Final Project for Biostat 202: Data Cleaning

Introduction

Background

Acute Renal Failure (ARF) after surgery is a severe complication and it affects patients' outcomes and hospital length. Especially, cardiovascular surgery is the second most likely to cause ARF following transplant surgery. Some biomarkers have been developed to predict ARF after surgery but have not been practical, and more easily access predict markers are desired in a clinical setting. American College of Surgeons National Surgical Quality Improvement Program® (ACS NSQIP®) data is developed by surgeons and includes patients' outcomes that are thirty days after their operations. The purpose of this study is to predict ARF after cardiac surgery by pre-operation lab tests and conditions of operation using the NSQIP data.

Research question

Do pre-operation lab tests and conditions of operation predict Acute Renal Failure after surgery among people with cardiovascular surgery?

Variables/Predictors

44 variables that are easily accessible by the time the operation is complete are included. We exclude weight, height, and days from lab tests to the operation because these variables are replaced by new variables, BMI, and the lab results within 14 days.

Subjects' background: Sex, Age, BMI, Race and ethnicity, Smoking history, Comorbidities such as diabetes and congestive heart failure. Lab results: BUN, serum creatinine, serum albumin, total bilirubin, and so on. The only lab results which are conducted within 14 days before the surgery are included. Conditions of operation: Duration of Anesthesia, total operation time, and so on.

In this project, Data cleaning and visualization were conducted using R. Other programming language was used for modeling.

Data preparation

1. Import packages

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(gt)
```

```
## Warning: package 'gt' was built under R version 4.3.2
```

```
library(webshot2)
library(gtsummary)
```

2. Load the dataset

```
NSQIP <- read_csv(file = "Project 4 - NSQIP Data.csv")
```

```
## New names:
## Rows: 162945 Columns: 145
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (66): SEX, RACE_NEW, ETHNICITY_HISPANIC, PRNCPTX, INOUT, TRANST, DISCHDE... dbl
## (77): ...1, CaseID, CPT, Age, AdmYR, HEIGHT, WEIGHT, DPRNA, DPRBUN, DPRC... lgl
## (2): DPRPT, DNEURODEF
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...1`
```

3. Exclude the subjects without cardiovascular surgery or with severe ARF with dialysis at the pre-operation time.

```
# Filter the subject with Cardiovascular Surgery by CPT code
dt_heart <- NSQIP %>%
  filter(33016 <= CPT & CPT <= 33999)
dt_arteries <- NSQIP%>%
  filter(34001 <= CPT & CPT<= 37799)
dt_cv <- rbind(dt_heart, dt_arteries)


# Exclude the patients with Dialysis at pre-op
dt_cv_no_dialysis <- dt_cv %>%
  filter(DIALYSIS != "Yes")
```

4. The lab results which are conducted within 14 days before the surgery are included. This is because the lab results performed more than 15 days before surgery do not reflect the subject's condition at the preoperative timing.

```
# Change the lab results to NA if the lab test was conducted before 15 days or more from the sur
gery. (DPRxxx: Days from the pre-lab test to the surgery)
dt_lab_updated_bmi <- dt_cv_no_dialysis %>%
  mutate(
    PRSODM1 = if_else(DPRNA > 14 | is.na(DPRNA), NA, PRSODM),
    PRBUN1 = if_else(DPRBUN > 14 | is.na(DPRBUN), NA, PRBUN),
    PRCREAT1 = if_else(DPRCREAT > 14 | is.na(DPRCREAT), NA, PRCREAT),
    PRALBUM1 = if_else(DPRALBUM > 14 | is.na(DPRALBUM), NA, PRALBUM),
    PRBILI1 = if_else(DPRBILI > 14 | is.na(DPRBILI), NA, PRBILI),
    PRSGOT1 = if_else(DPRSGOT > 14 | is.na(DPRSGOT), NA, PRSGOT),
    PRALKPH1 = if_else(DPRALKPH > 14 | is.na(DPRALKPH), NA, PRALKPH),
    PRWBC1 = if_else(DPRWBC > 14 | is.na(DPRWBC), NA, PRWBC),
    PRHCT1 = if_else(DPRHCT > 14 | is.na(DPRHCT), NA, PRHCT),
    PRPLATE1 = if_else(DPRPLATE > 14 | is.na(DPRPLATE), NA, PRPLATE),
    PRPTT1 = if_else(DPRPTT > 14 | is.na(DPRPTT), NA, PRPTT),
    PRINR1 = if_else(DPRINR > 14 | is.na(DPRINR), NA, PRINR)
  )%>%
  # Calculate BMI
  mutate(
    BMI = WEIGHT/(HEIGHT^2)*703
  )

