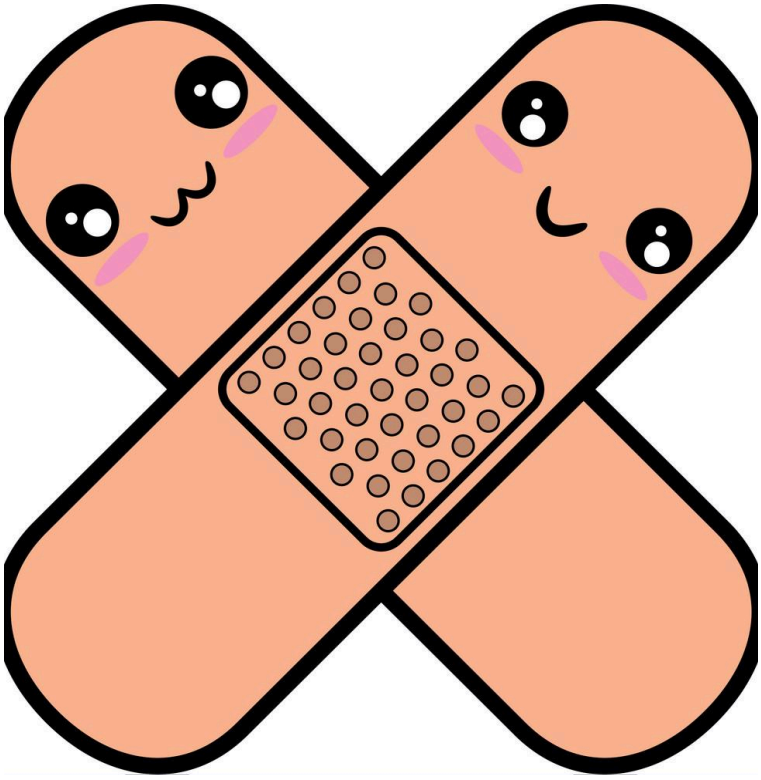
Carcinoma of the Oropharynx Treatment Analysis

STAT 590 Survival Analysis Final Report



California State University, Long Beach

Spring 2024

By:

Teah Thies

Luis Labra

Submitted to:

Dr. Tianni Zhou

Table of Contents

1. Abstract
2. Introduction
3. Variable Description
4. Exploratory Data Analysis
    - *I. Institution, Treatment, and Gender Distribution*
    - *II. Age of Patients*
    - *III. Censored Observations*
    - *IV. Time Plot*
5. Methods and Model Development
    - *I. Kaplan-Meier Estimator*
        - *i. Confidence Intervals*
    - *II. Log Rank Test*
    - *III. Stratified Log-Rank Test*
        - *i. Selection*
    - *IV. Constructing Cox Proportional Hazard Model*
        - *i. Proportional Hazard Assumption - Model 1*
        - *ii. Summary & Results - Model 1*
        - *iii. Summary & Results - Model 2*
        - *iv. Proportional Hazard Assumption - Model 3*
        - *v. Summary & Results - Model 3*
        - *vi. Summary & Results - Model 3*
6. Final Model
7. Results
8. Conclusion
9. Code
    - *I. R Code*
    - *II. SAS Code*

Abstract

We study a comprehensive clinical trial conducted by the Radiation Therapy Oncology Group in the United States. The investigation is to assess the efficacy of these two treatment approaches. In performing the investigation, a Kaplan-Meier estimation was employed extensively to reveal insufficient difference of the treatment. Further analysis of stratification proves helpful in investigating patient outcomes and in the development of a Cox Proportional Hazard model which quantifies factors the general public may already perceive play a pivotal role in predicting outcomes.

Introduction

The Trial conducted by Radiation Therapy Oncology group encompasses patients diagnosed with squamous carcinoma across 15 locations within the mouth and throat, involving 16 participating institutions. However, this analysis focuses solely on data pertaining to three specific sites within the oropharynx. Upon enrollment, patients were randomly allocated to one of two treatment cohorts: radiation therapy alone (Tx=1) or combined radiation therapy with a chemotherapeutic agent (Tx=2).In this study, around 30% of survival data is censored, predominantly due to patients surviving at the time of analysis. Although some individuals were lost to follow-up, typically because of relocation or transfer to non-participating institutions, such occurrences were infrequent. Despite efforts to establish eligibility criteria controlling disease severity, there remains considerable heterogeneity among the individuals studied. The dataset encompasses variables that are anticipated to influence survival, including sex, tumor size and lymph node involvement classifications, age, general health status, and tumor differentiation grade. Patients meeting specific T and N staging criteria, as well as those with distant metastases, were excluded from the analysis.

Alongside assessing whether combined treatment surpasses radiation therapy alone, the study aims to discern how these factors impact survival and to correct survival rates for potential covariate imbalances. Such adjustments coordinate methods found in linear regression and covariance analysis, while accounting for the challenge of censoring. Given the absence of robust empirical or theoretical frameworks, the study underscores the necessity for nonparametric and robust analytical approaches to accommodate various failure time distributions. These considerations emphasize the complexity of survival analysis within clinical trials and the importance of comprehensive statistical methodologies to derive meaningful insights.

Variable Description

Case: Case Number
Inst: Participating Institution
Sex: 1=male, 2=female
Tx: Treatment: 1=standard, 2=test
Grade: 1=well differentiated, 2=moderately differentiated, 3=poorly differentiated, 9=missing
Age: In years at time of diagnosis
Cond: Condition: 1=no disability, 2=restricted work, 3=requires assistance with self care, 4=bed confined, 9=missing
Site: 1=faucial arch, 2=tonsillar fossa, 3=posterior pillar, 4=pharyngeal tongue, 5=posterior wall
T_Stage: 1=primary tumor measuring 2 cm or less in largest diameter, 2=primary tumor measuring 2 cm to 4 cm in largest diameter with minimal infiltration in depth, 3=primary tumor measuring more than 4 cm, 4=massive invasive tumor
N_Stage: 0=no clinical evidence of node metastases, 1=single positive node 3 cm or less in diameter, not fixed, 2=single positive node more than 3 cm in diameter, not fixed, 3=multiple positive nodes or fixed positive nodes
Entry_Dt: Date of study entry: Day of year and year, dddyy
Status: 0=censored, 1=dead
Time: Survival time in days from day of diagnosis

Exploratory Data Analysis

We'll embark on exploratory analysis to comprehensively understand the data collected from a clinical trial. Our primary focus will be on examining predictors and how they potentially influence the outcome variable of survival time. We aim to gain an understanding of the dataset's intricacies, paving the way for informed statistical modeling and hypothesis testing.

I. Institution, Treatment, and Gender Distribution

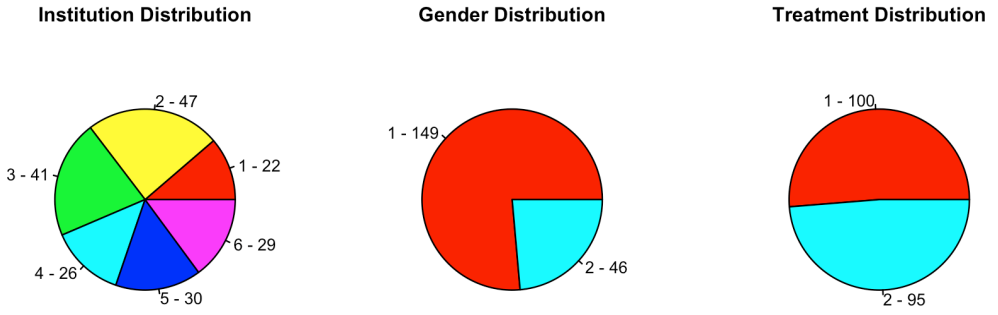


Figure 1: Pie Charts made in R

The graphs in Figure 1 show that females may be underrepresented compared to males within the study. This prompts the concern of why this is so. However, the distributions of institution and treatment distribution are evenly represented for a better sample. The approximately equal count of treatment types looks promising that it will yield significant results when comparing.

II. Age of Patients at Diagnosis

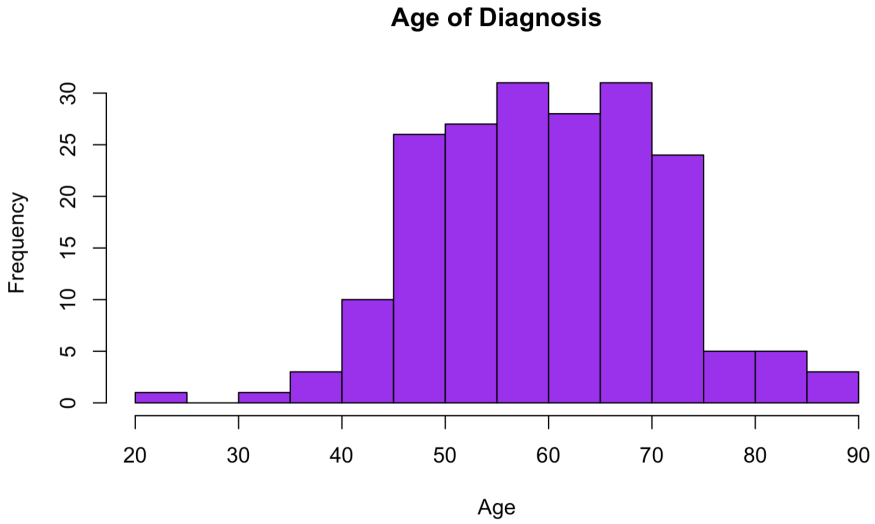


Figure 2: Histogram made in R

Shown in Figure 2, a majority of the patients were diagnosed between the ages of 45 and 75. This age distribution suggests that the condition may primarily affect individuals within this age range, highlighting the importance of screening and proactive measures for these individuals.

