Untitled

Adriana RM

10/2/2020

{r setup, include=FALSE} knitr::opts\_chunk$set(echo = TRUE)

#Load Packages

library(haven)  
library(dplyr)  
library(reshape2)  
library(ggplot2)  
library(esquisse)  
library(haven)  
library(stringr)  
library(broom)  
library(SmartEDA)  
library(summarytools)  
library(stats)  
library(kableExtra)  
library(DataExplorer)  
library(scales)  
library(RColorBrewer)  
library(ecm)

## Loading the Data

```{r Read Data, message=FALSE, warning=FALSE, include=FALSE}

#Read in data and save it in df object (original data for comparison) dataFile = ‘/mnt/workspace/GWTG/COVID19/data/covid19\_cvd\_aug20.sas7bdat’ formatsFile = ‘/mnt/workspace/GWTG/COVID19/data/R\_Python\_windows\_formats.sas7bcat’

## Assigning the dataframe to the dfor object

dfor <- haven::read\_sas(dataFile, formatsFile)

```

#Print Dimensions  
dim(dfor)  
  
#Print Unique Record ID  
cat("\n Original: Record, Patient, Hospital Counts \n")  
  
length(unique(dfor$RCRDNUM))   
length(unique(dfor$PATIENT\_ID))  
length(unique(dfor$SRC\_FAC\_ID))

#Exclude NA's on COVCLINTRIAL  
# 48 rows with NAs for COVCLIN Trail   
  
cat("\n Missing COVCLINTRIAL \n")  
sum(is.na(dfor$COVCLINTRIAL))  
  
# 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")  
length(unique(df$RCRDNUM))   
length(unique(df$PATIENT\_ID))  
length(unique(df$SRC\_FAC\_ID))

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

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

#Selecting those with more than 1 repeat or "drop"  
dupdff<-dup\_df %>% filter(DROPS > 1)  
  
dim(dupdff)  
head(dupdff)  
  
#using RCRDNUM  
drops <- dupdff["RCRDNUM"]  
head(drops)  
  
#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)  
ef2 <- df[df$RCRDNUM %notin% vecdrops$RCRDNUM, , drop = FALSE]  
dim(ef2)  
  
# #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$RCRDNUM))  
length(unique(ef2$PATIENT\_ID))  
length(unique(ef2$SRC\_FAC\_ID))

##Select Working Variables and Add Labels

```{r Select Working Variables and Add Labels, include=FALSE} #Select Working Variables and Exclude NAs in COVCLINTRIAL

library(dplyr)

dft1 <- ef2 %>% dplyr::select(CASE\_ID,PATIENT\_ID, SEX,AGEi,RACEi, HGBADM, WBCADM, PLATELET, INITSCR, TROPADM,ANTIHYPRTNSV, LIPLOWTHRP,ANTIPLT, ANTICOAG, ANTIHYPRGLYM, HISETHNI,DDMER, TROPADM, INITSCR, COVCLINTRIAL, MEDHISTO\_01, MEDHISTO\_02, MEDHISTO\_08, MEDHISTO\_09,MEDHISTO\_11, MEDHISTO\_12, PATMANICUDT,PATMANICU, SRC\_FAC\_ID, 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) %>% ## CONTINUE ADD SEVERETIY VARS dplyr::filter(!is.na(COVCLINTRIAL))

#Adding Labels for Factor Vars

#Sex

dft1SEX, levels = c(“1”, “2”), labels = c(“Male”, “Female”))

#Covclintrial

dft1COVCLINTRIAL, levels = c(“1”, “2”), labels =c( “Yes”, “No”))

#RACEi is AHA’s defition

dft1RACEi, levels = c(“1”, “2”, “3”, “4”, “5”, “6”, “7”), labels = c(“Hispanic”, “Non-Hispanic Black”, “Native American”, “Asian”, “Pacific Islander”, “Non-Hispanic White”, “UTD”)) #Hispanic Ethnicity dft1HISETHNI, levels = c(“1”, “2”), labels = c(“Yes”, “No/UTD”))

#Atrial Fibrilation dft1MEDHISTO\_01, levels = c(“0”, “1”), labels = c(“No”, “Yes”))

#Atrial Flutter dft1MEDHISTO\_02, levels = c(“0”, “1”), labels = c(“No”, “Yes”))

#Diabetes Mellitus  
dft1MEDHISTO\_08, levels = c(“0”, “1”), labels = c(“No”, “Yes”))

#Dyslipidemia dft1MEDHISTO\_09, levels = c(“0”, “1”), labels = c(“No”, “Yes”)) #Heart Failure dft1MEDHISTO\_11, levels = c(“0”, “1”), labels = c(“No”, “Yes”)) #Hypertension dft1MEDHISTO\_12, levels = c(“0”, “1”), labels = c(“No”, “Yes”))

## Medications Prior To Admission

dft1ANTIHYPRTNSV, levels = c(“1”, “2”), labels = c(“Yes”, “No”))

dft1LIPLOWTHRP, levels = c(“1”, “2”), labels = c(“Yes”, “No/ND”))

dft1ANTIPLT, levels = c(“1”, “2”), labels = c(“Yes”, “No/ND”))

dft1ANTICOAG, levels = c(“1”, “2”), labels = c(“Yes”, “No/ND”))

dft1ANTIHYPRGLYM, levels = c(“1”, “2”), labels = c(“Yes”, “No/ND”))

dft1PATMANICU)

## Payment Sources

dft1PSOURCE\_01) dft1PSOURCE\_02) dft1PSOURCE\_03) dft1PSOURCE\_04) dft1PSOURCE\_05) dft1PSOURCE\_06) dft1PSOURCE\_07) dft1PSOURCE\_09)

```

```{r Recode and Combine Variables}

#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)) dft1death)

# 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

dft2Covid\_Sev)

## Create New Race Ethnicity Variable: RACEgroup

library(forcats) dft1RACEi, 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)

```

#Using Arsenal  
library(arsenal)  
library(magrittr)  
tmpdir <- tempdir()  
   
  
#Add Labels  
  
dft1$psour  
  
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, t1, 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)  
  
  
write2word(  
 latabla, paste0(tmpdir, "/test.tableby.doc"), quiet = TRUE,  
 title = "My table 1", # passed to summary.tableby  
 total = FALSE # passed to summary.tableby  
)