University of California, Santa Barbara
Department of Probability and Statistics
PSTAT 175 Final Project

Retinopathy Survival Analysis

Report by:

Anthony Castillo, Xin Chen, Vanessa Ho

Instructor:

Adam Tashmann: December 5th, 2019

Abstract

With the combination of lack of exercise and unhealthy meal choices, the amount of diabetic individuals have increased significantly from 108 million people in 1980 to 422 million people in 2014, according to the World Health Organization. Within the group of people with both type 1 and type 2 diabetes, diabetic retinopathy could be developed when blood sugar is poorly controlled. In this project, we built a Cox Proportional Hazard model to see how the type of laser used for surgery, type of diabetes, time to loss of vision or last follow-up and risk score for the eye affects the successfulness of surgery (based on the length of extended vision time after surgery) over time with the use of a right-censored data set of 197 patients with diabetic retinopathy.

Background Information and Data Source

We have obtained our data set “Diabetic Retinopathy” from the built in data within the Survival package. This data set includes 197 patients with 50% of them having “high risk” diabetic retinopathy according to the scales defined by the Diabetic Retinopathy Study (DRS). Each patient had a control eye with no treatment and a laser treatment eye; and the period of event is from the start of treatment to the time when visual acuity dropped below 5/200 two visits in a row. A built-in lag time of around 6 months is also included since the time difference between each visit in 3 months. Survival times are recorded by the following equation: actual time to vision loss in months minus the minimum possible time to event (around 6.5 months) and the censored data is either death, dropout or end of study.

The fixed covariates include: type of laser used, which eye was treated, type of diabetes, and whether or not eye was treated.

The variables are:

- laser (type of laser used): coded as 1 for argon and 2 for xenon
- eye (which eye was treated): coded as 1 for left eye and 2 for right eye
- age: age at diagnosis of diabetes
- type (type of diabetes): coded as 1 for adults and 2 for juvenile
- trt (control or treated eye): coded as 0 for control and 1 for treated
- futime (duration to vision loss): time to loss of vision or last follow up
- status (censored or loss of vision): coded as 0 for censored and 1 for loss of vision in eye
- risk (level of diabetic retinopathy diagnosed): a risk score of eye

```
head(rp)
```

	X <int>	id <int>	laser <fctr>	eye <fctr>	age <int>	type <fctr>	trt <int>	futime <dbl>	status <int>	risk <int>
1	1	5	argon	left	28	adult	1	46.23	0	9
2	2	5	argon	left	28	adult	0	46.23	0	9
3	3	14	argon	right	12	juvenile	1	42.50	0	8
4	4	14	argon	right	12	juvenile	0	31.30	1	6
5	5	16	xenon	right	9	juvenile	1	42.27	0	11
6	6	16	xenon	right	9	juvenile	0	42.27	0	11

6 rows

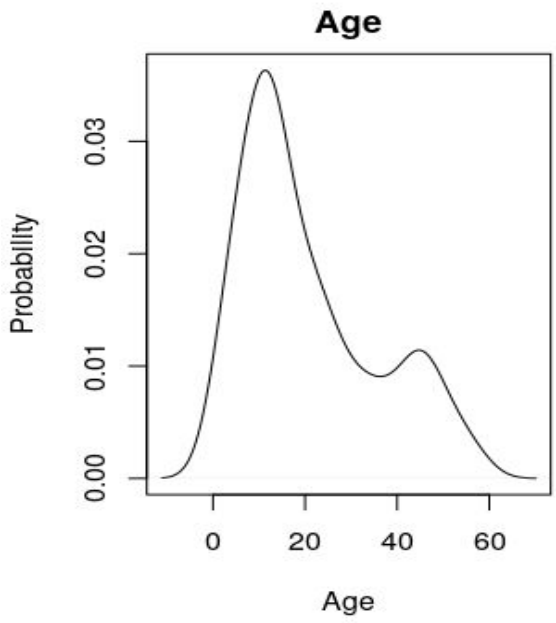
Research Question

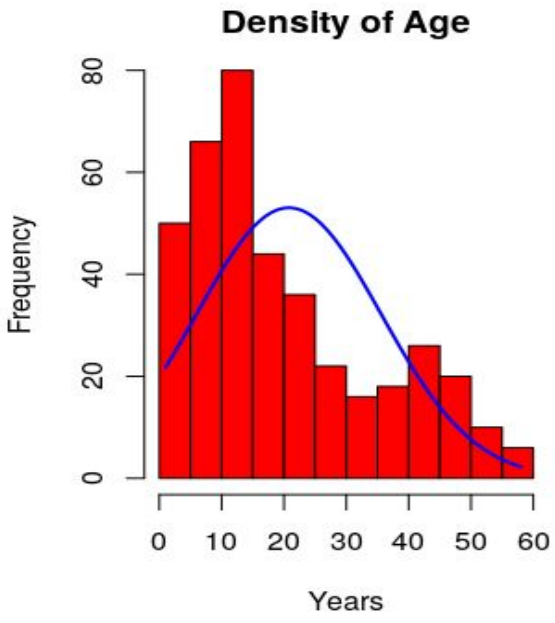
Our group is interested in whether treatment status, type of laser, right or left eye, type of diabetes, and the risk score would affect the patient's extended vision time upon surgery.

