BST260 final-project

1. Abstract

This project provides a comprehensive analysis of the impact of the COVID-19 pandemic across the United States from January 2020 to May 2023. To better understand the evolution of the pandemic, I divided the pandemic timeline into distinct periods based on crucial time points and trends by case and mortality rates. By visualizing the data across country and state levels, I was able to identify disparities in case and mortality trends, highlighting the impact of regional differences and policy decisions. Death rates are computed for each selected period, revealing variations in state-level performance in responding to COVID-19. States with better vaccination coverage and healthcare systems outperformed those with lower access to public health resources, highlighting the critical role of public health intervention. The finding also discovered that COVID-19 became less virulent over time, evidenced by declining death rates across different periods. This is likely due to widespread vaccine adoption, improved treatments, and increased population immunity.

2. Introduction

The COVID-19 pandemic is one of the most significant global public health crises in modern history. Since appearing in late 2019, COVID-19 has caused millions of infections and deaths globally. The pandemic profoundly affects societies, economies, and healthcare systems worldwide. The United States has been one of the countries most significantly impacted by the COVID-19 pandemic, experiencing millions of confirmed cases and hundreds of thousands of deaths, along with widespread economic and social disruptions. The different impact of the COVID-19 pandemic provides an opportunity to analyze the factors that influence health outcomes (e.g., death rate) and assess the evolution of the pandemic over time. To understand the evolution of the COVID-19 pandemic in the United States, it is important to explore its different waves with unique patterns of infection and death rates. These waves often correspond with the emergence of new variants (e.g., "Delta" variant, "Omicron" variant), changes in public health policies, and shifts in population behavior. By dividing the COVID-19 pandemic into specific time periods based on these waves, it is possible to assess trends in mortality and morbidity across the states. This approach offers valuable insights into the effectiveness of public health strategies and identifies areas requiring improvement. Understanding the success and shortcomings of the intervention at each wave provides a clear roadmap for future pandemic preparedness.

This study focuses on three critical aspects related to the COVID-19 pandemic in the United States: identifying distinct waves of the pandemic, calculating death rates across states during these periods, and assessing changes in the virus's virulence over time. Correctly identifying the time periods for distinct waves of the pandemic can provide a clear pattern of how COVID-19 evolved and how people responded with measures like vaccines and public health actions. Identifying distinct waves of the pandemic not only helps in understanding the evolution of COVID-19 but also provides a basis for evaluating the timing and effectiveness of public health responses. This categorization allows researchers and policymakers to identify critical moments where interventions were most effective. By calculating and comparing mortality rates across states in each wave, the results of the analysis will highlight which states are performing better or worse in managing pandemics. This analysis provides insights into the disparities in healthcare resources, infrastructure, and public policy approaches between different states. For example, states with complete health facilities and instance response to the pandemic may show lower death rates. In contrast, states with poor medical facilities and delayed responses may show higher death rates. At last, evaluating whether COVID-19 became virulent across different periods can provide valuable insights into the relationship between viral evolution, healthcare interventions, and public health strategies. This analysis can help determine whether COVID-19 has become less severe due to medical advances or more dangerous due to evolution. The significance of this study extends not only focus on understanding the pandemic's past impact but also provides insights to inform future pandemic preparedness and response strategies. By identifying state-level disparities and evaluating the effectiveness of public health interventions, policymakers can develop effective approaches to address weaknesses and build more resilient healthcare systems.

3. Method

3.1 Data description and collection:
This study utilized multiple datasets to analyze the COVID-19 pandemic in the United States from January 2020 to July 2023. The population census data was from the U.S. Census Bureau's National Population Estimates database, including yearly population estimates for all 50 states and the District of Columbia and Puerto Rico from 2020 to 2023. COVID-19 case and mortality data were retrieved from the Centers for Disease Control and Prevention (CDC), including weekly reports that provided detailed counts of new cases and deaths for each state.

