Final: COVID-19 pandemic in the US

Abstract

This study investigates excess mortality across U.S. states from 2020 to 2024 in the context of the COVID-19 pandemic. While official COVID-19 death counts provide a baseline estimate, they may underestimate the true impact due to reporting delays and indirect effects. To better capture the pandemic's full mortality burden, we analyzed excess deaths, defined as observed deaths above historical expectations. We used CDC mortality data from 2015–2024, calculating weekly excess deaths and stratifying them by state and pandemic wave. Regression models were fitted at the state level to evaluate the relationship between COVID-19 deaths and excess mortality, with R^2 values used to assess model fit. We further explored the role of vaccination coverage and obesity prevalence in predicting death rates. Our results show that COVID-19 deaths account for a substantial portion of excess mortality, though unexplained variation remains in some states. States like California and Texas experienced consistently high excess deaths, while states such as Vermont and Wyoming remained low. Patterns varied significantly by wave, reflecting changes in virus virulence and public health responses. This analysis enhances understanding of geographic and temporal patterns of pandemic-related mortality and suggests areas for targeted health system strengthening and future research.

Introduction

The COVID-19 pandemic has had a profound impact on public health, society, and the economy, with the United States emerging as one of the countries most severely affected. Since the virus was first detected in early 2020, over one million confirmed COVID-19-related deaths have been reported in the U.S. alone (CDC, 2024). These deaths, while staggering, may only represent a fraction of the true toll of the pandemic due to underreporting, limited testing early in the outbreak, and variations in death certification practices across jurisdictions. Additionally, COVID-19 not only caused direct mortality through infection but also indirectly contributed to increased deaths due to disruptions in healthcare access, delayed medical procedures, mental health deterioration, and socioeconomic stressors (Woolf et al., 2021).

To assess the mortality burden of the pandemic, two related but distinct measures are commonly used: death rate and excess mortality. The death rate is defined as the number of deaths in a population per unit of population, typically per 100,000 individuals. This measure allows for standardized comparisons across populations of different sizes and is widely used to track the severity of outbreaks over time and across regions. However, death rates derived solely from COVID-19-confirmed deaths may significantly underestimate the true impact of the pandemic. Excess mortality addresses this limitation by estimating the number of deaths above what would have been expected based on historical trends. It provides a more comprehensive indicator of the pandemic's full mortality burden, capturing both direct and indirect effects. According to the World Health Organization (2022), excess mortality reflects not only COVID-19 deaths but also deaths from other causes that may have been exacerbated by the pandemic's strain on healthcare systems. This makes excess mortality a valuable metric for retrospective analysis of the pandemic's impact, especially when official death counts are incomplete or delayed.

To assess the mortality burden of the pandemic, two related but distinct measures are commonly used: death rate and excess mortality. The death rate is defined as the number of deaths in a population per unit of population, typically per 100,000 individuals. This measure allows for standardized comparisons across populations of different sizes and is widely used to track the severity of outbreaks over time and across regions. However, death rates derived solely from COVID-19-confirmed deaths may significantly underestimate the true impact of the pandemic. Excess mortality addresses this limitation by estimating the number of deaths above what would have been expected based on historical trends. It provides a more comprehensive indicator of the pandemic's full mortality burden, capturing both direct and indirect effects. According to the World Health Organization (2022), excess mortality reflects not only COVID-19 deaths but also deaths from other causes that may have been exacerbated by the pandemic's strain on healthcare systems. This makes excess mortality a valuable metric for retrospective analysis of the pandemic's impact, especially when official death counts are incomplete or delayed.

CDC. (2024). COVID-19 Data Tracker. U.S. Centers for Disease Control and Prevention. https://covid.cdc.gov

Woolf, S. H., Chapman, D. A., Sabo, R. T., Zimmerman, E. B. (2021). Excess Deaths From COVID-19 and Other Causes, March–July 2020. JAMA, 324(15), 1562–1564. https://doi.org/10.1001/jama.2020.19545

WHO. (2022). Global excess deaths associated with COVID-19. https://www.who.int/data/stories/global-excess-deaths-associated-with-covid-19

Methods

Overview and Purpose

This study examines geographic and temporal variation in excess mortality associated with the COVID-19 pandemic in the United States. Our methodological approach integrates multiple national datasets to assess both direct and indirect mortality effects of the pandemic. We focused on reproducibility, transparency, and robustness throughout the data pipeline.

Data Sources and Collection

We obtained six primary datasets covering the period from **January 1, 2015 to March 31, 2025**:

- **All-cause mortality and COVID-19 deaths**: CDC National Center for Health Statistics (NCHS), accessed via the weekly deaths API. This dataset includes weekly state-level death counts, disaggregated by cause.
- **COVID-19 case counts**: Sourced from the CDC COVID Data Tracker, reporting weekly laboratory-confirmed new infections.
- **Hospitalizations**: Weekly counts of COVID-19-related admissions from CDC's COVID-NET system.
- **Vaccination data**: Weekly cumulative counts of primary series and booster doses, obtained from the CDC's vaccine coverage dashboard.
- **Obesity prevalence**: Annual, state-level adult obesity rates from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), processed to align with weekly temporal resolution.
- **Population estimates**: U.S. Census Bureau mid-year state population estimates for 2020–2024, linearly interpolated to weekly intervals.

All data were accessed via public APIs or downloaded as CSVs and JSON files. Download scripts were written in R using `httr2`, and files were archived with reproducible time-stamps.

```r
library(httr2)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2     
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(lubridate)
library(dplyr)
library(ggplot2)
library(ISOweek)
library(broom)

get_cdc_data <- function(endpoint) {
  request(endpoint) |>
    req_url_query("$limit" = 10000000) |>
    req_perform() |>
    resp_body_json(simplifyVector = TRUE)
}
cases_raw <- get_cdc_data("https://data.cdc.gov/resource/pwn4-m3yp.json")
hosp_raw <- get_cdc_data("https://data.cdc.gov/resource/39z2-9zu6.json")
deaths_raw <- get_cdc_data("https://data.cdc.gov/resource/r8kw-7aab.json")
vax_raw <- get_cdc_data("https://data.cdc.gov/resource/rh2h-3yt2.json")
obesity <- read.csv("obesity_rate_2020_2023_cleaned.csv")
deaths_1519 <- read.csv("death1519.csv")
```

Data Cleaning and Preparation

Data preprocessing was conducted using **R version 4.3.3** and the **tidyverse 2.0.0** collection. Major cleaning steps included:

1. **Standardizing geographic identifiers**: Full state names were mapped to USPS abbreviations, including special handling for D.C.
2. **Temporal alignment**: Year-week combinations were converted to ISO 8601-compliant dates using the `ISOweek2date()` function, ensuring consistency across datasets.
3. **Type coercion and validation**: All numeric variables were cast to appropriate types. Implausible or negative values were treated as missing.
4. **Population interpolation**: Annual estimates were linearly interpolated to weekly resolution to enable rate calculations.
5. **Integration**: Cleaned datasets were merged on state and week. Additional derived fields, such as case-fatality ratios and hospitalizations per capita, were computed for analysis.

