BST 260 Final Project

Abstract

This project examines the COVID-19 pandemic from January 2020 to December 2024, focusing on trends in COVID-19 deaths, cases, and death rates across eight defined pandemic waves. Weekly mortality data were collected for each state, with the pandemic period segmented into waves based on fluctuations in COVID-19 case and death counts. COVID-19 death rates were analyzed for each state across these waves, revealing variations in mortality outcomes and identifying states that performed comparatively better or worse during specific periods. Additionally, we assessed the virulence of COVID-19, measured as the case fatality rate, across pandemic waves to evaluate changes in the virus's severity over time. Excess mortality was estimated for each week and state using a quasi-Poisson model, comparing observed weekly deaths to baseline pre-pandemic mortality levels (2014-2019). While significant excess mortality was observed during early waves, we also detected negative excess mortality in later periods, suggesting lower-than-expected deaths during specific timeframes. Results highlight variability in mortality patterns between states, a general decline in virulence across waves, and periods of both positive and negative excess mortality, underscoring the complexity of mortality trends during the pandemic.

Introduction

The COVID-19 pandemic has had far-reaching effects across various aspects of society, including the economy, social dynamics, and public health systems. Economically, the pandemic caused unprecedented disruptions to global markets, with significant job losses, business closures, and shifts in labor markets that have reshaped industries. Socially, it altered daily life through widespread lock- downs, remote work, and changes in interpersonal interactions, while exacerbating existing inequalities [1]. Public health systems were strained to their limits, grappling with overwhelming caseloads, shortages of resources, and the need for rapid adaptation to a continuously evolving crisis. The pandemic highlighted systemic vulnerabilities and the importance of preparedness in mitigating such global threats.

In the United States, COVID-19 posed unique challenges given the country's size, diversity, and decentralized governance. The pandemic's impact varied widely across states, shaped by differences in demographic characteristics, healthcare infrastructure, and state-level policies. State-specific responses, including mask mandates, business closures, vaccination rollouts, and social distancing measures, significantly influenced the trajectory of the pandemic within each jurisdiction [2]. These policies, combined with varying public attitudes and behaviors toward compliance, led to notable disparities in COVID-19 outcomes, such as cases, deaths, and economic recovery.

Understanding the trends and variations in COVID-19's impact across states is critical for informing future public health strategies. By identifying patterns in mortality outcomes, case trends, and excess mortality, researchers and policymakers can evaluate the effectiveness of interventions and develop targeted measures to address disparities. This study aims to analyze COVID-19 outcomes in the U.S., focusing on differences across states, to uncover key factors that shaped the pandemic's trajectory and provide insights for improving resilience against future public health crises.

To achieve these objectives, this project adopts a multifaceted methodlogical approach. Eight distinct waves of the pandemic were identified using a combination of COVID-19 case and mortality trends by data visualization. Weekly death rates per 100,000 population were calculated for each state to compare regional outcomes and highlight states that mitigated mortality effectively. COVID-19 virulence was assessed by measuring the case fatality rate (CFR) across waves, allowing for an analysis of how disease severity changed over time, particularly as vaccination campaigns progressed and new viral variants emerged.Finally, excess mortality was estimated using a quasi-Poisson model to compare observed weekly deaths during the pandemic with the expected weekly mortality based on data from 2014 to 2019 which is an estimation of morality during 2020-2024 assuming no pandemic. This approach provides a comprehensive assessment of mortality trends, both directly attributed to COVID-19 and those exceeding expected levels.

The motivation for this study stems from the need to develop a clearer understanding of the pandemic's progression and its uneven impact across the United States. By combining mortality, case, and excess death analyses, this project offers insights that are crucial for evaluating policy responses, identifying systemic vulnerabilities, and informing future strategies for public health preparedness. The findings presented here contribute to the broader understanding of COVID-19's societal impact and aim to assist policymakers, public health officials, and researchers in mitigating the effects of future health crises.

Methods

Dataset

Data for this study were sourced from publicly available repositories, primarily focusing on COVID-19 deaths, cases, and population estimates. Weekly COVID-19 mortality data spanning from December 29, 2019, to November 30, 2024, were retrieved from the Centers for Disease Control and Prevention (CDC) through its open data API, specifically the dataset titled Provisional COVID-19 Death Counts by Week Ending Date and State [3].

To establish pre-pandemic baseline mortality, weekly death counts for all causes from 2014 to 2019 were obtained from the CDC API endpoint, Weekly Counts of Deaths by State and Select Causes [4]. These data provided the foundation for estimating excess mortality during the pandemic.

Population estimates were essential for calculating mortality rates and standardizing comparisons across states. Data for the years 2014 to 2019 were retrieved using the U.S. Census Bureau's API [5], while updated population estimates for 2020 to 2023 were sourced from the Census Bureau's NST-EST2023- POP dataset [6]. These datasets were harmonized into a single, continuous population table for integration with the mortality data.

COVID-19 case data were obtained from the CDC's COVID-19 Cases and Deaths by State over Time dataset [7]. Weekly case counts were matched to mortality data by state, year, and MMWR week to ensure temporal alignment for analysis.

The datasets were cleaned and merged using the statistical programming language R. In the data cleaning process, we notice the lack of population and COVID-19 cases in 2024 data. We excluded this year from some parts of our analysis. This is an limitation of the data gathering process. All variable names were modified and validated for consistency. Year and mmwr week were added to explicitly model time. Missing values were excluded from the final analysis after careful consideration.

Analysis Framework

Identify Waves

The pandemic timeline was divided into distinct waves based on visual inspection of trends in COVID-19 deaths and cases. In the downstream analysis, for each wave, total excess deaths were aggregated, and excess mortality rates were ranked by state. Average state populations during each wave were used to compute standardized death rates, allowing for cross-state comparisons. The method allows us to more comprehensively define waves since it combines both deaths and cases. However, visual inspection lacks of quantitative explanation of our wave definition.

