Predict the efficiency of COVID-19 vaccinates around the world

Student: Angela Amador

TMU Student Number: 500259095

Supervisor: Tamer Abdou, PhD

Submission Date: Oct 16th, 2023

**Table of Contents**

[**Abstract 2**](#_heading=h.30j0zll)

[Literature Review 2](#_heading=h.3znysh7)

[Predicting the incidence of COVID-19 using data mining 2](#_heading=h.usjlvgxlmtb)

[Article 2 2](#_heading=h.mfnzcaubntsx)

[Article 3 2](#_heading=h.isviujlh9q77)

[Article 3 3](#_heading=h.mojrfkornyci)

[Article 5 3](#_heading=h.uw6p3gxmilxy)

[Article 6 3](#_heading=h.6lh3zokadmyr)

[**Exploration Data Analysis 3**](#_heading=h.3vkygbp5fn5s)

[Dataset 3](#_heading=h.w88kulit1yoo)

[Data Dictionary 3](#_heading=h.ot5fgaf92fz2)

[Metadata 5](#_heading=h.3b23vrqpjytm)

[Heading 4 5](#_heading=h.j9nhshl4a376)

[Heading 4 5](#_heading=h.hnktu5wp95ca)

[Heading 4 5](#_heading=h.6lsdxlycu5d6)

[Heading 4 5](#_heading=h.tihrxjrl5l7b)

[Heading 4 6](#_heading=h.bgrzamdzk1wf)

[Sample Chart 6](#_heading=h.tyjcwt)

[Sample Table 6](#_heading=h.3dy6vkm)

# Abstract

The pandemic of the COVID-19 extended from China across the world. It not only shutdown the world as we knew it but also challenged the scientific community to find a vaccine to control the virus. In a record time, laboratories around the world rushed to find a solution to control the virus and save lives. In this study, I aim to predict the efficiency of COVID-19 vaccinates around the world.

# Literature Review

Several studies have been conducted to analyze and evaluate the effectiveness of COVID-19 vaccine to control the pandemic and help humanity to overcome the pandemic.

## Predicting the incidence of COVID-19 using data mining

Sample body text.

## Article 2

Sample body text.

## Article 3

Sample body text.

## Article 3

Sample body text.

## Article 5

Sample body text.

## Article 6

Sample body text.

# Exploration Data Analysis

## Dataset

The dataset, provided by Our World in Data, provides COVID-19 vaccination information collected by **Our World in Data** available to **Kaggle community** <https://www.kaggle.com/datasets/caesarmario/our-world-in-data-covid19-dataset/download?datasetVersionNumber=418>. This dataset is updated daily, for the purpose of this study I am analyzing the data with information up to Oct 7th, 2023.

The dataset is a csv file, comma delimited with 67 variables and 346,567 observations.

## Data Dictionary

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 346567 entries, 0 to 346566

Data columns (total 67 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 iso\_code 346567 non-null object

