

Cox Proportional Hazards Regression in SAS

...

Yale Paulsen

VA Lung Dataset

Data representing 137 lung cancer patients treated with either the “standard” therapy or an experimental one. Data were collected prior to 1980 and contain 8 variables.

Table 1: valung dataset: Contains data for 137 lung cancer patients on one of two treatments

Therapy	Cell	Time	Status	KPS	Diagnosis Time	Age	Prior
standard	Squamous	72	dead	60	7	69	no
standard	Squamous	411	dead	70	5	64	yes
standard	Squamous	228	dead	60	3	38	no
standard	Squamous	126	dead	60	9	63	yes
standard	Squamous	118	dead	70	11	65	yes
standard	Squamous	10	dead	20	5	49	no

Therapy:

- Test
- Standard

VA Lung Dataset

Data representing 137 lung cancer patients treated with either the “standard” therapy or an experimental one. Data were collected prior to 1980 and contain 8 variables.

Table 1: valung dataset: Contains data for 137 lung cancer patients on one of two treatments

Therapy	Cell	Time	Status	KPS	Diagnosis Time	Age	Prior
standard	Squamous	72	dead	60	7	69	no
standard	Squamous	411	dead	70	5	64	yes
standard	Squamous	228	dead	60	3	38	no
standard	Squamous	126	dead	60	9	63	yes
standard	Squamous	118	dead	70	11	65	yes
standard	Squamous	10	dead	20	5	49	no

Cell Type:

- Squamous
- Large
- Small
- Adeno

VA Lung Dataset

Data representing 137 lung cancer patients treated with either the “standard” therapy or an experimental one. Data were collected prior to 1980 and contain 8 variables.

Table 1: valung dataset: Contains data for 137 lung cancer patients on one of two treatments

Therapy	Cell	Time	Status	KPS	Diagnosis Time	Age	Prior
standard	Squamous	72	dead	60	7	69	no
standard	Squamous	411	dead	70	5	64	yes
standard	Squamous	228	dead	60	3	38	no
standard	Squamous	126	dead	60	9	63	yes
standard	Squamous	118	dead	70	11	65	yes
standard	Squamous	10	dead	20	5	49	no

Time and Status:

- Time = y
- Status = dead, censored

VA Lung Dataset

Data representing 137 lung cancer patients treated with either the “standard” therapy or an experimental one. Data were collected prior to 1980 and contain 8 variables.

Table 1: valung dataset: Contains data for 137 lung cancer patients on one of two treatments

Therapy	Cell	Time	Status	KPS	Diagnosis Time	Age	Prior
standard	Squamous	72	dead	60	7	69	no
standard	Squamous	411	dead	70	5	64	yes
standard	Squamous	228	dead	60	3	38	no
standard	Squamous	126	dead	60	9	63	yes
standard	Squamous	118	dead	70	11	65	yes
standard	Squamous	10	dead	20	5	49	no

Prior: yes, no

Several numerical covariates:

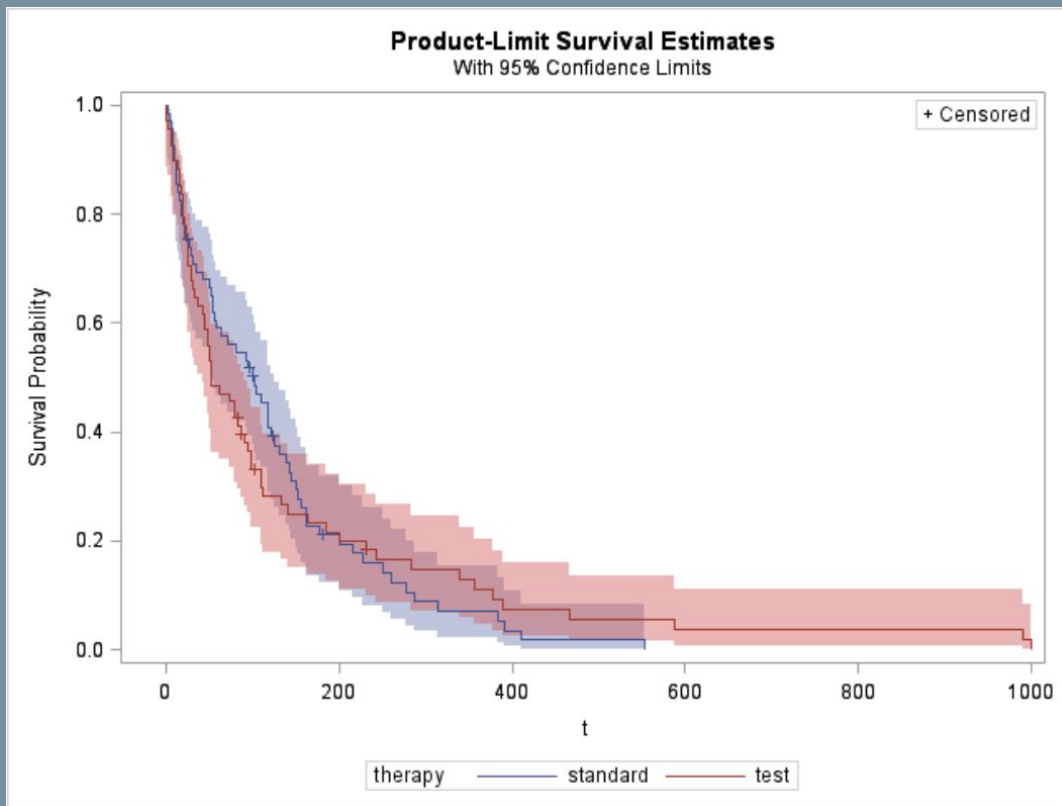
- kps
- Diagnosis time
- Age

Kaplan Meier Survival Estimation

Kaplan Meier Estimates of Survival

Here we see the survival curves for patients on the test therapy compared to the standard therapy.

There doesn't appear to be a difference here.



...

