Appendix

Yixuan Jiao, Hongpu Min, Qingzhen Sun, Yining Chen

2023-12-10

```
library(tidyverse)
library(survival)
library(survminer)
library(biostat3)
library(finalfit)
```

```
## Warning: package 'finalfit' was built under R version 4.2.3
```

```
library(dplyr)
library(kableExtra)
library(ggplot2)
library(ggpubr)
library(riskRegression)
library(regclass)
```

EDA

```
aids <- read.csv('data/AIDS_Clinical_Trials_Group175.csv')
aids <- aids%>%
  mutate(cid = as.factor(cid),

         trt = as.factor(trt),treat=as.factor(treat),
                                                                    hemo = as.factor(hemo),
                                                                    homo = as.factor(homo),
                                                                    drugs = as.factor(drugs),
                                                                    race = as.factor(race),
                                                                    gender = as.factor(gender),
                                                                    str2 = as.factor(str2),
                                                                    symptom = as.factor(symptom))
```

```
aids <- aids%>%
  mutate(cid=factor(cid,
  levels = c(0,1),
  labels = c("Censoring",
             "Failure")),
  treat=factor(treat,
  levels = c(0,1),
  labels = c("ZDV only",
```

```
            "Others")),
  hemo=factor(hemo,
  levels = c(0,1),
  labels = c("No",
             "Yes")),
  homo=factor(homo,
  levels = c(0,1),
  labels = c("No",
             "Yes")),
  gender=factor(gender,
  levels = c(0,1),
  labels = c("Female",
             "Male")),
  race=factor(race,
  levels = c(0,1),
  labels = c("White",
             "Non-white")),
  drugs=factor(drugs,
  levels = c(0,1),
  labels = c("No",
             "Yes")),
  symptom=factor(symptom,
  levels = c(0,1),
  labels = c("No",
             "Yes"))
  )
```

Table 1: Variable description

Variable	Description
age	age (yrs) at baseline
race	race (0=White, 1=non-white)
gender	gender (0=F, 1=M)
trt	treatment indicator (0 = ZDV only; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI only)
drug	history of IV drug use ((0=no, 1=yes))
hemo	hemophilia (0=no, 1=yes)
homo	homosexual activity (0=no, 1=yes)
kanor	Karnofsky score (0-100)
symptom	symptomatic infection indicator (0=asymp, 1=symp)
cd40	CD4 count at baseline
cd80	CD8 count at baseline
str2	antiretroviral history (0=naive, 1=experienced)
oprior	Non-ZDV antiretroviral therapy pre-175 (0=no, 1=yes)
z30	ZDV in the 30 days prior to 175 (0=no, 1=yes)
zpiror	ZDV prior to 175 (0=no, 1=yes)

sex age race hemo homo drug kanor symptom cd40

```
explanatory = c("age", "hemo", "homo", "race","gender","drugs","karnof","cd40","symptom","cd80","wtkg")
dependent = "treat"
baseline <- aids %>%
  mutate(
        cd80 = ff_label(cd80, "CD8 Count"),
        wtkg = ff_label(wtkg, "Weight"),
        gender = ff_label(gender, "Gender"),
        hemo = ff_label(hemo, "Hemophilia"),
        homo = ff_label(homo, "Homosexuality"),
        race = ff_label(race, "Race"),
        drugs = ff_label(drugs, "History of IV drug use "),
        karnof = ff_label(karnof, "Karnofsky score of 100"),
        cd40 = ff_label(cd40, "CD4 count"),
        age = ff_label(age, "Age"),
        symptom = ff_label(symptom, "Symptomatic infection"),
        treat = ff_label(treat, "Treatment")
    )%>%
  summary_factorlist(dependent, explanatory,column = TRUE, total_col = TRUE,,col_totals_prefix = "N=",ac

baseline
```

```
##     Dependent: Treatment                ZDV only        Others         Total
##                         Age Mean (SD)    35.2 (8.9)    35.3 (8.7)    35.2 (8.7)
##                Hemophilia        No    490 (92.1)   1469 (91.4)   1959 (91.6)
##                                  Yes      42 (7.9)     138 (8.6)     180 (8.4)
##             Homosexuality        No    191 (35.9)    534 (33.2)    725 (33.9)
##                                  Yes    341 (64.1)   1073 (66.8)   1414 (66.1)
##                      Race     White    376 (70.7)   1146 (71.3)   1522 (71.2)
##                           Non-white    156 (29.3)    461 (28.7)    617 (28.8)
##                    Gender    Female    100 (18.8)    268 (16.7)    368 (17.2)
##                                 Male    432 (81.2)   1339 (83.3)   1771 (82.8)
##  History of IV drug use         No    469 (88.2)   1389 (86.4)   1858 (86.9)
##                                  Yes     63 (11.8)    218 (13.6)    281 (13.1)
##   Karnofsky score of 100        70       4 (0.8)       5 (0.3)       9 (0.4)
##                                 80      17 (3.2)      63 (3.9)      80 (3.7)
##                                 90    197 (37.0)    590 (36.7)    787 (36.8)
##                                100    314 (59.0)    949 (59.1)   1263 (59.0)
##                CD4 count Mean (SD) 353.2 (114.1) 349.6 (120.0) 350.5 (118.6)
##    Symptomatic infection        No    443 (83.3)   1326 (82.5)   1769 (82.7)
##                                  Yes     89 (16.7)    281 (17.5)    370 (17.3)
##                CD8 Count Mean (SD) 987.2 (475.2) 986.4 (482.0) 986.6 (480.2)
##                   Weight Mean (SD)   76.1 (13.2)   74.8 (13.3)   75.1 (13.3)
```

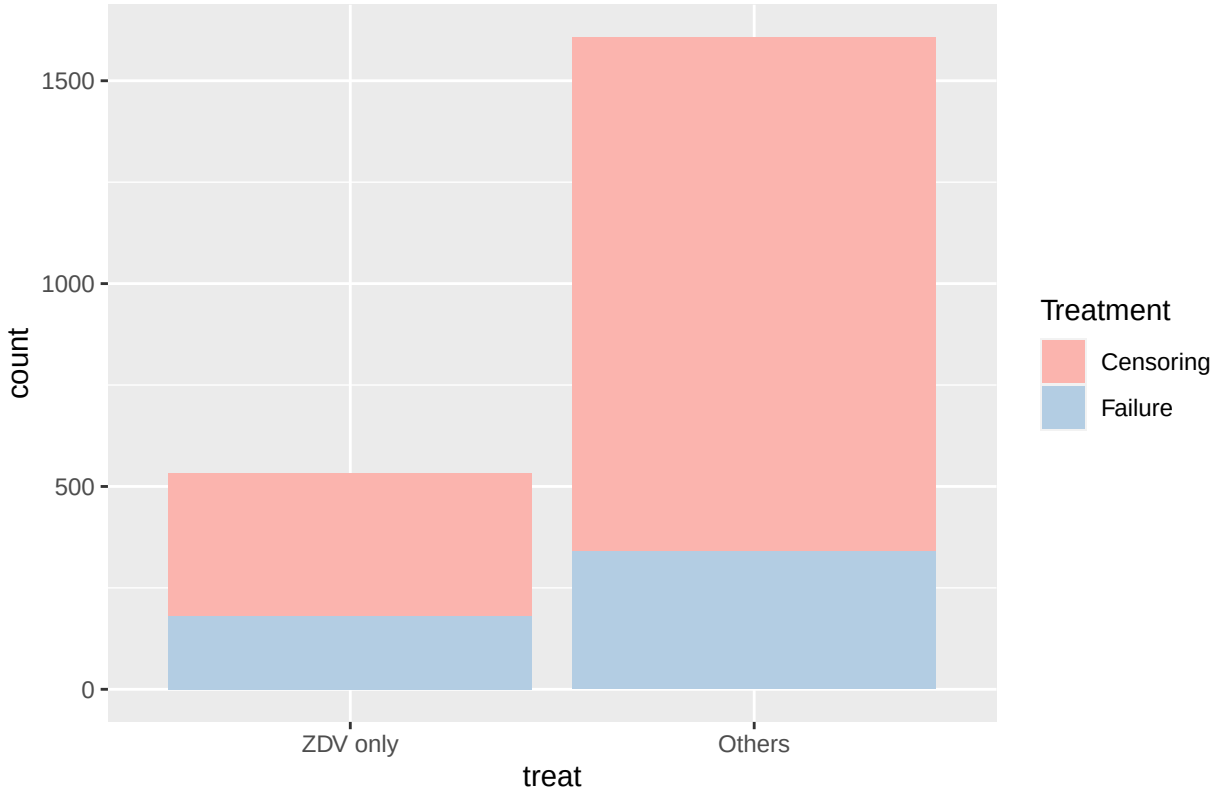
```
kable(baseline,caption = "Base-Line Characteristics of the Patients According to the Treatment Indicato
```

```
ggplot(data=aids, aes(x=treat, fill=cid)) +
  geom_bar()+
  scale_fill_brewer(palette="Pastel1")+
  ggtitle("Figure 1: Distribution of  Treatment VS. Patient Status")+
  guides(fill = guide_legend(title = "Treatment"))
```

Table 2: Base-Line Characteristics of the Patients According to the Treatment Indicator

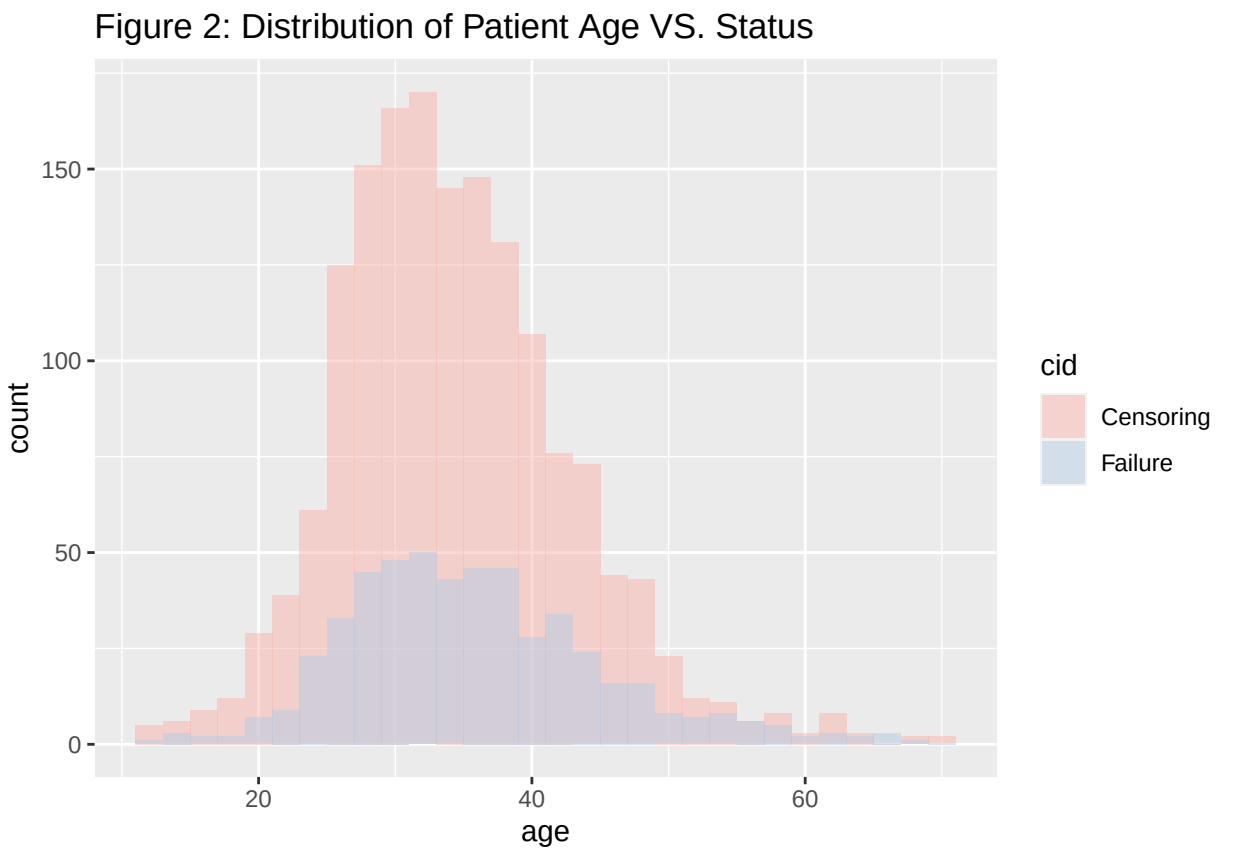
Dependent: Treatment		ZDV only	Others	Total
Age	Mean (SD)	35.2 (8.9)	35.3 (8.7)	35.2 (8.7)
Hemophilia	No	490 (92.1)	1469 (91.4)	1959 (91.6)
	Yes	42 (7.9)	138 (8.6)	180 (8.4)
Homosexuality	No	191 (35.9)	534 (33.2)	725 (33.9)
	Yes	341 (64.1)	1073 (66.8)	1414 (66.1)
Race	White	376 (70.7)	1146 (71.3)	1522 (71.2)
	Non-white	156 (29.3)	461 (28.7)	617 (28.8)
Gender	Female	100 (18.8)	268 (16.7)	368 (17.2)
	Male	432 (81.2)	1339 (83.3)	1771 (82.8)
History of IV drug use	No	469 (88.2)	1389 (86.4)	1858 (86.9)
	Yes	63 (11.8)	218 (13.6)	281 (13.1)
Karnofsky score of 100	70	4 (0.8)	5 (0.3)	9 (0.4)
	80	17 (3.2)	63 (3.9)	80 (3.7)
	90	197 (37.0)	590 (36.7)	787 (36.8)
	100	314 (59.0)	949 (59.1)	1263 (59.0)
CD4 count	Mean (SD)	353.2 (114.1)	349.6 (120.0)	350.5 (118.6)
Symptomatic infection	No	443 (83.3)	1326 (82.5)	1769 (82.7)
	Yes	89 (16.7)	281 (17.5)	370 (17.3)
CD8 Count	Mean (SD)	987.2 (475.2)	986.4 (482.0)	986.6 (480.2)
Weight	Mean (SD)	76.1 (13.2)	74.8 (13.3)	75.1 (13.3)

