Data Preprocessing for the VQI TEVAR Dataset

Jennifer Ci, Thu Vu, Lily Hanyi Wang

merge the datasets

Compare the data from July 2021 and September 2021. Keep the most updated ones.

There are 6 rows from LTF July data missing in September dataset. And also 6 missing from the PROC dataset.

Only received info of PROC dataset variables? No info about LTF variables?

Variables that exists in both LTF and PROC datasets are: PATIENTID, PRIMPROCID, DEAD, PROC_SURVIVALDAYS, IDE_OTHER. Merge by these variables.

3327 participants have more than one records in this dataset.

data cleaning based on inclusion, exclusion criteria

Exclusion criteria:

- `PRESENTATION` exclude rupture patients

- `PATHOLOGY` exclude groups with pathology: 4=trauma, 8 = Aortic Thrombus,9 = Other (Retired) (retired since 09/30/2014),10 = Aorto-esophageal Fistula (Retired) (retired since 09/30/2014),11 = Aorto-bronchial Fistula (Retired) (retired since 09/30/2014)

- `URGENCY`: exclude rupture. (elective is same to asymptomatic)

- `PROXZONE_DISEASE`: exclude 0 and 1

- `DISTZONE_DISEASE`: exclude 0

- `PROXZONE_DISEASE < DISTZONE_DISEASE`: disease starting point should be earlier than ending point.

Wrong data points with distal zone < proximal zone:

```
##    PATIENTID PROXZONE_DISEASE DISTZONE_DISEASE
## 1      85423                5                4
## 2     184672                3                0
## 3     188991                8                7
## 4     192399                3                0
## 5     198402                4                3
## 6     199759                4                0
## 7     251502                2                1
## 8     287173                4                1
## 9     288150                5                2
## 10    288150                5                2
## 11    290463                5                3
## 12    307868                3                2
## 13    347630                4                3
## 14    349408                9                8
## 15    375257                8                6
```

```
## 16    376442           8                6
## 17    405193           5                4
## 18    420868           5                0
## 19    443562           4                3
## 20    445629           3                2
## 21    455349           8                4
## 22    599099           2                1
## 23    603976           4                2
## 24    603976           4                2
## 25    640888           4                3
## 26    666627           3                0
## 27    666627           3                0
## 28    689701           2                1
## 29    714961           3                1
## 30    760981           4                3
## 31    793532           4                2
## 32    793634           5                2
## 33    793704           5                4
## 34    794699           5                4
## 35    801346           5                4
```

After exclsion, there are in total 15424 objectives in the final `TEVAR` dataset. 1995 participants have more than one records in this dataset.

Data cleaning: comorbidities changed to No/Yes scale.

What is the point of filtering `BRANCH_POST`. *Filter `BRANCH_POST` based on the same criteria?*

Patient demographic and co-morbidities

`R_CR_PRESENT`: a lot of missing values, probably because Retired December 2014.

strange unit (normally 0.7 to 1.3 mg/dL)

```
## R_CR_PRESENT
##  1  2
## 10 15
```

Operative Variables

7 variables related to details about `PRIOR_AORSURG`. *include?*

`GENHIST`, `R_GENHIST`: *merge and change to 0/1 scale?*

`DEV1_PROXZONE`: *exclude align with PROXZONE?*

`DEV1_GRAFTDET`: *shouldn't be included in operative variables?* This field is required if Number of Aortic Devices has one of [1, 2, 3, 4, 5, 6]. To protect the identity of the device a random number is assigned.

many variables related to details about `PATHOLOGY`. *include?*

deal with 'conditional' variables

Outcomes

long term outcomes vs procedure outcomes?

Primary (Long term) outcomes: `DEAD` and `PROC_SURVIVALDAYS`.

Cox model for estimation of hazard ratios.

Secondary (Procedure) outcomes: others

Volume Variables

Volume Variables: `REGIONID`, `CENTERID`, `PHYSICIANID`

how to do cluster on centers and physicians? Dr. Dansey's instructions

Report mean or median: based on outliners?

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE,message = FALSE,warning = FALSE)
library(tidyverse)
library(table1)
library(Hmisc)
library(ggplot2)
## ------------- working directories for Lily ----------
wd_lily = '/Users/hanyiwang/Desktop/Comparative-analysis-of-treatments-of-CAA'
path_lily = c(
  "../data/TEVAR_International_20210712/TEVAR_International_LTF_r12_2_14_20210701.csv",
  "../data/TEVAR_International_20210712/TEVAR_International_PROC_r12_2_14_20210701.csv",
  "../data/TEVAR_International_20210901/TEVAR_International_LTF_r12_2_14_20210901.csv",
  "../data/TEVAR_International_20210901/TEVAR_International_PROC_r12_2_14_20210901.csv",
  "../data/FBVAR.csv")



## ------------- working directories for Jenn ----------
#wd_jenn = '/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA'
#path_jenn = c(
#  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210712/TE
#  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210712/TE
#  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210901/TE
#  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210901/TE


