Sta_Project_1

Yuxin Li

2025-12-14

Full code and results are presented on https://github.com/yuxin188/Effectiveness-of-SGLT2-Inhibitors-in-Secondary-MI-Prevention.git

```r
# import packages for analysis and plots
library(survival)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```r
library(survminer)
```

```
## Loading required package: ggpubr

##
## Attaching package: 'survminer'

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

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```r
# import data
sglt2_data <- read.csv("/Users/yuxin/Desktop/ic/intro to sat/project1/mi_sglt2.csv")

# checking missing value
vars_to_check <- c("age", "sex", "bmi_bl", "sbp_bl", "sbp_fup", "mi_flg", "alloc")
# Number of missing values for each variable
for (v in vars_to_check) {
  cat(v, ":", sum(is.na(sglt2_data[[v]])), "\n")
}
```

```
## age : 0
## sex : 0
## bmi_bl : 0
## sbp_bl : 0
## sbp_fup : 0
## mi_flg : 0
## alloc : 0
```

```r
# Check total number of rows and unique IDs
nrow(sglt2_data)
```

```
## [1] 1988
```

```r
length(unique(sglt2_data$id))
```

```
## [1] 1988
```

Aim1

```r
# create a new dataset which only contains participants allocated to SGLT2 inhibitors for analyses of a
sglt2_arm <- subset(sglt2_data, alloc == "SGLT2")

# check adherence variable
#table(sglt2_arm$adherence_flg, useNA = "ifany")
#table(sglt2_arm$sex, useNA = "ifany")
#summary(sglt2_arm$age)
#sum(is.na(sglt2_arm$age))

# fit the logistic regression
fit_1 <- glm(adherence_flg ~ age + sex + bmi_bl, family=binomial(), data=sglt2_arm)
summary(fit_1)
```

part1: Does participant age or sex affect adherence to SGLT2 inhibitors?

```
##
## Call:
## glm(formula = adherence_flg ~ age + sex + bmi_bl, family = binomial(),
##     data = sglt2_arm)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.513401   0.798209   0.643  0.52010
## age         -0.009815   0.011236  -0.874  0.38237
## sexMale     -0.214216   0.238242  -0.899  0.36857
## bmi_bl       0.057946   0.020414   2.839  0.00453 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
## 
##     Null deviance: 953.59  on 996  degrees of freedom
## Residual deviance: 945.16  on 993  degrees of freedom
## AIC: 953.16
## 
## Number of Fisher Scoring iterations: 4
```

```r
exp(coef(fit_1))
```

```
## (Intercept)         age      sexMale      bmi_bl
##   1.6709645   0.9902327   0.8071741   1.0596574
```

```r
exp(confint(fit_1))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %    97.5 %
## (Intercept) 0.3553282 8.160159
## age         0.9682805 1.011954
## sexMale     0.4971277 1.268993
## bmi_bl      1.0184736 1.103407
```

Aim2

```r
# factor as date
sglt2_data$randomisation_dte <- as.Date(sglt2_data$randomisation_dte)
sglt2_data$mi_dte <- as.Date(sglt2_data$mi_dte)
sglt2_data$endfup_dte <- as.Date(sglt2_data$endfup_dte)

# duration for participants who develop MI
time_to_mi <- difftime(sglt2_data$mi_dte, sglt2_data$randomisation_dte, units = "days")

# duration for participants who do not develop MI
time_to_censor <- difftime(sglt2_data$endfup_dte, sglt2_data$randomisation_dte, units = "days")

# duration for two possible outcomes
sglt2_data$time_to_mi_or_censor <- as.numeric(ifelse(sglt2_data$mi_flg == 1, time_to_mi, time_to_censor)

# median time
median(sglt2_data$time_to_mi_or_censor)
```

