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.

Merge the long term follow-up dataset and the pre-procedure dataset by `PATIENTID`.

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)

include the indication, merge the two columns

- `URGENCY`: exclude rupture. (elective is same to asymptomatic)
- `PROXZONE_DISEASE`: exclude 0 and 1

population of interest: the asymptomatic and symptomatics groups.

	Overall
	(N=13129)
PRESENTATION	
Asymptomatic	8105 (61.7%)
Symptomatic	5024 (38.3%)

Demographic history

GENDER, BMI(USING HEIGHT AND WEIGHT)?, R_PREOP_AMBUL(Preop ambulatory status), AGE (Patient age at procedure), AGECAT, PATHOLOGY (exclude 4, 8-11)

`AGECAT`: 1 = <40,2 = 40-49,3 = 50-59,4 = 60-69,5 = 70-79,6 = 80-89,7 = >89

`GENDER`: 1 = Male,2 = Female

	Asymptomatic	Symptomatic	Overall
	(N=8105)	(N=5024)	(N=13129)
GENDER			
male	5578 (68.8%)	2931 (58.3%)	8509 (64.8%)
female	2527 (31.2%)	2093 (41.7%)	4620 (35.2%)
AGE			
Mean (SD)	71.4 (10.1)	65.3 (13.6)	69.1 (11.9)

	Asymptomatic	Symptomatic	Overall
Median [Min, Max]	72.0 [0, 90.0]	67.0 [0, 90.0]	71.0 [0, 90.0]
AGECAT			
<40	78 (1.0%)	202 (4.0%)	280 (2.1%)
40-49	180 (2.2%)	479 (9.5%)	659 (5.0%)
50-59	621 (7.7%)	904 (18.0%)	1525 (11.6%)
60-69	2144 (26.5%)	1346 (26.8%)	3490 (26.6%)
70-79	3450 (42.6%)	1320 (26.3%)	4770 (36.3%)
80-89	1536 (19.0%)	704 (14.0%)	2240 (17.1%)
>89	96 (1.2%)	69 (1.4%)	165 (1.3%)

anatomy

	Asymptomatic	Symptomatic	Overall
	(N=8105)	(N=5024)	(N=13129)
PATHOLOGY			
Aneurysm	6423 (79.2%)	1459 (29.0%)	7882 (60.0%)
Dissection	743 (9.2%)	2333 (46.4%)	3076 (23.4%)
Aneurysm from dissection	532 (6.6%)	310 (6.2%)	842 (6.4%)
PAU	307 (3.8%)	424 (8.4%)	731 (5.6%)
IMH	47 (0.6%)	242 (4.8%)	289 (2.2%)
PAU with IMH	53 (0.7%)	256 (5.1%)	309 (2.4%)
PROXZONE_DISEASE			
2	855 (10.5%)	1165 (23.2%)	2020 (15.4%)
3	1734 (21.4%)	2368 (47.1%)	4102 (31.2%)
4	1216 (15.0%)	716 (14.3%)	1932 (14.7%)
5	701 (8.6%)	301 (6.0%)	1002 (7.6%)
6	369 (4.6%)	86 (1.7%)	455 (3.5%)
7	661 (8.2%)	91 (1.8%)	752 (5.7%)
8	1681 (20.7%)	187 (3.7%)	1868 (14.2%)
9	888 (11.0%)	110 (2.2%)	998 (7.6%)

patient condition variables:

`PRIOR_CVD`: 0 = None,1 = hx stroke, asymptomatic,2 = hx stroke, minor deficit,3 = hx stroke, major deficit

`PRIOR_CAD`: 0 = None,1 = hx MI but no sx,2 = Stable angina,3 = Unstable angina or MI < 6 mos (retired since 09/12/2012),4 = MI < 6 mos,5 = Unstable angina

`PRIOR_CHF`: 0 = None,1 = Asymp, hx CHF,2 = Mild,3 = Moderate,4 = Severe

`COPD`: 0 = No,1 = Not Treated,2 = On Meds,3 = On Home Oxygen

`DIABETES`: 0 = None,1 = Diet,2 = Non-insulin Meds,3 = Insulin

`HTN`: History of hypertension; 0 = No, 1 = Yes (>=140/90 or history) (retired since 11/15/2016), 2 = Yes, controlled [added on 04/13/2020] , 3 = Yes, uncontrolled [added on 04/13/2020]

`PREOP_SMOKING`: 0 = Never,1 = Prior,2 = Current

`PRIOR_AORSURG`: Any aortic procedures performed on a separate date prior to the index procedure; 0 = None,1 = Open,2 = Endo,3 = Both,4 = Other (retired since 09/30/2014)

