Survival Analysis of Skin Cutaneous Melanoma Time to Event Data Mined from The Cancer Genome Atlas Program

Lin Zou

Introduction and Background

Skin cutaneous melanoma causes more than 10,000 deaths each year; however, the cancer has a 98% five-year survival rate when it is diagnosed early (*Cutaneous Melanoma*, n.d.). Therefore, it is important to identify the characteristics, preferably those collected at baseline, that increase the chance of survival in order to help clinicians develop patient-specific treatments. To identify these characteristics, we performed Kaplan-Meier curve estimation, which is a nonparametric method used to model the relationship between the time of occurrence of the event of interest, called a death, and the previous occurrence of some other event called a loss (Kaplan & Meier, 1958). To assess the curve estimations generated from the different levels of the covariates for statistically significant differences in survival, we implemented a log-rank test. This method tests the null hypothesis that survival curves of two populations do not differ against the alternative hypothesis that they do differ (Mantel, 1985). Lastly, we considered all of our chosen covariates at the same time using a Cox proportional hazard model, which is a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time (Cox, 1972).

In contribution to the literature, our research objective is to demonstrate the use of survival analysis techniques in identifying demographic and clinical variables that could give more insight into early diagnosis and treatment of skin cutaneous melanoma.

Methods and Materials

Data Exploration

We used observational skin cutaneous melanoma time-to-event data collected by the Cancer Genome Atlas program (TCGA). The data is right censored, so the response is a variable that consists of the days to death and the vital status of the patients. Censoring could have been due to the termination of the study before the event occurred or the patient failed to participate in a follow-up medical exam. If the patient's death was observed (vital status = 1), the days to death for that patient was reported. If the patient's survival time was censored (vital status = 0), the days to death for that patient was marked with a cross symbol to indicate censoring. The covariates we chose to include in our analysis are shown in Table 1. Age at index was dichotomized based on the median age in the dataset, and tumor stage was dichotomized due to having a small number of patients with stage 0, I/II, or IV tumors.

Table 1. Demographics table of all the predictors by deceased patients and censored patients.

	Deceased Patients (n = 223)	**Censored Patients (n = 246)**
Age at Index (Less than 60 : Over 60)	124 : 99	133 : 113
Gender (Female : Male)	75 : 148	105 : 141
Tumor Stage [Early Stages (Stage 0, I, I/II, II) : Later Stages (Stage III and IV)]	122 : 101	142 : 104
Treatment (None : Radiation : Pharmaceutical : Both)	88 : 35 : 45 : 55	127 : 29 : 62 : 28
Prior Malignancy (No : Yes)	210 : 13	237 : 9

Table 1 shows that the 469 patients are primarily represented by those that are less than 60 years old, male, in the early stages of the cancer, or not previously diagnosed with a malignant tumor. We can observe that the majority of the patients received some form of therapy, whether it was radiation, pharmaceutical, or both. Our preliminary exploration of the data revealed some potential associations between the predictors and skin cancer survival. Specifically, being young or being female is associated with better survival. In addition, treatment of the tumor in its early stages may be linked to patient survival. Survival may also be associated with undertaking either radiation or pharmaceutical therapies, but not both therapies. On the other hand, it appears that prior malignancy does not affect patient survival. Furthermore, missing data was imputed using the Multivariate Imputation via Chained Equations (MICE) package in R.

Data Analysis

Kaplan-Meier estimation is a nonparametric method used to estimate the probability of survival over time. To perform this estimation, we have $S(t) = \exp[-H(t)]$ for a continuous random variable and $S(t) = \prod_{s=0}^{t}[1 - d\widehat{H}(s)]$ for a discrete random variable, where t is time, $S(t)$ is the survival function, $H(t)$ is the cumulative hazard rate based on right-censored data, $\widehat{H}(s)$ is the Nelson-Aalen nonparametric estimator of $H(t)$, and $d\widehat{H}(s)$ is the change over time in the Nelson-Aalen estimator. Hazard is the probability that an individual who is under observation at a specific time has an event at that time. Note that $S(t)$ is the product integral of $1 - d\widehat{H}(t)$. To obtain the Kaplan-Meier estimator of the survival function, we take the product integral of $1 - d\widehat{H}(t)$ and get

$$\hat{S}(t) = \prod_{s=0}^{t}[1 - d\widehat{H}(t)] = \prod_{s=0}^{t}[1 - \frac{dN(s)}{Y(s)}]$$

where $\hat{S}(t)$ is the estimated survival function and $d\hat{H}(t) = \frac{dN(s)}{Y(s)}$ is the change over time in $N(s)$, an increasing step function with jumps at the observed death times, divided by $Y(s)$, a decreasing step function with steps of size one at each censoring time. $\hat{S}(t)$ is a step function with steps at the death times where $dN(t) > 0$ (Klein & Moeschberger, 2003). Kaplan-Meier estimation has the following statistical assumptions: the probability of censoring is unrelated to the outcome of interest, the survival probabilities must remain constant for participants throughout the study, and the events actually occurred at the times specified (Koletsi & Pandis, 2017). To assess the curve estimations generated from the different levels of the predictors for statistically significant differences in survival, we implemented the log-rank test.

The log-rank test is a nonparametric method used to identify significant differences between two or more survival curve estimations, where each curve estimates the survival probability of each group. The null hypothesis of the test is that there is no difference in survival between the two curves. The alternative hypothesis is that there is a significant difference in survival between the two curves (Klein & Moeschberger, 2003). We tested this hypothesis with significance level $\alpha = 0.05$. The log-rank test compares the observed number of events in each group to the expected frequencies under the null. The test statistic approximately follows a χ^2 test statistic on $g - 1$ degrees of freedom, where $g = 1,\ldots,$ j is the total number of groups. The equation of the log-rank test statistic for g groups is

$$\chi^2 = \Sigma \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}}$$

where ΣO_{jt} is the sum of the observed number of events in the jth group over time and ΣE_{jt} is the sum of the expected number of events in the jth group over time (Klein & Moeschberger, 2003). Note that the log-rank test makes no assumptions about the shape of the survival curve. In

addition to univariate analyses, we also used a Cox proportional hazard model to explore additional associations of survival predictors.

The Cox proportional hazard model is a regression model used to describe the association between the survival time and multiple predictors. It assumes proportionality of the hazard rates of individuals with distinct values of a covariate, absence of influential observations and outliers, and linear continuous variables (Klein & Moeschberger, 2003). To assess the validity of the statistical assumptions in our Cox proportional hazard model, we performed χ^2 tests of independence between Schoenfeld residuals and time for each covariate. Schoenfeld residuals are the differences between the observed covariates and the expected covariates at each occurrence of an event. The null hypothesis is that the residuals and time are independent, and the alternative hypothesis is that they are dependent. We also performed a global test for the model. Next, we checked the diagnostic plots of the Schoenfeld residuals for proportional hazards in each covariate. These residuals are independent of time and have expectation zero under the Cox model, so any trends would violate this assumption. Lastly, we checked the deviance residuals for influential observations and outliers. The positive deviance residuals correspond to patients that died before the expected survival time, while the negative deviance residuals correspond to patients that lived beyond the expected survival time. Since we do not have any continuous variables in our model, we did not check the Martingale residuals for the linearity of continuous variables.

Examining the diagnostic plot of the deviance residuals, the points appear to be symmetrically distributed around zero and there does not appear to be any influential observations or outliers (see Figure Aiii in Appendix). Also, recall that the proportional hazard assumption is satisfied if there is a nonsignificant relationship between residuals and time. From

the χ^2 independence tests, age at index, gender, and prior malignancy were not statistically significant, but tumor stage, treatment, and the overall model were statistically significant, at the 0.05 significance level (see Table Ai in Appendix). Moreover, there were no trends in the diagnostic plots (see Figure Aiv in Appendix).

Since tumor stage and treatment violate the proportional hazard assumption, we incorporated these nonproportional effects into the model as stratification factors instead of predictors. The strata are composed of the eight factor/level combinations of tumor stage and treatment. Now the observations in the z = 1, … , 8th stratum have an arbitrary estimated baseline hazard function $h_{0z}(t)$ and the effect of age at index, gender, and prior malignancy on the estimated hazard function is represented by the proportional hazard model in that stratum, such that

$$\widehat{h_z(t, \boldsymbol{X})} = \widehat{h_{0z}(t)}\, e^{\sum_{j=1}^{p} \hat{\beta}_j X_j}$$

where t is time, $X = x_1, \ldots, x_3$, p = 1, …, 3, and $\hat{\beta}$ are the estimated coefficients (Klein & Moeschberger, 2003). The exponentiated estimated coefficients can be interpreted as hazard ratios. In the stratified proportional hazard model, the regression coefficients are assumed to be the same for each stratum, whereas the baseline hazard functions can be different for each stratum. After the stratification, we can observe that the estimated hazard ratios for age at index, gender, and prior malignancy are similar to those before the stratification. Note that by stratifying tumor stage and treatment, we are no longer able to estimate their hazard ratios. However, we were able to satisfy the proportional hazard assumption, as all the covariates in the test of independence were nonsignificant and there were no trends in the diagnostic plots (see Table Aii, Figure Av, and Figure Avi in Appendix). All the statistical assumptions of the Cox proportional hazard model were satisfied, so we proceeded with the analysis.

Results

Our analysis started with examining whether or not there was any association between the groups for each predictor by comparing their survival distributions. Table 2 highlights the results of three Kaplan-Meier models fitted with only age at index, tumor stage, or treatment. For each group within each covariate, we show the sample size (Records), the number of observed deaths (Events), and the median survival in the group along with its estimated 95% confidence interval. We wanted to discuss these models in particular due to the clear differences in median survival between the groups of each covariate. The results for all the Kaplan-Meier models can be found in Table A2 of the Appendix.

Table 2. Results from Kaplan-Meier models. Only the covariates with large median survival differences between levels are displayed.

	Groups	**Records**	**Events**	**Median Survival (Days) (95% Confidence Interval)**
Age at Index	Less than 60	257	124	3195 (2421, 4601)
	Over 60	212	99	1832 (1487, 2071)
Tumor Stage	Early Stages (Stage 0, I, I/II, II)	264	122	3379 (2273, 4222)
	Later Stages (Stage III and IV)	205	101	1832 (1124, 2192)
Treatment	None	215	88	2192 (1780, 5237)
	Radiation	64	35	3424 (2711, 6164)
	Pharmaceutical	107	45	3141 (1628, 5110)
	Both	83	55	1857 (1446, 2184)

From Table 2, we can observe that patients less than 60 years old survived longer than patients over 60 years old. Taking this into consideration, the overall health of older patients may not be able to withstand the effects of skin cancer. In addition, patients in the early stages of their tumor survived longer than patients that were in the later stages of their tumor. This makes sense, as early stage tumors have not spread to other parts of the body and are easier to treat than later stage tumors that have spread to distant parts of the body. Lastly, patients treated with either radiation or pharmaceutical therapy tended to survive longer than those who did not participate in any therapy or were treated with both therapies. Specifically, patients that did radiation therapy survived the longest, while patients that did both therapies survived the least amount of days. From this information, we suspect that the therapies by themselves work well in keeping patients healthy; however, undergoing both therapies may lead to harmful side effects from the combination of radiation and drugs, and may even be worse than doing no therapy. Lastly, we did not detect any large differences in median survival between the levels of gender and between the levels of prior malignancy. This suggests that being a female or male offered no physiological advantages in terms of skin cancer survival. However, note that prior malignancy status is overrepresented by no prior malignancy, so the absence of a difference in median survival between the levels of prior malignancy may be misleading.

As an example, we will showcase and discuss the Kaplan-Meier model fitted only with age at index using Figure 1. Plots of the four Kaplan-Meier models fitted with each of the other four categorical covariates and the Kaplan-Meier model fitted with only the intercept can be found in Figures Avii – Axi of the Appendix.

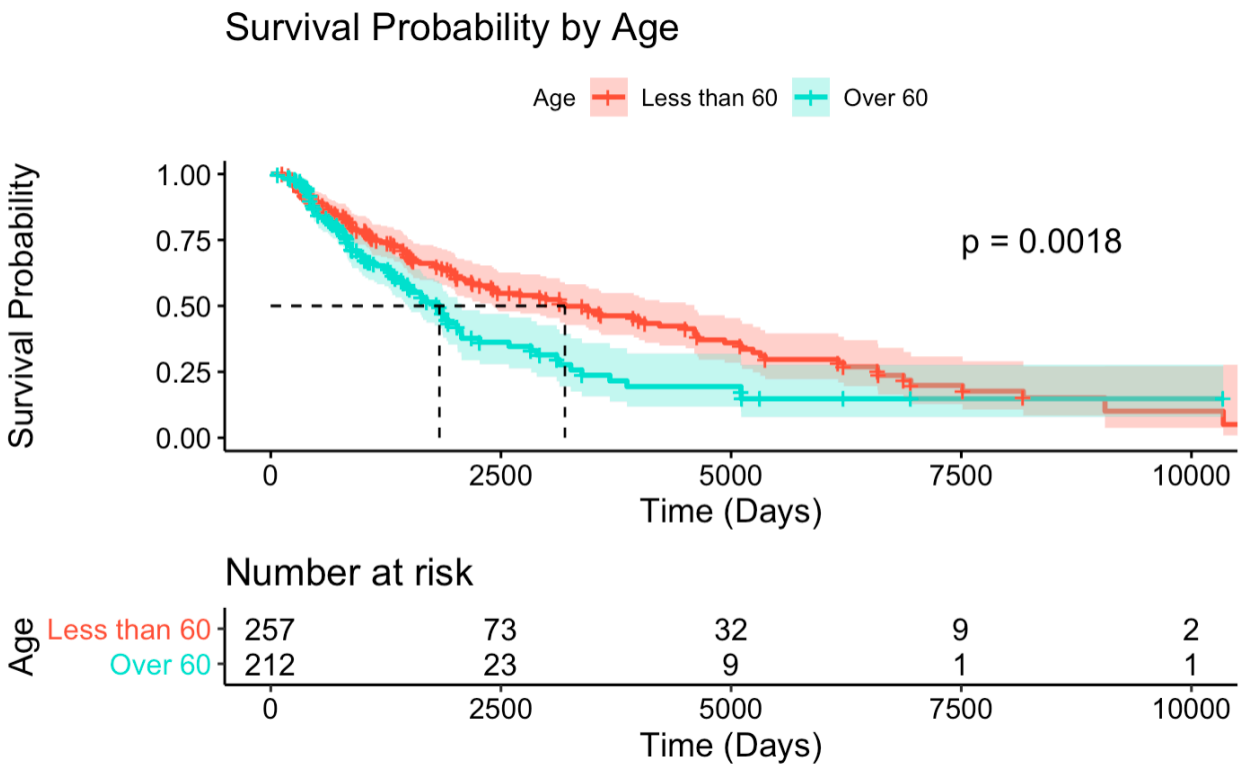


Figure 1. Estimated survival probability over time computed from the Kaplan-Meier model fitted with only age. Red corresponds to patients less than 60 years old and blue corresponds to patients over 60 years old. The p-value is computed from a log-rank test and the table shows the number of patients at risk of dying from skin cancer at each time point for both groups.

Figure 1 shows the estimated survival probability over time computed from the Kaplan-Meier model fitted with only age. Survival probability is the probability that an individual survives from the first diagnosis to a specified future time. The reported p-value is the log-rank test p-value for the age covariate. We can observe that patients over the age of 60 have lower survival probability compared to patients less than 60 years old. This was expected because we saw that patients less than 60 years old survived longer than patients over 60 years old. From the plots of the other models, we observed that patients with later tumor stages had lower survival probability compared to patients with early tumor stages. Moreover, patients that had both therapies or no therapy have lower survival probability compared to patients that had either radiation or pharmaceutical therapy. Contrarily, survival probability was indistinguishable when

comparing female and male patients and when comparing patients with and without a prior malignancy, as we expected.

