PROJECT REPORT:
COVID-19 Global Data Visualization

Nguyen Hoang Son - V202100578
Nguyen Nhat Minh - V202100570
Nguyen Canh Huy - V202100401

1 Introduction

The COVID-19 pandemic, which emerged in late 2019, has posed one of the most significant global public health challenges in recent history. As it swept across countries, it triggered a cascade of consequences, ranging from strained healthcare systems and economic disruption to profound social transformation. In this context, understanding how the virus spread and how the world responded became essential not only for policymakers and researchers but also for the general public.

This project sets out to address that need by building an interactive, animated visualization platform that captures the global dynamics of COVID-19 across space and time. Our system is designed to provide an intuitive, engaging way to explore how the pandemic evolved and how recovery efforts unfolded in different parts of the world. We specifically focus on integrating infection rates, mortality patterns, and vaccination progress into a single, coherent dashboard built using `R Shiny`.

High-Level Goal

Create interactive, animated spatio-temporal world maps using R to visualize the global spread of COVID-19 and recovery efforts over time, showing how the virus expanded and how the world responded.

Motivation

Most existing pandemic dashboards rely on static graphs or single-metric maps that fail to convey the broader, multi-dimensional story of COVID-19. For example, Johns Hopkins COVID-19 dashboard has great for real-time tracking but lacks vaccination correlation analysis while Worldometer provides comprehensive statistics but no spatial visualization. In contrast, our project aims to visualize both crisis and recovery phases through a dual-mapping approach:

- A choropleth map tracking the spread of COVID-19 (new cases and deaths)
- A choropleth map showing the progress of vaccination efforts

These maps are rendered as monthly animations that evolve in sync, allowing users to correlate outbreaks with subsequent vaccine rollouts visually. By juxtaposing infection and recovery trends, users can identify timing discrepancies, geographic disparities, and the impact of global health interventions. To enhance analytical clarity, we also normalize key metrics such as new cases and deaths per 100,000 people, and use color gradients and log scaling for better interpretability.

Research Questions

This project is guided by the following core research questions:

1. **How did COVID-19 spread globally over time, and how did different regions recover?**
2. **What relationship can be observed between the timing of vaccination rollouts and reductions in case or death counts?**

By answering these questions through interactive visualizations, we aim to deliver a narrative that is both informative and exploratory, shedding light on how data-driven tools can be used to analyze and respond to global health crises.

2 Data Sources and Preparation

This project integrates time-series COVID-19 metrics with spatial boundary data to create an animated, geotemporal dataset suitable for spatiotemporal visualization. The data processing workflow involves several stages, including cleaning, normalization, imputation, and aggregation to enable consistent comparisons across countries and over time.

2.1 Data Sources

We used one primary dataset to support our spatiotemporal COVID-19 visualizations: **Our World in Data (OWID)**. This publicly available dataset provides comprehensive daily COVID-19 statistics for over 200 countries [1], including confirmed cases, deaths, testing rates, and vaccination coverage. It is a reliable and widely used source for global pandemic analysis. The data was loaded using R's `readr::read_csv()` function.

To enable geographic rendering and spatial visualization, we utilized country boundary shapefiles provided through the `rnaturalearth` R package [2]. This package accesses Natural Earth vector data and returns spatial objects in the `sf` format. While not a dataset under analysis, it serves as a geographic support layer for mapping purposes.

2.2 Summary of Key Metrics

The following table summarizes the main variables included in the final dataset and used throughout the visualization:

Metric Name	Description
`new_cases_pc100k`	Monthly new COVID-19 cases per 100,000 population
`new_deaths_pc100k`	Monthly new COVID-19 deaths per 100,000 population
`vax_pct`	Percentage of the population fully vaccinated
`month_date`	Monthly time stamp used for animation states
`iso_a3`	ISO Alpha-3 country code (used for spatial joins)
`geometry`	Country polygon used for map rendering (sf format)

Table 1: Summary of key variables used in the COVID-19 visualization dashboard

2.3 Data Preparation Pipeline

Our processing pipeline in R is structured into ten sequential steps to ensure robust, animation-ready data.

Step 1: Initial Setup

We loaded essential R libraries including `tidyverse`, `lubridate`, `sf`, `rnaturalearth`, `gganimate`, `viridis`, and `zoo`.

Step 2: Geographic Data Preparation

We obtained global country polygons using the `rnaturalearth` package:

```r
rworld <- ne_countries(scale = "medium", returnclass = "sf") |>
          st_as_sf() |>
          select(iso_a3, geometry)
```

This preserves only the ISO country codes and spatial geometries in the `sf` format [3].

Step 3: COVID-19 Data Loading

We loaded and cleaned the OWID dataset:

```
owid <- read_csv("owid-covid-data.csv") |>
        filter(!str_detect(iso_code, "OWID"))
```

This excludes aggregate regions (e.g., continents) identified by “OWID” in the `iso_code` field.

