Survival Analysis of "VA Lung Cancer Data"

Sabrina Enriquez

University of California- Davis

Fall 2020

Author Note

Correspondence concerning this article should be addressed to Sabrina Enriquez, Department of Mathematics, University of California, Davis. Address: 1 Shields Ave, Davis, CA 95616. E-mail: seenriquez@ucdavis.edu

Abstract

In this analysis we will be using “VA lung cancer data" which can be found in *The statistical analysis of failure time data* (Kalbfleisch, 1980). We are using methods of survival analysis to determine how covariates are related to survival outcome. This data has N=137 patients and 8 covariates. This data comes from a clinical trial titled “Veteran’s Administration Lung Cancer Trial", which consisted of patients with advanced, inoperable lung cancer who were treated with chemotherapy and standard treatment. Their outcomes were documented to assess the efficacy of chemotherapy for this class of lung cancer. Using various methods learned in BST 222 we will analyze the data for trends, investigate associations between covariates and survival, and report our findings.

Survival Analysis of "VA Lung Cancer Data"

Introduction

The Veteran's Administration Lung Cancer Trial was conducted to study survival outcomes between veterans who received standard treatment: treatment =1 or test treatment (chemotherapy): treatment= 2."VA lung cancer data" can be found online[1] and is originally taken from *The statistical analysis of failure time data*, pages 223-224 (Kalbfleisch, 1980). The data has N=137 patients with advanced, inoperable lung cancer who were either in treatment group 1 or 2, and there are 8 total covariates. The covariates are as follows:

Variables

1. Treatment 1=standard, 2=test
2. Cell type 1=squamous, 2=small cell, 3=adeno, 4=large
3. Survival in days
4. Status 1=dead, 0=censored
5. Karnofsky score (measure of general performance, 100=best)
6. Months from Diagnosis
7. Age in years
8. Prior therapy 0=no, 10=yes

The event of interest is death and the survival time measures from the start of treatment to death in days with random censoring on the right for those who left the study before death. The time origin is from the first treatment and time scale is in days on study. We have the following driving questions:

[1] http://www.stat.rice.edu/~sneeley/STAT553/Datasets/survivaldata.txt

1. Is standard or test treatment more successful in extending survival time?
   - H_0: There is not a significant difference in survival outcomes between the treatment groups. $S_1(t) = S_2(t)$ for all t< τ= final survival time.
   - H_A: There is a significant difference in the survival outcomes between the treatment groups. $S_1(t) \neq S_2(t)$ for all t< τ= final survival time.
2. What relationship does each covariate have with survival status?
   - For each covariate the null hypothesis is H_0: there is no significant association with survival outcomes.$HR_i = 1$ for all i covariates where HR denotes the hazard ratio for a given covariate.
   - For each covariate H_A: there is a significant association between some covariate and survival outcomes.$HR_i \neq 1$ for some i covariates where HR denotes the hazard ratio for a given covariate.
3. If we group participants by time from diagnosis will we find any survival patterns?
   - H_0: If we analyze survival for patients diagnosed before the median diagnosis time and after the median diagnosis time separately, we will observe no significant difference in survival patterns from the whole population. Denote the survival of those diagnosed before the median time as $S_{M1}(t)$ and those after as $S_{M2}(t)$ at time t. $S_{M1}(t) = S_{M2}(t) = S(t)$ for all $t < \tau$.
   - H_A: If we analyze survival for patients diagnosed before the median diagnosis time and after the median diagnosis time separately, we will observe a significant difference in survival patterns from the whole population.$S_{Mi}(t) \neq S(t)$ for some $t < \tau$ and $i \in \{1, 2\}$.

These hypotheses reflect the trial's focus on determining the efficacy of chemotherapy and lead us to use survival analysis methods learned in BST 222. In particular, we hope to find prognostic indicators by answering these driving questions.

Background

The following definitions and concepts will be utilized throughout this project and all definitions are sourced from *Survival Analysis Techniques for Censored and Truncated Data* by Klein and Moeschberger, 2003 unless otherwise stated.

Definition. Survival Function: The basic quantity employed to describe time-to-event phenomena is the survival function, the probability of an individual surviving beyond time x (experiencing the event after time x). It is defined as

$$S(x) = Pr(X > x).$$

Definition. Hazard Function: This function is fundamental in survival analysis and is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate. The hazard rate is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

If X is a continuous random variable, then,

$$h(x) = f(x)/S(x) = -d \log[S(x)]/dx.$$

Definition. Cumulative Hazard Function: The cumulative hazard function tells us the total risk that has been accumulated at time x and is defined by

$$H(x) = \int_0^x h(u)du = -\log[S(x)]$$

Definition. Kaplan- Meier estimator: The standard estimator of the survival function, proposed by Kaplan and Meier (1958), is called the Product-Limit estimator. This estimator is defined as follows for all values of t in the range where there is data:

$$\hat{S}(t) = \begin{cases} 1 & t \leq t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & t_i \leq t \end{cases}$$

where d_i is number of deaths and Y_i is the number of individuals at risk at time t_i

Definition. Cox Proportional Hazards Model (Qi, 2020a): Cox(1972) proposed to model the hazard function by

$$h(t|\mathbf{Z}) = h_0(t)c(\beta^t\mathbf{Z}) = h_0(t)\exp(\sum \beta_k\mathbf{Z}_k)$$

where $h_0(t)$ is an arbitrary baseline hazard rate, $\beta = (\beta_1, ..., \beta_p)^t$ is a parametric vector and $c(\beta^t\mathbf{Z})$ is a known function. This is called a semi-parametric model because a parametric form is assumed only for the covariate effect and the baseline hazard rate is treated nonparametrically.

