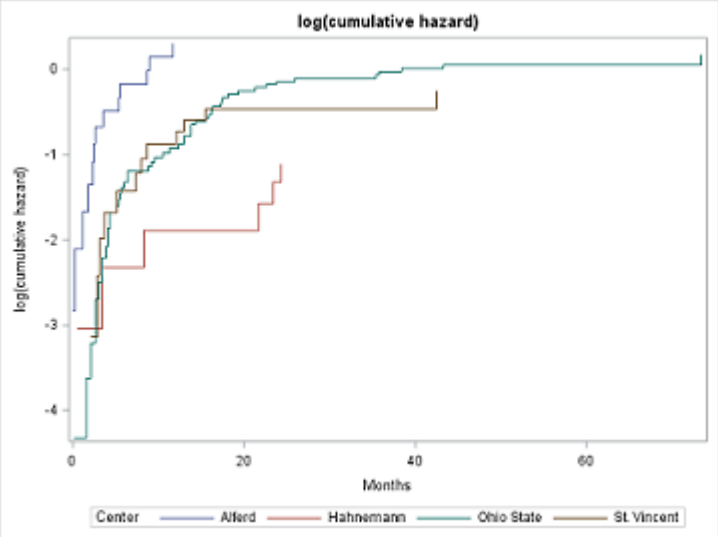
Homework 3: BST 665: Survival Analysis

Devin Koehl

Spring Semester 2019

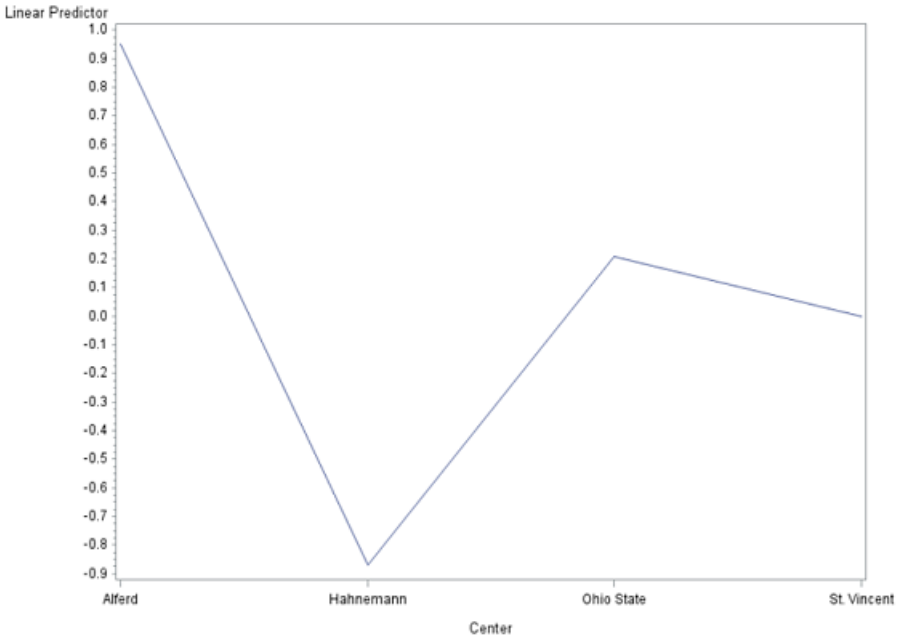
1. A physician is studying survival following bone marrow transplant in leukemia patients. She has conducted a multicenter trial of 137 patients.

A. Using the length of the follow-up period as the survival time and the status at last follow-up as the censoring variable, plot the log of the estimated cumulative hazard function for each treatment center. What does this figure tell us?



This figure shows us that Alfred has the highest hazard, followed by Ohio State, St. Vincent, and lastly Hahnemann has the lowest.