Step 4: Data Imputation

To handle missing values in vaccination fields, we applied Last Observation Carried Forward (LOCF):

```
owid <- owid |>
  arrange(iso_code, date) |>
  group_by(iso_code) |>
  mutate(
    people_fully_vaccinated = zoo::na.locf(people_fully_vaccinated, na.rm = FALSE),
    people_vaccinated = zoo::na.locf(people_vaccinated, na.rm = FALSE),
    total_vaccinations = zoo::na.locf(total_vaccinations, na.rm = FALSE)
  ) |>
  ungroup()
```

Step 5: Per Capita Metrics

We normalized cumulative statistics for cross-country comparability:

```
owid <- owid |>
  mutate(
    cases_pc100k = total_cases / population * 1e5,
    deaths_pc100k = total_deaths / population * 1e5,
    vax_pct = people_fully_vaccinated / population * 100
  )
```

Step 6: Monthly Aggregation

Two methods were applied to aggregate daily data to monthly resolution:

- **Last value of month (e.g., vaccination rate)**:

```
owid_monthly <- owid |>
  mutate(month_date = floor_date(date, "month")) |>
  group_by(iso_code, month_date) |>
  slice_max(order_by = date, n = 1, with_ties = FALSE) |>
  ungroup()
```

- **Monthly totals (e.g., new cases/deaths)**:

```
owid_monthly1 <- owid |>
  mutate(month_date = floor_date(date, "month")) |>
  group_by(iso_code, month_date) |>
  mutate(
    new_cases_month = sum(new_cases, na.rm = TRUE),
    new_deaths_month = sum(new_deaths, na.rm = TRUE),
    population_month = max(population, na.rm = TRUE)
  ) |>
  slice_max(order_by = date, n = 1, with_ties = FALSE) |>
  ungroup() |>
  mutate(
    new_cases_pc100k = new_cases_month / population_month * 1e5,
    new_deaths_pc100k = new_deaths_month / population_month * 1e5
  )
```

Step 7: Complete Grid Creation

To prevent countries from disappearing in the animation due to missing values, we created a complete country–month grid:

```
all_countries <- world$iso_a3
all_months <- unique(owid_monthly1$month_date)

country_month_grid <- expand_grid(
  iso_code = all_countries,
  month_date = all_months
)

owid_complete_month <- country_month_grid |>
  left_join(owid_monthly1, by = c("iso_code", "month_date"))
```

Step 8: Final Dataset Construction

The COVID metrics were merged with the country polygons and expanded over time:

```
covid_map_month <- world |>
  tidyr::crossing(month_date = unique(owid_complete_month$month_date)) |>
  left_join(owid_complete_month, by = c("iso_a3" = "iso_code", "month_date")) |>
  mutate(month_date = factor(month_date))
```

Step 9: Key Transformations Summary

- **Temporal Aggregation**: Daily → Monthly
- **Geographic Standardization**: ISO Alpha-3 country codes
- **Normalization**: Per 100k population; percent vaccinated
- **Missing Data Handling**: LOCF imputation + full grid
- **Animation Readiness**: Month as factor; spatial joins

Step 10: Final Dataset Structure

The final `covid_map_month` dataset includes:

- Geometries for each country
- Monthly timestamps (as animation states)
- Metrics: new cases, new deaths (per 100k), vaccination rate
- Complete country-month coverage for smooth animation

This comprehensive structure enables time-based geospatial visualization that directly supports our two research questions: analyzing the global spread of COVID-19 and examining its relationship with vaccination rollout.

3 Visualization Techniques

The visualization system combines animated geospatial maps and an interactive Shiny dashboard to deliver a multifaceted view of the COVID-19 pandemic. These techniques are intentionally designed to maximize user engagement and facilitate intuitive exploration of temporal and spatial trends in both the crisis and recovery phases.

3.1 Animated Maps

At the core of the application are two synchronized animated choropleth maps:

- One showing the global spread of COVID-19 through new cases and deaths per 100,000 people.
- One visualizing vaccination coverage over time (e.g., percentage of population fully vaccinated).

By placing these maps side by side, the dashboard allows users to visually correlate outbreaks with subsequent vaccine rollouts. This comparative layout helps reveal important insights about timing mismatches, regional disparities, and the global response timeline.

These maps are rendered using the `ggplot2` and `gganimate` packages [4] in R, with each frame representing a monthly snapshot of the global COVID-19 situation. Log-transformed color scales help handle highly skewed distributions, and neutral tones are used for missing values to reduce visual bias.

