COVID-19 Virus: Analysis of Chararcteristics

Shiyu Shen

1 Introduction

1.1 Backgroud

When talking about 2020, "COVID-19", a novel coronavirus firstly found in Wuhan, China, in the end of 2019 (Mclntosh, 2020), must be the most popular word that appears on the Internet, newspapers and daily conversations. Since COVID-19 has been discovered and claimed as human pandemic by Chinese government, there are more than 36 millions of people being confirmed with COVID-19 and about 1 million deaths by now (2020). COVID-19 is definitely raising a huge concern of public health around the world.

1.2 Problem Statement

When facing a new virus, people usually care about its characteristics. It is clinically important to have a overview about the virus. Thus this report is going to analyze basic characterisistics of COVID-19 virus by addressing the following 3 questions: 1. What is the incubation period of CHOVID-19? 2. What are some of the most popular symptoms for COVID-19? 3. Is China the original place for COVID-19?

1.3 Data Discription

This data is obtained from Kaggle: Novel Corona Virus 2019 Data set. These COVID-19 case data are collected from 20th January to 28th February, 2020, which is the early pandemic period of COVID-19. This data set consists of 10 qualitative attributes, 7 quantitative attributes and 3 other attributes (Table1.1).

2 Methodology

The data set used in this report is downloaded from a Kaggle webpage. It offers day-level information and data set on COVID-19 affected cases. The first step to deal with this data set is to clean it. Specifically, the following method will be used:

1. If a row has all NA in most of the columns, it will be deleted;
2. Missing values in "location", "gender" and "country" will not be imputed;
3. Missing values in "age" will be generated using the mean of "age";
4. Missing values in "hosp visit date" will be imputed by it's "reporting date";
5. Missing value in "symptom onset date" will be imputed by the average of the difference between "symptom onset data" and "hosp visit date";
6. Missing value in "exposure start data" will be imputed by the average of the difference between "exposure start data" and "symptom onset date";
7. Missing value in "exposure end data" will be imputed by "hosp visit data".

After cleaning the data set, incubation period of COVID-19 will be approximate by computing the difference between symptom onset data and exposure start date. Then, word cloud technique will be used to analyze the symptoms of COVID-19. Last, number of new COVID-19 cases in countries outside of China will be computed and ranked. This will help to see how new cases is located around the world.

Table 1: Table1.1 Description of COVID-19 data set

Variable Name	Description
id	Patient ID
case_in_county	Number of case in county
reporting date	Case reporting date
summary	Patient summary discription
location	Location
country	Country
gender	Gender
age	Age
sympton_onset	Sympton onset date
if_onset_approximated	Whether symptom onset is approximated
hosp_visit_date	hospotal visiting date
exposure_start	Exposure start date
exposure_end	Exposure end date
visiting Wuhan	Whether the patient visted Wuhan
from Wuhan	Whether the patient is from Wuhan
death	Whether the patient died
recovered	Whether the patient recovered
symptom	Symptom
source	Source of the patient's information
link	Source link

