Survival Analysis in R

The survival analysis or time-to-event is a clinical course duration variable for each subject having a beginning and an end anywhere along the time line of the complete study. In this chronic kidney disease (CKD) data, it begins when the patient is enrolled into the study or when the treatment begins, and ends when the event of interest, the progression of CKD, is reached or the patient is censored from the study. This report focuses on the survival analysis on the patients' progression of chronic kidney disease.

Load CKD data

First of all, I use read.csv ( ) function to load all csv files.

```
creatinine <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_creatinine.csv")
DBP <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_DBP.csv")
demo <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_demo.csv")
glucose <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_glucose.csv")
HGB <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_HGB.csv")
ldl <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_ldl.csv")
meds <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_meds.csv")
SBP <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_SBP.csv")
stage <- read.csv("/Users/yunzhao/Desktop/CKDdata/T_stage.csv")
```

Import libraries

This block of code loads the required packages for survival analysis.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyv
erse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.0     ✓ dplyr   1.0.5
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_c
onflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

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

```
library(ggfortify)
library(ranger)
library(caTools)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Data pre-processing

Before the analysis, I join all csv files with left_join function to create the full CKD data set.

```
# Join csv data
CKDdata <- left_join(creatinine,DBP,by=c("id","time")) %>%
  left_join(glucose, by=c("id","time")) %>%
  left_join(HGB, by=c("id","time")) %>%
  left_join(ldl, by=c("id","time")) %>%
  left_join(SBP, by=c("id","time")) %>%
  left_join(demo,by=c("id")) %>%
  left_join(stage,by=c("id")) %>%
  left_join(meds,by=c("id"))

# Rename columns
names(CKDdata) <- c("id","creatinine","time","DBP","glucose",
                    "HGB","ldl","SBP","race","gender","age",
                    "stage","drug","dosage","start_day","end_day")

# stage-True = 1: the patient would progress to CKD
# stage-False = 0: the patient would not progress to CKD
CKDdata$stage <- ifelse(CKDdata$stage=="True",1,0)

# Exclude the "id" column
CKDdata$id <- NULL

head(CKDdata)
```

```
##   creatinine time   DBP glucose   HGB    ldl    SBP    race gender age stage
## 1       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
## 2       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
## 3       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
## 4       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
## 5       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
## 6       1.29    0 95.32    6.24 13.51 161.49 134.11 Unknown   Male  70     1
##           drug dosage start_day end_day
## 1 atorvastatin     10        19     109
## 2 atorvastatin     10       117     207
## 3     losartan    100        19     289
## 4     losartan    100       403     493
## 5     losartan    100       587     677
## 6    metformin   1000        19     109
```

Kaplan-Meier Model

1. Kaplan-Meier Model

The Kaplan-Meier method is the most common statistical method to estimate survival times and probabilities. In preparing Kaplan-Meier survival analysis, each subject is characterized by two variables: (1) serial time; (2) status at the end of the serial time, which means the event of interest (CKD progression) occurrence or censored.

In the first step, I use Surv( ) to build the standard survival object: (1) The variable ‘time’ records all survival time; (2) ‘stage’ indicates whether a patient’s progression of CKD was observed (stage = 1) or that survival time was censored (stage = 0); (3) a “+” after the time in the print out of ‘kmf’ indicates censoring.

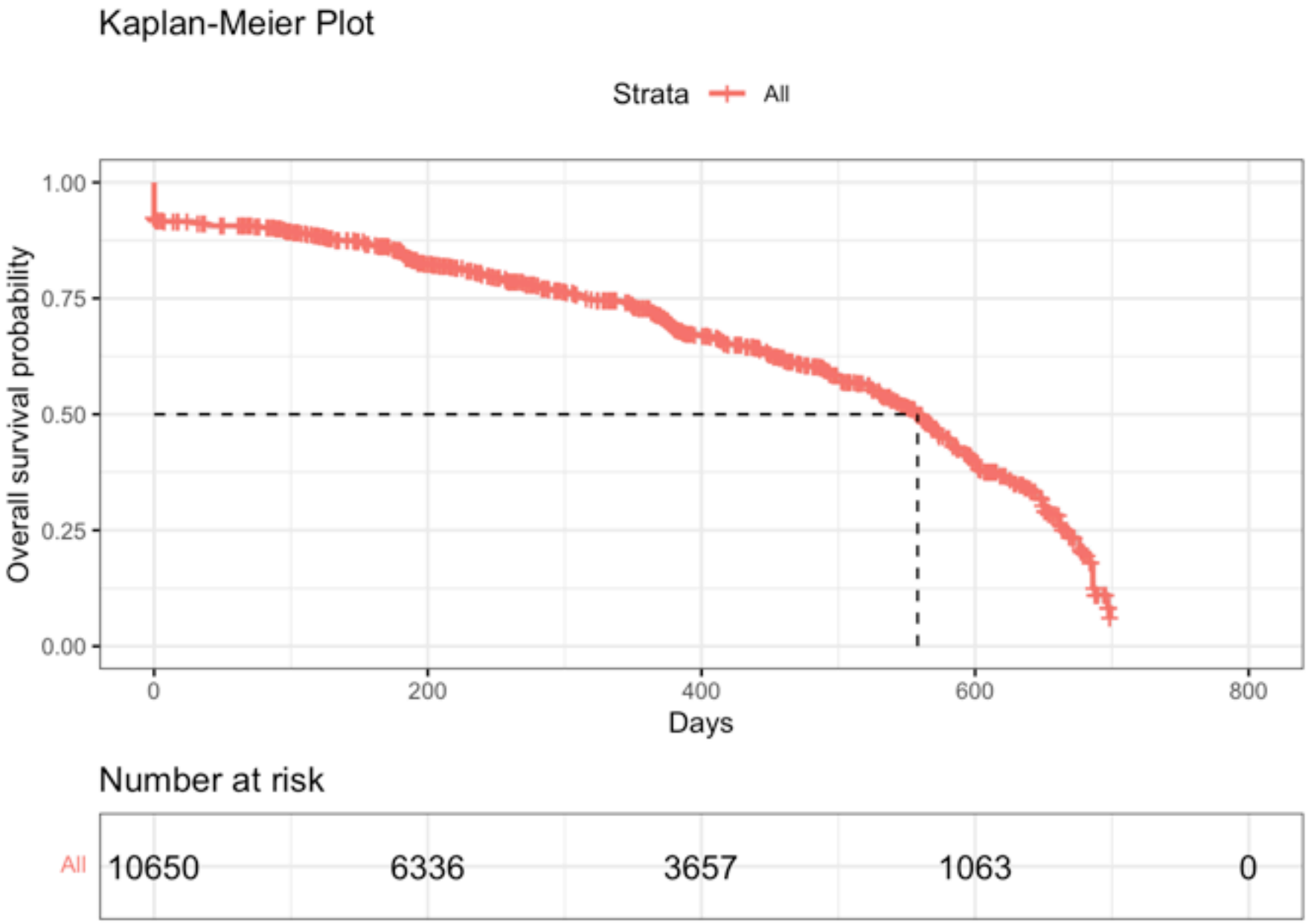
```
kmf <- with(CKDdata, Surv(time, stage))
head(kmf, 100)
```

```
##   [1]   0    0    0    0    0    0    0    0    0    0    0  107  107  107  107
##  [16] 107  107  107  107  107  107  107  286  286  286  286  286  286  286  286
##  [31] 286  286  286  382  382  382  382  382  382  382  382  382  382  382  580
##  [46] 580  580  580  580  580  580  580  580  580  580  688  688  688  688  688
##  [61] 688  688  688  688  688  688    0+   0+   0+   0+  68+  68+  68+  68+ 289
+
##  [76] 289+ 289+ 289+ 387+ 387+ 387+ 387+ 470+ 470+ 470+ 470+   0    0  184  184
##  [91] 430  430  502  502  621  621    0+   0+   0+   0+
```

Then I use Surv(time,stage) ~ 1 and the survfit() function to produce the Kaplan-Meier estimates of the probability of survival over time.

```
kmf_fit <- survfit(Surv(time, stage) ~ 1, data = CKDdata)

