Final Project: COVID-19 pandemic in the US

Jingxi Zhang

Abstract

The COVID-19 pandemic significantly impacted the United States and it influence people's life. This study investigates COVID-19 trends from January 2020 to May 2023, focusing on three objectives: divide and identify the pandemic period, compute the deaths rates by state for each period, and discussing changes in virulence. By using CDC data on cases and deaths, combined with U.S. Census population data, three major waves were identified after making datil visualizations: Wave 1 (October 2020–March 2021), Wave 2 (August 2021–April 2022), and Wave 3 (November 2022–March 2023). Then, average monthly death rates per 100,000 population were calculated for each wave, allowing for a comparison of outcomes across states and periods. Differences in death rates between waves revealed significant regional disparities, with some states achieving improvements while others experienced worsening trends. In result, South Dakota (SD) demonstrated the most substantial improvement, with a decrease of 19.7 deaths per 100,000 population between wave 1 and wave 2. In overall, all the state has decrease in death rate in wave 3, reflecting the combined effects of increased vaccination coverage, reduced virulence of circulating variants, and enhanced public health responses.

Introduction

The COVID-19 pandemic has worldwide impacted on people's life since 2020, leading to millions of deaths and putting unprecedented strain on healthcare systems. In the United States, the pandemic period can be classified into 3 waves, driven by emerging variants, vaccination efforts, and public health measures. These periods have differed in their intensity, duration, and regional impact. So, detailed analyses are crucial to understanding the factors behind these variations. Such insights are essential for shaping effective public health policies and responses in the future.

Understanding state-level COVID-19 mortality rates and case patterns during three pandemic period is the main goal of this study. Pandemic wave are identified through case trends and real-world events, such as the emergence of new variants and the implementation of large-scale vaccination campaigns. Each period represents distinct phases of the pandemic, characterized by shifts in the virus's virulence, healthcare system capacity, and public compliance with health measures. Second, it investigates which states showed consistent progress or setbacks in death rates across these periods. Metrics such as average death rate differences, and trends in cases were analyzed to identify patterns. Understanding these variances is crucial for informing public health interventions. States differ significantly in demographics, healthcare infrastructure, and policy responses, influencing COVID-19 outcomes. For instance,

early vaccination efforts, the availability of intensive care resources, and the timeliness of public health measures varied widely across states. These factors, combined with the virus's evolving nature, likely contributed to the observed disparities in death rates.

Lastly, the study analyzes the factors that might responsible for these variations at the state level, focusing to policy choices, healthcare-related impacts, and demographic traits. In this study, we assume that states with higher vaccination rates during each wave experienced lower COVID-19 death rates, particularly in wave 2 and wave 3, when vaccines were widely available. Also, policy measures, such as mask mandates, stay-at-home orders, and other non-pharmaceutical interventions, were associated with improved outcomes, with states implementing stricter measures showing lower death rates. Furthermore, demographic and healthcare-related factors, including population density, median age, prevalence of chronic diseases, and ICU bed availability, contributed to variations in death rates across states and pandemic waves.

The significance of this study lies in its potential to help guide evidence-based policymaking. By identifying patterns of success and failure, the findings offer useful insights for policymakers to strengthen public health responses in future pandemics. By analyzing the interplay of policy, healthcare, and demographic factors, this research contributes to the broader understanding of pandemic management and resilience.

In summary, this study analyzes COVID-19 death rates across different pandemic periods to understand regional variations and trends. By addressing the research questions outlined above, it aims to uncover the drivers of success and failure at the state level and to provide valuable lessons for improving public health preparedness. The findings emphasize the need for adaptive, data-driven strategies to mitigate the impact of pandemics and to protect public health in the face of evolving challenges.