III. Censored Observations

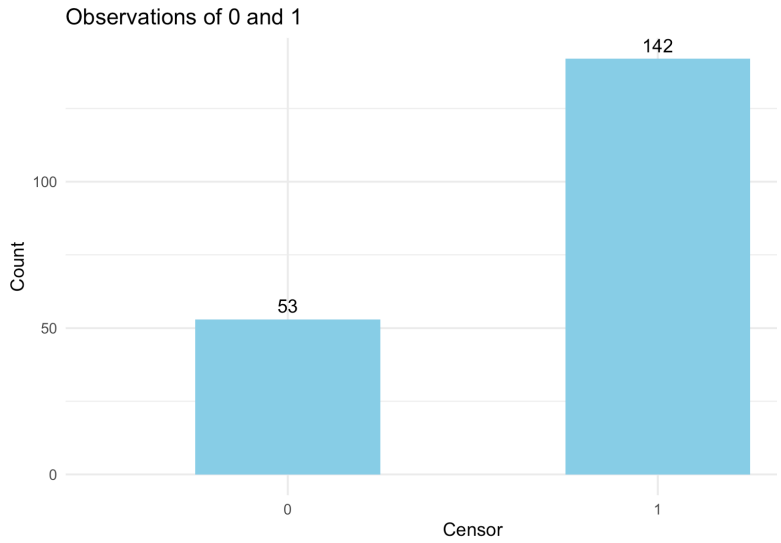


Figure 3: Bar Graph made in R

Given our background information on the data, it stated that censorship was infrequent. Alternatively, the bar graph shows that about a fourth of our observations are censored. That is not too many, but it will alter our approach to methodology. Censored observations can still be informative if using certain models accordingly.

IV. Time Plot

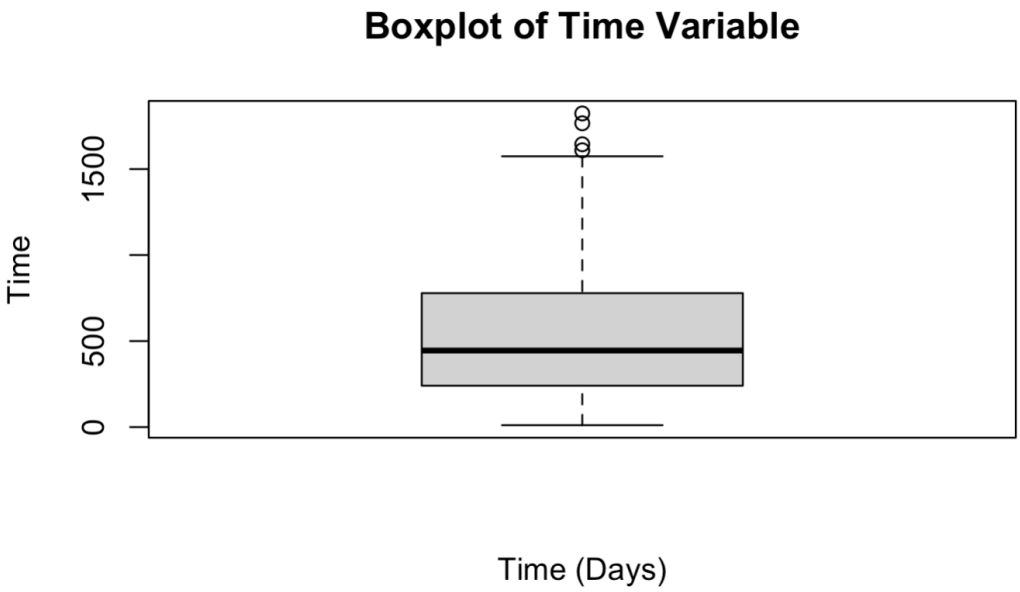


Figure 4: Box Plot made in R

The minimum value stands at 11 days while the maximum value stands at 1823 days. The interquartile range–which captures 50% of our patients' time of event–spans from the first quartile 240.5 days to the third quartile of 778.5 days. The median is 445.0, indicating that half of the observations fall below this value. Furthermore, the mean, calculated at 558.7 which is not an ideal summary measurement to focus on due to having censored observations in the data. These summary statistics collectively provide a quick glance of the survival time distribution.

Methods and Model Development

I. Kaplan-Meier Estimator

A Kaplan-Meier (KM) estimator–a non-parametric method–is carried out to estimate the survival function from censored data. The data is separated by treatment group. Recall that "Treatment 1" represents the patients who received radiation therapy alone, and "Treatment 2" represents the patients who received combined radiation therapy with a chemotherapeutic agent. The KM

survival curve shown in Figure 4 represents the probability of survival given how many days the patient has been monitored by the study per each treatment.

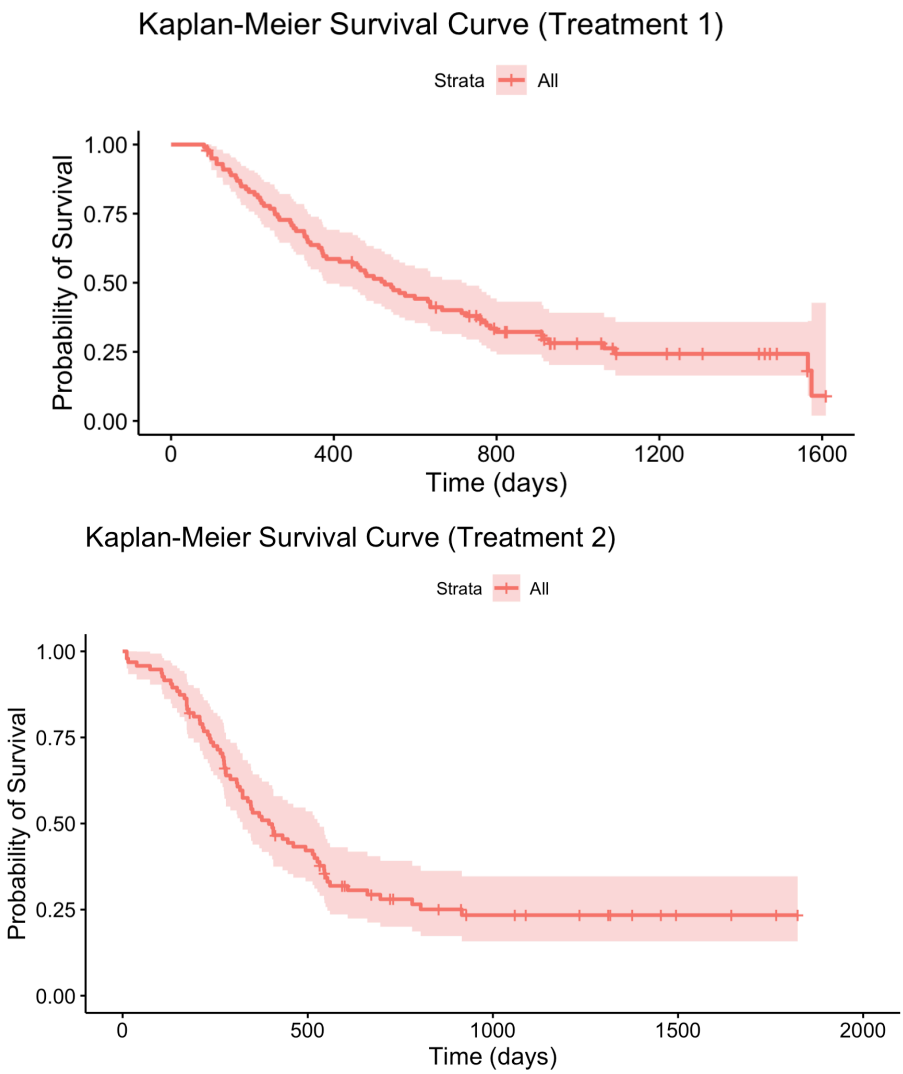


Figure 5: KM Curve made in R for each treatment

The results in Figure 5 show that the median survival time for standar treatment (Tx=1) is 525 days. At just below 1600 days, the survival rate drops below 25%. Alternatively, the median survival time for test treatment (Tx=2) is 395 days. This is much lower than the standard treatment.

Before moving onto more statistical and quantitative analysis, we note the difference in shape of the survival functions between these two groups. Treatment 1 appears to decrease linearly whereas Treatment 2 has a faster decrease early on before leveling-off. Where the slope is steepest, the hazard rate for experiencing an event in the given time frame will be higher than a time frame where the slope appears flatter. In other words, it appears the survival curve for

Treatment 2 (the test treatment) harms our participants' survival early on. This could be attributed to Treatment 2 being more aggressive with the inclusion of chemotherapeutic agents. This qualitative trait is of great importance when considering the quality of life of patients.

These qualitative observations allude to the standard treatment yielding higher survival times. As statisticians, we know that the trends observed above need to be explored more precisely.

I. Confidence Interval

Calculating 95% confidence intervals for the first year using Greenwood's Variance for both treatment cohorts, the results show…

Tx1 := (0.548, 0.739)
Tx2 := (0.439, 0.643)

Figure 6: Calculations in R

Therefore, it is 95% certain that the survival probability estimates at 1 year are (0.548, 0.739) for treatment 1 and (0.439, 0.643) for treatment 2. The confidence intervals have overlap. Therefore we cannot say for certain that one treatment is better than the other based on this information.

Calculating a 95% confidence interval using Log-Log Transformation for both treatment cohorts, the results show…

Tx1 := (0.381, 0.809)
Tx2 := (0.282, 0.729)

Figure 7: Calculations in R

Using the Log-Log transformation, it is 95% certain that the survival probability estimates at 1 year are (0.381, 0.809) for treatment 1 and (0.282, 0.729) for treatment 2. Again, The confidence intervals have overlap. Therefore, it is not certain that one treatment is better than the other based on this information.