Figure 1: Distribution of Treatment VS. Patient Status



```
ggplot(aids, aes(x=age, fill=cid)) +
  geom_histogram(alpha=0.5, position="identity")+
  scale_fill_brewer(palette ="Pastel1")+
  ggtitle("Figure 2: Distribution of Patient Age VS. Status")
```

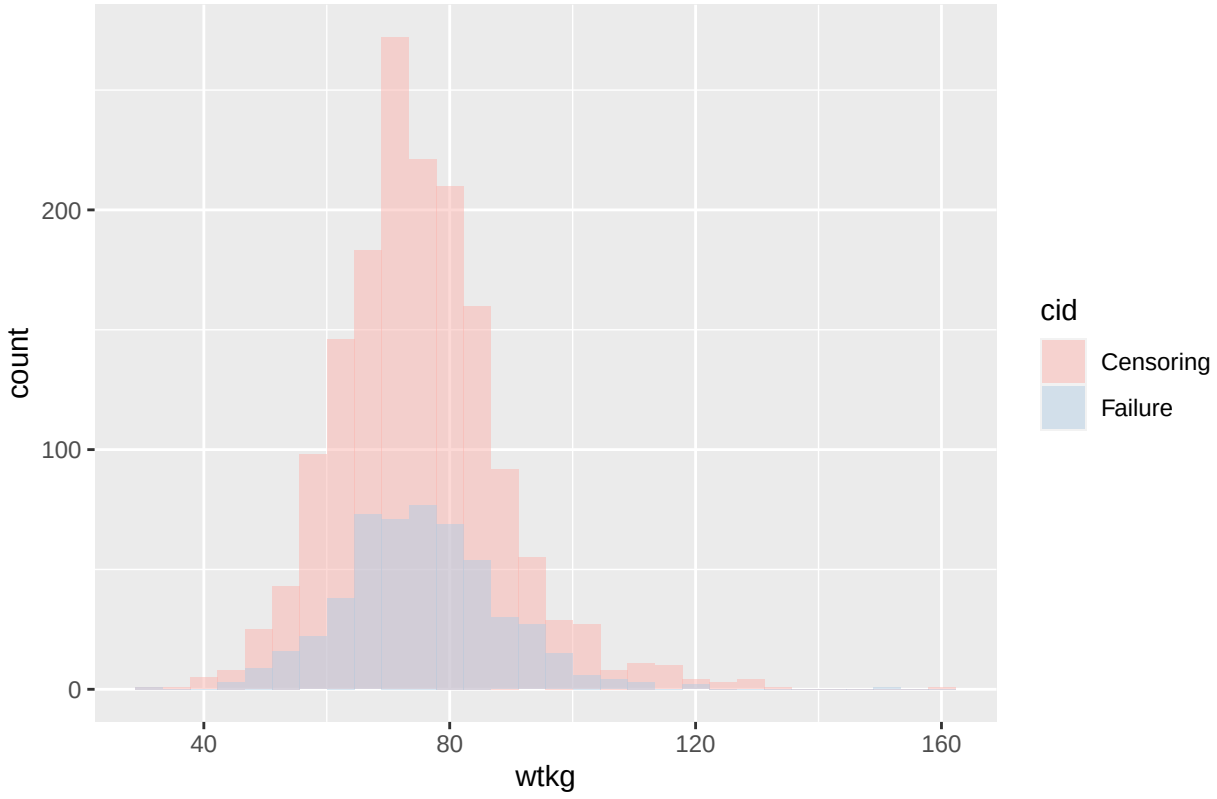
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(aids, aes(x=wtkg, fill=cid)) +
  geom_histogram(alpha=0.5, position="identity")+
  scale_fill_brewer(palette ="Pastel1")+
  ggtitle("Figure 3: Distribution of Patient Weight VS. Status")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

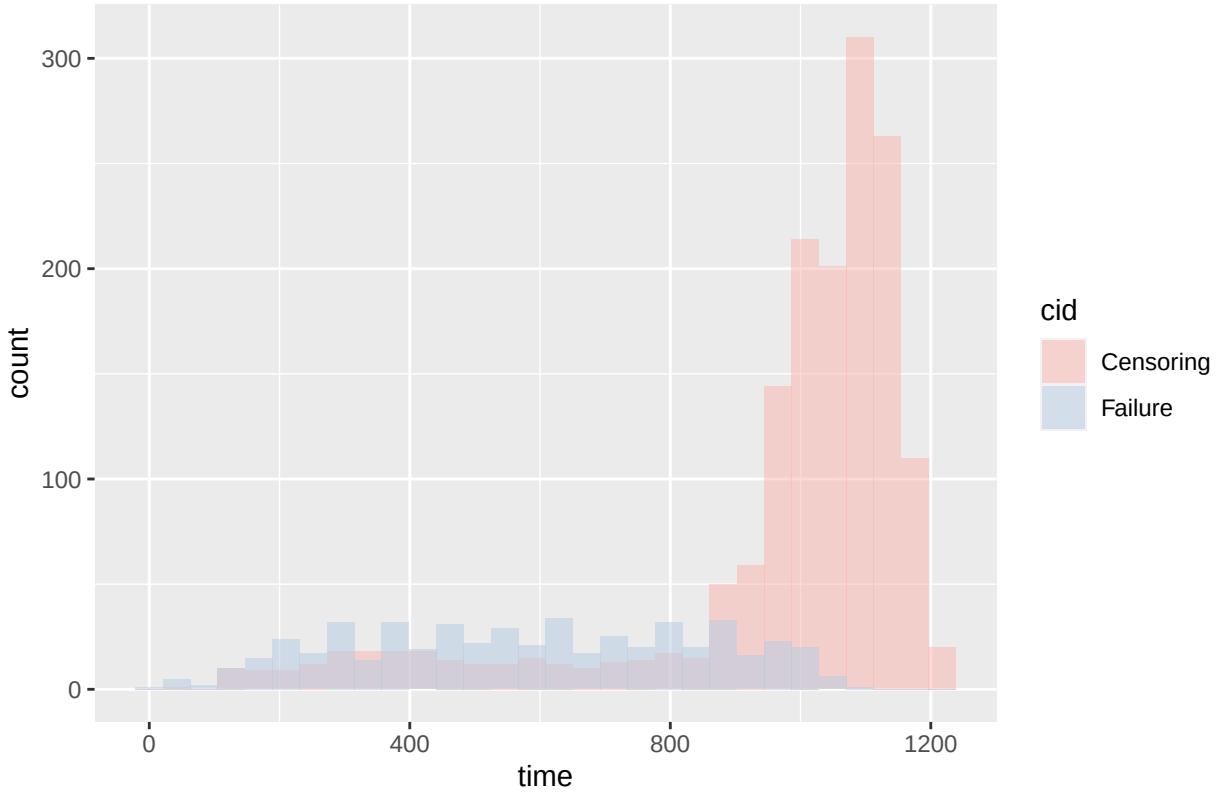
Figure 3: Distribution of Patient Weight VS. Status




```
ggplot(aids, aes(x=time, fill=cid)) +
  geom_histogram(alpha=0.5, position="identity")+
  scale_fill_brewer(palette ="Pastel1")+
  ggtitle("Figure 4: Distribution of Time to Failure or Censoring")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

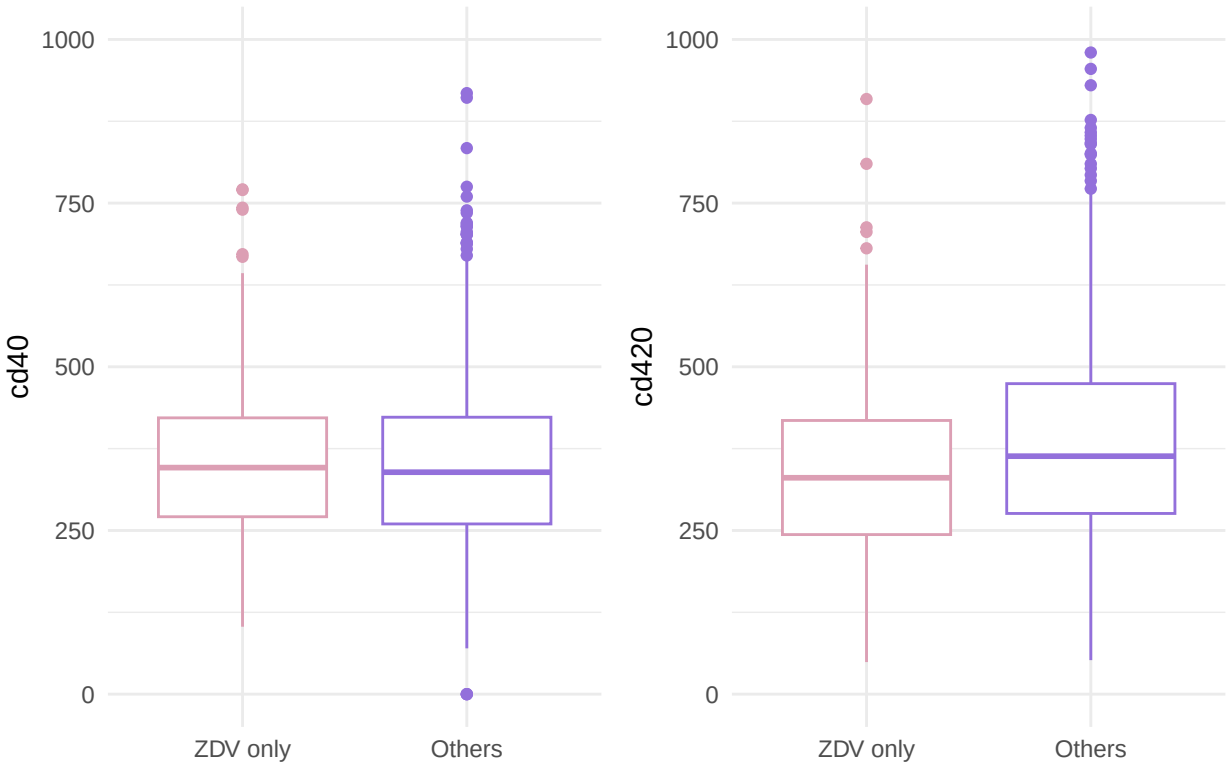
Figure 4: Distribution of Time to Failure or Censoring




```
p1 <- ggplot(aids,aes(x= treat, y=cd40,color=treat)) + geom_boxplot(show.legend = FALSE)+labs(x="")+scal
  theme_minimal()+
 scale_y_continuous(limits = c(0, 1000))


p2 <- ggplot(aids,aes(x= treat, y=cd420,color=treat)) + geom_boxplot(show.legend = FALSE)+labs(x="")+sca
  theme_minimal()+
 scale_y_continuous(limits = c(0, 1000))

p <- ggpubr::ggarrange(p1, p2, ncol=2,nrow = 1,common.legend = TRUE)
```

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

## Warning: Removed 3 rows containing non-finite values (‘stat_boxplot()‘).

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

## Warning: Removed 3 rows containing non-finite values (‘stat_boxplot()‘).
```

```
ggpubr::annotate_figure(p, top = ggpubr::text_grob("Figure 5: CD4 Count Change VS. Treatment Groups", c
```

