Analysis of COVID-19 phases and virulence 2020-2024*

Charles Bray†

20 December, 2024

Abstract

This study aimed to assess the evolution of the SARS-CoV 2 virus, its associated disease burden, and the public health response. Data from the CDC on hospitalizations, deaths, and cases were analyzed and divided into phases based on these metrics, and a novel approach using a sliding window of 9 weeks was developed to define pandemic waves. Seven waves of the pandemic were identified from January 2020 to December 2024, over which case, hospitalization, and death rates varied across states and regions. The virulence of COVID-19 decreased over time, with the mortality-to-case ratio dropping from 0.06 in the first wave to 0.01 in later waves while the mortality-to-hospitalization ratio dropped from 0.5 to 0.1. This study highlights the importance of standardized definitions for pandemic phases and provides insights into regional public health responses and their consequences. These definitions and associated state-level performance measures provide valuable historical context for future pandemic readiness.

Keywords: R, LaTeX, Quarto

*Thank you, BST 260 teaching team, for your hard work throughout the semester.

†SM Candidate, Department of Biostatistics, Harvard TH Chan School of Public Health, bray@hcp.med.harvard.edu

1 Introduction

The COVID-19 pandemic, in which the world saw the emergence of an acute upper respiratory infection caused by the novel coronavirus SARS-CoV-2, has been a major cause of death and disability in the United States and globally since the first months of 2020. Over 1 million deaths are reported through time of writing in the United States, according to the World Health Organization, while millions are suspected to have diminished quality of life from the still poorly-understood set of conditions associated with "long COVID". (Organization [7]) (Mirin [5]). It is popular to describe the net consequences of the pandemic but less so to assess the evolution of the virus and the evolving ability of our public health and medical sectors to respond to the morbidity and mortality it caused, as measured through deaths, hospitalizations, and rates of infection.

Systematizing the way in which we assess individual phases of a pandemic is important for several reasons:

Standardized definitions: Researchers can compare data and findings in a consistent way rather than contrasting conclusions that rely on novel estimations of pandemic phases.

Inter-regional comparability: With consistent wave definitions, it is possible to compare regional or state-level responses within the U.S. This allows for the identifcation of what works well and what needs to improve as far as local public health responses (e.g., were school lock-downs too long or too short in certain areas, what was the effect of masking during each phase of the pandemic, etc.).

Clear historical record of pandemic trajectory: By clearly defining waves, a framework is created for understanding how the pandemic changed over time, which allows us to contextualize the public health response based on how the virus's behavior and human responses evolved.

Future pandemic readiness: The identification of distinct waves gives public health officials and other relevant decisionmakers historical context that leads to better planning and resource allocation in anticipation of future COVID-19 surges or other viruses that may behave similarly.

Some investigators have aimed to address this gap in knowledge by defining "waves", or distinct phases characterizing the COVID-19 pandemic by the pattern or trajectory of infection as well as the

predominant viral strain. For instance, Iran's experience of the pandemic was characterized by five distinct phases according to Amin et al. (2022) through the end of 2022, as defined by public health communications and increases in case counts following times of stability. (Rozhin Amin [10]). Another group of researchers in Chile defined the start and end of waves based on sustained increase or decrease as well as case rates achieving a certain threshold. (Andres Ayala [1]). Other researchers employed more sophisticated methods, including Break Least Squares, to define two waves in 2020 in North America each lasting roughly 50 days. (Ranjula Bali Swain [9]). However, little research exists to clarify times of transition between waves of COVID-19 in the United States, strictly speaking.

To this end, it is necessary to provide a conceptualization of the distinct stages of the COVID-19 pandemic as it affected the US population, as well as the relative performance of different areas within the US. This paper aims to describe the changing dynamics of the novel coronavirus as such.

2 Methods

Data were obtained from the Center for Disease Control and Prevention (CDC) via API, comprising hospitalizations, deaths, and cases associated with COVID-19 infection and reported weekly at the US state-level. (CDC [2]). States were defined as the 50 standard US states along with DC and Puerto Rico, for a total of 52 geographical units; states were assigned to regions of the US based on a classification scheme provided on the BST 260 course GitHub repository. (Rafael Irizarry [8]).

