Analysis of the COVID-19 Case Rate and Excess Mortality Rate by Pandemic Wave

Yiqiao Zhu[1], Shuoyuan Gao[1]

[1] Department of Biostatistics at the University of Michigan School of Public Health

Abstract :

COVID-19 pandemics are organized into multiple waves, each with temporally and geographically variable transmission and mortality. Understanding these dynamics is critical for assessing public health responses and future preparedness. We obtained weekly state-level data on COVID-19 cases, deaths, and excess mortality spanning January 2020 to April 2025. We calculated population-standardized rates and visualized temporal trends to identify three major pandemic waves. We then modeled mortality rates and evaluated the performance of forecasting models across different waves. Wave 1 showed high excess mortality and the highest case fatality rate (CFR), especially in the Northeast, reflecting early transmision. Wave 2 was the deadliest nationally, with widespread impact across regions. Wave 3 showed a mixed pattern, with some states improving due to vaccination and better preparedness. The cross-wave model fitting performance, nonlinear models is better linear models in predicting excess mortality.

Index Terms: COVID-19, Excess Mortality, Case Fatality Rate, Mortality Mode

1 Introduction

This report is organized in a left-to-right reading direction, consistent with standard English-language academic formatting.[1]

The COVID-19 pandemic has had a significant global impact over the past five years. COVID-19 was first detected in Wuhan, China, in late 2019 and quickly became a global pandemic. By early 2020, cases had been reported in virtually every country, prompting the World Health Organization to officially declare a pandemic. The pandemic was a major success, and it is still ongoing. The pandemic is still ongoing. As the number of infections and deaths exceeded expectations, societies were also devastated on multiple levels, including a hard-hit healthcare system and a severe economic downturn.

In the US, the first confirmed case was reported on January 21, 2020, in the state of Washington. In just three months, the United States became the center of the global epidemic, with a sharp rise in cumulative cases and deaths. The pandemic's pattern in the United States has been shown by multiple waves of infection, each influenced by regional factors, public health responses, and the emergence of new variants and etc. These waves also represent the effects of virus transmission and the level of preparedness of the population for the virus. As such, they provide a useful framework for analyzing the progress of pandemics and various solutions.

COVID-19 Pandemics occur in distinct temporal waves, each with different levels of virus transmission, severity, and public health response. Understanding these temporal patterns is critical to evaluating the pandemic's progression and the effectiveness of mitigation strategies. Splitting the pandemic period into well-defined waves creates a framework for analyzing changes in State population-level outcomes such as viral virulence, health-care system strain, and excess mortality.

The purpose of this study was to Analysis of the COVID-19 case rate and The purpose of this study was to Analysis of the COVID-19 case rate and the Excess Mortality Rate by Pandemic Wave. A data-driven approach was utilized to divide the U.S. pandemic timeline (April 2020 to July 2024) into three distinct waves. Analysis relying on these waves summarizes state-to-state differences. Reveal how geographic differences affect outbreak outcomes. Finally, quantitatively assess the generalizability of statistical models trained on one wave to other waves in predicting excess mortality. We used linear regression, localized polynomial regression (LOESS), and cubic spline methods to capture temporal variation in COVID-19 case rates and mortality outcomes and to assess the performance of these models across pandemic phases.

1.1 Report Organisation

The remainder of this paper is organized as follows: Section 2 describes the data sources and methods. Section 3 presents the overall results of mortality trends and model comparisons. Section 4 discusses the findings, limitations, and Future work.

2 Method

This study integrated data from three primary sources to construct a comprehensive panel of weekly state-level population, COVID-19 case, and death outcomes.

Annual population estimates were obtained from the U.S. Census Bureau's *National and State Population Estimates 2024* dataset U.S. Census Bureau (2024)., which provides data for all 50 states and the District of Columbia, and Puerto Rico from April 1, 2020, through July 1, 2024. Weekly counts of COVID-19 confirmed cases were obtained from the State Case Surveillance Dataset, as provided by Centers for Disease Control and Prevention (2024b). Death records, including the percentage of total deaths, COVID-19 attributable deaths, and expected deaths, were obtained from the CDC (Centers for Disease Control and Prevention, 2024a). All data were aligned by state and pandemic week based on the MMWR calendar. The case count variable was converted to an integer format and standardized as cases. The year variable in the mortality dataset was also harmonized to accommodate cross-year formats (e.g., "2021/2022"), and all numeric columns were explicitly converted to ensure consistency across datasets.