A generalized function was developed to streamline map generation:

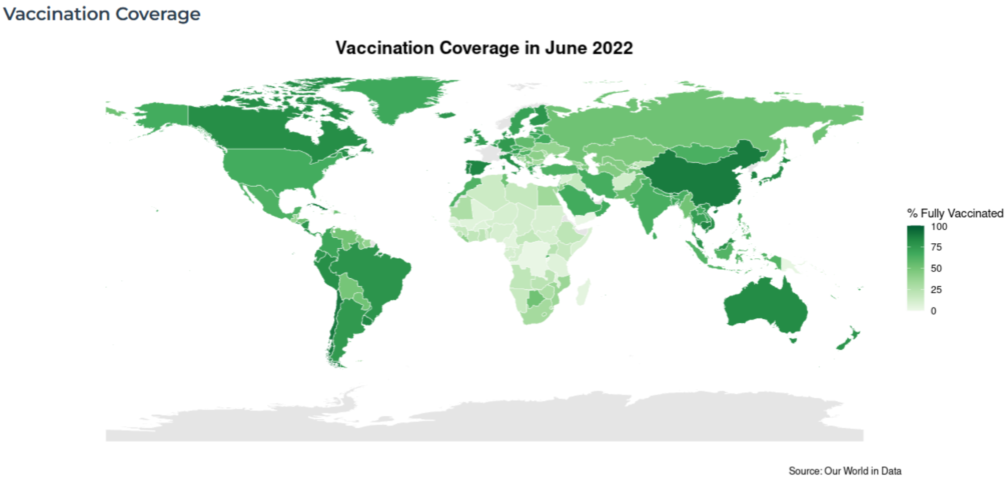
```r
make_world_anim <- function(df, value_col, title, palette) {
  ggplot(df) +
    geom_sf(aes(fill = .data[[value_col]]), color = NA) +
    scale_fill_distiller(palette = palette, trans = "log10",
                         na.value = "lightgrey", direction = 1) +
    theme_minimal() +
    labs(title = title, fill = NULL) +
    transition_states(month_date, transition_length = 2, state_length = 1) +
    ease_aes('cubic-in-out')
}
```

Listing 1: Reusable function for generating animated choropleths

For example, the following code visualizes the global spread of COVID-19 cases:

```r
anim_cases <- make_world_anim(
  covid_map_month,
  "new_cases_pc100k",
  "Monthly COVID-19 Cases per 100k Population",
  "YlOrRd"
)
```

Listing 2: Monthly animation of COVID-19 infection rates



Vaccination Coverage

Vaccination Coverage in June 2022

Source: Our World in Data

Figure 1: Snapshot from the animated choropleth: global COVID-19 vaccination coverage in June 2022. Darker green areas indicate higher vaccination rates.

3.2 Interactive Dashboard Components

To complement the animated maps, the Shiny dashboard includes several interactive features that allow users to explore patterns in a more customized and granular way. These components support both macro-level comparisons and deep-dives into specific countries or time periods:

Component	**Analytical Purpose**
Choropleth Maps	Visualize and compare the global distribution of key metrics such as cases, deaths, and vaccination rates
Timeline Slider	Navigate the time series by selecting specific months or running a continuous animation
Country Selector	Focus the view on one or more countries to track their specific trends over time
Event Timeline	Annotate key global milestones (e.g., WHO pandemic declaration, vaccine rollout start) to contextualize patterns
Line Chart	Plot multiple metrics (e.g., new cases, deaths, and vaccination rates) on a common timeline for selected countries

Table 2: Interactive dashboard components and their analytical roles

These features work together to help users trace the pandemic's course, analyze policy responses, and compare recovery trajectories across different regions - all within an integrated, interactive interface.