The study proceeds in three parts. First, the COVID-19 pandemic was split broadly into different phases based on counts of cases, deaths, and hospitalizations across regions of the United States. Following this, the performance of individual US states was described along these measures within each determined wave. Finally, the nature of COVID-19 strains (their virulence and/or strain on hospitals) is determined by comparing the evolution of different measures from early to later waves.

In order to describe the trajectory of the COVID-19 pandemic, a novel approach to defining break-points between infection regimes (or "waves" of COVID-19) is developed. A sliding window of 9 weeks (or roughly 2 months) was applied to each measure (hospitalizations, cases, and deaths) within region, in which window were determined if the center value (the 5th index) of the window was a maximum or minimum value. Then, if at least half of the regions reporting data (typically, 10 regions) experienced

a local minimum within 5 weeks of each other (or, half the minimum/maximum search window), the current wave was determined to have ended. Instead of over-weighting high-population regions, this approach appreciates potential regional segmentation of the pandemic when evaluating country-level phases of viral spread.

Prior to algorithm development, (1) break least squares as well as (2) Markov Regime Switching (MRS) models were considered (Ranjula Bali Swain [9]). Ultimately, the complexity of these models–and in the case of MRS, the infeasibility of constraining wave definitions to just two regimes–motivated the development of this data-driven identification strategy. Other, more simple, strategies such as that featured in Ayala et al., 2021 (Andres Ayala [1]), were also considered but determined to be too inflexible. Namely, hard thresholds for number of cases per population were used in defining transitions in infectious regimes, which were seen here as not considerate of changes in testing ubiquity and changing virulence of different generations of COVID-19.

3 Results

Defining waves

The COVID-19 pandemic in the US was divided into 7 waves from January 1, 2020 through December 1, 2024 according to case and hospitalization rates. The choice of these two measures as opposed to deaths was determined to be appropriate based on the persistent lag of deaths data relative to cases and hospitalizations, and the fact that death timing relative to time of infection is known to exhibit much greater variance than symptoms or hospitalization timing relative to time of infection. (David Baud [3]). See Figure 1 below for a visualization of wave boundaries, where solid black dots indicate local minima in per-population rates of cases or hospitalizations by region, while black crosses indicate local maxima. Each dotted red line shows the inter-regional wave cutoff (where one wave ends and another begins); there is no assumed interim period between waves (i.e., one wave ends the day before another begins).

