Final Project - Covid-19 Analysis

Kesav Adithya Venkidusamy

10/29/2021

Introduction

Coronavirus disease or COVID-19 is a global pandemic infectious disease caused by virus called sars-cov-2. Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are most likely to develop serious complications from COVID-19 illness. The center for disease control and prevention (CDC) is the national public health agency of the United States. The agency's main goal is the protection of public health and safety through control and prevention of disease, injury, and disability in US and worldwide. CDC plays an essential role in the response to COVID-19. The agency collects the data on regular basis and provide for public use. Among numerous datasets available in CDC, below are the ones considered for analysis

1. Provisional COVID-19 deaths by sex and age
2. Provisional COVID-19 deaths by week, sex and age
3. Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21

The COVID-19 pandemic also pushed many companies to develop new vaccines to minimize the severity of symptoms. These vaccines were developed rapidly and underwent clinical trials rigorous enough to meet FDA (Food and Drug Administration) requirement for emergency use. The government played a role in monitoring the adverse reactions of these newly developed vaccines with Vaccine Adverse Event Report System (VAERS). VARES is co-managed by the Central for Disease Control and Prevention (CDC) and U.S Food and Drug Administration (FDA). VARES accepts reports from people who have received vaccines and experienced adverse effects or from heathcare providers who are required by law to report:

1. Any adverse event listed in the VARES table of reportable events following vaccination that occurs within the specified time period after vaccinations
2. An adverse event listed by the vaccine manufacturer as a contradiction to further doses of vaccine

VARES data is accessible by two mechanisms: by downloading raw data in comma-separated values (CSV) files for import into a database, spreadsheet or, by use of CDC WONDER online search tool. For this project, below datasets from VARES is considered.

1. VARESDATA.csv
2. VARESVAX.csv
3. VARESSYMPTOMS.csv

Research questions

Below are the questions will be explored as part of this project using CDC and VARES data sets.

1. Did elder people and those having underlying medical conditions are affected mostly by COVID-19?
2. Does the death caused by COVID-19 have any correlation with age of the person?
3. What role did age and sex play in COVID-19 infection?
4. Of the deaths occurred, what is the average age?
5. How many deaths occurred after vaccination?
6. Is there evidence of more vaccinated states reporting less deaths due to COVID-19?
7. Did vaccination cause any adverse effect leading to death?
8. How much is the death rate after first and second dose of vaccine?

Approach

My hypothesis is that elderly people and those who are having underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more infected by COVID-19 leading to death compared to other people. I am also trying to analyze the role played by vaccine in preventing deaths by COVID-19. This approach will look at several variables present across the dataset to determine causes of the deaths by covid. Looking at the variables like age, sex and those having preconditions will help to give a better understanding of those frequently affected by the virus which led to death. I would also want to analyze if vaccine plays any major role in preventing the infection by virus to elder people and those having preconditions. The variables like state from VARES dataset is used to analyze this scenario. This project is not intended to and will not be able to provide any sort of proof that other people are not infected by COVID-19 virus; This project compares the age and sex of people those are infected with virus. This report will not provide information on adverse effect caused by covid vaccine. This will just analyze if the vaccination has any correlation with infection or death

How your approach addresses (fully or partially) the problem

The problem will only be able to partially be addressed because the data is not conclusive proof one way or another. There are also ethical problems with using these data for concrete evidence in either direction. My intention is to observe and analyze the COVID-19 impacts and check if elderly people and those having preconditions are impacted more by the virus compared to other people based on the dataset available in CDC portal. In addition, I would also want to analyze if vaccine plays any role in preventing the spread of virus using dataset available in VARES.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

CDC Datasets

`Numerous Covid-19 related datasets are available for public use in CDC website. Those datasets feature`

Provisional COVID-19 deaths by week, sex and age + Data as of – Date of Analysis + State - Jurisdiction of occurrence + MMWR Week – MMWR week number + End Week - Last week-ending date of data period + Sex - Sex + Age Group - Age group + Total Deaths – Deaths from all causes of deaths + COVID-19 Deaths - Deaths Involving COVID-19

Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21 + Start Date - First week-ending date of data period + End Date - Last week-ending date of data period + Group - Time-period Indicator for record: by Month, by Year, Total + State - Jurisdiction of occurrence + Condition - Condition contributing to deaths involving COVID-19 + Age Group - Age group + COVID-19 Deaths - COVID 19 Deaths

VARES Dataset

VARES data are distributed in three data sets, VARESVAX, VARESDATA and VARESSYMPTOMS. Data sets belong to year 2020 and 2021 will be used for this project. The code book for this data set is available in the below link

Code Book

```
covid19_week <- read.csv("Provisional_COVID-19_Deaths_by_Week_Sex_and_Age.csv")
covid19_condition <- read.csv("COVID-19_Deaths_by_State_and_Age.csv")
data20 <- read.csv("2020VAERSDATA.csv")
data21 <- read.csv("2021VAERSDATA.csv")
symptoms20 <- read.csv("2020VAERSSYMPTOMS.csv")
symptoms21 <- read.csv("2021VAERSSYMPTOMS.csv")

print(dim(covid19_week))
```

```
## [1] 3276    8
```

```
print(dim(covid19_condition))
```

```
## [1] 310500     14
```

```
print(dim(data20))
```

```
## [1] 50204    35
```

```
print(dim(data21))
```

```
## [1] 624237     35
```

```
print(dim(symptoms20))
```

```
## [1] 61565    11
```

```
print(dim(symptoms21))
```

```
## [1] 834197     11
```

Required Packages

1. Readxl
2. Dplyr
3. Broom
4. Stringr
5. ggplot2
6. pastecs
7. GGally
8. Scales
9. CoefPlot

Plots and Table Needs

1. Histogram - Look for Normality
2. Scatterplots - Identify relationships
3. Residual Plots - Look for outliers
4. Density Plot - Observe distributions
5. Box Plot - Look for outliers
6. Tables
    - Covid19 weekly count by state
    - Preconditions
    - Vaccines

Questions for future steps

I want to do research on how age and pre-existing conditions plays a major role in count of Covid-19 deaths. Also, I would want to analyze if vaccination plays any role in minimizing or preventing death due to covid-19.