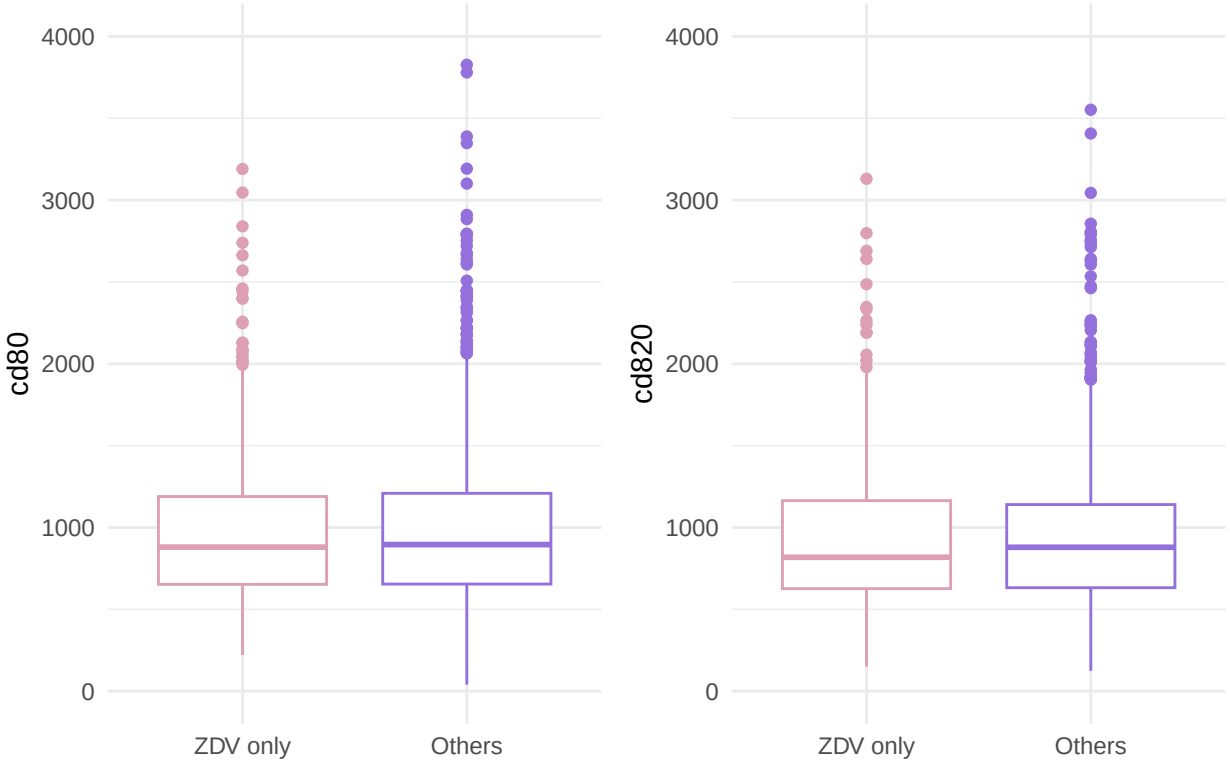



```
p3 <- ggplot(aids,aes(x= treat, y=cd80,color=treat)) + geom_boxplot(show.legend = FALSE)+labs(x="")+scal
  theme_minimal()+
 scale_y_continuous(limits = c(0, 4000))

p4 <- ggplot(aids,aes(x= treat, y=cd820,color=treat)) + geom_boxplot(show.legend = FALSE)+labs(x="")+sca
  theme_minimal()+
 scale_y_continuous(limits = c(0, 4000))

p<- ggarrange(p3,p4,ncol=2,align = "hv",nrow = 1,common.legend = TRUE)
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
## Removed 2 rows containing non-finite values (`stat_boxplot()`).
## Removed 2 rows containing non-finite values (`stat_boxplot()`).
## Removed 2 rows containing non-finite values (`stat_boxplot()`).
```

```
annotate_figure(p, top = text_grob("Figure 6: CD8 Count Change VS. Treatment Groups", color = "#0F2540"
```

Figure 6: CD8 Count Change VS. Treatment Groups



Non-parametric tests

```
aids <- read_csv("data/AIDS_Clinical_Trials_Group175.csv")
```

```
## New names:
## Rows: 2139 Columns: 25
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (25): ...1, time, trt, age, wtkg, hemo, homo, drugs, karnof, oprior, z30...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

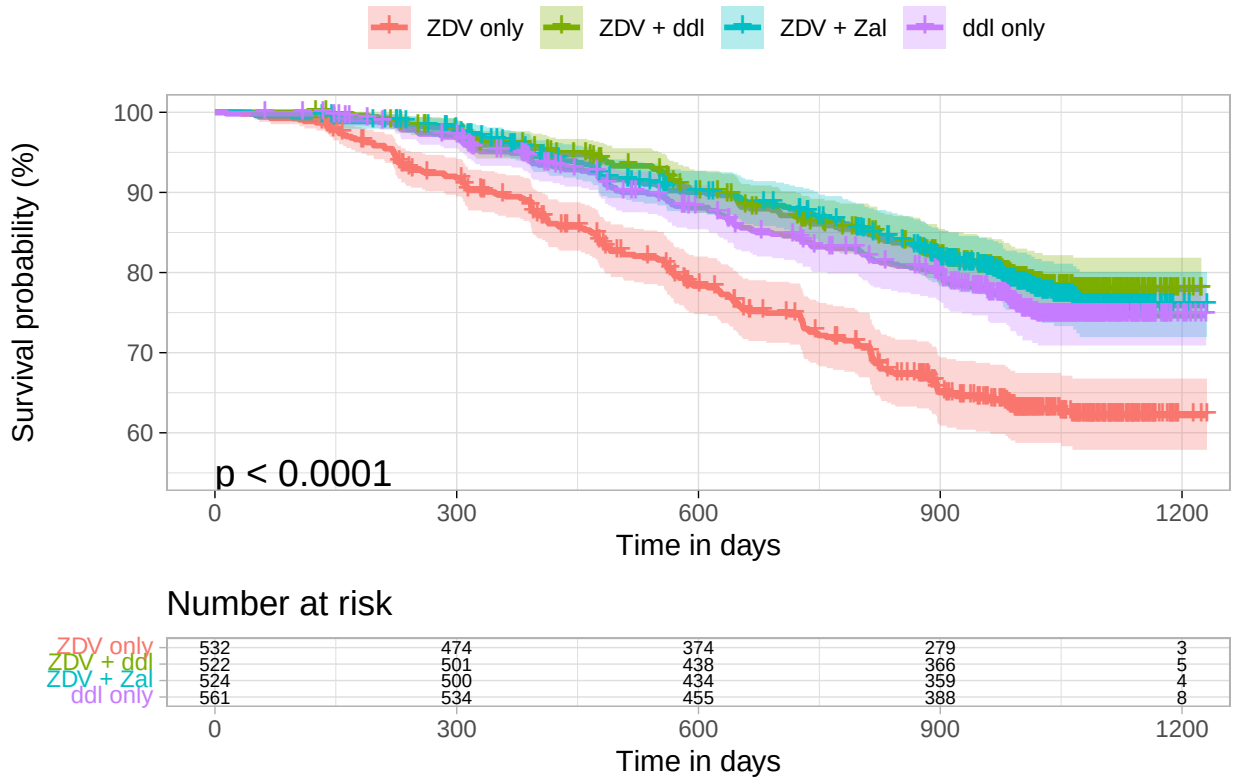
```
aids %>% head()
```

```
## # A tibble: 6 x 25
##    ...1  time   trt   age  wtkg  hemo  homo drugs karnof oprior   z30 zprior
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>  <dbl>
## 1     0   948     2    48  89.8     0     0     0    100      0     0      1
## 2     1  1002     3    61  49.4     0     0     0     90      0     1      1
## 3     2   961     3    45  88.5     0     1     1     90      0     1      1
## 4     3  1166     3    47  85.3     0     1     0    100      0     1      1
```

```
## 5     4  1090     0    43  66.7     0     1     0    100      0     1      1
## 6     5  1181     1    46  88.9     0     1     1    100      0     1      1
## # i 13 more variables: preanti <dbl>, race <dbl>, gender <dbl>, str2 <dbl>,
## #   strat <dbl>, symptom <dbl>, treat <dbl>, offtrt <dbl>, cd40 <dbl>,
## #   cd420 <dbl>, cd80 <dbl>, cd820 <dbl>, cid <dbl>
```

```r
#KM estimate
km_fit_trt <- survfit(Surv(time, cid) ~ trt, data = aids)
pkm <- km_fit_trt %>% ggsurvplot(data = aids,
                      fun = "pct", #can be replaced by cum hazard
                      conf.int = TRUE,
                      risk.table = TRUE,
                      pval = TRUE,
                      pval.coord = c(0,55),
                      fontsize = 2.5,
                      ggtheme = theme_light(),
                      xlab =  "Time in days",
                      title = "Figure 4: Kaplan-Meier Survival Function Estimate",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                      ylim = c(55, 100))

hkm <- km_fit_trt %>% ggsurvplot(data = aids,
                      fun = "cumhaz", #can be replaced by cum hazard
                      conf.int = TRUE,
                      risk.table = TRUE,
                      pval = TRUE,
                      fontsize = 2.5,
                      ggtheme = theme_light(),
                      xlab =  "Time in days",
                      title = "Figure 5: Kaplan-Meier Cumulative Hazard Function",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"))
pkm
```

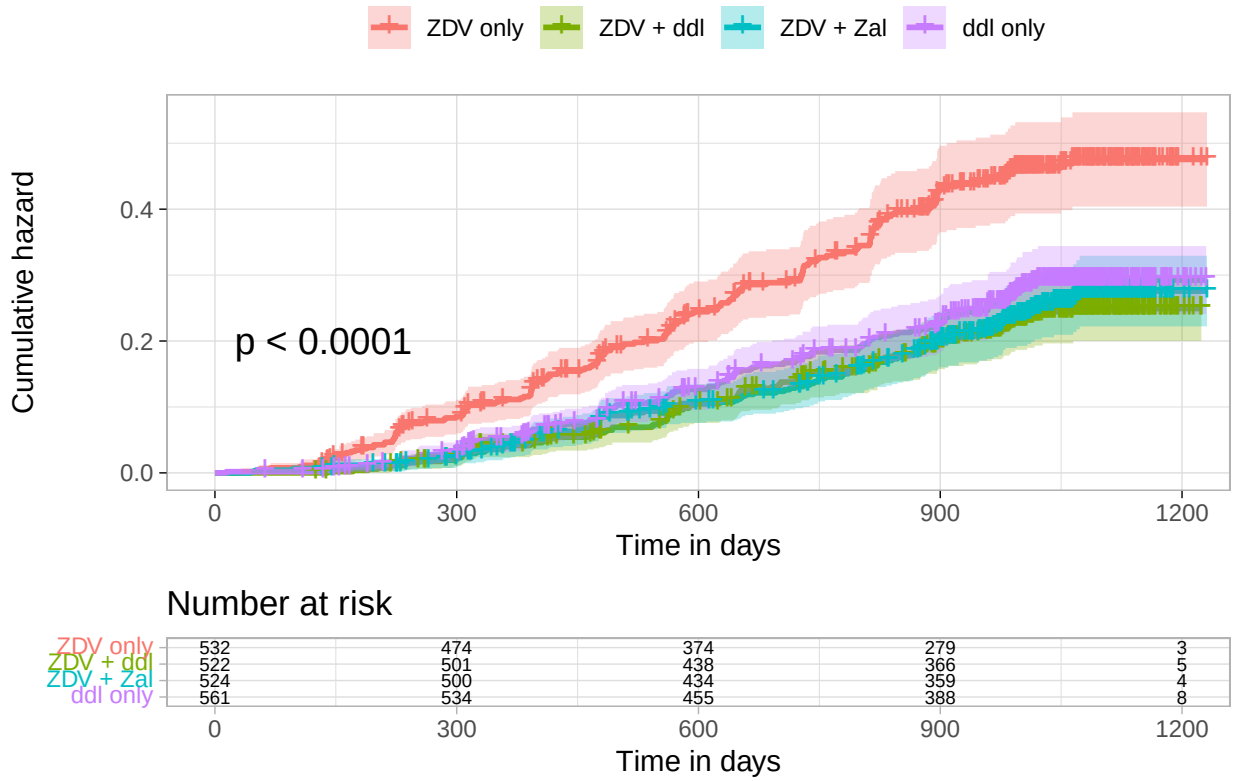
Figure 4: Kaplan−Meier Survival Function Estimate



Number at risk


```
hkm
```

Figure 5: Kaplan−Meier Cumulative Hazard Function



Number at risk


```
quantile(km_fit_trt,probs = c(0.1,0.15,0.2))
```

```
## $quantile
##        10  15  20
## trt=0 347 468 569
## trt=1 626 822 986
## trt=2 610 806 972
## trt=3 537 672 898
##
## $lower
##        10  15  20
## trt=0 284 394 484
## trt=1 559 691 876
## trt=2 476 720 867
## trt=3 476 613 760
##
## $upper
##        10  15  20
## trt=0 406 557 649
## trt=1 721 929  NA
## trt=2 748 910  NA
## trt=3 631 813 994
```

Table 3: Quantile Survival Time (in days) by Treatments

Treatment	90^{th}	85^{th}	80^{th}
ZDV only	347 (284,406)	468 (394,557)	569 (484,649)
ZDV + ddl	626 (559,721)	822 (691,929)	986 (876,NA)
ZDV + Zal	610 (476,748)	806 (720,910	972 (867,NA)
ddl only	537 (476,631)	672 (613,813)	898 (760,994)

```
survdiff(Surv(time, cid) ~ trt, data = aids)
```

```
## Call:
## survdiff(formula = Surv(time, cid) ~ trt, data = aids)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 532      181      116    37.030     47.67
## trt=1 522      103      134     6.988      9.40
## trt=2 524      109      132     4.158      5.58
## trt=3 561      128      139     0.933      1.27
##
##  Chisq= 49.2  on 3 degrees of freedom, p= 1e-10
```

