Mining electronic medical records

QUESTIONS

---

40 marks

The scientific context for this question can be found in the "MedicalRecordsContext" file. All questions should be answered with respect to this context. (**Note:** This is an entirely fictional data set created by R.W. Oldford.)

a. **(3 marks)** Populations and sample

   i. *(1 mark)* What is the target population here?

      - People treat with Treatment A or B

   ii. *(1 mark)* What is the study population?

      - Past people who had treatment A or B that have record

   iii. *(1 mark)* What is the sample?

      - The 3200 patients

In the following questions, **all** code **must** be written using magrittr pipes in combination with functions from dplyr or others from the tidyverse.

To compare treatments, in each part you need to determine the proportion of patients who recovered for each treatment.

b. **(10 marks)** The analysis will begin by comparing the recovery rates of the treatments ignoring the sex and age of the patients.

   i. *(4 marks)* Beginning with `medicalRecords %>%`, form a pipeline which will produce a data set having two variables, `Treatment` and `recoveryRate` (in that order), which shows the **percentage** of **all** patients (ignoring sex and age) who recovered from each treatment.

   - Save this data as the value of `medRecovery`
   - Show your code and the result `medRecovery`.

```
medicalRecords %>%
transmute(Treatment = Treatment,
          recoveryRate = if_else(Treatment == "A",
                                 medicalRecords%>%
                                    subset(Treatment == "A" & Outcome == "Recovered") %>%
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(1600)%>%
                                    multiply_by(100),
                                 medicalRecords %>%
                                    subset(Treatment == "B" & Outcome == "Recovered") %>%
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(1600)%>%
                                    multiply_by(100))) %>%
unique->
medRecovery
medRecovery
```

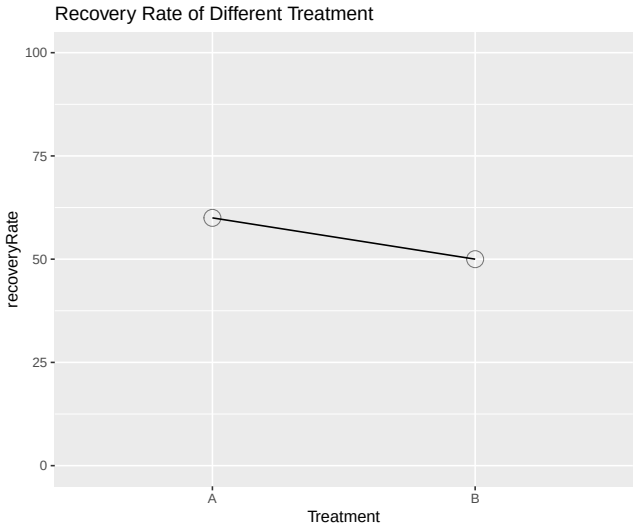
```
##   Treatment recoveryRate
## 1         A           60
## 5         B           50
```

   i. *(5 marks)* Using the dataset `medRecovery` just created, produce a `ggplot` that has

   - `Treatment` on the horizontal axis, `recoveryRate` on the vertical axis,
   - a point of size 5 at the different values of `(Treatment, recoveryRate)`
   - a line connecting the above two points (**hint:** an aesthetic of `group = 1` is required to make this work)
   - vertical limits between 0 and 100
   - suitable title and axis labels

   Show your code and the resulting `ggplot`

```
ggplot(data = medRecovery, mapping = aes(x = Treatment, y = recoveryRate)) +
geom_point(mapping = aes(group = 1),alpha = 0.5, size = 5, pch = 21) +
geom_line(aes(group = 1)) +
ylim(0,100) +
ggtitle("Recovery Rate of Different Treatment")
```



i. *(1 mark)* On the basis of these results, which treatment would you recommend for this disease, `A` or `B`?
   - Based on the above results, I would recommend treatment A for this disease because it has a higher recovery rate.

