A Survival Analysis of Patients with Cirrhosis

Ryan DeStefano, Robert Goebel

STAT 417: Survival Analysis Methods

March 20, 2022

Introduction

The “cirrhosis dataset” contains variables on 424 primary biliary cirrhosis patients referred to a specific clinic. For context, cirrhosis is a disease of the liver. The dataset was taken from the website *Kaggle* and includes data on 424 individuals. The goal of the study was to assess the time from registration to the clinic, until death. This clinic occurred over the duration of 10 years and the outcomes of the study for specific patients were death, liver transplant, or survival through 10 years. Survival in the context of this study, is not dying from cirrhosis. The variables analyzed in this report are censoring status, drug taken, age, sex, cholesterol, and cirrhosis stage. The time to event variable is “N_days” which represents the time (in days) from the start of the study until patient death. The censoring variable takes values of 0 or 1. 0 represents a censored observation, meaning that patient either survived the length of the study or received a liver transplant during the study and 1 represents a complete observation, meaning this patient died due to cirrhosis during the study. The quantitative predictors in this analysis are age (in years at the start of the study) and cholesterol (cholesterol level at the start of the study). The three categorical predictors are drug (D-penicillamine or placebo), sex (M or F), and stage (1, 2, 3, or 4)

Also of note, we removed any observations with incomplete values for explanatory variables. This causes the sample size to decrease from 424 to 276. The age variable was transformed from days to years for ease of use in interpretation in the analysis. And the censoring variable was transformed to be encoded as either a 0 or 1 depending on the outcome of the study (death coded to 1 and transplant or survival coded as 0).

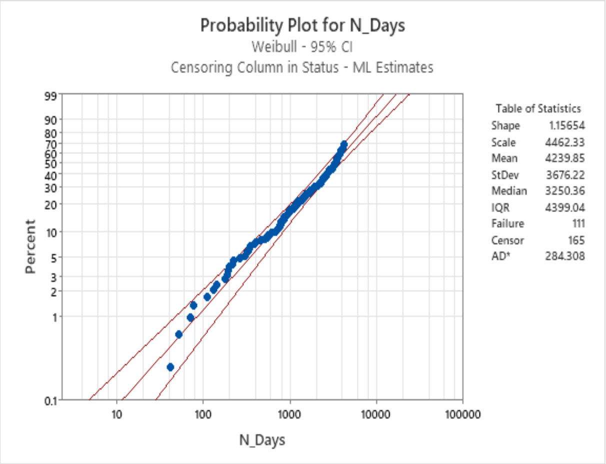
Parametric Survival Analysis

The first step in the parametric analysis is to find the best fitting probability distribution for the time to event variable, t. To find this distribution for time until death from cirrhosis we must examine the Anderson Darling statistic and examine the probability plot for the time to event variable (N_days), fit to an exponential, logistic, lognormal, and weibull probability distributions.

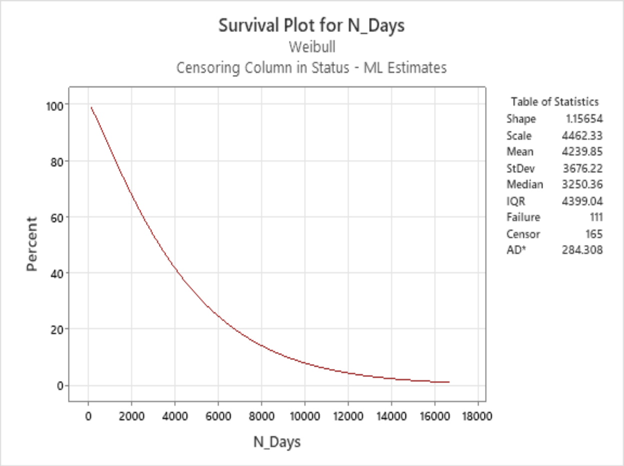
The Anderson Darling statistics and description of probability plots are as follows for the 4 probability distributions:

Exponential: AD = 284.54	Points follow a straight line fairy closely
Logistic: AD = 285.18	Points stray from a straight line
Lognormal: AD = 284.89	Points stray from a straight line
Weibull: AD = 284.31	Points follow a straight line very closely

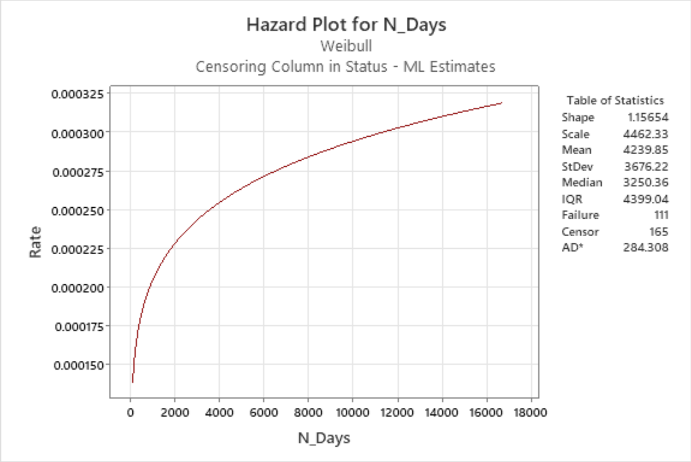
We will continue this parametric analysis with the weibull distribution because when fit to the data, the analysis with the weibull probability distribution has the smallest AD statistic and the points fall very close to a straight line on the probability plot as seen in Graph 1. We then constructed survival and hazard plots for patients time from registration to clinic until death from cirrhosis. Examining Graph 2 we see that the probability of surviving past t days decreases fairly rapidly and constantly from 0 to 6000 days then begins to level out after. This leveling out is due to the probability of surviving after about 6000 days not changing much because almost all the probability is already exhausted by this time, so this probability can’t decrease much more. The expected number of days to death as seen on Graph 2 is 4239.85 days and the median time to death is 3250.36 days. As seen in graph 3, the hazard of death is constantly increasing. Increasing rapidly from 0 to 2000 days then more slowly after, this makes sense because in general, risk of death from a disease should remain constantly increasing. If you still have the disease after a certain amount of time it means the risk of death is always there.