3.2 Data Cleaning:
All of the processes are in the R studio environment. Population data was imported from an Excel file, removing unnecessary columns and rows. The dataset was transformed into a tidy format using the tidyr package. Specifically, the yearly population census dataset was reshaped into a long format (pivot_longer() function), where each row represents a unique combination of year, state, and population value. State abbreviations were added using standard mapping functions (match() function), and Puerto Rico and the District of Columbia were added separately. Missing data were excluded from the dataset. COVID-19 case and

death data were accessed from the CDC's public API. Data was retrieved in JSON format using the httr2 package. Both datasets are cleaned to include only relevant variables: state, date, and counts of cases and deaths. Dates in both datasets were transformed into a standard format, and weeks and years were calculated using the lubridate package(as_Date(), epiweek(), epiyear()). COVID-19 cases and deaths were filtered to include only states present in the population dataset to ensure consistency. Missing values are also excluded. All datasets (population census, Covid-19 case, Covid-19 death) were combined together (left_join() function) using state, year, and weeks as key variables.

3.3 Data visualization:
The population census dataset was grouped by year, and the total population was calculated by summing up (summarise() function) the individual state populations for each year with the tidyr package. This summarization provided the annual national population for the period from 2020 to 2023. In order to visualize the trend of COVID-19 based on monthly, the floor_date() function from the lubridate package is used to aggregate dates into monthly intervals. After this transformation, data were grouped by month, and the total cases and deaths for each month at the state level were calculated. Monthly case and death rates were calculated using the following formula: (Total Cases or Deaths/ Total population) * 100,000. The data was visualized to highlight trends in case and death rates over time. Two primary visualizations were created: A line plot (geom_line()) combined with a bar chart (geom_col()) was generated to display trends in case rates per 100,000 population using the ggplot2 package. This visualization highlighted significant variation in case numbers across 2021 to 2023 in units of month. This graph provides insight into how the disease spreads nationally. A similar visualization was created for death rates. It shows the progression of COVID-19 severity and its impact on the death rate. Peaks in death rates were aligned with key events such as the emergence of COVID-19 or the emergence of vaccines, providing a clear narrative of the pandemic's impact.

3.4 Separating Periods Based on Visualization:
To analyze the progression of COVID-19, the pandemic was divided into five distinct periods based on trends observed in monthly case and death rate visualizations. A literature review was also conducted to validate and support this periodization (CDC Museum COVID-19 Timeline, 2023). These periods all reflect significant key time points, such as the emergence of COVID-19, vaccines, and COVID-19 variant.

3.5 Death rates by state for each period:
For each period, data cleaning involved filtering dates (filter() function) to select the specific time frames of interest by the dplyr package. Within each period, I grouped data by state and calculated the total number of deaths (sum() function) while averaging population data (mean() function) to account for cross-year variations. Death rates per 100,000 individuals were then calculated as (Total Deaths/ Total population) * 100,000. The ggplot2 package was used to create horizontal bar plots of death rates, where states were ordered by their death rates using the reorder() function.

3.6 Trends in COVID-19 Virulence:
To analyze trends in COVID-19 virulence over time, I calculated case and death rates across five distinct periods defined in 3.4. Using the dplyr package in R, we filtered data for each period and computed the total number of cases and deaths, as well as the average population during each time frame. The death and case rate for each period is calculated as (Total Deaths or Cases/ Total population) * 100,000. The results were visualized using ggplot2, with line charts (geom_line() function) illustrating the trends in case and death rates over different periods.

4. Results

4.1 Data Collection and Cleaning:
The combined dataset includes yearly population data from the U.S. Census Bureau's National Population Estimates database and COVID-19 case data and death data from the CDC. The population dataset includes the annual population census for all 50 states, the District of Columbia, and Puerto Rico. The CDC dataset provides detailed counts of COVID-19 cases and deaths at the state level. Data cleaning involved standardizing population data, reshaping it into a tidy format, and aligning state abbreviations. Key variables, including state, state name, date, year, week, cases, deaths, population, and total population, were selected. Table 1 provides an overview of the data structure (in the Supplementary Methods):

4.2 Data visualization:
The total population of the United States from 2020 to 2023 was summarized by year, providing a baseline for calculating COVID-19 cases and death rates. Monthly totals for cases and deaths were calculated, and case and death rates per 100,000 population were calculated by dividing the monthly totals by the annual population and scaling with 100,000. Data visualizations were created to illustrate these trends:

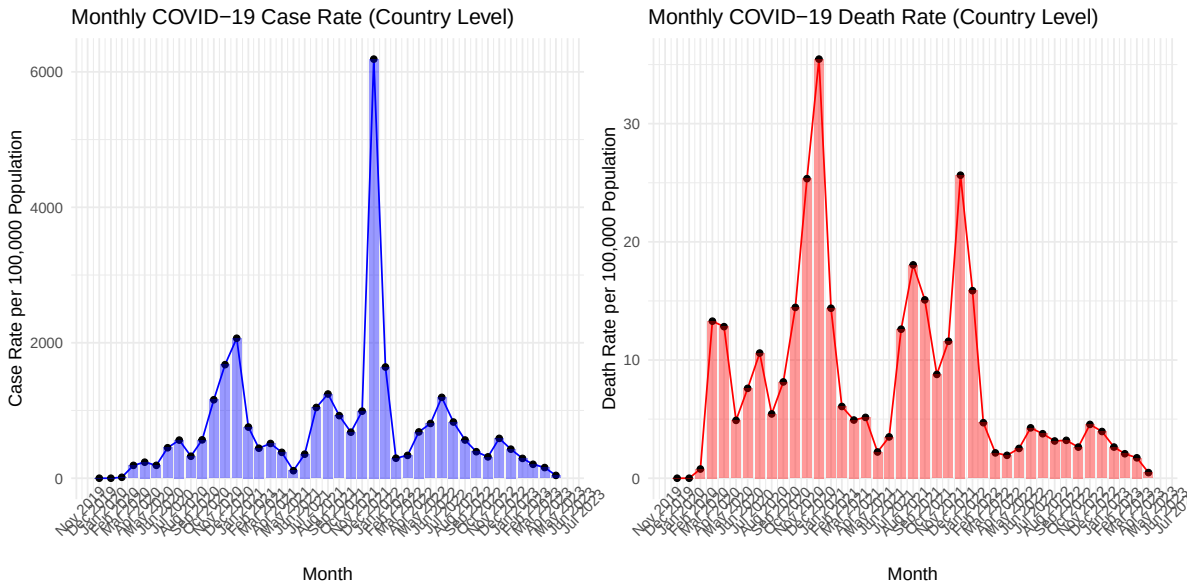
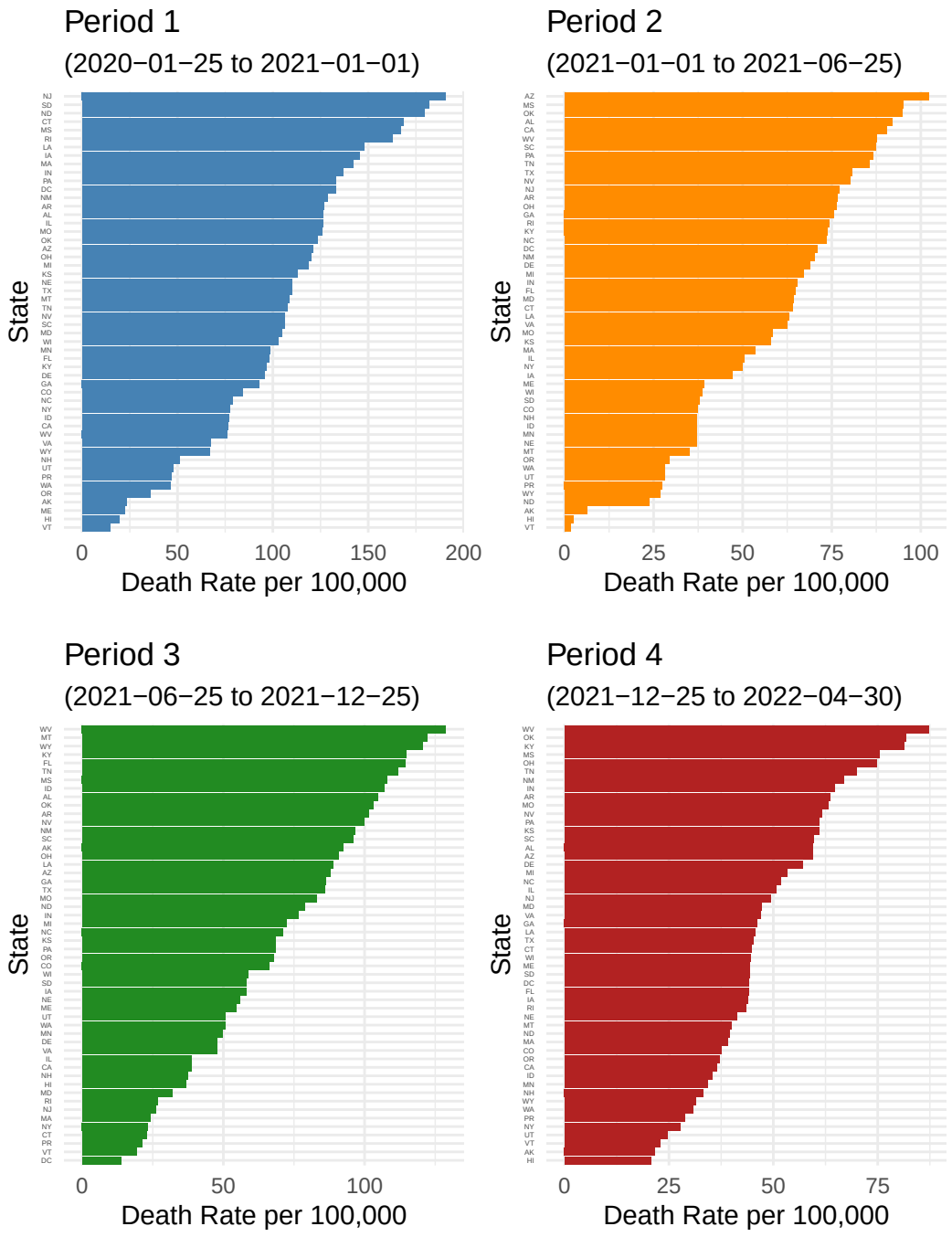