## ------------- read data ----------
setwd(wd_lily)
TEVAR_LTF_07 = read.csv(path_lily[1])
TEVAR_PROC_07 = read.csv(path_lily[2])
TEVAR_LTF_09 = read.csv(path_lily[3])
TEVAR_PROC_09 = read.csv(path_lily[4])

#setwd(wd_jenn)
#TEVAR_LTF_07 = read.csv(path_jenn[1])
#TEVAR_PROC_07 = read.csv(path_jenn[2])
#TEVAR_LTF_09 = read.csv(path_jenn[3])
#TEVAR_PROC_09 = read.csv(path_jenn[4])
## ------------- merge July and September data ----------

# find data in LTF July data but not in LTF September data by `PATIENTID`
# add these data points to the September data
TEVAR_LTF <- rbind(TEVAR_LTF_07[! TEVAR_LTF_07$PATIENTID  %in% TEVAR_LTF_09$PATIENTID,],
                   TEVAR_LTF_09)

# Similar for PROC data
TEVAR_PROC <-rbind(TEVAR_PROC_07[! TEVAR_PROC_07$PATIENTID  %in% TEVAR_PROC_09$PATIENTID,],
                   TEVAR_PROC_09)

## ------------- merge LTF and PROC data----------
# same variables in LTF and PROC data
#colnames(TEVAR_PROC)[colnames(TEVAR_PROC) %in% colnames(TEVAR_LTF)]

TEVAR <- merge(TEVAR_LTF,TEVAR_PROC, all = TRUE,
               by=c("PATIENTID","PRIMPROCID","DEAD","PROC_SURVIVALDAYS","IDE_OTHER"))
```

```r
# diagnosing duplicate id, decide which record to keep?
n_occur <- data.frame(table(TEVAR$PATIENTID))
#n_occur[n_occur$Freq > 1,]%>%n_distinct

TEVAR %>% filter(DISTZONE_DISEASE < PROXZONE_DISEASE) %>%
  select(PATIENTID,PROXZONE_DISEASE,DISTZONE_DISEASE)

## ------------- inclusion and exclusion----------
TEVAR = TEVAR %>%
  filter(PRESENTATION !=2) %>%
  filter(PATHOLOGY %in% c(1,2,3,5,6,7)) %>%
  filter(URGENCY %in% c(1,2,3)) %>%
  filter(PROXZONE_DISEASE %in% c(2,3,4,5,6,7,8,9)) %>%
  filter(DISTZONE_DISEASE > PROXZONE_DISEASE)

n_occur <- data.frame(table(TEVAR$PATIENTID))
## ------------- variables cleaning, patient demographic and co-morbidities----------
TEVAR = TEVAR %>% mutate(
  #SURGYEAR=factor(SURGYEAR)
  PRESENTATION = factor(PRESENTATION,levels = c(0,1),labels = c('Asymptomatic','Symptomatic')),
  extent = ifelse((PROXZONE_DISEASE %in% c(2,3,4,5)) & DISTZONE_DISEASE<6, 1,
              ifelse((PROXZONE_DISEASE %in% c(2,3)) & DISTZONE_DISEASE>=8, 2,
                  ifelse((PROXZONE_DISEASE %in% c(4,5)) & DISTZONE_DISEASE>8, 3,
                      ifelse((PROXZONE_DISEASE %in% c(6,7,8,3)),4,
                          ifelse((PROXZONE_DISEASE %in% c(4,5)) & DISTZONE_DISEASE<9, 5,
                              ifelse( DISTZONE_DISEASE==8, 6,
                                  ifelse( DISTZONE_DISEASE==9, 7,NA))))))),
  extent = factor(extent))
TEVAR = TEVAR %>%
  mutate(DEAD=factor(DEAD)) %>%
  mutate(PRESENTATION = factor(PRESENTATION,levels = c(0,1),
                          labels = c('Asymptomatic','Symptomatic'))) %>%
  mutate(AGECAT = factor(AGECAT,levels = c(1,2,3,4,5,6,7),
                     labels = c('<40','40-49','50-59','60-69','70-79','80-89','>89'))) %>%
  mutate(GENDER=factor(GENDER,levels=c(1,2),
                   labels=c('male','female'))) %>%
  mutate(PROXZONE_DISEASE=factor(PROXZONE_DISEASE)) %>%
  mutate(URGENCY=factor(URGENCY,levels = c(1,2,3),labels = c('Elective','Urgent','Emergent'))) %>%
  mutate(PATHOLOGY=factor(PATHOLOGY,levels=c(1,2,3,5,6,7),
                       labels = c('Aneurysm','Dissection','Aneurysm from dissection','PAU',
                                'IMH','PAU with IMH'))) %>%
  mutate(R_PREOP_AMBUL = factor(R_PREOP_AMBUL,levels = c(1,2,3,4),
                            labels=c("Amb","Amb w/ Assistance","Wheelchair","Bedridden"))) %>%
  mutate(AGECAT = factor(AGECAT,levels = c(1,2,3,4,5,6,7),
                     labels = c('<40','40-49','50-59','60-69','70-79','80-89','>89')))  %>%
  mutate(GENDER=factor(GENDER,levels=c(1,2),labels=c('male','female'))) %>%
  mutate(ETHNICITY = factor(ETHNICITY,levels=c(0,1),
                        labels = c('None Hispanic or Latino','Hispanic or Latino'))) %>%
  mutate(RACE=factor(RACE,levels = c(5,3,2,1,4,6,7),
                 labels = c('White','Black or African American','Asian',
                        'American Indian or Alaskan Native',
                        'Native Hawaiian or other Pacific Islander','More than 1 race',
                        'Unknown/Other'))) %>%
```