Data Exploration

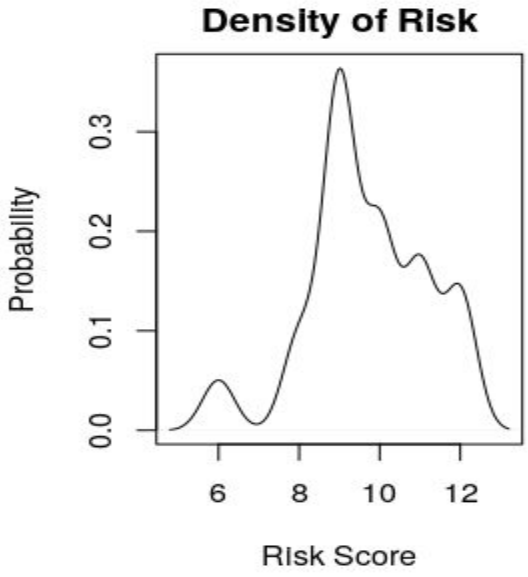
Below, we can see quantiles and kernel density estimates for the variables age, risk, and time. The bar plots match up with the kernel density estimates for the most part. Age, for the most part, is front-loaded with the maximum at around 20 years (one can call it bimodal with that second peak at around 42.5 years of age). The risk variable has its mode at around the 8-9 bin, with most of the data having high risk scores. The time variable has peaks at the 0-10 bin and the 40-50 bin. None of these density estimates are easily normal (Gaussian) distributions.

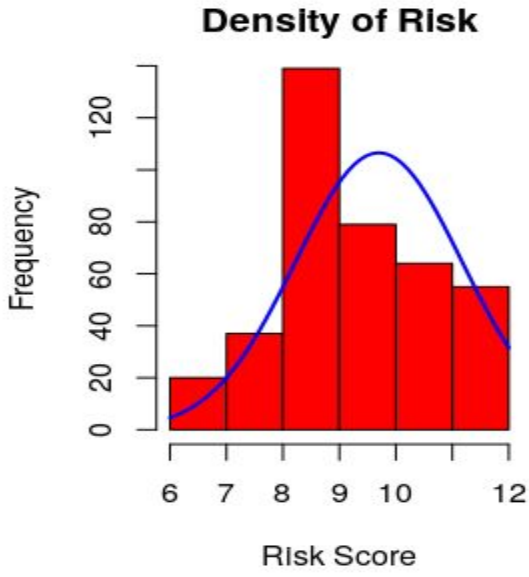
```
quantile(age)
0%  25%  50%  75% 100%
1   10   16   30   58
```





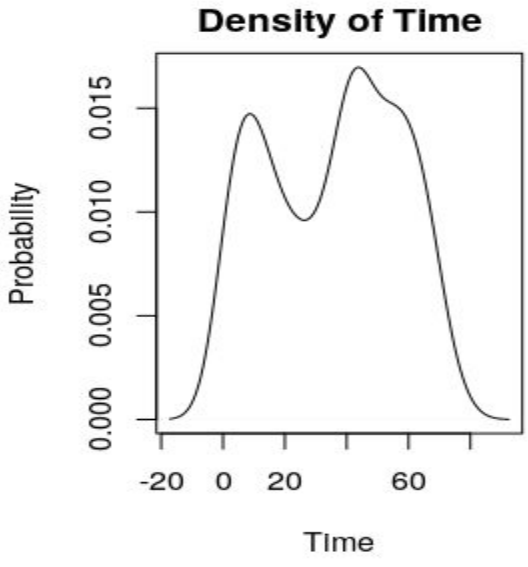
```
quantile(risk)
0%  25%  50%  75% 100%
6    9   10   11   12
```

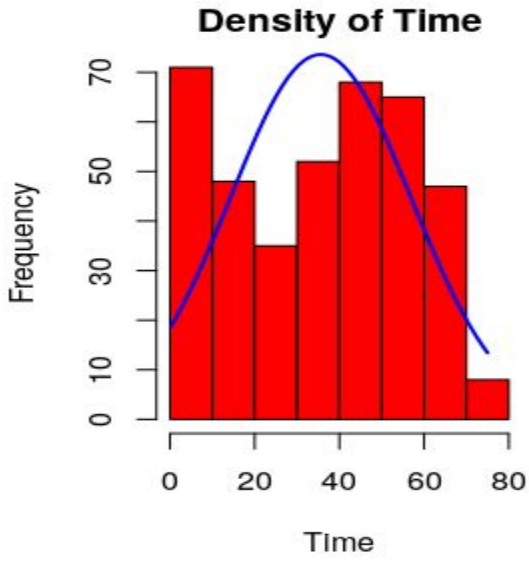





```
quantile(time)

0%      25%      50%      75%     100%
0.3000 13.9775 38.8000 54.2525 74.9700
```