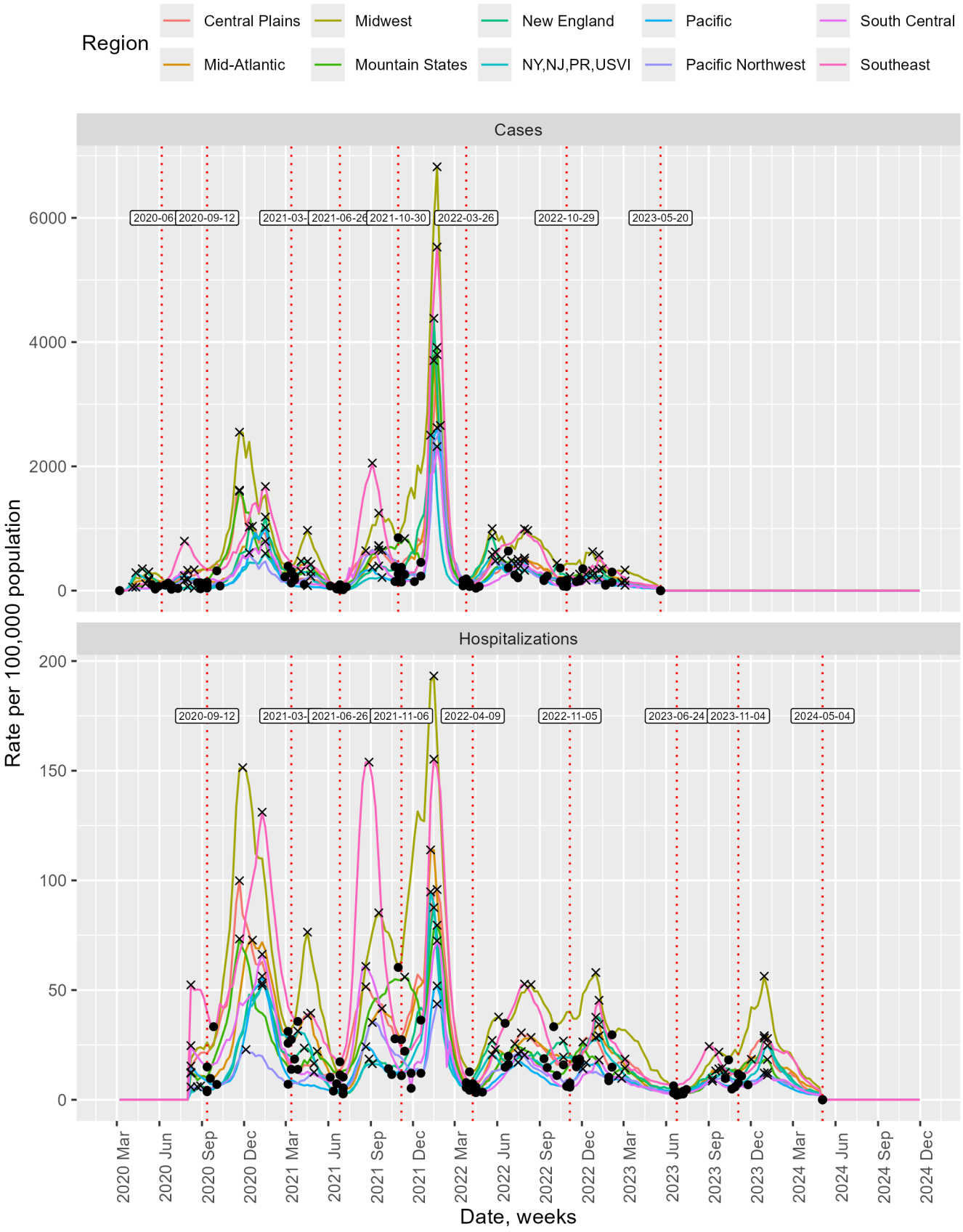


Figure 1: Definition of candidate COVID-19 waves based on regional minima and maxima

The first wave occurred roughly March 2020 to June 6, 2020, which we classify as the "Originator" wave. This was succeeded by the Second and Third waves, which occurred 2020-06-07 to 2020-09-12 and 2020-09-13 to 2021-03-13, respectively. The fourth wave, which we can call the Delta wave based on known speciation ("Early Emergence Phase of SARS-CoV-2 Delta Variant in Florida, US" [4]), occurred 2021-03-14 to 2021-06-26. This was eventually followed by the extremely transmissible Omicron variant, which characterized the wave lasting 2021-10-31 to 2022-03-26. Finally, while there continued to be local minima and maxima that could be considered to delineate further waves, it was decided that the low number of deaths along with moderation in number of hospitalizations, and regional segmentation in terms of dispersion around nationwide wave end dates (see Figure 2), collectively characterized a "post-acute" phase of the pandemic, starting at 2024-12-31 and extending through the end of data collection, December 1, 2024. This was driven by vaccination efforts and milder Omicron subvariants prevalent at the time and to this day. (*Omicron and its Subvariants: A Guide to What We Know* [6]).