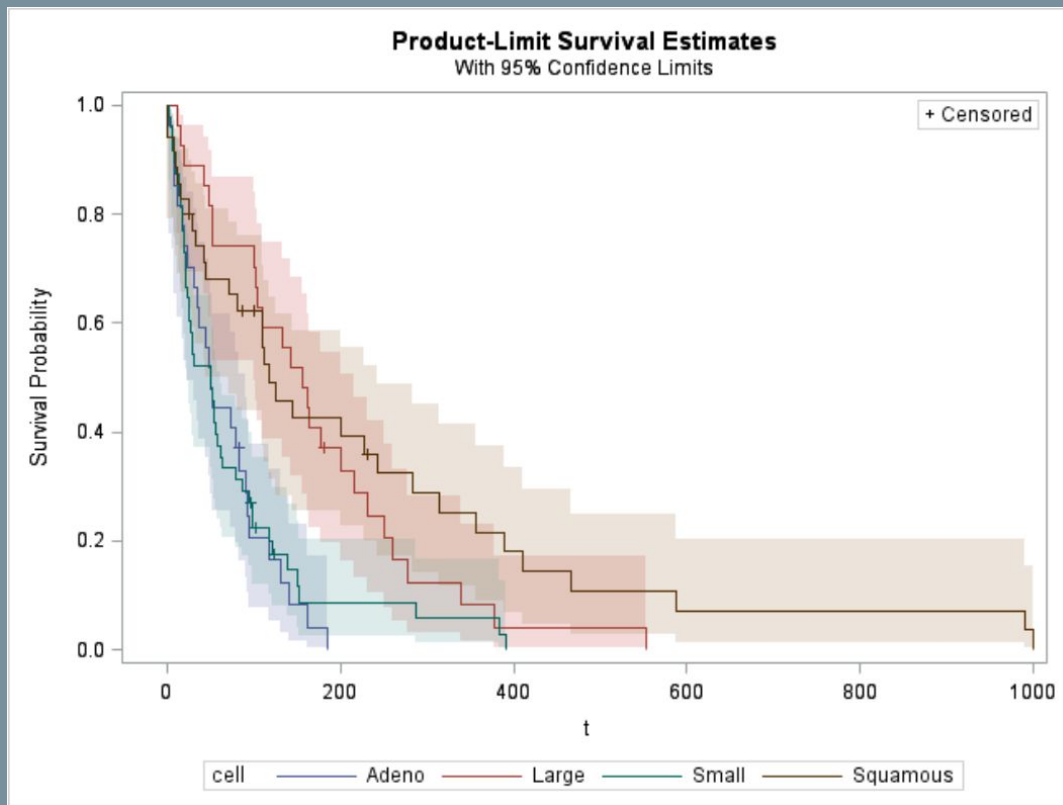
Looking next at some covariates

...

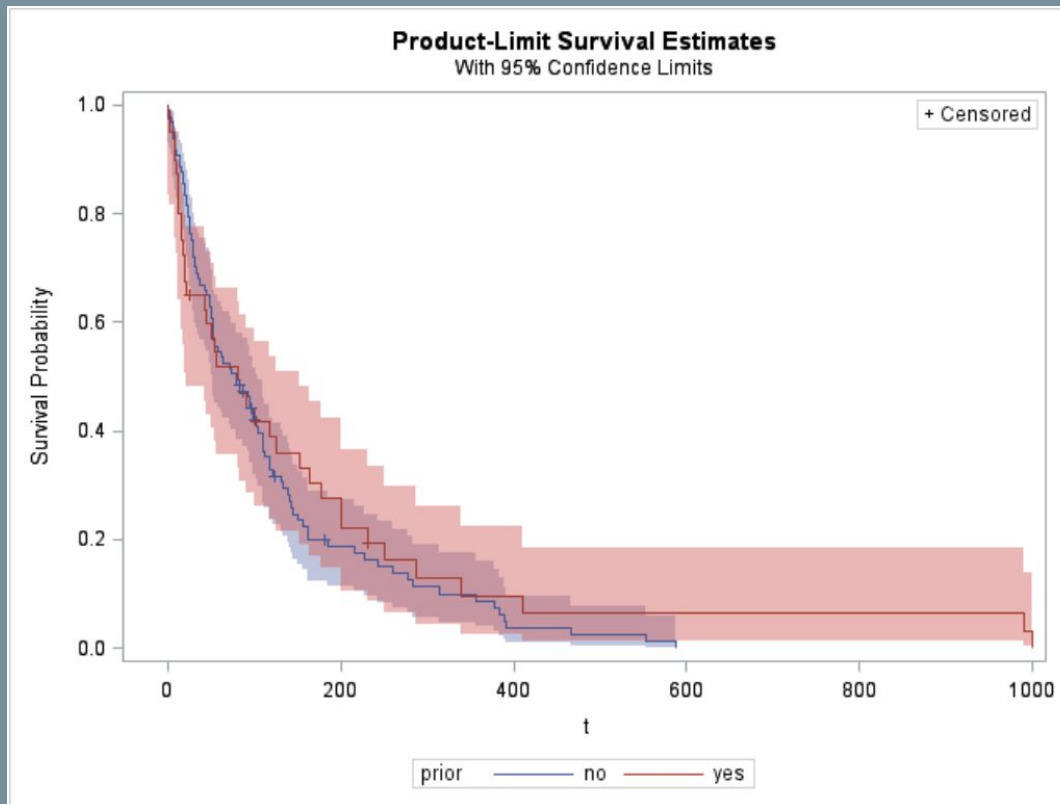
Kaplan Meier Estimates of Survival

Some differences here

We'll look more closely
at this later.



Visualizing other categorical covariates



Hypothesis Tests of Covariates

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
kps	1	44.8525	<.0001	44.8525	<.0001
age	2	44.8526	<.0001	0.000030	0.9956
diagtime	3	44.8526	<.0001	6.537E-7	0.9994

Looking at numerical variables together

kps is the only significant numerical variable in this analysis.

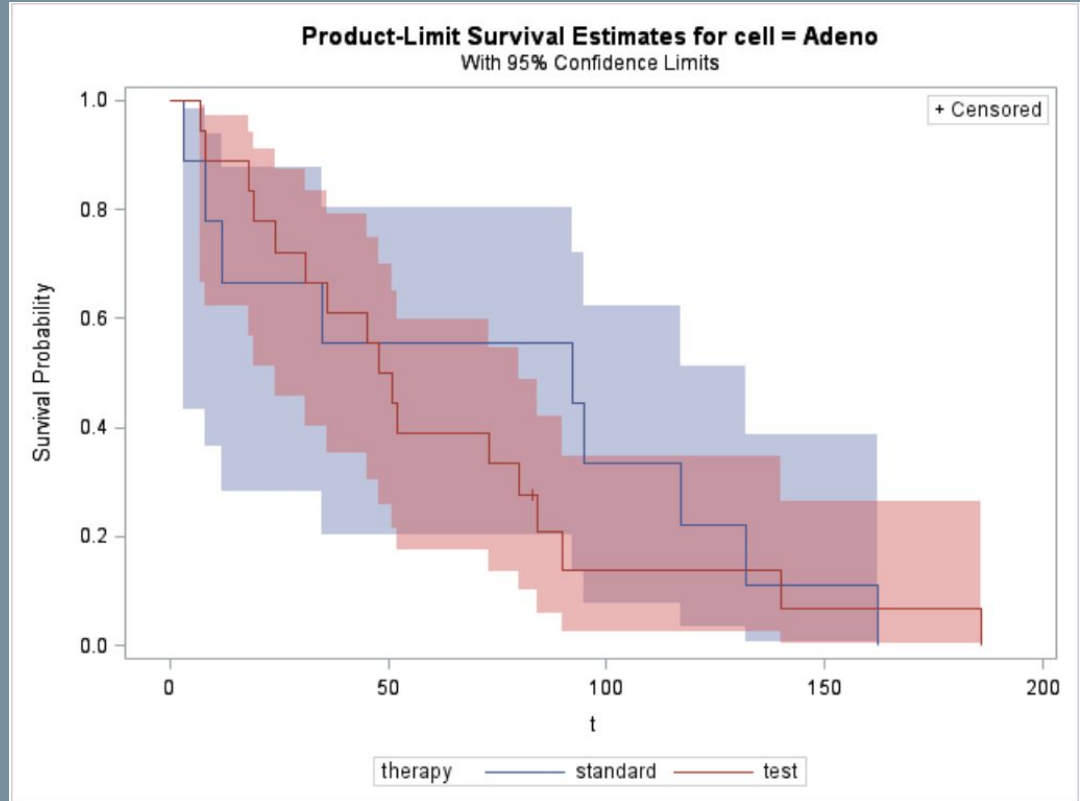
...

