The effect of IDV on patients with HIV

Gbeke Fawehinmi and Janie Neal

4/25/2017

In this paper we will analyze the effect of a drug treatment, IDV, on the time it takes patients to die or develop aids. We hope to build a model that accurately estimates the effect or non-effect of the drug, and analyze that model to gather useful information about how IDV should be used in the future.

Meet our Data

Our data set is a collection of observations taken from a sample of 1156 people infected with HIV in order to test the effect of IDV. We have a control group who did not use medicine with IDV and another group who did. These groups are of about equal size. The sample has a much larger proportion of men than women with 738 men and 148 women (Figure 1). The mean age is 38.7 with 75% of the people being 44 or younger, giving us a sample of people mostly in late young adulthood (Figure 2). Our sample has significant numbers of white non-hispanic, black non-hispanic, and hispanic people (Figure 3). Due to the small amount of Asian/Pacific Islander and American Indian people included in the sample we will not make conclusions about the effect of these ancestries on the effectiveness of IDV.

Figure 1

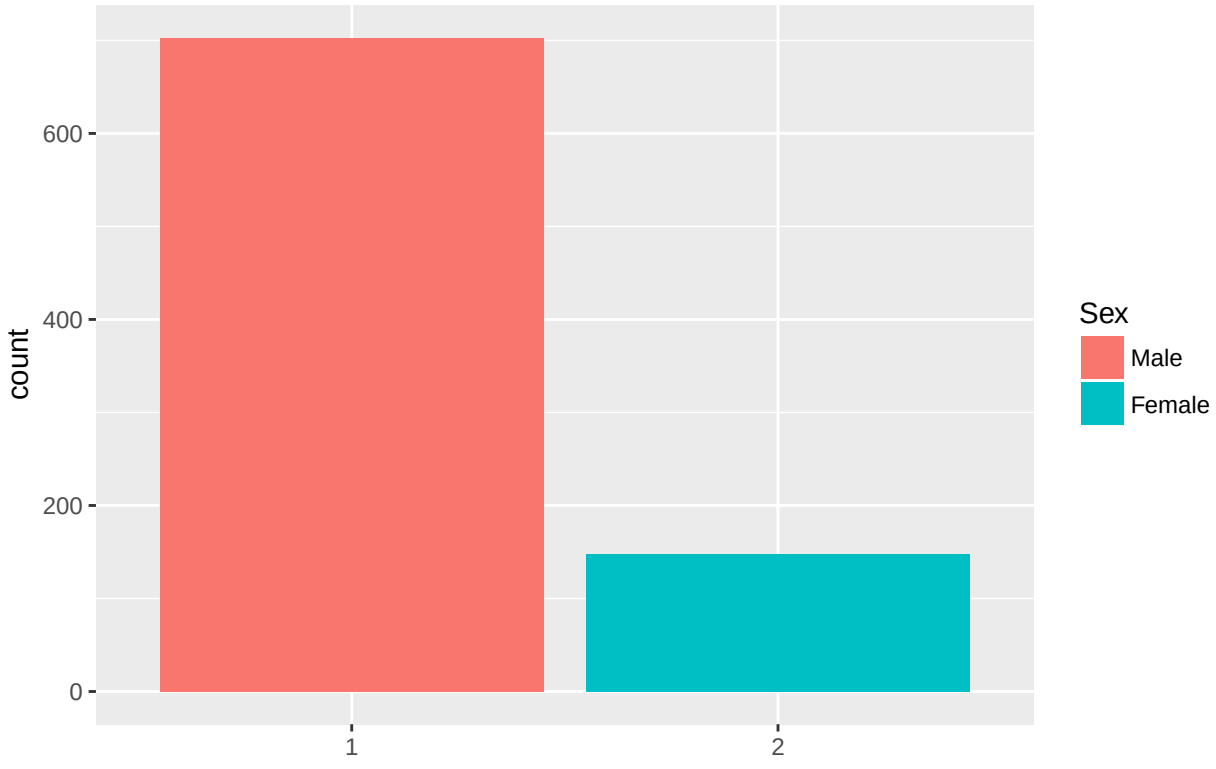


Figure 2

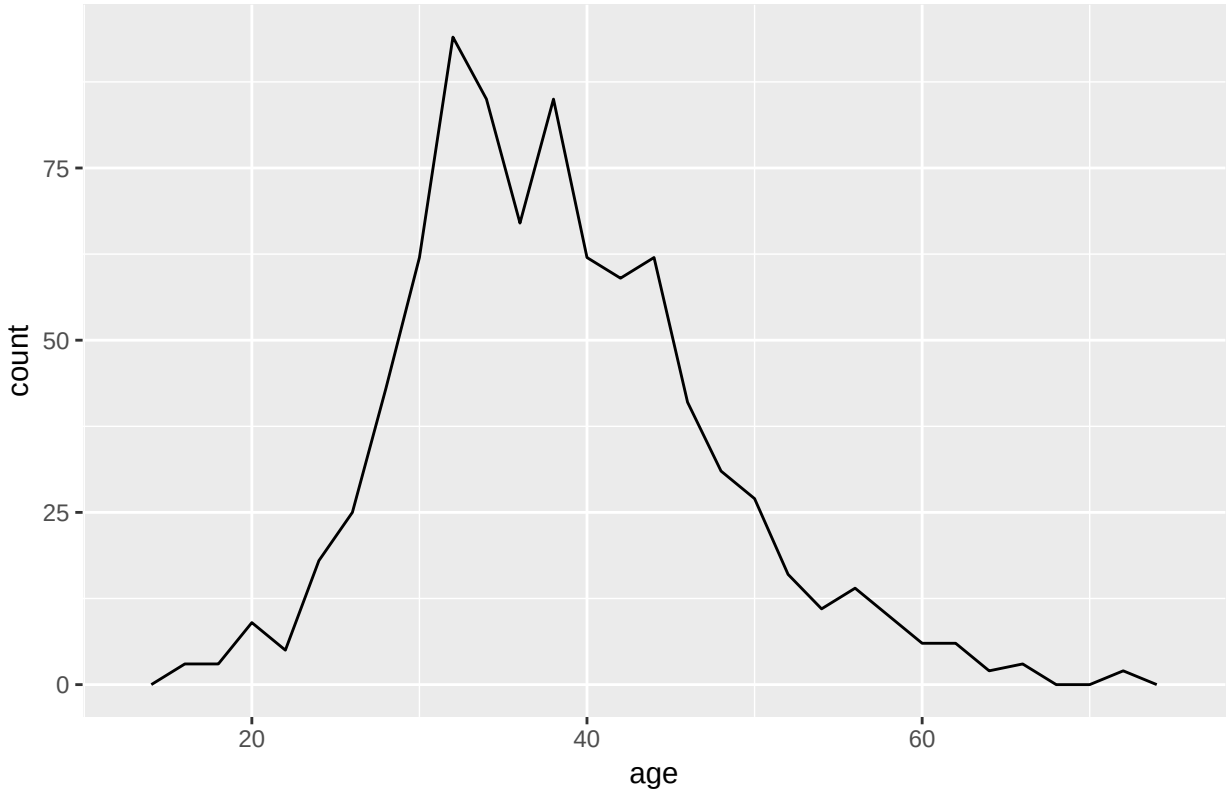
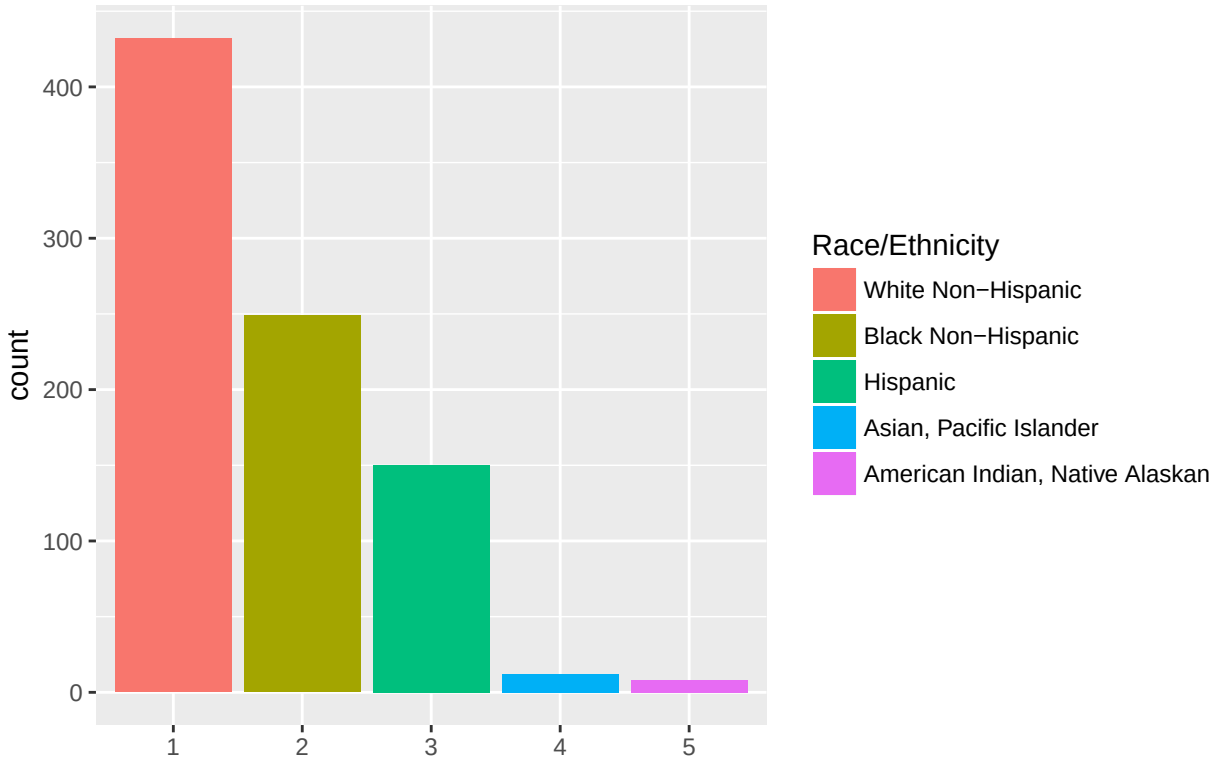


