HDS 5230 High performance computing

Homework Week 2

Miao Cai[*]

2019-01-30

You will use these datasets to answer some questions listed below. You must be careful to think about what the appropriate denominator is for each question. As you code the answers, be mindful to use the 'high performance' coding approaches in data.table.

```r
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```r
data_path = "healthcare2/"
csv_files = list.files(path = data_path, pattern = "*.csv")

readallcsv = function(i){
  assign(
    gsub(".csv", "", csv_files[i]),
    fread(paste0(data_path, csv_files[i])),
    envir = parent.frame()
  )
}

for (i in seq_along(csv_files))  readallcsv(i)
```

1) Are men more likely to die than women in this group of patients? Assume people without a date of death in the mortality table are still alive.

Here I recode all patients without a `Gender` to other.

```r
Patient[!Gender %in% c("female", "male"), Gender := "other"]
q1 = Mortality[Patient, on = "PatientID"]
q1[, .(death_percent = sum(!is.na(DateOfDeath))/.N), by = Gender][order(-death_percent)]
```

```
##    Gender death_percent
## 1:   male     0.3594713
## 2: female     0.3511153
## 3:  other     0.3492670
```

According to the returned data, it seems that males do have a little bit higher chance to die than women in this group of patients, although the difference is nominal.

2) I am interested to know if there are patterns in the disease groups across gender. For every patient with at least one outpatient visit, identify if they have been diagnosed with any of the 22 conditions listed in the diseaseMap table at any time point. You will need to consider all three ICD columns in the outpatientVisit file (not just one). Create a table with the rate of disease for each condition for men, women, and all. It should look like this, where the XX% is the percent with the condition:

```r
OutpatientVisit = melt(
  OutpatientVisit,
  measure.vars = patterns("^ICD10"),
  id.vars = c("VisitID", "PatientID"),
```

