

# P8108 Group 2 Survival Analysis Project

Yiming Zhao (yz3955)      Wenshan Qu (wq2160)      Tucker Morgan (t1m2152)  
Junzhe Shao (js5959)      Benjamin Goebel (bpg2118)

2022-12-1

## Introduction

Breast cancer is a leading cause of cancer-related morbidity and mortality worldwide, and the choice of treatment can have a significant impact on patient outcomes. Hormone therapy and chemotherapy are two common treatment options for breast cancer, but their effectiveness in improving survival may vary depending on the specific characteristics of the patient and the tumor and there is plenty of literature comparing their effectiveness.

In this study, we aim to evaluate the effectiveness of hormone therapy and chemotherapy in improving breast cancer survival for our target population using the data from the Rotterdam tumor bank. Specifically, we will examine the treatment effect and taking into account various patient and tumor characteristics that may be the confounders of the treatment.

In addition to evaluating the effectiveness of different treatments, we will also compare the performance of non-parametric modeling approaches, such as random forests, to that of semi-parametric approaches, such as the Cox proportional hazards model, in predicting breast cancer survival. Understanding the performance of these different modeling approaches can inform the selection of the most appropriate method for predicting breast cancer survival in different settings.

# Exploratory Data Analysis

**Table 1: Distribution of covariates between hormone therapy treatment groups**

	0 (N=2643)	1 (N=339)	Total (N=2982)	p value
<b>age</b>				< 0.001
Mean (SD)	54.098 (12.984)	62.549 (9.921)	55.058 (12.953)	
Range	24.000 - 90.000	28.000 - 88.000	24.000 - 90.000	
<b>meno</b>				< 0.001
Mean (SD)	0.519 (0.500)	0.879 (0.327)	0.560 (0.496)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
<b>size</b>				< 0.001
<=20	1283 (48.5%)	104 (30.7%)	1387 (46.5%)	
20-50	1119 (42.3%)	172 (50.7%)	1291 (43.3%)	
>50	241 (9.1%)	63 (18.6%)	304 (10.2%)	
<b>grade</b>				< 0.001
Mean (SD)	2.722 (0.448)	2.826 (0.380)	2.734 (0.442)	
Range	2.000 - 3.000	2.000 - 3.000	2.000 - 3.000	
<b>nodes</b>				< 0.001
Mean (SD)	2.327 (4.207)	5.720 (4.576)	2.712 (4.384)	
Range	0.000 - 34.000	1.000 - 24.000	0.000 - 34.000	
<b>pgr</b>				< 0.001
Mean (SD)	168.706 (300.337)	108.233 (200.302)	161.831 (291.311)	
Range	0.000 - 5004.000	0.000 - 1497.000	0.000 - 5004.000	
<b>er</b>				0.069
Mean (SD)	164.792 (272.563)	180.608 (271.693)	166.590 (272.465)	
Range	0.000 - 3275.000	0.000 - 2444.000	0.000 - 3275.000	
<b>chemo</b>				< 0.001
0	2091 (79.1%)	311 (91.7%)	2402 (80.5%)	
1	552 (20.9%)	28 (8.3%)	580 (19.5%)	
<b>dtime</b>				< 0.001
Mean (SD)	2679.067 (1309.178)	2030.534 (1043.971)	2605.340 (1298.078)	
Range	36.000 - 7043.000	45.000 - 6270.000	36.000 - 7043.000	
<b>death</b>				0.093
Mean (SD)	0.421 (0.494)	0.469 (0.500)	0.427 (0.495)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	

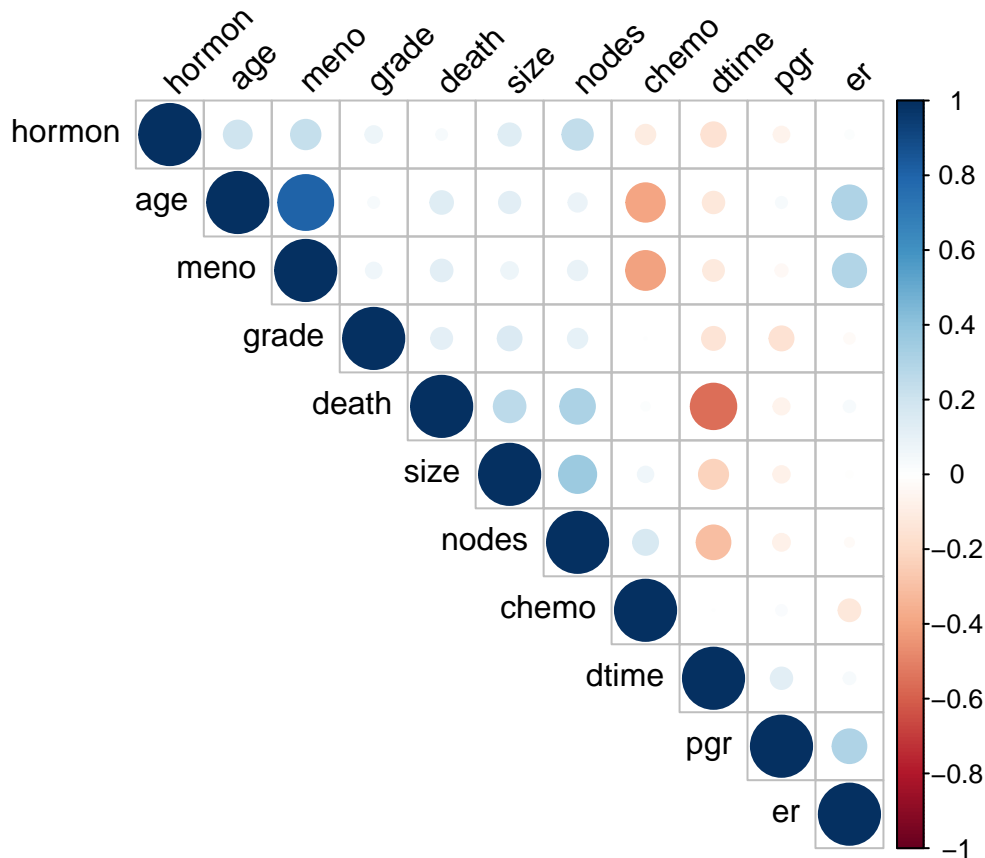
Add description here...

**Table 2: Distribution of covariates between chemotherapy treatment groups**

	0 (N=2402)	1 (N=580)	Total (N=2982)	p value
<b>age</b>				< 0.001
Mean (SD)	57.560 (12.775)	44.698 (7.322)	55.058 (12.953)	
Range	25.000 - 90.000	24.000 - 73.000	24.000 - 90.000	
<b>meno</b>				< 0.001
Mean (SD)	0.658 (0.474)	0.153 (0.361)	0.560 (0.496)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
<b>size</b>				0.002
<=20	1148 (47.8%)	239 (41.2%)	1387 (46.5%)	
20-50	1028 (42.8%)	263 (45.3%)	1291 (43.3%)	
>50	226 (9.4%)	78 (13.4%)	304 (10.2%)	
<b>grade</b>				0.964
Mean (SD)	2.734 (0.442)	2.734 (0.442)	2.734 (0.442)	
Range	2.000 - 3.000	2.000 - 3.000	2.000 - 3.000	
<b>nodes</b>				< 0.001
Mean (SD)	2.353 (4.240)	4.198 (4.651)	2.712 (4.384)	
Range	0.000 - 34.000	1.000 - 34.000	0.000 - 34.000	
<b>pgr</b>				0.002
Mean (SD)	157.556 (283.077)	179.536 (322.848)	161.831 (291.311)	
Range	0.000 - 3000.000	0.000 - 5004.000	0.000 - 5004.000	
<b>er</b>				< 0.001
Mean (SD)	183.599 (286.924)	96.148 (186.163)	166.590 (272.465)	
Range	0.000 - 3275.000	0.000 - 1929.000	0.000 - 3275.000	
<b>hormon</b>				< 0.001
0	2091 (87.1%)	552 (95.2%)	2643 (88.6%)	
1	311 (12.9%)	28 (4.8%)	339 (11.4%)	
<b>dtime</b>				0.805
Mean (SD)	2605.028 (1286.877)	2606.633 (1344.609)	2605.340 (1298.078)	
Range	36.000 - 7043.000	164.000 - 6270.000	36.000 - 7043.000	
<b>death</b>				0.322
Mean (SD)	0.422 (0.494)	0.445 (0.497)	0.427 (0.495)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	

Add description here...