Graph 1: Probability Plot

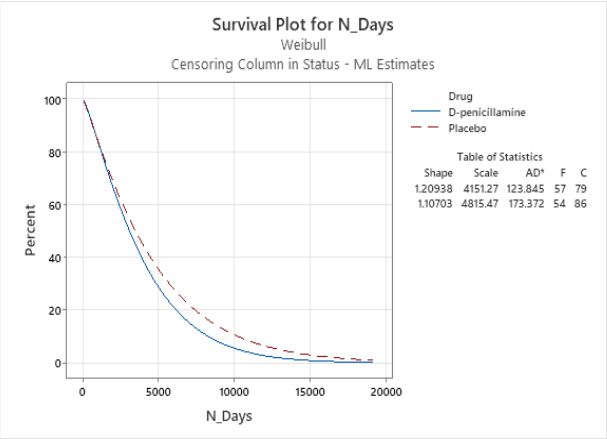


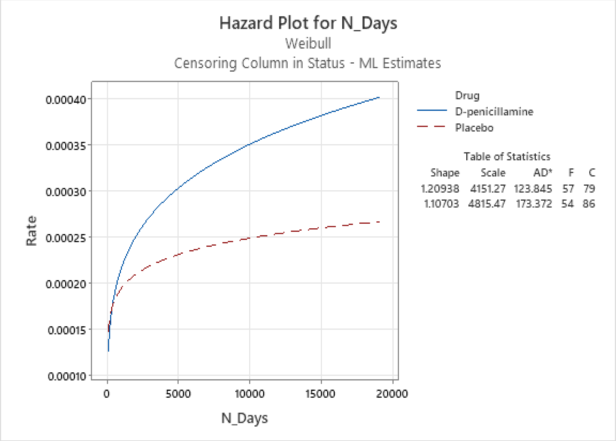
Graph 2: Survival Plot



Graph 3: Hazard Plot

We will now examine the difference in survival experiences for each of the levels of the categorical predictors (drug, sex, stage). Looking at the expected at mean survival times for patients grouped by drug type we see that patients taking the placebo have a mean survival time of 4636.84 days and median 3458.23 days while those patients taking the D-penicillamine drug have mean survival time 3897.18 days and median 3065.94 days. From these summary statistics it appears that patients taking the treatment drug D-penicillamine tend to survive less than those patients taking the placebo. Furthermore, examining the survival of the groups as seen in Graph 4, it appears that the survival probability beyond time, t, is always less for those taking the D-penicillamine compared to those taking the placebo. That is, the survival curve for the patients taking the D-penicillamine drug is lower than the curve for the placebo curve at all times, t. Patients taking the placebo tend to survive longer than those taking the D-penicillamine drug. Looking at the hazard curves in Graph 5 we see that the hazard of death for those patients taking D-penicillamine increase much faster than the hazard of death for those taking the placebo, and the hazard of death from cirrhosis is always higher for those in the D-penicillamine group.

