Effectiveness of COVID-19 Vaccines in the US

December 10, 2024

This project provides an analysis of the effectiveness of COVID-19 vaccines in reducing the number of infections and deaths. The analysis is based on data from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

Install and load R packages

```
knitr::opts_chunk$set(echo = TRUE)

if (!requireNamespace("pacman", quietly = TRUE)) {
  install.packages("pacman")
}

library(pacman)
p_load(tidyverse,ggplot2, scales, nlme, texreg)
```

Data preparation

The data are reshaped and cleaned, restricting the dataset to the 50 states and the District of Columbia. Observations with missing values in key variables, such as the date, are also excluded. Since the analysis is at the national level, the key variables are infection cases and deaths.

```
US_cases <- US_cases %>%
  pivot_longer(cols= -(UID: Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  rename(state = Province_State, county = Admin2) %>%
  mutate(date=mdy(date),
         month = format(date, "%Y-%m")) %>%
  filter(!is.na(date)) %>%
  filter(!state %in% c("American Samoa","Diamond Princess","Grand Princess","Guam",
                       "Northern Mariana Islands","Puerto Rico","Virgin Islands")) %>%
  select(-c(Lat, Long_))


US_deaths <- US_deaths %>%
  pivot_longer(cols= -(UID: Combined_Key),
               names_to = "date",
               values_to = "deaths") %>%
  select(Admin2:deaths) %>%
```

```
  rename(state = Province_State, county = Admin2) %>%
  mutate(date=mdy(date),
         month = format(date, "%Y-%m")) %>%
  filter(!is.na(date)) %>%
  filter(!state %in% c("American Samoa","Diamond Princess","Grand Princess","Guam",
                       "Northern Mariana Islands","Puerto Rico","Virgin Islands")) %>%
  select(-c(Lat, Long_))


US_data <- US_cases %>% full_join(US_deaths)
rm(US_cases, US_deaths)

US_data <- US_data %>% select(Combined_Key,Country_Region, state, county, month, date, everything())

US_data <- US_data[order(US_data$date),]
```

Descriptive summary

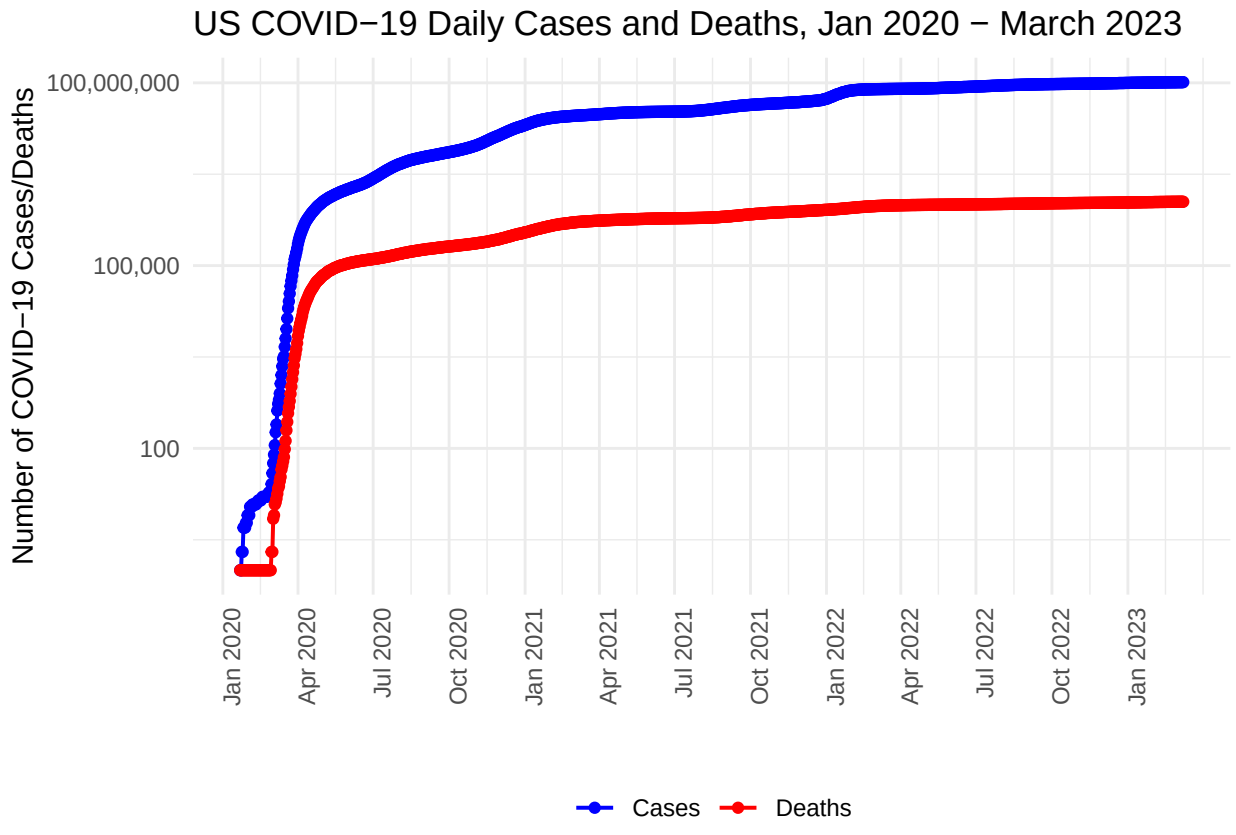
```
US_by_date <- US_data %>%
  group_by(date) %>%
  summarise(cases = sum(cases), deaths = sum(deaths)) %>%
  mutate(new_deaths = deaths - lag(deaths),
         new_cases = cases - lag(cases))

US_by_state<- US_data %>%
  group_by(state, date) %>%
  summarise(cases = sum(cases), deaths = sum(deaths)) %>%
  mutate(new_deaths = deaths - lag(deaths),
         new_cases = cases - lag(cases))

#US_by_date %>% slice_max(new_cases) %>% select(date, cases, new_cases)
#US_by_date %>% slice_max(new_deaths)  %>% select(date, deaths, new_deaths)
#US_by_state %>% slice_max(new_cases) %>% select(state, date, cases, new_cases)
#US_by_state %>% slice_max(new_deaths) %>% select(state, date, deaths, new_deaths)

summary(US_by_date)
```

```
##       date                cases               deaths          new_deaths
##  Min.   :2020-01-22   Min.   :        1   Min.   :      1   Min.   :-264.0
##  1st Qu.:2020-11-02   1st Qu.:  9359922   1st Qu.: 231618   1st Qu.: 321.2
##  Median :2021-08-15   Median : 36672979   Median : 615173   Median : 703.5
##  Mean   :2021-08-15   Mean   : 46690024   Mean   : 621520   Mean   : 978.4
##  3rd Qu.:2022-05-27   3rd Qu.: 83337084   3rd Qu.:1001733   3rd Qu.:1408.2
##  Max.   :2023-03-09   Max.   :102593255   Max.   :1117385   Max.   :4372.0
##                                                             NA's   :1
##    new_cases
##  Min.   :  -4260
##  1st Qu.:  25442
##  Median :  55566
##  Mean   :  89837
##  3rd Qu.: 111740
```

```
##  Max.   :1347188
##  NA's   :1
```

