COG pediatric AML EDA - Metadata Version 2

Jenea Adams

Jan 10 2022

Contents

References: - https://resources.github.com/github-and-rstudio/

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.2.1     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# library(mmtable2)
library(ggplot2)
library(formatR)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 40), tidy = TRUE)
```

Read in the data

```
setwd("~/OneDrive - Children's Hospital of Philadelphia/COG-pediatric-AML/metadata_v2")

pat.met = read.csv("patient-met.csv", header = T)
wgs = read.csv("WGS.csv", header = T)
rna = read.csv("RNA.csv", header = T)
# regno =
# read.csv('regno-to-caseID.csv',
# header = T)
```

```
setwd("~/OneDrive - Children's Hospital of Philadelphia/COG-pediatric-AML/metadata_v2")
new_pat.met = read.csv("NCT01371981-D4-Dataset.csv")
```

#Inspect the imported data

```
# head(new_pat.met)
```

```
# head(rna)
```

Change the name of rna case_id –> usi

```
colnames(rna)[1] = "usi"
# head(rna)
```

same for wgs

```
colnames(wgs)[1] = "usi"
# head(wgs)
```

#Parse subsets and intersections of the datasets

RNAseq and matched WGS –> 390

```
length(intersect(rna$usi, wgs$usi))
```

```
## [1] 390
```

Same as above but using the filter function from dplyr

```
x = filter(rna, rna$usi %in% wgs$usi)

nrow(x)
```

```
## [1] 390
```

RNA-seq data and patient metadata –> 1040

```
length(intersect(rna$usi, new_pat.met$usi))
```

```
## [1] 1040
```

Sanity check: How many tumor WGS do we have? –> 0

```
wgs.tum.pat.met = filter(wgs, wgs$usi %in%
    new_pat.met$usi & !wgs$sample_type ==
    "Normal")
nrow(wgs.tum.pat.met)
```

```
## [1] 0
```

N individuals with WGS (normal) x RNA-seq tumor x Pat.Met –> 390

```
rna_norm = filter(rna, sample_type == "Normal")

rna_tum = filter(rna, sample_type == "Tumor")

# x2 = filter(rna_tum, rna_tum$usi %in%
# wgs$usi & rna_tum$usi %in% regno$usi
# & !rna_tum %in% rna_norm)

rna.wgs = merge(rna, wgs, by = "usi")
rna_tum.wgs = merge(rna_tum, wgs, by = "usi")
```

```
x2 = filter(new_pat.met, new_pat.met$usi %in%
    rna_tum$usi & new_pat.met$usi %in% wgs$usi)

nrow(x2)
```

```
## [1] 390
```

It seems to be that the only intersections of RNA x WGS and Pat.Met are limited by the amount of available WGS data

How many patients have metadata and EITHER WGS or RNA data? –> 1058 – We will use this lateer for downstream visualizations

```
x3 = filter(new_pat.met, new_pat.met$usi %in%
    rna$usi | new_pat.met$usi %in% wgs$usi)
nrow(x3)
```

```
## [1] 1058
```

How many individuals have patient metadata AND WGS?

```
x4 = filter(new_pat.met, new_pat.met$usi %in%
    wgs$usi)
nrow(x4)
```

```
## [1] 408
```

How many individuals have RNA-seq tumor AND patient metadata

```
x5 = filter(new_pat.met, new_pat.met$usi %in%
    rna_tum$usi)
nrow(x5)
```

```
## [1] 1040
```

How many patients have WGS (normal) data but NO RNA-seq or Pat.Met?

```
x6 = filter(wgs, !wgs$usi %in% rna$usi &
    !wgs$usi %in% new_pat.met$usi)
nrow(x6)
```

```
## [1] 0
```

How many patients have RNA-seq tumor but NO RNA-sew tumor or WGS or Pat.Met?

```
x7 = filter(rna_tum, !rna_tum$usi %in% wgs$usi &
    !rna_tum$usi %in% new_pat.met$usi & !rna_tum$usi %in%
    rna_norm$usi)
nrow(x7)
```

```
## [1] 11
```

How many patients have WGS and RNA-seq tumor only?

```
x8 = filter(rna_tum, rna_tum$usi %in% wgs$usi &
    !rna_tum$usi %in% new_pat.met$usi & !rna_tum$usi %in%
    rna_norm$usi)
nrow(x8)
```

```
## [1] 0
```

How many individuals do we have RNA-seq and pat.met for?

```
x9 = filter(rna, rna$usi %in% new_pat.met$usi)
nrow(x9)
```

```
## [1] 1040
```

why is this the same number as row 2 in the table? are all of these tumor data? –> yes

```
table(x9$sample_type)
```

```
##
## Tumor
##  1040
```

Visualizing patient demographic data

We will use x3 to represent patients with metadata and some kind of cavatica data (RNA or WGS)

```
cavatica.aaml2 = x3
```

We will look at Race, AgeYr, Risk Group, CBF Identification, years to event free survival, and in the future WBCl_i for leukemic burden. There are 130+ features by which to model variation across the cohort.

Each of the these is now represented by number indicator variables, and will need to be switched to strings for visualization.

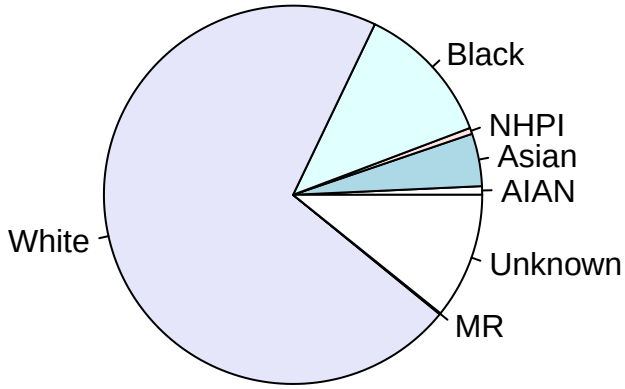
First let's just define the categories.

```
race = cavatica.aaml2$race_cat

age = cavatica.aaml2$ageyr

risk = cavatica.aaml2$riskgrp

cbf = cavatica.aaml2$cbf_pt

yrsefs = cavatica.aaml2$yrsefs
```

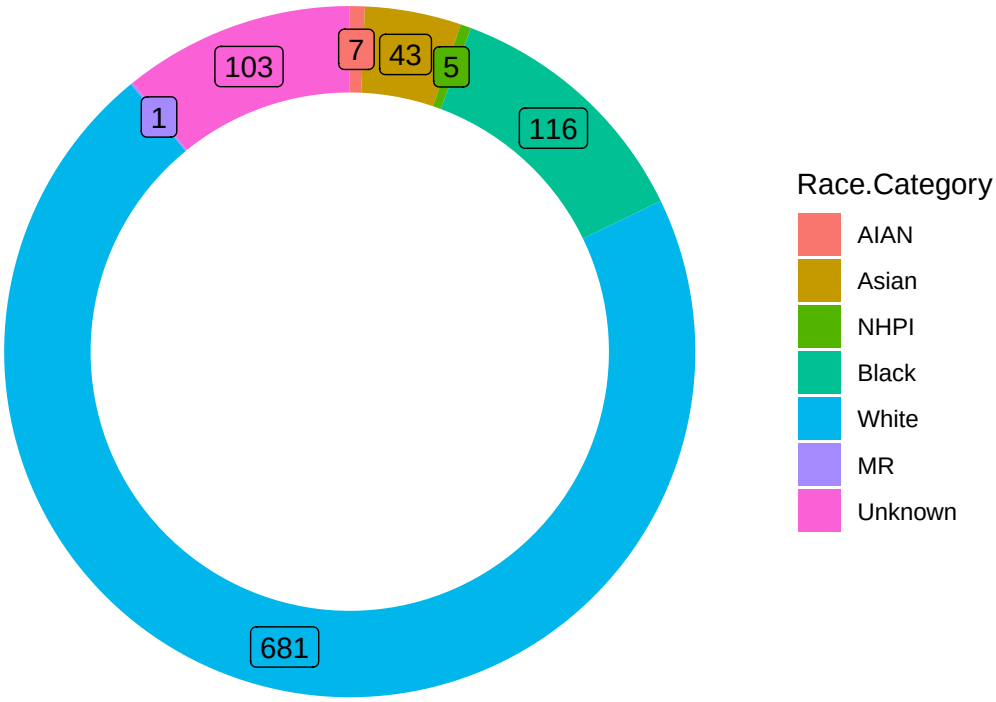
Race

```
race.table = data.frame(table(race))
pie(race.table$Freq, labels = c("AIAN", "Asian",
    "NHPI", "Black", "White", "MR", "Unknown"),
    main = "Pie chart of updated race categories")
```

Pie chart of updated race categories




```
colnames(race.table)[1] = "Race.Category"
colnames(race.table)[2] = "Count"

cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

race.table$fraction = round(race.table$Count/sum(race.table$Count), 3)

race.table$ymax = cumsum(race.table$fraction)

race.table$ymin = c(0, head(race.table$ymax, n=-1))

race_donut_plot = ggplot(race.table, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Race.Category)) +
  geom_rect() +
  geom_label(aes(label = Count, x = 3.5, y = (ymin+ymax)/2), inherit.aes = T, show.legend = F) +
  coord_polar(theta="y") + # Try to remove that to understand how the chart is built initially
  xlim(c(0, 4)) + theme_void() + scale_fill_discrete(labels = c("AIAN", "Asian", "NHPI", "Black", "Whit



race_donut_plot
```



Of the known race categories, the proporation of white is

```
681/(7 + 43 + 5 + 116 + 1 + 681)
```

```
## [1] 0.7983587
```

non-white is

```
1 - 0.7983587
```

```
## [1] 0.2016413
```

Survival Time By Race

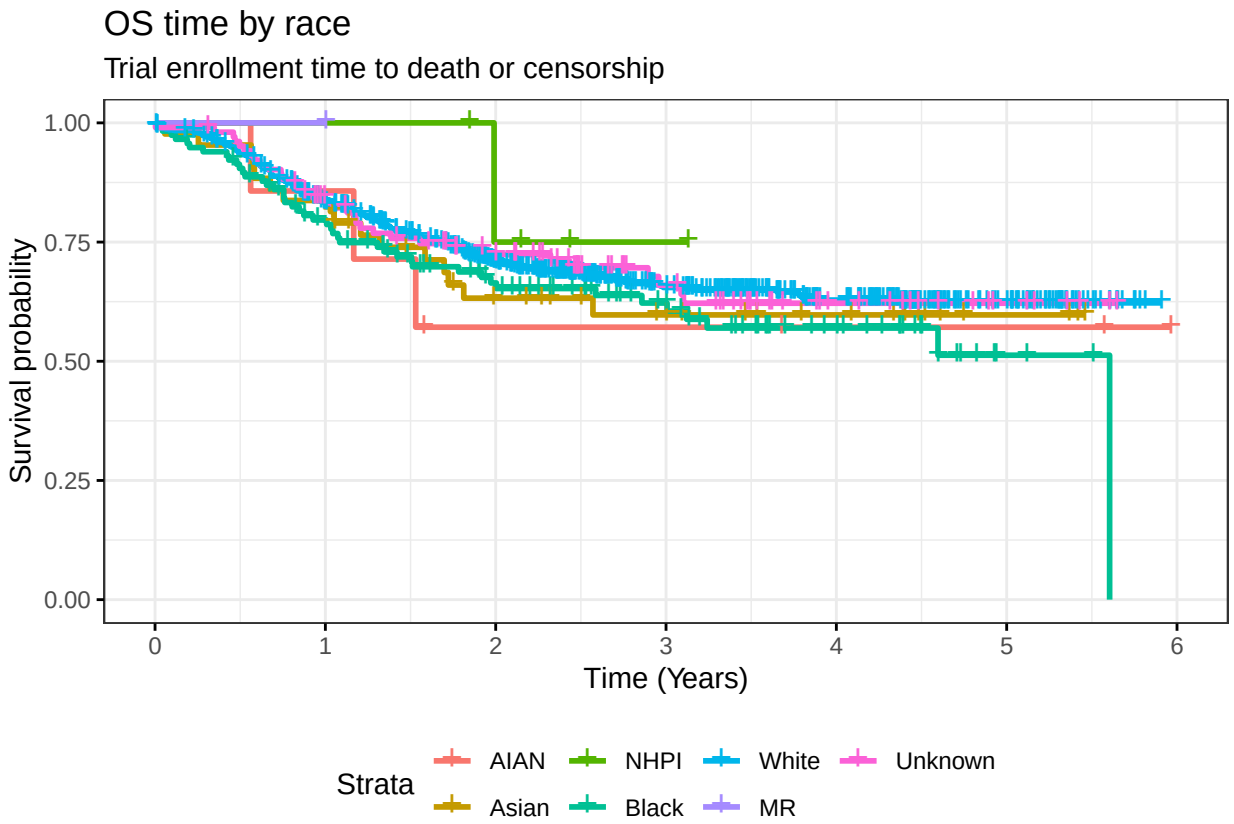
```
library(survival)
library(survminer)
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```r
# library(ggplot2)



# fit0 =
# survfit(Surv(enroll.death.censored,
# Death.Status) ~1, type =
# 'kaplan-meier', data =
# cavatica.aaml2) summary(fit0)

fit_race = survfit(Surv(yrsos, osi) ~ race_cat,
    type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_race))

ggsurvplot(fit_race, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("AIAN",
        "Asian", "NHPI", "Black", "White",
        "MR", "Unknown"), ggtheme = theme_bw(),
    xlab = "Time (Years)", title = "OS time by race",
    subtitle = "Trial enrollment time to death or censorship")
```



Statistical test of significance between these groups

```r
survdiff(Surv(yrsos, osi) ~ race_cat, data = cavatica.aaml2)
```

```
## Call:
## survdiff(formula = Surv(yrsos, osi) ~ race_cat, data = cavatica.aaml2)
```

```
##
## n=956, 102 observations deleted due to missingness.
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## race_cat=1   7        3    2.377    0.1636    0.1651
## race_cat=2  43       16   14.129    0.2477    0.2595
## race_cat=3   5        1    1.898    0.4247    0.4277
## race_cat=4 116       44   35.558    2.0040    2.2608
## race_cat=5 681      218  226.403    0.3119    1.1095
## race_cat=6   1        0    0.187    0.1874    0.1878
## race_cat=9 103       33   34.448    0.0609    0.0684
##
##  Chisq= 3.4  on 6 degrees of freedom, p= 0.8
```

```
coxph(Surv(yrsos, osi) ~ race_cat, data = cavatica.aaml2)
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ race_cat, data = cavatica.aaml2)
##
##              coef exp(coef) se(coef)      z    p
## race_cat -0.02819   0.97220  0.03732 -0.756 0.45
##
## Likelihood ratio test=0.58  on 1 df, p=0.4464
## n= 956, number of events= 315
##    (102 observations deleted due to missingness)
```

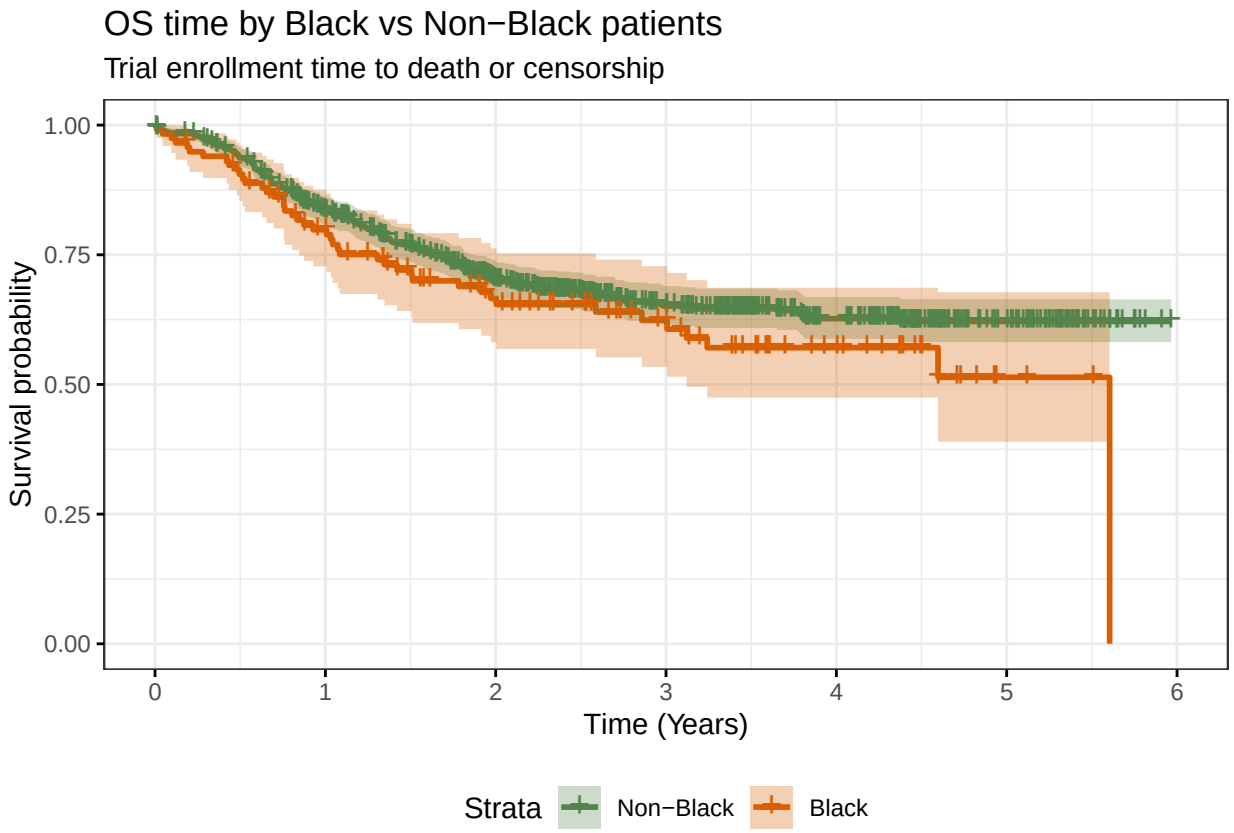
```
summary(coxph(Surv(yrsos, osi) ~ race_cat,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ race_cat, data = cavatica.aaml2)
##
##   n= 956, number of events= 315
##    (102 observations deleted due to missingness)
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## race_cat -0.02819   0.97220  0.03732 -0.756     0.45
##
##          exp(coef) exp(-coef) lower .95 upper .95
## race_cat    0.9722      1.029    0.9036     1.046
##
## Concordance= 0.516  (se = 0.014 )
## Likelihood ratio test= 0.58  on 1 df,   p=0.4
## Wald test            = 0.57  on 1 df,   p=0.4
## Score (logrank) test = 0.57  on 1 df,   p=0.4
```

Black vs non-black OS survival

