

Veteran's Administration Lung Cancer  
Data Analysis Report

Sam Oliszewski

Oregon State University

## Introduction

In a study conducted by the US Veterans Administration, male patients with advanced inoperable lung cancer were randomly assigned to two treatments of either a standard therapy or a test chemotherapy. Time to death was recorded for 137 patients, while 9 left the study before death. Various covariates were also documented for each patient including: tumor cell type, Karnofsky performance score, time between diagnosis and start of study (in months), age of patient, and an indicator of prior therapy. This data set has been published in D. Kalbfleisch and R.L. Prentice (1980), *The Statistical Analysis of Failure Time Data*. Wiley, New York.

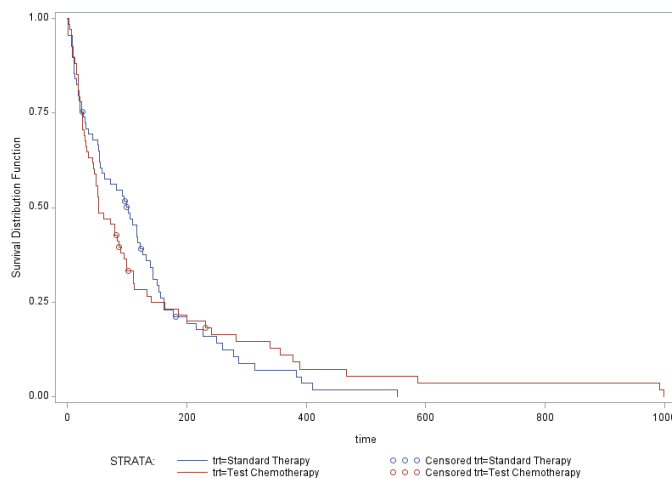
## Methods and Results

The primary goal of the study was to assess if the test chemotherapy is beneficial. Secondary goals included the analysis of the additional covariates as prognostic variables.

The dataset of interest contains the following variables:

- trt: treatment type; 1=Standard Therapy, 2=Test Chemotherapy
- celltype: histological type of the tumor; 1=Squamous, 2=Smallcell, 3=Adeno, 4=Large
- time: survival time
- status: censoring status; 0=start of the observation period or censoring, 1= death
- karno: Karnofsky performance score that describes the overall patients' status at the beginning of the study (discretely scored)
- diagtime: time between diagnosis and start of the study (in months)
- age: age of the patient (in years)
- prior: indicates if the patient has received another therapy before the current one; 0=No Prior Therapy, 10=Prior Therapy

In a preliminary effort to determine if there appears to be a treatment effect on time to death for the V.A. lung cancer patients, a plot of the Kaplan Meier survival curves for each treatment group was generated. This plot is shown below:

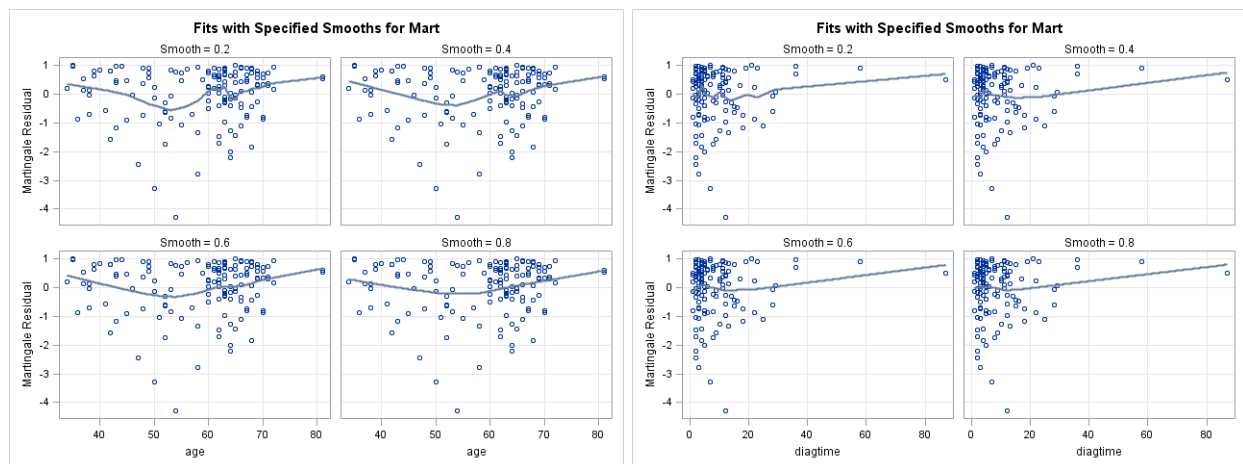


Based on these curves, there is not an obvious difference in the survival functions for the two treatment groups. Further analysis can confirm this intuition.