These steps were scripted and executed reproducibly via `quarto render` and version-controlled in a public GitHub repository.

```r
##cases
cases_clean <- cases_raw |> select(state, start_date, cases = new_cases) |>
  mutate(week = epiweek(start_date), year = epiyear(start_date))|>
  mutate(cases = as.integer(cases)) |> select(-start_date)

##deaths
deaths_clean <- deaths_raw |> select(state, year, week = mmwr_week, deaths = covid_19_deaths
  mutate(year = case_when(str_detect(year, '/') ~
           str_extract(year, "(?<=/)[0-9]+"), # if year contains /, e.g. 2019/2020, then onl
         TRUE ~ year)) |> mutate(year = as.numeric(year))  |>
  mutate(week = as.numeric(week)) |> mutate(deaths = as.numeric(deaths)) |>filter(state != "

##vax
vax_clean <- vax_raw |> filter(date_type == "Admin") |>
  select(
    state = location,
    date,
    series_complete_daily,
    booster_daily
  ) |>
  mutate(
    date = ymd_hms(date),
    series_complete_daily = as.numeric(series_complete_daily),
    booster_daily = as.numeric(booster_daily)
  ) #standardize format

##hosp
hosp_clean <- hosp_raw |> select(jurisdiction, collection_date, hosp = new_covid_19_hospital
  mutate(week = epiweek(collection_date), year = epiyear(collection_date))|>
  mutate(hosp = as.integer(hosp)) |> select(-collection_date)|>
  group_by(jurisdiction, week, year) |>
  summarise(hosp = sum(hosp))  |> ungroup() |> rename(state = jurisdiction)
```

```
`summarise()` has grouped output by 'jurisdiction', 'week'. You can override
using the `.groups` argument.
```

```
##population 2020-2024
pop <- read.csv("NST-EST2024-ALLDATA.csv")
pop <- pop |> select(STATE, NAME, POPESTIMATE2020, POPESTIMATE2021, POPESTIMATE2022, POPESTI
  rename(
    state = NAME,
    `2020` = POPESTIMATE2020,
    `2021` = POPESTIMATE2021,
    `2022` = POPESTIMATE2022,
```

```r
    `2023` = POPESTIMATE2023,
    `2024` = POPESTIMATE2024
  ) |>
  select(state, `2020`, `2021`, `2022`, `2023`, `2024`) |>
  pivot_longer(
    cols = c("2020", "2021", "2022", "2023", "2024"),
    names_to = "year",
    values_to = "population"
  ) |>
  mutate(year = as.integer(year))
```

```r
# Data clean
state_mapping <- tibble(
  full = c(state.name, "District of Columbia", "New York City", "Puerto Rico",
           "Guam", "American Samoa", "Virgin Islands",
           "Federated States of Micronesia", "Marshall Islands",
           "Northern Mariana Islands", "Palau"),
  abb = c(state.abb, "DC", "NY", "PR", "GU", "AS", "VI", "FSM", "RMI", "MP", "PW")
)

clean_state <- function(df, state_col) {
  df |>
    mutate(state_temp = !!sym(state_col)) |>
    left_join(state_mapping, by = c("state_temp" = "full")) |>
    mutate(state = coalesce(abb, state_temp)) |>
    select(-state_temp, -abb)
}

# finally cases, deaths, vax, hosp, and population data
cases_final <- clean_state(cases_clean, "state")
deaths_final <- clean_state(deaths_clean, "state")
pop_final <- clean_state(pop, "state")
vax_final <- vax_clean |>
  mutate(
    year = year(date),
    week = epiweek(date)
  ) |>
  select(-date) |>
  group_by(state, year,week) |>
  summarise(
    series_complete = sum(series_complete_daily, na.rm = TRUE),
    booster = sum(booster_daily, na.rm = TRUE),
    .groups = "drop"
  )
hosp_final <- clean_state(hosp_clean, "state")

##merged data
merged_data <- cases_final |>
  left_join(deaths_final, by = c("state", "year", "week"))
merged_data <- merged_data |>
  left_join(pop_final, by = c("state", "year"))
merged_data <- merged_data |>
  arrange(state, year, week)
```

Pandemic Wave Definitions

We classified the pandemic into four major temporal waves based on CDC reports and dominant variant periods:

- **Initial outbreak** (2020-01-01 to 2021-06-30)
- **Pre-Omicron** (2021-07-01 to 2021-11-30)
- **Omicron peak** (2021-12-01 to 2022-03-31)
- **Rebound** (2022-04-01 to 2023-03-31)

Each observation was assigned to a wave using date cutoffs.

Outcome Measures

We evaluated two key outcomes:

- **COVID-19 death rate**: Defined as weekly confirmed COVID-19 deaths per 100,000 population.
- **Excess mortality**: Calculated as the difference between observed all-cause deaths and expected deaths, where expected values were the five-year weekly average for 2015–2019 by state and week.

Analytical Techniques

1. **Descriptive analysis**: Trends in cases, deaths, and hospitalizations were visualized with time-series plots. Boxplots and dot plots summarized distributional differences across states and waves.

2. **Regression modeling**: We estimated the association between weekly COVID-19 death rate and key covariates—vaccination coverage and obesity prevalence—using a linear model with robust standard errors:

   `lm(death_rate ~ series_complete + ObesityRate, data = df)`

3. **Attribution analysis**: To quantify alignment between reported COVID-19 deaths and total excess deaths, we fitted state-level linear models:

   `excess_deaths ~ covid_deaths`

   (Slope coefficients and R^2 values were used to interpret explanatory power.)

4. **Virulence estimation**: Case fatality ratios (CFR) and hospitalization-to-case ratios were calculated by wave. Differences were tested using Kruskal–Wallis tests due to non-normality.