```
fit_black.nb = survfit(Surv(yrsos, osi) ~
    raceb, type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_black.nb))
```

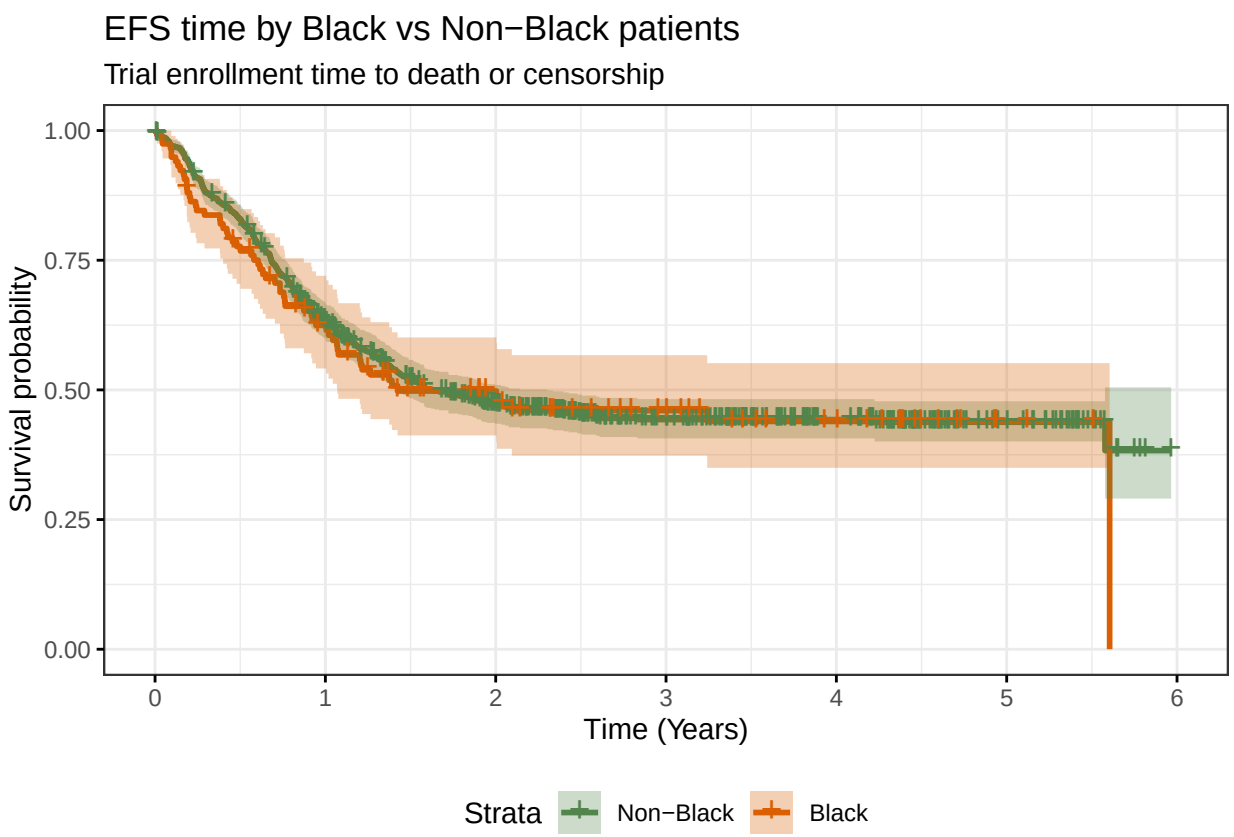
```
ggsurvplot(fit_black.nb, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("Non-Black",
        "Black"), ggtheme = theme_bw(), xlab = "Time (Years)",
    palette = c("#52854C", "#D95F02"), conf.int = T,
    title = "OS time by Black vs Non-Black patients",
    subtitle = "Trial enrollment time to death or censorship")
```




```
summary(coxph(Surv(yrsos, osi) ~ raceb, data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ raceb, data = cavatica.aaml2)
##
##   n= 853, number of events= 282
##    (205 observations deleted due to missingness)
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## raceb 0.2367    1.2671   0.1641 1.442    0.149
##
##       exp(coef) exp(-coef) lower .95 upper .95
## raceb     1.267     0.7892    0.9185     1.748
##
## Concordance= 0.513  (se = 0.011 )
## Likelihood ratio test= 1.97  on 1 df,   p=0.2
## Wald test            = 2.08  on 1 df,   p=0.1
## Score (logrank) test = 2.09  on 1 df,   p=0.1
```

```
fit_black.nb_efs = survfit(Surv(yrsefs, efsi) ~
    raceb, type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_black.nb_efs))

ggsurvplot(fit_black.nb_efs, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("Non-Black",
        "Black"), ggtheme = theme_bw(), xlab = "Time (Years)",
    palette = c("#52854C", "#D95F02"), conf.int = T,
    title = "EFS time by Black vs Non-Black patients",
    subtitle = "Trial enrollment time to death or censorship")
```



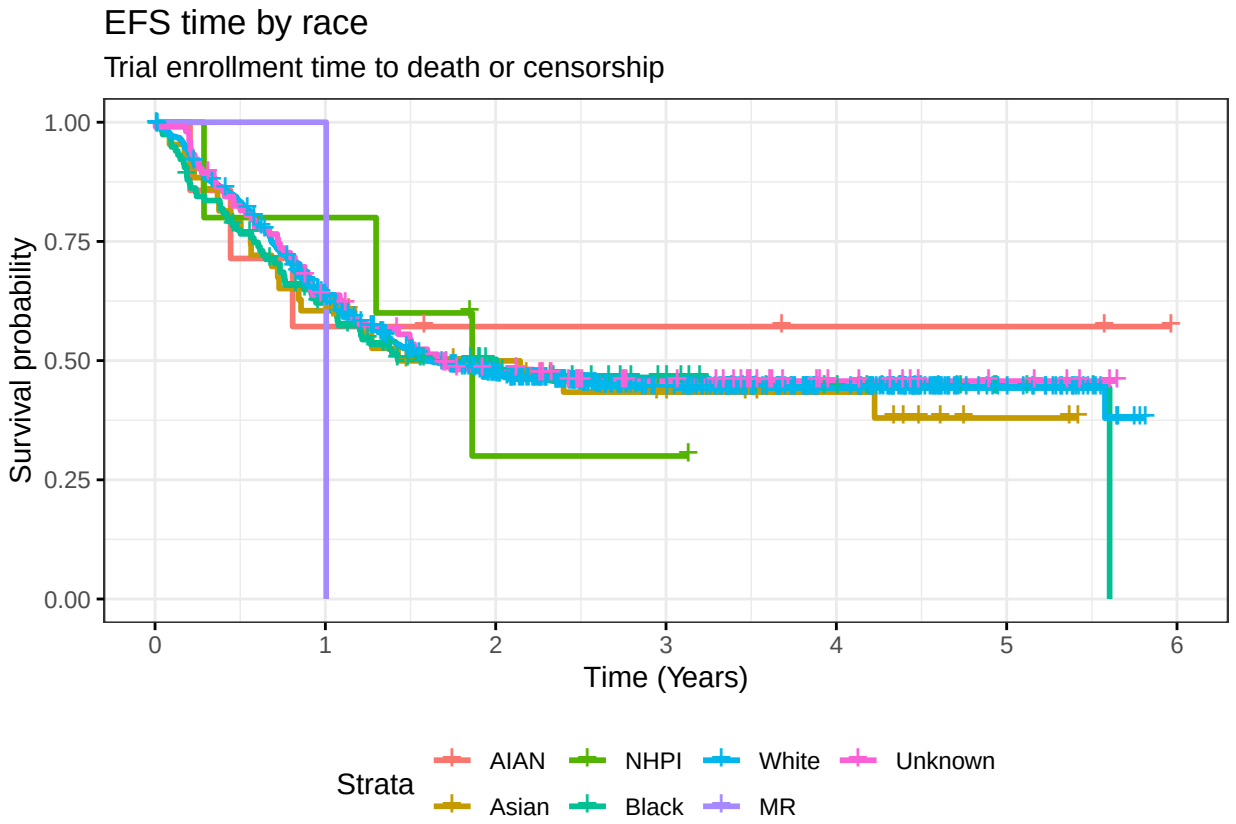

```
summary(coxph(Surv(yrsefs, efsi) ~ raceb,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ raceb, data = cavatica.aaml2)
##
##   n= 853, number of events= 451
##    (205 observations deleted due to missingness)
##
##          coef exp(coef) se(coef)    z Pr(>|z|)
## raceb 0.05915   1.06093  0.13770 0.43    0.668
##
##       exp(coef) exp(-coef) lower .95 upper .95
## raceb     1.061     0.9426      0.81      1.39
##
```

```
## Concordance= 0.506  (se = 0.009 )
## Likelihood ratio test= 0.18  on 1 df,   p=0.7
## Wald test            = 0.18  on 1 df,   p=0.7
## Score (logrank) test = 0.18  on 1 df,   p=0.7
```

EFS survival race

```
fit_race_efs = survfit(Surv(yrsefs, efsi) ~
    race_cat, type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_race_efs))

ggsurvplot(fit_race_efs, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("AIAN",
        "Asian", "NHPI", "Black", "White",
        "MR", "Unknown"), ggtheme = theme_bw(),
    xlab = "Time (Years)", title = "EFS time by race",
    subtitle = "Trial enrollment time to death or censorship")
```



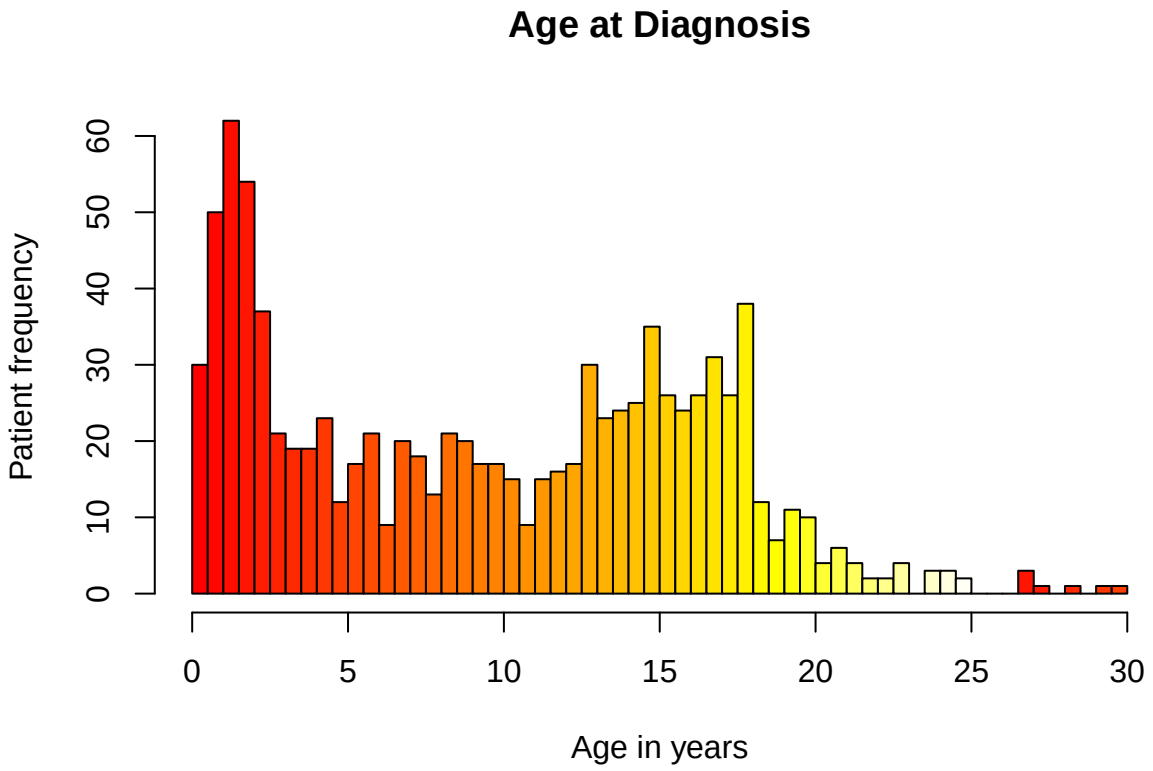

```
summary(coxph(Surv(yrsefs, efsi) ~ race_cat,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ race_cat, data = cavatica.aaml2)
##
##   n= 956, number of events= 505
##    (102 observations deleted due to missingness)
```

```
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## race_cat -0.01226   0.98782  0.02891 -0.424    0.672
##
##          exp(coef) exp(-coef) lower .95 upper .95
## race_cat    0.9878      1.012    0.9334     1.045
##
## Concordance= 0.507  (se = 0.011 )
## Likelihood ratio test= 0.18  on 1 df,   p=0.7
## Wald test            = 0.18  on 1 df,   p=0.7
## Score (logrank) test = 0.18  on 1 df,   p=0.7
```

Age

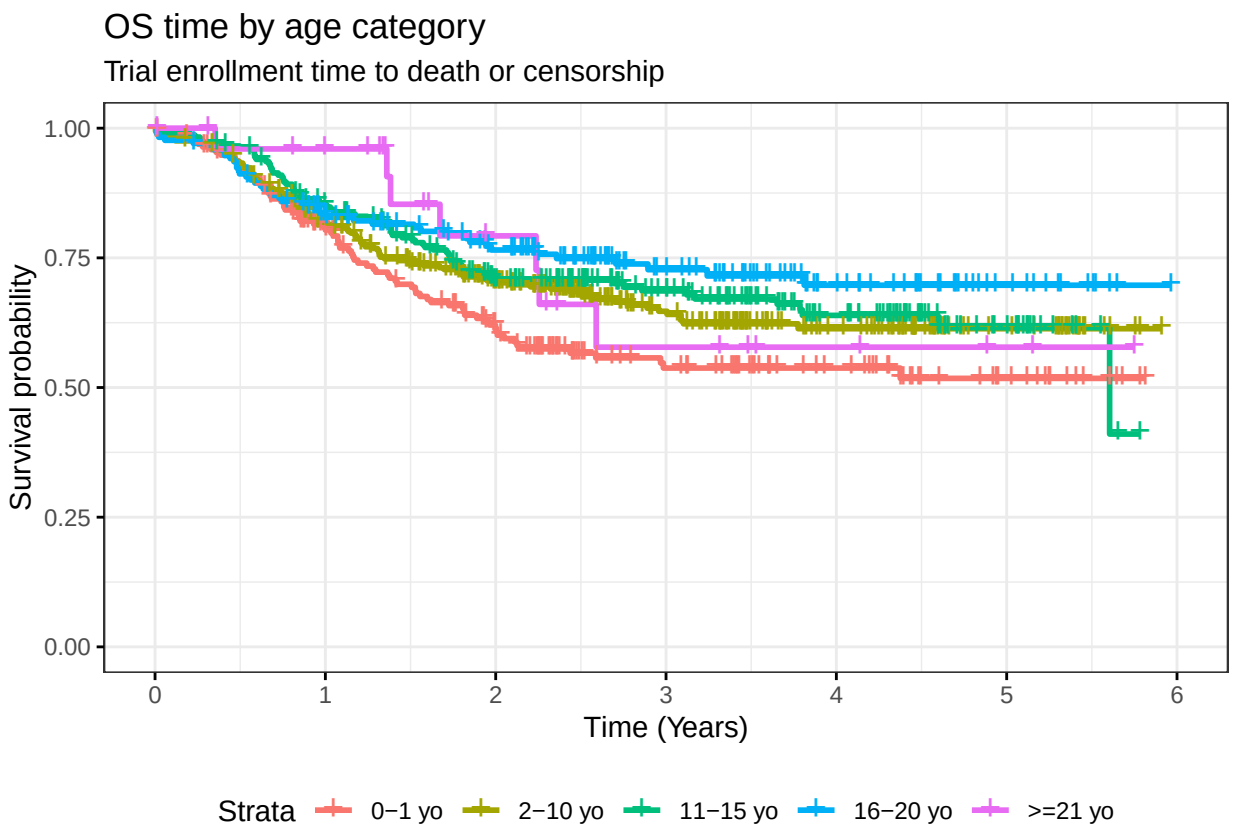
```
hist(age, breaks = 100, main = "Age at Diagnosis",
    col = heat.colors(50), xlim = c(0, 30),
    xlab = "Age in years", ylab = "Patient frequency")
```



Survival time by age category

```
fit_age = survfit(Surv(yrsos, osi) ~ agecateg_,
    type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_age))

ggsurvplot(fit_age, data = cavatica.aaml2,
```

```
    legend = "bottom", legend.labs = c("0-1 yo",
        "2-10 yo", "11-15 yo", "16-20 yo",
        ">=21 yo"), ggtheme = theme_bw(),
    xlab = "Time (Years)", title = "OS time by age category",
    subtitle = "Trial enrollment time to death or censorship")
```



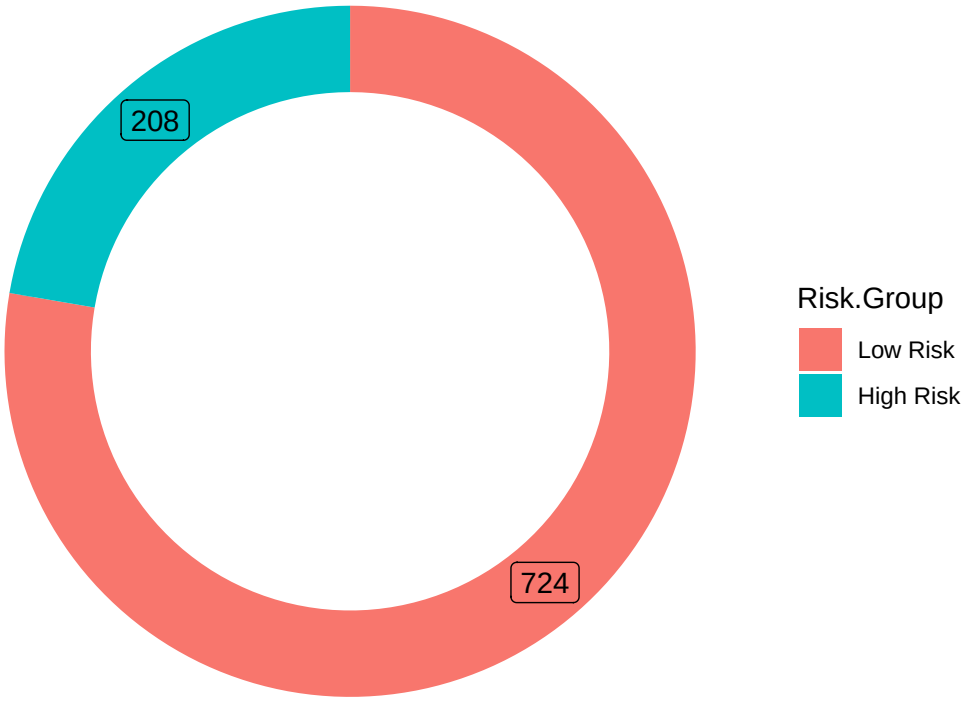