This model assumes

1. IID observations
2. Noninformative/independent censoring
3. Hazard ratio (HR) is independent of time
4. Hazard ratio for two Z's are proportional

The partial likelihood function is

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp[\sum_k \beta_k Z_{(i)k}]}{\sum_{j\in R(t_i)} \exp[\sum_k \beta_k Z_{jk}]}$$

where $R(t_i)$ is the risk set at time t_i which includes all individuals who are still under study at a time just prior to time t_i. Here we assume no ties between the event times and $Z_{(i)k}$ is the k-th covariate associated with the individual whose failure time is t_i.

Definition. Hazard Ratio (HR): The hazard ratio compares two hazard rates and the Cox-Mantel estimate of HR for two groups A and B is given by:

$$HR = \frac{H_A}{H_B} = \frac{O_A/E_A}{O_B/E_B}$$

where O_i is the observed number of events (deaths) in group i , E_i is the expected number of events (deaths) in group i, and H_i is the overall hazard rate for the ith group.

Definition. Stratified Cox Model (Qi, 2020b): When the proportional hazards assumption is not met it may be possible to stratify on that variable and employ the proportional hazards model within each stratum for the other covariates.

$$h_p(t|\mathbf{Z}) = h_{0p}(t)c(\beta^t\mathbf{Z}) = h_{0p}(t)\exp(\sum \beta_k\mathbf{Z}_k), p = 1, 2, ...G$$

where $h_{0p}(t)$ is an arbitrary baseline hazard rate for each p strata, $\beta = (\beta_1, ..., \beta_p)^t$ is a parametric vector and $c(\beta^t\mathbf{Z})$ is a known function. This model has the following assumptions:

1. Each of the G groups has it's own baseline hazard function.
2. But with same regression coefficients, i.e. hazard ratio same for each stratum.
3. No interactions between variables.

Hypothesis testing for the Cox PH model (Qi, 2020a)

The global null hypothesis is defined by $\beta = \beta_0$ and the alternative global hypothesis is $\beta \neq \beta_0$. We can test our hypothesis using the Wald test, likelihood ratio, and score test. These tests will allow us to determine the validity of our Cox model estimates.

1. **Definition.** Wald Test: based on the partial likelihood function, find the partial MLE $\mathbf{b}$, which has a p-variate normal distribution with mean β and variance-covariance estimated by the inverse of the information matrix $I^{-1}(\mathbf{b})$.

   $$\chi^2_W = (\mathbf{b} - \beta_0)^t\mathbf{I}(\mathbf{b})(\mathbf{b} - \beta_0)$$

   where the information matrix $\mathbf{I}(\beta)$ is defined by:

   $$I_{gh} = -\frac{\partial^2 \log L(\beta)}{\partial\beta_g\partial\beta_h}$$

   This statistic follows a chi-squared distribution with p degrees of freedom if H_0 is true for large samples. For a single covariariate it follows a standard normal distribution.

2. **Definition.** Likelihood ratio test: calculate the difference in -2log-likelihood

$$\chi^2_{LR} = 2[LL(\mathbf{b}) - LL(\beta_0)].$$

This also follows a chi-squared distribution with p degrees of freedom under H_0 for large n.

3. **Definition.** Score Test: the score function is defined as

$$U_k = \frac{\partial \log L(\beta)}{\partial \beta_k}.$$

The score test has $H_0 : \beta = \beta_0$ and the test statistic is

$$U(\beta_0)^t I^{-1}(\beta_0) U(\beta_0)$$

. This statistic follows the chi-squared distribution with pdf under the null hypothesis.

Methods

We begin by checking the data set for data quality control. We find that there is no missing data and only 6% of participants were censored. Furthermore, we do not have any suspicious data points that have values inconsistent with the rest. We compute the descriptive statistics and provide plots of our distributions for reference. First we notice that amongst our categorical variables: treatment type, cell type, prior therapy, and censor status we have distributions shown in Figure 1. Then we show the distributions of the continuous variables and interpret the basic statistics of our data. This step is important for familiarizing ourselves with the broader shapes of our data and ensuring our questions are appropriate to the data. Since this trial was conducted to analyze treatment type we see that the population is evenly split and we can continue to pursue our questions around treatment with certainty that the data is not skewed toward a treatment.

We will begin our analysis by estimating the survival curves and cumulative hazard functions for treatment type using Kaplan-Meier and Nelson- Aalen estimates respectively.

These estimators are appropriate since the data is right censored and will give a broad perspective of survival by treatment type. We will overlay the survival and cumulative hazard curves for both treatment types producing 2 plots showing how survival and hazard compare for each treatment. Depending on if the data satisfies the PH assumption, we will either conduct Cox PH analysis or use the stratified Cox model to compute the hazard ratio between treatment types.These methods are appropriate for answering our first question because they satisfy the assumptions for each method and we can reject or accept our null hypothesis using these results.