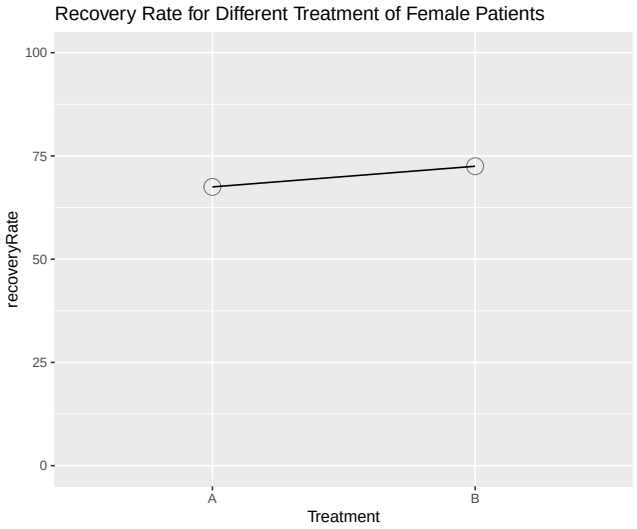
c. **(8 marks)** Objections are raised that the previous analysis did not distinguish between the sexes. It might be, for example, that the female physiology reacts different than does the male one.

i. *(4 marks)* Again, beginning with `medicalRecords %>%`, now form a pipeline which will produce a data set having two variables, `Treatment` and `recoveryRate` (in that order), which shows the **percentage** of **female** patients (ignoring age) who recovered from each treatment.

- Save this data as the value of `medRecoveryFemale`.
- Using `medRecoveryFemale` produce a 'ggplot just as was done for all patients before but now only for females.
- **Do not** show the code for your ggplot construction. Just the plot itself.
- **Do** show your pipeline code that constructs the `medRecoveryFemale`, and print `medRecoveryFemale`
- On the basis of these results, would you make the same recommendation for females as you did before when you ignored the patient sex? Why, or why not?

```
medicalRecords %>%
transmute(Treatment = Treatment,
          recoveryRate = if_else(Treatment == "A",
                                 medicalRecords %>%
                                    subset(Treatment == "A" & Outcome == "Recovered" & Sex == "Fema
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(medicalRecords %>%
                                                subset(Treatment == "A" & Sex == "Female") %>%
                                                extract(, "Freq") %>%
                                                sum())%>%
                                    multiply_by(100),
                                 medicalRecords %>%
                                    subset(Treatment == "B" & Outcome == "Recovered" & Sex == "Fema
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(medicalRecords %>%
                                                subset(Treatment == "B" & Sex == "Female") %>%
                                                extract(, "Freq") %>%
                                                sum())%>%
                                    multiply_by(100)))%>%
unique->
medRecoveryFemale
medRecoveryFemale
```

```
##   Treatment recoveryRate
## 1         A         67.5
## 5         B         72.5
```



- No, I would not recommend the same treatment as before. Previously, our analysis shows that treatment A has higher recovery rate. However, when we consider only the female patients, treatment B has a higher recovery rate.

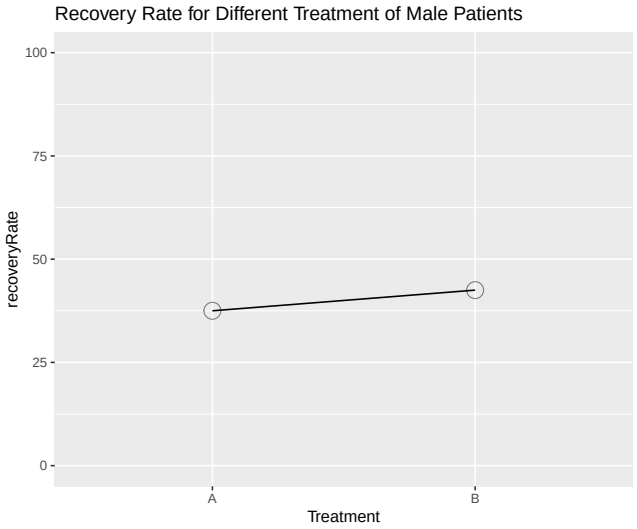
i. *(3 marks)* Repeat the previous part, but now for male patients only, saving the results as `medRecoveryMale`

