Case Study 2 Report 1

Altamash Rafiq, Annie Tang, Sonya Zhang

October 4th, 2017

Contribution Notes

This paper was a group effort with all team members meeting outside of class to discuss and work together. Altamash contributed to creating the survival plot analysis. Annie contributed to the histogram exploratory data analysis. Sonya contributed to the overall write up and description as well as the data distribution analysis.

Introduction

The goal of this case study was to be able to make inferences on the impact of clinical presentation, gender, and race on time to neurological assessment for patients with stroke-like symptoms who are admitted to the emergency room. In this case clincial presentation was the number of major stroke symptoms. The symptoms considered in this study were headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems. Race was recorded as black, hispanic, or neither, and gender was marked as female or male. The reason the time to neurological assessment is of interest is because many stroke treatments are most effective if administered quickly after the stroke symptoms begin to occur (within 3 hours). Because time to assessment and treatment are critical to long-term patient prognosis, it is helpful to understand how variables such as race, gender, and number of symptoms impacts this.

Data

The dataset used in this analysis gave the time to critical neurological assessment following admission to the emergency room. This sample included 335 patients with mild to moderate motor impairment. No missing data was discovered so the data was able to be used as is. The data consisted of nctdel as the response variable. This was the time (in hours) to either neurological assessment and CT scan or the time to which the patient dropped out of the study. The data included an indicator variable, fail, which indicated a 1 if the patient received neurological assessment and a CT scan and indicated a 0 otherwise. The exact reason for each patient's departure from the study is unknown, but a few of the reasons may include leaving the emergency room, refusing treatment, or death. The data also included male, black, and hispanic as indicator variables with 1 indicating that the patient was recorded as part of the racial or gender group. Four other indicator variables were present in the dataset - sn1, sn2, sn3, and all4. These variables represented the patient presenting one, two, three, or four of the total symptoms respectively. Therefore, if a patient received a 0 indicator for all four of these variables, it would indicate the patient had none of the four stroke symptoms of interest.

Data Distribution

Some analysis was performed on the data to determine the distribution of the data entries among different category groupings.

```r
data <- read.table("kellydat.txt", header = TRUE)
suppressMessages(library(survminer))
suppressMessages(library(survival))
suppressMessages(library(plyr))
count(data,c("sn1","sn2","sn3","all4"))
```

```
##   sn1 sn2 sn3 all4 freq
## 1   0   0   0    0   69
## 2   0   0   0    1    6
## 3   0   0   1    0   21
## 4   0   1   0    0   77
## 5   1   0   0    0  162
```

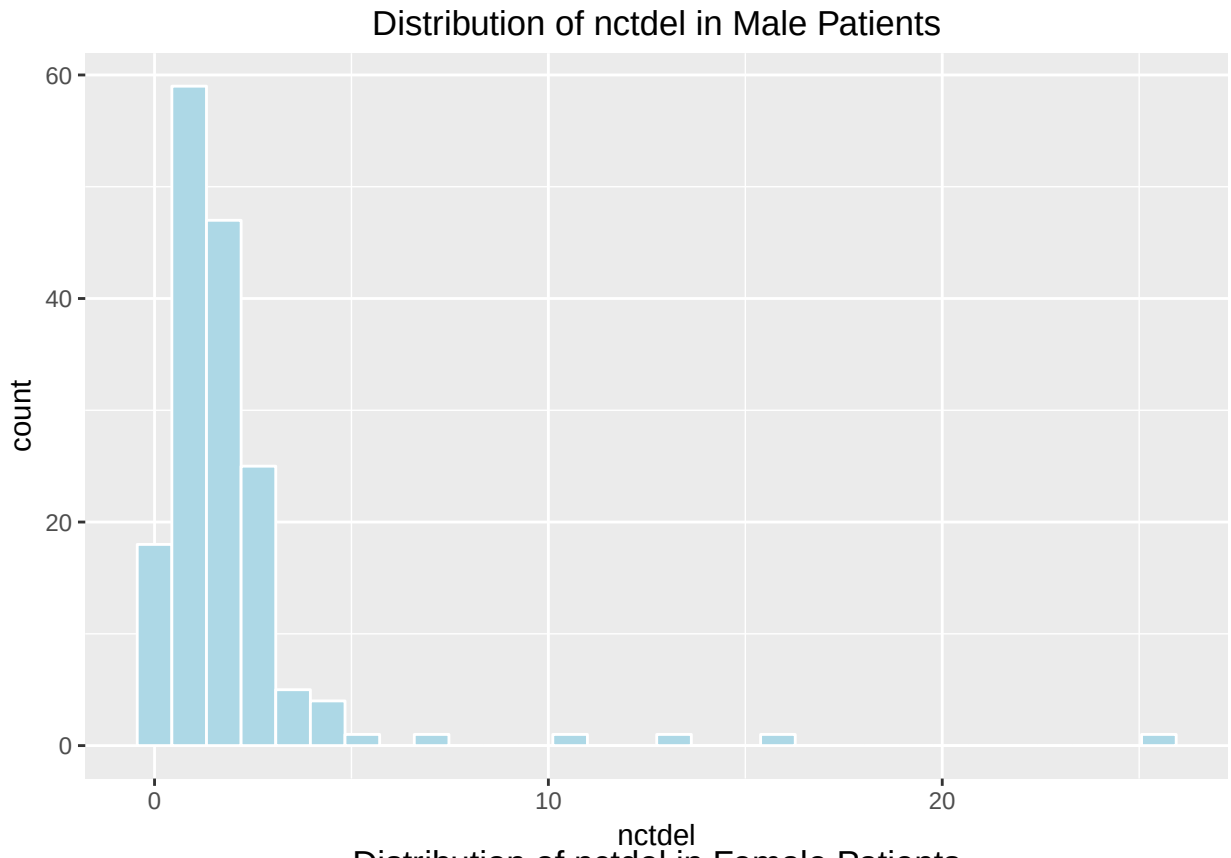
This chart indicates that most patients only exhibited one symptom, and very few patients exhibited all four symptoms. This is important to keep in mind when later analyzing the data. It may be hard to draw conclusions or make inferences about patients who exhibit all 4 symptoms since the sample of patients with all 4 symptoms in this dataset is fairly small.

