BST 260 Report

Irisa Jin, Xu Han, Cenjia Zeng

Abstract

...

Introduction

The COVID-19 pandemic has deeply impacted global health and mortality, displaying some vulnerabilities in healthcare systems worldwide. Official statistics report millions of deaths, but evidence suggests the actual toll is 1.6 times higher than reported due to underreporting and the indirect effects of the pandemic (Karlinsky & Kobak, 2021; Shang et al., 2022). Excess mortality, the difference between observed and expected deaths, has become a critical measure for understanding the pandemic's full impact. This metric includes both direct effects, such as deaths caused by the virus, and indirect consequences, like delayed medical care and strained healthcare systems (Kung et al., 2020). By offering a comprehensive view, excess mortality can guide public health policies and improve preparedness for future crises (Araja, 2022; Bech et al., 2021).

Globally, the impact of COVID-19 on mortality has varied. Some regions, like South America and parts of Asia, experienced high excess mortality due to severe outbreaks and healthcare challenges (Shang et al., 2022). In contrast, other areas saw reductions in non-COVID deaths, as measures like lockdowns limited the spread of other infectious diseases (Karlinsky & Kobak, 2021). In the United States, the pandemic showed significant health disparities. African American and Hispanic communities suffered disproportionately high death rates, largely due to structural inequalities like limited healthcare access and economic barriers (Siegel et al., 2021; Goldstein & Atherwood, 2020). These inequities underscore the importance of state-level analyses to identify patterns and develop equitable health interventions (Kung et al., 2020).

Evidence suggests that COVID-19 became less severe over time in some populations. For instance, by early 2022, case-hospitalization rates had decreased by 90% in locations like Spain

and Boston compared to the start of the pandemic (Amer et al., 2022). Similarly, Massachusetts healthcare workers experienced fewer severe complications after April 2020 (Lan et al., 2020). Nationwide U.S. data from April 2020 to April 2021 showed that severe cases were a minority of hospital admissions, with their proportion decreasing over time (Whitfield et al., 2021). However, these improvements were not evenly distributed. Vulnerable groups, including racial and ethnic minorities, older adults, and socioeconomically disadvantaged populations, continued to experience higher rates of severe illness and death (Bergmann et al., 2022). This emphasizes the need to study changes in COVID-19 severity across different waves and populations to introduce adaptive public health strategies.

This study aims to analyze mortality trends in the United States from January 2020 to December 2024. Using daily and weekly mortality data, it will divide the pandemic into distinct waves based on trends observed through data visualizations. For each wave, the study will show the calculation of state-level death rates to identify which states were most and least affected. It will also assess whether COVID-19 became less severe over time by examining hospitalization and death rates. Additionally, this study will estimate weekly excess mortality for each state and evaluate how much of it can be attributed to COVID-19 deaths. By exploring both COVID-19-related and unrelated mortality, this study seeks to provide a clearer picture of the pandemic's full impact. Through careful analysis and visualization, the study's findings aim to inform public health strategies and improve readiness for future health emergencies (Ramírez-Soto & Ortega-Cáceres, 2022; Stokes et al., 2021).

Methods

The data used in this project are from multiple sources, including COVID-19 cases, deaths, hospitalizations, and vaccination records from CDC APIs, state-level population estimates from the US Census Bureau, and state-to-region mappings from publicly available JSON files. The estimated mortality was built based on the 2014-2019 mortality data which was sourced from CDC. Specifically, the time series approach was implemented by using state-level total death with 52 weeks per year. However, it was found that 2014 had 53 weeks which was considered as an outlier and was omitted from the analysis. The ARIMA model was used for fitting procedures in order to identify patterns and seasonality. A 208-week data frame was generated to forecast 2020 - 2023 mortality-related data. The results were aggregated to produce estimated mortality for each state without interruption from the COVID pandemic. The main data preprocessing involved cleaning, aggregating, and joining datasets by state and week using the CDC's MMWR format and saved into "dat" data frame in the "WrangledData.rdata" file. Rates per 100,000 people were calculated for comparability across states and time periods. Then, the estimated data was merged with "dat" by year, week, and state. Later analysis was conducted based on these cleaned data. All programming was conducted in the R language. When dividing wave periods, data visualization packages were used to examine the data. The visualized graphs depicted a peak sandwiched between two flat zones; this was considered the epidemic wave. The pandemic was divided into three periods with more accurate time nodes