```r
  mutate(TRANSFER=factor(TRANSFER,levels = c(0,1,2),labels = c('No','Hospital','Rehab Unit'))) %>%
  mutate(PRIMARYINSURER=factor(PRIMARYINSURER,levels=c(1,2,3,4,5,6),
                              labels = c('Medicare','Medicaid','Commercial', 'Military/VA',
                                        'Non US Insurance','Self Pay'))) %>%
  mutate(LIVINGSTATUS=factor(LIVINGSTATUS,levels=c(1,2,3),
                            labels=c('Home','Nursing home','Homeless'))) %>%
  mutate(PREOP_FUNCSTATUS=factor(PREOP_FUNCSTATUS,levels = c(0,1,2,3,4),
                                labels = c('Full','Light work','Self care','Assisted care',
                                          'Bed bound'))) %>%
  mutate(PRIOR_CVD = factor(PRIOR_CVD,levels =c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
  mutate(PRIOR_CAD = factor(PRIOR_CAD,levels =c(0,1,2,3,4,5),
                           labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
  mutate(PRIOR_CHF = factor(PRIOR_CHF,levels =c(0,1,2,3,4),
                           labels = c('No','Yes','Yes','Yes','Yes'))) %>%
  mutate(COPD = factor(COPD,levels = c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
  mutate(DIABETES=factor(DIABETES,levels = c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
  mutate(PREOP_DIALYSIS=factor(PREOP_DIALYSIS,levels=c(0,1,2),labels=c('No','Yes','Yes'))) %>%
  mutate(HTN=factor(HTN,levels = c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
  mutate(PREOP_SMOKING=factor(PREOP_SMOKING,levels=c(0,1,2),labels=c('No','Yes','Yes'))) %>%
  mutate(PRIOR_CABG = factor(PRIOR_CABG,levels = c(0,1,2),labels = c('No','Yes','Yes'))) %>%
  mutate(PRIOR_PCI = factor(PRIOR_PCI,levels = c(0,1,2),labels = c('No','Yes','Yes'))) %>%
  mutate(PRIOR_ANEURREP = factor(PRIOR_ANEURREP,levels =c(0,1,2,3,4,5),
                                labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
  mutate(STRESS = factor(STRESS,levels =c(0,1,2,3,4),
                        labels = c('No','Yes','Yes','Yes','Yes'))) %>%
  mutate(DC_ASA = factor(DC_ASA,levels = c(0,1,2,3),labels = c('No','Yes','No','No'))) %>%
  mutate(DC_P2Y = factor(DC_P2Y,levels =c(0,1,2,3,4,5,6,7),
                        labels = c('No','Yes','Yes','Yes','Yes','Yes','No','No'))) %>%
  mutate(DC_STATIN = factor(DC_STATIN,levels = c(0,1,2,3),labels = c('No','Yes','No','No')))

## ------------- variables cleaning, operative variables----------

TEVAR = TEVAR %>%
  mutate(PRIOR_AORSURG=factor(PRIOR_AORSURG,levels=c(0,1,2,3,4),
                             labels=c('None','Open','Endo','Both','Other'))) %>%
  mutate(PRIOR_AORSURG_OPENLOC1=factor(PRIOR_AORSURG_OPENLOC1,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_OPENLOC2=factor(PRIOR_AORSURG_OPENLOC2,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_OPENLOC3=factor(PRIOR_AORSURG_OPENLOC3,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_OPENLOC4=factor(PRIOR_AORSURG_OPENLOC4,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_ENDOLOC1=factor(PRIOR_AORSURG_ENDOLOC1,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_ENDOLOC2=factor(PRIOR_AORSURG_ENDOLOC2,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_ENDOLOC3=factor(PRIOR_AORSURG_ENDOLOC3,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PRIOR_AORSURG_ENDOLOC4=factor(PRIOR_AORSURG_ENDOLOC4,
                                      levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(PATHOLOGY=factor(PATHOLOGY,levels=c(1,2,3,5,6,7),
                         labels = c('Aneurysm','Dissection','Aneurysm from dissection','PAU',
```