To confirm the difference in survival between the two age groups using a formal statistical test, we implemented the log-rank test, and discovered that there are statistically significant differences for the age covariate, in addition to tumor stage and treatment (p-values < 0.05). Based on what we observed from our Kaplan-Meier curves, we expected these results to occur. The log-rank test results for all the covariates can be found in Table A3 of the Appendix.

The next step in our analysis utilized the stratified Cox proportional hazard model to assess the differences in survival function between the groups for each predictor after adjusting for all other predictors in each of the tumor stage and treatment factor/level combinations. Figure 2 shows the estimated survival probability over time for each stratum.

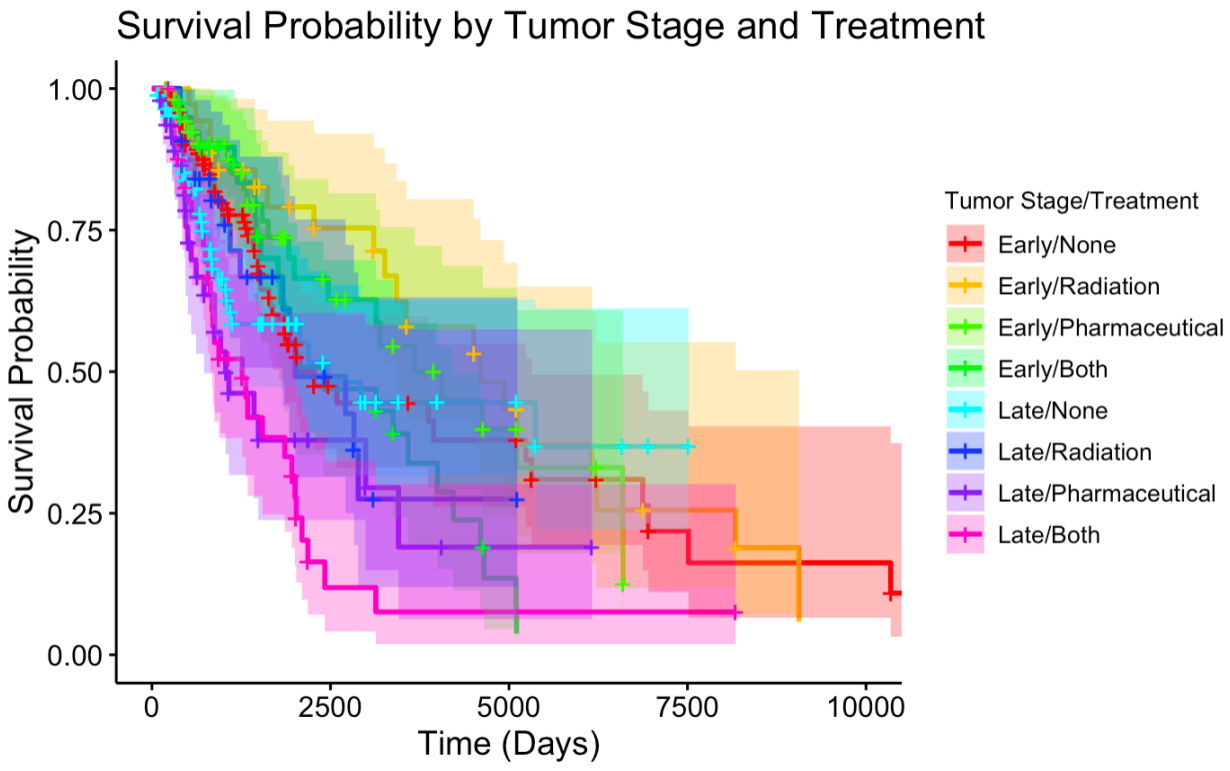


Figure 2. Estimated survival probability over time from stratified Cox proportional hazard model.

From Figure 2, we can observe that using just pharmaceutical or using both therapies to treat late stage tumors had the lowest survival probability, while no therapy or radiation therapy for late stage tumors had higher survival probability. This suggests that using just pharmaceutical or using both therapies to treat late stage tumors is even more detrimental than if the patient did not undergo any therapy. On the other hand, radiation therapy may be the most effective treatment for late stage tumors.

We can also see that no therapy or using both therapies for early stage tumors had lower survival probability compared to pharmaceutical or radiation therapy for early stage tumors. This suggests that using both therapies to treat early stage tumors is excessive and may cause more harm than benefit. Contrarily, using either therapies on their own may be the most effective method of treatment for early stage tumors.

A log-rank test of our Cox proportional hazard model suggested that we had enough statistical evidence to reject the null hypothesis and conclude that there is at least one difference in the survival function when comparing the groups for each predictor after adjusting for all other predictors (χ_3^2 = 10.55, p-value < 0.05). After confirming the presence of significant differences, we turned our attention to the covariates and discovered that only age [Estimated hazard ratio (95% confidence interval) = 1.6 (1.1965, 2.164)] was a statistically significant predictor of skin cancer survival (p-value < 0.05). To interpret the hazard ratio for age, we say adjusting for all other predictors, patients over the age of 60 have 1.6 times the hazard of dying from skin cancer in comparison to patients under the age of 60. This makes practical sense, as older patients may not be able to withstand the adverse health effects of skin cancer or the side effects of the various therapy treatments. Lastly, we expected gender and prior malignancy to be

nonsignificant based on our results from the Kaplan-Meier curves and the log-rank tests. The estimated hazard ratios for all the covariates can be found in Table A4 of the Appendix.

Discussion and Summary

Survival analysis is a useful methodology for handling the bias present in censored time to event data. In particular, we found that the Cox proportional hazard model results were easy to interpret because it was similar to interpreting the estimated coefficients from a logistic regression as odds ratios. Furthermore, the consistencies in the results throughout the survival analysis indicate that these techniques are reliable in identifying demographic and clinical variables that could give more insight into early diagnosis and treatment of skin cutaneous melanoma. Based on our results, we recommend clinicians to develop alternative treatment procedures for patients that are over the age of 60, have stage III or IV tumors, or undergo both radiation and pharmaceutical therapy, as these characteristics are associated with a higher risk of dying from skin cancer.

The potential issues of our analysis include the way we dichotomized age at index using the median as the cutoff and the way we addressed nonproportionality in the Cox proportional hazard model using stratification. Future studies would have more clinical relevance if a reason was established for the dichotomization of the age data at a specific cutoff. Moreover, future studies may explore other ways of handling covariates with nonproportional effects, such as adding covariate and time interaction terms. Although the interpretability of such a model would suffer from the higher order terms, we may observe improvements in the accuracy of its predictions.

References

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220. Retrieved March 6, 2020, from www.jstor.org/stable/2985181

Cutaneous Melanoma. (n.d.). UPMC Hillman Cancer Center. Retrieved March 6, 2020, from https://hillman.upmc.com/cancer-care/melanoma-skin/types/cutaneous-melanoma

Kaplan, E., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi:10.2307/2281868

Klein, J.P., & Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data: Vol. Second edition.* Springer.

Koletsi, D., & Pandis, N. (2017). Survival analysis, part 2: Kaplan-Meier method and the log-rank test. *American Journal of Orthodontics and Dentofacial Orthopedics*, *152*(4), 569–571. https://doi.org/10.1016/j.ajodo.2017.07.008

Mantel, N. (1985). Propriety of the Mantel-Haenszel Variance for the Log Rank Test. *Biometrika, 72*(2), 471-472. doi:10.2307/2336101

Statistical tools for high-throughput data analysis. (n.d.). *Cox Proportional-Hazards Model*. Retrieved March 6, 2020, from http://www.sthda.com/english/wiki/cox-proportional-hazards-model

Statistical tools for high-throughput data analysis. (n.d.). *Survival Analysis Basics*. Retrieved March 6, 2020, from http://www.sthda.com/english/wiki/survival-analysis-basics#log-rank-test-comparing-survival-curves-survdiff

The Cancer Genome Atlas Program (nciglobal,ncienterprise). (2018, June 13). [CgvMiniLanding]. National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Views, R. (2017, September 25). *Survival Analysis with R*. /2017/09/25/survival-analysis-with-r/

Appendix

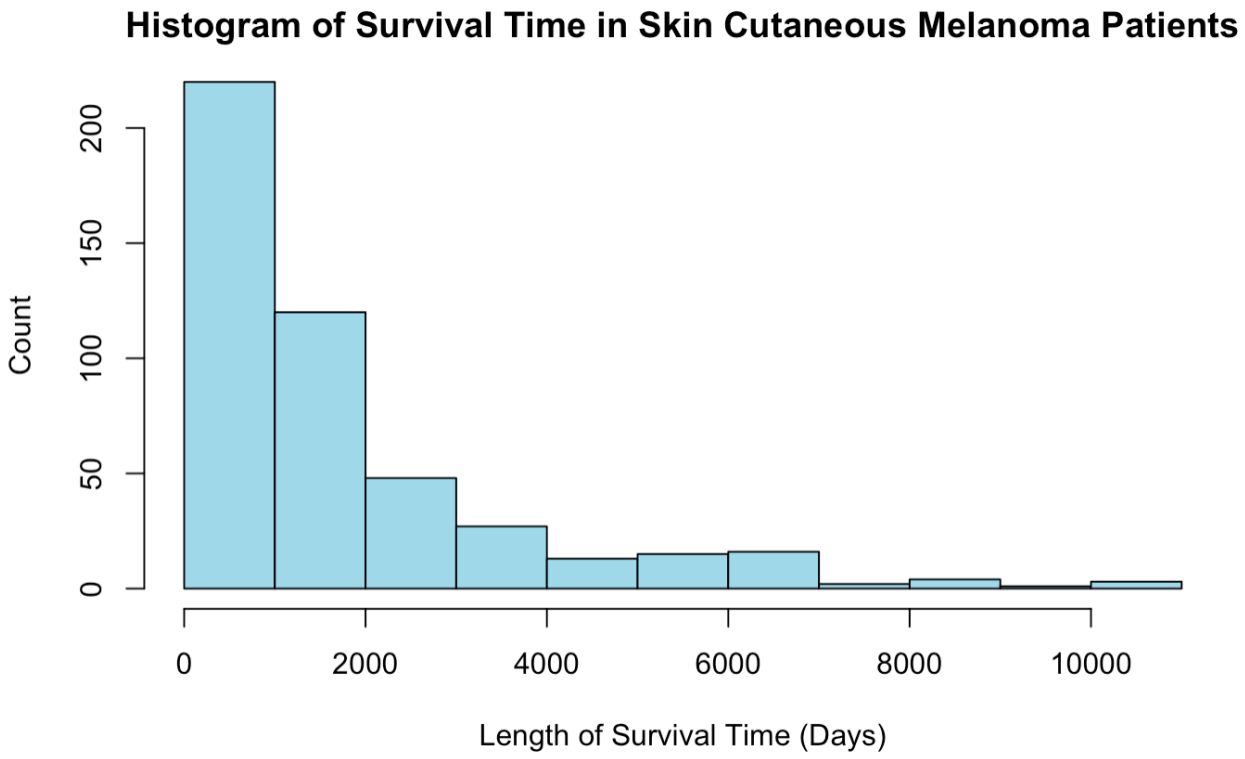


Figure Ai. Histogram of survival times in skin cutaneous melanoma patients.

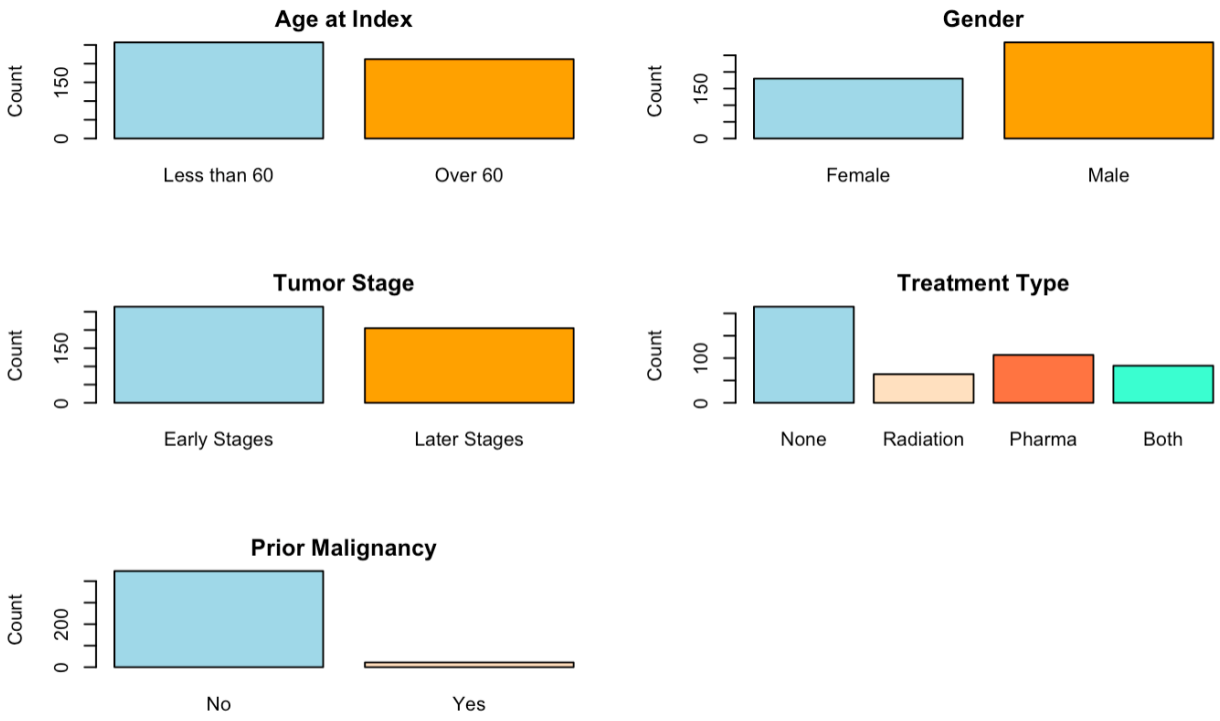


Figure Aii. Distributions of all the covariates.

Table Ai. χ^2 test of independence of Schoenfeld residuals and time for each covariate before stratifying tumor stage and treatment. Nonproportional effects are bolded.

	χ^2 Test of Independence
Age at Index	χ_1^2 = 0.00481 p-value = 0.94
Gender	χ_1^2 = 0.88194 p-value = 0.35
Prior Malignancy	χ_1^2 = 0.75924 p-value = 0.38
Tumor Stage	χ_1^2 = 13.43347 p-value = 0.00025
Treatment	χ_3^2 = 11.52062 p-value = 0.00922
Global	χ_7^2 = 26.15812 p-value = 0.00047

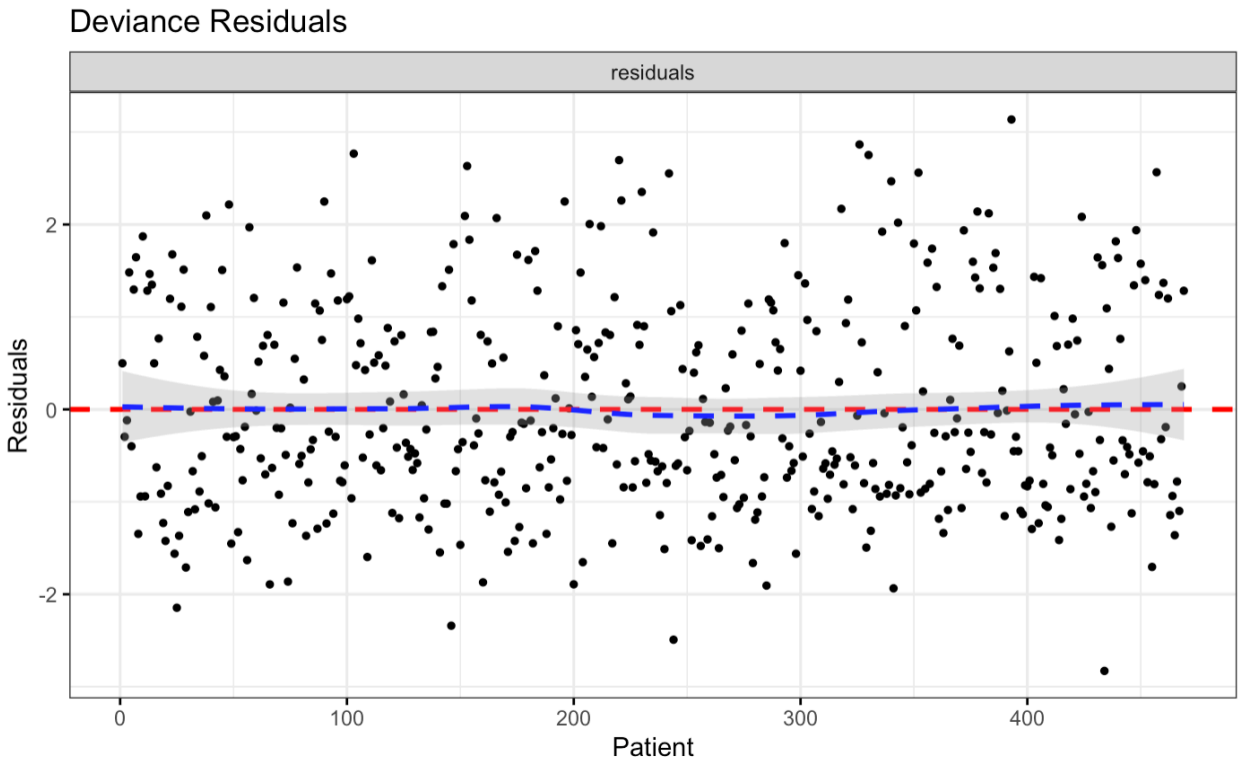


Figure Aiii. Deviance residuals for each patient before stratification of tumor stage and treatment. Positive deviance residuals correspond to patients that died before the expected survival time, while the negative deviance residuals correspond to patients that lived beyond the expected survival time.