Two key metrics were derived from mortality records:

[1] This document follows a two-column format. Please read each page from top to bottom of the left column, then continue at the top of the right column.

$$\text{Excess Deaths} = \text{Total Deaths} - \left(\frac{\text{Total Deaths}}{\frac{\text{Percent of Expected Deaths}}{100}} \right) \quad (1)$$

$$\text{Excess Mortality Rate} = \text{Percent of Expected Deaths} - 100 \quad (2)$$

To construct a unified dataset, we first generated a full grid of weekly dates and U.S. states using a cross join. This allowed us to merge population estimates, COVID-19 cases, and mortality data using consistent identifiers—state, epidemiological week, and year. Observations with missing values in key metrics were imputed with zeros to ensure temporal completeness. The resulting dataset includes one row per state-week combination, with fully aligned demographic, case, and mortality information.

2.1: Defining Pandemic Waves Using Data Visualization

To delineate pandemic waves, we calculated weekly COVID-19 case rates, death rates, and excess mortality rates per 100,000 population from January 2020 to April 2025, enabling standardized comparisons across states. These metrics were visualized over time and stratified by region. Wave boundaries were identified based on visual inflection points where all three indicators showed sustained peaks across multiple regions, marked by shaded areas in the time series plots.

2.2: State-Level COVID-19 Mortality and Excess Mortality Rate by Wave

Quantile-Based Classification and Heatmap Visualization. To analyze geographic patterns in COVID-19 mortality, we calculated the cumulative COVID-19 death rate and excess mortality per 100,000 population for each U.S. state within each pandemic wave. These values were summed across all weeks within each wave. To facilitate interpretation, we categorized states into Low, Medium, and High mortality groups based on the 33rd and 66th percentiles. These groupings were visualized via two heatmaps to highlight temporal and spatial mortality trends.

Ranking States by Mortality Extremes. To underscore disparities, we ranked states by total COVID-19 death rate and excess mortality within each wave. The top three and bottom three states were identified and visualized using a faceted bar chart, enabling direct wave-to-wave comparisons.

2.3: Virulence Across Waves

To estimate clinical severity across pandemic waves, we calculated the Case Fatality Rate (CFR), defined as:

$$\text{CFR} = \left(\frac{\text{Total Death Rate}}{\text{Total Case Rate}} \right) \times 100$$

Weekly case and death rates (per 100,000) were aggregated across all states within each wave. We excluded observations with missing or zero case rates.

2.4: Modeling the Relationship Between Case Rate and Excess Mortality

To evaluate the relationship between COVID-19 case rates and excess mortality rates, we filtered weekly state-level data to include only the relevant metrics and pivoted the dataset to a wide format. We first fitted linear models per wave and visualized trends with scatter plots and regression lines.

We then trained models on one wave and tested on another to evaluate generalizability using RMSE and R^2. Finally, we compared model fit using linear regression (Equation 3), LOESS smoothing (Equation 4), and cubic splines (Equation 5), particularly when predicting Wave 3 from the Wave 2 model. Additional comparison between models trained on Waves 1 and 3, applied to Wave 2 data, revealed differences in distribution and model behavior across phases.

$$\text{Excess Mortality Rate} = \beta_0 + \beta_1 \cdot \text{Cases Rate} \quad (3)$$

$$\hat{y}_{\text{LOESS}}(x_0) = \sum_{i=1}^{n} w_i(x_0) \cdot y_i \quad (4)$$

$$\hat{y}_{\text{spline}} = \beta_0 + \sum_{j=1}^{5} \beta_j B_j(x) \quad (5)$$

Where $w_i(x_0)$ are kernel-based weights around x_0, and $B_j(x)$ are spline basis functions constructed from the case rate using 6 df.

3 Results

3.1: Defining Pandemic Waves Using Data Visualization

Based on Figure 1 below, in case rates, death rates, and excess mortality rates, we identified three major pandemic waves and visualized them using shaded grey bands.

- **Wave 1:** March 1 – June 30, 2020
- **Wave 2:** October 1, 2020 – February 28, 2021
- **Wave 3:** July 1 – October 31, 2021