Evaluate Deaths Rates

To evaluate COVID-19 deaths rates across pandemic waves and states, the COVID-19 deaths in a specific wave and state were sum together, standardized by state population during that wave, and multiplied by 100,000 to compute death rates per 100,000 people.

$$\text{Death Rate}_t = \left(\frac{\text{COVID-19 deaths}_t}{\text{Population}} \right) \times 100,000.$$

Evaluate COVID-19 Virulence

To evaluate COVID-19 virulence across pandemic waves, the case fatality rate (CFR) was computed until case reporting was discontinued in 2023. The CFR was defined as:

$$\text{CFR} = \left(\frac{\text{COVID-19 Deaths}}{\text{Total Cases}} \right) \times 100.$$

The CFR served as a proxy for disease severity and enabled the assessment of changes in virulence across pandemic periods.

Estimation of Excess Mortality

The analysis of excess deaths during pandemic followed methods adapted from excessmort R package [8].Pre-pandemic baseline mortality was estimated using a quasi-Poisson regression model from excessmort. The relationship between COVID-19 deaths and estimated excess mortality was assessed using a linear regression model.

The expected counts of mortality from 2014 to 2024 were calculated by the excess model function from excessmort. Pre-pandemic weekly total deaths from 2014 to 2019 was used as a reference period to catch the normal mortality tendency. The pandemic period were excluded while computing the expected counts thus the excess mortality won't be considered as normal mortality. To address over dispersion in the weekly mortality data, a quasi-Poisson regression model was chosen.

The excess mortality was defined as:

$$\text{Excess Mortality}_t = \text{Observed Deaths}_t - \text{Expected Deaths}_t.$$

The cumulative excess mortality was the output of excess cumulative function from excessmort. The weekly excess mortality were calculated by taking difference between current-week cumulative excess mortality minus previous- week cumulative excess mortality.

Weekly excess deaths were standardized by state population to compute excess death rates per 100,000 people, facilitating meaningful comparisons across states and pandemic waves. The excess death rate was defined as:

$$\text{Excess Death Rate}_t = \left(\frac{\text{Excess Mortality}_t}{\text{Population}} \right) \times 100,000.$$

To assess whether COVID-19 deaths explained observed excess mortality, a linear regression model was fitted with weekly excess deaths as the dependent variable and weekly COVID-19 deaths as the independent predictor. The model was expressed as:

$$\text{Excess Mortality}_t = \beta_0 + \beta_1 \cdot \text{COVID-19 Deaths}_t + \epsilon_t,$$

where β_0 is the intercept, β_1 is the regression coefficient, and ϵ_t is the error term. Model performance was evaluated using the R^2 value to determine the proportion of variability in excess mortality explained by COVID-19 deaths.

Visualization

Time series plots were created to illustrate trends in observed deaths, expected deaths, excess mortality, and COVID-19 deaths over the pandemic period. Wave boundaries were overlaid as vertical lines to delineate distinct phases of the pandemic.

Bar plots were generated to compare excess death rates across states for each wave, with states ranked in descending order of mortality outcomes. These plots highlighted regional disparities in excess mortality and enabled comparisons across waves.

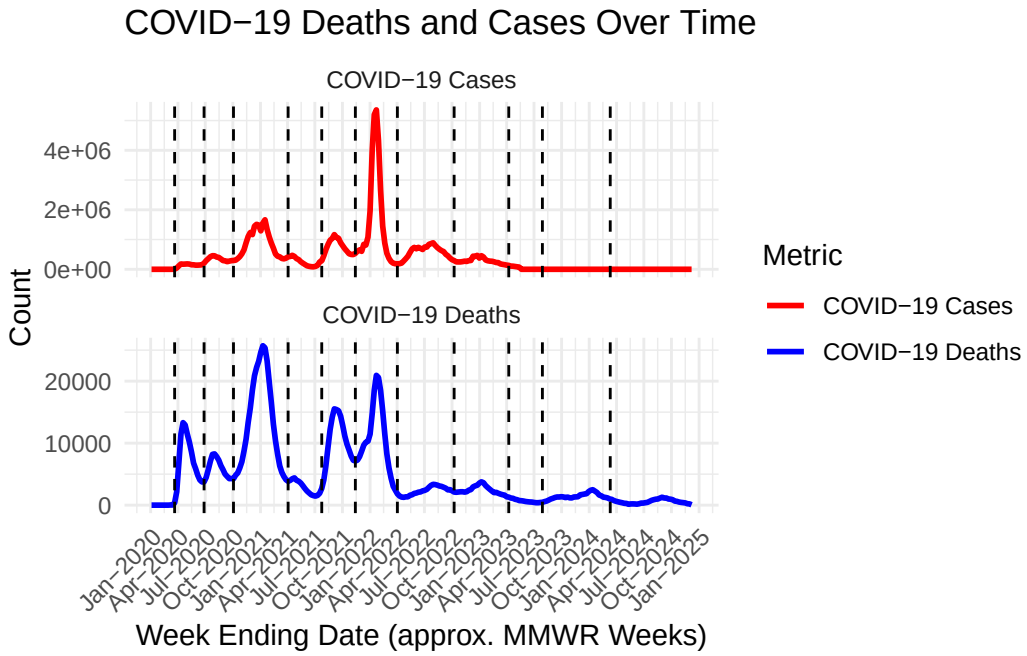
COVID-19 death rates per 100,000 population were visualized for each state and wave to identify trends in mortality outcomes. Similarly, case fatality rates were plotted across waves to evaluate changes in COVID-19 virulence over time. Scatter plots with linear regression lines were used to illustrate the relationship between weekly COVID-19 deaths and excess mortality, providing insights into the contribution of COVID-19 deaths to overall mortality patterns.

The combination of time series, bar, and scatter plots allowed for a comprehensive understanding of COVID-19 mortality, excess deaths, and disease severity across states and pandemic waves.

