DATA607 PROJECT

Gaurav Dnyanesh Mahajan

2025-11-16

Did India's COVID-19 Vaccination Rollout Causally Reduce Mortality?

A Bayesian Causal Impact Analysis Using COVID-19 Data Hub

1. Introduction

India experienced one of the world's most significant COVID-19 health burdens, characterized by multiple waves and substantial heterogeneity. A key national policy question is whether the vaccination rollout that began in mid-January 2021 produced a measurable causal reduction in COVID-19 mortality. This study constructs a complete, reproducible data science pipeline, beginning with the extraction of messy state-level COVID-19 data and ending with a Bayesian structural time-series causal impact estimate of the vaccination rollout on weekly deaths per 100,000 people in India. The report includes data cleaning, exploratory data analysis, weekly feature construction, causal modelling, uncertainty quantification, diagnostics, and interpretation.

2. Data Sources and Extraction

Data were obtained from the COVID-19 Data Hub at level -2 resolution (states and union territories) spanning March 2020 to December 2022. The raw datslevel-2et comprises 20,998 rows and 47 variables, including cumuative confirmed cases, cumulative deaths, testing counts, vaccination coverage, stringency index, and population estimates.

A missingness audit demonstrated a structural absence of vaccination indicators before 2021, intermittent gaps in stringency data, and some downward correction in cumulative case and death counts. These issues required systematic cleaning before analysis.

3. Data Cleaning and Feature Construction

Daily counts of cases, deaths, and tests were computed from cumulative data. Negative values caused by retrospective corrections were set to zero. All epidemiological indicators were scale per 100,000 population. Vaccination coverage and stringency were forward-filled to ensure continuity, consistent with their gradual evolution over time.

Weekly aggregation was performed to reduce daily noise and reporting variability. State-level weekly indicators were combined into a national time series using population-weighted means. Weekly deaths per 100,000 serve as the outcome variable, while weekly cases per 100,000 and the mean stringency index serve as covariates.

4. Exploratory Data Analysis

Nationally aggregated weekly deaths per 100,000 show clear peaks corresponding to major epidemic waves. The maximum occurs during the Delta wave of 2021, after which a substantial decline begins. Vaccination coverage rises steadily beginning mid-January 2021. The temporal patterns are visible in the national weekly mortality plot.

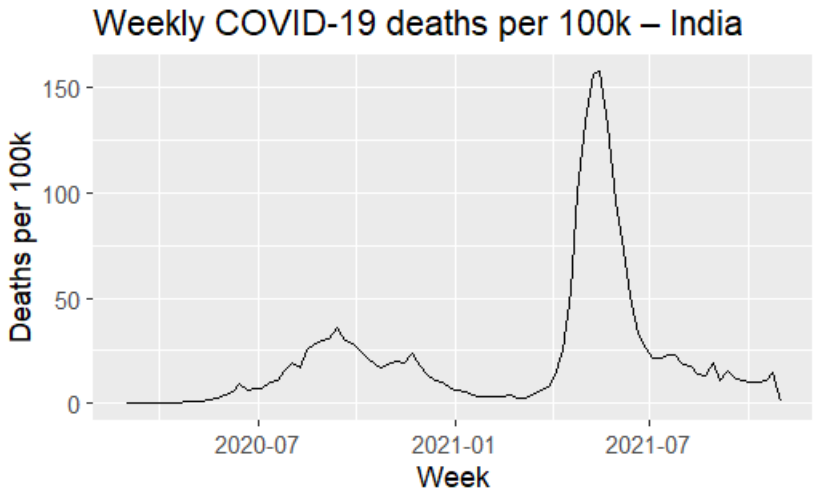


Figure 1: Weekly COVID-19 deaths per 100,000 people in India, 2020–2022.

State-level box-plots reveal substantial differences across states, with some consistently experiencing higher weekly death rates. This heterogeneity motivates national-level modeling but emphasizes the importance of covariates that capture infection pressure and policy intensity.

