Survival Analysis Project: Preliminary EDA

Madison Hobbs & Lathan Liou

4/2/2019

Introduction

Our project seeks to understand time of survival until an AIDS defining event or death. In the first phase of our project, we want to understand the distributions of our time-to-event variables and the remaining explanatory variables. Further, we are curious to see whether there exists correlated relationships between any of our explanatory variables. This will be important during our modeling phase when we have to consider whether to include correlated variables or not.

Exploratory Data Analysis

A Note About Treatments

According to the variable information table, we note that `txgrp` could have four levels (1: ZDV + 3TC, 2: ZDV + 3TC + IDV, 3: d4T + 3TC, and 4: d4T + 3TC + IDV). However, this dataset contains only two levels of `txgrp` (1: ZDV + 3TC, 2: ZDV + 3TC + IDV), as shown below:

```
data %>% group_by(txgrp) %>% summarise(n())
```

```
## # A tibble: 2 x 2
##   txgrp `n()`
##   <int> <int>
## 1     1   422
## 2     2   429
```

In fact, since the variable `tx` is supposed to indicate whether the treatment contained IDV, we might assume that `txgrp` and `tx` are redundant information in this dataset and that a 1 in `txgrp` is equivalent to a 0 in `tx` while a 2 in `txgrp` is equivalent to a 1 in `tx`. We confirm this hunch below.

The following code says: create a new dataframe by taking all the rows in `data` where `txgrp` is 1 and `tx` is 0 *or* `txgrp` is 2 and `tx` is 1. Now, make sure that new dataframe is identical to the original data frame, and return `TRUE` if this is indeed the case.

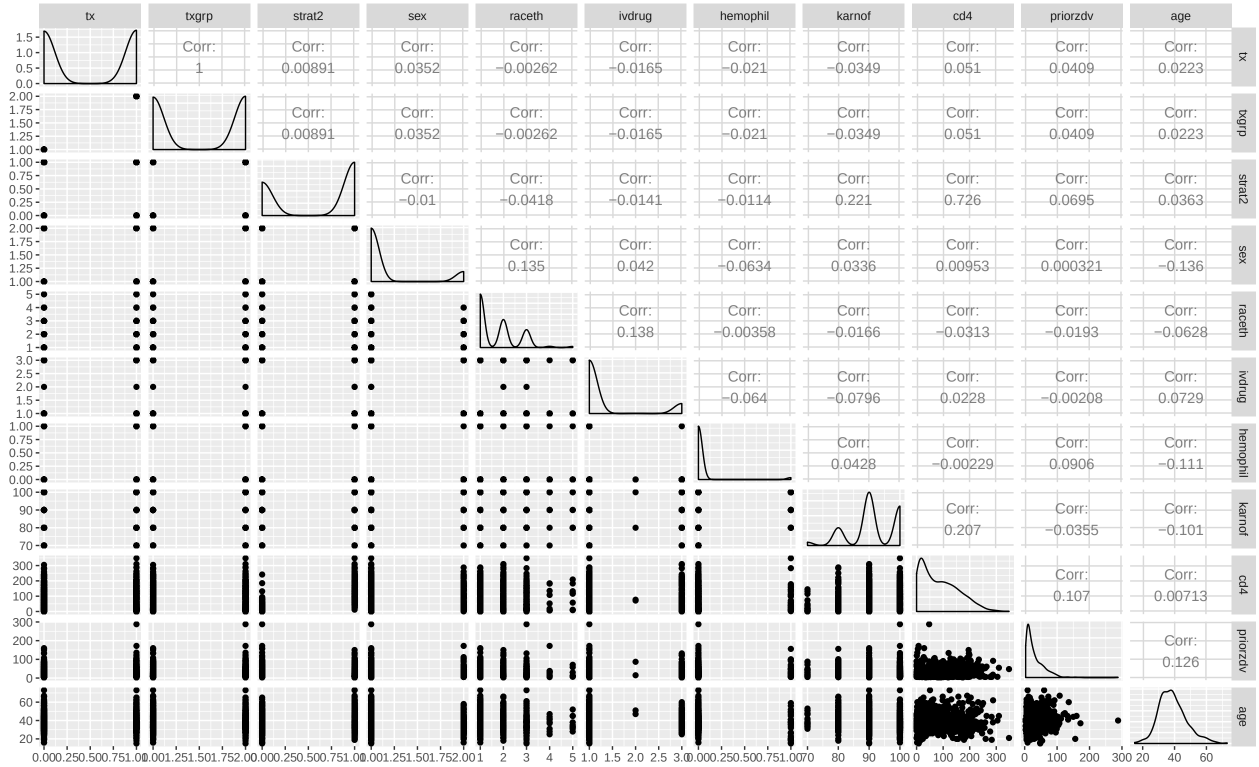
```
# Is it true that for every entry in `data`
all(
  (data %>%
     filter(((txgrp == 1 && tx == 0) || (txgrp == 2 && tx == 1))))
  == data
) == TRUE
```

```
## [1] TRUE
```

