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 "orders"","administrations" and "patients".
   Note: Data processing code included in "code_for_partII.R"

Part II Questions

1. Average time elapsed between initial diagnosis and first treatment

Assumption: We assumed earliest administered_date as the patient's first treatment date. Observed from the merged table df2 (joined table administrations, patients and demographics), there're a few cases where the first administered_date is before the initial diagnosis_date. Under such cases, I will use the earlist administered_date, which must be after initial diagnosis_date, for a patient to generate a new column named 'date_diff'. The column 'date_diff' is days interval from initial diagnosis_date to earliest administered_date.

This is the quick view of merged results:

patient_id	date_diff	gender
5402v	17011	male
a3169y	27786	male
a6230g	18385	female
a6991u	11894	unknown
a7481m	26457	female

Conclusion:

The averaged time between a patient's initial diagnosis date and first treatment date is approximately 14025.56 days. The average time differs dramatically by gender, separately 12693.5 days in male and 15114.65 days in female.

2. Patients on drug nivolumab from 2012-2016

This is the quick view of manipulated drug 'nivolumab' information:

patient_id	external_patient_id	order_id	administered_date	drug_name	year
h9993d	70	r90118525421e	2015-08-07	nivolumab	2015
h9993d	70	o81913333209g	2014-01-08	nivolumab	2014
w9986a	46	l36534942738h	2014-12-15	nivolumab	2014
s9417f	29	s10845324909e	2014-02-20	nivolumab	2014
s9417f	29	h7336725593a	2015-04-09	nivolumab	2015

Conclusion:

There were totally 62 unique patients on drug "nivolumab" during year 2012-2016. While, some patients made multiple orders for the drug "nivolumab". So there were totally 100 unique orders for drug "nivolumab" 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. In stratification 'High_to_Medium_Risk', 'High' refers to the unknown gender is female, 'Medium' refers to the unknown gender is male. In stratification 'Medium_to_Low_Risk', 'Medium' refers to the unknown gender is female, 'Low' refers to the unknown gender is male.

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

patient_id	gender	age	race	Cancer_Risk_Level
h9993d	female	69	NON_WHITE	High_Risk
w9986a	female	46	WHITE	Low_Risk
n9925d	female	84	NON_WHITE	High_RIsk

This is the categorical table with risk level as category:

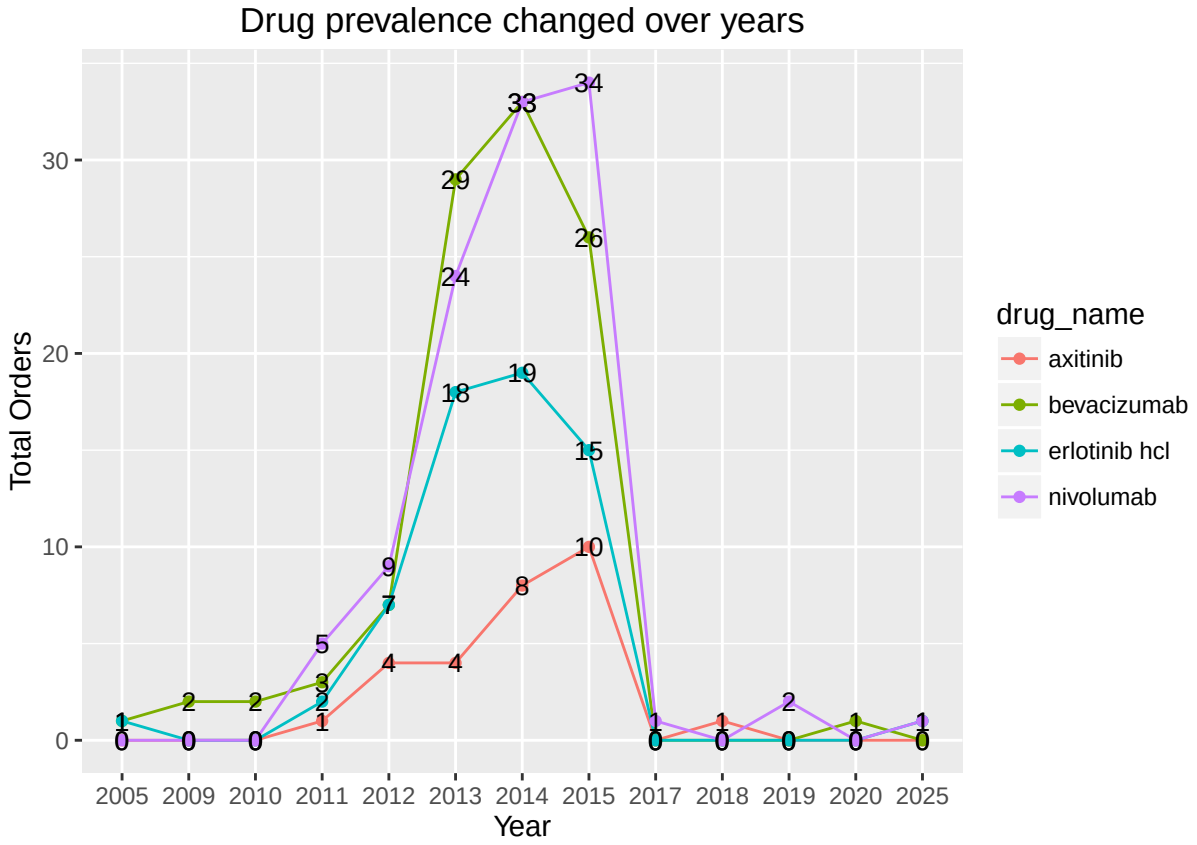
Cancer_Risk_Level	Frequency
Low_Risk	40
High_Risk	33
Medium_Risk	24
Medium_to_Low	1
High_to_Medium	1

Conclusion:

Among 99 unique patients (with 2 duplicated patients deleted from table 'demographics'), 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 even drugs administered in the future, namely 2017, 2018, 2019, 2020 and 2025. Since the medical researcher may only be interested in historical data, therefore I also created a ggplot using drug data during 2005-2015.