```
summary(coxph(Surv(yrsos, osi) ~ agecateg_,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ agecateg_, data = cavatica.aaml2)
##
##   n= 956, number of events= 315
##    (102 observations deleted due to missingness)
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## agecateg_ -0.16597   0.84708  0.05404 -3.071  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## agecateg_    0.8471      1.181    0.7619    0.9417
##
## Concordance= 0.55  (se = 0.016 )
## Likelihood ratio test= 9.65  on 1 df,   p=0.002
## Wald test            = 9.43  on 1 df,   p=0.002
## Score (logrank) test = 9.48  on 1 df,   p=0.002
```

Risk

```
risk.table = data.frame(table(risk))
```

```
colnames(risk.table)[1] = "Risk.Group"
colnames(risk.table)[2] = "Count"

cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

risk.table$fraction = round(risk.table$Count/sum(risk.table$Count), 3)

risk.table$ymax = cumsum(risk.table$fraction)

risk.table$ymin = c(0, head(risk.table$ymax, n=-1))

risk_donut_plot = ggplot(risk.table, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Risk.Group)) +
  geom_rect() +
  geom_label(aes(label = Count, x = 3.5, y = (ymin+ymax)/2), inherit.aes = T, show.legend = F) +
  coord_polar(theta="y") + # Try to remove that to understand how the chart is built initially
  xlim(c(0, 4)) + theme_void() + scale_fill_discrete(labels = c("Low Risk", "High Risk"))



risk_donut_plot
```



Creating a mosaic plot with ggplot across Race and Risk Group

```
# # install.packages('devtools') #
# devtools::install_github('haleyjeppson/ggmosaic')
# library(ggmosaic)
```

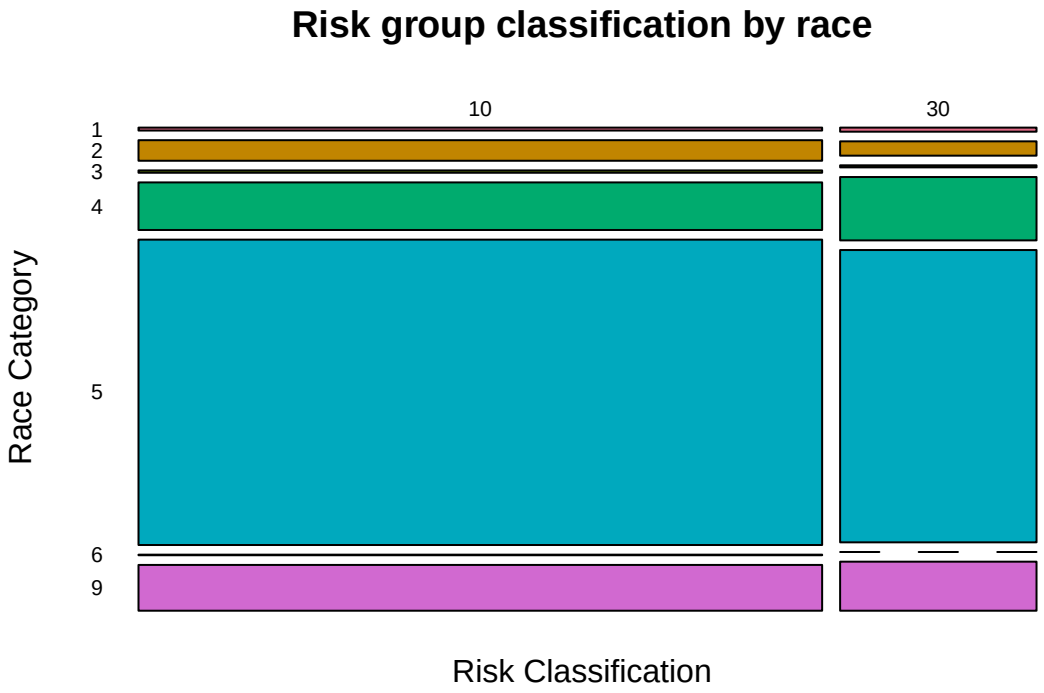
```
# ggplot(data = cavatica.aaml2) +
# geom_mosaic(aes(x = product(riskgrp),
# fill = race_cat)) + theme_mosaic()
```

```
race.risk = table(risk, race)
race.risk
```

```
##     race
## risk   1   2   3   4   5   6   9
##   10   5  35   4  81 520   1  78
##   30   2   7   1  31 143   0  24
```

```
race.risk_mosaic = mosaicplot(race.risk,
    xlab = "Risk Classification", ylab = "Race Category",
    main = "Risk group classification by race",
    col = colorspace::qualitative_hcl(7),
    clegend = T, las = 1)
```

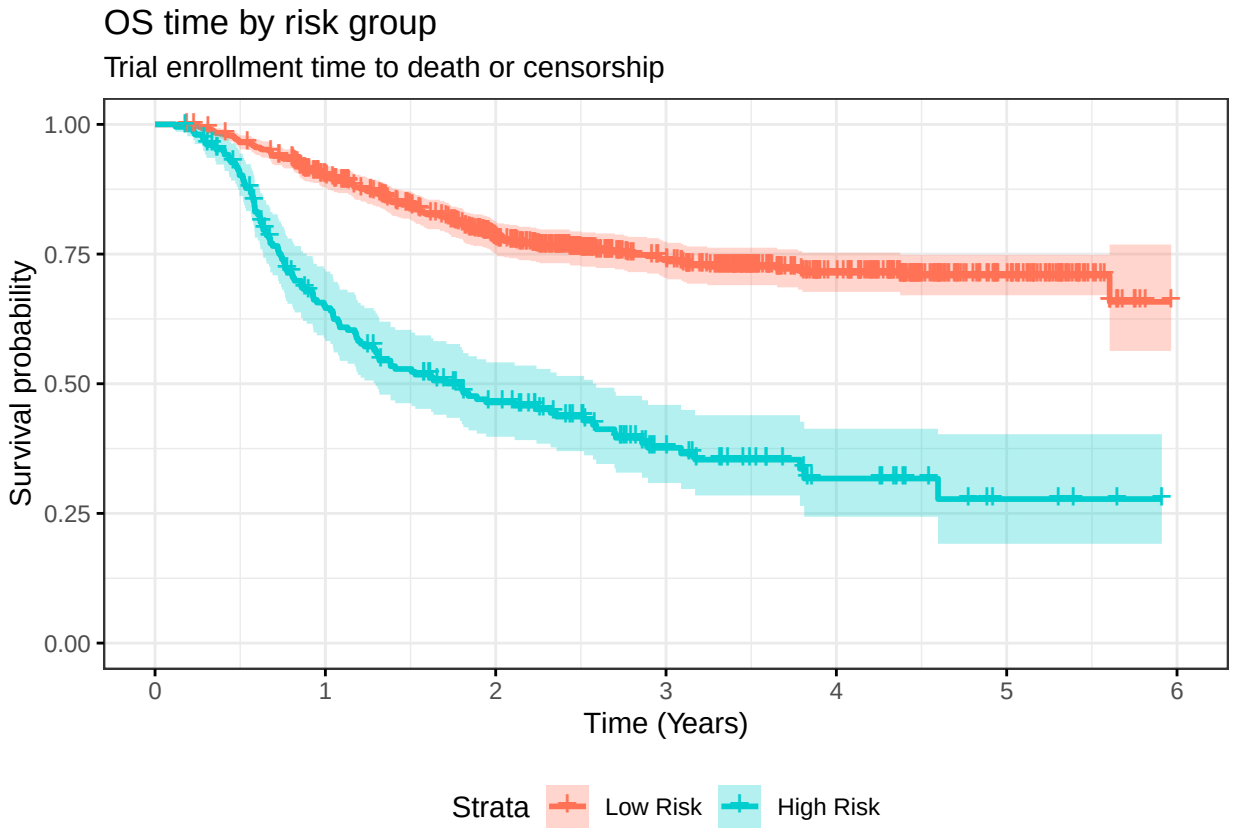
```
## Warning: In mosaicplot.default(race.risk, xlab = "Risk Classification", ylab = "Race Category",
##     main = "Risk group classification by race", col = colorspace::qualitative_hcl(7),
##     clegend = T, las = 1) :
##  extra argument 'clegend' will be disregarded
```



Survival time by risk group

```
fit_risk = survfit(Surv(yrsos, osi) ~ riskgrp,
    type = "kaplan-meier", data = cavatica.aaml2)
```

```
# head(summary(fit_risk))

ggsurvplot(fit_risk, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("Low Risk",
        "High Risk"), ggtheme = theme_bw(),
    xlab = "Time (Years)", conf.int = T,
    palette = c("coral1", "cyan3"), title = "OS time by risk group",
    subtitle = "Trial enrollment time to death or censorship")
```



Statistical test of significance between the two curves

```
survdiff(Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
```

```
## Call:
## survdiff(formula = Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
##
## n=932, 126 observations deleted due to missingness.
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## riskgrp=10 724      178    247.7      19.6       119
## riskgrp=30 208      119     49.3      98.6       119
##
##  Chisq= 119  on 1 degrees of freedom, p= <2e-16
```

```
coxph(Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
##
##             coef exp(coef) se(coef)     z      p
## riskgrp 0.061221  1.063134 0.005958 10.28 <2e-16
##
## Likelihood ratio test=93.03  on 1 df, p=< 2.2e-16
## n= 932, number of events= 297
##    (126 observations deleted due to missingness)
```

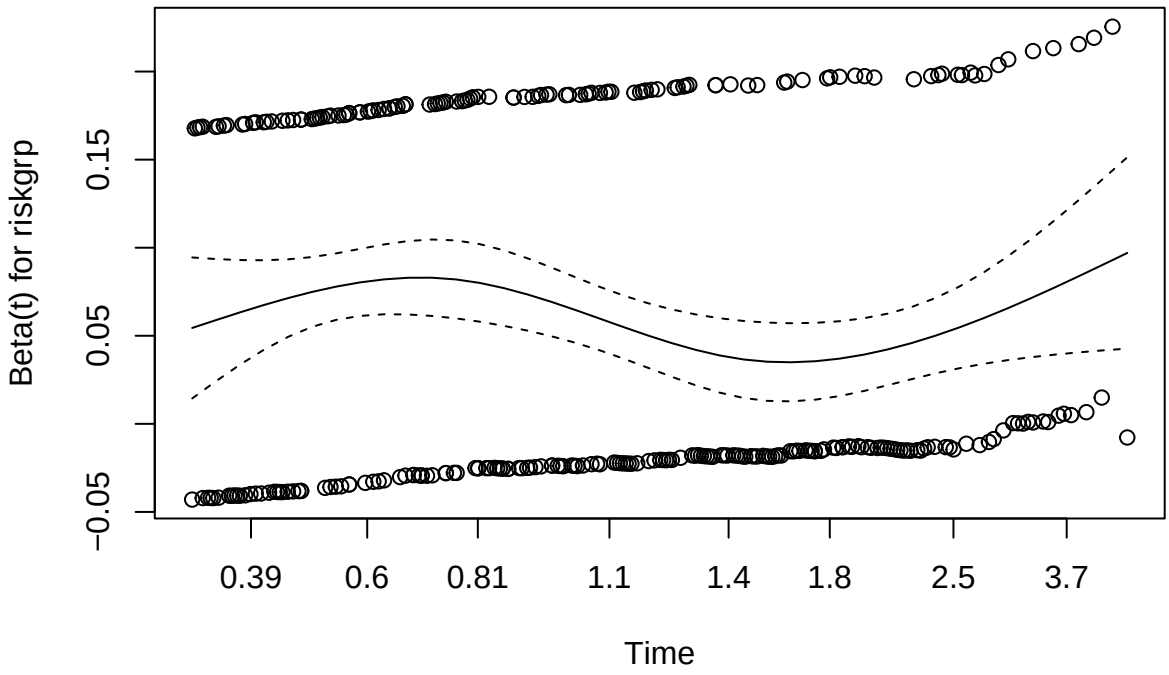
```
cph.fit.risk = coxph(Surv(yrsos, osi) ~ riskgrp,
    data = cavatica.aaml2)
summary(cph.fit.risk)
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
##
##   n= 932, number of events= 297
##    (126 observations deleted due to missingness)
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## riskgrp 0.061221  1.063134 0.005958 10.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## riskgrp     1.063     0.9406     1.051     1.076
##
## Concordance= 0.624  (se = 0.014 )
## Likelihood ratio test= 93.03  on 1 df,   p=<2e-16
## Wald test            = 105.6  on 1 df,   p=<2e-16
## Score (logrank) test = 119.2  on 1 df,   p=<2e-16
```

```
temp = cox.zph(cph.fit.risk)
temp
```

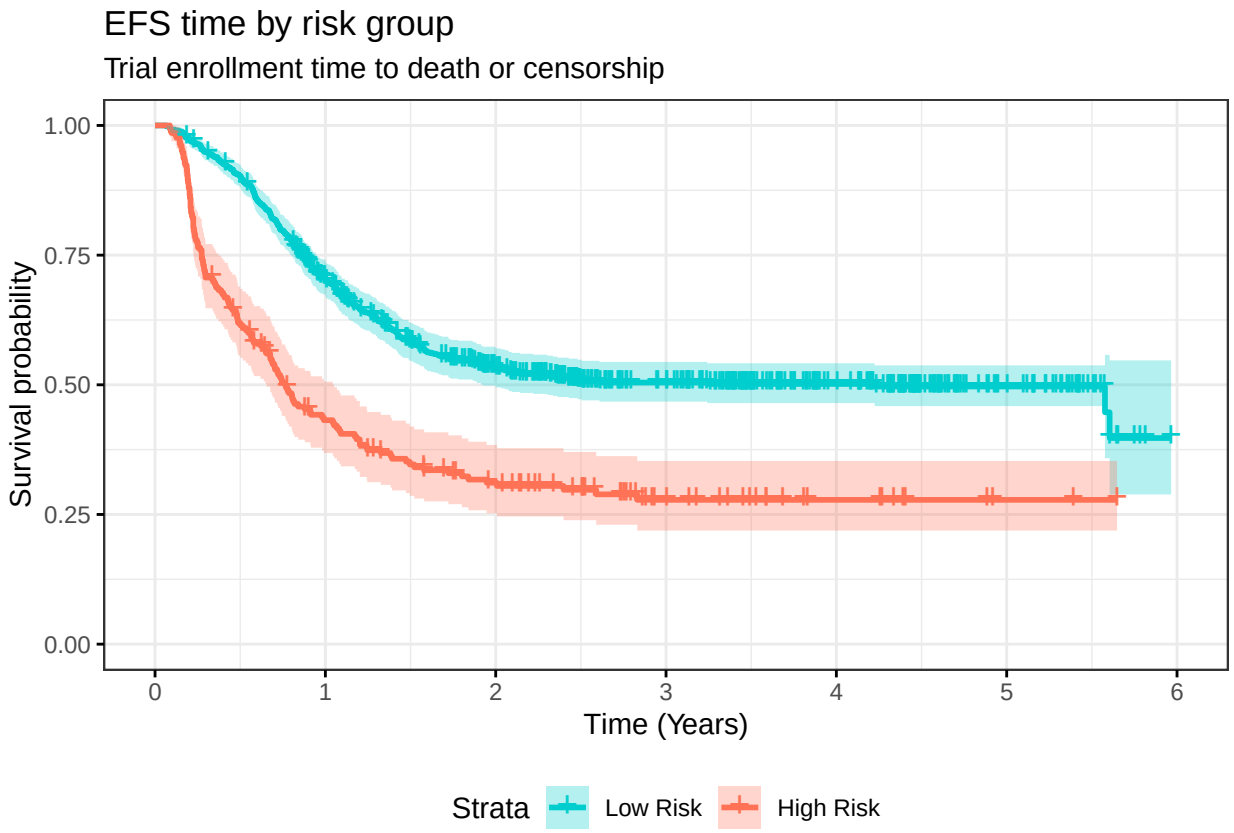
```
##         chisq df    p
## riskgrp  1.71  1 0.19
## GLOBAL   1.71  1 0.19
```

```
plot(temp)
```



Looking at EFS –> are there more types of events in patients across rislk groups?