Results

COVID-19 Waves

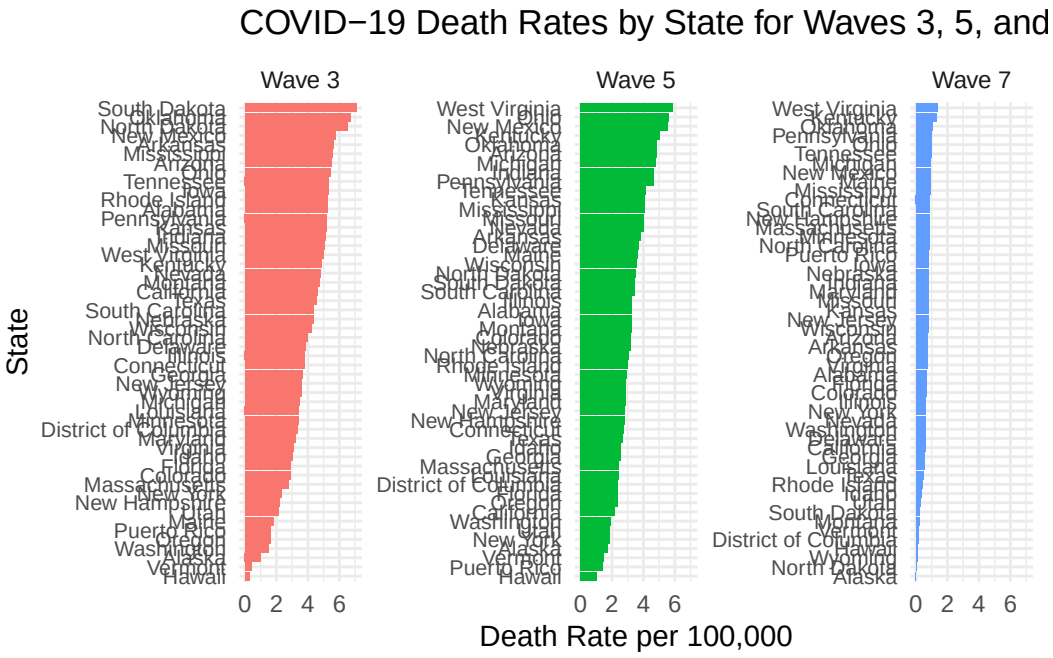
The COVID-19 pandemic timeline was divided into eight distinct waves based on visual inspection of trends in COVID-19 cases and deaths. Wave boundaries were determined by identifying peaks and troughs in the time series data for weekly reported cases and deaths. Figure 1 displays the COVID-19 case and death trends over time, with vertical lines indicating the boundaries of each wave. Each wave represents a specific period where COVID-19 dynamics exhibited significant changes in terms of infection rates and mortality patterns. This division provides a structured framework for analyzing the pandemic's impact over time and across states.



COVID-19 Death Rates by State for Waves 3, 5, and 7

Waves 3 and 5 were two major waves characterized by substantial peaks in COVID-19 deaths and cases across the United States. Wave 3 recorded the highest number of COVID-19 deaths nationwide, while Wave 5 exhibited the largest surge in COVID-19 cases. Wave 7, although much smaller in magnitude, was the last wave with relatively complete data before reporting inconsistencies emerged in Wave 8 due to the discontinuation of COVID-19 case tracking by the CDC.

Figure 2 presents the COVID-19 death rates per 100,000 population for each state during Waves 3, 5, and 7. States are ranked in descending order of death rates within each wave, facilitating comparisons across regions and time periods. During Wave 3, South Dakota, Oklahoma, and North Dakota reported the highest death rates, reflecting the severe impact of the pandemic in these states. In Wave 5, West Virginia, Ohio, and New Mexico experienced the most significant mortality outcomes. By Wave 7, the death rates were markedly lower, with West Virginia and Kentucky showing the highest mortality among the states.



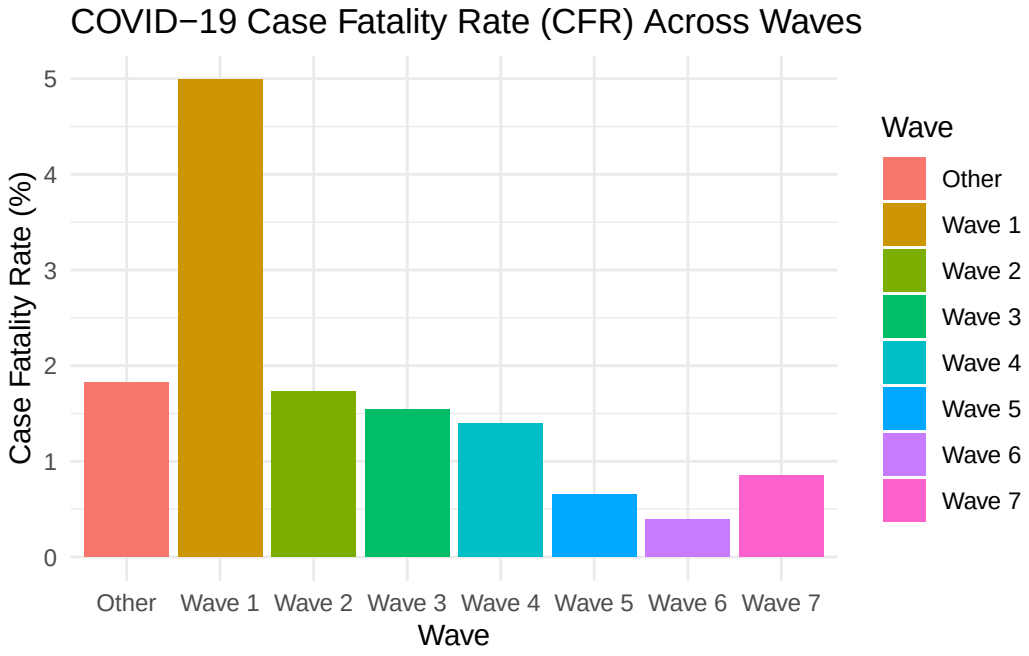
Changes in COVID-19 Virulence Across Waves

To evaluate changes in COVID-19 virulence across the pandemic, we examined the case fatality rate (CFR), defined as the ratio of COVID-19 deaths to reported cases, across the identified waves. Figure 3 presents the CFR for each wave, highlighting a clear decline in virulence over time. The first wave exhibited the highest CFR, reaching approximately 5%, reflecting both the absence of vaccination and limited understanding of effective treatment protocols during the early stages of the pandemic.

Following Wave 1, the CFR began to decline and remained consistently lower in subsequent waves. This reduction can likely be attributed to multiple factors, including the widespread rollout of vaccines, which reduced the likelihood of severe disease and mortality. Additionally, the accumulation of population immunity through both natural infection and vaccination contributed to a decrease in severe outcomes. The natural selection of viral variants may have

also played a role, as viruses with reduced virulence tend to propagate more effectively within populations.