Using the same methods to derive 95% confidence intervals for the second year, it yields the following results…

Tx1 := (0.305, 0.5)
Tx2 := (0.2, 0.392)

Figure 8: Calculations in R

Tx1 := (0.178,0.598)
Tx2 := (0.087, 0.516)

Figure 9: Calculations in R

Figures 8 and 9 represent the confidence intervals using Greenwood's variance and log-log transformation respectively. Thus, using the same interpretation as the confidence intervals for survival times in year 1, there is not significant evidence from the confidence interval that concludes a difference in treatments.

II. Log Rank Test

So far, there is a case to argue there is statistically significant evidence to support the no difference hypothesis regarding the survival functions of treatment vs. no treatment but only at specific times. However, are the differences over the entirety of the survival functions sufficient to now claim a difference in survival between treatment?

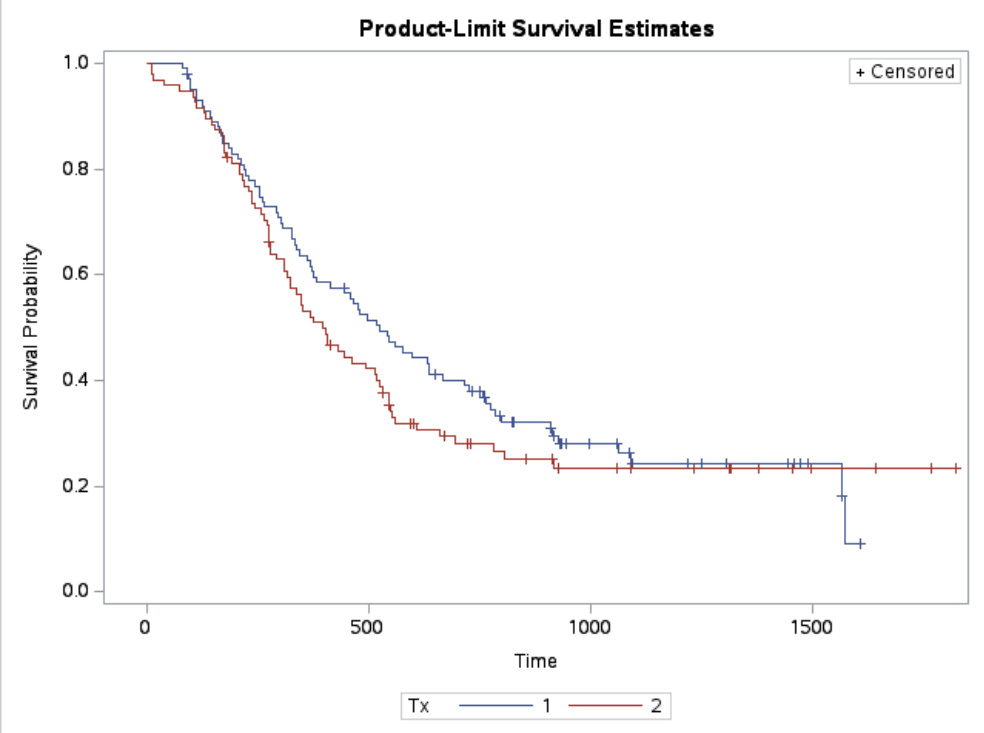


Figure 10: Overlay of KM Curve Stratified by Treatment

Consider the hazard, $\lambda_{Tx=i}(t)$, and survival functions, $s_{Tx=i}(t)$, of the treatment group i. We test the null hypothesis that these two groups have the same hazard function (equivalent to testing that the survival functions are the same). The results of the Log Rank Test comparing the two treatments is shown in Figure 11.

The LIFETEST Procedure
Testing Homogeneity of Survival Curves for Time over Strata

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.9264	1	0.3358
Wilcoxon	1.8095	1	0.1786
-2Log(LR)	0.7544	1	0.3851

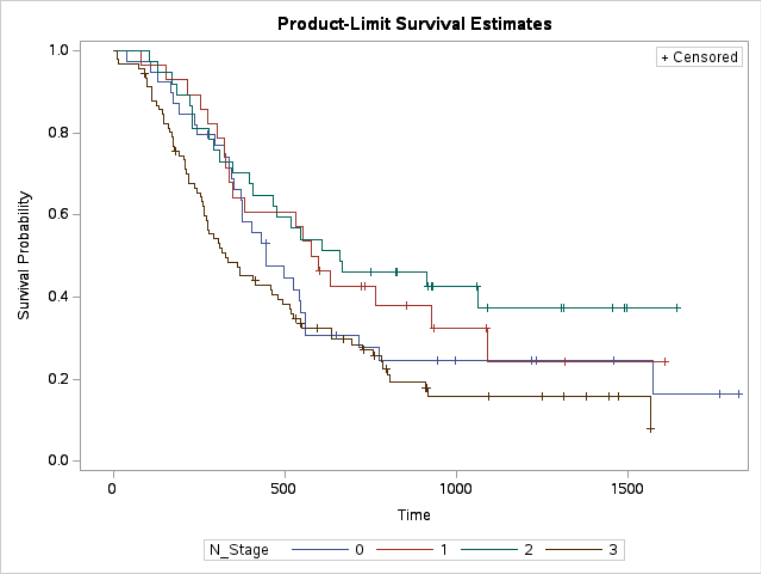
Figure 11: Summary Statistics of the Log Rank Test for Treatment vs. No Treatment

The large p-value indicates that any differences observed are not statistically significant; there is insufficient evidence that the survival and hazard functions are different

III. Stratified Log Rank Test

Recall that there are several potentially confounding variables such as N_Stage, a category quantifying the spread of cancer to nearby lymph nodes. To ensure that the lack of significance identified by the Log Rank Test, we will stratify to explore the possibility that there is a hidden treatment effect for different prognostic factors. For example, treatments may be significantly different for individuals in specific groups than in others. Consider N_Stage. This factor records the category describing the spread of the initial cancer to nearby lymph nodes: 0 is no evidence of cancer spread while a category of 1 to 3 indicates the level of lymph node metastasis. It makes sense that knowing the group a patient is in, with respect to N_Stage, provides insight on their survival since a prognosis at 3 is for a patient exhibiting metastases.

The reasoning just described is justified through Figure 12 showing a plot the KM curve with N_Stage as the strata. Figure 12 displays more variability between the KM curves coming from different values for N_Stage. By grouping our data properly, we see better discrimination between the treatments for some of the groups. Testing Homogeneity of Survival Curves for Time over Strata=N_Stage given by SAS is included as part of Figure 12. The test statistic corresponds to a small p-value (0.0122) indicating that this will be significant at the common 0.05 level of significance; there is evidence to contradict the hypothesis that the homogeneity of the survival/hazard curves.



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	10.9188	3	0.0122
Wilcoxon	11.7946	3	0.0081
-2Log(LR)	10.9196	3	0.0122

Figure 12: KM Survival functions & Stratified Log Rank Summary for Strata = N_Stage

Our data set contains several other variables over which the survivorship or treatment may be different just as in the case of N_Stage; the explanatory power in survival could explain the results observed for the treatment. Figure 13 demonstrates that consideration of a proper strata may better provide evidence of a difference in Treatment outcomes, a crucial fact to observe in our data if it is in fact present.

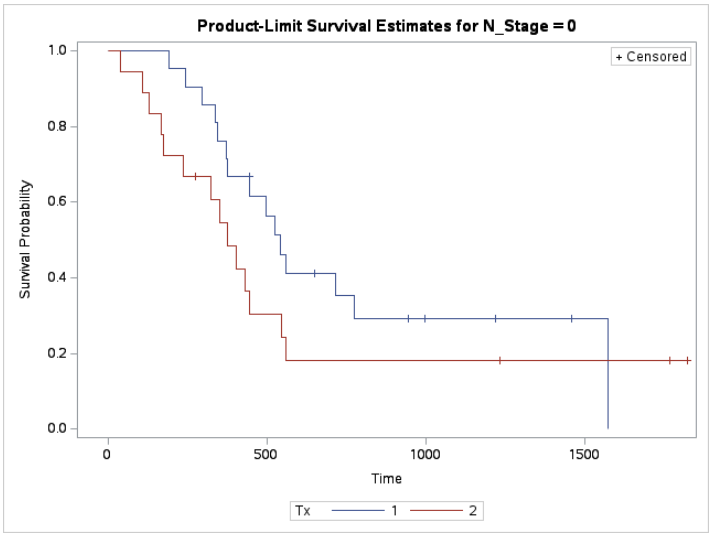
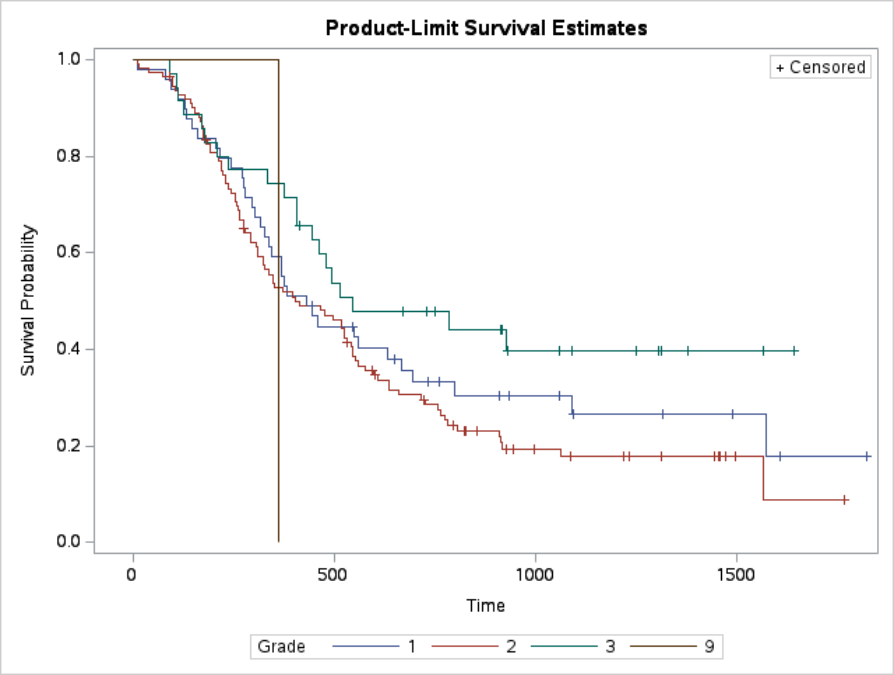


Figure 13: KM for Treatment when we Stratify by N_Stage

Knowing that there is a potentially different survival/hazard function explained solely by proper strata selection, requires exploring the remaining data sets and testing for Homogeneity over Strata. Due to the natural influence of certain variables, INST, SEX, GRADE, AGE, T_STAGE stand out as candidates suitable to explain differences.