Figure 1: Correlation of variables in Rotterdam data set



Add description here... what stands out?

## Methods

The dataset of interest for this analysis comes from the Rotterdam tumor bank, including data from 2982 breast cancer patients. Follow up time for patients varied from just 1 month to as long as 231 months. Several prognostic variables are recorded including year of surgery, age at surgery, menopausal status (pre- or post-), tumor size (mm), differentiation grade, number of positive lymph nodes, progesterone receptors (fmol/l), estrogen receptors (fmol/l), and indicators for hormonal treatment and chemotherapy treatment. The outcome considered in this analysis was patient death. The censoring mechanism is right censoring, which we assume to be non-informative.

### Log-rank test

Add description of method... Likely can be added from work

### Regularized Cox Proportional Hazard Model

As part of this analysis, we consider the Cox Proportional Hazard (Cox PH) model, which allows us to model the hazard ratio based on covariates to understand their impact on the survival function. The Cox

PH typically takes the form:

$$h(t|Z = z) = h_0(t)e^{\beta'z}.$$

In this application, we use the Least Absolute Shrinkage and Selection Operator (LASSO), an application of the  $\ell_1$  norm regularization penalty. In the Cox PH framework, this penalty term takes the form of:

$$\lambda \sum |\beta_i|$$

where  $\lambda$  represents our penalty coefficient. This penalty helps to avoid over-fitting of our data in an effort to improve the generalizability of our predictive model in different settings. The algorithm used here in `glmnet` uses the Breslow approximation to handle ties. For more details on the derivation of this term and the algorithm used to fit the penalized Cox PH model, see Simon et al. (2011).

## Random Survival Forest

The survival tree and the corresponding random survival forest (RSF) are highly favorable non-parametric methods when studying survival data. Generally, for a single survival tree, it will assign subjects to groups based on certain splitting rules regarding their covariates, and the subjects in each group will share a similar survival behavior. This model, being non-parametric, makes no additional assumptions. However, we continue to assume that the censoring is non-informative.

Add a little more description of method, doesn't have to be very long...

## Brier score

We compare the Cox proportional-hazards model with the random survival forest by calculating the Brier score for each model on the test data set. The formula for the Brier score is as follows.

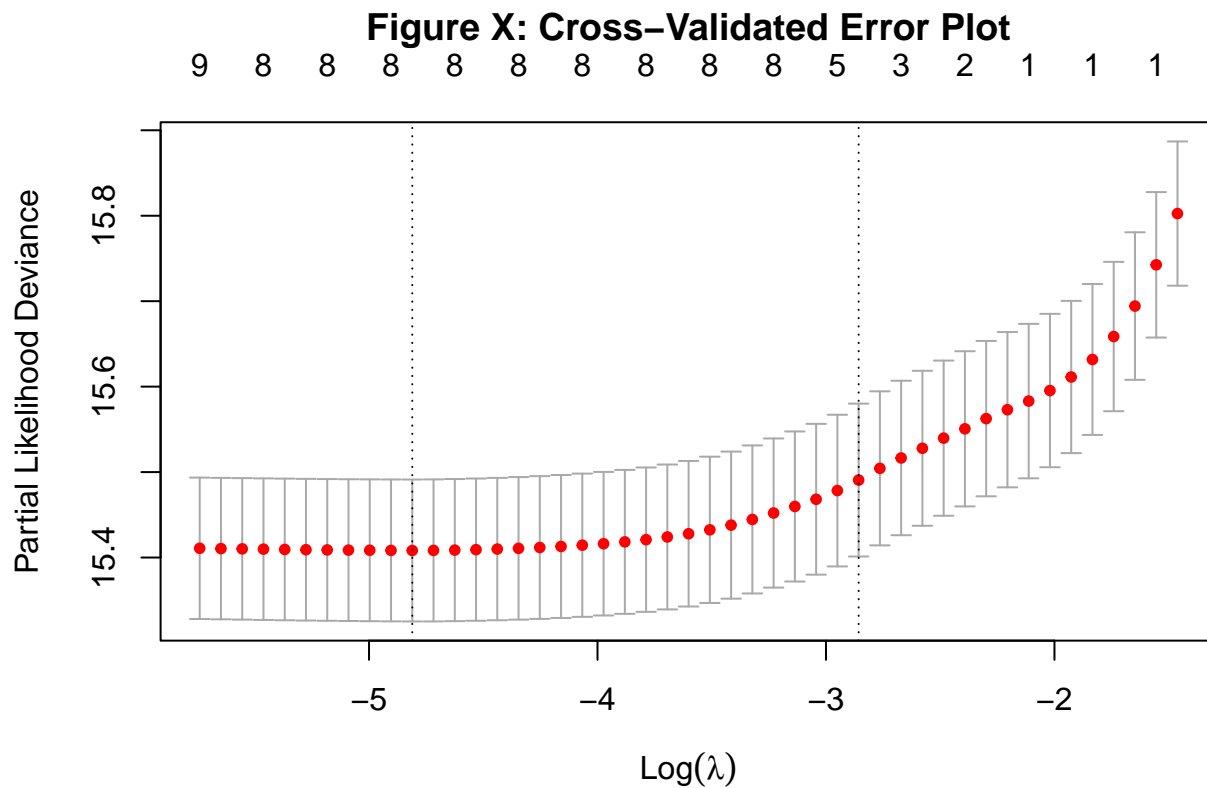
$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2$$

The Brier score is used to evaluate the accuracy of probabilistic predictions from a model; its value ranges from 0 to 1 with 0 being perfect and 1 being the opposite. We calculate the Brier score using the test data set. For each observation in the test data set, predictions are made at the observed time of censoring or event.

## Results

Log-rank test

Regularized Cox Proportional Hazard Model



In the plot above, we see the results of our cross-validation with the partial likelihood deviance, a measure of error, on the y-axis, and the natural log of the penalty term on the x-axis. The far-left dashed line represents the  $\log(\lambda)$  value associated with the model that produced the minimum cross-validation error. The far-right dashed line represents the  $\log(\lambda)$  value associated with the “1se” model that results in a cross-validation error one standard error from the minimum. We choose the “1se” model in this application because it increases the penalization on the model, enforces parsimony, and helps to reduce overfitting of our model on the training data set.

