State-Level and Country-Level Analysis of COVID-19 in the United States

Abstract

COVID-19 has significantly impacted global health and society since its outbreak in 2019. This study investigates the dynamics effects of COVID-19 by analyzing death and case rates across four pandemic periods in the United States. Using weekly case and death data from the CDC and annual population data from the U.S. Census Bureau, we calculated rates per 100,000 people and divided the pandemic into distinct phases based on national trends and key events. Our analysis reveals the progression of the pandemic, from the initial high lethality of the virus to the influence of the Alpha and Omicron variants. State-level analyses highlight disparities, such as Alaska's low death rates due to sparse population and West Virginia's high rates influenced by an older population and limited healthcare resources. This research underscores the importance of tailoring public health strategies to regional characteristics and demonstrates how structured data analysis can provide actionable insights for pandemic preparedness and response.

Key words: COVID-19, death rates, case rates, data analysis

Introduction

In December 2019, a case of pneumonia of unknown etiology was detected in Wuhan, Hubei Province, China. Due to the high similarity between this novel virus and SARS-CoV (i.e., SARS coronavirus), the novel coronavirus (2019-nCoV) was named severe acute respiratory syndrome coronavirus 2, SARS-CoV-2 [1]. At the same time, the World Health Organization (WHO) announced that the disease caused by the SARS-CoV-2 virus is officially named COVID-19 (Corona Virus Disease 2019), which means "Coronavirus Disease 2019". The virus is spreading rapidly around the world, it has infected 4,806,299 people and caused 318,599 deaths in just five months [2]. As of November 2022, there have been about 600 million confirmed cases of COVID-19, including around 6 million deaths [3]. COVID-19 is highly contagious and has a high capacity for mutation, making it a constant challenge for humanity.

In the four years from the beginning of COVID-19 to the present, it has undergone multiple mutations and has been the leading cause of death. Although scientists and medical organizations have developed several vaccines and treatments, the rapid spread of the virus and the emergence of new variants often undermine the effectiveness of existing measures [4]. To date, although COVID-19 has been largely contained globally, it is still critical to gain an in-depth understanding of its impact on human health and social systems through rigorous scientific research and data analysis.

The death rate is a measure of the ability of a pathogen or virus to infect or destroy a host in an infectious disease and is described as the proportion of deaths in a defined relevant population, i.e., the percentage of cases that result in death. The death rate specifies the severity of the disease, and it is necessary for determining public health priorities and target interventions to reduce the severity of the risk [5]. This project aims to examine the entire COVID-19 pandemic by computing the death rate during different states in the United States. As the COVID-19 pandemic has gone through multiple stages of development since its outbreak, it is necessary to conduct a phased study of the different periods to analyze in depth the differentiated impacts it has had at each period. Meanwhile, given the significant differences in infectiousness and lethality among different strains in different period, we further introduced the measurement of the cases rate to more comprehensively support the relevant studies. This multidimensional analytical approach can help us to gain a deeper understanding of the impact of the COVID-19 on human health and society. To achieve the above goal, we calculate the death rate and case rate for each period based on the dataset from the U.S. Centers for Disease Control and Prevention (CDC), which records various deaths and cases for each U.S. state weekly, and the United States Census Bureau provides the population totals for each state yearly. We hypothesize that there would be significant differences in the death and case rates of different periods of the COVID-19 epidemic across U.S. states. Such research not only helps to comprehensively assess the immediate threat posed by COVID-19, but also provides science-based prevention and control strategies and theoretical support for future responses to similar public health crises.

Methods

Data Source

We conducted the study which was built on data from two different sources. The US COVID-19 Public dataset was obtained from the U.S. Centers for Disease Control and Prevention (CDC), which has the number of new cases and deaths between January 2020 and May 2023 due to COVID-19 recorded weekly at state level [6]. To further calculate for case and death rates, we used a dataset from U.S. National Population Totals and Components of Change: 2020-2023, created by the United States Census Bureau [7]. This dataset was compiled to provide annual population totals for each U.S. state from 2020 to 2023, which can be used to calculate the percentage of cases and deaths in the whole population.