part1 Whether SGLT2 inhibitors affected incidence of secondary MI

```
## [1] 1826
```

```r
# change ref to Usual Care
sglt2_data$alloc <- factor(sglt2_data$alloc)
sglt2_data$alloc <- relevel(sglt2_data$alloc, ref = "Usual Care")
```

```
fit_2 <- coxph(formula = Surv(time_to_mi_or_censor, mi_flg) ~ alloc, data = sglt2_data)

summary(fit_2)
```

```
## Call:
## coxph(formula = Surv(time_to_mi_or_censor, mi_flg) ~ alloc, data = sglt2_data)
##
##    n= 1988, number of events= 623
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## allocSGLT2 -0.24122   0.78567  0.08045 -2.998  0.00272 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## allocSGLT2    0.7857      1.273    0.6711    0.9199
##
## Concordance= 0.532  (se = 0.01 )
## Likelihood ratio test= 9.03  on 1 df,   p=0.003
## Wald test            = 8.99  on 1 df,   p=0.003
## Score (logrank) test = 9.03  on 1 df,   p=0.003
```

```
exp(confint((fit_2)))
```

```
##                 2.5 %    97.5 %
## allocSGLT2 0.6710573 0.9198617
```

```
km_fit <- survfit(Surv(time_to_mi_or_censor, mi_flg) ~ alloc, data = sglt2_data)
km_plot <- ggsurvplot(
  km_fit,
  data = sglt2_data,
  conf.int = FALSE,
  censor = TRUE,

  xlab = "Time since randomisation (years)",
  ylab = "MI-free probability",
  break.time.by = 365,
  xscale = "d_y",

  legend.title = "Treatment group",
  legend.labs = c("Usual care", "SGLT2 inhibitors"),
  legend = "top",

  palette = c("blue", "red"),
  ggtheme = theme_bw(base_size = 12)
)
```
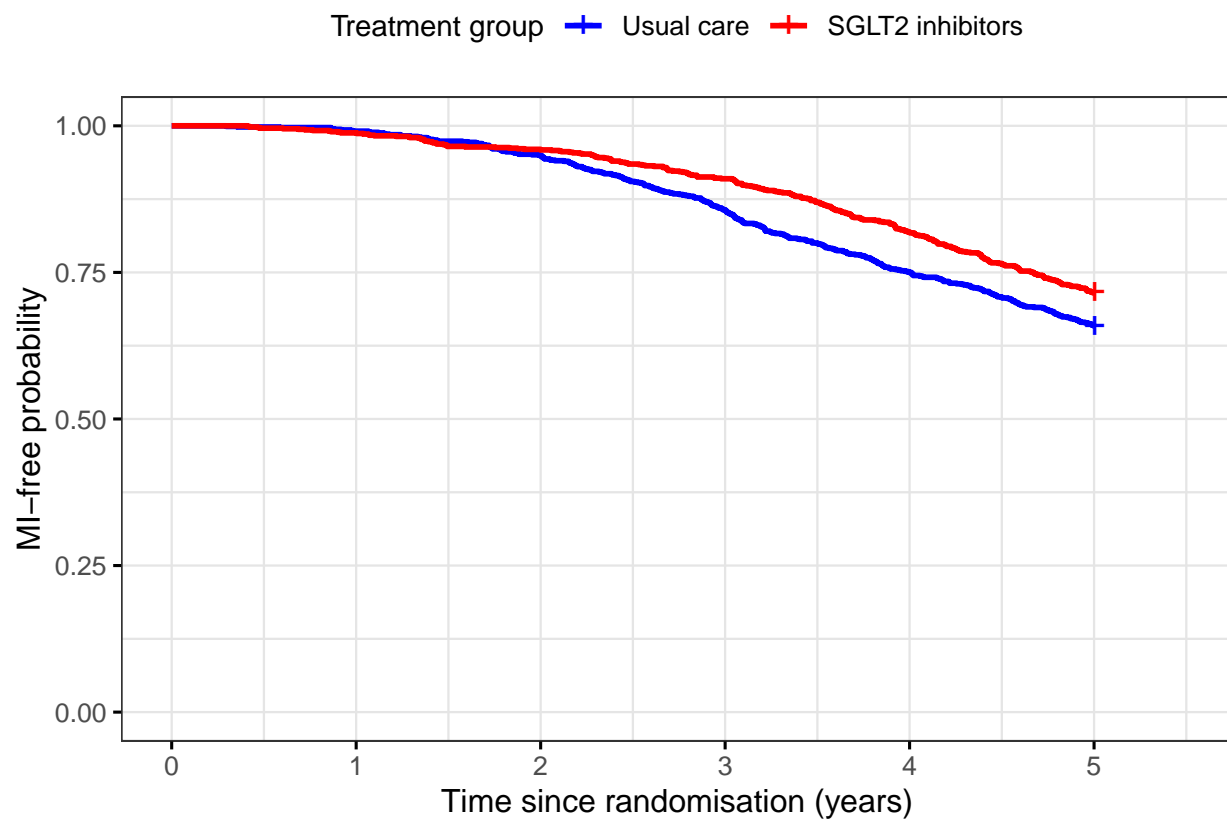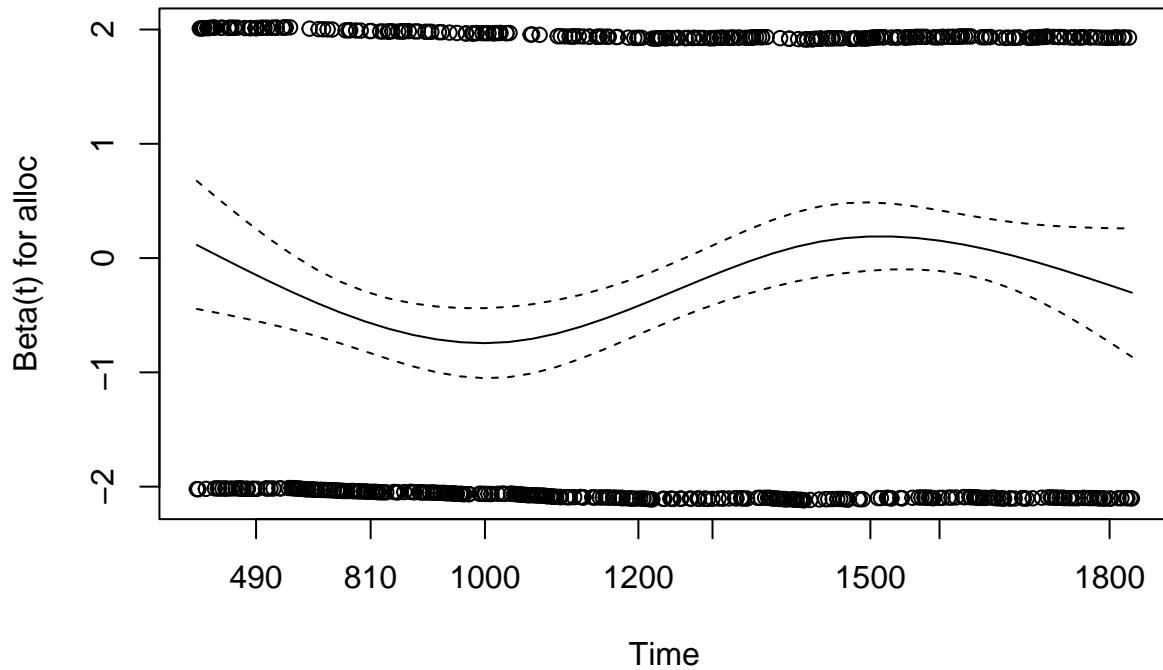
K-M curve