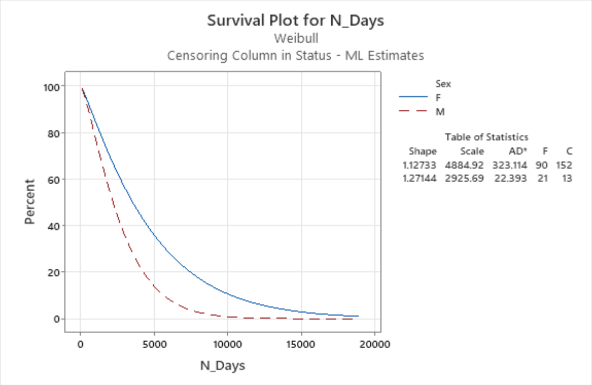


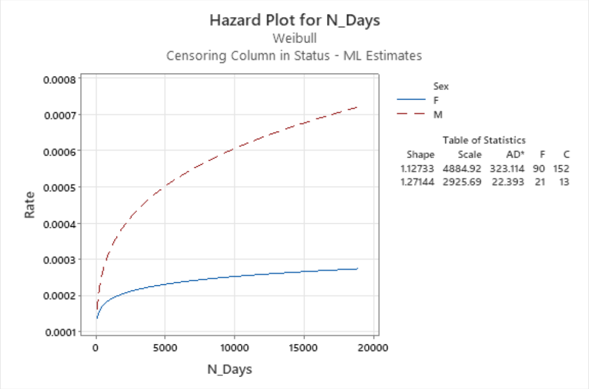


Graph 4: Survival Plot by Drug Type

Graph 5: Hazard Plot by Drug Type

Examining the summary statistics for survival grouped by sex we see that the mean survival time for male patients is 2714.66 days and median 2192.98 days while the mean survival time for female patients is 4676.80 days and median 3529.08 days. From these summary statistics it appears that female patients tend to survive longer than male patients. This is further evidenced by the survival plots. Looking at Graph 6 we see that the survival curve for female patients is always above the survival curve for male patients. This means that the probability of survival beyond t days at any time, t, for female patients is larger than the survival probability for males. Looking at the hazard plot in Graph 7 we see that the hazard of death for males is higher than the hazard of death for females at all times, t.





Graph 6: Survival Plot by Sex

Graph 7: Hazard Plot by Sex

Examining the summary statistics for survival grouped by cirrhosis stage we see that the mean survival time for patients with stage 1 cirrhosis is 7406.82 days and median 7228.64 days, the mean survival time for patients with cirrhosis stage 2 is 4367.36 days and median 4210.79 days, the mean survival time for patients with cirrhosis stage 3 is 3743.29 days and median 3329.16 days, and the mean survival time for patients with cirrhosis stage 4 is 2560.18 days and median 1719.50 days. From these summary statistics it appears that as the stage of cirrhosis increases, the length of survival tends to decrease. This makes sense because the worse the stage of cirrhosis the quicker you would expect said patients to die and this is what appears to be the case. This is further evidenced by the survival plots. Looking at Graph 8 we see that the survival curve for those patients with stage 1 cirrhosis is always above the other 3 curves. However, the relationship is a bit different between stage 2, 3, and 4 cirrhosis and survival. The curves start off as expected, with the probability of stage 2 being higher than stage 3, and stage 3 being higher than stage 4, but at 8000 days the 3 curves converge. This means that the probability of survival is the same for all 3 stages (excluding stage 1) after 8000 days. Looking at Graph 9 we see some interesting trends. The hazard of death for stage 4 patients stays constantly fairly low, the hazard for stage 1 and 3 patients aways is fairly low throughout time and are very similar, and the hazard for stage 2 patients starts off the lowest but grows consistently over time to have the largest hazard at many times.

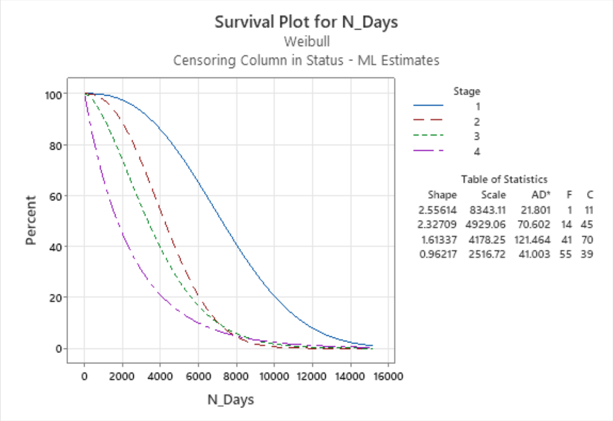


Table of Statistics

Shape	Scale	AD*	F	C
2.55614	8343.11	21.801	1	11
2.32709	4929.06	70.602	14	45
1.61337	4178.25	121.464	41	70
0.96217	2516.72	41.003	55	39

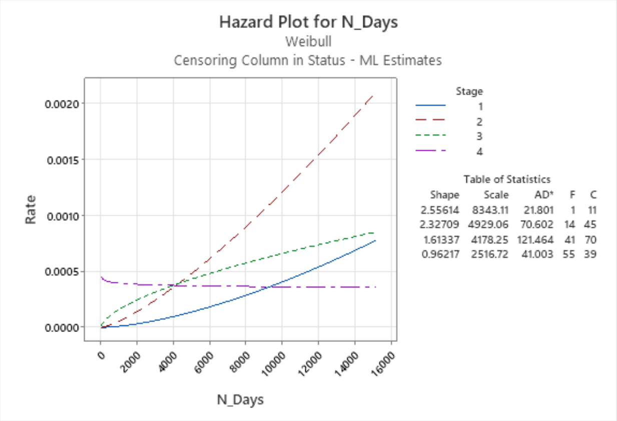


Table of Statistics

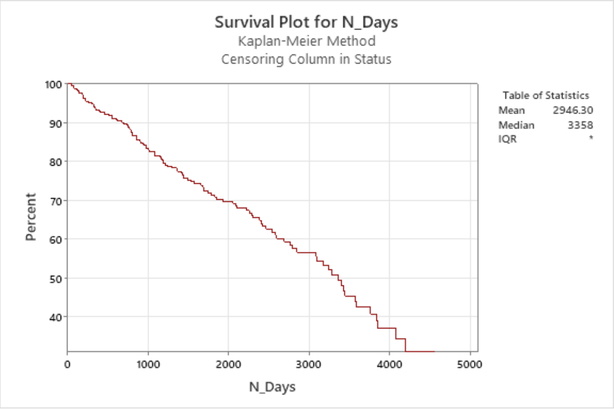
Shape	Scale	AD*	F	C
2.55614	8343.11	21.801	1	11
2.32709	4929.06	70.602	14	45
1.61337	4178.25	121.464	41	70
0.96217	2516.72	41.003	55	39