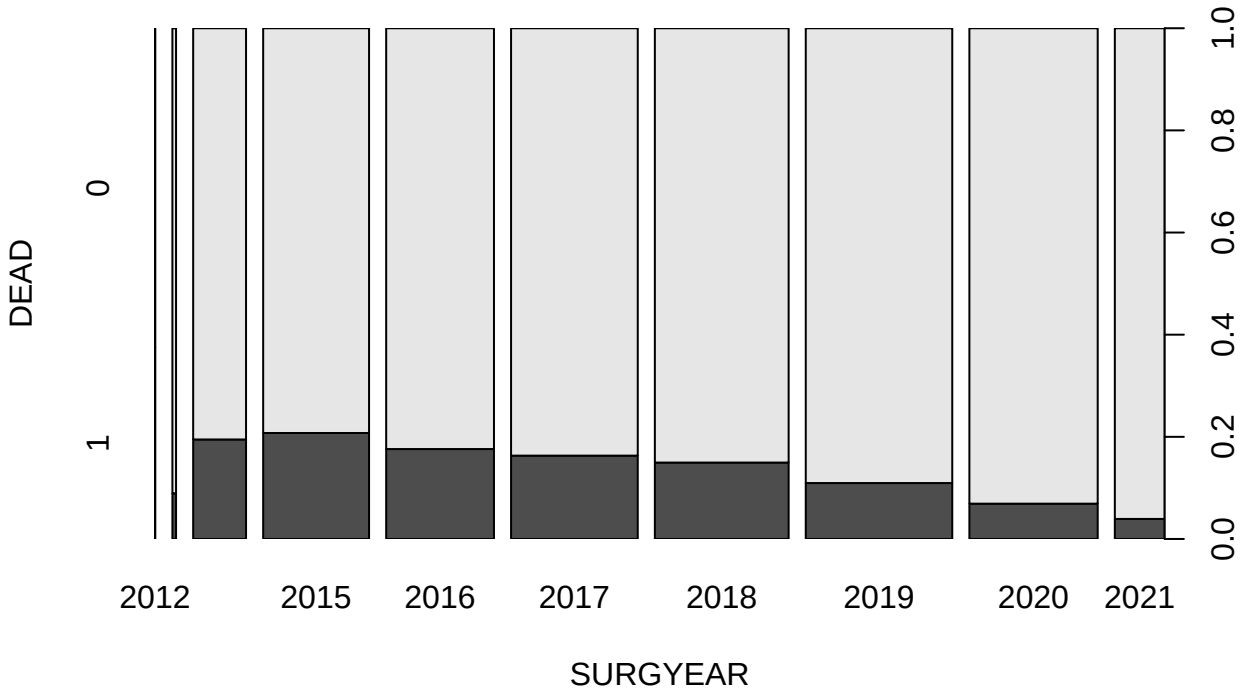
	Asymptomatic	Symptomatic	Overall

	Asymptomatic	Symptomatic	Overall
	(N=8105)	(N=5024)	(N=13129)
factor(R_PREOP_AMBUL)			
1	73 (0.9%)	49 (1.0%)	122 (0.9%)
2	5 (0.1%)	1 (0.0%)	6 (0.0%)
4	0 (0%)	1 (0.0%)	1 (0.0%)
Missing	8027 (99.0%)	4973 (99.0%)	13000 (99.0%)
factor(PREOP_SMOKING)			
0	1433 (17.7%)	1563 (31.1%)	2996 (22.8%)
1	4354 (53.7%)	1644 (32.7%)	5998 (45.7%)
2	2317 (28.6%)	1807 (36.0%)	4124 (31.4%)
Missing	1 (0.0%)	10 (0.2%)	11 (0.1%)
factor(PRIOR_CVD)			
0	7153 (88.3%)	4487 (89.3%)	11640 (88.7%)
1	592 (7.3%)	332 (6.6%)	924 (7.0%)
2	290 (3.6%)	159 (3.2%)	449 (3.4%)
3	67 (0.8%)	45 (0.9%)	112 (0.9%)
Missing	3 (0.0%)	1 (0.0%)	4 (0.0%)
factor(PRIOR_CAD)			
0	6111 (75.4%)	4158 (82.8%)	10269 (78.2%)
1	1479 (18.2%)	569 (11.3%)	2048 (15.6%)
2	409 (5.0%)	179 (3.6%)	588 (4.5%)
4	63 (0.8%)	56 (1.1%)	119 (0.9%)
5	39 (0.5%)	59 (1.2%)	98 (0.7%)
Missing	4 (0.0%)	3 (0.1%)	7 (0.1%)
factor(PRIOR_CHF)			
0	6946 (85.7%)	4433 (88.2%)	11379 (86.7%)
1	716 (8.8%)	338 (6.7%)	1054 (8.0%)
2	259 (3.2%)	139 (2.8%)	398 (3.0%)
3	155 (1.9%)	85 (1.7%)	240 (1.8%)
4	29 (0.4%)	29 (0.6%)	58 (0.4%)
factor(COPD)			
0	5265 (65.0%)	3760 (74.8%)	9025 (68.7%)
1	724 (8.9%)	374 (7.4%)	1098 (8.4%)
2	1680 (20.7%)	721 (14.4%)	2401 (18.3%)
3	435 (5.4%)	169 (3.4%)	604 (4.6%)
Missing	1 (0.0%)	0 (0%)	1 (0.0%)
factor(DIABETES)			
0	6701 (82.7%)	4232 (84.2%)	10933 (83.3%)
1	333 (4.1%)	218 (4.3%)	551 (4.2%)
2	845 (10.4%)	397 (7.9%)	1242 (9.5%)
3	226 (2.8%)	176 (3.5%)	402 (3.1%)
Missing	0 (0%)	1 (0.0%)	1 (0.0%)
factor(HTN)			
0	846 (10.4%)	507 (10.1%)	1353 (10.3%)
1	5684 (70.1%)	3470 (69.1%)	9154 (69.7%)
2	1103 (13.6%)	536 (10.7%)	1639 (12.5%)
3	443 (5.5%)	486 (9.7%)	929 (7.1%)
Missing	29 (0.4%)	25 (0.5%)	54 (0.4%)
factor(PRIOR_AORSURG)			
0	5948 (73.4%)	3965 (78.9%)	9913 (75.5%)