Another visualiization with the reference group:

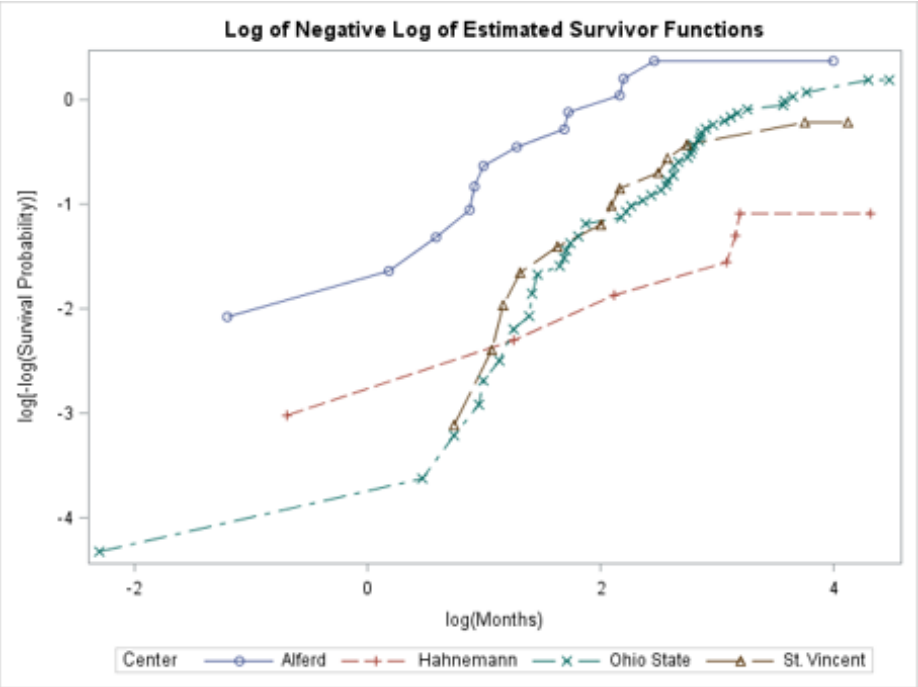


This figure shows us that Alfred has the highest, followed by Ohio State. Hahnemann has the lowest. Since St. Vincent is the reference group log of 0 is 1.

B. Is it reasonable to assume proportional hazards for treatment center? Why or why not? [Note: In class, we discussed several ways of checking the proportional hazards assumption. For this problem, you should use at least two different methods to assess the proportional hazards assumption for treatment center].

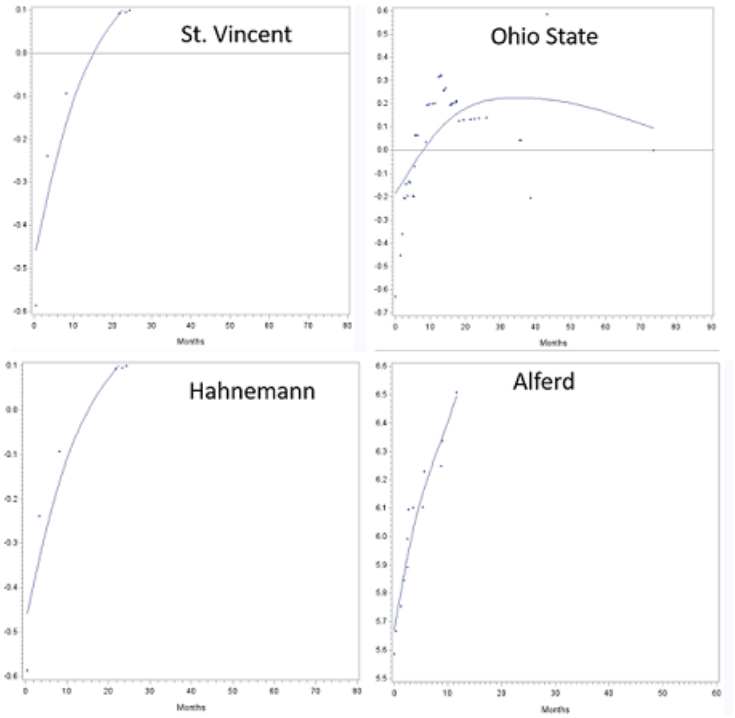
I would conclude it is not reasonable to assume the proportional hazards treatment. I plotted the log-log survival plot, the Schoenfeld residuals, and also checked the correlation of transformations of the time variable.