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once per session.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```r
km_plot$plot <- km_plot$plot +
  theme(
    legend.title = element_text(size = 11),
    legend.text  = element_text(size = 10),
    axis.title   = element_text(size = 12),
    axis.text    = element_text(size = 10),
    panel.grid.major = element_line(colour = "grey90", linewidth = 0.4),
    panel.grid.minor = element_line(colour = "grey90",linewidth = 0.4)
  )

km_plot
```



```r
plot(cox.zph(fit_2))
```

```
fit_3 <- coxph(formula = Surv(time_to_mi_or_censor, mi_flg) ~ alloc* sex, data = sglt2_data)
summary(fit_3)
```

part2 analysis of the main trial outcomes by sex

```
## Call:
## coxph(formula = Surv(time_to_mi_or_censor, mi_flg) ~ alloc *
##     sex, data = sglt2_data)
##
##   n= 1988, number of events= 623
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## allocSGLT2        -0.1172    0.8894   0.2501 -0.468   0.6395
## sexMale            0.5635    1.7568   0.1832  3.075   0.0021 **
## allocSGLT2:sexMale -0.1391   0.8701   0.2642 -0.527   0.5985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## allocSGLT2           0.8894     1.1243    0.5448     1.452
## sexMale              1.7568     0.5692    1.2267     2.516
## allocSGLT2:sexMale   0.8701     1.1492    0.5185     1.460
##
```

```
## Concordance= 0.553  (se = 0.011 )
## Likelihood ratio test= 25.58  on 3 df,   p=1e-05
## Wald test            = 23.91  on 3 df,   p=3e-05
## Score (logrank) test = 24.41  on 3 df,   p=2e-05
```

```r
# time split (0 to 1.5), (1.5 to 5)
tcuts <- c(1.5*365.25) #transfer to days
time_split <- survSplit(Surv(time_to_mi_or_censor, mi_flg) ~ .,
                        data = sglt2_data,
                        cut = tcuts,
                        episode = "period")

time_split$period <- factor(time_split$period, levels = c(1,2),
                            labels = c("0-1.5y","1.5-5y"))

fit_4 <- coxph(formula = Surv(tstart, time_to_mi_or_censor, mi_flg) ~ alloc * period, data = time_split)

summary(fit_4)
```

part3 interaction with time

```
## Call:
## coxph(formula = Surv(tstart, time_to_mi_or_censor, mi_flg) ~
##     alloc * period, data = time_split)
##
##   n= 3915, number of events= 623
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## allocSGLT2              0.2950    1.3431   0.2589  1.139   0.2545
## period1.5-5y                NA        NA   0.0000     NA       NA
## allocSGLT2:period1.5-5y -0.5947    0.5518   0.2725 -2.182   0.0291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## allocSGLT2                 1.3431     0.7445    0.8086    2.2310
## period1.5-5y                   NA         NA        NA        NA
## allocSGLT2:period1.5-5y    0.5518     1.8124    0.3235    0.9412
##
## Concordance= 0.541  (se = 0.01 )
## Likelihood ratio test= 13.86  on 2 df,   p=0.001
## Wald test            = 13.75  on 2 df,   p=0.001
## Score (logrank) test = 13.85  on 2 df,   p=0.001
```

