Regression Analysis Midterm - R Code

Sanne Glastra

2024-10-20

Data Preparation

setwd

```
setwd("/Users/sanneglastra/Documents/school/columbia/fall 2025/regression analysis II/midterm")
```

read libraries

```
library(tableone)
library(officer)
library(flextable)
library(ggsurvfit)
```

```
## Loading required package: ggplot2
```

```
library(survival)
library(tibble)
library(dplyr)
```

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

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

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

read csv

```
data <- read.csv("midtermdata.csv")
```

recode variables

```
# raceeth variable -- combine variables
data$raceth <- ifelse(data$raceth %in% c(3, 4, 5, 6), 3, data$raceth)

# ivdrug variable -- combine variables
data$ivdrug <- ifelse(data$ivdrug %in% c(2, 3), 2, data$ivdrug)

# karnof variable -- combine variables
data$karnof <- ifelse(data$karnof %in% c(80, 70), 80, data$karnof)
```

define outcome variables

```
# define pfs variables
data$time_pfs <- data$time
data$censor_pfs <- data$censor

# define os variables
data$time_os <- data$time_d
data$censor_os <- data$censor_d
```

Create Table 1: Baseline Characteristics

```
# specify which variables want to use in table 1
allVars = c("sex", "raceth", "age", "karnof","strat2", "cd4", "ivdrug", "hemophil","priorzdv")

# specify which variables are categorical variables
catVars = c("sex", "raceth", "ivdrug", "hemophil", "strat2", "karnof")

# Define labels for categorical variables
sex_levels <- c("1" = "Male", "2" = "Female")
raceth_levels <- c(
  "1" = "White Non-Hispanic",
  "2" = "Black Non-Hispanic",
  "3" = "Other/unknown")
ivdrug_levels <- c("1" = "Never", "2" = "Currently/Previously")
hemophil_levels <- c("1" = "Yes", "0" = "No")
strat2_levels <- c("0" = "CD4   50", "1" = "CD4 > 50")
karnof_levels <- c(
  "100" = "No evidence of disease",
  "90" = "Minor signs/symptoms of disease",
  "80" = "Some signs/symptoms of disease or cares for self")
tx_levels <- c("1" = "Three-drug regimen including IDV", "0" = "Two-drug regimen without IDV")

# Convert categorical variables to factors
data1 <- data
data1$sex <- factor(data$sex, levels = names(sex_levels), labels = sex_levels)
data1$raceth <- factor(data$raceth, levels = names(raceth_levels), labels = raceth_levels)
```

```r
data1$ivdrug <- factor(data$ivdrug, levels = names(ivdrug_levels), labels = ivdrug_levels)
data1$hemophil <- factor(data$hemophil, levels = names(hemophil_levels), labels = hemophil_levels)
data1$strat2 <- factor(data$strat2, levels = names(strat2_levels), labels = strat2_levels)
data1$karnof <- factor(data$karnof, levels = names(karnof_levels), labels = karnof_levels)
data1$tx <- factor(data$tx, levels = names(tx_levels), labels = tx_levels)


# create baseline table
descriptives_tableone <- tableone::CreateTableOne(data = data1,
                                                  vars = allVars,
                                                  factorVars = catVars,
                                                  strata = "tx",
                                                  test = T,
                                                  addOverall = TRUE)

# print baseline table (showing all levels and using median(iqr))
descriptives_tableone <- print(descriptives_tableone,
                               nonnormal = allVars, # specify which variables to use median(iqr)
                               showAllLevels = TRUE)
```

```
##                          Stratified by tx
##                           level
##   n
##   sex (%)                 Male
##                           Female
##   raceth (%)              White Non-Hispanic
##                           Black Non-Hispanic
##                           Other/unknown
##   age (median [IQR])
##   karnof (%)              No evidence of disease
##                           Minor signs/symptoms of disease
##                           Some signs/symptoms of disease or cares for self
##   strat2 (%)              CD4   50
##                           CD4 > 50
##   cd4 (median [IQR])
##   ivdrug (%)              Never
##                           Currently/Previously
##   hemophil (%)            Yes
##                           No
##   priorzdv (median [IQR])
##                          Stratified by tx
##                           Overall
##   n                        1151
##   sex (%)                   951 (82.6)
##                             200 (17.4)
##   raceth (%)                596 (51.8)
##                             327 (28.4)
##                             228 (19.8)
##   age (median [IQR])       38.00 [33.00, 44.00]
##   karnof (%)                396 (34.4)
##                             541 (47.0)
##                             214 (18.6)
##   strat2 (%)                439 (38.1)
```

```
##                              712 (61.9)
##   cd4 (median [IQR])      74.50 [23.00, 136.50]
##   ivdrug (%)                968 (84.1)
##                             183 (15.9)
##   hemophil (%)               35 ( 3.0)
##                            1116 (97.0)
##   priorzdv (median [IQR]) 21.00 [10.00, 42.00]
##                           Stratified by tx
##                            Three-drug regimen including IDV
##   n                          574
##   sex (%)                    468 (81.5)
##                              106 (18.5)
##   raceth (%)                 302 (52.6)
##                              162 (28.2)
##                              110 (19.2)
##   age (median [IQR])      38.00 [33.00, 44.00]
##   karnof (%)                 194 (33.8)
##                              274 (47.7)
##                              106 (18.5)
##   strat2 (%)                 219 (38.2)
##                              355 (61.8)
##   cd4 (median [IQR])      79.50 [23.62, 138.75]
##   ivdrug (%)                 484 (84.3)
##                               90 (15.7)
##   hemophil (%)               14 ( 2.4)
##                              560 (97.6)
##   priorzdv (median [IQR]) 22.00 [11.00, 42.00]
##                           Stratified by tx
##                            Two-drug regimen without IDV p      test
##   n                          577
##   sex (%)                    483 (83.7)                  0.370
##                               94 (16.3)
##   raceth (%)                 294 (51.0)                  0.816
##                              165 (28.6)
##                              118 (20.5)
##   age (median [IQR])      38.00 [33.00, 44.00]          0.915 nonnorm
##   karnof (%)                 202 (35.0)                  0.877
##                              267 (46.3)
##                              108 (18.7)
##   strat2 (%)                 220 (38.1)                  1.000
##                              357 (61.9)
##   cd4 (median [IQR])      69.50 [22.50, 134.50]         0.257 nonnorm
##   ivdrug (%)                 484 (83.9)                  0.902
##                               93 (16.1)
##   hemophil (%)               21 ( 3.6)                  0.310
##                              556 (96.4)
##   priorzdv (median [IQR]) 19.00 [10.00, 42.00]          0.246 nonnorm
```