Figure 3



The Model

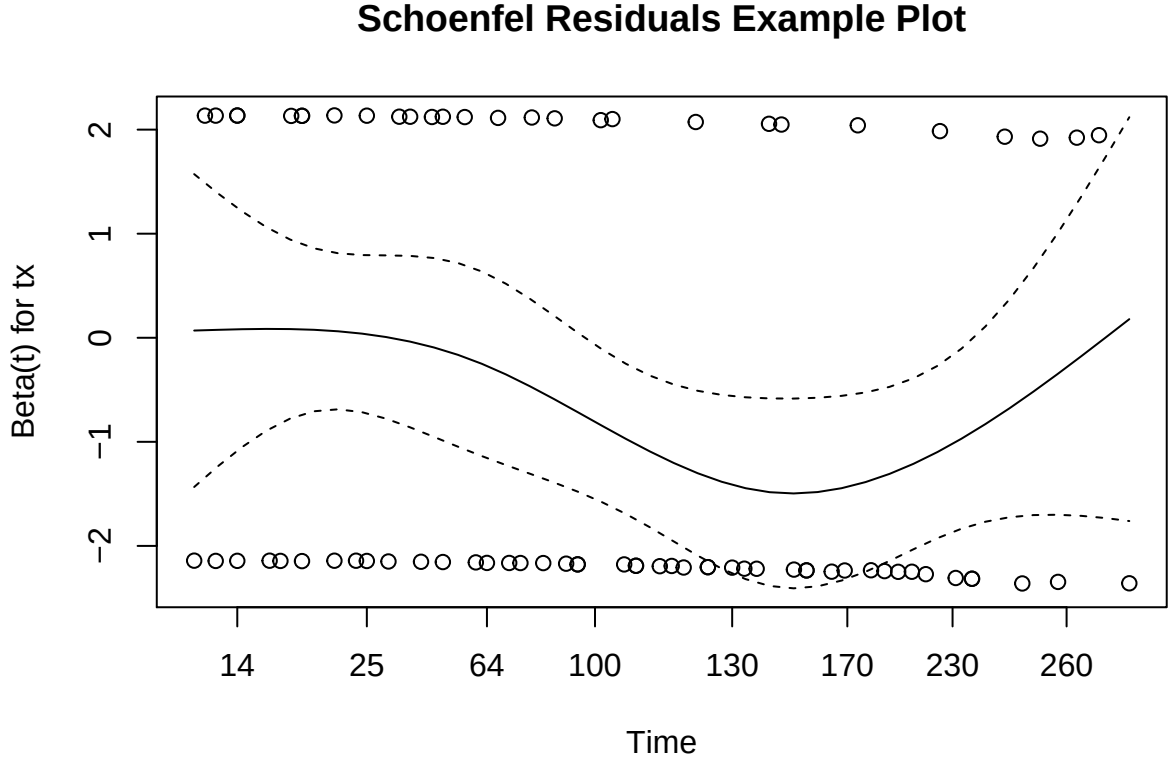
We are using the Cox proportional hazard regression model, a method for investigating the effect of several variables upon the time a specified event takes to happen.