A multivariate scatterplot of weekly cases versus weekly deaths per 100,000 illustrates how the relationship between infection pressure and mortality changed after vaccination began. In the pre-intervention period, the points follow a compact, upward-sloping pattern, showing that increases in case rates were closely matched by increases in deaths. After the vaccination rollout, the relationship shifts: even during weeks with very high case rates, deaths rise more slowly, and the smoothing curve becomes flatter and lower. This decoupling suggests that vaccination weakened the link between infections and severe outcomes, providing visual support for the patterns later quantified in the causal model.

These exploratory patterns reinforce the need for a causal approach capable of distinguishing changes driven by vaccination from those driven by shifting case dynamics or policy interventions. These exploratory patterns reinforce the need for a causal approach capable of distinguishing changes driven by vaccination from those driven by shifting case dynamics or policy interventions.

5. Causal Modeling Framework

The intervention date is defined as 16 January 2021, corresponding to India's nationwide vaccination campaign. The BSTS model decomposes weekly mortality into latent trend and regression components using pre-intervention data to learn relationships among deaths, cases, and policy stringency. The CausalImpact package then projects a counterfactual mortality trajectory for the post-intervention period.

The model specification includes weekly deaths per 100,000 as the response and weekly cases per 100,000 and mean stringency index as covariates. This approach allows the counterfactual to reflect plausible epidemiological dynamics in the absence of vaccination.

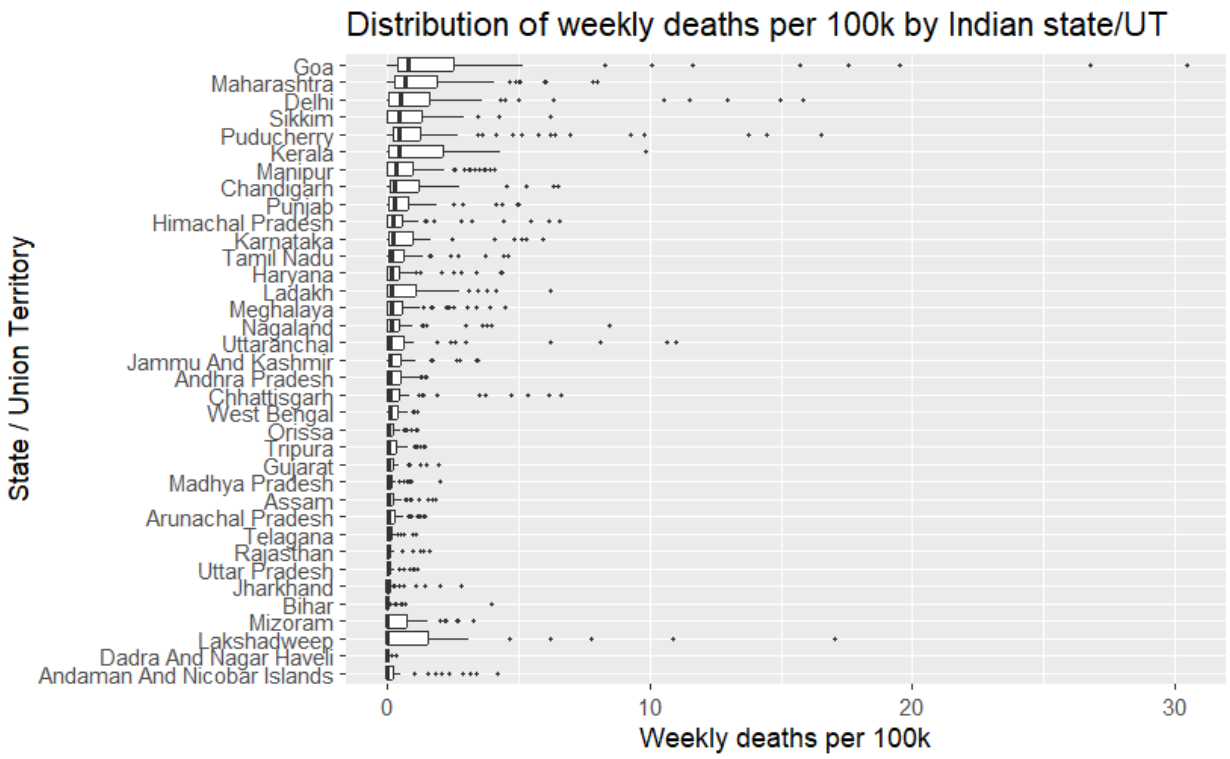


Figure 2: Distribution of weekly deaths per 100,000 across Indian states and union territories