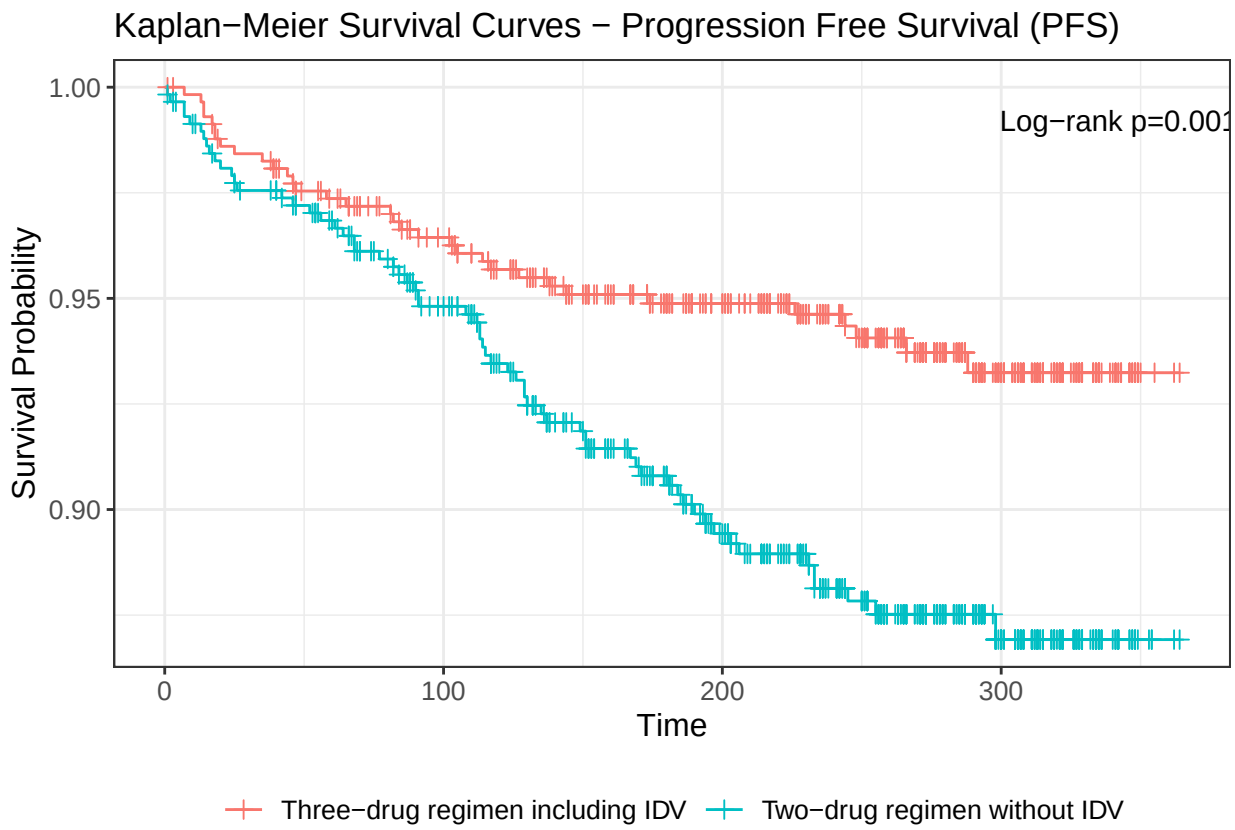
```
# print table to word
descriptives_tableone %>%
  as.data.frame() %>%
  rownames_to_column("Characteristic") %>%
  flextable() %>%
  # Bold headers
```

```
  bold(part = "header") %>%
  # Set font to Times New Roman, size to 12
  font(fontname = "Times New Roman", part = "all") %>%
  fontsize(size = 11, part = "all") %>%
  # Add borders only below the header and in specific locations
  hline_top(border = fp_border(color = "black", width = 1.5), part = "header") %>%
  hline_bottom(border = fp_border(color = "black", width = 1.5), part = "body") %>% # Bottom border
  # Add padding for readability
  padding(padding = 5, part = "all") %>%
  flextable::save_as_docx(path = "descriptives_tableone.docx")
```