Assumptions

The first assumption is that censoring of individuals must not be related to the probability of an event occurring. For example, participants should not be censored when they leave the study as a result of bad effects of the drug. We have no certain way of checking this but believe that the data fulfills this condition.

The Cox proportional hazard model also, as the name indicates, assumes proportional hazards. We use the R function cox.zph to assess whether our data fulfills this assumption and can be accurately fitted to a proportional hazard model. For each covariate, the function correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.(Easy Guides) If the function returns a significant p value for the relationship between any covariate's Schoenfeld residuals and time then we must assume the proportional assumption is violated. If we are reasonably confident that no relationships exist, as we are for this data due to high p values, then we can proceed.

```
##            rho  chisq      p
## tx     -0.2144 3.5560 0.0593
## sex    -0.1163 1.0758 0.2996
## raceth -0.0170 0.0257 0.8727
## karnof -0.0696 0.3499 0.5542
## age     0.0443 0.1244 0.7243
## ivdrug -0.1115 0.9221 0.3369
## cd4     0.2291 3.4130 0.0647
## GLOBAL      NA 9.1437 0.2425
```



Building and Final Product

In order to create the best fitting Cox Ph model for this data, we calculated the AIC, BIC, and a new AIC for survival analysis.

AIC stands for Akaike information criterion, and it measures the quality of each model in a collection of statistical models. It measures the quality by assigning each model with a numerical value, and the best model has the lowest value. The formula is:

$$AIC = -2ln(likelihood) + 2p$$

In this case, we used forward model building, which meant we started with a model that had no covariates (just the intercept) and built more models by adding one covariate at a time. This covariate is only added if the AIC calculation with this variable was the lowest one in relation to those of the other variables, and also lower than if no variable was added at all. After adding the explanatory variables that make the AIC the lowest possible, those variables are used in our best fitted model.

BIC stands for Bayesian Information Criterion, and it works in the same way as AIC, however, because the likelihood grows with more parameters, then we penalize the BIC by adding ln(n)p instead of just 2p (as it is in the AIC calculation). So the formula for BIC is:

$$BIC = -2ln(likelihood) + ln(n)p$$

This limits the amount of parameters that are in the model.

From further research on AIC, BIC, and survival analysis, we found an equation from the article "Improved AIC Selection Strategy for Survival Analysis", which extends the traditional AIC to apply to survival analysis. The formula is:

$$AIC(SUR) = AIC + \frac{2(p+2)(p+3)}{n-p-3}$$

where p is the number of covariates in the model, and n is the total number of observations. This calculation is a better representation of the quality of the model, and is applied to each of the AIC calculations in order to compare them. We will still calculate AIC and BIC, however, these are inferior (especially BIC) to the AIC(SUR).

