# Predicting mortality of SAH

```
knitr::opts_chunk$set(
    echo = TRUE,
    message = FALSE,
    warning = FALSE,
    root.dir = "C:/Users/yiy43/Desktop/PH1976 Final Project/predicts_mortality/"
)
demo_tr <- read.csv("./demo_train.csv", header = T)
demo_te <- read.csv("./demo_test.csv", header = T)
med_tr <- read.csv("./medication_train.csv", header = T)
med_te <- read.csv("./medication_test.csv", header = T)
proc_tr <- read.csv("./procedure_train.csv", header = T)
proc_te <- read.csv("./procedure_test.csv", header = T)

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(anytime)
```

```
## Warning: package 'anytime' was built under R version 3.6.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v stringr 1.4.0
## v tidyr   1.0.2     v forcats 0.4.0
## v readr   1.3.1
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(compareGroups)
```

```
## Warning: package 'compareGroups' was built under R version 3.6.3
```

**Factor-to-time conversion** also created LOS variable

```
demo_tr <- demo_tr %>%
  mutate(New_admitted_dt_tm = anytime(New_admitted_dt_tm),
         New_discharge_dt_tm = anytime(New_discharge_dt_tm),
         los = difftime(New_discharge_dt_tm, New_admitted_dt_tm,
                        units = "days"))
demo_te <- demo_te %>%
  mutate(New_admitted_dt_tm = anytime(New_admitted_dt_tm),
         New_discharge_dt_tm = anytime(New_discharge_dt_tm),
         los = difftime(New_discharge_dt_tm, New_admitted_dt_tm,
                        units = "days")
         )

med_tr$med_started_dt_tm <- anytime(med_tr$med_started_dt_tm)
med_te$med_started_dt_tm <- anytime(med_te$med_started_dt_tm)

proc_tr$procedure_dt_tm <- anytime(proc_tr$procedure_dt_tm)
proc_te$procedure_dt_tm <- anytime(proc_te$procedure_dt_tm)
```

```
cg <- compareGroups(death ~ gender+race+age_in_years+los,
                    data = demo_tr,
                    method = c(3,3,1,2),
                    max.xlev = 12,
                    Q1 = 0, Q3 = 1)
createTable(cg, show.n = F, show.p.overall = T)
```

**Descriptive of demographic training data**

```
##
## --------Summary descriptives table by 'death'---------
##
## ____________________________________________________________________
##                                False           True       p.overall
##                               N=3659          N=1252
## ____________________________________________________________________
## gender:                                                        0.196
##     Female                  2168 (59.3%)     709 (56.6%)
##     Male                    1490 (40.7%)     543 (43.4%)
##     Unknown                   1 (0.03%)       0 (0.00%)
## race:                                                             .
##     African American         835 (22.8%)     246 (19.6%)
##     Asian                    61 (1.67%)      33 (2.64%)
##     Asian/Pacific Islander    1 (0.03%)       0 (0.00%)
##     Biracial                  4 (0.11%)       1 (0.08%)
##     Caucasian               2411 (65.9%)     832 (66.5%)
```

```
##     Hispanic                 36 (0.98%)      15 (1.20%)
##     Mid Eastern Indian        3 (0.08%)       0 (0.00%)
##     Native American          36 (0.98%)      17 (1.36%)
##     Other                   171 (4.67%)      47 (3.75%)
##     Pacific Islander          3 (0.08%)       1 (0.08%)
##     Unknown                  98 (2.68%)      60 (4.79%)
## age_in_years                 56.4 (15.3)     63.3 (15.9)    <0.001
## los                        13.2 [1.00;330] 5.47 [1.01;224]  <0.001
## ------------------------------------------------------------------
```

- delete the one observation that has unknown gender in demo train
- race has too many categories, need to combine some of them.
  - Combine Asian, Pacific Islander to category Asian/Pacific Islander.
  - Combine the minority groups Biracial, Hispanic, Mid Eastern Indian, Native American, and Other to a new category, Others.
  - We may want to keep the Unknown category in race. (Are Unknowns missing values?)

```
demo_tr_new <- demo_tr %>% filter(gender != "Unknown") %>%
  mutate(race = fct_recode(race,
                           `Asian/Pacific Islander` = 'Asian',
                           `Asian/Pacific Islander` = 'Pacific Islander',
                           `Others` = 'Biracial',
                           `Others` = 'Hispanic',
                           `Others` = 'Mid Eastern Indian',
                           `Others` = 'Native American',
                           `Others` = 'Other'))
```