Looking at therapy within cell type

...

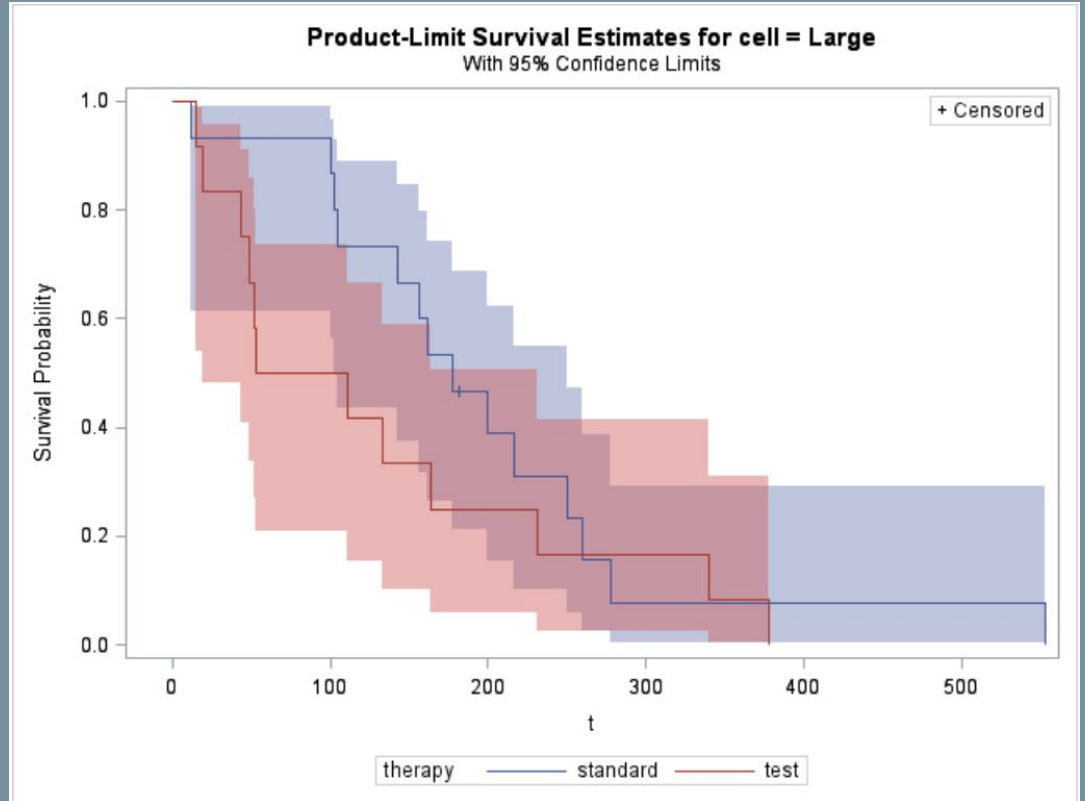
Kaplan Meier Estimates of Survival

No difference here



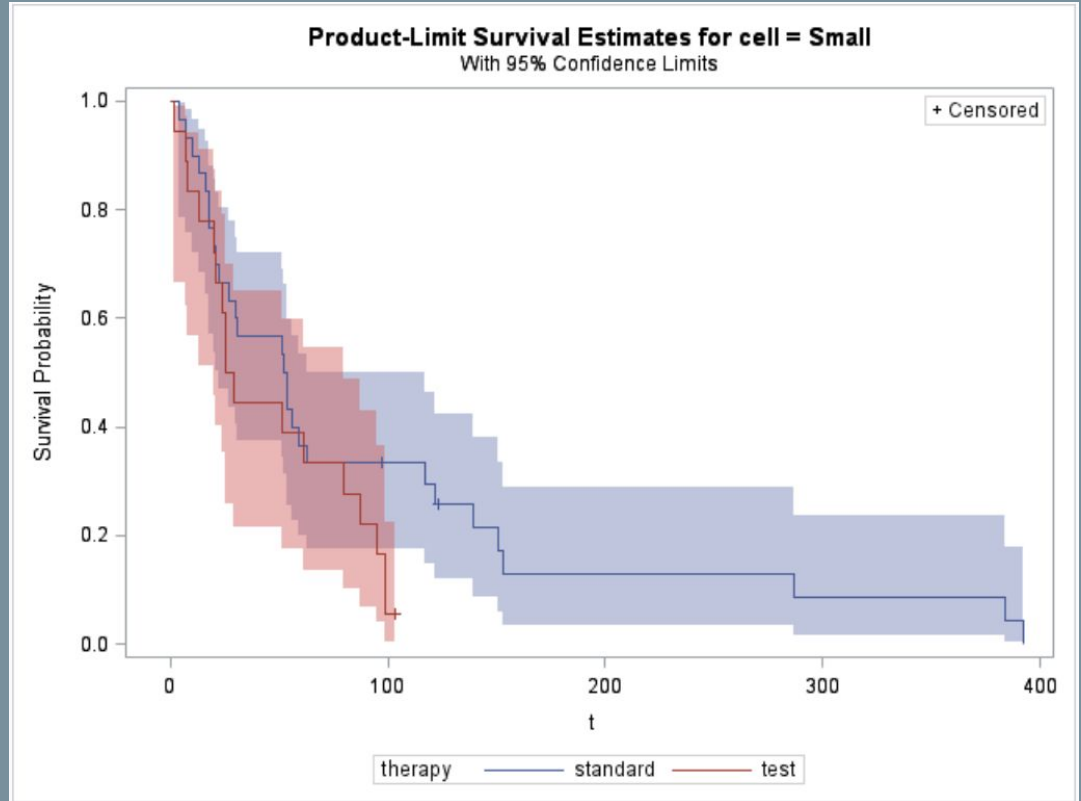
Kaplan Meier Estimates of Survival