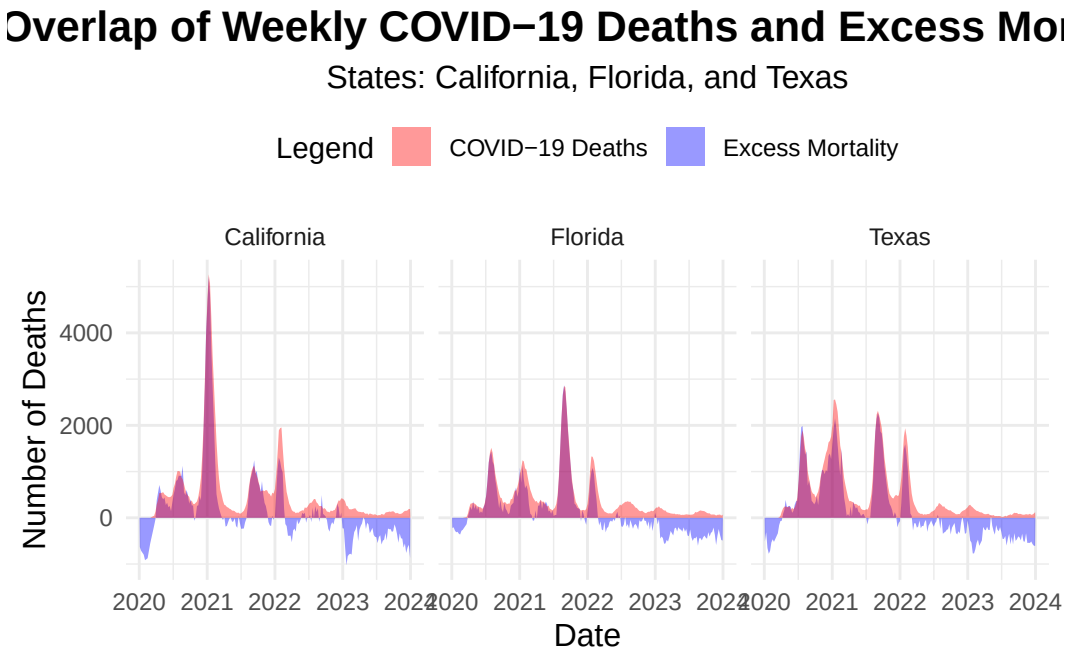
Wave 5, despite having the highest number of reported cases, exhibited a relatively low CFR. This period coincided with the rapid expansion of vaccination coverage, which significantly mitigated the severity of infections. By Wave 7, the CFR remained low, although reporting inconsistencies and reduced case tracking in certain states may have impacted the completeness of the data. Wave 8 was excluded from this analysis due to a complete absence of case data, as the CDC ceased tracking COVID-19 cases during this period.



Excess Mortality and the Role of COVID-19 Deaths

To examine excess mortality, we estimated the weekly deviation of observed deaths from expected mortality levels for each state, where expected deaths were derived using a quasi-Poisson regression model. Excess mortality rep- resents the number of deaths exceeding the predicted baseline, accounting for seasonal trends and long-term mortality patterns. The linear regression between estimated excess mortality and COVID-19 deaths shows an estimated coefficient equals to 0.907945 and a adjusted R square equals to 0.7865. Pearson's product moment correlation equals 0.88686. All the results suggests a strong correlation. Figure 4 highlights California, Florida, and Texas as examples of states where COVID-19 deaths closely overlapped with excess mortality, particularly during the major waves of the pandemic. The full visualization for all states is included in the supplementary material.

The results demonstrate a strong overlap between COVID-19 deaths and excess mortality from 2020 through early 2022. This suggests that COVID-19 was the primary driver of excess deaths during this period, underscoring the significant toll the pandemic took on overall mortality. States with higher peaks in COVID-19 deaths exhibited corresponding increases in excess mortality, particularly during major waves such as Wave 3 and Wave 5.



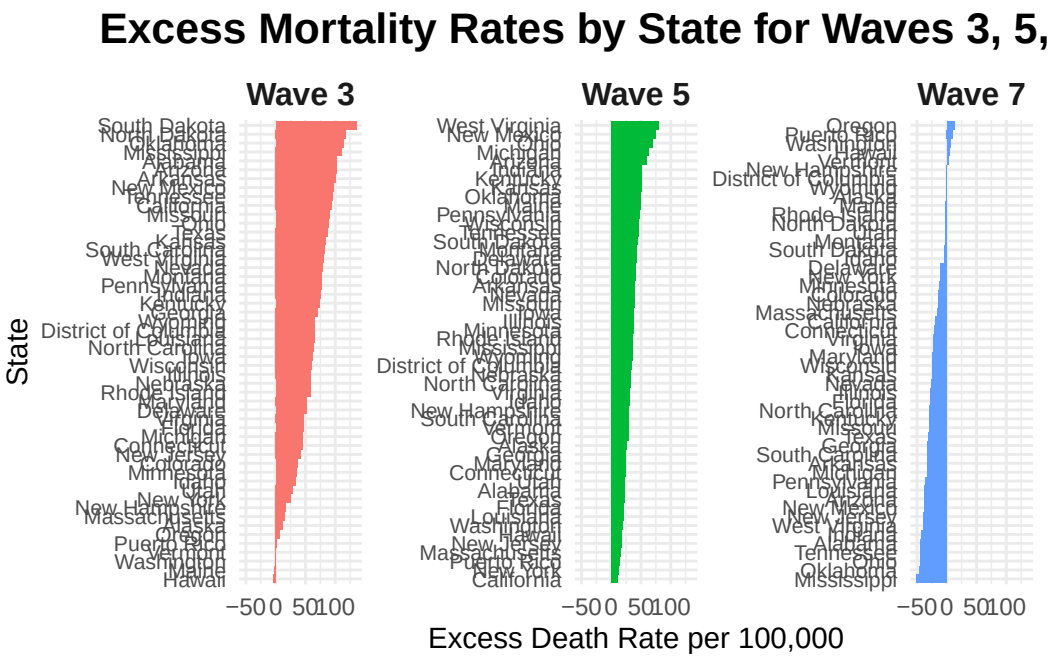
Excess Mortality Rates by State for Waves 3, 5, and 7

To further examine state-level differences in excess mortality, we computed excess mortality rates for Waves 3, 5, and 7. Excess mortality rates were defined as the number of weekly excess deaths per 100,000 population. Figure 5 dis- plays the ranked excess mortality rates by state for these waves, providing a clear comparison of the pandemic's impact across states during key periods.

