preprocessing

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.

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

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

After excluding some data points, there are in total 17214 objectives in the final overall dataset.

population of interest: the asymptomatic and symptomatics groups.

	Overall
	(N=17214)
PRESENTATION	
Asymptomatic	10232 (59.4%)
Symptomatic	6982 (40.6%)

Demographic history

Under procedure tab, history and demographic variables

`R_PREOP_AMBUL`: Preop ambulatory status; 1 = Amb,2 = Amb w/ Assistance,3 = Wheelchair,4 = Bedridden

`TRANSFER`: Transferred From?; 0 = No,1 = Hospital,2 = Rehab Unit

`PRIMARYINSURER`: Primary Insurer; 1 = Medicare,2 = Medicaid,3 = Commercial,4 = Military/VA,5 = Non US Insurance,6 = Self Pay

`HTCM`: Min/max range: 137 to 203 cm.

`WTKG`: Min/max range: 18.1 to 227 kgs.

Preference, inch/cm, lb/kg?

`LIVINGSTATUS`: Living Status; 1 = Home,2 = Nursing home,3 = Homeless

`PREOP_FUNCSTATUS`: Functional Status; 0 = Full,1 = Light work,2 = Self care,3 = Assisted care,4 = Bed bound

`PREOP_DIALYSIS`: Dialysis status; 0 = No,1 = Functioning Transplant,2 = On Dialysis

	Asymptomatic	Symptomatic	Overall
	(N=10232)	(N=6982)	(N=17214)
GENDER			
male	7134 (69.7%)	4207 (60.3%)	11341 (65.9%)
female	3098 (30.3%)	2775 (39.7%)	5873 (34.1%)
ETHNICITY			
None Hispanic or Latino	9777 (95.6%)	6452 (92.4%)	16229 (94.3%)
Hispanic or Latino	440 (4.3%)	520 (7.4%)	960 (5.6%)
Missing	15 (0.1%)	10 (0.1%)	25 (0.1%)
RACE			
White	8011 (78.3%)	4241 (60.7%)	12252 (71.2%)
Black or African American	1310 (12.8%)	1945 (27.9%)	3255 (18.9%)
Asian	244 (2.4%)	180 (2.6%)	424 (2.5%)
American Indian or Alaskan Native	22 (0.2%)	17 (0.2%)	39 (0.2%)
Native Hawaiian or other Pacific Islander	20 (0.2%)	23 (0.3%)	43 (0.2%)
More than 1 race	27 (0.3%)	11 (0.2%)	38 (0.2%)
Unknown/Other	597 (5.8%)	562 (8.0%)	1159 (6.7%)
Missing	1 (0.0%)	3 (0.0%)	4 (0.0%)
HTCM			
Mean (SD)	172 (10.7)	172 (11.6)	172 (11.1)
Median [Min, Max]	173 [0, 419]	172 [0, 213]	173 [0, 419]
Missing	1 (0.0%)	29 (0.4%)	30 (0.2%)
WTKG			
Mean (SD)	83.5 (22.1)	84.9 (23.1)	84.1 (22.5)
Median [Min, Max]	81.2 [24.0, 962]	82.0 [23.0, 205]	81.7 [23.0, 962]
Missing	0 (0%)	4 (0.1%)	4 (0.0%)

patient condition variables, pathway demographic:

Prior diseases history all changed to 0/1 scale.

`PRIOR_CVD`, `PRIOR_CAD`, `PRIOR_CHF`, `COPD`, `PRIOR_CABG`, `PRIOR_PCI`, `R_PRIOR_CABGPTCA`, `PRIOR_CEACAS`, `R_PRIOR_CEA`, `PRIOR_ANEURREP`, `PRIOR_BYPASS`, `PRIOR_PVI`.

only use one variable for past heart disease? but forgot which to use

`DIABETES`,`PREOP_DIALYSIS`, `HTN`,`PREOP_SMOKING`, `STRESS`, `HEMO` (Pre op Hemoglobin: range 4-20(g/dl)),

Which to include?