based on time-series trends publish by the CDC: (1) starting from the initial global outbreak (Jan 25, 2020 – Feb 28, 2021), characterized by the first significant wave of COVID-19; (2) the beginning of the Delta Variant Surge (Mar 1, 2021 – Nov 30, 2021), marked by a rise in cases and deaths associated with the Delta variant; and (3) seeing the Omicron wave (Dec 1, 2021 – Apr 29, 2023), notable for a sharp increase in cases but a lower mortality rate due to widespread vaccinations.

State-level death rates were calculated for each period, and trends in hospitalization and death rates were analyzed to assess changes in virulence over time. Visualizations, including time-series plots and bar charts, were created using ggplot2 to support these analyses.

To test the relationship between COVID-19 and the excess mortality, the linear regression was conducted to examine the association between the COVID-related mortality and the excess mortality. To make the results more visible, a scatter plot with fitted lines was generated. A box plot visualizes the distribution of excess deaths across states, highlighting the median, interquartile range, and potential outliers, which helps identify variability and patterns in excess deaths and provides insight into the excess mortality in different states. The box plot also supports hypothesis testing and comparisons among states. A line chart was used to look at how well the actual number of mortality and the number of estimated deaths plus the number of COVID deaths fit. The closer fitted lines represent more similar trends, indicating that COVID may be used to explain excess mortality.

Lastly, the above processes regarding states' performance in COVID were repeated for excess mortality. After separating the results by state, the weekly average mortality rate and total mortality rate were examined. The rate calculations were based on rate per 100,000 people.

Results

The COVID-19 pandemic was divided into three periods based on the observed trends in case and death rates across the US states, as well as outside information. From Figure 1, three major peaks were defined to be both present in case and death rate. They corresponded to and included the initial global outbreak, Delta Variant surge, and the Omicron wave; this was matched to information given by the official CDC website. The initial global outbreak (Jan 25, 2020 – Feb 28, 2021) is marked by the first large peak in cases and deaths, with a steep rise in both metrics during late 2020 and early 2021. The Delta Variant surge (Mar 1, 2021 – Nov 30, 2021) shows a distinct increase in both case and death rates, albeit with a broader and steady peak. Lastly, the Omicron wave (Dec 1, 2021 – Apr 29, 2023) features a sharp, high spike in case rates but a comparatively smaller rise in death rates, suggesting a shift in the pandemic dynamics and the advancements in scientific discoveries (widespread of vaccines).

State-level death rates varied significantly across the defined periods. From Figure 2, it can be clearly seen that in period (1), states like Rhode Island, New Jersey, and Mississippi reported the highest death rates, reflecting the early and severe impact of the pandemic in these regions,

with Vermont and Hawaii at the lowest death rates. During the (2) phase, West Virginia, Florida, and Tennessee experienced the highest death rates, while Vermont, New Hampshire, and Rhode Island showed the lowest death rates. In (3) time period, West Virginia, Kentucky, and Ohio had the highest death rates, while Alaska, Hawaii, and Vermont had the lowest death rates. Interestingly, Vermont, Hawaii, and Alaska remains having the lowest death rates throughout the pandemic, possibly due to a small population and less ability to spread the disease.

The virulence of COVID-19 appeared to decrease over time, as evidenced by Figure 3. In (1), the vaccination (regular and booster, which has not been available yet) remains to be almost 0; the hospitalization rate increases rapidly in this time period. In (2), the regular series is increasing and people also begin to take booster vaccines; the hospitalization rate is not as high as (1). There is a large peak of hospitalization in (3) but declines quickly, possibly due to high vaccination (including booster) rates. This conclusion can also be explained by a decreasing death rate.

The time series analysis produced estimated data for 2020 - 2023 as expected. The result from linear regression is shown in Table 1. From the result, the excess mortality can be explained by the COVID pandemic. The LINE assumptions check passed by checking Q-Q plot and residual plot. A scatterplot was used to more clearly depict the relationship between the variables. It can be seen in the scatterplot that the dots form a straight line. In the Figure 4, it can also be seen that when COVID mortality was added to the actual mortality, it was very close to the estimated mortality, especially in 2020 - 2022.