Survival Curve Analysis and Log-Rank

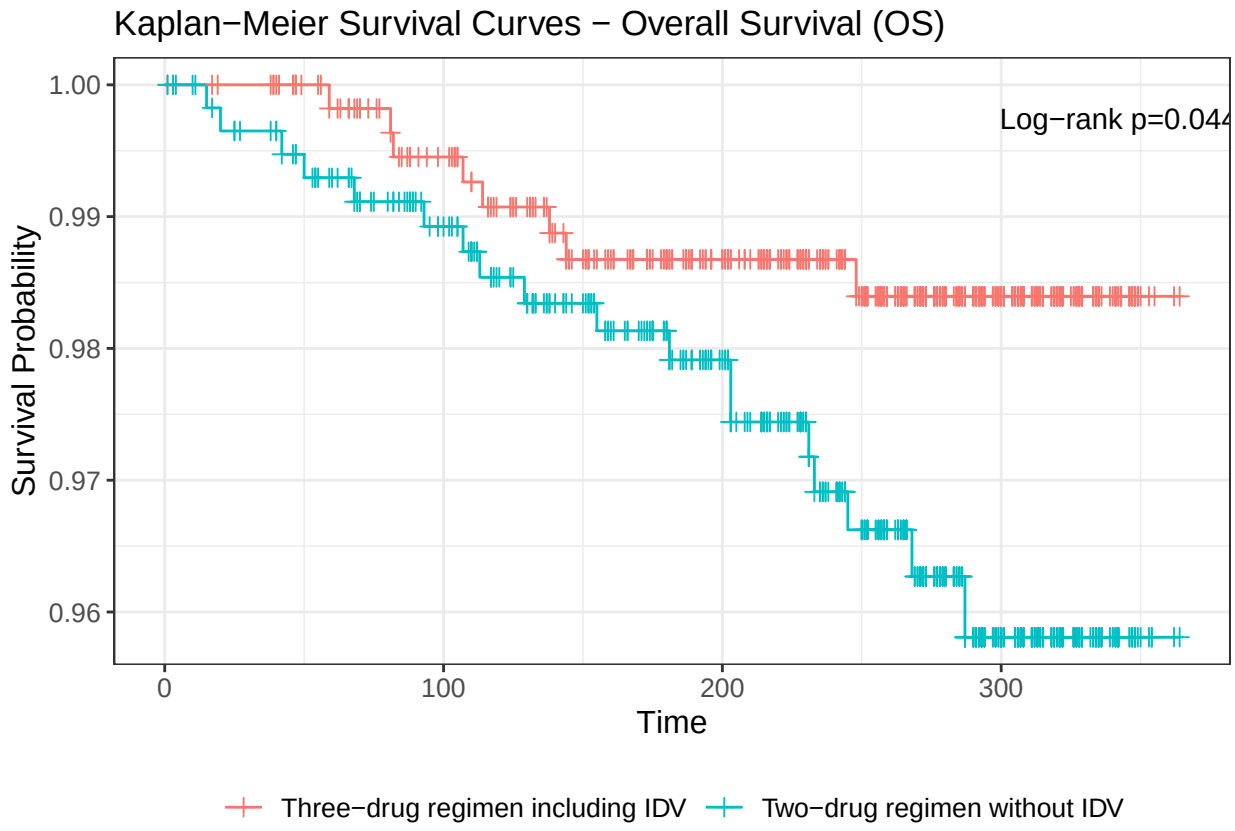
PFS treatment comparison

```
survfit2(Surv(time_pfs, censor_pfs)~tx, data=data1) %>%
  ggsurvfit() +
  add_censor_mark() +
  add_pvalue(location="annotation",
             caption="Log-rank {p.value}") +
  labs(title = "Kaplan-Meier Survival Curves - Progression Free Survival (PFS)")
```



OS treatment comparison

```
survfit2(Surv(time_os, censor_os)~tx, data=data1) %>%
  ggsurvfit() +
  add_censor_mark() +
  add_pvalue(location="annotation",
             caption="Log-rank {p.value}") +
  labs(title = "Kaplan-Meier Survival Curves - Overall Survival (OS)")
```



Univariate Analysis - PFS

process categorical data so it can be read correctly by cox model

```
# make all categorical data into factors
data = data %>%
  mutate(sex = factor(sex,levels=c("1","2")),
         tx = factor(tx,levels=c("0","1")),
         raceth = factor(raceth,levels=c("3","1","2")),
         karnof = factor(karnof,levels=c("80","90", "100")),
         strat2 = factor(strat2,levels=c("0","1")),
         ivdrug = factor(ivdrug,levels=c("1","2")),
         hemophil = factor(hemophil,levels=c("0","1")))
```

fit cox model for each covariate individually

```
# treatment
fit_tx_pfs = coxph(Surv(time_pfs, censor_pfs)~tx,data=data,
            ties="efron")

# age
fit_age_pfs = coxph(Surv(time_pfs, censor_pfs)~age,data=data,
            ties="efron")
# sex
fit_sex_pfs = coxph(Surv(time_pfs, censor_pfs)~sex,data=data,
            ties="efron")

# race / ethnicity
fit_raceth_pfs = coxph(Surv(time_pfs, censor_pfs)~raceth,data=data,
            ties="efron")

# karnofsky performance scale
fit_karnof_pfs = coxph(Surv(time_pfs, censor_pfs)~karnof,data=data,
            ties="efron")

# cd4 stratum at screening
fit_strat2_pfs = coxph(Surv(time_pfs, censor_pfs)~strat2,data=data,
            ties="efron")

# baseline cd4 count
fit_cd4_pfs = coxph(Surv(time_pfs, censor_pfs)~cd4,data=data,
            ties="efron")

# iv drug use history
fit_ivdrug_pfs = coxph(Surv(time_pfs, censor_pfs)~ivdrug,data=data,
            ties="efron")

# hemophiliac
fit_hemophil_pfs = coxph(Surv(time_pfs, censor_pfs)~hemophil,data=data,
            ties="efron")

# months of prior zdv use
fit_priorzdv_pfs = coxph(Surv(time_pfs, censor_pfs)~priorzdv,data=data,
            ties="efron")
```

