Untitled

Adriana RM

10/2/2020

#Load Packages

library(haven)  
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

library(reshape2)  
library(ggplot2)  
library(esquisse)  
library(haven)  
library(stringr)  
library(broom)  
library(SmartEDA)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(summarytools)

## Registered S3 method overwritten by 'pryr':  
## method from  
## print.bytes Rcpp

## Warning in fun(libname, pkgname): couldn't connect to display ":0"

## system might not have X11 capabilities; in case of errors when using dfSummary(), set st\_options(use.x11 = FALSE)

## For best results, restart R session and update pander using devtools:: or remotes::install\_github('rapporter/pander')

library(stats)  
library(kableExtra)

##   
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':  
##   
## group\_rows

library(DataExplorer)  
library(scales)  
library(RColorBrewer)  
library(ecm)

## Loading the Data

#Print Dimensions  
dim(dfor)

## [1] 8920 431

#Print Unique Record ID  
cat("\n Original: Record, Patient, Hospital Counts \n")

##   
## Original: Record, Patient, Hospital Counts

length(unique(dfor$RCRDNUM))

## [1] 8920

length(unique(dfor$PATIENT\_ID))

## [1] 8592

length(unique(dfor$SRC\_FAC\_ID))

## [1] 88

#Exclude NA's on COVCLINTRIAL  
# 48 rows with NAs for COVCLIN Trail   
  
cat("\n Missing COVCLINTRIAL \n")

##   
## Missing COVCLINTRIAL

sum(is.na(dfor$COVCLINTRIAL))

## [1] 48

# 48 rows with NAs for COVCLIN Trail   
df <- dfor %>% dplyr::filter(!is.na(COVCLINTRIAL))  
  
#Print Unique  
cat("\n Complete COVCLINTRIAL Record, Patient, Hospital Counts \n")

##   
## Complete COVCLINTRIAL Record, Patient, Hospital Counts

length(unique(df$RCRDNUM))

## [1] 8872

length(unique(df$PATIENT\_ID))

## [1] 8545

length(unique(df$SRC\_FAC\_ID))

## [1] 86

#   
# 8920-8545 #number of records with duplicate patient IDs  
# 8920-48-431 #dimensions using Adriana's counts  
# 8920-48-327 #dimensions using Laura's counts

## Identify Duplicates  
##Searches once from the top down and once from the bottom up.  
##Adds "dups" variable to the df where duplicate is TRUE   
  
df$dups <- duplicated(pull(df, PATIENT\_ID)) |  
duplicated(pull(df,PATIENT\_ID),fromLast=TRUE)  
  
  
#Extracts Duplicates where TRUE  
dup\_df <- df[df$dups == "TRUE", ]  
  
#431 duplicates identified  
dim(dup\_df)

## [1] 431 432

#Number of Distinct Patient\_IDs and RCRDNUM in the dataframe with all duplicates  
n\_distinct(dup\_df$PATIENT\_ID)

## [1] 104

n\_distinct(dup\_df$RCRDNUM)

## [1] 431

# nrow(dup\_df)  
# length(unique(dup\_df$PATIENT\_ID))  
  
#Number of rows with at least one repeating PATIENT\_ID and   
#number of unique repeating PATIENT\_IDs row(dup\_df)  
  
length(unique(dup\_df$PATIENT\_ID))

## [1] 104

dim(df[duplicated(df$PATIENT\_ID, fromLast = TRUE),])

## [1] 327 432

dim(df[duplicated(df$PATIENT\_ID),])

## [1] 327 432

# dup\_summary <- dup\_df %>%   
# group\_by(PATIENT\_ID) %>%   
# summarise(n\_records = length(PATIENT\_ID))  
#   
# n\_patients\_per\_dup\_n <- dup\_summary %>%   
# group\_by(n\_records) %>%  
# summarise(n\_patients = n(n\_records))  
 # n\_patients\_per\_dup\_n

#Arrange by these Vars  
dup\_df <-dup\_df %>% arrange(PATIENT\_ID, COVCLINTRIAL, SRC\_FAC\_ID, CASE\_ID)  
  