No difference here



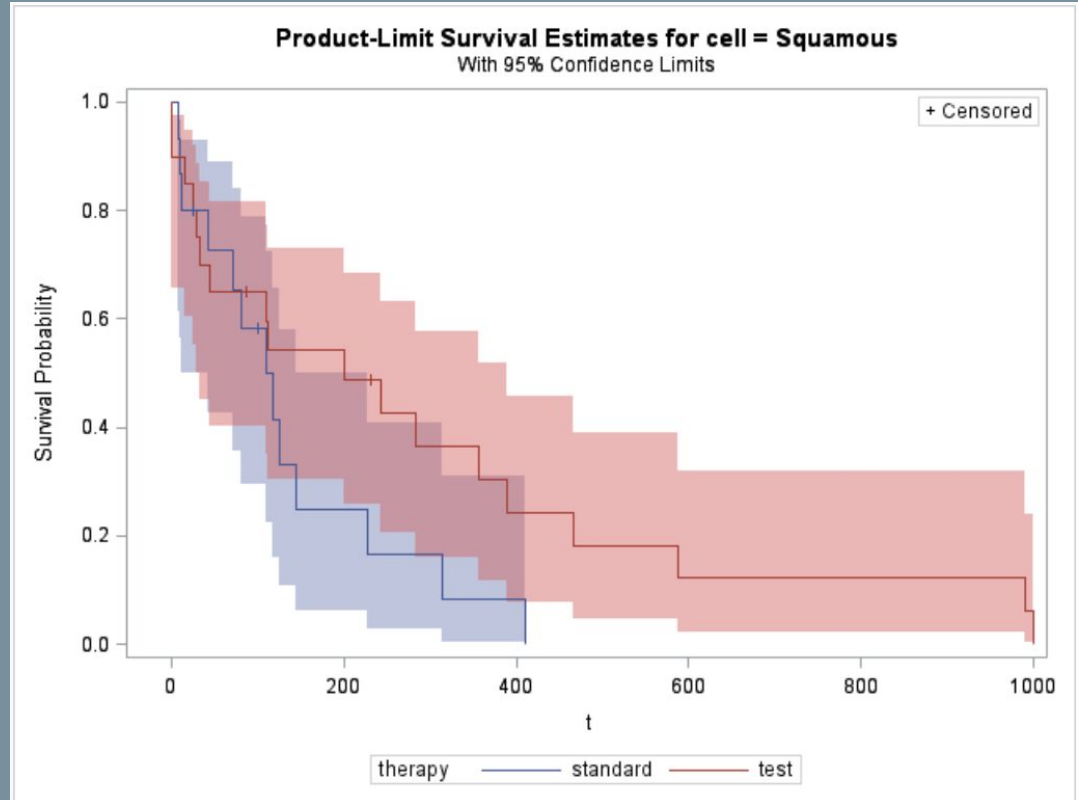
Kaplan Meier Estimates of Survival

No difference here?

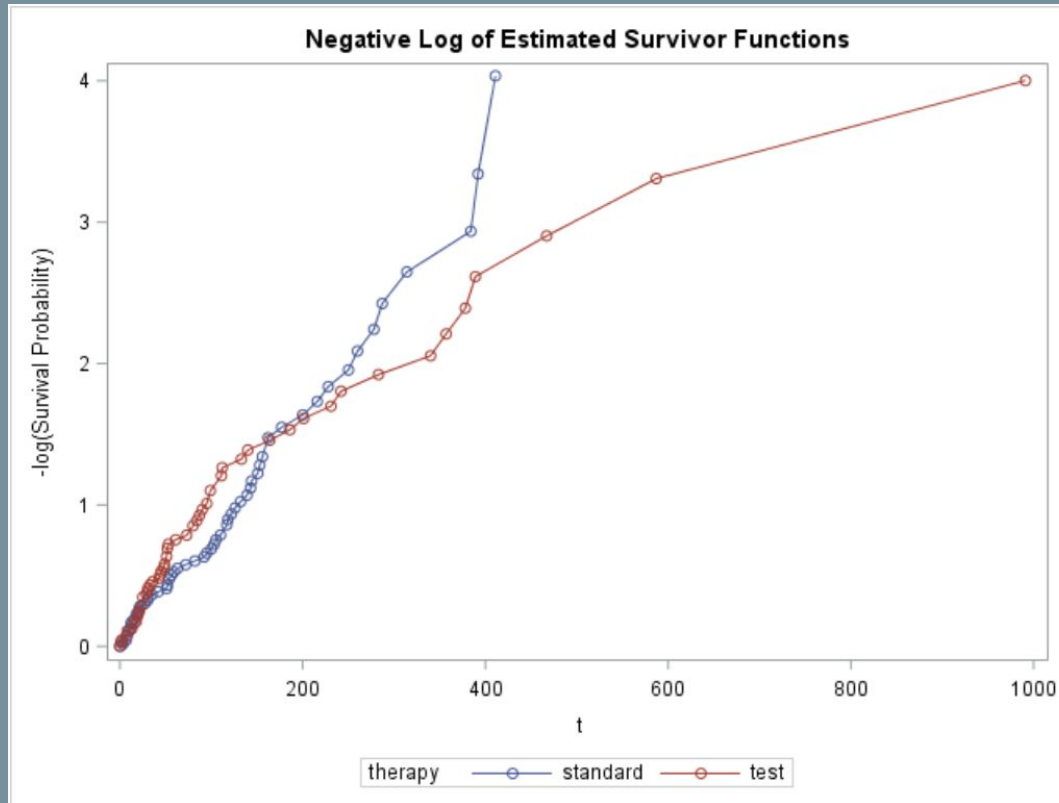


Kaplan Meier Estimates of Survival

No difference here?