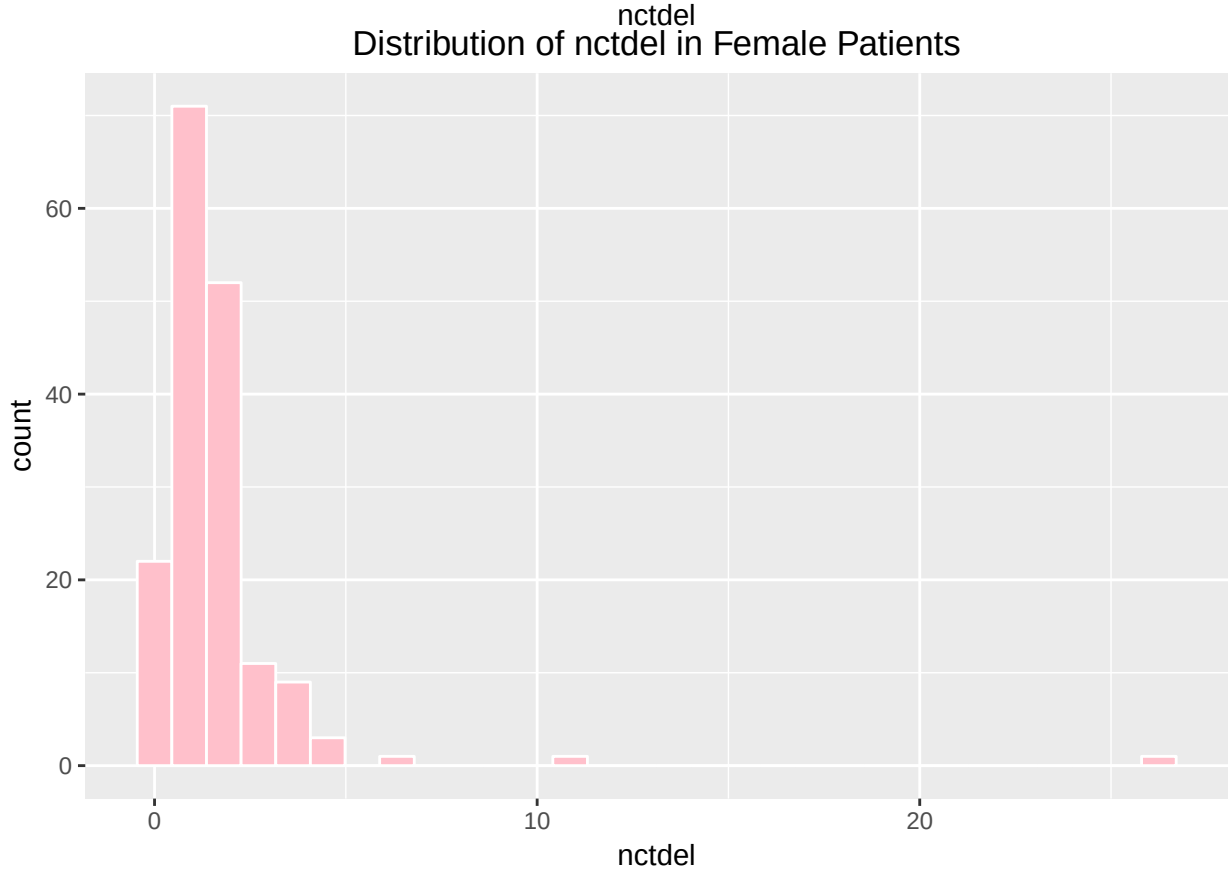
```r
count(data, c("black", "hisp", "male"))
```

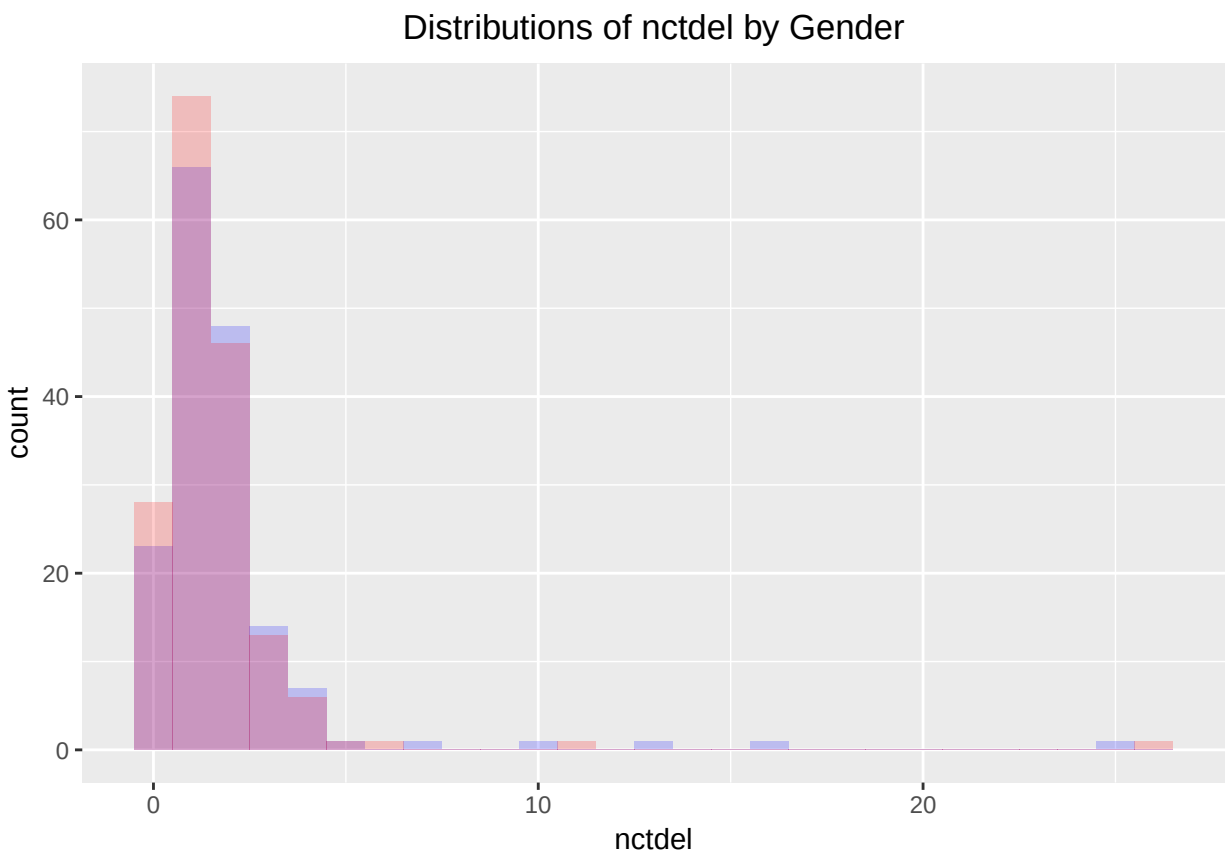
```
##   black hisp male freq
## 1     0    0    0  124
## 2     0    0    1  114
## 3     0    1    0    4
## 4     0    1    1    5
## 5     1    0    0   43
## 6     1    0    1   45
```

This chart indicates the distribution of patients in this sample by race and gender. In this sample, within each race and in the overall dataset, gender is split fairly evenly between male and female. However, the sample is majority non-hispanic and non-black, and there is very few datapoints of hispanic patients. Again, it may be hard to draw conclusions about how being hispanic affects time to critical neurological assessment since the sample size is so small.