- Show your code and results

- On the basis of these results, would you make the same recommendation for males as you

```
medicalRecords %>%
transmute(Treatment = Treatment,
          recoveryRate = if_else(Treatment == "A",
                                 medicalRecords %>%
                                    subset(Treatment == "A" & Outcome == "Recovered" & Sex == "Male
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(medicalRecords %>%
                                                subset(Treatment == "A" & Sex == "Male") %>%
                                                extract(, "Freq") %>%
                                                sum())%>%
                                    multiply_by(100),
                                 medicalRecords %>%
                                    subset(Treatment == "B" & Outcome == "Recovered" & Sex == "Male
                                    extract(, "Freq") %>%
                                    sum() %>%
                                    divide_by(medicalRecords %>%
                                                subset(Treatment == "B" & Sex == "Male") %>%
                                                extract(, "Freq") %>%
                                                sum())%>%
                                    multiply_by(100)))%>%
unique->
medRecoveryMale
medRecoveryMale
```

```
##   Treatment recoveryRate
## 1         A         37.5
## 5         B         42.5
```



- No, I would not recommend the same treatment as before. Previously, our analysis shows that treatment A has higher recovery rate. However, when we consider only the male patients, treatment B has a higher recovery rate.

i. *(1 mark)* Based on all analyses so far, if the patient's sex were unknown at the time treatment had to be given, which would you recommend? Why?

- Based on the all analyses so far, I would recommend Treatment B if the patient's sex were unknown. This is because when we consider the recovery rate separately for female and male, Treatment B has the higher recovery rate in both female and male.

d. **(12 marks)** Objections are now raised that all previous analysis failed to account for possible differences in age as well as sex. Older people might have different reactions than younger, and that too might depend upon their sex.

i. *(5 marks)* Again, beginning with `medicalRecords %>%`, form a pipeline which will produce a **single** data set having four variables, `Age`, `Sex`, `Treatment` and `recoveryRate` (in that order), which shows the **percentage** of patients (of every age and sex group) who recovered from each treatment.

- Save the result as `medRecoveryAll`
- Show your code and the value of `medRecoveryAll`

```
medicalRecords%>%
   group_by(Age,Sex,Treatment) %>%
   summarize(Outcome = Outcome, Freq = Freq, denom = sum(Freq)) %>%
   filter(Outcome == "Recovered") %>%
   summarize(Age = Age,
             Sex = Sex,
             Treatment = Treatment,
             recoveryRate = Freq*100/denom) ->
   medRecoveryAll
medRecoveryAll
```

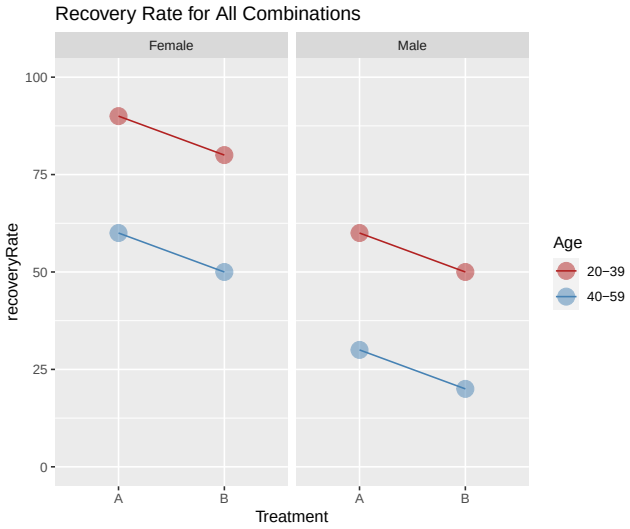
```
## # A tibble: 8 x 4
## # Groups:   Age, Sex [4]
##   Age   Sex    Treatment recoveryRate
##   <fct> <fct>  <fct>            <dbl>
## 1 20-39 Female A                   90
## 2 20-39 Female B                   80
## 3 20-39 Male   A                   60
## 4 20-39 Male   B                   50
## 5 40-59 Female A                   60
## 6 40-59 Female B                   50
## 7 40-59 Male   A                   30
## 8 40-59 Male   B                   20
```