```
fit_risk_efs = survfit(Surv(yrsefs, efsi) ~
    riskgrp, type = "kaplan-meier", data = cavatica.aaml2)
# head(summary(fit_risk_efs))

ggsurvplot(fit_risk_efs, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("Low Risk",
        "High Risk"), ggtheme = theme_bw(),
    xlab = "Time (Years)", conf.int = T,
    palette = c("cyan3", "coral1"), title = "EFS time by risk group",
    subtitle = "Trial enrollment time to death or censorship")
```




```
summary(coxph(Surv(yrsos, osi) ~ riskgrp,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ riskgrp, data = cavatica.aaml2)
##
##   n= 932, number of events= 297
##    (126 observations deleted due to missingness)
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## riskgrp 0.061221  1.063134 0.005958 10.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## riskgrp     1.063     0.9406     1.051     1.076
##
## Concordance= 0.624  (se = 0.014 )
## Likelihood ratio test= 93.03  on 1 df,   p=<2e-16
## Wald test            = 105.6  on 1 df,   p=<2e-16
## Score (logrank) test = 119.2  on 1 df,   p=<2e-16
```

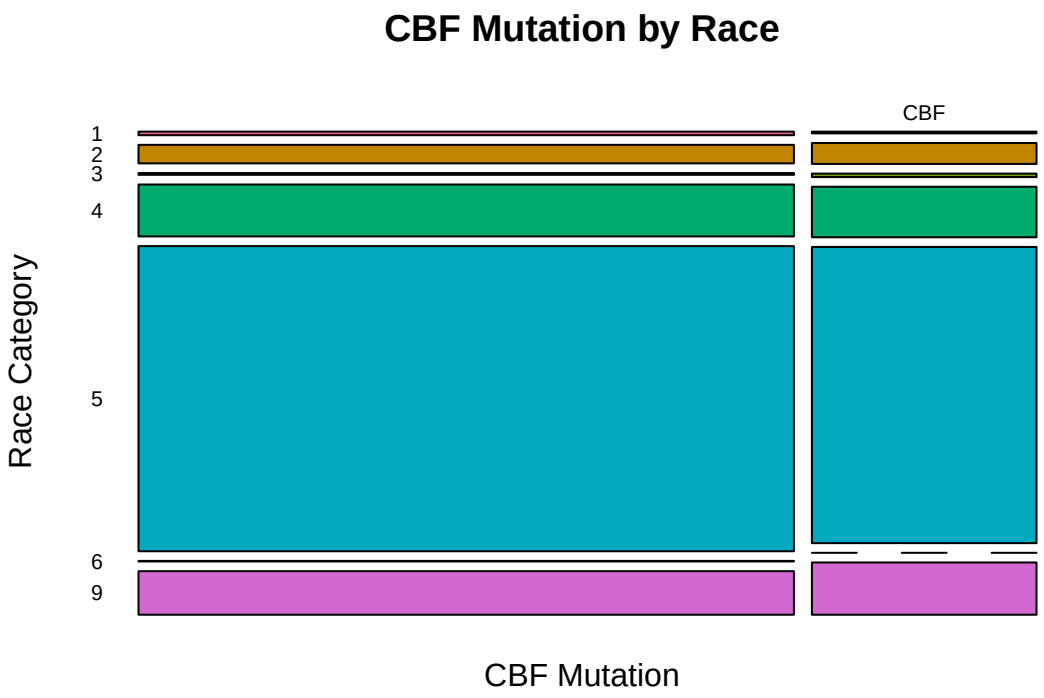
CBF

```
table(cbf)
```

```
## cbf
##     CBF
## 814 244
```

```
race.cbf = table(cbf, race)

race..cbf_mosaic = mosaicplot(race.cbf, xlab = "CBF Mutation",
    ylab = "Race Category", main = "CBF Mutation by Race",
    col = colorspace::qualitative_hcl(7),
    clegend = T, las = 1)
```

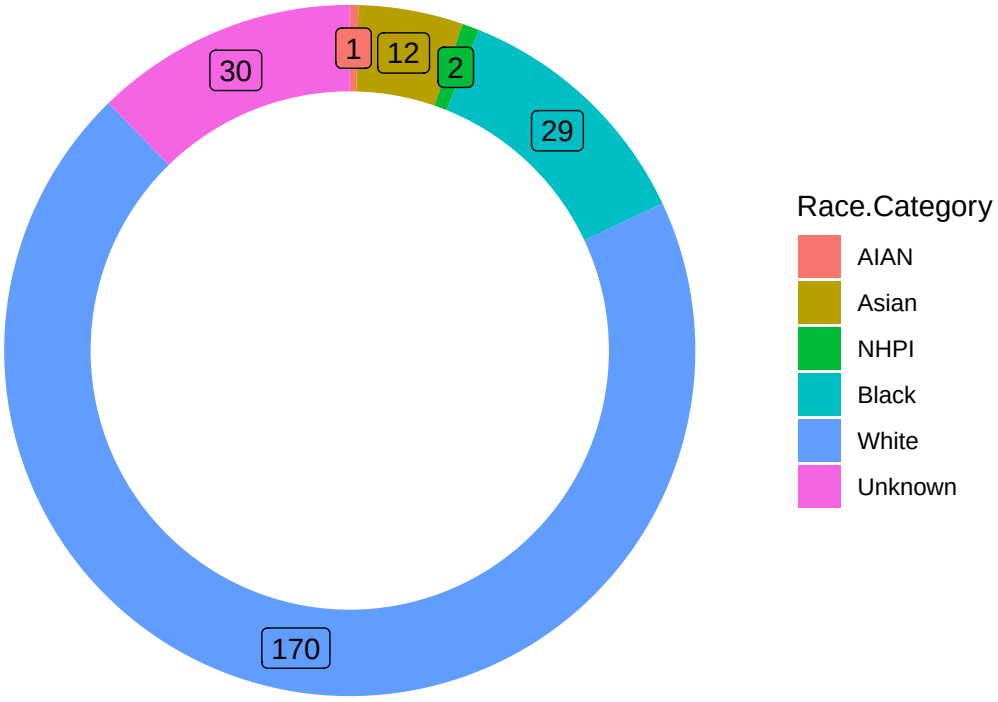
```
## Warning: In mosaicplot.default(race.cbf, xlab = "CBF Mutation", ylab = "Race Category",
##     main = "CBF Mutation by Race", col = colorspace::qualitative_hcl(7),
##     clegend = T, las = 1) :
##  extra argument 'clegend' will be disregarded
```



Filtering the matients that have a CBF mutation –> 244 patients

```
cbf.mut.pts = filter(cavatica.aaml2, cbf =="CBF")

cbf.race.table = data.frame(table(cbf.mut.pts$race_cat))

colnames(cbf.race.table)[1] = "Race.Category"
colnames(cbf.race.table)[2] = "Count"
```

```
cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

cbf.race.table$fraction = round(cbf.race.table$Count/sum(cbf.race.table$Count), 3)

cbf.race.table$ymax = cumsum(cbf.race.table$fraction)

cbf.race.table$ymin = c(0, head(cbf.race.table$ymax, n=-1))

cbf.race_donut_plot = ggplot(cbf.race.table, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Race.Categor
  geom_rect() +
  geom_label(aes(label = Count, x = 3.5, y = (ymin+ymax)/2), inherit.aes = T, show.legend = F) +
  coord_polar(theta="y") + # Try to remove that to understand how the chart is built initially
  xlim(c(0, 4)) + theme_void() + scale_fill_discrete(labels = c("AIAN", "Asian", "NHPI", "Black", "White



cbf.race_donut_plot
```



This shows that no MR patients have a CBF mutation. The proportion of white from the known are

```
cbf.race.table
```

```
##   Race.Category Count fraction  ymax  ymin
## 1             1     1    0.004 0.004 0.000
## 2             2    12    0.049 0.053 0.004
## 3             3     2    0.008 0.061 0.053
## 4             4    29    0.119 0.180 0.061
## 5             5   170    0.697 0.877 0.180
## 6             9    30    0.123 1.000 0.877
```

```
170/(1 + 12 + 2 + 29 + 170)
```

```
## [1] 0.7943925
```

Survival Time by CBF status

First, we need to create a dummy variable for CBF mutation status in the main dataframe

```
cavatica.aaml2$cbf_ind = ifelse(cavatica.aaml2$cbf_pt ==
    "CBF", 1, 0)
head(cavatica.aaml2)
```

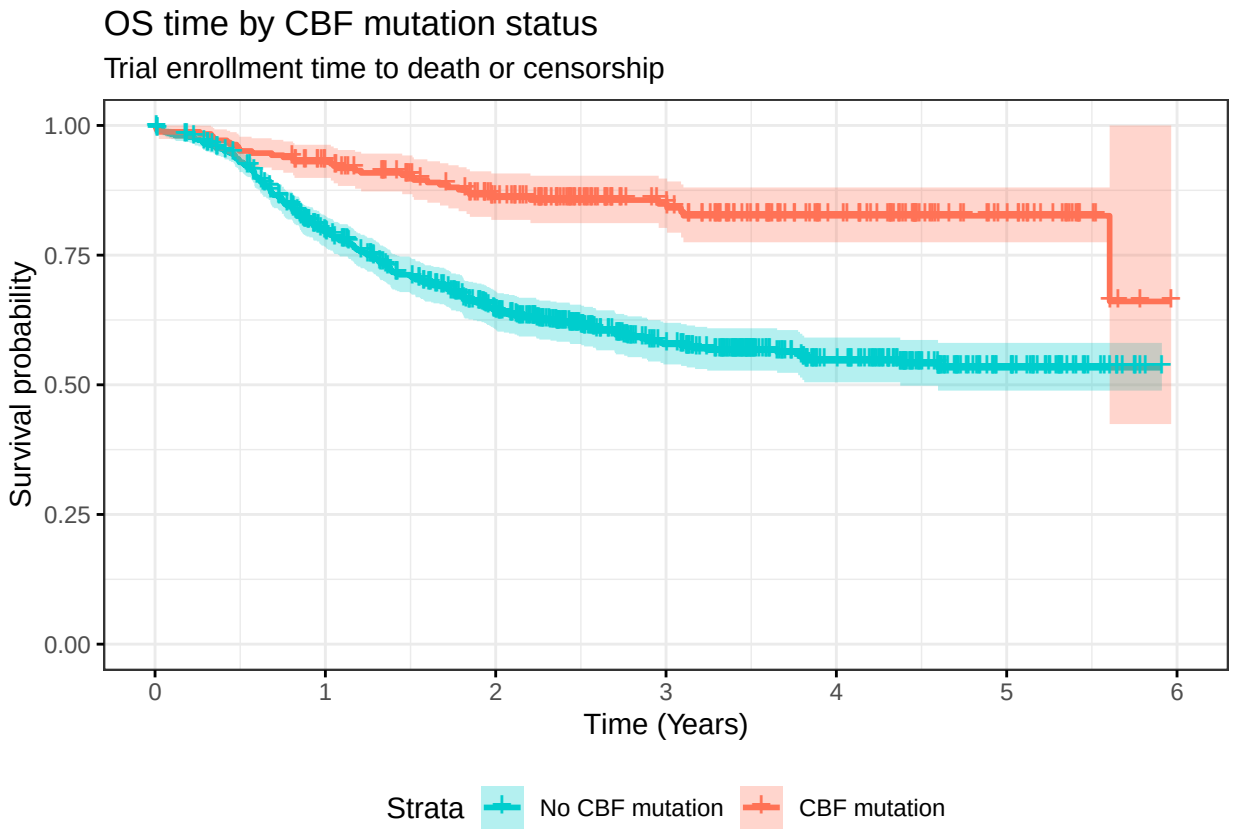
```
##      usi                                    consort_clsf trt_arm gender
## 1 PAXWMS                  01-Excluded: Post amendment 7A      NA     NA
## 2 PAXXBC                  01-Excluded: Post amendment 7A      NA     NA
## 3 PAXXCX                  01-Excluded: Post amendment 7A      NA     NA
## 4 PAYASV                  01-Excluded: Post amendment 7A      NA     NA
## 5 PAYLLP                  01-Excluded: Post amendment 7A      NA     NA
## 6 PAUIIB 03-Excluded: FLT3/ITD high AR, Arm C enrollment      10     NA
##   agecateg_ ageyr race_cat raceb ethnic_cat wbc1 wbc1_i cnsct_ noncnsex
## 1        NA    NA       NA    NA         NA   NA     NA     NA       NA
## 2        NA    NA       NA    NA         NA   NA     NA     NA       NA
## 3        NA    NA       NA    NA         NA   NA     NA     NA       NA
## 4        NA    NA       NA    NA         NA   NA     NA     NA       NA
## 5        NA    NA       NA    NA         NA   NA     NA     NA       NA
## 6        NA    NA       NA    NA         NA   NA     NA     NA       NA
##   p2offtx_ at821 ainv16 cbf_pt mll_pt amono7 cyt5q_ itdlowHAR npmstat_ cebpast_
## 1       NA    NA     NA                NA     NA         NA       NA       NA
## 2       NA    NA     NA                NA     NA         NA       NA       NA
## 3       NA    NA     NA                NA     NA         NA       NA       NA
## 4       NA    NA     NA                NA     NA         NA       NA       NA
## 5       NA    NA     NA                NA     NA         NA       NA       NA
## 6       NA    NA     NA                NA     NA         NA       NA       NA
##   mrdstat_ mrdpct_pos riskgrp who_f folup_yrs yrsefs dysefs efsi yrsos dysos
## 1       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
## 2       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
## 3       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
## 4       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
## 5       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
## 6       NA         NA      NA    NA        NA     NA     NA   NA    NA    NA
##   osi yrscir ciri yrstrm_ons dystrm_ons trmi_ons yrsdfs1 dfsi1 yrsos1 osi1
## 1  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
## 2  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
## 3  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
## 4  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
## 5  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
## 6  NA     NA   NA         NA         NA       NA      NA    NA     NA   NA
##   yrscir1 ciri1 yrsdfs2 dysdfs2 dfsi2 yrsos2 dysos2 osi2 yrstrm2 dystrm2 trmi2
## 1      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
## 2      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
## 3      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
## 4      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
```

```
## 5      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
## 6      NA    NA       NA       NA    NA     NA     NA   NA       NA       NA    NA
##   yrsrr2 dysrr2 rri2 trm_p1 trm_p2 trm_p3 trm_p4 trtarm_p1 card_hf_p1
## 1     NA     NA   NA     NA     NA     NA     NA         NA         NA
## 2     NA     NA   NA     NA     NA     NA     NA         NA         NA
## 3     NA     NA   NA     NA     NA     NA     NA         NA         NA
## 4     NA     NA   NA     NA     NA     NA     NA         NA         NA
## 5     NA     NA   NA     NA     NA     NA     NA         NA         NA
## 6     NA     NA   NA     NA     NA     NA     NA         10          0
##   card_ef_p1 card_lvsd_p1 nero_pnpn_p1 nero_seiz_p1 pulm_ards_p1 pulm_hypx_p1
## 1         NA           NA           NA           NA           NA           NA
## 2         NA           NA           NA           NA           NA           NA
## 3         NA           NA           NA           NA           NA           NA
## 4         NA           NA           NA           NA           NA           NA
## 5         NA           NA           NA           NA           NA           NA
## 6          0            0            0            0            0            0
##   pulm_rf_p1 ren_kid_p1 ren_creat_p1 p1_inf_vgs p1_inf_gnb p1_inf_fungi
## 1         NA          NA           NA         NA         NA           NA
## 2         NA          NA           NA         NA         NA           NA
## 3         NA          NA           NA         NA         NA           NA
## 4         NA          NA           NA         NA         NA           NA
## 5         NA          NA           NA         NA         NA           NA
## 6          0           0            0          0          0            0
##   p1doserd p1icud p1efract p1sfract trt_arm_p2 card_hf_p2 card_ef_p2
## 1       NA     NA       NA       NA         NA         NA         NA
## 2       NA     NA       NA       NA         NA         NA         NA
## 3       NA     NA       NA       NA         NA         NA         NA
## 4       NA     NA       NA       NA         NA         NA         NA
## 5       NA     NA       NA       NA         NA         NA         NA
## 6        0      0       71       41         NA         NA         NA
##   card_lvsd_p2 nero_pnpn_p2 nero_seiz_p2 pulm_ards_p2 pulm_hypx_p2 pulm_rf_p2
## 1           NA           NA           NA           NA           NA         NA
## 2           NA           NA           NA           NA           NA         NA
## 3           NA           NA           NA           NA           NA         NA
## 4           NA           NA           NA           NA           NA         NA
## 5           NA           NA           NA           NA           NA         NA
## 6           NA           NA           NA           NA           NA         NA
##   ren_kid_p2 ren_creat_p2 p2_inf_vgs p2_inf_gnb p2_inf_fungi p2doserd p2icud
## 1         NA           NA         NA         NA           NA       NA     NA
## 2         NA           NA         NA         NA           NA       NA     NA
## 3         NA           NA         NA         NA           NA       NA     NA
## 4         NA           NA         NA         NA           NA       NA     NA
## 5         NA           NA         NA         NA           NA       NA     NA
## 6         NA           NA         NA         NA           NA       NA     NA
##   p2efract p2sfract trt_arm_p3 card_hf_p3 card_ef_p3 card_lvsd_p3 nero_pnpn_p3
## 1       NA       NA         NA         NA         NA           NA           NA
## 2       NA       NA         NA         NA         NA           NA           NA
## 3       NA       NA         NA         NA         NA           NA           NA
## 4       NA       NA         NA         NA         NA           NA           NA
## 5       NA       NA         NA         NA         NA           NA           NA
## 6       NA       NA         NA         NA         NA           NA           NA
##   nero_seiz_p3 pulm_ards_p3 pulm_hypx_p3 pulm_rf_p3 ren_kid_p3 ren_creat_p3
## 1           NA           NA           NA         NA         NA           NA
## 2           NA           NA           NA         NA         NA           NA
```

```
## 3           NA           NA           NA         NA         NA           NA
## 4           NA           NA           NA         NA         NA           NA
## 5           NA           NA           NA         NA         NA           NA
## 6           NA           NA           NA         NA         NA           NA
##   p3_inf_vgs p3_inf_gnb p3_inf_fungi p3doserd p3icud p3efract p3sfract
## 1         NA         NA           NA       NA     NA       NA       NA
## 2         NA         NA           NA       NA     NA       NA       NA
## 3         NA         NA           NA       NA     NA       NA       NA
## 4         NA         NA           NA       NA     NA       NA       NA
## 5         NA         NA           NA       NA     NA       NA       NA
## 6         NA         NA           NA       NA     NA       NA       NA
##   trt_arm_p4 card_hf_p4 card_ef_p4 card_lvsd_p4 nero_pnpn_p4 nero_seiz_p4
## 1         NA         NA         NA           NA           NA           NA
## 2         NA         NA         NA           NA           NA           NA
## 3         NA         NA         NA           NA           NA           NA
## 4         NA         NA         NA           NA           NA           NA
## 5         NA         NA         NA           NA           NA           NA
## 6         NA         NA         NA           NA           NA           NA
##   pulm_ards_p4 pulm_hypx_p4 pulm_rf_p4 ren_kid_p4 ren_creat_p4 p4_inf_vgs
## 1           NA           NA         NA         NA           NA         NA
## 2           NA           NA         NA         NA           NA         NA
## 3           NA           NA         NA         NA           NA         NA
## 4           NA           NA         NA         NA           NA         NA
## 5           NA           NA         NA         NA           NA         NA
## 6           NA           NA         NA         NA           NA         NA
##   p4_inf_gnb p4_inf_fungi p4doserd p4icud p4efract p4sfract rem_response
## 1         NA           NA       NA     NA       NA       NA           NA
## 2         NA           NA       NA     NA       NA       NA           NA
## 3         NA           NA       NA     NA       NA       NA           NA
## 4         NA           NA       NA     NA       NA       NA           NA
## 5         NA           NA       NA     NA       NA       NA           NA
## 6         NA           NA       NA     NA       NA       NA           NA
##   rcv_bortz_notif path_rev cyto_rev cbf_ind
## 1                                         0
## 2                                         0
## 3                                         0
## 4                                         0
## 5                                         0
## 6                                         0
```

Now that we have an indicator variable for cbf status, we can create the survival plot against this feature