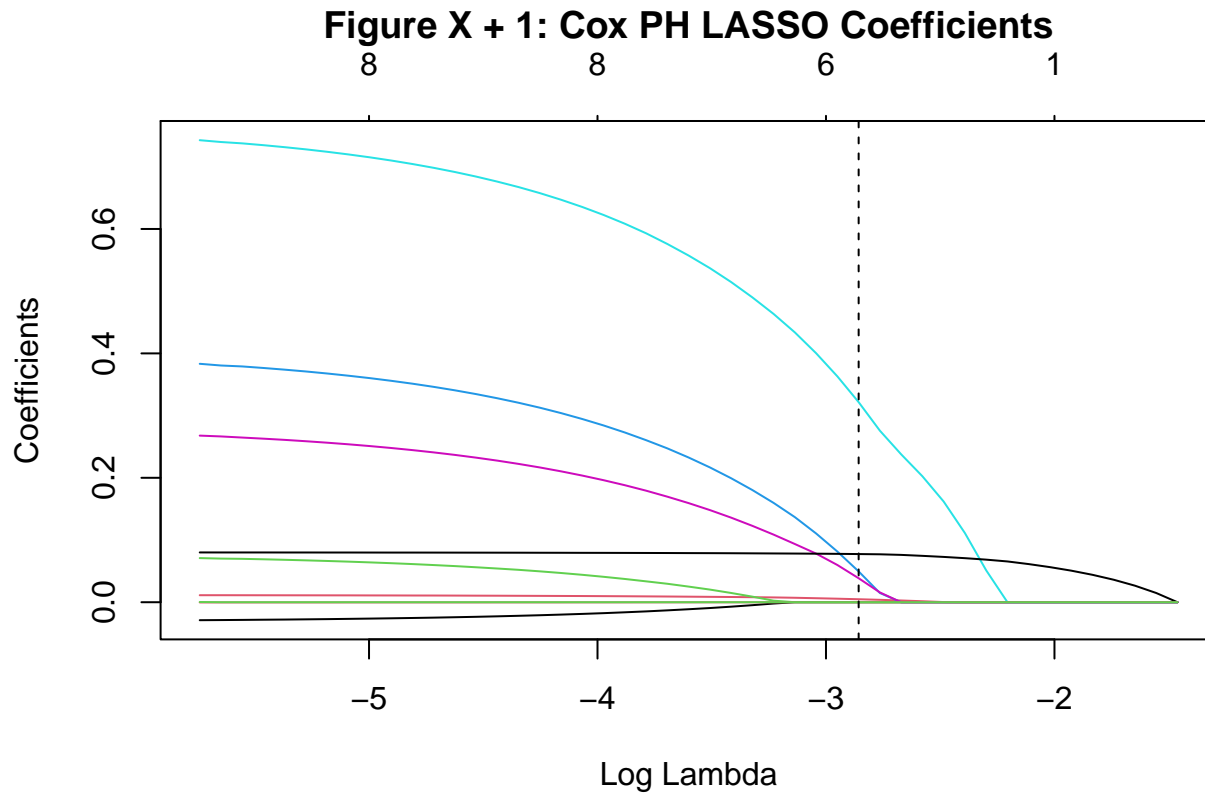


Figure X + 1 shows the values of each coefficient in the model as the penalty term increases. The dashed line shows the “1se” penalty term, and so the coefficient values at this point can be seen in the model. The cross-validated modeling results are shown below.

Table 3: Cox Proportion Hazard LASSO Coefficients (1se)

	coef	exp_coef
year	0.0000	1.0000
age	0.0048	1.0048
meno	0.0000	1.0000
size20-50	0.0495	1.0508
size>50	0.3216	1.3793
grade	0.0386	1.0393
nodes	0.0774	1.0805
pgr	0.0000	1.0000
er	0.0000	1.0000
hormon1	0.0000	1.0000
chemo1	0.0000	1.0000

Here, the regularization procedure removes **meno**, **er**, **hormon** and **chemo**. However, we are still interested in assessing the treatment effects of hormone therapy and chemotherapy, so we will add these back to fit our final model.

	2.5 %	97.5 %
age	1.0096113	1.0209412
size20-50	1.3250933	1.7610328
size>50	1.7728357	2.6497651
grade	1.1207981	1.5335347
nodes	1.0717882	1.0964130
pgr	0.9993588	0.9998718
hormon1	0.7818093	1.1517154
chemo1	0.8503259	1.2090508

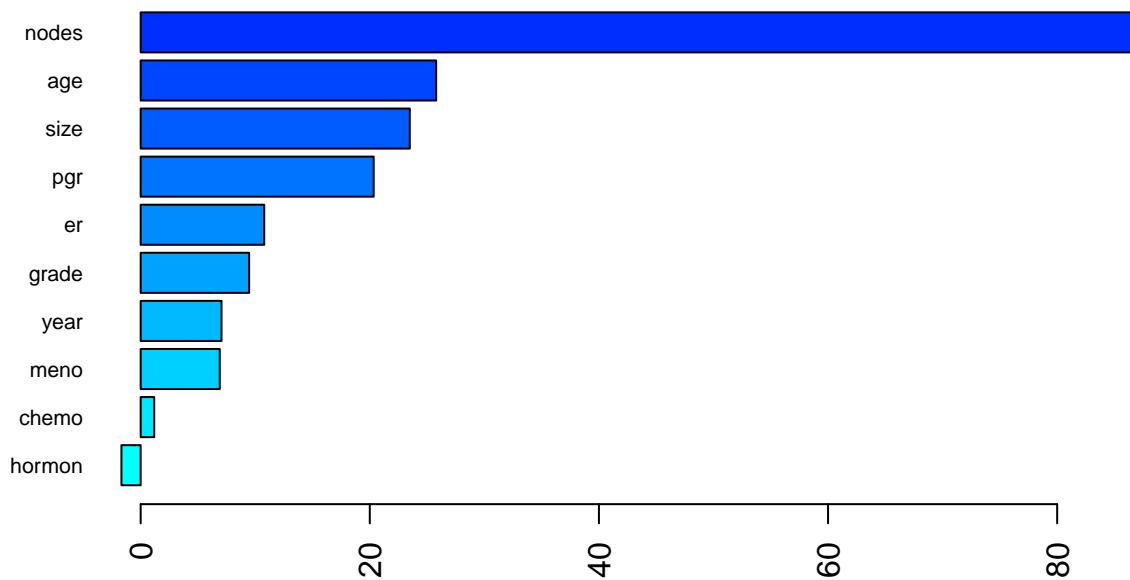
And we find significant effects for age at surgery, size of tumor, differentiation grade, number of positive lymph nodes, and pgr. We find non-significant effects for each of the two treatments of interest.

The main assumption of this model is that the hazards are proportional between subgroups. We can check this assumption by calculating and plotting Schoenfeld residuals, which can be found in the appendix. From the plots we see violations in number of positive lymph nodes and minor violations in age, tumor size, differentiation grade, and progesterone receptors where 0 does not lie in the Schoenfeld residual 95% confidence interval.

