Testicular Cancer EDA

Saul, Jhet, Frankie, and Chase

Research Questions:

- Github Repo: https://github.com/Smoorad99/485-TT

The questions we aim to solve with our data are the following:

- How does race affect survival rate/months survived for testi cancer?
  - Exploring the role race plays in testicular cancer survival rates could help us address disparities between different races. If we find that survival rate of testicular cancer is impacted by race, it would push us to explore why this may be. For example, it may indicate inequality in healthcare received by different races.
- How does survival rate/months change based on treatment options for testi cancer?
  - Exploring the relationship between survival rate/months and treatment method helps us understand which treatment methods are most effective. Investigating the quality of life patients experience while undergoing different treatments may also help us better understand the effectiveness of each treatment.
- How does the survival rate/months change based on the marital status of the patient?

Abstract:

Testicular cancer is one of the most common types of cancer in both adolescent and young adult males. Fortunately, it has a high five-year survival rate of about 95%. The data used in our analysis comes from the Surveillance, Epidemiology, and End Results Program (SEER) database. Multiple survival analysis methodologies were applied including Kaplan-Meier estimator and Cox proportional hazards regression. Key factors such as race, marital status, months from diagnosis to treatment, and order in which radiation and surgery were performed were all included in the analysis. Our model was found to be significant by three tests. It was found treatment, diagnosis timing, marital status, and race to be significant in predicting survival outcomes among those with testicular cancer. Survival analysis serves as a significant tool in investigating the history of testicular cancer. The findings of this study can assist in clinical decision making, improving patient care, and ultimately working towards a greater survival rate for those with this disease.

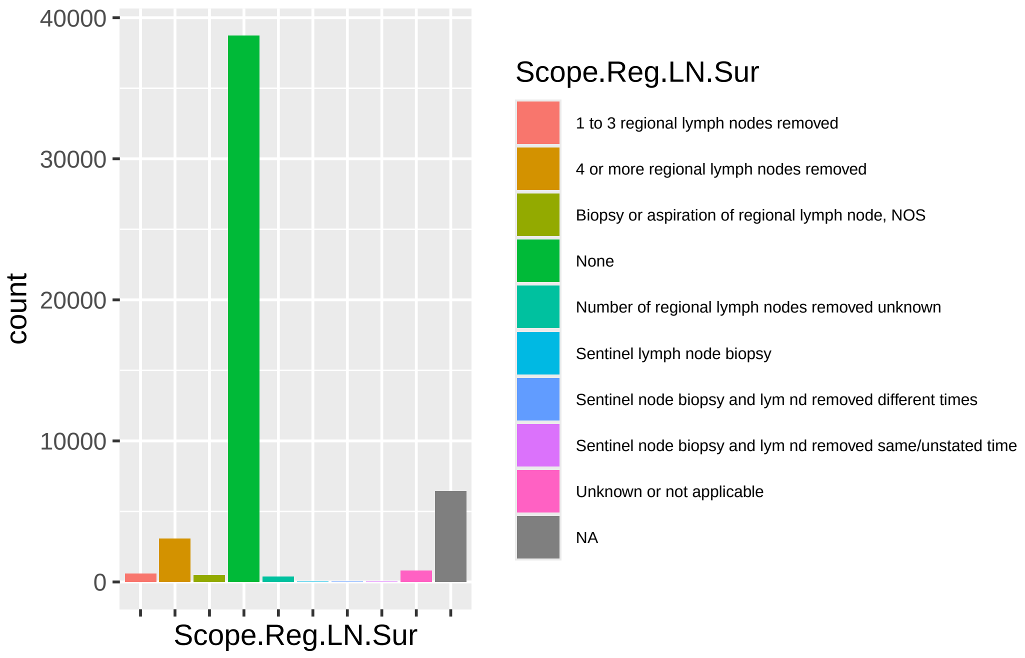
Exploratory Data Analysis

Jhet

After cleaning and managing the data, all of the variables I am in charge of analyzing are now categorical, meaning statistics such as mean, standard deviation and median are not very useful.

Instead, we can see the frequency distributions above in count form, as well as visualized below.

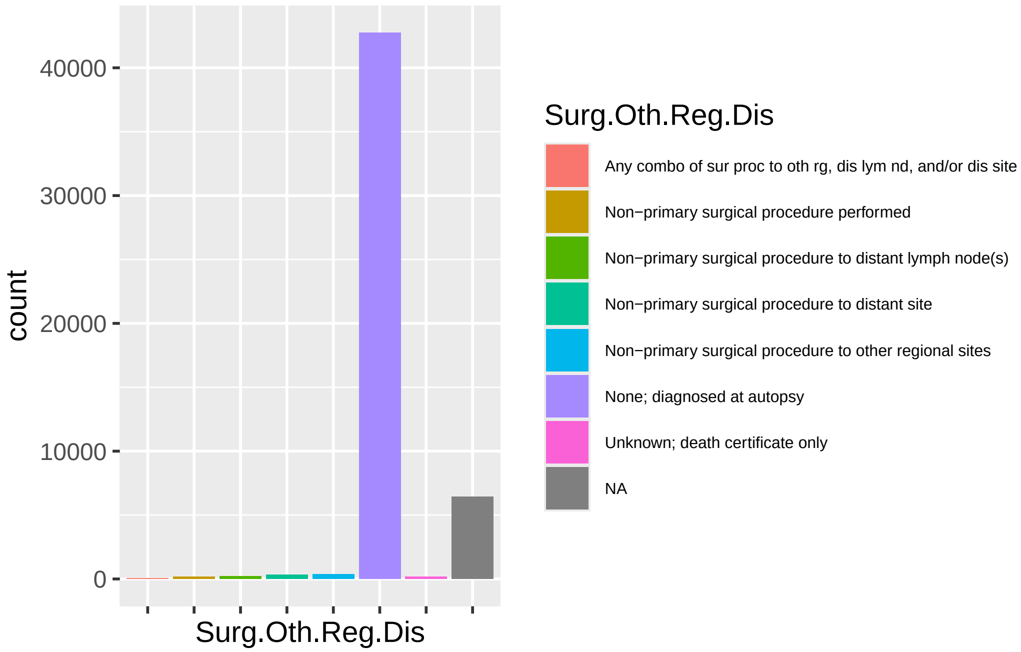
```
                        1 to 3 regional lymph nodes removed
                                                        593
                     4 or more regional lymph nodes removed
                                                       3078
           Biopsy or aspiration of regional lymph node, NOS
                                                        495
                                                       None
                                                      38710
             Number of regional lymph nodes removed unknown
                                                        366
                                 Sentinel lymph node biopsy
                                                         11
   Sentinel node biopsy and lym nd removed different times
                                                         16
Sentinel node biopsy and lym nd removed same/unstated time
                                                         13
                                  Unknown or not applicable
                                                        801
```