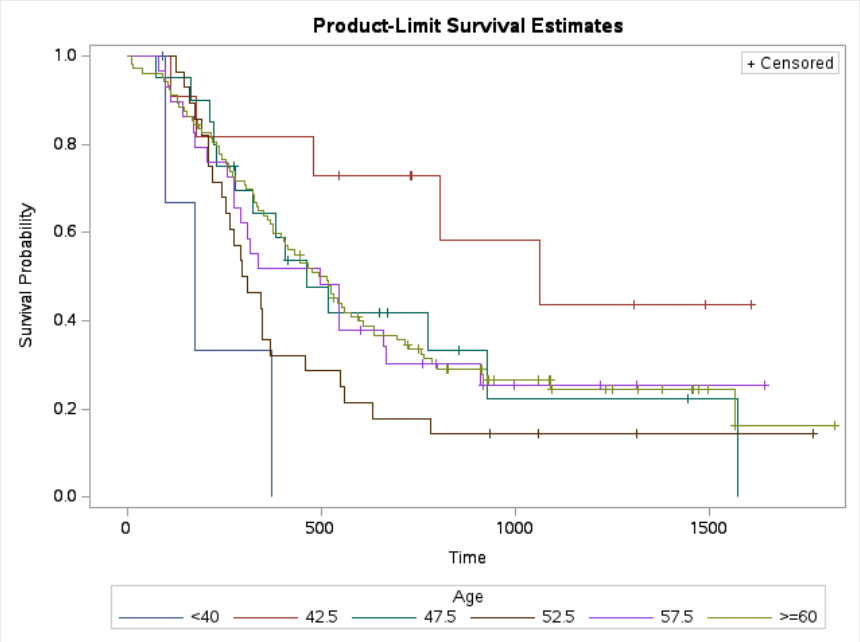
Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	5.5857	3	0.1336
Wilcoxon	3.1658	3	0.3668
-2Log(LR)	6.4115	3	0.0932

Figure 14: KM Survival functions & Stratified Log Rank Summary for Strata=Grade

Of special note were the results of the Grade Strata summarized in Figure 14. When Stratifying with Grade, there are certain data cleansing strategies to implement that may improve results. In the example of Grade, expulsion of recombination of certain strata, especially of grade=9 is justified.

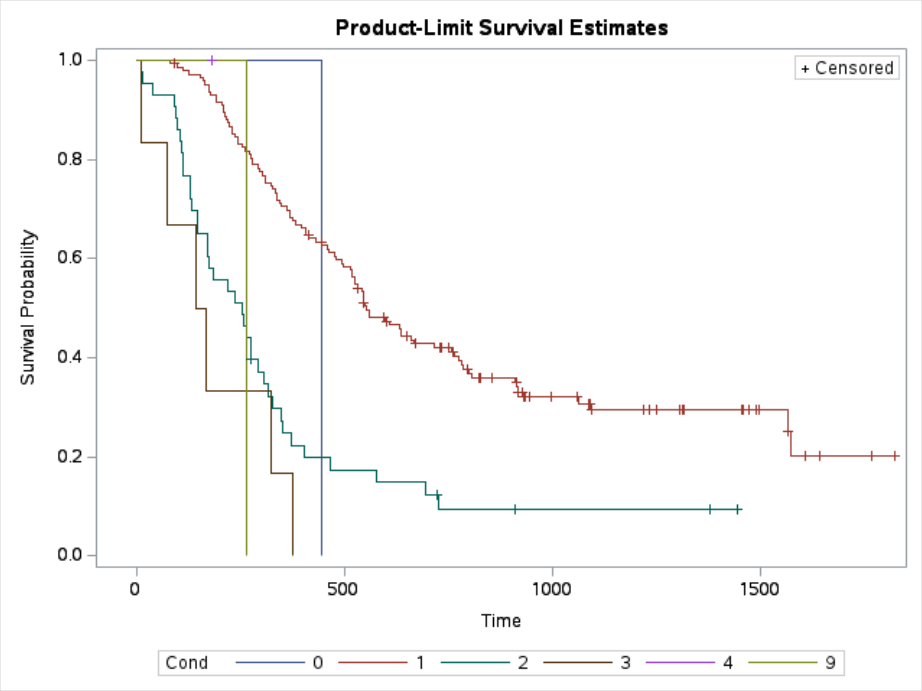
Another variable to note is AGE. As described this variable is one of the genuine quantitative variables in our data. However, to make use of the Cox Proportional Hazard model's interpretability of parameters, adjusting quantitative variables such as age may be helpful. We

consider this approach by stratifying age into categories, <40, [40,45), [45,50), [50,55), [55,60) >= 60. These strata are not unusual but some trial and error may lead to improvements in results.



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	12.8958	5	0.0244
Wilcoxon	11.8645	5	0.0367
-2Log(LR)	10.7058	5	0.0575

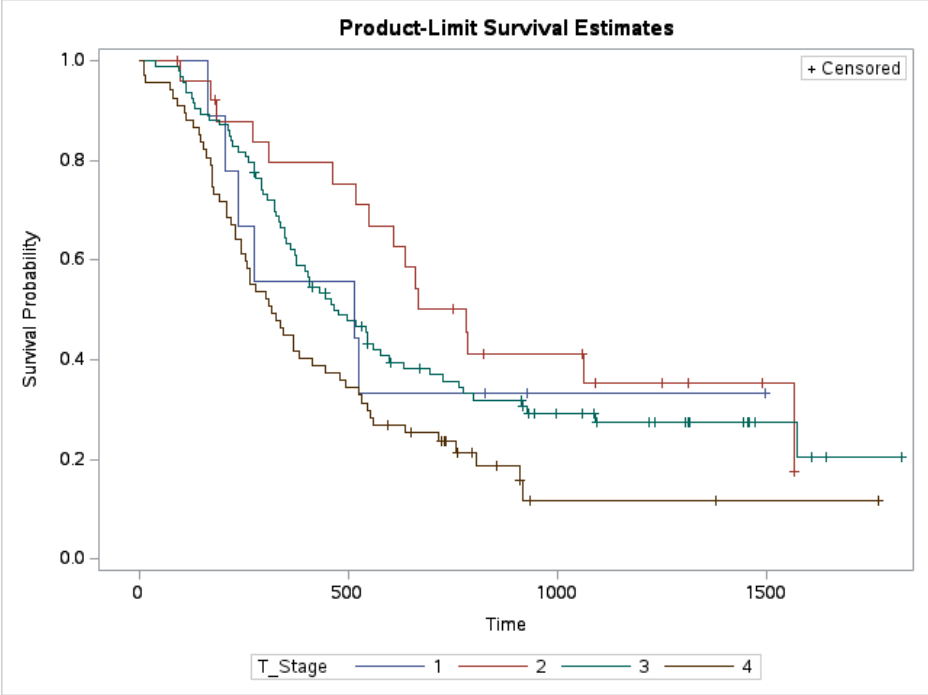
Figure 15: KM Survival functions & Stratified Log Rank Summary for Strata=Age



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	50.6832	5	<.0001

Figure 16: KM Survival functions & Stratified Log Rank Summary for Strata=Cond

Figure 16 gives the most significant p-value in support of differing survival functions using Cond, an ordinal measure of the degree of a patient's need for external support.



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	10.8580	3	0.0125
Wilcoxon	13.7376	3	0.0033
-2Log(LR)	12.2619	3	0.0065

Figure 17: KM Survival functions & Stratified Log Rank Summary for Strata=T_Stage

The final results are for strata of T_Stage. This result has the second lowest p-value from the Log-Rank Test.

Selection

The previous section emphasized the existence of covariates allowing significant changes in the survival curve of patients. This observation leads to consideration of predicting our patient's survival/hazard. The Log-Rank Test helped identify covariates with great discriminating power; aCox Proportional Hazards Model utilizing these covariates may prove useful.

