Analyzing the Impact of COVID-19 Control Measures on Public Health: A Data-Driven Approach

By Kevin Arnold, Brittany Roberts, and Sebastian Rolett

Abstract:

This report analyzes the CDC's COVID-19 case surveillance database, which includes a large amount of data collected during the pandemic. Our study aims to show how COVID-19 control measures, such as vaccination, affect public health outcomes. Even though we had some issues with missing data, we still think our results will be helpful. Our goal is to uncover correlations and causal relationships between different factors and COVID-19-related outcomes by thoroughly examining all available data.

1. Introduction:

The COVID-19 pandemic has had a profound impact on global public health, prompting a wide array of control and prevention measures. By studying a vast dataset from the CDC's COVID-19 case surveillance database, the report aims to understand the potential effects of these measures. Our objective is to track the pandemic's timeline and demographic factors and evaluate how preventive measures have affected them.

2. Data Collection and Challenges:

Our analysis is based on a dataset containing over 100 million entries. Despite the impressive size of the dataset, we have encountered challenges related to data integrity and missing values. The lack of data in some categories, such as "case_onset_interval" and "case_positive_specimen_interval," limits our ability to make precise conclusions. Additionally,

the presence of a substantial number of zeros in key areas further reduces the pool of viable data. Also due to the large size, the amount of time it takes to work with the data is significant and can make working with the dataset tedious. The data still has promise to give us important information, even with the challenges.

3. Methodology:

Our method involves a supervised learning linear regression task to anticipate the impact of COVID-19 on infections, hospitalizations, and deaths. To achieve this, we will undertake the following steps:

Data Preprocessing: We will conduct exploratory data analysis and data cleansing to prepare the dataset for analysis. This step is crucial for ensuring that the data is usable and accurate.

Feature Engineering: We will consider various factors, such as location, age, sex, and vaccination status, in our analysis. These factors will be used to create meaningful features that capture the nuances of the dataset.

Model Development: Predictive models will be built to estimate the number of cases, hospitalizations, and deaths that could have occurred without control measures. The exact choice of model and algorithm has yet to be decided to perform the predictions.

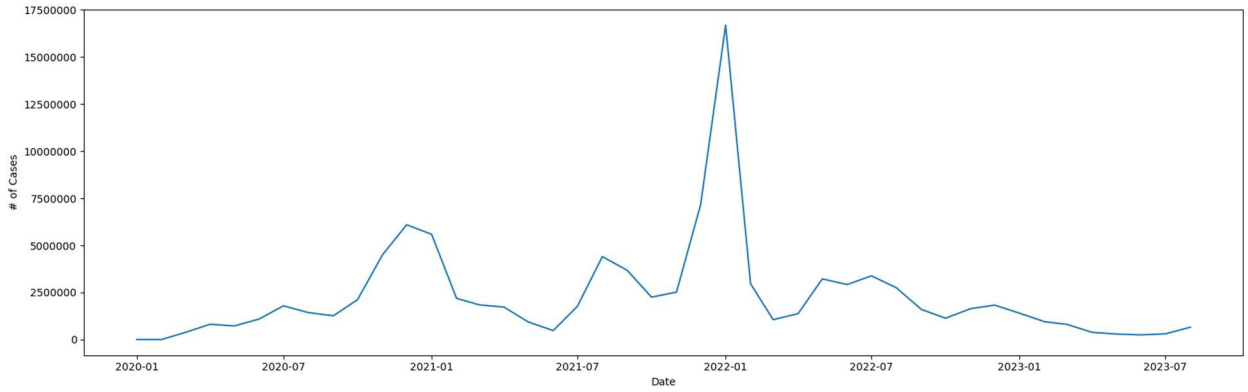
Evaluation: The models' performance will be assessed using appropriate metrics to gauge their accuracy and reliability.