Then for each covariate we will plot their respective Kaplan Meier survival curves and inspect them to determine which covariates are suitable for Cox Proportional Hazard analysis. Cox PH analysis requires that the data satisfy the proportional hazard assumption and by checking if the KM curves intersect we can easily determine whether Cox PH model is appropriate for a given covariate. We will use the Cox Proportional Hazard model to answer the second inquiry: what relationships do each of the covariates have with survival? We will determine if a covariate has a significant relationship by a p-value threshold of 5% and perform forward selection to determine the final Cox PH model. If we notice that a variable meets our selection threshold but the PH assumption is not satisfied we will use the Stratified Cox PH model to adjust for the violation. Moreover, we will use judgement when it comes to certain variables, since as discussed in BST 222 there are certain covariates such as age that are always important to keep in this context, regardless of its p-value. We will address when and where we deviate from our forward selection protocol and present our final model along with our interpretation of the results. Included in that model, we will consider interactions between variables in an effort to find more predictive factors.

For the third question we will group the participants by "months from diagnosis" and using the observed distribution we will choose intervals to compare survival curves between the groups. It is appropriate to look at the distribution of the months from

diagnosis variable to choose the intervals, because "early" or "late" diagnosis only makes sense in the context of the 137 participants in the data. Thus, we choose to group according to the median time which is 5 months and create a dummy variable assigning all participants with diagnosis time ≤ 5 as "early" and the rest as "late". Using this new dummy variable we once again consider the survival and hazard curves and interpret the results of incorporating it into the Cox PH model.

Results

As outlined we start with KM survival curves and cumulative hazard functions for treatment type using Kaplan-Meier and Nelson- Aalen estimates respectively.

We used Breslow method for breaking ties in our KM estimation and these plots allow us to partially answer our first question.

Question 1:

First glance at figure 3 would lead us to believe that treatment 2 is worse with lower survival and higher hazard. Note the PH assumption is not satisfied so we need a stratified Cox model. Using the stratified cox model:

$$h_k(t, z_1, ..., z_p) = h_{0k}(t)e^{\beta_1 z_1 + ... + \beta_p z_p}, k = 1, ..., G$$

We find that treatment type is NOT significant for survival with likelihood ratio, score, and wald test p-values=0.928! This result is surprising considering the trial was designed around measuring the effects of the treatment type. We can now answer our first question: Is standard or test treatment more successful in extending survival time? Neither test is significantly more successful than the other in extending survival time. We accept the null hypothesis and continue with forward selection and consider our other categorical variables.

Question 2:

Now we analyze cell type for our model and once again see in figure 4 that this variable does not satisfy the PH assumption and we must use a stratified Cox model. This time our preliminary tests suggest that cell type is a significant variable. Using rank statistics from the log-rank, Wilcoxon, and -2Log(LR) tests we find their respective p-values to be P<.0001, P=.0002, and P<.0001. These test inspire us to continue our investigation of cell type as a prognostic factor.

Using a univariate Cox model we are able to estimate the β values for each cell type. Setting squamous cell type as reference we find that hazard increases in the following order: squamous, large, small cell, and adeno. We give the detailed output in Table 1. Likelihood ratio, Score and Wald tests for this univariate model are all P<0.0001 indicating that the estimates reported in the table fit the cox model well. Moreover, the AIC decreases from 1011.768 to 993.197 with cell type included as a covariate. This result leads us to continuing with cell type as a stratified variable for our final Cox Model.

We continue to add covariates to this model and consider age. Now that we are stratifying over cell type we run Cox PH analysis using the reference set shown in Table 2.

Our AIC increases from 678.283 to 680.133 when we add age to our model indicating we should not continue with it. However, here is where knowing the data becomes important and we need to exercise judgement. Although the model diagnostics tell us to throw age out of our model, we know that age is very important to consider when studying survival and therefore keep it in the model. Age does not bring down our accuracy very much and it is more important that we include variables that we know to be important if we want our results to be meaningful. Continuing with this analysis we find the hazard ratio for age to be 1.004 with p-value=0.7 and we recognize that it is an insignificant value.

Next we consider adding karnofsky score to our existing stratified Cox model. When we do this we analyze with the reference set shown in table 3.

First we look to our AIC score and find that the score decreases from 678.283 to

640.943! That decrease tells us that the model fit is improving and that including Karnofsky score is a good idea. Then looking at the analysis of maximum likelihood estimates, we find that the hazard ratio for Karnofsky score in our model including age and stratified over cell type is 0.964 with P<0.0001. It is clear that this covariate should be included, but the hazard ratio implies that at the reference Karnofsky scores the hazard is almost constant. However as we vary the scores we find that with lower Karnofsky score hazard increases and with higher Karnofsky score it decreases. This finding is consistent with our understanding of Karnofsky score being an indicator for better outcome if scores are high.

Next we consider adding diagnosis time to our model as a continuous variable. Once again we see the AIC of our model decrease from 678.283 to 642.923, but when we look at the p-value for the MLE we find that diagnosis time has p-value=0.8874 disqualifying it from our model which has cutoff set to $\alpha = 0.05$. We move on to considering prior therapy for our model.

When we seek to include prior therapy into our stratified cox model we first check its KM curves to see if it satisfies the PH assumption. In figure 5 we see that the PH assumption is not satisfied and therefore if we wish to include it into our model we will need to stratify again. However, the log-rank, Wilcoxon, and -2Log(LR) tests have P-values =0.4789, 0.8107, and 0.1615 which tell us that this covariate is insignificant. Therefore, we check how it does as a newly stratified variable just to make sure there is no relationship, but find that this covariate should, in fact, be excluded from our model. We observe that the survival curves for both prior groups are very close to each other and therefore it is unnecessarily complicating our model without increasing accuracy.