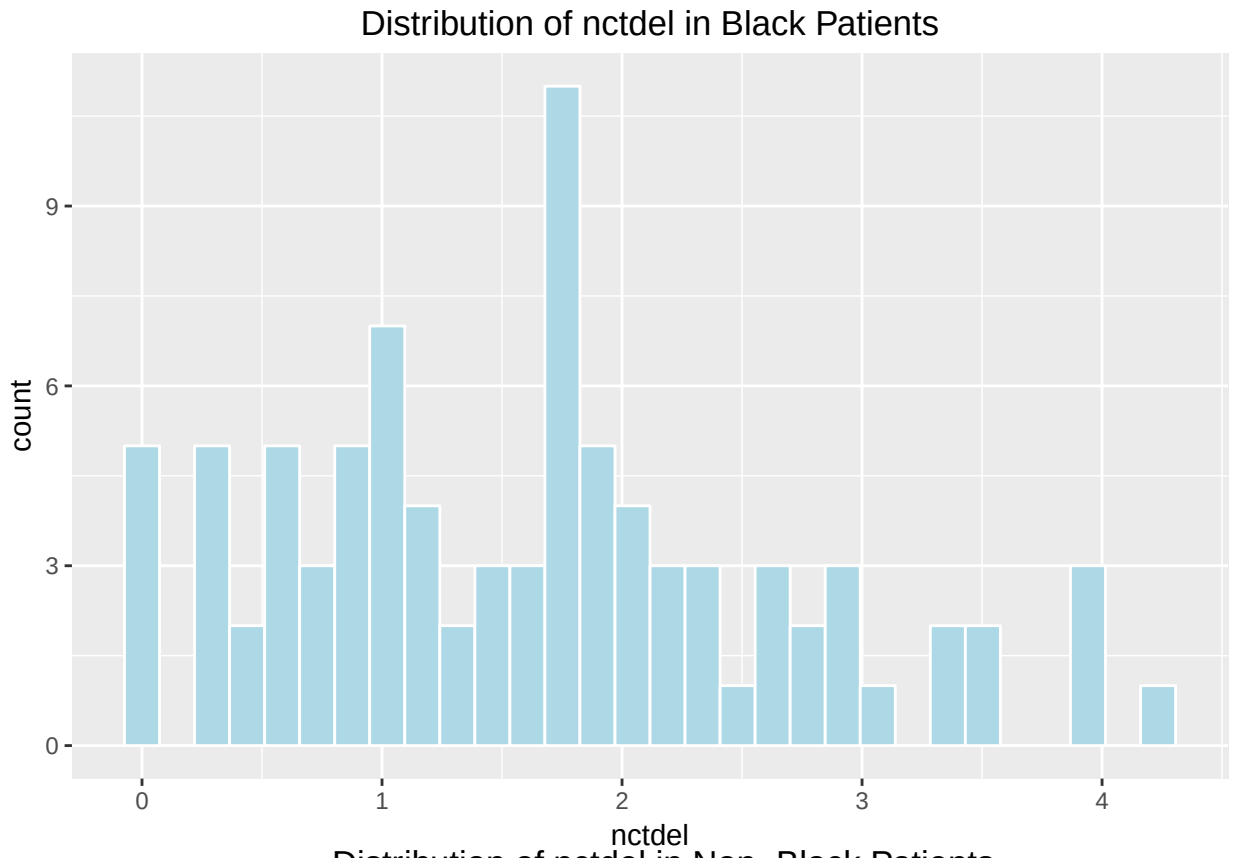
The distribution of time to neurological assessment within race and gender groups was also of interest. To get a general idea of the data distribution, histograms were created.