# Add an Accumulator Var for the times a PATIENT\_ID repeats   
dup\_df<-dup\_df %>% group\_by(PATIENT\_ID) %>% mutate(DROPS = cumcount(PATIENT\_ID))  
  
#Arrange by vars Descend by Number of Drops   
#Keep rows of 1 (Yes) for Covclintrial and randomly select   
dup\_df<-dup\_df %>% group\_by(PATIENT\_ID) %>% select(PATIENT\_ID, COVCLINTRIAL, SRC\_FAC\_ID, CASE\_ID, DROPS, everything()) %>%   
arrange(desc(DROPS))  
  
head(dup\_df)

## # A tibble: 6 x 433  
## # Groups: PATIENT\_ID [1]  
## PATIENT\_ID COVCLINTRIAL SRC\_FAC\_ID CASE\_ID DROPS RCRDNUM ANTCOGDVTPRO  
## <chr> <dbl+lbl> <dbl> <dbl> <dbl> <chr> <dbl+lbl>  
## 1 9.00064E+… 2 [2 = No/N… 4806 265 196 PT0007… 1 [1 = Yes]  
## 2 9.00064E+… 2 [2 = No/N… 4806 264 195 PT0007… 1 [1 = Yes]  
## 3 9.00064E+… 2 [2 = No/N… 4806 263 194 PT0007… 1 [1 = Yes]  
## 4 9.00064E+… 2 [2 = No/N… 4806 262 193 PT0007… 2 [2 = No]   
## 5 9.00064E+… 2 [2 = No/N… 4806 261 192 PT0007… 1 [1 = Yes]  
## 6 9.00064E+… 2 [2 = No/N… 4806 260 191 PT0007… 1 [1 = Yes]  
## # … with 426 more variables: ANTICOAGDVT <dbl+lbl>,  
## # ANTICOAGDVTDT\_PRC <dbl+lbl>, SEX <dbl>, DOB\_PRC <dbl+lbl>, HOMELESS <dbl>,  
## # PSOURCE\_01 <dbl>, PSOURCE\_02 <dbl>, PSOURCE\_03 <dbl>, PSOURCE\_04 <dbl>,  
## # PSOURCE\_05 <dbl>, PSOURCE\_06 <dbl>, PSOURCE\_07 <dbl>, PSOURCE\_09 <dbl>,  
## # RACE\_01 <dbl>, RACE\_02 <dbl>, RACE\_03 <dbl>, RACE\_04 <dbl>, RACE\_05 <dbl>,  
## # RACE\_06 <dbl>, ASIAN\_01 <dbl>, ASIAN\_02 <dbl>, ASIAN\_03 <dbl>,  
## # ASIAN\_04 <dbl>, ASIAN\_05 <dbl>, ASIAN\_06 <dbl>, ASIAN\_07 <dbl>,  
## # HAWAIIAN\_01 <dbl>, HAWAIIAN\_02 <dbl>, HAWAIIAN\_03 <dbl>, HAWAIIAN\_04 <dbl>,  
## # HISETHNI <dbl+lbl>, ETHNICYS\_01 <dbl>, ETHNICYS\_02 <dbl>,  
## # ETHNICYS\_03 <dbl>, ETHNICYS\_04 <dbl>, ARRDT\_PRC <dbl+lbl>,  
## # ADMDT\_PRC <dbl+lbl>, DISDATE\_PRC <dbl+lbl>, DNRPTNDECL <dbl+lbl>,  
## # DNRORDRDT\_PRC <dbl+lbl>, DSCHSTAT <dbl+lbl>, DSCHOTHFAC <dbl+lbl>,  
## # CMFTMEASNLY <dbl+lbl>, CMFTMEASNLYDT\_PRC <dbl+lbl>, MEANSTRANS <dbl+lbl>,  
## # MEDHISTO\_01 <dbl>, MEDHISTO\_02 <dbl>, MEDHISTO\_03 <dbl>, MEDHISTO\_04 <dbl>,  
## # MEDHISTO\_05 <dbl>, MEDHISTO\_06 <dbl>, MEDHISTO\_07 <dbl>, MEDHISTO\_08 <dbl>,  
## # MEDHISTO\_09 <dbl>, MEDHISTO\_10 <dbl>, MEDHISTO\_11 <dbl>, MEDHISTO\_12 <dbl>,  
## # MEDHISTO\_13 <dbl>, MEDHISTO\_14 <dbl>, MEDHISTO\_15 <dbl>, MEDHISTO\_16 <dbl>,  
## # MEDHISTO\_17 <dbl>, MEDHISTO\_18 <dbl>, MEDHISTO\_19 <dbl>, MEDHISTO\_20 <dbl>,  
## # MEDHISTO\_21 <dbl>, MEDHISTO\_22 <dbl>, MEDHISTO\_23 <dbl>, MEDHISTO\_24 <dbl>,  
## # MEDHISTO\_25 <dbl>, MEDHISTO\_26 <dbl>, MEDHISTO\_27 <dbl>, MEDHISTO\_28 <dbl>,  
## # MEDHISTO\_29 <dbl>, MEDHISTO\_30 <dbl>, MEDHISTO\_31 <dbl>, MEDHISTO\_32 <dbl>,  
## # MEDHISTO\_33 <dbl>, MEDHISTO\_34 <dbl>, MEDHISTO\_35 <dbl>, COVDIAG <dbl>,  
## # METHDIAG <dbl+lbl>, DIAGDT\_PRC <dbl+lbl>, SYMONSTDT\_PRC <dbl+lbl>,  
## # DOCUSYMP\_01 <dbl>, DOCUSYMP\_02 <dbl>, DOCUSYMP\_03 <dbl>, DOCUSYMP\_04 <dbl>,  
## # DOCUSYMP\_05 <dbl>, DOCUSYMP\_06 <dbl>, DOCUSYMP\_07 <dbl>, DOCUSYMP\_08 <dbl>,  
## # DOCUSYMP\_09 <dbl>, DOCUSYMP\_14 <dbl>, DOCUSYMP\_11 <dbl>, DOCUSYMP\_12 <dbl>,  
## # DOCUSYMP\_13 <dbl>, SPEOTHSYMP <chr>, INTERINFIL <dbl+lbl>,  
## # ANTIHYPRTNSV <dbl+lbl>, …