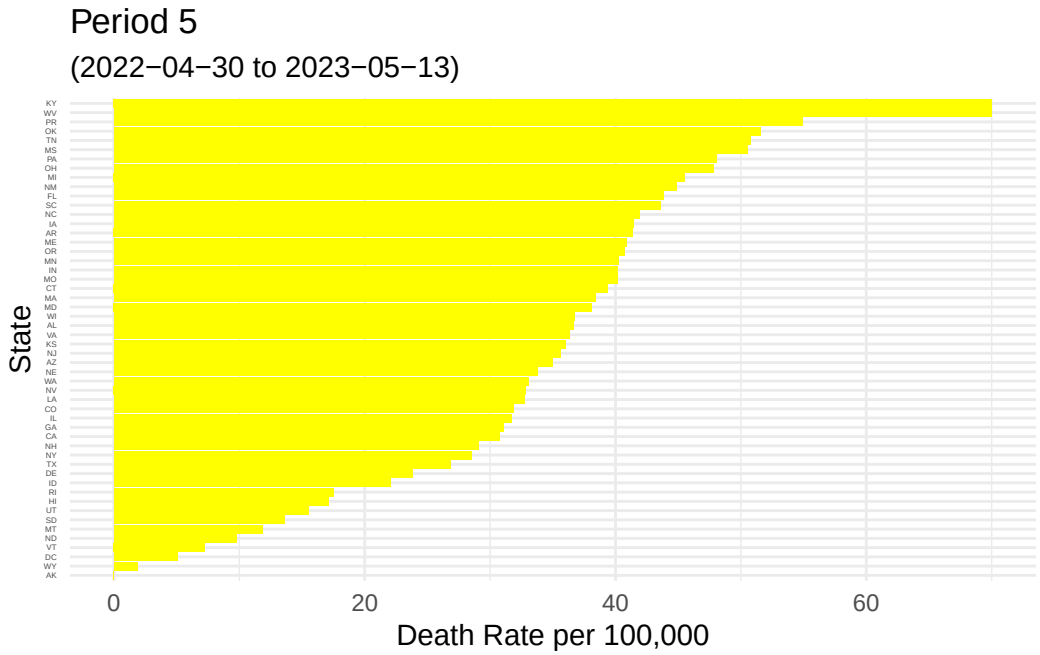
Figure 1: Monthly COVID-19 Case and Death Rates

The left data visualization in Figure 1 displays the monthly COVID-19 case rate per 100,000 population. The line plot highlights the monthly trends, with peaks in case rates corresponding to key periods in the pandemic. The bar chart provides additional clarity on the trend. The right data visualization in Figure 1 represents the monthly COVID-19 death rate per 100,000 population. The line plot highlights peaks in death rates. The bar chart provides insight into the overall trend of mortality rates across the pandemic.