`PREOP_CREAT`, `PREOP_ASA`, `PREOP_P2Y`, `PREOP_STATIN`, `PREOP_BETABLOCKER`, `PREOP_ACE`, `PREOP_ANTICOAG`,

Retired variables? Are the info transferred to new variables?

	Asymptomatic	Symptomatic	Overall
	(N=10232)	(N=6982)	(N=17214)
R_PREOP_AMBUL			
Amb	171 (1.7%)	149 (2.1%)	320 (1.9%)
Amb w/ Assistance	10 (0.1%)	3 (0.0%)	13 (0.1%)
Wheelchair	0 (0%)	0 (0%)	0 (0%)
Bedridden	0 (0%)	6 (0.1%)	6 (0.0%)
Missing	10051 (98.2%)	6824 (97.7%)	16875 (98.0%)
TRANSFER			
No	9849 (96.3%)	3117 (44.6%)	12966 (75.3%)
Hospital	360 (3.5%)	3848 (55.1%)	4208 (24.4%)
Rehab Unit	18 (0.2%)	16 (0.2%)	34 (0.2%)
Missing	5 (0.0%)	1 (0.0%)	6 (0.0%)
PRIMARYINSURER			
Medicare	5559 (54.3%)	2698 (38.6%)	8257 (48.0%)
Medicaid	411 (4.0%)	767 (11.0%)	1178 (6.8%)
Commercial	3054 (29.8%)	2520 (36.1%)	5574 (32.4%)
Military/VA	263 (2.6%)	135 (1.9%)	398 (2.3%)
Non US Insurance	406 (4.0%)	94 (1.3%)	500 (2.9%)
Self Pay	109 (1.1%)	523 (7.5%)	632 (3.7%)
Missing	430 (4.2%)	245 (3.5%)	675 (3.9%)
LIVINGSTATUS			
Home	10117 (98.9%)	6879 (98.5%)	16996 (98.7%)
Nursing home	99 (1.0%)	72 (1.0%)	171 (1.0%)
Homeless	14 (0.1%)	29 (0.4%)	43 (0.2%)
Missing	2 (0.0%)	2 (0.0%)	4 (0.0%)
PREOP_FUNCSTATUS			
Full	6619 (64.7%)	4874 (69.8%)	11493 (66.8%)
Light work	2095 (20.5%)	1157 (16.6%)	3252 (18.9%)
Self care	1255 (12.3%)	739 (10.6%)	1994 (11.6%)
Assisted care	205 (2.0%)	158 (2.3%)	363 (2.1%)
Bed bound	12 (0.1%)	23 (0.3%)	35 (0.2%)
Missing	46 (0.4%)	31 (0.4%)	77 (0.4%)
PREOP_DIALYSIS			
No	9988 (97.6%)	6699 (95.9%)	16687 (96.9%)
Yes	241 (2.4%)	282 (4.0%)	523 (3.0%)
Missing	3 (0.0%)	1 (0.0%)	4 (0.0%)
PRIOR_CVD			
No	9079 (88.7%)	6316 (90.5%)	15395 (89.4%)
Yes	1150 (11.2%)	665 (9.5%)	1815 (10.5%)
Missing	3 (0.0%)	1 (0.0%)	4 (0.0%)
DIABETES			
No	8467 (82.8%)	5949 (85.2%)	14416 (83.7%)
Yes	1765 (17.2%)	1032 (14.8%)	2797 (16.2%)
Missing	0 (0%)	1 (0.0%)	1 (0.0%)
HTN			
No	1077 (10.5%)	713 (10.2%)	1790 (10.4%)
Yes	9097 (88.9%)	6170 (88.4%)	15267 (88.7%)
Missing	58 (0.6%)	99 (1.4%)	157 (0.9%)
PREOP_SMOKING			
No	1959 (19.1%)	2261 (32.4%)	4220 (24.5%)
Yes	8272 (80.8%)	4709 (67.4%)	12981 (75.4%)
Missing	1 (0.0%)	12 (0.2%)	13 (0.1%)