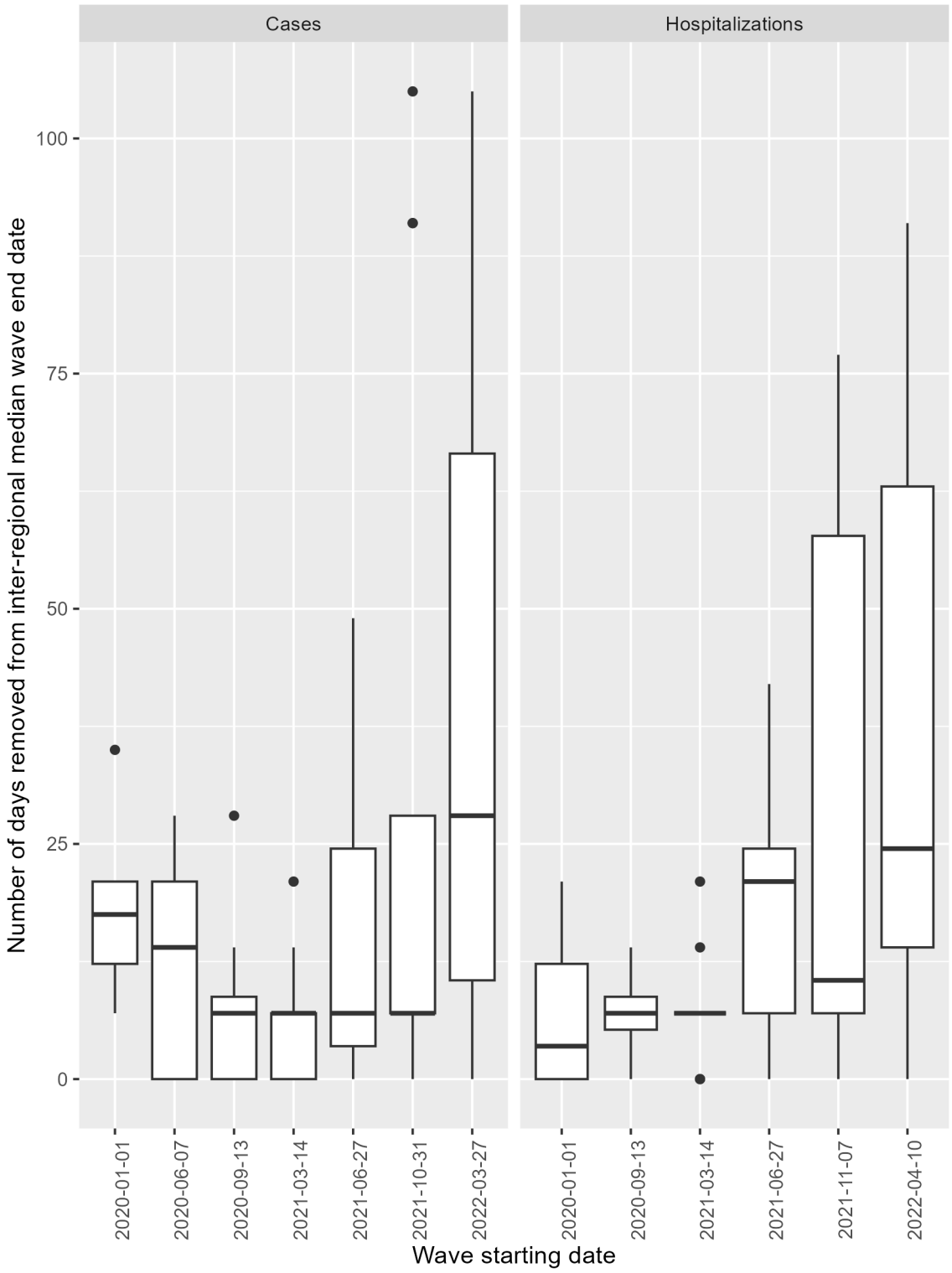


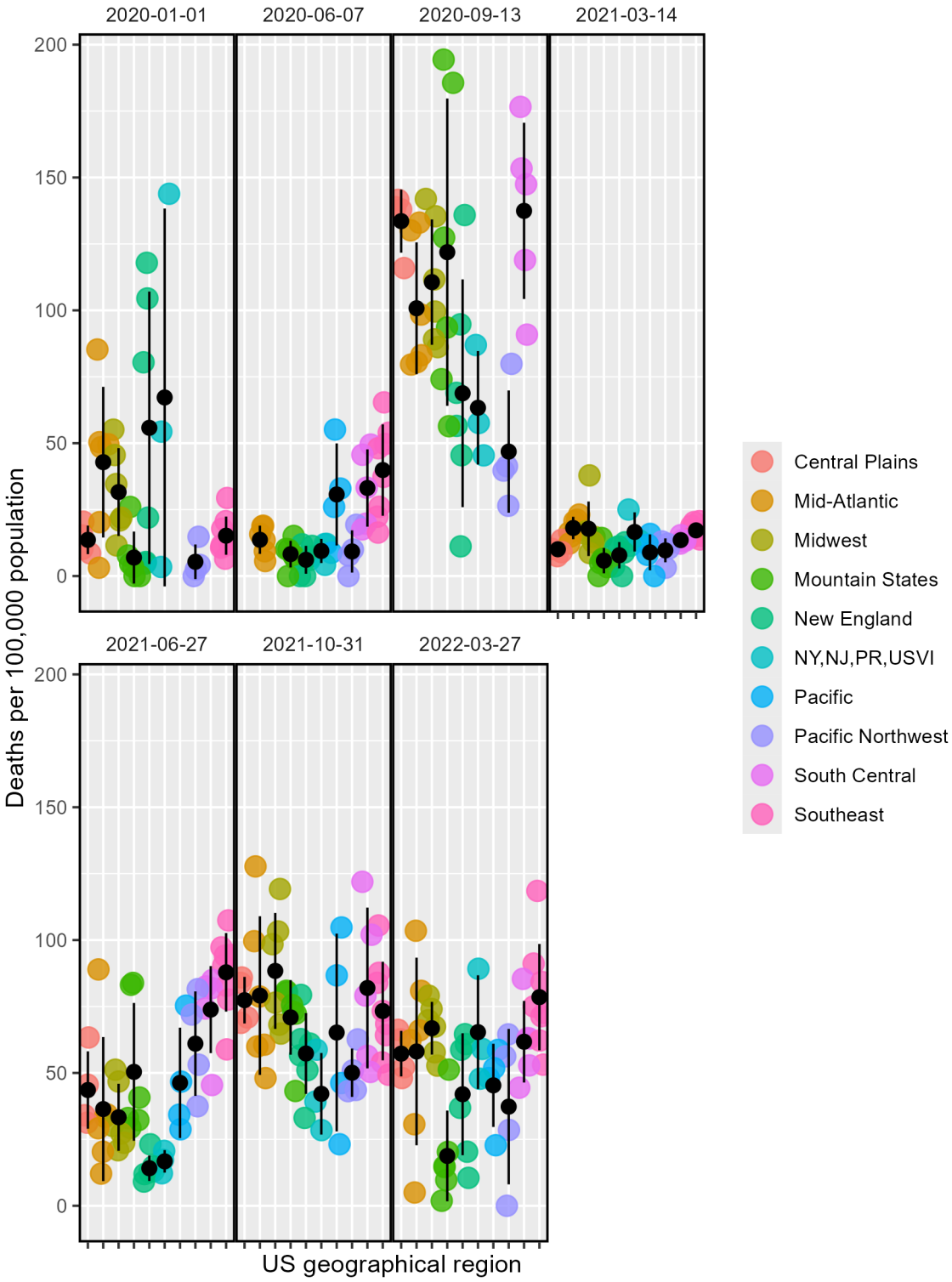
Figure 2: Dispersion of regional minima and maxima about national wave end dates (cases)

State performance

U.S. states exhibited wide variation in case, hospitalization, and death rates relative to each other in

each wave, with the relative performance changing from earlier to later waves as regional public health responses diverged.

For instance, New England states and mid-Atlantic states experienced the greatest burden in terms of deaths and hospitalizations during the first wave of the pandemic, being highly urban, older, and globally connected than many other areas of the country. The states of the South Central and Southeast, on the other hand, had the worst mortality in the second as well as latest waves (see Figure 3 below).



The top five states in terms of deaths per population were, in order, NJ, CT, MA, DC, RI. Seasonal change–i.e., people moving inside at greater rates in hotter months–brought higher mortality to southern and southwestern states in the subsequent wave of 2020, with MS, AZ, FL, SC, and TX occupying the top five spots through September of that year. The next wave that coincided with the winter of 2020/21 saw

SD, ND, OK, AR, and NM experience the worst state death rates, suggesting that state-level pandemic response was beginning to play a role in outcomes. By the final phase of the pandemic (the Omicron subvariants that began to predominate in spring 2022) this pattern was shown to persist, with states such as Kentucky and West Virginia consistently occupying the top five rankings of COVID-19-related mortality while states initially affected the most–many in New England and the mid-Atlantic–no longer appearing.

