Text mining for exploration of COVID-19 severity factors

LAI Khang Duy
Mariia KLIMINA

Université Paris Cité
UFR des Sciences Fondamentales et Biomédicales

Université
Paris Cité

03-05-2022

Contents

1 Abstract

COVID-19 is the disease that cause by the Sar-COV-2 virus, commence in China at the end of the year 2019. Over the time, studies show that there is some form of background diseases and risk factors that can hugely affect the severity cases rate of COVID-19. This study will apply NLP as well as text mining and exploration on the CORD-19 dataset to extract which background diseases and risk factors play the main role for the problem, hence researcher may produce a method to reduce rate.

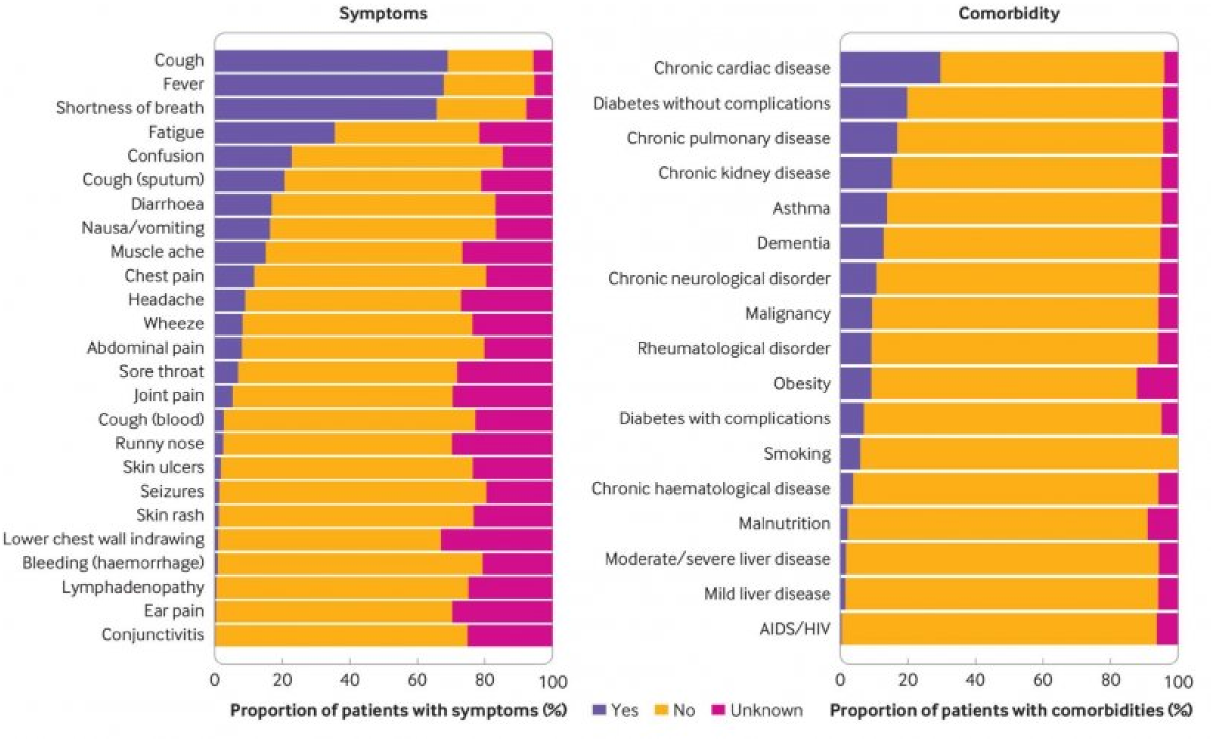

Figure 1: Cormobility and Symtom of COVID-19 cases

2 Introduction

- Explore with data
- Preprocessing
  - Reformat data json to dataframe
  - Remove all non-english paper
  - Tokenize
  - Remove stopword
  - Stemming

    - Lemmatisation

- Data selection
- Choose only paper that talk about risk factor and/or severity
- Apply NER (Named-entity recognition)

3 Data exploration

4 Preprocessing

something here

4.1 Handling multiple languages

4.2 Special characters and number remove

4.3 Tokenization

4.4 Stemming

4.5 Lemmatisation

5 Clustering

6 Extract medical term