This variable contains data relating to the scope of surgery done to regional lymph nodes. Most of the cases are "none", meaning no surgery was done. Besides NA, the next highest number of observations is in the 4+ lymph nodes removed category, followed by 1-3.

From this we can discern that removal of lymph nodes is uncommon, but becomes more necessary the more positive nodes there are.

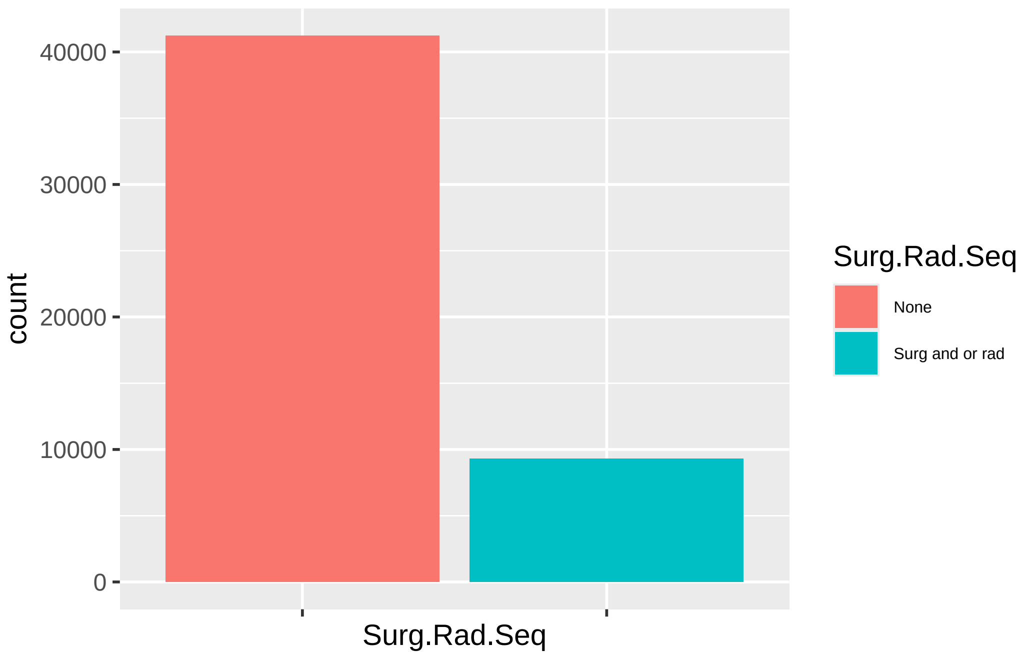
```
Any combo of sur proc to oth rg, dis lym nd, and/or dis site
                                                          63
                    Non-primary surgical procedure performed
                                                         182
     Non-primary surgical procedure to distant lymph node(s)
                                                         206
              Non-primary surgical procedure to distant site
                                                         348
      Non-primary surgical procedure to other regional sites
                                                         377
                                  None; diagnosed at autopsy
                                                       42730
                             Unknown; death certificate only
                                                         177
```



Surg.Oth.Reg.Dis stands for “surgery to other distant regions”, and over ~42,000 out of 50,000 observations are in the “none/diagnosed at autopsy” category.

Another 6000 of the remaining 8000 are NA’s, meaning the variable contains little meaningful data for our investigations.

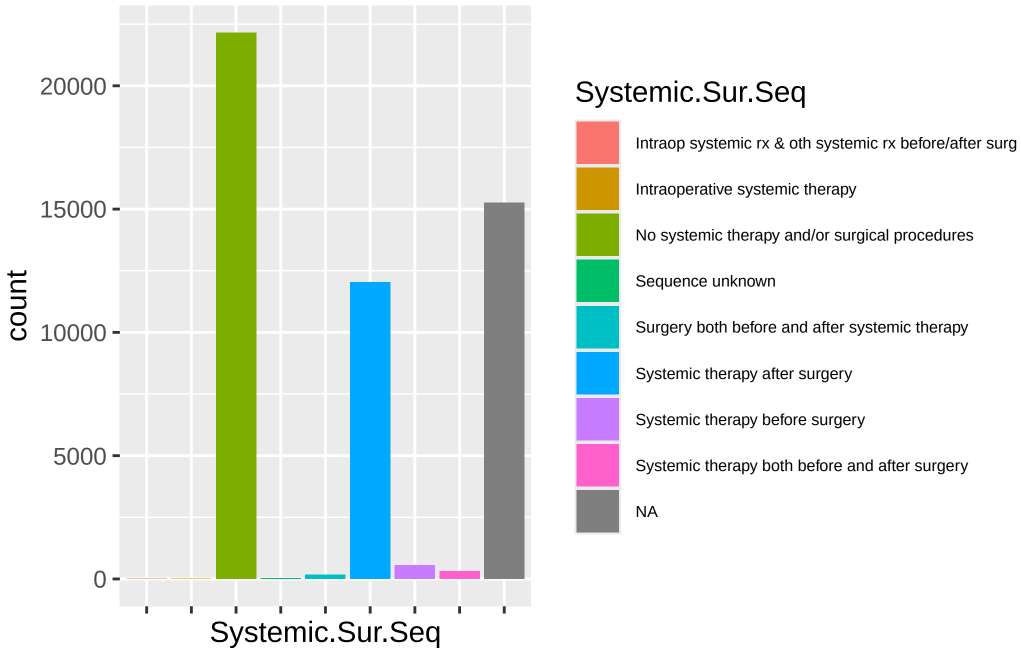
```
 None Surg and or rad 
41212            9310 
```