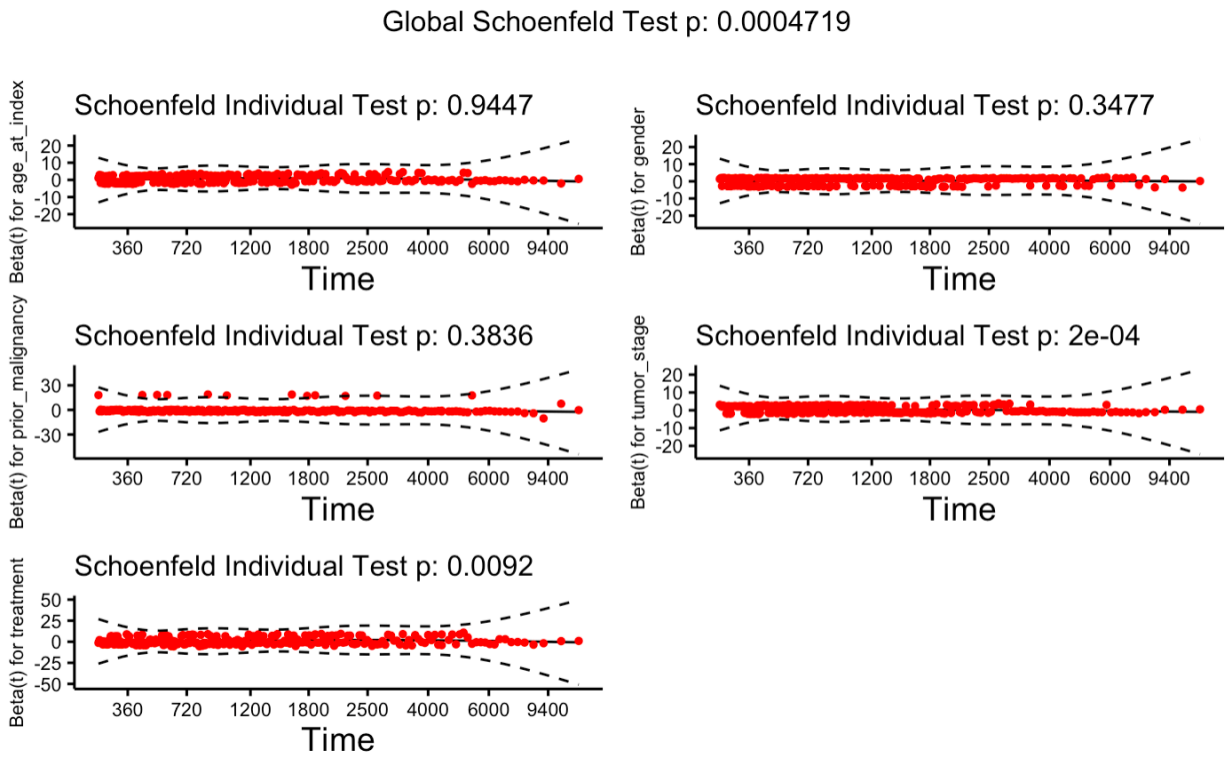


Figure Aiv. Schoenfeld residuals over time for each covariate before stratification of tumor stage and treatment. Departures from a horizontal line or signs of a trend suggest violation of proportional hazard assumption.

Table Aii. χ^2 test of independence of Schoenfeld residuals and time for each covariate after stratifying tumor stage and treatment. There are no nonproportional effects.

	χ^2 Test of Independence
Age at Index	χ_1^2 = 0.188 p-value = 0.66
Gender	χ_1^2 = 0.308 p-value = 0.58
Prior Malignancy	χ_1^2 = 0.509 p-value = 0.48
Global	χ_3^2 = 0.89 p-value = 0.83

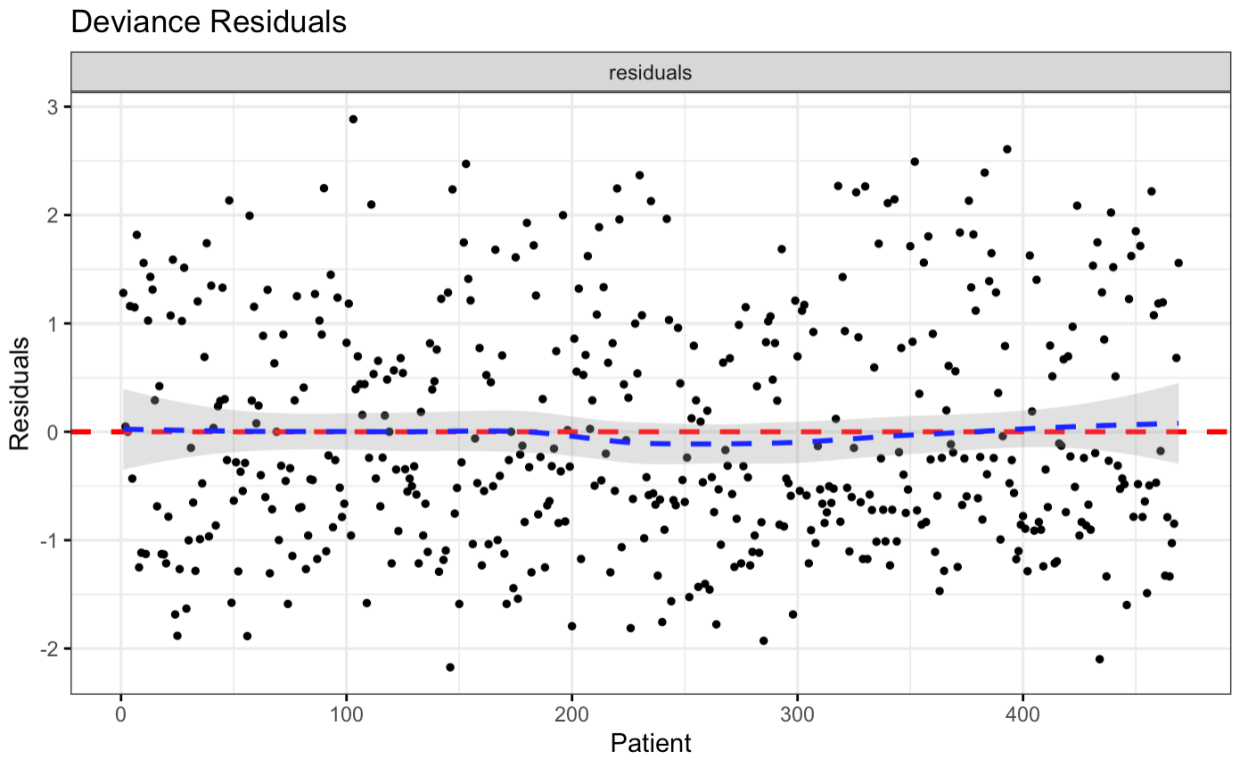


Figure Av. Deviance residuals for each patient after stratification of tumor stage and treatment. Positive deviance residuals correspond to patients that died before the expected survival time, while the negative deviance residuals correspond to patients that lived beyond the expected survival time.

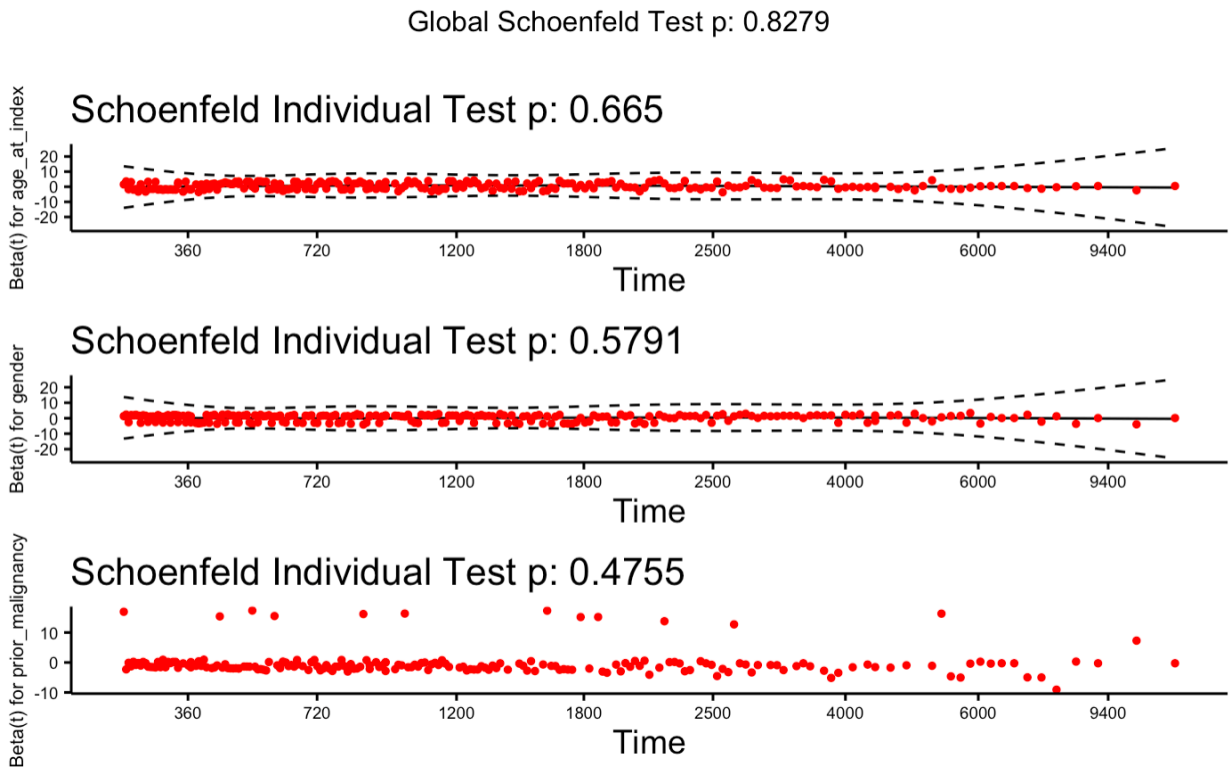


Figure Avi. Schoenfeld residuals over time for each covariate after stratification of tumor stage and treatment. Departures from a horizontal line or signs of a trend suggest violation of proportional hazard assumption.

Table A2. Results from Kaplan-Meier models including the number of observations in the data, the number of uncensored observations, and the median survival with an estimated 95% confidence interval. Covariates with large differences in median survival between levels are bolded.

Covariate	**Groups**	**Records**	**Events**	**Median Survival (Days) (95% Confidence Interval)**
Intercept		469	223	2270 (1960, 3136)
Age at Index	Less than 60	257	124	3195 (2421, 4601)
	Over 60	212	99	1832 (1487, 2071)
Gender	Female	180	75	2030 (1807, 4930)
	Male	289	148	2273 (1927, 3136)
Tumor Stage	Early Stages (Stage 0, I, I/II, II)	264	122	3379 (2273, 4222)
	Later Stages (Stage III and IV)	205	101	1832 (1124, 2192)
Treatment	None	215	88	2192 (1780, 5237)
	Radiation	64	35	3424 (2711, 6164)
	Pharmaceutical	107	45	3141 (1628, 5110)
	Both	83	55	1857 (1446, 2184)
Prior Malignancy	No	447	210	2273 (1960, 3139)
	Yes	22	13	2073 (1544, ∞)

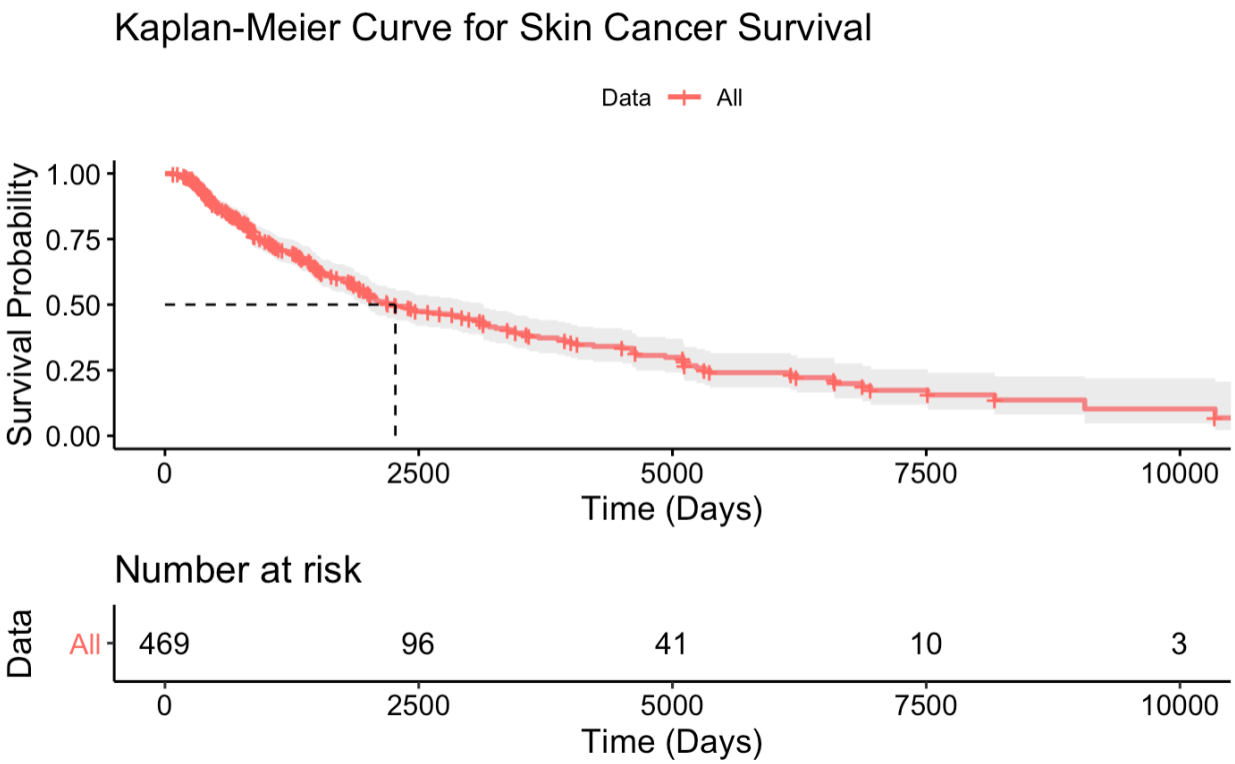


Figure Avii. Estimated survival probability over time computed from Kaplan-Meier model fitted with only the intercept. The dashed line indicates median survival.