First if we plot the log-log survival plot of each Center:



We can see the survival lines cross each other. If the lines are not parallel, the proportional hazards assumption may not be reasonable.

Plotting the the scaled Schoenfeld residuals against the time variable. If the residuals do not appear to be randomly scattered about the y=0 line, the proportional hazards assumption may not be justified. This can be seen below for each center:



The residuals do not appear to be randomly scattered for the Ohio State center or Hanheman. Again, confirming the log-log plot that we see above.

Lastly, just for my own double checking and wonderment I wanted to look at the correlation of the time variables and use the Supremum Test for Proportionals Hazards Assumption.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	schage	sschage
Months	-0.20665 0.0642 81	-0.10520 0.3499 81
months_sq	-0.10065 0.3713 81	-0.05844 0.6043 81
logmonths	-0.23527 0.0357 80	-0.13728 0.2246 80

Which demonstrate the residuals for the logmonths is correlated with time.

And lastly

Supremum Test for Proportionals Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
CenterAlferd	1.6406	1000	234	0.0560
CenterHahnemann	0.9423	1000	234	0.3700
CenterOhio_State	1.8134	1000	234	0.0280

Which shows that Ohio State is in violation of the proportional hazards assumption

2. The physician from Exercise 1 would like to compare the survival for the three disease groups. Using the Bone marrow.sas7bdat dataset, fit a stratified Cox proportional hazards model using the length of the follow-up period as the survival time and the status at last follow-up as the censoring variable. Stratify on treatment center and include effects for age, sex, and disease group in your model. Assume that the effects of each covariate are constant across strata.

A. Write the model you just fit using mathematical notation (i.e., using β coefficients). Be sure to define all variables used in the model,

including any dummy variables.

From notes: When you stratify using a covariate, you are essentially taking its effect and its interaction with time and absorbing them into the baseline hazard function. Each stratum will have it's own baseline hazard function.

$$log(h_r(t|\beta, x)) = log(h_{r_0}(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Where r = 1,2,3,4 to represent the stratum of Center with 1.Alferd, 2.Hahneman, 3.Ohio State, 4. St. Vincent

So for Alferd

$$log(h_r(t|\beta, x)) = log(h_{1_0}(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Etc, etc for the other centers.

Also,

x_1 = Age and

$$x_2 = \begin{cases} 1 & \text{Female} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$x_3 = \begin{cases} 1 & \text{AML Low Risk} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$x_4 = \begin{cases} 1 & \text{AML High Risk} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

B. Create a table reporting the estimated hazard ratios (and 95%confidence intervals) for each of the effects in this model. This table should include all pairwise hazard ratios comparing the three disease groups.

AML High to AML Low:

HR (t | AML high vs AML low) $= \frac{e^{\beta_4}}{e^{\beta_3}} = e^{\beta_4 - \beta_3} = e^{0.38991-(-0.45420)} = e^{0.84411} = 2.3259$.

Confidence intervals are given by (using covb option in SAS)

0.84411 +/- 1.96*$\sqrt{(Var(\beta_4) + Var(\beta_3) - 2cov(\beta_4, \beta_3))}$

0.84411 +/- 1.96*$\sqrt{0.0855947846 + 0.1012686919 - 2 * 0.0549745064}$

0.84411 +/- 1.96*0.277334

0.84411 - 1.96*0.277334 = 0.30053 = $e^{0.30053}$ = 1.35

0.84411 + 1.96*0.277334 = 1.387 = $e^{1.387}$ = 4.01

For high vs low, we have a HR of 2.33 with 95% CI of (1.35,4.01).

For the remaining Hazard ratios, since the reference group hazard will be log of 0 which is 1, they will simply be the same parameter estimates that we would get exponentiating each beta value.

HR (t | AML high vs ALL) = $e^{\beta_4} = e^{0.38991} = 1.476$.

Confidence Intervals are given by:

$0.38991 + / - 1.96 * 0.29257$

$0.38991 + 1.96 * 0.29257 = 0.9633472 = e^{0.9633472} = 2.62$.