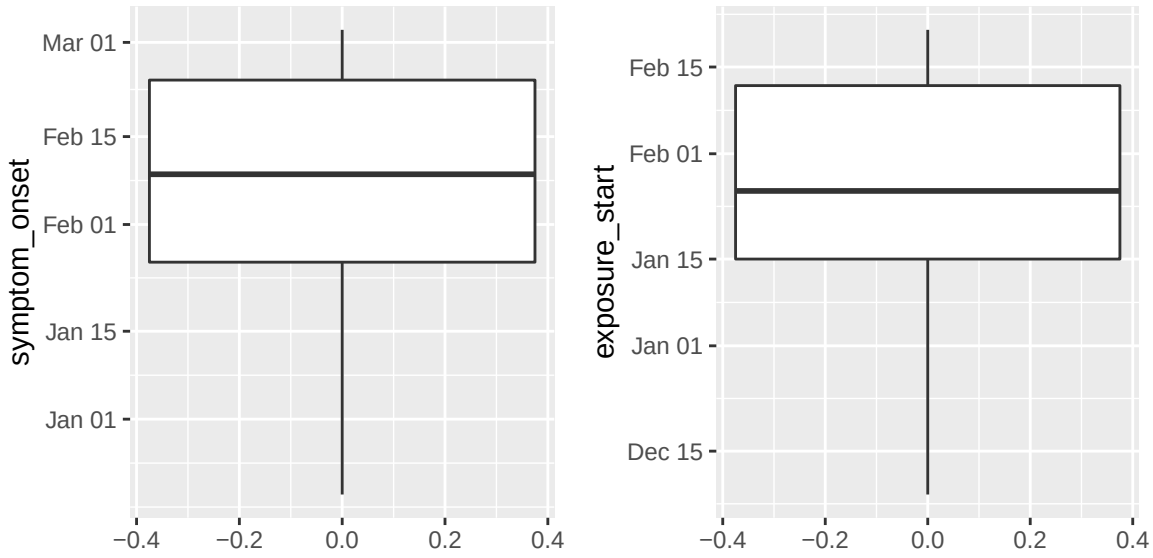
3 Results

In the beginning, 3 descriptive variables "summary", "source" and "link" will be excluded and the main data set will have 17 variables and 1085 observations.The following table listed some descriptive statistics about main variables in COVID-19 data set.

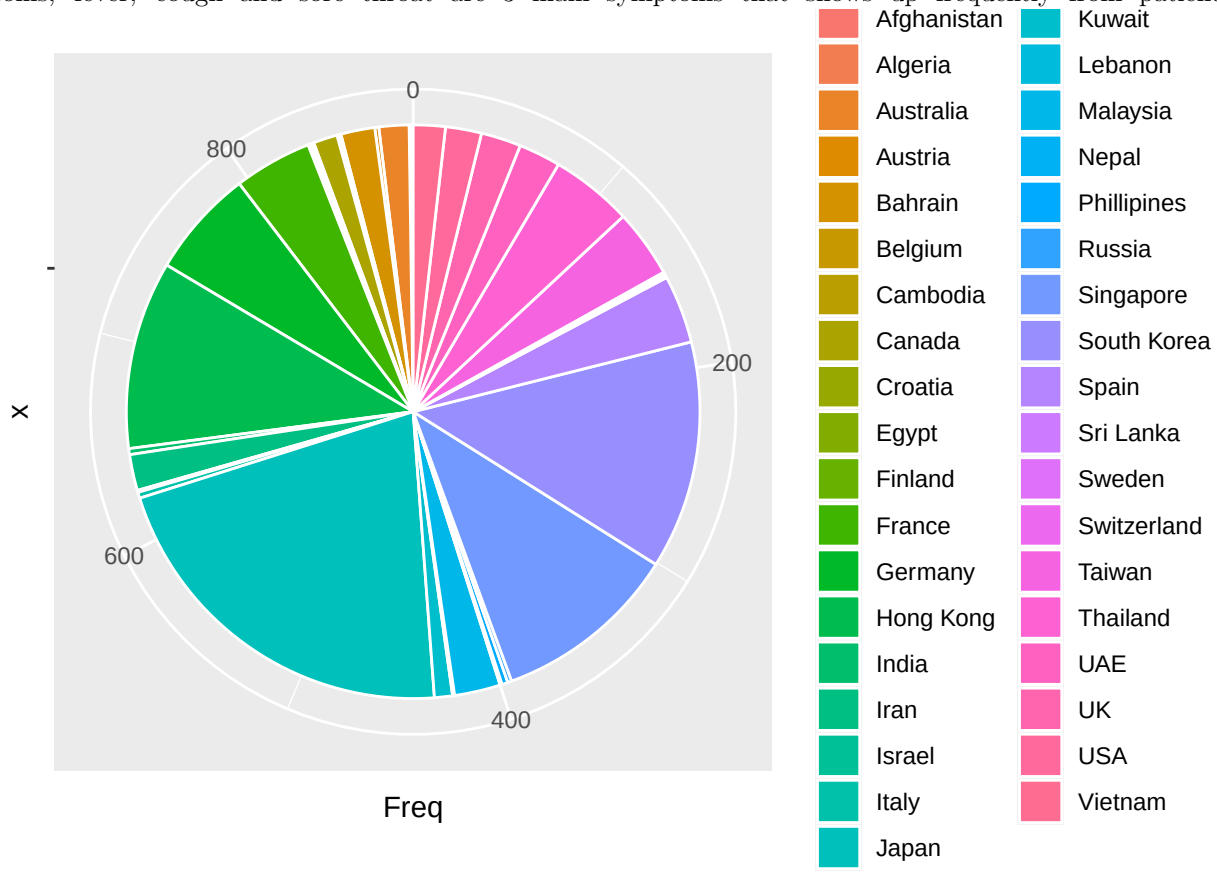
Using the cleaned data set, we draw a box plot to compare the medians of exposure start date and symptom onset date.

