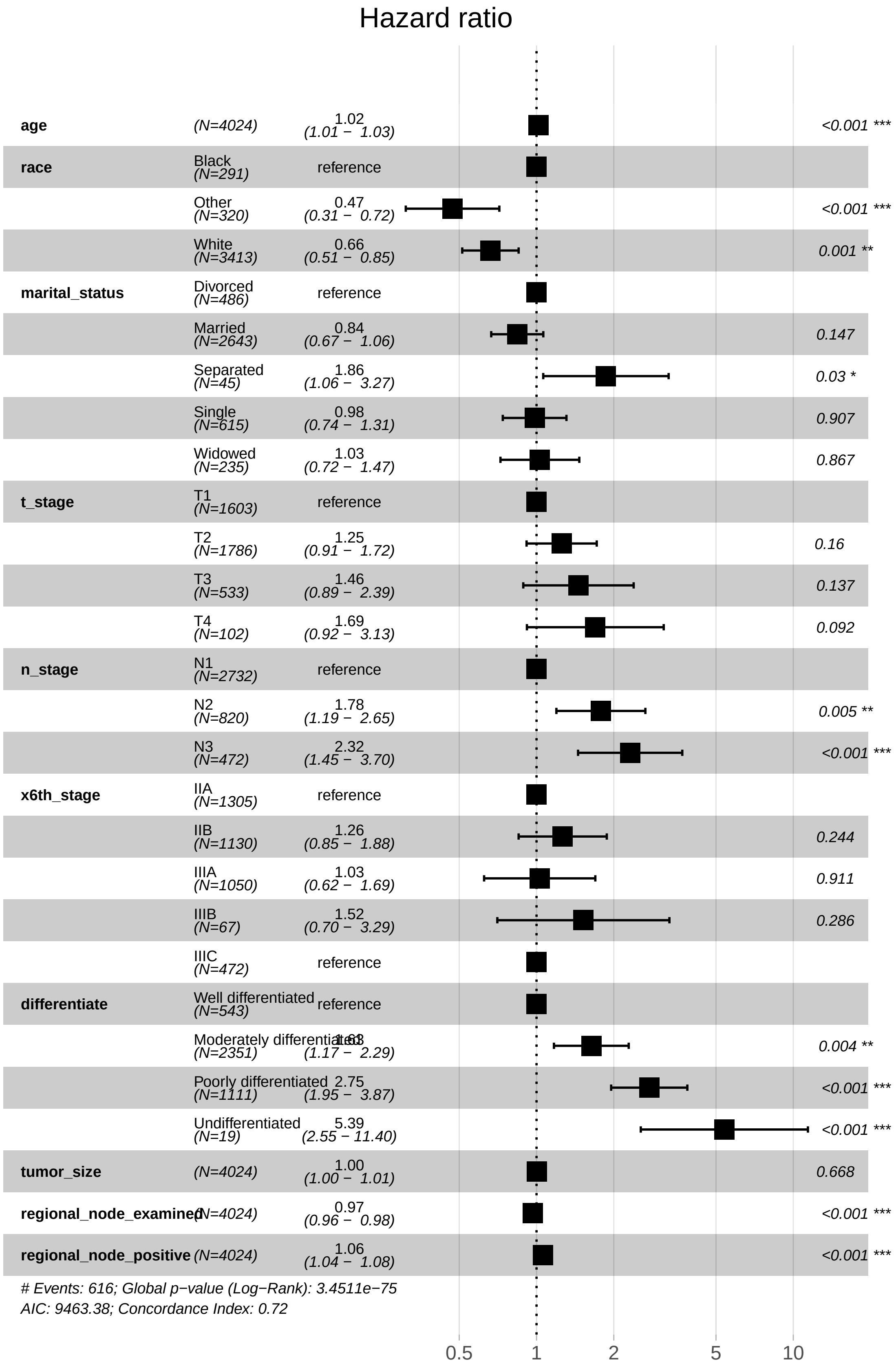

Figure 13: Forest Plot of Hazard Ratios