```
fit_cbf = survfit(Surv(yrsos, osi) ~ cbf_ind,
    type = "kaplan-meier", data = cavatica.aaml2)


# head(summary(fit_cbf))

ggsurvplot(fit_cbf, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("No CBF mutation",
        "CBF mutation"), ggtheme = theme_bw(),
    xlab = "Time (Years)", conf.int = T,
    palette = c("cyan3", "coral1"), title = "OS time by CBF mutation status",
    subtitle = "Trial enrollment time to death or censorship")
```

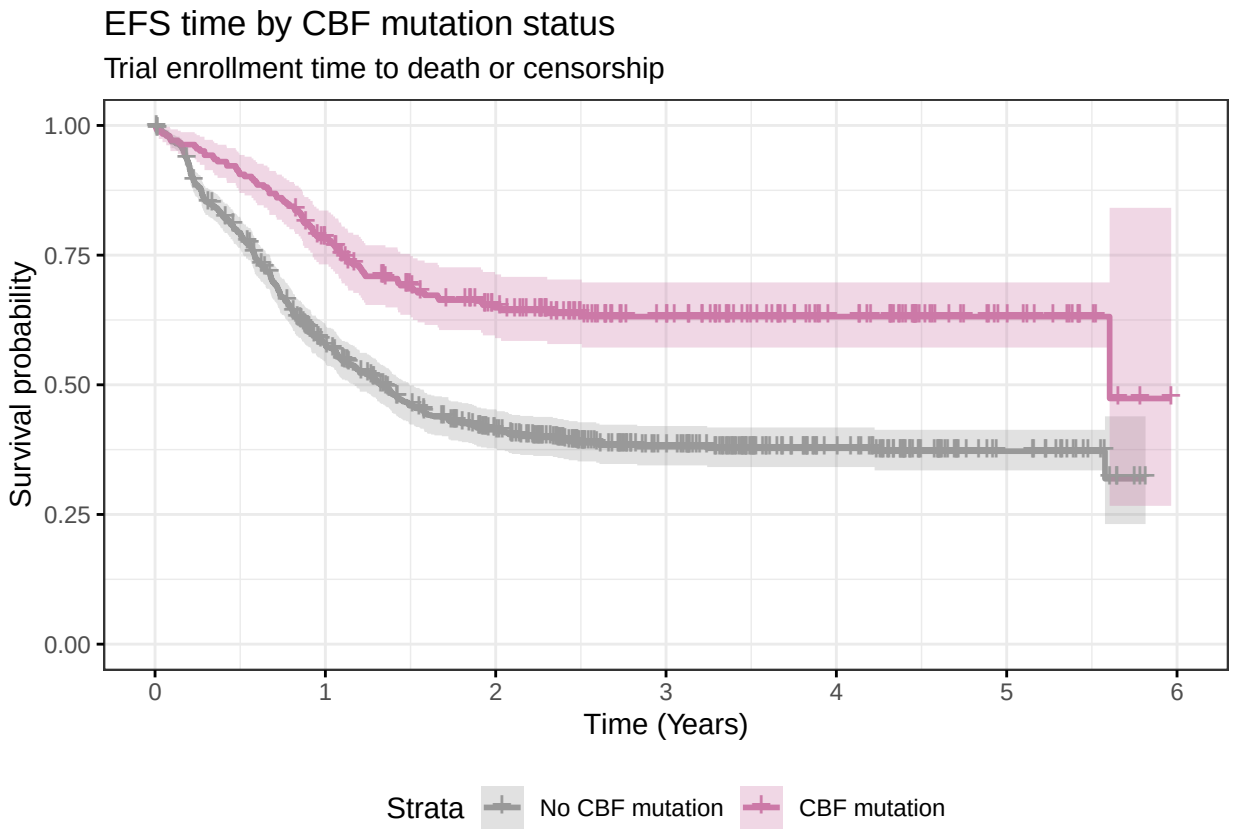


Per findings by Rau et al, CBf mutations seem to predict better survival (I'm looking at OS, but they compare EFS).

Let's look at EFS to make sure.

```
library(wesanderson)
```

```
fit_cbf_efs = survfit(Surv(yrsefs, efsi) ~
    cbf_ind, type = "kaplan-meier", data = cavatica.aaml2)


# head(summary(fit_cbf_efs))

ggsurvplot(fit_cbf_efs, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("No CBF mutation",
        "CBF mutation"), ggtheme = theme_bw(),
    xlab = "Time (Years)", conf.int = T,
    palette = c("#999999", "#CC79A7"), title = "EFS time by CBF mutation status",
    subtitle = "Trial enrollment time to death or censorship")
```



Performing statistical tests to see if there is a significant difference between the curves https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html

- We can conduct between-group significance tests using a log-rank test
- The log-rank test equally weights observations over the entire follow-up time and is the most common way to compare survival times between groups
- There are versions that more heavily weight the early or late follow-up that could be more appropriate depending on the research question (see ?survdiff for different test options)

We get the log-rank p-value using the survdiff function.

```
sd_cbf_efs = survdiff(Surv(yrsefs, efsi) ~
    cbf_ind, data = cavatica.aaml2)
sd_cbf_efs
```

```
## Call:
## survdiff(formula = Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
##
## n=956, 102 observations deleted due to missingness.
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## cbf_ind=0 712      418      351      12.8      42.2
## cbf_ind=1 244       87      154      29.1      42.2
##
##  Chisq= 42.2  on 1 degrees of freedom, p= 8e-11
```

extracting the actual p-value

```
1 - pchisq(sd_cbf_efs$chisq, length(sd_cbf_efs$n) -
    1)
```

```
## [1] 8.383882e-11
```

Cox regression modelling

We may want to quantify an effect size for a single variable, or include more than one variable into a regression model to account for the effects of multiple variables.

The Cox regression model is a semi-parametric model that can be used to fit univariable and multivariable regression models that have survival outcomes.

h(t|Xi)=h0(t)exp(B1Xi1+. . . +BpXip)

h(t): hazard, or the instantaneous rate at which events occur h0(t): underlying baseline hazard

Some key assumptions of the model:

- non-informative censoring
- proportional hazards

Note: parametric regression models for survival outcomes are also available, but they are not addressed in this training

We can fit regression models for survival data using the coxph function, which takes a Surv object on the left hand side and has standard syntax for regression formulas in R on the right hand side.

```
coxph(Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
##
##            coef exp(coef) se(coef)      z        p
## cbf_ind -0.7490    0.4729   0.1180 -6.346 2.22e-10
##
## Likelihood ratio test=46.92  on 1 df, p=7.41e-12
## n= 956, number of events= 505
##    (102 observations deleted due to missingness)
```

```
summary(coxph(Surv(yrsefs, efsi) ~ cbf_ind,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
##
##   n= 956, number of events= 505
##    (102 observations deleted due to missingness)
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## cbf_ind -0.7490    0.4729   0.1180 -6.346 2.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##         exp(coef) exp(-coef) lower .95 upper .95
## cbf_ind    0.4729      2.115    0.3752    0.5959
##
## Concordance= 0.567  (se = 0.01 )
## Likelihood ratio test= 46.92  on 1 df,   p=7e-12
## Wald test            = 40.27  on 1 df,   p=2e-10
## Score (logrank) test = 42.17  on 1 df,   p=8e-11
```

Let's visualize in nicer table?

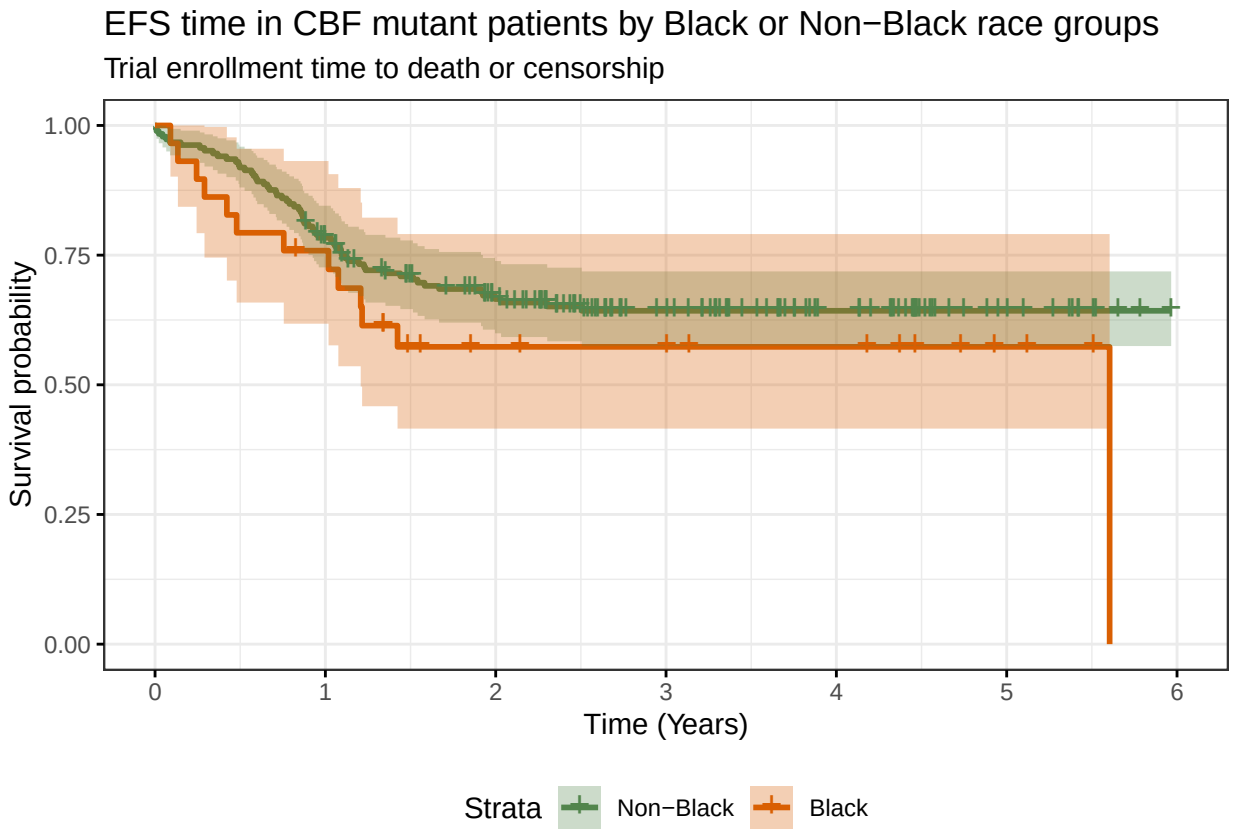
```
# coxph(Surv(yrsefs, efsi) ~ cbf_ind,
# data = cavatica.aaml2) %>%
# gtsummary::tbl_regression(exp = TRUE)
```

```
survdiff(Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
```

```
## Call:
## survdiff(formula = Surv(yrsefs, efsi) ~ cbf_ind, data = cavatica.aaml2)
##
## n=956, 102 observations deleted due to missingness.
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## cbf_ind=0 712      418      351      12.8      42.2
## cbf_ind=1 244       87      154      29.1      42.2
##
##  Chisq= 42.2  on 1 degrees of freedom, p= 8e-11
```

Similar to Rau et al, I'd like to take a look patients with CBF mutations to see if survival differs between Black and non-Black patients

```
fit_cbf.only_efs = survfit(Surv(yrsefs, efsi) ~
    raceb, type = "kaplan-meier", data = cbf.mut.pts)


# head(summary(fit_cbf.only_efs))

ggsurvplot(fit_cbf.only_efs, data = cbf.mut.pts,
    legend = "bottom", legend.labs = c("Non-Black",
        "Black"), ggtheme = theme_bw(), xlab = "Time (Years)",
    conf.int = T, palette = c("#52854C",
        "#D95F02"), title = "EFS time in CBF mutant patients by Black or Non-Black race groups",
    subtitle = "Trial enrollment time to death or censorship")
```




```
nrow(cbf.mut.pts)
```

```
## [1] 244
```

Statistical test on this subset

```
summary(coxph(Surv(yrsefs, efsi) ~ raceb,
    data = cbf.mut.pts))
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ raceb, data = cbf.mut.pts)
##
##   n= 214, number of events= 76
##    (30 observations deleted due to missingness)
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## raceb 0.3852    1.4699   0.3052 1.262    0.207
##
##       exp(coef) exp(-coef) lower .95 upper .95
## raceb      1.47     0.6803    0.8082     2.674
##
## Concordance= 0.522  (se = 0.022 )
## Likelihood ratio test= 1.46  on 1 df,   p=0.2
## Wald test            = 1.59  on 1 df,   p=0.2
## Score (logrank) test = 1.61  on 1 df,   p=0.2
```

```
survdiff(Surv(yrsefs, efsi) ~ raceb, data = cbf.mut.pts)
```

```
## Call:
## survdiff(formula = Surv(yrsefs, efsi) ~ raceb, data = cbf.mut.pts)
##
## n=214, 30 observations deleted due to missingness.
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## raceb=0 185       63    66.63     0.198      1.61
## raceb=1  29       13     9.37     1.409      1.61
##
##  Chisq= 1.6  on 1 degrees of freedom, p= 0.2
```

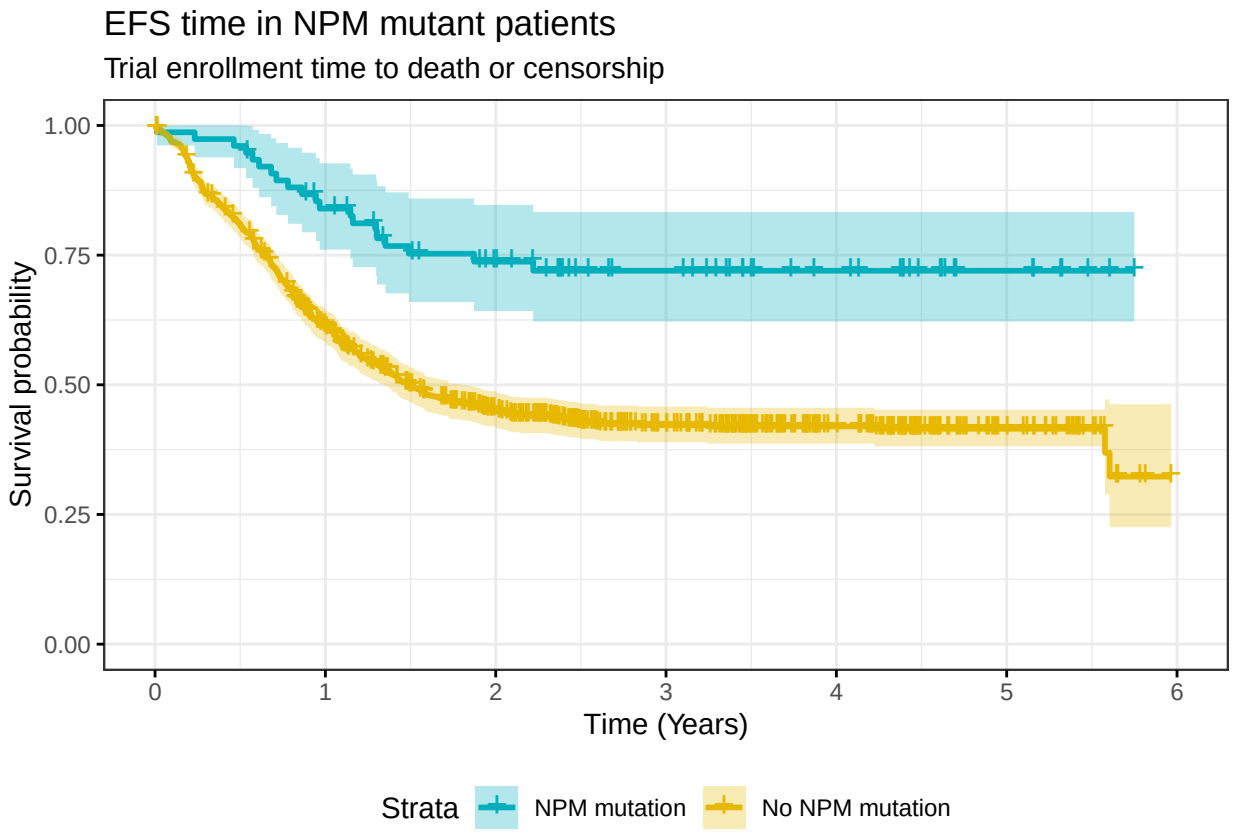
With this smaller group of patients, it does not appear that there is a significant difference of EFS between Black and Non-Black patients

NPM

```
table(cavatica.aaml2$npmstat_)
```

```
##
##   1   2
##  76 880
```

```
npm.mut.pts = filter(cavatica.aaml2, cavatica.aaml2$npmstat_ ==
    1)

fit_npm_efs = survfit(Surv(yrsefs, efsi) ~
    npmstat_, type = "kaplan-meier", data = cavatica.aaml2)


# head(summary(fit_npm_efs))

ggsurvplot(fit_npm_efs, data = cavatica.aaml2,
    legend = "bottom", legend.labs = c("NPM mutation",
        "No NPM mutation"), ggtheme = theme_bw(),
    xlab = "Time (Years)", conf.int = T,
    palette = c("#00AFBB", "#E7B800"), title = "EFS time in NPM mutant patients",
    subtitle = "Trial enrollment time to death or censorship")
```