place all HR / CI results into table and export to word

```
# make a list of all models
models <- list(
  tx = fit_tx_pfs,
  age = fit_age_pfs,
  sex = fit_sex_pfs,
  raceth = fit_raceth_pfs,
  karnof = fit_karnof_pfs,
  strat2 = fit_strat2_pfs,
```

```
  cd4 = fit_cd4_pfs,
  ivdrug = fit_ivdrug_pfs,
  hemophil = fit_hemophil_pfs,
  priorzdv = fit_priorzdv_pfs
)

# function to extract exp(coef) and 95% CI
extract_coefs <- function(fit) {
  summary_fit <- summary(fit)
  exp_coef <- summary_fit$coefficients[, "exp(coef)"]
  lower_95 <- summary_fit$conf.int[, "lower .95"]
  upper_95 <- summary_fit$conf.int[, "upper .95"]
  p_value <- summary(fit)$coefficients[, "Pr(>|z|)"]

  return(data.frame(HR = round(exp_coef,2), Lower_CI = round(lower_95,2), Upper_CI = round(upper_95, 2)
                    P_Value = round(p_value,3)))
}

# apply the extraction function and combine into single dataframe
results_list <- lapply(models, extract_coefs)
results_df <- do.call(rbind, results_list)
results_df$CI <- paste0("[", results_df$Lower_CI, ", ", results_df$Upper_CI, "]")

# print table
univariate_table_pfs <- print(results_df)
```

```
##                    HR Lower_CI Upper_CI P_Value           CI
## tx               0.50     0.33     0.77   0.001 [0.33, 0.77]
## age              1.02     1.00     1.04   0.061    [1, 1.04]
## sex              0.92     0.53     1.60   0.778  [0.53, 1.6]
## raceth.raceth1   0.77     0.47     1.25   0.292 [0.47, 1.25]
## raceth.raceth2   0.61     0.34     1.10   0.104  [0.34, 1.1]
## karnof.karnof90  0.34     0.22     0.53   0.000 [0.22, 0.53]
## karnof.karnof100 0.21     0.12     0.37   0.000 [0.12, 0.37]
## strat2           0.26     0.17     0.40   0.000  [0.17, 0.4]
## cd4              0.98     0.98     0.99   0.000 [0.98, 0.99]
## ivdrug           0.67     0.36     1.25   0.209 [0.36, 1.25]
## hemophil         1.02     0.32     3.22   0.972 [0.32, 3.22]
## priorzdv         1.00     0.99     1.01   0.511 [0.99, 1.01]
```

```
# print table to word
univariate_table_pfs %>%
  select(HR, CI, P_Value) %>%
  as.data.frame() %>%
  rownames_to_column("Characteristic") %>%
  flextable() %>%
  # Bold headers
  bold(part = "header") %>%
  # Set font to Times New Roman, size to 12
  font(fontname = "Times New Roman", part = "all") %>%
  fontsize(size = 11, part = "all") %>%
  # Add borders only below the header and in specific locations
  hline_top(border = fp_border(color = "black", width = 1.5), part = "header") %>%
```

```
  hline_bottom(border = fp_border(color = "black", width = 1.5), part = "body") %>% # Bottom border
  # Add padding for readability
  padding(padding = 5, part = "all") %>%
  flextable::save_as_docx(path = "univariate_table_pfs.docx")
```

Univariate Analysis - OS

fit cox model for each covariate individually

```
# treatment
fit_tx_os = coxph(Surv(time_os, censor_os)~tx,data=data,
            ties="efron")

# age
fit_age_os = coxph(Surv(time_os, censor_os)~age,data=data,
            ties="efron")
# sex
fit_sex_os = coxph(Surv(time_os, censor_os)~sex,data=data,
            ties="efron")

# race / ethnicity
fit_raceth_os = coxph(Surv(time_os, censor_os)~raceth,data=data,
            ties="efron")

# karnofsky performance scale
fit_karnof_os = coxph(Surv(time_os, censor_os)~karnof,data=data,
            ties="efron")

# cd4 stratum at screening
fit_strat2_os = coxph(Surv(time_os, censor_os)~strat2,data=data,
            ties="efron")

# baseline cd4 count
fit_cd4_os = coxph(Surv(time_os, censor_os)~cd4,data=data,
            ties="efron")

# iv drug use history
fit_ivdrug_os = coxph(Surv(time_os, censor_os)~ivdrug,data=data,
            ties="efron")

# hemophiliac
fit_hemophil_os = coxph(Surv(time_os, censor_os)~hemophil,data=data,
            ties="efron")

# months of prior zdv use
fit_priorzdv_os = coxph(Surv(time_os, censor_os)~priorzdv,data=data,
            ties="efron")
```