```r
                                            'IMH','PAU with IMH'))) %>%
    mutate(URGENCY=factor(URGENCY,levels = c(1,2,3),labels = c('Elective','Urgent','Emergent'))) %>%
    mutate(PATHOLOGY_ANEURYSM_TYPE=
              factor(PATHOLOGY_ANEURYSM_TYPE,levels = c(1,2,3,4,5),
                    labels = c('Degenerative, fusiform','Degenerative, saccular','Anastomotic',
                            'Prior trauma','Intercostal or visceral patch'))) %>%
    mutate(PATHOLOGY_DISSECT_TYPE=
              factor(PATHOLOGY_DISSECT_TYPE,levels = c(1,2),
                    labels = c('Acute, <= 30 days','Chronic, >30 daysr'))) %>%
    mutate(GENHIST = factor(GENHIST,levels = c(0,1,2,3,4,5),
                            labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
    mutate(R_GENHIST = factor(R_GENHIST,levels = c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
    mutate(Genetic=GENHIST) %>%  ############
    mutate(PROXZONE_DISEASE=factor(PROXZONE_DISEASE)) %>%
    mutate(DISTZONE_DISEASE=
              factor(DISTZONE_DISEASE, levels = c(0:15),
                    labels = c(0,1,2,3,4,5,6,7,8,9,'10R','10L','10B','11R','11L','11B'))) %>%
    mutate(ANESTHESIA=factor(ANESTHESIA,levels = c(1,2,3),labels = c('Local','Regional','General')))%>%
    mutate(IVUSTEE=factor(IVUSTEE,levels = c(0:5),labels = c('No','IVUS','TEE','Both','No','IVUS')))%>%
    mutate(ACCESS_R=factor(ACCESS_R,levels = c(0,1,2,3,4,5),
                            labels = c('None','Percutaneous femoral','Open femoral, transverse',
                                    'Open femoral, vertical','Open femoral, failed percutanous',
                                    'Iliac, via retroperitoneal'))) %>%
    mutate(ACCESS_L=factor(ACCESS_L,levels = c(0,1,2,3,4,5),
                            labels = c('None','Percutaneous femoral','Open femoral, transverse',
                                    'Open femoral, vertical','Open femoral, failed percutanous',
                                    'Iliac, via retroperitoneal')))  %>%
    mutate(ARMNECK_ACCESS=factor(ARMNECK_ACCESS,levels =c(0,1,2,3),
                                    labels = c('No','For branch treatment',
                                            'For femoral-brachial wire','For both')))%>%
    mutate(ARMNECK_ACCESS_LOC=factor(ARMNECK_ACCESS_LOC,levels =c(1:7),
                                    labels = c('Right arm','Left arm','Right axillary',
                                            'Left axillary','Right carotid',
                                            'Left carotid','Multiple')))%>%
    mutate(AORDEV_NUM=factor(AORDEV_NUM))%>%
    mutate(AORDEV_CMOD=factor(AORDEV_CMOD,levels=c(0,1),labels=c('No','Yes'))) %>%
    mutate(STAGEDAORTRT=factor(STAGEDAORTRT,levels=c(0,1),labels=c('No','Yes'))) %>%
    mutate(DEV1_GTYPE=factor(DEV1_GTYPE,levels=c(1,2,3),
                            labels=c('Standard','Custom','Physician modified'))) %>%
    mutate(DEV2_GTYPE=factor(DEV2_GTYPE,levels=c(1,2,3),
                            labels=c('Standard','Custom','Physician modified'))) %>%
    mutate(DEV3_GTYPE=factor(DEV3_GTYPE,levels=c(1,2,3),
                            labels=c('Standard','Custom','Physician modified')))


## Dr. Dansey's second filtering
TEVAR<-TEVAR %>%
  mutate(distal_seal= DEV1_DISTZONE,prox_seal= DEV1_PROXZONE)

for (i in 1:nrow(TEVAR)){
  if ((!(is.na(TEVAR$DEV2_DISTZONE[i])) && (!is.na(TEVAR$DEV2_DISTZONE[i]>TEVAR$DEV1_DISTZONE[i]))&&
      (TEVAR$DEV2_DISTZONE[i]>TEVAR$DEV1_DISTZONE[i]))) {
    TEVAR$distal_seal[i] = TEVAR$DEV2_DISTZONE[i]}
  else if((is.na(TEVAR$DEV1_DISTZONE[i])) && (!(is.na(TEVAR$DEV2_DISTZONE[i])))){
```

```r
        TEVAR$distal_seal[i] = TEVAR$DEV2_DISTZONE[i]}
  if ((!(is.na(TEVAR$DEV3_DISTZONE[i])) && (!is.na(TEVAR$DEV3_DISTZONE[i]>TEVAR$DEV2_DISTZONE[i]))&&
      (TEVAR$DEV3_DISTZONE[i]>TEVAR$DEV2_DISTZONE[i]))) {
    TEVAR$distal_seal[i] = TEVAR$DEV3_DISTZONE[i]}
  if ((!(is.na(TEVAR$DEV4_DISTZONE[i])) && (!is.na(TEVAR$DEV4_DISTZONE[i]>TEVAR$DEV3_DISTZONE[i]))&&
      (TEVAR$DEV4_DISTZONE[i]>TEVAR$DEV3_DISTZONE[i]))) {
    TEVAR$distal_seal[i] = TEVAR$DEV4_DISTZONE[i]}
  if ((!(is.na(TEVAR$DEV5_DISTZONE[i])) && (!is.na(TEVAR$DEV5_DISTZONE[i]>TEVAR$DEV4_DISTZONE[i]))&&
      (TEVAR$DEV5_DISTZONE[i]>TEVAR$DEV4_DISTZONE[i]))) {
    TEVAR$distal_seal[i] = TEVAR$DEV5_DISTZONE[i]}
  if ((!(is.na(TEVAR$DEV6_DISTZONE[i])) && (!is.na(TEVAR$DEV6_DISTZONE[i]>TEVAR$DEV5_DISTZONE[i]))&&
      (TEVAR$DEV6_DISTZONE[i]>TEVAR$DEV5_DISTZONE[i]))) {
    TEVAR$distal_seal[i] = TEVAR$DEV6_DISTZONE[i]}
  }