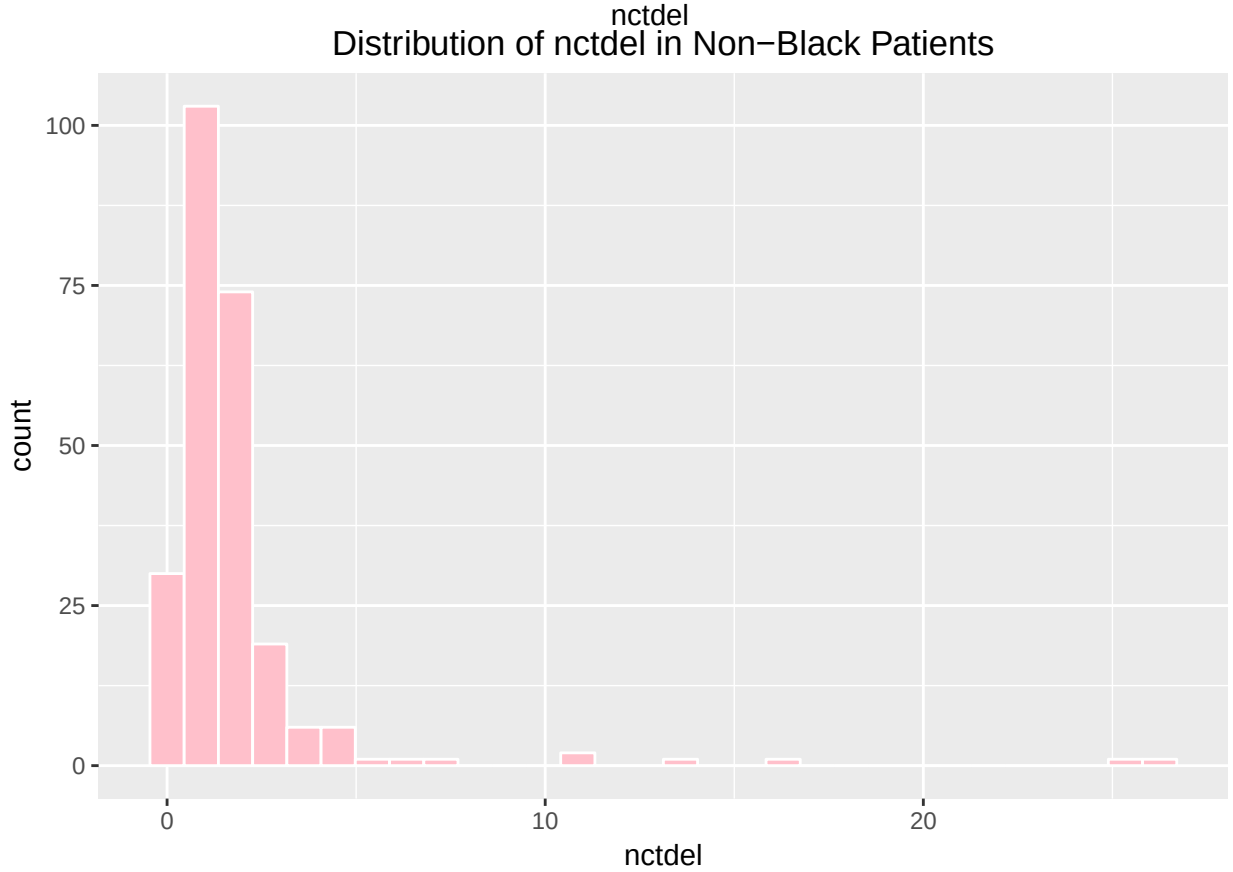

Distribution of nctdel in Male Patients
60
40
20
0
count
0
10
20
nctdel

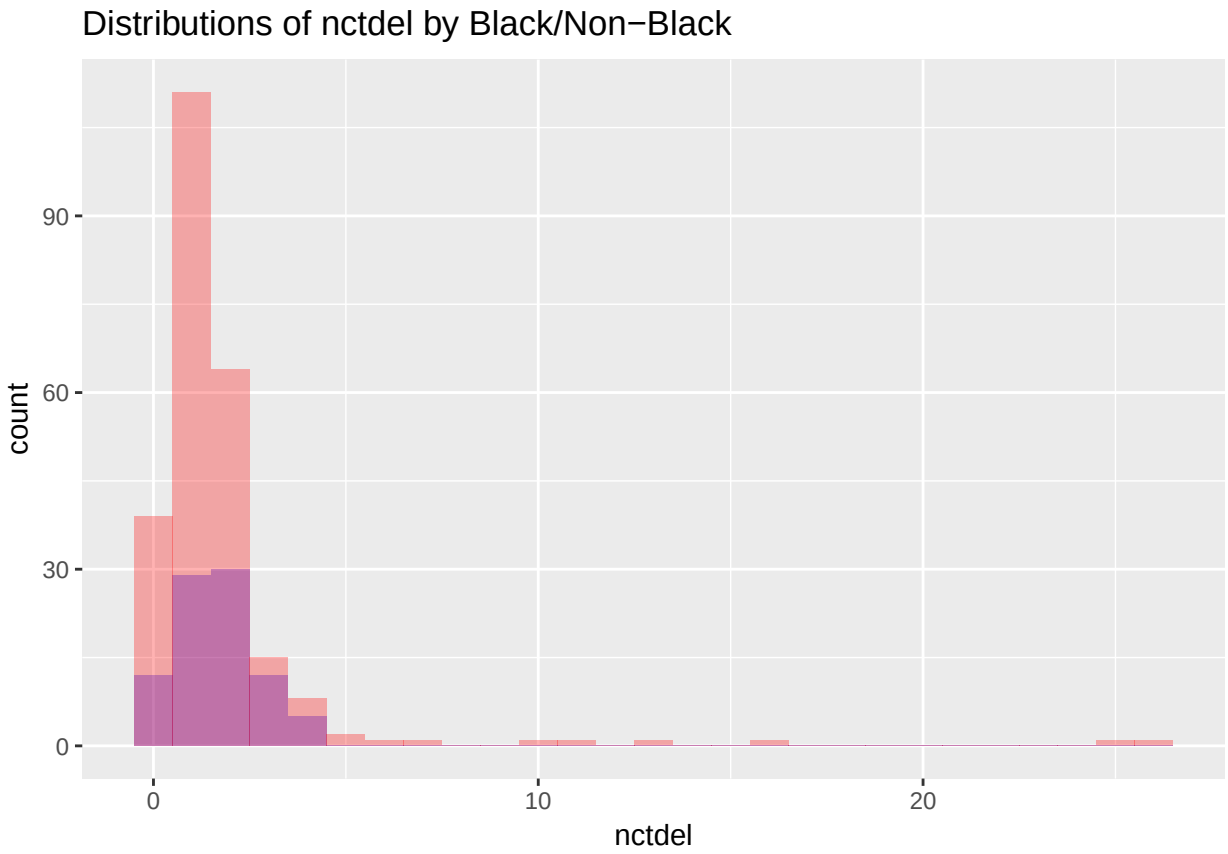

Distribution of nctdel in Female Patients
60
40
20
0
count
0
10
20
nctdel



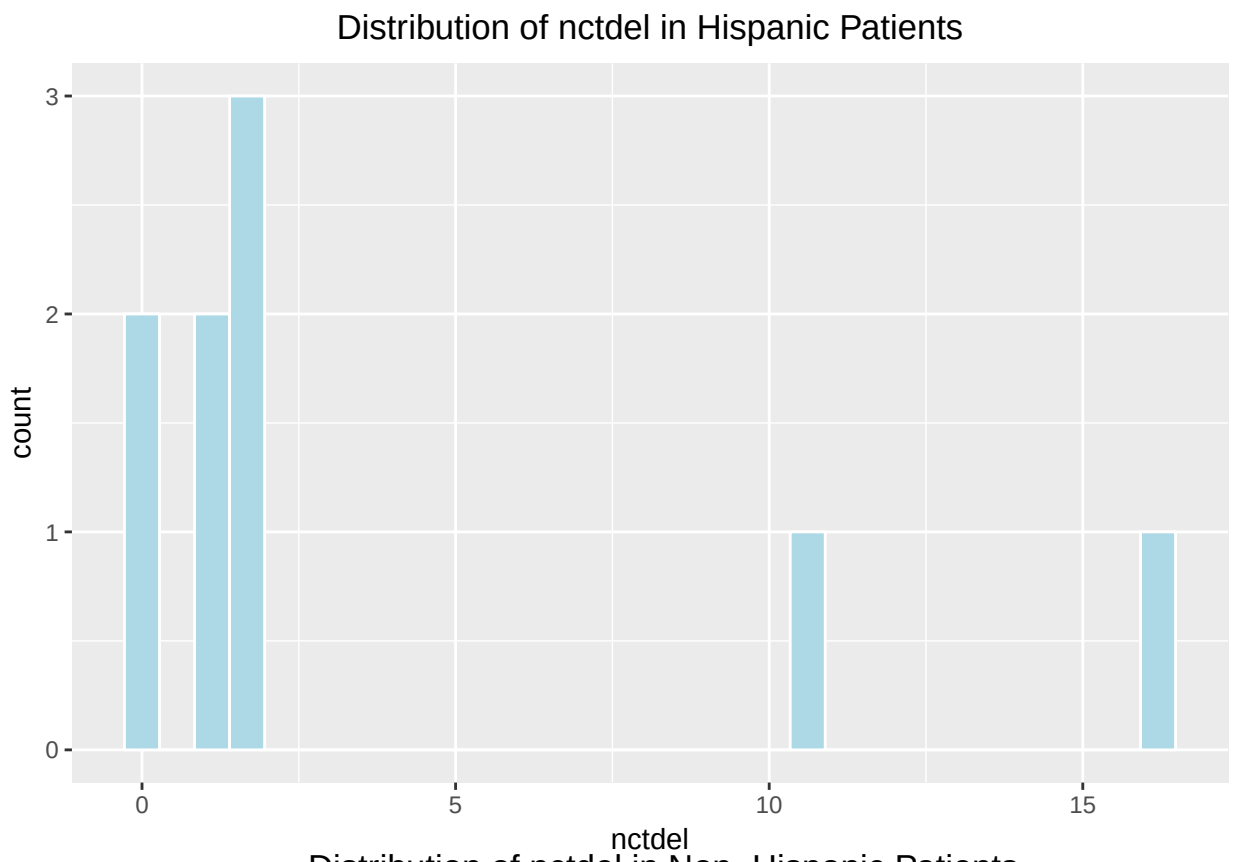
These plots show that the distributions of nctdel by gender (male/female) are about the same in shape and scale.