```
aids %>%
  mutate(trt = case_when(trt == "0" ~ "ZDV only",
                         trt == "1" ~ "ZDV + ddl",
                         trt == "2" ~ "ZDV + Zal",
                         trt == "3" ~ "ddl only")) %>%
  pairwise_survdiff(Surv(time, cid) ~ trt, data = .,p.adjust.method = "BH") %>%
  broom::tidy()
```

```
## # A tibble: 6 x 3
##   group1    group2         p.value
##   <chr>     <chr>            <dbl>
## 1 ZDV + ddl ddl only  0.278
## 2 ZDV + Zal ddl only  0.478
## 3 ZDV + Zal ZDV + ddl 0.636
## 4 ZDV only  ddl only  0.00000750
## 5 ZDV only  ZDV + ddl 0.0000000364
## 6 ZDV only  ZDV + Zal 0.000000242
```

Table 4: Pairwise Log-rank Tests by Treatments

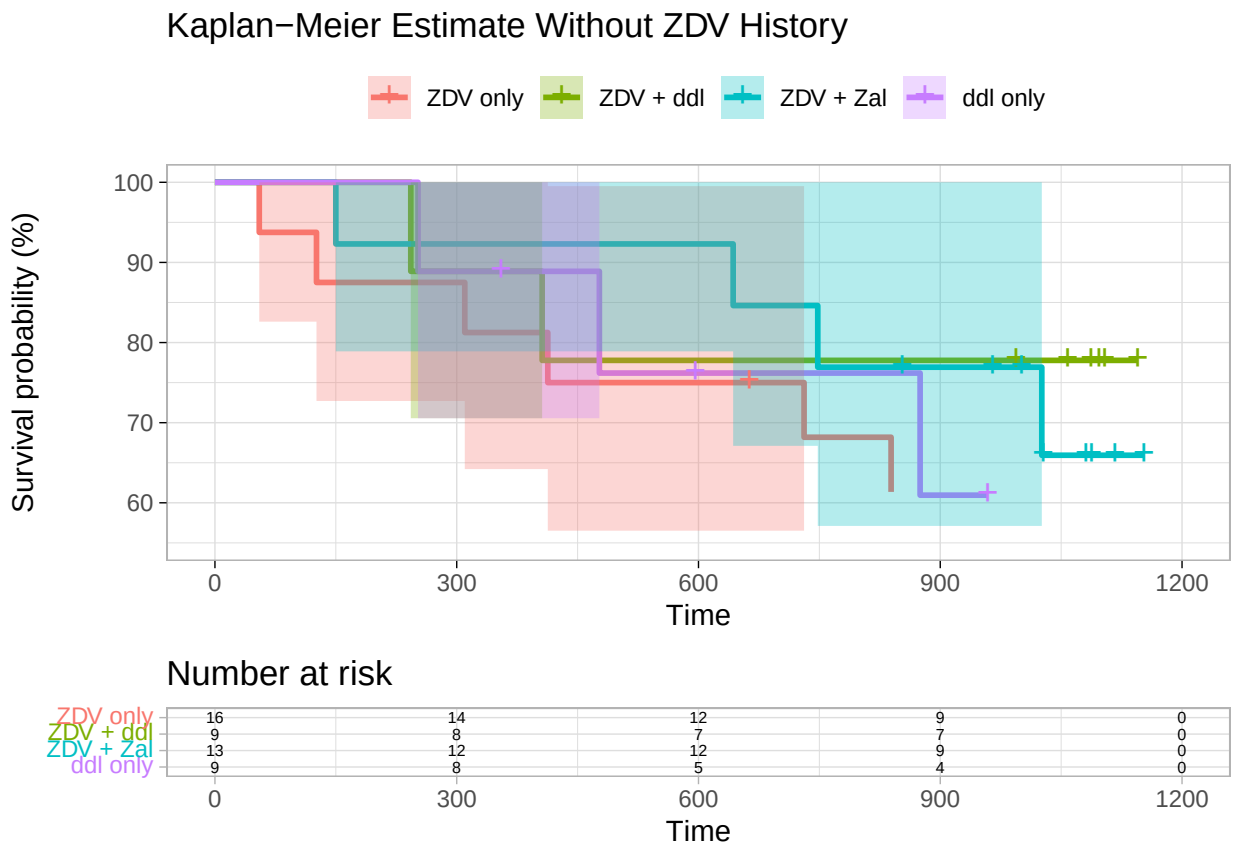
group1	group2	P-value
ZDV + ddl	ddl only	0.2784441
ZDV + Zal	ddl only	0.4776330
ZDV + Zal	ZDV + ddl	0.6362565
ZDV only	ddl only	0.0000075***
ZDV only	ZDV + ddl	0.0000000***
ZDV only	ZDV + Zal	0.0000002***

KM curve for zdv history stratify

```
# non-ZDV treatment history
aids_non_zdv <- aids %>%
  filter(oprior == 1)
```

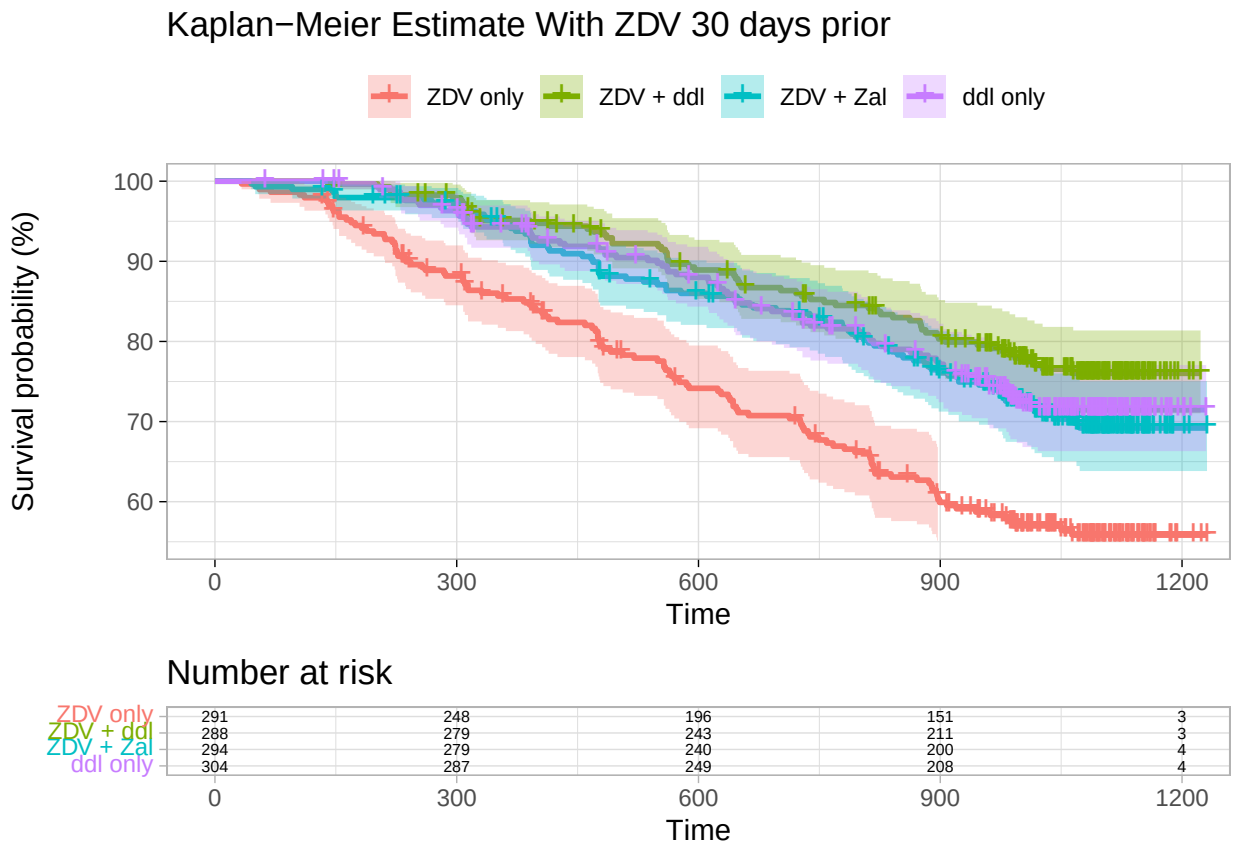
```
km_fit_trt1 <- survfit(Surv(time, cid) ~ trt, data = aids_non_zdv)
km_fit_trt1 %>% ggsurvplot(data = aids_non_zdv,
                       fun = "pct",
                       conf.int = TRUE,
                       risk.table = TRUE,
                       fontsize = 2,
                       ggtheme = theme_light(),
                       title = "Kaplan-Meier Estimate Without ZDV History",
                       legend.title = "",
                       legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                       ylim = c(55, 100))
```

```
## Warning: Removed 12 rows containing missing values (`geom_step()`).

## Warning: Removed 10 rows containing missing values (`geom_point()`).

## Warning: Removed 12 rows containing missing values (`geom_step()`).

## Warning: Removed 10 rows containing missing values (`geom_point()`).
```



```r
# ZDV in the 30 days prior to 175
aids_z30 = aids%>%
  filter(z30 == 1)
```

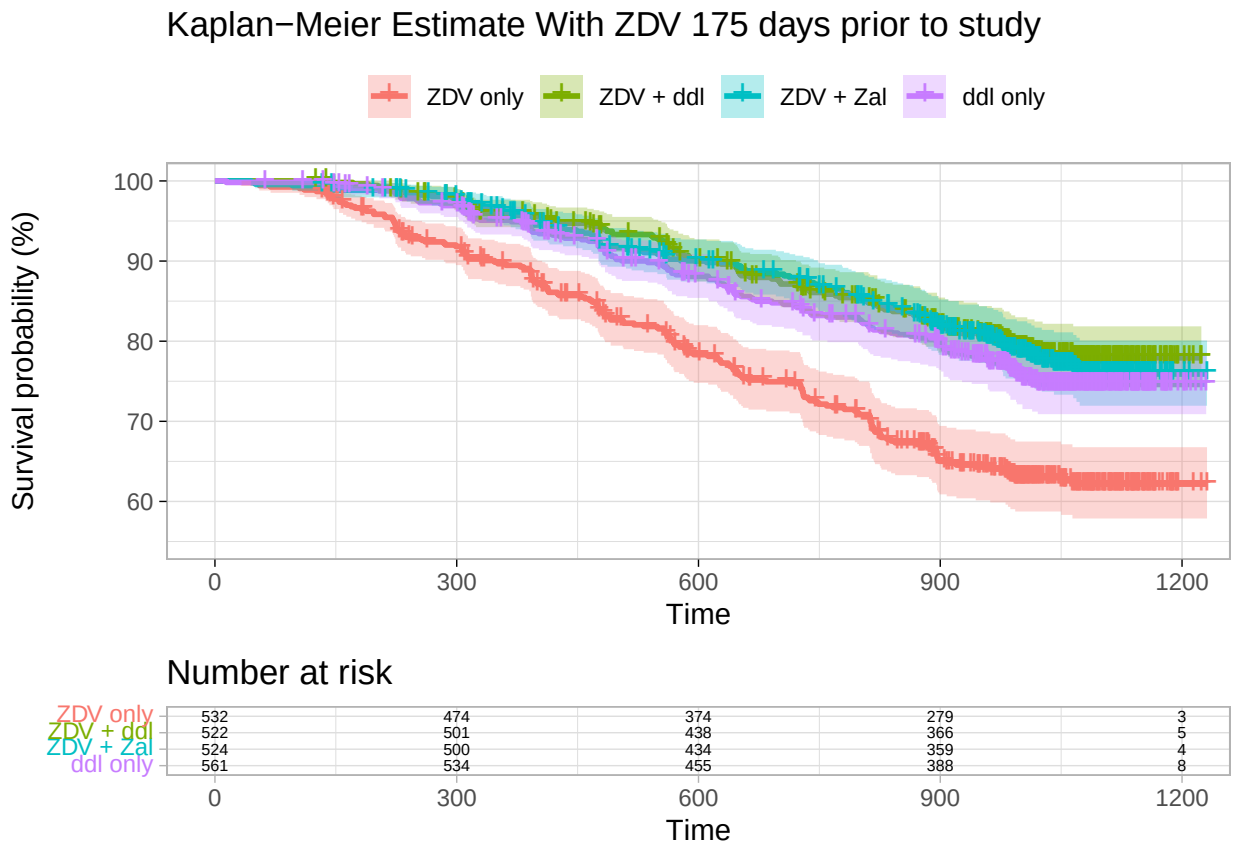
```r
km_fit_trt2 <- survfit(Surv(time, cid) ~ trt, data = aids_z30)
km_fit_trt2 %>% ggsurvplot(data = aids_z30,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Kaplan-Meier Estimate With ZDV 30 days prior",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                      ylim = c(55, 100))
```




```r
# ZDV prior to 175
aids_zprior = aids%>%
  filter(zprior == 1)
```

```r
km_fit_trt3 <- survfit(Surv(time, cid) ~ trt, data = aids_zprior)
km_fit_trt3 %>% ggsurvplot(data = aids_zprior,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
```

```
                    fontsize = 2,
                    ggtheme = theme_light(),
                    title = "Kaplan-Meier Estimate With ZDV 175 days prior to study",
                    legend.title = "",
                    legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                    ylim = c(55, 100))
```