Calculating AIC

```
fmod.aic <- coxph(Surv(time,censor) ~ 1, data=aidsdata) #Intercept
step(fmod.aic, ~ (tx + as.factor(txgrp) + strat2 + as.factor(sex) + as.factor(raceth) + as.factor(ivdru
```

```
## Start:  AIC=981.12
## Surv(time, censor) ~ 1
##
##                      Df    AIC
## + cd4                 1 944.81
## + karnof              1 952.20
## + strat2              1 962.66
## + tx                  1 976.44
## + as.factor(txgrp)    1 976.44
## + age                 1 979.22
## <none>                  981.12
```

```
## + hemophil           1 981.77
## + priorzdv           1 982.71
## + as.factor(sex)     1 982.93
## + as.factor(ivdrug)  2 983.03
## + as.factor(raceth)  4 983.20
##
## Step:  AIC=944.81
## Surv(time, censor) ~ cd4
##
##                     Df    AIC
## + karnof             1 927.06
## + tx                 1 940.88
## + as.factor(txgrp)   1 940.88
## + age                1 941.76
## <none>                 944.81
## + hemophil           1 945.05
## + as.factor(sex)     1 946.40
## + strat2             1 946.44
## + as.factor(ivdrug)  2 946.66
## + priorzdv           1 946.78
## + as.factor(raceth)  4 946.79
##
## Step:  AIC=927.06
## Surv(time, censor) ~ cd4 + karnof
##
##                     Df    AIC
## + tx                 1 922.48
## + as.factor(txgrp)   1 922.48
## + age                1 926.21
## <none>                 927.06
## + hemophil           1 928.09
## + as.factor(sex)     1 928.34
## + strat2             1 928.54
## + as.factor(ivdrug)  2 928.66
## + priorzdv           1 929.01
## + as.factor(raceth)  4 929.27
##
## Step:  AIC=922.48
## Surv(time, censor) ~ cd4 + karnof + tx
##
##                     Df    AIC
## + age                1 921.44
## <none>                 922.48
## + hemophil           1 923.38
## + as.factor(ivdrug)  2 923.58
## + as.factor(sex)     1 923.94
## + strat2             1 924.06
## + priorzdv           1 924.42
## + as.factor(raceth)  4 925.43
##
## Step:  AIC=921.44
## Surv(time, censor) ~ cd4 + karnof + tx + age
##
##                     Df    AIC
```

```
## <none>                  921.44
## + as.factor(sex)     1 922.37
## + hemophil           1 922.62
## + as.factor(ivdrug)  2 922.82
## + strat2             1 923.16
## + priorzdv           1 923.33
## + as.factor(raceth)  4 924.14
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ cd4 + karnof + tx + age,
##     data = aidsdata)
##
##             coef exp(coef) se(coef)     z       p
## cd4    -0.01178   0.98829  0.00262 -4.51 6.6e-06
## karnof -0.05791   0.94373  0.01338 -4.33 1.5e-05
## tx     -0.61120   0.54270  0.23922 -2.55   0.011
## age     0.02205   1.02229  0.01244  1.77   0.076
##
## Likelihood ratio test=67.7  on 4 df, p=7.01e-14
## n= 851, number of events= 75
```

By calculating the AIC at all of the steps of the drop-in deviance, we find that the best fitted Cox PH Model uses the variables cd4, karnof, tx, and age.