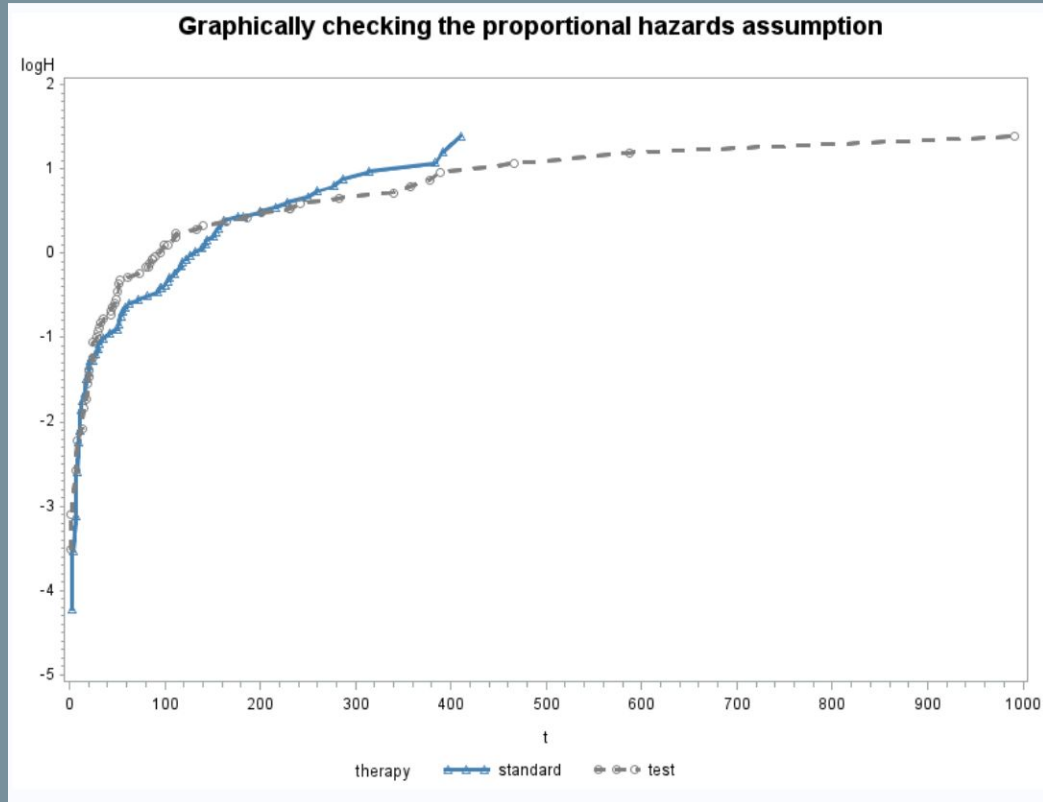


Nelson Cumulative Hazard Plot



Therapy:
Standard vs. Test

Checking Proportional Hazards Assumption for Therapy



Assumption doesn't hold

Proportional Hazards Regression

Full Model

$$h(y | x) = h_0(y) * \exp \{ \beta_{Therapy} * Therapy + \beta_{kps} * kps + \beta_{DiagTime} * DiagTime + \beta_{Age} * Age + \beta_{Prior} * Prior \\ + \beta_{Adeno} * Adeno + \beta_{Small} * Small + \beta_{Large} * Large \}$$

(Adeno, Small, Large) are indicator functions.
Squamous is baseline.

Full Model PL Estimation by Efron Method

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
therapy	1	2.0148	0.1558
kps	1	35.4979	<.0001
diagtime	1	0.0001	0.9929
age	1	0.8763	0.3492
prior	1	0.0950	0.7580
cell	3	18.1444	0.0004

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
therapy	standard	1	-0.29461	0.20755	2.0148	0.1558	0.745	therapy standard
kps		1	-0.03282	0.00551	35.4979	<.0001	0.968	
diagtime		1	0.0000818	0.00914	0.0001	0.9929	1.000	
age		1	-0.00871	0.00930	0.8763	0.3492	0.991	
prior	no	1	-0.07159	0.23231	0.0950	0.7580	0.931	prior no
cell	Adeno	1	1.19607	0.30092	15.7986	<.0001	3.307	cell Adeno
cell	Large	1	0.40129	0.28269	2.0151	0.1557	1.494	cell Large
cell	Small	1	0.86156	0.27528	9.7950	0.0017	2.367	cell Small

Full Model PL Estimation by Efron Method With Backward Stepwise Selection

Backwards Selection
eliminates all
covariates except for
kps and cell type.

Therapy is unimportant
given this dataset.

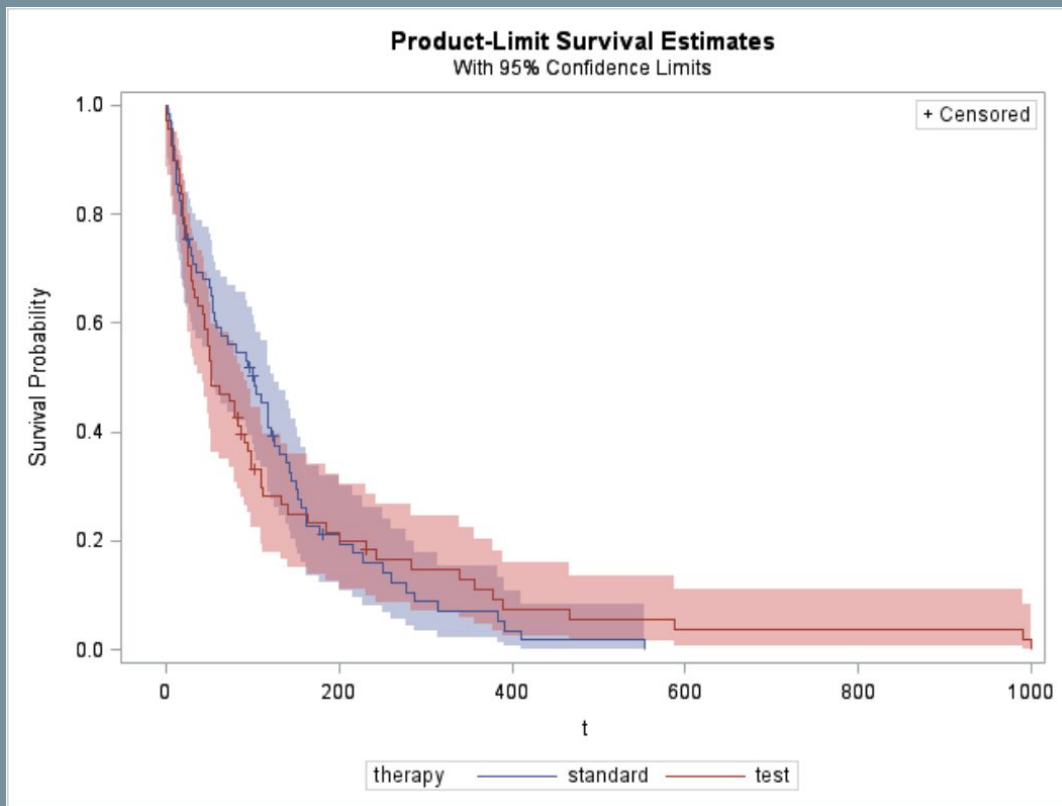
Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
kps	1	35.9851	<.0001
cell	3	17.2820	0.0006

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9851	<.0001	0.969	
cell	Adeno	1	1.15773	0.29294	15.6196	<.0001	3.183	cell Adeno
cell	Large	1	0.32565	0.27668	1.3853	0.2392	1.385	cell Large
cell	Small	1	0.71534	0.25269	8.0142	0.0046	2.045	cell Small

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	diagtime	1	5	0.0001	0.9929
2	prior	1	4	0.1224	0.7265
3	age	1	3	0.9315	0.3345
4	therapy	1	2	1.6971	0.1927

Kaplan Meier Estimates of Survival

Unsurprising given the estimates for survival we saw earlier.