Assumption and limitations:

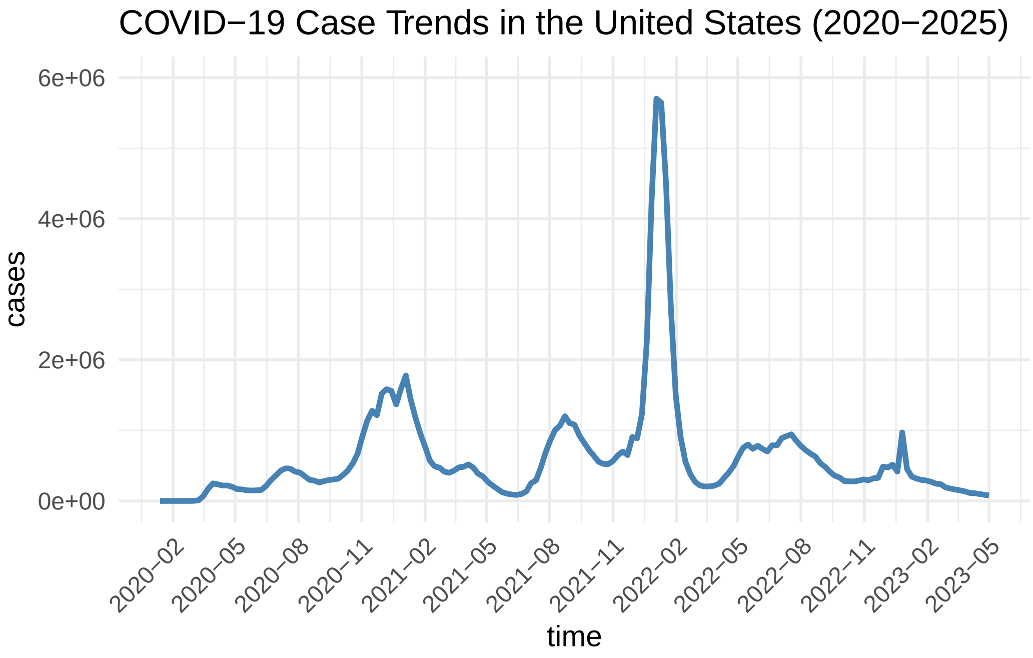
We assume stability in baseline mortality patterns from 2015–2019, and that underreporting biases are relatively constant within each state. Limitations include potential lag in death reporting, especially in less populous states, and residual confounding from unmeasured variables such as healthcare access or comorbidities.

Results

1. Wave Definition

```r
merged_data <- merged_data |>
  mutate(
    iso_week = paste0(year, "-W", sprintf("%02d", week), "-1"),
    date = ISOweek2date(iso_week))

summary_data <- merged_data |>
  group_by(date) |>
  summarise(total_cases = sum(cases, na.rm = TRUE)) |>
  ungroup()

ggplot(summary_data, aes(x = date)) +
  geom_line(aes(y = total_cases), color = "steelblue", size = 1, na.rm = TRUE) +
  labs(
    title = "COVID-19 Case Trends in the United States (2020-2025)",
    x = "time",
    y = "cases"
  ) +
  theme_minimal() +
  scale_x_date(date_breaks = "3 months", date_labels = "%Y-%m") +
  scale_y_continuous(limits = c(0, 6000000)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Weekly COVID-19 case counts in the United States from January 2020 through March 2025 displayed distinct temporal patterns aligned with the progression of the pandemic. An initial surge began in early 2020, peaking during the winter of 2020–2021, followed by a moderate but sustained wave through late 2021.

The most prominent spike occurred during the Omicron wave, which began in December 2021 and peaked in early 2022. During this period, weekly reported cases exceeded 5.5 million, making it the most intense transmission phase throughout the study period. After the Omicron peak, the rebound period from April 2022 to early 2023 featured several smaller waves, with lower but still notable levels of transmission. By late 2024, case counts had declined substantially, with minimal residual activity observed into early 2025.

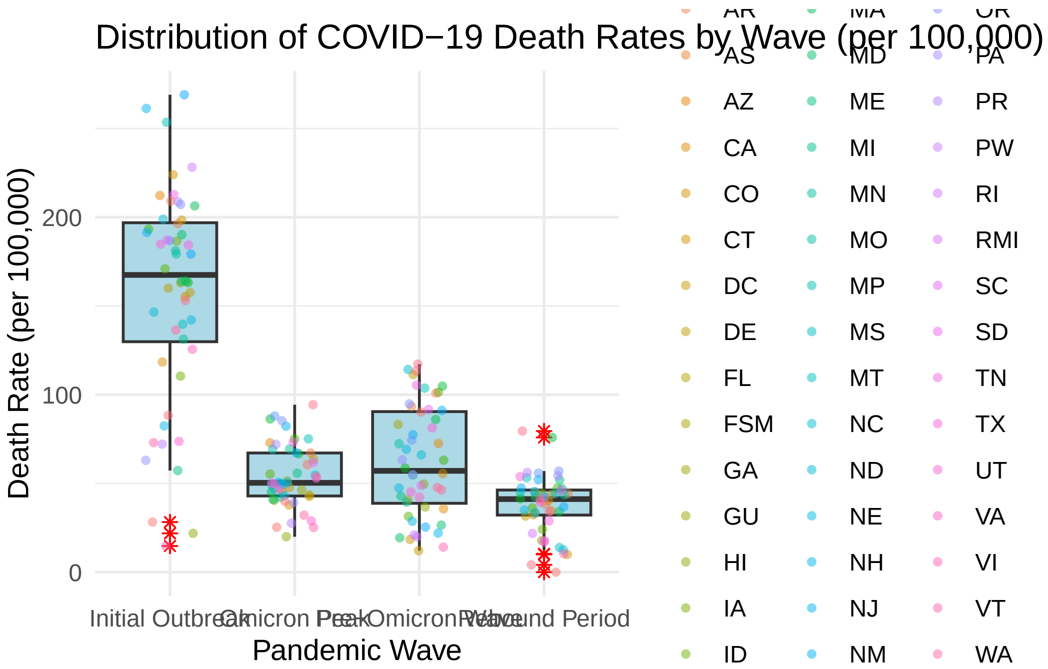
These trends are visually presented in the figure above, which illustrates the temporal distribution of cases.

2. COVID Death Rates

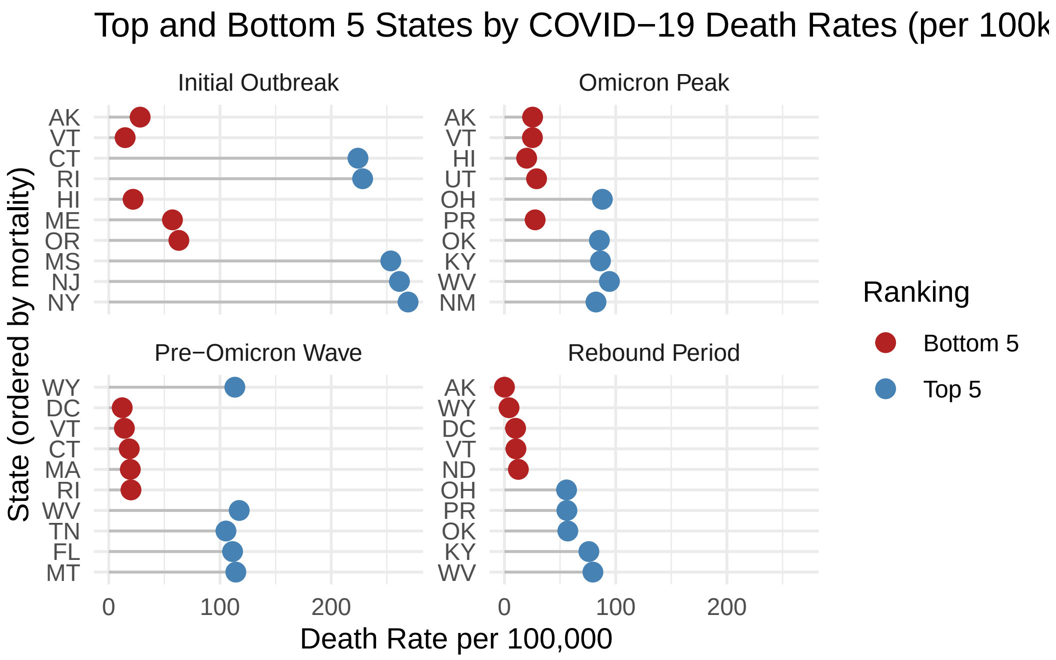
```
## Period
merged_data <- merged_data |>
  mutate(wave = case_when(
    date >= ymd("2020-01-01") & date <= ymd("2021-06-30") ~ "Initial Outbreak",
    date >= ymd("2021-07-01") & date <= ymd("2021-11-30") ~ "Pre-Omicron Wave",
    date >= ymd("2021-12-01") & date <= ymd("2022-03-31") ~ "Omicron Peak",
    date >= ymd("2022-04-01") & date <= ymd("2023-03-31") ~ "Rebound Period",
    TRUE ~ NA_character_
  )) |>
  filter(!is.na(wave))