#Selecting those with more than 1 repeat or "drop"  
dupdff<-dup\_df %>% filter(DROPS > 1)  
  
dim(dupdff)

## [1] 327 433

head(dupdff)

## # A tibble: 6 x 433  
## # Groups: PATIENT\_ID [1]  
## PATIENT\_ID COVCLINTRIAL SRC\_FAC\_ID CASE\_ID DROPS RCRDNUM ANTCOGDVTPRO  
## <chr> <dbl+lbl> <dbl> <dbl> <dbl> <chr> <dbl+lbl>  
## 1 9.00064E+… 2 [2 = No/N… 4806 265 196 PT0007… 1 [1 = Yes]  
## 2 9.00064E+… 2 [2 = No/N… 4806 264 195 PT0007… 1 [1 = Yes]  
## 3 9.00064E+… 2 [2 = No/N… 4806 263 194 PT0007… 1 [1 = Yes]  
## 4 9.00064E+… 2 [2 = No/N… 4806 262 193 PT0007… 2 [2 = No]   
## 5 9.00064E+… 2 [2 = No/N… 4806 261 192 PT0007… 1 [1 = Yes]  
## 6 9.00064E+… 2 [2 = No/N… 4806 260 191 PT0007… 1 [1 = Yes]  
## # … with 426 more variables: ANTICOAGDVT <dbl+lbl>,  
## # ANTICOAGDVTDT\_PRC <dbl+lbl>, SEX <dbl>, DOB\_PRC <dbl+lbl>, HOMELESS <dbl>,  
## # PSOURCE\_01 <dbl>, PSOURCE\_02 <dbl>, PSOURCE\_03 <dbl>, PSOURCE\_04 <dbl>,  
## # PSOURCE\_05 <dbl>, PSOURCE\_06 <dbl>, PSOURCE\_07 <dbl>, PSOURCE\_09 <dbl>,  
## # RACE\_01 <dbl>, RACE\_02 <dbl>, RACE\_03 <dbl>, RACE\_04 <dbl>, RACE\_05 <dbl>,  
## # RACE\_06 <dbl>, ASIAN\_01 <dbl>, ASIAN\_02 <dbl>, ASIAN\_03 <dbl>,  
## # ASIAN\_04 <dbl>, ASIAN\_05 <dbl>, ASIAN\_06 <dbl>, ASIAN\_07 <dbl>,  
## # HAWAIIAN\_01 <dbl>, HAWAIIAN\_02 <dbl>, HAWAIIAN\_03 <dbl>, HAWAIIAN\_04 <dbl>,  
## # HISETHNI <dbl+lbl>, ETHNICYS\_01 <dbl>, ETHNICYS\_02 <dbl>,  
## # ETHNICYS\_03 <dbl>, ETHNICYS\_04 <dbl>, ARRDT\_PRC <dbl+lbl>,  
## # ADMDT\_PRC <dbl+lbl>, DISDATE\_PRC <dbl+lbl>, DNRPTNDECL <dbl+lbl>,  
## # DNRORDRDT\_PRC <dbl+lbl>, DSCHSTAT <dbl+lbl>, DSCHOTHFAC <dbl+lbl>,  
## # CMFTMEASNLY <dbl+lbl>, CMFTMEASNLYDT\_PRC <dbl+lbl>, MEANSTRANS <dbl+lbl>,  
## # MEDHISTO\_01 <dbl>, MEDHISTO\_02 <dbl>, MEDHISTO\_03 <dbl>, MEDHISTO\_04 <dbl>,  
## # MEDHISTO\_05 <dbl>, MEDHISTO\_06 <dbl>, MEDHISTO\_07 <dbl>, MEDHISTO\_08 <dbl>,  
## # MEDHISTO\_09 <dbl>, MEDHISTO\_10 <dbl>, MEDHISTO\_11 <dbl>, MEDHISTO\_12 <dbl>,  
## # MEDHISTO\_13 <dbl>, MEDHISTO\_14 <dbl>, MEDHISTO\_15 <dbl>, MEDHISTO\_16 <dbl>,  
## # MEDHISTO\_17 <dbl>, MEDHISTO\_18 <dbl>, MEDHISTO\_19 <dbl>, MEDHISTO\_20 <dbl>,  
## # MEDHISTO\_21 <dbl>, MEDHISTO\_22 <dbl>, MEDHISTO\_23 <dbl>, MEDHISTO\_24 <dbl>,  
## # MEDHISTO\_25 <dbl>, MEDHISTO\_26 <dbl>, MEDHISTO\_27 <dbl>, MEDHISTO\_28 <dbl>,  
## # MEDHISTO\_29 <dbl>, MEDHISTO\_30 <dbl>, MEDHISTO\_31 <dbl>, MEDHISTO\_32 <dbl>,  
## # MEDHISTO\_33 <dbl>, MEDHISTO\_34 <dbl>, MEDHISTO\_35 <dbl>, COVDIAG <dbl>,  
## # METHDIAG <dbl+lbl>, DIAGDT\_PRC <dbl+lbl>, SYMONSTDT\_PRC <dbl+lbl>,  
## # DOCUSYMP\_01 <dbl>, DOCUSYMP\_02 <dbl>, DOCUSYMP\_03 <dbl>, DOCUSYMP\_04 <dbl>,  
## # DOCUSYMP\_05 <dbl>, DOCUSYMP\_06 <dbl>, DOCUSYMP\_07 <dbl>, DOCUSYMP\_08 <dbl>,  
## # DOCUSYMP\_09 <dbl>, DOCUSYMP\_14 <dbl>, DOCUSYMP\_11 <dbl>, DOCUSYMP\_12 <dbl>,  
## # DOCUSYMP\_13 <dbl>, SPEOTHSYMP <chr>, INTERINFIL <dbl+lbl>,  
## # ANTIHYPRTNSV <dbl+lbl>, …