# Plot kmf_fit
ggsurvplot(kmf_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           risk.table = TRUE, # add risk table
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           ylab = "Overall survival probability",
           title="Kaplan-Meier Plot")
```

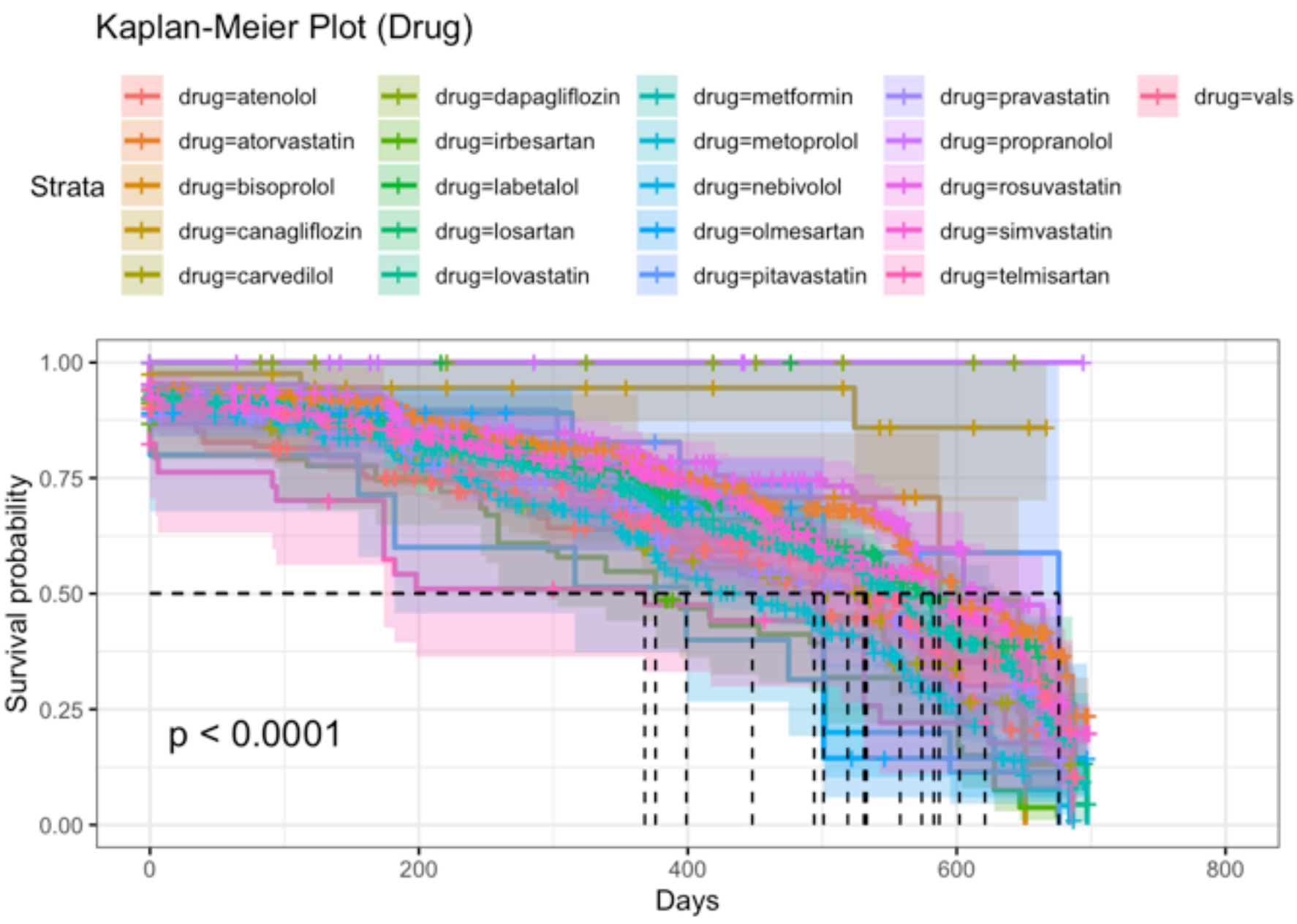
```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord =
pval.coord, : There are no survival curves to be compared.
##  This is a null model.
```



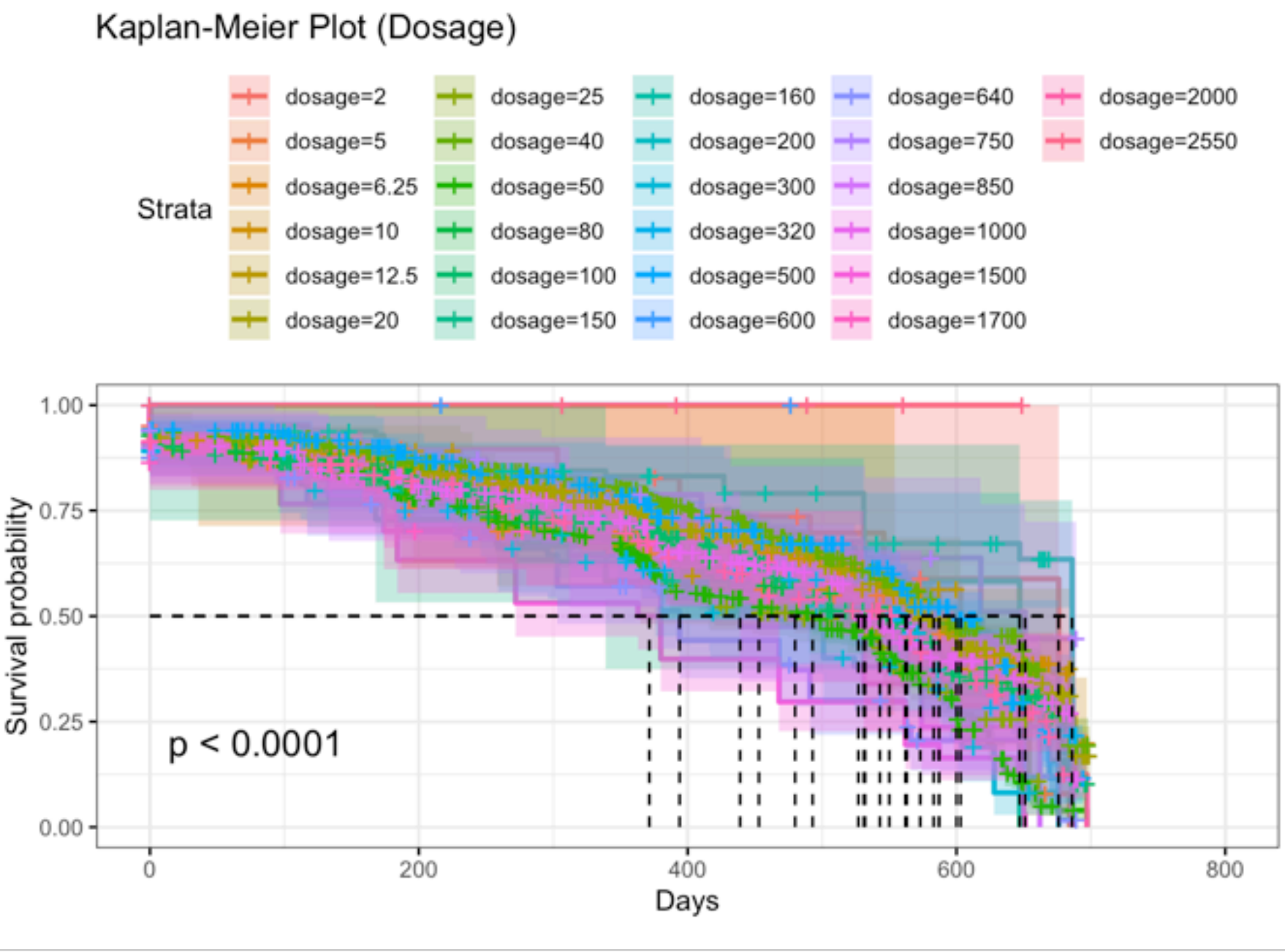
The plot produced by ggsurvplot shows the step function (solid line) with associated confidence bands (shaded area). The horizontal x-axis represents the survival duration for the interval. The vertical y-axis is the change in cumulative probability. Censored observations, inducated by tick marks, reduce the cumulative survival between intervals. The plotting result suggests that the overall patients' survival probability decreases with the increase in time.

Next, I look at survival curves by treatment (drug and dosage), gender and race, respectively. For age, I do a little data munging to look at survival by age (the age is lower than 60 years old and over 60 years old).

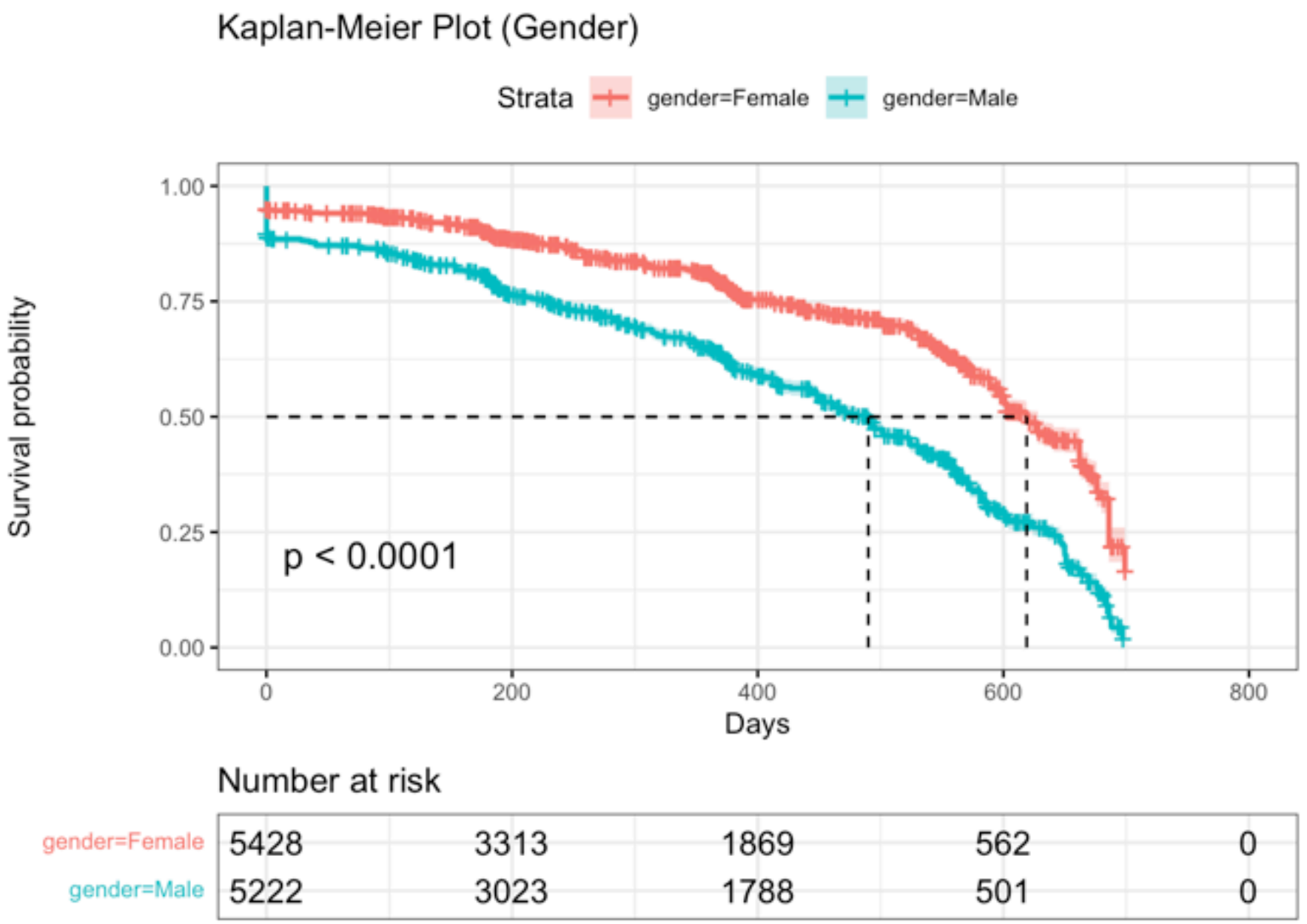
```
# drug
kmf_drug_fit <- survfit(Surv(time, stage) ~ drug, data = CKDdata)
ggsurvplot(kmf_drug_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           tables.height = 0.2, tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           title="Kaplan-Meier Plot (Drug)")
```



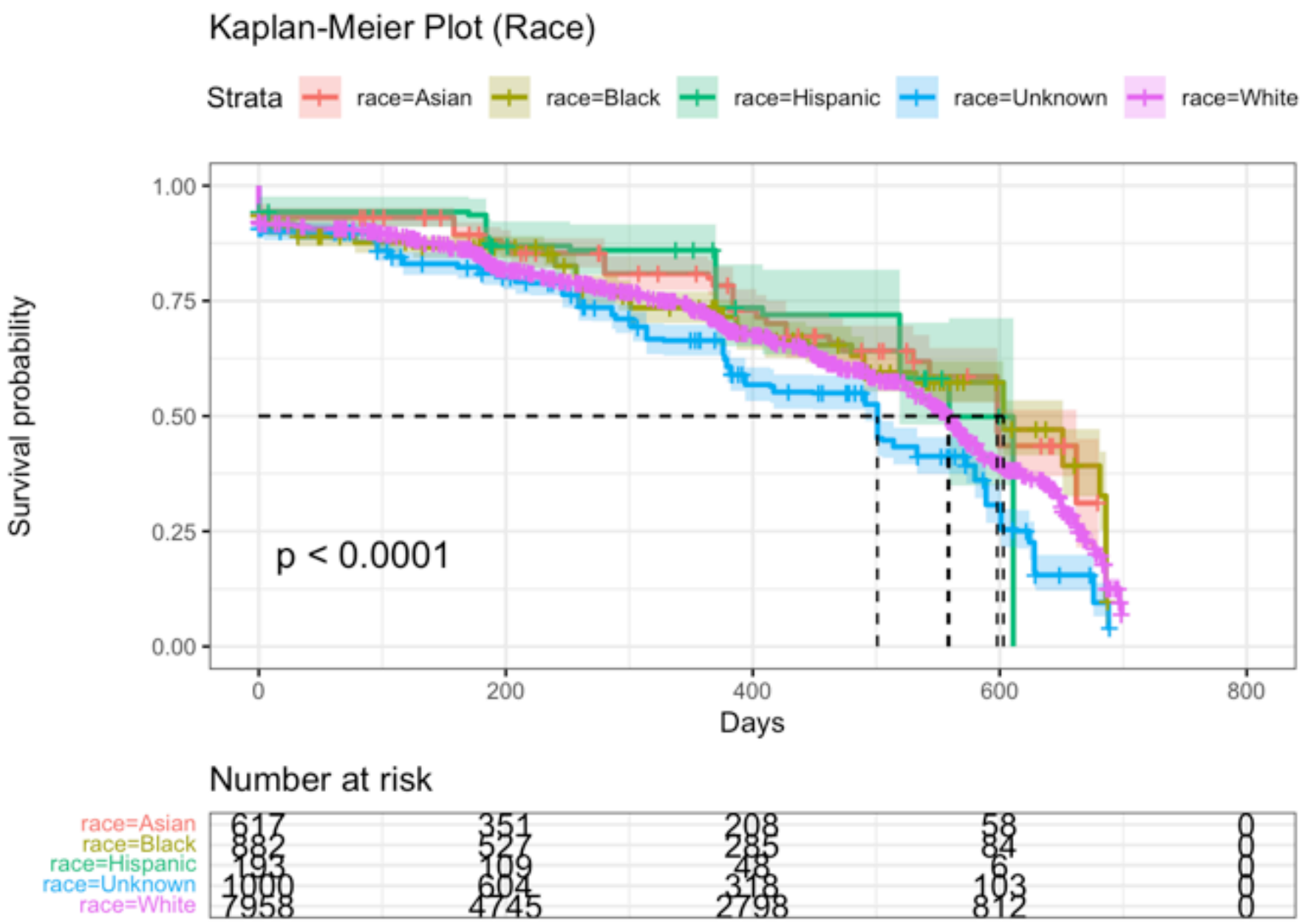