i. *(4 marks)* Using `medRecoveryAll`, produce a facetted `ggplot` which

- shows a point of size 5 for every combination of `(Treatment, recoveryRate)`
- has a line connecting the two treatments for each group
- is grouped by `Age`
- is coloured by `Age`
- is facetted by `Sex`
- has vertical limits from 0 to 100
- has suitable title and labels

Show your code and your plot

```
ggplot(data = medRecoveryAll, mapping = aes(x = Treatment, y = recoveryRate, group = Age, colour = Age))
scale_colour_manual(values = c("firebrick", "steelblue")) +
geom_line() +
geom_point(mapping = aes(group = Age), alpha = 0.5, size = 5) +
ylim(0,100) +
facet_wrap(~ Sex) +
ggtitle("Recovery Rate for All Combinations")
```



i. *(1 mark)* What do you conclude about each of the following (and why)?

- the recovery rates of males versus females
- the recovery rates of young versus old
  - From the above faceted ggplot, we can easily see that females have much higher recoveryRate compare to males in any combination of Age and Treatment. And it is obvious that the young have a much higher recoveryRate compare to the old group in every combination of Treatment and Sex.

ii. *(2 marks)* Based on this more detailed analysis, which treatment would you now recommend? Would this change if you did not know the age group of the patient?

- Based on this more detailed analysis, I would now recommend Treatment A. This is because when we condition only on treatment and have the same Age group and Sex, we can see that Treatment A always perform better than Treatment B. This would not change if I did not know the age grouup of the patient because in the two different Age groups of males and females, Treatment A always perform better than Treatment B.

e. **(5 marks)** Instead of using pipelines, a multi-way *contingency* table can be constructed using `xtabs()` and the results plotted using `eikos()` from the eikosograms package.

i. *(2 marks)* Use `xtabs()` to construct the $2 \times 2 \times 2 \times 2$ table of counts for all combinations of `Age`, `Sex`, `Treatment`, and `Outcome`.

- save the result as `medRecordsTable`
- Show your code
- **Do not** show the resulting table, but **do show** its `dimnames()` and its `sum()`

```
medRecordsTable <- xtabs(Freq~Age+Sex+Treatment+Outcome, data=medicalRecords)
dimnames(medRecordsTable)
```

```
## $Age
## [1] "20-39" "40-59"
##
## $Sex
## [1] "Female" "Male"
##
## $Treatment
## [1] "A" "B"
##
## $Outcome
## [1] "Died"      "Recovered"
```

```
sum(medRecordsTable)
```

```
## [1] 3200
```