#using RCRDNUM  
drops <- dupdff["RCRDNUM"]  
head(drops)

## # A tibble: 6 x 1  
## RCRDNUM   
## <chr>   
## 1 PT0007646  
## 2 PT0007645  
## 3 PT0007644  
## 4 PT0007643  
## 5 PT0007642  
## 6 PT0007641

#Using Patient ID drops  
#drops<-dupdff["PATIENT\_ID"]  
#   
# head(drops)  
# nrow(drops)  
  
#Make drops into a vector   
vecdrops <- as.vector(drops)

#Create Not In Operator  
`%notin%` <- Negate(`%in%`)  
  
#Using RCRDNUM  
dim(df)

## [1] 8872 432

ef2 <- df[df$RCRDNUM %notin% vecdrops$RCRDNUM, , drop = FALSE]  
dim(ef2)

## [1] 8545 432

# #Filtering DF, leaving only PATIENT\_IDs that are not in vecdrops(list of duplicates)  
# dim(df)  
# ef <- df[df$PATIENT\_ID %notin% vecdrops$PATIENT\_ID, , drop = FALSE]  
# dim(ef)  
  
#Dropping the 195 repeats of patient with 196 repeats  
#d2 <-ef[!(ef$PATIENT\_ID=="9.00064E+15004806"),]  
  
##Check  
length(unique(ef2$RCRDNUM))

## [1] 8545

length(unique(ef2$PATIENT\_ID))

## [1] 8545

length(unique(ef2$SRC\_FAC\_ID))

## [1] 86

##Select Working Variables and Add Labels

#Create age\_group variable  
  
dft1 <- dft1 %>% mutate(age\_group = case\_when(AGEi > 64 ~ '>64',  
 AGEi >= 50 & AGEi <= 64 ~ '50-64',  
 AGEi < 50 ~ '<50'))  
#Create psource\_grouped  
  
dft1 <- dft1 %>% mutate(psource\_group = case\_when(PSOURCE\_01 == 1 | PSOURCE\_02 == 1 |PSOURCE\_06 ==1 ~ "Medicaid",  
 PSOURCE\_05 == 1 ~ "Medicare",  
 PSOURCE\_03 ==1 | PSOURCE\_07 == 1 ~ "Other",  
 PSOURCE\_04 == 1 ~ "Self pay",  
 is.na(PSOURCE\_04) == TRUE | PSOURCE\_09 == 1 ~ "Unknown"))  
# Create Death Vars with AHA Logic  
  
dft1 <- dft1 %>% mutate(death = ifelse(!is.na(DISDATE) == TRUE & DSCHSTAT == 6 , 1, 0))  
dft1$death <- as.factor(dft1$death)  
  
  
# Discharge Data (DISDATE) is not missing and Discharge Status is 6 or "Expired"   
  
dft2<- dft1 %>% mutate(Covid\_Sev = case\_when(DSCHSTAT == 6 ~ "Level 1",  
 CARDARR == 1 ~ "Level 2",  
 SHKMGMT\_01 == 1 ~ "Level 3",  
 SHKMGMT\_02 == 1 ~ "Level 3",  
 SHKMGMT\_04 == 1 ~ "Level 3",  
 SHKMGMT\_05 == 1 ~ "Level 3",  
 SHKMGMT\_03 == 1 ~ "Level 4",  
 HOSPVENT == 1 ~ "Level 5",  
 DSCHSTAT != 6 &  
 CARDARR != 1 &  
 SHKMGMT\_01 != 1 &  
 SHKMGMT\_02 != 1 &  
 SHKMGMT\_04 != 1 &  
 SHKMGMT\_05 != 1 &  
 SHKMGMT\_03 != 1 &  
 HOSPVENT != 1 ~ "Level 6")) ## do else, or nested else   
  
dft2$Covid\_Sev <- as.factor(dft2$Covid\_Sev)  
  