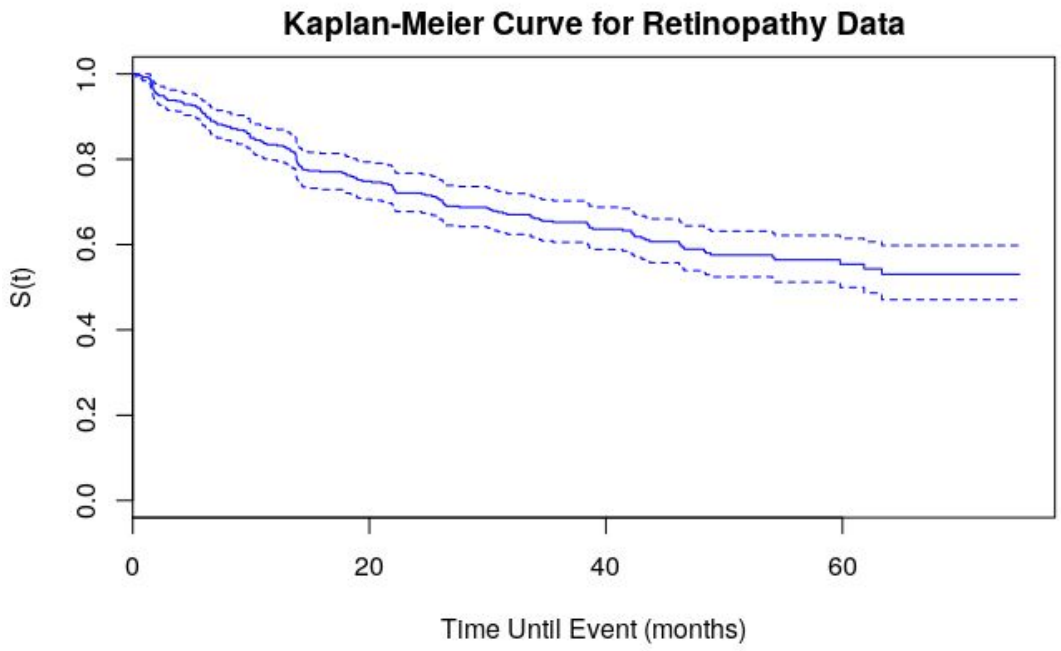
Our data summary clearly shows tallies for laser, eye, and type, along with number summaries for all other relevant factors.

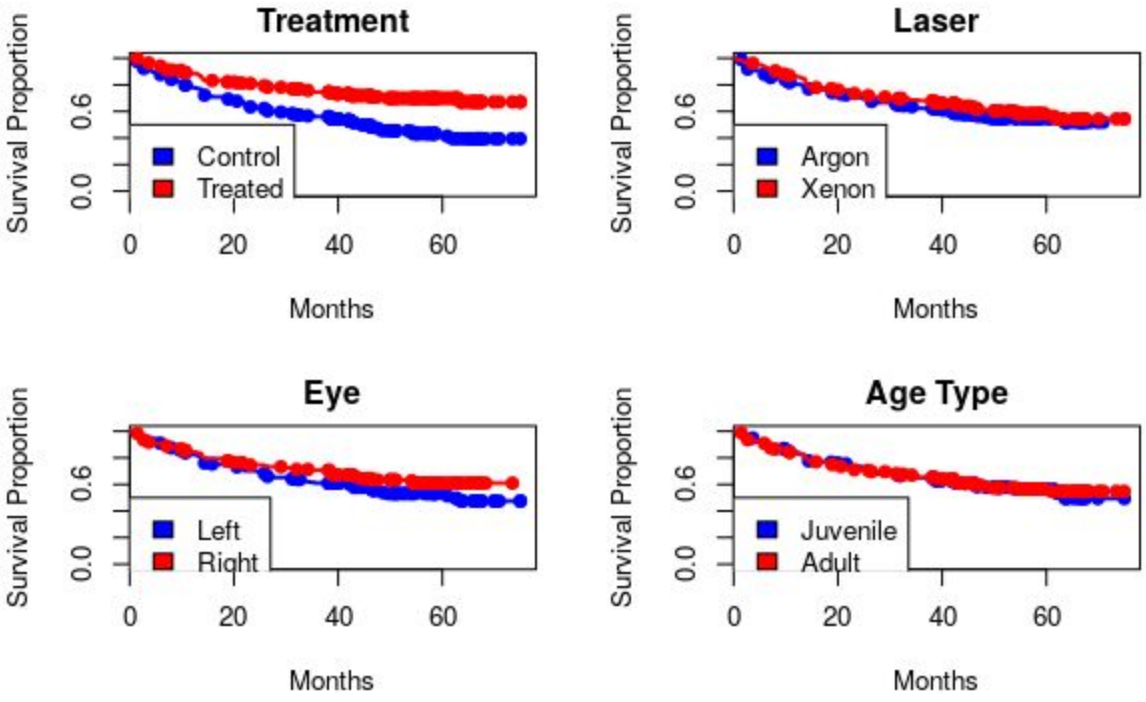
```
summary(rp)

      X                id              laser        eye            age               type
 Min.   :  1.00   Min.   :   5.0   argon:194   left :216   Min.   : 1.00   adult     :166
 1st Qu.: 99.25   1st Qu.: 480.0   xenon:200   right:178   1st Qu.:10.00   juvenile:228
 Median :197.50   Median : 834.0                           Median :16.00
 Mean   :197.50   Mean   : 873.2                           Mean   :20.78
 3rd Qu.:295.75   3rd Qu.:1296.0                           3rd Qu.:30.00
 Max.   :394.00   Max.   :1749.0                           Max.   :58.00
      trt           futime           status            risk
 Min.   :0.0   Min.   : 0.30   Min.   :0.0000   Min.   : 6.000
 1st Qu.:0.0   1st Qu.:13.98   1st Qu.:0.0000   1st Qu.: 9.000
 Median :0.5   Median :38.80   Median :0.0000   Median :10.000
 Mean   :0.5   Mean   :35.58   Mean   :0.3934   Mean   : 9.698
 3rd Qu.:1.0   3rd Qu.:54.25   3rd Qu.:1.0000   3rd Qu.:11.000
 Max.   :1.0   Max.   :74.97   Max.   :1.0000   Max.   :12.000
```