## Randomized Survival Forest

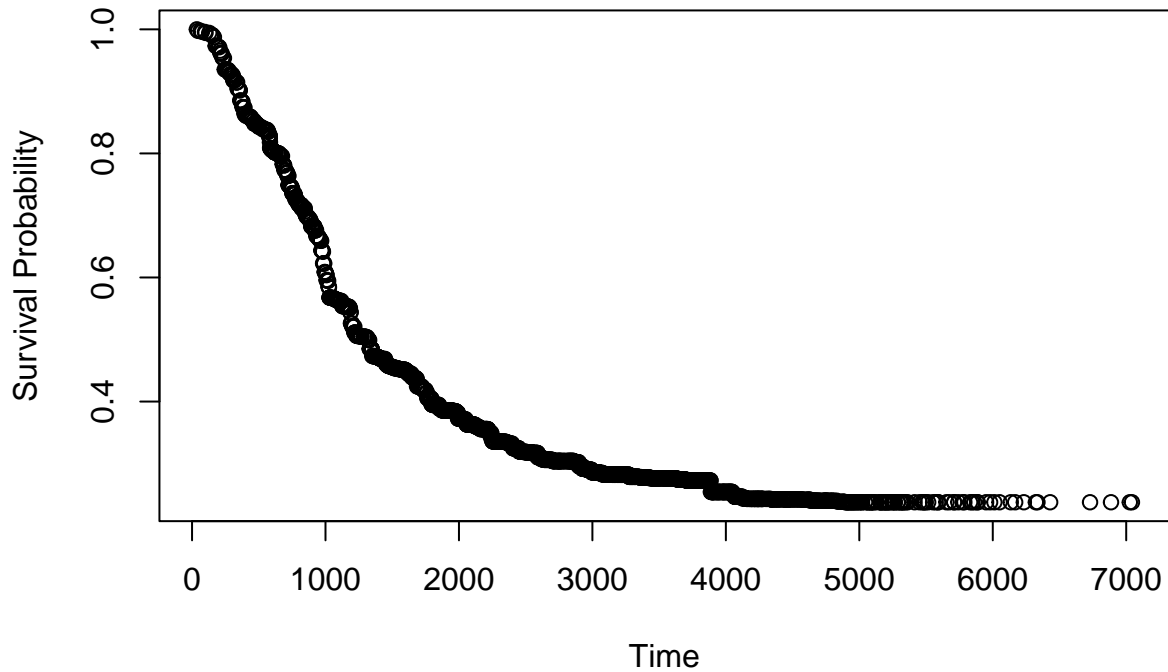
time	survival
36	1.0000000
45	0.9975471
64	0.9975471
74	0.9960829
97	0.9957616
101	0.9942716





With **ranger** package, we trained the random survival forest (RSF) with training dataset used for survival prediction. As a non-parametric method, there are no parameters in RSF that could be interpreted. The ultimate goal of RSF is to predict the survival probability function of a given data point based on its covariate vector. Compared to semi-parametric Cox-PH model which forces the outcome and the covariates to have a special connection, the RSF makes prediction based on the survival time of training data points that shares similar propensity with the given input data point.

## Survival Prediction for Patient 7



	time	survival
382	1319	0.4994471
383	1322	0.4994471
384	1323	0.4990825
385	1325	0.4990825
386	1327	0.4990101
387	1331	0.4849616

	year	age	meno	size	grade	nodes	pgr	er	hormon	chemo	dtime	death
2463	1992	69	1	20-50	2	8	5	6	1	0	1869	0

Since the “truth” of test data point (a single survival time) and the prediction we made here (a survival probability function) are not comparable, here we show the prediction result of the 7th test data point (pid = 58). The survival curve has been shown above, and the estimated median survival time is 1217 days.

## Brier Score

First, we calculated the Brier score for the Cox proportional-hazards model.

```
## [1] 0.329
```

The Brier score for the Cox proportional-hazards model is 0.329.

Second, let's calculate the Brier score for the random survival forest model.

```
## [1] 0.319
```

The Brier score for the random survival forest model is 0.319. The two models have very similar Brier scores. While we did not perform any statistical tests, given how close the two values are, it is fair to conclude that the Cox proportional-hazards model and the random survival forest are comparable in terms of predicting the outcome. However, if the goal of an analysis is inference, the semi-parametric Cox proportional-hazards model would be preferred to the non-parametric random survival forest.

An alternative metric that could be used instead of the Brier score is the area under the receiver operating characteristic curve.

## Discussion

### Conclusions

Hormone therapy and chemotherapy appear to have non-significant treatment effects on outcomes after adjusting for covariates. Tumor size and differentiation grade had the largest magnitude effects on hazard and were significant. The Cox proportional-hazards and random survival forest models performed similarly in predicting the outcome.

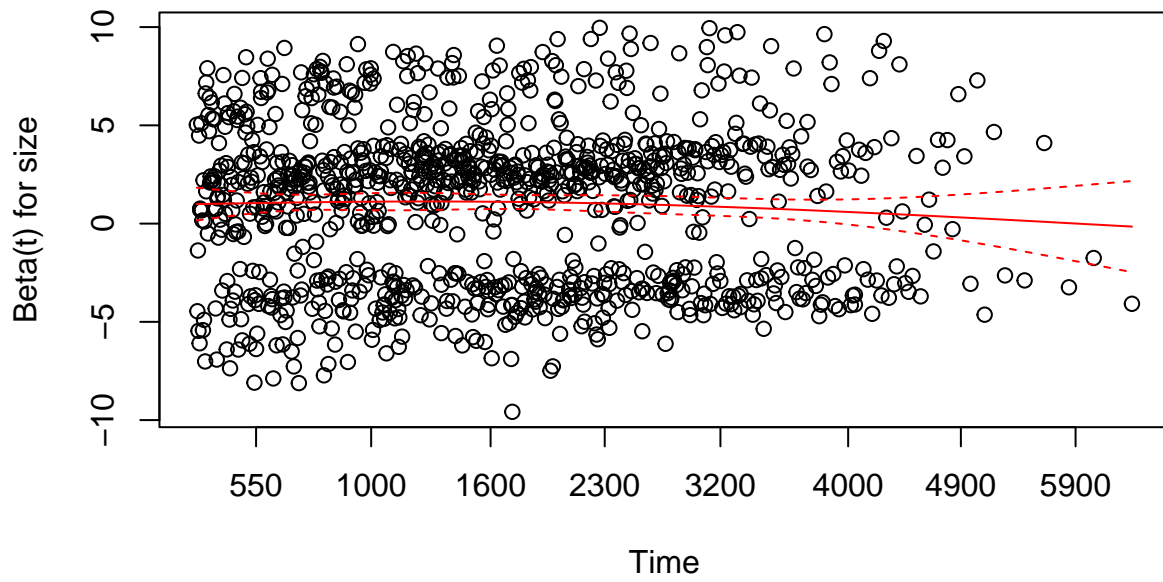
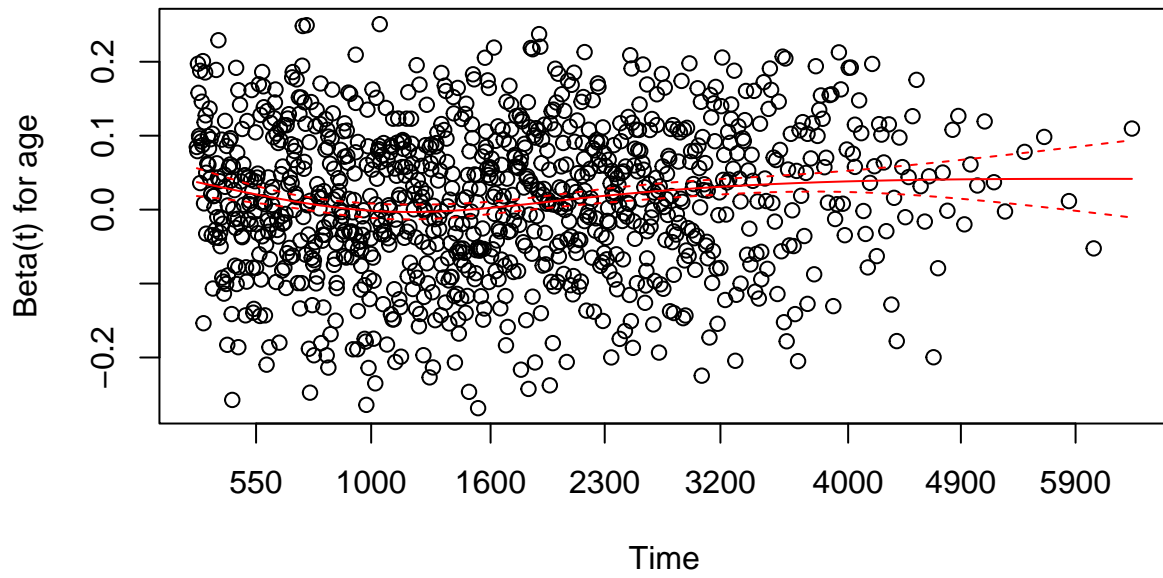
Some limitations are that there was a small sample size for the treated, especially when it comes to treatment subgroup of receiving both chemo and hormone therapy (only 1% of samples). Additionally, being non-parametric, there is no valid inference methods for random survival forest with potential bias for estimated individual survival curve. We ignored remission. Additionally, there were some violations in the proportional-hazards assumption for the Cox proportional-hazards model.