The categories in this variable originally contained the order of surgery and or radiation. However, ~40,000 were again contained in the "no surgery" category.

Thus, the variable is collapsed into just two levels, one in which no surgery or radiation was used, and another where one or both were used.

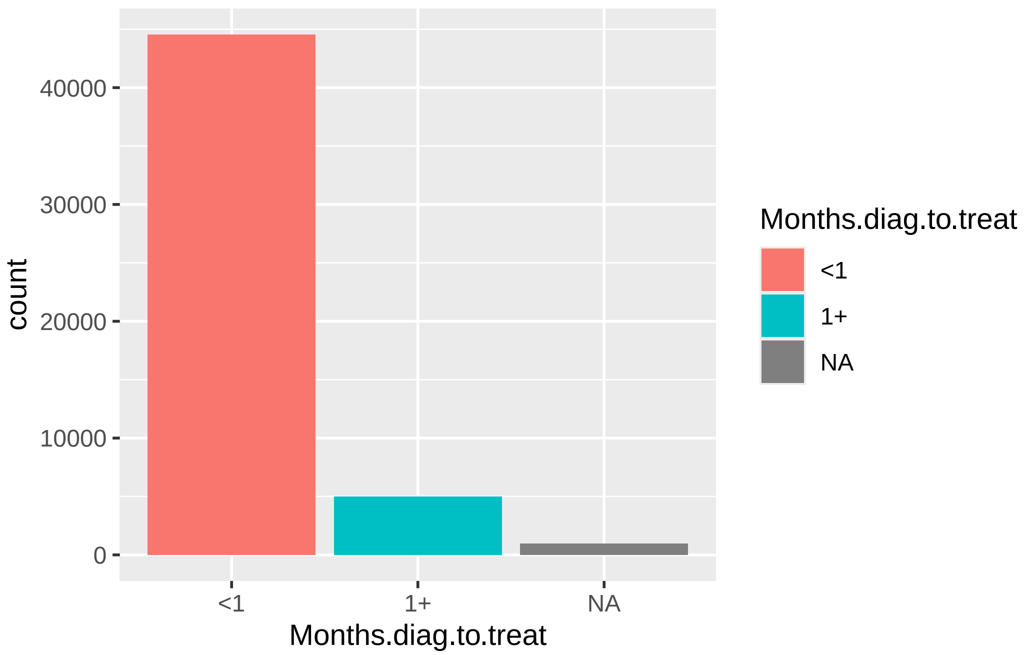
```
Intraop systemic rx & oth systemic rx before/after surg 
                                                      3 
                        Intraoperative systemic therapy 
                                                      9 
         No systemic therapy and/or surgical procedures 
                                                  22145 
                                       Sequence unknown 
                                                     30 
         Surgery both before and after systemic therapy 
                                                    167 
                         Systemic therapy after surgery 
                                                  12038 
                        Systemic therapy before surgery 
                                                    557 
         Systemic therapy both before and after surgery 
                                                    309 
```



Similar to the previous variable, this one contains data involving the sequence of systemic surgery and therapy. Most of the observations are in the none level, another large portion are patients that got systemic therapy after surgery, and another large chunk is NA's.

The presence of two substantial categories makes this a useful variable for our treatment related data science question.

```
   <1    1+
44540  5003
```



Just like the Surg.Rad.Seq variable, almost all of the patients received treatment within 1 month of their diagnosis.

Thus, the only way for the variable to be of any use is to make it binary, where one category is treatment within a month, and the other is one month or more.

Chase

```
unique(df$CS_tumor_size)
```

```
  [1]  NA  25  15  20  70 988  65  90   9  27  66  45  39 999  75  30 160  60
 [19]  35  42  18  80  12  92  23  50  55  28  43  85  26  10  40  52 100   4
 [37]  67  54   0  21  22  16  11   8  34  13  47  38 110  37  68  48  36 104
 [55]  56  24  32  17  19   7  62  58  57   6  63  73  78 989   2 130  72  44
 [73] 120  46  53 170  94  31  49  95   5 118 128 145  61  33  29 115  76  41
 [91]  81 150  59 109  89  84  83  69  77 135 140 670  14 520  64  51 105 125
[109] 102 195 123   3 161  71 280  74  82 112  88  87 993  93  98 992 994 180
[127]  96   1  97 124 226 650  86 250 420 152 205 320  79 114 108 990 103 260
[145] 158 107 210 119 239  91 132  99 270 146 111 920 129 122 450 121 200 550
[163] 888 155 113 117 700 137 101 151 220 165 116 133 190 127 400 162 181 141
[181] 800 185 188 215 390 189 950 230 470 177 995 580 620 138 350 142 157 991
[199] 106 255 126 271 139 263 134 225 172 148 202 174 690 411 156 168 164 175
```

```
[217] 201 131 780 600 720 187 154 204 193 300 560
```