Kaplan-Meier Estimation Curves

First, we see a Kaplan-Meier Curve for retinopathy survival as it pertains to time without controlling for any of our covariates. In our first KM curve, we see survival gradually decrease until S(t) reaches a value of about .55 (which is after the 60-month mark). Afterwards, we plotted Kaplan-Meier curves that control for each individual binomial covariate (meaning: covariates whose outputs are only one of two possible results). In these individual KM curves, we see the only variable that experiences any real separation of survival probabilities is the treatment covariate (all others only diverge slightly or overlap for the span of time).





Log Rank Test to Confirm KM

After plotting Kaplan-Meier curves, we conduct a log rank test on each variable. We see that the p-value of risk and treatment are smaller than 0.05, which means these variables have significant effects on the successfulness of surgery.

```
Call:
survdiff(formula = rp.surv ~ type, data = rp)

                N Observed Expected (O-E)^2/E (O-E)^2/V
type=adult    166       68     67.2   0.00988    0.0175
type=juvenile 228       87     87.8   0.00756    0.0175

 Chisq= 0  on 1 degrees of freedom, p= 0.9
Call:
survdiff(formula = rp.surv ~ eye, data = rp)

            N Observed Expected (O-E)^2/E (O-E)^2/V
eye=left  216       95     84.7      1.26      2.79
eye=right 178       60     70.3      1.52      2.79

 Chisq= 2.8  on 1 degrees of freedom, p= 0.09
Call:
survdiff(formula = rp.surv ~ trt, data = rp)

        N Observed Expected (O-E)^2/E (O-E)^2/V
trt=0 197      101     71.8      11.9      22.2
trt=1 197       54     83.2      10.3      22.2

 Chisq= 22.2  on 1 degrees of freedom, p= 2e-06
Call:
survdiff(formula = rp.surv ~ laser, data = rp)

              N Observed Expected (O-E)^2/E (O-E)^2/V
laser=argon 194       79     72.8     0.534      1.01
laser=xenon 200       76     82.2     0.473      1.01

 Chisq= 1  on 1 degrees of freedom, p= 0.3


Call:
survdiff(formula = rp.surv ~ risk, data = rp)
          N Observed Expected (O-E)^2/E (O-E)^2/V
risk=6   20        6     8.68    0.8285    0.8791
risk=8   37       11    14.73    0.9467    1.0480
risk=9  139       41    61.67    6.9299   11.5767
```

```
risk=10  79       47    23.81   22.5717   26.8303
risk=11  64       25    26.33    0.0668    0.0806
risk=12  55       25    19.77    1.3842    1.5887

 Chisq= 33  on 5 degrees of freedom, p= 4e-06
Call:
survdiff(formula = rp.surv ~ age, data = rp)

        N Observed Expected (O-E)^2/E (O-E)^2/V
age=0 116       45     43.7    0.0388    0.0542
age=1  90       36     33.3    0.2129    0.2720
age=2  92       35     40.9    0.8641    1.1795
age=3  96       39     37.0    0.1061    0.1396

 Chisq= 1.2  on 3 degrees of freedom, p= 0.7
```