Table 1: State-level COVID-19 mortality rank (top 10), by wave start date

Mortality (rank)	Wave starting 2020-01-01	Wave starting 2020-06-07	Wave starting 2020-09-13	Wave starting 2021-03-14	Wave starting 2021-06-27	Wave starting 2021-10-31	Wave starting 2022-03-27
1	NJ	MS	SD	MI	FL	WV	KY
2	CT	AZ	ND	NJ	MS	NM	WV
3	MA	FL	OK	PA	AL	OH	MS
4	DC	SC	AR	WV	TN	KY	PR
5	RI	TX	NM	FL	KY	AZ	OK
6	MI	AL	OH	MD	WV	MI	TN
7	NY	LA	IA	TN	AR	OK	PA
8	PA	GA	KS	MS	MT	PA	OH
9	MD	AR	MO	KY	WY	IN	FL
10	DE	NV	RI	IL	SC	TN	MI

Virulence

The predominant strain of COVID-19 that characterized each wave became less virulent, comparing earlier waves to later waves. The original wave resulted in a mortality : case ratio of 0.06 while the first wave of Omicron exhibited a ratio of 0.01. By the approaching-endemicity phase of the pandemic, which we call post-acute, the ratio had become 0.01 when observing the period March 26, 2022 to October 29, 2022, for which data reporting is robust. The corresponding hospitalization : case rate ratios were, respectively, ∞, 0.17, and 0.09. See Figure 4 below for trendlines of virulence.

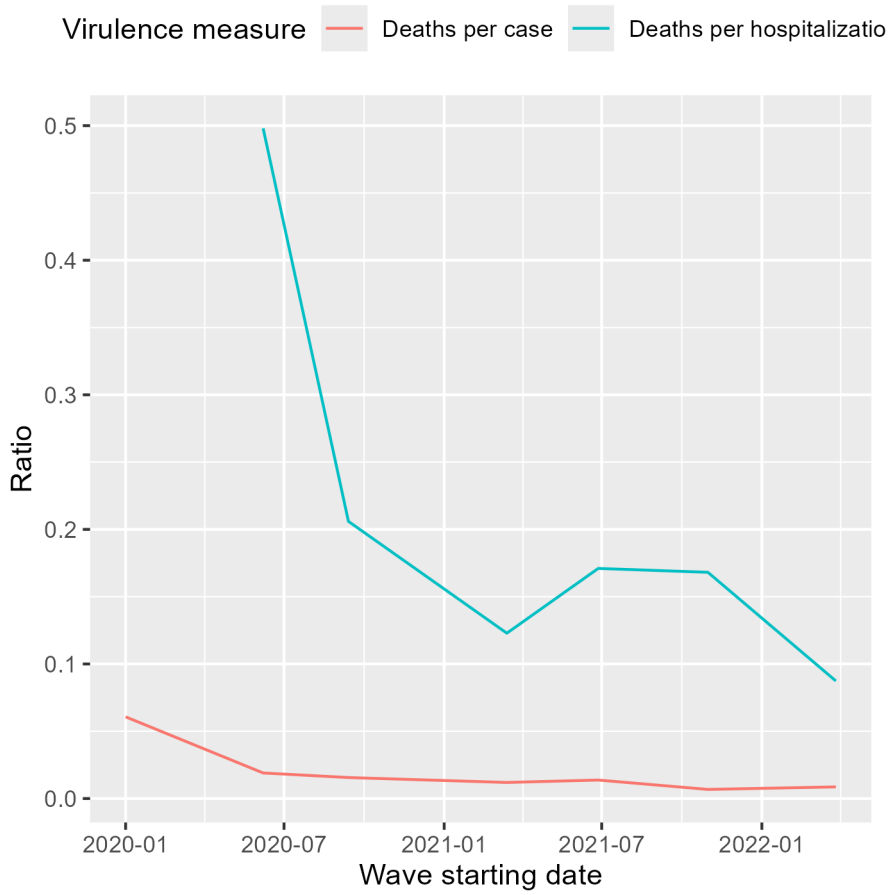


Figure 3: COVID-19 virulence by wave

4 Discussion

This study provides a novel framework for understanding the progression of the COVID-19 pandemic in the United States by defining waves using a data-driven approach and evaluating regional and state-level variations in pandemic dynamics. By identifying seven distinct waves between January 2020 and December 2024, the findings contribute to a nuanced understanding of how COVID-19 evolved over time in terms of case rates, hospitalizations, and virulence.

Strength of approach

The definition of pandemic waves using a sliding window approach highlights the importance of incorporating regional heterogeneity when evaluating national-level patterns. Unlike prior studies that used fixed thresholds (e.g., Ayala et al., 2021) or complex statistical models (e.g., Bali Swain, 2020), this method allowed for a more adaptable characterization of transitions between waves. The observed

alignment of regional minima and maxima underscores the dynamic relationship between viral evolution, public health interventions, and sociocultural factors across the United States.