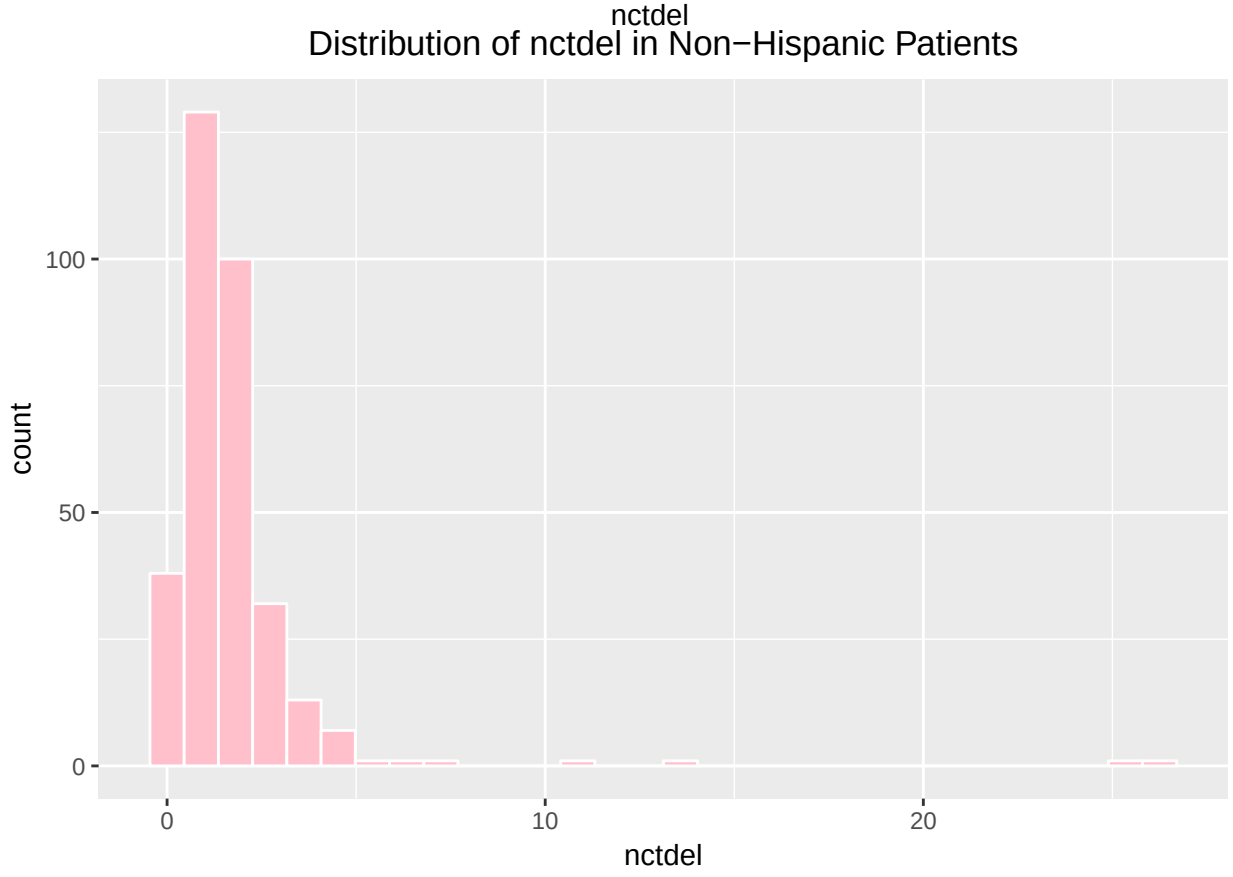

Distribution of nctdel in Black Patients
count
9
6
3
0
0
1
2
3
4
nctdel

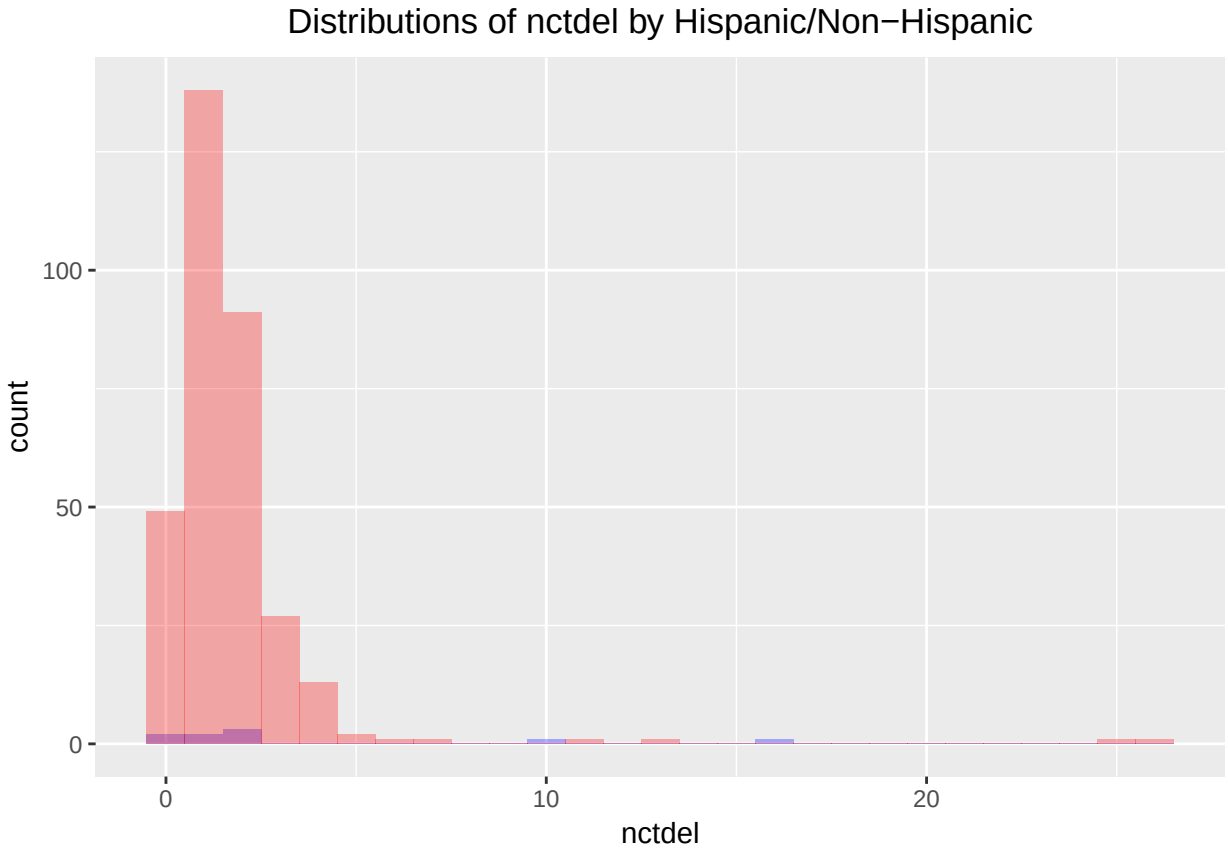

Distribution of nctdel in Non−Black Patients
count
100
75
50
25
0
0
10
20
nctdel



In these plots, it appears that the distribution of nctdel in black patients is less skewed and has a shorter tail than that of nctdel in non-black patients, but it is difficult to come to any concrete conclusions because there is much more data for non-black patients than black patients.


Distribution of nctdel in Hispanic Patients
count
nctdel
0
1
2
3
0
5
10
15


Distribution of nctdel in Non−Hispanic Patients
count
nctdel
0
50
100
0
10
20