Descriptive table for the new demographic training data:

```
cg.new <- compareGroups(death ~ gender+race+age_in_years+los,
                    data = demo_tr_new,
                    method = c(3,3,1,2),
                    Q1 = 0, Q3 = 1)
createTable(cg.new, show.n = F, show.p.overall = T)
```

```
##
## --------Summary descriptives table by 'death'---------
##
## ------------------------------------------------------------------
##                                False           True       p.overall
##                               N=3658          N=1252
## ------------------------------------------------------------------
## gender:                                                      0.109
##     Female                  2168 (59.3%)     709 (56.6%)
##     Male                    1490 (40.7%)     543 (43.4%)
## race:                                                       <0.001
##     African American         835 (22.8%)     246 (19.6%)
##     Asian/Pacific Islander    65 (1.78%)      34 (2.72%)
##     Others                   250 (6.83%)     80 (6.39%)
##     Caucasian               2411 (65.9%)     832 (66.5%)
##     Unknown                   97 (2.65%)      60 (4.79%)
## age_in_years                 56.4 (15.3)     63.3 (15.9)    <0.001
## los                        13.2 [1.00;330] 5.47 [1.01;224]  <0.001
## ------------------------------------------------------------------
```

Should we combine Unknown into Others?

```
med.tmp1 <- med_tr %>% merge(demo_tr, by = "patient_sk", all.x = T) %>%
  mutate(time_from_admit = difftime(med_started_dt_tm, New_admitted_dt_tm, units = "days"))

med.tmp1.tb <- med.tmp1 %>% group_by(patient_sk) %>%
  summarise(n_med = n(),
            min_time_from_admit = min(time_from_admit),
            max_time_from_admit = max(time_from_admit),
            any_vaso = any(generic_name %in% c("dopamine", "phenylephrine",  "norepinephrine")),
            n_vaso = sum(generic_name %in% c("dopamine", "phenylephrine",  "norepinephrine")))
```

```
head(med.tmp1.tb, 10)
```

**Descriptives of medication training data**

```
## # A tibble: 10 x 6
##    patient_sk n_med min_time_from_admit max_time_from_admit any_vaso n_vaso
##         <int> <int> <drtn>              <drtn>              <lgl>     <int>
##  1  105443638    25 0.341666667 days    0.6666667 days      FALSE         0
##  2  105449238     8 0.572222222 days    0.5895833 days      FALSE         0
##  3  105450304     1 0.416666667 days    0.4166667 days      TRUE          1
##  4  105566847    18 0.147222222 days    0.9368056 days      FALSE         0
##  5  105587629    13 0.007638889 days    0.9493056 days      FALSE         0
##  6  105592791    15 0.273611111 days    0.6666667 days      FALSE         0
##  7  105626642    12 0.291666667 days    0.7916667 days      FALSE         0
##  8  105689998    29 0.175694444 days    1.0000000 days      FALSE         0
##  9  105694554    22 0.214583333 days    0.6666667 days      FALSE         0
## 10  105837639    31 0.527083333 days    0.7840278 days      FALSE         0
```

```
cg2 <- compareGroups(~ n_med + n_med + min_time_from_admit + max_time_from_admit + any_vaso + n_vaso,
                     data = med.tmp1.tb,
                     method = c(1,2,2,2,3,2), Q1 = 0, Q3 = 1)
createTable(cg2, show.n = F)
```

```
##
## --------Summary descriptives table ---------
##
## ------------------------------------
##                          [ALL]
##                          N=4911
## ------------------------------------
## n_med                  26.0 (18.3)
## n_med                 23.0 [1.00;182]
## min_time_from_admit 0.25 [0.00;1.00]
## max_time_from_admit 0.89 [0.00;1.02]
## any_vaso:
##     FALSE              4022 (81.9%)
##     TRUE               889 (18.1%)
## n_vaso                0.00 [0.00;20.0]
## ------------------------------------
```

n_med: number of medications per patient.

min_time_from_admit, max_time_from_admit: minimal/maximum time of medication administration

after admission to hospital for each patient.

any_vaso: Did this patient receive at least 1 vassopressor (dopamin, phenylephrine, norepinephrine)?

n_vaso: Number of vassopressors the patient took.

The number of medications administered to a patient averages at 26 (median = 23) and ranges from 1 to 182. The majority of medications were administered within one day of admission to the hospital. About 18.1% of patient received at lease one vassopressor.