place all HR / CI results into table and export to word

```
# make a list of all models
models <- list(
  tx = fit_tx_os,
  age = fit_age_os,
  sex = fit_sex_os,
  raceth = fit_raceth_os,
  karnof = fit_karnof_os,
  strat2 = fit_strat2_os,
  cd4 = fit_cd4_os,
  ivdrug = fit_ivdrug_os,
  hemophil = fit_hemophil_os,
  priorzdv = fit_priorzdv_os
)

# function to extract exp(coef) and 95% CI
extract_coefs <- function(fit) {
  summary_fit <- summary(fit)
  exp_coef <- summary_fit$coefficients[, "exp(coef)"]
  lower_95 <- summary_fit$conf.int[, "lower .95"]
  upper_95 <- summary_fit$conf.int[, "upper .95"]
  p_value <- summary(fit)$coefficients[, "Pr(>|z|)"]

  return(data.frame(HR = round(exp_coef,2), Lower_CI = round(lower_95,2), Upper_CI = round(upper_95, 2)
                    P_Value = round(p_value,3)))
}

# apply the extraction function and combine into single dataframe
results_list <- lapply(models, extract_coefs)
results_df <- do.call(rbind, results_list)
results_df$CI <- paste0("[", results_df$Lower_CI, ", ", results_df$Upper_CI, "]")

# print table
univariate_table_os <- print(results_df)
```

```
##                    HR Lower_CI Upper_CI P_Value           CI
## tx               0.43     0.19     1.00   0.050    [0.19, 1]
## age              1.07     1.03     1.11   0.000 [1.03, 1.11]
## sex              1.22     0.46     3.24   0.686 [0.46, 3.24]
## raceth.raceth1   0.67     0.25     1.81   0.429 [0.25, 1.81]
## raceth.raceth2   1.12     0.40     3.14   0.836  [0.4, 3.14]
## karnof.karnof90  0.23     0.10     0.53   0.001  [0.1, 0.53]
## karnof.karnof100 0.07     0.02     0.32   0.000 [0.02, 0.32]
## strat2           0.29     0.12     0.66   0.003 [0.12, 0.66]
## cd4              0.99     0.98     1.00   0.004    [0.98, 1]
## ivdrug           1.25     0.47     3.32   0.652 [0.47, 3.32]
## hemophil         1.29     0.17     9.51   0.804 [0.17, 9.51]
## priorzdv         0.99     0.97     1.01   0.211 [0.97, 1.01]
```

```
# print table to word
univariate_table_os %>%
```

```
  select(HR, CI, P_Value) %>%
  as.data.frame() %>%
  rownames_to_column("Characteristic") %>%
  flextable() %>%
  # Bold headers
  bold(part = "header") %>%
  # Set font to Times New Roman, size to 12
  font(fontname = "Times New Roman", part = "all") %>%
  fontsize(size = 11, part = "all") %>%
  # Add borders only below the header and in specific locations
  hline_top(border = fp_border(color = "black", width = 1.5), part = "header") %>%
  hline_bottom(border = fp_border(color = "black", width = 1.5), part = "body") %>% # Bottom border
  # Add padding for readability
  padding(padding = 5, part = "all") %>%
  flextable::save_as_docx(path = "univariate_table_os.docx")
```

Multivariate Analysis - PFS

create different multivariate models (full and reduced)

```
# the following variables were significant in the univariate pfs analysis:
## karnof
## strat2
## cd4

# we will always add tx, age, and sex to our multivariate model

# first, we will fit the full model with all significant / relevant variables from univariate analysis
full_model <- coxph(Surv(time_pfs, censor_pfs)~tx+age+sex+karnof+strat2+cd4,
                    data = data, ties="efron")

# next, we will try reduced models by removing one variable and comparing it to the full model
## without karnof
reduced_model1 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + strat2 + cd4, data = data)
## without strat2
reduced_model2 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4, data = data)
## without cd4
reduced_model3 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + strat2, data = data)
## without karnof and strat2
reduced_model4 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + cd4, data = data)
## without karnof and cd4
reduced_model5 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + strat2, data = data)
## without cd4 and strat2
reduced_model6 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof , data = data)
## without cd4, strat2, and karnof
reduced_model7 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex , data = data)
```

create function to compare full and reduced models

```
# Function to compare full and reduced models
compare_cox_models <- function(full_model, reduced_model) {
  # Calculate log-likelihood difference
  loglik_diff <- 2 * (logLik(full_model) - logLik(reduced_model))

  # Calculate degrees of freedom difference
  df_diff <- attr(logLik(full_model), "df") - attr(logLik(reduced_model), "df")

  # Get p-value from chi-squared distribution
  p_value <- pchisq(loglik_diff, df = df_diff, lower.tail = F)

  # Print result
  cat("Log-Likelihood Difference: ", loglik_diff, "\n")
  cat("Degrees of Freedom Difference: ", df_diff, "\n")
  cat("p-value: ", p_value, "\n")
}
```

compare models

```
# full model vs. reduced model 1
compare_cox_models(full_model, reduced_model1) # sign so full model is better
```

```
## Log-Likelihood Difference:  18.99871
## Degrees of Freedom Difference:  2
## p-value:  7.490028e-05
```

```
# full model vs. reduced model 2
compare_cox_models(full_model, reduced_model2) # not sign, so can use reduced model
```

```
## Log-Likelihood Difference:  0.006084998
## Degrees of Freedom Difference:  1
## p-value:  0.937823
```

```
# full model vs. reduced model 3
compare_cox_models(full_model, reduced_model3) # sign so full model is better
```

```
## Log-Likelihood Difference:  19.44341
## Degrees of Freedom Difference:  1
## p-value:  1.036248e-05
```

```
# full model vs. reduced model 4
compare_cox_models(full_model, reduced_model4) # sign so full model is better
```

```
## Log-Likelihood Difference:  19.04777
## Degrees of Freedom Difference:  3
## p-value:  0.0002672504
```

```r
# full model vs. reduced model 5
compare_cox_models(full_model, reduced_model5) # sign so full model is better
```

```
## Log-Likelihood Difference:  41.08419
## Degrees of Freedom Difference:  3
## p-value:  6.275799e-09
```

```r
# full model vs. reduced model 6
compare_cox_models(full_model, reduced_model6) # sign so full model is better
```

```
## Log-Likelihood Difference:  49.81861
## Degrees of Freedom Difference:  2
## p-value:  1.520637e-11
```

```r
# full model vs. reduced model 7
compare_cox_models(full_model, reduced_model7) # sign so full model is better
```

```
## Log-Likelihood Difference:  84.26235
## Degrees of Freedom Difference:  4
## p-value:  2.174971e-17
```

```r
# therefore, we will continue with reduced model 2
```