...

Fitting the final model

...

Reduced model: backwards selection in SAS using the Efron method for ties

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9851	<.0001	0.969	
cell	Adeno	1	1.15773	0.29294	15.6196	<.0001	3.183	cell Adeno
cell	Large	1	0.32565	0.27668	1.3853	0.2392	1.385	cell Large
cell	Small	1	0.71534	0.25269	8.0142	0.0046	2.045	cell Small

$$h(y | x) = h_0(y) * \exp \{ \beta_{kps} * kps + \beta_{Adeno} * Adeno + \beta_{Small} * Small + \beta_{Large} * Large \}$$

Final Model With Exact PL

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.3764	4	<.0001
Score	63.9408	4	<.0001
Wald	61.2591	4	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
kps	1	35.9865	<.0001
cell	3	17.2835	0.0006

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9865	<.0001	0.969	
cell	Adeno	1	1.15793	0.29297	15.6211	<.0001	3.183	cell Adeno
cell	Large	1	0.32566	0.27669	1.3854	0.2392	1.385	cell Large
cell	Small	1	0.71533	0.25269	8.0135	0.0046	2.045	cell Small

Global tests show that there is some significant effect from one or more covariate.

kps and cell are significant

These results will be used to fit the final model.

Final Model Fitted with exact Partial Likelihood

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9865	<.0001	0.969	
cell	Adeno	1	1.15793	0.29297	15.6211	<.0001	3.183	cell Adeno
cell	Large	1	0.32566	0.27669	1.3854	0.2392	1.385	cell Large
cell	Small	1	0.71533	0.25269	8.0135	0.0046	2.045	cell Small

$$h(y | x) = h_0(y) * \exp \{ \beta_{kps} * kps + \beta_{Adeno} * Adeno + \beta_{Small} * Small + \beta_{Large} * Large \}$$

Final Model

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9865	<.0001	0.969	
cell	Adeno	1	1.15793	0.29297	15.6211	<.0001	3.183	cell Adeno
cell	Large	1	0.32566	0.27669	1.3854	0.2392	1.385	cell Large
cell	Small	1	0.71533	0.25269	8.0135	0.0046	2.045	cell Small

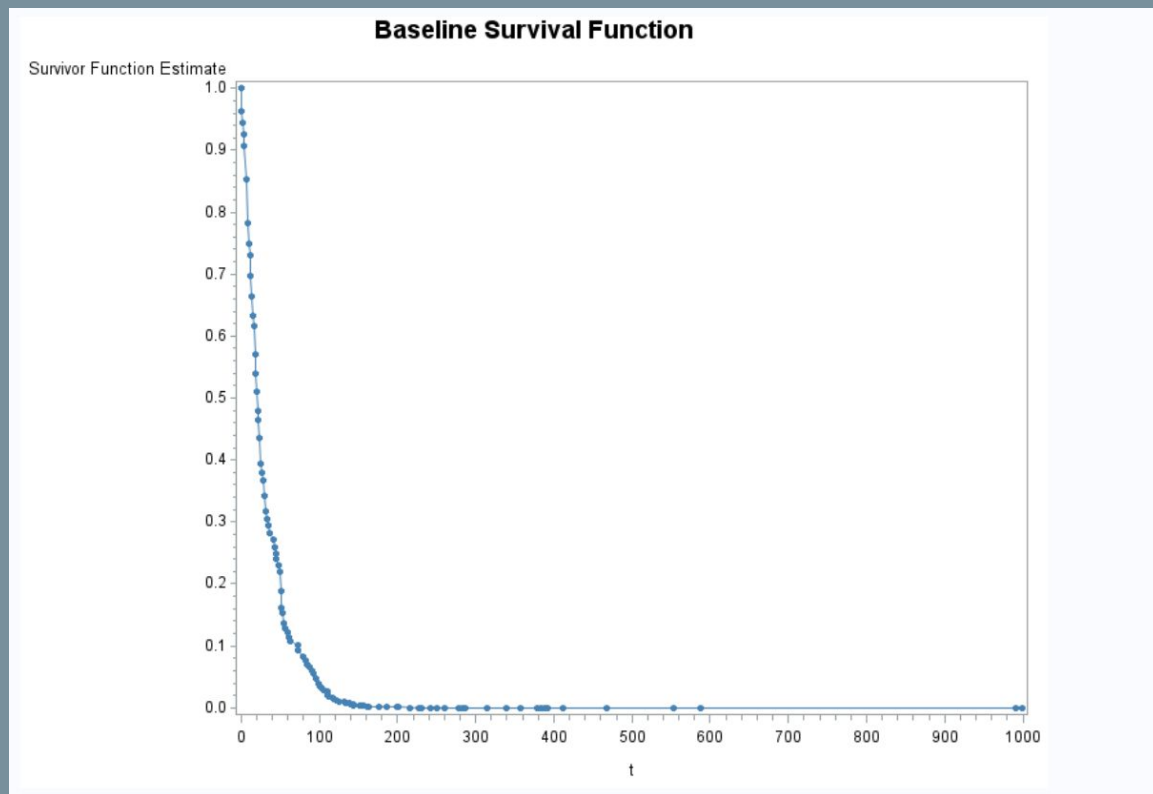
$$h(y | x) = h_0(y) * \exp \{ - 0.031 * kps + 1.158 * Adeno + 0.326 * Large + 0.715 * Small \}$$

Baseline Survival function

$h_0(y)$:

kps = 0

Cell = Squamous



...

Effects of important covariates

...

Looking at effects: kps

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9865	<.0001	0.969	
cell	Adeno	1	1.15793	0.29297	15.6211	<.0001	3.183	cell Adeno
cell	Large	1	0.32566	0.27669	1.3854	0.2392	1.385	cell Large
cell	Small	1	0.71533	0.25269	8.0135	0.0046	2.045	cell Small

Karnofsky Performance Status is a numerical value from 0 to 100. This is a prognostic score where 100 indicates no presence of disease and 0 indicates death. Typically scores are given in intervals of ten (i.e. 10, 20, 30), but researchers sometimes assign scores in between the normal intervals. E.g. our data has scores of 75, 85, 99.