Finally we consider if there may be any interaction terms with treatment 2= test since it was the purpose of this study. We use Breslow method again for ties and include cell type in the Cox model to get heuristics. We get Tables 4 which show that treatment does not interact with any of the term in a significant way. The closest we get are Wald

test p-values 0.0961 for squamous cell type and treatment 2 and p-value=0.0718 for no prior therapy and treatment 2. However, those do not meet our threshold and we are able to conlcude that therapy does not interact with another covariate in a way that makes it a reliable prognostic factor becuase those p-values are greater than 0.05.

The final model according to forward selection is a Cox model stratified over cell type and adjusted for age and Karnofsky score.

The final model MLE analysis and reference set is shown in Table 5. We can interpret our findings to mean that cell type, age, and karnofsky score are the prognostic and potentially predictive factors for survival and reject our second null hypothesis.

Now we offer a response to our second driving question. Q: What relationship does each covariate have with survival time? A: Simply reading from our results we found that for adeno cell type, hazard increases with age after 57.4 and with karnofsky score lower than 58.1. For large cell type hazard increases after age 56.2 and k-scores lower than 65. For small cell those values are 59.875 and 53.54 respectively and for squamous those values are 58.45 and 60.857. Moreover, if only considering cell type hazard increases in the order of squamous, large, small cell, and adeno. That result is surprising since one would expect that small cell has better odds of survival, but we also note that the p-value for large cell is 0.4065 so that result is not reliable. The rest of the covariates are not significant factors for survival, including treatment as discussed in question 1. This implies that the test treatment is no better or worst than standard treatment.

Question 3

Finally, we separate the data into groups “early” diagnosis for patients with diagnosis time <=5 months and “late” for the rest (5 months is the median time) and look for trends but find none. We then plot their KM curves and show them in figure 6. We see that the curves largely coincide indicating no significant difference in survival outcome. We perform univariate log-rank, Tarone, Peto, Modified Peto, and Fleming(0,1) trend tests and

find no significant trends. We report our results in Table 6.

The answer to our final question is that there is not a better chance of survival for those diagnosed "early" or "late". This coincides with finding that diagnosis time is not significant and therefore is not predictive when model building.

Conclusion

We conclude this paper by reiterating our findings:

1. Is standard or test treatment more successful in extending survival time?
   (a) Not significant in predicting survival. We accept H_0.
2. What relationship does each covariate have with survival time?
   (a) Cell type, age, and karnofsky score are our final stratified Cox model variables. All other covariates are not significant and were rejected through forward selection criteria.
3. If we group participants by time from diagnosis will we find any survival patterns?
   (a) Consistent with diagnosis time not being significant, "early" and "late" diagnosis is not predictive.

References

Kalbfleisch, J. G. (1980). *The statistical analysis of failure time data*. Wiley.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis techniques for censored and truncated data* (Second).

Qi, L. (2020a). Sta/bst 222 survival analysis lecture 12.

Qi, L. (2020b). Sta/bst 222 survival analysis lecture 17.

Table 1

Univariate Cox model for cell type with squamous as reference.

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr >ChiSq	Hazard Ratio	Label
celltype	**adeno**	1	1.14137	0.29285	15.1904	<.0001	3.131	celltype adeno
celltype	**large**	1	0.23018	0.27731	0.6890	0.4065	1.259	celltype large
celltype	**smallcell**	1	0.99643	0.25355	15.4442	<.0001	2.709	celltype smallcell

Table 2

Reference Set of Covariates for Plotting

Stratum	celltype	age
1	adeno	57.407407407
2	large	56.222222222
3	smallcell	59.875
4	squamous	58.457142857

Table 3

Reference Set of Covariates for Plotting

Stratum	celltype	age	karno
1	adeno	57.407407407	58.111111111
2	large	56.222222222	65
3	smallcell	59.875	53.541666667
4	squamous	58.457142857	60.857142857

Table 4

MLE analysis of covariates interactions with treatment.

Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr >ChiSq	Label
age			1	-0.00478	0.01308	0.1332	0.7151	
trt	**2**		1	1.61488	1.51891	1.1304	0.2877	trt 2
age*trt	**2**		1	-0.01700	0.01868	0.8289	0.3626	trt 2 * age
karno			1	-0.02804	0.00910	9.4883	0.0021	
karno*trt	**2**		1	-0.00910	0.01112	0.6696	0.4132	trt 2 * karno
diagtime			1	0.00234	0.02043	0.0131	0.9087	
diagtime*trt	**2**		1	-0.00281	0.02305	0.0148	0.9030	trt 2 * diagtime
celltype	**adeno**		1	1.38987	0.45060	9.5141	0.0020	celltype adeno
celltype	**smallcell**		1	0.52179	0.35946	2.1071	0.1466	celltype smallcell
celltype	**squamous**		1	0.08623	0.40081	0.0463	0.8297	celltype squamous
celltype*trt	**adeno**	**2**	1	-0.94233	0.59117	2.5408	0.1109	celltype adeno * trt 2
celltype*trt	**smallcell**	**2**	1	0.11879	0.54085	0.0482	0.8262	celltype smallcell * trt 2
celltype*trt	**squamous**	**2**	1	-0.95350	0.57304	2.7686	0.0961	celltype squamous * trt 2
prior	**0**		1	-0.43544	0.32258	1.8221	0.1771	prior 0
prior*trt	**0**	**2**	1	0.89204	0.49548	3.2413	0.0718	prior 0 * trt 2