To determine the effect of treatment on time to death for the V.A. lung cancer patients, a Cox Proportional Hazard model was fit to the data. To verify that the Cox Proportional Hazard model is appropriate for this data, assessment of residual plots was required. Specifically, evaluation of the Martingale and Schoenfeld

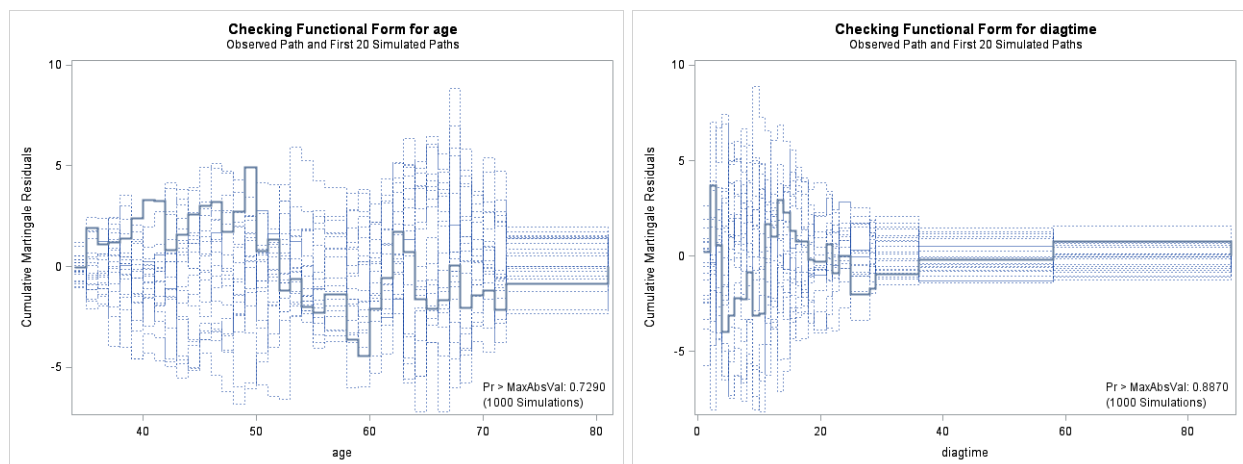
residual plots was performed. These residual plots would indicate whether the covariates in the data had linear contributions to the model and whether proportional hazards were observed in the data. These are two important assumptions to satisfy in order to consider inference from the Cox Proportional Hazard model reliable.

To begin, the Martingale residuals for each continuous covariate in the dataset were assessed. This included plots for the variables *age* and *diagtime*. These plots are shown below:



To verify the linear functional form of these covariates, we aim to see a flat line represented on the figure. We observe that as the smoothness increases for the fitted line, the curve flattens for both plots; however, neither is definitively flat nor strikingly curvy. Therefore, an additional test for the functional form of these covariates can help verify the linearity and satisfy this model assumption.

One assessment of the functional form of the covariates involves an evaluation of the cumulative Martingale residuals against each continuous covariate. These plots are shown below:

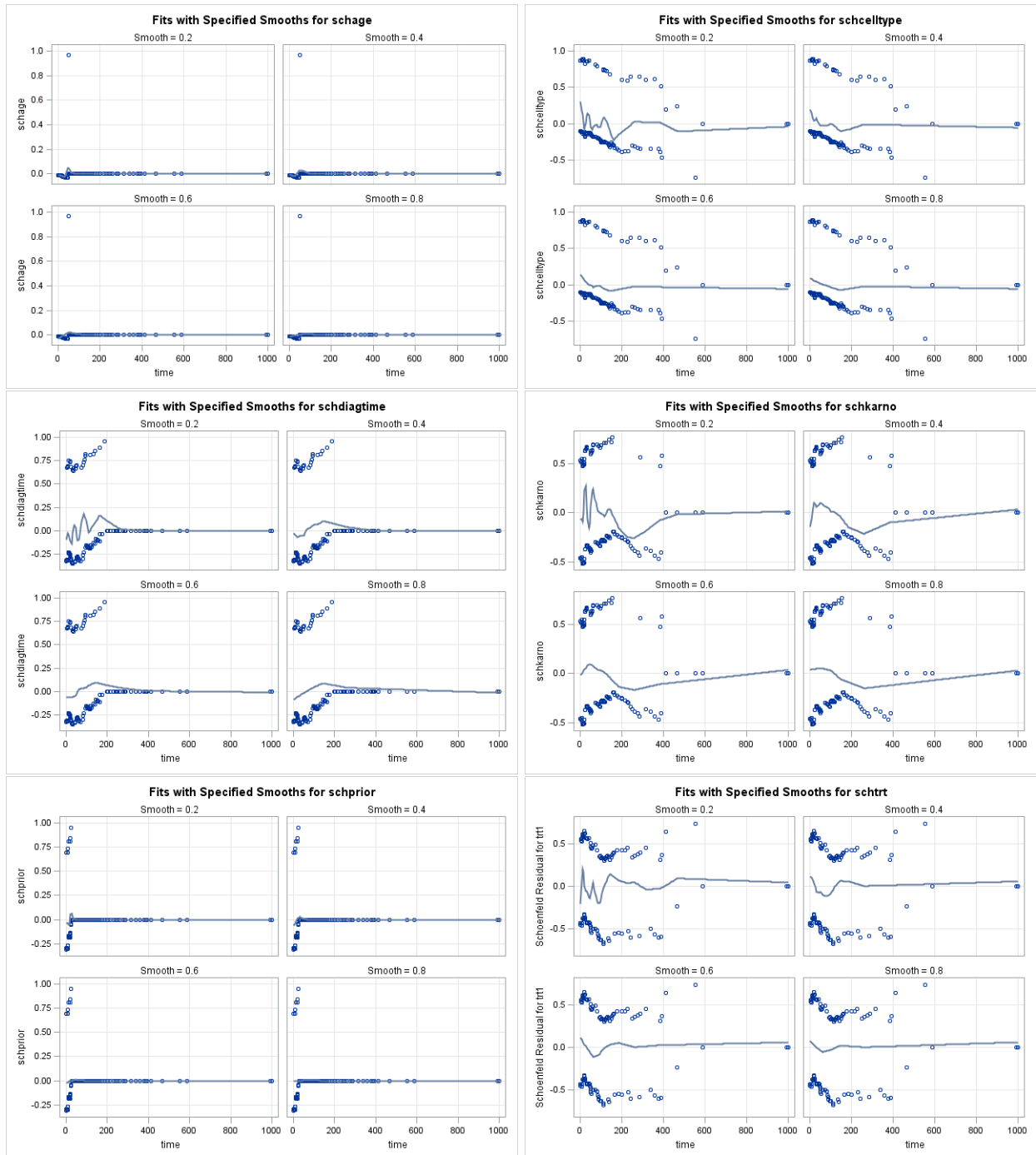


Confirmation of a linear functional form is depicted by the bolded line falling within the range of the dotted lines in the figure. Since both bolded lines in the two plots satisfy this, we have visual confirmation of a linear functional form of these continuous covariates.

A Kolmogorov-type supremum test for the functional form of each covariate can provide a quantitative measure for determining if there is evidence of a non-linear functional form. Therefore, computing such a test, in combination with information provided from the above plots, will allow for confirmation of the functional form of the continuous covariates in this dataset. The resulting p-values from the Kolmogorov-

type supremum test for the covariates *age* and *diagtime* were 0.73 and 0.89, respectively. These high p-values (p-value > 0.05) indicate that there is no evidence that these covariates do not have a linear functional form. Therefore, this assumption of the Cox Proportional Hazard model is satisfied.

The other important assumption for the Cox Proportional Hazard model is that there are proportional hazards in the data. To check this assumption, the plots for the Schoenfeld residuals must be considered. These plots are shown below:



This assumption is satisfied if the Schoenfeld residual plots for each covariate depict a flat smooth fitted line. The fitted lines for *age* and *prior* are very clearly flat lines and the fitted lines for *celltype*, *diagtime*, and *trt* are