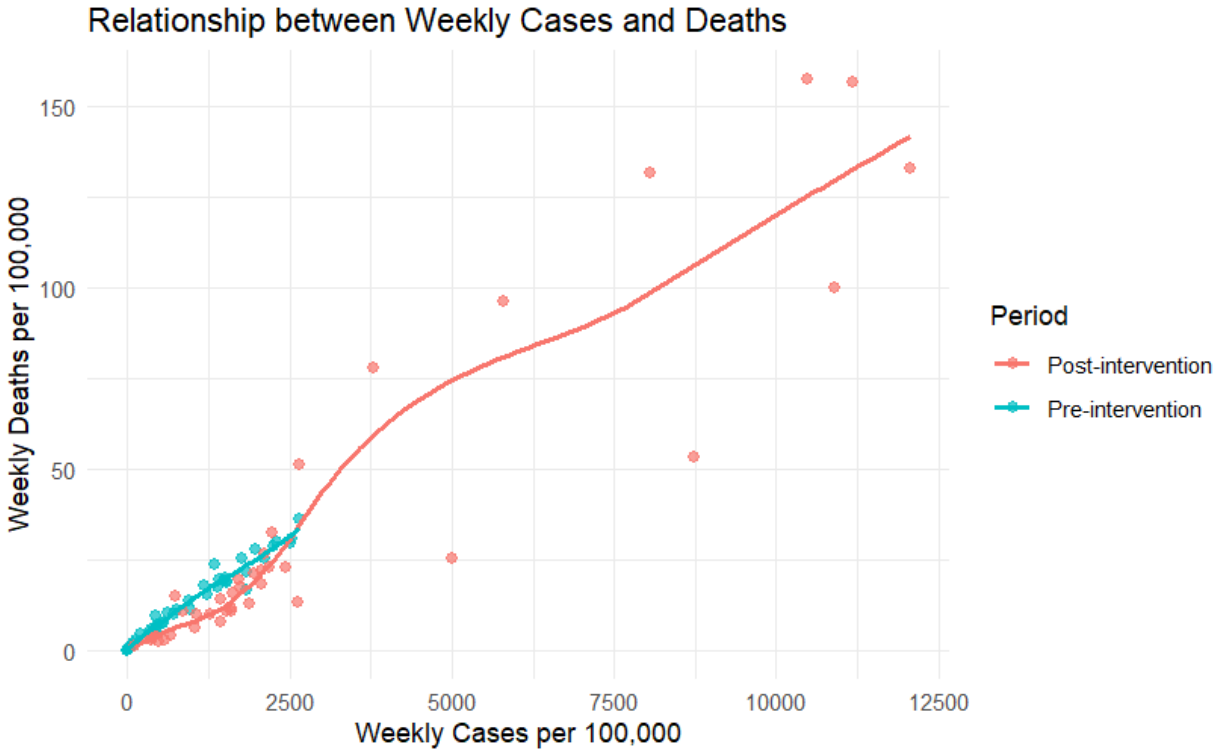


Figure 3: Multivariate scatterplot for pre and post intervention

6. Results

6.1 Counterfactual vs. Observed

The CausalImpact output quantifies the causal effect of vaccination on weekly mortality.

During the post-intervention period, the observed weekly mortality averaged 32.50 deaths per 100,000, whereas the model predicts that in the absence of vaccination, the average would have been 36.44, with a 95% credible interval from 33.03 to 40.42. The estimated average causal effect of vaccination is therefore -3.94 deaths per 100,000 per week, with a 95% credible interval from -7.92 to -0.53. This interval does not include zero, providing strong evidence of the mortality-reducing effect.