Data Cleaning

To ensure the dataset is consistent, accurate, and comprehensive for analysis, a thorough data cleaning process was conducted after selecting the datasets. First, we retrieved COVID-19 cases data and processed to retain only relevant columns, including state, date, year, MMWR week, and case counts. MMWR week standardizes reporting for epidemiological purposes, assigning each week of the year a sequential number [8] and dates were formatted into the ISO-8601 standard [9], which can enable consistent comparisons across timeframes between deaths and cases datasets. Besides, we filtered the dataset to include only state-level data for the 50 U.S. states, the District of Columbia (DC), and Puerto Rico (PR). Next, similar preprocessing steps for cases data were applied to the COVID-19 deaths data. The state column was standardized to state abbreviations using mappings from state names to ensure consistency with the cases data, while ensuring DC and PR were appropriately represented. We then combined the cleaned cases and deaths data by matching on state, year, and MMWR week to create a unified dataset that captures both cases and deaths for each state and week.

To incorporate population data, we first cleaned the raw yearly population data to remove region-level data and unused columns, then renaming and converting relevant columns to numeric types. State names also were converted to abbreviations consistent with the cases and deaths dataset. Then, the population data was reshaped into a long format with columns for state, year, and population. The final step involved merging the combined case and death dataset with the cleaned population dataset based on state and year. This integration resulted in a comprehensive dataset with COVID-19 cases, deaths, and population data for all states and DC, PR across January 2020 to May 2023, enabling robust analyses of death and case rates.

Analytical Method

To examine the different effects of different strains of the COVID-19 virus in the United States, we divided the study period into some distinct phases based on trends observed in national case and death rates. Monthly case and death rates per 100,000 people were calculated by aggregating cases and deaths at the national level and dividing by the total population. Then two trend plots were created to visualize these rates, and we divided the pandemic period by analyzing the trends which contain the higher peak and combine studies on the key periods in the development of COVID-19.

Given the varying levels of infection and differing policies across states, we then calculated state-level death rates for distinct periods and visualized these variations using horizontal bar chart. This analysis not only highlights the effectiveness of prevention and control measures in each state but also underscores the regional disparities in infection severity. Finally, to assess changes in virulence across the different phases, national death and case rates were computed by summing cases and deaths in each period. A comparative bar plot was created to visualize these rates across periods and observe if COVID-19 became less or more virulent across the

different periods. Notably, as the population data we utilized is recorded on an annual basis, we used the average population for periods spanning multiple years to ensure the accuracy and comparability of rate estimations.

Results

From Fig.1, we can see the trend of cases and deaths rate per 100,000 people in the national level across the whole pandemic period. Both trend lines exhibit notable peaks in January 2021, September 2021, and January 2022. Additionally, the deaths rate graph demonstrates a distinct peak in April 2020. Therefore, we can divide the pandemic period into 4 waves, January 2020 to September 2020, October 2020 to June 2021, July 2021 to April 2022, and May 2022 to May 2023. The final wave May 2022 to May 2023, while lacking any notable peaks, represents the concluding phase of the pandemic. But studying this phase is crucial as it provides insights into the long-term effects of the pandemic and the factors that contributed to its eventual resolution.