Table 5

Final stratified Cox analysis.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr >ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
age	1	-0.00854	0.00949	0.8100	0.3681	0.991	0.973	1.010
karno	1	-0.03656	0.00571	40.9441	<.0001	0.964	0.953	0.975

Reference Set of Covariates for Plotting

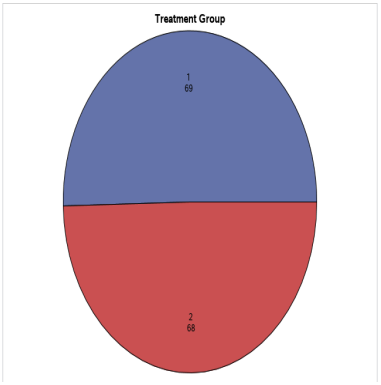
Stratum	celltype	age	karno
1	adeno	57.407407407	58.111111111
2	large	56.222222222	65
3	smallcell	59.875	53.541666667
4	squamous	58.457142857	60.857142857

Table 6

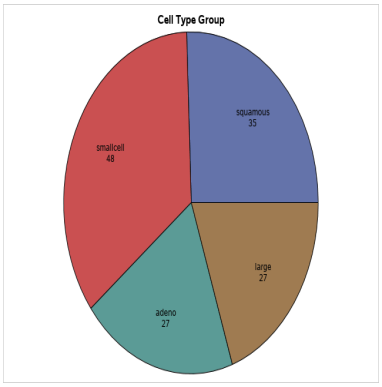
Trend tests for early and late diagnosis time.

Trend Tests

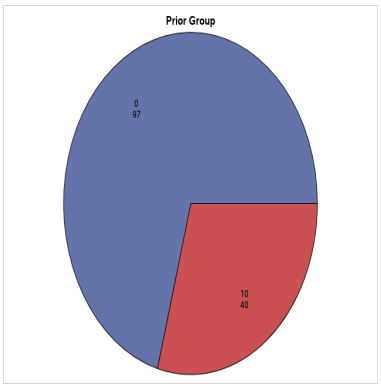
Test	TestStatistic	Standard Error	z-Score	Pr >|z|	Pr <z	Pr >z
Log-Rank	-1.9890	5.5481	-0.3585	0.7200	0.3600	0.6400
Tarone	-11.1325	46.9839	-0.2369	0.8127	0.4064	0.5936
Peto	-0.4139	3.3011	-0.1254	0.9002	0.4501	0.5499
Modified Peto	-0.4026	3.2653	-0.1233	0.9019	0.4509	0.5491
Fleming(0,1)	-1.5023	3.0000	-0.5008	0.6165	0.3083	0.6917



(a) *Treatment types are split evenly amongst patients.*



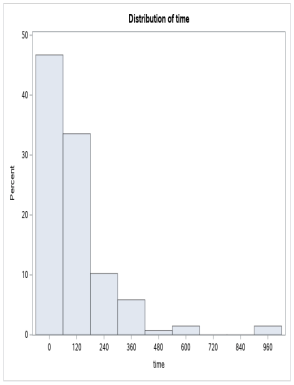
(b) *Cell types are distributed by: squamous 25%, small cell 35%, adeno 20%, and large 20%.*



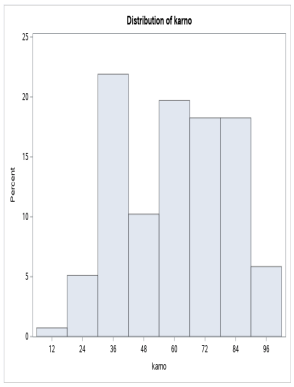
(c) *Prior therapy is distributed by: 0=no: 71%, 10= yes: 29%.*

Figure 1

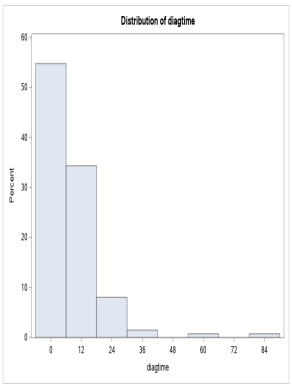
Distribution of categorical variables except status because only 6% were censored.



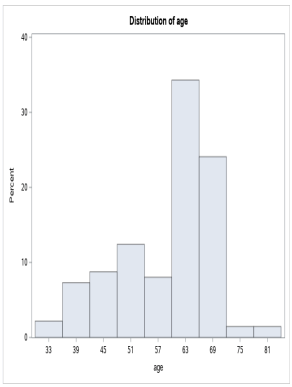
(a) *Most survival times occur before 120 days.*



(b) *There is much more variation in karnofsky scores with few at very low or very high scores.*



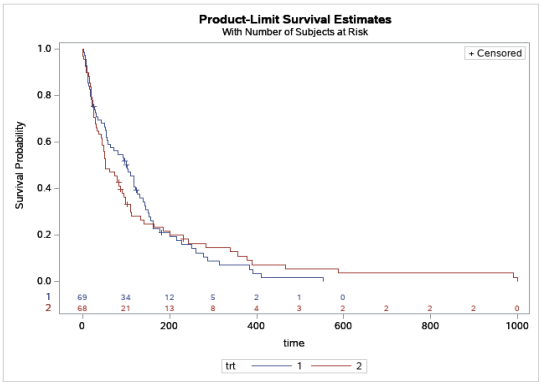
(c) *Most participants were diagnosed within the 6 months before the start of the trial.*



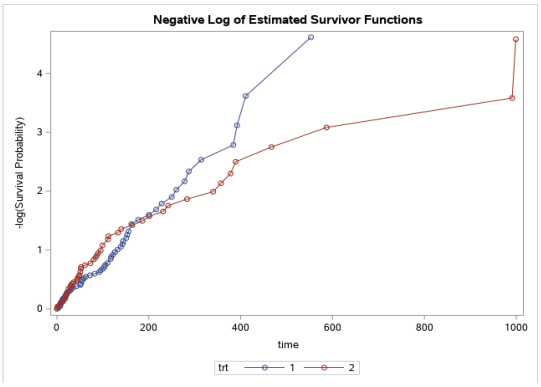
(d) *There is decent variation in age but the participants are mostly in their 60's.*

Figure 2

Distributions of continuous variables.