$0.38991 - 1.96 * 0.29257 = -0.1835272 = e^{-0.1835272} = 0.8323$.

HR (t | AML low vs ALL) = $e^{\beta_3} = e^{-0.45420} = 0.63495$.

$-0.45420 + 1.96 * 0.31823 = 0.1695308 = e^{0.1695308} = 1.1847$.

$-0.45420 - 1.96 * 0.31823 = -1.0779 = e^{-1.0779} = 0.3403$.

Results are summarized below:

Please note sex in the table refers to Female

Table: Cox Proportional Hazards Model

Covariate	HR	Lower 95% CI	Upper 95% CI	p-value
Age	1.02	0.99	1.05	0.1
Sex	1.35	0.84	2.16	0.2
AML High	1.48	0.83	2.62	0.2
AML Low	0.64	0.34	1.19	0.2
AML High vs. AML Low	2.33	1.35	4.01	0.002
AML High vs. ALL	1.48	0.83	2.62	0.2
AML Low vs. ALL	0.64	0.34	1.19	0.2

HR = Hazard Ratio; CI = Confidence Interval

*p-value of <0.05 is considered statistically significant

Results for age, sex, and disease group were given by Cox output.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
BP	High	1	0.23836	0.15605	2.3331	0.1266	1.269	0.935	1.723	BP High
AFib	Yes	1	0.96086	0.19479	24.3320	<.0001	2.614	1.784	3.829	AFib Yes

Checking pairwise comparisons, a contrast statement was used:

Contrast Estimation and Testing Results by Row

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
AML High vs AML Low	PARM	1	0.8441	0.2773	0.05	0.3005	1.3877	9.2639	0.0023
AML High vs ALL	PARM	1	0.3899	0.2926	0.05	-0.1835	0.9633	1.7762	0.1826
AML Low vs ALL	PARM	1	-0.4542	0.3182	0.05	-1.0779	0.1695	2.0372	0.1535

C. Does the effect of disease group on survival significantly differ by treatment center? Revise your model to answer this question and justify your answer using a hypothesis test. (Make sure you state the null and alternative hypotheses, level of significance used, test statistic, p-value and conclusion in terms of the problem).

To determine this, we need to add an interaction term between disease group and treatment center. This will assess if disease group varies by center.

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	2.4961	0.1141
Sex	1	1.4339	0.2311
Group	2	5.0765	0.0790
Center*Group	5	8.4542	0.1329

H_0 : The effect of disease group does not differ by treatment group. $(\beta_i = 0)$

H_A : The effect of disease group does differ by treatment group. $(\beta_i \neq 0)$

Where (β_i) represents interaction terms in each of the stratum models.

The level of signifcance is p<0.05 and the Wald Chi-Square test statistic is 8.4542. The p-value is 0.1329. Note that the center*group interaction is not statistically significant. This indicates that there is no evidence that the disease groups vary by center. We fail to reject the null hypothesis and conclude that the effect of disease group on survival does not differ by treatment center.

D. Write a short summary (1-2 paragraphs) of the results of this study. Be sure to include your interpretation of the results and to take the stated goal of the study into consideration.

A physician is studying survival following bone marrow transplant in leukemia patients. A Stratified Cox Proportional Hazard model was fit to assess the survival for the three disease groups defined as ALL, AML Low Risk, and AML High Risk by Center. SAS 9.4 was used to generate these outcomes. A p-value of < 0.05 was considered statistically significant. The Centers of interest were Alferd, Hahnemann, Ohio State, and St. Vincent. The effects included in the model were age, sex, and disease group. The effects of the covariates were assumed constant across center strata. When assessing the effect of age, sex, and disease group, nothing came into the model as statistically significant, although AML High Risk had the highest ratio of 1.5, indicating this group had a 50% increase in the risk of death. Being female (HR 1.3, CI 0.8-2.2) and older age (HR 1.0, CI 0.9-1.0) also had hazard ratios > 1 in the model. From the type 3 test, group appeared to have the strongest effect on the model (p<0.05). When comparing disease groups, the AML High risk compared to the AML Low Risk showed the biggest increase in risk and was statistically significant (HR 2.3, CI 1.4-4.01, p=0.002), demonstrating that that being AML High Risk had a 130% increase in the risk of death .