## mortality calculation
state_wave_mortality <- merged_data|>
  group_by(state, wave) |>
  summarise(
    total_deaths = sum(deaths, na.rm = TRUE),
    avg_population = mean(population, na.rm = TRUE),
    .groups = "drop"
  ) |>
  mutate(death_rate_per_100k = total_deaths / avg_population * 100000)

## 1. Boxplot
ggplot(state_wave_mortality, aes(x = wave, y = death_rate_per_100k)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red", outlier.shape = 8) +
  geom_jitter(aes(color = state), width = 0.2, alpha = 0.5, size = 1) +
  labs(
    title = "Distribution of COVID-19 Death Rates by Wave (per 100,000)",
    x = "Pandemic Wave",
    y = "Death Rate (per 100,000)"
  ) +
  theme_minimal()
```

```
Warning: Removed 32 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

```
Warning: Removed 32 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```r
## 2. Dot plot
top_states_plot <- state_wave_mortality |>
  group_by(wave) |>
  mutate(
    top_rank = min_rank(desc(death_rate_per_100k)),
    bottom_rank = min_rank(death_rate_per_100k),
    group = case_when(
      top_rank <= 5 ~ "Top 5",
      bottom_rank <= 5 ~ "Bottom 5",
      TRUE ~ NA_character_
    )
  ) |>
  filter(!is.na(group)) |>
  ungroup()

ggplot(top_states_plot, aes(x = fct_reorder2(state, wave, death_rate_per_100k),
                            y = death_rate_per_100k)) +
  geom_segment(aes(xend = state, y = 0, yend = death_rate_per_100k), color = "gray") +
  geom_point(aes(color = group), size = 3) +
  facet_wrap(~ wave, scales = "free_y") +
  coord_flip() +
  scale_color_manual(values = c("Top 5" = "steelblue", "Bottom 5" = "firebrick")) +
  labs(
    title = "Top and Bottom 5 States by COVID-19 Death Rates (per 100k)",
    x = "State (ordered by mortality)",
    y = "Death Rate per 100,000",
    color = "Ranking"
  ) +
  theme_minimal()
```



```
##3.  Table for best and worst per wave
top_table <- state_wave_mortality |>
  group_by(wave) |>
  summarise(
    highest = state[which.max(death_rate_per_100k)],
    highest_rate = max(death_rate_per_100k, na.rm = TRUE),
    lowest = state[which(death_rate_per_100k == min(death_rate_per_100k[death_rate_per_100k
    lowest_rate = min(death_rate_per_100k[death_rate_per_100k > 0], na.rm = TRUE)
  )
top_table
```

```
# A tibble: 4 x 5
  wave              highest highest_rate lowest lowest_rate
  <chr>             <chr>          <dbl> <chr>        <dbl>
1 Initial Outbreak NY             269.   VT           14.7
2 Omicron Peak     WV              94.4 HI           20.0
3 Pre-Omicron Wave WV             117.   DC           12.1
4 Rebound Period   WV              79.5 WY            4.12
```

```
# Merge vax, and obesity data
model_data <- merged_data |>
  mutate(
    death_rate = deaths / population * 100000  # death rate: Per 100,000 people
  )

df1 <- model_data |>
  left_join(vax_final, by = c("state", "year", "week"))
obesity_clean <- obesity |>
  rename(state = State_Abbrev, year = Year)
df2 <- df1 |>
  left_join(obesity_clean, by = c("state", "year"))

# Fit model
model_final <- df2 |>
  filter(!is.na(death_rate), !is.na(series_complete), !is.na(ObesityRate))

model <- lm(death_rate ~ series_complete + ObesityRate, data = model_final)
summary(model)
```

```

Call:
lm(formula = death_rate ~ series_complete + ObesityRate, data = model_final)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4070 -1.4145 -0.9438  0.6006 11.7883

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.449e-01  2.700e-01  -0.537    0.592
series_complete -1.777e-07  2.970e-07  -0.598    0.550
ObesityRate      6.647e-02  7.861e-03   8.456   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.23 on 5358 degrees of freedom
Multiple R-squared:  0.01378,   Adjusted R-squared:  0.01341
F-statistic: 37.43 on 2 and 5358 DF,  p-value: < 2.2e-16
```

COVID-19-attributed death rates varied substantially across both time and geography. Nationally, death rates were highest during the initial outbreak and the winter surge of 2020–2021, followed by a decline during the Pre-Omicron period. A secondary rise occurred during the Omicron peak, although the corresponding death rate increase was less pronounced relative to the spike in cases.

The above boxplot presents the distribution of weekly COVID-19 death rates by pandemic wave. The boxplots highlight a wide range of mortality experiences across states, with notable outliers during the early stages of the pandemic. States such as New York and New Jersey reported exceptionally high death rates during the initial outbreak, while several others maintained relatively low rates across all waves.