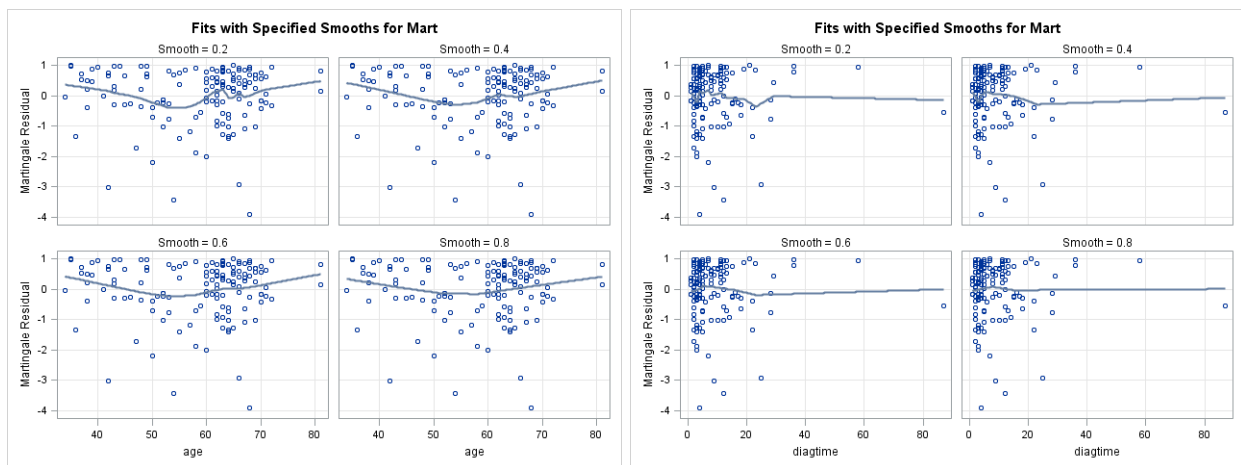
mostly very flat and thus give no indication that the proportional hazard assumption is violated. However, the fitted line for *karno* shows some evidence of curving and thus suggests that the proportional hazard assumption may be violated for this covariate.

Since the Schoenfeld residual plot for *karno* is inconclusive about violation of the proportional hazard assumption additional testing can be useful to determine whether the model assumption is satisfied. A Kolmogorov-type supremum test can be used to determine whether the proportional hazard assumption has been violated. Since the Karnofsky score variable is treated as categorical in the model, p-values are calculated for each increment of the metric. The associated p-values for each score are fairly large (p-value > 0.5), except for the scores of 40 and 80, which resulted in p-values of 0.02 and 0.00, respectively. The high p-values indicate there is no violation of the proportional hazard assumption, but since some of the Karnofsky scores indicated a violation of the proportional hazard assumption, either this covariate needs to be excluded from the model or the Cox Proportional Hazard model is inappropriate for the data. This analysis will therefore proceed with the exclusion of the Karnofsky score in the Cox Proportional Hazard model.

After excluding the *karno* covariate from the Cox Proportional Hazard model, we can run the Kolmogorov-type supremum test again to verify the model assumption again. This test indicated, with p-values for each covariate all larger than 0.05, that the proportional hazard assumption has been satisfied with this model.

With the exclusion of the *karno* covariate, it is wise to recheck the Martingale residuals to make sure the model is still appropriate.

The Martingale residuals for the model excluding *karno* are shown below:



We note that there are no significant changes to the Martingale residual plots.

Now that the Cox Proportional Hazard model has been justified to use for this data, a univariate analysis of the covariates in the model is useful to understand the influence of each covariate on the response. The model fit yielded the following regression coefficient output:

Parameter	DF	Estimate	Standard Error	Chi-Square	p-value	Hazard Ratio
trt=Standard	1	-0.16	0.20	0.66	0.41	0.85
celltype=Squamous	1	-0.29	0.29	1.03	0.31	0.75
celltype=Smallcell	1	0.75	0.26	8.20	0.00	2.11
celltype=Adeno	1	0.89	0.30	9.08	0.00	2.44
diagtime	1	0.01	0.01	1.16	0.28	1.01
age	1	0.00	0.01	0.26	0.61	1.01
prior=No Prior	1	0.07	0.23	0.08	0.77	1.07

The chi-square values reported above reflect the Wald test statistic and the p-value results from comparing the Wald statistic to a chi-square distribution with one degree of freedom (and thus reflect the results from a Wald test).