4.3 Separating Periods Based on Visualization:
The pandemic was divided into five periods based on data visualization in 4.2. The first period: from 2020-01-25 to 2021-01-01, the second period: from 2021-01-01 to 2021-06-25, the third period: from 2021-06-25 to 2021-12-25, the fourth period: from 2021-12-25 to 2022-04-30, and the fifth period: from 2022-04-30 to 2023-05-13.

4.4 Death rates by state for each period:
State-level COVID-19 death rates were analyzed across five distinct periods defined by key pandemic milestones. For each period, death rates per 100,000 population were calculated based on the total number of deaths and the average population. Horizontal bar plots in Figure 2 (also in Supplementary Methods) were created to visualize the state-wise death rates for each period, showing variations across states and over time.


Period 1
(2020−01−25 to 2021−01−01)
State
0
50
100
150
200
Death Rate per 100,000
Period 2
(2021−01−01 to 2021−06−25)
State
0
25
50
75
100
Death Rate per 100,000
Period 3
(2021−06−25 to 2021−12−25)
State
0
50
100
Death Rate per 100,000
Period 4
(2021−12−25 to 2022−04−30)
State
0
25
50
75
Death Rate per 100,000



Period 1: 2020-01-25 – 2021-01-01 shows the initial impact of COVID-19. New Jersey reported the highest death rate, about 180 deaths per 100,000 population. States like Vermont, with smaller populations or early interventions, experienced lower death rates.

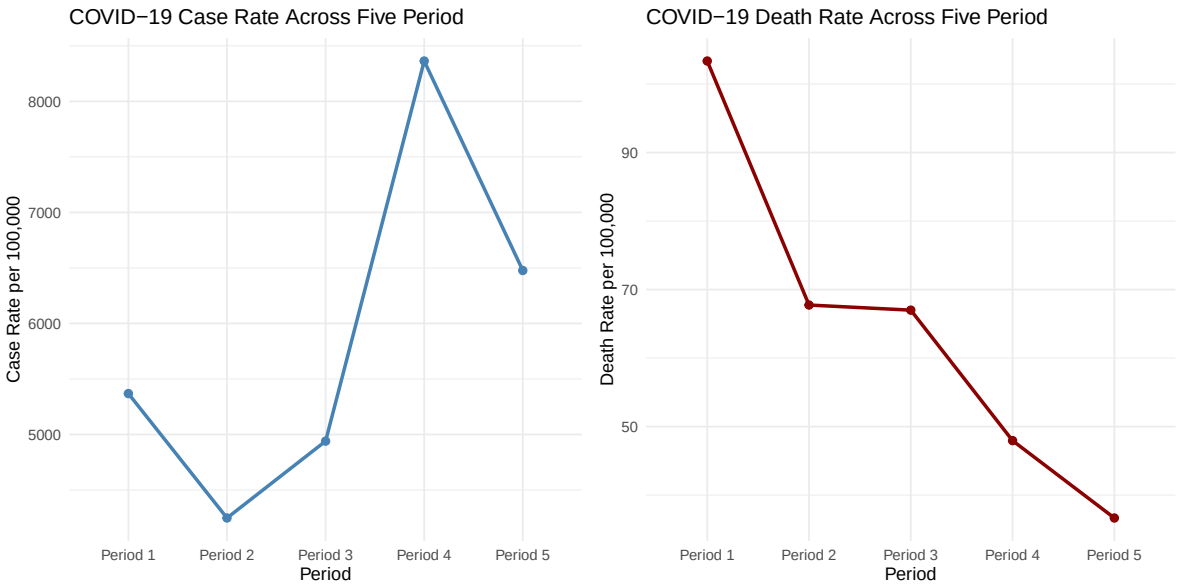
Period 2: 2021-01-01 – 2021-06-25 shows that death rates declined in most states because of the availability of vaccines. However, states such as Arizona and Mississippi experienced relatively high death rates (about 120 deaths per 100,000 population) due to vulnerable communities (Shen et al., 2023).