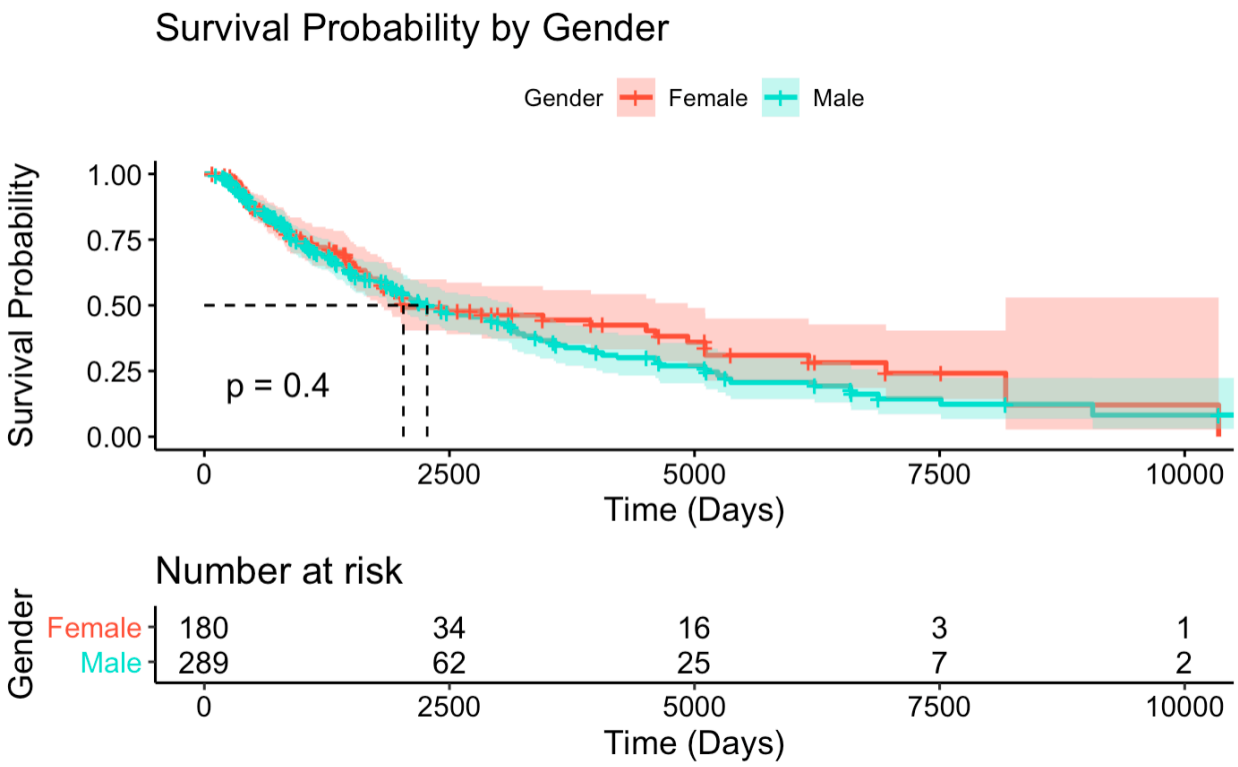


Figure Aviii. Estimated survival probability over time computed from the Kaplan-Meier model fitted with only gender. The dashed line indicates median survival.

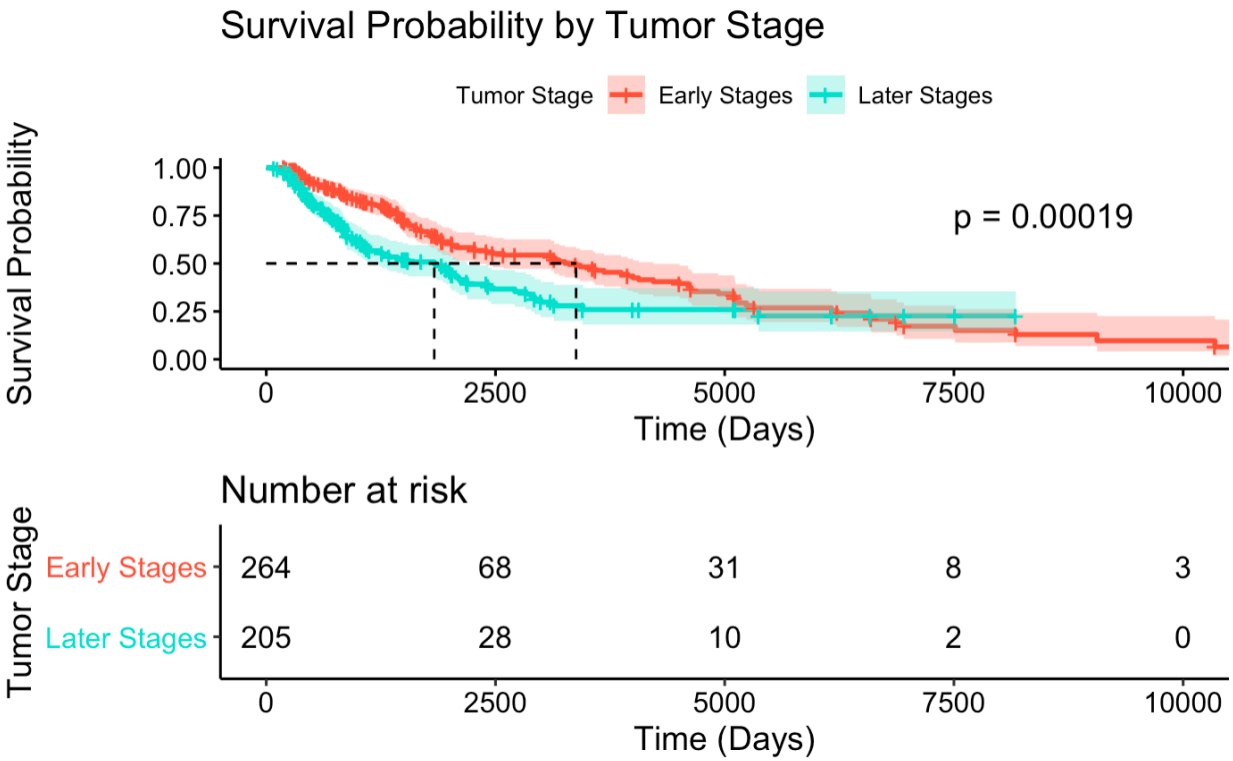


Figure Aix. Estimated survival probability over time computed from the Kaplan-Meier model fitted with only tumor stage. The dashed line indicates median survival.

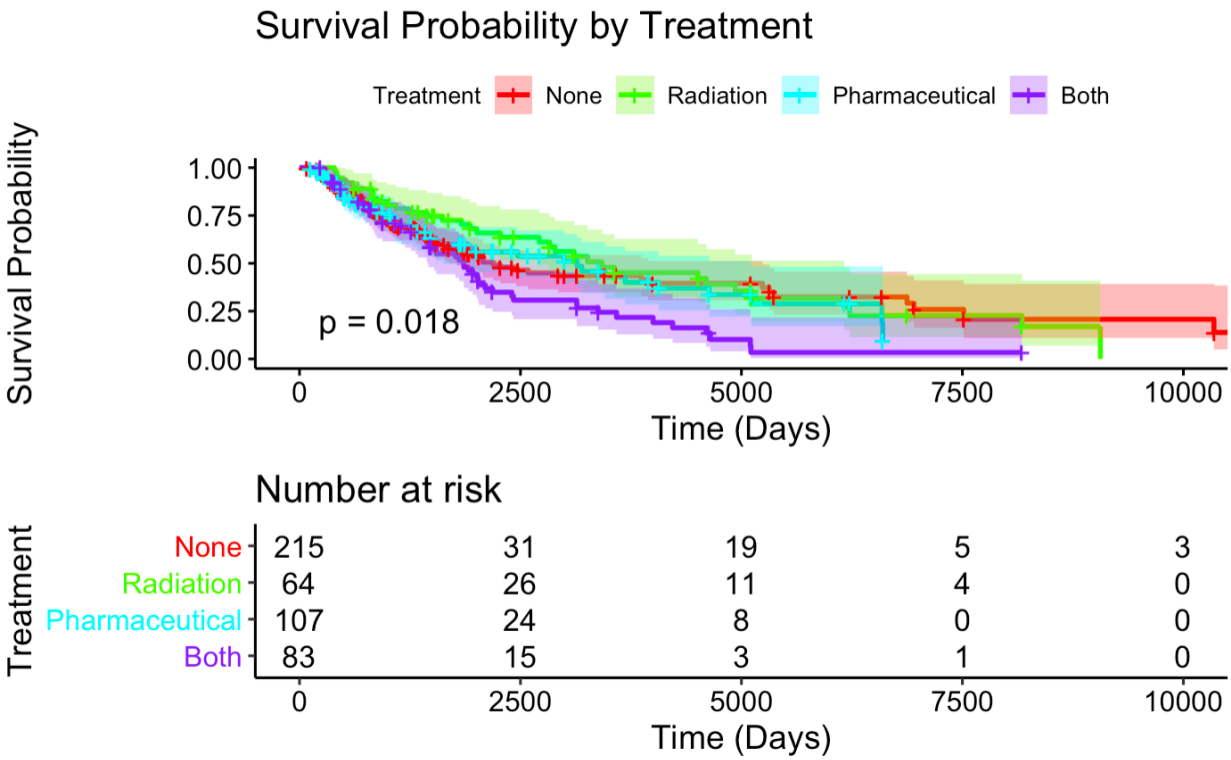


Figure Ax. Estimated survival probability over time computed from the Kaplan-Meier model fitted with only treatment.

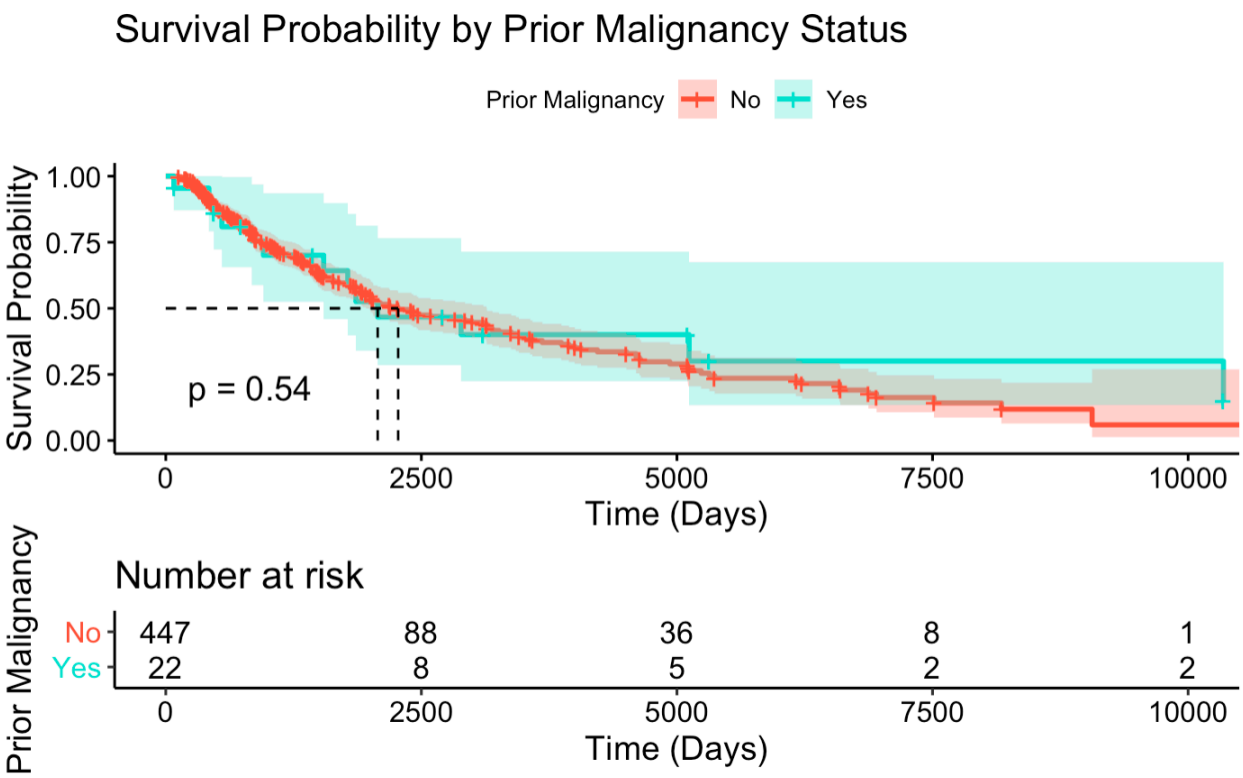


Figure Axi. Estimated survival probability over time computed from the Kaplan-Meier model fitted with only prior malignancy. The dashed line indicates median survival.

Table A3. Log-rank test results with significant differences detected in the bolded covariates.

Covariate	Log-Rank Test Results
Age at Index	$\chi_1^2 = 9.7$ p-value = 0.002
Gender	$\chi_1^2 = 0.7$ p-value = 0.4
Tumor Stage	$\chi_5^2 = 17.3$ p-value = 0.004
Treatment	$\chi_3^2 = 10$ p-value = 0.02
Prior Malignancy	$\chi_1^2 = 0.4$ p-value = 0.5

Table A4. Estimated hazard ratios of each covariate computed from the stratified Cox proportional hazard model with the censored days to death data as the response and all the covariates as predictors. The statistically significant predictors are bolded.