## Create New Race Ethnicity Variable: RACEgroup   
  
library(forcats)  
dft1$RACEgroup <- fct\_collapse(dft1$RACEi,  
 UTD = "UTD",  
 Other = "Native American",  
 NHWhite = "Non-Hispanic White",  
 Asian\_PI = c("Asian", "Pacific Islander"),  
 Black = "Non-Hispanic Black",  
 Hispanic = "Hispanic"  
)  
  
  
  
# Create Length of Hospitalization Variable  
# DISDATE is already length of admission  
#Derived date variables tend to have the "Studyday" tag  
#in the description and end in DT (ex. ANCHORDT).  
#These variables are deidentified date variables from the raw regsitry data,  
#where the anchor date, or baseline (day 0) is hospital admission date.   
#All derived dates are in the context of the Admission date (0) + day.  
  
table(dft2$Covid\_Sev)

##   
## Level 1 Level 2 Level 3 Level 4 Level 5   
## 1554 33 17 331 538

#Using Arsenal  
library(arsenal)

##   
## Attaching package: 'arsenal'

## The following object is masked from 'package:scales':  
##   
## ordinal

library(magrittr)  
tmpdir <- tempdir()  
   
#Add Labels  
  
labels(dft1) <- c( AGEi = 'Age', SEX = "Sex",  
 age\_group = "Age Group", RACEi = "Race/Ethnicity",  
 HISETHNI = "Hispanic Ethnicity", psource\_group = "Payment Source",  
 MEDHISTO\_01 = "Atrial Fibrilation",  
 MEDHISTO\_02 = "Atrial Flutter",  
 MEDHISTO\_08 = "Diabetes Mellitus",  
 MEDHISTO\_09 = "Dyslipidemia",  
 MEDHISTO\_11 = "Heart Failure",  
 MEDHISTO\_12 = "Hypertension",  
 ANTIHYPRTNSV = "Antihypertensive",  
 LIPLOWTHRP = "Lipid-lowering Therapy",  
 ANTIPLT = "Antiplatelet",  
 ANTICOAG = "Anticoagulant",  
 ANTIHYPRGLYM = "Anti-Hyperglycemic",   
 PATMANICU = "Managed in ICU",  
 death = "In-Hospital Death")  
  
tabla <- select(dft1, -c(CASE\_ID, PATIENT\_ID, SRC\_FAC\_ID, TROPADM, - TROPUADM, HGBADM, WBCADM, PLATELET, INITSCR, TROPADM,DDMERU, DDMER, SCRUADM,WBCUADM, PSOURCE\_01, PSOURCE\_02,PSOURCE\_03,PSOURCE\_04,PSOURCE\_05,PSOURCE\_07, PSOURCE\_09,  
 DISDATE,DSCHSTAT, ADMDT,DEATHDT,PSOURCE\_01,  
 PSOURCE\_02,PSOURCE\_03,PSOURCE\_04,  
 PSOURCE\_05, PSOURCE\_06, PSOURCE\_07,  
 PSOURCE\_09, CARDARR, SHKMGMT\_01,  
 SHKMGMT\_02, SHKMGMT\_04, SHKMGMT\_05, SHKMGMT\_03,   
 HOSPVENT, SCRUADM,TROPUADM, DDMERU, HGBUADM,  
 WBCUADM, RACEgroup, PATMANICUDT))  
  
table\_1 <-tableby(COVCLINTRIAL ~ ., data = tabla)  
  
latabla<-summary(table\_1, title = "Table 1", pfootnote = TRUE, digits=3, digits.test=2, nsmall.pct=1)

## Warning: Using 'digits.test = ' is deprecated. Use 'digits.p = ' instead.

## Warning: Using 'nsmall.pct = ' is deprecated. Use 'digits.pct = ' instead.

write2word(  
 latabla, paste0(tmpdir, "/test.tableby.doc"), quiet = TRUE,  
 title = "My table 1", # passed to summary.tableby  
 total = FALSE # passed to summary.tableby  
)