When accounting for the effect of disease group by center, this interaction was not statistically significant. This indicates that there is no evidence that the disease groups vary by center. Overall, when assessing risk factors of age, sex, and disease group by center, it did not appear to have a drastic effect on the model. Analyzing disease group by center may not give the full story to this data.

3. Using the length of the follow-up period as the survival time and the status at last follow-up as the censoring variable, fit a Cox proportional hazards model to assess the effects of disease group and platelet recovery on survival. [Disease group and platelet recovery should be the only effects in your model. For this exercise, do not stratify by treatment center].

$$log(h(t|\beta, x)) = log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x(t)_3$$

Where

$$x_1 = \begin{cases} 1 & \text{AML Low Risk} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$x_2 = \begin{cases} 1 & \text{AML High Risk} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$x(t)_3 = \begin{cases} 1 & \text{Platelet Recovery by time t} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

A. Report a hazard ratio (and 95% confidence interval) for the effect of platelet recovery. Give your interpretation of this hazard ratio. .

Discrete time-varying covariates start at a particular value and stay at that value until some intermediate event occurs. The value of the covariate then changes. Using counting process syntax, I created a new variable called RECOVERY STATUS to take the place of the RECOVERY variable. Now, we have a time-varying version of the recovery variable. Before, the "recovery" variable was just did the patient ever recovery and was not taking into the account the period of recovery.

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
RECOVERY_STATUS		1	-1.04385	0.32464	10.3386	0.0013	0.352	Recovery Status
Group	AML High Risk	1	0.30262	0.26894	1.2661	0.2605	1.353	Group AML High Risk
Group	AML Low Risk	1	-0.59004	0.29509	3.9981	0.0456	0.554	Group AML Low Risk

The HR for Recovery Status of Platelet Count is 0.4 (0.2-0.7)

The hazard ratio is protective, demonstrating there is a decrease in risk of 60% if you recover after your bone marrow transplant after adjusting for disease group.

B. Does the effect of platelet recovery vary by disease group? Justify your answer using a hypothesis test. (Make sure you state the null and alternative hypotheses, level of significance used, test statistic, p-value and conclusion in terms of the problem).

$$log(h(t|\beta, x)) = log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x(t)_3 + \beta_4 x_4 + \beta_5 x_5$$

Where

$$x_1 = \begin{cases} 1 & \text{AML Low Risk} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$x_2 = \begin{cases} 1 & \text{AML High Risk} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$x(t)_3 = \begin{cases} 1 & \text{Platelet Recovery by time t} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$x_4 = \begin{cases} 1 & x_1{*}x_3 \text{ if AML Low Risk and Platelet Recovery by time t} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$x_5 = \begin{cases} 1 & x_2{*}x_3 = 1 \text{ if AML Low High and Platelet Recovery by time t} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
RECOVERY_STATUS	1	23.3986	<.0001
Group	2	2.8453	0.2411
RECOVERY_STATU*Group	2	10.4046	0.0055

H_0 : The effect of platelet recovery does not vary by treatment group. (β_4,β_5 =

0)

H_A : The effect of platelet recovery does vary by treatment group (β_4, $\beta_5 \neq 0$)

We can examine the joint test seen above. The level of signifcance is p<0.05 and the Wald Chi-Square test statistic is 10.4046. The p-value is 0.0055 Note that the recovery status*group interaction is statistically significant. This indicates that there is evidence that the platelet recovery does vary by treatment groups. We reject the null hypothesis and conclude that the effect of platelet recovery varies by treatment group

c. Report hazard ratios and 95% confidence intervals for the effect of platelet recovery for each disease group. Discuss these results.

log(h(t | AML Low Risk with Platelet Recovery)) = $log(h_0(t))+\beta_1 x_1+\beta_3 x(t)_3+\beta_4 x_4$

From above, $e^{x_1+x_3+x_4} = e^{1.12800-0.66276-2.02235} = e^{-1.55711} = 0.2107$, matches what I see in SAS. Hooray.