```
# dosage
kmf_dosage_fit <- survfit(Surv(time, stage) ~ dosage, data = CKDdata)
ggsurvplot(kmf_dosage_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           tables.height = 0.2, tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           title="Kaplan-Meier Plot (Dosage)")
```



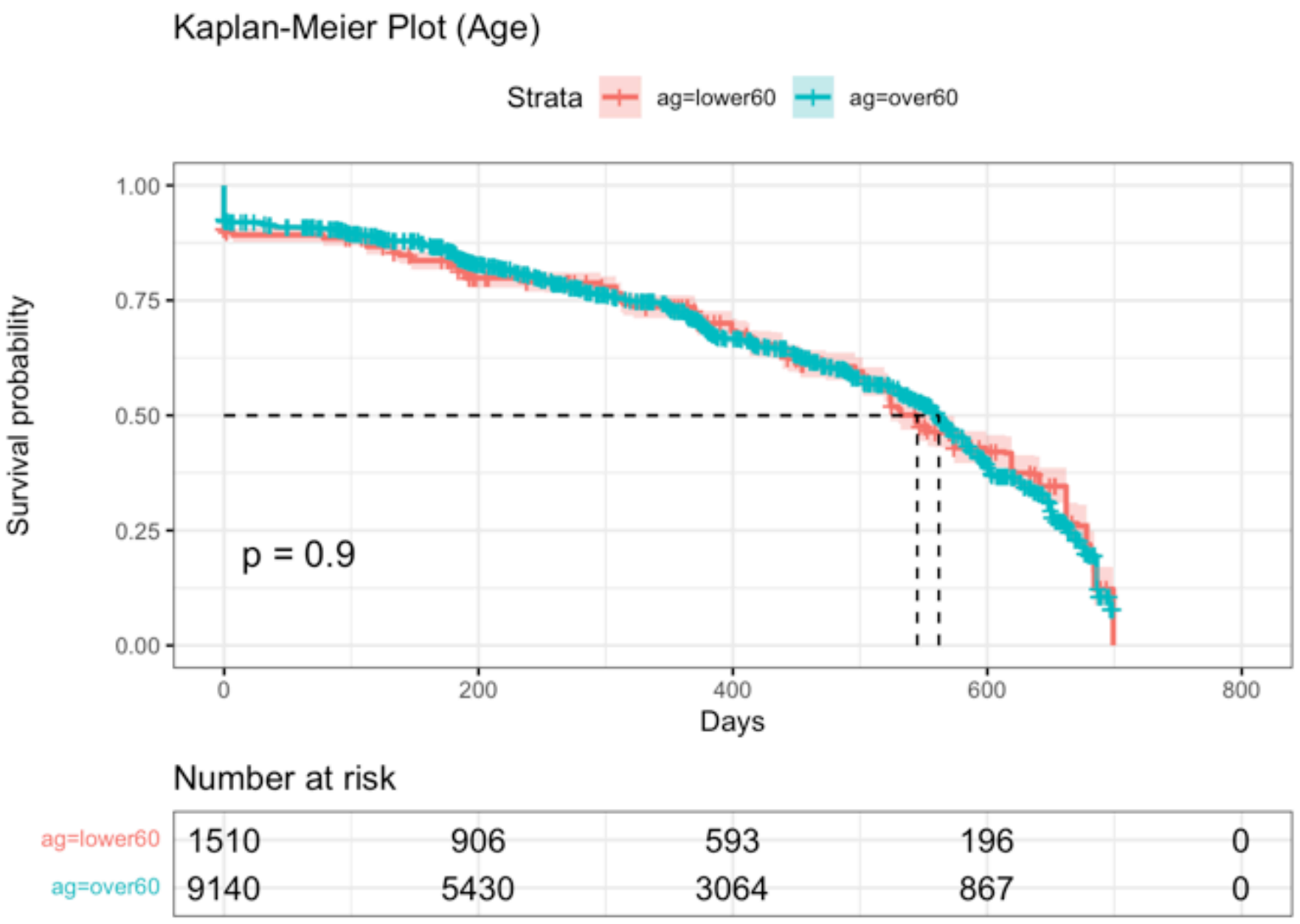

```
# gender
kmf_gender_fit <- survfit(Surv(time, stage) ~ gender, data = CKDdata)
ggsurvplot(kmf_gender_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           risk.table = TRUE, # add risk table
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           title="Kaplan-Meier Plot (Gender)")
```




```
# race
kmf_race_fit <- survfit(Surv(time, stage) ~ race, data = CKDdata)
ggsurvplot(kmf_race_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           risk.table = TRUE, # add risk table
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           title="Kaplan-Meier Plot (Race)")
```




```
# age
dat_age <- mutate(CKDdata, ag=ifelse((age < 60), "lower60", "over60"),
                  ag=factor(ag))
kmf_age_fit <- survfit(Surv(time, stage) ~ ag, data = dat_age)
ggsurvplot(kmf_age_fit,
           surv.median.line = "hv", # add medians survival
           pval = TRUE, # add p-value and tervals
           conf.int = TRUE,
           risk.table = TRUE, # add risk table
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           ggtheme = theme_bw(),
           xlab = "Days",
           title="Kaplan-Meier Plot (Age)")
```



Through all ploting results, it only can be concluded that female patients clearly have a better chance of surviving than male patients. Curves among other groups appear to overlap.

2. Estimating x-year survival

One quantity of interest in a survival analysis is the probability of surviving beyond a certain number (x) of years. In this CKD data, the aim is to estimate the probability of patients' surviving to one year before they progress to CKD. I use summary ( ) function with "time" argument (The "time" variables are in days, so I use times=365.25).

```
summary(survfit(Surv(time, stage) ~ 1, data=CKDdata), times = 365.25)
```

```
## Call: survfit(formula = Surv(time, stage) ~ 1, data = CKDdata)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365   4465    2371     0.72 0.00504         0.71         0.73
```

The summary result suggests that the one-year probability of survival before CKD progression in this dataset is 72%.

3. Estimating median survival time

The second quantity of interest in a survival analysis is the average survival time, which I quantify using the median. Survival times are not expected to be normally distributed, so the mean is not an appropriate summary. I obtain the median survival time of 558 days from survfit object. The median survival time is the time corresponding to a survival probability of 0.5.

```
survfit(Surv(time, stage) ~ 1, data = CKDdata)
```

```
## Call: survfit(formula = Surv(time, stage) ~ 1, data = CKDdata)
##
##       n  events  median 0.95LCL 0.95UCL
##   10650    4153     558     553     563
```

4. Comparing survival times between groups

I conduct the between-group significance test 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. I get the log-rank p-value using the survdiff function according to gender, race, age, and treatment (drug and dosage) in the CKD data set.

```
# Survival times between groups
survdiff(Surv(time, stage) ~ gender, data=CKDdata)
```

```
## Call:
## survdiff(formula = Surv(time, stage) ~ gender, data = CKDdata)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=Female 5428     1457     2157       227       486
## gender=Male   5222     2696     1996       245       486
##
##  Chisq= 486  on 1 degrees of freedom, p= <2e-16
```

```
survdiff(Surv(time, stage) ~ race, data=CKDdata)
```

```
## Call:
## survdiff(formula = Surv(time, stage) ~ race, data = CKDdata)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## race=Asian       617      169    225.0   13.9219   15.0787
## race=Black       882      287    335.4    6.9898    7.8249
## race=Hispanic    193       44     58.5    3.6106    3.7579
## race=Unknown    1000      515    389.5   40.4631   45.9193
## race=White      7958     3138   3144.6    0.0139    0.0589
##
##  Chisq= 66.8  on 4 degrees of freedom, p= 1e-13
```