State-level excess death rates varied significantly across the defined periods. A more negative excess mortality rate represents a larger rate. From Figure 4, it can be clearly seen that in period (1), states like Mississippi, Arizona, and Louisiana reported the highest excess death rates, with Vermont and Hawaii at the lowest death rates. During the (2) phase, West Virginia, Florida, and Montana experienced the highest excess death rates, while Massachusetts, District of Columbias and Minnesota showed the lowest excess death rates. In (3) time period, West Virginia, New Mexico, and Puerto Rico had the highest death rates, while District of Columbia, New Jersey, and Minnesota had the lowest death rates. Unlike results found for question two, no states remained having consistently the lowest death rate.

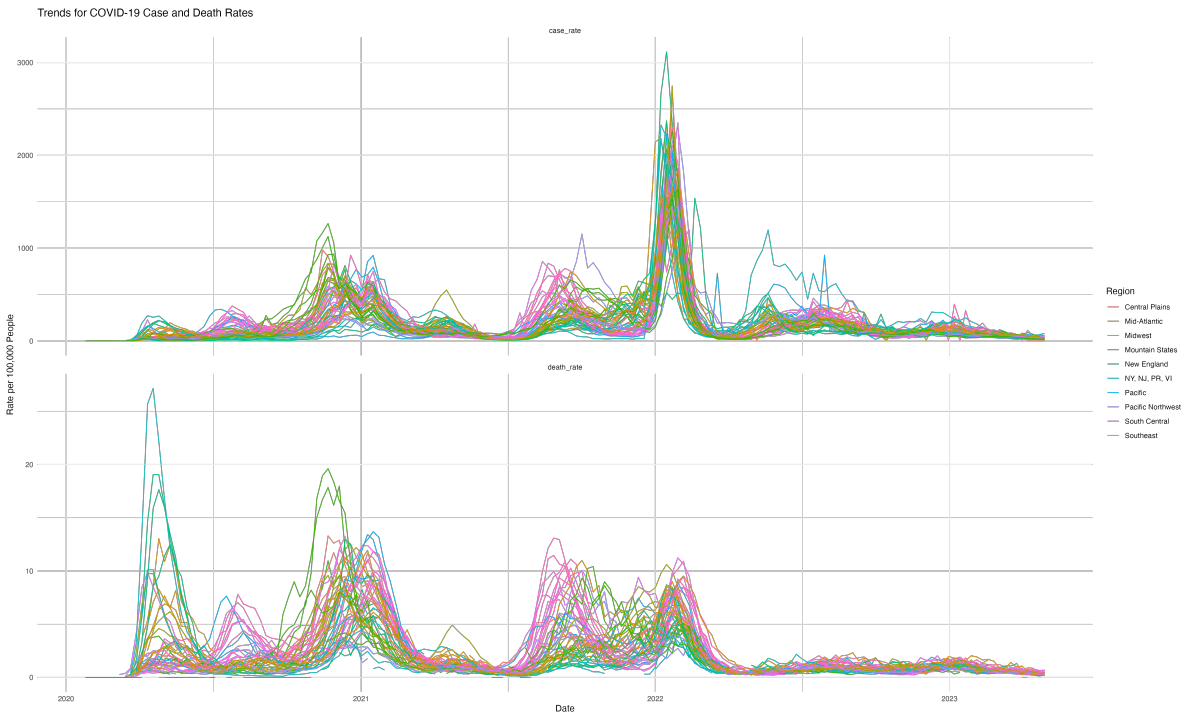
Trends for COVID-19 Case and Death Rates
case_rate
death_rate
Rate per 100,000 People
Date
Region
Central Plains
Mid-Atlantic
Midwest
Mountain States
New England
NY, NJ, PR, VI
Pacific
Pacific Northwest
South Central
Southeast