Calculating BIC

```
fmod.bic <- coxph(Surv(time,censor) ~ 1, data=aidsdata) #Intercept
step(fmod.bic, ~ (tx + as.factor(txgrp) + strat2 + as.factor(sex) + as.factor(raceth) + as.factor(ivdru
```

```
## Start:  AIC=981.12
## Surv(time, censor) ~ 1
##
##                      Df     AIC
## + cd4                 1  949.55
## + karnof              1  956.95
## + strat2              1  967.41
## <none>                   981.12
## + tx                  1  981.18
## + as.factor(txgrp)    1  981.18
## + age                 1  983.97
## + hemophil            1  986.52
## + priorzdv            1  987.46
## + as.factor(sex)      1  987.67
## + as.factor(ivdrug)   2  992.53
## + as.factor(raceth)   4 1002.18
##
## Step:  AIC=949.55
## Surv(time, censor) ~ cd4
##
##                      Df    AIC
## + karnof              1 936.55
## <none>                  949.55
## + tx                  1 950.37
## + as.factor(txgrp)    1 950.37
```

```
## + age                1 951.25
## + hemophil           1 954.55
## + as.factor(sex)     1 955.89
## + strat2             1 955.93
## + priorzdv           1 956.27
## + as.factor(ivdrug)  2 960.90
## + as.factor(raceth)  4 970.53
##
## Step:  AIC=936.55
## Surv(time, censor) ~ cd4 + karnof
##
##                     Df    AIC
## <none>                 936.55
## + tx                 1 936.72
## + as.factor(txgrp)   1 936.72
## + age                1 940.45
## + hemophil           1 942.33
## + as.factor(sex)     1 942.58
## + strat2             1 942.78
## + priorzdv           1 943.25
## + as.factor(ivdrug)  2 947.64
## + as.factor(raceth)  4 957.74
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ cd4 + karnof, data = aidsdata)
##
##            coef exp(coef) se(coef)     z       p
## cd4    -0.01164   0.98843  0.00258 -4.52 6.2e-06
## karnof -0.06047   0.94132  0.01331 -4.54 5.5e-06
##
## Likelihood ratio test=58.1  on 2 df, p=2.47e-13
## n= 851, number of events= 75
```

According to the comparisons of the BIC at all the steps of the drop-in deviance, we find that the best fitted Cox PH model uses the variables cd4 and karnof. This makes sense in relation to the AIC because we are penalizing the likelihood much more now, so the model doesn't allow for as many parameters.

Calculating AIC(SUR)

By using the Survival Analysis correction formula, we get the following numbers for each of the steps respectively.

```
981.12
```

```
## [1] 981.12
```

```
944.81 + ((2*3*4)/(851-1-3))
```

```
## [1] 944.8383
```

```
927.06 + ((2*4*5)/(851-2-3))
```

```
## [1] 927.1073
```

```
922.48 + ((2*5*6)/(851-3-3))
```

```
## [1] 922.551
```

```
921.44 + ((2*6*7)/(851-4-3))
```

```
## [1] 921.5395
```

```
922.37 + ((2*7*8)/(851-5-3)) #AIC calculation if we added the next variable, which is as.factor(sex)
```

```
## [1] 922.5029
```

We see that the last corrected AIC is in fact still the smallest compared to the other steps. Since this AIC calculation supports the model created by the original AIC, we will use that model as our best fit model.

Final Model

The coefficients for the covariates used in our final model are:

```
##         cd4      karnof          tx         age
## -0.01171470 -0.05626561 -0.45729811  0.02229333
```

Written in equation form that is $\lambda(t|X_{1i}, X_{2i}, X_{3i}, X_{4i}) = \lambda_0(t)e^{-0.0118X_{1i}+-0.0579X_{2i}+-0.611X_{3i}+0.022X_{4i}}$.