```
survdiff(Surv(time, stage) ~ ag, data = dat_age)
```

```
## Call:
## survdiff(formula = Surv(time, stage) ~ ag, data = dat_age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## ag=lower60 1510      655      652   0.01400    0.0171
## ag=over60  9140     3498     3501   0.00261    0.0171
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

```
survdiff(Surv(time, stage) ~ drug, data=CKDdata)
```

```
## Call:
## survdiff(formula = Surv(time, stage) ~ drug, data = CKDdata)
##
## n=10516, 134 observations deleted due to missingness.
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## drug=atenolol       524      263   203.02  1.77e+01  1.91e+01
## drug=atorvastatin  1649      453   645.12  5.72e+01  6.97e+01
## drug=bisoprolol      66       30    29.98  1.04e-05  1.07e-05
## drug=canagliflozin   41        3    16.60  1.11e+01  1.14e+01
## drug=carvedilol     284      132   102.79  8.30e+00  8.72e+00
## drug=dapagliflozin   16        0     6.22  6.22e+00  6.36e+00
## drug=irbesartan      68       60    30.71  2.79e+01  2.87e+01
## drug=labetalol       15        0     3.88  3.88e+00  3.99e+00
## drug=losartan      1182      393   443.53  5.76e+00  6.62e+00
## drug=lovastatin     213       83    84.32  2.07e-02  2.18e-02
## drug=metformin     2357      965   943.18  5.05e-01  6.74e-01
## drug=metoprolol    1031      579   375.67  1.10e+02  1.24e+02
## drug=nebivolol       35       35    17.20  1.84e+01  1.89e+01
## drug=olmesartan     137       45    38.26  1.19e+00  1.24e+00
## drug=pitavastatin    21        6     8.57  7.73e-01  7.91e-01
## drug=pravastatin    466      240   198.75  8.56e+00  9.23e+00
## drug=propranolol     59        0    25.28  2.53e+01  2.71e+01
## drug=rosuvastatin   500      117   182.36  2.34e+01  2.51e+01
## drug=simvastatin   1305      445   520.04  1.08e+01  1.27e+01
## drug=telmisartan     34       29    17.15  8.18e+00  8.62e+00
## drug=valsartan      513      233   218.37  9.81e-01  1.07e+00
##
##  Chisq= 356  on 20 degrees of freedom, p= <2e-16
```

```
survdiff(Surv(time, stage) ~ dosage, data=CKDdata)
```

```
## Call:
## survdiff(formula = Surv(time, stage) ~ dosage, data = CKDdata)
##
## n=10516, 134 observations deleted due to missingness.
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## dosage=2      21        6     8.57  7.73e-01  7.91e-01
## dosage=5     137       71    59.11  2.39e+00  2.48e+00
## dosage=6.25   17        5     5.09  1.63e-03  1.67e-03
## dosage=10    827      244   297.34  9.57e+00  1.06e+01
## dosage=12.5  121       41    48.68  1.21e+00  1.26e+00
## dosage=20   1492      493   591.29  1.63e+01  1.97e+01
## dosage=25    414      193   145.28  1.57e+01  1.66e+01
## dosage=40   1714      529   689.08  3.72e+01  4.59e+01
## dosage=50   1158      573   402.72  7.20e+01  8.21e+01
## dosage=80    314      144   129.01  1.74e+00  1.84e+00
## dosage=100  1240      510   492.04  6.56e-01  7.64e-01
## dosage=150    16        8     5.76  8.67e-01  8.88e-01
## dosage=160   160       37    70.73  1.61e+01  1.69e+01
## dosage=200    87       45    38.80  9.89e-01  1.03e+00
## dosage=300    77       55    33.93  1.31e+01  1.35e+01
## dosage=320   237      107    98.98  6.49e-01  6.85e-01
## dosage=500   844      234   293.19  1.19e+01  1.32e+01
## dosage=600    15        0     3.88  3.88e+00  3.99e+00
## dosage=640   112       85    47.52  2.96e+01  3.07e+01
## dosage=750    51       16    31.09  7.32e+00  7.76e+00
## dosage=850    61       36    25.57  4.26e+00  4.38e+00
## dosage=1000  757      304   304.32  3.28e-04  3.66e-04
## dosage=1500  128      128    58.87  8.12e+01  8.41e+01
## dosage=1700   96       54    44.67  1.95e+00  2.08e+00
## dosage=2000  396      193   173.16  2.27e+00  2.43e+00
## dosage=2550   24        0    12.33  1.23e+01  1.26e+01
##
##  Chisq= 353  on 25 degrees of freedom, p= <2e-16
```

Cox Regression

Cox proportional hazards model

The Cox proportional hazards model is a semi-parametric model that can be used to fit univariable and multivariable regression models that have survival outcomes. In this part:

1. I fit a Cox proportional hazards model that makes use of all the covariates in the data. I split the CKD data set into training and test subsets to make prediction and visualize the results. I also compute Confusion Matrix and ROC curves to access the model performance.
2. Hazard ratios The quantity of interest from a Cox regression is a hazard ratio (HR). The HR represents the ratio of hazards between two groups at any particular point in time.The regression parameter (coef) from column estimate in the coxph, then HR = exp(coef). A HR ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

```
table(CKDdata$stage)
```

```
##
##    0    1
## 6497 4153
```

```
# Split CKD data set
split_data <- sample(1:nrow(CKDdata), 0.8 * nrow(CKDdata), FALSE)
training_set <- CKDdata[split_data,]
test_set <- CKDdata[-split_data,]

# check 1 by 1 for infinite coefficients
training_set$id <- NULL
names(training_set)
```

```
##  [1] "creatinine" "time"       "DBP"        "glucose"    "HGB"
##  [6] "ldl"        "SBP"        "race"       "gender"     "age"
## [11] "stage"      "drug"       "dosage"
```

```
coxph(Surv(time,stage) ~ DBP, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ DBP, data = training_set)
##
##         coef exp(coef) se(coef)     z     p
## DBP 0.002309  1.002312 0.002935 0.787 0.431
##
## Likelihood ratio test=0.62  on 1 df, p=0.432
## n= 2203, number of events= 866
##    (6317 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ glucose, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ glucose, data = training_set)
##
##             coef exp(coef) se(coef)     z      p
## glucose 0.15644   1.16934  0.01245 12.57 <2e-16
##
## Likelihood ratio test=145.9  on 1 df, p=< 2.2e-16
## n= 5764, number of events= 2087
##    (2756 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ HGB, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ HGB, data = training_set)
##
##         coef exp(coef) se(coef)      z      p
## HGB -0.13734   0.87168  0.01402 -9.799 <2e-16
##
## Likelihood ratio test=94.82  on 1 df, p=< 2.2e-16
## n= 4627, number of events= 1796
##    (3893 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ ldl, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ ldl, data = training_set)
##
##          coef exp(coef)  se(coef)     z        p
## ldl 0.0027724 1.0027762 0.0008009 3.461 0.000537
##
## Likelihood ratio test=11.66  on 1 df, p=0.000637
## n= 5961, number of events= 2136
##    (2559 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ SBP, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ SBP, data = training_set)
##
##           coef  exp(coef)   se(coef)      z     p
## SBP -0.0007691  0.9992312  0.0022778 -0.338 0.736
##
## Likelihood ratio test=0.11  on 1 df, p=0.7354
## n= 2191, number of events= 856
##    (6329 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ race, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ race, data = training_set)
##
##                   coef exp(coef)  se(coef)      z        p
## raceBlack     0.192729  1.212554  0.107479  1.793 0.072945
## raceHispanic -0.006723  0.993299  0.189426 -0.035 0.971687
## raceUnknown   0.613832  1.847498  0.098456  6.235 4.53e-10
## raceWhite     0.326306  1.385839  0.087484  3.730 0.000192
##
## Likelihood ratio test=56.02  on 4 df, p=1.985e-11
## n= 8520, number of events= 3338
```

```
coxph(Surv(time,stage) ~ gender, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ gender, data = training_set)
##
##                coef exp(coef) se(coef)     z      p
## genderMale 0.69857   2.01088  0.03628 19.25 <2e-16
##
## Likelihood ratio test=389.6  on 1 df, p=< 2.2e-16
## n= 8520, number of events= 3338
```

```
coxph(Surv(time,stage) ~ age, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ age, data = training_set)
##
##          coef exp(coef) se(coef)     z     p
## age 0.001107  1.001108 0.001795 0.617 0.537
##
## Likelihood ratio test=0.38  on 1 df, p=0.5371
## n= 8520, number of events= 3338
```

```
coxph(Surv(time,stage) ~ drug, data=training_set)
```

```
## Warning in fitter(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 5,7,16 ; coefficient may be infinite.
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ drug, data = training_set)
##
##                          coef  exp(coef)   se(coef)      z        p
## drugatorvastatin  -5.495e-01  5.772e-01  8.716e-02 -6.305 2.89e-10
## drugbisoprolol    -2.142e-01  8.072e-01  2.082e-01 -1.029  0.30365
## drugcanagliflozin -2.669e+00  6.931e-02  1.002e+00 -2.663  0.00775
## drugcarvedilol     3.334e-02  1.034e+00  1.178e-01  0.283  0.77711
## drugdapagliflozin -1.527e+01  2.336e-07  7.750e+02 -0.020  0.98428
## drugirbesartan     3.880e-01  1.474e+00  1.604e-01  2.419  0.01556
## druglabetalol     -1.527e+01  2.334e-07  9.541e+02 -0.016  0.98723
## druglosartan      -3.806e-01  6.834e-01  8.935e-02 -4.260 2.05e-05
## druglovastatin    -1.442e-01  8.657e-01  1.356e-01 -1.064  0.28737
## drugmetformin     -1.862e-01  8.301e-01  7.823e-02 -2.379  0.01734
## drugmetoprolol     2.378e-01  1.268e+00  8.383e-02  2.837  0.00456
## drugnebivolol      4.542e-01  1.575e+00  2.050e-01  2.216  0.02672
## drugolmesartan    -6.645e-02  9.357e-01  1.854e-01 -0.358  0.72007
## drugpitavastatin  -1.179e+00  3.074e-01  5.816e-01 -2.028  0.04255
## drugpravastatin   -5.980e-02  9.419e-01  1.014e-01 -0.590  0.55518
## drugpropranolol   -1.528e+01  2.310e-07  4.404e+02 -0.035  0.97232
## drugrosuvastatin  -6.692e-01  5.121e-01  1.241e-01 -5.393 6.93e-08
## drugsimvastatin   -3.996e-01  6.706e-01  8.736e-02 -4.574 4.77e-06
## drugtelmisartan    4.346e-01  1.544e+00  2.292e-01  1.896  0.05800
## drugvalsartan     -1.355e-01  8.732e-01  1.027e-01 -1.320  0.18692
##
## Likelihood ratio test=301.6  on 20 df, p=< 2.2e-16
## n= 8417, number of events= 3307
##    (103 observations deleted due to missingness)
```

```
coxph(Surv(time,stage) ~ dosage, data=training_set)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ dosage, data = training_set)
##
##              coef exp(coef)  se(coef)     z        p
## dosage 1.255e-04 1.000e+00 3.118e-05 4.024 5.73e-05
##
## Likelihood ratio test=15.46  on 1 df, p=8.413e-05
## n= 8417, number of events= 3307
##    (103 observations deleted due to missingness)
```

```
# Cox regression model_1st round
cox1 <- coxph(Surv(time, stage) ~ ., data=training_set)
```

```
## Warning in fitter(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 17,19,26,28 ; coefficient may be infinite.
```

```
summary(cox1)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ ., data = training_set)
##
##   n= 1836, number of events= 686
##    (6684 observations deleted due to missingness)
##
##                        coef  exp(coef)   se(coef)      z Pr(>|z|)
## creatinine         4.180e-03  1.004e+00  1.378e-01  0.030 0.975804
## DBP                9.806e-03  1.010e+00  4.337e-03  2.261 0.023768 *
## glucose            5.507e-02  1.057e+00  2.725e-02  2.021 0.043290 *
## HGB               -1.072e-01  8.983e-01  3.032e-02 -3.536 0.000406 ***
## ldl               -9.647e-03  9.904e-01  1.632e-03 -5.910 3.42e-09 ***
## SBP               -1.107e-02  9.890e-01  3.431e-03 -3.225 0.001260 **
## raceBlack         -6.071e-01  5.449e-01  2.505e-01 -2.424 0.015363 *
## raceHispanic      -2.808e-01  7.552e-01  4.086e-01 -0.687 0.491912
## raceUnknown        8.411e-01  2.319e+00  2.319e-01  3.627 0.000287 ***
## raceWhite          2.167e-01  1.242e+00  2.037e-01  1.064 0.287461
## genderMale         8.830e-01  2.418e+00  1.055e-01  8.367  < 2e-16 ***
## age               -2.044e-02  9.798e-01  4.417e-03 -4.628 3.69e-06 ***
## drugatorvastatin  -4.526e-01  6.360e-01  2.033e-01 -2.226 0.026006 *
## drugbisoprolol     2.121e-01  1.236e+00  5.319e-01  0.399 0.690103
## drugcanagliflozin -1.365e+00  2.553e-01  1.019e+00 -1.340 0.180154
## drugcarvedilol    -9.429e-02  9.100e-01  2.859e-01 -0.330 0.741527
## drugdapagliflozin -1.512e+01  2.718e-07  1.581e+03 -0.010 0.992370
## drugirbesartan     1.076e+00  2.932e+00  4.556e-01  2.361 0.018213 *
## druglabetalol     -1.580e+01  1.372e-07  1.898e+03 -0.008 0.993359
## druglosartan      -4.459e-01  6.402e-01  2.166e-01 -2.058 0.039550 *
## druglovastatin     2.150e-01  1.240e+00  3.069e-01  0.701 0.483534
## drugmetformin     -5.186e-01  5.954e-01  2.385e-01 -2.174 0.029698 *
## drugmetoprolol     2.423e-01  1.274e+00  2.042e-01  1.187 0.235392
## drugnebivolol     -3.884e-01  6.781e-01  3.500e-01 -1.110 0.267127
## drugolmesartan    -5.379e-01  5.840e-01  3.650e-01 -1.474 0.140584
## drugpitavastatin  -1.462e+01  4.461e-07  2.325e+03 -0.006 0.994982
## drugpravastatin    6.747e-01  1.964e+00  2.335e-01  2.890 0.003851 **
## drugpropranolol   -1.623e+01  8.938e-08  1.237e+03 -0.013 0.989531
## drugrosuvastatin  -9.156e-01  4.003e-01  2.927e-01 -3.129 0.001756 **
## drugsimvastatin   -4.384e-01  6.450e-01  2.154e-01 -2.036 0.041774 *
## drugtelmisartan    1.154e+00  3.172e+00  4.867e-01  2.372 0.017710 *
## drugvalsartan     -2.212e-02  9.781e-01  2.399e-01 -0.092 0.926544
## dosage             4.103e-04  1.000e+00  1.343e-04  3.056 0.002245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## creatinine        1.004e+00  9.958e-01   0.76646    1.3156
## DBP               1.010e+00  9.902e-01   1.00131    1.0185
## glucose           1.057e+00  9.464e-01   1.00166    1.1146
## HGB               8.983e-01  1.113e+00   0.84649    0.9533
## ldl               9.904e-01  1.010e+00   0.98724    0.9936
## SBP               9.890e-01  1.011e+00   0.98237    0.9957
## raceBlack         5.449e-01  1.835e+00   0.33352    0.8903
## raceHispanic      7.552e-01  1.324e+00   0.33902    1.6821
## raceUnknown       2.319e+00  4.312e-01   1.47192    3.6531
```

```
## raceWhite          1.242e+00  8.052e-01   0.83314    1.8513
## genderMale         2.418e+00  4.135e-01   1.96635    2.9740
## age                9.798e-01  1.021e+00   0.97132    0.9883
## drugatorvastatin   6.360e-01  1.572e+00   0.42698    0.9473
## drugbisoprolol     1.236e+00  8.089e-01   0.43584    3.5066
## drugcanagliflozin  2.553e-01  3.916e+00   0.03468    1.8799
## drugcarvedilol     9.100e-01  1.099e+00   0.51965    1.5936
## drugdapagliflozin  2.718e-07  3.679e+06   0.00000       Inf
## drugirbesartan     2.932e+00  3.410e-01   1.20062    7.1622
## druglabetalol      1.372e-07  7.288e+06   0.00000       Inf
## druglosartan       6.402e-01  1.562e+00   0.41872    0.9789
## druglovastatin     1.240e+00  8.065e-01   0.67943    2.2627
## drugmetformin      5.954e-01  1.680e+00   0.37304    0.9502
## drugmetoprolol     1.274e+00  7.848e-01   0.85389    1.9015
## drugnebivolol      6.781e-01  1.475e+00   0.34151    1.3466
## drugolmesartan     5.840e-01  1.712e+00   0.28555    1.1943
## drugpitavastatin   4.461e-07  2.242e+06   0.00000       Inf
## drugpravastatin    1.964e+00  5.093e-01   1.24253    3.1028
## drugpropranolol    8.938e-08  1.119e+07   0.00000       Inf
## drugrosuvastatin   4.003e-01  2.498e+00   0.22555    0.7103
## drugsimvastatin    6.450e-01  1.550e+00   0.42293    0.9838
## drugtelmisartan    3.172e+00  3.153e-01   1.22185    8.2343
## drugvalsartan      9.781e-01  1.022e+00   0.61121    1.5653
## dosage             1.000e+00  9.996e-01   1.00015    1.0007
##
## Concordance= 0.751  (se = 0.012 )
## Likelihood ratio test= 360.5  on 33 df,   p=<2e-16
## Wald test            = 304  on 33 df,   p=<2e-16
## Score (logrank) test = 373.9  on 33 df,   p=<2e-16
```