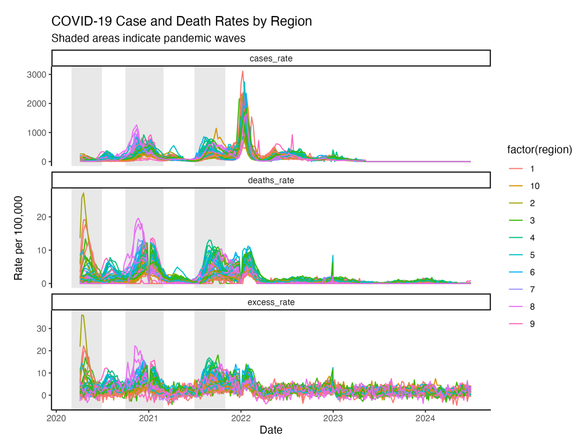

Figure 1. COVID-19 Case and Death Rates by Region. Shaded areas indicate pandemic waves. Region labels correspond to U.S. Census-defined groupings: **1** = New England, **2** = New York and New Jersey, Puerto Rico, Virgin Islands, **3** = Mid-Atlantic, **4** = Southeast, **5** = Midwest, **6** = South Central, **7** = Central Plains, **8** = Mountain States, **9** = Pacific, **10** = Pacific Northwest.

These intervals were chosen based on synchronized national surges across all three health metrics—case rates, death rates, and excess mortality—as follows:

- **Wave 1 (March–June 2020):** A sharp rise in death rates occurred in early 2020, particularly in Regions 9(Pacific) and 2(New York and New Jersey, Puerto Rico, Virgin Islands). Although case rates were relatively low, the excess mortality spiked, confirming the severity of the initial wave.
- **Wave 2 (October 2020–February 2021):** Both case and death rates rose dramatically, peaking around January 2021. Excess mortality reached its highest level during this period, and all three metrics showed synchronized spikes across all regions. Mountain States, which is region 8, is the peak region.
- **Wave 3 (July–October 2021):** In the summer of 2021, case rates rose again, peaking during August and September. The Pacific Northwest leads the case trends. The Mid-Atlantic has the highest excess mortality rate. While the increase in mortality was not as dramatic as the second wave, excess mortality continued to rise, marking a clear third wave in most regions.

3.2: State-Level COVID-19 Mortality and Excess Mortality by Wave

Figure 2 visualizes state-level COVID-19 death rates and excess mortality rates per 100,000 population across the three defined pandemic waves. States are grouped into low (blue), medium (gold), and high (red) categories based on quantiles of total death rate within each wave.

During **Wave 1**, mortality levels varied widely across states, with most falling in the low or medium categories.

In **Wave 2**, a notable shift occurred, with a large number of states moving into the high mortality group.

By **Wave 3**, the pattern became more mixed. While some states continued to experience high death rates and high excess mortality rates, others showed improvement and shifted into lower mortality categories.

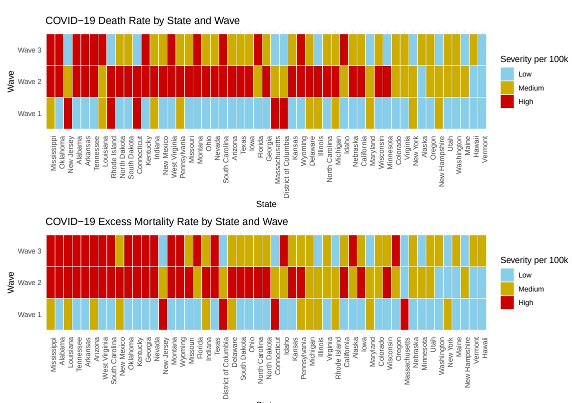

Figure 2. State-level COVID-19 death rates (top) and excess mortality rates (bottom) per 100,000 population across three pandemic waves. States are grouped into low (blue), medium (gold), and high (red) severity based on wave-specific quantiles.

Wave 1: The highest mortality was concentrated in the Northeast: New Jersey, Connecticut, and Massachusetts. In contrast, Alaska, Hawaii, and Montana reported zero COVID-19 deaths and very low excess mortality during this early phase.

Wave 2: the deadliest period nationally, states such as South Dakota and Arizona exhibited the highest excess mortality, reflecting widespread transmission in central regions.

Wave 3: Southern and remote states, including Florida and Alaska, emerged as mortality hotspots. Meanwhile, states such as New Jersey and Massachusetts, which had high mortality in Wave 1, recorded among the lowest rates in Wave 3.

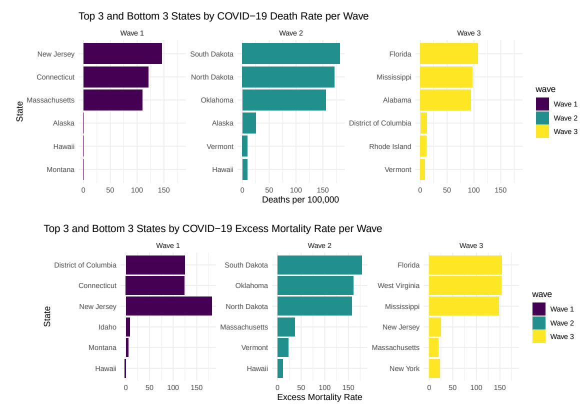

Figure 3. Top and bottom three U.S. states by COVID-19 death rate (top) and excess mortality rate (bottom) per 100,000 population across pandemic waves. States are grouped by wave, with bar color indicating wave membership.