```
survfit_result_zdv <- survfit(Surv(time, cid) ~ strata(oprior) + trt, data = aids)
survfit_result_zdv
```

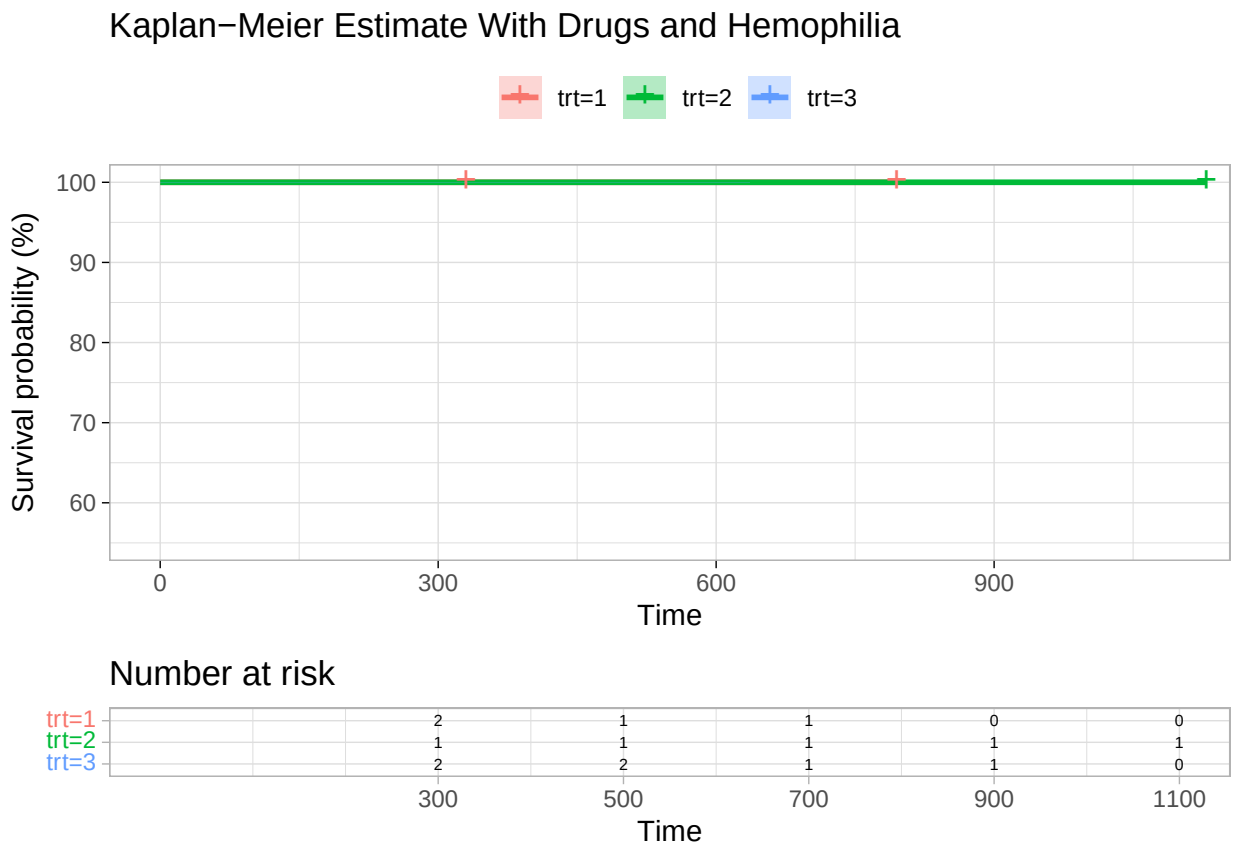
```
## Call: survfit(formula = Surv(time, cid) ~ strata(oprior) + trt, data = aids)
##
##                                   n events median 0.95LCL 0.95UCL
## strata(oprior)=oprior=0, trt=0 516    174     NA      NA      NA
## strata(oprior)=oprior=0, trt=1 513    101     NA      NA      NA
## strata(oprior)=oprior=0, trt=2 511    105     NA      NA      NA
## strata(oprior)=oprior=0, trt=3 552    124     NA      NA      NA
## strata(oprior)=oprior=1, trt=0  16      7     NA     731      NA
## strata(oprior)=oprior=1, trt=1   9      2     NA      NA      NA
## strata(oprior)=oprior=1, trt=2  13      4     NA    1026      NA
## strata(oprior)=oprior=1, trt=3   9      4    994     875      NA
```

KM Curve for medical history stratify

```
# patient with drug used and has hemophilia
aids_hemo_drug = aids%>%
  filter(hemo==1)%>%
  filter(drugs ==1)
```

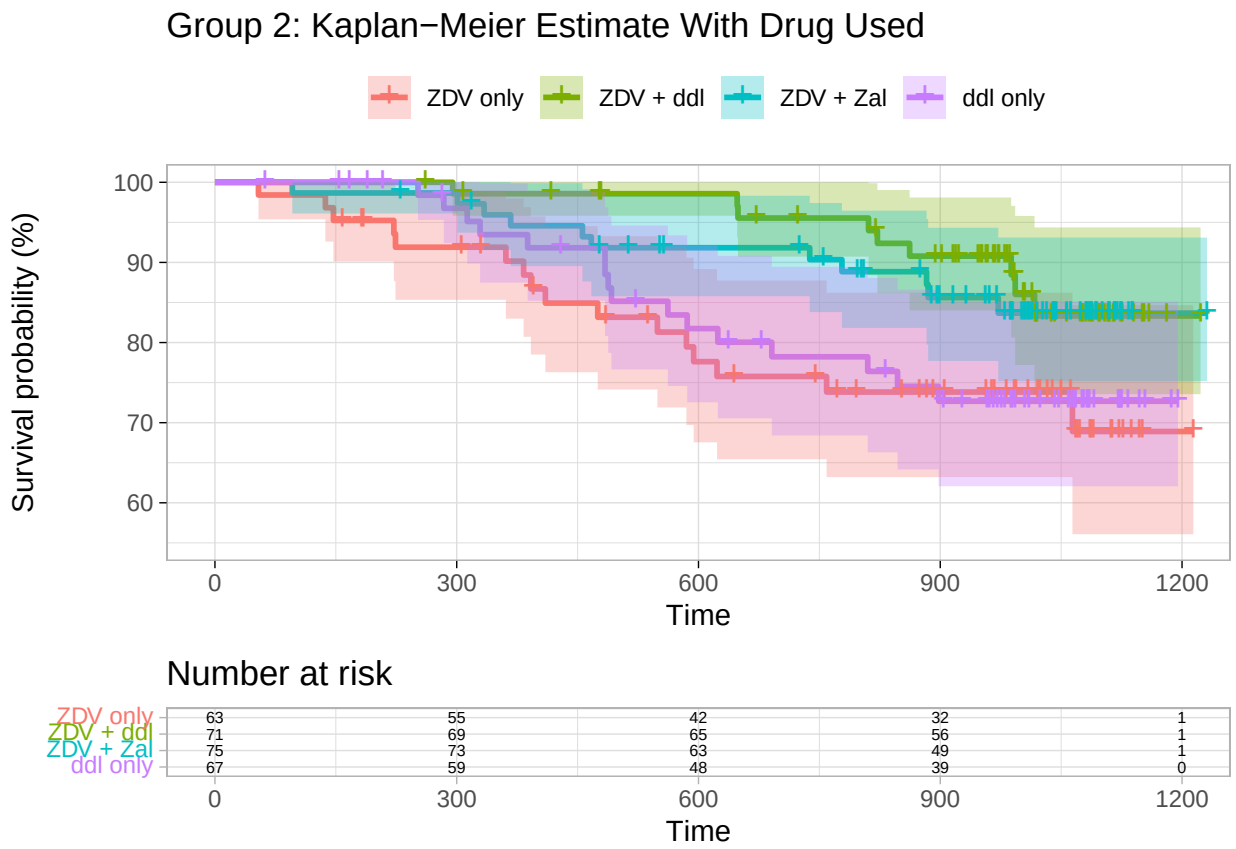
```
km_fit_trt4 <- survfit(Surv(time, cid) ~ trt, data = aids_hemo_drug)
km_fit_trt4 %>% ggsurvplot(data = aids_hemo_drug,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Kaplan-Meier Estimate With Drugs and Hemophilia",
                      legend.title = "",
                      legend.labs = levels(aids$trt),
                      ylim = c(55, 100))
```

```
## Warning: Removed 2 rows containing missing values (`geom_step()`).

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

## Warning: Removed 2 rows containing missing values (`geom_step()`).

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



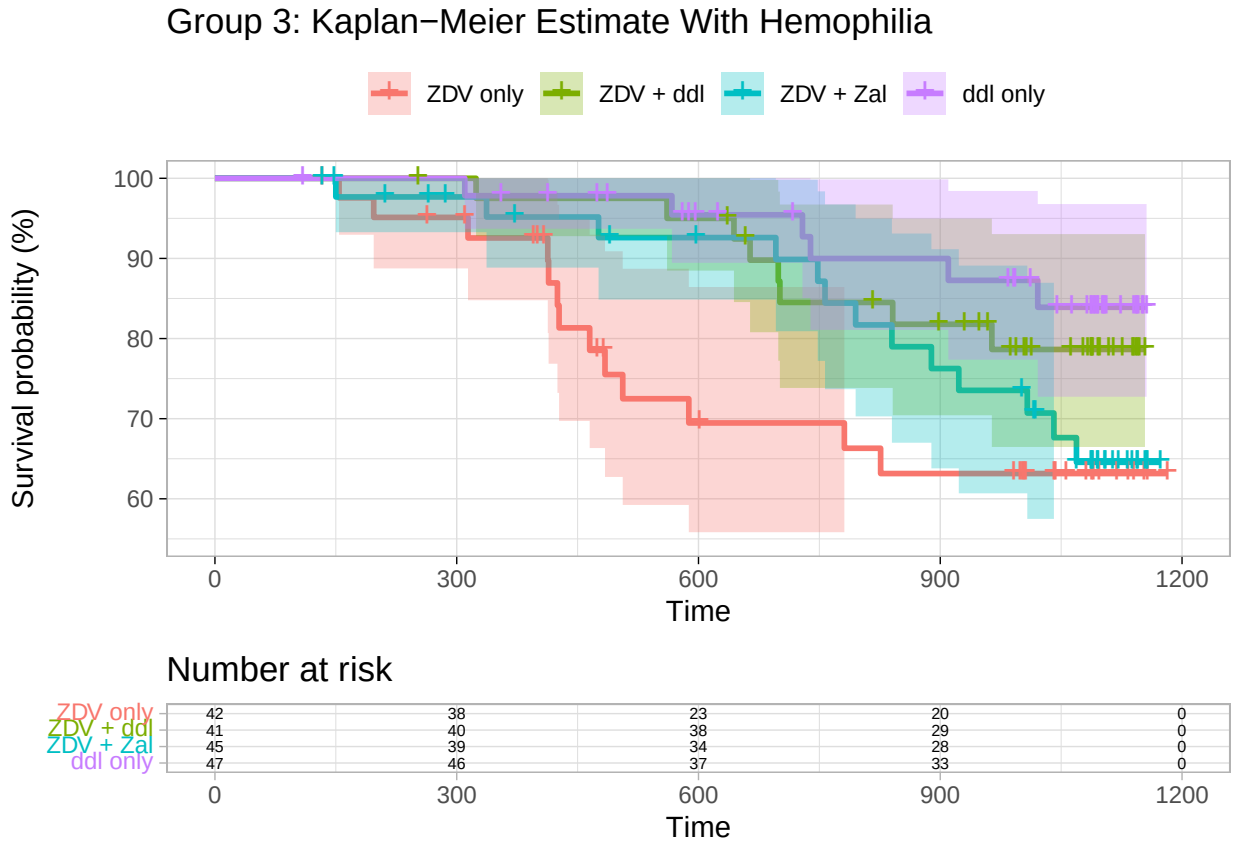

```
# patient with drug used and doesn't have hemophilia
aids_drug = aids%>%
  filter(hemo==0)%>%
  filter(drugs ==1)
```

```
km_fit_trt5 <- survfit(Surv(time, cid) ~ trt, data = aids_drug)
km_fit_trt5 %>% ggsurvplot(data = aids_drug,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Group 2: Kaplan-Meier Estimate With Drug Used",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                      ylim = c(55, 100))
```