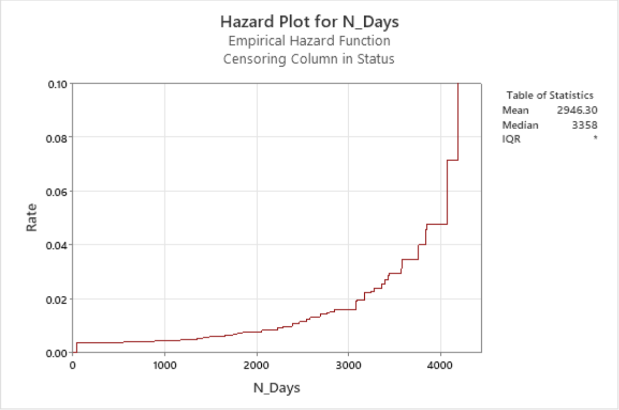
Graph 8: Survival Plot by Cirrhosis Stage

Graph 9: Hazard Plot by Cirrhosis Stage

Non Parametric Survival Analysis

By constructing Kaplan-Meier estimates for the cirrhosis dataset, we were able to analyze the overall survival experience for the study participants as well as assess the effects of multiple factors on the probability of survival. The overall plot of the Kaplan-Meier estimates represents the general survival experience for all subjects regardless of factors such as age, sex, and stage. The cumulative hazard plot is also provided to aid with analyzing the change in survival likelihood with respect to time.

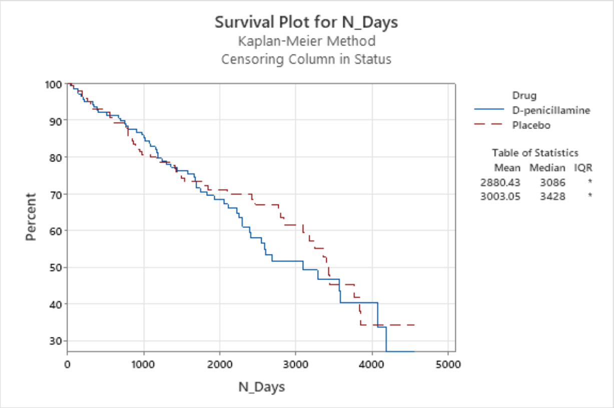


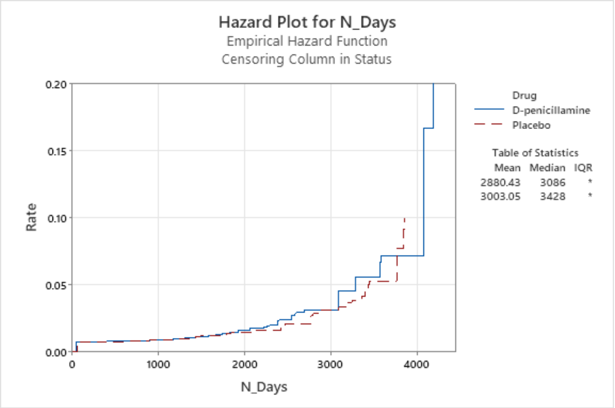


Graph 1a: Kaplan-Meier Method Survival Plot

Graph 1b: Estimated Cumulative Hazard Plot

According to the survival plot based on the Kaplan-Meier estimates, the probability of survival past time t appears to decrease at a roughly consistent pace throughout the duration of the study. There is a slight variation in the survival curve around 3000 days in which it appears to flatten out slightly, but this is easily attributable to the inherent variability of the data. The cumulative hazard plot provides similar insights, indicating that the hazard of death appears to accelerate upwards as time passes.