	Asymptomatic	Symptomatic	Overall
factor(STRESS)			
0	6148 (60.1%)	5920 (84.8%)	12068 (70.1%)
1	3366 (32.9%)	900 (12.9%)	4266 (24.8%)
2	356 (3.5%)	80 (1.1%)	436 (2.5%)
3	247 (2.4%)	62 (0.9%)	309 (1.8%)
4	107 (1.0%)	16 (0.2%)	123 (0.7%)
Missing	8 (0.1%)	4 (0.1%)	12 (0.1%)
HEMO			
Mean (SD)	12.8 (2.26)	11.7 (2.12)	12.4 (2.27)
Median [Min, Max]	13.0 [0.700, 116]	11.8 [1.20, 19.6]	12.5 [0.700, 116]
Missing	69 (0.7%)	11 (0.2%)	80 (0.5%)
PREOP_CREAT			
Mean (SD)	1.16 (0.726)	1.21 (0.811)	1.18 (0.762)
Median [Min, Max]	1.03 [0, 32.0]	1.01 [0.290, 19.8]	1.03 [0, 32.0]
Missing	268 (2.6%)	266 (3.8%)	534 (3.1%)

patient condition variables, pathway history:

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

`PREOP_EF`: Ejection Fraction; 1 = <30%,2 = 30-50%,3 = >50%,4 = Not Done,5 = Unknown

`PREOP_MAXAAADIA`: Maximum Aortic Diameter; *include?*

`LEG_MOTOR_FUNCTION`: Leg Motor Function; 1 = Normal,2 = Mild weakness,3 = Moderate weakness,4 = Severe weakness,5 = Paralysis *include?*

`DISTZONE_DISEASE`: Distal Zone of Disease *include?*

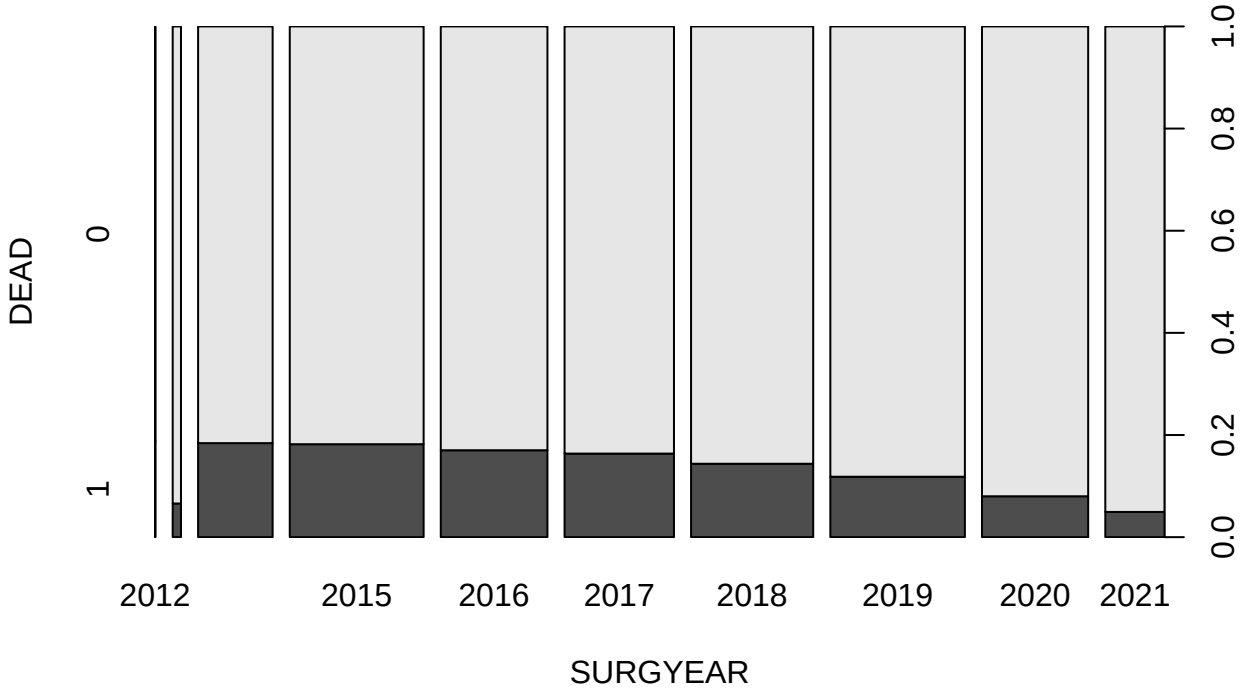
many variables related to details about `PATHOLOGY`. *include?*