Correlation

We present a pairs plot of our explanatory variables, excluding id, our time-to-event variables, and our censoring variables to 1) visualize the distribution of the variables and 2) identify potential pair-wise correlation. `cd4` and `strat2` have a correlation coefficient of 0.74, which indicates moderate to strong correlation. This makes sense since `strat2` is the indicator variable for the continuous variable, `cd4`. Additionally, as noted just above, `tx` and `txgrp` provide the exact same information and therefore are perfectly correlated. Lastly, we would like to note that `sex`, `ivdrug`, and `hemophil` are highly unbalanced variables, meaning that one level of the variables is disproportionately represented relative to the other level(s).

```
data %>%
  select(tx:age) %>%
  ggpairs()
```



Censored vs. Non-Censored

It's worth noting that there are, in fact, two censored time-to-event variables. The primary variable of interest is `time` which is time in days to AIDs diagnosis or death, and this is informed by `censor`, which is 1 (true) if an individual was either diagnosed with AIDS *or* died during the course of the study and 0 otherwise. The other censored variable is `time_d` which is the time in days to death alone, governed by `censor_d` which is 1 if the person died during the study and 0 if not.

Since the primary variable of interest is time to AIDs diagnosis or death, we examine the complete (non-censored) individuals - those who were either diagnosed with AIDS *or* who died over the course of the study. The only caveat is that there are only 69 such individuals out of a study of 851 - most of the participants did not die or get diagnosed before the study's end.

```
non_censored <- data %>% filter(censor == 1) %>%
  mutate(tx=ifelse(tx == 0, "Control", "IDV"))
View(non_censored)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
## Resources/modules/R_de.so'' had status 1
```

```
dim(non_censored) # complete AIDS or death
```

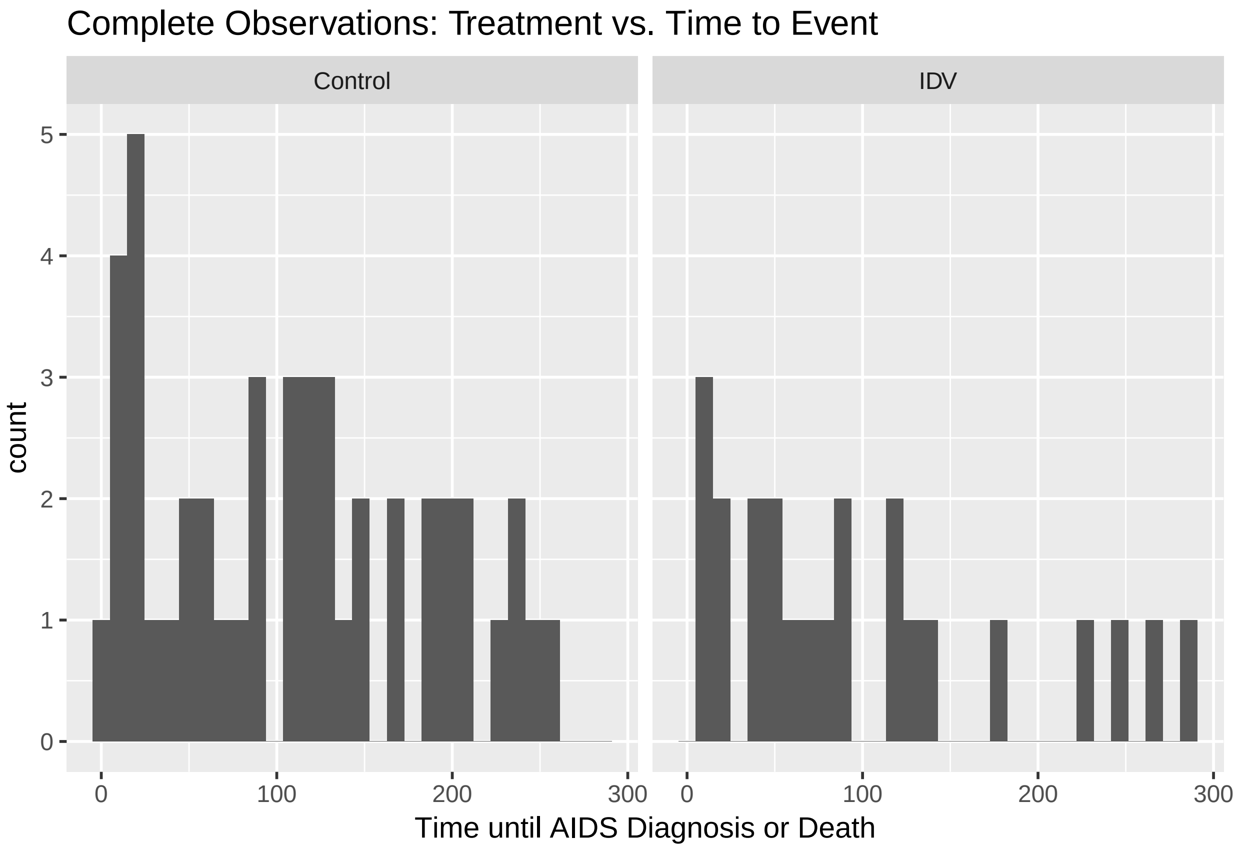
```
## [1] 69 16
```

```
dim(data)         # everyone
```

```
## [1] 851  16
```