Method

1. Data Wrangling:

Covid-19 Data:
The COVID-19 cases and death data were sourced from the CDC's publicly available dataset via its API: https://data.cdc.gov/resource/pwn4-m3yp.json and https://data.cdc.gov/resource/r8kw-7aab.json. Those data include detailed information on daily new cases and deaths reported at the state level, as well as their corresponding dates. For this study, data from January 2020 to May 2023 were collected. Key variables extracted from this dataset included the number of new cases (new_cases), deaths (new_deaths), and their respective reporting dates (end_date) and state.

Above dataset required several cleaning steps to ensure uniformity. Non-numeric entries and missing values in case and death counts were converted to NA, converted number data to a numeric format, date values were parsed into a standardized format, added additional columns such as epidemiological week and epidemiological year and filtered data include U.S. states and standardized the state name to their abbreviations. These steps enabled the aggregation of daily data into monthly summaries for trend analysis.

Population Data:
The national population data is collected from U.S. Census Bureau which is authoritative. State-level population estimates for the years 2020 to 2023 were used to calculate death rates.

The population data for 2020 and 2021 were download through API: https://api.census.gov/data/2021/pep/population. It was cleaned and restructured to improve its usability for analysis and to enable integration with other datasets. First, the column names were standardized by promoting the first row as headers, and the data was converted into a tibble for better handling. The unnecessary state column was removed, and the 'NAME' column was renamed to 'state_name' for consistency. To prepare for numeric analysis, the "POP_" prefix was removed from the year column, and all relevant columns were parsed into numeric format. Also to enhance geographic usability, state abbreviations were added by matching state names to built-in R datasets.

The population data for 2022 and 2023 is downloaded separately from U.S. Census Bureau as an Excel file. Data cleaning directly performs in Excel such as changing column headers appropriately and removed redundant rows that were part of the downloaded dataset's formatting. The dataset was then transformed into a long format using pivot_longer to alien with 2020 and 2021 data. This comprehensive structure enables easier analysis of population trends, facilitates integration with COVID-19 cases and deaths data, and allows for state-level and regional comparisons in subsequent analyses.

2. Divide the Pandemic Period:

After cleaning the data, processes and visualizes national-level COVID-19 case data is performed to identify trends and divide the pandemic period, from January 2020 to May 202. The first step involves aggregating weekly case data at the national level. The data then is grouped by 'mmwr_year' (epidemiological year) and 'mmwr_week' (epidemiological week), and the total weekly cases are calculated by summing cases across all states. This aggregation provides a comprehensive view of COVID-19 trends. To make plot in time-series analysis, the 'date' column is created by combining the mmwr_week and mmwr_year, allowing visualization in the form of date vs. total cases for a given week. A blue line is generated using ggplot2,

highlighting the trends in case numbers and providing a clear visual representation of fluctuations in cases over the pandemic period. This visualization allows for the identification of distinct pandemic waves by observing the patterns and peaks in the case data. From this analysis, three wave periods were identified: Wave 1 (October 2020–March 2021), Wave 2 (August 2021–April 2022), and Wave 3 (November 2022–March 2023), offering valuable insights into the evolution of COVID-19 spread over time.

3. Compute Death Rates and Compare

To calculate state-level death rates for each identified pandemic wave, the data was processed and analyzed in several steps. First, the deaths data was segmented into three defined wave periods: Wave 1 (October 2020–March 2021), Wave 2 (August 2021–April 2022), and Wave 3 (November 2022–March 2023). This was accomplished using a 'case_when' statement. Records outside these wave periods were excluded from further analysis. The resulting dataset was then merged with population data by matching the state and year columns, ensuring that death rates could be calculated relative to the population size.

Since different period have different length in month, average death rate is calculated. To compute that, the number of months in each wave period was predefined and added to the death data (6 months for Wave 1, 9 months for Wave 2, and 5 months for Wave 3). The dataset was then grouped by state and wave to aggregate total deaths and calculate the average population for each period. These values were used to compute the average monthly death rate per 100,000 population for each state and wave. The metric, average monthly death rate, was derived by dividing total deaths by the average population, normalizing per 100,000 population, and dividing the number of months in the wave. This approach allowed for accounting for differences in population size and wave duration.