	Asymptomatic	Symptomatic	Overall
	(N=10232)	(N=6982)	(N=17214)
factor(PREOP_EF)			
1	183 (1.8%)	104 (1.5%)	287 (1.7%)
2	1227 (12.0%)	536 (7.7%)	1763 (10.2%)
3	5967 (58.3%)	4125 (59.1%)	10092 (58.6%)
4	2145 (21.0%)	1600 (22.9%)	3745 (21.8%)
5	705 (6.9%)	610 (8.7%)	1315 (7.6%)
Missing	5 (0.0%)	7 (0.1%)	12 (0.1%)
PATHOLOGY			
Aneurysm	7722 (75.5%)	1707 (24.4%)	9429 (54.8%)
Dissection	1230 (12.0%)	3717 (53.2%)	4947 (28.7%)
Aneurysm from dissection	784 (7.7%)	478 (6.8%)	1262 (7.3%)
PAU	379 (3.7%)	499 (7.1%)	878 (5.1%)
IMH	58 (0.6%)	284 (4.1%)	342 (2.0%)
PAU with IMH	59 (0.6%)	297 (4.3%)	356 (2.1%)
PREOP_MAXAAADIA			
Mean (SD)	58.1 (13.3)	48.5 (16.6)	54.3 (15.4)
Median [Min, Max]	58.0 [0, 410]	45.0 [0, 160]	55.0 [0, 410]
Missing	110 (1.1%)	392 (5.6%)	502 (2.9%)
URGENCY			
Elective	9964 (97.4%)	3228 (46.2%)	13192 (76.6%)
Urgent	239 (2.3%)	2708 (38.8%)	2947 (17.1%)

	Asymptomatic	Symptomatic	Overall
Emergent	29 (0.3%)	1046 (15.0%)	1075 (6.2%)
factor(LEG_MOTOR_FUNCTION)			
1	9760 (95.4%)	6071 (87.0%)	15831 (92.0%)
2	320 (3.1%)	490 (7.0%)	810 (4.7%)
3	71 (0.7%)	154 (2.2%)	225 (1.3%)
4	16 (0.2%)	108 (1.5%)	124 (0.7%)
5	24 (0.2%)	129 (1.8%)	153 (0.9%)
Missing	41 (0.4%)	30 (0.4%)	71 (0.4%)
factor(DISTZONE_DISEASE)			
0	5 (0.0%)	0 (0%)	5 (0.0%)
1	2 (0.0%)	1 (0.0%)	3 (0.0%)
2	21 (0.2%)	26 (0.4%)	47 (0.3%)
3	222 (2.2%)	171 (2.4%)	393 (2.3%)
4	887 (8.7%)	834 (11.9%)	1721 (10.0%)
5	2129 (20.8%)	2038 (29.2%)	4167 (24.2%)
6	259 (2.5%)	312 (4.5%)	571 (3.3%)
7	171 (1.7%)	218 (3.1%)	389 (2.3%)
8	378 (3.7%)	338 (4.8%)	716 (4.2%)
9	3705 (36.2%)	1089 (15.6%)	4794 (27.8%)
10	442 (4.3%)	302 (4.3%)	744 (4.3%)
11	353 (3.5%)	362 (5.2%)	715 (4.2%)
12	1148 (11.2%)	573 (8.2%)	1721 (10.0%)
13	151 (1.5%)	233 (3.3%)	384 (2.2%)
14	130 (1.3%)	195 (2.8%)	325 (1.9%)
15	184 (1.8%)	266 (3.8%)	450 (2.6%)
Missing	45 (0.4%)	24 (0.3%)	69 (0.4%)

other variables

Surgery year would affect outcome, since surgeons got more familiar with the surgery.