Coefficients of drugs: dapagliflozin, labetalol, and propranolol were infinite, so these drugs are removed in the second round of modeling.

```
# Cox regression model_2nd round
CKDdata_1 <- CKDdata %>% filter(drug != "dapagliflozin") %>%
  filter(drug != "labetalol") %>%
  filter(drug != "propranolol")

CKDdata_1$id <- NULL

split_data_1 <- sample(1:nrow(CKDdata_1), 0.8 * nrow(CKDdata_1), FALSE)
training_set_1 <- CKDdata_1[split_data_1,]
test_set_1 <- CKDdata_1[-split_data_1,]

cox2 <- coxph(Surv(time, stage) ~ ., data=training_set_1)
summary(cox2)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ ., data = training_set_1)
##
##   n= 1840, number of events= 693
##    (6500 observations deleted due to missingness)
##
```

```
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## creatinine        -0.1873018  0.8291935  0.1405656 -1.332 0.182700
## DBP                0.0046173  1.0046280  0.0043437  1.063 0.287784
## glucose            0.0631977  1.0652374  0.0262537  2.407 0.016076 *
## HGB               -0.1043295  0.9009284  0.0305725 -3.413 0.000644 ***
## ldl               -0.0077277  0.9923021  0.0015865 -4.871 1.11e-06 ***
## SBP               -0.0091333  0.9909083  0.0033929 -2.692 0.007105 **
## raceBlack         -0.6305567  0.5322954  0.2475837 -2.547 0.010870 *
## raceHispanic      -0.2142482  0.8071480  0.3882506 -0.552 0.581065
## raceUnknown        0.8690459  2.3846347  0.2257596  3.849 0.000118 ***
## raceWhite          0.1812216  1.1986808  0.1967775  0.921 0.357078
## genderMale         0.8431662  2.3237126  0.1055051  7.992 1.33e-15 ***
## age               -0.0209139  0.9793033  0.0044827 -4.665 3.08e-06 ***
## drugatorvastatin  -0.4743815  0.6222698  0.2017368 -2.351 0.018699 *
## drugbisoprolol    -0.3606787  0.6972030  0.6068910 -0.594 0.552308
## drugcanagliflozin -1.3791645  0.2517888  1.0179030 -1.355 0.175447
## drugcarvedilol    -0.2893375  0.7487595  0.2801755 -1.033 0.301744
## drugirbesartan     0.8541435  2.3493612  0.3838822  2.225 0.026080 *
## druglosartan      -0.4504142  0.6373641  0.2131929 -2.113 0.034626 *
## druglovastatin     0.0442138  1.0452058  0.3119029  0.142 0.887274
## drugmetformin     -0.6148793  0.5407062  0.2387602 -2.575 0.010015 *
## drugmetoprolol     0.2652512  1.3037584  0.1991676  1.332 0.182926
## drugnebivolol     -0.3594678  0.6980477  0.3393514 -1.059 0.289473
## drugolmesartan    -0.4472911  0.6393577  0.3415994 -1.309 0.190398
## drugpitavastatin   0.5801418  1.7862918  1.0204787  0.568 0.569696
## drugpravastatin    0.5650318  1.7595037  0.2273793  2.485 0.012956 *
## drugrosuvastatin  -0.9568025  0.3841191  0.2959231 -3.233 0.001224 **
## drugsimvastatin   -0.4773472  0.6204271  0.2114312 -2.258 0.023965 *
## drugtelmisartan    1.2593050  3.5229722  0.5360226  2.349 0.018806 *
## drugvalsartan     -0.0939141  0.9103610  0.2364209 -0.397 0.691196
## dosage             0.0003745  1.0003745  0.0001363  2.748 0.006002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## creatinine           0.8292     1.2060   0.62952    1.0922
## DBP                  1.0046     0.9954   0.99611    1.0132
## glucose              1.0652     0.9388   1.01181    1.1215
## HGB                  0.9009     1.1100   0.84853    0.9566
## ldl                  0.9923     1.0078   0.98922    0.9954
## SBP                  0.9909     1.0092   0.98434    0.9975
## raceBlack            0.5323     1.8787   0.32765    0.8648
## raceHispanic         0.8071     1.2389   0.37711    1.7276
## raceUnknown          2.3846     0.4194   1.53199    3.7118
## raceWhite            1.1987     0.8343   0.81509    1.7628
## genderMale           2.3237     0.4303   1.88963    2.8575
## age                  0.9793     1.0211   0.97074    0.9879
## drugatorvastatin     0.6223     1.6070   0.41904    0.9241
## drugbisoprolol       0.6972     1.4343   0.21221    2.2906
## drugcanagliflozin    0.2518     3.9716   0.03424    1.8513
## drugcarvedilol       0.7488     1.3355   0.43237    1.2967
## drugirbesartan       2.3494     0.4256   1.10710    4.9855
## druglosartan         0.6374     1.5690   0.41968    0.9680
## druglovastatin       1.0452     0.9567   0.56717    1.9262
```

```
## drugmetformin         0.5407     1.8494   0.33863    0.8634
## drugmetoprolol        1.3038     0.7670   0.88240    1.9263
## drugnebivolol         0.6980     1.4326   0.35895    1.3575
## drugolmesartan        0.6394     1.5641   0.32732    1.2489
## drugpitavastatin      1.7863     0.5598   0.24172   13.2004
## drugpravastatin       1.7595     0.5683   1.12679    2.7475
## drugrosuvastatin      0.3841     2.6034   0.21507    0.6861
## drugsimvastatin       0.6204     1.6118   0.40994    0.9390
## drugtelmisartan       3.5230     0.2839   1.23210   10.0733
## drugvalsartan         0.9104     1.0985   0.57276    1.4470
## dosage                1.0004     0.9996   1.00011    1.0006
##
## Concordance= 0.742  (se = 0.012 )
## Likelihood ratio test= 340.6  on 30 df,   p=<2e-16
## Wald test            = 338.1  on 30 df,   p=<2e-16
## Score (logrank) test = 365.4  on 30 df,   p=<2e-16
```