3.3: Virulence Across Waves

Figure 4 presents the Case Fatality Rate (CFR) for each of the three pandemic waves.

Wave 1 had the highest CFR(4.5%), likely due to limited testing and early clinical uncertainty. In subsequent waves, the CFR declined to 1.41% in **Wave 2** and 1.22% in **Wave 3**, suggesting reduced disease severity. This decline aligns with vaccine rollout and increasing population-level immunity.

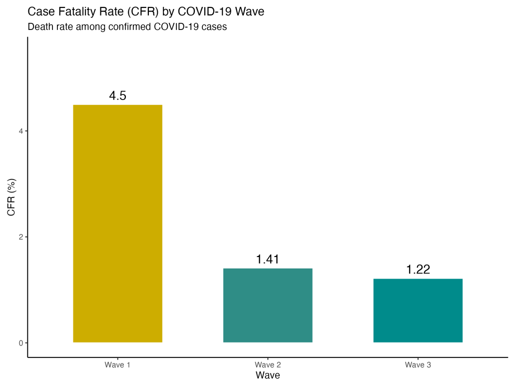

Figure 4. Case Fatality Rate (CFR) by COVID-19 wave.

3.4: Analyzing and Predicting the Relationship Between Case Rate and Excess Mortality Rate

Figure 5 illustrates how different models, trained on Wave 2 data, perform in predicting excess mortality during Wave 3. The goal is to evaluate how well case rates explain variation in excess mortality under different modeling approaches.

- **Observed data (yellow dots):** Actual excess mortality rates plotted against case rates for Wave 3.
- **Linear model (dashed cyan line):** Predictions of the linear regression model.
- **LOESS model (blue line):** Predictions of Wave 3 using LOESS local regression.
- **Cubic spline model (red line):** Predictions using cubic spline regression with six degrees of freedom.

These models highlight the limitations of purely linear approaches and demonstrate the value of flexible modeling (e.g., LOESS, splines) in capturing pandemic dynamics over time.

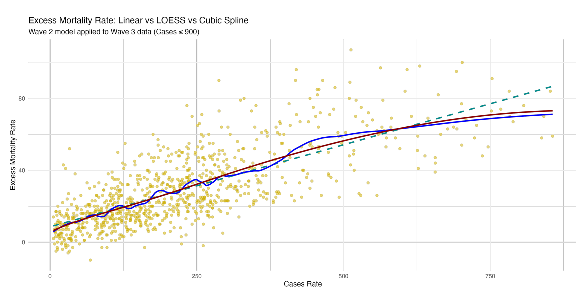

Figure 5. Model predictions of excess mortality in Wave 3 using models trained on Wave 2 data.

Observations:

- The linear model tends to over-predict excess mortality when the case rate exceeds 700. This is due to the actual data flattening and trending downward in the upper range of case rates.
- Loess and spline models can better capture the nonlinear trend of Wave 3, especially in the mid-to-high range.
- Among the three approaches, the spline model produces the smoothest curve and demonstrates the best fit in the upper range of case rates.

3.5: Model Performance Summary, Table 1

- The model trained on **Wave 2** data performs well in predicting excess mortality in **Wave 3** (*RMSE* = 16.40, R^2 = 0.558), indicating that the underlying data structures of these two waves are closely aligned.
- In contrast, the **Wave 1** model cannot effectively predict either Wave 2 (*RMSE* = 111.21) or Wave 3 (*RMSE* = 78.68), suggesting that the mortality risk structure in the early and middle stages of the epidemic is very different.
- The **Wave 3** model has the best fit on its own data (*RMSE* = 13.70), which can be regarded as a near-true model benchmark

Table 1
Cross-wave prediction performance for linear models. Each entry shows RMSE and R^2 for a model trained on one wave and tested on another.

Trained on	Tested on	RMSE	R^2
Wave 1	Wave 1	18.64	0.479
Wave 1	Wave 2	111.21	0.450
Wave 1	Wave 3	78.68	0.558
Wave 2	Wave 1	26.11	0.479
Wave 2	Wave 2	15.06	0.450
Wave 2	Wave 3	16.40	0.558
Wave 3	Wave 1	24.03	0.479
Wave 3	Wave 2	18.50	0.450
Wave 3	Wave 3	13.70	0.558

4 Discussion

This study aimed to characterize the relationship between COVID-19 case rates and excess mortality across distinct pandemic waves in the United States and to evaluate the generalizability of predictive models across these waves.