Figure 1: Trend Plot of Cases, and Deaths

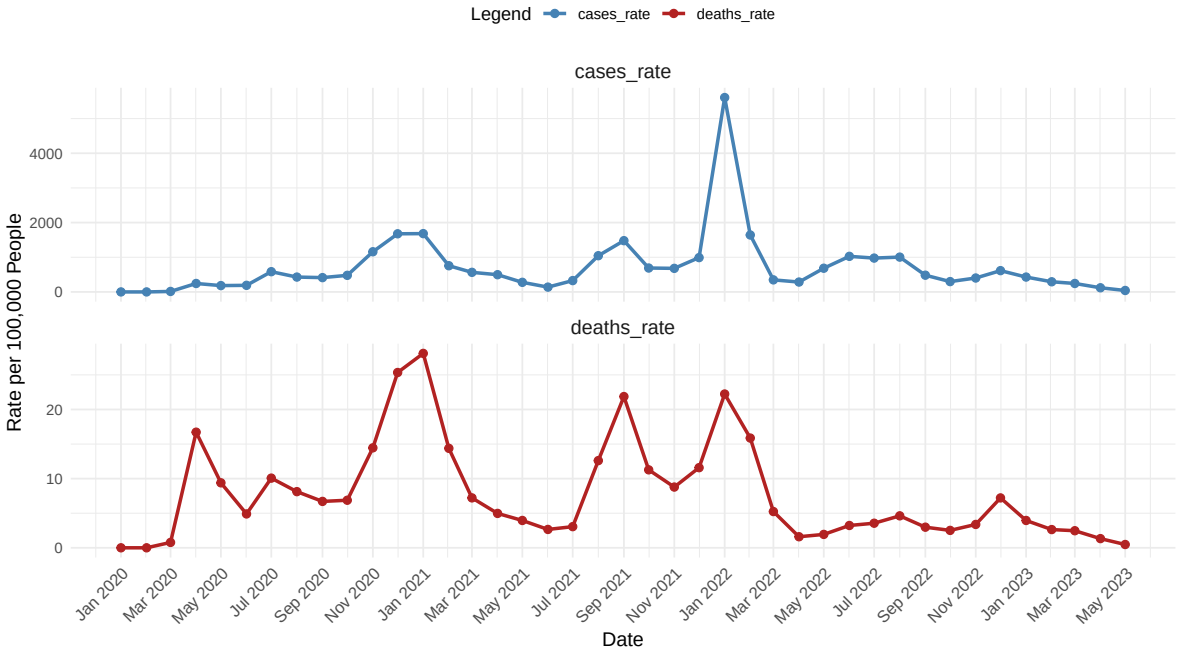


Fig.2 and Table 1 show the deaths rate per 100,000 people in the state level for each period: Period 1 (January 2020 to September 2020), Period 2 (October 2020 to June 2021), Period 3 (July 2021 to April 2022), and Period 4 (May 2022 to May 2023). For Period 1, the death rates are generally lower across most states, even AK and WY have the zero death rates. But with a few states MS, DC, LA, MA, CT, NJ, which exhibited relatively higher death rates (greater

than 100) compared to other states. NJ has the largest deaths rate and exceeds 150. For Period 2, it experiences a significant increase in death rates across most states, more than half of the states have death rates greater than 100. Only five states HI, VT, AK, WA, OR have death rates below 50. Notably, SD recorded the highest death rate, which is 181.49. Period 3 remains the high death rates compared to Period 2, and it also marks the highest rates of the whole pandemic. The state VT with the lowest death rate has reached 40.95, and the state WV with the highest death rate has reached over 200. The final period (Period 4) shows a marked reduction in death rates across all states, reflecting the decline in pandemic severity. The death rates across all states have dropped below 80, with AK once again reporting a zero death rate, similar to Period 1. Meanwhile, WV, despite a significant reduction in its death rate, remains the state with the highest mortality rate, as was the case in Period 3.