The coefficient of drug canagliflozin is infinite, so it is further removed in the third round of modeling.

```
# Cox regression model_3rd round
CKDdata_2 <- CKDdata_1 %>% filter(drug != "canagliflozin")

CKDdata_2$id <- NULL

split_data_2 <- sample(1:nrow(CKDdata_2), 0.8 * nrow(CKDdata_2), FALSE)
training_set_2 <- CKDdata_2[split_data_2,]
test_set_2 <- CKDdata_2[-split_data_2,]

cox3 <- coxph(Surv(time, stage) ~ ., data=training_set_2)
```

```
## Warning in fitter(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 23 ; coefficient may be infinite.
```

```
summary(cox3)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ ., data = training_set_2)
##
##   n= 1793, number of events= 674
##    (6515 observations deleted due to missingness)
##
##                      coef  exp(coef)   se(coef)      z Pr(>|z|)
## creatinine      -1.201e-02  9.881e-01  1.434e-01 -0.084 0.933256
## DBP              3.559e-03  1.004e+00  4.386e-03  0.811 0.417169
## glucose          3.645e-02  1.037e+00  2.766e-02  1.318 0.187594
## HGB             -8.767e-02  9.161e-01  3.125e-02 -2.805 0.005026 **
## ldl             -9.560e-03  9.905e-01  1.599e-03 -5.980 2.24e-09 ***
## SBP             -8.897e-03  9.911e-01  3.444e-03 -2.583 0.009786 **
## raceBlack       -7.059e-01  4.937e-01  2.499e-01 -2.825 0.004731 **
## raceHispanic    -4.349e-02  9.574e-01  3.762e-01 -0.116 0.907981
## raceUnknown      8.367e-01  2.309e+00  2.308e-01  3.624 0.000290 ***
```

```
## raceWhite          1.860e-01  1.204e+00  1.998e-01  0.931 0.351772
## genderMale         9.572e-01  2.604e+00  1.091e-01  8.774  < 2e-16 ***
## age               -2.604e-02  9.743e-01  4.482e-03 -5.810 6.23e-09 ***
## drugatorvastatin -3.628e-01  6.957e-01  2.092e-01 -1.734 0.082930 .
## drugbisoprolol   -4.506e-01  6.372e-01  7.329e-01 -0.615 0.538660
## drugcarvedilol   -7.113e-02  9.313e-01  2.893e-01 -0.246 0.805796
## drugirbesartan    9.555e-01  2.600e+00  4.276e-01  2.234 0.025455 *
## druglosartan     -4.161e-01  6.596e-01  2.214e-01 -1.879 0.060191 .
## druglovastatin    2.383e-01  1.269e+00  3.335e-01  0.715 0.474911
## drugmetformin    -5.724e-01  5.642e-01  2.428e-01 -2.357 0.018400 *
## drugmetoprolol    3.958e-01  1.486e+00  2.074e-01  1.908 0.056367 .
## drugnebivolol     5.686e-02  1.059e+00  3.675e-01  0.155 0.877032
## drugolmesartan   -3.221e-01  7.246e-01  3.297e-01 -0.977 0.328583
## drugpitavastatin -1.264e+01  3.239e-06  7.315e+02 -0.017 0.986212
## drugpravastatin   7.518e-01  2.121e+00  2.355e-01  3.193 0.001407 **
## drugrosuvastatin -9.109e-01  4.021e-01  3.004e-01 -3.033 0.002424 **
## drugsimvastatin  -3.573e-01  6.995e-01  2.149e-01 -1.663 0.096302 .
## drugtelmisartan   1.459e+00  4.302e+00  4.520e-01  3.228 0.001246 **
## drugvalsartan    -7.203e-02  9.305e-01  2.440e-01 -0.295 0.767820
## dosage            4.617e-04  1.000e+00  1.376e-04  3.356 0.000792 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## creatinine       9.881e-01  1.012e+00    0.7460    1.3086
## DBP              1.004e+00  9.964e-01    0.9950    1.0122
## glucose          1.037e+00  9.642e-01    0.9824    1.0949
## HGB              9.161e-01  1.092e+00    0.8616    0.9739
## ldl              9.905e-01  1.010e+00    0.9874    0.9936
## SBP              9.911e-01  1.009e+00    0.9845    0.9979
## raceBlack        4.937e-01  2.026e+00    0.3025    0.8056
## raceHispanic     9.574e-01  1.044e+00    0.4580    2.0016
## raceUnknown      2.309e+00  4.332e-01    1.4685    3.6295
## raceWhite        1.204e+00  8.303e-01    0.8142    1.7816
## genderMale       2.604e+00  3.840e-01    2.1030    3.2252
## age              9.743e-01  1.026e+00    0.9658    0.9829
## drugatorvastatin 6.957e-01  1.437e+00    0.4617    1.0484
## drugbisoprolol   6.372e-01  1.569e+00    0.1515    2.6800
## drugcarvedilol   9.313e-01  1.074e+00    0.5282    1.6420
## drugirbesartan   2.600e+00  3.846e-01    1.1245    6.0114
## druglosartan     6.596e-01  1.516e+00    0.4274    1.0180
## druglovastatin   1.269e+00  7.880e-01    0.6601    2.4399
## drugmetformin    5.642e-01  1.773e+00    0.3505    0.9080
## drugmetoprolol   1.486e+00  6.731e-01    0.9893    2.2308
## drugnebivolol    1.059e+00  9.447e-01    0.5151    2.1751
## drugolmesartan   7.246e-01  1.380e+00    0.3797    1.3828
## drugpitavastatin 3.239e-06  3.087e+05    0.0000       Inf
## drugpravastatin  2.121e+00  4.715e-01    1.3369    3.3646
## drugrosuvastatin 4.021e-01  2.487e+00    0.2232    0.7245
## drugsimvastatin  6.995e-01  1.430e+00    0.4591    1.0659
## drugtelmisartan  4.302e+00  2.325e-01    1.7739   10.4314
## drugvalsartan    9.305e-01  1.075e+00    0.5768    1.5010
## dosage           1.000e+00  9.995e-01    1.0002    1.0007
##
```

```
## Concordance= 0.759  (se = 0.012 )
## Likelihood ratio test= 368.9  on 29 df,   p=<2e-16
## Wald test            = 362.4  on 29 df,   p=<2e-16
## Score (logrank) test = 394.7  on 29 df,   p=<2e-16
```

The coefficient of drug pitavastatin is infinite, so it is further removed in the fourth round of modeling.

```
# Cox regression model_4th round
CKDdata_3 <- CKDdata_2 %>% filter(drug != "pitavastatin")

CKDdata_3$id <- NULL

split_data_3 <- sample(1:nrow(CKDdata_3), 0.8 * nrow(CKDdata_3), FALSE)
training_set_3 <- CKDdata_3[split_data_3,]
test_set_3 <- CKDdata_3[-split_data_3,]

cox4 <- coxph(Surv(time, stage) ~ ., data=training_set_3)
summary(cox4)
```

```
## Call:
## coxph(formula = Surv(time, stage) ~ ., data = training_set_3)
##
##   n= 1816, number of events= 691
##    (6475 observations deleted due to missingness)
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## creatinine         0.0852566  1.0889965  0.1381620  0.617 0.537184
## DBP                0.0052815  1.0052955  0.0043563  1.212 0.225364
## glucose            0.0484911  1.0496861  0.0279198  1.737 0.082423 .
## HGB               -0.0638281  0.9381663  0.0309825 -2.060 0.039386 *
## ldl               -0.0106863  0.9893706  0.0016130 -6.625 3.47e-11 ***
## SBP               -0.0110804  0.9889808  0.0033912 -3.267 0.001086 **
## raceBlack         -0.7970705  0.4506472  0.2434908 -3.274 0.001062 **
## raceHispanic      -0.5643523  0.5687284  0.4260968 -1.324 0.185347
## raceUnknown        0.7652322  2.1494934  0.2235508  3.423 0.000619 ***
## raceWhite          0.0949779  1.0996346  0.1935699  0.491 0.623664
## genderMale         1.0433515  2.8387150  0.1069388  9.757  < 2e-16 ***
## age               -0.0225666  0.9776861  0.0044037 -5.124 2.98e-07 ***
## drugatorvastatin  -0.5453730  0.5796256  0.1957016 -2.787 0.005324 **
## drugbisoprolol     0.0425695  1.0434885  0.4801761  0.089 0.929357
## drugcarvedilol    -0.1895310  0.8273471  0.2768631 -0.685 0.493618
## drugirbesartan     0.9029015  2.4667500  0.4239600  2.130 0.033198 *
## druglosartan      -0.4824727  0.6172552  0.2078016 -2.322 0.020244 *
## druglovastatin     0.1313415  1.1403572  0.3153147  0.417 0.677014
## drugmetformin     -0.5210906  0.5938725  0.2310645 -2.255 0.024122 *
## drugmetoprolol     0.1674465  1.1822820  0.1942005  0.862 0.388558
## drugnebivolol     -0.2695511  0.7637223  0.3448429 -0.782 0.434412
## drugolmesartan    -0.5590404  0.5717575  0.3368348 -1.660 0.096977 .
## drugpravastatin    0.6762680  1.9665250  0.2272476  2.976 0.002921 **
## drugrosuvastatin  -1.1322407  0.3223103  0.2982102 -3.797 0.000147 ***
## drugsimvastatin   -0.4696975  0.6251913  0.2025900 -2.318 0.020424 *
## drugtelmisartan    1.0655842  2.9025342  0.4458482  2.390 0.016848 *
## drugvalsartan     -0.0879622  0.9157955  0.2264612 -0.388 0.697705
```

```
## dosage            0.0003100  1.0003101  0.0001358  2.284 0.022390 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## creatinine         1.0890     0.9183    0.8307    1.4277
## DBP                1.0053     0.9947    0.9967    1.0139
## glucose            1.0497     0.9527    0.9938    1.1087
## HGB                0.9382     1.0659    0.8829    0.9969
## ldl                0.9894     1.0107    0.9862    0.9925
## SBP                0.9890     1.0111    0.9824    0.9956
## raceBlack          0.4506     2.2190    0.2796    0.7263
## raceHispanic       0.5687     1.7583    0.2467    1.3110
## raceUnknown        2.1495     0.4652    1.3869    3.3314
## raceWhite          1.0996     0.9094    0.7525    1.6070
## genderMale         2.8387     0.3523    2.3019    3.5006
## age                0.9777     1.0228    0.9693    0.9862
## drugatorvastatin   0.5796     1.7253    0.3950    0.8506
## drugbisoprolol     1.0435     0.9583    0.4072    2.6743
## drugcarvedilol     0.8273     1.2087    0.4809    1.4235
## drugirbesartan     2.4667     0.4054    1.0746    5.6624
## druglosartan       0.6173     1.6201    0.4108    0.9276
## druglovastatin     1.1404     0.8769    0.6147    2.1156
## drugmetformin      0.5939     1.6839    0.3776    0.9341
## drugmetoprolol     1.1823     0.8458    0.8080    1.7299
## drugnebivolol      0.7637     1.3094    0.3885    1.5013
## drugolmesartan     0.5718     1.7490    0.2955    1.1064
## drugpravastatin    1.9665     0.5085    1.2597    3.0700
## drugrosuvastatin   0.3223     3.1026    0.1797    0.5782
## drugsimvastatin    0.6252     1.5995    0.4203    0.9299
## drugtelmisartan    2.9025     0.3445    1.2114    6.9548
## drugvalsartan      0.9158     1.0919    0.5875    1.4275
## dosage             1.0003     0.9997    1.0000    1.0006
##
## Concordance= 0.759  (se = 0.012 )
## Likelihood ratio test= 382.5  on 28 df,   p=<2e-16
## Wald test            = 376  on 28 df,   p=<2e-16
## Score (logrank) test = 402.5  on 28 df,   p=<2e-16
```