```
summary(coxph(Surv(yrsefs, efsi) ~ npmstat_,
    data = cavatica.aaml2))
```

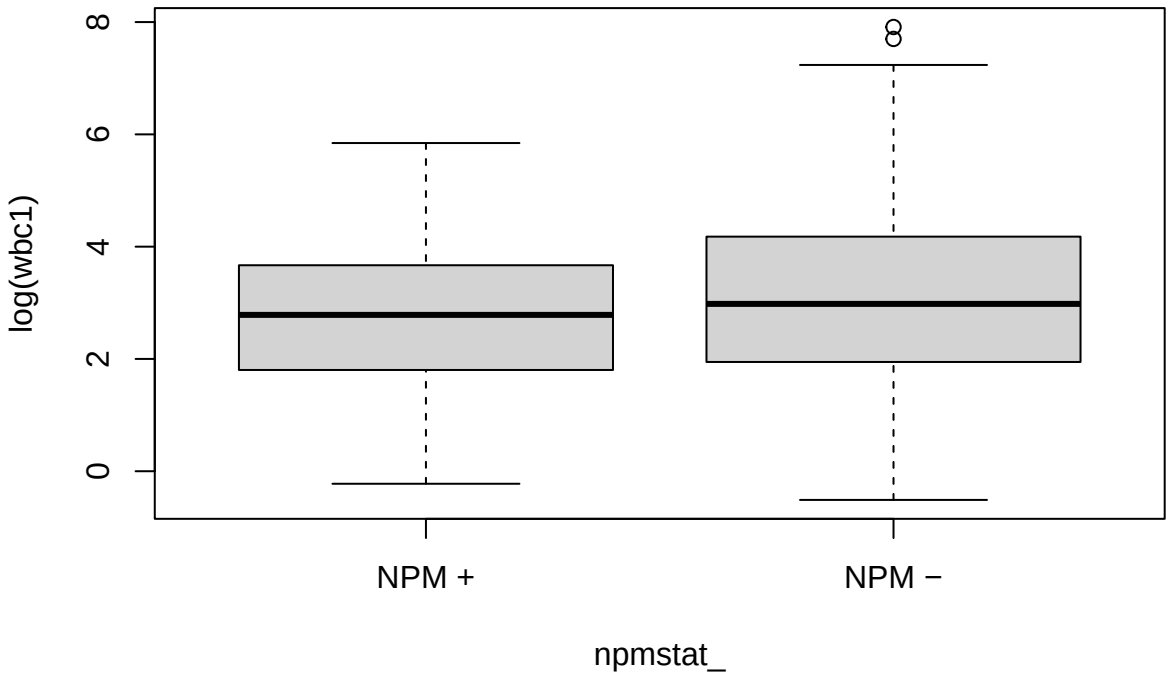
```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ npmstat_, data = cavatica.aaml2)
##
##   n= 956, number of events= 505
##    (102 observations deleted due to missingness)
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## npmstat_ 0.9991    2.7157   0.2283 4.376 1.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## npmstat_     2.716     0.3682     1.736     4.248
##
## Concordance= 0.53  (se = 0.006 )
## Likelihood ratio test= 26.39  on 1 df,   p=3e-07
## Wald test            = 19.15  on 1 df,   p=1e-05
## Score (logrank) test = 20.8  on 1 df,   p=5e-06
```

This paper on the 1031 trial (https://chop365-my.sharepoint.com/personal/adamsj8_chop_edu/Documents/COG-pediatric-AML/literature/AML_Pediatric_Sorafenib_AAML1031_Pollard_JCO_2022.pdf) reports that co-ocurring mutations of FLT3 and NPM seem to improve survival. Does NPM+ mutation improve survival of those with FLT mutations?

```
# cavatica.aaml2$npm.flt =
# ifelse(cavatica.aaml2$npmstat_ == 1,
# 1, (ifelse(cavatica.aaml2$itdlowHAR
# ==)))


# cavatica.aaml2$npm.flt = NA for(i in
# length(cavatica.aaml2)) {
# if(cavatica.aaml2$npmstat_ == 1 &&
# cavatica.aaml2$itdlowHAR ==
# 2){cavatica.aaml2$npm.flt = 0} else
# if(cavatica.aaml2$npmstat_ == 2 &&
# cavatica.aaml2$itdlowHAR ==
# 1){cavatica.aaml2$npm.flt = 1} else
# {cavatica.aaml2$npm.flt = 2} }

# cavatica.aaml2 %>% mutate(npm.flt =
# case_when( npmstat_ == 1 && itdlowHAR
# == 2 ~ 0, npmstat_ == 2 && itdlowHAR
# == 1 ~ 1, npmstat_ == 1 && itdlowHAR
# == 1 ~ 2, npmstat_ == 2 && itdlowHAR
# == 2 ~ 0, ))
# head(cavatica.aaml2$npm.flt)
```

This paper on the 1031 trial (https://chop365-my.sharepoint.com/personal/adamsj8_chop_edu/Documents/COG-pediatric-AML/literature/AML_Pediatric_Sorafenib_AAML1031_Pollard_JCO_2022.pdf) reports NPM mutations being associated with higher WBC, is this true? –> not exactly?

```
# library(vioplot)
boxplot(data = cavatica.aaml2, log(wbc1) ~
    npmstat_, names = c("NPM +", "NPM -"))
```

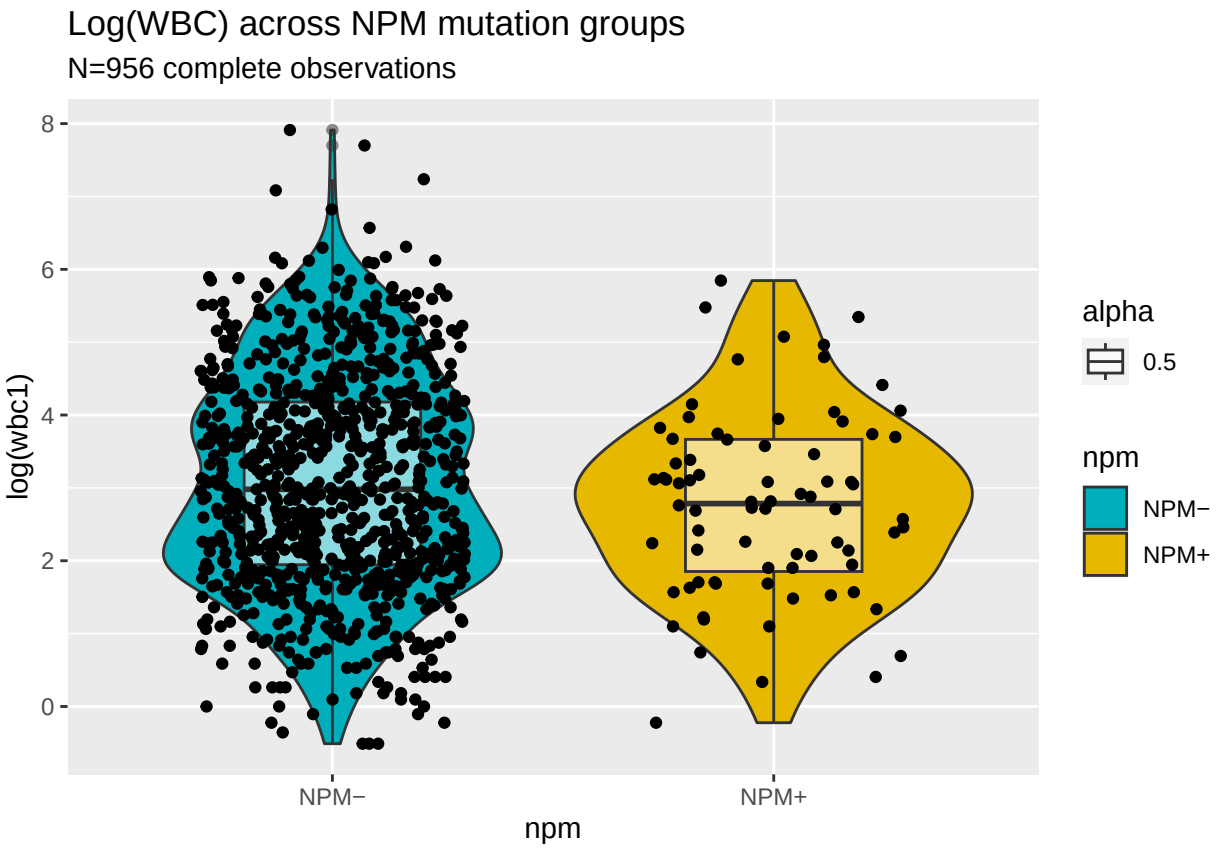


```
cavatica.aaml2$npm = ifelse(cavatica.aaml2$npmstat_ ==
    1, "NPM+", "NPM-")

ggplot(cavatica.aaml2 %>%
    filter(!is.na(npm)), aes(x = npm, y = log(wbc1))) +
    geom_violin(aes(fill = npm)) + geom_boxplot(aes(alpha = 0.5),
    width = 0.4) + geom_jitter(width = 0.3) +
    scale_fill_manual(values = c("#00AFBB",
        "#E7B800")) + ggtitle("Log(WBC) across NPM mutation groups",
    subtitle = "N=956 complete observations")
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

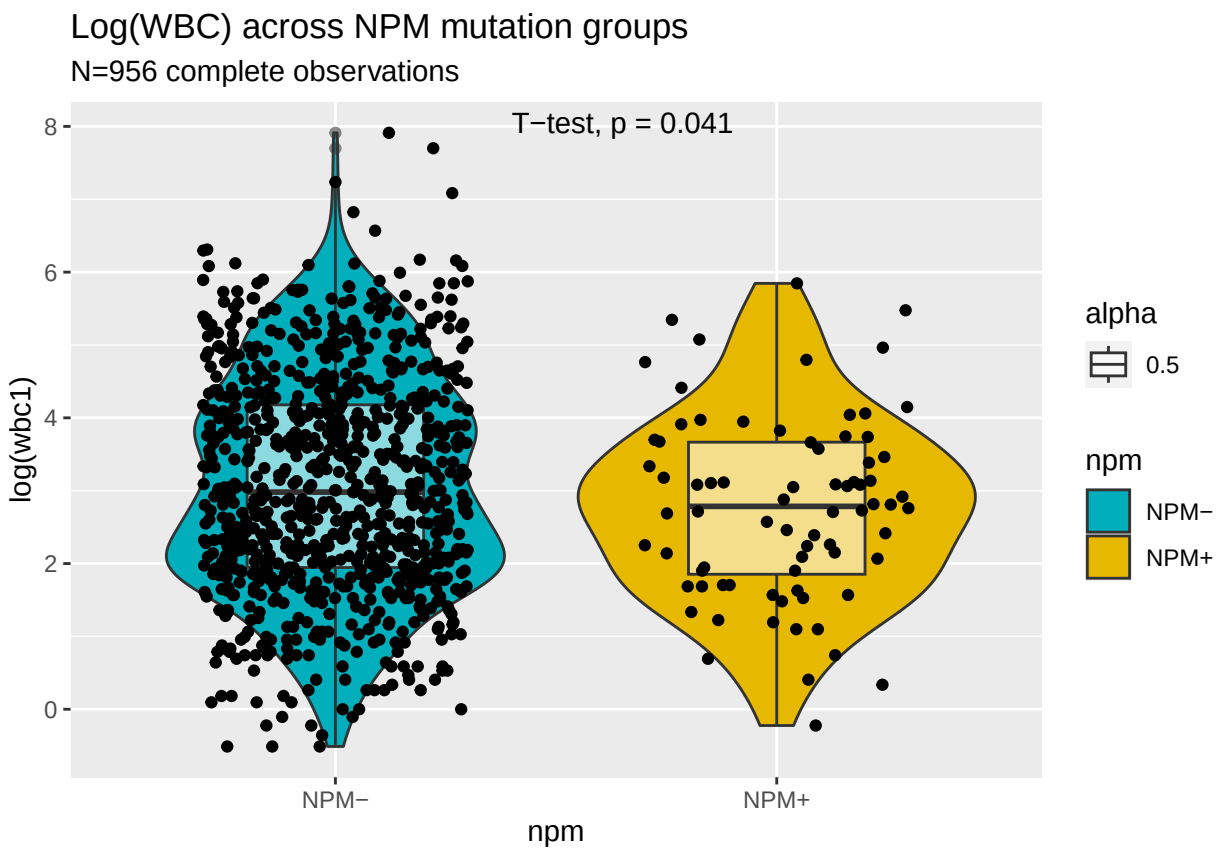



```
# ggplot(cavatica.aaml2, aes(x=npm,
# y=wbc1)) + geom_dotplot(binaxis='y',
# stackdir='center', dotsize = 1)
```

```
ggplot(cavatica.aaml2 %>%
    filter(!is.na(npm)), aes(x = npm, y = log(wbc1))) +
    geom_violin(aes(fill = npm)) + geom_boxplot(aes(alpha = 0.5),
```

```
    width = 0.4) + geom_jitter(width = 0.3) +
    scale_fill_manual(values = c("#00AFBB",
        "#E7B800")) + ggtitle("Log(WBC) across NPM mutation groups",
    subtitle = "N=956 complete observations") +
    stat_compare_means(method = "t.test",
        label.x = 1.5)
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_ydensity()`).

## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).

## Warning: Removed 1 rows containing non-finite values (`stat_compare_means()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

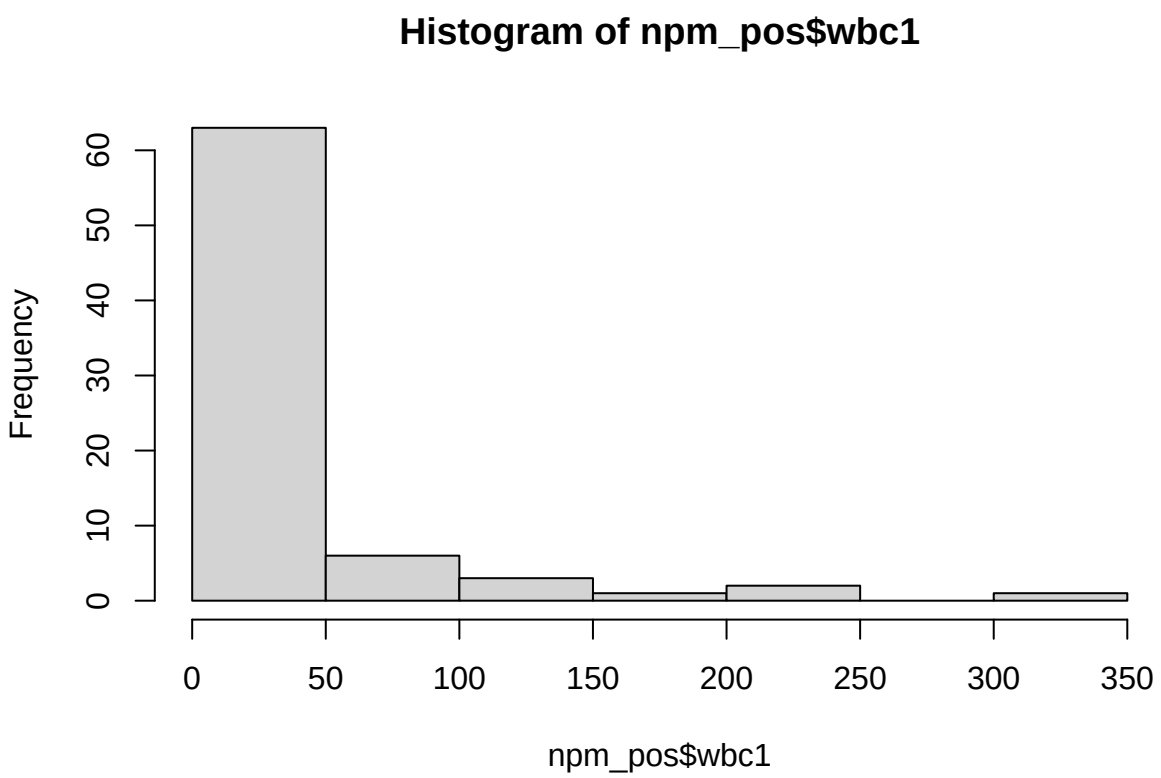



```
table(cavatica.aaml2$npm)
```

```
##
## NPM- NPM+
##  880   76
```

```
npm_pos = filter(cavatica.aaml2, cavatica.aaml2$npm ==
    "NPM+")
npm_neg = filter(cavatica.aaml2, cavatica.aaml2$npm ==
    "NPM-")
```

```
hist(npm_pos$wbc1)
```




```
mean(npm_pos$wbc1)
```

```
## [1] 35.42105
```

```
mean(npm_neg$wbc1)
```

```
## [1] NA
```

```
t.test(npm_neg$wbc1, npm_pos$wbc1, paired = F,
    var.equal = F, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  npm_neg$wbc1 and npm_pos$wbc1
## t = 3.6096, df = 198.93, p-value = 0.9998
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 44.72479
## sample estimates:
## mean of x mean of y
##  66.10034  35.42105
```