Figure 2: Deaths Rates by State for Each Period

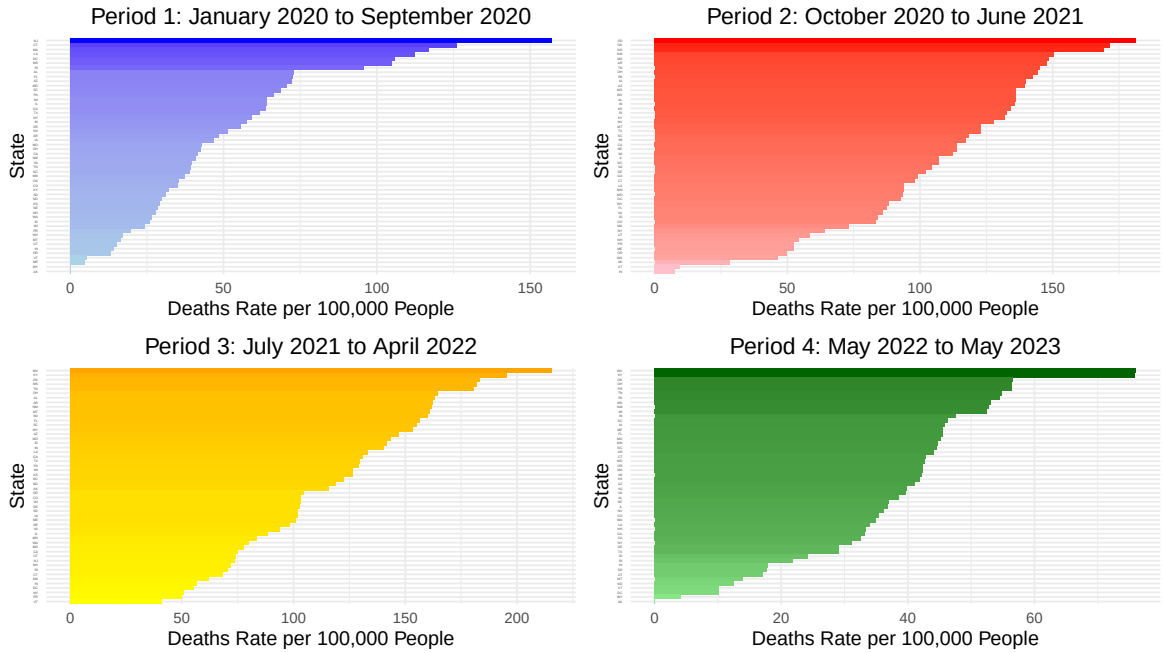


Table 1: Death Rate per 100,000 People for Each State in Each Period

Period 1		**Period 2**		**Period 3**		**Period 4**	
state	**death rate**	**state**	**death rate**	**state**	**death rate**	**state**	**death rate**
AK	0.00	HI	7.73	VT	40.95	AK	0.00
WY	0.00	VT	9.47	PR	49.92	WY	4.13
ME	4.69	AK	28.22	NY	50.82	DC	10.11
VT	5.29	WA	46.38	DC	55.08	VT	10.20
OR	13.22	OR	49.75	HI	56.72	ND	12.58
HI	14.13	ME	52.36	MA	62.16	MT	13.86

(Continued on next page)

(Continued from previous page)

Period 1		Period 2		Period 3		Period 4	
state	**death rate**	**state**	**death rate**	**state**	**death rate**	**state**	**death rate**
UT	15.04	PR	52.49	CT	68.11	UT	17.16
MT	16.37	NH	54.41	RI	70.34	SD	17.77
WV	17.08	UT	58.44	NH	71.73	HI	17.86
PR	19.72	NY	64.31	NJ	73.45	RI	21.84
WI	24.22	MA	73.26	UT	74.13	ID	24.24
ID	25.79	CO	83.19	CA	74.78	TX	29.07
WA	26.60	ID	84.12	MD	77.69	DE	29.07
NH	28.00	VA	86.01	WA	80.00	NY	31.10
NE	28.42	FL	87.54	MN	83.37	CA	32.59
KS	29.10	WY	88.35	IL	88.19	GA	33.29
ND	29.76	DC	92.74	VA	93.66	NH	33.29
SD	31.20	MD	93.43	NE	98.13	LA	34.02
KY	32.25	MN	93.99	ME	100.84	WA	34.87
CO	35.21	LA	94.00	IA	101.79	CO	35.43
OK	35.31	CT	98.20	SD	101.84	NV	36.20
MN	37.30	GA	99.34	DE	102.62	IL	36.90
NC	38.99	DE	102.02	WI	102.97	NE	36.97
TN	39.24	NJ	104.28	CO	102.98	AL	38.60
VA	39.54	NC	107.26	OR	104.51	VA	39.68
NM	40.93	IL	107.26	AK	115.60	NJ	39.91
CA	41.71	WI	112.36	ND	118.85	AZ	41.17
OH	42.57	NE	113.70	NC	122.37	KS	41.97
MO	42.91	CA	113.81	KS	126.33	WI	42.14
IA	46.70	MI	117.17	MI	126.37	MA	42.34
AR	48.30	SC	118.26	PA	129.18	OR	42.43
NV	51.22	TX	122.75	TX	129.48	MD	42.66
DE	55.45	MT	122.93	GA	131.11	CT	42.81
IN	57.62	NV	127.80	LA	133.11	AR	44.18
NY	59.03	KY	131.84	IN	140.18	NC	44.60
TX	61.92	RI	132.57	ID	141.70	MN	44.75
GA	63.68	KS	134.14	MO	143.43	MO	45.28
IL	63.93	IN	135.63	AZ	147.01	FL	45.50
MI	64.09	AL	136.17	WY	153.23	ME	45.56
PA	66.30	WV	136.18	SC	155.01	IA	45.85
SC	68.70	MO	136.28	FL	156.39	SC	46.29
MD	70.56	AZ	139.66	NV	160.14	IN	47.51
AZ	72.34	IA	139.85	MT	161.08	MI	52.53
FL	72.44	PA	142.40	NM	161.80	NM	52.75