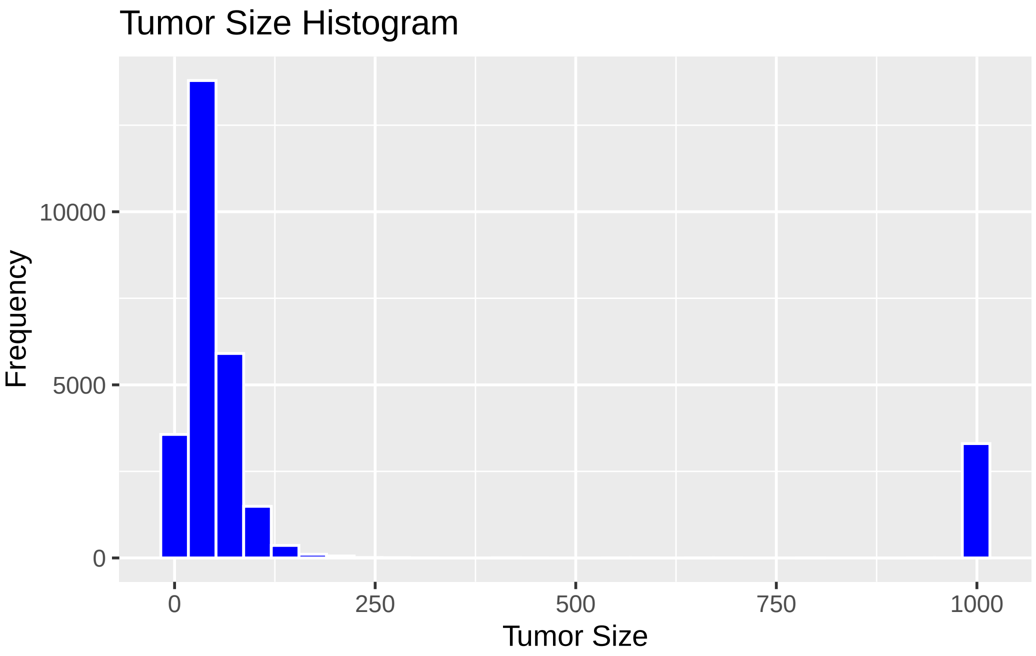
```
summary(df$CS_tumor_size, na.rm = TRUE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    0.0    25.0    43.0   154.9    70.0   999.0   21888 
```

```
sd(df$CS_tumor_size, na.rm = TRUE)
```

```
[1] 305.4416
```

```
ggplot(data = df, aes(x = CS_tumor_size))+
  geom_histogram(fill="blue",color="white")+
  labs(x='Tumor Size', y='Frequency', title = 'Tumor Size Histogram')
```



Frankie

```r
#Numerical variable of nodes examined
summary(df$nodes_examined_num)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   1.00    6.00   17.00   21.07   30.00   90.00   46157 
```

```r
#Categorical varailable for nodes examined
table(df$nodes_examined_cat)
```

```

      Aspiration performed Dissection, number unknown 
                       579                        326 
              Exact number          No nodes examined 
                      4365                      42032 
   Removed, number unknown   Sampling, number unknown 
                       508                         17 
                   Unknown 
                      2695 
```

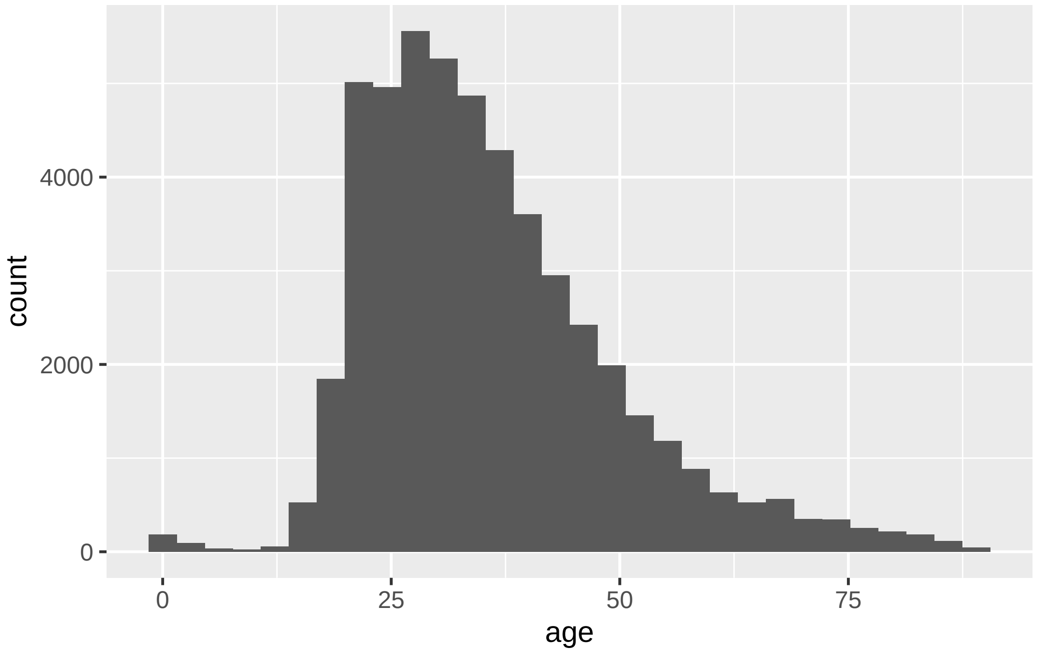
```r
#positive nodes numerical variable
summary(df$positive_nodes_num)
```