```
t.test(npm_neg$wbc1, npm_pos$wbc1, paired = F,
    var.equal = F, alternative = "greater")
```

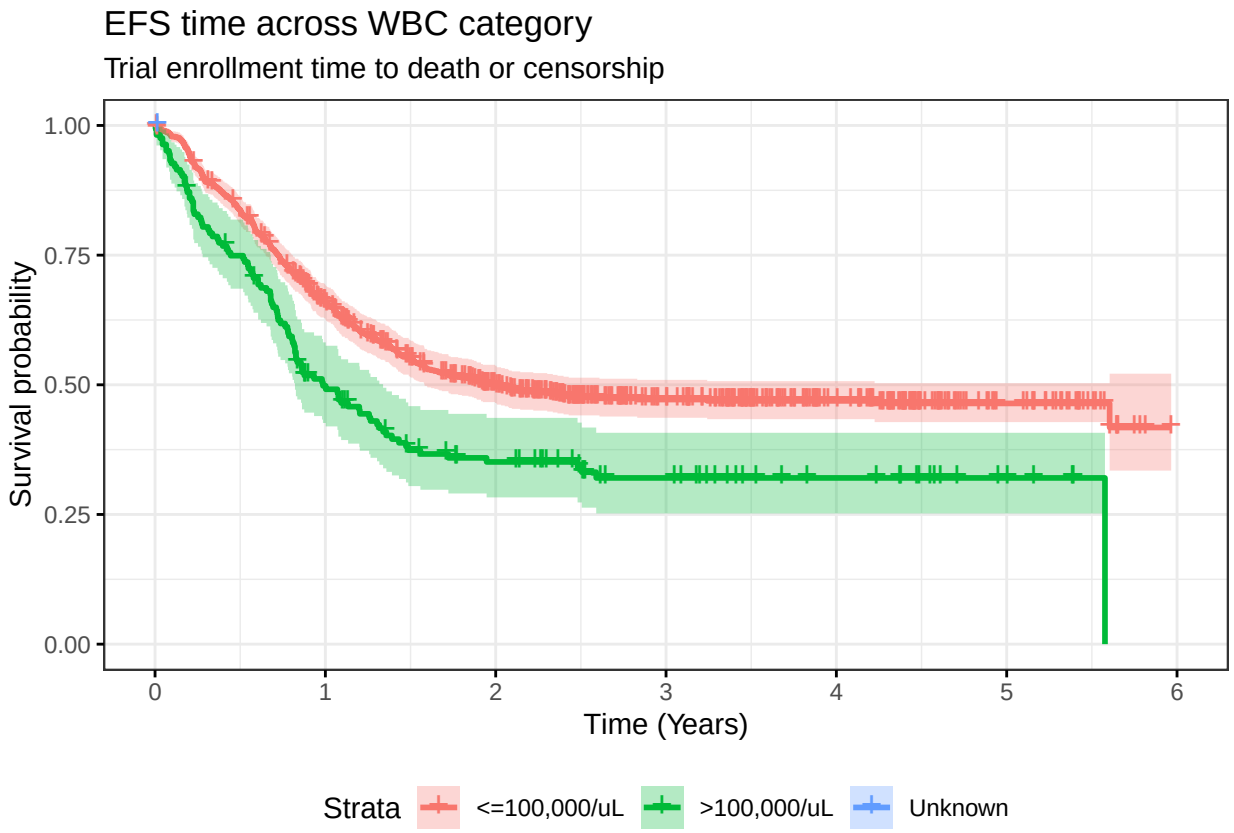
```
##
##  Welch Two Sample t-test
##
## data:  npm_neg$wbc1 and npm_pos$wbc1
## t = 3.6096, df = 198.93, p-value = 0.000194
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  16.63379      Inf
## sample estimates:
## mean of x mean of y
##  66.10034  35.42105
```

Leukemic Burden –> WBC groups

```
table(cavatica.aaml2$wbc1_i)
```

```
##
##   1   2   3
## 791 164   1
```

```
fit_wbc_efs = survfit(Surv(yrsefs, efsi) ~ wbc1_i,
               type = "kaplan-meier",
               data = cavatica.aaml2)


#head(summary(fit_wbc_efs))

ggsurvplot(fit_wbc_efs, data = cavatica.aaml2,
           legend="bottom",
           legend.labs = c("<=100,000/uL", ">100,000/uL", "Unknown"),
           ggtheme = theme_bw(),
           xlab = "Time (Years)",
           conf.int = T,
           # palette = c("cyan3", "coral1"),
           title = "EFS time across WBC category", subtitle = "Trial enrollment time to death or censors
```



High white blood cell count at presentation is an unfavorable prognostic factor for treatment outcome (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166101/). In this case, this higher burden seems to result in decrease median survival between these groups.

```
survdiff(Surv(yrsefs, efsi) ~ wbc1_i, data = cavatica.aaml2)
```

```
## Call:
## survdiff(formula = Surv(yrsefs, efsi) ~ wbc1_i, data = cavatica.aaml2)
##
## n=956, 102 observations deleted due to missingness.
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## wbc1_i=1 791      399 4.32e+02   2.58853  18.08102
## wbc1_i=2 164      106 7.25e+01  15.43981  18.09037
## wbc1_i=3   1        0 7.34e-03   0.00734   0.00736
##
##  Chisq= 18.1  on 2 degrees of freedom, p= 1e-04
```

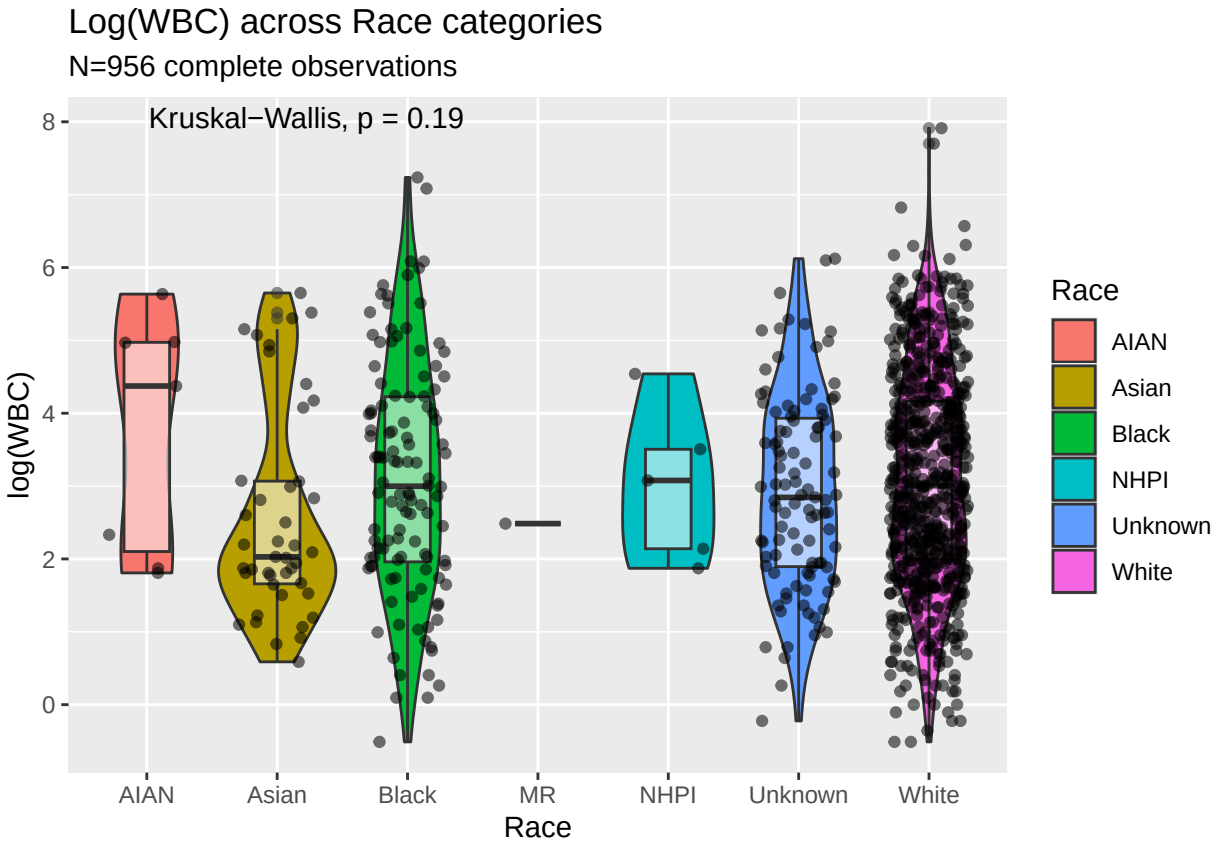
This seems to be supported by a significant p-value between the groups.

Leukemic burden across racial groups

```
cavatica.aaml2 = cavatica.aaml2 %>%
    mutate(race_str = case_when(race_cat ==
        1 ~ "AIAN", race_cat == 2 ~ "Asian",
        race_cat == 3 ~ "NHPI", race_cat ==
            4 ~ "Black", race_cat == 5 ~
```

```
            "White", race_cat == 6 ~ "MR",
        race_cat == 9 ~ "Unknown"))

ggplot(cavatica.aaml2 %>%
    filter(!is.na(race_str)), aes(x = race_str,
    y = log(wbc1))) + geom_violin(aes(fill = race_str)) +
    geom_boxplot(aes(alpha = 0.4), width = 0.35) +
    geom_jitter(width = 0.3, aes(alpha = 0.4)) +
    # scale_fill_manual(values =
    # c('#00AFBB', '#E7B800')) +
ggtitle("Log(WBC) across Race categories",
    subtitle = "N=956 complete observations") +
    stat_compare_means(method = "kruskal.test",
        label.x = 1.5) + guides(alpha = F) +
    scale_fill_discrete(name = "Race") +
    xlab("Race") + ylab("log(WBC)")
```

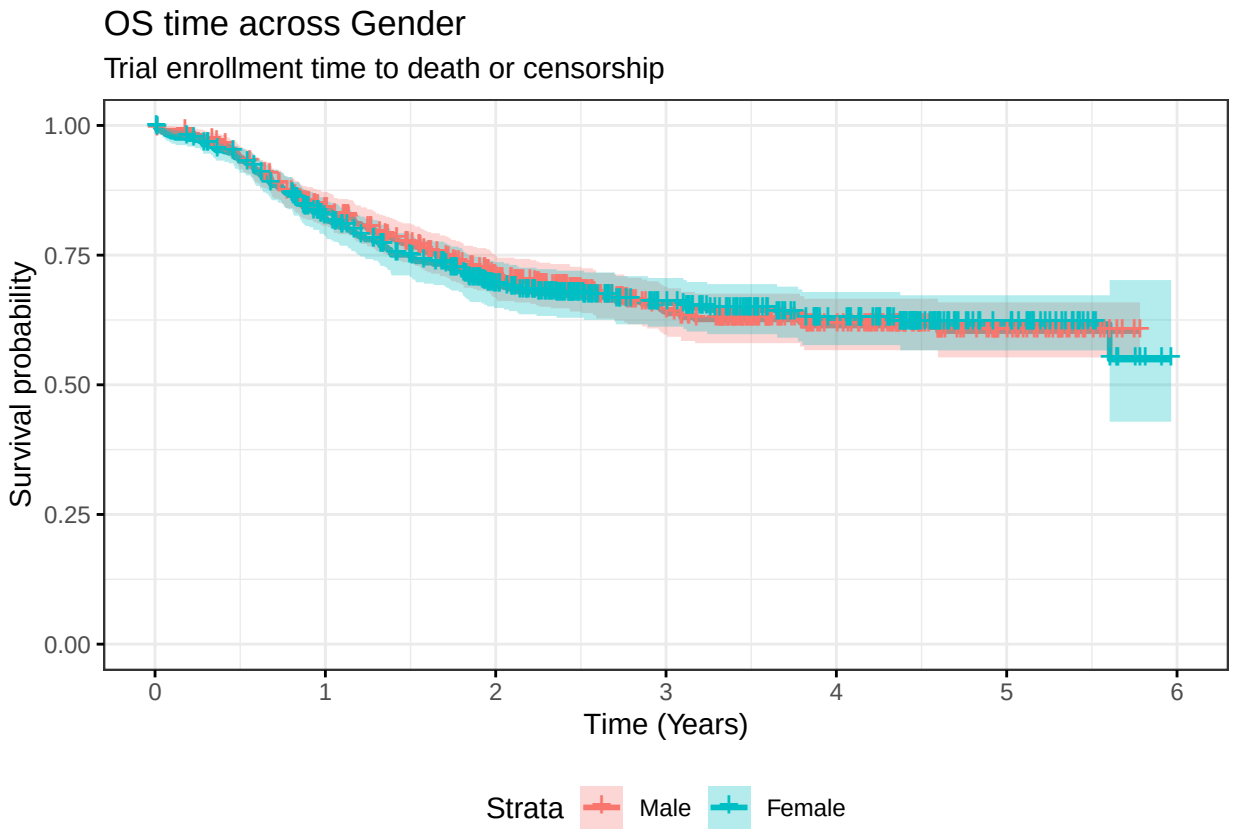
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.

## Warning: Removed 1 rows containing non-finite values (`stat_ydensity()`).

## Warning: Groups with fewer than two data points have been dropped.

## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).

## Warning: Removed 1 rows containing non-finite values (`stat_compare_means()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



I decided to stick with the Kruskal-Wallis test because it's non-parametrics and does not assume homoscedasticity within the data. It assumes that the groups come from the same distribution.

Gender

```
fit_gender_os = survfit(Surv(yrsos, osi) ~ gender,
                type = "kaplan-meier",
                data = cavatica.aaml2)


#head(summary(fit_gender_os))

ggsurvplot(fit_gender_os, data = cavatica.aaml2,
           legend="bottom",
           legend.labs = c("Male", "Female"),
           ggtheme = theme_bw(),
           xlab = "Time (Years)",
           conf.int = T,
           # palette = c("cyan3", "coral1"),
           title = "OS time across Gender", subtitle = "Trial enrollment time to death or censorship")
```



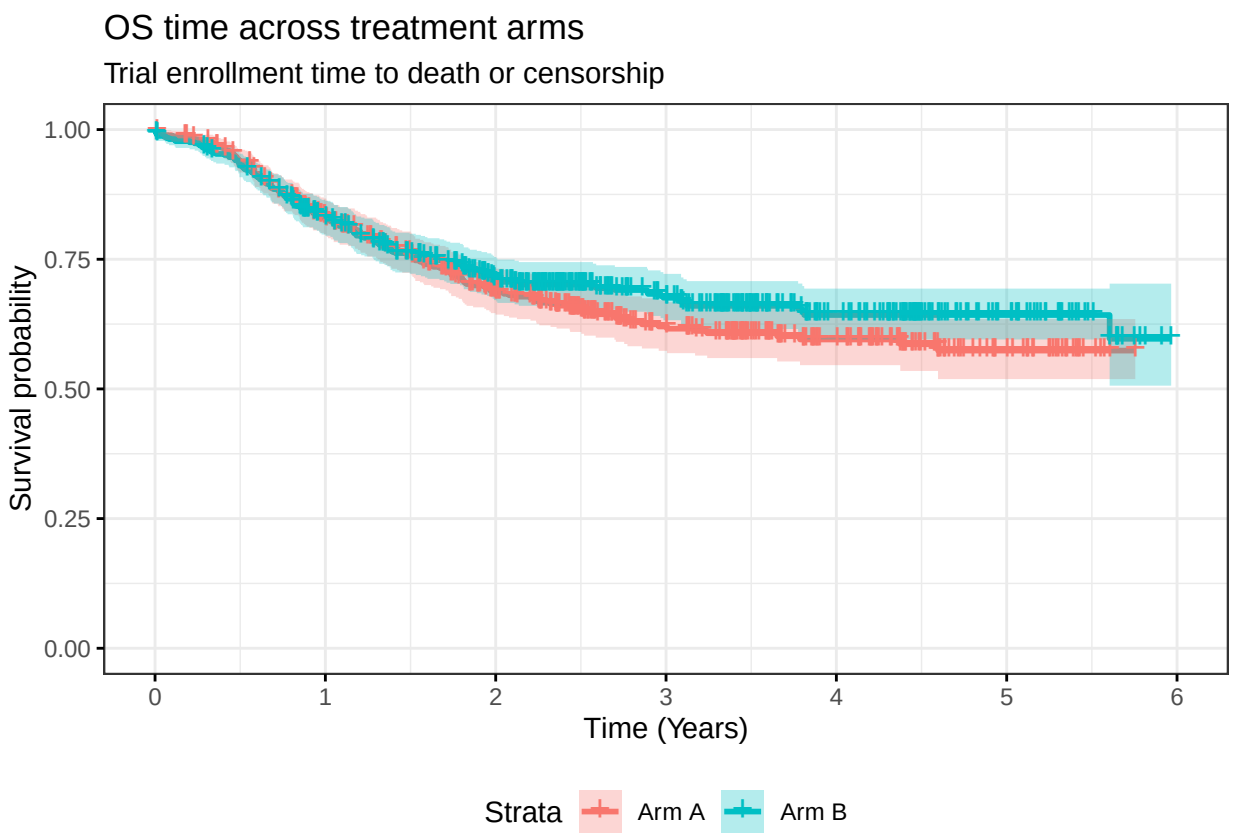

```
summary(coxph(Surv(yrsos, osi) ~ gender,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ gender, data = cavatica.aaml2)
##
##   n= 956, number of events= 315
##    (102 observations deleted due to missingness)
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## gender 0.01652   1.01665  0.11283 0.146    0.884
##
##        exp(coef) exp(-coef) lower .95 upper .95
## gender     1.017     0.9836    0.8149     1.268
##
## Concordance= 0.506  (se = 0.015 )
## Likelihood ratio test= 0.02  on 1 df,   p=0.9
## Wald test            = 0.02  on 1 df,   p=0.9
## Score (logrank) test = 0.02  on 1 df,   p=0.9
```

There is not enough evidnece to support a significant difference in overall survival across gender groups.

Treatment Arm

```
fit_trtArm_os = survfit(Surv(yrsos, osi) ~ trt_arm,
                type = "kaplan-meier",
                data = cavatica.aaml2)


#head(summary(fit_trtArm_os))

ggsurvplot(fit_trtArm_os, data = cavatica.aaml2,
           legend="bottom",
           legend.labs = c("Arm A", "Arm B"),
           ggtheme = theme_bw(),
           xlab = "Time (Years)",
           conf.int = T,
           # palette = c("cyan3", "coral1"),
           title = "OS time across treatment arms", subtitle = "Trial enrollment time to death or censo
```




```
summary(coxph(Surv(yrsos, osi) ~ trt_arm,
    data = cavatica.aaml2))
```

```
## Call:
## coxph(formula = Surv(yrsos, osi) ~ trt_arm, data = cavatica.aaml2)
##
##   n= 956, number of events= 315
##    (102 observations deleted due to missingness)
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## trt_arm -0.01390   0.98620  0.01128 -1.232    0.218
##
##         exp(coef) exp(-coef) lower .95 upper .95
## trt_arm    0.9862      1.014    0.9646     1.008
##
## Concordance= 0.511  (se = 0.015 )
## Likelihood ratio test= 1.52  on 1 df,   p=0.2
## Wald test            = 1.52  on 1 df,   p=0.2
## Score (logrank) test = 1.52  on 1 df,   p=0.2
```

Multivariate cox regression on survival times across features

```
res.cox1 = coxph(Surv(yrsefs, efsi) ~ race_cat +
    riskgrp, data = cavatica.aaml2)
summary(res.cox1)
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ race_cat + riskgrp, data = cavatica.aaml2)
##
##   n= 932, number of events= 485
##    (126 observations deleted due to missingness)
##
##               coef exp(coef)  se(coef)      z Pr(>|z|)
## race_cat -0.012906  0.987177  0.029201 -0.442    0.658
## riskgrp   0.038935  1.039703  0.005008  7.774 7.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## race_cat    0.9872     1.0130    0.9323     1.045
## riskgrp     1.0397     0.9618    1.0295     1.050
##
## Concordance= 0.585  (se = 0.013 )
## Likelihood ratio test= 53.64  on 2 df,   p=2e-12
## Wald test            = 60.5  on 2 df,   p=7e-14
## Score (logrank) test = 63.56  on 2 df,   p=2e-14
```