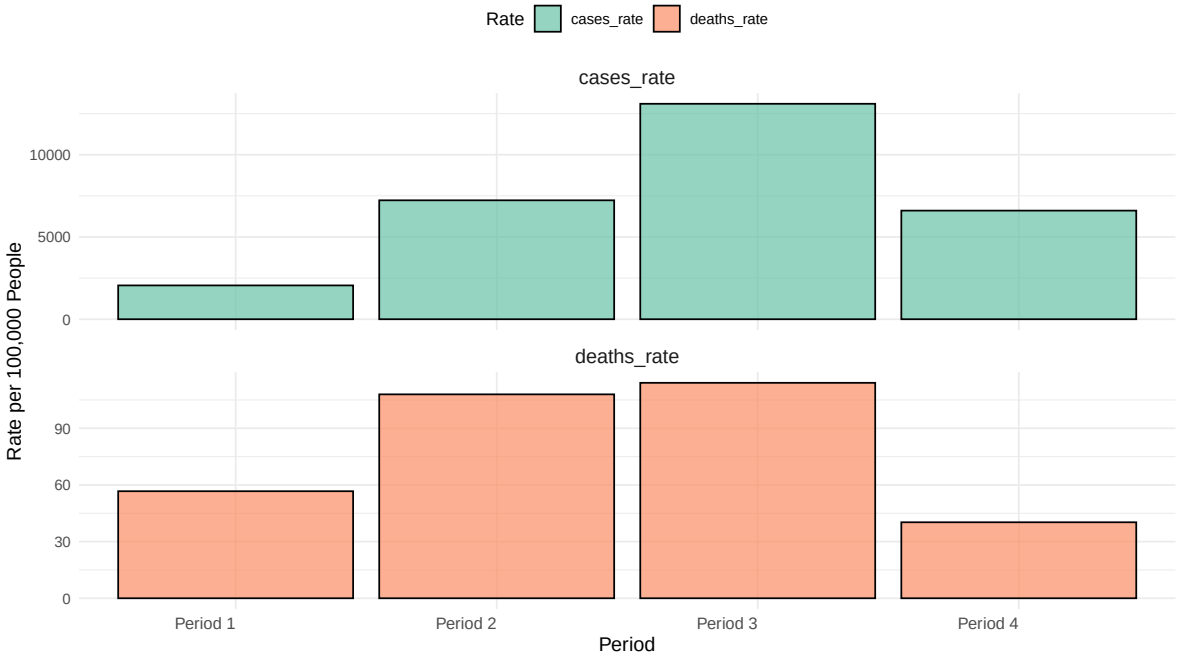
(Continued on next page)

(Continued from previous page)