Prediction

```
# Remove NA
test_set_3_n <- na.omit(test_set_3)

# Make prediction
pred <- predict(cox4, newdata=test_set_3_n, type="survival")
observed_result <- test_set_3_n$stage
predict_result <- ifelse(pred>0.5,0,1)
table(predict_result, observed_result)
```

```
##                observed_result
## predict_result   0   1
##              0 240  94
##              1  49  76
```

Confusion Matrix and ROC

```
# Confusion Matrix
MLmetrics::Accuracy(predict_result, test_set_3_n$stage)
```

```
## [1] 0.6884532
```

```
MLmetrics::Specificity(predict_result, test_set_3_n$stage)
```

```
## [1] 0.608
```

```
MLmetrics::Sensitivity(predict_result, test_set_3_n$stage)
```

```
## [1] 0.7185629
```

```
# ROC
require(ROCR)
```

```
## Loading required package: ROCR
```

```
require(pROC)
```

```
## Loading required package: pROC
```

```
## Type 'citation("pROC")' for a citation.
```

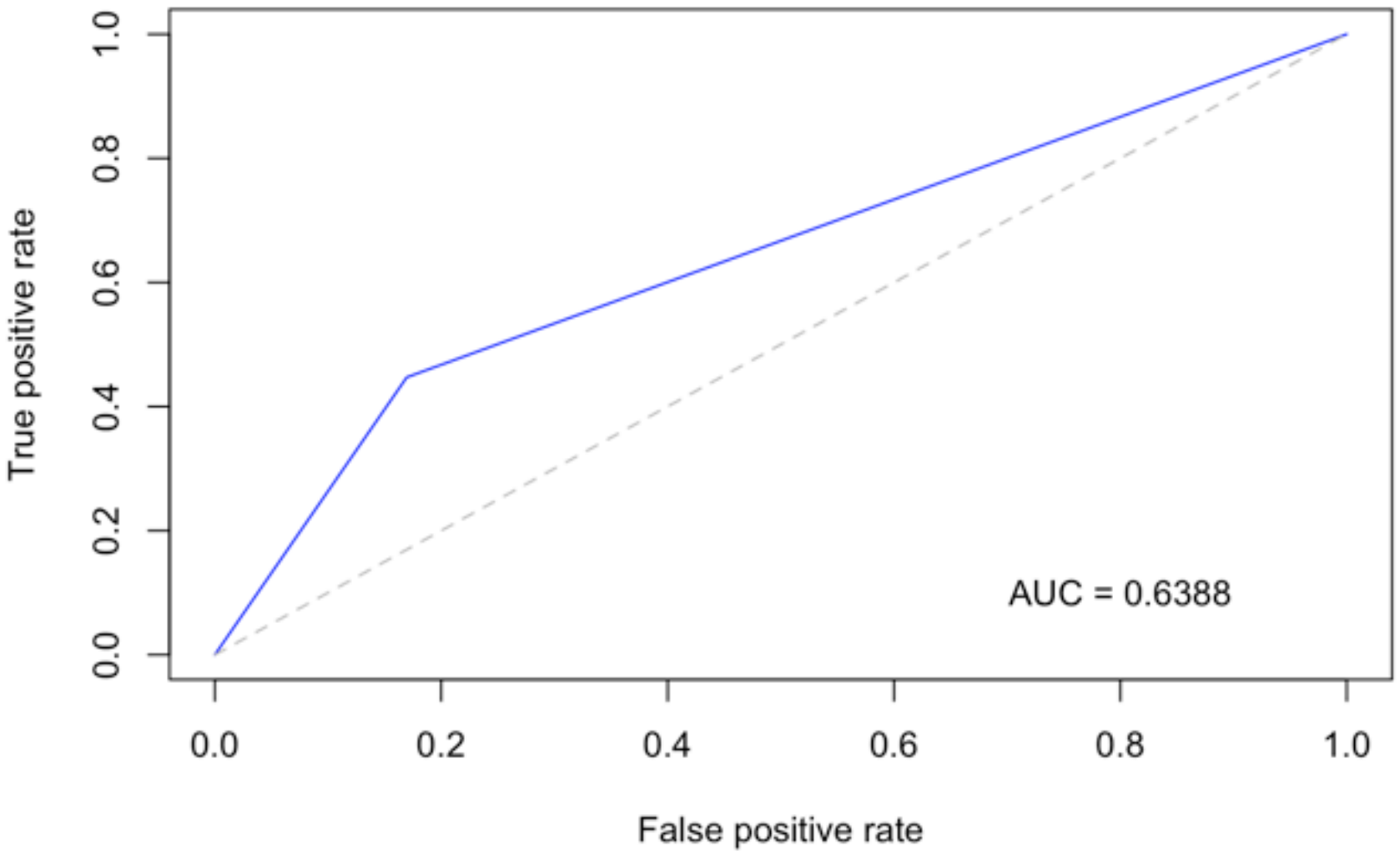
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
rocplot <- function(pred, truth, ...) {
  predob = prediction(pred, truth)
  perf = performance(predob, "tpr", "fpr")
  plot(perf, ...)
  area <- auc(truth, pred)
  area <- format(round(area, 4), nsmall = 4)
  text(x=0.8, y=0.1, labels = paste("AUC =", area))

  # the reference x=y line
  segments(x0=0, y0=0, x1=1, y1=1, col="gray", lty=2)
}

rocplot(predict_result, observed_result, col="blue")
```

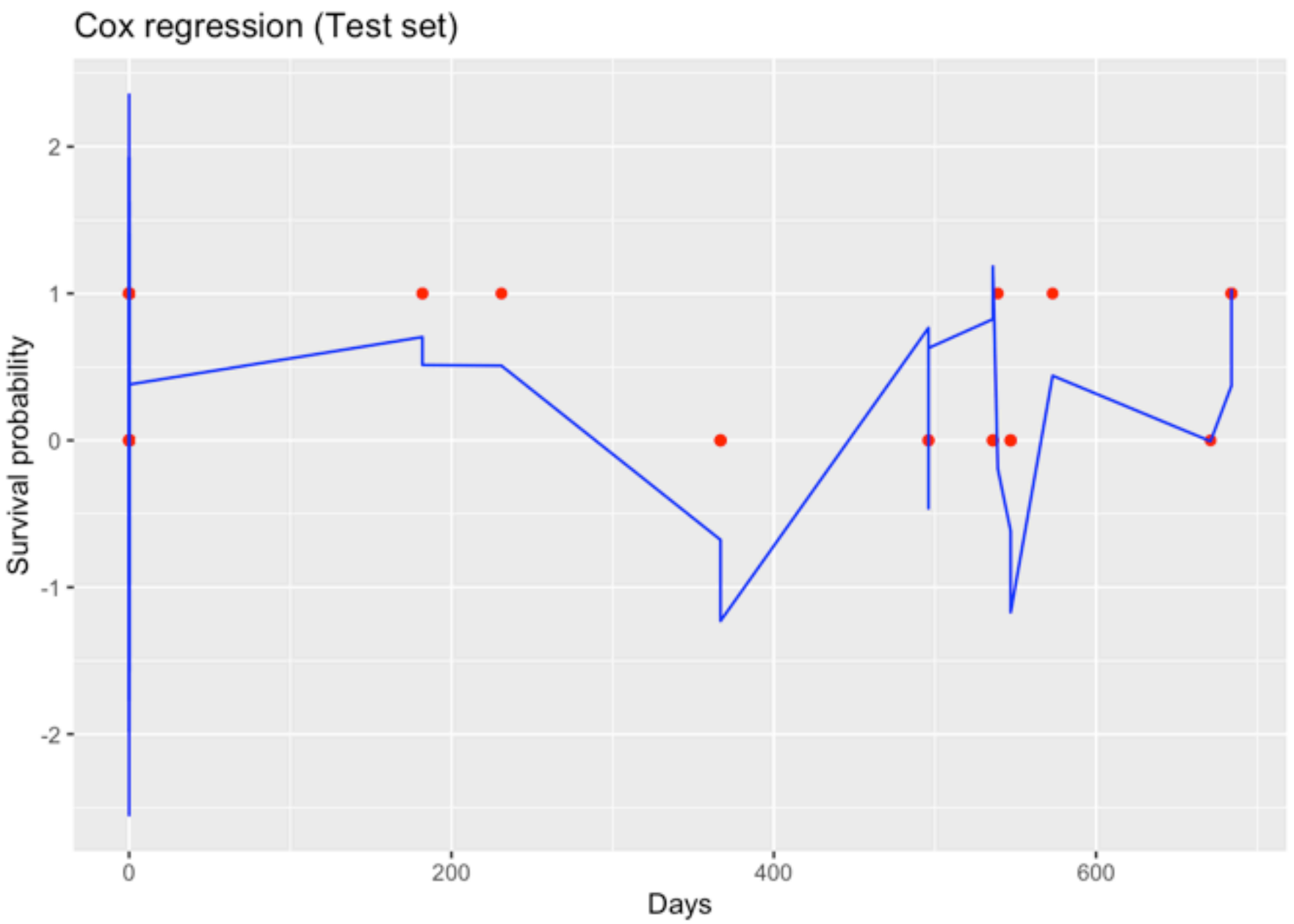
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Plot test set results