Going back to SAS for CI

$e^{-2.8123} = 0.0600$ and $e^{-0.3019} = 0.7394$. AML with Low Risk with Platelet Recovery has a HR of 0.2107 and CI (0.0600,0.7394).

log(h(t | AML High Risk with Platelet Recovery)) = $log(h_0(t))+\beta_2 x_2+\beta_3 x(t)_3+\beta_5 x_5$

$e^{x_2+x_3+x_5} = e^{0.36525-0.66276-0.03227} = e^{-0.32978} = 0.71908$

$e^{-1.5476} = 0.2127$ and $e^{0.8880} = 2.43$. AML with High Risk with Platelet Recovery has a HR of 0.719 and CI (0.2127 ,2.43).

log(h(t | ALL with Platelet Recovery)) = $log(h_0(t)) + \beta_3 x(t)_3$

$e^{x_3} = e^{-0.66276} = e^{-0.66276} = 0.5154$

$e^{-1.8974} = 0.1499$ and $e^{0.5719} = 1.771$. AML with High Risk with Platelet Recovery has a HR of 0.5154 and CI (0.1499,1.771).

Results from SAS:

Contrast Estimation and Testing Results by Row									
Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
Low AML with Platelet Recovery	PARM	1	-1.5571	0.6404	0.05	-2.8123	-0.3019	5.9113	0.0150
High AML with Platelet Recovery	PARM	1	-0.3298	0.6213	0.05	-1.5476	0.8880	0.2817	0.5956
ALL with Platelet Recovery	PARM	1	-0.6628	0.6300	0.05	-1.8974	0.5719	1.1069	0.2928

c. Write a short summary (1-3 paragraphs) of the results of this study. Be sure to include your interpretation of the results and to take the stated goal of the study into consideration. The physician has also asked that, in your summary, you briefly explain the concept of a time-varying covariate and how it relates to this data set. Include this explanation in your summary. (This explanation should be written so that it can be understood by the physician, who has not taken a survival analysis course).

A physician was interested in the effet of the platelet recovery process on survival and whether that effect varies by disease group.A Cox Proportional Hazard Model with Time-Varying Covariates was fit to assess the affects of disease group and platelet recovery on survival. SAS 9.4 was used to generate these outcomes. A p-value of < 0.05 was considered statistically significant. After a patient has a bone marrow transplant, there is a period where they may recover. This means that platelet recovery is not constant (invariable or unchanging) through the whole study since timing has an effect on the platelet status. The patient may recovery at 1 month or may recovery at3 months. This is referred to a time-varying covariate since it varies with time. The key rule for time-varying dovariates is that you cannot look into the future. Covariates may change in any way based on past data but it may not reach forward in time. For the patients receiving a bone marriw transplant, the patients "time zero" or beginning of transplant up until this recovery time, we know the patient did not recover because they were not at risk for the recovery event. Since we are given the time the patient did recover, we can calcute the interval from their time zero (or starting transplant date) until recovery, and then from the recovery time until end of follow-up. We are basically starting a new time zero, with the start time being when the recovery event occured until the end of their follow-up. It is within this time period we know the event occured.

When examining platelet recovery time, this came out as statistically significant and protective in the model with a HR of 0.4. This shows that if a patient undergoes platelet recovery therapy then the risk of death is reduced by 60%. After adjusting for platelet recovery, AML Low Risk and High Risk come into the model but are not statistically significant (HR 0.6 and 1.4 respectively). Assessing whether platelet recovery varies by disease group, the joint test for the interaction between recovery status and disease group is significant (p=0.006). This means there is evidence to conclude that the platelet recovery does vary

by treatment groups. Examining platelet recovery for each disease group, we can see that this was protective for each disease group, thus showing the risk decreased. The Low Risk group with platelet recovery was statistically significant (p=0.02). This shows that the risk decreases by close to 80% if you have platelet recovery in the low risk AML group. Comparing this to the High Risk AML risk, the risk decreases by only 29%. Lastly, platelet recovery in the ALL group showed a 49% decrease in risk. This demonstrates that platelet recovery has a singificant effect on disease group and the decrease in risk for death.