for (i in 1:nrow(TEVAR)){
  if ((!(is.na(TEVAR$DEV2_PROXZONE[i])) && (!is.na(TEVAR$DEV2_PROXZONE[i]<TEVAR$DEV1_PROXZONE[i]))&&
      (TEVAR$DEV2_PROXZONE[i]<TEVAR$DEV1_PROXZONE[i]))) {
    TEVAR$prox_seal[i] = TEVAR$DEV2_PROXZONE[i]}
  else if((is.na(TEVAR$DEV1_PROXZONE[i])) && (!(is.na(TEVAR$DEV2_PROXZONE[i])))){
      TEVAR$prox_seal[i] = TEVAR$DEV2_PROXZONE[i]}
  if ((!(is.na(TEVAR$DEV3_PROXZONE[i])) && (!is.na(TEVAR$DEV3_PROXZONE[i]<TEVAR$DEV2_PROXZONE[i]))&&
      (TEVAR$DEV3_PROXZONE[i]<TEVAR$DEV2_PROXZONE[i]))) {
    TEVAR$prox_seal[i] = TEVAR$DEV3_PROXZONE[i]}
  if ((!(is.na(TEVAR$DEV4_PROXZONE[i])) && (!is.na(TEVAR$DEV4_PROXZONE[i]<TEVAR$DEV3_PROXZONE[i]))&&
      (TEVAR$DEV4_PROXZONE[i]<TEVAR$DEV3_PROXZONE[i]))) {
    TEVAR$prox_seal[i] = TEVAR$DEV4_PROXZONE[i]}
  if ((!(is.na(TEVAR$DEV5_PROXZONE[i])) && (!is.na(TEVAR$DEV5_PROXZONE[i]<TEVAR$DEV4_PROXZONE[i]))&&
      (TEVAR$DEV5_PROXZONE[i]<TEVAR$DEV4_PROXZONE[i]))) {
    TEVAR$prox_seal[i] = TEVAR$DEV5_PROXZONE[i]}
  if ((!(is.na(TEVAR$DEV6_PROXZONE[i])) && (!is.na(TEVAR$DEV6_PROXZONE[i]<TEVAR$DEV5_PROXZONE[i]))&&
      (TEVAR$DEV6_PROXZONE[i]<TEVAR$DEV5_PROXZONE[i]))) {
    TEVAR$prox_seal[i] = TEVAR$DEV6_PROXZONE[i]}
  }

TEVAR<-TEVAR %>%
  mutate(ILIACDEV_END_R= factor(ILIACDEV_END_R, levels = c(0,1,2,3),
                                labels = c('None','Common',
                                           'External,Intended','External, Unintended'))) %>%
  mutate(ILIACDEV_END_L= factor(ILIACDEV_END_L, levels = c(0,1,2,3),
                                labels = c('None','Common',
                                           'External,Intended','External, Unintended'))) %>%
  mutate(BRANCH_STAGED=factor(BRANCH_STAGED,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_LSUB=factor(BRANCH_LSUB,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_CELIAC=factor(BRANCH_CELIAC,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_SMA=factor(BRANCH_SMA,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_RRENAL=factor(BRANCH_RRENAL,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_LRENAL=factor(BRANCH_LRENAL,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(BRANCH_INNO_POST=factor(BRANCH_INNO_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_LSUB_VERTPAT=
           factor(BRANCH_LSUB_VERTPAT,levels=c(1:7),
                  labels=c('Patent bilat','Patent bilat, L dominant','Patent bilat, R dominant',
```

```r
                                        'Occluded L,  patent R','Occluded R, patent L','Occluded bilat',
                                        'Not imaged'))) %>%
    mutate(ANESTHESIA_GEN_TIMEEXT=factor(ANESTHESIA_GEN_TIMEEXT,levels=c(1,2,4,5),
                                        labels=c('In OR','<12 hrs','12-24 hrs','>24 hrs'))) %>%
    mutate(POSTOP_SPINALDRAIN=factor(POSTOP_SPINALDRAIN,levels=c(0,1,2,3),
                                    labels=c('None','Pre-op','Post-op, prophylactic',
                                            'Post-op for spinal ischemia'))) %>%
    mutate(R_CELIAC=factor(R_CELIAC,levels=c(0:9),
                          labels=c('Patent, no intervention','Chronically Occluded',
                                  'Purposely Occluded','De-branch','Stent Only','Chimney',
                                  'Fen/scallop Only','Stented-fen',
                                  'Fenestrated Stentgraft Branch (Branched TEVAR)',
                                  'Side-arm Stent-graft Branch'))) %>%