```
ggplot() +
  geom_point(aes(x = test_set_3_n$time, y = test_set_3_n$stage),
             colour = 'red') +
  geom_line(aes(x = test_set_3_n$time, y = predict(cox4, newdata = test_set_3_n)),
            colour = 'blue') +
  ggtitle('Cox regression (Test set)') +
  xlab('Days') +
  ylab('Survival probability')
```



The Cox regression results can be interpreted as follows:

1. Statistical significance: the column marked “z” gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef / se(coef)). From the output above, I can conclude that the variables glucose, HGB, LDL, SBP, race_Black, race_Unknown, gender_Male, age, and drug (atorvastatin, irbesartan, losartan, pravastatin, rosuvastatin, simvastatin, telmisartan) have highly statistically significant coefficients.
2. Regression coefficients: the second important feature in the Cox model results is the the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable.
3. Hazard ratios: the exponentiated coefficients, also known as hazard ratios, give the effect size of covariates.
4. Confidence intervals of the hazard ratios: the summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)).

5. Global statistical significance of the model: finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough sample size, they will give similar results. For small sample size, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

Based on the results, glucose, HGB, LDL, SBP, race, gender, age, some drugs, and drug doseage are significant risk factors ($p < 0.05$) contributing to the progression of CKD. The Cox regression model is a good predictor in this CKD data set. A significant advantage of this Cox model is its ease of use. The algorithm automatically calculates simultaneously the effect of several risk factors on survival time.

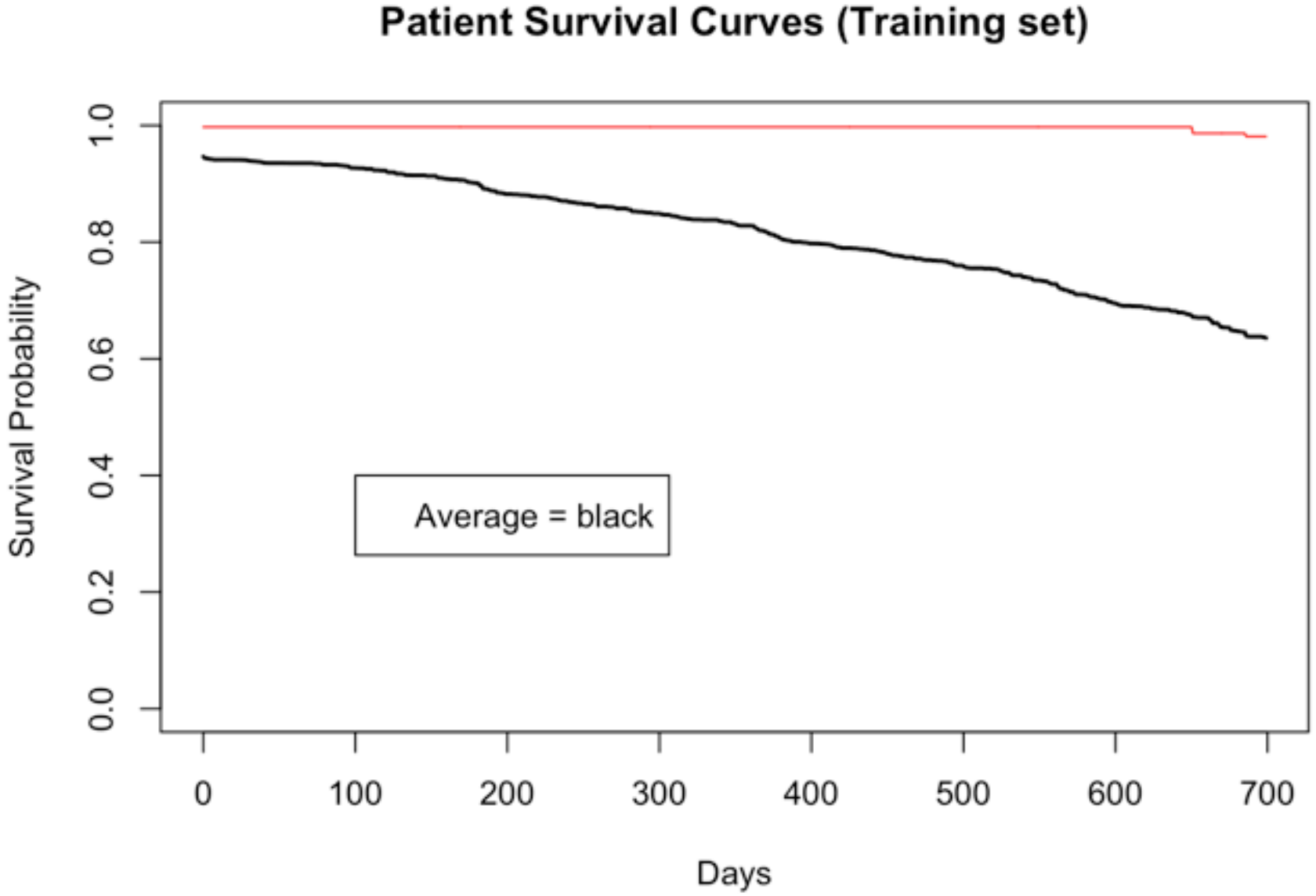
Random Forest Model

Finally, I use the ranger ( ) function to fit a Random Forest Ensemble model to the CKD dataset. ranger ( ) builds a model for each observation in the dataset. The next block of code plots random curves in the training and test sets, along with a curve that represents the global average for all of the patients.

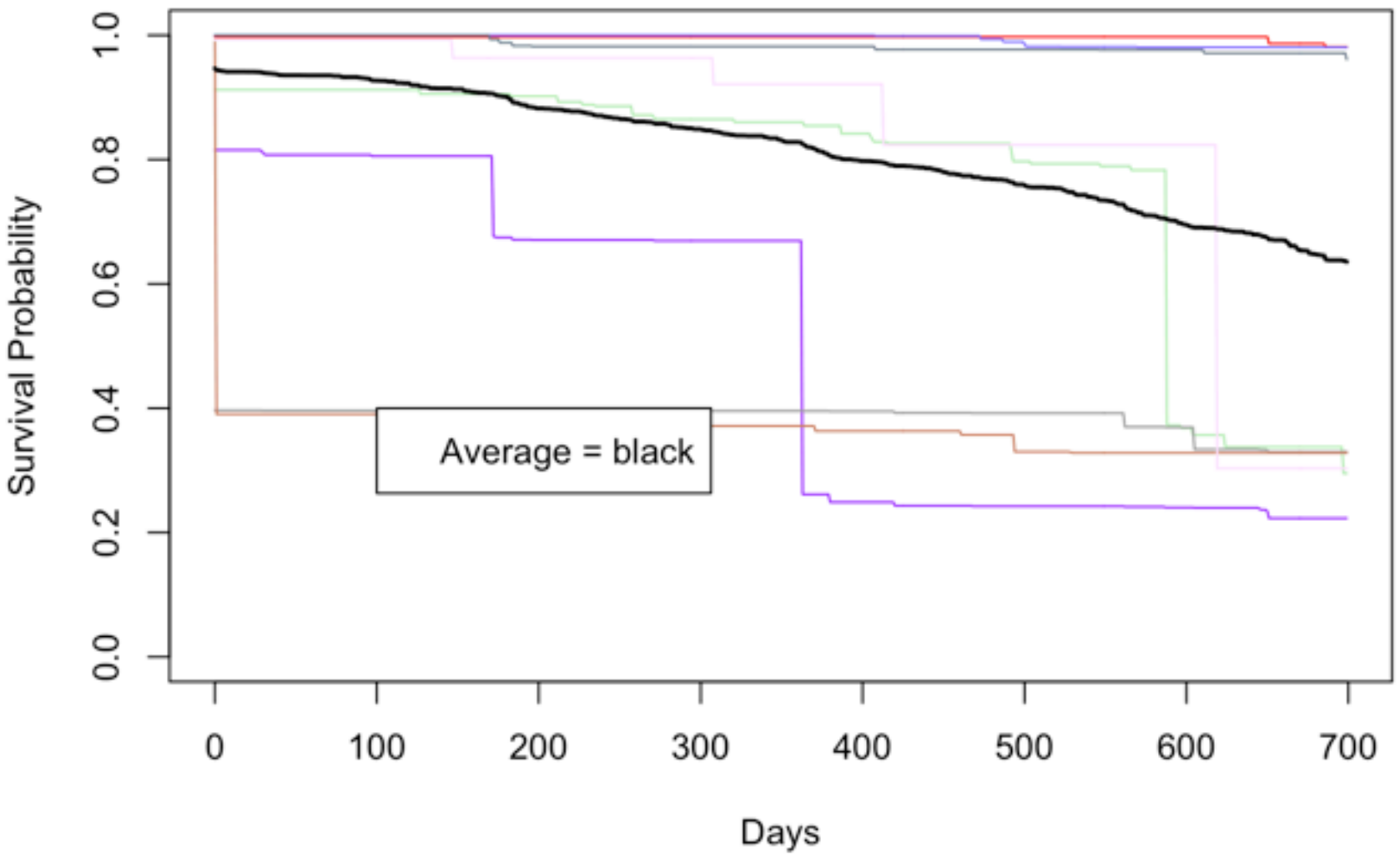
```
# Split dataset into training and test sets
split_data <- sample(1:nrow(CKDdata), 0.8 * nrow(CKDdata), FALSE)
training_set <- CKDdata[split_data,]
test_set <- CKDdata[-split_data,]

# Build random forest model
r_fit <- ranger(Surv(time, stage) ~ creatinine+gender+race+age, data=training_set,
mtry = 4, importance="permutation", splitrule="extratrees",verbose=TRUE)
```

```
## Computing permutation importance.. Progress: 35%. Estimated remaining time: 58
seconds.
## Computing permutation importance.. Progress: 69%. Estimated remaining time: 28
seconds.
```

```
prob_pred <- predict(r_fit, data=test_set, mtry = 4, importance="permutation", spl
itrule="extratrees",verbose=TRUE)

# Average the survival models
CKDprogress_time <- r_fit$unique.death.times
surv_probability <- data.frame(r_fit$survival)
avg_probability <- sapply(surv_probability, mean)

# Plot the survival models
# Training set
plot(r_fit$unique.death.times,r_fit$survival[1,],
     type = "l",
     ylim = c(0,1),
     col = "red",
     xlab = "Days",
     ylab = "Survival Probability",
     main = "Patient Survival Curves (Training set)")

cols <- colors()
for (n in sample(c(2:dim(training_set)[1]), 20)){
  lines(r_fit$unique.death.times, r_fit$survival[n,], type = "l", col = cols[n])
}
lines(CKDprogress_time, avg_probability, lwd = 2)
legend(100, 0.4, legend = c('Average = black'))
```




```
# Test set
plot(r_fit$unique.death.times,r_fit$survival[1,],
     type = "l",
     ylim = c(0,1),
     col = "red",
     xlab = "Days",
     ylab = "Survival Probability",
     main = "Patient Survival Curves (Test set)")

cols <- colors()
for (n in sample(c(2:dim(test_set)[1]), 20)){
  lines(r_fit$unique.death.times, r_fit$survival[n,], type = "l", col = cols[n])
}
lines(CKDprogress_time, avg_probability, lwd = 2)
legend(100, 0.4, legend = c('Average = black'))
```



The next block of illustrates how ranger ( ) ranks variable importance.

```
# Variable importance
vi <- data.frame(sort(round(r_fit$variable.importance, 4), decreasing = TRUE))
names(vi) <- "importance"
head(vi)
```

```
##            importance
## age            0.3348
## creatinine     0.2762
## gender         0.2242
## race           0.1649
```

```
cat("Prediction Error = 1 - Harrell's c-index = ", r_fit$prediction.error)
```

```
## Prediction Error = 1 - Harrell's c-index =  0.1076468
```

I notice that ranger ( ) flags age, creatinine, gender, and race are the most important factors of CKD progression. Age, gender, and race are the same variables with the p-values less than 0.05 in the Cox regression model. Furthermore, ranger( ) also computes Harrell's c-index. This is a generalization of the ROC curve, which reduces to the Wilcoxon-Mann-Whiteney statistics for binary variables, which in turn, is

equivalent to computing the area under the ROC curve. Here, the prediction error is 0.11 and the ROC value of 0.89 would normally be pretty good for a first try. But ranger ( ) doesn't do anything to address the time varying coefficients, which is apparently a challenge.

In summary, for this CKD data set, I would choose a carefully constructed Cox Regression model that takes into account all varying coefficients. Tree-based models for survival analysis will be useful in dealing with very large data sets.