IV. Constructing Cox Proportional Hazard Model

i. Proportional Hazard Assumption - Model 1

Before fitting a Cox Proportional Hazard (Cox PH) model, it's necessary to check the assumptions of proportional hazard rates. After fitting the model using all of the predictors in the data set, the proportional hazard function tests whether the hazard ratio for each predictor is

constant over time. Recall that hypothesis testing evaluates the p-value under the assumption that the null hypothesis is true. In this case, the null hypothesis is "the hazard rate for the predictor variable is constant over time." Therefore, if the p-value associated with the test for a predictor is less than 0.05 (our chosen alpha value), the null hypothesis will be rejected. This indicates sufficient evidence that there is a violation in the assumption and cannot be included in the final model. Results of this test are below.

	chisq	df	p
Inst	0.37565	1	0.540
Sex	0.03242	1	0.857
Tx	1.74422	1	0.187
Grade	0.00763	1	0.930
Age	4.06057	1	0.044
Cond	0.03812	1	0.845
Site	0.31281	1	0.576
T_Stage	1.18867	1	0.276
N_Stage	1.87343	1	0.171
Entry_Dt	1.23531	1	0.266
GLOBAL	9.71509	10	0.466

Figure 18: PH Assumption in R

The data in Figure 18 shows that the variable "Age" violates the PH assumption. It is essential to refit the model by stratifying the model by the variable "Age." Below is the PH assumption checked again with this adjustment.

	chisq	df	p
Inst	0.00609	1	0.938
Sex	0.01030	1	0.919
Tx	0.63335	1	0.426
Grade	0.04018	1	0.841
Cond	1.43493	1	0.231
Site	3.36886	1	0.066
T_Stage	1.76655	1	0.184
N_Stage	1.21831	1	0.270
Entry_Dt	0.00068	1	0.979
GLOBAL	9.09805	9	0.428

Figure 19: PH Assumption in R

The results, in Figure 19, now suggest that the PH assumption holds true for all predictors in the Cox PH model based on the p-values.

ii. Summary & Results - Model 1

The summary of the stratified Cox PH model is below.

```
               coef  exp(coef)   se(coef)      z Pr(>|z|)
Inst     -1.261e-02  9.875e-01  7.305e-02 -0.173    0.863
Sex      -3.842e-01  6.810e-01  2.810e-01 -1.367    0.172
Tx        2.136e-01  1.238e+00  2.290e-01  0.932    0.351
Grade    -1.550e-01  8.564e-01  1.817e-01 -0.853    0.394
Cond      1.076e+00  2.932e+00  2.390e-01  4.500 6.79e-06 ***
Site     -9.837e-02  9.063e-01  9.260e-02 -1.062    0.288
T_Stage   2.545e-01  1.290e+00  1.585e-01  1.606    0.108
N_Stage   1.397e-01  1.150e+00  1.096e-01  1.275    0.202
Entry_Dt -4.957e-06  1.000e+00  1.172e-05 -0.423    0.672
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
          exp(coef) exp(-coef) lower .95 upper .95
Inst         0.9875     1.0127    0.8557     1.139
Sex          0.6810     1.4684    0.3926     1.181
Tx           1.2381     0.8077    0.7903     1.940
Grade        0.8564     1.1676    0.5998     1.223
Cond         2.9316     0.3411    1.8351     4.683
Site         0.9063     1.1034    0.7559     1.087
T_Stage      1.2898     0.7753    0.9454     1.760
N_Stage      1.1500     0.8696    0.9276     1.426
Entry_Dt     1.0000     1.0000    1.0000     1.000

Concordance= 0.696  (se = 0.037 )
Likelihood ratio test= 33.06  on 9 df,   p=1e-04
Wald test            = 30.02  on 9 df,   p=4e-04
Score (logrank) test = 35.43  on 9 df,   p=5e-05
```

Figure 20: Cox PH Model in R

The model summary shows that the predictor “Cond” is significantly associated with the hazard rate of the event. Recall, the observation values indicate the patient’s condition with possible outcomes being 1 (no disability), 2 (restricted work), 3 (requires assistance with self care), 4 (bed confined), and 9 (missing/NA). In simpler terms…the higher the value, the worse the condition. Thus, for every one-unit increase in Cond, the hazard of the event increases by approximately 193.16% (HR = 2.9316).

The associated HR 95% confidence interval is (1.84, 4.68). This means that we are 95% confident that the true hazard ratio lies within this range of values.

Considering clinical implications, these findings highlight the importance of monitoring patient's conditions, especially if they progress to a more severe state that requires more restriction and assistance. The 193% increase in hazard rate is alarming for patients whose condition worsens. Healthcare providers should prioritize proactive strategies such as treatment plans and frequent evaluations. Ensuring a patient's health does not deteriorate into disabilities is absolutely crucial.

iii. Summary & Results - Model 2

Without further processing of features, the only significant predictor identified is "Cond." By isolating this predictor, healthcare providers and researchers can delve deeper into its impact on hazard rates without the confounding influence of other variables. A model utilizing just this covariate is summarized in Figure 21 below.

```
        coef exp(coef) se(coef)     z Pr(>|z|)
Cond 0.32208   1.38000  0.06491 4.962 6.97e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
Cond      1.38     0.7246     1.215     1.567
```

Figure 21: Significant Model in R

With a significant hazard ratio of 1.38 and a 95% confidence interval spanning from 1.22 to 1.567, it's evident that worsening conditions substantially elevate the risk of adverse events. This insight emphasizes the importance of proactive management and early intervention strategies for patients with deteriorating health conditions. Additionally, by narrowing their focus, healthcare providers can speed up decision-making processes, allowing for more efficient strategization and prioritization.

iv. Proportional Hazard Assumption - Model 3

It is noticed that the values 0 and 9 are some of the observations in the column. With 9=missing data and ‘0’ most likely being a mistake because it is not defined in the background, we will remove those observations and run the model again to see how this impacts the results. The PH assumption for the updated column is passed. Assumption and model results for the model are in Figure 22..

```
        chisq df    p
Cond     2.17  1 0.14
GLOBAL   2.17  1 0.14
```

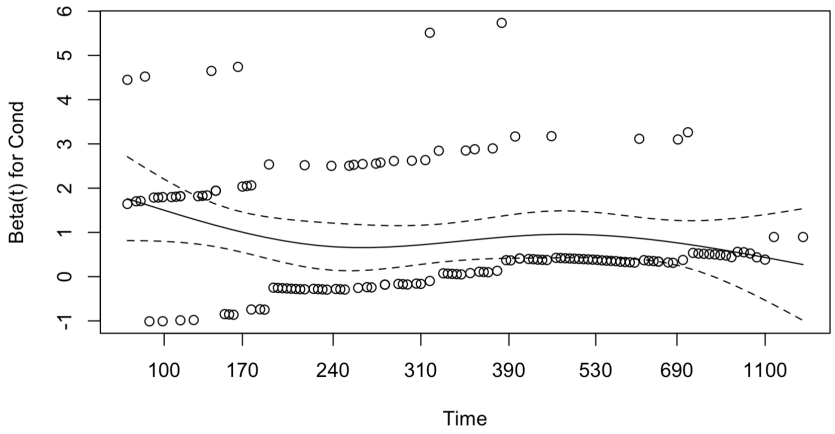


Figure 22: PH Assumption Check in R

v. Summary & Results - Model 3

```
       coef exp(coef) se(coef)     z Pr(>|z|)
Cond 0.8957    2.4490   0.1415 6.328 2.48e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
Cond     2.449     0.4083     1.856     3.232
```

Figure 23: Updated Observation Model in R

The new update yields aggressive results compared to the model with missing data. With 9=missing data in the previous model, those specific observations undermine the coefficient pertaining to how severely the condition impacts hazard rate. Data shows that for every one-unit increase in condition, the hazard of the event increases by approximately 145% (HR = 2.449). The 95% confidence interval for the hazard rate is (1.856, and 3.232) and therefore significant.

To Code or Not To Code, That is the Question…..
Our analysis thus far shows that an attempt to find a more meaningful model demands manipulation of the covariates. One initial change is to format our data to consist entirely of indicator variables; if there are p distinct values in one feature, we will generate $p - 1$ indicators (where one distinct value will serve as a reference).

vi. Summary & Results - Model 4

Initial screening signaled Cond as *the* variable to use.Coding the data into two variables Cond1 and Cond2 since we will use any value different from 1 or 2 as our reference. However, since Cond is already recorded as an *ordinal* variable, with larger values quantifying a greater extent of degrading condition, we question whether this is truly necessary. We examine a model in which Cond is used as given in our data (no coding) against a model coding the data.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1324.530	1310.178
AIC	1324.530	1312.178
SBC	1324.530	1315.134

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Cond	1	0.32215	0.06492	24.6269	<.0001	1.380	1.215	1.567

Figure 24 : Summary of CoxPH Cond variable as given

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1324.530	1290.599
AIC	1324.530	1294.599
SBC	1324.530	1300.510

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Cond1	1	-1.49163	0.37709	15.6470	<.0001	0.225	0.107	0.471
Cond2	1	-0.40483	0.39196	1.0668	0.3017	0.667	0.309	1.438

Figure 25 : Summary of CoxPH Cond variable Coded

With the aid of the different models shown in Figure 24 and Figure 25 we can compare the log likelihoods. With no coding, our likelihood is larger with 1310.178 vs 1290.599 in the coding model in Figure 25. Note that although AIC is lower with the coded model (1290.599 vs 1312.178 in the no coding model), under the coded model, Cond2 is not significant in the CoxPH

model! These two observations force us to leave the Cond data as given and utilize it as a quantitative variable rather than categorical.

Final (additions to the) Model

Having accepted Cond as given in the date, other variables (to potentially add to this mode) need to be constructed/coded.

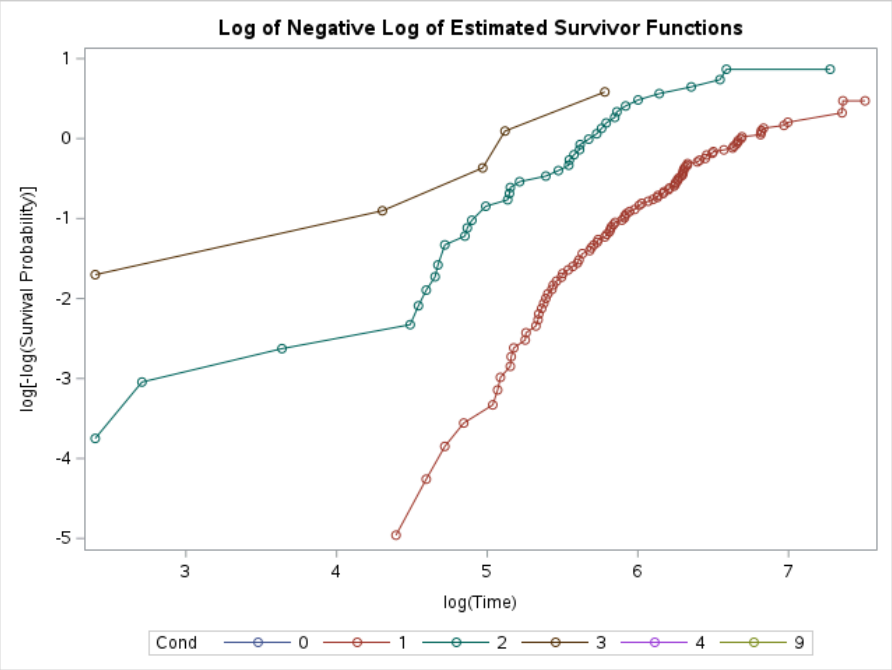


Figure 26 :

To start, recall the results of the Kaplan-Meier curves when stratified. Figures 12 -17 demonstrate significant differences in survival among the data under stratification. It stands to reason that including indicator variables will be significant when added to our model.