[*] Department of Epidemiology and Biostatistics, Saint Louis University. Email address miao.cai@slu.edu

```
  value.name = "ICD10"
)

Npatients = nrow(Patient)

q2 = DiseaseMap[
  OutpatientVisit, on = "ICD10"
  ][, .N, by = .(PatientID, Condition)][Patient, on = "PatientID"]

q2[
  ,.(condition_N = .N), by = .(Condition, Gender)
  ][,condition_percent := paste0(condition_N*100/nrow(Patient), "%")][
    order(Condition, Gender)]
```

```
##                           Condition Gender condition_N condition_percent
##  1:                        Alcohol female         737            3.685%
##  2:                        Alcohol   male         713            3.565%
##  3:                        Alcohol  other         127            0.635%
##  4:                         Cancer female         475            2.375%
##  5:                         Cancer   male         446             2.23%
##  6:                         Cancer  other          74             0.37%
##  7:       Congestive_heart_failure female         291            1.455%
##  8:       Congestive_heart_failure   male         501            2.505%
##  9:       Congestive_heart_failure  other          72             0.36%
## 10:                       Dementia female         303            1.515%
## 11:                       Dementia   male         266             1.33%
## 12:                       Dementia  other          48             0.24%
## 13:                     Depression female        1182             5.91%
## 14:                     Depression   male         751            3.755%
## 15:                     Depression  other         173            0.865%
## 16:    Diabetes_with_complications female         397            1.985%
## 17:    Diabetes_with_complications   male         344             1.72%
## 18:    Diabetes_with_complications  other          70             0.35%
## 19: Diabetes_without_complications female         974             4.87%
## 20: Diabetes_without_complications   male         875            4.375%
## 21: Diabetes_without_complications  other         146             0.73%
## 22:                          Drugs female         387            1.935%
## 23:                          Drugs   male         343            1.715%
## 24:                          Drugs  other          59            0.295%
## 25:                            HIV female          54             0.27%
## 26:                            HIV   male          56             0.28%
## 27:                            HIV  other          15            0.075%
## 28:                   Hypertension female        2712            13.56%
## 29:                   Hypertension   male        2866            14.33%
## 30:                   Hypertension  other         441            2.205%
## 31:                      LiverMild female          90             0.45%
## 32:                      LiverMild   male          83            0.415%
## 33:                      LiverMild  other          11            0.055%
## 34:                    LiverSevere female         460              2.3%
## 35:                    LiverSevere   male         466             2.33%
## 36:                    LiverSevere  other          87            0.435%
## 37:        Metastatic_solid_tumour female         312             1.56%
## 38:        Metastatic_solid_tumour   male         309            1.545%
## 39:        Metastatic_solid_tumour  other          38             0.19%
## 40:          Myocardial_infarction female         300              1.5%
## 41:          Myocardial_infarction   male         526             2.63%
## 42:          Myocardial_infarction  other          74             0.37%
## 43:                        Obesity female        1745            8.725%
## 44:                        Obesity   male        1247            6.235%
## 45:                        Obesity  other         250             1.25%
## 46:                      Paralysis female         139            0.695%
## 47:                      Paralysis   male         104             0.52%
## 48:                      Paralysis  other          25            0.125%
## 49:           Peptic_ulcer_disease female          97            0.485%
## 50:           Peptic_ulcer_disease   male          80              0.4%
## 51:           Peptic_ulcer_disease  other          14             0.07%
## 52:    Peripheral_vascular_disease female         229            1.145%
## 53:    Peripheral_vascular_disease   male         200                1%
## 54:    Peripheral_vascular_disease  other          47            0.235%
## 55:                      Pulmonary female         672             3.36%
## 56:                      Pulmonary   male         647            3.235%
## 57:                      Pulmonary  other         114             0.57%
## 58:                          Renal female         345            1.725%
## 59:                          Renal   male         305            1.525%
## 60:                          Renal  other          48             0.24%
## 61:                      Rheumatic female         124             0.62%
## 62:                      Rheumatic   male          98             0.49%
## 63:                      Rheumatic  other          21            0.105%
## 64:                         Stroke female         253            1.265%
## 65:                         Stroke   male         271            1.355%
## 66:                         Stroke  other          39            0.195%
## 67:                           <NA> female        8229           41.145%
## 68:                           <NA>   male        7743           38.715%
## 69:                           <NA>  other        1370             6.85%
##                          Condition Gender condition_N condition_percent
```

I assume the denominator here is the number of patients : 20000

3) Calculate the mortality rate for every year between 2005 and 2018. Is it generally increasing, or decreasing? Assume patients are only at risk of death as of their first visit (in the outpatient Visit file). Once they have died, they are no longer at risk in subsequent years

```
q3 = Mortality[Patient, on = "PatientID"]
q3[, year := as.integer(substr(DateOfDeath, 1, 4))]
q3 = q3[, .(N_death = .N), by = year][order(year)][!is.na(year)]

q3[, cum_death := cumsum(N_death)
  ][, atrisk := 20000 - shift(cum_death, fill = 0, type = "lag")
    ][, mortality_rate := N_death*100/atrisk]

q3
```