assess confounding

```r
# Function to assess confounding based on model coefficients
assess_confounding <- function(full_model, reduced_model) {
  # Extract the coefficients for tx from both models
  tx_full <- full_model$coefficients[1]
  tx_reduced <- reduced_model$coefficients[1]

  # Calculate the percent change
  percent_change <- 100 * (tx_reduced - tx_full) / tx_full

  # Return the results
  percent_change
}
```

```r
# compare reduced_model2 with and without cd4
assess_confounding(reduced_model2, reduced_model6) # percent change is 7.1%
```

```
##      tx1
## 7.099693
```

```r
# compare reduced_model2 with and without karnof
assess_confounding(reduced_model2, reduced_model4) # percent change is 0.01%
```

```
##        tx1
## -0.0122251
```

```
# both of these percentages are small enough, so we will move on with both cd4 and karnof in our model
```

assess interaction terms (effect modifiers)

```
reduced_model2_1 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4 + tx*cd4,
                          data = data)
# not sign

reduced_model2_2 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4 + tx*karnof,
                          data = data)
# not sign

reduced_model2_3 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4 + tx*sex,
                          data = data)
# not sign

reduced_model2_4 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4 + tx*age,
                          data = data)
# not sign
```

now we will proceed with our final model

```
reduced_model2 <- coxph(Surv(time_pfs, censor_pfs) ~ tx + age + sex + karnof + cd4, data = data)
summary(reduced_model2)
```

```
## Call:
## coxph(formula = Surv(time_pfs, censor_pfs) ~ tx + age + sex +
##     karnof + cd4, data = data)
##
##   n= 1151, number of events= 96
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## tx1       -0.659987  0.516858  0.215498 -3.063  0.00219 **
## age        0.022901  1.023166  0.011405  2.008  0.04465 *
## sex2       0.089814  1.093970  0.283388  0.317  0.75130
## karnof90  -0.739323  0.477437  0.229315 -3.224  0.00126 **
## karnof100 -1.173882  0.309165  0.293252 -4.003 6.25e-05 ***
## cd4       -0.014720  0.985388  0.002522 -5.838 5.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx1          0.5169     1.9348    0.3388    0.7885
## age          1.0232     0.9774    1.0005    1.0463
## sex2         1.0940     0.9141    0.6277    1.9065
## karnof90     0.4774     2.0945    0.3046    0.7484
## karnof100    0.3092     3.2345    0.1740    0.5493
## cd4          0.9854     1.0148    0.9805    0.9903
##
```

```
## Concordance= 0.778  (se = 0.023 )
## Likelihood ratio test= 98.36  on 6 df,   p=<2e-16
## Wald test            = 78.89  on 6 df,   p=6e-15
## Score (logrank) test = 93.13  on 6 df,   p=<2e-16
```

place all HR / CI results into table and export to word

```
# function to extract exp(coef) and 95% CI
extract_coefs <- function(fit) {
  summary_fit <- summary(fit)
  exp_coef <- summary_fit$coefficients[, "exp(coef)"]
  lower_95 <- summary_fit$conf.int[, "lower .95"]
  upper_95 <- summary_fit$conf.int[, "upper .95"]
  p_value <- summary(fit)$coefficients[, "Pr(>|z|)"]

  return(data.frame(HR = round(exp_coef,2), Lower_CI = round(lower_95,2), Upper_CI = round(upper_95, 2)
                    P_Value = round(p_value,3)))
}

# apply the extraction function and combine into single dataframe
results <- extract_coefs(reduced_model2)
results$CI <- paste0("[", results$Lower_CI, ", ", results$Upper_CI, "]")

# print table
multivariate_table_pfs <- print(results)
```

```
##             HR Lower_CI Upper_CI P_Value           CI
## tx1       0.52     0.34     0.79   0.002 [0.34, 0.79]
## age       1.02     1.00     1.05   0.045    [1, 1.05]
## sex2      1.09     0.63     1.91   0.751 [0.63, 1.91]
## karnof90  0.48     0.30     0.75   0.001  [0.3, 0.75]
## karnof100 0.31     0.17     0.55   0.000 [0.17, 0.55]
## cd4       0.99     0.98     0.99   0.000 [0.98, 0.99]
```

```
# print table to word
multivariate_table_pfs %>%
  select(HR, CI, P_Value) %>%
  as.data.frame() %>%
  rownames_to_column("Characteristic") %>%
  flextable() %>%
  # Bold headers
  bold(part = "header") %>%
  # Set font to Times New Roman, size to 12
  font(fontname = "Times New Roman", part = "all") %>%
  fontsize(size = 11, part = "all") %>%
  # Add borders only below the header and in specific locations
  hline_top(border = fp_border(color = "black", width = 1.5), part = "header") %>%
  hline_bottom(border = fp_border(color = "black", width = 1.5), part = "body") %>% # Bottom border
  # Add padding for readability
  padding(padding = 5, part = "all") %>%
  flextable::save_as_docx(path = "multivariate_table_pfs.docx")
```