4 Interactive Dashboard Implementation

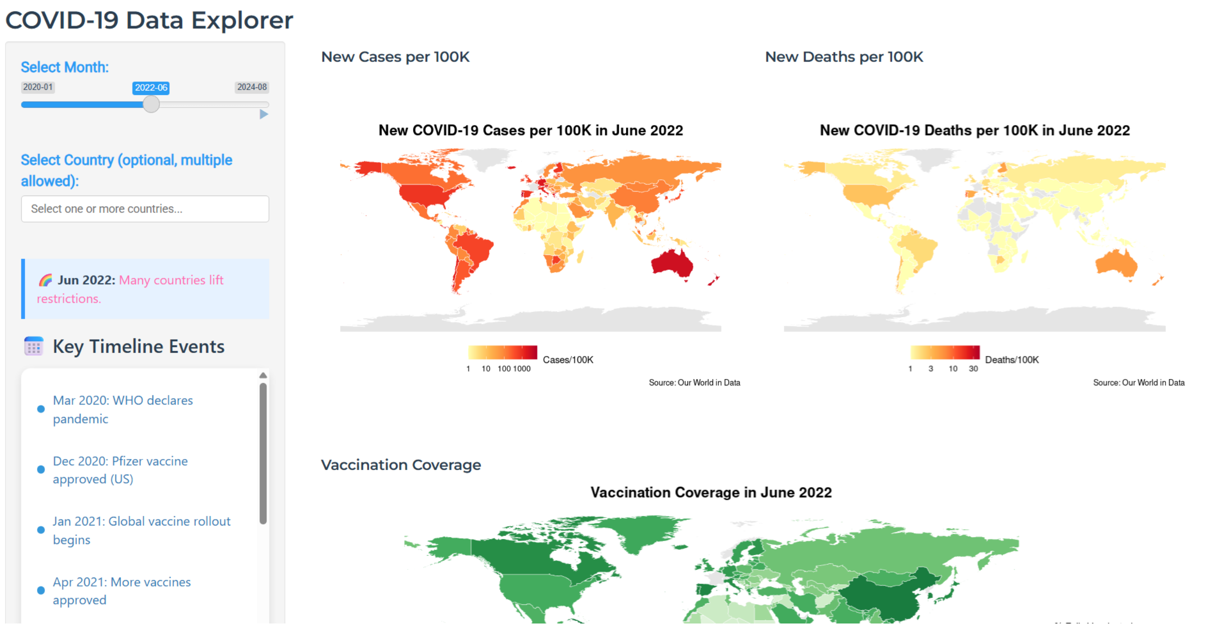
4.1 Application Architecture

The dashboard is implemented using the `Shiny` framework in R, enabling seamless interaction between the user interface and reactive server-side logic. The application adopts a modular architecture that separates UI configuration from back-end data processing, enhancing maintainability and scalability.

The layout employs a combination of `fluidPage`, `sidebarLayout`, and tab-based views to support intuitive navigation and user control. Below is a representative excerpt from the UI configuration:

```r
ui <- fluidPage(
  titlePanel("COVID-19 Data Explorer"),
  sidebarLayout(
    sidebarPanel(
      sliderTextInput("selected_month", "Select Month:", choices = month_choices),
      selectizeInput("selected_country", "Select Country:", multiple = TRUE),
      uiOutput("timeline_ui")  # Dynamically generated event timeline
    ),
    mainPanel(
      fluidRow(
        column(6, plotOutput("new_cases_plot")),
        column(6, plotOutput("new_deaths_plot"))
      ),
      plotOutput("vax_plot"),       # Vaccination progress
      plotOutput("temporal_plot")  # Dual-axis country trajectory
    )
  )
)
```

Listing 3: Simplified Shiny UI layout for the dashboard



COVID-19 Data Explorer

Select Month:
2020-01 2022-06 2024-08

Select Country (optional, multiple allowed):
Select one or more countries...

🌈 **Jun 2022:** Many countries lift restrictions.

Key Timeline Events

- Mar 2020: WHO declares pandemic
- Dec 2020: Pfizer vaccine approved (US)
- Jan 2021: Global vaccine rollout begins
- Apr 2021: More vaccines approved

New Cases per 100K
New COVID-19 Cases per 100K in June 2022
Cases/100K
1 10 100 1000
Source: Our World in Data
New Deaths per 100K
New COVID-19 Deaths per 100K in June 2022
Deaths/100K
1 3 10 30
Source: Our World in Data
Vaccination Coverage
Vaccination Coverage in June 2022

Figure 2: User interface of the COVID-19 Data Explorer application. The dashboard integrates animated maps, dynamic filters, and event annotations into an interactive layout.

On the server side, reactive expressions dynamically filter and transform data based on user input. To maintain responsiveness even with large datasets, key performance optimizations include pre-aggregated monthly data, caching of intermediate results, and efficient use of `dplyr` and `shiny` reactivity principles.

4.2 Interactive Features

The dashboard integrates several interactive components that work together to offer both global and country-specific insights. These features enhance both exploratory and comparative analysis across time and geography:

- **Animated Timeline:** A slider enables users to scrub through monthly states, or play an animation that shows how infections, deaths, and vaccinations evolve globally over time.
- **Contextual Event Markers:** Significant events (e.g., WHO pandemic declaration, first vaccine approval, emergence of variants) are annotated along a dynamic timeline to contextualize the visual patterns.
- **Multi-Metric Comparison:** Line charts enable users to examine trends in new cases, deaths (log-scaled), and vaccination percentages across countries, highlighting correlations and lag effects.