Model Building

Now, we start to build our Cox PH model. We are using both backward elimination method and forward stepwise selection method to pick the right set of covariates. First we build a full model with all covariates. Then we can use the function "step" in R to apply backward elimination. It shows that the risk and trt are the two significant variables.

```
full = coxph(Surv(time, cns) ~ age+ risk+laser + type + trt + eye)
step(full, direction = "backward")


Start:  AIC=1714.53
Surv(time, cns) ~ age + risk + laser + type + trt + eye

        Df    AIC
- type   1 1712.8
- laser  1 1713.6
- age    1 1713.7
- eye    1 1714.2
<none>     1714.5
- risk   1 1719.0
- trt    1 1735.4

Step:  AIC=1712.82
Surv(time, cns) ~ age + risk + laser + trt + eye
```

```
        Df    AIC
- laser  1 1711.9
- age    1 1712.0
- eye    1 1712.5
<none>     1712.8
- risk   1 1717.3
- trt    1 1733.7


Step:  AIC=1711.87
Surv(time, cns) ~ age + risk + trt + eye

       Df    AIC
- age   1 1711.0
- eye   1 1711.9
<none>    1711.9
- risk  1 1716.5
- trt   1 1732.4

Step:  AIC=1710.98
Surv(time, cns) ~ risk + trt + eye

       Df    AIC
- eye   1 1710.6
<none>    1711.0
- risk  1 1715.3
- trt   1 1731.1

Step:  AIC=1710.64
Surv(time, cns) ~ risk + trt

       Df    AIC
<none>    1710.6
- risk  1 1715.6
- trt   1 1731.1
Call:
coxph(formula = Surv(time, cns) ~ risk + trt)

         coef exp(coef) se(coef)      z        p
risk  0.14600   1.15720  0.05588  2.613  0.00898
trt  -0.77792   0.45936  0.16881 -4.608 4.06e-06

Likelihood ratio test=29.33  on 2 df, p=4.282e-07
n= 394, number of events= 155
```

Model Checking

Now, we use the likelihood tests to select covariates. First, we check the anova table for the full model with all covariates. From the two variables that are significant, we choose treatment (trt) as the most significant one and then add risk to build the model.

```
anova(full)
Analysis of Deviance Table
Cox model: response is Surv(time, cns)
Terms added sequentially (first to last)

       loglik   Chisq Df Pr(>|Chi|)
NULL  -867.99
age   -867.85  0.2656  1   0.606281
risk  -864.33  7.0419  1   0.007963 **
laser -863.83  1.0043  1   0.316283
type  -863.64  0.3840  1   0.535483
trt   -852.12 23.0396  1  1.587e-06 ***
eye   -851.27  1.7035  1   0.191832
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

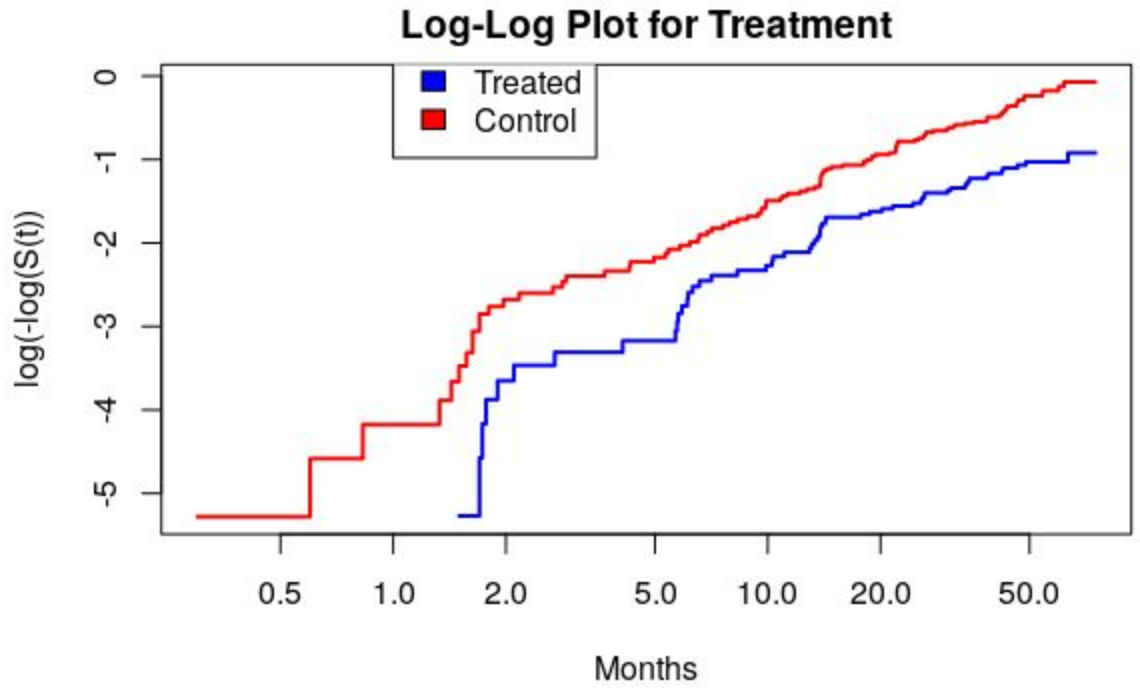
Residual Tests (cox.zph)

As p value > 0.05 , we see no need to stratify.