Multivariate Analysis - OS

create different multivariate models (full and reduced)

```
# the following variables were significant in the univariate pfs analysis:
## karnof
## strat2
## cd4

# we will always add tx, age, and sex to our multivariate model

# first, we will fit the full model with all significant / relevant variables from univariate analysis
full_model <- coxph(Surv(time_os, censor_os)~tx+age+sex+karnof+strat2+cd4,
                    data = data, ties="efron")

# next, we will try reduced models by removing one variable and comparing it to the full model
## without karnof
reduced_model1 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + strat2 + cd4, data = data)
## without strat2
reduced_model2 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + karnof + cd4, data = data)
## without cd4
reduced_model3 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + karnof + strat2, data = data)
## without karnof and strat2
reduced_model4 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + cd4, data = data)
## without karnof and cd4
reduced_model5 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + strat2, data = data)
## without cd4 and strat2
reduced_model6 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex + karnof, data = data)
## without cd4, strat2, and karnof
reduced_model7 <- coxph(Surv(time_os, censor_os) ~ tx + age + sex , data = data)
```

compare models (using previously created function)

```
# full model vs. reduced model 1
compare_cox_models(full_model, reduced_model1) # sign so full model is better
```

```
## Log-Likelihood Difference:  13.08342
## Degrees of Freedom Difference:  2
## p-value:  0.001442019
```

```
# full model vs. reduced model 2
compare_cox_models(full_model, reduced_model2) # not sign, so can use reduced model
```

```
## Log-Likelihood Difference:  0.9691912
## Degrees of Freedom Difference:  1
## p-value:  0.324882
```

```
# full model vs. reduced model 3
compare_cox_models(full_model, reduced_model3) # not sign, so can use reduced model
```

```
## Log-Likelihood Difference:  1.271633
## Degrees of Freedom Difference:  1
## p-value:  0.2594606
```

```
# full model vs. reduced model 4
compare_cox_models(full_model, reduced_model4) # sign so full model is better
```

```
## Log-Likelihood Difference:  14.42371
## Degrees of Freedom Difference:  3
## p-value:  0.002381634
```

```
# full model vs. reduced model 5
compare_cox_models(full_model, reduced_model5) # sign so full model is better
```

```
## Log-Likelihood Difference:  15.22228
## Degrees of Freedom Difference:  3
## p-value:  0.001636216
```

```
# full model vs. reduced model 6
compare_cox_models(full_model, reduced_model6) # sign so full model is better
```

```
## Log-Likelihood Difference:  9.489914
## Degrees of Freedom Difference:  2
## p-value:  0.008695435
```

```
# full model vs. reduced model 7
compare_cox_models(full_model, reduced_model7) # sign so full model is better
```

```
## Log-Likelihood Difference:  28.35475
## Degrees of Freedom Difference:  4
## p-value:  1.056916e-05
```

```
# therefore, we will continue with reduced model 2 and reduced model 3
```

assess confounding (using previously created function)

```
# compare reduced_model2 with and without cd4
assess_confounding(reduced_model2, reduced_model6) # percent change is 2.0%
```

```
##      tx1
## 2.009869
```

```
# compare reduced_model2 with and without karnof
assess_confounding(reduced_model2, reduced_model4) # percent change is 4.955%
```

```
##      tx1
## 4.955818
```

```
# both of these percentages are small enough, so we will move on with both cd4 and karnof in our model

# compare reduced_model2 with and without strat2
assess_confounding(reduced_model3, reduced_model6) # percent change is 2.53%
```

```
##       tx1
## -2.534817
```

```
# compare reduced_model2 with and without karnof
assess_confounding(reduced_model3, reduced_model4) # percent change is 0.027%
```

```
##       tx1
## 0.2798851
```

```
# therefore, both reduced_model2 and reduced_model3 don't have a significant confounders. however, redu
```

check the significance of covariates in both models to see which ones are more significant

```
summary(reduced_model2)
```

```
## Call:
## coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof +
##     cd4, data = data)
##
##   n= 1151, number of events= 26
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## tx1       -0.811817  0.444050  0.425777 -1.907 0.056563 .
## age        0.073123  1.075863  0.020033  3.650 0.000262 ***
## sex2       0.468703  1.597921  0.500027  0.937 0.348576
## karnof90  -1.056678  0.347609  0.431565 -2.448 0.014346 *
## karnof100 -2.145312  0.117031  0.760539 -2.821 0.004791 **
## cd4       -0.011095  0.988967  0.004346 -2.553 0.010688 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx1          0.4441     2.2520   0.19276    1.0229
## age          1.0759     0.9295   1.03444    1.1189
## sex2         1.5979     0.6258   0.59970    4.2577
## karnof90     0.3476     2.8768   0.14919    0.8099
## karnof100    0.1170     8.5447   0.02636    0.5196
## cd4          0.9890     1.0112   0.98058    0.9974
##
## Concordance= 0.846  (se = 0.035 )
## Likelihood ratio test= 45.18  on 6 df,   p=4e-08
## Wald test            = 39.43  on 6 df,   p=6e-07
## Score (logrank) test = 49.46  on 6 df,   p=6e-09
```

```r
summary(reduced_model3)
```

```
## Call:
## coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof +
##     strat2, data = data)
##
##   n= 1151, number of events= 26
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## tx1       -0.84967   0.42756  0.42662 -1.992 0.046409 *
## age        0.07511   1.07800  0.01992  3.770 0.000163 ***
## sex2       0.48615   1.62605  0.50149  0.969 0.332340
## karnof90  -1.09663   0.33400  0.43006 -2.550 0.010773 *
## karnof100 -2.16241   0.11505  0.75952 -2.847 0.004412 **
## strat21   -1.22680   0.29323  0.44779 -2.740 0.006151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx1          0.4276     2.3389   0.18529    0.9866
## age          1.0780     0.9276   1.03672    1.1209
## sex2         1.6260     0.6150   0.60850    4.3452
## karnof90     0.3340     2.9940   0.14377    0.7759
## karnof100    0.1150     8.6921   0.02596    0.5098
## strat21      0.2932     3.4103   0.12191    0.7053
##
## Concordance= 0.856  (se = 0.03 )
## Likelihood ratio test= 44.88  on 6 df,   p=5e-08
## Wald test            = 40.21  on 6 df,   p=4e-07
## Score (logrank) test = 50.52  on 6 df,   p=4e-09
```

```r
# covariates in reduced_model3 have overall lower p values than in reduced_model3
```