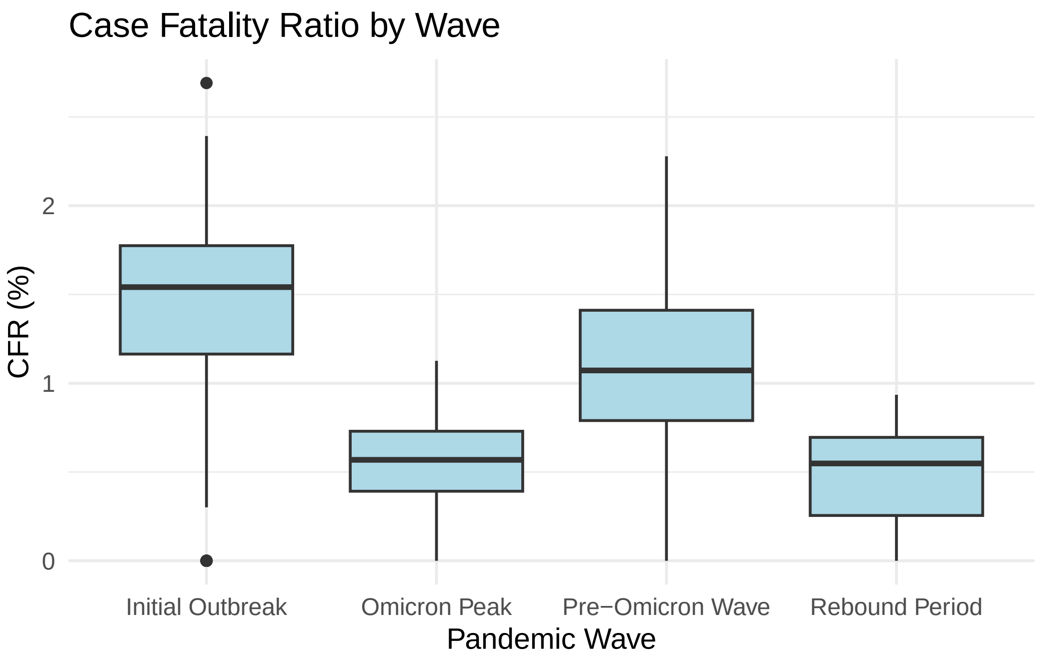
A ranking analysis of state-level death rates is shown in the above dot plot, identifying the top and bottom five states by mortality within each wave. For example, West Virginia consistently appeared among the highest-mortality states during the Omicron and rebound periods, whereas states like Vermont and Hawaii remained among the lowest throughout.

The above table provides a summary of the states with the highest and lowest COVID-19 death rates per 100,000 population for each wave. The variation in death burden across regions reflects considerable heterogeneity in exposure, health infrastructure, population vulnerability, and possibly reporting practices.

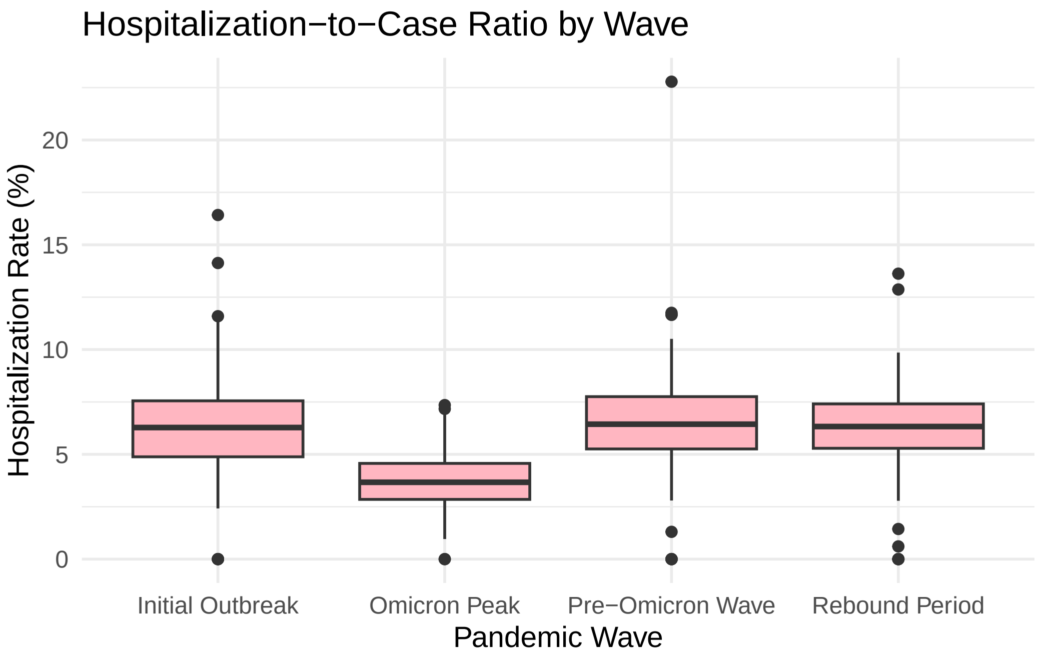
3. Virulence Trends

```r
# merge cases, death, and hosp
merged_data_hosp <- merged_data |>
  left_join(hosp_final, by = c("state", "year", "week")) |>
  mutate(date = as.Date(date)) |>
  mutate(wave = case_when(
    date >= ymd("2020-01-01") & date <= ymd("2021-06-30") ~ "Initial Outbreak",
    date >= ymd("2021-07-01") & date <= ymd("2021-11-30") ~ "Pre-Omicron Wave",
    date >= ymd("2021-12-01") & date <= ymd("2022-03-31") ~ "Omicron Peak",
    date >= ymd("2022-04-01") & date <= ymd("2023-03-31") ~ "Rebound Period",
    TRUE ~ NA_character_
  )) |>
  filter(!is.na(wave))
```

```
# 1. Compute CFR and hospitalization ratio
virulence_df <- merged_data_hosp |>
  group_by(state, wave) |>
  summarise(
    total_cases = sum(cases, na.rm = TRUE),
    total_deaths = sum(deaths, na.rm = TRUE),
    total_hosp = sum(hosp, na.rm = TRUE),
    .groups = "drop"
  ) |>
  filter(total_cases > 0) |>
  mutate(
    cfr = total_deaths / total_cases * 100,
    hosp_ratio = total_hosp / total_cases * 100
  )

# 2. CFR boxplot
ggplot(virulence_df, aes(x = wave, y = cfr)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Case Fatality Ratio by Wave", y = "CFR (%)", x = "Pandemic Wave") +
  theme_minimal()
```



```r
# 3. Hospitalization ratio boxplot
ggplot(virulence_df, aes(x = wave, y = hosp_ratio)) +
  geom_boxplot(fill = "lightpink") +
  labs(title = "Hospitalization-to-Case Ratio by Wave", y = "Hospitalization Rate (%)", x = "
  theme_minimal()
```



```r
# 4. calculate average metrics table
virulence_summary <- virulence_df |>
  group_by(wave) |>
  summarise(
    avg_cfr = mean(cfr, na.rm = TRUE),
    avg_hosp = mean(hosp_ratio, na.rm = TRUE)
  )

print(virulence_summary)
```

```
# A tibble: 4 x 3
  wave             avg_cfr avg_hosp
  <chr>              <dbl>    <dbl>
1 Initial Outbreak   1.38      6.38
2 Omicron Peak       0.546     3.66
3 Pre-Omicron Wave   1.03      6.49
```

```
4 Rebound Period     0.486     6.11
```