Table 2: Table3.1 Descriptive statistics of main variables in COVID-19 data set

Variable Name	Obs/frequency	#NA
country	38 countries	183
gender	F/M=382/520	242
age	Min:25 Max:96 Mean:49.48	222
if_onset_approximated	Yes:24 No:536	577
visiting Wuhan	Yes:192 No 893	
from Wuhan	Yes:156 No:925	4
death	Yes:63, No:1022	
recovered	Yes:159 No:926	
symptom		815



As you can see, the median date of the exposure start date for the patients are around Jan 26th; the median date of the exposure start date for the patients are around Feb 9th. The median difference is around 14 days, which tells us the incubation period for COVID-19 is around 14 days. Then, a pie chart of new COVID-19 cases is drawn. We can see that, Japan, Singapore, South Korean and Hong Kong composed large part of the new cases in countries outside of China. Last, a word cloud is drawn to show the symptoms collected from patients. Within those symptoms, fever, cough and sore throat are 3 main symptoms that shows up frequently from patients.



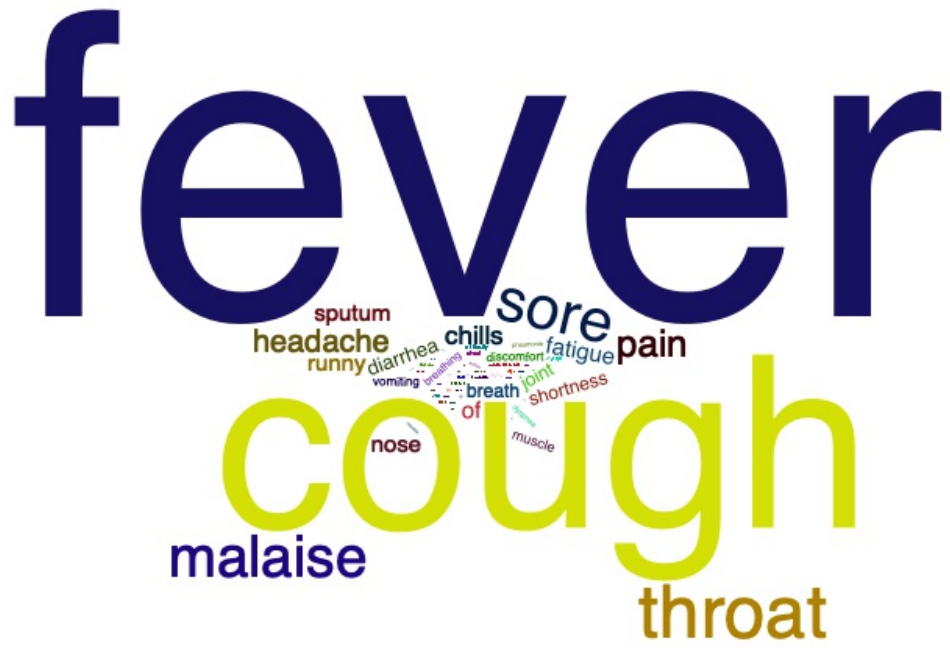


Figure 1: wordcloud

Conclusion

First, we found that the difference between the date of exposure and date of symptom onset is around 14 days. This results is consistent with the claimed incubation period of COVID-19. Second, fever and cough are the most frequent symptoms that shows up on the word cloud, which explained how these symptoms are considered as main symptoms of COVID-19 when inspecting the potential patient. Last, countries that we found having large percentage of new cases those countries who are closely located around and closely related with China, which indicate that China is the most possible original place for COVID-19 virus.