    mutate(R_GRFTCONFIG=factor(R_GRFTCONFIG,levels=c(1,2,3),
                              labels=c('Standard','Plus Bare Stent','Branched/fenestrated'))) %>%
    mutate(R_LT_RENAL=factor(R_LT_RENAL,levels=c(0:9),
                            labels=c('Patent, no intervention','Chronically Occluded',
                                    'Purposely Occluded','De-branch','Stent Only','Chimney',
                                    'Fen/scallop Only','Stented-fen',
                                    'Fenestrated Stentgraft Branch (Branched TEVAR)',
                                    'Side-arm Stent-graft Branch'))) %>%
    mutate(R_PRATTZONE=factor(R_PRATTZONE)) %>%
    mutate(R_RT_RENAL=factor(R_RT_RENAL,levels=c(0:9),
                            labels=c('Patent, no intervention','Chronically Occluded',
                                    'Purposely Occluded','De-branch','Stent Only','Chimney',
                                    'Fen/scallop Only','Stented-fen',
                                    'Fenestrated Stentgraft Branch (Branched TEVAR)',
                                    'Side-arm Stent-graft Branch'))) %>%
    mutate(R_SMA=factor(R_SMA,levels=c(0:9),
                        labels=c('Patent, no intervention','Chronically Occluded',
                                'Purposely Occluded','De-branch','Stent Only','Chimney',
                                'Fen/scallop Only','Stented-fen',
                                'Fenestrated Stentgraft Branch (Branched TEVAR)',
                                'Side-arm Stent-graft Branch'))) %>%
    mutate(R_SPINAL_DRAIN=factor(R_SPINAL_DRAIN,levels=c(0,1,2),
                                labels=c('No Drain','Pre-op Drain','Post-op Drain')))

## Dr. Dansey's first filtering
# 0"0.None" 1"1.Scallop/Fen/Branch" 2"2.Occluded/Covered" 3"3.Chimney"
TEVAR<-TEVAR %>% mutate(lrenal = ifelse(BRANCH_LRENAL_TRT %in% c(0,6,7), 0,
                                    ifelse(BRANCH_LRENAL_TRT %in% c(9,10,11,12,13,14), 1,
                                        ifelse(BRANCH_LRENAL_TRT %in% c(1,2,3,4), 2,
                                            ifelse(BRANCH_LRENAL_TRT == 8, 3,NA)))),
                        rrenal = ifelse(BRANCH_RRENAL_TRT %in% c(0,6,7), 0,
                                    ifelse(BRANCH_LRENAL_TRT %in% c(9,10,11,12,13,14), 1,
                                        ifelse(BRANCH_LRENAL_TRT %in% c(1,2,3,4), 2,
                                            ifelse(BRANCH_LRENAL_TRT == 8, 3,NA)))),
                        sma = ifelse(BRANCH_SMA_TRT %in% c(0,6,7), 0,
                                ifelse(BRANCH_SMA_TRT %in% c(9,10,11,12,13,14), 1,
                                    ifelse(BRANCH_SMA_TRT %in% c(1,2,3,4), 2,
                                        ifelse(BRANCH_SMA_TRT == 8,3,NA)))),
                        celiac = ifelse(BRANCH_CELIAC_TRT %in% c(0,6,7), 0,
```

```r
                                          ifelse(BRANCH_CELIAC_TRT %in% c(9,10,11,12,13,14), 1,
                                                 ifelse(BRANCH_CELIAC_TRT %in% c(1,2,3,4), 2,
                                                        ifelse(BRANCH_CELIAC_TRT == 8,3,NA)))),
                            lsub = ifelse(BRANCH_LSUB_TRT %in% c(0,6,7), 0,
                                          ifelse(BRANCH_LSUB_TRT %in% c(9,10,11,12,13,14), 1,
                                                 ifelse(BRANCH_LSUB_TRT %in% c(1,2,3,4), 2,
                                                        ifelse(BRANCH_LSUB_TRT == 8,3,NA)))),
                            )

# drop missing data
TEVAR = TEVAR %>% filter(lrenal %in% c(1,2,3)
                         |rrenal %in% c(1,2,3)
                         |celiac %in% c(1,2,3)
                         |lsub %in% c(1,2,3)
                         |sma %in% c(1,2,3))