During Wave 3, South Dakota, North Dakota, and Oklahoma recorded the highest excess mortality rates, with rates exceeding 100 deaths per 100,000 population. This wave, occurring in the winter of 2020-2021, was marked by widespread viral transmission and significant strain on healthcare systems, particularly in rural and underserved areas.

Wave 5, driven largely by the Delta variant surge in mid-2021, exhibited high excess mortality in West Virginia, New Mexico, and Ohio. Although vaccination campaigns were underway, disparities in vaccination rates and healthcare access likely exacerbated the mortality burden in these states.

By Wave 7, excess mortality rates declined significantly across most states, reflecting reduced virulence, improved immunity, and advancements in treatment. Despite this trend, Oregon and Puerto Rico reported higher-than-expected excess mortality rates compared to other states, contrasting with their previously lower mortality outcomes in earlier waves.



Discussion

The comparison of state-level COVID-19 death rates across waves highlights regional disparities in mortality and the evolving dynamics of the pandemic over time. The stark differences in death rates between Wave 3 and Wave 7 reflect several factors, including improved treatment protocols, the widespread roll- out of vaccines, and the accumulation of population immunity through natural infection and vaccination efforts. Reduced virus virulence in later waves and the disproportionate impact of earlier waves on vulnerable populations likely contributed to these trends. State-specific factors, such as healthcare capacity, policy responses, and population demographics, shaped the underlying condi- tions influencing how each state experienced and responded to different phases of the pandemic.

Overall, the results indicate that COVID-19 became less virulent over time, as evidenced by the declining case fatality rate. The combined effects of vaccination, accumulated immunity, and the natural selection of less severe viral variants contributed to this trend. However, differences in reporting practices and data availability across waves should be considered when interpreting these results.

Negative excess mortality values were observed in many states following major waves, particularly after mid-2022. These negative values may reflect increased public awareness of health risks, improved healthcare responses, and preventive public policies such as social distancing and vaccination campaigns. Additionally, reduced mortality may have been influenced by selection effects, where vulnerable populations disproportionately affected during earlier waves reduced the pool of individuals at elevated risk. Other contributors include improvements in medical care, fewer deaths from other causes due to behavioral changes (e.g., reduced traffic accidents), and potential misclassification of deaths in official reporting systems.

While COVID-19 deaths largely explain the trends in excess mortality during the peak waves, further analysis is warranted to explore the complex interplay of factors contributing to negative excess mortality observed in the later phases of the pandemic. The comparison across Waves 3, 5, and 7 highlights how ex- cess mortality evolved alongside the pandemic. Earlier waves exhibited higher mortality due to limited immunity and healthcare strain, whereas later waves saw substantial decreases in excess mortality, indicating improvements in pandemic management and a shift in the virus's impact.Nevertheless, regional disparities persisted, underscoring the importance of state-specific factors, including healthcare access, demographics, and policy responses, in shaping excess mortality outcomes.

Some have argued that deaths occurring during the COVID-19 pandemic were not directly caused by the virus but rather by other underlying conditions or indirect factors such as delayed medical care. However, the clear overlap between COVID-19 deaths and excess mortality observed in our anal- ysis strongly suggests otherwise. The temporal alignment and magnitude of COVID-19 deaths closely mirror patterns in excess mortality, providing compelling evidence that these deaths were indeed a direct consequence of the virus. While other factors may have contributed to some deaths, the data overwhelmingly supports COVID-19 as the primary driver of the observed excess mortality during the pandemic.

There are some limitations in our analysis, such as the lack of data for 2024, the use of this year's population for each week, and the use of visual segmentation for wave detection.

References

[1] Nwosu, C. O., Kollamparambil, U., and Oyenubi, A., "Socio-economic inequalities in ability to work from home during the coronavirus pandemic," The Economic and Labour Relations Review, vol. 33, no. 2, pp.290–307, 2022. [Online]. Available:https://www.cambridge.org/core/journals/the-economic-and-labour-relations-review/article/socioeconomic-inequalities-in-ability-to-work-from-home-during-the-coronavirus-pandemic 8693F0AA6332E958DC7A7563E52708AB. [Accessed: Dec. 18, 2024].

[2] Kerpen, P., Moore, S., and Mulligan, C. B., "A final report card on the states' response to COVID-19," National Bureau of Economic Research, 2023. [Online]. Available: https://www.nber.org/papers/w29928. [Accessed: Dec. 20, 2024].

[3] Centers for Disease Control and Prevention, "Provisional COVID-19 Death Counts by Week Ending Date and State," 2024. [Online]. Available: https: //data.cdc.gov/resource/r8kw-7aab.json. [Accessed: Dec. 18, 2024].

[4] Centers for Disease Control and Prevention, "Weekly Counts of Deaths by State and Select Causes," 2024. [Online]. Available: https://data.cdc. gov/resource/3yf8-kanr.json. [Accessed: Dec. 18, 2024].

[5] U.S. Census Bureau, "Population Estimates Program (PEP), Annual Population Estimates," 2024. [Online]. Available: https://api.census.gov/. [Accessed: Dec. 18, 2024].

[6] U.S. Census Bureau, "NST-EST2023-POP: Annual State Resident Population Estimates." [Online]. Available: https://www.census.gov. [Accessed: Dec. 18, 2024].

[7] Centers for Disease Control and Prevention, "United States COVID-19 Cases and Deaths by State over Time," 2024. [Online]. Available:https: //data.cdc.gov/resource/pwn4-m3yp.json. [Accessed: Dec. 18, 2024].

[8] R. J. Acosta and R. A. Irizarry, "A Flexible Statistical Framework for Estimating Excess Mortality," Epidemiology, vol. 33, no. 3, pp. 346–353, Apr. 2022, doi: https://doi.org/10.1097/ede.0000000000001445.