```
rp.coxzph <- cox.zph(rp.coxph)
            rho  chisq     p
trt      -0.0303 0.1237 0.725
risk     -0.0932 1.3812 0.240
trt:risk  0.0204 0.0556 0.814
GLOBAL        NA 2.2959 0.513
```

Log-log plot

From the graph for the log-log plot for 'trt,' we see no need to stratify our data.



Interaction term

Because `risk:trt` is not statistically significant (p-value = .552 > .05), our model is conclusively `formula = Surv(time, cns) ~ risk + trt.`

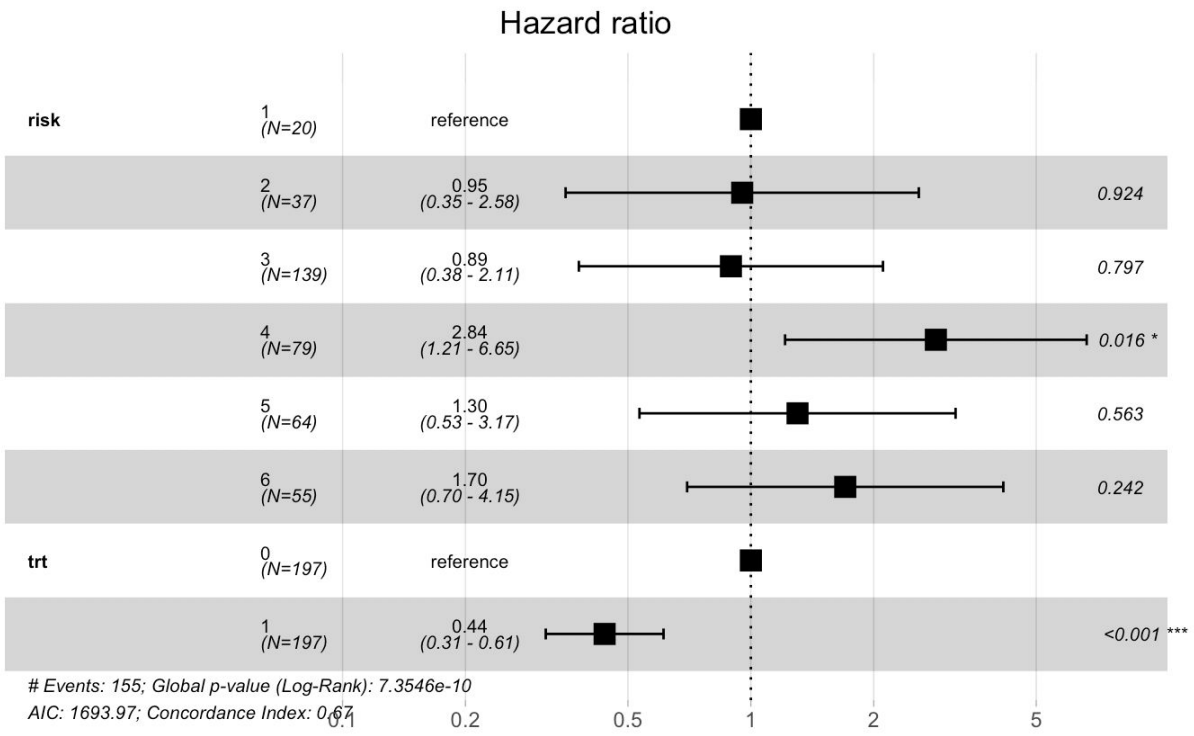
```
rp.coxph <- coxph(Surv(time, cns) ~ risk*trt, data=rp)

Call:
coxph(formula = Surv(time, cns) ~ risk * trt, data = rp)

             coef exp(coef) se(coef)     z     p
risk       0.1705    1.1859   0.0696  2.45 0.014
trt       -0.0932    0.9110   1.1617 -0.08 0.936
risk:trt -0.0691    0.9332   0.1163 -0.59 0.552