Figure 1

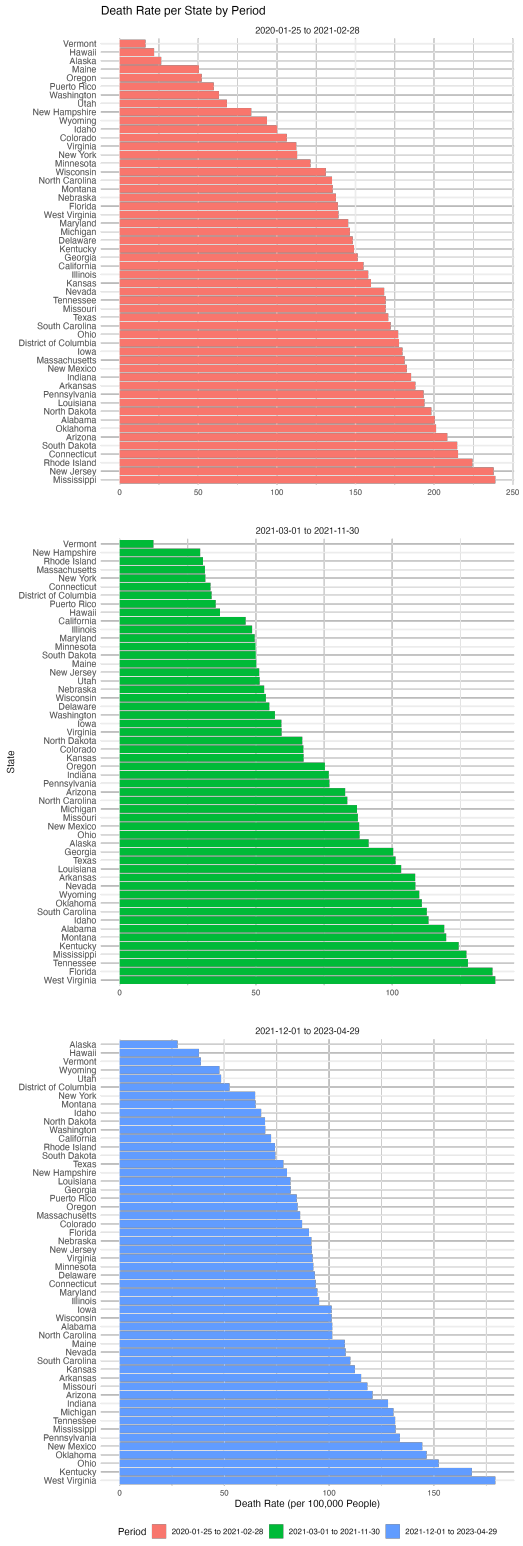
Death Rate per State by Period
2020-01-25 to 2021-02-28
Vermont
Hawaii
Alaska
Maine
Oregon
Puerto Rico
Washington
Utah
New Hampshire
Wyoming
Idaho
Colorado
Virginia
New York
Minnesota
Wisconsin
North Carolina
Montana
Nebraska
Florida
West Virginia
Maryland
Michigan
Delaware
Kentucky
Georgia
California
Illinois
Kansas
Nevada
Tennessee
Missouri
Texas
South Carolina
Ohio
District of Columbia
Iowa
Massachusetts
New Mexico
Indiana
Arkansas
Pennsylvania
Louisiana
North Dakota
Alabama
Oklahoma
Arizona
South Dakota
Connecticut
Rhode Island
New Jersey
Mississippi
0
50
100
150
200
250
2021-03-01 to 2021-11-30
Vermont
New Hampshire
Rhode Island
Massachusetts
New York
Connecticut
District of Columbia
Puerto Rico
Hawaii
California
Illinois
Maryland
Minnesota
South Dakota
Maine
New Jersey
Utah
Nebraska
Wisconsin
Delaware
Washington
Iowa
Virginia
North Dakota
Colorado
Kansas
Oregon
Indiana
Pennsylvania
Arizona
North Carolina
Michigan
Missouri
New Mexico
Ohio
Alaska
Georgia
Texas
Louisiana
Arkansas
Nevada
Wyoming
Oklahoma
South Carolina
Idaho
Alabama
Montana
Kentucky
Mississippi
Tennessee
Florida
West Virginia
0
50
100
State
2021-12-01 to 2023-04-29
Alaska
Hawaii
Vermont
Wyoming
Utah
District of Columbia
New York
Montana
Idaho
North Dakota
Washington
California
Rhode Island
South Dakota
Texas
New Hampshire
Louisiana
Georgia
Puerto Rico
Oregon
Massachusetts
Colorado
Florida
Nebraska
New Jersey
Virginia
Minnesota
Delaware
Connecticut
Maryland
Illinois
Iowa
Wisconsin
Alabama
North Carolina
Maine
Nevada
South Carolina
Kansas
Arkansas
Missouri
Arizona
Indiana
Michigan
Tennessee
Mississippi
Pennsylvania
New Mexico
Oklahoma
Ohio
Kentucky
West Virginia
0
50
100
150
Death Rate (per 100,000 People)
Period
2020-01-25 to 2021-02-28
2021-03-01 to 2021-11-30
2021-12-01 to 2023-04-29