# Remove the duplicated cases
dt_unique_cases <- dt_lab_updated_bmi[duplicated(dt_lab_updated_bmi$CaseID)==FALSE,]
```

5. Set the target

```
dt_target <- dt_unique_cases%>%
  mutate(renaloutcome = ifelse(RENAINSF=="Progressive Renal Insufficiency"|OPRENAFL=="Acute Rena
l Failure",1,0))
table(dt_target$renaloutcome)
```

```
##
##    0    1
## 5282  154
```

```
write_csv(dt_target, "NSQIPdataset_cleaned.csv")
```

6. Change the columns' names

```
dt_renamed <- dt_target %>%
  rename("Sex" = "SEX", "Race"="RACE_NEW",
         "Ethnicity"="ETHNICITY_HISPANIC" , "Smoke"="SMOKE",
         "Diabetes"="DIABETES" ,"COPD" ="HXCOPD",
         "Dyspnea"="DYSPNEA", "Congestive Heart Failure" ="HXCHF",
         "Hypertension"="HYPERMED", "Ascites"="ASCITES",
         "Disseminated Cancer"="DISCANCR", "Bleeding disorders"="BLEEDDIS",
         "Severe Acure Renal Failure"="OPRENAFL",

         "Total operation time"="OPTIME", "Sodium"="PRSODM1",
         "BUN"="PRBUN1",
         "Serum Creatinine"="PRCREAT1",
         "Serum Albumin"="PRALBUM1",
         "Total Bilirubin"="PRBILI1",
         "SGOT"="PRSGOT1",
         "Alkaline Phosphatase"="PRALKPH1",
         "WBC"="PRWBC1",
         "Hematocrit"="PRHCT1",
         "Platelet count"="PRPLATE1",
         "PTT"="PRPTT1",
         "INR"="PRINR1"  )
```

Data visualization

1. Make the demographic table

```
demographics <- dt_renamed %>%
  select("Sex","Race",
         "Ethnicity",  "Smoke",
         "Diabetes", "COPD",
         "Dyspnea", "Congestive Heart Failure",
         "Hypertension", "Ascites",
         "Disseminated Cancer","Bleeding disorders",
         "Severe Acure Renal Failure")%>%
  tbl_summary(by=  "Severe Acure Renal Failure",
              sort = list(everything() ~ "frequency"))%>%
  add_overall()

demographics
```

Characteristic	**Overall**, N = 5,436[1]	**Acute Renal Failure**, N = 101[1]	**No Complication**, N = 5,335[1]
Sex			
male	3,464 (64%)	76 (75%)	3,388 (64%)

[1] n (%)

Characteristic	**Overall**, N = 5,436[1]	**Acute Renal Failure**, N = 101[1]	**No Complication**, N = 5,335[1]
female	1,972 (36%)	25 (25%)	1,947 (36%)
Race			
White	4,075 (75%)	78 (77%)	3,997 (75%)
Unknown/Not Reported	700 (13%)	10 (9.9%)	690 (13%)
Black or African American	514 (9.5%)	10 (9.9%)	504 (9.4%)
Asian	117 (2.2%)	3 (3.0%)	114 (2.1%)
American Indian or Alaska Native	17 (0.3%)	0 (0%)	17 (0.3%)
Native Hawaiian or Pacific Islander	13 (0.2%)	0 (0%)	13 (0.2%)
Ethnicity			
No	4,525 (83%)	87 (86%)	4,438 (83%)
Unknown	628 (12%)	9 (8.9%)	619 (12%)
Yes	283 (5.2%)	5 (5.0%)	278 (5.2%)
Smoke	1,746 (32%)	39 (39%)	1,707 (32%)
Diabetes			
NO	3,931 (72%)	59 (58%)	3,872 (73%)
NON-INSULIN	817 (15%)	22 (22%)	795 (15%)
INSULIN	688 (13%)	20 (20%)	668 (13%)
COPD	673 (12%)	22 (22%)	651 (12%)
Dyspnea			
No	4,580 (84%)	81 (80%)	4,499 (84%)
MODERATE EXERTION	782 (14%)	19 (19%)	763 (14%)
AT REST	74 (1.4%)	1 (1.0%)	73 (1.4%)

[1] n (%)

Characteristic	**Overall**, N = 5,436[1]	**Acute Renal Failure**, N = 101[1]	**No Complication**, N = 5,335[1]
Congestive Heart Failure	235 (4.3%)	8 (7.9%)	227 (4.3%)
Hypertension	3,983 (73%)	87 (86%)	3,896 (73%)
Ascites	12 (0.2%)	1 (1.0%)	11 (0.2%)
Disseminated Cancer	42 (0.8%)	0 (0%)	42 (0.8%)
Bleeding disorders	1,118 (21%)	27 (27%)	1,091 (20%)

[1] n (%)

2. Make the summary of the lab results