Likelihood ratio test=29.68  on 3 df, p=2e-06
n= 394, number of events= 155
```

Hazard Ratios and CI



The chart above shows that the hazard ratio of treatment is centered at 0.44 with a 95% confidence interval of 0.31 to 0.61. This value indicates that individuals' eyes that received treatment have 56% less likelihood to extend actual time to vision loss (in months) than the eyes that did not receive treatment. In the representation of this table, the labeling for risk has also been altered to {1, 2, 3, 4, 5, 6} with:

1 as risk factor of 6

2 as risk factor of 8

3 as risk factor of 9

4 as risk factor of 10

5 as risk factor of 11

6 as risk factor of 12

in order to use the ggforest() function. There is no collected data with a risk factor of 7. Through analyzing the hazard ratio chart, we can see that the hazard ratio of risk factor of 8 is centered at 0.95 with its 95% confidence interval between 0.35 and 2.58, which indicates that compared to a risk factor of 6, there is a 5% less likelihood to extend

actual time to vision loss. The hazard ratio of risk factor of 9 is centered at 0.89 (95% CI [0.38, 2.11]) had 11% less likelihood to extend actual time to vision loss compared to risk factor of 6. While the hazard ratio of risk factor 10, centered at 2.84 (95% CI [1.21, 6.65]) had 184% more likelihood to extend time to vision loss and risk factor 11, centered at 1.30 (95% CI [0.53, 3.17]) had 30% more likelihood to extend time to vision loss. Lastly, for risk factor 12, its hazard ratio is centered at 1.70 (95% CI [0.7, 4.15] had 70% more likelihood to extend time to actual vision loss. With a lower risk score, the act of extending vision time through surgery may not be necessary as it may lead to the same length of vision without surgery. Therefore, there is less of a likelihood for risks scores below but not including 10. With risk scores that are 10 and above, surgery would be a better option because they would, in general, have less vision time left. Performing the surgery, in this case, could extend their vision time in comparison to not receiving the surgery.

Extension - AFT models

A summary of a lognormal model of our data (with treatment and risk as the covariates) shows statistical significance for both trt and risk, and thus our original model check is verified by this metric.

```
rpfit <- survreg(Surv(time,cns) ~ trt + risk, data=rp, dist="lognormal")

summary(rpfit)

Call:
survreg(formula = Surv(time, cns) ~ trt + risk, data = rp, dist = "lognormal")
              Value Std. Error     z       p
(Intercept)  5.9082     0.7838  7.54 4.8e-14
trt          1.0042     0.2291  4.38 1.2e-05
risk        -0.2123     0.0775 -2.74  0.0062
Log(scale)   0.6248     0.0633  9.87 < 2e-16

Scale= 1.87

Log Normal distribution
Loglik(model)= -828.7   Loglik(intercept only)= -842.9
       Chisq= 28.3 on 2 degrees of freedom, p= 7.2e-07
Number of Newton-Raphson Iterations: 3
n= 394
```

A summary of a weibull model of our data (with treatment and risk as the covariates) shows statistical significance for both trt and risk, and thus our original model check is verified by this metric.

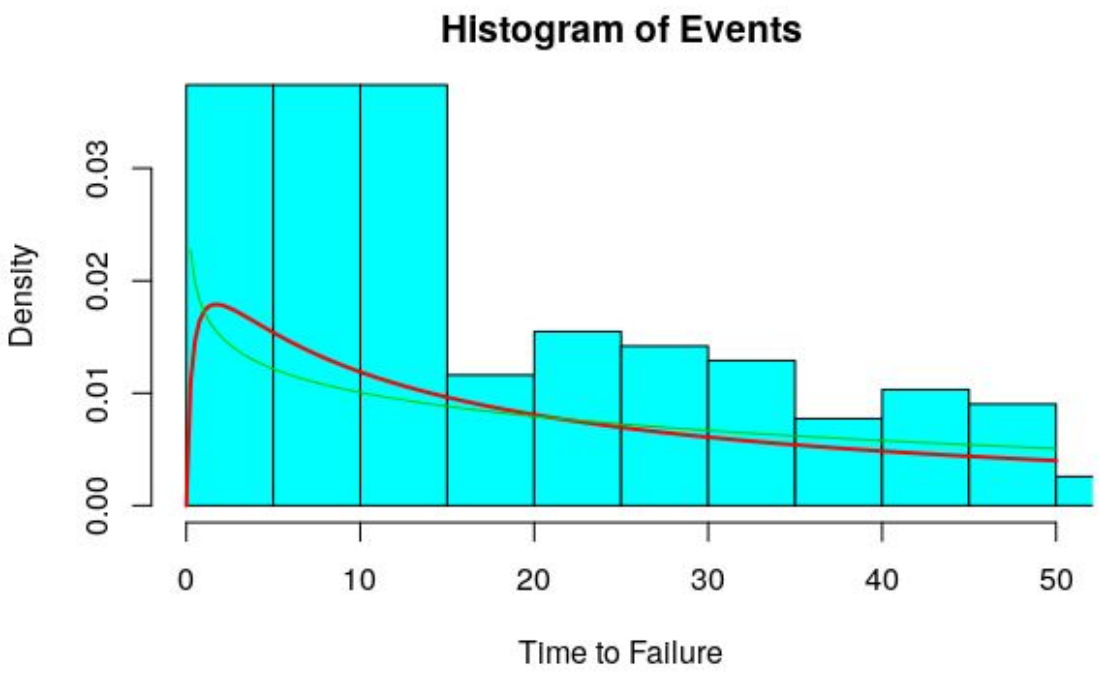
```
rpwfit <- survreg(Surv(time,cns) ~ trt + risk, data=rp, dist="weibull")

summary(rpwfit)