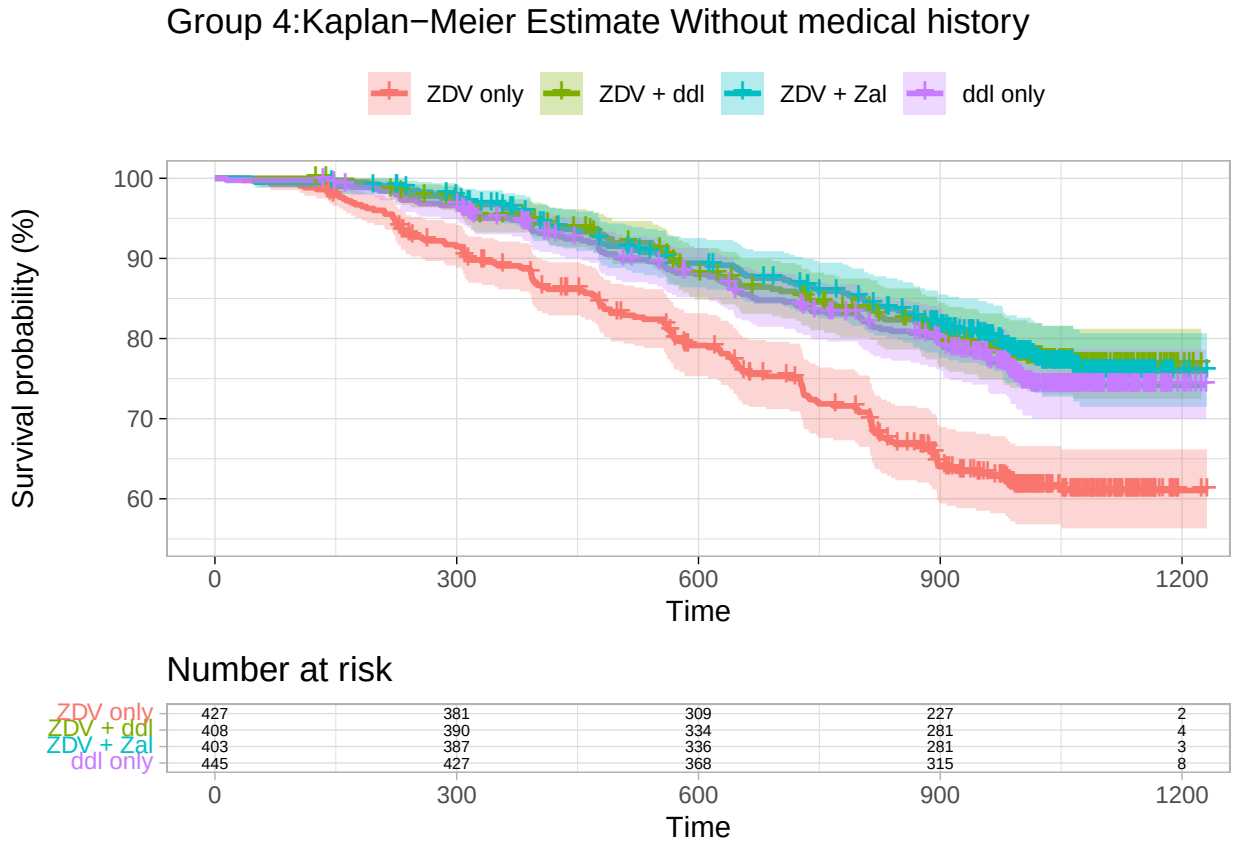

```
# patient without drug used but has hemophilia
aids_hemo = aids%>%
  filter(hemo==1)%>%
  filter(drugs ==0)
```

```
km_fit_trt6 <- survfit(Surv(time, cid) ~ trt, data = aids_hemo)
km_fit_trt6 %>% ggsurvplot(data = aids_hemo,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Group 3: Kaplan-Meier Estimate With Hemophilia",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddI", "ZDV + Zal", "ddI only"),
                      ylim = c(55, 100))
```




```
# patient without drug used and doesn't have hemophilia
aids_nomedical = aids%>%
  filter(hemo==0)%>%
  filter(drugs ==0)
```

```
km_fit_trt7 <- survfit(Surv(time, cid) ~ trt, data = aids_nomedical)
km_fit_trt7 %>% ggsurvplot(data = aids_nomedical,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Group 4:Kaplan-Meier Estimate Without medical history",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddI", "ZDV + Zal", "ddI only"),
                      ylim = c(55, 100))
```




```
survfit_result_drug <- survfit(Surv(time, cid) ~ strata(drugs) + trt, data = aids)
survfit_result_drug
```

```
## Call: survfit(formula = Surv(time, cid) ~ strata(drugs) + trt, data = aids)
##
##                                 n events median 0.95LCL 0.95UCL
## strata(drugs)=drugs=0, trt=0 469    165     NA      NA      NA
## strata(drugs)=drugs=0, trt=1 449     94     NA      NA      NA
## strata(drugs)=drugs=0, trt=2 448     98     NA      NA      NA
## strata(drugs)=drugs=0, trt=3 492    111     NA      NA      NA
## strata(drugs)=drugs=1, trt=0  63     16     NA      NA      NA
## strata(drugs)=drugs=1, trt=1  73      9     NA      NA      NA
## strata(drugs)=drugs=1, trt=2  76     11     NA      NA      NA
## strata(drugs)=drugs=1, trt=3  69     17     NA      NA      NA
```

```
survfit_result_hemo <- survfit(Surv(time, cid) ~ strata(hemo) + trt, data = aids)
survfit_result_hemo
```

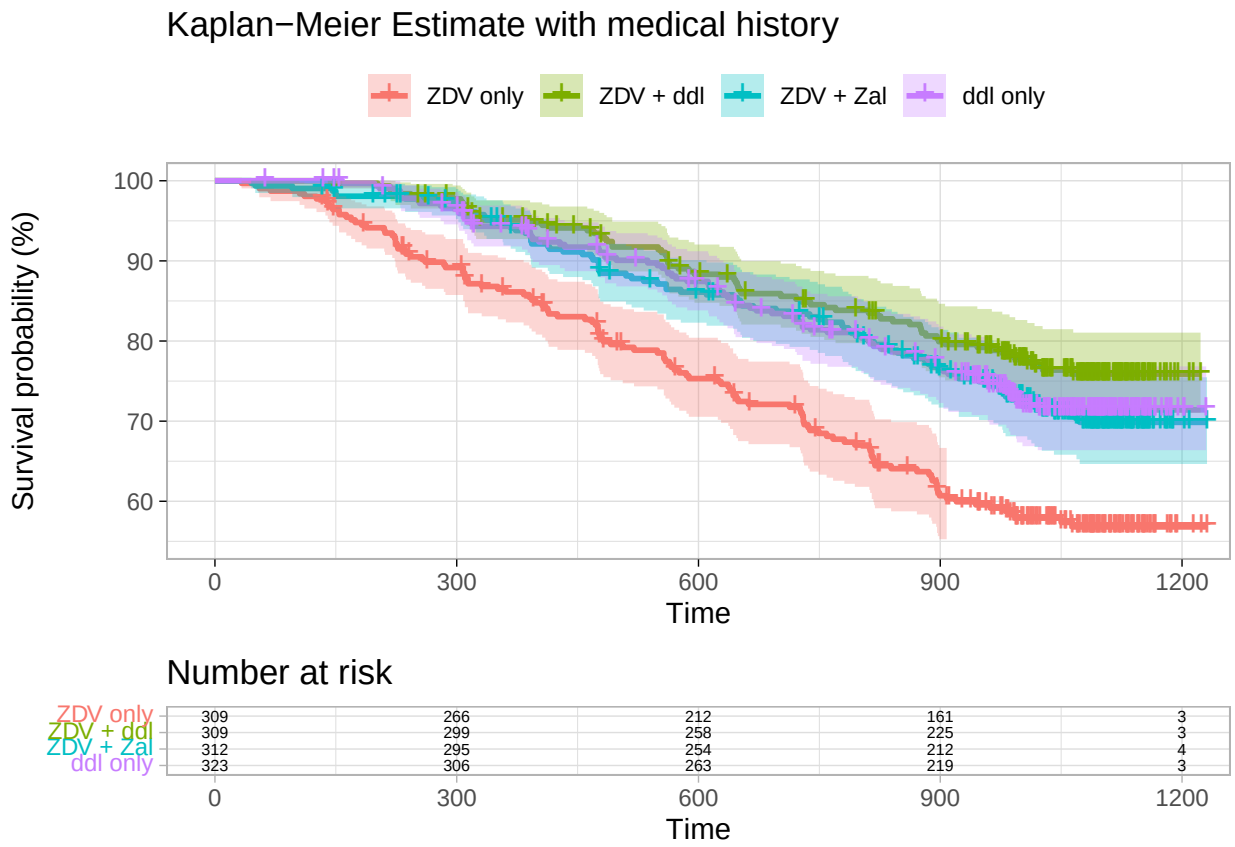
```
## Call: survfit(formula = Surv(time, cid) ~ strata(hemo) + trt, data = aids)
##
##                                n events median 0.95LCL 0.95UCL
## strata(hemo)=hemo=0, trt=0 490    168     NA      NA      NA
## strata(hemo)=hemo=0, trt=1 479     95     NA      NA      NA
## strata(hemo)=hemo=0, trt=2 478     96     NA      NA      NA
## strata(hemo)=hemo=0, trt=3 512    121     NA      NA      NA
```

```
## strata(hemo)=hemo=1, trt=0  42     13     NA     826      NA
## strata(hemo)=hemo=1, trt=1  43      8     NA      NA      NA
## strata(hemo)=hemo=1, trt=2  46     13     NA      NA      NA
## strata(hemo)=hemo=1, trt=3  49      7     NA      NA      NA
```

KM curve for patient with or without azt therapy before

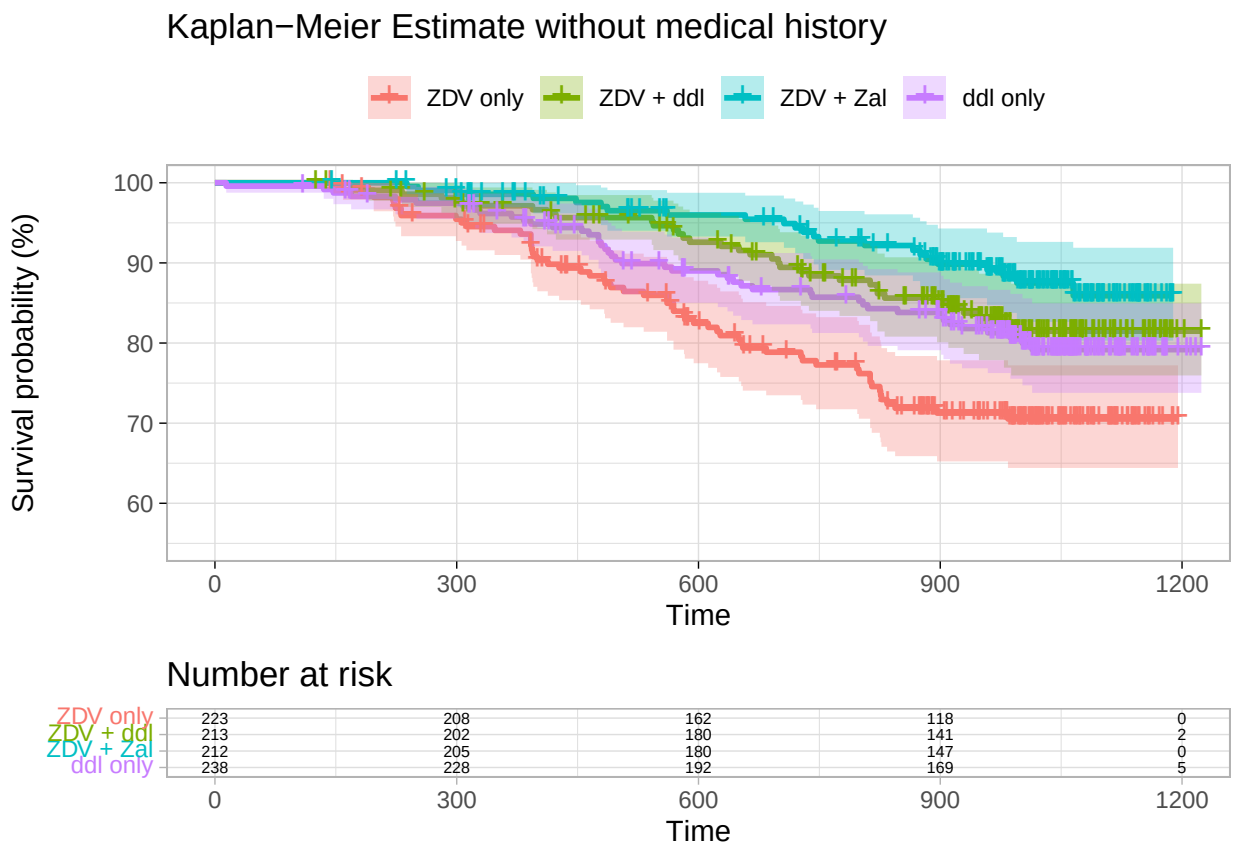
```
# patient with azt theorpy before
aids_azt_yes = aids%>%
  filter(str2==1)
```

```
km_fit_trt8 <- survfit(Surv(time, cid) ~ trt, data = aids_azt_yes)
km_fit_trt8 %>% ggsurvplot(data = aids_azt_yes,
                      fun = "pct",
                      conf.int = TRUE,
                      risk.table = TRUE,
                      fontsize = 2,
                      ggtheme = theme_light(),
                      title = "Kaplan-Meier Estimate with medical history",
                      legend.title = "",
                      legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                      ylim = c(55, 100))
```



```
# patient without azt therapy before
aids_azt_no <-
  aids %>%
  filter(str2==0)
```

```
km_fit_trt9 <- survfit(Surv(time, cid) ~ trt, data = aids_azt_no)
km_fit_trt9 %>% ggsurvplot(data = aids_azt_no,
                       fun = "pct",
                       conf.int = TRUE,
                       risk.table = TRUE,
                       fontsize = 2,
                       ggtheme = theme_light(),
                       title = "Kaplan-Meier Estimate without medical history",
                       legend.title = "",
                       legend.labs = c("ZDV only","ZDV + ddl", "ZDV + Zal", "ddl only"),
                       ylim = c(55, 100))
```




```
survfit_result_azt <- survfit(Surv(time, cid) ~ strata(str2) + trt, data = aids)
survfit_result_azt
```

```
## Call: survfit(formula = Surv(time, cid) ~ strata(str2) + trt, data = aids)
##
##                               n events median 0.95LCL 0.95UCL
## strata(str2)=str2=0, trt=0 223     59     NA      NA      NA
```

```
## strata(str2)=str2=0, trt=1 213     34     NA      NA      NA
## strata(str2)=str2=0, trt=2 212     23     NA      NA      NA
## strata(str2)=str2=0, trt=3 238     44     NA      NA      NA
## strata(str2)=str2=1, trt=0 309    122     NA      NA      NA
## strata(str2)=str2=1, trt=1 309     69     NA      NA      NA
## strata(str2)=str2=1, trt=2 312     86     NA      NA      NA
## strata(str2)=str2=1, trt=3 323     84     NA      NA      NA
```