	Cox Proportional Hazard Model (Log-Rank Test: χ_3^2 = 10.55, p-value = 0.01)	
Covariate	**Hazard (95% Confidence Interval)**	**P-value**
Age at Index: Over 60 **(Base group: Less than 60)**	1.61 (1.1965, 2.164)	< 0.01
Gender: Male (Base group: Female)	1.12 (0.8395, 1.500)	0.44
Prior Malignancy: Yes (Base group: No prior malignancy)	0.78 (0.4260, 1.415)	0.41

R Code and Output

```
library(mice)

## Loading required package: lattice

##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind

library(survival)
library(ggplot2)
library(ggfortify)
library(ranger)
library(gridExtra)
library(survminer)

## Loading required package: ggpubr

## Loading required package: magrittr

# Data Preparation
exposure <- read.table("~/Desktop/STAT 4893W/exposure.tsv", sep = "\t", heade
r = TRUE)
exposure$bmi <- as.numeric(as.character(exposure$bmi))

## Warning: NAs introduced by coercion

clinical <- read.table("~/Desktop/STAT 4893W/clinical.tsv", sep = "\t", heade
r = TRUE)
skin <- clinical[,c("case_id", "age_at_index", "gender",
                    "primary_diagnosis", "treatment_or_therapy", "treatment_t
ype", "tumor_stage",
                    "prior_malignancy", "days_to_death", "vital_status")]

skin$days_to_death <- ifelse(skin$days_to_death == "--", NA, as.numeric(as.ch
aracter(skin$days_to_death)))

## Warning in ifelse(skin$days_to_death == "--", NA,
## as.numeric(as.character(skin$days_to_death))): NAs introduced by coercion

skin$age_at_index <- as.numeric(as.character(skin$age_at_index))

## Warning: NAs introduced by coercion

skin$tumor_stage <- ifelse(skin$tumor_stage == "i/ii nos", "i/ii",
                           ifelse(skin$tumor_stage == "not reported", NA,
                                  ifelse(skin$tumor_stage == "stage 0", "stag
e 0",
                                         ifelse(skin$tumor_stage == "stage i"
```

```
| skin$tumor_stage == "stage ia" | skin$tumor_stage == "stage ib", "stage i",
                                                    ifelse(skin$tumor_stage == "s
tage ii" | skin$tumor_stage == "stage iia" | skin$tumor_stage == "stage iib"
| skin$tumor_stage == "stage iic", "stage ii",
                                                           ifelse(skin$tumor_stag
e == "stage iii" | skin$tumor_stage == "stage iiia" | skin$tumor_stage == "st
age iiib" | skin$tumor_stage == "stage iiic", "stage iii",
                                                                  ifelse(skin$tum
or_stage == "stage iv", "stage iv", NA)))))))
skin$tumor_stage <- as.factor(skin$tumor_stage)

treatment_or_therapy.A <- as.character(skin$treatment_or_therapy[seq(1, lengt
h(skin$treatment_or_therapy), 2)])
treatment_or_therapy.B <- as.character(skin$treatment_or_therapy[seq(2, lengt
h(skin$treatment_or_therapy), 2)])
treatment_type.A <- as.character(skin$treatment_type[seq(1, length(skin$treat
ment_type), 2)])
treatment_type.B <- as.character(skin$treatment_type[seq(2, length(skin$treat
ment_type), 2)])

treatment.df <- data.frame(cbind(treatment_or_therapy.A, treatment_or_therapy
.B,
                                 treatment_type.A, treatment_type.B))

treatment <- ifelse(treatment_or_therapy.A == "yes" & treatment_or_therapy.B
== "yes", "Radiation and Pharmaceutical Therapy, NOS",
                    ifelse(treatment_or_therapy.A == "yes" & treatment_or_the
rapy.B == "no", treatment_type.A,
                           ifelse(treatment_or_therapy.A == "no" & treatment_
or_therapy.B == "yes", treatment_type.B,
                                  ifelse(treatment_or_therapy.A == "no" & tre
atment_or_therapy.B == "no", "None", NA))))

skin <- skin[seq(1, nrow(skin), 2),]

skin <- data.frame(cbind(skin[,c("case_id", "age_at_index",
                                 "gender", "primary_diagnosis", "tumor_stage"
)],
                         treatment,
                         skin[,c("prior_malignancy", "days_to_death", "vital_
status")]))

skin <- skin[-which(skin$vital_status == "Not Reported"),]
skin$vital_status <- droplevels(skin$vital_status)
skin <- data.frame((cbind(skin, bmi=exposure$bmi[-294])))

imp.days.to.death <- mice(skin[,-1], meth = c("norm.predict", "", "", "polr",
"polyreg", "", "rf", "", "norm.predict"), seed = 427)
```

```
##
##  iter imp variable
##   1   1  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   1   2  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   1   3  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   1   4  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   1   5  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   2   1  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   2   2  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   2   3  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   2   4  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   2   5  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   3   1  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   3   2  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   3   3  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   3   4  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   3   5  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   4   1  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   4   2  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   4   3  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   4   4  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   4   5  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   5   1  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   5   2  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   5   3  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   5   4  age_at_index  tumor_stage  treatment  days_to_death  bmi
##   5   5  age_at_index  tumor_stage  treatment  days_to_death  bmi

## Warning: Number of logged events: 25

imp.days.to.death <- complete(imp.days.to.death)

skin <- data.frame(cbind(skin$case_id,imp.days.to.death))
skin$vital_status <- ifelse(skin$vital_status == "Dead", 1, 0)
skin$age_at_index <- as.factor(ifelse(skin$age_at_index > 60, "Over 60", "Les
s than 60"))

# Reorder levels of factors
skin$tumor_stage <- factor(skin$tumor_stage, levels = c("stage 0", "stage i",
"i/ii",
                                                        "stage ii", "stage ii
i", "stage iv"))
skin$treatment <- factor(skin$treatment, levels = c("None", "Radiation Therap
y, NOS",
                                                    "Pharmaceutical Therapy,
NOS",
                                                    "Radiation and Pharmaceut
ical Therapy, NOS"))

# Group tumor stage 0 through stage II
```

```r
# Group tumor stage III and stage IV
skin$tumor_stage <- as.factor(ifelse(skin$tumor_stage == "stage 0" | skin$tum
or_stage == "stage i" | skin$tumor_stage == "i/ii" | skin$tumor_stage == "sta
ge ii", "Early Stages", "Later Stages"))

# Data Exploration

# All data
summary(skin)
```

```
##                                skin.case_id       age_at_index    gender
##  0153f141-625e-4623-9f8a-296678002c63:  1   Less than 60:257   female:180
##  015ba831-106b-4b84-9e8c-243a9eeeebf6:  1   Over 60     :212   male  :289
##  01ad975d-c2ed-4e4d-bd3b-c9512fc9073c:  1
##  01cb0004-fc1e-4da5-9d27-f458f8d711ee:  1
##  01cfbfae-f344-439d-aeab-a9e15d636325:  1
##  021a9431-7fdc-42b8-9ff0-e17ded79ff1d:  1
##  (Other)                             :463
##                       primary_diagnosis       tumor_stage
##  Malignant melanoma, NOS       :419     Early Stages:264
##  Nodular melanoma              : 21     Later Stages:205
##  Epithelioid cell melanoma     :  8
##  Amelanotic melanoma           :  7
##  Superficial spreading melanoma:  5
##  Spindle cell melanoma, NOS    :  3
##  (Other)                       :  6
##                                     treatment   prior_malignancy
##  None                                    :215   no :447
##  Radiation Therapy, NOS                  : 64   yes: 22
##  Pharmaceutical Therapy, NOS             :107
##  Radiation and Pharmaceutical Therapy, NOS: 83
##
##
##
##  days_to_death    vital_status         bmi
##  Min.   :   79   Min.   :0.0000   Min.   :17.63
##  1st Qu.:  519   1st Qu.:0.0000   1st Qu.:25.73
##  Median : 1078   Median :0.0000   Median :27.54
##  Mean   : 1777   Mean   :0.4755   Mean   :28.02
##  3rd Qu.: 2030   3rd Qu.:1.0000   3rd Qu.:30.11
##  Max.   :10870   Max.   :1.0000   Max.   :55.47
##
```

```r
# Dead
summary(skin[skin$vital_status==1,])
```

```
##                                skin.case_id       age_at_index    gender
##  0153f141-625e-4623-9f8a-296678002c63:  1   Less than 60:124   female: 75
##  015ba831-106b-4b84-9e8c-243a9eeeebf6:  1   Over 60     : 99   male  :148
##  01ad975d-c2ed-4e4d-bd3b-c9512fc9073c:  1
```

```
##  01cb0004-fc1e-4da5-9d27-f458f8d711ee:  1
##  01cfbfae-f344-439d-aeab-a9e15d636325:  1
##  05f4058d-3066-4d16-8320-cf92f122945f:  1
##  (Other)                              :217
##                               primary_diagnosis       tumor_stage
##  Malignant melanoma, NOS              :207     Early Stages:122
##  Nodular melanoma                     :  7     Later Stages:101
##  Amelanotic melanoma                  :  3
##  Lentigo maligna melanoma             :  2
##  Acral lentiginous melanoma, malignant:  1
##  Desmoplastic melanoma, malignant     :  1
##  (Other)                              :  2
##                                     treatment  prior_malignancy
##  None                                    :88   no :210
##  Radiation Therapy, NOS                  :35   yes: 13
##  Pharmaceutical Therapy, NOS             :45
##  Radiation and Pharmaceutical Therapy, NOS:55
##
##
##
##  days_to_death      vital_status      bmi
##  Min.   :   79.0   Min.   :1     Min.   :17.78
##  1st Qu.:  518.5   1st Qu.:1     1st Qu.:25.97
##  Median : 1093.0   Median :1     Median :27.83
##  Mean   : 1793.6   Mean   :1     Mean   :28.14
##  3rd Qu.: 2087.0   3rd Qu.:1     3rd Qu.:30.06
##  Max.   :10870.0   Max.   :1     Max.   :49.15
##
```

```
sd(skin[skin$vital_status==1,]$bmi)
```

```
## [1] 4.155457
```

```
# Censored
summary(skin[skin$vital_status==0,])
```

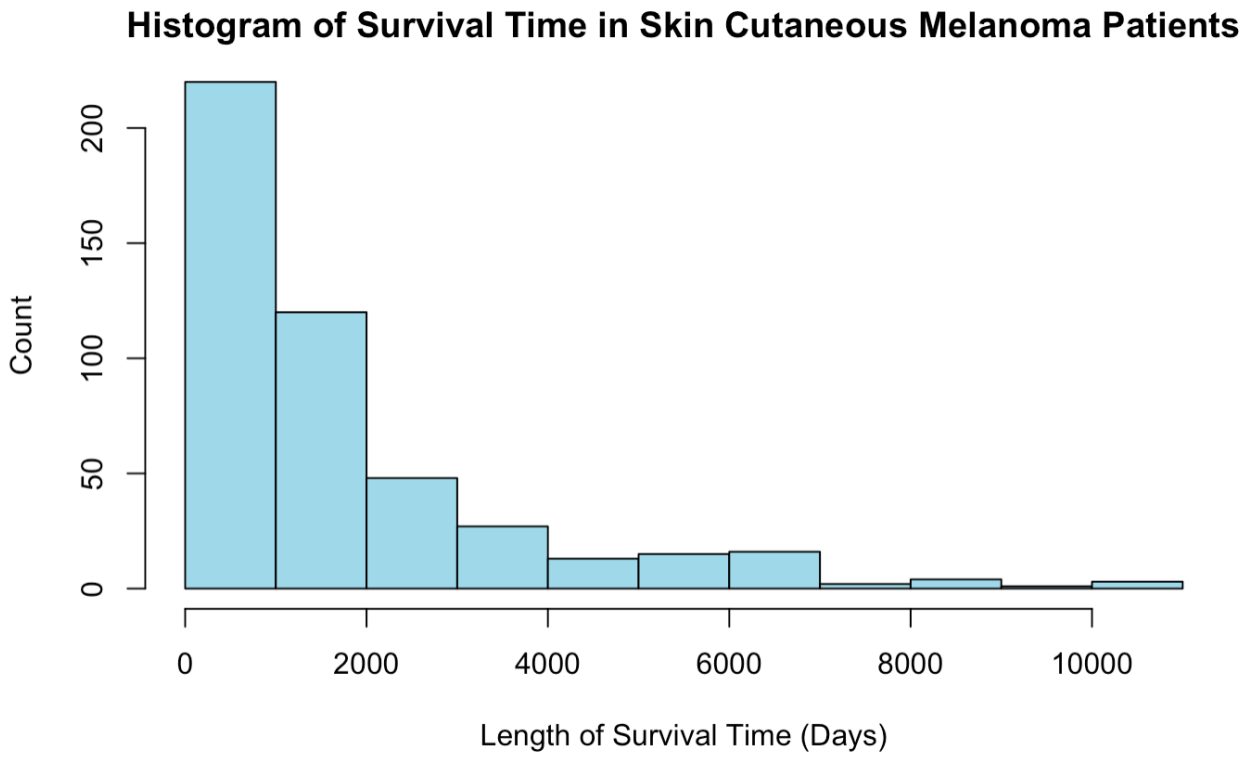
```
##                                skin.case_id       age_at_index    gender
##  021a9431-7fdc-42b8-9ff0-e17ded79ff1d:  1   Less than 60:133   female:105
##  021d32b0-a94a-4dad-95f4-8eb0abd894bf:  1   Over 60     :113   male  :141
##  04add7f0-d212-486e-ae70-d41d8112f523:  1
##  04c3d01d-a949-4dc7-9829-a8c99180dba0:  1
##  04f7ecc8-4f6b-41ac-a984-9ea604698a21:  1
##  080bb1f8-f20c-4daa-9de3-b5782b69583c:  1
##  (Other)                             :240
##                       primary_diagnosis       tumor_stage
##  Malignant melanoma, NOS       :212     Early Stages:142
##  Nodular melanoma              : 14     Later Stages:104
##  Epithelioid cell melanoma     :  8
##  Amelanotic melanoma           :  4
##  Superficial spreading melanoma:  4
##  Spindle cell melanoma, NOS    :  2
```

```
##  (Other)                          :  2
##                                      treatment   prior_malignancy
##  None                                     :127   no :237
##  Radiation Therapy, NOS                   : 29   yes:  9
##  Pharmaceutical Therapy, NOS              : 62
##  Radiation and Pharmaceutical Therapy, NOS: 28
##
##
##
##  days_to_death      vital_status      bmi
##  Min.   :   79.0   Min.   :0     Min.   :17.63
##  1st Qu.:  530.8   1st Qu.:0     1st Qu.:25.26
##  Median : 1078.0   Median :0     Median :27.34
##  Mean   : 1761.5   Mean   :0     Mean   :27.91
##  3rd Qu.: 2026.5   3rd Qu.:0     3rd Qu.:30.12
##  Max.   :10346.0   Max.   :0     Max.   :55.47
##
```

```
sd(skin[skin$vital_status==0,]$bmi)
```