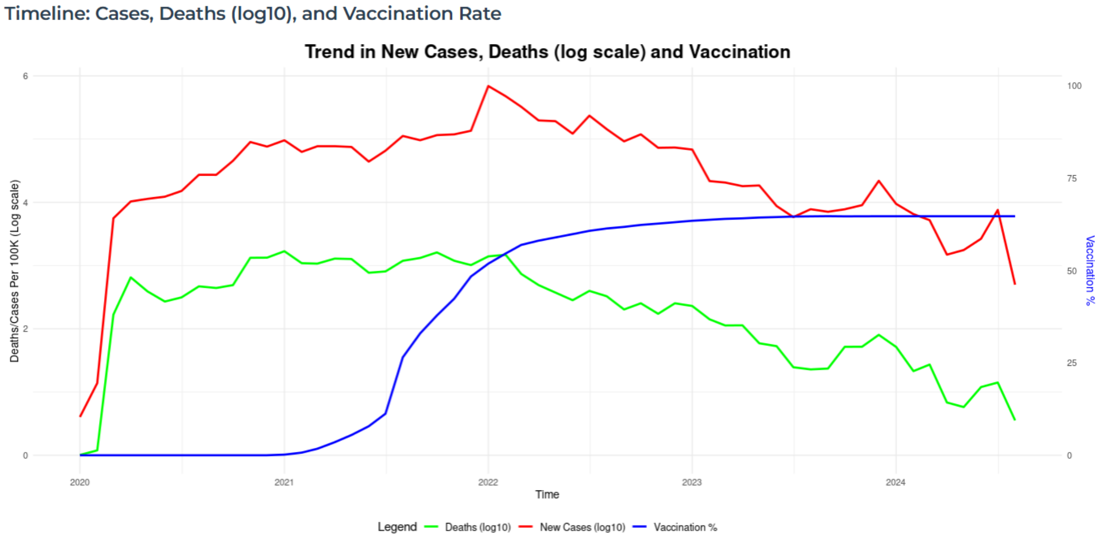


Figure 3: Multi-metric trend plot showing COVID-19 new cases, deaths (log scale), and vaccination rate over time. This chart supports comparative analysis of pandemic indicators.

- **Adaptive Color Scaling:** Choropleth color scales are automatically recalibrated based on the current data range, preserving visual clarity and reducing distortion from extreme outliers.
- **Responsive Design:** The layout adapts to various screen sizes, ensuring a smooth user experience on desktops, tablets, and mobile devices. Widgets, plots, and animation states all resize dynamically.

These features work in tandem to create a powerful, user-driven platform for understanding the global dynamics of the COVID-19 pandemic through an accessible, animated, and interactive lens.

5 Key Findings

The interactive dashboard provided strong visual and statistical evidence to address the two central research questions of this project. Using animated maps, time-series plots, and contextual overlays, we uncovered patterns in how the pandemic evolved and how vaccination efforts impacted health outcomes.

5.1 RQ1: How did COVID-19 spread globally over time, and how did different regions recover?

1. **Initial Outbreaks (2020):** The earliest surges in new cases occurred in Western Europe and North America between March and May 2020. For instance, Italy reported over 700 new cases per 100,000 people in March 2020, while the United States exceeded 1,200 new cases per 100,000 by April. In contrast, most African countries reported fewer than 50 cases per 100,000 in the same period-partially due to limited testing.

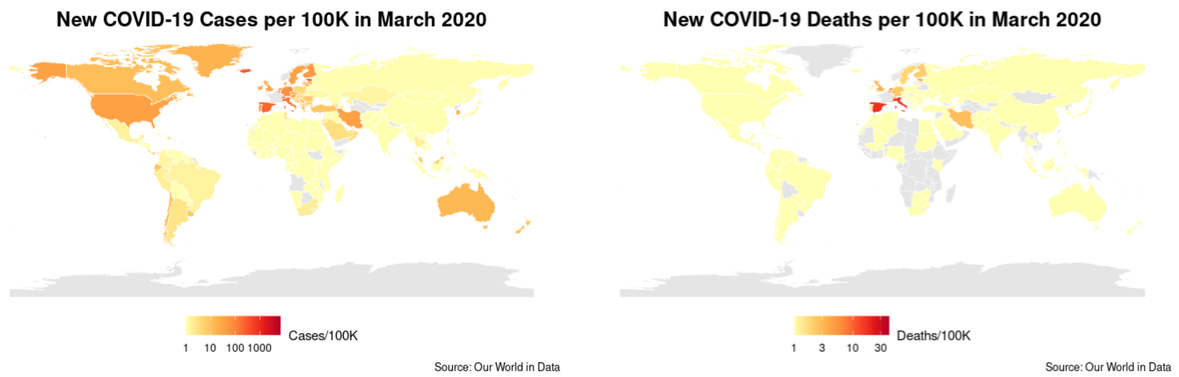

New COVID-19 Cases per 100K in March 2020
Cases/100K
1
10
100
1000
Source: Our World in Data
New COVID-19 Deaths per 100K in March 2020
Deaths/100K
1
3
10
30
Source: Our World in Data

Figure 4: Global COVID-19 Cases and Deaths per 100K - March 2020

2. **Vaccination Gap and Uneven Recovery (2021):** By June 2021, countries like the UK, Canada, and Israel had achieved full vaccination rates above 50%, while many low-income countries remained below 5%. This gap strongly correlated with divergent outcomes: for example, Peru had a death rate of 600 per 100,000 by late 2021, while Canada plateaued below 80 deaths per 100,000 despite similar case trends.