To analyze death rates across states COVID-19, we began by visualizing above calculated data. Each state was represented by a unique line and color to highlight changes across waves. While the line plot provided an overview of trends, the large number of states resulted in a cluttered and difficult-to-interpret visualization. To overcome this limitation and better summarize the state-level changes, we calculated the differences in death rates between consecutive waves. The data was first reshaped into a wide format, allowing for the computation of two key metrics: Difference between Wave 1 and Wave 2 and Difference between Wave 2 and Wave 3. Negative values in these metrics indicate a reduction in death rates (improvement), while positive values suggest an increase (worsening outcomes). The results were sorted to identify states with the most significant improvements or setbacks for each comparison. This method provided a concise and interpretable overview of state-level changes in death rates during the pandemic, allowing for the identification of trends that were otherwise obscured in the line plot.

4. Assumptions and Limitation

COVID-19 data are assumed to be accurate despite issues like underreporting or delays. Besides, state population estimates are considered constant within each year, ignoring migration or demographic changes.

Results

The blue line in figure 1 shows how the number of cases increased over time. The first notable increase occurs in late 2020 and early 2021, followed by a significant decline. In comparison to the first wave, a second wave that peaks in mid-late-2021 has a small growth. The most noticeable spike is observed in early 2022. After this peak, the total case numbers drop sharply and remain relatively stable with smaller fluctuations through the rest of 2022 and into 2023. Hence, three primary pandemic periods is identified: Wave 1 (October 2020–March 2021), Wave 2 (August 2021–April 2022), and Wave 3 (November 2022–March 2023).

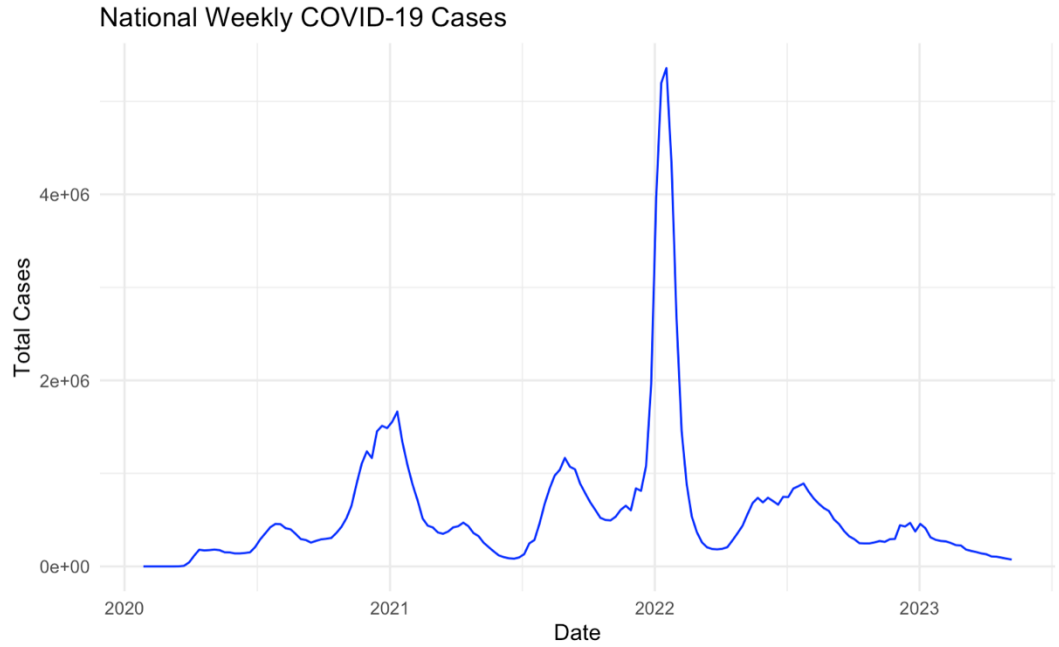

Figure 1: National Weekly COVID-19 Cases Over Time (2020-2023)

Figure 2 shows the total COVID-19 deaths over time in the U.S. The initial sharp increase in deaths occurred in early 2020, followed by multiple subsequent waves. Notable peaks are observed in late 2020, mid-2021, and early 2022. After early 2022, a general downward trend is observed, with smaller fluctuations continuing into 2023 and 2024.