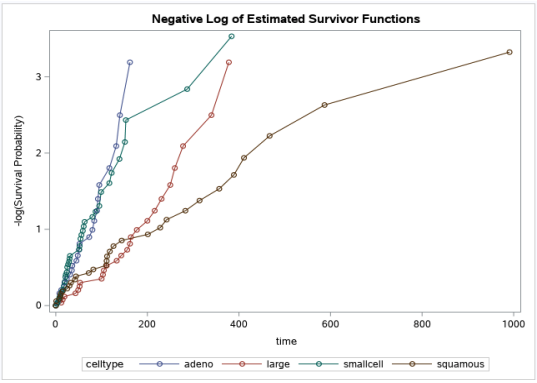
(a) *KM survival plot by treatment type.*



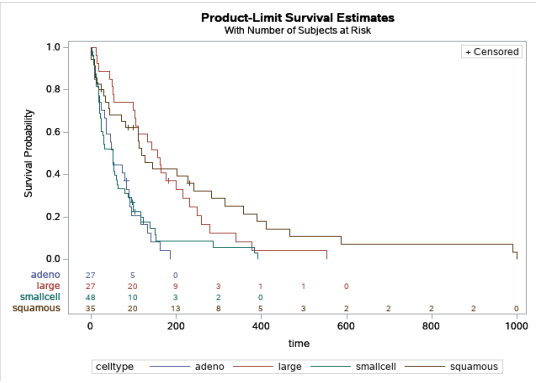
(b) *N-A Cumulative Hazard plot for treatment type.*

Figure 3

We notice that the treatment types are showing similar survival and cumulative hazard. They intersect one another thereby making them violate the PH assumption.



(a) *KM survival plot by cell type.*



(b) *N-A Cumulative Hazard plot for cell type.*

Figure 4

Observe once again that cell type intersect one another thereby making them violate the PH assumption. However, rank tests indicate that cell type is significant with P<.0001