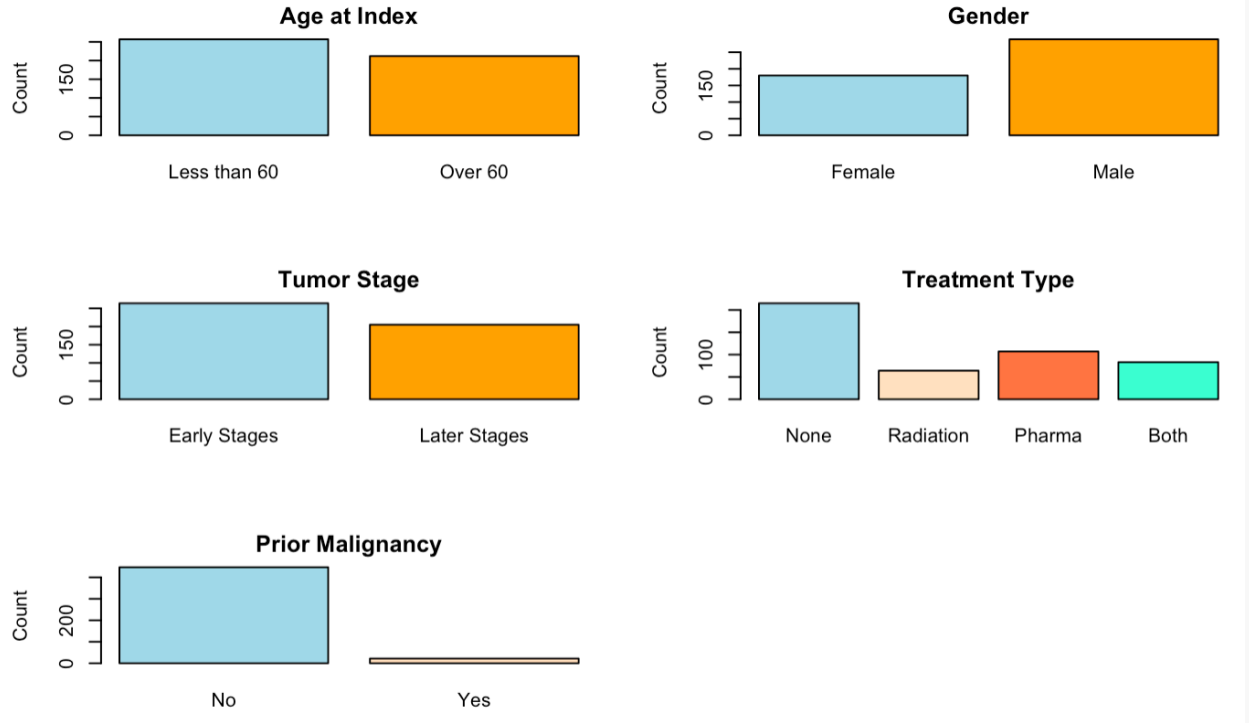
```
## [1] 5.101129
```

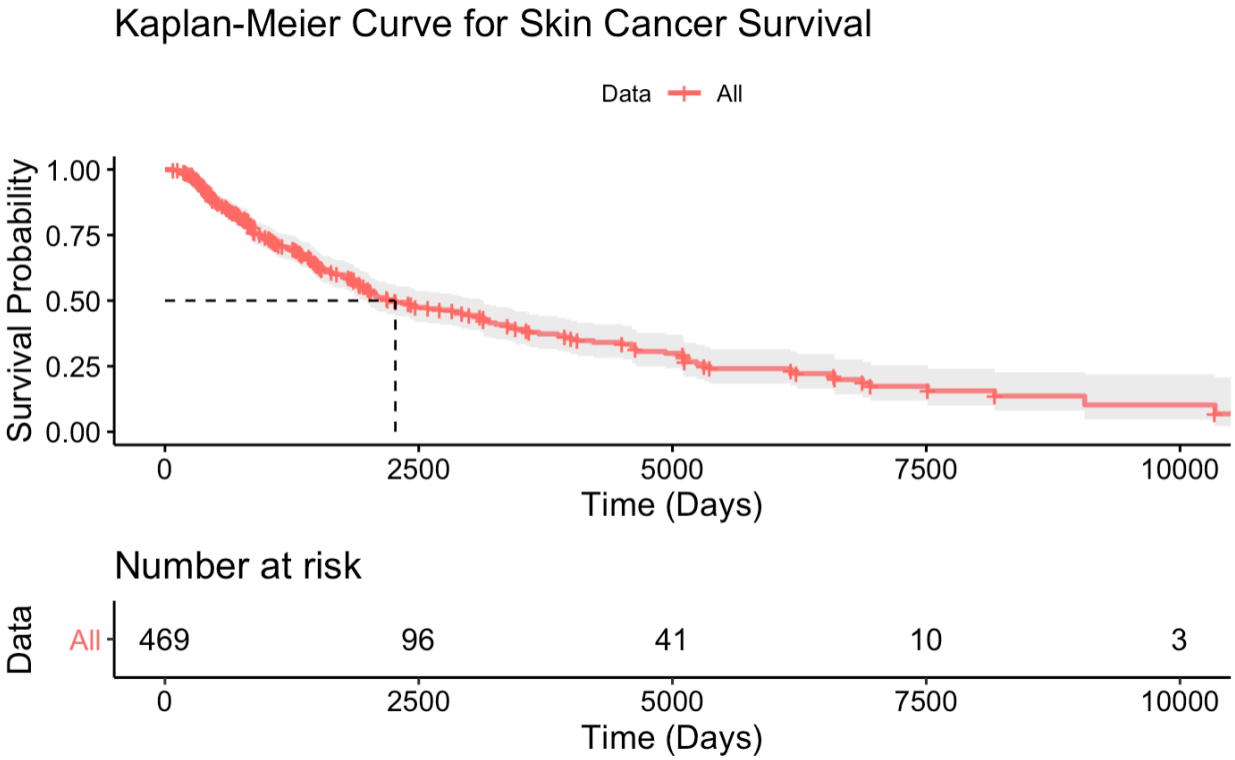
```
# Distribution of survival time
hist(skin$days_to_death, xlab = "Length of Survival Time (Days)", ylab = "Cou
nt", col = "lightblue",
     main = "Histogram of Survival Time in Skin Cutaneous Melanoma Patients")
```



```r
# Distribution of predictors
par(mfrow=c(3,2))
barplot(table(skin$age_at_index), col=c("lightblue", "orange"),
        ylab = "Count", main = "Age at Index")
barplot(table(skin$gender), col=c("lightblue", "orange"), names.arg = c("Fema
le", "Male"),
        ylab = "Count", main = "Gender")
barplot(table(skin$tumor_stage), col=c("lightblue", "orange"),
        ylab = "Count", main = "Tumor Stage")
barplot(table(skin$treatment), col=c("lightblue","bisque","coral","aquamarine
"),
        ylab = "Count", main = "Treatment Type",
        names.arg = c("None", "Radiation", "Pharma", "Both"))
barplot(table(skin$prior_malignancy), col=c("lightblue","bisque"), names.arg
= c("No", "Yes"),
        ylab = "Count", main = "Prior Malignancy")
hist(skin$bmi, xlab = "Body Mass Index", main = "Body Mass Index", col = "lig
htblue")
abline(v = median(skin$bmi), col = "red", lwd = 2, lty = 2)
legend(x = 40, y = 200, col = "red", lwd = 2, lty = 2, "Median")

# Kaplan-Meier Curve Estimations
km_fit <- survfit(Surv(days_to_death, vital_status) ~ 1, data = skin)
km_fit

## Call: survfit(formula = Surv(days_to_death, vital_status) ~ 1, data = skin
)
##
##       n  events  median 0.95LCL 0.95UCL
##     469     223    2270    1960    3136

ggsurvplot(km_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE, legend.title =
"Data",
           title="Kaplan-Meier Curve for Skin Cancer Survival",
           risk.table.height=.3, xlab = "Time (Days)", ylab = "Survival Proba
bility",
           surv.median.line = "hv")

## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.co
ord = pval.coord, : There are no survival curves to be compared.
```

```
##  This is a null model.
```






```
km_age_fit <- survfit(Surv(days_to_death, vital_status) ~ age_at_index, data
= skin)
km_age_fit
```

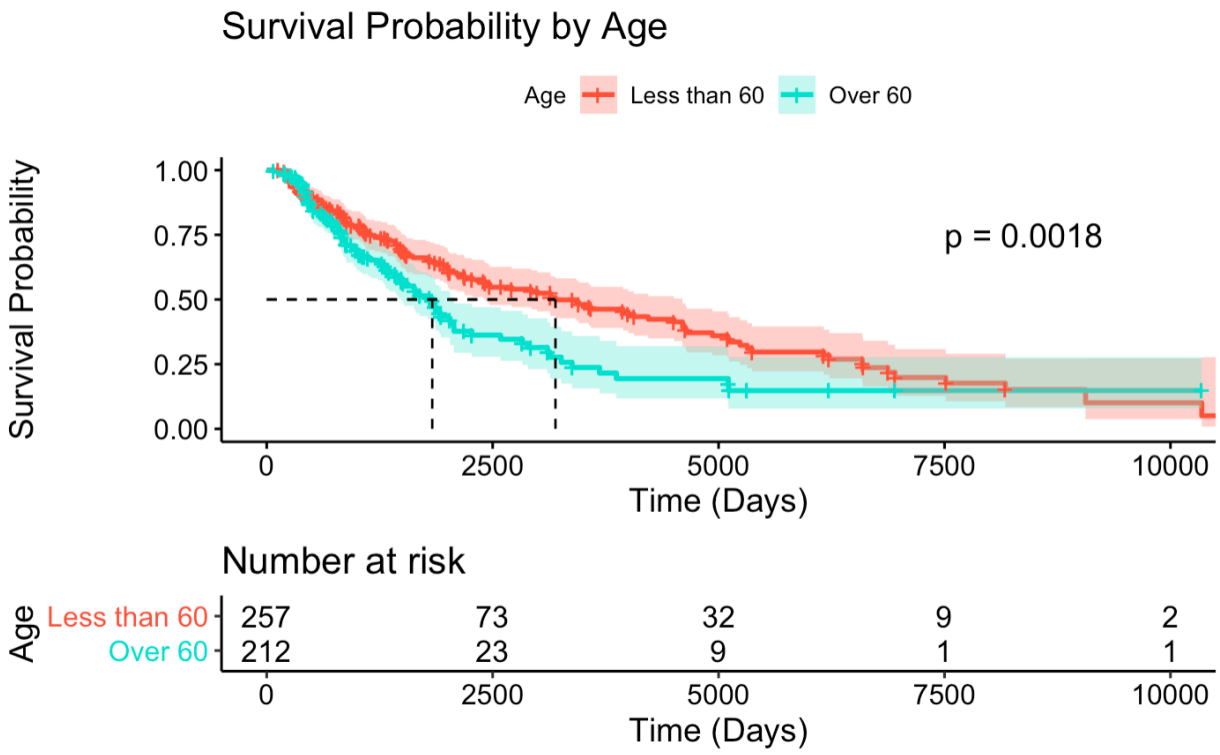
```
## Call: survfit(formula = Surv(days_to_death, vital_status) ~ age_at_index,
##     data = skin)
##
##                              n events median 0.95LCL 0.95UCL
## age_at_index=Less than 60 257    124   3195    2421    4601
## age_at_index=Over 60      212     99   1832    1487    2071
```

```
survdiff(Surv(days_to_death, vital_status) ~ age_at_index, data = skin)
```

```
## Call:
## survdiff(formula = Surv(days_to_death, vital_status) ~ age_at_index,
##     data = skin)
##
##                              N Observed Expected (O-E)^2/E (O-E)^2/V
## age_at_index=Less than 60 257      124    145.6      3.22      9.75
## age_at_index=Over 60      212       99     77.4      6.05      9.75
##
##  Chisq= 9.7  on 1 degrees of freedom, p= 0.002
```

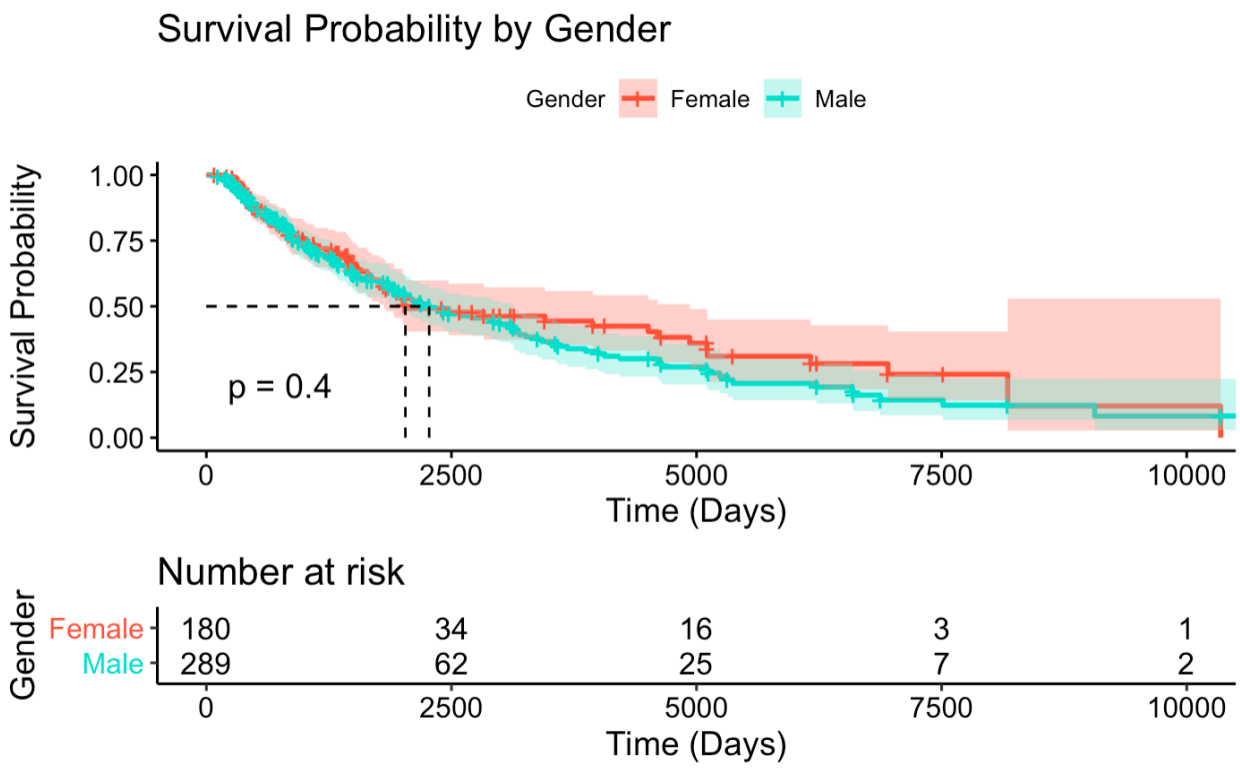
```
ggsurvplot(km_age_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs = c("Less than 60", "Over 60"), legend.title="Age",
           palette=c("tomato", "turquoise"),
           title="Survival Probability by Age",
           risk.table.height=.3, xlab = "Time (Days)", ylab = "Survival Proba
bility",
           surv.median.line = "hv", pval.coord = c(7500,0.75))
```

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```




```
km_gender_fit <- survfit(Surv(days_to_death, vital_status) ~ gender, data = s
kin)
km_gender_fit

## Call: survfit(formula = Surv(days_to_death, vital_status) ~ gender,
##     data = skin)
##
##                  n events median 0.95LCL 0.95UCL
## gender=female 180     75   2030    1807    4930
## gender=male   289    148   2273    1927    3136

survdiff(Surv(days_to_death, vital_status) ~ gender, data = skin)

## Call:
## survdiff(formula = Surv(days_to_death, vital_status) ~ gender,
##     data = skin)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=female 180       75       81     0.446     0.704
## gender=male   289      148      142     0.255     0.704
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.4

ggsurvplot(km_gender_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs = c("Female", "Male"), legend.title="Gender",
           palette=c("tomato", "turquoise"),
           title="Survival Probability by Gender",
           risk.table.height=.3, xlab = "Time (Days)", ylab = "Survival Proba
```

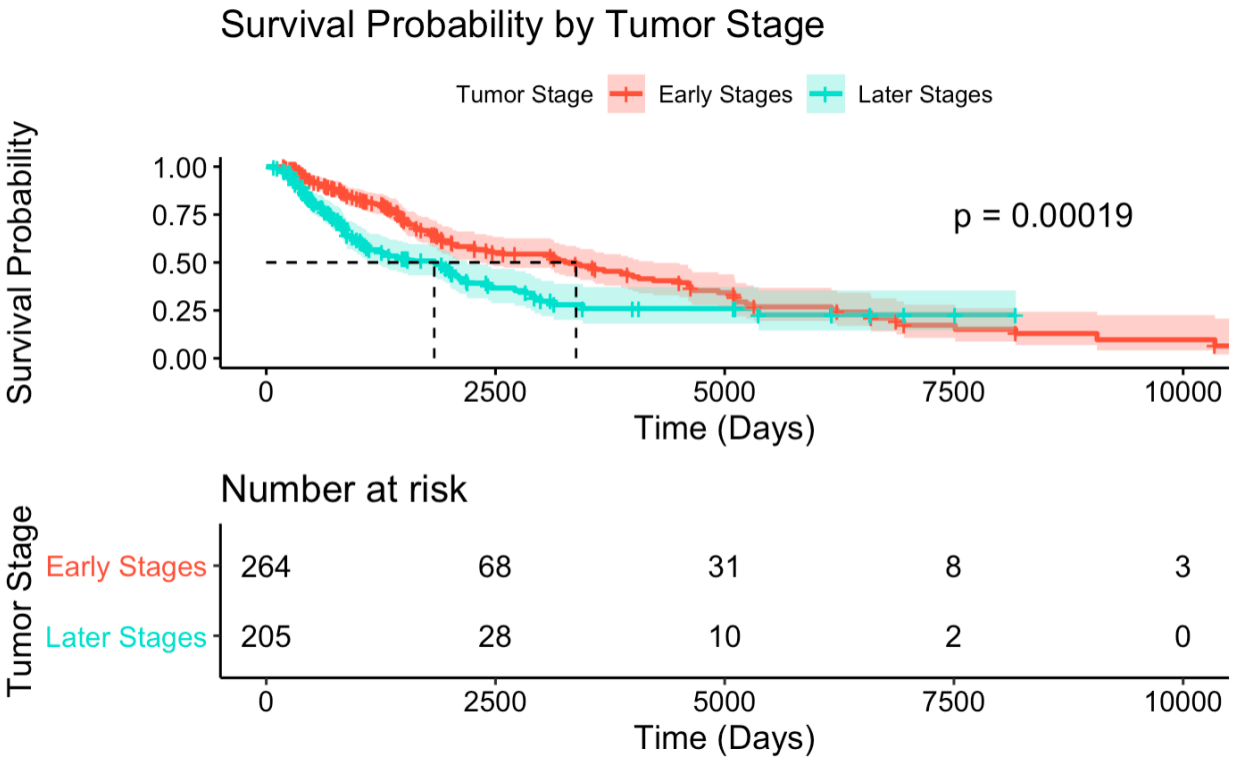
```
bility",
           surv.median.line = "hv")

## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```