Summing across all post-rollout weeks, the observed cumulative mortality is 1365.02 deaths per 100,000, compared to a counterfactual estimate of 1530.66 deaths per 100,000, with credible bounds from 1387.07 to 1697.47. The estimated cumulative mortality reduction due to vaccination is -165.65 deaths per 100,000, with a credible interval from -332.46 to -22.05.

These values represent a relative reduction of 11% in weekly deaths (95% CI: -20% to -2%). The Bayesian posterior probability that the intervention had a real effect is 98.7%, with a tail-area probability of 0.013.

These findings support the conclusion that vaccination causally reduced COVID-19 mortality in India.

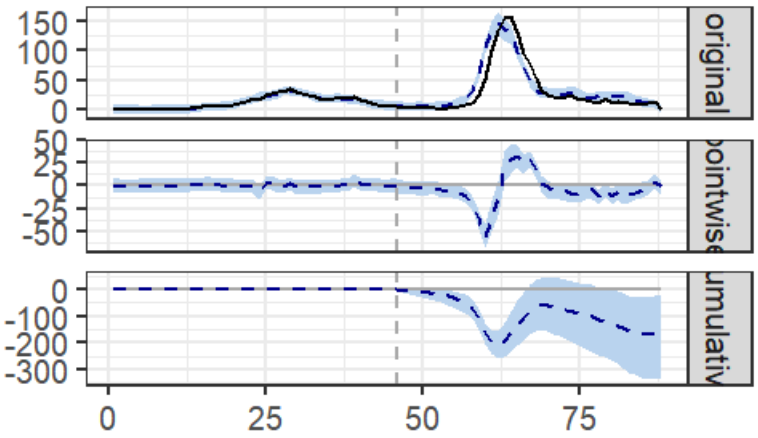


Figure 4: CausalImpact plot showing observed mortality, predicted counterfactual, pointwise effects, and cumulative effects.

6.2 Temporal Pattern and interpretation of the effect

Beyond the overall magnitude, the temporal pattern of the causal effect is informative. In the weeks immediately following the intervention date, the observed series and the predicted counterfactual series remain relatively close, reflecting the gradual nature of the vaccination rollout and the fact that immunity develops with some delay. As more of the population becomes fully vaccinated, the observed mortality increasingly falls below the counterfactual trajectory. This divergence is sustained over much of the post-intervention period, indicating a persistent mortality benefit rather than a short-lived fluctuation.

The pointwise causal effect series shows that weekly reductions in deaths per 100,000 are generally negative across most of the post-intervention period, with some variation in magnitude that likely reflects changing epidemic waves and variant dynamics. Importantly, the cumulative effect curve decreases steadily over time and does not return to zero, which implies that the gains from vaccination compound as the post-intervention

period progresses. Taken together, the average, cumulative, and relative effects indicate that vaccination not only reduced mortality on a typical week but also produced a substantial long-run reduction in the total number of deaths per 100,000 people.

From a substantive perspective, an 11% relative reduction in weekly mortality at the national level is large when scaled to India's population size. Even under the conservative bounds of the credible interval, the model suggests that vaccination prevented a meaningful number of deaths that would otherwise have occurred under similar case and policy conditions without vaccine-induced protection.

6.3 Model diagnostics and adequacy

The credibility of the causal conclusions depends on how well the model fits the pre-intervention period and how reasonable its assumptions appear in light of the residual behavior. The residual-over-time plot for the pre-intervention window shows that residuals fluctuate around zero without visible drift, with most values lying within a narrow band of approximately -4 to +4 deaths per 100,000. There is no apparent long-term trend in residuals, suggesting that the latent trend and regression components capture the main structure of the pre-intervention series. This is important because the counterfactual predictions in the post-intervention period are based on extrapolating the relationships learned in this pre-intervention window.