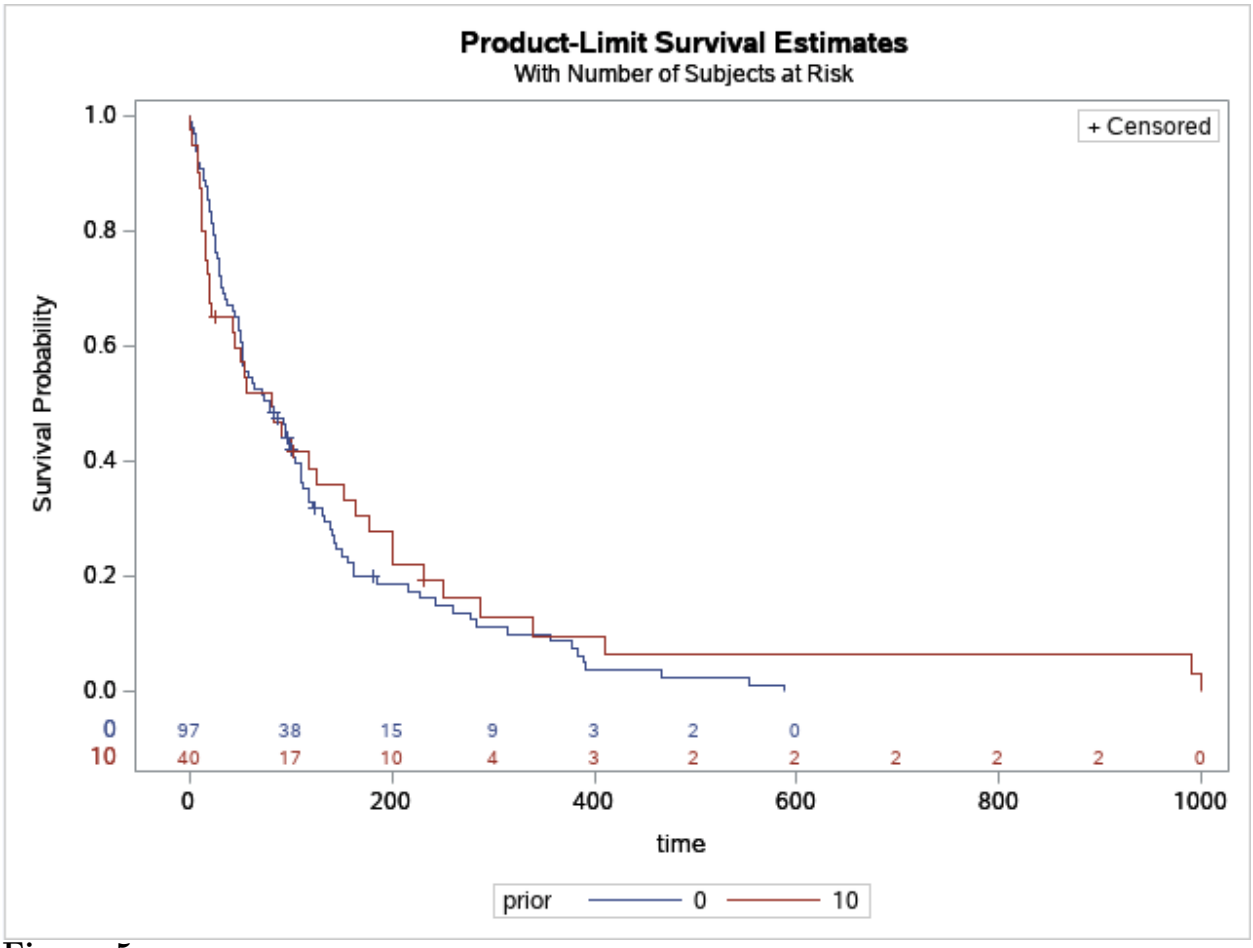


Figure 5

KM curve for prior therapy.