```
km_stage_fit <- survfit(Surv(days_to_death, vital_status) ~ tumor_stage, data
= skin)
km_stage_fit

## Call: survfit(formula = Surv(days_to_death, vital_status) ~ tumor_stage,
##     data = skin)
##
##                             n events median 0.95LCL 0.95UCL
## tumor_stage=Early Stages 264    122   3379    2273    4222
## tumor_stage=Later Stages 205    101   1832    1124    2192

survdiff(Surv(days_to_death, vital_status) ~ tumor_stage, data = skin)

## Call:
## survdiff(formula = Surv(days_to_death, vital_status) ~ tumor_stage,
##     data = skin)
##
##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## tumor_stage=Early Stages 264      122    147.9      4.55      13.9
## tumor_stage=Later Stages 205      101     75.1      8.97      13.9
##
##  Chisq= 13.9  on 1 degrees of freedom, p= 2e-04
```

```
ggsurvplot(km_stage_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs = c("Early Stages", "Later Stages"),
           legend.title="Tumor Stage",
           palette=c("tomato", "turquoise"),
           title="Survival Probability by Tumor Stage",
           risk.table.height=.4, xlab = "Time (Days)", ylab = "Survival Proba
bility",
           surv.median.line = "hv", pval.coord = c(7500,0.75))
```

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```




```
km_trt_fit <- survfit(Surv(days_to_death, vital_status) ~ treatment, data = s
kin)
km_trt_fit
```

```
## Call: survfit(formula = Surv(days_to_death, vital_status) ~ treatment,
##     data = skin)
##
##                                                    n events median 0.95
LCL
## treatment=None                                   215     88   2192    1
780
## treatment=Radiation Therapy, NOS                  64     35   3424    2
711
## treatment=Pharmaceutical Therapy, NOS            107     45   3141    1
628
## treatment=Radiation and Pharmaceutical Therapy, NOS  83     55   1857    1
```

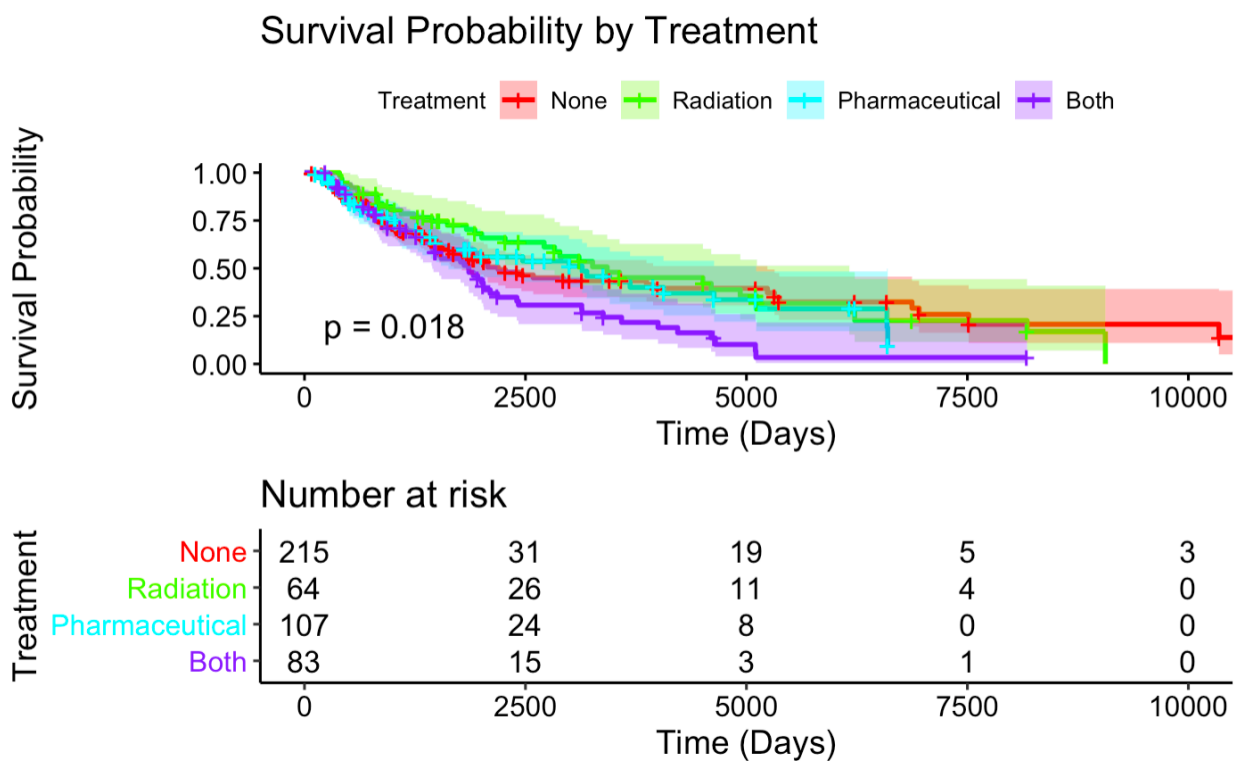
```
446
##                                                    0.95UCL
## treatment=None                                        5237
## treatment=Radiation Therapy, NOS                      6164
## treatment=Pharmaceutical Therapy, NOS                 5110
## treatment=Radiation and Pharmaceutical Therapy, NOS   2184
```

```
survdiff(Surv(days_to_death, vital_status) ~ treatment, data = skin)
```

```
## Call:
## survdiff(formula = Surv(days_to_death, vital_status) ~ treatment,
##     data = skin)
##
##                                                     N Observed Expected
## treatment=None                                    215       88     93.7
## treatment=Radiation Therapy, NOS                   64       35     44.3
## treatment=Pharmaceutical Therapy, NOS             107       45     46.8
## treatment=Radiation and Pharmaceutical Therapy, NOS  83       55     38.1
##                                                   (O-E)^2/E (O-E)^2/V
## treatment=None                                       0.3482    0.6160
## treatment=Radiation Therapy, NOS                     1.9692    2.5027
## treatment=Pharmaceutical Therapy, NOS                0.0711    0.0909
## treatment=Radiation and Pharmaceutical Therapy, NOS  7.4759    9.1377
##
##  Chisq= 10  on 3 degrees of freedom, p= 0.02
```

```
ggsurvplot(km_trt_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs = c("None", "Radiation",
                           "Pharmaceutical", "Both"),
           legend.title="Treatment",
           palette=rainbow(4),
           title="Survival Probability by Treatment",
           risk.table.height=.4, xlab = "Time (Days)", ylab = "Survival Proba
bility")
```

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```




```
km_prior_fit <- survfit(Surv(days_to_death, vital_status) ~ prior_malignancy,
data = skin)
km_prior_fit
```

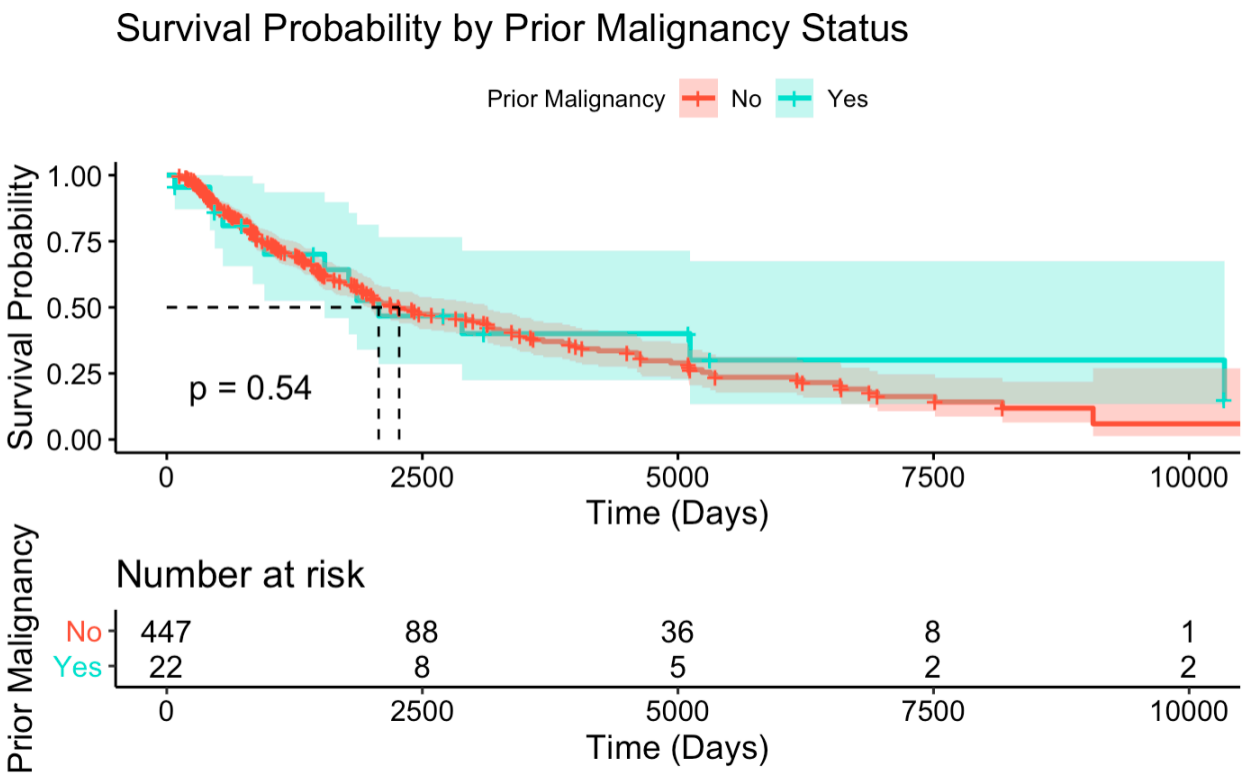
```
## Call: survfit(formula = Surv(days_to_death, vital_status) ~ prior_malignan
cy,
##     data = skin)
##
##                          n events median 0.95LCL 0.95UCL
## prior_malignancy=no  447    210   2273    1960    3139
## prior_malignancy=yes  22     13   2073    1544      NA
```

```
survdiff(Surv(days_to_death, vital_status) ~ prior_malignancy, data = skin)
```

```
## Call:
## survdiff(formula = Surv(days_to_death, vital_status) ~ prior_malignancy,
##     data = skin)
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## prior_malignancy=no  447      210    207.8    0.0239     0.368
## prior_malignancy=yes  22       13     15.2    0.3264     0.368
##
##  Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

```
ggsurvplot(km_prior_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs = c("No", "Yes"), legend.title="Prior Malignancy",
           palette=c("tomato", "turquoise"),
           title="Survival Probability by Prior Malignancy Status",
```

```
           risk.table.height=.3, xlab = "Time (Days)", ylab = "Survival Proba
bility",
           surv.median.line = "hv")

## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```

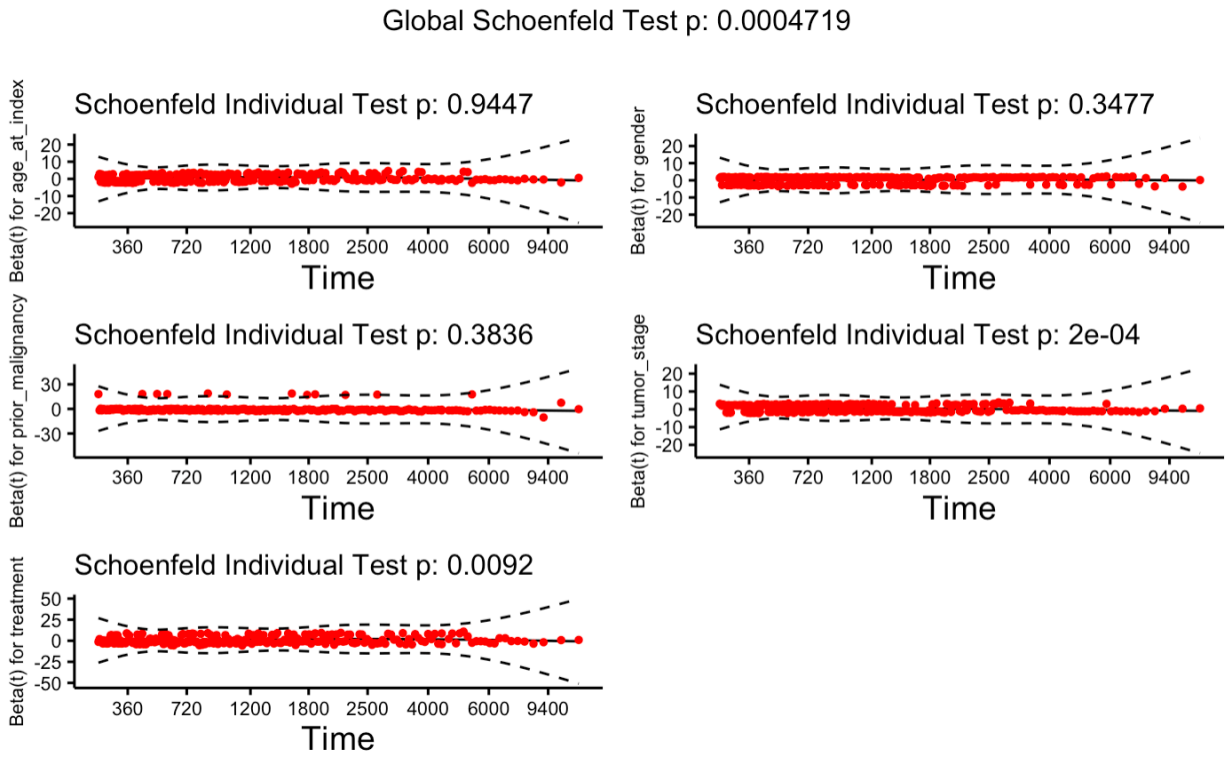



```
# Cox Proportional Hazard Model
cox <- coxph(Surv(days_to_death, vital_status) ~ age_at_index + gender + prio
r_malignancy +
               tumor_stage + treatment, data = skin)
summary(cox)

## Call:
## coxph(formula = Surv(days_to_death, vital_status) ~ age_at_index +
##     gender + prior_malignancy + tumor_stage + treatment, data = skin)
##
##   n= 469, number of events= 223
##
##                                                    coef exp(coef) se(co
ef)
## age_at_indexOver 60                              0.5733    1.7741   0.1
441
## gendermale                                       0.1566    1.1696   0.1
436
## prior_malignancyyes                             -0.2444    0.7832   0.2
996
```

```
## tumor_stageLater Stages                                0.5358    1.7089   0.1
384
## treatmentRadiation Therapy, NOS                       -0.1313    0.8769   0.2
015
## treatmentPharmaceutical Therapy, NOS                   0.1961    1.2167   0.1
880
## treatmentRadiation and Pharmaceutical Therapy, NOS     0.5218    1.6850   0.1
767
##                                                            z Pr(>|z|)
## age_at_indexOver 60                                    3.979 6.93e-05 ***
## gendermale                                             1.091 0.275360
## prior_malignancyyes                                   -0.816 0.414660
## tumor_stageLater Stages                                3.872 0.000108 ***
## treatmentRadiation Therapy, NOS                       -0.652 0.514558
## treatmentPharmaceutical Therapy, NOS                   1.043 0.296734
## treatmentRadiation and Pharmaceutical Therapy, NOS     2.953 0.003142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                       exp(coef) exp(-coef)
## age_at_indexOver 60                                      1.7741     0.5637
## gendermale                                               1.1696     0.8550
## prior_malignancyyes                                      0.7832     1.2768
## tumor_stageLater Stages                                  1.7089     0.5852
## treatmentRadiation Therapy, NOS                          0.8769     1.1403
## treatmentPharmaceutical Therapy, NOS                     1.2167     0.8219
## treatmentRadiation and Pharmaceutical Therapy, NOS       1.6850     0.5935
##                                                       lower .95 upper .95
## age_at_indexOver 60                                      1.3376     2.353
## gendermale                                               0.8827     1.550
## prior_malignancyyes                                      0.4353     1.409
## tumor_stageLater Stages                                  1.3029     2.241
## treatmentRadiation Therapy, NOS                          0.5909     1.302
## treatmentPharmaceutical Therapy, NOS                     0.8417     1.759
## treatmentRadiation and Pharmaceutical Therapy, NOS       1.1919     2.382
##
## Concordance= 0.627  (se = 0.021 )
## Likelihood ratio test= 38.05  on 7 df,   p=3e-06
## Wald test            = 39.07  on 7 df,   p=2e-06
## Score (logrank) test = 39.66  on 7 df,   p=1e-06

cox_fit <- survfit(cox)