```
lab <- dt_renamed %>%
  select( "Age", "BMI","Total operation time","Sodium",
          "BUN",
          "Serum Creatinine",
          "Serum Albumin",
          "Total Bilirubin",
          "SGOT",
          "Alkaline Phosphatase",
          "WBC",
          "Hematocrit",
          "Platelet count",
          "PTT",
          "INR", "Severe Acure Renal Failure")%>%
  tbl_summary(by= "Severe Acure Renal Failure",
              missing = "no",
              digits = list(everything() ~ c(1, 1)),
              statistic = list(all_continuous() ~ "{mean} ({sd}) / ({max}-{min})"))%>%
  add_overall()

lab
```

Characteristic	**Overall**, N = 5,436[1]	**Acute Renal Failure**, N = 101[1]	**No Complication**, N = 5,335[1]
Age	66.5 (12.6) / (89.0-21.0)	69.1 (9.9) / (87.0-32.0)	66.5 (12.6) / (89.0-21.0)
BMI	28.5 (6.3) / (84.1-11.2)	29.8 (7.6) / (59.2-16.6)	28.5 (6.3) / (84.1-11.2)
Total operation time	189.3 (122.8) / (983.0-0.0)	307.6 (160.4) / (859.0-65.0)	187.0 (120.9) / (983.0-0.0)

[1] Mean (SD) / (Maximum-Minimum)

Characteristic	**Overall**, N = 5,436[1]	**Acute Renal Failure**, N = 101[1]	**No Complication**, N = 5,335[1]
Sodium	138.4 (3.4) / (156.0-119.0)	137.6 (4.4) / (156.0-125.0)	138.5 (3.4) / (156.0-119.0)
BUN	19.7 (11.3) / (149.0-1.0)	30.6 (20.4) / (127.0-8.0)	19.5 (10.9) / (149.0-1.0)
Serum Creatinine	1.1 (0.7) / (15.0-0.1)	1.9 (1.5) / (9.8-0.6)	1.1 (0.7) / (15.0-0.1)
Serum Albumin	3.6 (0.6) / (7.2-1.0)	3.1 (0.8) / (4.5-1.2)	3.6 (0.6) / (7.2-1.0)
Total Bilirubin	0.7 (0.7) / (14.7-0.1)	0.9 (1.9) / (14.7-0.2)	0.7 (0.6) / (11.5-0.1)
SGOT	33.1 (48.6) / (949.0-1.9)	38.7 (40.1) / (223.0-8.0)	32.9 (48.8) / (949.0-1.9)
Alkaline Phosphatase	86.0 (47.2) / (907.0-1.1)	81.6 (25.8) / (185.0-31.0)	86.1 (47.7) / (907.0-1.1)
WBC	8.7 (3.9) / (50.0-0.9)	11.8 (7.0) / (50.0-4.3)	8.6 (3.7) / (48.9-0.9)
Hematocrit	38.3 (6.1) / (60.0-10.0)	34.3 (7.2) / (53.2-13.2)	38.4 (6.0) / (60.0-10.0)
Platelet count	231.2 (86.5) / (864.0-1.0)	232.2 (90.9) / (501.0-26.0)	231.2 (86.4) / (864.0-1.0)
PTT	35.3 (15.0) / (120.0-7.1)	36.8 (15.1) / (115.7-21.0)	35.2 (15.0) / (120.0-7.1)
INR	1.1 (0.4) / (10.0-0.8)	1.3 (0.6) / (4.3-0.9)	1.1 (0.4) / (10.0-0.8)

[1] Mean (SD) / (Maximum-Minimum)