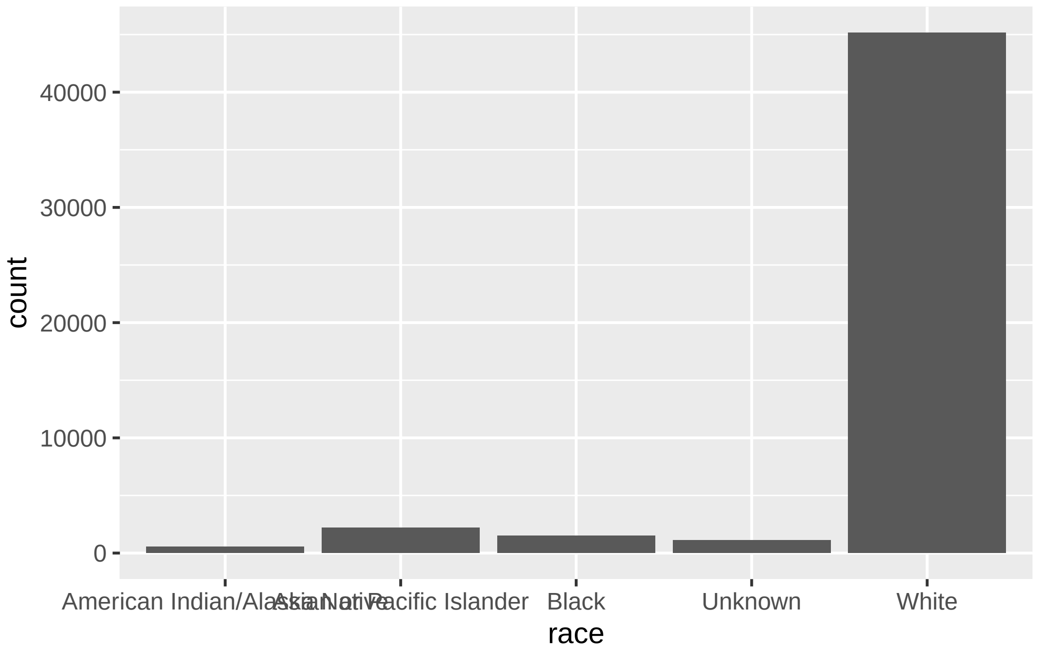
```
Length  Class   Mode 
     0   NULL   NULL 
```

```r
#categorical nodes numerical variable
table(df$positive_nodes_cat)
```