Period 1		Period 2		Period 3		Period 4	
state	**death rate**	**state**	**death rate**	**state**	**death rate**	**state**	**death rate**
AL	72.96	OH	144.34	AR	162.05	MS	53.14
RI	95.67	TN	144.97	AL	163.07	PA	54.59
MS	104.82	AR	147.95	OH	164.65	TN	54.82
DC	105.69	MS	148.69	TN	180.72	PR	56.41
LA	112.40	NM	150.56	MS	181.82	OH	56.51
MA	116.94	ND	169.23	OK	183.12	OK	56.66
CT	126.06	OK	171.62	KY	195.29	KY	75.92
NJ	157.06	SD	181.49	WV	215.62	WV	76.09

Fig.3 illustrates the death rates and case rates per 100,000 people at the national level across four distinct pandemic periods. Both plots reveal that Period 3 recorded the highest case rate and death rate, followed by Period 2 as the second highest for both rates. However, the two plots differ in that Period 1 exhibits a lower case rate than Period 4, whereas Period 4 shows a lower death rate compared to Period 1. Additionally, the case rate experienced a significant increase from Period 2 to Period 3, while the death rate showed a comparatively smaller rise during the same period. Moreover, both the case rate and death rate exhibit a significant increase from Period 1 to Period 2, followed by a notable decrease from Period 3 to Period 4. Therefore, at Period 1 and Period 2, COVID-19 became more virulent, and when entering Period 3 to Period 4, it became less virulent.

Figure 3: Deaths Rates and Cases Rates in the Country Level for Each Period



Discussion

Our findings align with the progression of the COVID-19 vaccine development. Each of the four periods represents a distinct stage in the evolution of the pandemic. Period 1 corresponds to the early phase of the outbreak, marked by a lack of effective countermeasures and the absence of a developed vaccine. During this period, the high lethality of the initial virus led to a pronounced peak in the death rate, despite a relatively low case rate [10]. Period 2 reflects the phase dominated by the Alpha variant. Although the outbreak was temporarily contained following the measures implemented in Period 1, the Alpha variant's higher transmissibility caused a significant rise in both the case rate and the death rate [11]. Notably, this period also saw the development and rollout of the Pfizer-BioNTech and Moderna vaccines, which played a pivotal role in regaining control over the pandemic [12]. As we entered Period 3, the outbreak's complexity and variability became increasingly evident. This phase witnessed a new wave of infections driven by the Omicron variant, characterized by its higher transmissibility and immune escape capabilities [13]. The rapid global spread of the Omicron strain led to a sharp surge in case rate. However, the Omicron variant is not highly lethal [14], so the death rate increased at a significantly slower rate compared to earlier periods. Period 4 marked the winding down of the pandemic, as a significant portion of the population had already developed antibodies [15], so case rate and death rate both have the significant decrease, many countries began to relax their preventive measures, and economic and social activities gradually returned to normal.

In addition, the state-level analysis highlights significant disparities in demographic characteristics and healthcare resources across states. For instance, Alaska (AK) exhibits a notably low overall death rate, with no deaths recorded during Periods 1 and 4. This can likely be attributed to Alaska is one of the most sparsely populated states in the U.S. [16], its low population density effectively reduces opportunities for virus transmission. Moreover, Alaska's remote geographic location may have delayed the introduction of the virus, affording residents and healthcare systems valuable time to learn from COVID-19 in other states and implement more effective response strategies. Conversely, West Virginia (WV) demonstrates a contrasting trend, with a relatively low death rate in Period 1 but consistently high death rates in the subsequent three periods and has the highest death rates during the last two periods. This can be explained by West Virginia includes a disproportionately high percentage of elderly individuals [17], they are particularly vulnerable to severe illness and death from COVID-19 [18]. Furthermore, many areas within the state are rural and remote, with limited access to healthcare resources and lower levels of economic development [19]. These challenges likely overwhelmed the local healthcare system during COVID-19, contributing to a higher death rate. These state-level variations underscore the unique challenges faced by different regions in addressing the COVID-19 pandemic and highlight the critical need for tailored public health strategies.