Figure 2

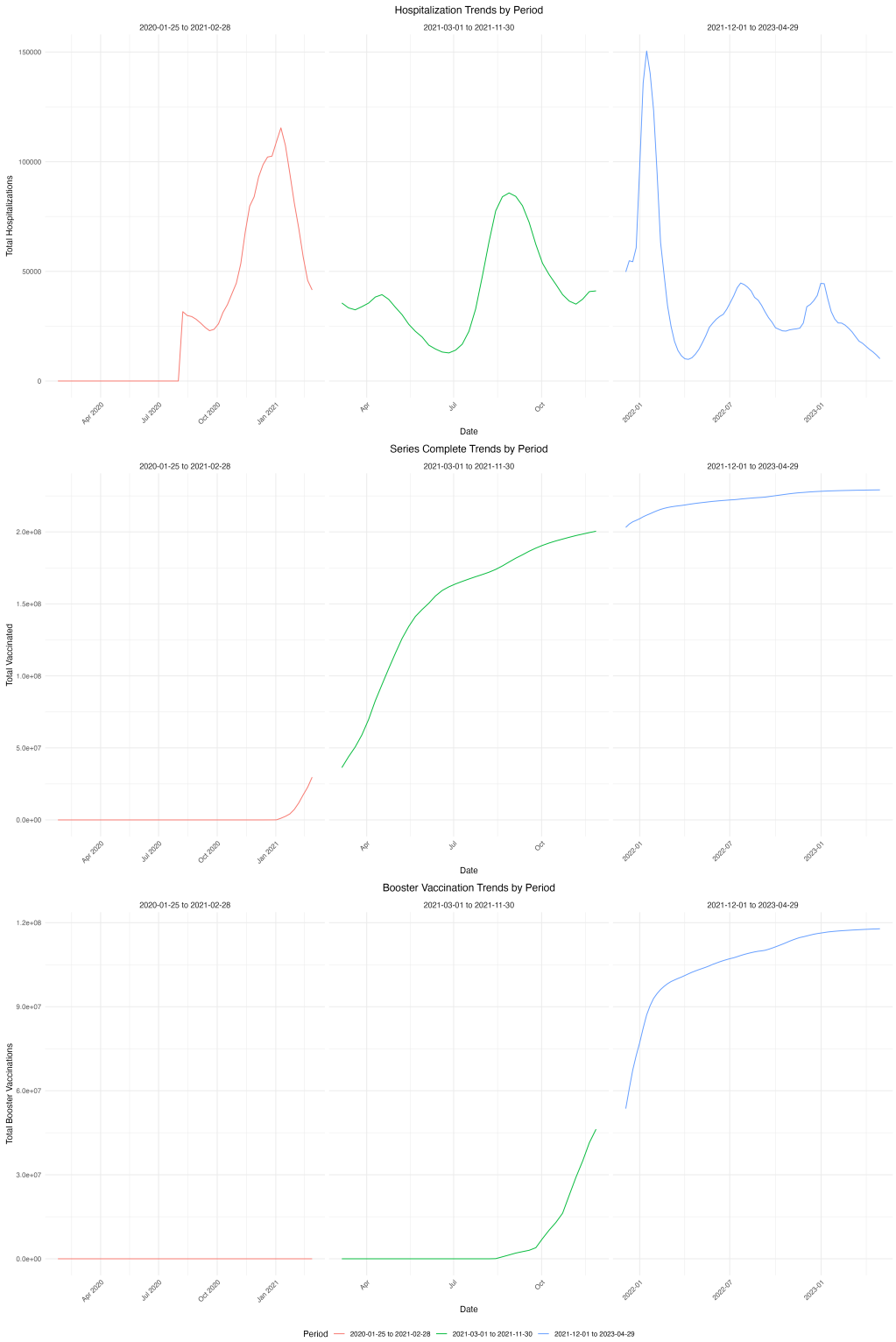
Hospitalization Trends by Period
2020-01-25 to 2021-02-28
2021-03-01 to 2021-11-30
2021-12-01 to 2023-04-29
Total Hospitalizations
150000
100000
50000
0
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr
Jul
Oct
2022-01
2022-07
2023-01
Date
Series Complete Trends by Period
Total Vaccinated
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Booster Vaccination Trends by Period
Total Booster Vaccinations
1.2e+08
9.0e+07
6.0e+07
3.0e+07
0.0e+00
Period
2020-01-25 to 2021-02-28
2021-03-01 to 2021-11-30
2021-12-01 to 2023-04-29