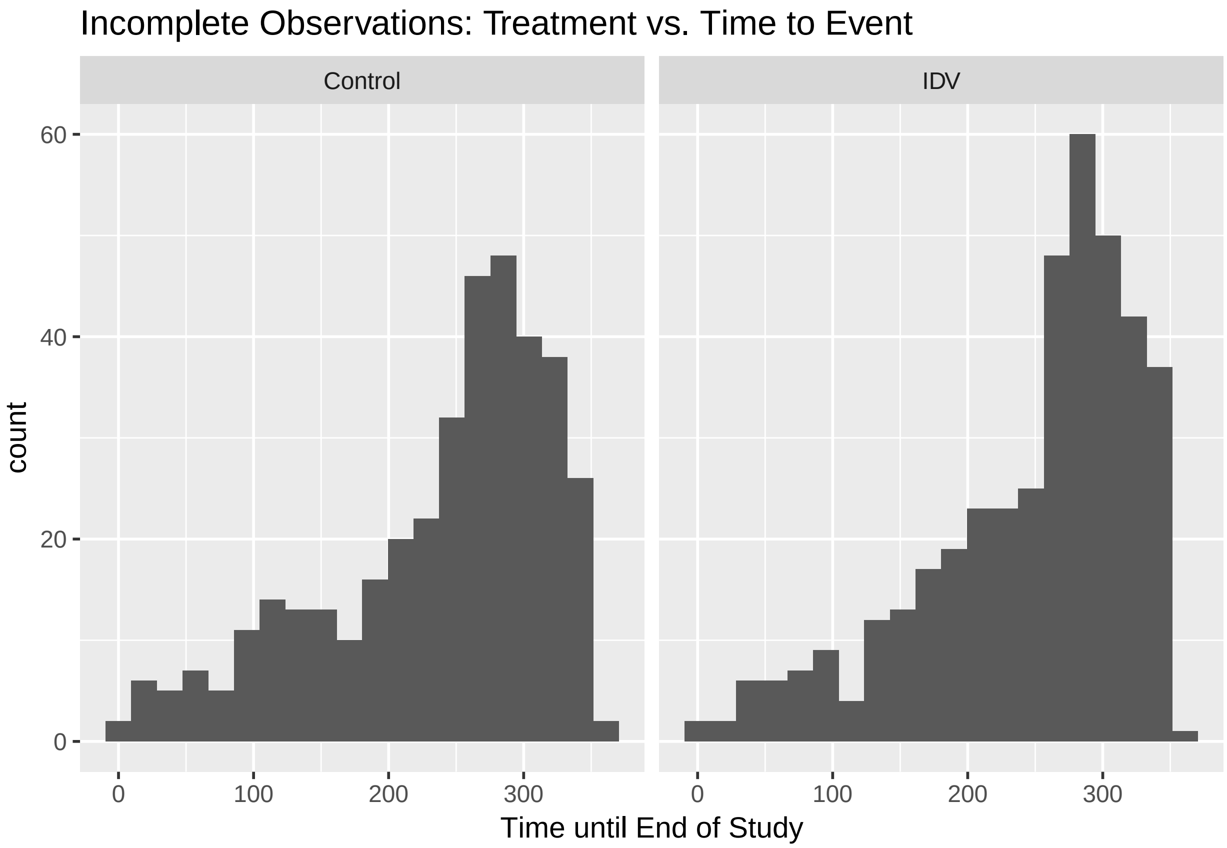
Among those with complete times, we notice from the side-by-side histograms below that both the control and IDV groups are skewed right. This makes sense - for complete observations, it's probably less common for people to last a long time without being diagnosed or dying. The distributions between the control and IDV groups don't look that different however, especially given the tiny sample sizes.

```
ggplot(non_censored, aes(x = time)) + geom_histogram(bins = 30) + facet_grid(.~tx) + ggtitle("Complete
```