The first indicator variables considered to accept into a Cox PH Model are those coming from the T_Stage, N_Stage, and Grade covariates - these strata showed significant impact under the stratified Log Rank Test on survival curves mentioned earlier. To continue, code each of the different observations into the appropriate number of coding variables, demonstrated in the supplementary code, and apply a Forward-Step selection Process in SAS with default parameters. Deployment gives us the results in Figure 27.

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Cond	1	0.28672	0.06538	19.2308	<.0001	1.332
T_Stage4	1	0.47365	0.17730	7.1366	0.0076	1.606

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	T_Stage4		1	2	7.2636		0.0070

Obs	Step	Criterion	WithoutCovariates	WithCovariates
1	0	AIC	1324.530	1312.178
2	1	AIC	1324.530	1307.359

Figure 27 : Initial Forward Selection Output, required Cond for all models

The default Forwards Selection requires a significance value more significant than 0.05 to be included ensuring that the most significant variables are identified. To verify that T_Stage4 is significant we adjust the forward selection process to: START with the first 2 variables in our model (Cond and T_Stage2) and proceed as normal. This adjusted forward selection process results in a model with T_Stage4 and removes T_Stage2. This adjusted forward selection also produces a model with Cond and T_Stage4.

To attempt to confirm that Cond and T_Stage4 should be the covariates in the Cox PH, perform another adjustment to the Forward selection process: START with the first 3 variables (Cond, T_Stage2, and T_Stage3) and the remaining variables are considered for inclusion ingot the model one-by-one. This change in initial model actually ends with this initialization - no other variable is significant when T_Stage2 is included with T_Stage3.

There are two models under consideration for adoption.We choose between these two outcomes by analyzing the model statistics with T_Stage4 included (shown in Figure28) and the model statistics with T_Stage3 and T_Stage3 shown in Figure 29.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1324.530	1303.359
AIC	1324.530	1307.359
SBC	1324.530	1313.270

Figure 28: Model statistics for the model including Cond and T_Stage4

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1324.530	1302.977
AIC	1324.530	1308.977
SBC	1324.530	1317.845

Figure 29: Model statistics for the model including Cond, T_Stage2, and T_Stage3

To select a model, we use Likelihood. Since the log likelihood criterion is higher with the single variable T_Stage4, we therefore accept T_Stage 4 into our model (AIC also supports choosing to accept T_Stage4 over taking both T_Stage2 and T_Stage3). It is worthwhile noting that Backward Elimination also produces a model with covariates in Figure 29.

The next model makes use of the quantitative variable Cond and the indicator variable T_Stage4. With this model accepted, we worry about the interaction between Cond and T_Stage4. To analyze the interaction, fit a Cox PH model with the new variable CondStage, the results of this potential final modal are shown in Figure 30.

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Cond	1	0.88630	0.16920	27.4381	<.0001	2.426	1.741	3.380
T_Stage4	1	1.37307	0.31537	18.9563	<.0001	3.947	2.128	7.324
CondStage	1	-0.68436	0.18896	13.1164	0.0003	0.504	0.348	0.731

Figure 30: Model statistics for Final Mode with Cond, T_Stage2, and T_Stage3

Results

Using the Kaplan-Meier estimator, Figure 5 depicts differences between standard (Tx=1) and test (Tx=2) treatments: median survival times are 525 days and 395 days, respectively. Treatment 1 shows a linear decline, while Treatment 2 initially declines rapidly before stabilizing, potentially indicating a higher early hazard rate, possibly due to its aggressiveness. The confidence intervals from the KM estimator did not indicate that there is a significant difference in survival times between the two treatments tested.

After fitting a Cox PH model, results yield that all predictors in the Cox PH model were found to uphold the PH assumption based on p-values and parallelism of the log likelihood curves. The summary of the Cox PH model indicates that the predictor "Cond" significantly influences event hazard rates, with a one-unit increase correlating to approximately a 145% rise in hazard (HR = 2.9316), accompanied by a 95% confidence interval of (1.856, and 3.232), signifying the range within which the true hazard ratio is likely to lie.

The final model indicates that the coding variable T_Stage4 has an even greater influence estimated at 3.95, providing an increase of risk several times over if this variable is deemed true. This fact is not surprising considering that T_Stage4 indicates a massive tumor. Consideration of the interaction between the patient's condition and the condition is also significant, providing us with even more predictive power. For quick comparison to earlier models, the Fit statistics we achieve are shown in Figure 30.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1324.530	1291.848
AIC	1324.530	1297.848
SBC	1324.530	1306.716

Figure 30: Model Fit for Final Mode with Cond, T_Stage2, and T_Stage3

Conclusion

The KM estimator findings hold implications for medical practice and healthcare decision-making, particularly in determining the optimal treatment strategy for patients. While Treatment 2 may show an initial aggressive response, its ultimate efficacy in prolonging survival remains comparable to standard treatment. This must lead into further investigation into the long-term outcomes and potential adverse effects associated with each treatment.

Results using a stratified Cox PH model confirmed that the condition predictor is significant and alarmingly impacts the hazard rate. For instance, for every one-unit increase in condition severity, the patient is nearly three times as likely to experience the event. This information is crucial for healthcare providers as it helps them better understand and anticipate the progression of medical conditions, allowing for more targeted interventions and improved patient care.

The final model clarifies that the condition of a patient impacts hazard rates without other variables. Its hazard ratio of 1.38 and 95% confidence interval (1.22 to 1.567) highlight increased risk with worsening conditions, stressing proactive management and early assessments. This focused approach accelerates the efficiency of patient care.

Such insights are invaluable for clinicians in weighing the benefits and risks to optimize patient outcomes and enhance overall quality of care. Almost predictably, the size of the tumor plays another significant role in determining patient outcomes. Though the final predictor variables are not unexpected, quantifying the risk for patience and identifying areas of concern for patients who meet the criteria of higher hazard risk may be better candidates for aggressive treatments moving forward.

Code

I. R Code

Import Data

```
library(readxl)
data <- read_excel("~/Desktop/pharynx.xls")
```

Figure 1: Pie Charts

```
patient_data <- data[, c("Inst", "Sex", "Tx")]

## pie chart
# Function to create pie chart for each group
create_pie_chart <- function(data, variable, title) {
  color_palette <- rainbow(length(unique(data[[variable]])))
  variable_counts <- table(data[[variable]])
  pie(variable_counts, labels = paste(names(variable_counts), "-", variable_counts),
      col = color_palette,
      main = title)
}

par(mfrow=c(1, 3))
create_pie_chart(patient_data, "Inst", "Institution Distribution")
create_pie_chart(patient_data, "Sex", "Gender Distribution")
create_pie_chart(patient_data, "Tx", "Treatment Distribution")
```

Figure 2: Histogram - Age

```
hist(data$Age,
     col = "purple",
     main = "Age of Diagnosis",
     xlab = "Age",
     ylab = "Frequency",
     border = "black",
     breaks = 10)
```

Figure 3: Bar Graph - Censored Observations

```
library(ggplot2)

# Count occurrences of 0 and 1 in the 'Time' column
counts <- table(data$Status)

# Convert counts to a data frame
counts_df <- as.data.frame(counts)
names(counts_df) <- c("Censor", "Count")

# Create the bar graph
ggplot(counts_df, aes(x = factor(Censor), y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.5) +
  geom_text(aes(label = Count), vjust = -0.5) +
```

```
  labs(x = "Censor", y = "Count", title = "Observations of 0 and 1") +
  theme_minimal()
```

Figure 4: Box Plot - Time

```
boxplot(data$Time,
        main = "Boxplot of Time Variable",
        xlab = "Time (Days)",
        ylab = "Time")
```

```
summary(data$Time)
```

Figure 5: KM Estimate - Treatment

```
library(survival)
library(survminer)
```

```
# Fit survival models for each treatment group
t1 <- survfit(Surv(Time, Status) ~ Tx, data = subset(data, Tx == 1))
t2 <- survfit(Surv(Time, Status) ~ Tx, data = subset(data, Tx == 2))

# Create Kaplan-Meier plots for each treatment group
ggsurvplot(t1, data = data,
           xlab = "Time (days)", ylab = "Probability of Survival",
           title = "Kaplan-Meier Survival Curve (Treatment 1)")
```

```
ggsurvplot(t2, data = data,
           xlab = "Time (days)", ylab = "Probability of Survival",
           title = "Kaplan-Meier Survival Curve (Treatment 2)")
```

```
surv_median(t1)
```

```
surv_median(t2)
```