1 continent 330089 non-null object

2 location 346567 non-null object

3 date 346567 non-null object

4 total\_cases 308672 non-null float64

5 new\_cases 337028 non-null float64

6 new\_cases\_smoothed 335769 non-null float64

7 total\_deaths 287169 non-null float64

8 new\_deaths 337072 non-null float64

9 new\_deaths\_smoothed 335842 non-null float64

10 total\_cases\_per\_million 308672 non-null float64

11 new\_cases\_per\_million 337028 non-null float64

12 new\_cases\_smoothed\_per\_million 335769 non-null float64

13 total\_deaths\_per\_million 287169 non-null float64

14 new\_deaths\_per\_million 337072 non-null float64

15 new\_deaths\_smoothed\_per\_million 335842 non-null float64

16 reproduction\_rate 184817 non-null float64

17 icu\_patients 37509 non-null float64

18 icu\_patients\_per\_million 37509 non-null float64

19 hosp\_patients 38759 non-null float64

20 hosp\_patients\_per\_million 38759 non-null float64

21 weekly\_icu\_admissions 10160 non-null float64

22 weekly\_icu\_admissions\_per\_million 10160 non-null float64

23 weekly\_hosp\_admissions 23145 non-null float64

24 weekly\_hosp\_admissions\_per\_million 23145 non-null float64

25 total\_tests 79387 non-null float64

26 new\_tests 75403 non-null float64

27 total\_tests\_per\_thousand 79387 non-null float64

28 new\_tests\_per\_thousand 75403 non-null float64

29 new\_tests\_smoothed 103965 non-null float64

30 new\_tests\_smoothed\_per\_thousand 103965 non-null float64

31 positive\_rate 95927 non-null float64

32 tests\_per\_case 94348 non-null float64

33 tests\_units 106788 non-null object

34 total\_vaccinations 78953 non-null float64

35 people\_vaccinated 75575 non-null float64

36 people\_fully\_vaccinated 72224 non-null float64

37 total\_boosters 47234 non-null float64

38 new\_vaccinations 65019 non-null float64

39 new\_vaccinations\_smoothed 180079 non-null float64

40 total\_vaccinations\_per\_hundred 78953 non-null float64

41 people\_vaccinated\_per\_hundred 75575 non-null float64

42 people\_fully\_vaccinated\_per\_hundred 72224 non-null float64

43 total\_boosters\_per\_hundred 47234 non-null float64

44 new\_vaccinations\_smoothed\_per\_million 180079 non-null float64

45 new\_people\_vaccinated\_smoothed 179887 non-null float64

46 new\_people\_vaccinated\_smoothed\_per\_hundred 179887 non-null float64

47 stringency\_index 197651 non-null float64

48 population\_density 294167 non-null float64

49 median\_age 273580 non-null float64

50 aged\_65\_older 264005 non-null float64

51 aged\_70\_older 270838 non-null float64

52 gdp\_per\_capita 268118 non-null float64

53 extreme\_poverty 172778 non-null float64

54 cardiovasc\_death\_rate 268731 non-null float64

55 diabetes\_prevalence 282404 non-null float64

56 female\_smokers 201575 non-null float64

57 male\_smokers 198833 non-null float64

58 handwashing\_facilities 131627 non-null float64

59 hospital\_beds\_per\_thousand 237221 non-null float64

60 life\_expectancy 318823 non-null float64

61 human\_development\_index 260466 non-null float64

62 population 346567 non-null float64

63 excess\_mortality\_cumulative\_absolute 11953 non-null float64

64 excess\_mortality\_cumulative 11953 non-null float64

65 excess\_mortality 11953 non-null float64

66 excess\_mortality\_cumulative\_per\_million 11953 non-null float64

dtypes: float64(62), object(5)

memory usage: 177.2+ MB

## Metadata

The dataset size is 91.1 MB, the Pandas data profiling is almost 300 MB. I have added a column called Dropped to identify the variables that I will be removing from the dataset for the analysis.