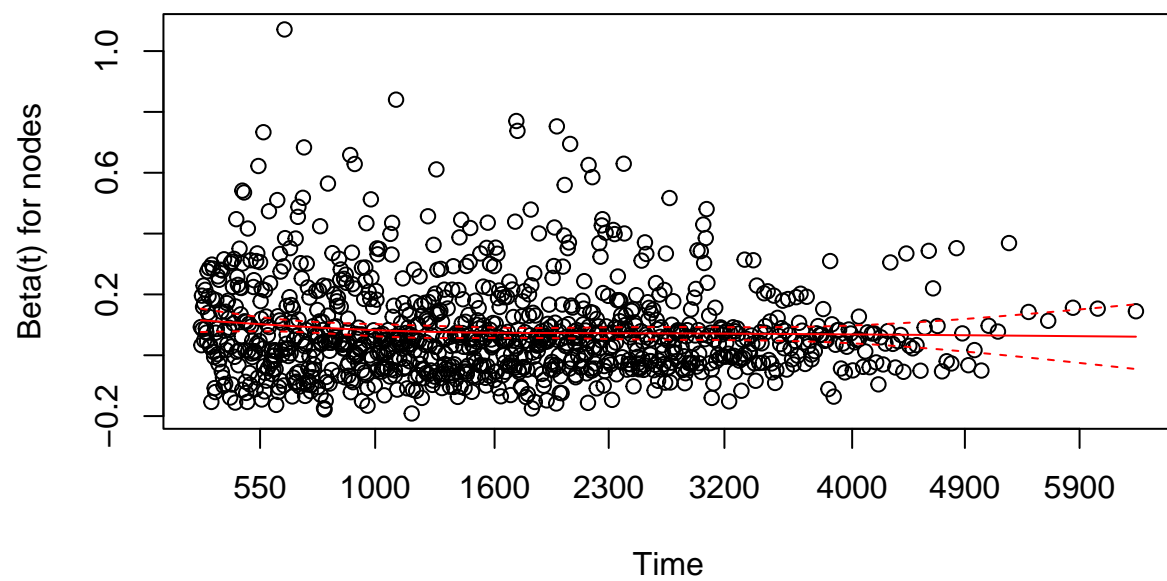
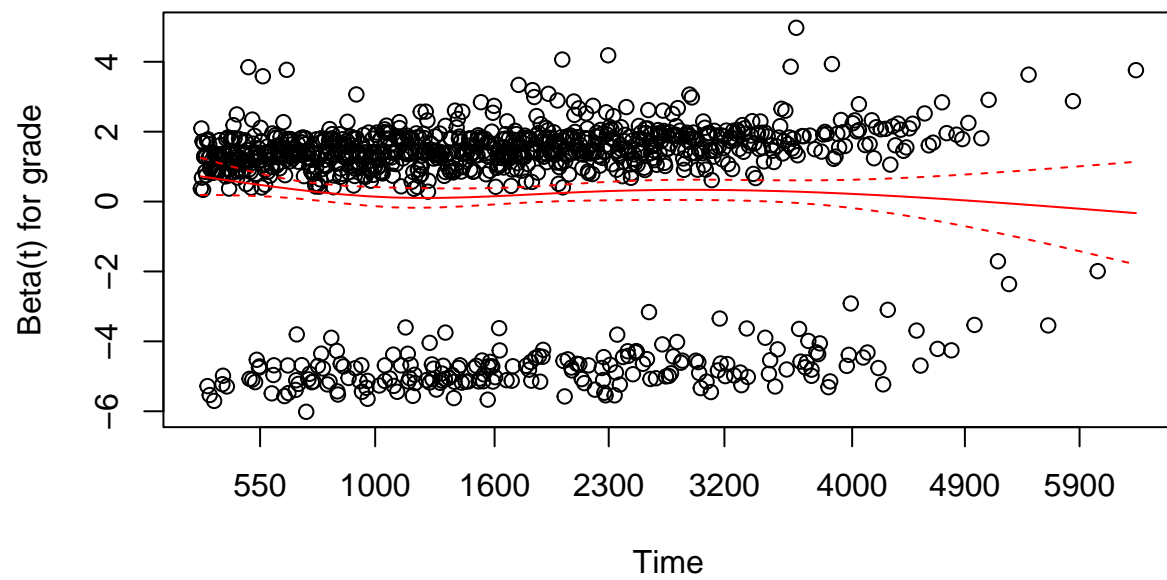
### Limitations and Future Work