# Schoenfeld residuals - Proportional hazards assumption
test.ph <- cox.zph(cox)
test.ph

##                    chisq df       p
## age_at_index     0.00481  1 0.94470
## gender           0.88194  1 0.34767
```

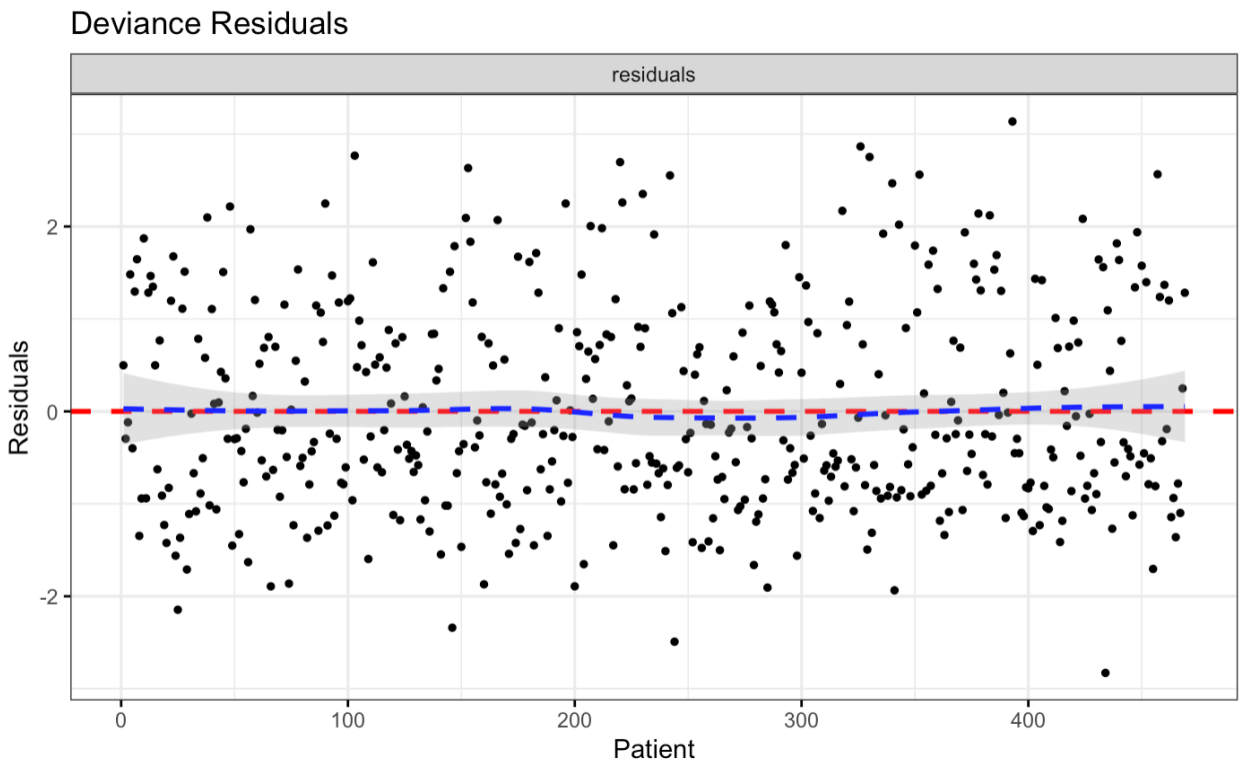
```
## prior_malignancy  0.75924  1 0.38357
## tumor_stage      13.43347  1 0.00025
## treatment        11.52062  3 0.00922
## GLOBAL           26.15812  7 0.00047
```

```
ggcoxzph(test.ph)
```




```
# Deviance residuals - Influential observations or outliers
ggcoxdiagnostics(cox, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```

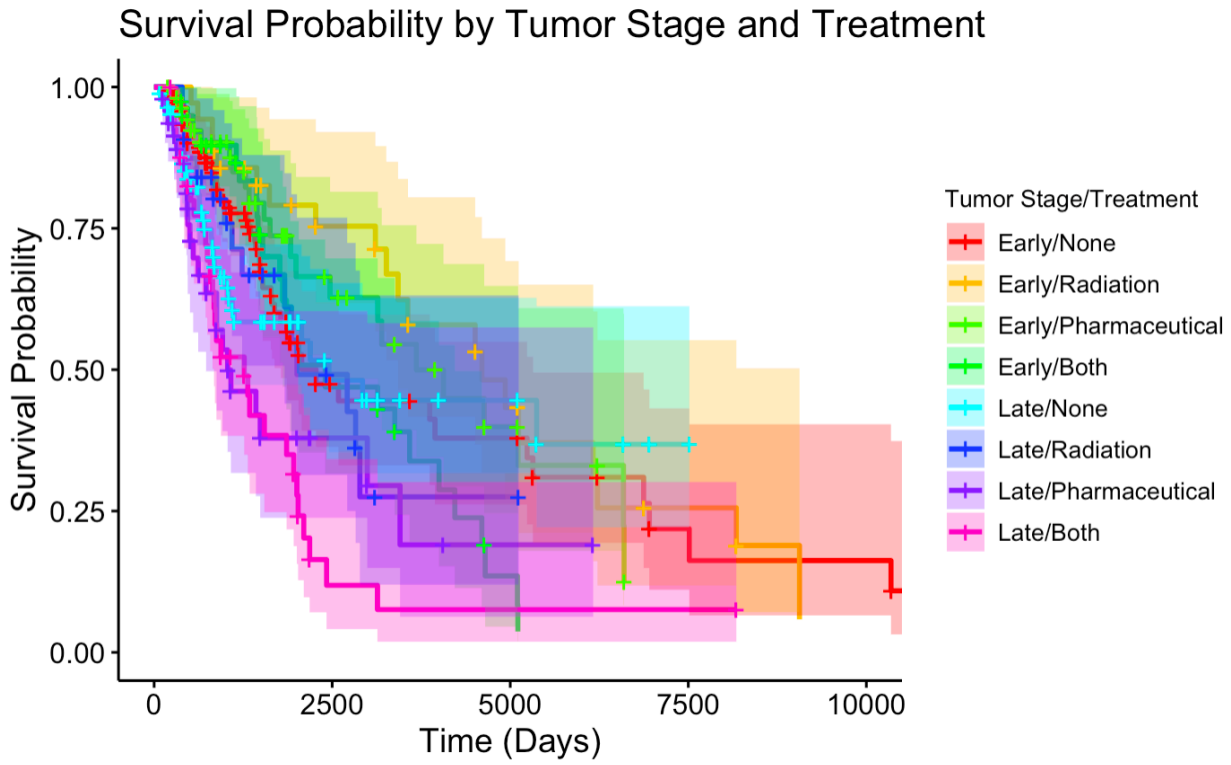
```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Cox Proportional Hazard Model With Stratification
cox <- coxph(Surv(days_to_death, vital_status) ~ age_at_index + gender + prio
r_malignancy +
               strata(tumor_stage, treatment), data = skin)
summary(cox)

## Call:
## coxph(formula = Surv(days_to_death, vital_status) ~ age_at_index +
##     gender + prior_malignancy + strata(tumor_stage, treatment),
##     data = skin)
##
##   n= 469, number of events= 223
##
##                         coef exp(coef) se(coef)      z Pr(>|z|)
## age_at_indexOver 60  0.4756    1.6090   0.1511  3.147  0.00165 **
## gendermale           0.1152    1.1221   0.1481  0.778  0.43661
## prior_malignancyyes -0.2532    0.7763   0.3062 -0.827  0.40830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## age_at_indexOver 60    1.6090     0.6215    1.1965     2.164
## gendermale             1.1221     0.8912    0.8395     1.500
## prior_malignancyyes    0.7763     1.2882    0.4260     1.415
##
## Concordance= 0.554  (se = 0.024 )
## Likelihood ratio test= 10.3  on 3 df,   p=0.02
```

```
## Wald test            = 10.44  on 3 df,   p=0.02
## Score (logrank) test = 10.55  on 3 df,   p=0.01
```

```
cox_fit <- surv_fit(cox, data = skin)
ggsurvplot(cox_fit, conf.int=TRUE,
           legend.labs = c("Early/None", "Early/Radiation",
                           "Early/Pharmaceutical", "Early/Both",
                           "Late/None", "Late/Radiation",
                           "Late/Pharmaceutical", "Late/Both"),
           legend.title="Tumor Stage/Treatment",
           palette=rainbow(8),
           title="Survival Probability by Tumor Stage and Treatment",
           xlab = "Time (Days)", ylab = "Survival Probability",
           legend = "right")
```



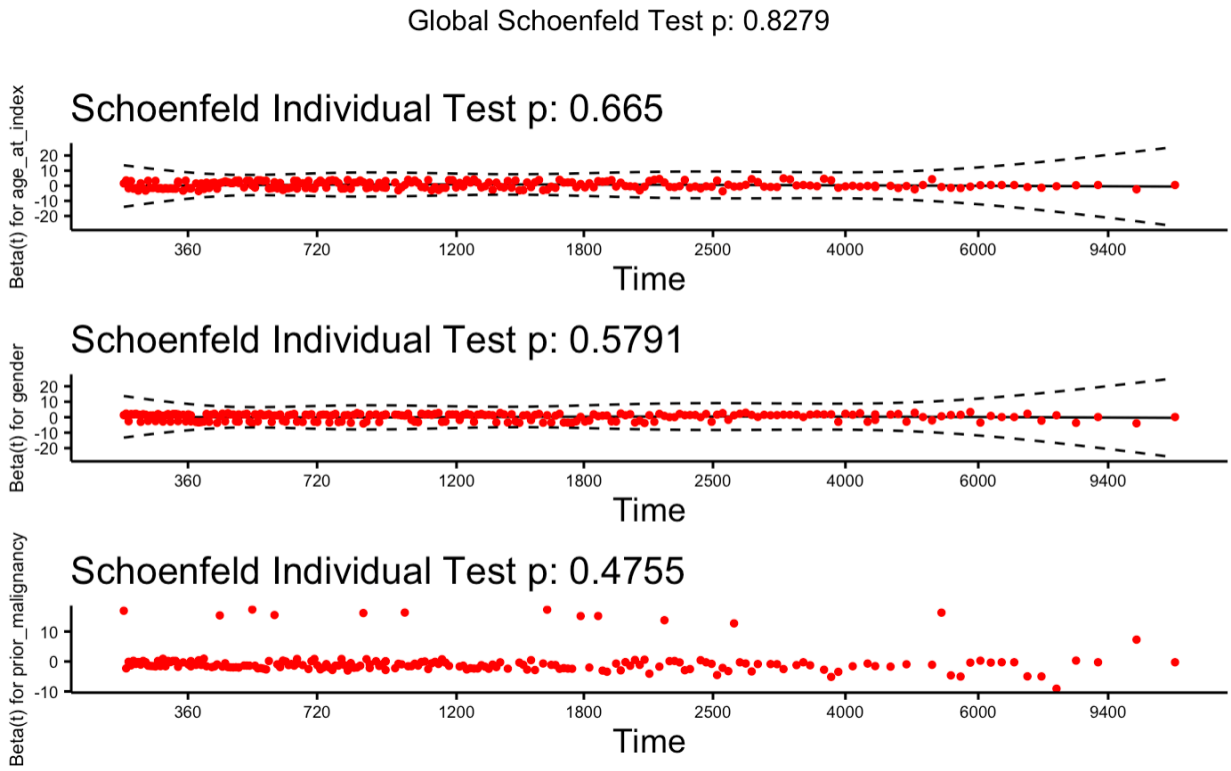

```
# Schoenfeld residuals - Proportional hazards assumption
test.ph <- cox.zph(cox)
test.ph
```

```
##                  chisq df    p
## age_at_index     0.188  1 0.66
## gender           0.308  1 0.58
## prior_malignancy 0.509  1 0.48
## GLOBAL           0.890  3 0.83
```

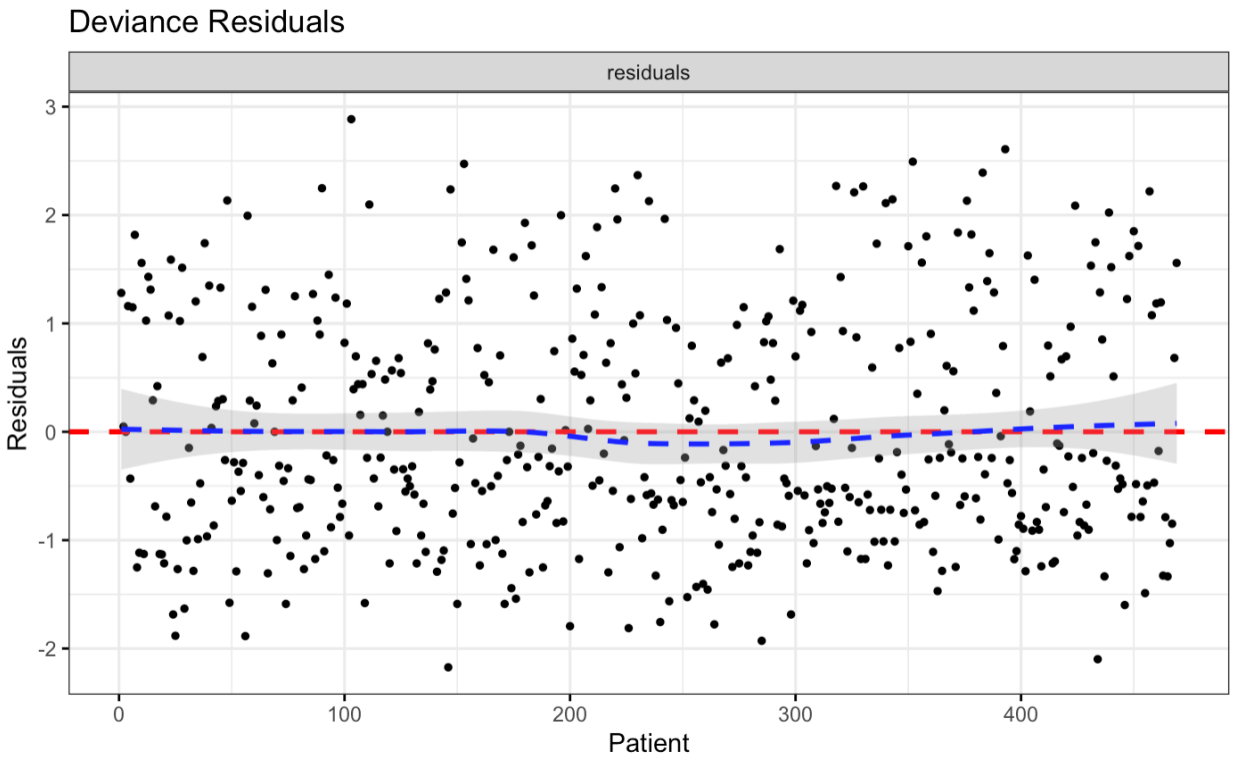
```
ggcoxzph(test.ph)
```

```
## Warning: Removed 40 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 25 rows containing missing values (geom_point).

## Warning: Removed 40 row(s) containing missing values (geom_path).

## Warning: Removed 40 row(s) containing missing values (geom_path).
```




```
# Deviance residuals - Influential observations or outliers
ggcoxdiagnostics(cox, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw(),
                 main = "Deviance Residuals", xlab = "Patient", ylab = "Resid
uals")

## `geom_smooth()` using formula 'y ~ x'
```


Deviance Residuals
residuals
Residuals
3
2
1
0
-1
-2
0
100
200
300
400
Patient