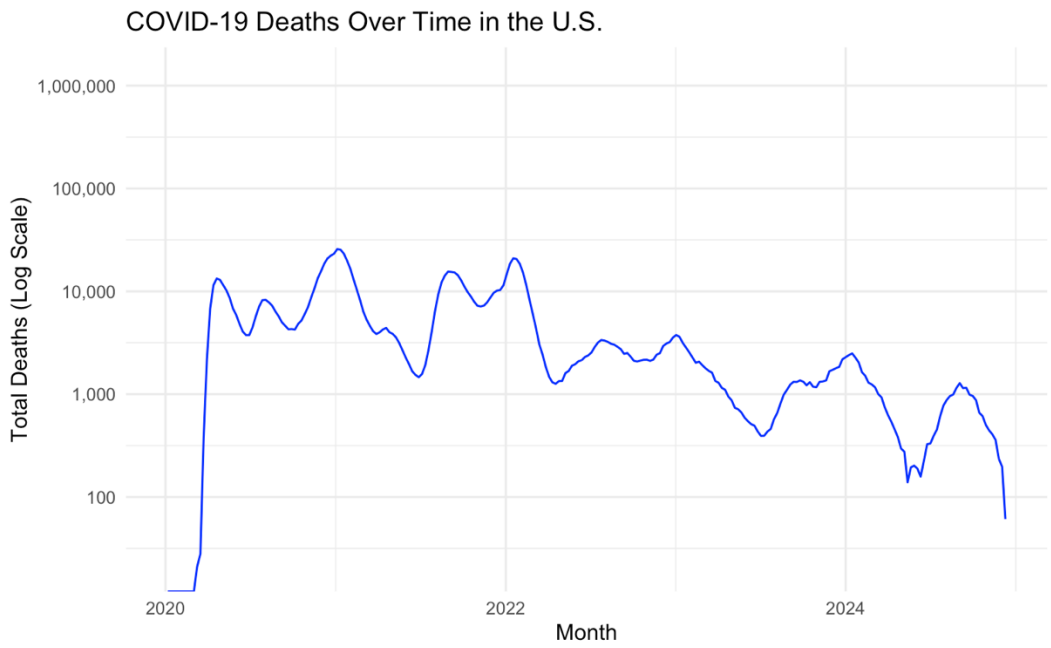

Figure 2: COVID-19 Death Over Time (2020-2023)

Figure 3 illustrates the average monthly death rates per 100,000 population across the three pandemic periods (Wave 1, Wave 2, and Wave 3) for all U.S. states. Each colored line represents a state, showing its death rates across the periods. Notably, the death rates varied significantly between states and periods. Some states experienced an increase in death rates from Wave 1 to Wave 2. In contrast, Wave 2 to Wave 3 shows big decline trend in death rates across most states. By Wave 3, some states such as West Virginia, which had high death rates in Wave 2, had significantly improved.