Call:
survreg(formula = Surv(time, cns) ~ trt + risk, data = rp, dist = "weibull")
              Value Std. Error     z       p
(Intercept)  5.9942     0.7042  8.51 < 2e-16
trt          0.9707     0.2154  4.51 6.6e-06
risk        -0.1793     0.0692 -2.59  0.0095
Log(scale)   0.2039     0.0722  2.82  0.0048

Scale= 1.23
Weibull distribution
Loglik(model)= -832.9   Loglik(intercept only)= -848
Chisq= 30.18 on 2 degrees of freedom, p= 2.8e-07
Number of Newton-Raphson Iterations: 5
n= 394
```

Below, we see our time to failure in the form of a density histogram. According to this histogram, most of our data is front-loaded in the first three bins. We also see this in the weibull and log-normal kernel density estimate curves as their respective statistical modes are both within the first bin (time to failure = 0 - 5). Afterwards the data tapers off over time, which is expected since failure must happen eventually and since observations with high failure times would be a statistical rarity.



Our ANOVA test of the weibull model shoes z-scores for trt that are both less than .05, and thus both covariates are statistically significant (as demonstrated earlier in this report).

```
anova(rpwfit)
```

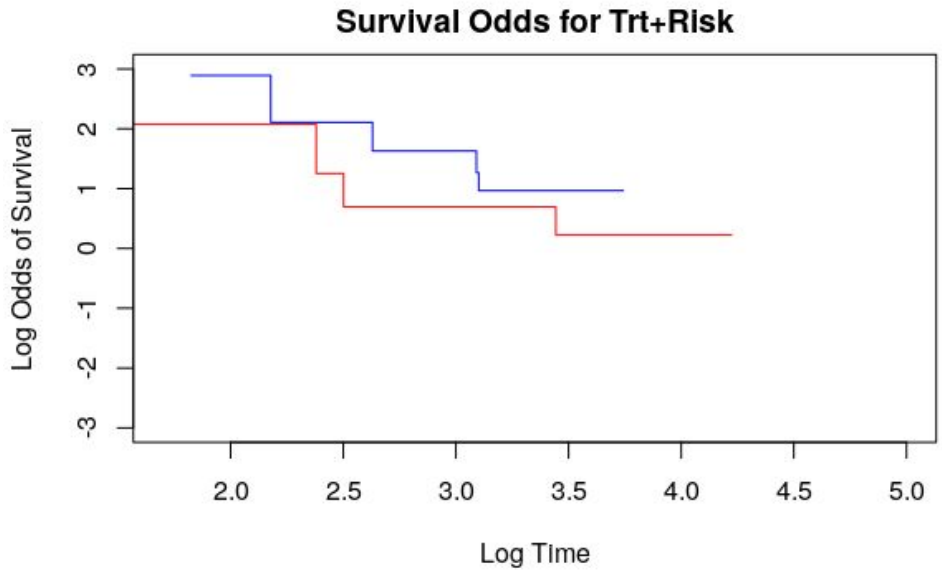
	Df	Deviance	Resid. Df	-2*LL	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
NULL	*NA*	*NA*	392	1695.939	*NA*
trt	1	23.180392	391	1672.758	1.474935e-06
risk	1	7.004184	390	1665.754	8.131945e-03

Lastly, we performed an exponential fit on a survival model of our data and chosen factors, and then took an ANOVA summary of that new model. As seen below, treatment and risk are both statistically significant (with their respective p-values being less than .05), and thus, as seen before, we keep them in our final model.

```
rpfit.e <- survreg(Surv(time,cns) ~ trt+risk, data=rp, dist="exponential")
anova(rpfit.e)
```

	Df	Deviance	Resid. Df	-2*LL	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
NULL	*NA*	*NA*	393	1706.454	*NA*
trt	1	24.531838	392	1681.922	7.309203e-07
risk	1	7.533084	391	1674.389	6.057614e-03

Lastly, we plot a log-odds model of our data with the chosen factors of treatment plus risk (trt + risk), Below, we see Kaplan-Meiers curves of our chosen model in terms of log-odds.



Conclusion

In this project, we have the data that includes 197 patients to investigate whether treatment status, type of laser, right or left eye, type of diabetes, and the risk score would affect the patient's extended vision time upon surgery. First of all, we plotted Kaplan-Meier curves to check for the significance of each covariate on survival probabilities on an individual basis. Afterwards, we built our model using AIC quantities (meaning: forward and backwards selection), and then tested our model for accuracy and whether the proportional hazards assumption was violated (it was not). We then tested for the statistical significance of an interaction term between the two significant covariates (trt and risk) to find such interaction is statistically insignificant. Lastly, plotted confidence intervals for our hazard ratios and then extended our research into an AFT

model extension, where weibull and log-normal models were plotted and tested, only to confirm our original results. Thus, our final model is:

```
Surv ~ trt + risk
```

Data Source

https://vincentarelbundock.github.io/Rdatasets/datasets.html
https://vincentarelbundock.github.io/Rdatasets/csv/survival/retinopathy.csv
https://vincentarelbundock.github.io/Rdatasets/doc/survival/retinopathy.html