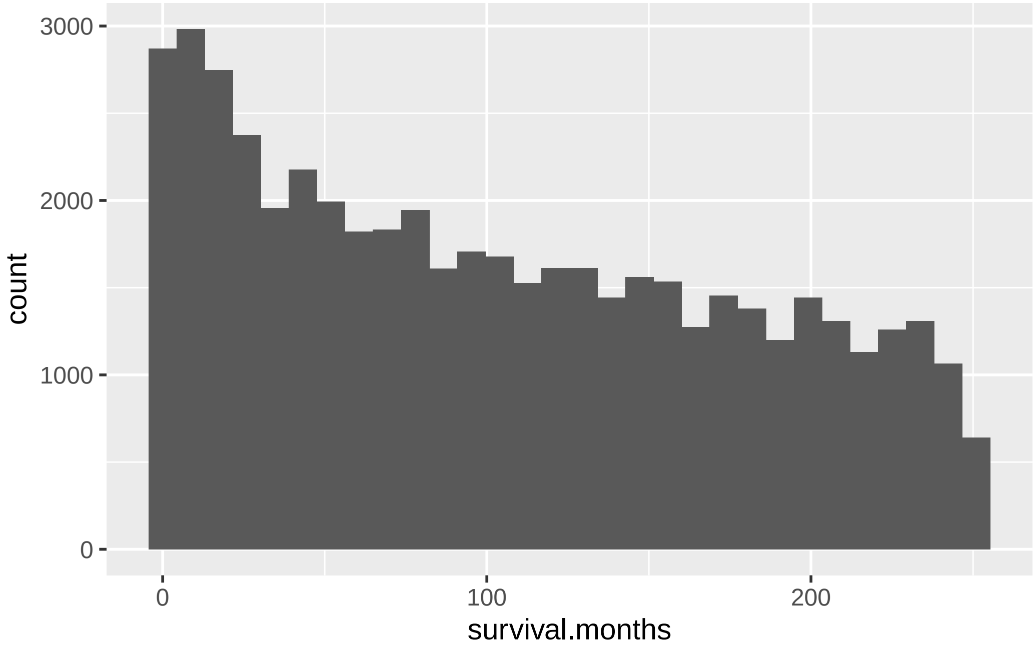
```
< table of extent 0 >
```

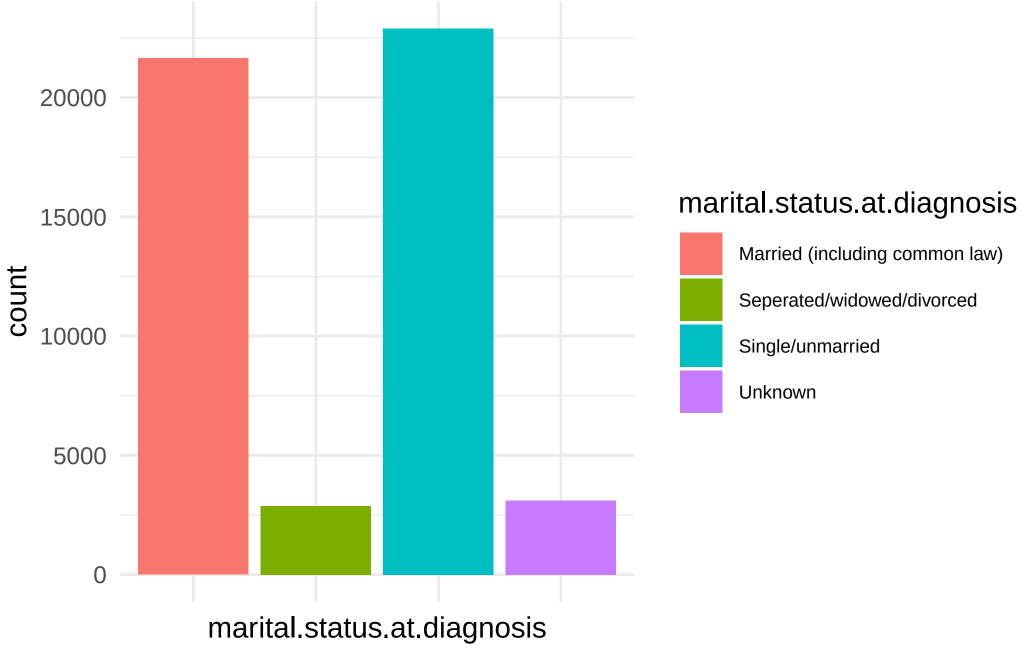
Saul



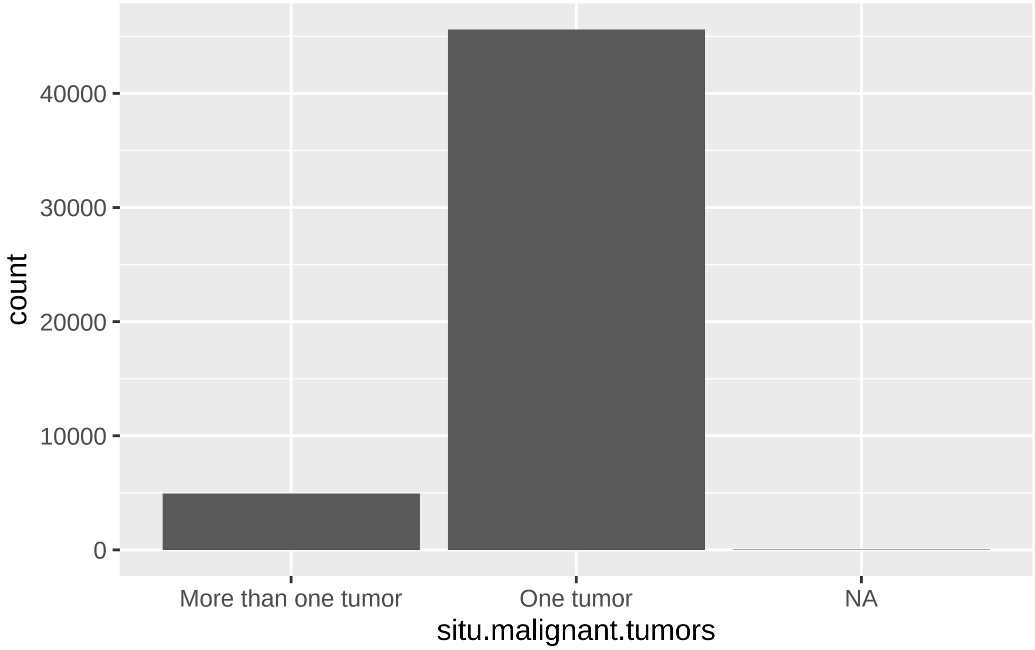


The vast majority of individuals with a tumor in their testis are white. We checked the overall demographics of the seer data and found it was primarily white. Because we do not have access to the counties in which the individuals reported from, it is difficult to gauge whether this is an issue or not.

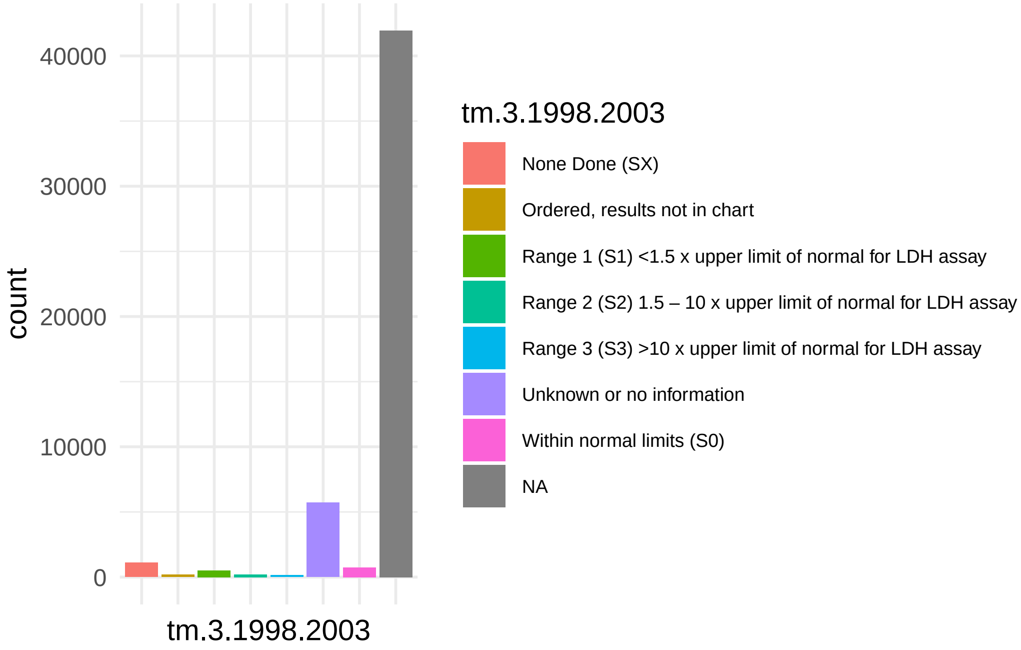




Most patients were either married, or single (never married)