```r
# HR in each period from the time-split interaction model
b  <- coef(fit_4)
V  <- vcov(fit_4)

# 0-1.5 years: exp(beta_alloc)
hr_0_15 <- exp(b["allocSGLT2"])
ci_0_15 <- exp(confint(fit_4)["allocSGLT2", ])
```

```
p_0_15  <- summary(fit_4)$coefficients["allocSGLT2", "Pr(>|z|)"]

# 1.5-5 years: exp(beta_alloc + beta_interaction)
L <- c(allocSGLT2 = 1, `allocSGLT2:period1.5-5y` = 1)
beta <- b[names(L)]
Vsub <- V[names(L), names(L)]
est  <- sum(L * beta)
se   <- sqrt(t(L) %*% Vsub %*% L)

hr_15_5 <- exp(est)
ci_15_5 <- exp(c(est - 1.96 * se, est + 1.96 * se))
p_15_5  <- 2 * pnorm(-abs(est / se))

list(
  `0-1.5y` = c(HR = hr_0_15,  LCL = ci_0_15[1], UCL = ci_0_15[2], p = p_0_15),
  `1.5-5y` = c(HR = hr_15_5, LCL = ci_15_5[1], UCL = ci_15_5[2], p = p_15_5)
)
```

```
## $`0-1.5y`
## HR.allocSGLT2    LCL.2.5 %    UCL.97.5 %            p
##     1.3431282    0.8086025     2.2310014    0.2545327
##
## $`1.5-5y`
##           HR           LCL           UCL            p
## 0.7410766204 0.6274424941 0.8752906639 0.0004180112
```

AIm3

```
sglt2_data$alloc <- relevel(factor(sglt2_data$alloc), ref="Usual Care")

fit_5 <- lm(sbp_fup ~ alloc + sbp_bl, data = sglt2_data)
summary(fit_5)
```

part1 Whether SGLT2 inhibitors affected systolic blood pressure at the interim assessment.

```
##
## Call:
## lm(formula = sbp_fup ~ alloc + sbp_bl, data = sglt2_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.0186  -5.5414  -0.0836   5.5835  27.1038
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.73463    2.13975   0.343    0.731
## allocSGLT2  -1.91564    0.35897  -5.336 1.06e-07 ***
## sbp_bl       0.94388    0.01495  63.117  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 8.002 on 1985 degrees of freedom
## Multiple R-squared:  0.6694, Adjusted R-squared:  0.6691
## F-statistic:  2010 on 2 and 1985 DF,  p-value: < 2.2e-16
```

```
confint((fit_5))
```

```
##                  2.5 %      97.5 %
## (Intercept) -3.4617518  4.9310155
## allocSGLT2  -2.6196462 -1.2116381
## sbp_bl       0.9145541  0.9732099
```

```
fit_6 <- lm(sbp_fup ~ alloc * sex + sbp_bl, data = sglt2_data)
summary(fit_6)
```

part2 analysis of the main trial outcomes by sex

```
## 
## Call:
## lm(formula = sbp_fup ~ alloc * sex + sbp_bl, data = sglt2_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.9874  -5.5340  -0.0434   5.5660  26.9275
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.36877    2.21804   0.617  0.53723
## allocSGLT2        -2.40017    0.92575  -2.593  0.00959 **
## sexMale           -0.77437    0.71139  -1.089  0.27650
## sbp_bl             0.94405    0.01496  63.114  < 2e-16 ***
## allocSGLT2:sexMale 0.57029    1.00438   0.568  0.57023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.004 on 1983 degrees of freedom
## Multiple R-squared:  0.6696, Adjusted R-squared:  0.669
## F-statistic:  1005 on 4 and 1983 DF,  p-value: < 2.2e-16
```

```
confint(fit_6)
```

```
##                        2.5 %      97.5 %
## (Intercept)        -2.9811620  5.7187010
## allocSGLT2         -4.2157197 -0.5846142
## sexMale            -2.1695150  0.6207846
## sbp_bl              0.9147146  0.9733844
## allocSGLT2:sexMale -1.3994627  2.5400417
```