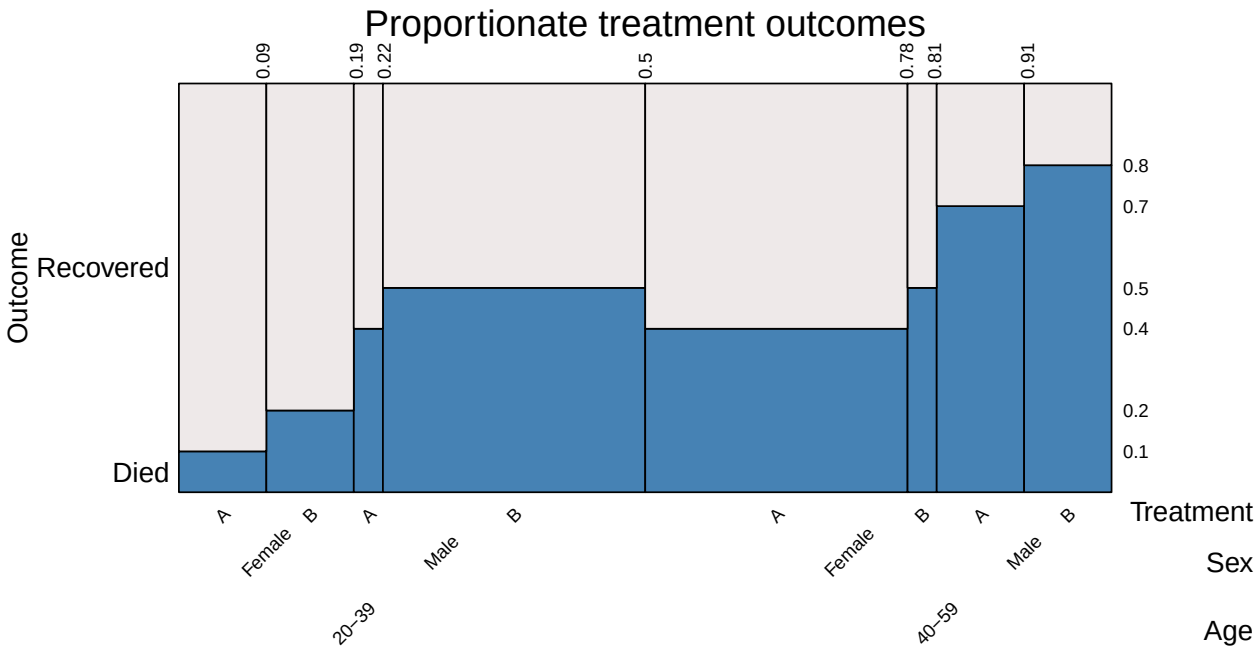
i. *(3 marks)* Using the multi-way table `medRecordsTable`, complete the following R `chunk` using the `RMarkdown` parameters (e.g. `fig.width`, etc.) appearing in the R `chunk` header (of course remove or change `eval = FALSE` to `eval = TRUE`) to show an eikosogram having

- `Outcome` as response and
- `Treatment`, `Sex`, and `Age` (in that order) as explanatory variates.

```
eikos(...,   # <- complete this part
      main = "Proportionate treatment outcomes",
      xlab_rot = 45, xvals_size = 8, lock_aspect = FALSE)
```

- Show your completed code.
- Which treatments are now suggested for each of the four groups: Young women, older women, young men, older men?
  - The treatments now suggested for all four groups will still be Treatment A, because in the eikosogram, treatment A still has the higher protion of recovered compare to treatment B. However, for female with age 40-59, there are too little cases to suggest that Treatment B is superior than treatment A

```
eikos(Outcome ~ Treatment + Sex + Age,    # <- complete this part
      data = medRecordsTable,
      main = "Proportionate treatment outcomes",
      xlab_rot = 45, xvals_size = 8, lock_aspect = FALSE)
```


Proportionate treatment outcomes
0.09
0.19
0.22
0.5
0.78
0.81
0.91
Outcome
Recovered
Died
0.8
0.7
0.5
0.4
0.2
0.1
A
B
A
B
A
B
A
B
Female
Male
Female
Male
20–39
40–59
Treatment
Sex
Age

f. **(2 marks)** The **Conclusions** stage.

i. *(1 mark)* Is it possible that a fourth (after `Treatment`, `Age`, and `Sex`) binary variate could be found which would reverse which treatment had the higher recovery rate? Or not?

- Yes, I think its possible that an addition of a fourth binary variate would reverse which treatment that have a higher recovery rate. Such as, one treatment may not be effective at all if a patient has a certain blood type, then it would change which treatment has a higher recovery rate

ii. *(1 mark)* What two principal recommendations would you have the health scientists undertake in the future to ensure that more reliable conclusions might be drawn about which treatment is superior?

- 1. Separate the patients into smaller categories because each categories may have different results.
  2. Plot the data and have a clearly visualization of the treatments.