A descriptive table grouped by any_vaso with

```
med.tmp1.tb$any_vaso <- as.factor(med.tmp1.tb$any_vaso)
cg3 <- compareGroups(any_vaso ~ n_med + n_med + min_time_from_admit + max_time_from_admit + any_vaso + n
                     data = med.tmp1.tb,
                     method = c(1,2,2,2,2,1,2), Q1 = 0, Q3 = 1)
createTable(cg3, show.n = F)
```

```
##
## --------Summary descriptives table by 'any_vaso'---------
##
## _____________________________________________________________________
##                           FALSE             TRUE       p.overall
##                           N=4022           N=889
## _____________________________________________________________________
## n_med                   23.2 (15.8)      38.8 (22.6)     <0.001
## n_med                 21.0 [1.00;148]  36.0 [1.00;182]   <0.001
## min_time_from_admit 0.26 [0.00;1.00] 0.23 [0.00;0.98]  <0.001
## max_time_from_admit 0.88 [0.00;1.00] 0.93 [0.11;1.02]  <0.001
## any_vaso:                                               0.000
##     FALSE             4022 (100%)       0 (0.00%)
##     TRUE               0 (0.00%)        889 (100%)
## n_vaso                 0.00 (0.00)      2.06 (1.88)     <0.001
## n_vaso               0.00 [0.00;0.00] 1.00 [1.00;20.0]   0.000
## _____________________________________________________________________
```

Patients who have at least one vassopressors tend to have more medications and larger range of medication administration time.

```
med_freq <- med_tr %>% group_by(generic_name) %>%
  summarise(n = n()) %>%
  arrange(desc(n))
```

The top 10 most freqeuntly administered medications:

```
head(med_freq, 10)
```

```
## # A tibble: 10 x 2
##    generic_name           n
##    <fct>              <int>
##  1 lvp solution       12674
##  2 sodium chloride     5823
##  3 ondansetron         5540
##  4 fentanyl            5499
##  5 potassium chloride  4319
##  6 acetaminophen       4196
##  7 propofol            4021
##  8 nicardipine         3778
##  9 labetalol           3574
```

```
## 10 morphine            3362
```

Those most frequently used medications might not be informative about the mortality.

```
proc.tmp1 <- proc_tr %>% merge(demo_tr, by = "patient_sk", all.x = T) %>%
  mutate(time_from_admit = difftime(procedure_dt_tm, New_admitted_dt_tm, units = "days"))

proc.tmp1.tb <- proc.tmp1 %>% group_by(patient_sk) %>%
  summarise(n_proc = n(),
            min_time_from_admit = min(time_from_admit),
            max_time_from_admit = max(time_from_admit))
```

```
head(proc.tmp1.tb, 100)
```

**Descriptives of procedure training data**

```
## # A tibble: 100 x 4
##    patient_sk n_proc min_time_from_admit max_time_from_admit
##         <int>  <int> <drtn>              <drtn>
##  1  105443638      1 0.00000000 days     0.00000000 days
##  2  105449238      3 0.00000000 days     0.00000000 days
##  3  105450304      3 0.00000000 days     0.00000000 days
##  4  105566847      3 0.42638889 days     0.42638889 days
##  5  105587629      4 0.05902778 days     0.05902778 days
##  6  105592791      2 0.00000000 days     0.00000000 days
##  7  105626642      1 0.00000000 days     0.00000000 days
##  8  105689998      5 0.00000000 days     1.00000000 days
##  9  105694554      2 0.00000000 days     1.00000000 days
## 10  105837639      4 0.00000000 days     0.00000000 days
## # ... with 90 more rows
```

```
cg3 <- compareGroups(~ n_proc + n_proc + min_time_from_admit + max_time_from_admit,
                     data = proc.tmp1.tb,
                     method = c(1,2,2,2),
                     Q1 = 0, Q3 = 1)
createTable(cg3, show.n = F)
```

```
##
## --------Summary descriptives table ---------
##
## ------------------------------------
##                          [ALL]
##                          N=4911
## ------------------------------------
## n_proc                 3.45 (4.56)
## n_proc               2.00 [1.00;95.0]
## min_time_from_admit 0.00 [0.00;1.00]
## max_time_from_admit 0.17 [0.00;1.04]
## ------------------------------------
```

The top 10 most freqeuntly used procedure:

```
proc_freq <- proc_tr %>% group_by(procedure_id) %>%
  summarise(n = n(),
            procedure_description = first(procedure_description)) %>%
```

```
  arrange(desc(n))
```

```
head(proc_freq, 10)
```

```
## # A tibble: 10 x 3
##    procedure_id     n procedure_description
##           <int> <int> <fct>
##  1           44  1753 arteriography of cerebral arteries
##  2          141   837 insertion of endotracheal tube
##  3         2548   798 continuous invasive mechanical ventilation for less than ~
##  4         2549   607 continuous invasive mechanical ventilation for 96 consecu~
##  5           43   546 venous catheterization, not elsewhere classified
##  6         3006   494 clipping of aneurysm
##  7         4756   442 endovascular (total) embolization or occlusion of head an~
##  8          337   342 ventriculostomy
##  9       122879   263 insertion of endotracheal airway into trachea, via natura~
## 10           42   261 arterial catheterization
```

Are any of the procedures predictive of death?