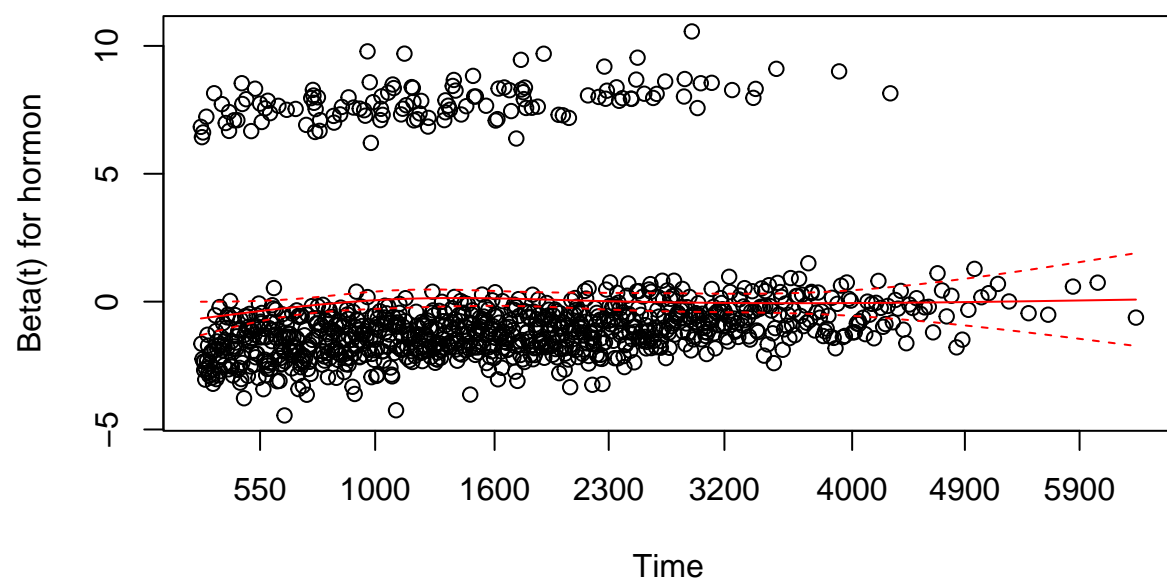
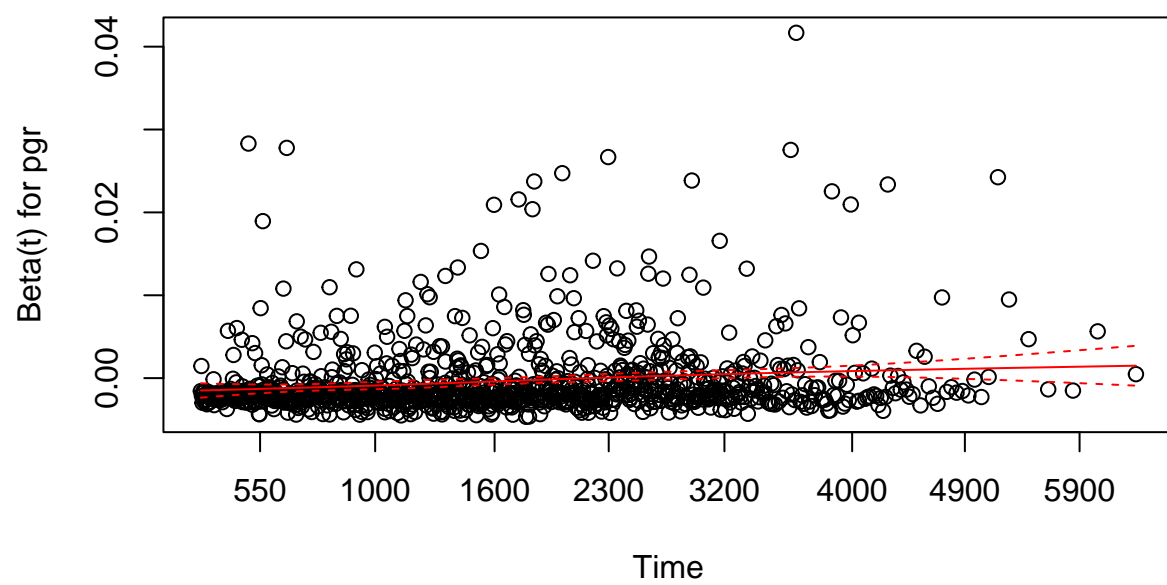
- talk about proportional hazard assumption violation here, ways to mitigate

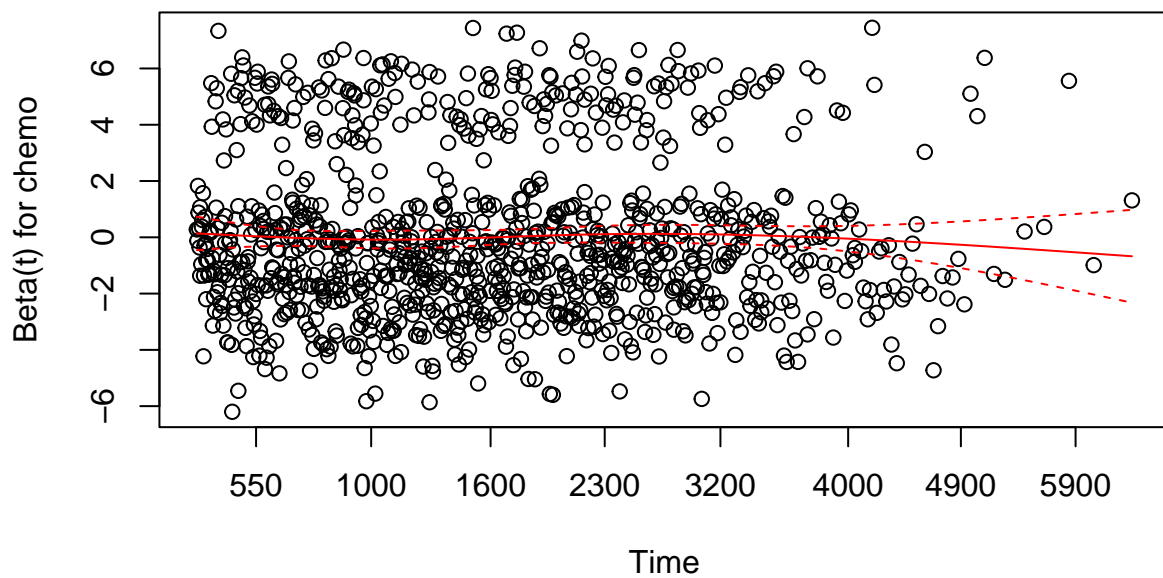
## Appendix

Calculating and plotting Schoenfeld residuals.









From these plots, we see violations in number of positive lymph nodes and minor violations in age, tumor size, differentiation grade, and progesterone receptors where 0 does not lie in the Schoenfeld residual 95% confidence interval.

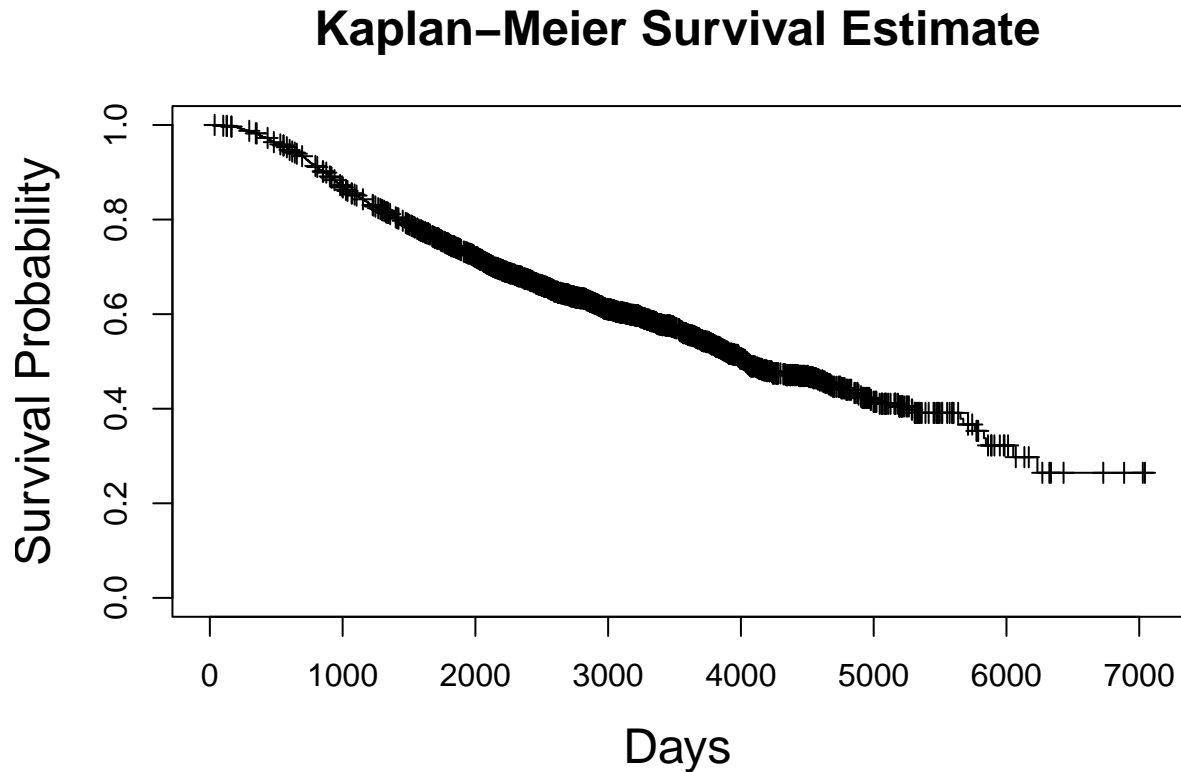
— final report ends here, move final report items to sections above this line —

## Cross-Validation

TM - do we need to explain cross-validation? Seems like it's more important to explain other methods. If we can, let's remove this section.