Analysis and Discussion

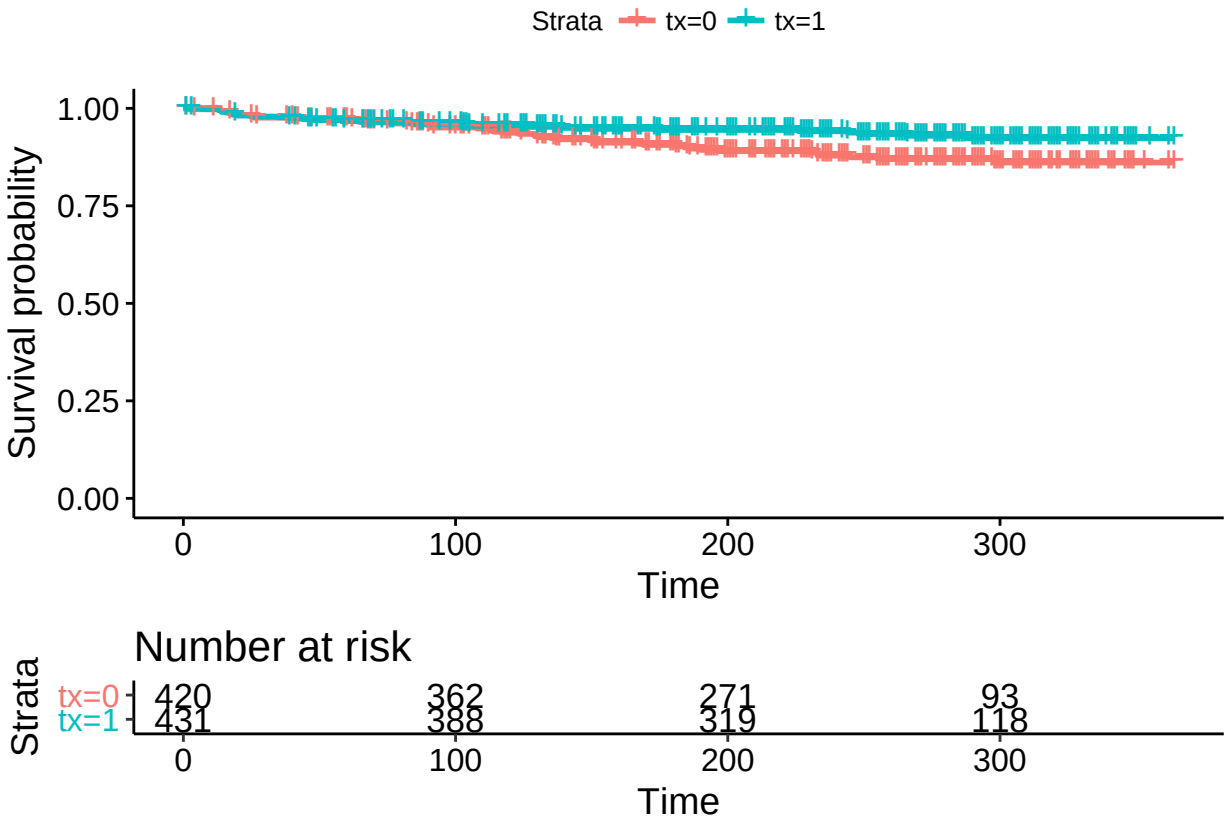
By analyzing the best fit model, we find that rate of survival of the patients in this trial is related to the following variables: if the patient had IDV in their treatment, the patient's number on the Karnofsky Performance Scale, the baseline CD4 cell count of the patient, and the age of the patient. We can interpret the coefficients in the following ways:

1. For cd4, there is an 1.17% decrease in the expected hazard for each 1 unit increase in baseline CD4 count.
2. For Karnofsky Performance scale, there is an 5.63% decrease in the expected hazard for each 1 unit increase.
3. For treatment, there is an 45.7% decrease in the expected hazard when receiving the treatment verus not.
4. For age, there is an 2.23% increase in the expected hazard for each year increase in age.