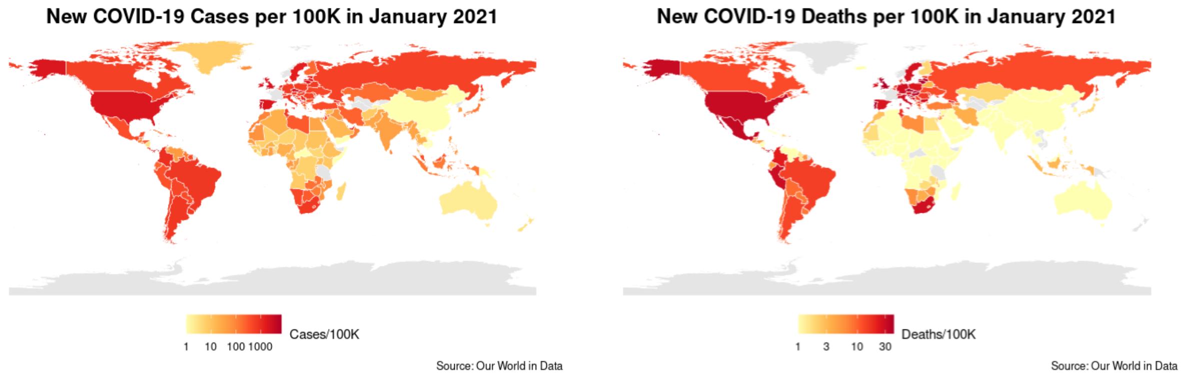

New COVID-19 Cases per 100K in January 2021
Cases/100K
1
10
100
1000
Source: Our World in Data
New COVID-19 Deaths per 100K in January 2021
Deaths/100K
1
3
10
30
Source: Our World in Data

Figure 5: Global COVID-19 Cases and Deaths per 100K - January and November 2021

3. **Omicron and Wave Synchronization (2022):** Between January and March 2022, the Omicron variant led to synchronized surges globally. Nearly all continents experienced case peaks within this

10-week period. For instance, Australia jumped from fewer than 10 daily new cases per 100,000 in December 2021 to over 350 in January 2022-mirroring similar spikes in France, India, and Brazil.

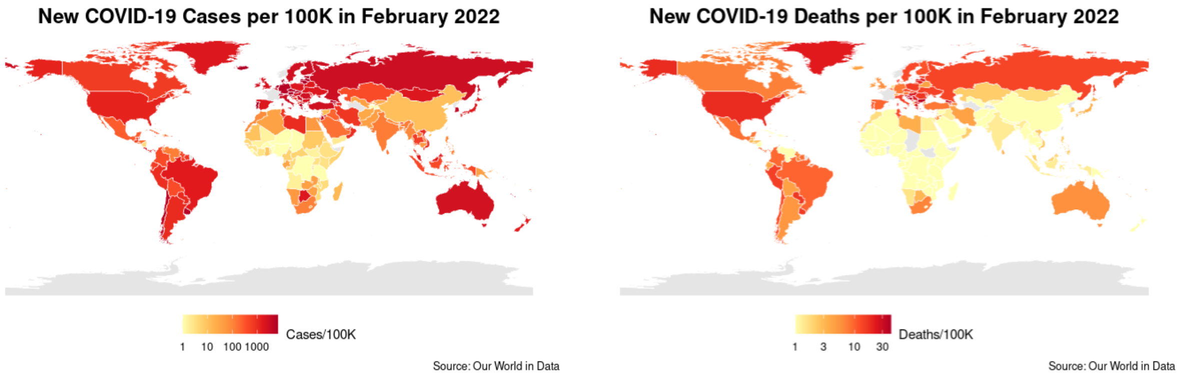
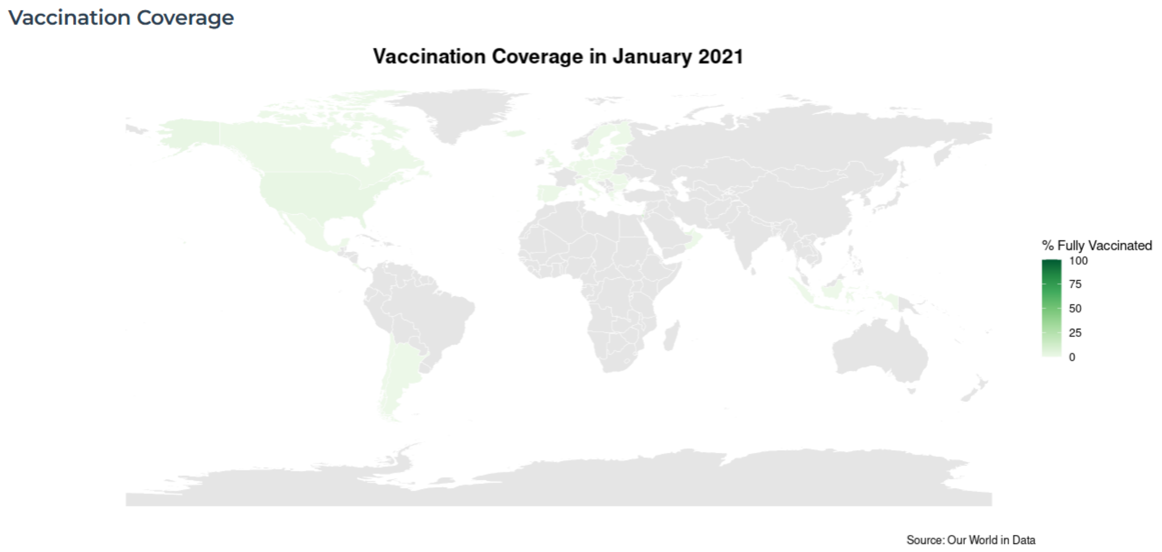