## Supplemental analyses

### Kaplan-Meier Survival Estimate



### Log-rank Tests

To answer the research questions of whether hormone treatment and chemotherapy are effective to breast cancer, the initial step is to test whether the survival probability functions of treatment group and control group are identical with log-rank test. In this report, we conducted 3 log-rank tests.

#### Hormone Therapy

First, we test whether hormone therapy is effective. The null hypothesis of this test is  $H_0 : S_1(t) = S_0(t)$ , where  $S_1(t)$  is the survival function of hormone treatment group,  $S_0(t)$  is the survival function of control group, and the alternative hypothesis is  $H_1 : S_1(t) \neq S_0(t)$ .

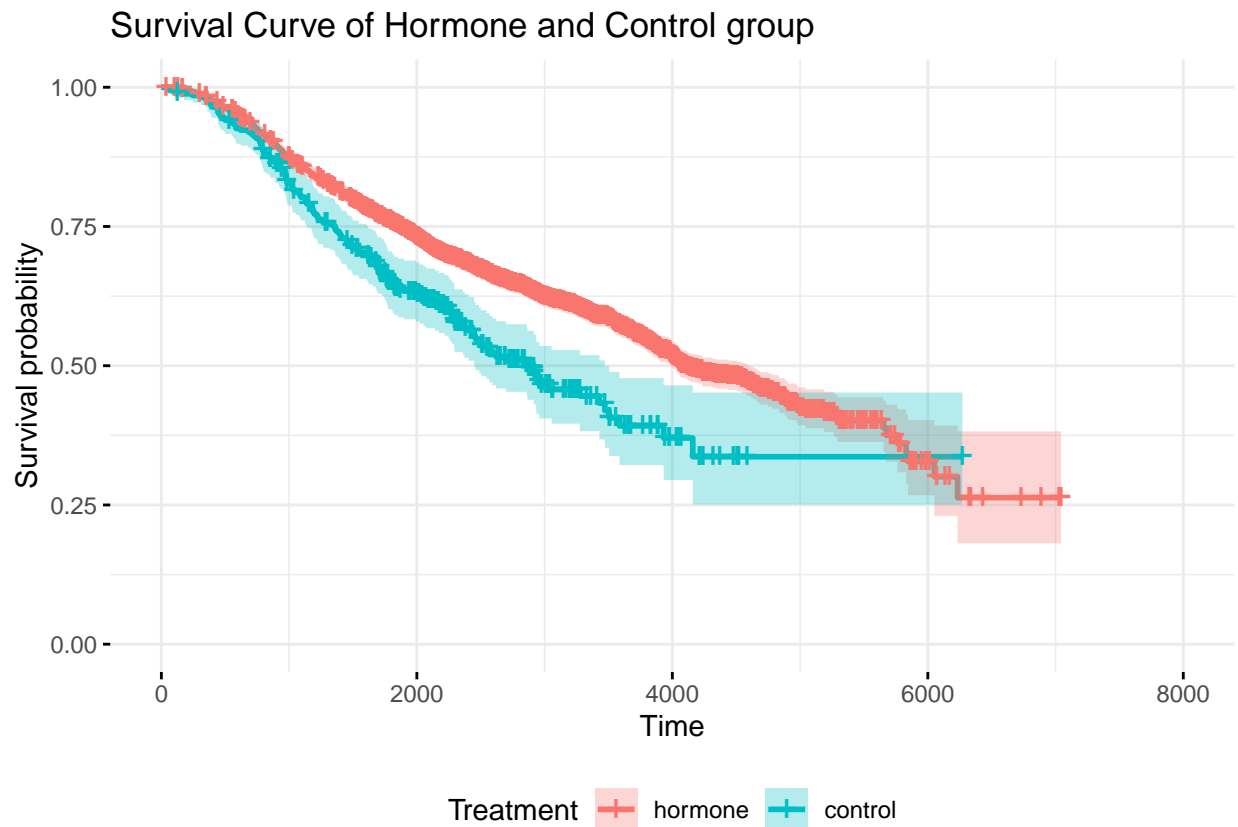
```
## Call:
## survdiff(formula = Surv(dtime, death) ~ hormon, data = rotterdam)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
```



```
## hormon=0 2643      1113      1162      2.04      23.7
## hormon=1  339       159       110     21.43     23.7
##
## Chisq= 23.7  on 1 degrees of freedom, p= 1e-06

## [1] 1.133649e-06
```

The test statistic is 23.7, and the corresponding p-value is  $1.133^{-6} \ll 0.05$ , thus we reject the null and conclude that we are 95% confident that  $S_1(t) \neq S_0(t)$ .



And based on the survival plot above, we can further conclude that the hormone treatment is effective to breast cancer on average.