Period 3: 2021-06-25 – 2021-12-25 shows that the appearance of the Delta variant increases the death rate. States such as West Virginia and Montana show high death rates (about 130 deaths per 100,000 population) and the District of Columbia shows a relatively low death rate.

Period 4: 2021-12-25 – 2022-04-30 shows the appearance of the Omicron variant. Despite the high transmissibility of the Omicron variant, death rates during this period (about 80 deaths per 100,000 population) were generally lower than during the Delta wave.

Period 5: 2022-04-30 – 2023-05-13 is the last period of the COVID-19 pandemic. Death rates continued to decline across most states. This is because of the widespread administration of booster vaccines and improved public health measures (Lazarus et al., 2023).

4.5 Trends in COVID-19 Virulence:
Figure 3 provides insights into trends in COVID-19 case and death rates per 100,000 population across five periods, allowing an evaluation of whether the virus became more or less virulent over time.



Case Rate: The case rate varies across the periods. Period 1 had a moderate case rate, which declined significantly in Period 2, corresponding with the initial vaccine appearance. However, the case rate increased again during Period 3 with the emergence of the Delta variant. The case rate arrived at a peak in Period 4 during the Omicron variant. A marked decline occurred in Period 5, suggesting increased adaption in COVID-19 vaccination boost and increased immunity. Death Rate: The death rates showed a consistent downward trend over the five periods. Period 1 had the highest death rate, reflecting the early severity of the virus. Death rates declined significantly in Period 2 and stabilized during Period 3. Period 4 showed a continued reduction in the death rate. Finally, Period 5 had the lowest death rate of all periods. COVID-19 became less virulent over time, as indicated by the consistent decline in death rates across the periods, despite fluctuations in case rates. The reduced death rate in later periods suggests improvements in treatment, increased vaccine coverage, and increased immunity.

5. Discussion

The findings of this report provide significant insights into the evolution of the COVID-19 pandemic in the United States. It highlights the patterns of disparities, the effectiveness of public health interventions, and changes in the virus's virulence over time at both the state and country levels. These observations not only provide insights into key factors that influenced the outcomes during each of the five distinct periods of the pandemic but also provide a reference for future preparedness and response strategies.

The COVID-19 pandemic period was divided into waves based on key trends observed in case and death rates visualized through data analysis and visualization. Period 1 corresponds to the initial appearance and rapid spread of COVID-19. It is characterized by high case and death rates due to limited knowledge and strategies corresponding to COVID-19 (Lai et al., 2022).

Data visualization shows sharp peaks in both case and death rates during this period. Period 2 is characterized by the appearance of the vaccine. During this period, both the case rate and death rate declined due to the efficacy and effectiveness of the COVID-19 vaccines (Mohammed et al., 2022). Period 3 is marked by the appearance of the Delta variant. Delta variant not only has higher transmissibility compared with the original COVID-19 but also lower vaccine effectiveness against Infection (Bian et al., 2021). This results in an increase in cases and death rates again. Period 4 is highlighted by the Omicron variant. This period revealed a sharp increase in case rates due to its high transmissibility. However, data visualization also shows a consistent decline in death rates. This is because the Omicron variant has the highest transmissibility but causes milder disease compared with previous variants (Yu et al., 2022). Period 5 is the final period of the pandemic. It shows a continued decline in both case and death rates in data visualization due to widespread vaccine boosters, improved public health measures, and increased public immunity.