When looking at the censored (incomplete) times for diagnosis/death, both control and IDV groups are in the opposite direction (left).

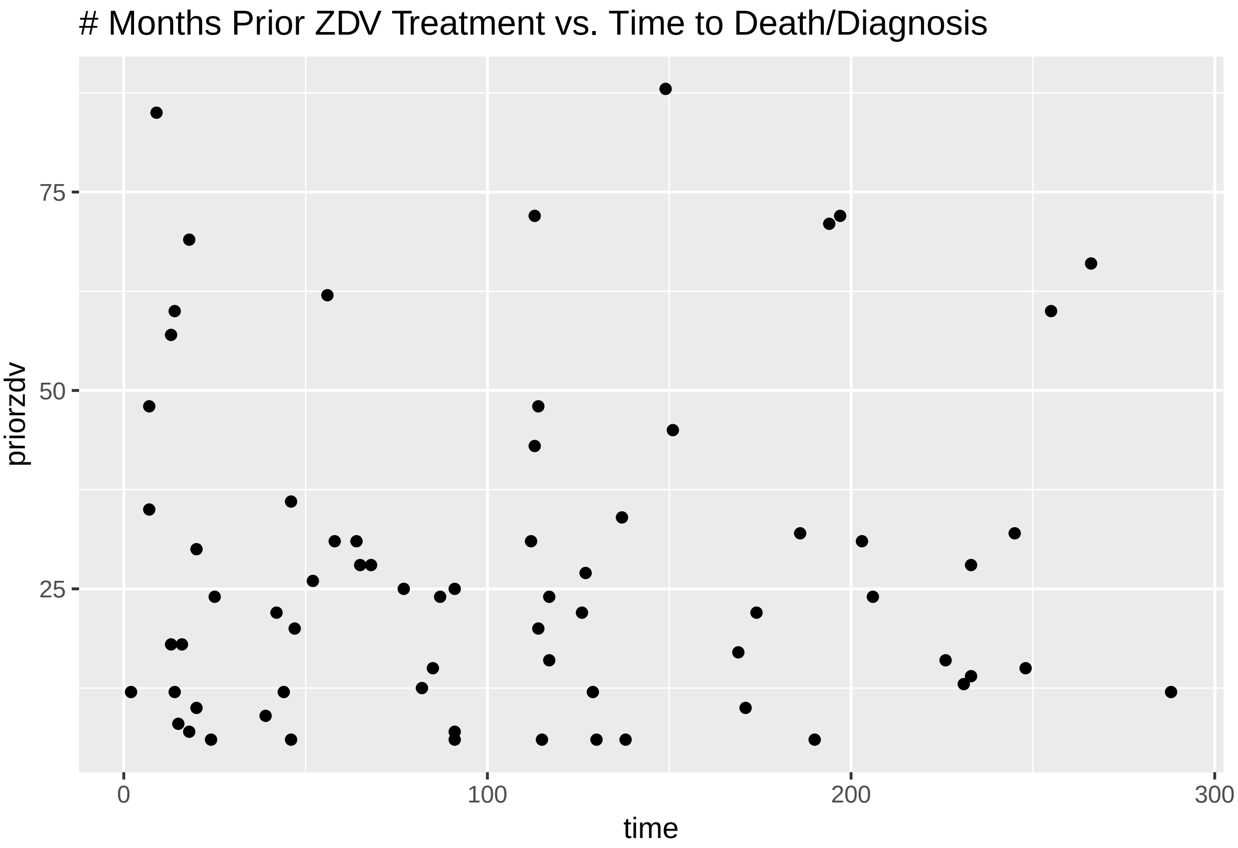
```
ggplot(data %>% filter(censor == 0) %>% mutate(tx=ifelse(tx == 0, "Control", "IDV")), aes(x = time_d))
```