4. Results and Discussion:

Our analysis aims to uncover valuable insights regarding the potential impact of COVID-19 control measures. We will investigate how changes in data correlate with key events in the

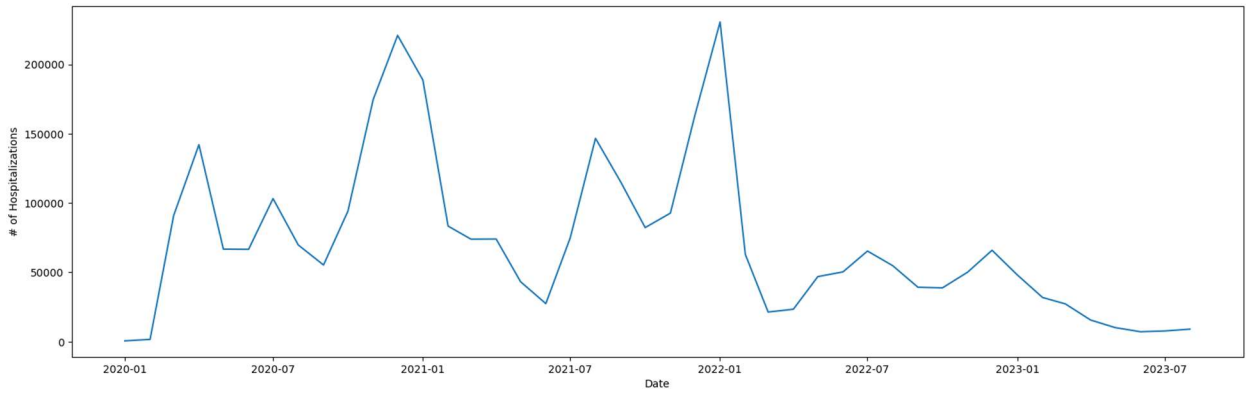
pandemic, such as vaccine rollouts and variations in vaccine efficacy. Additionally, we will explore the impact of factors such as mask mandates, public closures, and hospital capacity on COVID-19 outcomes. We will use visualizations to make it easier to understand the trends and relationships we find.

We decided to train the regression model using the data that was collected between 2020-01-01 and 2021-01-01. We decided on these dates based on the following graph showing the trends in Covid-19 cases (*Covid-19 case surveillance public use data with geography*):

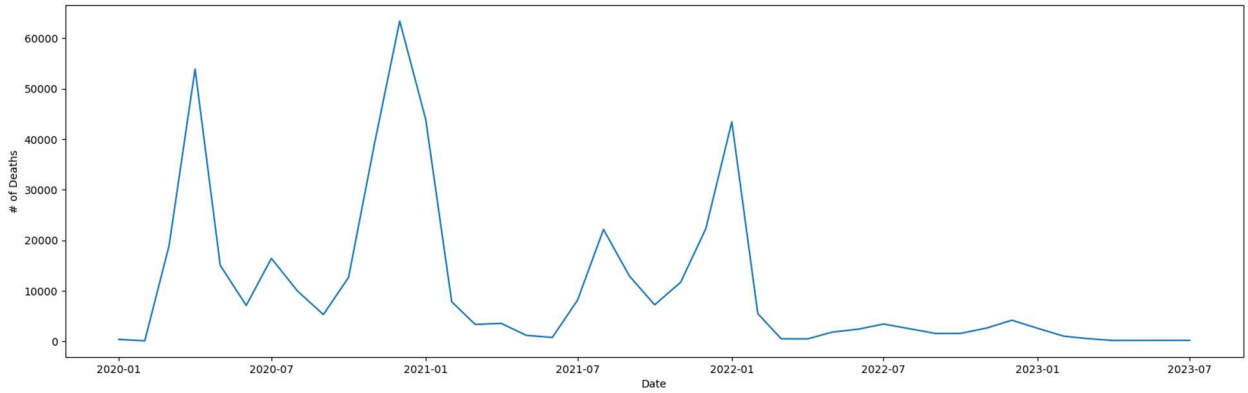


As shown above, there is a progressive increase in cases leading up to the end of 2020 and then cases start to decline in the new year, this is due to starting of administration of Covid-19 vaccines (Assistant Secretary for Public Affairs (ASPA), 2023). This same trend can also be seen in the number of hospitalizations (Figure A) and number of deaths (Figure B).

