'''

C18 ~ C20 condition\_concept\_id in NCC: 74582, 436635, 4247719, 438699, 432257, 441800, 197500, 432837, 438979, 433143

Chemotherapy drug concept id in NCC: 955632, 43259208, 1318011, 42959014, 1367268, 35202021, 40244266, 42921872, 1397141, 1338512, 1315411, 1378382, 43269829, 1344905, 1388798, 1308290, 42936747

antithromboticstics: 1301025, 40228152, 1350310, 1112807, 40241331, 1301065, 19042778, 40241186, 1322184, 1315865, 45892847, 42935771, 43013024, 1310149, 42960517, 42962067

Antihypertensive: 1332418, 19078080, 42950421, 1373225, 1347384, 19023453, 1308842, 40167849, 1307046, 42950393, 40167852, 19102171, 42960031, 1334456, 1363053, 1318853, 956874, 1346823, 1367500, 40235487, 19078101, 40184217, 978555, 19127434, 40224172, 1317640, 1332527, 1332497, 1318137, 19127432, 19102491, 42950354, 21130969, 19023454, 19096740, 970250, 974166, 1314002, 1328165, 1386957, 1319133, 1346686, 21091811, 42950567, 974474, 1314577, 19127433, 1353766, 40184187, 40235491, 19096752, 42960035, 974473, 1307863, 19122209, 1340128, 21141332, 1341927, 1332525

DM:1516766, 40164922, 19059796, 42953740, 42942982, 1503297, 45775620, 42968754, 36887702, 42961500, 1525215, 1597756, 42961331, 42942973, 40166035, 42708168, 1502826, 40231403, 1529331, 44785829, 43013884, 40239216, 40164892, 21081251, 45774751, 45775456, 42960653, 40231394, 1580747, 42942980, 42942984

statin:19022956, 1539403, 1545958, 42969306, 1551860, 42969173, 1592085, 1551803, 40165636, 43527029, 1526479, 1510813, 42969232, 19122209, 43527032, 1332497

'''

'''

1.PERSON 테이블 -> pperson

2-1.VISIT\_OCCURRENCE 테이블 -> visit\_occ

3.DEATH 테이블 -> death

4.CONDITION\_OCCURRENCE -> cond\_occ

6.DRUG 테이블 -> drug

'''

library(skimr)

library(tidyverse)

library(lubridate)

library(survival)

library(survminer)

## read table

pperson <- read.csv(file = "PERSON.csv")

visit\_occ <- read.csv(file = "VISIT\_OCCURRENCE.csv")

death <- read.csv(file = "DEATH.csv")

cond\_occ <- read.csv(file = "CONDITION\_OCCURRENCE.csv")

drug <- read.csv(file = "DRUG.csv")

## C18 ~ C20에 해당하는 환자 추출

colon\_list <- c('74582', '436635', '4247719', '438699', '432257', '441800', '197500', '432837', '438979', '433143')

colon\_cond <- subset(cond\_occ, cond\_occ$condition\_concept\_id %in% colon\_list)

colon\_cond <- arrange(colon\_cond, condition\_start\_date)

colon\_cond <- colon\_cond[-which(duplicated(colon\_cond$person\_id)),]

## colon cancer 환자의 성별,나이,사망 기록 join

visit\_occ <- arrange(visit\_occ, -visit\_end\_date)

visit\_occ <- visit\_occ[-which(duplicated(visit\_occ$person\_id)),]

visit\_occ <- subset(visit\_occ, select = c(person\_id, visit\_end\_date))

colon\_cond <- left\_join(colon\_cond, pperson, by = 'person\_id')

colon\_cond <- left\_join(colon\_cond, visit\_occ, by = 'person\_id')

colon\_cond <- left\_join(colon\_cond, death, by = 'person\_id')

colon\_cond$outcomeCount <- ifelse(is.na(colon\_cond$death\_time) == TRUE, 0 , 1)