Below is the metadata of the whole dataset.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Dropped** |
| iso\_code | ISO 3166-1 alpha-3 – three-letter country codes. Note that OWID-defined regions (e.g. continents like 'Europe') contain prefix 'OWID\_'. | Y |
| continent | Continent of the geographical location | N |
| location | Geographical location | N |
| date | Date of observation | N |
| total\_cases | Total confirmed cases of COVID-19. Counts can include probable cases, where reported. | N |
| new\_cases | New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. | N |
| new\_cases\_smoothed | New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported. | Y |
| total\_deaths | Total deaths attributed to COVID-19. Counts can include probable deaths, where reported. | N |
| new\_deaths | New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. | N |
| new\_deaths\_smoothed | New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported. | Y |
| total\_cases\_per\_million | Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported. | Y |
| new\_cases\_per\_million | New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported. | Y |
| new\_cases\_smoothed\_per\_million | New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported. | Y |
| total\_deaths\_per\_million | Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported. | Y |
| new\_deaths\_per\_million | New deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported. | Y |
| new\_deaths\_smoothed\_per\_million | New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable deaths, where reported. | Y |
| reproduction\_rate | Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database | Y |
| icu\_patients | Number of COVID-19 patients in intensive care units (ICUs) on a given day | N |
| icu\_patients\_per\_million | Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people | Y |
| hosp\_patients | Number of COVID-19 patients in hospital on a given day | N |
| hosp\_patients\_per\_million | Number of COVID-19 patients in hospital on a given day per 1,000,000 people | Y |
| weekly\_icu\_admissions | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week (reporting date and the preceeding 6 days) | Y |
| weekly\_icu\_admissions\_per\_million | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people (reporting date and the preceeding 6 days) | Y |
| weekly\_hosp\_admissions | Number of COVID-19 patients newly admitted to hospitals in a given week (reporting date and the preceeding 6 days) | Y |
| weekly\_hosp\_admissions\_per\_million | Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people (reporting date and the preceeding 6 days) | Y |
| total\_tests | Total tests for COVID-19 | N |
| new\_tests | New tests for COVID-19 (only calculated for consecutive days) | N |
| total\_tests\_per\_thousand | Total tests for COVID-19 per 1,000 people | Y |
| new\_tests\_per\_thousand | New tests for COVID-19 per 1,000 people | Y |
| new\_tests\_smoothed | New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window | Y |
| new\_tests\_smoothed\_per\_thousand | New tests for COVID-19 (7-day smoothed) per 1,000 people | Y |
| positive\_rate | The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests\_per\_case) | N |
| tests\_per\_case | Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive\_rate) | N |
| tests\_units | Units used by the location to report its testing data. A country file can't contain mixed units. All metrics concerning testing data use the specified test unit. Valid units are 'people tested' (number of people tested), 'tests performed' (number of tests performed. a single person can be tested more than once in a given day) and 'samples tested' (number of samples tested. In some cases, more than one sample may be required to perform a given test.) | N |
| total\_vaccinations | Total number of COVID-19 vaccination doses administered | N |
| people\_vaccinated | Total number of people who received at least one vaccine dose | N |
| people\_fully\_vaccinated | Total number of people who received all doses prescribed by the initial vaccination protocol | N |
| total\_boosters | Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol) | N |
| new\_vaccinations | New COVID-19 vaccination doses administered (only calculated for consecutive days) | N |
| new\_vaccinations\_smoothed | New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window | Y |
| total\_vaccinations\_per\_hundred | Total number of COVID-19 vaccination doses administered per 100 people in the total population | Y |
| people\_vaccinated\_per\_hundred | Total number of people who received at least one vaccine dose per 100 people in the total population | Y |
| people\_fully\_vaccinated\_per\_hundred | Total number of people who received all doses prescribed by the initial vaccination protocol per 100 people in the total population | Y |
| total\_boosters\_per\_hundred | Total number of COVID-19 vaccination booster doses administered per 100 people in the total population | Y |
| new\_vaccinations\_smoothed\_per\_million | New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population | Y |
| new\_people\_vaccinated\_smoothed | Daily number of people receiving their first vaccine dose (7-day smoothed) | Y |
| new\_people\_vaccinated\_smoothed\_per\_hundred | Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population | Y |
| stringency\_index | Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) | Y |
| population\_density | Number of people divided by land area, measured in square kilometers, most recent year available | Y |
| median\_age | Median age of the population, UN projection for 2020 | N |
| aged\_65\_older | Share of the population that is 65 years and older, most recent year available | N |
| aged\_70\_older | Share of the population that is 70 years and older in 2015 | N |
| gdp\_per\_capita | Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available | Y |
| extreme\_poverty | Share of the population living in extreme poverty, most recent year available since 2010 | N |
| cardiovasc\_death\_rate | Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people) | N |
| diabetes\_prevalence | Diabetes prevalence (% of population aged 20 to 79) in 2017 | N |
| female\_smokers | Share of women who smoke, most recent year available | N |
| male\_smokers | Share of men who smoke, most recent year available | N |
| handwashing\_facilities | Share of the population with basic handwashing facilities on premises, most recent year available | N |
| hospital\_beds\_per\_thousand | Hospital beds per 1,000 people, most recent year available since 2010 | Y |
| life\_expectancy | Life expectancy at birth in 2019 | Y |
| human\_development\_index | A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506 | Y |
| population | Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population\_latest.csv for full list of sources | N |
| excess\_mortality\_cumulative\_absolute | Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality | Y |
| excess\_mortality\_cumulative | Percentage difference between the cumulative number of deaths since 1 January 2020 and the cumulative projected deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality | Y |
| excess\_mortality | Percentage difference between the reported number of weekly or monthly deaths in 2020–2021 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality | Y |
| excess\_mortality\_cumulative\_per\_million | Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years, per million people. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess\_mortality | Y |