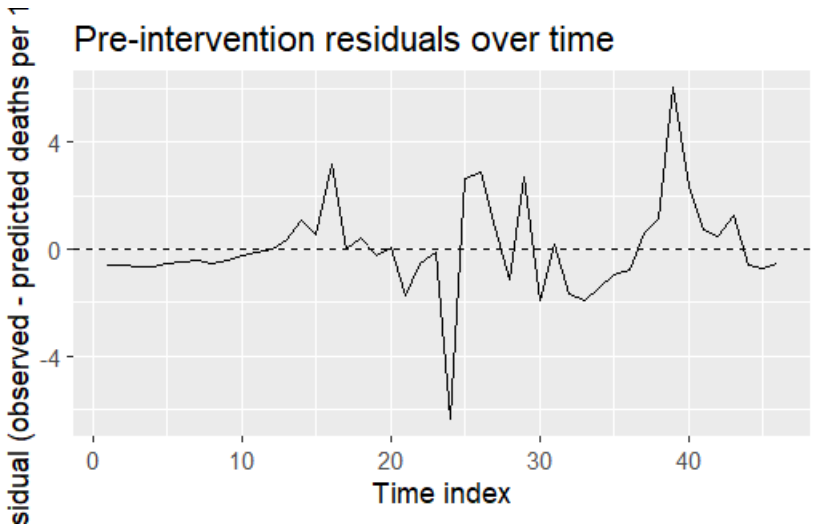


Figure 5: Pre-intervention residuals over time

The Q-Q plot of pre-intervention residuals indicates approximate normality. Points fall close to the reference line over most of the distribution, with slight deviations in the upper tail. This behavior is consistent with the Gaussian error assumptions. There is no evidence of extreme skewness or heavy tails that would raise concerns about the model.

Taken together, the residual-time and Q–Q diagnostics support the adequacy of the BSTS model for this application. The model appears to capture the dynamics of the pre-intervention mortality series well, which strengthens confidence that the counterfactual estimates for the post-intervention period provide a plausible basis for estimating the causal effect of vaccination.

7. Discussion

The findings support the conclusion that India's vaccination rollout causally reduced COVID-19 mortality at the national level. This conclusion aligns with global studies showing similar protective effects of vaccines across high and middle-income settings. Beyond its statistical significance, the estimated magnitude of the

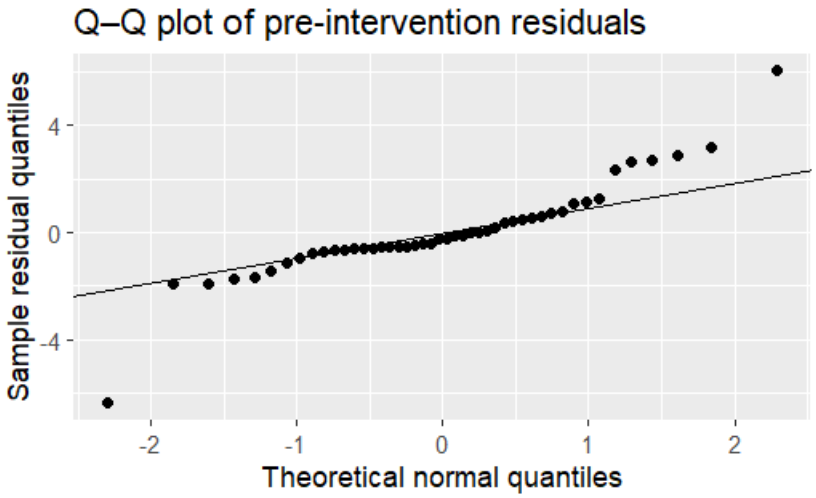


Figure 6: Q–Q plot of pre-intervention residuals

effect suggests that the vaccination campaign prevented a non-trivial number of deaths, particularly high-mortality Delta wave. The results demonstrate that vaccination remains a central tool for reducing mortality, especially when combined with public-health measures such as movement restrictions, masking, and testing.

Despite the robustness of the BSTS approach, several limitations must also be addressed. The model's validity depends o the assumption that relationships between covariates and mortality in the pre-intervention period would have persisted had vaccination not been introduced. This assumption, though standard in causal time-series analysis, cannot be fully verified. Moreover, the model does not explicitly incorporate the emergence of SARS-CoV-2 variants, which may alter the transmission dynamics in ways that covariates only partially capture. Vaccination rollout was gradual rather than instantaneous, and the intervention was modeled as a step function for analytical tractability. Finally, the analysis focuses on national effects and does not quantify heterogeneity across states, which may vary widely in healthcare infrastructure and vaccination updates.