## ------------- variables cleaning, outcomes----------
TEVAR = TEVAR %>%
  mutate(DEAD=factor(DEAD)) %>%
  mutate(AORDEV_TECHSUCC=factor(AORDEV_TECHSUCC,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(CONVTOOPEN=factor(CONVTOOPEN,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(R_ENDOLEAK_AT_COMPLETION=
           factor(R_ENDOLEAK_AT_COMPLETION,levels=c(0:4),
                  labels=c('No','Attachment Site (type I)','Branch (type II)',
                           'Mid Graft (type III)','Indeterminate'))) %>%
  mutate(BRANCH_LSUB_POST=factor(BRANCH_LSUB_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_CELIAC_POST=factor(BRANCH_CELIAC_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_SMA_POST=factor(BRANCH_SMA_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_RRENAL_POST=factor(BRANCH_RRENAL_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_LRENAL_POST=factor(BRANCH_LRENAL_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_RCOMILI_POST=factor(BRANCH_RCOMILI_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(BRANCH_LCOMILI_POST=factor(BRANCH_LCOMILI_POST,levels=c(1,2,3),
                                 labels=c('Patent','Stenosis/Partial Coverage > 50%','Occluded'))) %>%
  mutate(POSTOP_VASO=factor(POSTOP_VASO,levels=c(0,1,2,3),labels=c('No','Yes','Yes','Yes'))) %>%
  mutate(POSTOP_COMPLICATIONS=factor(POSTOP_COMPLICATIONS,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(R_POSTOP_HEMATOMA=factor(R_POSTOP_HEMATOMA,levels=c(0,1,2,3,4),
                                 labels=c('No','Minor','Transfusion','Thrombin injection',
                                          'Surgical Rx'))) %>%
  mutate(R_POSTOP_SITEOCC=factor(R_POSTOP_SITEOCC,levels=c(0:5),
                                 labels=c('No','Medical Rx','Interventional Rx','Surgical Rx',
                                          'Stenosis','Occlusion'))) %>%
  mutate(POSTOP_MI=factor(POSTOP_MI,levels=c(0,1,2),
                                 labels=c('No','Troponin only','EKG or clinical'))) %>%
  mutate(POSTOP_DYSRHYTHMIA=factor(POSTOP_DYSRHYTHMIA,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(POSTOP_CHF=factor(POSTOP_CHF,levels=c(0,1),labels=c('No','Yes'))) %>%
  mutate(POSTOP_RESPIRATORY=factor(POSTOP_RESPIRATORY,levels = c(0:3),
```

```r
                                               labels = c('No','Yes','Yes','Yes'))) %>%
    mutate(POSTOP_DIALYSIS=factor(POSTOP_DIALYSIS,levels = c(0:2),
                                  labels = c('No','Yes','Yes'))) %>%
    mutate(POSTOP_ARMEMBO=factor(POSTOP_ARMEMBO,levels = c(0:5),
                                 labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
    mutate(POSTOP_LEGEMBO=factor(POSTOP_LEGEMBO,levels = c(0:5),
                                 labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
    mutate(POSTOP_LEGCOMPART=factor(POSTOP_LEGCOMPART,levels = c(0:4),
                                    labels = c('No','Yes','Yes','Yes','Yes'))) %>%
    mutate(R_POSTOP_BOWELISCH=factor(R_POSTOP_BOWELISCH,levels = c(0:2),labels = c('No','Yes','Yes'))) %>%
    mutate(POSTOP_RENALISCH=factor(POSTOP_RENALISCH,levels = c(0:3),
                                   labels = c('No','Yes','Yes','Yes'))) %>%
    mutate(R_POSTOP_SSI=factor(R_POSTOP_SSI,levels = c(0:4),labels = c('No','Yes','Yes','Yes','Yes'))) %>%
    mutate(POSTOP_SPINAL_ISCHEMIA=factor(POSTOP_SPINAL_ISCHEMIA,levels = c(0:2),
                                         labels = c('No','Yes','Yes'))) %>%
    mutate(RETX=factor(RETX,levels = c(0:2),labels = c('No','Yes','Yes'))) %>%
    mutate(DC_STATUS=factor(DC_STATUS,levels = c(1:6),
                            labels = c('Home','Rehab Unit','Nursing Home',
                                       'Dead','Other Hospital','Homeless'))) %>%
    mutate(R_LE_ISCH=factor(R_LE_ISCH,levels = c(0,1),labels = c('No','Yes'))) %>%
    mutate(R_POSTOP_RENAL=factor(R_POSTOP_RENAL,levels = c(0:5),
                                 labels = c('No','Yes','Yes','Yes','Yes','Yes'))) %>%
    mutate(R_RTOR=factor(R_RTOR,levels = c(0:5),labels = c('No','Yes','Yes','Yes','Yes','Yes')))



## ------------- variables cleaning, volume variables----------
TEVAR = TEVAR %>%
  mutate(REGIONID=factor(REGIONID)) %>%
  mutate(CENTERID=factor(CENTERID)) %>%
  mutate(PHYSICIANID=factor(PHYSICIANID))