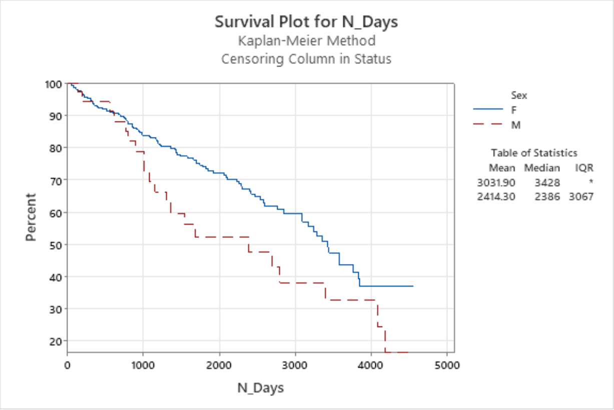




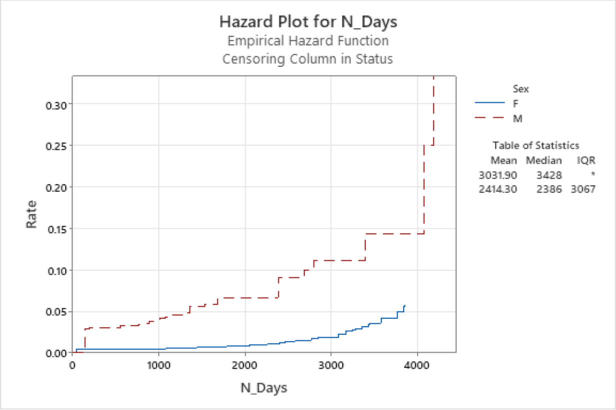
Graph 1c: Survival Plot by Drug

Graph 1d: Cumulative Hazard Plot by Drug

Graph 1c demonstrates the variation in survival experiences for subjects who were administered the D-penicillamine drug compared to those who received the placebo drug. The survival experiences appear to be consistent for around the first 1500 days, after which point the probability of survival appears to be substantially lower for participants in the placebo group. The survival curves reconvene at around 3500 days, although the trends are more difficult to discern due to the small quantity of entries with complete event times past this point. The hazard plot in Graph 1d demonstrates that the cumulative hazard for participants who took either type of drug were at roughly the same instantaneous risk of death due to cancer at each time t regardless of the type of drug they received.

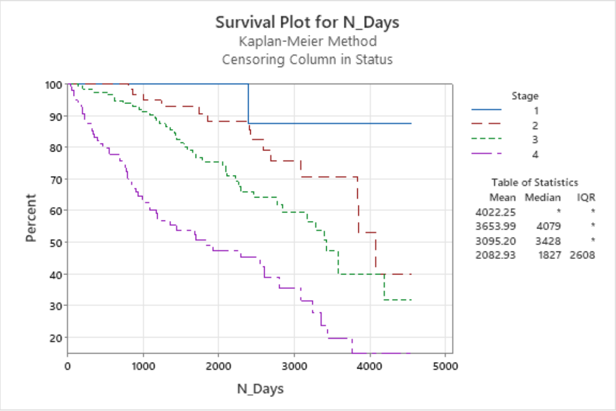


Graph 1e: Survival Plot by Sex

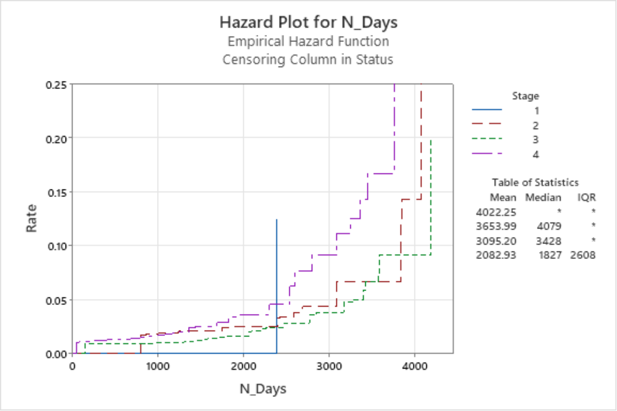


Graph 1f: Cumulative Hazard Plot by Sex

The estimated survival curves for males and females appear to differ drastically throughout time, with the first differences occurring at around 500 days. After this point, males appear to have a significantly lower probability of survival compared to females who also survived up until the time of comparison. In accordance with Graph 1e, Graph 1f demonstrates that the instantaneous risk of death at time t given survival until time t is significantly higher for males at all points in time compared to females.



Graph 1g: Survival Plot by Stage



Graph 1h: Cumulative Hazard Plot by Stage

Based on Graph 1g, the survival plot shows that the probability of survival varies widely based on the stage of the cancer with which the participant was diagnosed. The graph indicates that patients with stage 4 cancer are estimated to be least likely to survive beyond time t, with stage 3 being the next lowest in terms of survival probability. Patients with stage 2 cancer appear to have the second highest probability of survival, with stage 1 having the highest probability of survival overall out of the four groups. Graph 1h provides similar insights, suggesting that the instantaneous risk of death due to cancer at time t given survival until time t is roughly the same for all groups until around 500 days, with subjects with stage 4 cancer having the highest hazard of death after this point. There is a significant spike in the hazard for stage 1 cancer patients at around 2500 days which is likely to be an outlier as a result of the small number of patients with stage 1 cancer included in the study. Patients with stage 2 or stage 3 cancer appear to have roughly similar hazards of death at all points in time based on Graph 1h.

In order to assess the statistical significance of the variation between the survival curves based on drug administration, sex, and cancer stage, log-rank tests were conducted. With a test statistic of 0.4 and a resulting p-value of 0.5, we do not have sufficient evidence to suggest that the estimated survival experiences differ for patients who received the D-penicillamine drug compared to those who received the placebo drug. A test statistic of 4.5 and a p-value of 0.03 means that we do have sufficient evidence to suggest that the estimated overall survival experiences differ for males and females. With a test statistic of 44.6 and a p value of approximately 0, we have strong evidence to suggest that patients have different survival experiences based on their cancer stage.

Factor	Test Statistic (Chi-Square)	Degrees of Freedom	P-value
Drug	0.4	1	0.5
Sex	4.5	1	0.03
Stage	44.6	3	<0.0001

Compared to the parametric analysis, the non-parametric analysis yielded similar results. Both analyses concluded that the survival probability for subjects in the D-penicillamine group was lower that that of subjects in the placebo group, although the nonparametric analysis determined that this difference was statistically insignificant. Both analyses also concluded that the overall survival probability for males was lower than for females. Both the parametric and non-parametric analyses were also in agreement that the likelihood of a patient's survival was likely to be higher for those with stage 1 or stage 2 cirrhosis compared to those with stage 3 or stage 4 cirrhosis.