The full model thus takes the form:

$$H(t) = h_0(t)\exp(-0.16trtStandard - 0.29celltypeSquamous + 0.75celltypeSmallcell + 0.89celltypeAdeno + 0.01diagtime + 0.00age + 0.07priorNoPrior)$$

This output can be interpreted as follows:

1. The risk of death for patients given the standard lung cancer therapy is 0.85 times the risk of death for patients given the test chemotherapy.
2. The risk of death for patients with squamous tumor cells is 0.75 times the risk of death for the reference group (patients with large tumor cells).
3. The risk of death for patients with smallcell tumor cells is 2.11 times the risk of death for the reference group (patients with large tumor cells).
4. The risk of death for patients with adeno tumor cells is 2.44 times the risk of death for the reference group (patients with large tumor cells).
5. A one-unit increase in the value of *diagtime* results an increase in the risk of death by 1% (1.01-1.01=1%).
6. A one-unit increase in the value of *age* results in an increase in the risk of death by 1% (1.01-1.00=1%).
7. The risk of death for patients with no prior therapy is 1.07 times the risk of death for patients with prior therapy.

This output indicates that only one term is clearly significant in the model— *celltype*. Specifically, smallcell and adeno tumor cell types are significant in the model.

We can perform backwards selection to then create a reduced model with only relevant covariates. Each iteration, we will remove the covariate with the highest p-value, which is greater than 0.15, until all covariates in the model are significant (and including *trt*). This process resulting in selecting to include the covariates *trt* and *celltype* in the reduced model, as these were the only significant terms.

We determined that the log likelihood of the full model is 984.975. Fitting a reduced model with only the significant covariates *celltype* and *trt* (*trt* is kept due to treatment distinguishing), we find the log likelihood to be 986.217. We can then perform a likelihood ratio test to determine whether the full or reduced model is preferred. The likelihood ratio test statistic is 1.24 and when this statistic is compared to a chi-square distribution with 3 degrees of freedom (p-k → 7-4=3), we obtain a p-value of 0.74. This indicates that there is insufficient evidence that the coefficients of the extra covariates existing only in the full model are not equal to 0. Therefore, the reduced model is preferred.

Fitting the reduced model yields the following regression output:

<i>Parameter</i>	<i>DF</i>	<i>Estimate</i>	<i>Standard Error</i>	<i>Chi-Square</i>	<i>p-value</i>	<i>Hazard Ratio</i>
trt=Standard	1	-0.19	0.20	0.98	0.32	0.82
celltype=Squamous	1	-0.30	0.29	1.07	0.30	0.74
celltype=Smallcell	1	0.79	0.25	9.72	0.00	2.21
celltype=Adeno	1	0.87	0.29	8.83	0.00	2.38

The reduced model takes the form:

$$H(t) = h_0(t)\exp(-0.19trtStandard - 0.30celltypeSquamous + 0.79celltypeSmallcell + 0.87celltypeAdeno)$$

The following inferences can be made from the reduced (“best”) model:

1. The risk of death for patients given the standard lung cancer therapy is 0.82 times the risk of death for patients given the test chemotherapy.
2. The risk of death for patients with squamous tumor cells is 0.74 times the risk of death for the reference group (patients with large tumor cells).
3. The risk of death for patients with smallcell tumor cells is 2.21 times the risk of death for the reference group (patients with large tumor cells).
4. The risk of death for patients with adeno tumor cells is 2.38 times the risk of death for the reference group (patients with large tumor cells).

### Discussion

This analysis aimed to determine the treatment effect on time to death for V.A. lung cancer patients, as well as to determine if any other covariates could be prognostic tools. Treatment type was not determined to be a significant covariate in the model of time to death for this data, but it was determined that the risk of death is slightly lower for patients given the standard therapy. Further, it was determined that tumor cell type could be used to determine the risk of death for lung cancer patients and that squamous tumor cells were attributed to lower risk of death, versus the higher risk associated with the presence of smallcell and adeno tumor cells.

The inferences drawn from this analysis have some caveats. The first is that Karnofsky score is not considered in this model and could be relevant to the risk of death. Further analysis the reintroduces this covariate is recommended to understand its influence on the time to death. Since the primary goal of this study was to assess the treatment effect on time to death, the exclusion of this covariate from the model should not impact the conclusion drawn about treatment effect, however.

This analysis involved through examination of model assumptions to ensure the validity of the inferences made. The scope of the study was refined to only examine variables that could accurately be interpreted using the Cox Proportional Hazard model. Overall, the goals for this study were achieved using the methods discussed in this paper.