## ------------- select only variables we need----------
FBVAR = TEVAR %>% select(
  # population of interest
  PRESENTATION,
  # Patient demographic and co-morbidities
  R_PREOP_AMBUL,AGE,AGECAT,GENDER,ETHNICITY,RACE,TRANSFER,PRIMARYINSURER,
  LIVINGSTATUS,PREOP_FUNCSTATUS,PRIOR_CVD,PRIOR_CAD,PRIOR_CHF,COPD,DIABETES,
  PREOP_DIALYSIS,HTN,PREOP_SMOKING,PRIOR_CABG,PRIOR_PCI,PRIOR_ANEURREP,STRESS,
  PREOP_CREAT,DC_ASA,DC_P2Y,DC_STATIN,R_CR_PRESENT,
  lrenal,rrenal,sma,celiac,lsub,####
  #operative variables
  PRIOR_AORSURG,PRIOR_AORSURG_OPENLOC1,PRIOR_AORSURG_OPENLOC2,PRIOR_AORSURG_OPENLOC3,
  PRIOR_AORSURG_OPENLOC4,PRIOR_AORSURG_ENDOLOC1,PRIOR_AORSURG_ENDOLOC2,
  PRIOR_AORSURG_ENDOLOC3,PRIOR_AORSURG_ENDOLOC4,PATHOLOGY,PREOP_MAXAAADIA,URGENCY,
  PATHOLOGY_ANEURYSM_TYPE,PATHOLOGY_DISSECT_TYPE,PATHOLOGY_DISSECT_ONSET_DAYS,
  Genetic,####
  PROXZONE_DISEASE,DISTZONE_DISEASE,ANESTHESIA,CONTRAST,EBL,FLUOROTIME,INTRAOP_PRBC,
  TOTALPROCTIME,IVUSTEE,ACCESS_R,ACCESS_L,LRGST_SHEATH_SIZE_R,
  LRGST_SHEATH_SIZE_L,ARMNECK_ACCESS,ARMNECK_ACCESS_LOC,AORDEV_NUM,
```

```r
AORDEV_CMOD,AORDEV_TECHSUCC,CONVTOOPEN,R_ENDOLEAK_AT_COMPLETION,
STAGEDAORTRT,DEV1_GTYPE,DEV2_GTYPE,DEV3_GTYPE,
distal_seal, prox_seal,extent,####
DEV1_GRAFTDET,ILIACDEV_END_R,ILIACDEV_END_L,BRANCH_STAGED,BRANCH_LSUB,
BRANCH_CELIAC,BRANCH_SMA,BRANCH_RRENAL,BRANCH_LRENAL,
BRANCH_INNO_POST,BRANCH_LSUB_POST,BRANCH_CELIAC_POST,BRANCH_SMA_POST,
BRANCH_RRENAL_POST,BRANCH_LRENAL_POST,BRANCH_RCOMILI_POST,BRANCH_LCOMILI_POST,
BRANCH_LSUB_VERTPAT,ANESTHESIA_GEN_TIMEEXT,ICUSTAY,POSTOP_PRBC,POSTOP_SPINALDRAIN,
POSTOP_VASO,POSTOP_HIGHCREAT,POSTOP_COMPLICATIONS,R_POSTOP_HEMATOMA,R_POSTOP_SITEOCC,
POSTOP_MI,POSTOP_DYSRHYTHMIA,POSTOP_CHF,POSTOP_CEREBROSX,POSTOP_RESPIRATORY,
POSTOP_DIALYSIS,POSTOP_ARMEMBO,POSTOP_LEGEMBO,POSTOP_LEGCOMPART,POSTOP_INTISCH,
R_POSTOP_BOWELISCH,POSTOP_RENALISCH,R_POSTOP_SSI,POSTOP_SPINAL_ISCHEMIA,RETX,DC_STATUS,
DC_ASA,DC_P2Y,DC_STATIN,R_CELIAC,R_CR_PRESENT,R_DISTATTZONE,R_GDPROXIMAL,R_GRFTCONFIG,
R_LE_ISCH,R_LT_RENAL,R_PRATTZONE,R_POSTOP_RENAL,R_RT_RENAL,R_RTOR,R_SMA,R_SPINAL_DRAIN,
# outcomes
DEAD,PROC_SURVIVALDAYS,TOTAL_LOS,POSTOP_LOS,AORDEV_TECHSUCC,CONVTOOPEN,
R_ENDOLEAK_AT_COMPLETION,BRANCH_LSUB_POST,BRANCH_CELIAC_POST,BRANCH_SMA_POST,
BRANCH_RRENAL_POST,BRANCH_LRENAL_POST,BRANCH_RCOMILI_POST,BRANCH_LCOMILI_POST,ICUSTAY,
POSTOP_PRBC,POSTOP_VASO,POSTOP_HIGHCREAT,POSTOP_COMPLICATIONS,R_POSTOP_HEMATOMA,
R_POSTOP_SITEOCC,POSTOP_MI,POSTOP_DYSRHYTHMIA,POSTOP_CHF,POSTOP_CEREBROSX,POSTOP_RESPIRATORY,
POSTOP_DIALYSIS,POSTOP_ARMEMBO,POSTOP_LEGEMBO,POSTOP_LEGCOMPART,POSTOP_INTISCH,
R_POSTOP_BOWELISCH,POSTOP_RENALISCH,R_POSTOP_SSI,POSTOP_SPINAL_ISCHEMIA,RETX,
DC_STATUS,R_LE_ISCH,R_POSTOP_RENAL,R_RTOR,
# clustering
REGIONID,CENTERID,PHYSICIANID
)

## ------------ variables labels----------
#var.labels = c(AGE="Age (years)", AGECAT="Age (category)")
#label(FBVAR) = as.list(var.labels[match(names(FBVAR), names(var.labels))])


## ------------ store as new dataset 'FBVAR' to ensure easier access for modelling----------
write.csv(FBVAR,path_lily[5])
TEVAR %>%
  filter(! is.na(TEVAR$R_CR_PRESENT)) %>%
  select(R_CR_PRESENT) %>%table()
```