Technical Test Quantitative Sciences

Mingzhu Ye

Part I Questions

1.When presented with a new dataset or database, what steps do you generally take to evaluate it prior to working with it?

1. Examine dataset's structure.
2. Estimate the dataset's scale.
3. Clearify relationsips of all the datasets.
4. Check tidiness of all datasets.
5. Summarize the datasets' semantics.

2.Based on the information provided above and the attached dataset, what 3 questions would you like to understand prior to conducting any analysis of the data?

1. What are the datasets about? By checking the dataset's info, we can know the datasets are about patient's information, orders and administration.
2. What columns are included in each dataset, and what are their meanings?
3. How are the datasets related? By checking the identifier, we can know that orders is related to administrations and patients. patients is related to demographics.

3.How would you prep the dataset provided for analysis? Please list steps taken and provide code used to prep the tables for analysis.

1. Read in the datasets.
2. Get a quick view of the datasets.
3. Examine the structures and dimensions of the datasets.
4. Check whether there are duplicated observations in each dataset.
5. Unify time formats in the datasets "order"","administration" and "patients". (Data processing code included in "code_for_partII.R")

Part II Questions

1. Average time elapsed between initial diagnosis and first treatment

Assumption: Observed from the merged table df2, there do exist cases where the first administered_date is before the initial diagnose date. Under such case, I will use the earlist administered_date after initial diagnose date for a patient to generate 'date_diff'

This is the quick view of merged results:

```
head(subdf2, 5)
```

```
##   patient_id date_diff  gender
## 1      5402v     17011    male
## 2     a3169y     27786    male
## 3     a6230g     18385  female
## 4     a6991u     11894 unknown
## 5     a7481m     26457  female
```

Conclusion:

The averaged time between a patient's initial diagnose and first treatment is 14025.56 days. And the average time differs in male and female, separately 12693.5 days and 15114.65 days.

2. Patients on drug nivolumab from 2012-2016

Conclusion:

There were totally 62 unique patients on drug "nibolumab" during year 2012-2016. While, some patients made multiple orders for the drug "nibolumab", so there were totally 100 unique orders for drug "nibolumab" during year 2012-2016.

3. Add new stratification to dataset demographics

Assumption: Since there are 4 patients without gender information, they couldn't be added risk levels using original stratification rules. New classification was inspired from the origial ones, with 'High_to_Medium_Risk','Medium_to_Low_Risk' being added.

This is the quick view of the demographics with stratification rules added:

```
head(demographics2,3)
```

```
##   patient_id gender age      race Cancer_Risk_Level
## 1     h9993d female  69 NON_WHITE         High_Risk
## 2     w9986a female  46     WHITE          Low_Risk
## 3     n9925d female  84 NON_WHITE         High_Risk
```

This is the categorical table with risk level as category:

```
table(demographics2$Cancer_Risk_Level)
```

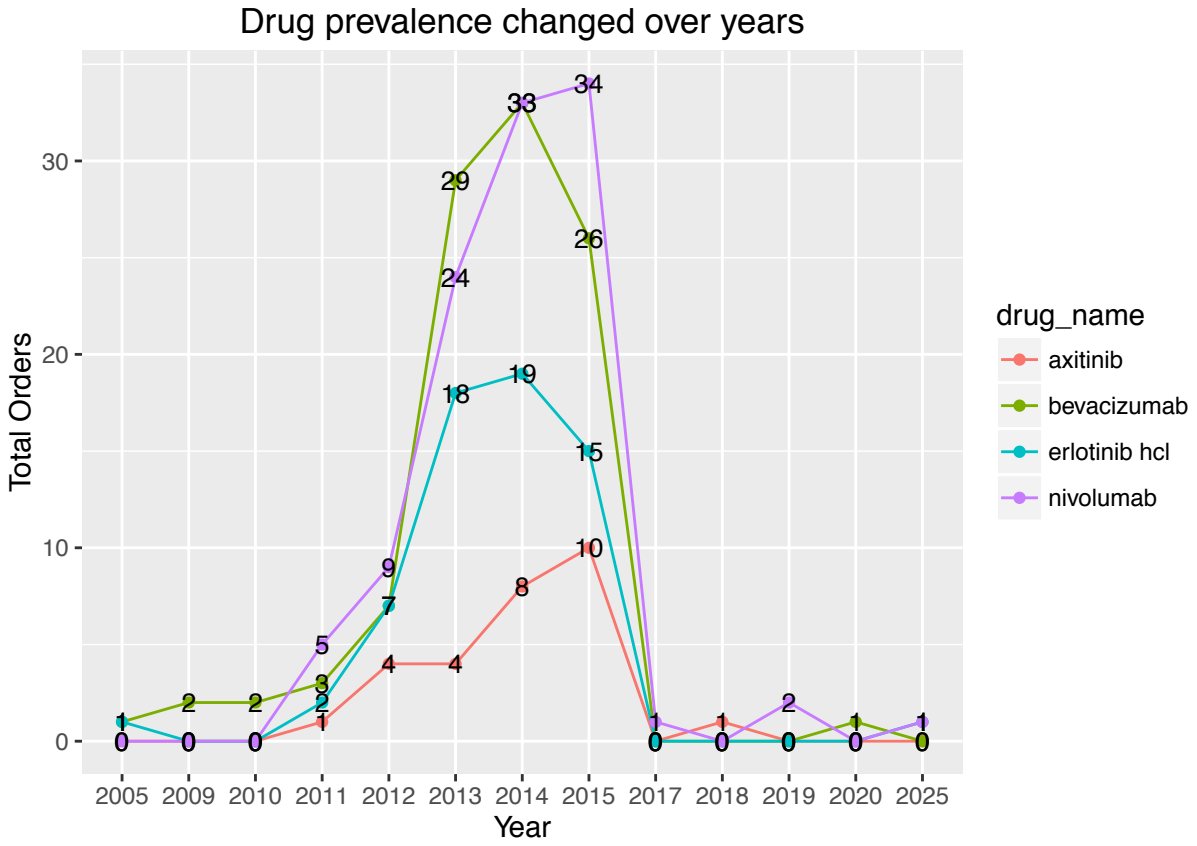
```
##
##           High_Risk High_to_Medium_Risk            Low_Risk
##                  33                   1                  40
##         Medium_Risk  Medium_to_Low_Risk
##                  24                   1
```