There also have some limitations in our analysis: The population data we collected is annual, which introduces potential inaccuracies when calculating rates, as population data can change

daily. In our analysis, we assumed that the population of each state remains constant throughout the year. Additionally, for periods spanning two different years, we simply used the average of the populations as the population in this period. We also assume that all the data collected is accurate and no errors, however, this is difficult to achieve in reality. Some deaths from other causes may have been mistakenly classified as COVID-related deaths. Future research could consider utilizing daily population data and incorporating methods to address issues of misclassification.

To summarize, in this project, we adopted a structured analytical approach comprising data cleaning, aggregation, and visualization. This method allows for a systematic analysis of case and death rates over time and across states, providing insights into the dynamics of the pandemic, informing public health responses and future preparedness planning.

Reference

[1] Yuki, K., Fujiogi, M., & Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. Clinical immunology, 215, 108427.

[2] World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 97. In Coronavirus disease 2019 (COVID-19): situation report, 97.

[3] World Health Organization. (n.d.). WHO Coronavirus (COVID-19) dashboard. Retrieved December 9, 2024, from https://covid19.who.int/

[4] Mohapatra, R. K., Pintilie, L., Kandi, V., Sarangi, A. K., Das, D., Sahu, R., & Perekhoda, L. (2020). The recent challenges of highly contagious COVID-19, causing respiratory infections: Symptoms, diagnosis, transmission, possible vaccines, animal models, and immunotherapy. Chemical biology & drug design, 96(5), 1187-1208.

[5] Khafaie, M. A., & Rahim, F. (2020). Cross-country comparison of case fatality rates of COVID-19/SARS-COV-2. Osong public health and research perspectives, 11(2), 74.

[6] Centers for Disease Control and Prevention. (n.d.). Data.CDC.gov: CDC open data portal. Retrieved December 12, 2024, from https://data.cdc.gov/

[7] U.S. Census Bureau. (n.d.). National population totals: 2020-2023. Retrieved December 12, 2024, from https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-total.html

[8] Centers for Disease Control (US), Centers for Disease Control, & Prevention (US). (2007). Morbidity and mortality weekly report: MMWR (Vol. 56, No. 30-50). US Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control.

[9] Houston, G. (1993, January). ISO 8601: 1988 Date/Time Representations.

[10] Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., & Ho, R. C. (2020). Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. International journal of environmental research and public health, 17(5), 1729.

[11] Eyre, D. W., Taylor, D., Purver, M., Chapman, D., Fowler, T., Pouwels, K. B., … & Peto, T. E. (2022). Effect of Covid-19 vaccination on transmission of alpha and delta variants. New England Journal of Medicine, 386(8), 744-756.

[12] Dighriri, I. M., Alhusayni, K. M., Mobarki, A. Y., Aljerary, I. S., Alqurashi, K. A., Aljuaid, F. A., … & Almutairi, A. N. (2022). Pfizer-BioNTech COVID-19 vaccine (BNT162b2) side effects: a systematic review. Cureus, 14(3).

[13] Andrews, N., Stowe, J., Kirsebom, F., Toffa, S., Rickeard, T., Gallagher, E., … & Lopez Bernal, J. (2022). Covid-19 vaccine effectiveness against the Omicron (B. 1.1. 529) variant. New England Journal of Medicine, 386(16), 1532-1546.

[14] Burki, T. K. (2022). Omicron variant and booster COVID-19 vaccines. The Lancet Respiratory Medicine, 10(2), e17.

[15] Ioannidis, J. P. (2022). The end of the COVID-19 pandemic. European journal of clinical investigation, 52(6), e13782.

[16] Fleming, M. D., Chapin III, F. S., Cramer, W., Hufford, G. L., & Serreze, M. C. (2000). Geographic patterns and dynamics of Alaskan climate interpolated from a sparse station record. Global Change Biology, 6(S1), 49-58.

[17] Sutphin, R. (2024). 2023 West Virginia Rural Health Conference Abstracts. Marshall Journal of Medicine, 10(2), 6.

[18] Pant, S., & Subedi, M. (2020). Impact of COVID-19 on the elderly. Journal of Patan Academy of Health Sciences, 7(2), 32-38.

[19] Brisbin, R. A., Kilwein, J. C., & Plein, L. C. (2024). West Virginia politics and government. U of Nebraska Press.