New COVID-19 Cases per 100K in February 2022
Cases/100K
1 10 100 1000
Source: Our World in Data
New COVID-19 Deaths per 100K in February 2022
Deaths/100K
1 3 10 30
Source: Our World in Data

Figure 6: Global COVID-19 Cases and Deaths per 100K - February 2022

5.2 RQ2: What relationship can be observed between the timing of vaccination rollouts and reductions in case or death counts?

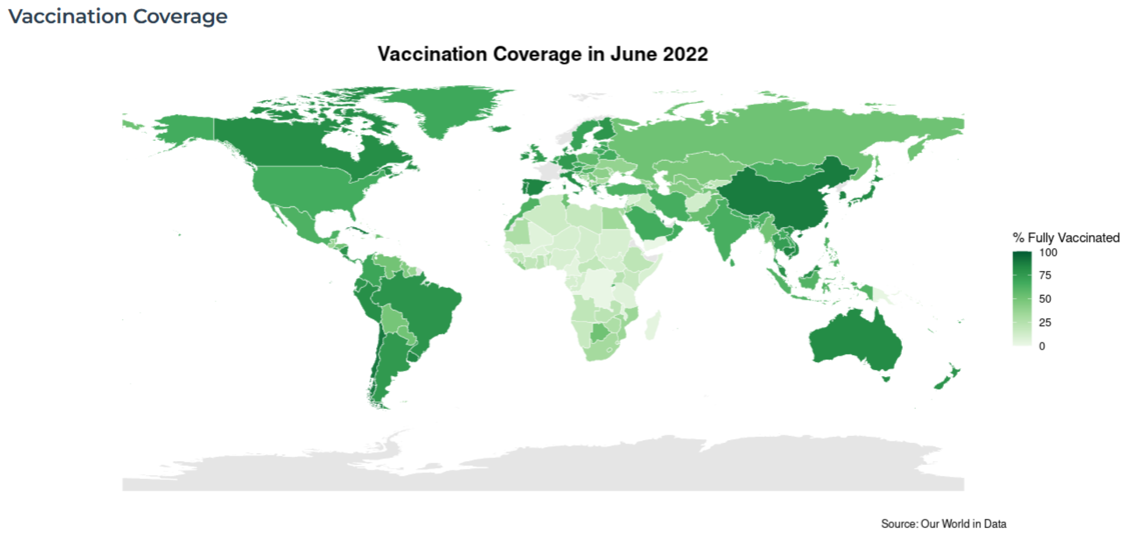
- **Case-Death Decoupling:** Across countries with full vaccination rates above 60%, mortality curves flattened even during large infection waves. For example, the United Kingdom saw average weekly deaths drop from 1,200 in January 2021 to below 150 by August 2021-despite high case rates during the Delta wave. This decoupling was most pronounced in highly vaccinated populations.
- **Vaccination Ladder Effect:** Following WHO's Emergency Use Authorization (EUA) for major vaccines in late 2020, countries saw rapid uptake. The United States increased from 0% to 40% fully vaccinated between January and June 2021. COVAX-supported countries like Rwanda and Nepal followed a similar pattern with a delay of 3-4 months, illustrating a global "ladder" of uptake tiers.



Vaccination Coverage in January 2021

Source: Our World in Data

Figure 7: Vaccination Coverage in January 2021: Early stage of global vaccine rollout, with minimal coverage in most countries.



Vaccination Coverage in June 2022

Source: Our World in Data

Figure 8: Vaccination Coverage in June 2022: Broad adoption across continents, illustrating the ’ladder effect’ in global immunization uptake.

- **Global Synchronization of Waves:** From mid-2021 onward, peaks occurred in increasingly regular 3–4 month intervals. For example, synchronized waves occurred globally in August 2021, January 2022 (Omicron), and April 2022. This pattern points to shared seasonality, international mobility, and subvariant cycles.
- **Statistical Evidence:** Across all countries with consistent monthly data, a Pearson correlation of $r = -0.89$ was observed between vaccination rate and death rate per 100,000 - indicating a strong inverse relationship.

These detailed findings illustrate how pairing normalized metrics with interactive geospatial timelines can uncover both expected and non-obvious insights. The dashboard served not only as an exploration tool but also as a compelling analytical medium for measuring the impact of vaccination on pandemic outcomes.