Regional stratification

The analysis of state-level mortality and hospitalization trends further illustrates significant regional disparities, with urbanized states like New Jersey and Connecticut experiencing the highest mortality during the early waves, while southern states such as Mississippi and Kentucky bore the brunt of later waves. This shift probably is reflective of the differences in pandemic response measures, and in particular vaccination rates, or healthcare infrastructure and population behaviors. The findings also demonstrate a clear decline in virulence over the course of the pandemic, with mortality-to-case and hospitalization-to-case ratios decreasing from the initial wave to the post-acute phase dominated by Omicron subvariants. This trend aligns with previous research suggesting that increased population immunity and the reduced severity of later variants contributed to a milder disease profile (Baud et al., 2020; CDC, 2022).

Implications of findings

These results have critical implications for public health planning and pandemic preparedness. The adaptability of the wave-detection framework offers a potential tool for monitoring future pandemics, particularly in scenarios where regional variations and evolving virus characteristics must be accounted for. Additionally, the state-level disparities in outcomes emphasize the need for targeted interventions to address vulnerabilities in specific regions, especially during the early phases of a pandemic when resources are limited.

Limitations of the approach

Several limitations of this study warrant discussion. First, the reliance on CDC-reported data may introduce biases related to under-reporting or inconsistencies in data collection across states and time periods. Second, the study's wave-detection algorithm, while robust, is inherently sensitive to the choice of parameters (namely, window size and local minima rather than other indicators of change or trend) and may not capture subtle transitions between waves. Third, this analysis does not directly incorporate external factors such as policy changes, socioeconomic variables, or mobility data, which could further contextualize the observed patterns.

Future directions for research

Future research should aim to refine the wave-detection methodology by utilizing regularly updated data streams or dashboards (such as the COVID-19 dashboard published by the CDC) as well as by relying on machine learning techniques to increase predictive power for wave initiation, peaking, and transition into further viral regimes. Additionally, longitudinal studies that can incorporate demographic, healthcare utilization, and even behavioral data could provide deeper insights into what drove regional disparities. For instance, preexisting relationships with primary care providers, masking adherence, and cultural practices like multigenerational living. Finally, examining the impact of specific public health interventions and vaccination campaigns on wave dynamics would offer valuable lessons for managing future pandemics.

For the reasons above, a robust, standardized definition of pandemic phases provides a useful framework to generate insights into regional public health responses and their consequences as well as gives a starting point for future studies of the progression of COVID-19 in the U.S. and globally.

References

[1] Matilde Maddaleno Andres Ayala. "Identification of COVID-19 Waves Considerations for Research and Policy." In: *International Journal of Environmental Research and Public Health* 18.11058 (2021).

[2] CDC. *Data — Centers for Disease Control and Prevention*. https://data.cdc.gov/. Accessed: 2024-11-1. 2024.

[3] Guillaume Favre David Baud. "Real estimates of mortality following COVID-19 infection." In: *The Lancet Infectious Diseases* 20.7 (2020).

[4] "Early Emergence Phase of SARS-CoV-2 Delta Variant in Florida, US." In: ().

[5] Arthur Mirin. "A preliminary estimate of the economic impact of long COVID in the United States." In: *Fatigue Biomedicine, Health & Behavior* 10.4 (2022).

[6] *Omicron and its Subvariants: A Guide to What We Know*.

[7] World Health Organization. *WHO COVID-19 dashboard*. https://data.who.int/dashboards/covid19/deaths. Accessed: 2024-12-20. 2024.

[8] Rafael Irizarry. *github.com/datasciencelabs*. https://github.com/datasciencelabs/2024/tree/refs/heads/main/data. Accessed: 2024-11-1. 2024.

[9] Fan Yang Wallentin Ranjula Bali Swain Xiang Lin. "COVID-19 pandemic waves Identification and interpretation of global data." In: *Heliyon* 10.3 (2024), e25090.

[10] Khatereh Hannani Rozhin Amin. "Five consecutive epidemiological waves of COVID-19 a population-based cross-sectional study on characteristics, policies, and health outcome." In: *BMC Infectious Diseases* 22.906 (2022).