	Asymptomatic	Symptomatic	Overall
1	1170 (14.4%)	568 (11.3%)	1738 (13.2%)
2	796 (9.8%)	433 (8.6%)	1229 (9.4%)
3	189 (2.3%)	56 (1.1%)	245 (1.9%)
Missing	2 (0.0%)	2 (0.0%)	4 (0.0%)

other variables

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



Outcome variables

`PROC_SURVIVALDAYS`: This should be the longest known time of survival data available for the patient. Survival days are calculated as the Last Date of Contact (or Date of Death) for the patient - Procedure date for a procedure. Please refer to included Death and Survival Days Logic.pdf for additional details."

`POSTOP_LOS`: Length of Stay in days calculated by DISCHARGE_DT - SURGERY_DT

	Asymptomatic	Symptomatic	Overall
	(N=8105)	(N=5024)	(N=13129)
factor(DEAD)			
0	7092 (87.5%)	4205 (83.7%)	11297 (86.0%)
1	1012 (12.5%)	818 (16.3%)	1830 (13.9%)
Missing	1 (0.0%)	1 (0.0%)	2 (0.0%)
PROC_SURVIVALDAYS			
Mean (SD)	722 (723)	683 (744)	707 (731)
Median [Min, Max]	459 [-355, 3360]	420 [0, 3200]	446 [-355, 3360]
Missing	1 (0.0%)	0 (0%)	1 (0.0%)
POSTOP_LOS			
Mean (SD)	5.98 (25.1)	8.76 (19.7)	7.04 (23.2)
Median [Min, Max]	3.00 [0, 1100]	6.00 [0, 1100]	4.00 [0, 1100]

	Asymptomatic	Symptomatic	Overall
Missing	1 (0.0%)	0 (0%)	1 (0.0%)

Clustering variables:

19 regions, 184 centers, 948 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.

Cluster on centers and physicians

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")


## ------------- 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])
## ------------- merge the data----------

## ------------- inclusion and exclusion----------
TEVAR_PROC_07 = TEVAR_PROC_07 %>%
  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_PROC_07 = TEVAR_PROC_07 %>%
  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')))

## ------------- population of interest ----------
table1(~  PRESENTATION, data = TEVAR_PROC_07)
## ------------- table1: demographic----------
table1(~ GENDER+AGE+AGECAT
         | PRESENTATION, data = TEVAR_PROC_07)
## ------------- table: anatomy ----------
table1(~ PATHOLOGY+PROXZONE_DISEASE | PRESENTATION, data = TEVAR_PROC_07)
## ------------- table: patients conditions ----------
```

```
table1(~ factor(R_PREOP_AMBUL)+factor(PREOP_SMOKING)+factor(PRIOR_CVD)+
         factor(PRIOR_CAD)+factor(PRIOR_CHF)+factor(COPD)+factor(DIABETES)+factor(HTN)+
         factor(PRIOR_AORSURG)
       | PRESENTATION, data = TEVAR_PROC_07)


plot(DEAD~SURGYEAR,data=TEVAR_PROC_07)
## ------------- table3: outcomes----------
table1(~ factor(DEAD) + PROC_SURVIVALDAYS+POSTOP_LOS | PRESENTATION, data = TEVAR_PROC_07)

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

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

#TEVAR_PROC_07 %>% select(REGIONID) %>% table()
#TEVAR_PROC_07 %>% select(CENTERID) %>% table()
#TEVAR_PROC_07 %>% select(PHYSICIANID) %>% table()
```