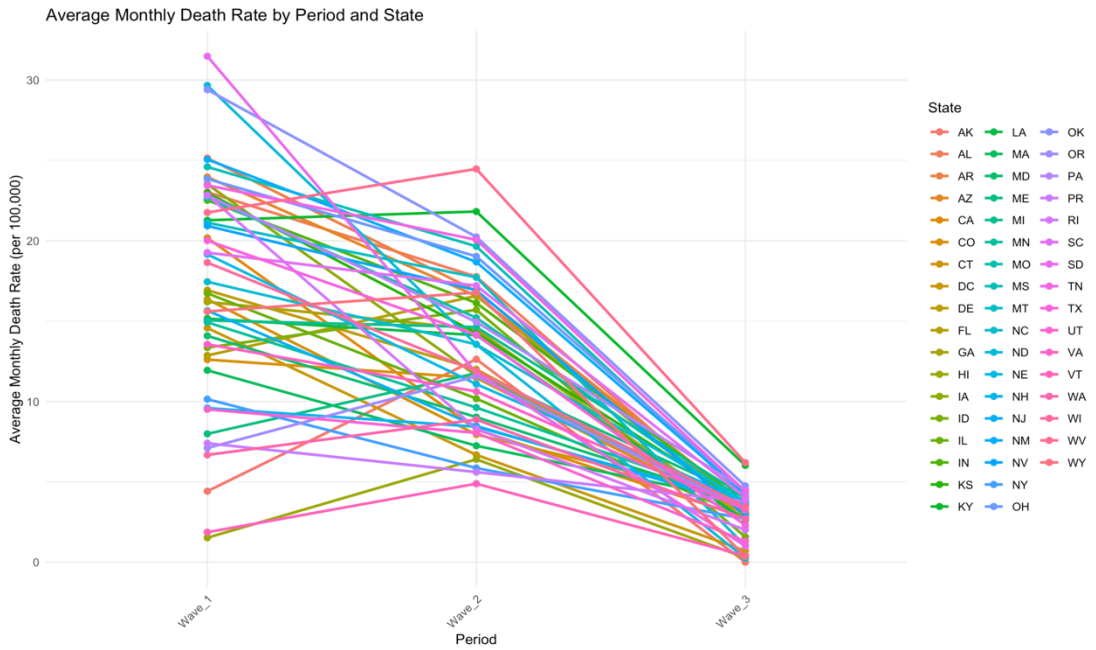

Figure 3: Average Monthly Death Rate (per 100,000)

Table 1 presents the sorted differences in average monthly death rates per 100,000 population between Wave 1 and Wave 2 for all U.S. states and territories. Negative values indicate a decrease in death rates, while positive values signify an increase in death rates. South Dakota (SD), North Dakota (ND), and Rhode Island (RI) exhibited the most significant reductions in death rates, with decreases of -19.7, -16.1, and -14.5 deaths per 100,000 population per month. Similarly, states like California (CA), Iowa (IA), and Oklahoma (OK) also showed considerable reductions, with declines ranging from -11.9 to -9.1. On the other hand, states like Alaska (AK), Hawaii (HI), and Oregon (OR) experienced the largest increases in death rates, with rises of 8.2, 4.9, and 4.5 deaths per 100,000 population per month. Other states, including Maine (ME) and Florida (FL), also showed notable increases.

Table 1: Sorted Differences Between Death Rates (Period 1 and 2)

#	State	Diff (1-2)	#	State	Diff (1-2)	#	State	Diff (1-2)	#	State	Diff (1-2)
1	SD	-19.6982	14	AZ	-7.4046	27	MA	-4.6968	40	LA	-1.0388
2	ND	-16.1165	15	NJ	-7.1779	28	NY	-4.2855	41	MI	-0.4141
3	RI	-14.5405	16	WI	-6.7582	29	NV	-4.0091	42	KY	0.5464
4	CA	-11.9249	17	IL	-6.5547	30	NC	-3.8943	43	WY	1.1810
5	IA	-11.9089	18	IN	-6.3953	31	MT	-3.4086	44	WA	2.1973
6	OK	-9.1630	19	NM	-6.3791	32	TN	-3.3927	45	ID	2.3493
7	KS	-8.7221	20	TX	-5.8656	33	VA	-2.9330	46	WV	2.7143
8	CT	-8.4084	21	MN	-5.3036	34	SC	-2.0575	47	VT	3.0195
9	AR	-8.2716	22	AL	-5.2634	35	PR	-1.7933	48	FL	3.7118
10	NE	-8.0794	23	MD	-5.0575	36	GA	-1.7062	49	ME	3.7724
11	DC	-7.8950	24	MS	-4.9457	37	UT	-1.4634	50	OR	4.4708
12	PA	-7.8717	25	DE	-4.9291	38	NH	-1.1623	51	HI	4.8811
13	MO	-7.4646	26	OH	-4.8218	39	CO	-1.1372	52	AK	8.2143