## Dataset subset

Metadata of the subset, 29 variables, and 346,567 observations:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| continent | Continent of the geographical location |
| location | Geographical location |
| date | Date of observation |
| total\_cases | Total confirmed cases of COVID-19. Counts can include probable cases, where reported. |
| new\_cases | New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. |
| total\_deaths | Total deaths attributed to COVID-19. Counts can include probable deaths, where reported. |
| new\_deaths | New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA. |
| icu\_patients | Number of COVID-19 patients in intensive care units (ICUs) on a given day |
| hosp\_patients | Number of COVID-19 patients in hospital on a given day |
| total\_tests | Total tests for COVID-19 |
| new\_tests | New tests for COVID-19 (only calculated for consecutive days) |
| positive\_rate | The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests\_per\_case) |
| tests\_per\_case | Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive\_rate) |
| tests\_units | Units used by the location to report its testing data. A country file can't contain mixed units. All metrics concerning testing data use the specified test unit. Valid units are 'people tested' (number of people tested), 'tests performed' (number of tests performed. a single person can be tested more than once in a given day) and 'samples tested' (number of samples tested. In some cases, more than one sample may be required to perform a given test.) |
| total\_vaccinations | Total number of COVID-19 vaccination doses administered |
| people\_vaccinated | Total number of people who received at least one vaccine dose |
| people\_fully\_vaccinated | Total number of people who received all doses prescribed by the initial vaccination protocol |
| total\_boosters | Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol) |
| new\_vaccinations | New COVID-19 vaccination doses administered (only calculated for consecutive days) |
| median\_age | Median age of the population, UN projection for 2020 |
| aged\_65\_older | Share of the population that is 65 years and older, most recent year available |
| aged\_70\_older | Share of the population that is 70 years and older in 2015 |
| extreme\_poverty | Share of the population living in extreme poverty, most recent year available since 2010 |
| cardiovasc\_death\_rate | Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people) |
| diabetes\_prevalence | Diabetes prevalence (% of population aged 20 to 79) in 2017 |
| female\_smokers | Share of women who smoke, most recent year available |
| male\_smokers | Share of men who smoke, most recent year available |
| handwashing\_facilities | Share of the population with basic handwashing facilities on premises, most recent year available |
| population | Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population\_latest.csv for full list of sources |

### Heading 4

Sample body text.

### Heading 4

Sample body text.

### Heading 4

Sample body text.

### Heading 4

Sample body text.

## Sample Chart

## Sample Table

|  | Header Row | Header Row | Header Row |
| --- | --- | --- | --- |
| First Column | Item 1 | Item 2 | Item 3 |
| First Column | Item 4 | Item 5 | Item 6 |

#### References

* Ahouz, F., Golabpour, A. Predicting the incidence of COVID-19 using data mining. *BMC Public Health* **21**, 1087 (2021). <https://doi.org/10.1186/s12889-021-11058-3>