Prior ZDV on Complete Observations

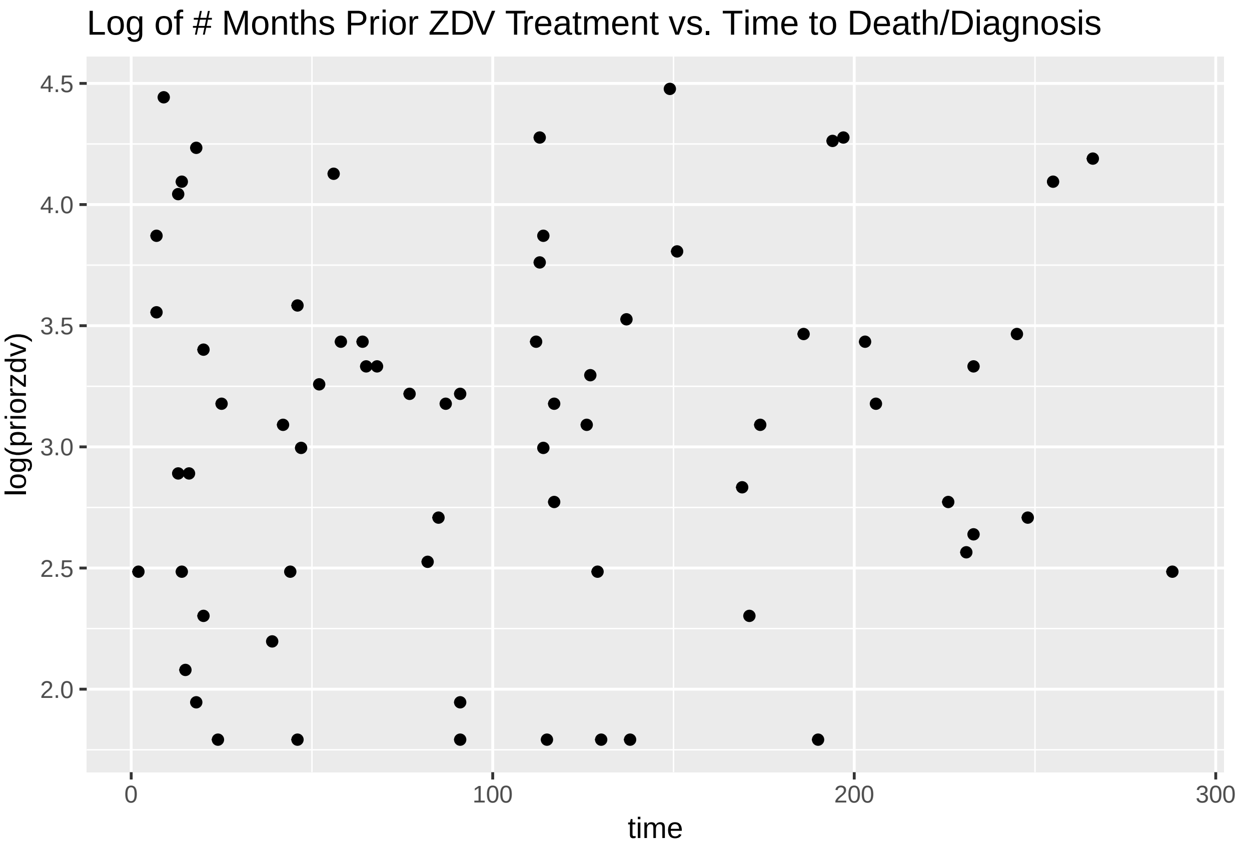
We were also curious about the relationship between time to diagnosis/death and number of months of prior ZDV use for non-censored participants. Interestingly, there appeared to be no relationship whatsoever, as evidenced by the following scatterplot:

```
ggplot(non_censored, aes(x = time, y= priorzdv)) +
  geom_point() +
  ggtitle("# Months Prior ZDV Treatment vs. Time to Death/Diagnosis")
```



This finding is made even clearer when we log the number of months of prior zdv:

```
ggplot(non_censored, aes(x = time, y= log(priorzdv))) +
  geom_point() +
  ggtitle("Log of # Months Prior ZDV Treatment vs. Time to Death/Diagnosis")
```


Log of # Months Prior ZDV Treatment vs. Time to Death/Diagnosis
log(priorzdv)
4.5
4.0
3.5
3.0
2.5
2.0
0
100
200
300
time