Table 1: Sorted Differences of Death Rates Between Wave 1 and Wave 2

Table 2 presents the sorted differences in average monthly death rates per 100,000 population between Wave 2 and Wave 3 across U.S. West Virginia (WV), Montana (MT), and Wyoming (WY) showed the most significant decreases in death rates, with reductions of -18.3, -16.7, and -16.4 deaths per 100,000 population per month. These states had substantial improvements in outcomes during Wave 3 compared to Wave 2. Similarly, states like Kentucky (KY), Tennessee (TN), and Mississippi (MS) also experienced large reductions, ranging from -15.8 to -15.5 deaths per 100,000 population. At the other end, Puerto Rico (PR), Massachusetts (MA), and New York (NY) exhibited the smallest improvements, with reductions of -1.8, -3.4, and -3.2 deaths per 100,000 population.

Table 1: Sorted Differences Between Death Rates (Period 2 and 3)

#	State	Diff (2-3)	#	State	Diff (2-3)	#	State	Diff (2-3)	#	State	Diff (2-3)
1	WV	-18.2723	14	AR	-13.3598	27	PA	-10.4429	40	WA	-6.1723
2	MT	-16.6920	15	ND	-13.2920	28	NC	-9.7780	41	HI	-6.1232
3	WY	-16.4033	16	SC	-13.1674	29	DE	-8.9395	42	DC	-5.9685
4	KY	-15.7971	17	AZ	-12.9372	30	CO	-8.4732	43	MN	-5.8336
5	TN	-15.6566	18	AK	-12.6281	31	WI	-8.4457	44	CA	-5.4733
6	MS	-15.5830	19	IN	-12.4510	32	OR	-8.3230	45	MD	-5.3624
7	OK	-15.4944	20	TX	-11.8491	33	IA	-7.9602	46	NJ	-4.9040
8	AL	-14.6366	21	GA	-11.8427	34	ME	-7.7237	47	NH	-4.7167
9	OH	-14.6107	22	MO	-11.6291	35	IL	-7.3588	48	VT	-4.4813
10	NM	-14.4445	23	LA	-11.5320	36	VA	-7.3452	49	CT	-4.0372
11	ID	-14.1271	24	SD	-10.7753	37	NE	-7.2905	50	MA	-3.4403
12	NV	-14.0117	25	KS	-10.6895	38	UT	-6.7477	51	NY	-3.1647
13	FL	-13.5817	26	MI	-10.5792	39	RI	-6.2926	52	PR	-1.8195

Table 2: Sorted Differences of Death Rates Between Wave 2 and Wave 3

Discussion

Using data visualization to analyze monthly case trends, the pandemic was divided into three distinct waves: Wave 1 (October 2020–March 2021), Wave 2 (August 2021–April 2022)**,** and Wave 3 (November 2022–March 2023). Each wave corresponds to significant events that influenced death rates of COVID-19 variants and the implementation of major public health interventions. For instance, the peak of Wave 2 coincided with the rapid spread of the Omicron variant, which caused a dramatic increase in infections but generally less severe outcomes due to widespread vaccination and prior immunity (Wikipedia contributors 2024). Dividing the pandemic into waves structured trends in death numbers influenced by healthcare impacts, facilitating a clearer understanding of the virus and the effectiveness of health strategies.