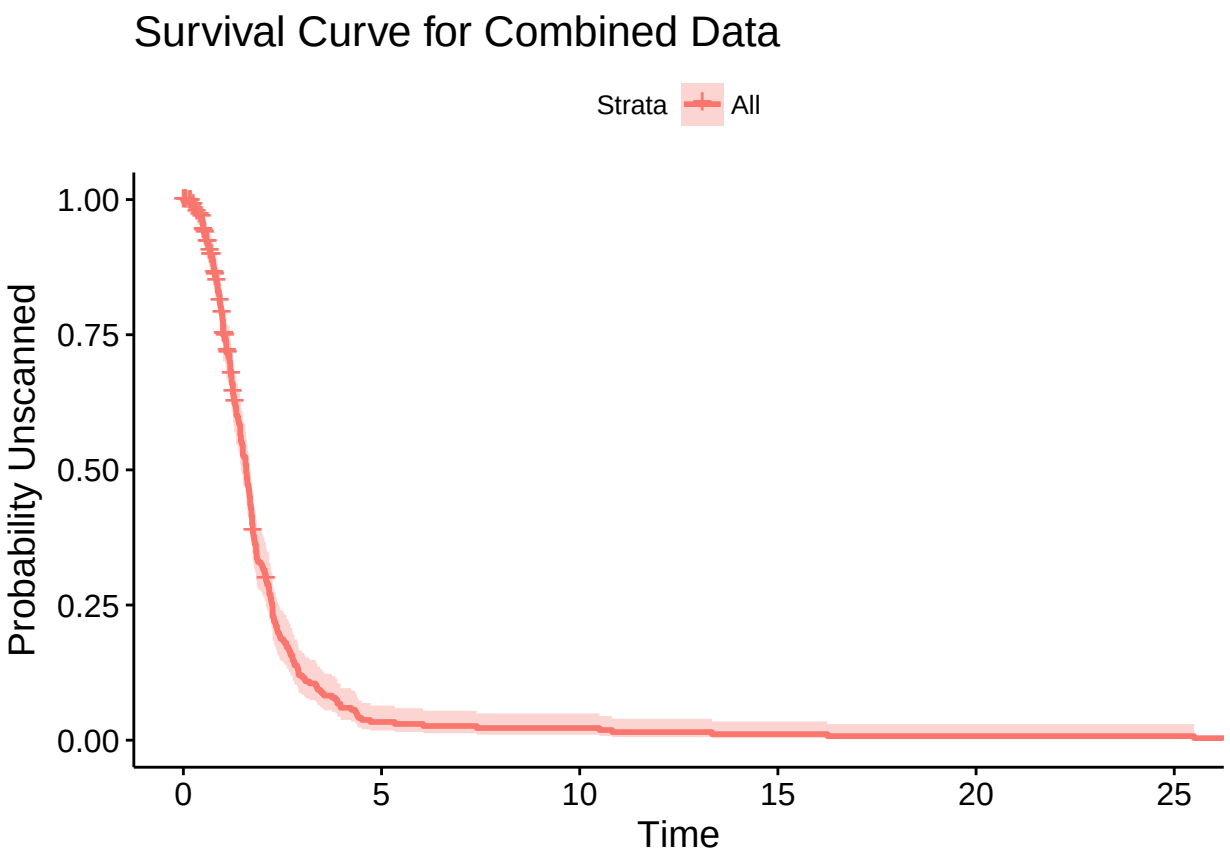


Again, there is very little data for hispanic patients, so the distribution shown here for ncntdel in hispanic patients is not representative.

Survival Curve Exploratory Data Analysis

Kaplan-Meier estimation was utilized to create survival curves of the data. Ggsurvplot in the survminer R package was used to perform this. The y-axis on the resulting plots indicates the probability that a given patient has not been scanned yet. This is plotted against time (nctdel) on the x-axis such that at any time x the probability that someone has not been scanned yet is y. Therefore, to determine the probability that a patient has been scanned, a calculation of 1-y must be performed.