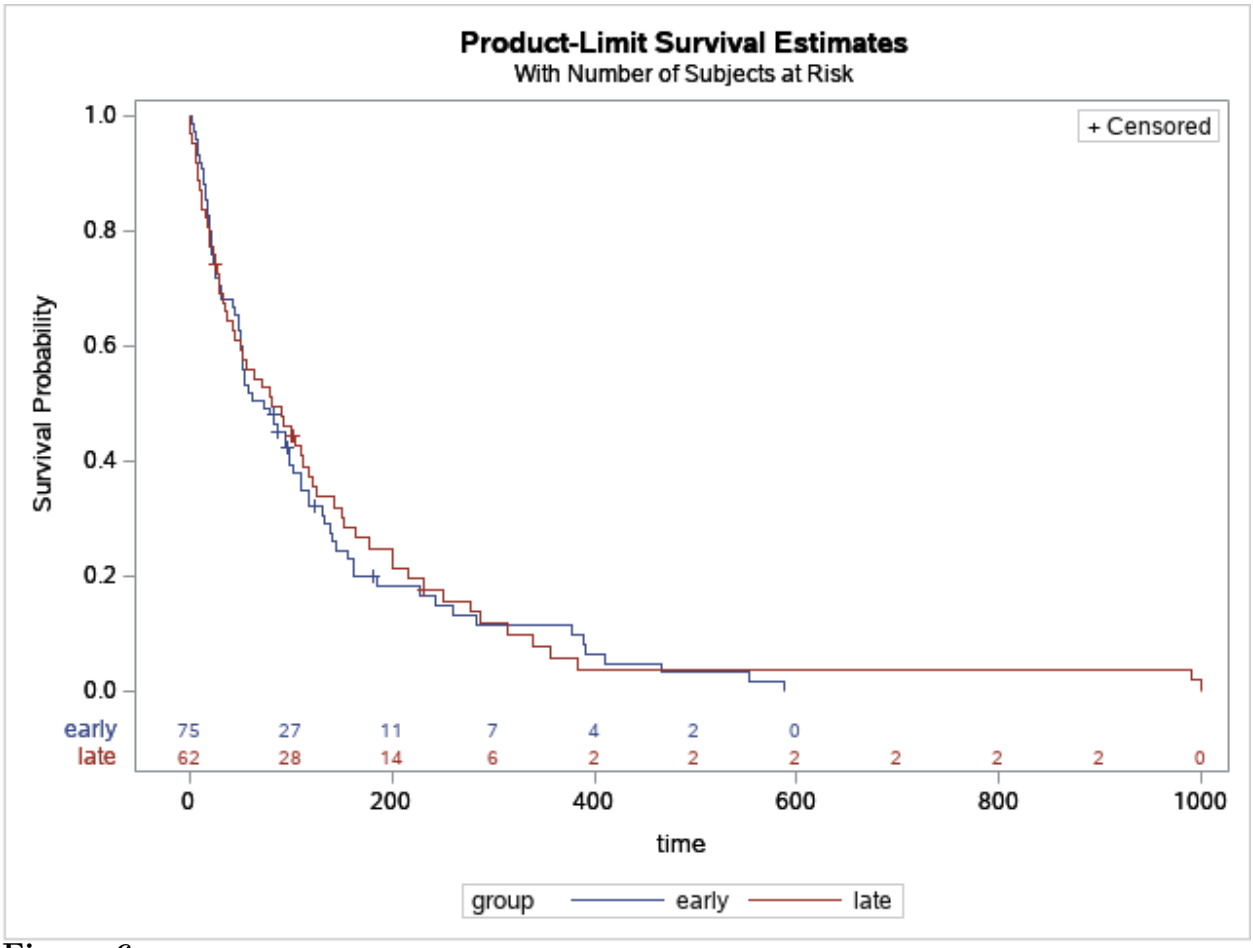


Figure 6

KM survival curves for patients with "early" and "late" diagnosis time.

```
/*Sabrina Enriquez project 1*/
LIBNAME p1 "/folders/myfolders/p1";
PROC import DATAFILE= "/folders/myfolders/p1/vaLung.csv"
    DBMS=csv
     out=p1.vets
        replace;

    GETNAMES=YES; /*Option GETNAMES determines whether to generate SAS variable names from the data values in the f
                   record of the imported file.*/
RUN;

*descriptive stats;
proc SGPLOT data=p1.vets;
   VBAR trt;
run;

proc SGPLOT data=p1.vets;
   VBAR karno;
run;

proc SGPLOT data=p1.vets;
   VBAR age;
run;

proc template;
define statgraph simplepie;
begingraph;
entrytitle "Treatment Group";
layout region;
piechart category=trt / datalabellocation=inside;
endlayout;
endgraph;
end;
run;
proc sgrender data=p1.vets
template=simplepie;
run;

proc template;
define statgraph simplepie2;
begingraph;
entrytitle "Prior Group";
layout region;
piechart category=prior / datalabellocation=inside;
endlayout;
endgraph;
end;
run;
proc sgrender data=p1.vets
template=simplepie2;
run;

proc template;
define statgraph simplepie3;
begingraph;
entrytitle "Cell Type Group";
layout region;
piechart category=celltype / datalabellocation=inside;
endlayout;
endgraph;
end;
run;
proc sgrender data=p1.vets
template=simplepie3;
run;


proc SGPLOT data=p1.vets;
   VBAR celltype;
run;

proc sgrender data=p1.vets
template=simplepie;
run;

proc univariate data=p1.vets;
   histogram;
run;
```

```
*KM plot;
ods listing gpath= '/folders/myfolders/p1/';
ods graphics / imagename= "lifetest" imagefmt=png;
proc lifetest data=p1.vets outs=p1.KM
plots=survival(atrisk= 0 to 1000 by 100) ;
time time*status(0); strata trt; run;

*Nelson-Aalen;
ods listing gpath= '/folders/myfolders/p1/';
ods graphics / imagename= "cumhaz" imagefmt=png;
proc phreg data=p1.vets
plots(overlay)=cumhaz ;
class trt;
model time*status(0)=trt; run;

*Nelson-Aalen;

proc lifetest data=p1.vets nelson method=breslow outs=p1.NA
plots=logsurv ;
time time*status(0); strata trt; run;




*K- sample test;
proc lifetest data=p1.vets
plots=(survival) ;
time time*status(0); strata trt /
test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;

proc lifetest data=p1.vets
plots=(hazard) ;
time time*status(0); strata trt /
test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;




*now trend test;
proc lifetest data=p1.vets
plots=survival(atrisk= 0 to 1000 by 100 ) ;
time time*status(0); strata trt /
trend test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;



*performing a proportional hazards regression with trt as the single covariate in the model;
proc phreg data =p1.vets
plots(overlay)=(survival);
class trt;
model time*status(0) = trt;
run;
*not significant

*adjusting for age;
proc phreg data= p1.vets plots(overlay) = (survival);
class trt;
model time*status(0) = trt age ;
run;

*
*
*
*
*
* objective 2 see how other covariates relate to survival;


*cell type ;

*KM plot;
ods listing gpath= '/folders/myfolders/p1/';
ods graphics / imagename= "lifetest" imagefmt=png;
proc lifetest data=p1.vets
plots=survival(atrisk= 0 to 1000 by 100) ;
time time*status(0); strata celltype; run;
```

```
proc lifetest data=p1.vets
plots=hazard ;
time time*status(0); strata celltype; run;

proc lifetest data=p1.vets
plots=logsurv ;
time time*status(0); strata celltype; run;

*doesn't pass ph assumption;
*performing a proportional hazards regression with celltype as the single covariate in the model;
proc phreg data=p1.vets plots(overlay) = (survival);
class celltype;
model time*status(0) = celltype;
run;




*adjusting cell type model for age;
proc phreg data= p1.vets plots(overlay) = (survival);
model time*status(0) = age;
strata celltype;
run;
*age not significant but still included;



*adjusting model for karno;
proc phreg data= p1.vets plots(overlay) = (survival);

model time*status(0) = age karno;
strata celltype;
run;
*karno is significant




*looking for karno hazard trend over values;
proc phreg data= p1.vets plots(overlay) = (survival);
class karno;
model time*status(0) =  karno;
run;
*karno is significant



*adjusting cell type model for diagtime;
proc phreg data= p1.vets plots(overlay) = (survival);

model time*status(0) = age karno diagtime;
strata celltype;
run;
*diagtime is not significant;



proc lifetest data=p1.vets
plots=survival(atrisk= 0 to 1000 by 100) ;
time time*status(0); strata prior; run;
*doesnt satisfy ph and is not significant.


*adjusting cell type model for strata prior just in case;
proc phreg data= p1.vets plots(overlay) = (survival);

model time*status(0) = age karno;
strata celltype prior;
run;


*prior is not significant;

proc phreg data= p1.vets plots(overlay) = (survival);
model time*status(0) =prior   ;

run;
```

```
*prio



*now we check for interaction terms- let's focus on trt since that was the purpose here;
proc phreg data=p1.vets;
   class prior celltype(ref='large') trt(desc);
   model time*status(0) = age|trt karno|trt diagtime|trt celltype|trt prior|trt
         / risklimits alpha=0.05;
run;

*final model is:;
proc phreg data= p1.vets plots(overlay) = (survival)  ;

model time*status(0) = age karno/ risklimits alpha=0.05;
strata celltype;
run;

*question 3 let's subset our diagnosis times to 2 groups from median = 5 months;


data p1.diag;
set p1.vets;
  if diagtime>5 then group = "late ";
  else group = "early";
run;


proc print data=p1.diag; run;


*now trend test;
proc lifetest data=p1.diag
plots=survival(atrisk= 0 to 1000 by 100 ) ;
time time*status(0); strata group /
trend test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;
```