Figure A



Figure



5. Conclusion:

Even though we faced challenges with missing data, our analysis of the CDC's COVID-19 case surveillance database can show how well control measures worked during the pandemic. Our goal is to find out how the pandemic could have been different without preventive measures by analyzing trends and correlations in the data. This research adds to the broader knowledge of the impact of COVID-19 on public health and the necessity of timely interventions.

Works Cited

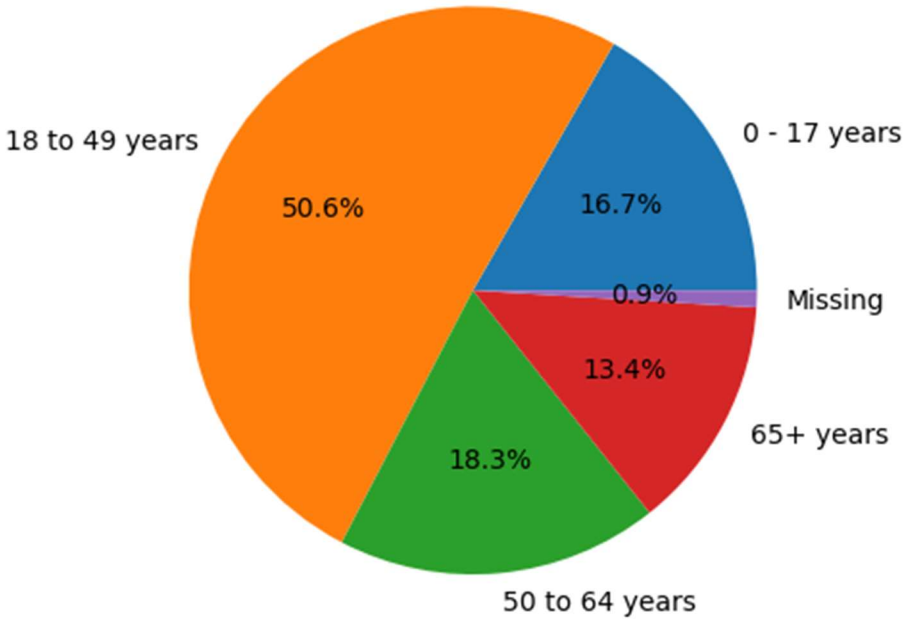
(Assistant Secretary for Public Affairs (ASPA), 2023) Assistant Secretary for Public Affairs (ASPA). (2023, September 13). *Covid-19 vaccines*. HHS.gov. https://www.hhs.gov/coronavirus/covid-19-vaccines/index.html#:~:text=August%2023%2C%202021,years%20of%20age%20and%20older.

Centers for Disease Control and Prevention. (n.d.). *Covid-19 case surveillance public use data with geography*. Centers for Disease Control and Prevention. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

Population distribution by sex. KFF. (2022, October 28). https://www.kff.org/other/state-indicator/distribution-by-sex/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D

Appendix A

While wrangling the data, we looked at some other statistics that were of interest but that did not necessarily fit for what we were looking at. We found that the age group with the largest number of Covid-19 cases were from the age of 18-49 years old and the age group with the smallest number were from the age of 65 and up. It is interesting because when the vaccines were approved and they were deciding on who to get them out to first, it was determined that the elderly and people with pre-existing conditions would receive the vaccines.



Appendix B

Another detail that was found while wrangling the data, were the number of cases between males and females. In this data set, we saw that 54.4% of cases were female and 45.4% were male. As of 2021, the United States population was 50.8% were female and 49.2% were male so it makes sense that there would be more cases for females over males (*Population distribution by sex* 2022).