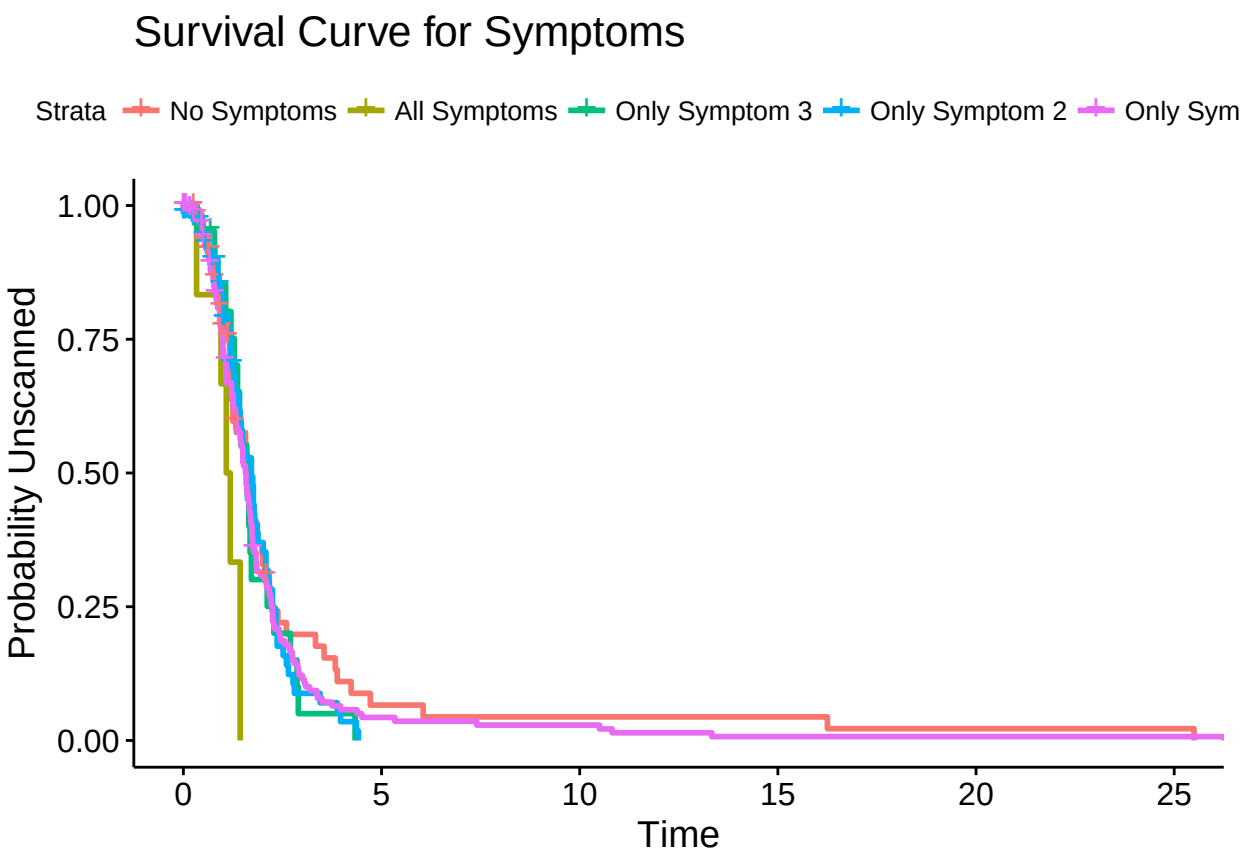
```
fitM <- survfit(Surv(nctdel, fail) ~ 1, data = data)
ggsurvplot(fitM, data = data, risk.table = FALSE) + labs(y = "Probability Unscanned") +
  ggtitle("Survival Curve for Combined Data")
```



First a survival curve for the entire dataset was plotted without subsetting. This plot and its summary suggest that based on this sample, the probability of getting scanned in under three hours is less than 0.88385.

Survival curves based on clinical presentation were also created.

```
fit <- survfit(Surv(nctdel, fail) ~ sn1 + sn2 + sn3 + all4, data = data)
ggsurvplot(fit, data = data, risk.table = FALSE, legend.labs=c("No Symptoms", "All Symptoms", "Only Symp
  ggtitle("Survival Curve for Symptoms")
```

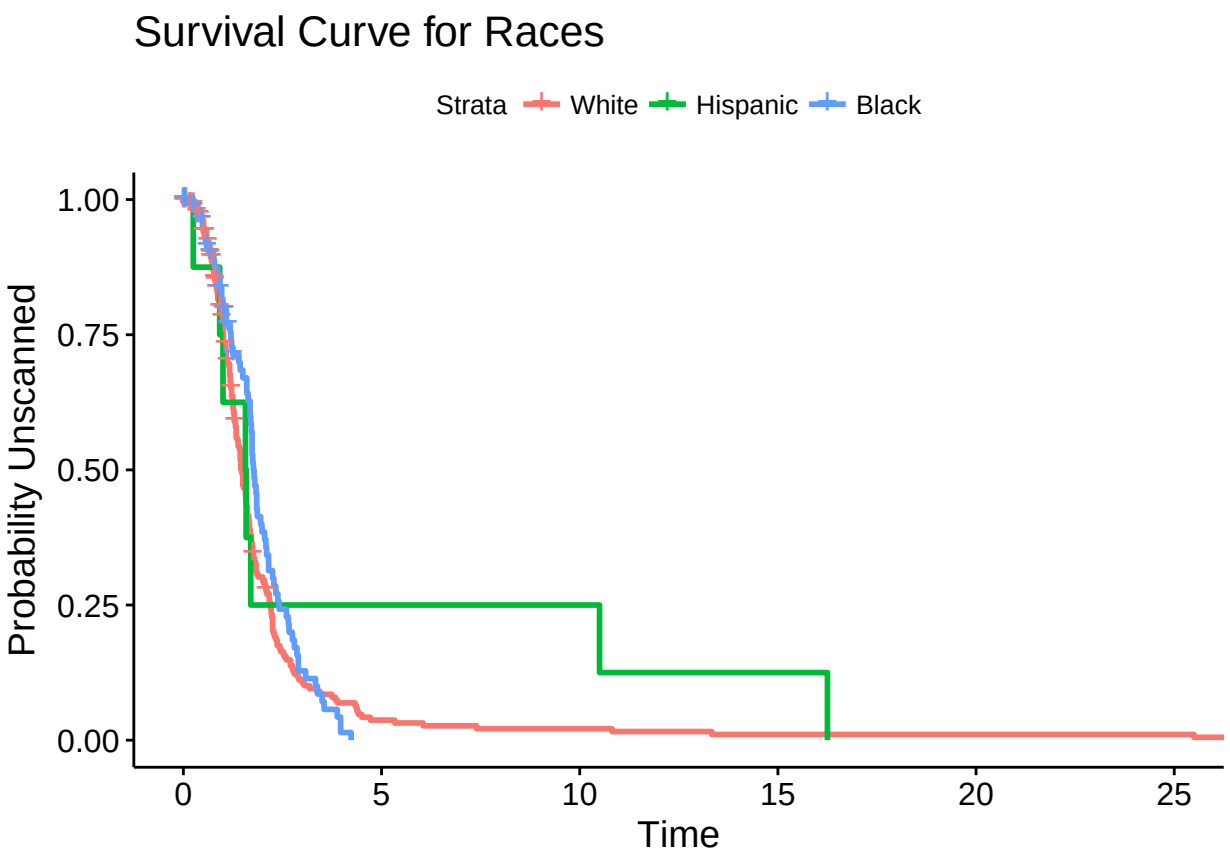


This plot and its summary suggest that based on this sample:

1) If a patient has none of the three symptoms, their probability of getting scanned in under three hours is 0.8237.
2) If a patient has 4 symptoms, their probability of getting scanned in under three hours is 1.
3) If a patient has only three symptoms, their probability of getting scanned in under three hours is between 0.9499 and 1.
4) If a patient has only two symptoms, their probability of getting scanned in under three hours is between 0.9119 and 0.9296.
5) If a patient has only one symptom, their probability of getting scanned in under three hours is 0.88517.

Based on this prelimary analysis, the results are similar to what one may intuitively expect. The more symptoms a patient possessed, the more likely they were to receive neurological assessment in under three hours. Three hours was used as benchmark because it is the time period in which thrombolytic therapy is most effective in treating ischemic stroke patients. However, in further analysis three hours will not be the only time period of interest as there are other treatment options and other benefits to understanding the overall time to neurological assessment.

```
fit1 <- survfit(Surv(nctdel, fail) ~ black + hisp, data = data)
ggsurvplot(fit1, data = data, risk.table = FALSE, legend.labs=c("White", "Hispanic", "Black")) + labs(y
  ggtitle("Survival Curve for Races")
```



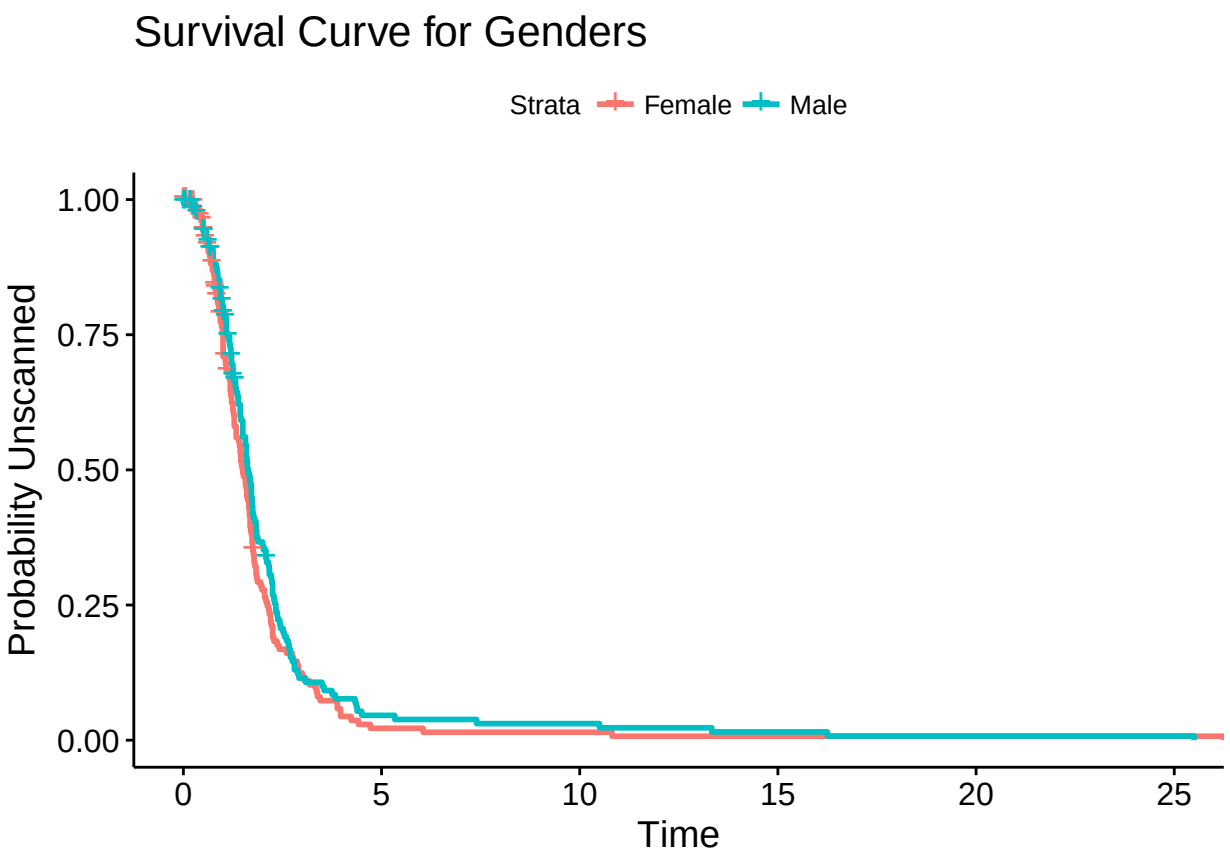
This plot and its summary suggest that in this sample:

1) If a patient is White, their probability of getting scanned in under three hours is 0.89389.
2) If a patient is Black, their probability of getting scanned in under three hours is between 0.8717 and 0.8859.
3) If a patient is Hispanic, their probability of getting scanned in under three hours is between 0.75 and 0.875. However, because there are only six hispanics in the sample, this estimate is not reliable and it is not possible to draw conclusions from this exploratory data analysis about the effect of being hispanic on time to neurological assessment.

As is visible from the graph, at almost any given time, the probability of a white person having received assessment is conspicuously higher than the probability of a black person having received assessment. The probability of getting scanned before the three hour mark for white people in the sample is also higher than that of black people.

Survival curves were also plotted for males and females.

```
fit2 <- survfit(Surv(nctdel, fail) ~ male, data = data)
ggsurvplot(fit2, data = data, risk.table = FALSE, legend.labs=c("Female", "Male")) + labs(y = "Probabil
  ggtitle("Survival Curve for Genders")
```



This plot and its summary suggest that in this sample:

1) If a patient is Male, their probability of getting scanned in under three hours is between 0.88518 and 0.89284.
2) If a patient is Female, their probability of getting scanned in under three hours is 0.88292.

Based on the graph, at most times it appears as if males and females have roughly equal probability of having been scanned. However, in additional analysis it may be of interest to subset into race and gender at the same time to see if this makes a difference in gender in time to neurological assessment.