Code Appendix

```
library(dplyr)
library(ggplot2)
library(readr)
library(ggrepel)
library(tidyverse)
library(excessmort)
library(purrr)
library(tidytext)

dat <- readRDS("../data/dat.rds")
dat <- dat %>%
  mutate(week_ending_date = as.Date(week_ending_date),
         month = floor_date(week_ending_date, "month"))%>%
  mutate(total_deaths = as.numeric(total_deaths),
         population = as.numeric(population))

# Subset data for COVID-19 period (2020-2024)
dat <- dat %>%
  mutate(week_ending_date = as.Date(week_ending_date),
```

```
         month = floor_date(week_ending_date, "month")) # Create month column

# Subset data for COVID-19 period (2020-2024)
dat_covid <- dat %>%
  filter(year >= 2020 & year <= 2024)

# Add COVID-19 death rate and virulence to dat_covid
dat_covid <- dat_covid %>%
  mutate(covid_death_rate = (covid_19_deaths / population) * 100000, # Death rate per 100,000
         virulence = (covid_19_deaths / cases) * 100) # Case Fatality Rate (CFR) as percentag

# Add total death rate to full dat dataset
dat <- dat %>%
  mutate(rate = (total_deaths / population) * 100000) # Total death rate per 100,000

# Summarize COVID-19 deaths and cases by week_ending_date to preserve details
covid_summary <- dat_covid %>%
  group_by(week_ending_date, month) %>%
  summarize(total_covid_deaths = sum(covid_19_deaths, na.rm = TRUE),
            total_cases = sum(cases, na.rm = TRUE),
            .groups = "drop")
# Plot COVID-19 deaths over time to identify waves by week
ggplot(covid_summary, aes(x = week_ending_date, y = total_covid_deaths)) +
  geom_line(color = "blue", linewidth = 1) +
  labs(title = "COVID-19 Deaths Over Time",
       x = "Week Ending Date",
       y = "Total COVID-19 Deaths") +
  theme_minimal()

# Plot cases over time to show another wave indication by week
ggplot(covid_summary, aes(x = week_ending_date, y = total_cases)) +
  geom_line(color = "red", size = 1) +
  labs(title = "COVID-19 Cases Over Time",
       x = "Week Ending Date",
       y = "Total Cases") +
  theme_minimal()
# Reshape the data to long format for faceting
covid_summary_long <- covid_summary %>%
  select(week_ending_date, total_covid_deaths, total_cases) %>%
  pivot_longer(cols = c(total_covid_deaths, total_cases),
               names_to = "metric",
               values_to = "value")
```

```r
# Clean metric names for labeling
covid_summary_long <- covid_summary_long %>%
  mutate(metric = case_when(
    metric == "total_covid_deaths" ~ "COVID-19 Deaths",
    metric == "total_cases" ~ "COVID-19 Cases"
  ))

# 2020w12 to w26, 2020 w26 to w40, 2020w40 to 2021 w13, 2021 w29 to w45, 2021w45 to 2022w13,

# Define example MMWR weeks for vertical lines
# Define example MMWR weeks for vertical lines
mmwr_weeks <- as.Date(c("2020-03-21", "2020-06-27", "2020-10-03", "2021-04-03",
                        "2021-07-24", "2021-11-13", "2022-04-02", "2022-10-08",
                        "2023-04-08", "2023-07-29", "2024-03-11"))

# Add waves to the dat_covid dataset based on MMWR weeks
dat_covid <- dat_covid %>%
  mutate(wave = case_when(
    week_ending_date >= as.Date("2020-03-21") & week_ending_date <= as.Date("2020-06-27") ~ "
    week_ending_date > as.Date("2020-06-27") & week_ending_date <= as.Date("2020-10-03") ~ "W
    week_ending_date > as.Date("2020-10-03") & week_ending_date <= as.Date("2021-04-03") ~ "W
    week_ending_date > as.Date("2021-07-24") & week_ending_date <= as.Date("2021-11-13") ~ "W
    week_ending_date > as.Date("2021-11-13") & week_ending_date <= as.Date("2022-04-02") ~ "W
    week_ending_date > as.Date("2022-04-02") & week_ending_date <= as.Date("2022-10-08") ~ "W
    week_ending_date > as.Date("2022-10-08") & week_ending_date <= as.Date("2023-04-08") ~ "W
    week_ending_date > as.Date("2023-07-29") & week_ending_date <= as.Date("2024-03-11") ~ "W
    TRUE ~ "Other"
  ))

# Create individual datasets for each wave
dat_w1 <- dat_covid %>% filter(wave == "Wave 1")
dat_w2 <- dat_covid %>% filter(wave == "Wave 2")
dat_w3 <- dat_covid %>% filter(wave == "Wave 3")
dat_w4 <- dat_covid %>% filter(wave == "Wave 4")
dat_w5 <- dat_covid %>% filter(wave == "Wave 5")
dat_w6 <- dat_covid %>% filter(wave == "Wave 6")
dat_w7 <- dat_covid %>% filter(wave == "Wave 7")
dat_w8 <- dat_covid %>% filter(wave == "Wave 8")

# Plot deaths and cases in separate panels but aligned on the x-axis
# Include mmwr_week on x-label but reduce crowding with breaks
# Add vertical lines for MMWR weeks
```

```r
fig_1 <- ggplot(covid_summary_long, aes(x = week_ending_date, y = value)) +
  geom_line(aes(color = metric), size = 1) +
  geom_vline(xintercept = as.numeric(mmwr_weeks), linetype = "dashed", color = "black") +
  scale_color_manual(values = c("COVID-19 Deaths" = "blue", "COVID-19 Cases" = "red")) +
  facet_wrap(~ metric, scales = "free_y", ncol = 1) + # Separate panels, independent scales
  scale_x_date(date_breaks = "3 months", date_labels = "%b-%Y") + # Reduce x-axis labels
  labs(title = "COVID-19 Deaths and Cases Over Time",
       x = "Week Ending Date (approx. MMWR Weeks)",
       y = "Count",
       color = "Metric") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
fig_1
# Define wave labels and positions
wave_labels <- data.frame(
  wave = c("1", "2", "3", "4", "5", "6", "7", "8"),
  x = as.Date(c("2020-05-10", "2020-08-15", "2021-01-01", "2021-09-01",
                "2022-02-01", "2022-07-01", "2023-01-01", "2023-11-15")),
  y = 40000 # Adjust y-position for visibility
)

# Base plot with wave numbers
ggplot(covid_summary_long, aes(x = week_ending_date, y = value)) +
  geom_line(aes(color = metric), size = 1) +
  geom_vline(xintercept = as.numeric(mmwr_weeks), linetype = "dashed", color = "black") +