Cox-PH model

```
aids = read_csv("data/AIDS_Clinical_Trials_Group175.csv") %>% mutate(trt = as.factor(trt),
                                                                     hemo = as.factor(hemo),
                                                                     homo = as.factor(homo),
                                                                     drugs = as.factor(drugs),
                                                                     race = as.factor(race),
                                                                     gender = as.factor(gender),
                                                                     str2 = as.factor(str2),
                                                                     symptom = as.factor(symptom)) %>% c
```

```
## New names:
## Rows: 2139 Columns: 25
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (25): ...1, time, trt, age, wtkg, hemo, homo, drugs, karnof, oprior, z30...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

model selection *selection on personal information*

```
selectCox(Surv(time, cid) ~ trt + age + wtkg + homo + race + gender, data = aids, rule = "aic")
```

```
## $fit
## Cox Proportional Hazards Model
##
## rms::cph(formula = newform, data = data, x = TRUE, y = TRUE,
##     surv = TRUE)
##
##                         Model Tests     Discrimination
##                                               Indexes
## Obs       2139    LR chi2     51.46     R2       0.024
## Events     521    d.f.            4    R2(4,2139)0.022
## Center -0.0148    Pr(> chi2) 0.0000     R2(4,521)0.087
##                   Score chi2  56.26     Dxy      0.179
##                   Pr(> chi2) 0.0000
##
##       Coef    S.E.   Wald Z Pr(>|Z|)
## trt=1 -0.7123 0.1235 -5.77  <0.0001
## trt=2 -0.6464 0.1213 -5.33  <0.0001
## trt=3 -0.5328 0.1155 -4.61  <0.0001
```

```
## age    0.0130 0.0049  2.66  0.0079
##
##
## $In
## [1] "trt" "age"
##
## $call
## selectCox(formula = Surv(time, cid) ~ trt + age + wtkg + homo +
##     race + gender, data = aids, rule = "aic")
##
## attr(,"class")
## [1] "selectCox"
```

$In [1] “trt” “age”

selection on medical history/treatment history/lab results

```
selectCox(Surv(time, cid) ~ trt + hemo + drugs + karnof + str2 + symptom + cd40 + cd80, data = aids, ru
```

```
## $fit
## Cox Proportional Hazards Model
##
## rms::cph(formula = newform, data = data, x = TRUE, y = TRUE,
##     surv = TRUE)
##
##                         Model Tests     Discrimination
##                                               Indexes
## Obs       2139    LR chi2    228.09     R2       0.104
## Events     521    d.f.            9    R2(9,2139)0.097
## Center -3.6773    Pr(> chi2) 0.0000     R2(9,521)0.343
##                   Score chi2 232.44     Dxy      0.393
##                   Pr(> chi2) 0.0000
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## trt=1     -0.7949 0.1241 -6.41  <0.0001
## trt=2     -0.6685 0.1216 -5.50  <0.0001
## trt=3     -0.5744 0.1158 -4.96  <0.0001
## drugs=1   -0.3177 0.1460 -2.18  0.0295
## karnof    -0.0253 0.0069 -3.65  0.0003
## str2=1     0.3769 0.0956  3.94  <0.0001
## symptom=1  0.4061 0.1024  3.97  <0.0001
## cd40      -0.0042 0.0004 -9.30  <0.0001
## cd80       0.0005 0.0001  5.52  <0.0001
##
##
## $In
## [1] "trt"     "drugs"   "karnof"  "str2"    "symptom" "cd40"    "cd80"
##
## $call
## selectCox(formula = Surv(time, cid) ~ trt + hemo + drugs + karnof +
##     str2 + symptom + cd40 + cd80, data = aids, rule = "aic")
##
## attr(,"class")
## [1] "selectCox"
```

$In [1] “trt” “drugs” “karnof” “str2” “symptom” “cd40” “cd80”

selection with interaction

```
selectCox(Surv(time, cid) ~ trt + drugs + karnof + str2 + symptom + cd40 + cd80 + age + age * drugs + a
```

```
## $fit
## Cox Proportional Hazards Model
##
## rms::cph(formula = newform, data = data, x = TRUE, y = TRUE,
##     surv = TRUE)
##
##                       Model Tests      Discrimination
##                                              Indexes
## Obs      2139    LR chi2    231.48      R2       0.105
## Events    521    d.f.           11    R2(11,2139)0.098
## Center -3.289    Pr(> chi2) 0.0000     R2(11,521)0.345
##                  Score chi2 235.70      Dxy      0.395
##                  Pr(> chi2) 0.0000
##
##               Coef    S.E.   Wald Z Pr(>|Z|)
## trt=1         -0.7943 0.1241 -6.40  <0.0001
## trt=2         -0.6671 0.1216 -5.49  <0.0001
## trt=3         -0.5709 0.1159 -4.93  <0.0001
## karnof        -0.0248 0.0070 -3.56  0.0004
## str2=1         0.3731 0.0956  3.90  <0.0001
## symptom=1      0.4048 0.1023  3.96  <0.0001
## cd40          -0.0041 0.0004 -9.23  <0.0001
## cd80           0.0005 0.0001  5.44  <0.0001
## drugs=1        0.4399 0.8198  0.54  0.5916
## age            0.0093 0.0052  1.78  0.0759
## drugs=1 * age -0.0206 0.0218 -0.95  0.3440
##
##
## $In
## [1] "trt"         "karnof"      "str2"        "symptom"     "cd40"
## [6] "cd80"        "drugs * age"
##
## $call
## selectCox(formula = Surv(time, cid) ~ trt + drugs + karnof +
##     str2 + symptom + cd40 + cd80 + age + age * drugs + age *
##     trt, data = aids, rule = "aic")
##
## attr(,"class")
## [1] "selectCox"
```

$In [1] “trt” “karnof” “str2” “symptom” “cd40” “cd80” “drugs * age”

final model

```
cox1 = coxph(Surv(time, cid) ~ trt + karnof + str2 + symptom + cd40 + cd80 + age * drugs, data = aids)
```

```
summary(cox1)
```

```
## Call:
## coxph(formula = Surv(time, cid) ~ trt + karnof + str2 + symptom +
##     cd40 + cd80 + age * drugs, data = aids)
##
##   n= 2139, number of events= 521
##
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt1       -7.943e-01  4.519e-01  1.241e-01 -6.401 1.54e-10 ***
## trt2       -6.671e-01  5.132e-01  1.216e-01 -5.487 4.09e-08 ***
## trt3       -5.709e-01  5.650e-01  1.159e-01 -4.927 8.34e-07 ***
## karnof     -2.478e-02  9.755e-01  6.971e-03 -3.555 0.000378 ***
## str21       3.731e-01  1.452e+00  9.563e-02  3.901 9.58e-05 ***
## symptom1    4.048e-01  1.499e+00  1.023e-01  3.956 7.63e-05 ***
## cd40       -4.108e-03  9.959e-01  4.451e-04 -9.228  < 2e-16 ***
## cd80        4.548e-04  1.000e+00  8.366e-05  5.436 5.45e-08 ***
## age         9.306e-03  1.009e+00  5.242e-03  1.775 0.075864 .
## drugs1      4.399e-01  1.553e+00  8.198e-01  0.537 0.591542
## age:drugs1 -2.061e-02  9.796e-01  2.178e-02 -0.946 0.343950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## trt1          0.4519     2.2129    0.3543    0.5763
## trt2          0.5132     1.9487    0.4044    0.6513
## trt3          0.5650     1.7698    0.4502    0.7091
## karnof        0.9755     1.0251    0.9623    0.9889
## str21         1.4522     0.6886    1.2040    1.7515
## symptom1      1.4990     0.6671    1.2266    1.8320
## cd40          0.9959     1.0041    0.9950    0.9968
## cd80          1.0005     0.9995    1.0003    1.0006
## age           1.0093     0.9907    0.9990    1.0198
## drugs1        1.5526     0.6441    0.3113    7.7419
## age:drugs1    0.9796     1.0208    0.9387    1.0223
##
## Concordance= 0.698  (se = 0.012 )
## Likelihood ratio test= 231.5  on 11 df,   p=<2e-16
## Wald test            = 229.8  on 11 df,   p=<2e-16
## Score (logrank) test = 235.7  on 11 df,   p=<2e-16
```

final model with age_group

```
aids = aids %>% mutate(age_group = as.factor(ifelse(age >= 11 & age <= 30, "11-30", ifelse(age >= 31 & a
```

```
cox2 = coxph(Surv(time, cid) ~ trt + age_group + drugs + karnof + str2 + symptom + cd40 + cd80 + age_gr
```

```
summary(cox2)
```

```
## Call:
## coxph(formula = Surv(time, cid) ~ trt + age_group + drugs + karnof +
```

```
##     str2 + symptom + cd40 + cd80 + age_group * drugs, data = aids)
##
##   n= 2139, number of events= 521
##
##                              coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt1                   -8.001e-01  4.493e-01  1.242e-01 -6.441 1.18e-10 ***
## trt2                   -6.658e-01  5.139e-01  1.216e-01 -5.473 4.43e-08 ***
## trt3                   -5.673e-01  5.670e-01  1.160e-01 -4.891 1.00e-06 ***
## age_group31-50         -9.626e-02  9.082e-01  1.037e-01 -0.928 0.353341
## age_group51-70          3.464e-01  1.414e+00  1.901e-01  1.822 0.068462 .
## drugs1                 -3.665e-01  6.932e-01  3.445e-01 -1.064 0.287443
## karnof                 -2.381e-02  9.765e-01  7.007e-03 -3.398 0.000678 ***
## str21                   3.772e-01  1.458e+00  9.591e-02  3.933 8.37e-05 ***
## symptom1                4.364e-01  1.547e+00  1.031e-01  4.233 2.31e-05 ***
## cd40                   -4.186e-03  9.958e-01  4.476e-04 -9.353  < 2e-16 ***
## cd80                    4.689e-04  1.000e+00  8.296e-05  5.652 1.58e-08 ***
## age_group31-50:drugs1  7.046e-02  1.073e+00  3.832e-01  0.184 0.854096
## age_group51-70:drugs1  4.513e-01  1.570e+00  6.976e-01  0.647 0.517674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## trt1                      0.4493     2.2258    0.3522    0.5731
## trt2                      0.5139     1.9460    0.4049    0.6522
## trt3                      0.5670     1.7635    0.4517    0.7118
## age_group31-50            0.9082     1.1010    0.7412    1.1129
## age_group51-70            1.4139     0.7072    0.9741    2.0523
## drugs1                    0.6932     1.4426    0.3529    1.3617
## karnof                    0.9765     1.0241    0.9632    0.9900
## str21                     1.4583     0.6858    1.2084    1.7598
## symptom1                  1.5471     0.6464    1.2640    1.8935
## cd40                      0.9958     1.0042    0.9949    0.9967
## cd80                      1.0005     0.9995    1.0003    1.0006
## age_group31-50:drugs1     1.0730     0.9320    0.5063    2.2738
## age_group51-70:drugs1     1.5703     0.6368    0.4002    6.1623
##
## Concordance= 0.698  (se = 0.011 )
## Likelihood ratio test= 235.2  on 13 df,   p=<2e-16
## Wald test            = 233.4  on 13 df,   p=<2e-16
## Score (logrank) test = 239.7  on 13 df,   p=<2e-16
```

model checking *check multicollinearity*