The analysis showed significant temporal and geographic differences in case and mortality rates, which highlights the importance of adapting public health interventions to the regional and temporal context. For each period, death rates were computed by state to describe which states performed better or worse during the different periods. For instance, the relatively low death rates observed in Vermont during the early periods of the pandemic reflect the success of proactive public health measures and robust healthcare systems (Vatovec & Hanley, 2022). In contrast, the high death rates in states like West Virginia and Arizona during specific periods highlight disparities in healthcare access and socioeconomic factors such as community-level deprivation and racial inequalities. (Hendricks et al., 2021). Policymakers should focus on strengthening healthcare systems in low-resourced areas, improving vaccine distribution, and addressing socioeconomic problems to reduce vulnerability during future pandemics. Another crucial finding is the decline in COVID-19's virulence over time. Even though earlier periods indicate high death rates due to the initial impact of COVID-19 and limited medical knowledge, subsequent periods showed declining death rates. This trend can be attributed to widespread vaccination campaigns, improved public health infrastructure, and increased public immunity (Yuan et al., 2023). The appearance of variants such as Delta and Omicron introduced significant challenges, but their impact on the death rate was mitigated by these advances. These observations emphasize the role of continued medical innovation and vaccine development in pandemic mitigation.

Despite the comprehensive approach used in this report, there are some limitations that should be addressed in future studies. First, the reliance on state-level data did not capture the within-state variations and disparities. Urban and rural areas in the same state often experience different outcomes, and capturing this variation could provide deeper insights into the changing dynamics of pandemics. Second, the study did not account for confounding variables such as age distribution and socioeconomic factors that might influence COVID-19 outcomes significantly.

In conclusion, this report highlights the complex interaction of factors affecting COVID-19's impact across the United States. The findings reveal significant disparities in state-level out-

comes (death rate), emphasizing the critical role of public health interventions. While the pandemic's evolution presented unprecedented challenges, the lessons learned provide a foundation for more effective responses in the future. Understanding and solving the identified limitations and building on these insights will be important for enhancing resilience against future public health crises.

Reference

Bian, L., Gao, Q., Gao, F., Wang, Q., He, Q., Wu, X., … & Liang, Z. (2021). Impact of the Delta variant on vaccine efficacy and response strategies. Expert review of vaccines, 20(10), 1201-1209.

CDC Museum COVID-19 timeline. (2023, March 15). Centers for Disease Control and Prevention. https://www.cdc.gov/museum/timeline/covid19.html

Hendricks, B., Paul, R., Smith, C., Wen, S., Kimble, W., Amjad, A., … & Hodder, S. (2021). Coronavirus testing disparities associated with community level deprivation, racial inequalities, and food insecurity in West Virginia. Annals of epidemiology, 59, 44-49.

Lazarus, J. V., Wyka, K., White, T. M., Picchio, C. A., Gostin, L. O., Larson, H. J., … & El-Mohandes, A. (2023). A survey of COVID-19 vaccine acceptance across 23 countries in 2022. Nature medicine, 29(2), 366-375.

Lai, S., Bogoch, I. I., Ruktanonchai, N. W., Watts, A., Lu, X., Yang, W., … & Tatem, A. J. (2022). Assessing spread risk of COVID-19 in early 2020. Data Science and Management, 5(4), 212-218.

Mohammed, I., Nauman, A., Paul, P., Ganesan, S., Chen, K. H., Jalil, S. M. S., … & Zakaria, D. (2022). The efficacy and effectiveness of the COVID-19 vaccines in reducing infection, severity, hospitalization, and mortality: a systematic review. Human vaccines & immunotherapeutics, 18(1), 2027160.

Shen, F. L., Shu, J., Lee, M., Oh, H., Li, M., Runger, G., … & Liu, L. (2023). Evolution of COVID-19 health disparities in Arizona. Journal of Immigrant and Minority Health, 25(4), 862-869.

Vatovec, C., & Hanley, J. (2022). Survey of awareness, attitudes, and compliance with COVID-19 measures among Vermont residents. Plos one, 17(3), e0265014.

Yu, W., Guo, Y., Zhang, S., Kong, Y., Shen, Z., & Zhang, J. (2022). Proportion of asymptomatic infection and nonsevere disease caused by SARS-CoV-2 Omicron variant: a systematic review and analysis. Journal of medical virology, 94(12), 5790-5801.

Yuan, Y., Jiao, B., Qu, L., Yang, D., & Liu, R. (2023). The development of COVID-19 treatment. Frontiers in immunology, 14, 1125246.