Indicators of viral virulence, including the case fatality ratio (CFR) and hospitalization-to-case ratio, varied across pandemic waves. These metrics provide insight into the clinical severity of infection during each major phase.

The first box plot displays the distribution of weekly case fatality ratios by wave. The CFR was highest during the initial outbreak, with a median ratio exceeding 1%. This value declined in subsequent waves, reaching its lowest point during the Omicron peak. The rebound period maintained low fatality ratios, consistent with broader access to vaccines and therapeutics.

Hospitalization-to-case ratios followed a similar trajectory. As shown in the second box plot, hospitalization rates were elevated during the initial and pre-Omicron periods, but decreased sharply during the Omicron peak and remained relatively low through the rebound phase. These trends reflect shifts in the clinical profile of COVID-19 infections over time.

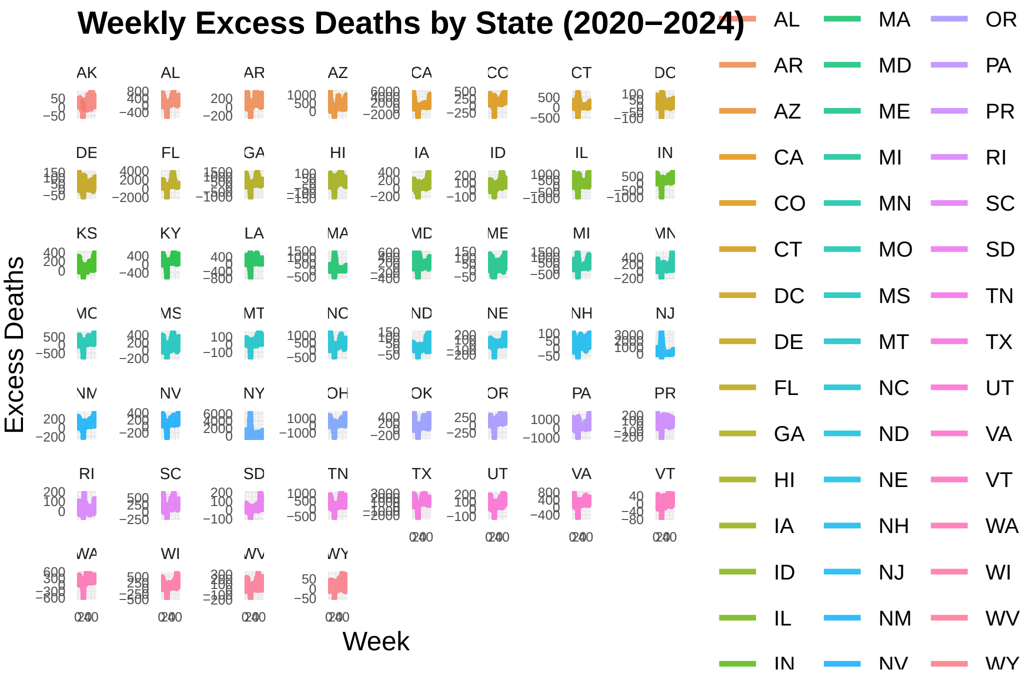
The following table summarizes the average case fatality and hospitalization ratios by wave. The data confirm substantial declines in both metrics after the initial surge, suggesting evolving patterns in disease presentation and health system response over the course of the pandemic.

4. Excess Mortality Estimation

```r
## deaths data 2015-2019
deaths_1519 <- deaths_1519 |>
  select(
    state = `Jurisdiction.of.Occurrence`,
    year = `MMWR.Year`,
    week = `MMWR.Week`,
    all_cause = `All..Cause`
  ) |>
  mutate(
    year = as.integer(year),
    week = as.integer(week),
    all_cause = as.integer(all_cause)
  ) |>
  filter(year %in% 2015:2019)

deaths_1519_final <- clean_state(deaths_1519, "state")


## excess mortality
# 1. calculated excess mortality
```

```r
library(tidyverse)
library(lubridate)

baseline <- deaths_1519_final |>
  group_by(state, week) |>
  summarise(expected_deaths = mean(all_cause, na.rm = TRUE), .groups = "drop")

deaths_excess <- deaths_final |>
  mutate(total = as.numeric(total)) |>
  left_join(baseline, by = c("state", "week")) |>
  mutate(
    excess_deaths = total - expected_deaths
  ) |>
  rename(total_deaths = total) |>
  select(state, year, week, excess_deaths, expected_deaths, total_deaths)

merged_excess <- deaths_excess |>
  left_join(deaths_final |>
              mutate(deaths = as.numeric(deaths)) |>
              select(state, year, week, covid_deaths = deaths) |>
              distinct(state, year, week, .keep_all = TRUE),
            by = c("state", "year", "week"))



# 2. visualization
ggplot(merged_excess, aes(x = week, y = excess_deaths, color = state)) +
  geom_line(alpha = 0.8, linewidth = 1) +
  facet_wrap(~ state, scales = "free_y") +
  labs(
    title = "Weekly Excess Deaths by State (2020-2024)",
    x = "Week",
    y = "Excess Deaths",
    color = "State"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    strip.text = element_text(size = 6),
    axis.text.x = element_text(size = 5),
    axis.text.y = element_text(size = 5),
    plot.title = element_text(size = 12, face = "bold"),
    strip.background = element_blank(),
    panel.spacing = unit(0.4, "lines")
  )
```

```
Warning: Removed 3816 rows containing missing values or values outside the scale range
(`geom_line()`).
```



```
# 3. fit

# Fit a model excess_deaths ~ covid_deaths for each state
lm_results <- merged_excess |>
  group_by(state) |>
  filter(!is.na(covid_deaths), !is.na(excess_deaths)) |>
  do(tidy(lm(excess_deaths ~ covid_deaths, data = .))) |>
  filter(term == "covid_deaths") |>
  rename(slope = estimate)
lm_results
```

```
# A tibble: 52 x 6
# Groups:   state [52]
   state term          slope std.error statistic   p.value
   <chr> <chr>         <dbl>     <dbl>     <dbl>     <dbl>
```

```
 1 AK    covid_deaths 1.33     0.0995      13.4 5.39e- 24
 2 AL    covid_deaths 1.12     0.0283      39.6 1.00e-106
 3 AR    covid_deaths 0.989    0.0355      27.9 1.29e- 72
 4 AZ    covid_deaths 1.03     0.0283      36.3 1.29e-103
 5 CA    covid_deaths 1.05     0.0297      35.5 1.96e-103
 6 CO    covid_deaths 0.956    0.0363      26.4 2.91e- 72
 7 CT    covid_deaths 1.02     0.0359      28.4 4.14e- 72
 8 DC    covid_deaths 1.22     0.0825      14.8 1.93e- 28
 9 DE    covid_deaths 1.05     0.0771      13.7 1.54e- 26
10 FL    covid_deaths 1.11     0.0227      49.1 5.69e-137
# i 42 more rows
```

```r
# R squared
rsq_results <- merged_excess |>
  group_by(state) |>
  filter(!is.na(covid_deaths), !is.na(excess_deaths)) |>
  do(glance(lm(excess_deaths ~ covid_deaths, data = .))) |>
  select(state, r.squared) |>
  arrange(desc(r.squared))
rsq_results
```

```
# A tibble: 52 x 2
# Groups:   state [52]
   state r.squared
   <chr>     <dbl>
 1 NJ        0.939
 2 FL        0.899
 3 TX        0.893
 4 MA        0.869
 5 AL        0.868
 6 PA        0.868
 7 MI        0.857
 8 OH        0.855
 9 MS        0.853
10 GA        0.849
# i 42 more rows
```