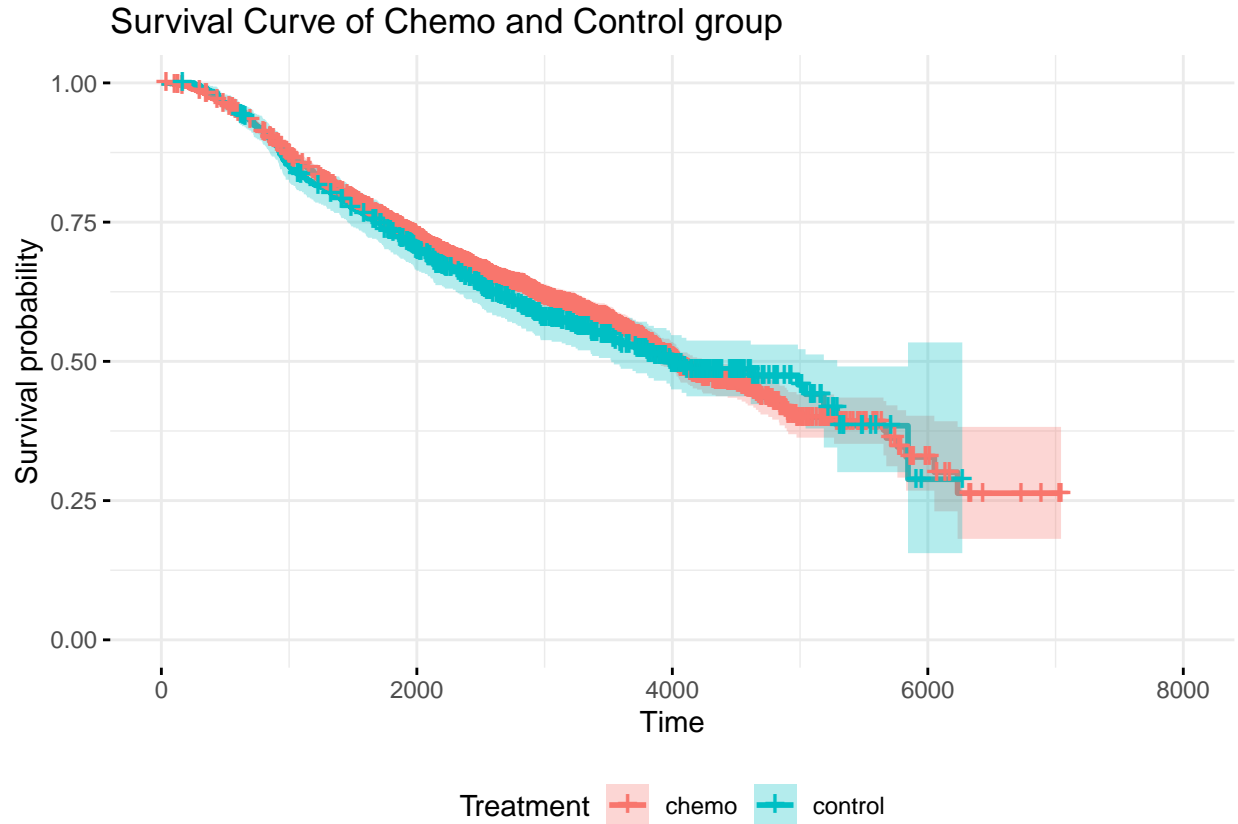
### Chemotherapy

Then, we test whether chemotherapy is effective. The null hypothesis of this test is  $H_0 : S_1(t) = S_0(t)$ , similarly,  $S_1(t)$  is the survival function of chemo treatment group,  $S_0(t)$  is the survival function of control group, and the alternative hypothesis is  $H_1 : S_1(t) \neq S_0(t)$ .

```
## Call:
## survdiff(formula = Surv(dtime, death) ~ chemo, data = rotterdam)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## chemo=0 2402      1014      1024   0.0963   0.495
## chemo=1  580       258       248   0.3977   0.495
##
## Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

```
## [1] 0.4818191
```

The test statistic is 0.495, and the corresponding p-value is  $0.48 > 0.05$ , thus we fail to reject the null and conclude that we are 95% confident that  $S_1(t) = S_0(t)$ . In other words, the chemotherapy is not effective to the treatment of breast cancer.



### Combined Analysis

After testing the effect of hormone and chemotherapy separately, we are interested in whether there are interaction (or confounding) between the two treatments. Hence, we reassign the data into 4 new treatment groups: 1) receive both hormone and chemotherapy; 2) only receive hormone therapy; 3) only receive chemotherapy; and 4) not receiving both of the therapies.

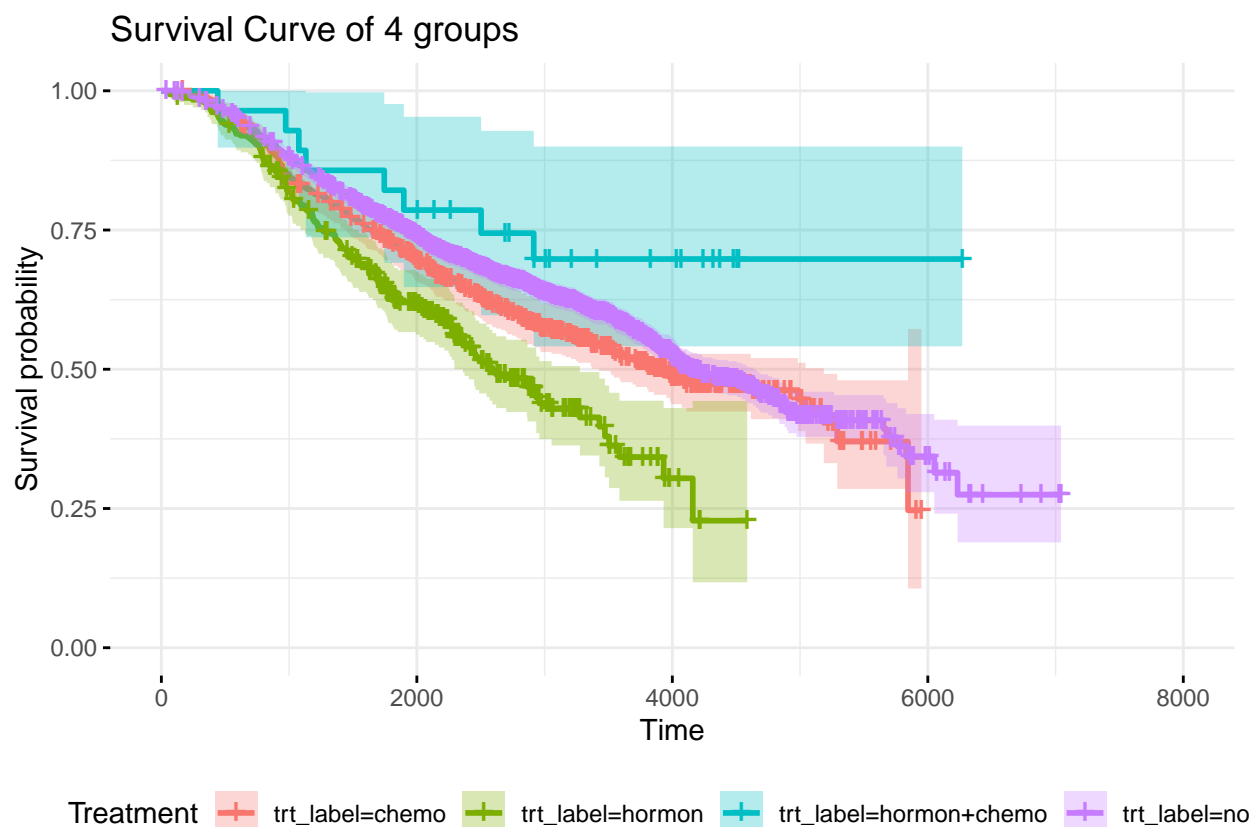
Then we tested the survival function of these 4 groups with log-rank test. Here, the null hypothesis is  $H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t)$ , corresponding to the survival function of 4 groups; and the alternative hypothesis is: the survival probability function of 4 groups are not identical.

```
## Call:
## survdiff(formula = Surv(dtime, death) ~ trt_label, data = rotterdam1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## trt_label=chemo      552      250      233.7      1.13      1.39
## trt_label=hormon     311      151      96.0     31.45     34.43
## trt_label=hormon+chemo   28       8      14.3      2.79      2.83
## trt_label=none     2091     863     927.9      4.54     16.84
##
## Chisq= 40.4  on 3 degrees of freedom, p= 9e-09
```

## [1] 8.838693e-09

The test statistics is 40.4, with p-value  $8.84^{-9} \ll 0.05$ , then reject the null and conclude that the survival function of 4 groups are not identical. But this result is not very meaningful, because we are more interested in the way they are different.

Consequently, we refer to the survival plot:



Based on the plot, it seems like: Hormone therapy or Chemotherapy itself cannot improve the survival of breast cancer, while the combination therapy of Hormone + Chemo can improve the survival.

Although it is a very interesting finding, but we should question whether this conclusion is valid. According to the summary data shown below, the sample size of hormone+chemo treatment group is only 28, which accounts for 8% of people who receive hormone therapy, and accounts for 1% of the whole dataset. As a result, the survival curve predicted for this combo treatment group can be unconvincing and biased comparing to other treatment groups, which makes the conclusion drawn above to be invalid.

Var1	Freq
chemo	552
hormon	311
hormon+chemo	28
none	2091

Although the results drawn from the third plot are not reliable, it is still a discovery that should be highlighted by researchers, since it is a common phenomenon biologically that two drugs can perform treatment effect only when taken together. More clinical trials and statistical analyses are suggested to further explore this issue.

Also, the test results of the first two log-ranks tests should be considered carefully for making decisions, because they didn't adjusted with the baseline covariates, then the observed treatment effects of hormone may be confounded by other covariates.

## **How our results compare with past research**

## **Conclusion**

## References

—Note this reference is in MLA format—

Simon, Noah et al. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent.”  
Journal of statistical software vol. 39,5 (2011): 1-13. doi:10.18637/jss.v039.i05