  geom_text(data = wave_labels, aes(x = x, y = y, label = wave),
            inherit.aes = FALSE, color = "black", size = 4, fontface = "bold") +
  scale_color_manual(values = c("COVID-19 Deaths" = "blue", "COVID-19 Cases" = "red")) +
  facet_wrap(~ metric, scales = "free_y", ncol = 1) +
  scale_x_date(date_breaks = "3 months", date_labels = "%b-%Y") +
  labs(title = "COVID-19 Deaths and Cases Over Time with Waves",
       x = "Week Ending Date (approx. MMWR Weeks)",
       y = "Count",
       color = "Metric") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))



# Compute death rates by state for each wave
death_rate_by_wave <- dat_covid %>%
  group_by(wave, state_name) %>%
```

```
  summarize(total_deaths = sum(covid_19_deaths, na.rm = TRUE),
            total_population = sum(population, na.rm = TRUE),
            death_rate = (total_deaths / total_population) * 100000, # Death rate per 100,000
            .groups = "drop")

# Rank states within each wave
death_rate_by_wave <- death_rate_by_wave %>%
  arrange(wave, desc(death_rate))

# View the summary table for each wave
#print(death_rate_by_wave)

# Plot separate death rates for each wave as individual plots
waves <- unique(death_rate_by_wave$wave)

for (w in waves) {
  wave_data <- death_rate_by_wave %>% filter(wave == w)
  p <- ggplot(wave_data, aes(x = reorder(state_name, -death_rate), y = death_rate, fill = wav
    geom_bar(stat = "identity") +
    coord_flip() +
    labs(title = paste("COVID-19 Death Rates by State -", w),
         x = "State",
         y = "Death Rate per 100,000",
         fill = "Wave") +
    theme_minimal()
  print(p)
}
# Filter data for waves 3, 5, and 7
selected_waves <- c("Wave 3", "Wave 5", "Wave 7")

# Compute death rates by state for selected waves
death_rate_selected <- dat_covid %>%
  filter(wave %in% selected_waves) %>%  # Select specific waves
  group_by(wave, state_name) %>%
  summarize(total_deaths = sum(covid_19_deaths, na.rm = TRUE),
            total_population = sum(population, na.rm = TRUE),
            death_rate = (total_deaths / total_population) * 100000, # Death rate per 100,000
            .groups = "drop")

# Plot death rates for Waves 3, 5, and 7 using facets
ggplot(death_rate_selected, aes(x = reorder(state_name, -death_rate), y = death_rate, fill =
  geom_bar(stat = "identity", show.legend = FALSE) +
```

```r
  coord_flip() +
  facet_wrap(~ wave, scales = "free_y", ncol = 3) +  # Create separate panels for each wave
  labs(title = "COVID-19 Death Rates by State for Waves 3, 5, and 7",
       x = "State",
       y = "Death Rate per 100,000") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8))  # Adjust text size for clarity
# Filter data for waves 3, 5, and 7
selected_waves <- c("Wave 3", "Wave 5", "Wave 7")

# Compute death rates by state for selected waves
death_rate_selected <- dat_covid %>%
  filter(wave %in% selected_waves) %>%  # Select specific waves
  group_by(wave, state_name) %>%
  summarize(total_deaths = sum(covid_19_deaths, na.rm = TRUE),
            total_population = sum(population, na.rm = TRUE),
            death_rate = (total_deaths / total_population) * 100000, # Death rate per 100,000
            .groups = "drop") %>%
  mutate(state_name = reorder_within(state_name, death_rate, wave))  # Order within facets

# Plot with proper ordering within facets
fig_2 <- ggplot(death_rate_selected, aes(x = state_name, y = death_rate, fill = wave)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +
  facet_wrap(~ wave, scales = "free_y", ncol = 3) +  # Separate panels for waves 3, 5, and 7
  scale_x_reordered() +  # Ensure proper ordering
  labs(title = "COVID-19 Death Rates by State for Waves 3, 5, and 7",
       x = "State",
       y = "Death Rate per 100,000") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8))  # Adjust text size for clarity
fig_2
# Compute average virulence (CFR) for each wave, excluding Wave 8
virulence_summary <- dat_covid %>%
  filter(wave != "Wave 8") %>%
  group_by(wave) %>%
  summarize(mean_virulence = mean(virulence, na.rm = TRUE),
            median_virulence = median(virulence, na.rm = TRUE),
            total_deaths = sum(covid_19_deaths, na.rm = TRUE),
            total_cases = sum(cases, na.rm = TRUE),
            overall_cfr = (total_deaths / total_cases) * 100, # Overall CFR for the wave
            .groups = "drop")
```

```r
# Print virulence summary
print(virulence_summary)

# Plot CFR across waves, excluding Wave 8
fig_3 <- ggplot(virulence_summary, aes(x = wave, y = overall_cfr, fill = wave)) +
  geom_bar(stat = "identity") +
  labs(title = "COVID-19 Case Fatality Rate (CFR) Across Waves",
       x = "Wave",
       y = "Case Fatality Rate (%)",
       fill = "Wave") +
  theme_minimal()
fig_3
# Load the dataset from the RDS file in the data directory
dat <- readRDS("../data/dat.rds")

dat <- dat |>
  mutate(across(c(total_deaths, covid_19_deaths, population, cases),as.numeric)) |>
  filter(!is.na(total_deaths), !is.na(population)) |>
  rename(
    date = week_ending_date,   # Rename weekly date column
    outcome = total_deaths     # Rename total deaths to outcome
  ) |>
  mutate(
    date = as.Date(date),      # Ensure date is in proper format
    outcome = as.integer(outcome)  # Ensure outcome is integer
  ) |>
  arrange(state_name, date)   # Sort by state and date
exclude_dates <- seq(as.Date("2020-01-01"), as.Date("2023-12-31"), by = "week")
state_fits <- dat |>
  group_by(state_name) |>
  group_split() |>
  set_names(unique(dat$state_name)) |>
  map(~ excess_model(
    counts = .x,
    start = min(.x$date),
    end = max(.x$date),
    exclude = exclude_dates,     # Exclude 2020-2024 dates