Figure 3

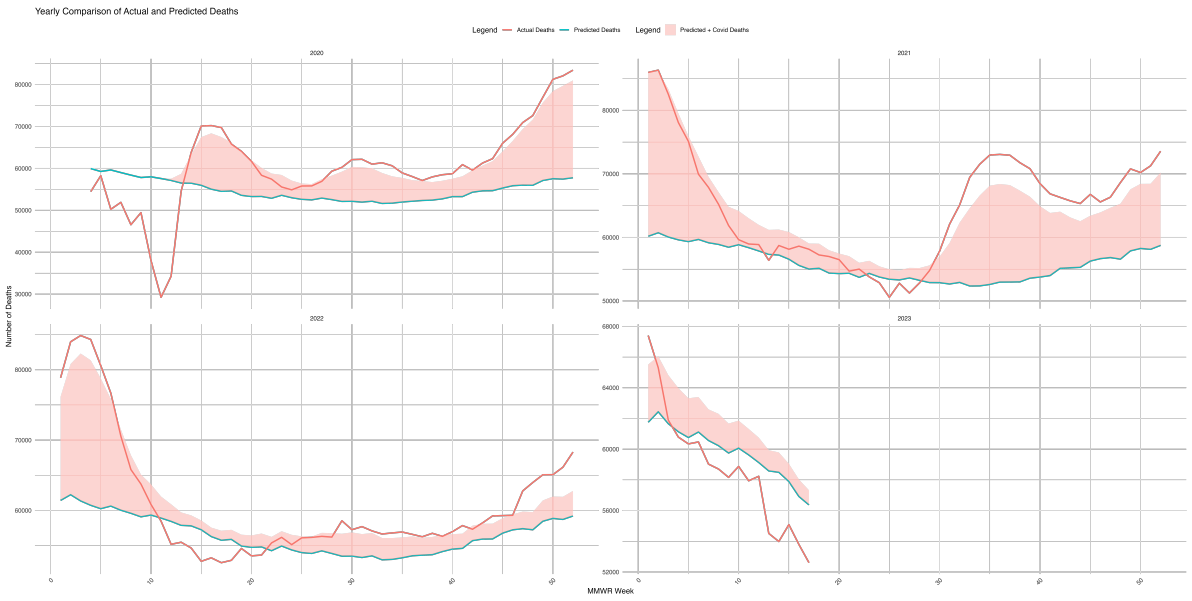

Figure 4

```
Call:
lm(formula = excess_deaths ~ deaths, data = q4)

Residuals:
     Min       1Q   Median       3Q      Max
-1036.77   -35.75     1.32    38.15   676.42

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -5.318621   1.241431   -4.284 1.86e-05 ***
deaths      -1.178869   0.004103 -287.307  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.66 on 7581 degrees of freedom
  (1257 observations deleted due to missingness)
Multiple R-squared:  0.9159,    Adjusted R-squared:  0.9159
F-statistic: 8.255e+04 on 1 and 7581 DF,  p-value: < 2.2e-16
```

Discussion

Reference