as a result, even though both reduced_model2 and reduced_model3 are effective models, reduced_model3 has slightly less confounding and slightly more significance

assess interaction terms (effect modifiers)

```r
reduced_model3_1 <- coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof + strat2 + tx*ka
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 8 ; coefficient may be infinite.
```

```r
# not sign

reduced_model3_2 <- coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof + strat2 + tx*st
# not sign

reduced_model3_3 <- coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof + strat2 + tx*ag
# not sign
```

```
reduced_model3_4 <- coxph(formula = Surv(time_os, censor_os) ~ tx + age + sex + karnof + strat2 + tx*se
# not sign
```

now we will proceed with our final model

```
reduced_model3 <- coxph(Surv(time, censor) ~ tx + age + sex + karnof + strat2, data = data)
summary(reduced_model3)
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ tx + age + sex + karnof +
##     strat2, data = data)
##
##   n= 1151, number of events= 96
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## tx1       -0.68845   0.50236  0.21556 -3.194 0.001404 **
## age        0.02169   1.02193  0.01139  1.904 0.056945 .
## sex2       0.06618   1.06842  0.28420  0.233 0.815877
## karnof90  -0.80946   0.44510  0.22978 -3.523 0.000427 ***
## karnof100 -1.24429   0.28815  0.29333 -4.242 2.22e-05 ***
## strat21   -1.20143   0.30076  0.22841 -5.260 1.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx1          0.5024     1.9906    0.3292    0.7665
## age          1.0219     0.9785    0.9994    1.0450
## sex2         1.0684     0.9360    0.6121    1.8649
## karnof90     0.4451     2.2467    0.2837    0.6983
## karnof100    0.2881     3.4705    0.1622    0.5120
## strat21      0.3008     3.3249    0.1922    0.4706
##
## Concordance= 0.746  (se = 0.026 )
## Likelihood ratio test= 78.92  on 6 df,   p=6e-15
## Wald test            = 76.01  on 6 df,   p=2e-14
## Score (logrank) test = 88.14  on 6 df,   p=<2e-16
```

place all HR / CI results into table and export to word

```
# function to extract exp(coef) and 95% CI
extract_coefs <- function(fit) {
  summary_fit <- summary(fit)
  exp_coef <- summary_fit$coefficients[, "exp(coef)"]
  lower_95 <- summary_fit$conf.int[, "lower .95"]
  upper_95 <- summary_fit$conf.int[, "upper .95"]
  p_value <- summary(fit)$coefficients[, "Pr(>|z|)"]

  return(data.frame(HR = round(exp_coef,2), Lower_CI = round(lower_95,2), Upper_CI = round(upper_95, 2)
                    P_Value = round(p_value,3)))
```

```
}

# apply the extraction function and combine into single dataframe
results <- extract_coefs(reduced_model3)
results$CI <- paste0("[", results$Lower_CI, ", ", results$Upper_CI, "]")

# print table
multivariate_table_os <- print(results)
```

```
##             HR Lower_CI Upper_CI P_Value           CI
## tx1       0.50     0.33     0.77   0.001 [0.33, 0.77]
## age       1.02     1.00     1.05   0.057    [1, 1.05]
## sex2      1.07     0.61     1.86   0.816 [0.61, 1.86]
## karnof90  0.45     0.28     0.70   0.000  [0.28, 0.7]
## karnof100 0.29     0.16     0.51   0.000 [0.16, 0.51]
## strat21   0.30     0.19     0.47   0.000 [0.19, 0.47]
```

```
# print table to word
multivariate_table_os %>%
  select(HR, CI, P_Value) %>%
  as.data.frame() %>%
  rownames_to_column("Covariate") %>%
  flextable() %>%
  # Bold headers
  bold(part = "header") %>%
  # Set font to Times New Roman, size to 12
  font(fontname = "Times New Roman", part = "all") %>%
  fontsize(size = 11, part = "all") %>%
  # Add borders only below the header and in specific locations
  hline_top(border = fp_border(color = "black", width = 1.5), part = "header") %>%
  hline_bottom(border = fp_border(color = "black", width = 1.5), part = "body") %>% # Bottom border
  # Add padding for readability
  padding(padding = 5, part = "all") %>%
  flextable::save_as_docx(path = "multivariate_table_os.docx")
```