Regression Analysis

We looked at a total of around 6 models with numerous variable combos and interactions but the model that fit the best (had the largest partial likelihood test statistic), was the model containing the variables gender, sex, stage, age, and cholesterol. As seen in Table 2 this model had a test statistic of 73.47 with degrees of freedom equal to 7. This computes to a p-value of approximately 0 for the model. Looking at Table 1 we see that the variables age, cholesterol, and stage are individually significantly useful in predicting the hazard of death from cirrhosis, after adjusting for all other variables**.** The variables sex and drug were not significantly useful predictors of hazard of death but as seen in the parametric analysis, there were differences in hazard for these groups. It's also worth noting that this combo resulted in the largest test statistic, an even larger statistic than when sex and drug were omitted from the model even with these two predictors not being significant on their own.

	Coefficient	exp(Coef)	95% CI	P-Value	Std. Error
Sex (Male)	-.090	.914	(.624, 1.339)	.643	.195
Drug (Placebo)	.203	1.225	(.739, 2.028)	.430	.257
Stage (2)	1.091	2.977	(.389, 22.759)	,293	1.038
Stage (3)	1.633	5.117	(.697, 37.551)	.108	1.017
Stage (4)	2.514	12.352	(1.691, 90.185)	.013	1.014
Age	.037	1.037	(1.016, 1.059)	<.001	.011
Cholesterol	.002	1.002	(1.001, 1.002)	<.001	.<.001

Table 1: Cox Regression Output

	Test Statistic	P-Value	Degrees of Freedom
Likelihood Ratio Test	73.47	<0.001	7

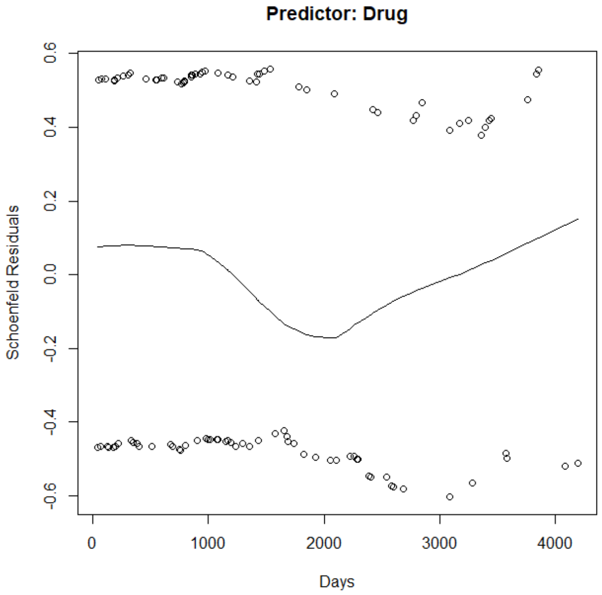
Table 2: Partial Likelihood Output

To check for any violations in the proportional hazard assumption we will examine the Schoenfeld residual plots as well as look at the output of the Schoenfeld residual test. Looking at the Schoenfeld residual plots below we see that the predictors drug, age, and cholesterol could have some issues with violating proportional hazards. Further analyzing, we see that in Table 3, the predictors stage, age, and cholesterol have p-values less than .05. This means that at a significance level of .05, evidence suggests that the proportional hazards assumption for these predictors after adjusting for all other predictors. The sex and drugs variables have large p-values so the proportional hazard assumption seems to be satisfied for those predictors. We did not remove any outliers because there doesn't appear to be any need because there weren't extreme cases in the data.

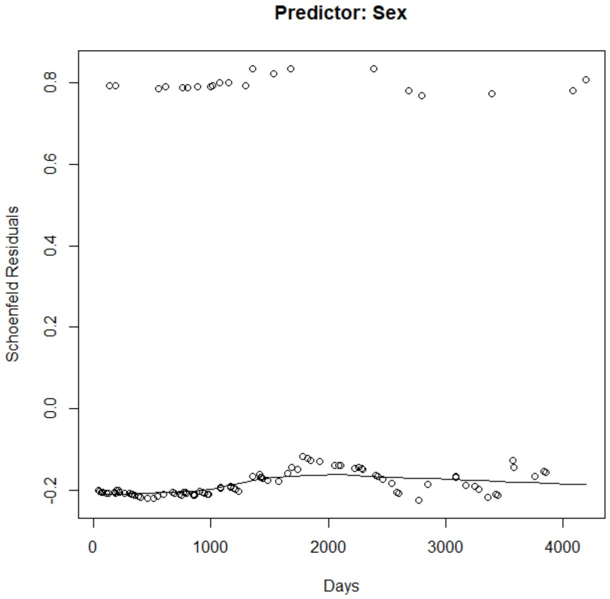
We now will look at the values of the coefficients and hazard ratios to explore some associations between hazard and some of the variables.