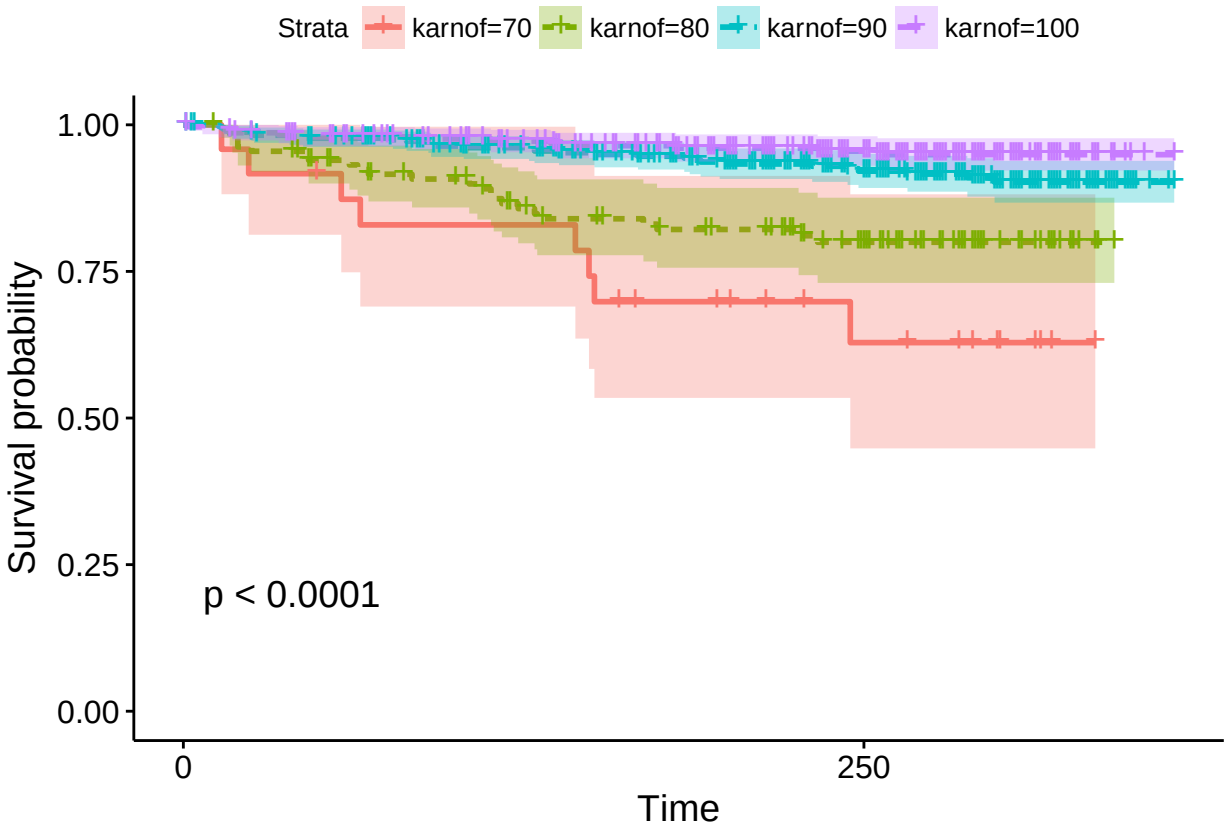
In less mathematical terms, younger, healthier patients with higher CD4 count are slightly more likely to survive or not develop AIDS at any time, and patients that recieved IDV are 45% more likely to survive or not develop AIDS than those who didn't. We are comfortable with generalizing this to American men between the ages of 30 and 50 yeaars old, because they were well represented in the study. Because the data comes from a double blind, placebo-controlled trial we are able to say IDV caused the decrease in the expected hazard of dying or developing AIDS.

Fun Visualizations

These survival plots show how cateogrizing the data in different ways can drastically effect how the estimated hazard is calculated. We see a major effect of Karnof Scores on the survival of patients.




```
## Warning in .get_data(fit, data = data): The `data` argument is not
## provided. Data will be extracted from model fit.
```



Credits

Bibliography

Liang, Hua, and Guohua Zou. "Improved AIC Selection Strategy for Survival Analysis." Computational statistics & data analysis 52.5 (2008): 2538–2548. PMC. Web. 27 Apr. 2017.

Easy Guides. "Cox Model Assumptions." Rbloggers. Tal Galili. 12 Dec. 2016. Web. 27 Apr. 2017 https://www.r-bloggers.com/cox-model-assumptions/

Who did what

Janie: Research on cox.zph Gbeke: Research on AIC and BIC for survival models