Excess deaths were calculated weekly as observed minus expected all-cause deaths, using 2015–2019 averages as baseline. This captures both direct COVID-19 fatalities and indirect pandemic effects. As a result, the above figure shows state-level excess death trends, with sharp peaks during the initial outbreak and Omicron wave. The rebound period showed more localized mortality surges.

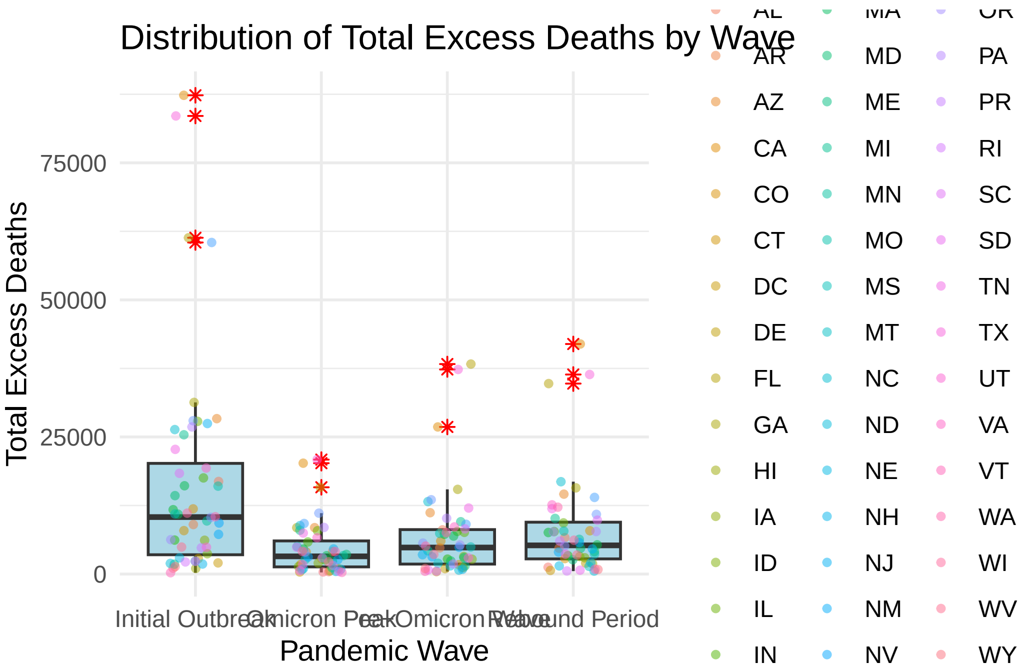
To assess attribution, we regressed excess deaths on reported COVID-19 deaths for each state. Most states showed strong alignment (β_1 1.0). In terms of model fit (R^2), states like New Jersey, Florida and Texas had high R^2 (>0.85), while others, such as New York (0.35) and Puerto Rico (0.27), had lower values, indicating greater unexplained excess mortality.

5. Excess Death Rates

```
# 1.
merged_excess <- merged_excess |>
  mutate(
    date = as.Date(paste(year, week, 1, sep = "-"), format = "%Y-%U-%u"),
    wave = case_when(
      date >= ymd("2020-01-01") & date <= ymd("2021-06-30") ~ "Initial Outbreak",
      date >= ymd("2021-07-01") & date <= ymd("2021-11-30") ~ "Pre-Omicron Wave",
      date >= ymd("2021-12-01") & date <= ymd("2022-03-31") ~ "Omicron Peak",
      date >= ymd("2022-04-01") & date <= ymd("2023-03-31") ~ "Rebound Period",
      TRUE ~ NA_character_
    )
  ) |>
  filter(!is.na(wave))
```