- We are 95% confident that the hazard of death from cirrhosis for male patients is between 38.6% lower and 33.9% higher than the hazard of death from cirrhosis for female patients, after adjusting for all other predictors.
- We are 95% confident that the hazard of death from cirrhosis for patients taking the placebo is between 26.1% lower and 102.8% higher than the hazard of death from cirrhosis for patients taking the D-penicillamine drug, after adjusting for all other predictors.

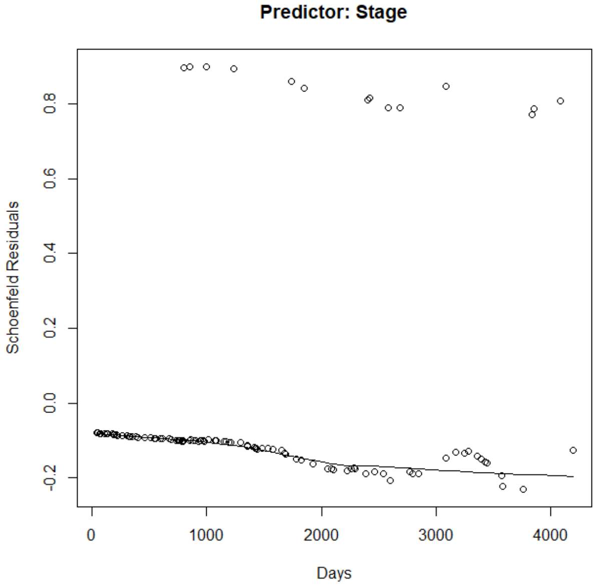
- We are 95% confident that the hazard of death from cirrhosis for patients with stage cirrhosis is between 69.1% and 801.8% higher than the hazard of death from cirrhosis for patients with stage 1 cirrhosis, after adjusting for all other predictors.
- We are 95% confident that the hazard of death from cirrhosis increases by between 6.1% and 5.9% for every increase in 1 year of the patient, after adjusting for other variables.
- We are 95% confident that the hazard of death from cirrhosis increases by between 0.1% and 0.2% for every increase in 1 unit of cholesterol in the patient, after adjusting for other variables.



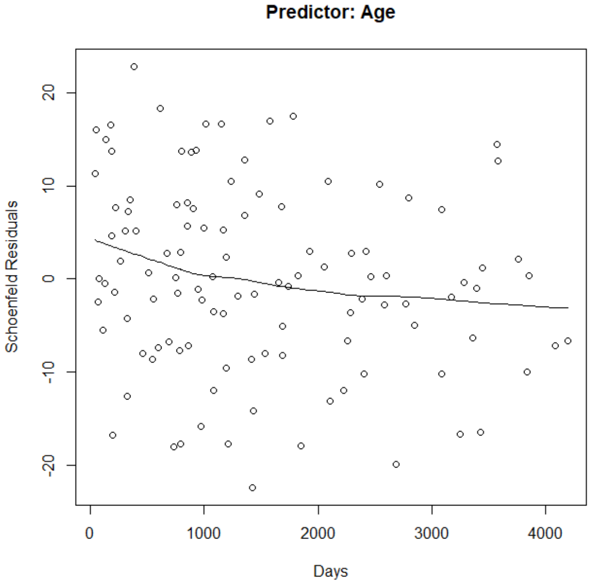
Graph 10: Drug Type Schoenfeld Residual Plot



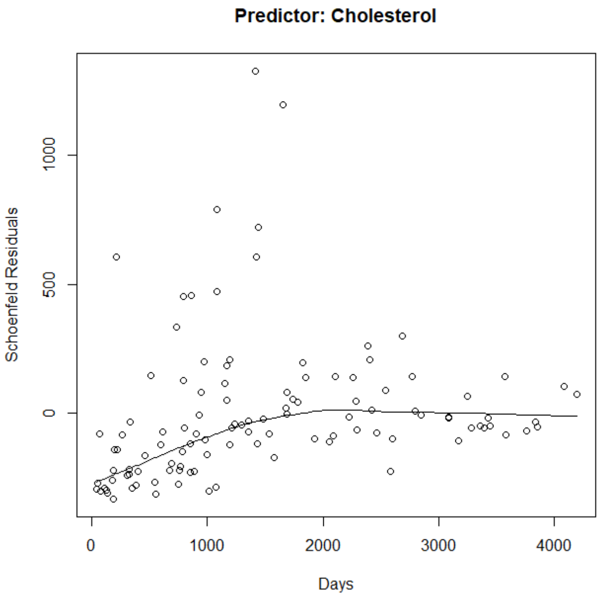
Graph 11: Sex Schoenfeld Residual Plot



Graph 12: Stage Schoenfeld Residual Plot



Graph 13: Age Schoenfeld Residual Plot



Graph 14: Cholesterol Residual Plot

	Chisq	P
Sex	.727	.294
Drug	.987	.321
Stage	10.421	.017
Age	5.598	.018
Cholesterol	10.648	.001

Table 3: Schoenfeld Residual Test

Some specific hazard ratio values:

Description	Hazard Ratio
50 year old males who took the placebo drug, have stage 4 cirrhosis, with a cholesterol level of 200 to a 50 year old female who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 200	15.13
40 year old males who took the placebo drug, have stage 2 cirrhosis, with a cholesterol level of 200 to a 50 year old males who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 200	2.056
50 year old males who took the placebo drug, has stage 1 cirrhosis, with a cholesterol level of 100 to a 50 year old female who took the placebo drug, has stage 1 cirrhosis, with a cholesterol level of 400	0.549

Table 4: Specific Hazard Ratios

Looking at table 4, we see that 50 year old males who took the placebo drug, have stage 4 cirrhosis, with a cholesterol level of 200 are estimated to have a hazard of death from cirrhosis 1513% higher than the hazard of death for 50 year old females who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 200. 40 year old males who took the placebo drug, have stage 2 cirrhosis, with a cholesterol level of 200 are estimated to have a hazard of death from cirrhosis 105.6% higher than the hazard of death for 50 year old males who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 200. 50 year old males who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 100 are estimated to have a hazard of death from cirrhosis 45.1% lower than the hazard of death for 50 year old females who took the placebo drug, have stage 1 cirrhosis, with a cholesterol level of 400.

Summary

After performing this analysis, we were able to uncover multiple factors which serve as valuable predictors of a patient's hazard of death to cancer and overall survival experiences. The parametric analysis matched the survival curve with a weibull distribution and produced multiple survival curves to explain the impact of the various predictors used. The patient's sex was a valuable predictor, with males being significantly more likely to die compared to females regardless of all other factors. There was a very strong association between cirrhosis stage and risk of death, with stage 4 cirrhosis providing the lowest probability of survival and stage 1 cirrhosis providing the highest probability of survival. No significant association was found between the type of drug administered (D-penicillamine versus placebo) and the survival experience.

For the regression analysis, a model was selected which was based on the aforementioned factors drug, sex, and cirrhosis stage as well as age and cholesterol. Based on the Schoenfeld residual plots, only the drug variable appeared to have a violation of the proportional hazards assumption. The model revealed a significant positive association between stage 4 cirrhosis and hazard of death. The model also suggested the presence of strong positive associations between age and hazard of death as well as cholesterol and hazard of death. Overall, these results are in line with what would be expected for general health trends, with age, cholesterol, and stage 4 cancer being generally associated with hazard of death even outside the scope of this study. However, the most surprising result was the ineffectiveness of the D-penicillamine drug on impacting survival, as the difference between the survival experience of patients who received the drug was statistically insignificant from the survival experience of patients who received the placebo drug. The insights provided by this analysis serve to explain the contributing factors which can lead to one's death due to cirrhosis and hopefully may guide progress towards helping to prevent as many deaths as possible.

Appendix - R Code

```
cir = na.omit(cirrhosis)

for (i in 1:nrow(cir)){
  if (cir$Status[i] == "D"){
    cir$Status[i] = 1
  }
  else{
    cir$Status[i] = 0
  }
}

cir$Status = as.integer(cir$Status)

cir$Age = cir$Age / 365

write.csv(cir, "cir.csv", row.names=FALSE)

# Non-Parametric Analysis

km.obj <-
survfit(Surv(N_Days,Status)~as.factor(Drug)+as.factor(Sex)+as.fa
ctor(Stage)+Age+Cholesterol, data=cir)

survdiff(Surv(N_Days,Status)~as.factor(Drug), data=cir)
survdiff(Surv(N_Days,Status)~as.factor(Sex), data=cir)
survdiff(Surv(N_Days,Status)~as.factor(Stage), data=cir)

# Regression Analysis

cr.obj
<-coxph(Surv(N_Days,Status)~as.factor(Drug)+as.factor(Sex)+as.fa
ctor(Stage)+Age+Cholesterol, data=cir)

summary(cr.obj)

schoen = residuals(cr.obj, type = "schoenfeld")

cir.times = sort(cir[cir$Status != 0 , ]$N_Days)
```

```
plot(cir.times, schoen[,1], xlab = "Days", ylab = "Schoenfeld
Residuals", main = "Predictor: Drug" )
smooth.sres = lowess(cir.times, schoen[,1])
lines(smooth.sres$x, smooth.sres$y, lty = 1)

plot(cir.times, schoen[,2], xlab = "Days", ylab = "Schoenfeld
Residuals", main = "Predictor: Sex" )
smooth.sres = lowess(cir.times, schoen[,2])
lines(smooth.sres$x, smooth.sres$y, lty = 1)

plot(cir.times, schoen[,3], xlab = "Days", ylab = "Schoenfeld
Residuals", main = "Predictor: Stage" )
smooth.sres = lowess(cir.times, schoen[,3])
lines(smooth.sres$x, smooth.sres$y, lty = 1)

plot(cir.times, schoen[,6], xlab = "Days", ylab = "Schoenfeld
Residuals", main = "Predictor: Age" )
smooth.sres = lowess(cir.times, schoen[,6])
lines(smooth.sres$x, smooth.sres$y, lty = 1)

plot(cir.times, schoen[,7], xlab = "Days", ylab = "Schoenfeld
Residuals", main = "Predictor: Cholesterol" )
smooth.sres = lowess(cir.times, schoen[,7])
lines(smooth.sres$x, smooth.sres$y, lty = 1)

cox.zph(cr.obj, transform = "log")
```