The result from Table 1 and Table 2 is different. States has high reduction in death rate of Table 1 is significantly different from the state has high reduction in death rate of Table 2. States that led in mortality reductions early may have faced new challenges in later waves, while others improved over time as they addressed gaps in vaccination or healthcare preparedness. For example, West Virginia (WV) ranked among the top six states with the largest increase in death rates from Wave 1 to Wave 2 but showed the most significant decrease in death rates from Wave 2 to Wave 3. These variations emphasize the dynamic and regionally specific nature of the COVID-19 pandemic response (Zhu, Jianyu, Chi Zhang, Mingqi Li, Li Zhou, Fengjiao Xu, and Yubin Zhang. 2024).

Significant differences between waves were found when state-level death rates were calculated. During Wave 1 to Wave 2, most of the state showed decline trend.

However, some states experienced worsening outcomes, such as AK with 8% of increasing death rate. From Wave 2 to Wave 3, all the state has decrease trend in death rate which shows the critical of state-level public health strategies, such as vaccination rates, healthcare infrastructure, and mitigation measures. States with robust vaccination coverage and healthcare capacity generally demonstrated improved performance across waves, while those facing challenges in public health response saw worsening trends. The findings demonstrated the necessity of specialized regional approaches to resolve inequalities in different states and enhance preparedness for upcoming public health crises (Wikipedia contributors. 2024).

The decreasing death rates in the majority of states from Wave 2 to Wave 3 suggested that virulence of COVID-19 was decreasing with time. This trend aligns with the availability of vaccines and advancements in treatment options during later period. However, the severity of the Omicron wave in Wave 2 demonstrated that new variants could increase virulence suddenly.

The result suggests that COVID-19 became less or more virulent across the different periods. This may due to increase in healthcare system capacity, vaccination rates, and public compliance played a critical role in determining outcomes. This indicates that virulence alone cannot fully explain the observed trends. It is necessary to consider with contextual factors (Island, September 15, 2021).

This study has also had several limitations. First, local differences in timing or intensity may not be captured by the pandemic's period division because it is based on national trends and population. Also, the population data is only for year to year. So, when calculate monthly death rate, it is not accurate. Second, the analysis depends on case and death statistics that have been recorded, which could not be accurate or consistent. Third, factors such as socioeconomic conditions, testing rates, and healthcare access were not explicitly modeled, limiting the ability to fully explain state-level disparities. Future research could address these limitations by incorporating bigger data and including some socioeconomic factors. Additionally, analyzing the effects of other variants would provide more deeper insights into pandemic management.

This study highlights the significance of analyzing trends and regional variances in pandemic outcomes. By dividing the pandemic into periods, analyzing state-level death rates, and evaluating changes in virulence, the study shows the need for adaptive, data-driven strategies to prevent pandemic crises. Governments can use these insights to design targeted interventions, ensuring more equitable and effective responses in future public health crises.

Reference

Island, Anna Sigridur, María Óskarsdóttir, Corentin Cot, Giacomo Cacciapaglia, and Francesco Sannino. 2021. *"Nordic Vaccination Strategies Face/Off via Age Range Comparative Analysis on Key Indicators of COVID-19 Severity and Healthcare Stress Level."* arXiv preprint arXiv:2109.11517. September 15, 2021. https://doi.org/10.48550/arXiv.2109.11517.

Wikipedia contributors. 2024. *"SARS-CoV-2 Omicron Variant."* Wikipedia. Accessed December 16, 2024. https://en.wikipedia.org/wiki/SARS-CoV-2_Omicron_variant.

Wikipedia contributors. 2024. *"COVID-19 Vaccine."* Wikipedia. Accessed December 16, 2024. https://en.wikipedia.org/wiki/COVID-19_vaccine.

Zhu, Jianyu, Chi Zhang, Mingqi Li, Li Zhou, Fengjiao Xu, and Yubin Zhang. 2024. *"Association Between Vaccination Rate and COVID-19 Case-Hospitalization Risk Across Variants: A Nationwide Analysis in the United States." BMC Public Health* 24: 17790. https://doi.org/10.1186/s12889-024-17790-w.