Outcome variables

Primary outcomes: `DEAD` and `PROC_SURVIVALDAYS`.

Secondary outcomes: `POSTOP_LOS`

other outcomes?

	Asymptomatic	Symptomatic	Overall
	(N=10232)	(N=6982)	(N=17214)
DEAD			
0	8934 (87.3%)	5868 (84.0%)	14802 (86.0%)
1	1295 (12.7%)	1113 (15.9%)	2408 (14.0%)
Missing	3 (0.0%)	1 (0.0%)	4 (0.0%)
PROC_SURVIVALDAYS			
Mean (SD)	829 (777)	930 (865)	870 (815)
Median [Min, Max]	545 [-355, 3450]	613 [0, 3290]	571 [-355, 3450]
Missing	2 (0.0%)	0 (0%)	2 (0.0%)
POSTOP_LOS			
Mean (SD)	5.90 (23.7)	8.73 (17.8)	7.05 (21.6)
Median [Min, Max]	3.00 [0, 1100]	6.00 [0, 1100]	4.00 [0, 1100]
Missing	2 (0.0%)	0 (0%)	2 (0.0%)

Clustering variables:

19 regions, 189 centers, 1094 physicians.

Most physicians only performed 1 or 2 procedures. Several performed over 100 procedures. Since the more surgeries a surgeon did, the more familiar he or she is. So we need to cluster on this.

how to do clustering on centers and physicians

mean and median: based on outliners?

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE,message = FALSE,warning = FALSE)
library(tidyverse)
library(table1)

## ------------- 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")


## ------------- 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/TEVA
  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210712/TEVA
  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210901/TEVA
  "/Users/jenniferci/Desktop/Comparative-analysis-of-treatments-of-CAA/TEVAR_International_20210901/TEVA

## ------------- 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"))

## ------------- 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))

## ------------- data cleaning----------
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(SURGYEAR=factor(SURGYEAR)) %>%
  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(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'))) %>%
  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(PRIOR_CVD = factor(PRIOR_CVD,levels =c(0,1,2,3),labels = c('No','Yes','Yes','Yes'))) %>%
  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(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')))



## ------------- population of interest ----------
table1(~  PRESENTATION, data = TEVAR)
## ------------- table: demographic----------
table1(~ GENDER+ETHNICITY+RACE+HTCM+WTKG
         | PRESENTATION, data = TEVAR,caption = 'Table 1- demographic')
```

```
## ------------- table: patient condition (pathway demographics) ----------
table1(~ R_PREOP_AMBUL+TRANSFER+PRIMARYINSURER+LIVINGSTATUS+PREOP_FUNCSTATUS+PREOP_DIALYSIS+
         PRIOR_CVD+DIABETES+PREOP_DIALYSIS+HTN+PREOP_SMOKING+factor(STRESS)+HEMO+PREOP_CREAT
         | PRESENTATION, data = TEVAR)
## ------------- table: patient condition anatomy ----------
table1(~ factor(PREOP_EF)+PATHOLOGY+PREOP_MAXAAADIA+URGENCY+
         factor(LEG_MOTOR_FUNCTION)+factor(DISTZONE_DISEASE)
         | PRESENTATION, data = TEVAR,caption = 'Table 2- Anatomy detail ')
plot(DEAD~SURGYEAR,data=TEVAR)
## ------------- table3: outcomes----------
table1(~ DEAD+PROC_SURVIVALDAYS+POSTOP_LOS | PRESENTATION, data = TEVAR,caption='Table 3- outcomes ')

## ------------- Survival curves----------

## ------------- clustering variables----------

#TEVAR %>% select(REGIONID) %>% table()
#TEVAR %>% select(CENTERID) %>% table()
#TEVAR %>% select(PHYSICIANID) %>% table()
```