The hazard ratio here suggests that a person with a kps score of $(k + 1)$ has about 97% the risk of death as someone with a score of k .

Looking at effects: cell type

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
kps		1	-0.03106	0.00518	35.9865	<.0001	0.969	
cell	Adeno	1	1.15793	0.29297	15.6211	<.0001	3.183	cell Adeno
cell	Large	1	0.32566	0.27669	1.3854	0.2392	1.385	cell Large
cell	Small	1	0.71533	0.25269	8.0135	0.0046	2.045	cell Small

Cell type is a description of the type of cancer. Different cell-types have different characteristics. E.g. Small cell carcinoma metastasizes easily.

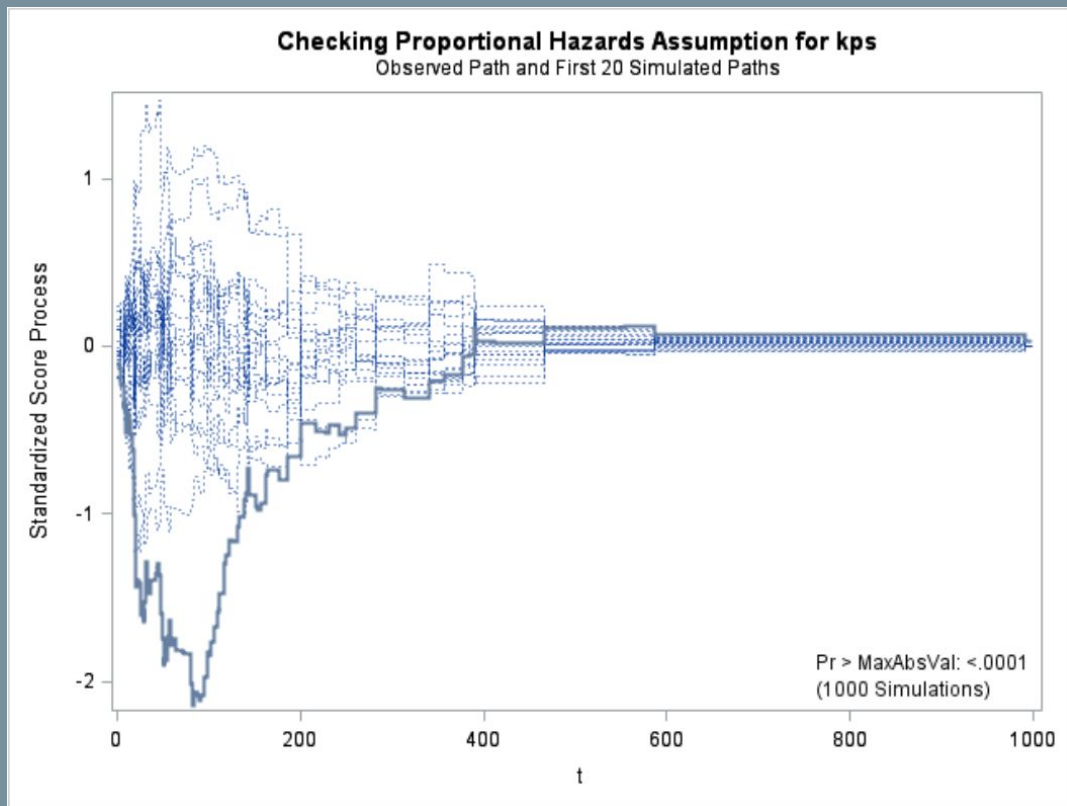
These hazard ratios show the relative effect of each cell-type against the baseline: Squamous cell.

All other things being equal, a person with an Adeno cell type cancer has a hazard of approximately 3.2X that of a Squamous cell patient. Similarly, Small and Large cell types have around 2.1X and 1.4X the hazard of Squamous cell cancers.

Checking the proportional hazards assumption

kps

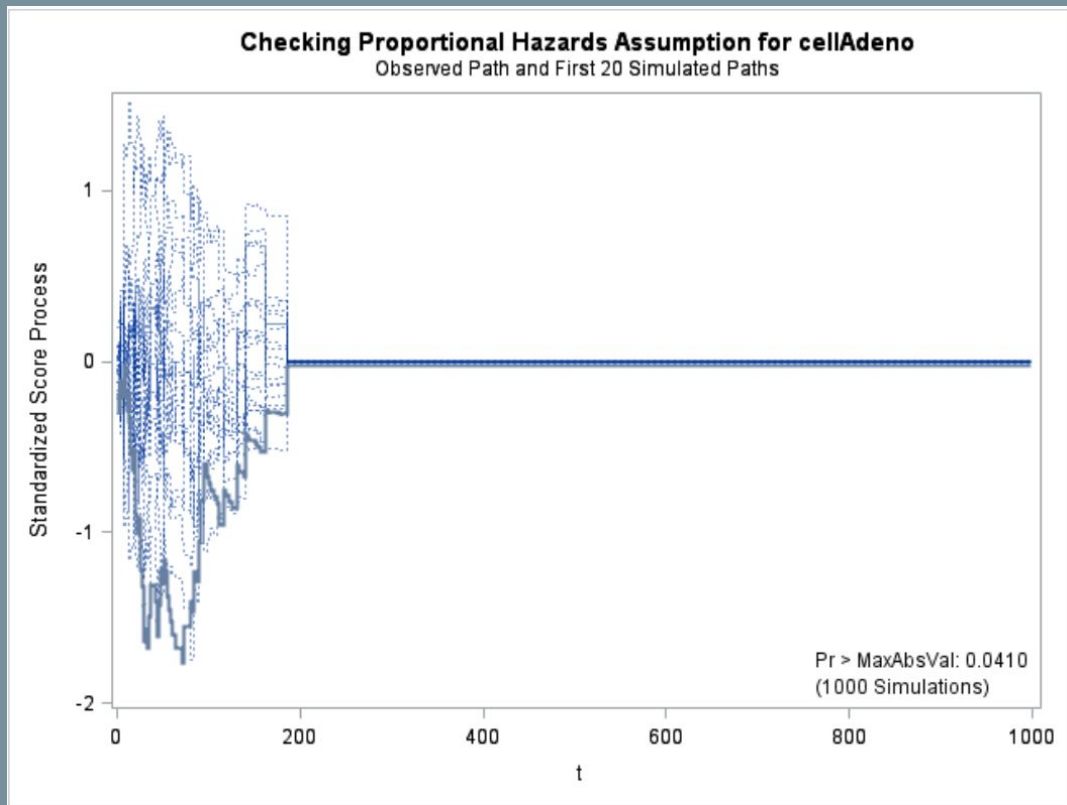
The large deviation from the typical path indicates a violation of the proportional hazards assumption.