colon\_cond$age <- as.integer(substr(colon\_cond$condition\_start\_date, 1, 4)) - colon\_cond$year\_of\_birth

colon\_cond$deathtime <- as.integer(colon\_cond$death\_date - colon\_cond$condition\_start\_date)

colon\_cond$followlosstime <- as.integer(colon\_cond$visit\_end\_date - colon\_cond$condition\_start\_date)

colon\_cond$surtime <- ifelse(colon\_cond$outcomeCount == 1, colon\_cond$deathtime, colon\_cond$followlosstime)

colon\_cond$survivalTime <- ifelse(colon\_cond$survtime>1825, 1825, colon\_cond$survtime)

## 각 약제 추출

drug$end\_date <- ifelse(is.na(drug$drug\_exposure\_end\_date) == FALSE, drug$drug\_exposure\_end\_date, ifelse(is.na(drug$days\_supply) == FALSE, drug$drug\_exposure\_start\_date+drug$days\_supply, drug$drug\_exposure\_start\_date+1))

drug$end\_date <- as.Date(drug$end\_date, origin="1970-01-01")

CTx\_list <- c('955632', '43259208', '1318011', '42959014', '1367268', '35202021', '40244266', '42921872', '1397141', '1338512', '1315411', '1378382', '43269829', '1344905', '1388798', '1308290', '42936747')

BPmed\_list <- c('1332418', '19078080', '42950421', '1373225', '1347384', '19023453', '1308842', '40167849', '1307046', '42950393', '40167852', '19102171', '42960031', '1334456', '1363053', '1318853', '956874', '1346823', '1367500', '40235487', '19078101', '40184217', '978555', '19127434', '40224172', '1317640', '1332527', '1332497', '1318137', '19127432', '19102491', '42950354', '21130969', '19023454', '19096740', '970250', '974166', '1314002', '1328165', '1386957', '1319133', '1346686', '21091811', '42950567', '974474', '1314577', '19127433', '1353766', '40184187', '40235491', '19096752', '42960035', '974473', '1307863', '19122209', '1340128', '21141332', '1341927', '1332525')

DMmed\_list <- c('1516766', '40164922', '19059796', '42953740', '42942982', '1503297', '45775620', '42968754', '36887702', '42961500', '1525215', '1597756', '42961331', '42942973', '40166035', '42708168', '1502826', '40231403', '1529331', '44785829', '43013884', '40239216', '40164892', '21081251', '45774751', '45775456', '42960653', '40231394', '1580747', '42942980', '42942984')

antithromboticstics\_list <- c('1301025', '40228152', '1350310', '1112807', '40241331', '1301065', '19042778', '40241186', '1322184', '1315865', '45892847', '42935771', '43013024', '1310149', '42960517', '42962067')

statin\_list <- c('19022956', '1539403', '1545958', '42969306', '1551860', '42969173', '1592085', '1551803', '40165636', '43527029', '1526479', '1510813', '42969232', '19122209', '43527032', '1332497')

CTx <- subset(drug, drug$drug\_concept\_id %in% CTx\_list)

CTx <- arrange(CTx, drug\_exposure\_start\_date)

CTx <- CTx[-which(duplicated(CTx$person\_id)),]

HTN <- subset(drug, drug$drug\_concept\_id %in% BPmed\_list)

HTN <- arrange(HTN, drug\_exposure\_start\_date)

HTN <- HTN[-which(duplicated(HTN$person\_id)),]

DM <- subset(drug, drug$drug\_concept\_id %in% DMmed\_list)

DM <- arrange(DM, drug\_exposure\_start\_date)

DM <- DM[-which(duplicated(DM$person\_id)),]

statin <- subset(drug, drug$drug\_concept\_id %in% statin\_list)

statin <- arrange(statin, drug\_exposure\_start\_date)

statin <- statin[-which(duplicated(statin$person\_id)),]

antithrombotics <- subset(drug, drug$drug\_concept\_id %in% antithromboticstics\_list)

antithrombotics <- arrange(antithrombotics, drug\_exposure\_start\_date)

antithrombotics <- antithrombotics[-which(duplicated(antithrombotics$person\_id)),]