Changed to binary (One tumor or More than one tumor)



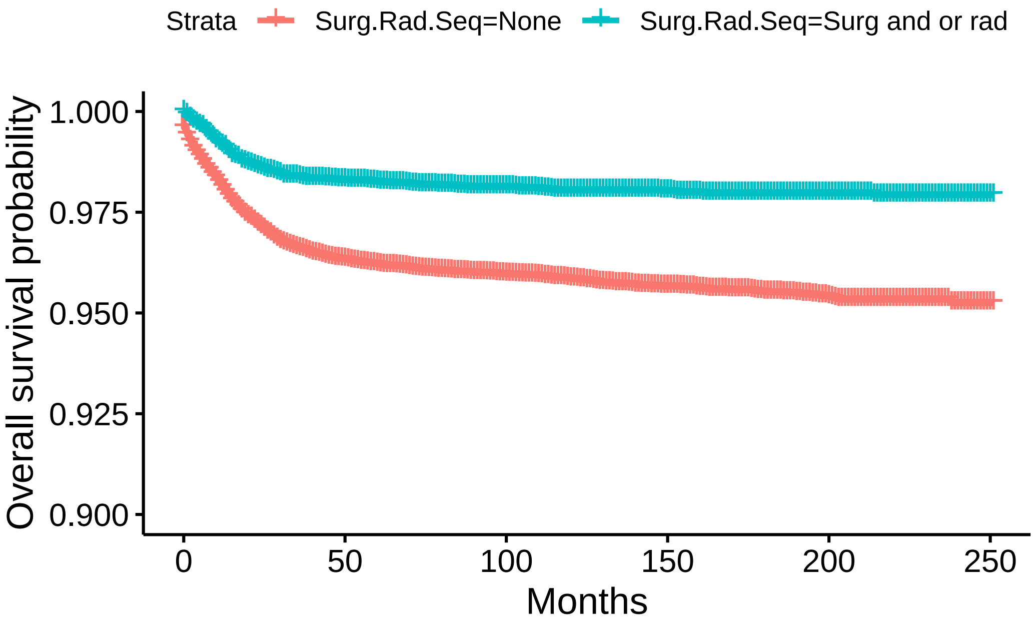
Tons of ‘NA’ values in tumor marker variables, probably in part do them not spanning all years.

Modeling

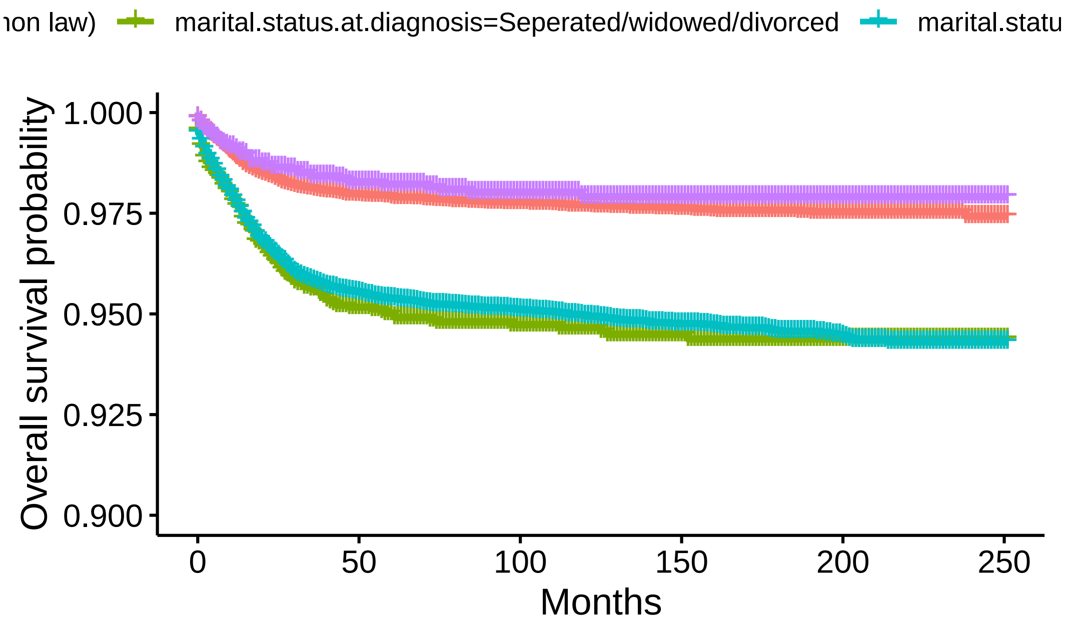
Survival Analysis

```
df$testi_death <- ifelse(df$death.site == "Testis", 1, 0)
df$SurvObj <- with(df, Surv(survival.months, testi_death == 1))
```

```
km <- survfit(SurvObj ~ 1, data = df, conf.type = "log-log")

km.trt <- survfit(SurvObj ~ Surg.Rad.Seq, data = df, conf.type = "log-log")

ggs <- ggsurvplot(
  fit = survfit(Surv(survival.months, testi_death) ~ Surg.Rad.Seq, data = df),
  xlab = "Months",
  ylab = "Overall survival probability")

ggs$plot <- ggs$plot + ylim(c(0.9, 1))

ggs
```



```
ggs <- ggsurvplot(
  fit = survfit(Surv(survival.months, testi_death) ~ marital.status.at.diagnosis, data = d
  xlab = "Months",
  ylab = "Overall survival probability")

ggs$plot <- ggs$plot + ylim(c(0.9, 1))

ggs
```



```
mod <- coxph(Surv(survival.months, testi_death) ~ race + Surg.Rad.Seq + Months.diag.to.tre
# %>% gtsummary::tbl_regression(exp = TRUE)
```