Checking the proportional hazards assumption

Cell type Adeno

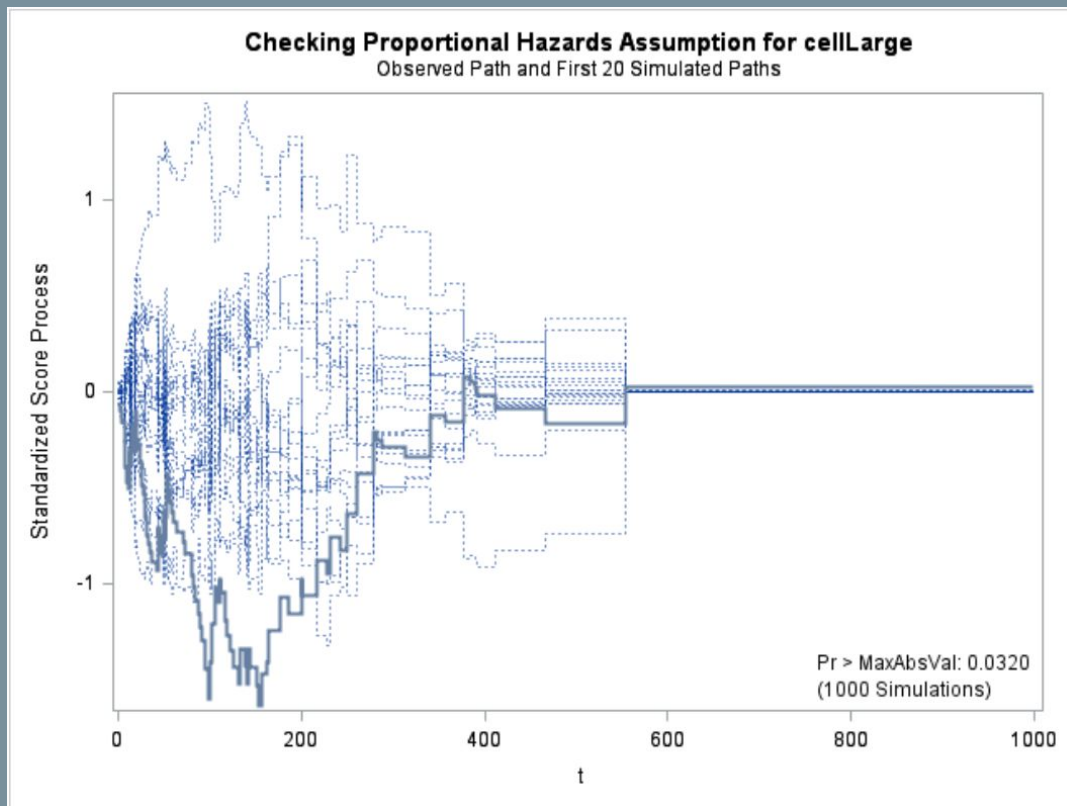
The large deviation from the typical path indicates a violation of the proportional hazards assumption.



Checking the proportional hazards assumption

Cell type Large

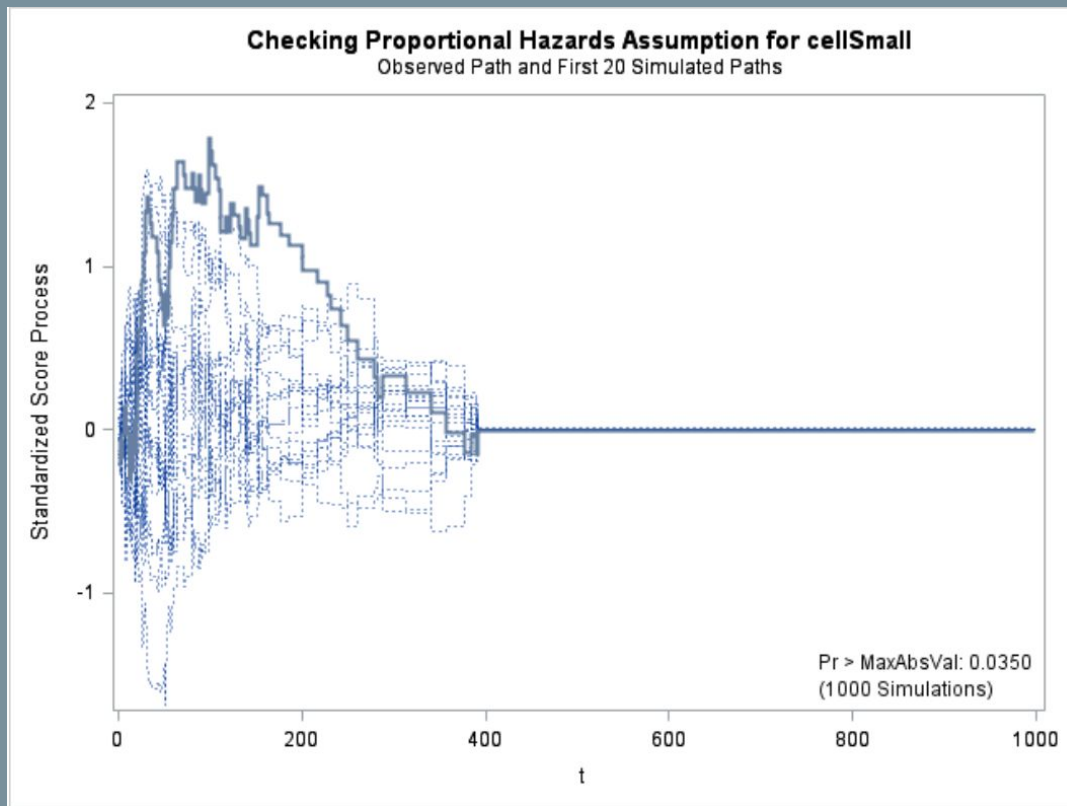
The large deviation from the typical path indicates a violation of the proportional hazards assumption.



Checking the proportional hazards assumption

Cell type Small

The large deviation from the typical path indicates a violation of the proportional hazards assumption.



p-values

Supremum Test for Proportionals Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
kps	2.1450	1000	1747787217	<.0001
cellAdeno	1.7607	1000	1747787217	0.0410
cellLarge	1.6373	1000	1747787217	0.0320
cellSmall	1.7837	1000	1747787217	0.0350

All covariates fail to meet the proportional hazards assumption.

Cox Proportional Hazards Regression in SAS

...

Yale Paulsen