HTN$HTN <- 1

DM$DM <- 1

statin$statin <- 1

antithrombotics$antithrombotics <- 1

HTN <- subset(HTN, select = c(person\_id, HTN))

DM <- subset(DM, select = c(person\_id, DM))

statin <- subset(statin, select = c(person\_id, statin))

antithrombotics <- subset(antithrombotics, select = c(person\_id, antithrombotics))

## condition table과 drug table 결합

colon\_CTx <- left\_join(colon\_cond, CTx, by = 'person\_id')

colon\_CTx <- subset(colon\_CTx, is.na(colon\_CTx$drug\_concept\_id) == FALSE)

colon\_CTx <- left\_join(colon\_CTx, HTN, by = 'person\_id')

colon\_CTx <- left\_join(colon\_CTx, DM, by = 'person\_id')

colon\_CTx <- left\_join(colon\_CTx, statin, by = 'person\_id')

colon\_CTx <- left\_join(colon\_CTx, antithrombotics, by = 'person \_id')

finaldata <- subset(colon\_CTx, select = c(person\_id, HTN, DM, statin, gender\_concept\_id, age, antithrombotics, survivalTime, outcomeCount))

finaldata[is.na(finaldata)] <- 0

finaldata$HTN <- as.factor(finaldata$HTN)

finaldata$DM <- as.factor(finaldata$DM)

finaldata$statin <- as.factor(finaldata$statin)

finaldata$antithrombotics <- as.factor(finaldata$antithrombotics)

finaldata$gender\_concept\_id <- as.factor(finaldata$gender\_concept\_id)

finaldata$outcomeCount <- as.factor(finaldata$outcomeCount)

write.csv(finaldata, file = "finaldata.csv", row.names = FALSE)

sink("demographic1.txt")

print(skim(finaldata))

sink()

sink("demographic2.txt")

print(summary(finaldata))

sink()

res.cox <- coxph(Surv(survivalTime, outcomeCount) ~ HTN + DM + statin + antithrombotics + age + gender\_concept\_id, data = finaldata)

sink("cox.txt")

print(summary(res.cox))

sink()

surv\_object <- Surv(time = finaldata$survivalTime, event = finaldata$outcomeCount)

fit\_HTN <- survfit(surv\_object ~ HTN, data=finaldata)

survplot\_HTN <- ggsurvplot(fit\_HTN, data = finaldata, pval=TRUE)

png("survplot\_HTN.png")

print(survplot\_HTN, newpage = FALSE)

dev.off()

fit\_DM <- survfit(surv\_object ~ DM, data=finaldata)

survplot\_DM <- ggsurvplot(fit\_DM, data = finaldata, pval=TRUE)

png("survplot\_DM.png")

print(survplot\_DM, newpage = FALSE)

dev.off()

fit\_statin <- survfit(surv\_object ~ statin, data=finaldata)

survplot\_statin <- ggsurvplot(fit\_statin, data = finaldata, pval=TRUE)

png("survplot\_statin.png")

print(survplot\_statin, newpage = FALSE)

dev.off()

fit\_antithrombotics <- survfit(surv\_object ~ antithrombotics, data=finaldata)

survplot\_antithrombotics <- ggsurvplot(fit\_antithrombotics, data = finaldata, pval=TRUE)

png("survplot\_antithrombotics.png")

print(survplot\_antithrombotics, newpage = FALSE)

dev.off()

fit\_gender\_concept\_id <- survfit(surv\_object ~ gender\_concept\_id, data=finaldata)

survplot\_gender\_concept\_id <- ggsurvplot(fit\_gender\_concept\_id, data = finaldata, pval=TRUE)

png("survplot\_gender\_concept\_id.png")

print(survplot\_gender\_concept\_id, newpage = FALSE)

dev.off()