```
VIF(cox2)
```

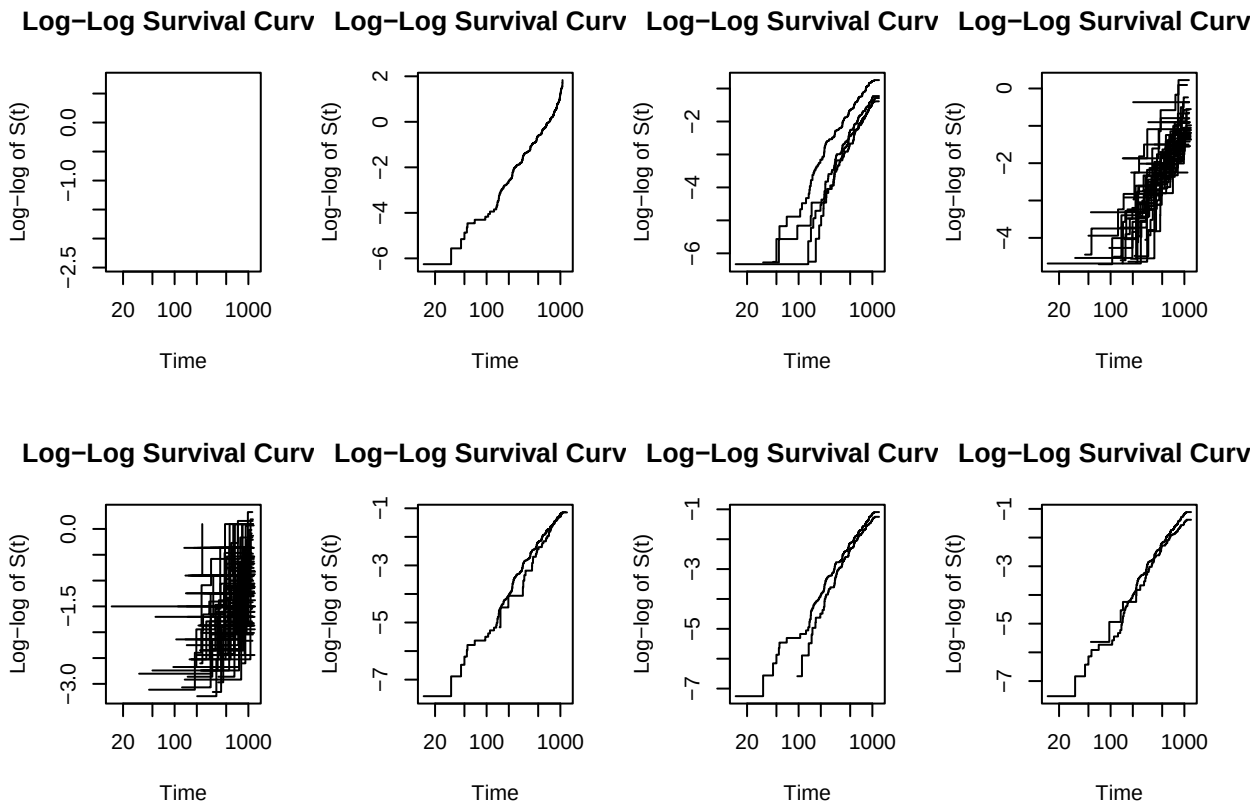
```
## Warning in VIF(cox2): No intercept: vifs may not be sensible.
```

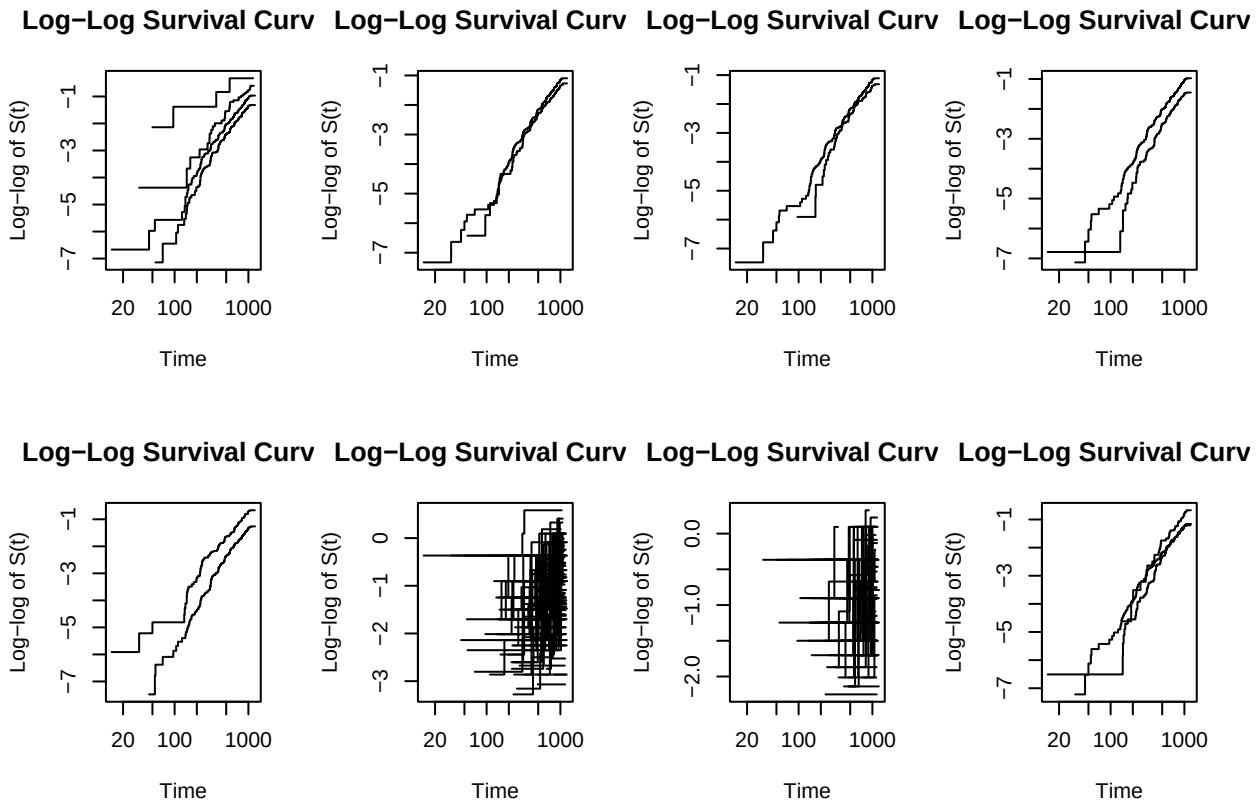
```
##                     GVIF Df GVIF^(1/(2*Df))
## trt             1.021160  3        1.003496
## age_group       1.221872  2        1.051372
## drugs           5.649854  1        2.376942
## karnof          1.065217  1        1.032094
## str2            1.018843  1        1.009378
```

```
## symptom          1.062425  1        1.030740
## cd40             1.106090  1        1.051708
## cd80             1.074564  1        1.036612
## age_group:drugs 6.442779  2        1.593193
```

Plot log-log survival curve

```
par(mfrow = c(2,4))
var_list = names(aids)

for (i in var_list) {
plot(survfit(Surv(time, cid) ~ aids[[i]], data = aids),
fun = 'cloglog',
conf.int = FALSE,
col = 1,
lty = 1,
xlab = "Time",
ylab = "Log-log of S(t)",
main = "Log-Log Survival Curves")
}
```





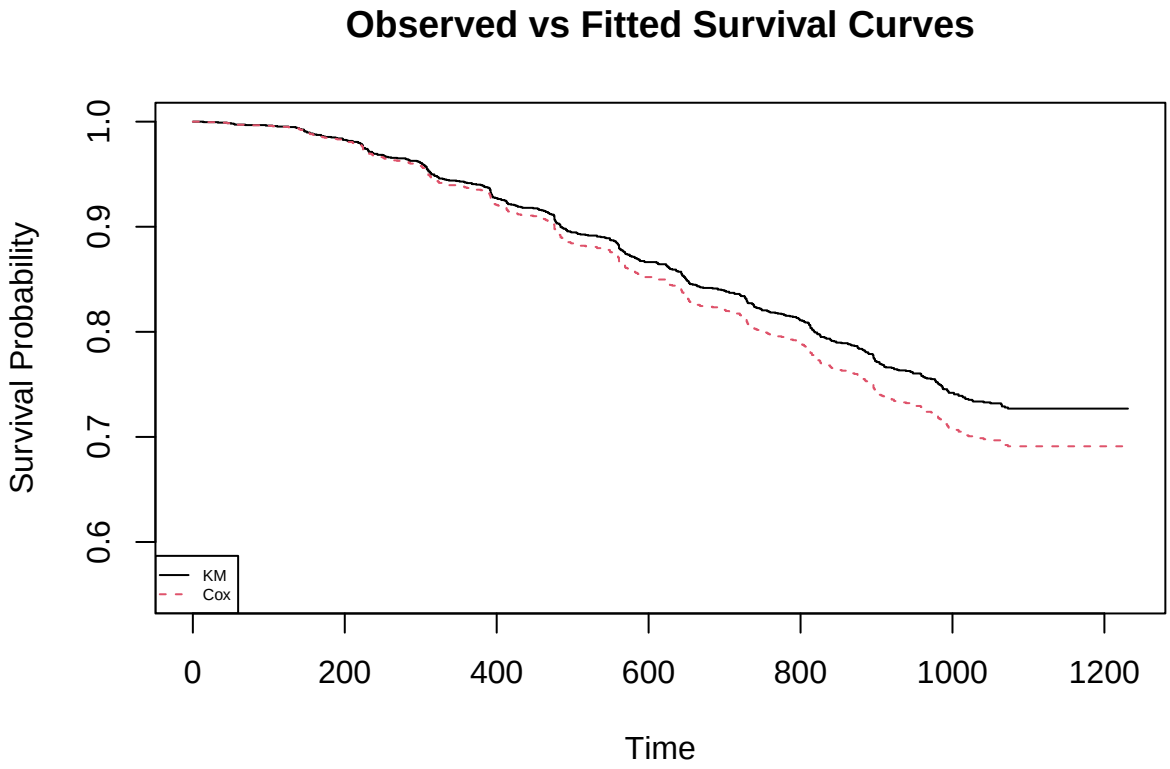
Plot the observed and fitted

```
par(mfrow = c(1,1))

plot(survfit(Surv(time, cid) ~ 1, data = aids),
conf.int = FALSE,
col = 1,
lty = 1,
ylim = c(0.55,1),
xlab = "Time",
ylab = "Survival Probability",
main = "Observed vs Fitted Survival Curves")

lines(survfit(cox2),
conf.int = FALSE,
col = 2,
lty = 2,
ylim = c(0.55,1))
```

```
## Warning in survfit.coxph(cox2): the model contains interactions; the default
## curve based on columm means of the X matrix is almost certainly not useful.
## Consider adding a newdata argument.
```

```
legend("bottomleft",
legend = c("KM", "Cox"),
```

```
col = 1:2,
lty = c(1, 2),
cex = 0.5,
merge = TRUE)
```




```
broom::tidy(cox2) %>%
  mutate(`exp(estimate)` = exp(estimate)) %>%
  relocate(`exp(estimate)`, .after = estimate) #%>%
```

```
## # A tibble: 13 x 6
##    term                   estimate `exp(estimate)` std.error statistic  p.value
##    <chr>                     <dbl>           <dbl>     <dbl>     <dbl>    <dbl>
##  1 trt1                   -0.800             0.449 0.124        -6.44  1.18e-10
##  2 trt2                   -0.666             0.514 0.122        -5.47  4.43e- 8
##  3 trt3                   -0.567             0.567 0.116        -4.89  1.00e- 6
##  4 age_group31-50         -0.0963            0.908 0.104        -0.928 3.53e- 1
##  5 age_group51-70          0.346             1.41  0.190         1.82  6.85e- 2
##  6 drugs1                 -0.366             0.693 0.345        -1.06  2.87e- 1
##  7 karnof                 -0.0238            0.976 0.00701      -3.40  6.78e- 4
##  8 str21                   0.377             1.46  0.0959        3.93  8.37e- 5
##  9 symptom1                0.436             1.55  0.103         4.23  2.31e- 5
## 10 cd40                   -0.00419           0.996 0.000448     -9.35  8.55e-21
## 11 cd80                    0.000469          1.00  0.0000830     5.65  1.58e- 8
## 12 age_group31-50:drugs1   0.0705            1.07  0.383         0.184 8.54e- 1
## 13 age_group51-70:drugs1   0.451             1.57  0.698         0.647 5.18e- 1
```

```
#kable(caption = "Summary of Final Cox-PH model")
```

Table 5: Summary of Final Cox-PH model

Term	Estimate	exp(Estimate)	std.error	Test Statistic	P-value
trt1	-0.8001201	0.4492750	0.1242171	-6.4413036	<0.0000001***
trt2	-0.6657643	0.5138806	0.1216476	-5.4728949	<0.0000001***
trt3	-0.5673093	0.5670492	0.1159872	-4.8911374	0.0000010***
age_group31-50	-0.0962569	0.9082307	0.1037108	-0.9281280	0.3533412
age_group51-70	0.3463719	1.4139283	0.1901102	1.8219527	0.0684622
drugs1	-0.3664675	0.6931787	0.3445059	-1.0637481	0.2874428
karnof	-0.0238122	0.9764691	0.0070067	-3.3984920	0.0006776***
str21	0.3772408	1.4582554	0.0959051	3.9334819	0.0000837***
symptom1	0.4363595	1.5470649	0.1030943	4.2326242	0.0000231***
cd40	-0.0041861	0.9958226	0.0004476	-9.3526642	<0.0000001***
cd80	0.0004689	1.0004690	0.0000830	5.6524859	<0.0000001***
age_group31-50:drugs1	0.0704626	1.0730045	0.3831692	0.1838943	0.8540964
age_group51-70:drugs1	0.4512700	1.5703052	0.6975511	0.6469347	0.5176742