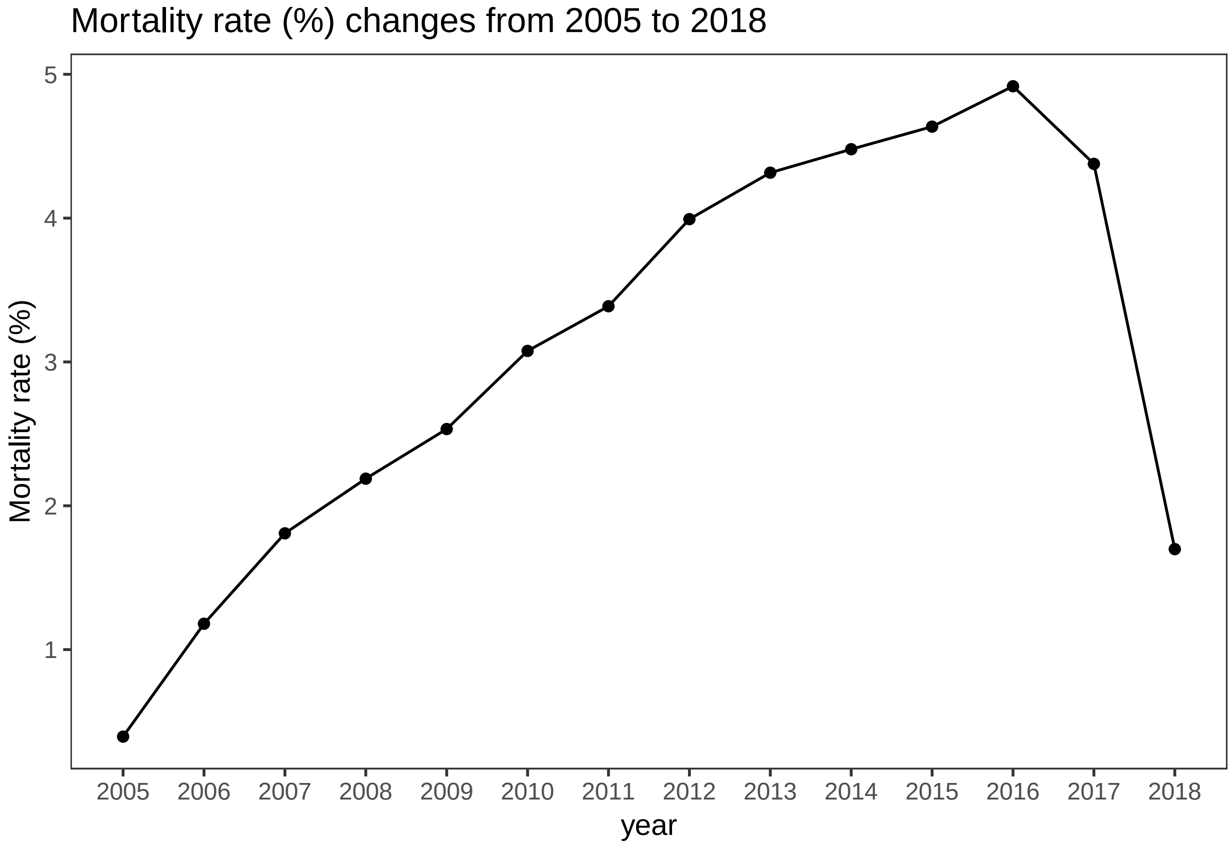
```
##     year N_death cum_death atrisk mortality_rate
##  1: 2005      79        79  20000       0.395000
##  2: 2006     235       314  19921       1.179660
##  3: 2007     356       670  19686       1.808392
##  4: 2008     423      1093  19330       2.188308
##  5: 2009     479      1572  18907       2.533453
##  6: 2010     567      2139  18428       3.076840
##  7: 2011     605      2744  17861       3.387268
##  8: 2012     689      3433  17256       3.992814
##  9: 2013     715      4148  16567       4.315809
## 10: 2014     710      4858  15852       4.478930
## 11: 2015     702      5560  15142       4.636111
## 12: 2016     710      6270  14440       4.916898
## 13: 2017     601      6871  13730       4.377276
## 14: 2018     223      7094  13129       1.698530
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
ggplot(q3, aes(year, mortality_rate)) +
  geom_point() + geom_line() +
  scale_x_continuous("year", labels = 2005:2018, breaks = 2005:2018) +
  labs(title = "Mortality rate (%) changes from 2005 to 2018") +
  ylab("Mortality rate (%)") + theme_test()
```



According to the time trend plot, the mortality rate has been generally increasing, while it experienced a major drop in the recent two years (2017 and 2018).

a. This is a harder question to answer than at first glance. What should the denominator of patients be for every year? How will you calculate it?

From my understanding, the denominator should be the patients at risk in the specific year (who were still alive). I calculated it by excluding the patients till the last year.