    model = "quasipoisson",      # Recommended for weekly data
    weekday.effect = FALSE,      # Turn off weekday effect for weekly data
    keep.counts = T
  ))
```

```r
excess_plot(state_fits$California)
expected_plot(state_fits$California)
expected_diagnostic(state_fits$California)
start_date <- as.Date("2020-01-01")
end_date <- as.Date("2023-12-31")

excess_mortality_all_states <- map_dfr(
  state_fits,
  ~ excess_cumulative(fit = .x, start = start_date, end = end_date),
  .id = "state_name"  # Add a column to identify the state
) |>
  group_by(state_name) |>
  mutate(
    weekly_observed = observed - lag(observed),
    weekly_observed = ifelse(is.na(weekly_observed), observed, weekly_observed)
  )|>
  rename(cum_excess = observed,
         weekly_excess = weekly_observed) |>
  select(-c(sd,fitted,se))
excess_mortality <- excess_mortality_all_states |>
  left_join(dat, by = c("state_name", "date")) |>
  filter(!is.na(weekly_excess), !is.na(covid_19_deaths))
#test with cor.test
cor.test(excess_mortality$weekly_excess, excess_mortality$covid_19_deaths)

#test with linear regression
regression_model <- lm(weekly_excess ~ covid_19_deaths, data = excess_mortality)

summary(regression_model)

# Filter data to include only up to the end of 2023
excess_mortality |>
  filter(date <= as.Date("2023-12-31")) |>
  ggplot(aes(x = date)) +
    geom_area(aes(y = weekly_excess, fill = "Excess Mortality"), linewidth = 1,alpha = 0.4) +
    geom_area(aes(y = covid_19_deaths, fill = "COVID-19 Deaths"), linewidth = 1,alpha = 0.4)
    labs(
      title = "Overlap of Weekly COVID-19 Deaths and Excess Mortality",
      x = "Date",
      y = "Number of Deaths",
      fill = "Legend"
    ) +
```

```
    scale_fill_manual(values = c("Excess Mortality" = "blue", "COVID-19 Deaths" = "red")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
      axis.title = element_text(size = 12),
      legend.position = "top"
    ) +
    facet_wrap(~state_name)

# Filter data to include only California, Florida, and Texas up to the end of 2023
selected_states <- c("California", "Florida", "Texas")

 fig_4 <- excess_mortality %>%
  filter(date <= as.Date("2023-12-31"), state_name %in% selected_states) %>%
  ggplot(aes(x = date)) +
    geom_area(aes(y = weekly_excess, fill = "Excess Mortality"), linewidth = 1, alpha = 0.4)
    geom_area(aes(y = covid_19_deaths, fill = "COVID-19 Deaths"), linewidth = 1, alpha = 0.4
    labs(
      title = "Overlap of Weekly COVID-19 Deaths and Excess Mortality",
      subtitle = "States: California, Florida, and Texas",
      x = "Date",
      y = "Number of Deaths",
      fill = "Legend"
    ) +
    scale_fill_manual(values = c("Excess Mortality" = "blue", "COVID-19 Deaths" = "red")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
      plot.subtitle = element_text(hjust = 0.5, size = 12),
      axis.title = element_text(size = 12),
      legend.position = "top"
    ) +
    facet_wrap(~state_name)
fig_4
temp <- dat_covid |>
  select(state_name,week_ending_date, wave) |>
  rename(date = week_ending_date)
excess_mortality <- excess_mortality |>
  left_join(temp, by = c("state_name", "date")) |>
  filter(!is.na(weekly_excess), !is.na(covid_19_deaths)) |>
  mutate(excess_death_rate = (weekly_excess / population) * 100000) # Death rate per 100,000
```

```r
excess_rate_by_wave <- excess_mortality %>%
  group_by(wave, state_name) %>%
  summarize(total_deaths = sum(weekly_excess, na.rm = TRUE),
            total_population = mean(population, na.rm = TRUE), #?! why sum
            death_rate = (total_deaths / total_population) * 100000, # Death rate per 100,000
            .groups = "drop")

# Rank states within each wave
excess_rate_by_wave <- excess_rate_by_wave %>%
  arrange(wave, desc(death_rate))

# View the summary table for each wave
#print(death_rate_by_wave)

# Plot separate death rates for each wave as individual plots
waves <- unique(excess_rate_by_wave$wave)

for (w in waves) {
  wave_data <- excess_rate_by_wave %>% filter(wave == w)
  p <- ggplot(wave_data, aes(x = reorder(state_name, -death_rate), y = death_rate, fill = wav
    geom_bar(stat = "identity") +
    coord_flip() +
    labs(title = paste("EXCESS Death Rates by State -", w),
         x = "State",
         y = "Death Rate per 100,000",
         fill = "Wave") +
    theme_minimal()
  print(p)
}
# Filter to include only Wave 3, Wave 5, and Wave 7
selected_waves <- c("Wave 3", "Wave 5", "Wave 7")

excess_rate_by_wave_filtered <- excess_mortality %>%
  filter(wave %in% selected_waves) %>%
  group_by(wave, state_name) %>%
  summarize(total_deaths = sum(weekly_excess, na.rm = TRUE),
            total_population = mean(population, na.rm = TRUE),
            death_rate = (total_deaths / total_population) * 100000, # Death rate per 100,000
            .groups = "drop") %>%
  arrange(wave, desc(death_rate))

# Plot death rates for Wave 3, Wave 5, and Wave 7 in one plot with facets
```

```r
fig_5 <- ggplot(excess_rate_by_wave_filtered,
       aes(x = reorder_within(state_name, death_rate, wave),
           y = death_rate, fill = wave)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_x_reordered() + # Properly orders within each facet
  facet_wrap(~ wave, ncol = 3, scales = "free_y") + # One row, three columns
  labs(title = "Excess Mortality Rates by State for Waves 3, 5, and 7",
       x = "State",
       y = "Excess Death Rate per 100,000",
       fill = "Wave") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 12, face = "bold"),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.text.y = element_text(size = 8),
    legend.position = "none"
  )
fig_5
fig_1
fig_2
fig_3
fig_4
fig_5
```