Figure 6: Confidence Interval: First Year - Greenwood's Variance

```
## Tx=1
summary(t1, times = 12*30)    #### (0.548, 0.739)
```

```
## Tx=2
summary(t2, times = 12*30)    #### (0.439, 0.643)
```

```r
## s(t=360) estimate
st1=summary(t1, times = 12*30)$surv
st2=summary(t2, times = 12*30)$surv

## standard error
se1=((1/(log10(st1))^2)*(0.0483^2/st1^2))^(1/2)
## where (0.0483^2/st1^2))^(1/2) is
se2=((1/(log10(st2))^2)*(0.0516^2/st2^2))^(1/2)

## A
A1=1.96*se1
A2=1.96*se2

## log-log transformation
lb1=st1^(exp(A1))
ub1=st1^(exp(-A1))

lb2=st2^(exp(A2))
ub2=st2^(exp(-A2))

lb1
```

```
ub1
```

```
lb2
```

```
ub2
```

Figure 7: Confidence Interval: First Year - Log-Log Transformation

Note that $Var_G\hat{S}(t) \times \hat{S}(t)^2 = \sum_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right)$

```r
## s(t=360) estimate
st1=summary(t1, times = 12*30)$surv
st2=summary(t2, times = 12*30)$surv

## standard error
se1=((1/(log10(st1))^2)*(0.0483^2/st1^2))^(1/2)
## where (0.0483^2/st1^2))^(1/2) is
se2=((1/(log10(st2))^2)*(0.0516^2/st2^2))^(1/2)

## A
A1=1.96*se1
A2=1.96*se2

## log-log transformation
lb1=st1^(exp(A1))
ub1=st1^(exp(-A1))

lb2=st2^(exp(A2))
ub2=st2^(exp(-A2))

lb1
```

```
ub1
```

```
lb2
```

```
ub2
```

Figure 8: Confidence Interval: Second Year - Greenwood's Variance

```
## Tx=1
summary(t1, times = 24*30)    #### (0.305, 0.5)
```

```
## Tx=2
summary(t2, times = 24*30)    #### (0.2, 0.392)
```

Figure 9: Confidence Interval: Second Year - Log-Log Transformation

```
## s(t=360) estimate
st1=summary(t1, times = 24*30)$surv
st2=summary(t2, times = 24*30)$surv

## standard error
se1=((1/(log10(st1))^2)*(0.0493^2/st1^2))^(1/2)
## where (0.0493^2/st1^2))^(1/2) is
se2=((1/(log10(st2))^2)*(0.0516^2/st2^2))^(1/2)

## A
A1=1.96*se1
A2=1.96*se2

## log-log transformation
lb1=st1^(exp(A1))
ub1=st1^(exp(-A1))

lb2=st2^(exp(A2))
ub2=st2^(exp(-A2))

lb1
```

```
ub1
```

```
lb2
```

```
ub2
```

Figure 21: Cox Proportional Hazard Model - All Predictors

```r
cox <- coxph(Surv(Time, Status) ~ Inst+Sex+Tx+Grade+Age+Cond+Site+T_Stage+N_Stage+Entry_Dt, data=data)

## test proportional assumption
cox.zph(cox)
```

```r
cox_strat <- coxph(Surv(Time, Status) ~ Inst+Sex+Tx+Grade+strata(Age)+Cond+Site+T_Stage+N_Stage+Entry

cox.zph(cox_strat)

## summary of model
summary(cox_strat)
```

Figure 22: Cox Proportional Hazard Model - Significant Model

```r
final <- coxph(Surv(Time, Status) ~ Cond, data=data)

## test proportional assumption
cox.zph(final)

plot(cox.zph(final))

## summary of model
summary(final)
```

Figure 23: Cox Proportional Hazard Model - Significant Model/Remove NA

```r
data2 <- data[data$Cond != 0 & data$Cond != 9, ]

mod <- coxph(Surv(Time, Status) ~ Cond, data=data2)

## test proportional assumption
cox.zph(mod)

plot(cox.zph(mod))

## summary of model
summary(mod)
```

II. SAS Code

```
/*create dataset*/
data pharynx;
    input Case Inst Sex Tx Grade Age Cond Site T_Stage N_Stage
Entry_Dt Status Time;
    datalines;
1 2 2 1 1 51 1 2 3 1 2468 1 631
2 2 1 2 1 65 1 4 2 3 2968 1 270
3 2 1 1 2 64 2 1 3 3 3368 1 327
4 2 1 1 1 73 1 1 4 0 5768 1 243
```

5 5 1 2 2 64 1 1 4 3 9568 1 916
6 4 1 2 1 61 1 2 3 0 10668 0 1823
7 4 1 1 2 65 1 2 4 3 10768 1 637
8 4 1 2 3 84 1 4 1 3 12068 1 235
9 6 1 1 2 54 2 1 3 3 13368 1 255
10 3 1 1 2 72 2 4 2 2 15468 1 184
11 3 1 1 2 42 1 4 2 2 15468 1 1064
12 2 1 1 2 61 1 1 4 3 18268 1 414
13 3 1 2 1 71 1 2 3 1 18468 1 216
14 4 1 2 2 83 3 4 3 1 19068 1 324
15 2 1 1 3 43 1 2 4 3 20768 1 480
16 5 1 2 2 52 1 4 4 3 21768 1 245
17 4 2 1 3 68 1 4 2 3 22768 0 1565
18 6 1 2 2 69 1 1 3 0 23368 1 560
19 3 2 2 3 65 3 1 3 0 25968 1 376
20 5 1 1 2 58 1 2 4 3 28068 1 911
21 2 1 2 2 63 1 2 4 3 28068 1 279
22 4 1 1 2 59 3 2 4 3 28268 1 144
23 3 1 1 1 75 1 2 3 1 28268 1 1092
24 6 1 1 1 65 2 1 3 3 28968 1 94
25 4 1 2 3 41 1 2 4 3 29468 1 177
26 3 1 1 2 60 1 4 3 3 29868 0 1472
27 3 1 2 2 72 1 4 1 3 30468 1 526
28 5 1 2 2 51 1 1 4 3 30868 1 173
29 2 2 1 2 72 2 2 3 1 30868 1 575
30 6 1 1 2 49 1 4 3 2 31068 1 222
31 3 1 2 2 82 3 1 3 0 31868 1 167
32 2 2 1 2 64 1 1 2 3 32468 1 1565
33 4 2 2 2 57 2 2 4 3 33568 1 256
34 3 1 2 1 67 2 2 3 3 33368 1 134
35 6 1 2 2 65 2 1 3 0 33868 1 404
36 3 1 2 2 62 1 4 1 2 369 0 1495
37 2 1 1 2 49 1 4 1 3 769 1 162
38 5 1 1 2 60 1 4 3 3 969 1 262
39 3 1 1 2 75 2 2 3 3 1769 1 307
40 2 1 2 2 54 1 2 2 3 2469 1 782
41 3 1 2 2 59 1 4 2 2 2469 1 661
42 5 1 1 2 58 1 1 3 2 3569 1 546
43 3 1 2 2 50 1 1 4 0 4469 0 1766
44 2 1 1 1 60 1 1 3 0 4569 1 374
45 3 2 1 1 43 1 2 2 2 4969 0 1489
46 4 1 1 2 48 2 2 3 3 5169 0 1446
47 4 1 2 2 49 3 1 4 3 5669 1 74
48 3 1 1 1 44 1 1 3 1 2769 0 1609
49 2 1 1 1 77 1 1 4 1 8369 1 301

50 2 1 1 1 75 1 2 4 1 9369 1 328
51 3 1 1 1 54 1 1 3 3 11869 1 459
52 3 1 1 1 68 1 1 4 0 12569 1 446
53 6 1 2 3 58 1 4 3 2 12769 0 1644
54 2 1 2 3 66 1 2 4 3 12969 1 494
55 3 1 2 1 47 1 1 3 2 13269 1 279
56 5 1 1 2 60 1 4 3 2 13569 1 915
57 2 1 1 2 66 1 4 4 2 14369 1 228
58 3 1 1 3 51 1 1 3 3 15569 1 127
59 2 2 1 1 49 1 1 3 0 15669 1 1574
60 6 1 1 1 50 1 2 4 0 16669 1 561
61 2 2 1 1 52 1 4 4 3 16769 1 370
62 2 1 2 2 40 1 4 4 3 17869 1 805
63 4 1 2 2 69 1 1 3 3 19969 1 192
64 5 1 2 2 56 1 2 1 3 20469 1 273
65 5 2 2 3 70 2 4 4 3 20469 0 1377
66 3 2 2 3 47 1 4 3 2 23069 1 407
67 3 1 1 3 46 1 2 3 1 24569 1 929
68 3 1 2 1 53 1 4 2 3 26669 1 548
69 3 1 2 1 67 1 4 3 1 27969 0 1317
70 3 2 2 1 68 1 4 3 1 26869 0 1317
71 2 1 1 2 90 1 4 3 3 28069 1 517
72 3 2 1 3 44 1 4 3 2 28969 0 1307
73 5 1 2 2 48 1 1 4 2 29069 1 230
74 4 1 1 2 67 1 2 3 1 30469 1 763
75 5 2 1 2 58 2 4 4 3 30469 1 172
76 4 1 2 2 69 1 1 3 2 32869 0 1455
77 4 1 2 2 75 1 4 3 0 32869 0 1234
78 6 1 2 3 58 1 2 3 3 33069 1 544
79 3 1 1 1 72 1 4 3 3 33269 1 800
80 6 1 1 2 72 1 1 3 0 33569 0 1460
81 6 1 1 3 70 1 4 2 3 33669 1 785
82 6 1 1 2 71 1 2 4 0 34469 1 714
83 1 1 2 2 55 1 1 3 1 35369 1 338
84 3 2 2 1 73 1 1 3 0 36369 1 432
85 1 1 2 2 50 1 4 3 3 870 0 1312
86 6 1 2 2 63 2 1 3 0 4270 1 351
87 2 2 1 1 58 1 2 1 3 4470 1 205
88 1 2 1 2 56 1 4 3 0 4870 0 1219
89 6 2 2 2 62 3 4 4 3 4970 1 11
90 3 2 1 1 55 1 4 2 2 5470 1 666
91 2 1 1 1 50 2 2 4 3 5770 1 147
92 3 1 2 1 77 1 4 2 2 7870 0 1060
93 1 2 1 2 67 1 2 3 2 8270 1 477
94 3 1 1 3 53 1 2 3 2 9670 0 1058