```
res.cox2 = coxph(Surv(yrsefs, efsi) ~ as.character(race_cat) +
    as.character(riskgrp) + as.character(npmstat_) +
    cbf_pt + as.character(itdlowHAR), data = cavatica.aaml2)
summary(res.cox2)
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ as.character(race_cat) +
##     as.character(riskgrp) + as.character(npmstat_) + cbf_pt +
##     as.character(itdlowHAR), data = cavatica.aaml2)
##
##   n= 932, number of events= 485
##    (126 observations deleted due to missingness)
##
```

```
##                              coef exp(coef) se(coef)      z Pr(>|z|)
## as.character(race_cat)2   0.45803   1.58096  0.61491  0.745    0.456
## as.character(race_cat)3   0.77049   2.16083  0.81949  0.940    0.347
## as.character(race_cat)4   0.34864   1.41714  0.59454  0.586    0.558
## as.character(race_cat)5   0.38528   1.47002  0.58143  0.663    0.508
## as.character(race_cat)6   1.06449   2.89937  1.15711  0.920    0.358
## as.character(race_cat)9   0.31331   1.36795  0.59515  0.526    0.599
## as.character(riskgrp)30   0.47195   1.60312  0.10650  4.432 9.35e-06 ***
## as.character(npmstat_)2   1.11714   3.05611  0.23946  4.665 3.08e-06 ***
## cbf_ptCBF                -0.73817   0.47799  0.12743 -5.793 6.93e-09 ***
## as.character(itdlowHAR)2 -0.06625   0.93589  0.16921 -0.392    0.695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## as.character(race_cat)2      1.5810     0.6325    0.4737    5.2764
## as.character(race_cat)3      2.1608     0.4628    0.4336   10.7690
## as.character(race_cat)4      1.4171     0.7056    0.4419    4.5445
## as.character(race_cat)5      1.4700     0.6803    0.4703    4.5945
## as.character(race_cat)6      2.8994     0.3449    0.3002   28.0050
## as.character(race_cat)9      1.3679     0.7310    0.4261    4.3920
## as.character(riskgrp)30      1.6031     0.6238    1.3011    1.9752
## as.character(npmstat_)2      3.0561     0.3272    1.9114    4.8865
## cbf_ptCBF                    0.4780     2.0921    0.3723    0.6136
## as.character(itdlowHAR)2     0.9359     1.0685    0.6717    1.3039
##
## Concordance= 0.642  (se = 0.013 )
## Likelihood ratio test= 109.9  on 10 df,   p=<2e-16
## Wald test            = 99.73  on 10 df,   p=<2e-16
## Score (logrank) test = 109.5  on 10 df,   p=<2e-16
```

visualizing these reuslts: https://rpkgs.datanovia.com/survminer/survminer_cheatsheet.pdf

https://shariq-mohammed.github.io/files/cbsa2019/1-intro-to-survival.html#6_cox_regression

```
res.cox2 = coxph(Surv(yrsefs, efsi) ~ as.character(race_cat) +
    as.character(riskgrp) + as.character(npmstat_) +
    cbf_pt + as.character(itdlowHAR), data = cavatica.aaml2)
summary(res.cox2)
```

```
## Call:
## coxph(formula = Surv(yrsefs, efsi) ~ as.character(race_cat) +
##     as.character(riskgrp) + as.character(npmstat_) + cbf_pt +
##     as.character(itdlowHAR), data = cavatica.aaml2)
##
##   n= 932, number of events= 485
##    (126 observations deleted due to missingness)
##
##                              coef exp(coef) se(coef)      z Pr(>|z|)
## as.character(race_cat)2   0.45803   1.58096  0.61491  0.745    0.456
## as.character(race_cat)3   0.77049   2.16083  0.81949  0.940    0.347
## as.character(race_cat)4   0.34864   1.41714  0.59454  0.586    0.558
## as.character(race_cat)5   0.38528   1.47002  0.58143  0.663    0.508
## as.character(race_cat)6   1.06449   2.89937  1.15711  0.920    0.358
```

```
## as.character(race_cat)9   0.31331   1.36795  0.59515  0.526    0.599
## as.character(riskgrp)30   0.47195   1.60312  0.10650  4.432 9.35e-06 ***
## as.character(npmstat_)2   1.11714   3.05611  0.23946  4.665 3.08e-06 ***
## cbf_ptCBF                -0.73817   0.47799  0.12743 -5.793 6.93e-09 ***
## as.character(itdlowHAR)2 -0.06625   0.93589  0.16921 -0.392    0.695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef) lower .95 upper .95
## as.character(race_cat)2     1.5810     0.6325    0.4737    5.2764
## as.character(race_cat)3     2.1608     0.4628    0.4336   10.7690
## as.character(race_cat)4     1.4171     0.7056    0.4419    4.5445
## as.character(race_cat)5     1.4700     0.6803    0.4703    4.5945
## as.character(race_cat)6     2.8994     0.3449    0.3002   28.0050
## as.character(race_cat)9     1.3679     0.7310    0.4261    4.3920
## as.character(riskgrp)30     1.6031     0.6238    1.3011    1.9752
## as.character(npmstat_)2     3.0561     0.3272    1.9114    4.8865
## cbf_ptCBF                   0.4780     2.0921    0.3723    0.6136
## as.character(itdlowHAR)2    0.9359     1.0685    0.6717    1.3039
##
## Concordance= 0.642  (se = 0.013 )
## Likelihood ratio test= 109.9  on 10 df,   p=<2e-16
## Wald test            = 99.73  on 10 df,   p=<2e-16
## Score (logrank) test = 109.5  on 10 df,   p=<2e-16
```

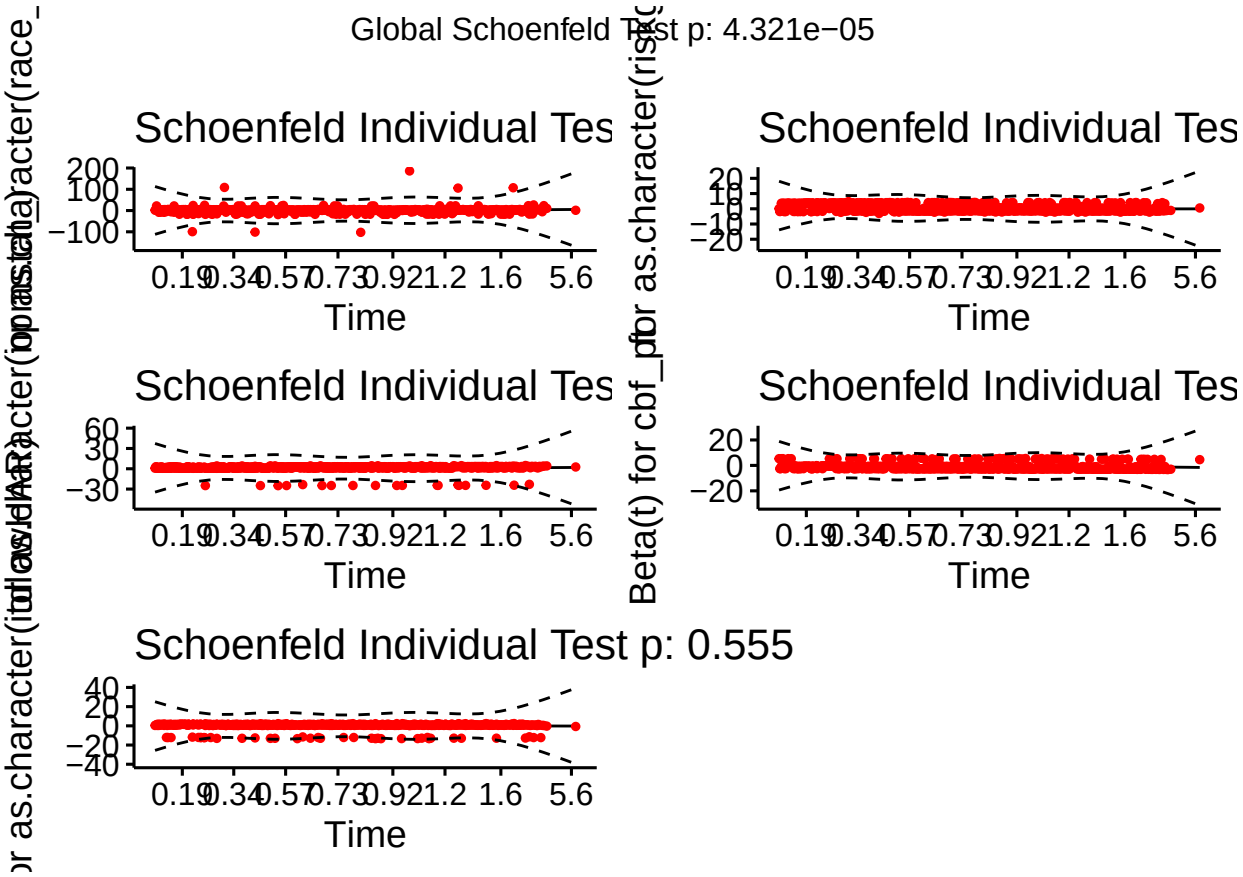
Testing the proportional hazards assumption : In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption. For each covariate, the function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole. The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship.

```
test.ph1 = cox.zph(res.cox2)
test.ph1
```

```
##                          chisq df       p
## as.character(race_cat)   6.486  6    0.37
## as.character(riskgrp)   31.636  1 1.9e-08
## as.character(npmstat_)   0.809  1    0.37
## cbf_pt                   0.854  1    0.36
## as.character(itdlowHAR)  0.349  1    0.55
## GLOBAL                  37.676 10 4.3e-05
```

The proportional hazard assumption is not supported for the risk group feature since there is a significant relationship between residuals and time. Proportional hazards assumption for this model is not reasonable (?)

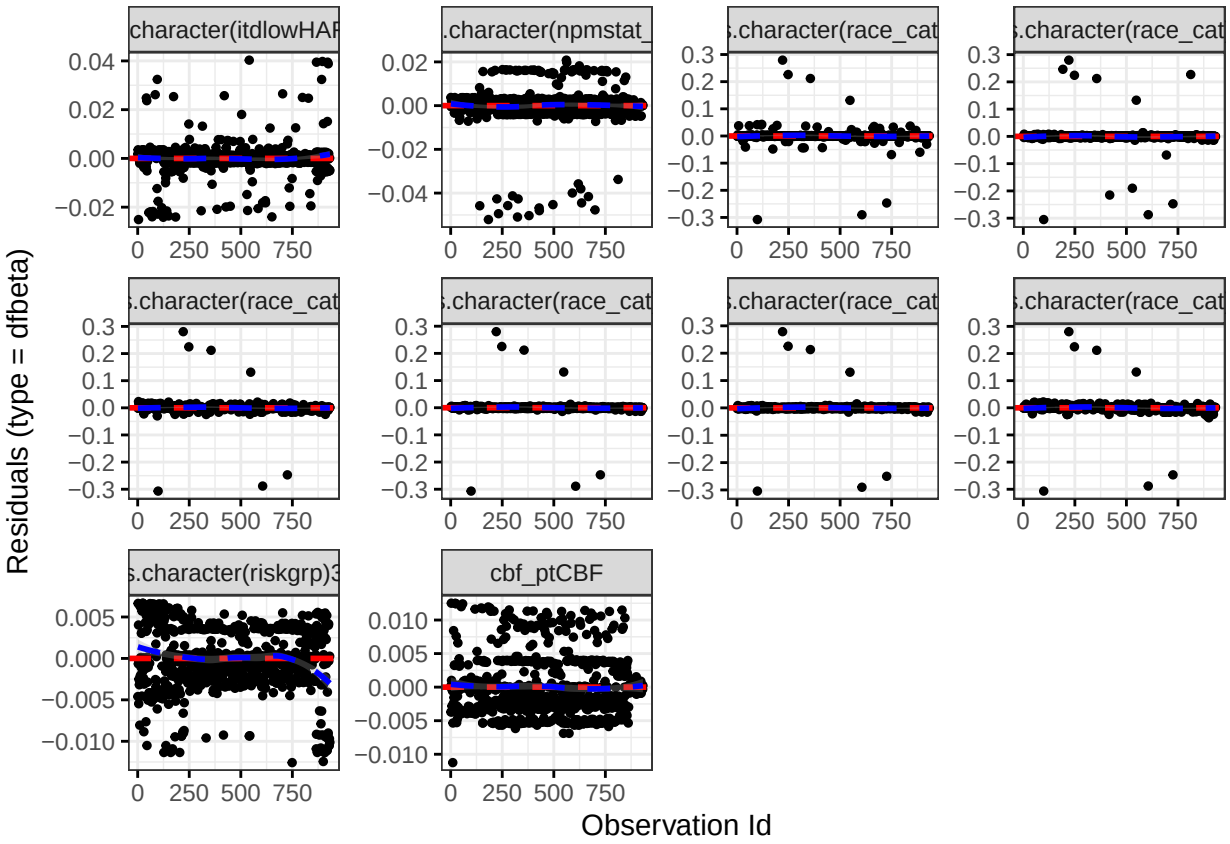
```
ggcoxzph(test.ph1)
```



Looking for outliers

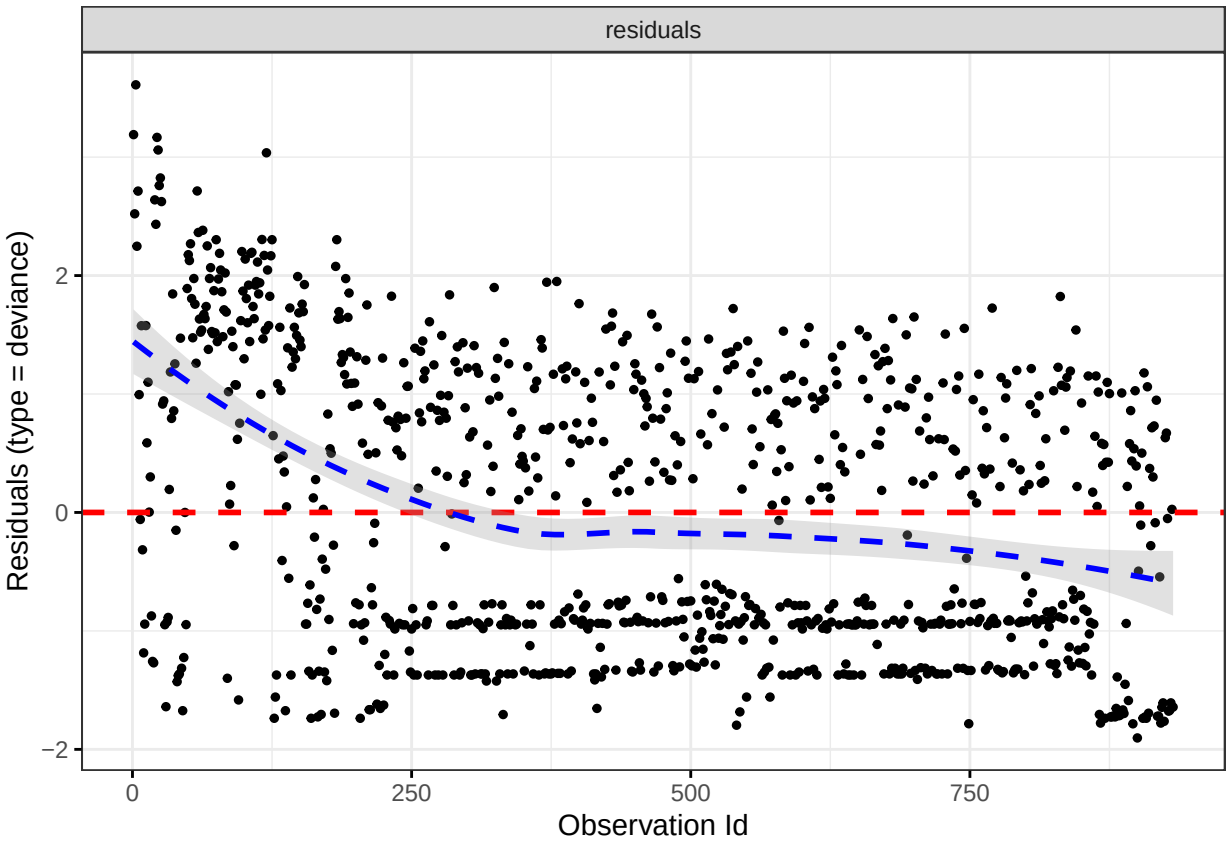
```
ggcoxdiagnostics(res.cox2, type = "dfbeta",
    linear.predictions = F, ggtheme = theme_bw())
```

```
## Warning: ‘gather_()‘ was deprecated in tidyr 1.2.0.
## i Please use ‘gather()‘ instead.
## i The deprecated feature was likely used in the survminer package.
##   Please report the issue at <https://github.com/kassambara/survminer/issues>.

## ‘geom_smooth()‘ using formula = ’y ~ x’
```




```
ggcoxdiagnostics(res.cox2, type = "deviance",
    linear.predictions = F, ggtheme = theme_bw())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



It is also possible to check outliers by visualizing the deviance residuals.

The deviance residual is a normalized transform of the martingale residual.

These residuals should be roughly symmetrically distributed about zero with a standard deviation of 1.

Positive values correspond to individuals that "died too soon" compared to expected survival times.

Negative values correspond to individual that "lived too long".

Very large or small values are outliers, which are poorly predicted by the model.

```
# res.cox3 = coxph(Surv(yrsefs, efsi) ~
# race_cat + riskgrp + npmstat_ +
# cbf_pt + itdlowHAR, data =
# cavatica.aaml2) summary(res.cox3)
```

A forest plot to summarize the results across these features for this particular model. How to read a forest plot: https://s4be.cochrane.org/blog/2016/07/11/tutorial-read-forest-plot/

```
# ggforest(res.cox2)
```

The bigger the black box, the more participants in this group. The horzontal line through the group represents the 95% confidence interval

adjusted survival curves

```
# ggcoxadjustedcurves(res.cox2, data =
# cavatica.aaml2)
```

Preparing for rMATS-turbo

Once exploratory data analysis is complete, Export the list of data subsets needed and group them in Cavatica for downstream analysis sort_files_cavatica.py