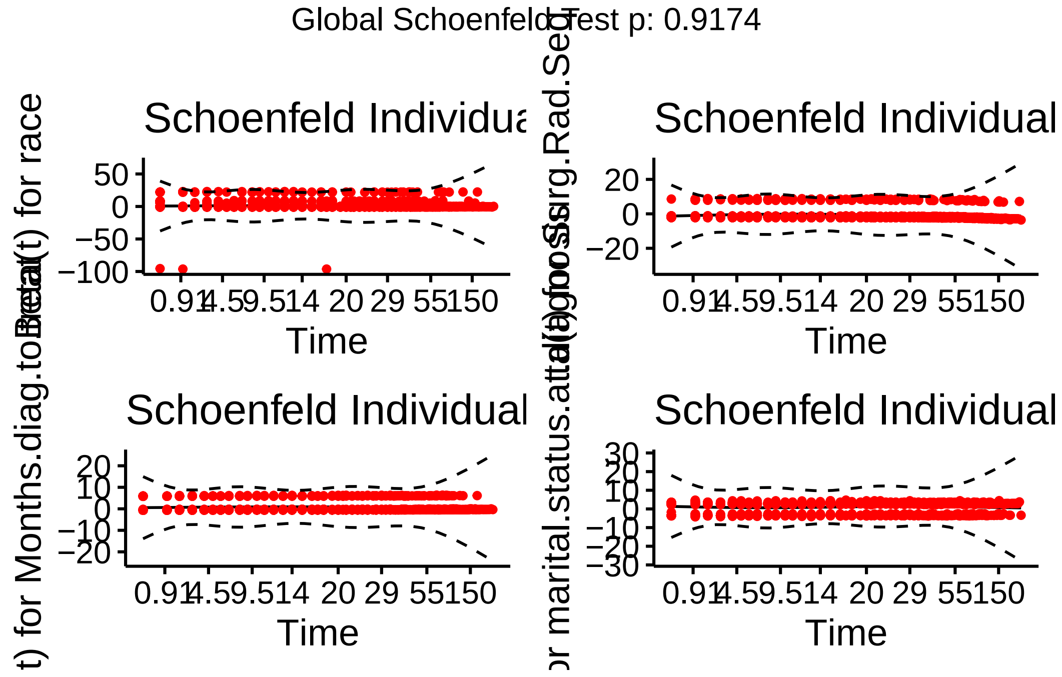
The p-values for all three overall tests (Likelihood, Wald, and Score) are less than .01, indicating the model is significant and rejecting the null hypothesis that all betas are 0.

Checking Cox Proportional Hazards assumptions

```
test.ph <- cox.zph(mod)
```

From the output above, the test is not statistically significant for the global test as well as each of the covariates. Therefore, we can assume the proportional hazards.

```
ggcoxzph(test.ph)
```



No pattern with time (supports the use of proportional hazards)

Interpretations

```r
gtsummary::tbl_regression(mod, exp = TRUE)
```

Characteristic	**HR**	**95% CI**	**p-value**
race			
American Indian/Alaska Native	—	—	
Asian or Pacific Islander	1.07	0.66, 1.75	0.8
Black	1.45	0.89, 2.37	0.14
Unknown	0.11	0.03, 0.37	<0.001
White	0.88	0.57, 1.37	0.6
Surg.Rad.Seq			
None	—	—	
Surg and or rad	0.53	0.45, 0.62	<0.001
Months.diag.to.treat			
<1	—	—	
1+	2.15	1.90, 2.44	<0.001
marital.status.at.diagnosis			
Married (including common law)	—	—	
Seperated/widowed/divorced	2.27	1.86, 2.77	<0.001
Single/unmarried	2.08	1.86, 2.34	<0.001
Unknown	0.98	0.73, 1.32	0.9

Based on the above regression table, we can interpret the following information from our model:

- *Marital status:*
  - Patients that were married at the time of diagnosis had half (0.4, 0.64) the risk of death due to testicular cancer compared to patients that were divorced at the time of diagnosis. (p < 0.001)
  - Patients separated from their partner at diagnosis had higher risk of death to testicular cancer than divorced patients by a factor of 1.56 (1.01, 2.39) (p = 0.044).
  - These observations point towards a trend where patients that are more lonely, or have less emotional support are at more risk of death to their diagnosed testicular cancer.
- *Months from diagnosis to treatment:* The hazard for death was higher for patients not receiving treatment for at least a month after being diagnosed. The rsik of death was higher by a factor of 2.16 (95% CI: 1.90 - 2.45, p= 0.001). The control for this variable were patients treated within a month of the diagnosis.

- *Race:* The coefficients of Black and Pacific Islander or Asian are both positive which indicates a higher hazard ratio, but both groups have p-values of $0.140318 > .05$ and $0.806804 > .05$ respectively signaling they are not statistically significant. The unknown race has a negative coefficient which indicates a lower hazard ratio and is statistically significant with a p-value $= 0.000360 < .05$. The coefficient of White is negative, which suggests a lower hazard ratio, but is not statistically significant p-value $= 0.575 > .05$.
- *Surgery and radiation:* Relative to patients not receiving surgery or radiation, the patiets who did recceive surgery or radiation had a lower risk of death by a factor of 0.53 (95% CI: 0.45 - 0.62, p= 0.001).