Conclusion:

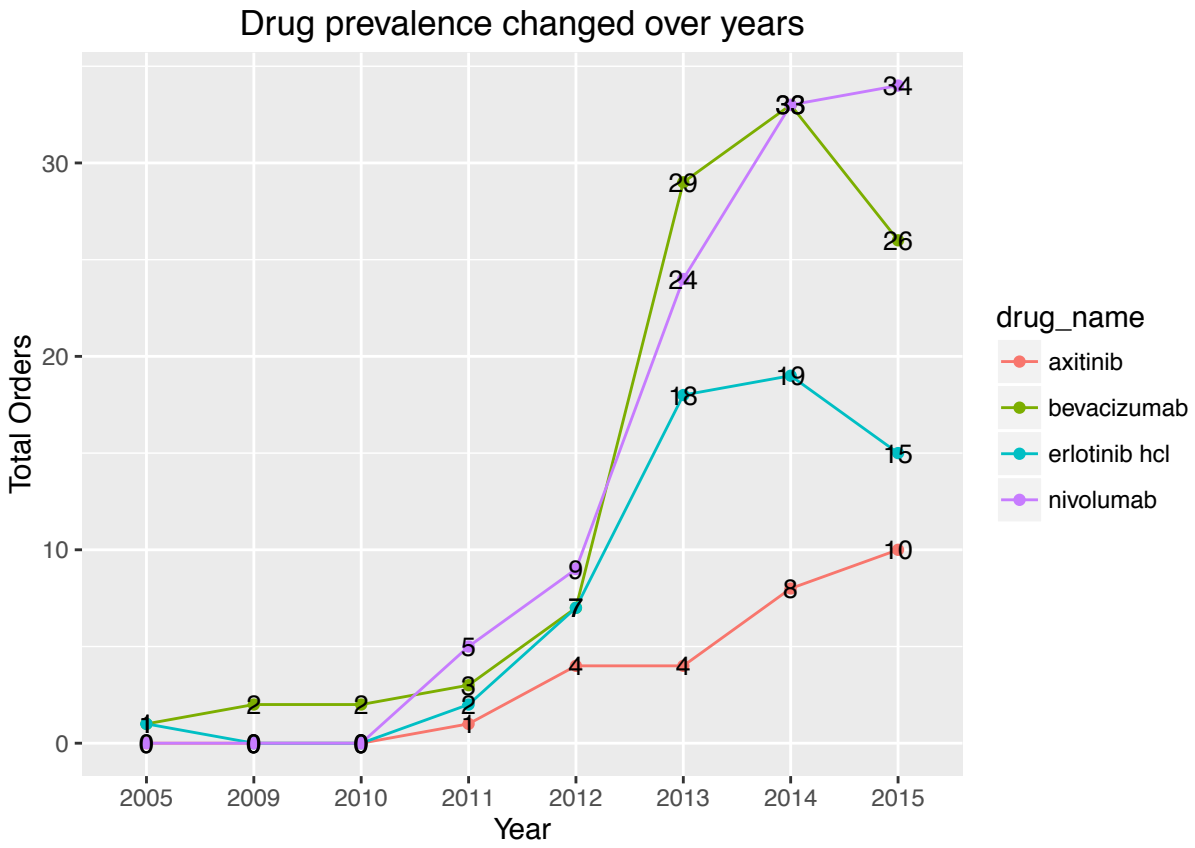
Among 99 unique patients, 40 are of Low_Risk, 33 are of High_Risk, 24 are of Medium_Risk, 1 is High_to_Medium_Risk and 1 is Medium_to_Low_Risk.

4. How drug prevalence has changed over time

Assumption 1:
Due to the small sample size, a ggplot of drug prevalence based on year has been created. This year-result may be more significant than results based on days or months.



Assumption 2:
As observed from plot1, there are drugs administered in the future, namely 2017, 2018, 2019, 2020 and 2025. Since the medical researcher may only want to look history data, therefore I also created a plot2 using drug data during 2005-2015.


Drug prevalence changed over years
Total Orders
Year
2005
2009
2010
2011
2012
2013
2014
2015
0
10
20
30
drug_name
axitinib
bevacizumab
erlotinib hcl
nivolumab