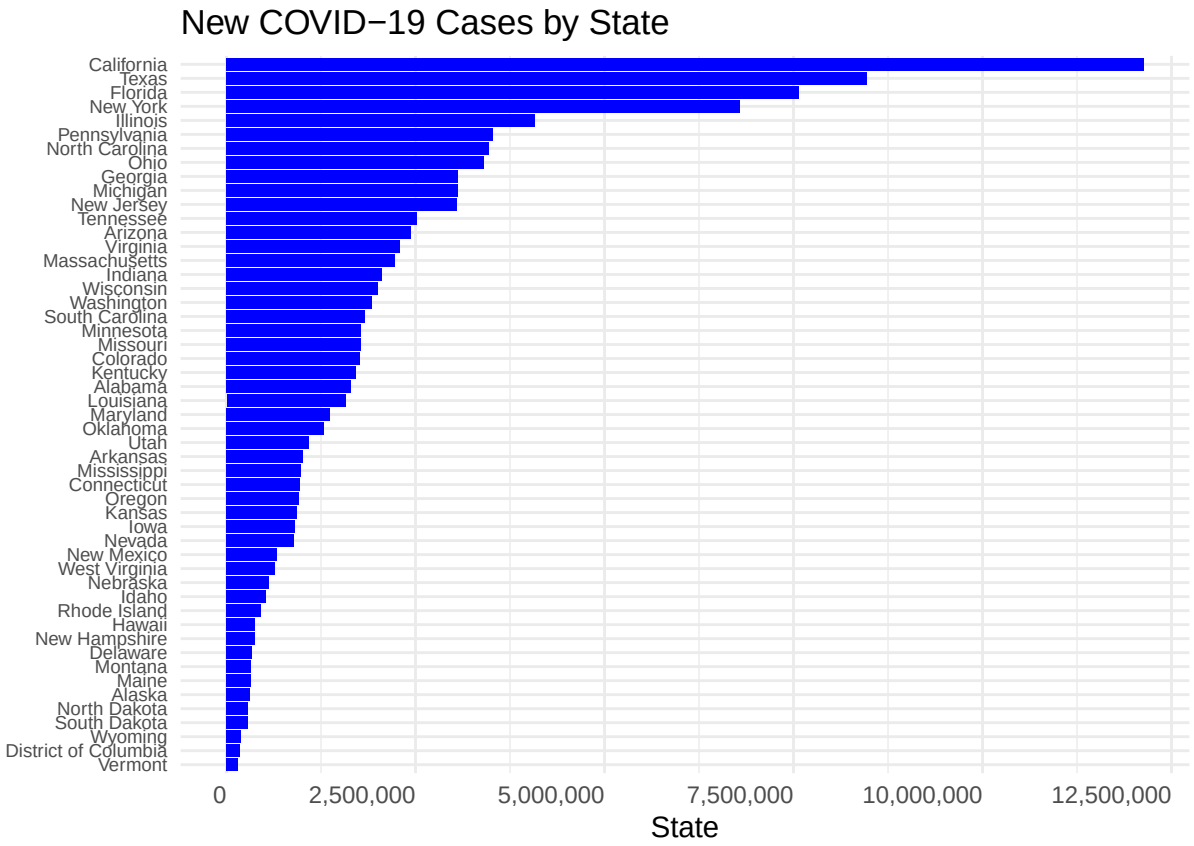
```
summary(US_by_state)
```

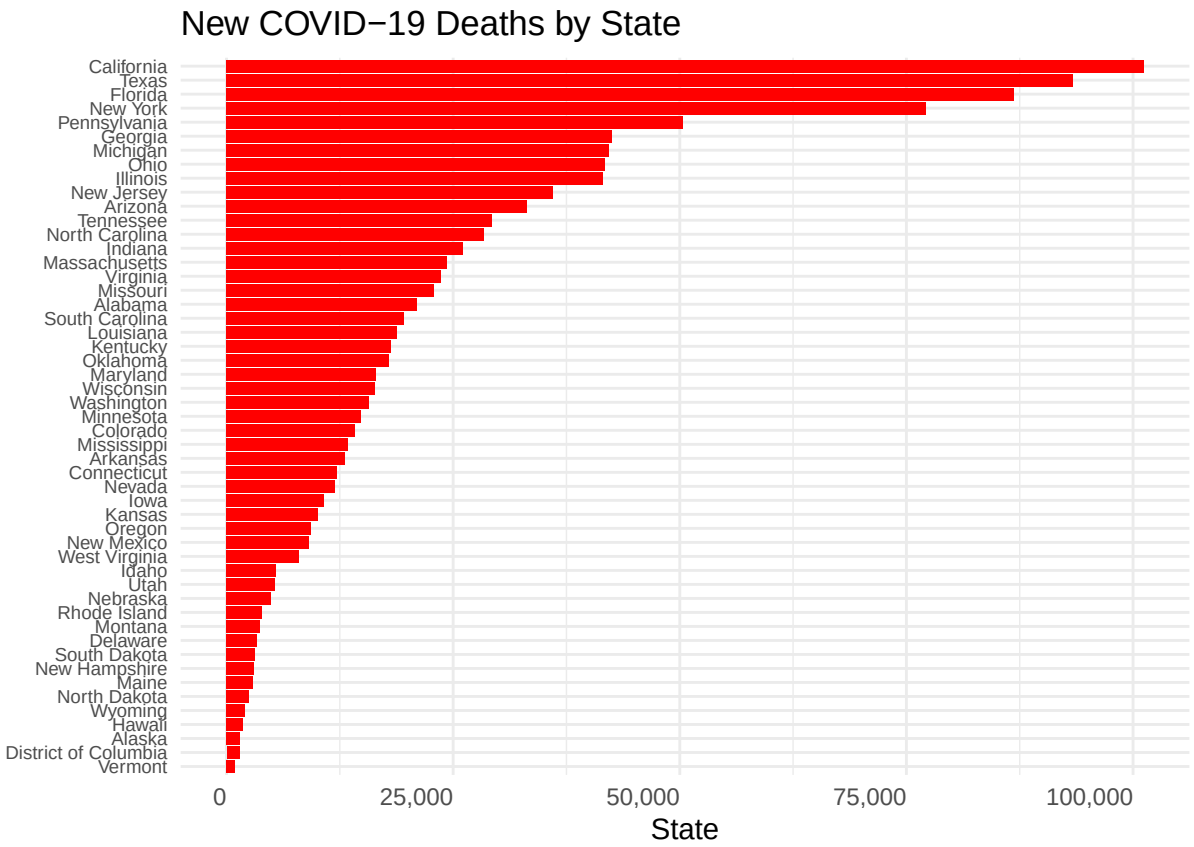
```
##     state               date                cases              deaths
##  Length:58293       Min.   :2020-01-22   Min.   :       0   Min.   :     0
##  Class :character   1st Qu.:2020-11-02   1st Qu.:   89838   1st Qu.:  1364
##  Mode  :character   Median :2021-08-15   Median :  375278   Median :  5554
##                     Mean   :2021-08-15   Mean   :  915491   Mean   : 12187
##                     3rd Qu.:2022-05-28   3rd Qu.: 1084488   3rd Qu.: 15522
##                     Max.   :2023-03-09   Max.   :12129699   Max.   :101159
##
##    new_deaths         new_cases
##  Min.   :-704.00   Min.   :-27000
##  1st Qu.:   0.00   1st Qu.:     0
##  Median :   3.00   Median :   339
##  Mean   :  19.19   Mean   :  1762
##  3rd Qu.:  17.00   3rd Qu.:  1462
##  Max.   :2441.00   Max.   :207110
##  NA's   :51        NA's   :51
```