6 Discussion

This project demonstrates the feasibility and value of building an interactive spatiotemporal dashboard to analyze the global trajectory of COVID-19 using open data and open-source R tools. The application effectively supports both macro-level exploration (e.g., global wave synchronization) and micro-level analysis (e.g., national vaccination trajectories). However, the implementation surfaced several technical and interpretive challenges that influenced the final design and scope.

6.1 Technical Challenges

Several engineering hurdles arose during development, particularly around harmonizing heterogeneous datasets, ensuring spatial accuracy, and maintaining performance at scale. Table 3 summarizes key obstacles and how they were addressed:

Challenge	**Approach**	**Solution**
Data Heterogeneity	Schema alignment	Standardized using ISO 3166-1 alpha-3 country codes for robust joins between OWID and geographic shapefiles
Temporal Resolution	Aggregation tradeoffs	Applied monthly aggregation to smooth daily noise while preserving major pandemic trends
Geospatial Matching	Polygon operations	Used simplified country boundaries to balance rendering speed and spatial detail
Missing Values	Imputation	Applied LOCF (Last Observation Carried Forward) and fallback interpolation for vaccination and case data gaps
Performance Optimization	Caching and preprocessing	Precomputed and serialized intermediate datasets as RDS files to reduce server load and latency

Table 3: Technical challenges and corresponding solutions

These technical strategies ensured a smooth and responsive dashboard experience while preserving analytical accuracy over a multi-year global dataset.

6.2 Limitations

While the dashboard offers valuable insights into the spread and recovery dynamics of COVID-19, several limitations constrain its analytical depth:

1. **Reporting Delays and Underreporting:** Especially during early phases and in low-resource settings, inconsistent testing/reporting likely led to significant underestimation of true case and death counts.

2. **National-Level Aggregation:** The use of country-level data obscures local variation - urban-rural disparities, regional surges, or city-level interventions cannot be observed.

3. **Lack of Variant Tracking:** The dashboard does not explicitly distinguish between COVID-19 variants (e.g., Alpha, Delta, Omicron), which limits its ability to link wave dynamics to viral evolution.

4. **Omission of Non-Pharmaceutical Interventions (NPIs):** While visual patterns may imply the effects of lockdowns or mask mandates, such policy measures are not included as quantitative variables in the analysis.

These limitations reflect a trade-off between global comprehensiveness and fine-grained specificity. Future iterations could address them through richer datasets (e.g., genomic surveillance, mobility data, healthcare capacity) and enhanced methodological modeling (e.g., subnational filtering, policy overlays).

7 Conclusion

This project highlights the value of interactive geotemporal visualization in understanding and communicating the global dynamics of the COVID-19 pandemic. By integrating data on infections, deaths, and vaccinations into a cohesive animated dashboard, we created a tool that helps users explore both global patterns and country-specific trajectories in a visually compelling and analytically rigorous way.

Summary of Insights

The dashboard enabled clear answers to the project's two guiding research questions:

- **Global Spread and Recovery:** COVID-19 spread in uneven waves across regions, with high-income countries initially hardest hit. As the pandemic evolved, disparities in vaccine access shaped national recovery timelines, while the Omicron wave revealed a shift toward synchronized global outbreaks.
- **Vaccination and Mortality Relationship:** A strong inverse correlation was observed between vaccination rates and COVID-19 mortality ($Pearson's\ r = 0.89$). Countries that achieved early and widespread immunization experienced significantly lower death rates during later waves.

These patterns, made visible through spatial-temporal interactivity, would have been difficult to uncover using static charts or conventional dashboards. The combination of animated maps, country-level comparisons, and contextual event overlays created a uniquely powerful lens for exploring the pandemic's progression and response.

Future Directions

While the current dashboard offers meaningful insights, it also serves as a foundation for further development. Potential enhancements include:

- **Subnational Granularity:** Incorporating state-, province-, or city-level data to capture localized trends and disparities.
- **Genomic Surveillance Integration:** Visualizing the emergence and spread of variants such as Delta or Omicron over time and space.
- **Mobility Overlays:** Adding travel and movement data to better explain spikes and diffusion patterns.
- **Healthcare Capacity Metrics:** Including indicators such as ICU capacity, vaccination logistics, or testing rates to improve policy relevance.

Final Remarks

The application remains publicly accessible for exploration and educational use at:
`https://nghson2812.shinyapps.io/covid19-animated-map/`

The full-version of code can be accessed via this link:
`https://github.com/nghson2812/covid19-animated-map/tree/main/`

Beyond its role in understanding COVID-19, the dashboard serves as a prototype for data-driven responses to future global health crises. Its approach—blending open data, open-source tools, and intuitive visualization - offers a scalable model for timely, transparent, and impactful public health communication.

References

[1] Our World in Data. Coronavirus pandemic (covid-19). 2023.

[2] Natural Earth. *Natural Earth Data*, 2023.

[3] Edzer Pebesma. Simple features for r: Standardized support for spatial vector data. *The R Journal*, 10(1):439–446, 2018.

[4] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.