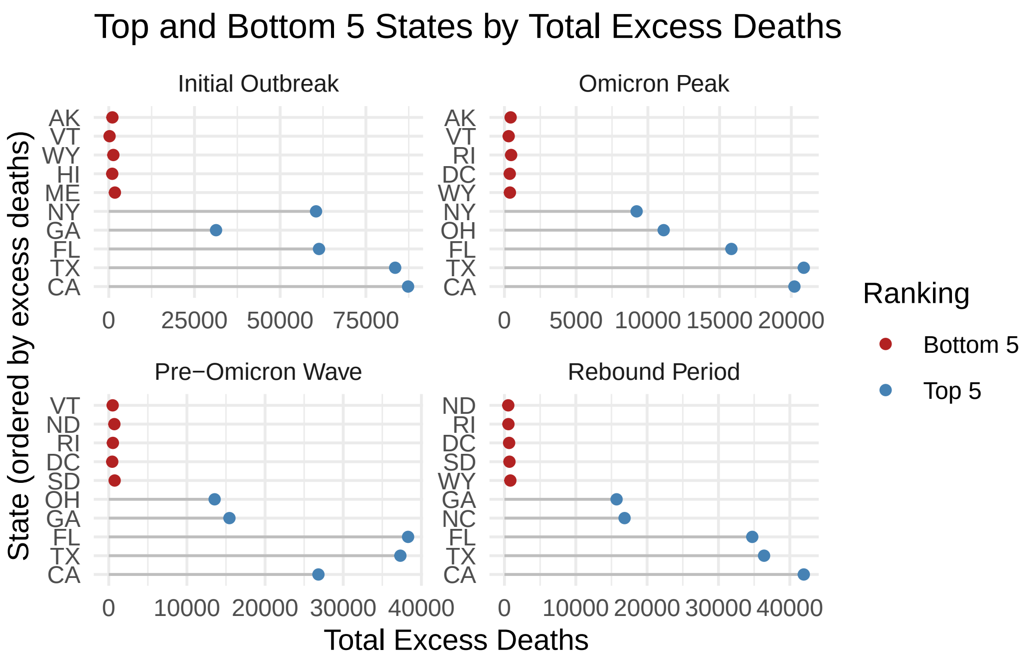
```
Warning: There were 53 warnings in `mutate()`.
The first warning was:
i In argument: `date = as.Date(paste(year, week, 1, sep = "-"), format =
  "%Y-%U-%u")`.
Caused by warning in `strptime()`:
! (0-based) yday 367 in year 2021 is invalid
i Run `dplyr::last_dplyr_warnings()` to see the 52 remaining warnings.
```

```
# 2. The number of excess deaths per wave in each state
state_wave_excess <- merged_excess |>
  group_by(state, wave) |>
  summarise(
    total_excess_deaths = sum(excess_deaths, na.rm = TRUE),
    .groups = "drop"
  )

# 3. Boxplot
ggplot(state_wave_excess, aes(x = wave, y = total_excess_deaths)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red", outlier.shape = 8) +
  geom_jitter(aes(color = state), width = 0.2, alpha = 0.5, size = 1) +
  labs(
    title = "Distribution of Total Excess Deaths by Wave",
    x = "Pandemic Wave",
    y = "Total Excess Deaths"
  ) +
  theme_minimal()
```



```
# 4. Top/Bottom 5 States Dot Plot
top_states_excess <- state_wave_excess |>
  group_by(wave) |>
  mutate(
    top_rank = min_rank(desc(total_excess_deaths)),
    bottom_rank = min_rank(total_excess_deaths),
    group = case_when(
      top_rank <= 5 ~ "Top 5",
      bottom_rank <= 5 ~ "Bottom 5",
      TRUE ~ NA_character_
    )
  ) |>
  filter(!is.na(group)) |>
  ungroup()
```

```r
ggplot(top_states_excess, aes(x = fct_reorder2(state, wave, total_excess_deaths),
                              y = total_excess_deaths)) +
  geom_segment(aes(xend = state, y = 0, yend = total_excess_deaths), color = "gray") +
  geom_point(aes(color = group), size = 1.5) +
  facet_wrap(~ wave, scales = "free") +
  coord_flip() +
  scale_color_manual(values = c("Top 5" = "steelblue", "Bottom 5" = "firebrick")) +
  labs(
    title = "Top and Bottom 5 States by Total Excess Deaths",
    x = "State (ordered by excess deaths)",
    y = "Total Excess Deaths",
    color = "Ranking"
  ) +
  theme_minimal()
```



```r
# 5. Table
top_excess_table <- state_wave_excess |>
  group_by(wave) |>
  summarise(
    highest = state[which.max(total_excess_deaths)],
    highest_deaths = max(total_excess_deaths, na.rm = TRUE),
    lowest = state[which.min(total_excess_deaths)],
    lowest_deaths = min(total_excess_deaths, na.rm = TRUE)
  )
top_excess_table
```

```
# A tibble: 4 x 5
  wave              highest highest_deaths lowest lowest_deaths
  <chr>             <chr>            <dbl> <chr>          <dbl>
1 Initial Outbreak CA              87315. VT               238.
2 Omicron Peak     TX              20868. VT               296.
3 Pre-Omicron Wave FL              38297. DC               445.
4 Rebound Period   CA              41956. ND               534.
```

The box plot summarizes the states with the highest and lowest total excess deaths per wave, with California, Texas, and Florida consistently topped the list; Vermont and Hawaii remained among the lowest. Additionally, these rates varied widely, particularly during the initial outbreak, where some states exceeded 75000 deaths.

The following dot plot illustrates the distribution of excess death by wave, showing the greatest dispersion during the first wave and more uniformity thereafter. Figure 10 identifies the five highest and lowest states in each wave. States like California and Texas frequently ranked among the highest, while Alaska and Vermont remained among the lowest.

Finally, the last table summarizes extremes in excess death rates by wave. These differences point to persistent geographic disparities in mortality burden over time.

Discussion

Our analysis reveals substantial variation in excess mortality across U.S. states and pandemic waves. States like California, Texas, and New York consistently experienced high peaks in excess deaths, especially during the initial outbreak and rebound periods, whereas states like Vermont, Alaska, and Hawaii maintained relatively low levels. This highlights the heterogeneity of the pandemic's impact across regions, reflecting differences in policy responses, healthcare capacity, population demographics, and social behaviors.

Moreover, regression models show that COVID-19 deaths can explain most of the variation in excess deaths in many states, with R^2 values above 0.85 in states such as New Jersey and Louisiana. However, other states such as New York and Puerto Rico exhibited much lower explanatory power ($R^2 < 0.4$), suggesting that excess mortality in these regions may also result from indirect effects like overwhelmed healthcare systems or misclassification in death reporting.

Excess mortality provides a more comprehensive and less biased view of the pandemic's total burden than reported COVID-19 deaths alone. It captures deaths due to underreporting, indirect impacts (e.g., delayed care, mental health issues), and comorbidities. Thus, states with unexplained excess mortality may benefit from targeted investigation and strengthened health surveillance systems. From a policy perspective, understanding the spatiotemporal distribution of excess mortality can inform future emergency preparedness, resource allocation, and targeted interventions during future pandemics or crises.

This study has several limitations. First, death reporting may be delayed or incomplete, especially in low-population states or during crisis periods. Second, differences in testing and death certification practices across states may bias both COVID and excess mortality estimates. Lastly, population mobility and undocumented infections might have distorted denominator estimates in mortality rates.

Future work could incorporate socioeconomic factors, vaccination coverage, healthcare accessibility, and behavioral data to better explain variation in excess mortality. Longitudinal models could assess delayed effects of policy interventions, and international comparisons may help benchmark national performance. Machine learning approaches may also improve early warning systems for excess deaths.