Findings

The first wave highlights the challenges of the early pandemic, where underreporting due to limited testing led to an apparent mismatch between case and death rates. Despite relatively low case counts, excess mortality reveals that the impact was more severe than reported case data alone would suggest.

The second wave reflects the compounded effect of widespread transmission, colder weather, and delayed public health responses. The simultaneous spikes across all three metrics underscore its severity and national reach.

The third wave, while still significant, occurred in a context of greater preparedness. The comparatively lower case fatality rate and somewhat contained mortality may reflect the protective effects of vaccination, improved treatment protocols, and greater public awareness. The persistent rise in excess mortality, however, suggests ongoing vulnerabilities in certain regions or populations.

These patterns underscore the shifting geographic burden of the pandemic over time. The early concentration of mortality in the Northeast gave way to central and southern regions in later waves. The decline in mortality rates in states like New Jersey and Massachusetts by Wave 3 likely reflects improved public health responses, accumulated experience, and broader population-level interventions.

Mortality Patterns and Model Performance

One of the most salient findings is the clear heterogeneity in both mortality burden and model performance across pandemic waves.

Wave 1 exhibited the highest case fatality rate (CFR), suggesting severe clinical outcomes among detected cases. However, this elevated CFR likely reflects the limited testing capacity and healthcare preparedness in early 2020, where mild or asymptomatic cases went undetected, inflating the apparent severity. In contrast, Waves 2 and 3 showed markedly lower CFRs, consistent with expanded testing, improved treatment protocols, and the initial impact of vaccination campaigns. These trends indicate a reduction in clinical virulence and demonstrate the adaptability of healthcare systems over time.

Geographic disparities in mortality outcomes were also evident. The Northeast bore the brunt of Wave 1, while central and southern states experienced greater mortality during later waves. Notably, some states that were severely impacted early on, such as New Jersey and Massachusetts, demonstrated significantly improved outcomes in Wave 3. This pattern likely reflects the benefits of accumulated public health experience, stronger mitigation measures, and targeted vaccination efforts.

Limitations

The cross-wave predictive analysis revealed important limitations in the temporal stability of statistical models.

Models trained on Wave 1 data performed poorly when applied to subsequent waves, particularly Wave 2, which had distinct epidemiological characteristics. This suggests that early pandemic conditions were fundamentally different from later phases in terms of virus spread, healthcare response, and population behavior.

Although approaches such as LOESS and cubic splines provided improved fits in scenarios where linear models failed to capture complex trends—particularly in Wave 3—the best-performing predictions were always obtained from models trained and tested on the same wave. This underscores the limited generalizability of pandemic forecasting models across distinct time periods.

Only three variables—COVID-19 cases, deaths, and the excess mortality rate—were considered. Important factors such as vaccination coverage and booster rates, which are likely to influence later trends, were not included. Additionally, time was not explicitly modeled as a variable in the fit.

The excess mortality rate, while valuable for capturing hidden pandemic impacts, may also be influenced by indirect effects such as healthcare disruptions, deferred treatment for non-COVID conditions, or population shifts due to migration.

Future Work

Future work should incorporate a time variable into the prediction model to better capture longitudinal effects.

To better understand excess mortality trends in later waves, future models should introduce vaccine-related variables. Comparing the excess mortality rates between Wave 1 and later waves (Wave 2 and 3) could help quantify the protective effect of vaccination at the state level.

Additionally, incorporating booster coverage, hospitalization data, and indicators of healthcare capacity could further improve model accuracy and interpretability.

Acknowledgements

This project report was conducted as part of the BIOSTAT620: Introduction to Health Data Science course offered by the Department of Biostatistics at the University of Michigan School of Public Health. We would like to thank Professor Dylan Cable and our Graduate Student Instructor (GSI), Yize Hao, for their instruction and guidance throughout the semester.

References

Centers for Disease Control and Prevention (2024a). *Provisional COVID-19 Death Counts by Week Ending Date and State.* Accessed April 2025. URL: https://data.cdc.gov/National-Center-for-Health-Statistics/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab.

— (2024b). *Weekly United States COVID-19 Cases and Deaths by State.* Accessed April 2025. URL: https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-Deaths-by-/pwn4-m3yp.

U.S. Census Bureau (2024). *National and State Population Estimates 2024.* Accessed April 2025. URL: https://www.census.gov/newsroom/press-kits/2024/national-state-population-estimates.html.