95 2 1 2 3 55 1 2 2 2 11070 0 1312
96 6 1 2 1 71 2 2 3 3 11870 1 696
97 2 1 1 1 65 1 1 4 3 12470 1 112
98 1 2 2 2 50 1 2 4 3 13170 1 308
99 5 1 2 2 61 2 4 4 3 14470 1 15
100 5 1 2 1 72 2 1 4 0 14670 1 130
101 4 1 1 1 51 1 2 3 0 15270 1 296
102 4 1 1 2 59 1 4 3 3 15870 1 293
103 2 1 2 2 56 1 1 4 0 16070 1 545
104 3 1 1 2 61 1 1 3 1 16670 0 1086
105 1 1 1 3 61 1 2 2 3 17470 0 1250
106 3 1 2 2 68 2 1 3 3 18770 1 147
107 5 2 1 2 71 2 1 3 3 18970 1 726
108 2 2 2 2 57 1 2 2 2 19070 1 310
109 2 1 1 2 72 1 4 3 1 20570 1 599
110 3 1 1 2 55 1 2 3 0 21170 0 998
111 4 2 2 3 61 1 2 2 2 21970 0 1089
112 5 1 1 1 47 1 4 4 1 23170 1 382
113 4 2 1 3 66 1 2 3 2 24370 0 932
114 1 1 2 2 52 2 4 4 3 25170 1 264
115 1 1 2 1 61 2 4 4 3 25470 1 11
116 5 1 1 1 66 2 2 4 3 25870 0 911
117 2 1 1 3 64 2 4 4 3 28570 1 89
118 5 1 1 2 73 1 1 4 0 28770 1 525
119 2 1 2 2 67 1 1 3 3 31670 0 532
120 3 1 1 2 68 1 1 2 3 32770 1 637
121 6 1 1 3 58 2 2 3 3 33370 1 112
122 6 2 1 1 68 1 4 3 3 33670 0 1095
123 1 1 1 3 85 2 4 2 2 34170 1 170
124 4 1 1 2 74 1 1 3 0 34270 0 943
125 5 1 1 2 53 1 1 4 0 34370 1 191
126 6 1 2 2 60 1 2 1 2 34470 0 928
127 3 1 1 3 58 1 2 3 2 35570 0 918
128 3 1 1 2 66 1 1 1 2 36270 0 825
129 1 1 1 2 58 2 2 2 3 1271 1 99
130 2 1 1 2 39 1 4 3 3 1571 1 99
131 2 1 1 1 54 1 1 4 1 1871 0 933
132 6 1 2 3 49 1 2 2 3 2271 1 461
133 6 1 2 2 52 1 1 3 1 2671 1 347
134 1 1 1 2 35 2 4 3 0 3371 1 372
135 5 1 2 3 44 1 2 4 3 4371 0 731
136 5 1 1 9 81 1 4 3 3 4971 1 363
137 1 1 2 2 74 2 1 3 0 6771 1 238
138 4 1 2 2 65 1 4 4 3 7571 0 593
139 2 2 1 2 66 2 4 4 3 7771 1 219

140 3 1 1 2 74 2 4 3 2 8871 1 465
141 1 2 2 3 90 0 2 3 0 10571 1 446
142 2 1 2 2 60 1 1 4 1 11371 1 553
143 5 1 2 2 63 1 1 4 1 15371 1 532
144 2 2 2 2 61 1 4 4 1 15471 1 154
145 2 1 2 1 67 1 1 4 3 15971 1 369
146 4 1 1 2 88 1 2 3 0 16171 1 541
147 5 2 2 3 69 2 4 4 0 18371 1 107
148 2 1 2 2 46 1 1 4 1 18871 0 854
149 1 1 1 2 69 1 1 2 2 20171 0 822
150 6 2 1 2 48 1 2 3 0 20271 1 775
151 2 1 1 1 77 1 1 4 0 20271 1 336
152 6 1 2 3 69 1 2 1 3 20271 1 513
153 6 1 2 3 75 1 4 3 3 20971 0 914
154 5 1 1 2 71 1 4 4 3 21671 1 757
155 5 2 1 2 58 1 1 4 3 21871 0 794
156 3 1 2 2 66 2 2 3 2 22171 1 105
157 5 2 1 1 44 1 2 4 1 23771 0 733
158 1 2 2 2 59 1 1 3 1 25371 0 600
159 2 1 1 2 78 9 4 4 3 26371 1 266
160 2 2 2 1 58 2 1 4 3 27371 1 317
161 2 1 2 3 65 1 4 3 3 28071 1 407
162 6 2 2 2 53 2 4 3 2 28471 1 346
163 3 1 2 2 49 1 4 2 2 29471 1 518
164 1 1 2 2 65 1 2 3 2 29971 1 395
165 5 1 1 1 59 1 1 4 1 31471 1 81
166 6 2 2 2 79 1 4 2 2 31971 1 608
167 6 1 1 1 57 1 2 4 3 32171 0 760
168 6 1 1 1 54 1 2 4 0 32371 1 343
169 3 1 2 2 47 1 1 3 0 32671 1 324
170 1 1 1 2 68 1 1 4 1 33071 1 254
171 2 1 1 3 63 1 4 2 2 34071 0 751
172 2 1 1 3 72 1 2 3 3 34271 1 334
173 6 2 2 1 51 2 2 3 1 34771 1 275
174 5 2 2 1 43 1 2 3 3 1272 0 546
175 6 2 2 2 43 2 4 3 3 3572 1 112
176 1 1 2 2 65 4 4 2 3 4672 0 182
177 1 1 2 2 54 1 1 4 3 5472 1 209
178 6 1 2 3 50 1 2 4 3 5572 1 208
179 2 1 2 2 39 1 1 4 3 5672 1 174
180 2 1 1 1 46 1 1 4 0 5972 0 651
181 2 1 2 3 49 1 2 3 3 8072 0 672
182 1 2 2 2 52 2 1 3 2 8272 1 291
183 1 2 2 2 69 2 4 4 1 13671 0 723
184 4 2 1 2 55 1 2 3 0 14372 1 498

```
185 4 1 2 2 48 2 2 3 0 14372 0 276
186 1 1 1 2 20 1 4 2 3 15672 0 90
187 4 2 1 2 47 1 4 3 3 15772 1 213
188 5 2 2 2 67 2 4 3 0 20572 1 38
189 4 2 1 2 66 2 2 3 2 20772 1 128
190 2 1 1 1 60 1 2 3 0 20972 0 445
191 2 1 1 1 54 1 2 4 3 22772 1 159
192 5 1 2 2 54 1 1 3 3 24372 1 219
193 4 1 2 2 59 2 1 4 0 24872 1 173
194 5 1 2 3 47 1 1 3 3 27672 0 413
195 3 1 2 2 57 2 1 3 3 12371 1 274
;
run;

/*view dataset properly recorded  --- checked and Working
proc print data=pharynx;
*/


/* For analyzing survival estimates & plot - Log Rank Test*/
proc lifetest data=pharynx;
     time Time*Status(0);
     strata Tx;
     title 'Logrank test for pharynx data by Treatment';
run;


/* For analyzing survival estimates & plot - Log Rank Test*/
proc lifetest data=pharynx;
     time Time*Status(0);
     test Tx;
     strata T_Stage;   /* INPUT STRATA */
run;


/* For analyzing survival estimates & plot - Log Rank Test*/
proc lifetest data=pharynx;
     time Time*Status(0);
     test Tx;
     strata Age (35 to 60 by 10);   /* INPUT STRATA */
run;

/* For analyzing survival estimates & plot - STRATIFIED Log Rank
Test*/
proc lifetest data=pharynx;
```

```
      time Time*Status(0);
      STRATA  N_Stage / GROUP = Tx;    /* INPUT STRATA */
      title 'stratified logrank test for pharynx data by Treatment';
run;
```

INITIAL MODEL SELECTION/COMPARISON

```
/* RECODING */
data pharynx_new;
  set pharynx;
  if Tx = 1 then Tx =0;
  else Tx = 1;
  if Sex = 2 then Sex =0;
run;

data pharynx_new;
      set pharynx_new;
      Cond1 = (Cond eq 1);
      Cond2 = (Cond eq 2);

      T_Stage1 = (T_Stage eq 1);
      T_Stage2 = (T_Stage eq 2);
      T_Stage3 = (T_Stage eq 3);
      T_Stage4 = (T_Stage eq 4);

      N_Stage1 = (N_Stage eq 1);
      N_Stage2 = (N_Stage eq 2);
      N_Stage3 = (N_Stage eq 3);

      Grade2 = (Grade eq 2);
      Grade3 = (Grade eq 3);
      run;


proc phreg data=pharynx_new;
      model Time*Status(0) = Cond/ RL;
run;

proc phreg data=pharynx_new;
      model Time*Status(0) = Cond1 Cond2 / RL;
Run;
```

FORWARD SELECTION - reference: http://www.math.wpi.edu/saspdf/stat/chap49.pdf

```
/* MODEL SELECTION */
ods output ModelBuildingSummary=Summary;
ods output FitStatistics=Fit;

proc phreg data=pharynx_new;
     model Time*Status(0) = Cond T_Stage2 T_Stage3 T_Stage4
N_Stage1 N_Stage2
          Grade2 Grade3/ RL selection=stepwise start=3
          include= 1;
run;
```

TESTING FOR INTERACTION

```
data pharynx_new;
     set pharynx_new;
     CondStage = Cond*T_Stage4;
     run;

proc phreg data=pharynx_new;
     model Time*Status(0) = Cond T_Stage4 CondStage/RL;
run;
```