As can be seen from the line graph, the number of COVID-19 cases and deaths sharply rose between January 2020 and July 2020. From July 2020 to April 2021, the rates continued to rise but at a slower pace, and largely leveled off through much of 2022. Despite the actual number of deaths, the fatality rate remained

relatively low. A notable exception is during the early stages of the pandemic when it was significantly higher.





On the other hand, the bar graph shows that the top three states for both new COVID-19 infections and deaths were California, Texas, and Florida. The least affected states for infections were the District of Columbia, Wyoming, and Vermont. However, for deaths, Wyoming is replaced by Alaska. It is important to note that these rankings reflect actual numbers, not proportions. Hence, these rankings partly reflect the population size of these states. As a side note, since the analysis pertains to human lives, presenting the actual numbers rather than only proportions is equally important.

Interrupted Time Series Model

The ideal approach to testing the effectiveness of COVID-19 vaccines is to use data from randomized controlled trials. In the absence of such data, the appropriate method is to use non-experimental data while controlling for confounding factors. In this context, an interrupted time series approach is applied. This approach fits the purpose at hand for at least two reasons. First, the data contain known date of policy change (December 14, 2020, when the COVID-19 vaccines became available). Second, because the data provide large number of observations before and after the vaccine rollout, it is possible to credibly establish the counterfactual (that is, the number of cases and deaths that would have occurred without the introduction of the vaccines) to estimate the vaccine efficacy. The estimation equation is given by:

$$y_t = \alpha_0 + \alpha_1 Time + \alpha_2 PostVaccineRollout + \alpha_3 TimeSinceVaccineRollout + e_t$$

where y_t is the number of new cases or deaths, $Time$ is number of days, $PostVaccineRollout$ is binary variable taking a value of 1 if the date is on or after December 14, 2020, and 0 otherwise, $TimeSinceVaccineRollout$ is the number of days since the vaccines rollout, and y_t is error term. The model is estimated using generalized least squares technique.

##

```
## ============================================================
##                                New Cases      New Deaths
## ------------------------------------------------------------
## Intercept                     -22522.49        443.53 ***
##                               (13343.30)       (91.27)
## Time (days)                      444.88 ***      2.94 ***
##                                  (70.73)        (0.48)
## Post-Vaccination                13259.76        407.30 ***
##                               (15770.76)      (107.88)
## Time Since Vaccination (days)    -519.85 ***     -4.94 ***
##                                  (72.95)        (0.50)
## ------------------------------------------------------------
## AIC                             29962.27      18576.73
## BIC                             29987.47      18601.93
## Log Likelihood                 -14976.13      -9283.36
## Num. obs.                        1142          1142
## ============================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The estimates suggest that vaccines are effective. Specifically, the time variable (444.88) is positive and significant, indicating that the number of new cases tends to increase over time. The post vaccine rollout variable is not significant, implying no immediate change in new cases following the rollout. Regarding the number of days since vaccination, we see that as time since vaccination increases, new cases decrease by 519.85, which is statistically significant.

In the case of new deaths, after the vaccine rollout, the number of deaths increases by 407.30, which might partly reflect time lag between infection and death. The longer the time since vaccine rollout, the lower the number of deaths (-4.94), which is highly significant. These results warrant careful interpretation, especially considering potential biases in data records due to the difficulty of accurately identifying the causes of deaths.

Conclusion

Overall, the estimates suggest that the vaccines helped lower the number of cases and deaths. However, there are important caveats that call for caution and further analysis. First, establishing causality is difficult when the data are not generated from randomized control trials. Therefore, it is necessary to control for confounders such as the timing of different variants like Delta and Omicron, as well as changes in testing rates, social distancing and lockdown measures, healthcare capacity, and similar factors.

Second, there may be potential sources of bias due to incomplete or inaccurate data. In particular, extra caution is needed when defining COVID-related deaths; that is, whether they are deaths caused by COVID or deaths of individuals with COVID. Additionally, biases may arise from personal beliefs about vaccines. For instance, some might interpret the results as both statistically significant and economically large, while others may view them as too small.

Additional material

GitHub: https://github.com/kgmaysd/COVID-19