Future research may extend this work by estimating state-level causal effects, comparing BSTS results with alternative causal frameworks such as synthetic controls or difference-in-differences, and incorporating additional covariates such as mobility, variant prevalence, or hospitalization data. More sophisticated models could jointly track infections, hospitalizations, and deaths to reflect the clinical progression of COVID-19 more accurately.

8. Conclusion

This study provides evidence that India's COVID-19 vaccination rollout causally reduced weekly COVID-19 mortality using a Bayesian structural time series model applied to cleaned, aggregated national epidemiological data. The counterfactual estimates produced by the CausalImpact framework indicate that mortality would have been substantially higher in the absence of vaccination. The analysis demonstrates the value of probabilistic causal modeling in public-health policy evaluation and showcases a rigorous, reproducible workflow from raw data acquisition through uncertainty quantification. As India and other nations prepare for future public-health emergencies, the lessons from COVID-19 underscore the importance of early vaccine deployment, robust data infrastructure, and scientifically grounded decision support.

APPENDIX

```
library(COVID19)
```

```
## Warning: package 'COVID19' was built under R version 4.4.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tidyr)
library(visdat)
```

```
## Warning: package 'visdat' was built under R version 4.4.3
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(CausalImpact)
```

```
## Warning: package 'CausalImpact' was built under R version 4.4.3

## Loading required package: bsts

## Warning: package 'bsts' was built under R version 4.4.3

## Loading required package: BoomSpikeSlab

## Warning: package 'BoomSpikeSlab' was built under R version 4.4.3

## Loading required package: Boom

## Warning: package 'Boom' was built under R version 4.4.3

##
## Attaching package: 'Boom'

## The following object is masked from 'package:stats':
##
##     rWishart

##
## Attaching package: 'BoomSpikeSlab'

## The following object is masked from 'package:stats':
##
##     knots

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.4.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.4.3
```

```
##
## ######################### Warning from 'xts' package #########################
## #                                                                            #
## # The dplyr lag() function breaks how base R's lag() function is supposed to #
## # work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or      #
## # source() into this session won't work correctly.                           #
## #                                                                            #
## # Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
## # conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop          #
## # dplyr from breaking base R's lag() function.                               #
## #                                                                            #
## # Code in packages is not affected. It's protected by R's namespace mechanism #
## # Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning. #
## #                                                                            #
## ##############################################################################
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'bsts'
```

```
## The following object is masked from 'package:BoomSpikeSlab':
##
##     SuggestBurn
```

Extract India, level 2 (states/Union Territories)

```
india_raw <- covid19(
  country = "India",
  level   = 2,          # states / union territories
  start   = "2020-03-01",
  end     = "2022-12-31",
  verbose = FALSE
)

dim(india_raw)
```

```
## [1] 20998    47
```

```
glimpse(india_raw)
```

```
## Rows: 20,998
## Columns: 47
## $ id                                  <chr> "0e833256", "0e833256", "0e833256"~
```

```
## $ date                                  <date> 2020-04-09, 2020-04-10, 2020-04-1~
## $ confirmed                             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ deaths                                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ recovered                             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ tests                                 <int> 80, 130, 211, 211, 211, 211, 356, ~
## $ vaccines                              <int64> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ people_vaccinated                     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ people_fully_vaccinated               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ hosp                                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ icu                                   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ vent                                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ school_closing                        <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
## $ workplace_closing                     <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2~
## $ cancel_events                         <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ gatherings_restrictions               <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ transport_closing                     <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ stay_home_restrictions                <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
## $ internal_movement_restrictions        <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ international_movement_restrictions <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ information_campaigns                 <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ testing_policy                        <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
## $ contact_tracing                       <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ facial_coverings                      <int> -4, -4, -4, -4, -4, -4, -4, -4, -4~
## $ vaccination_policy                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ elderly_people_protection             <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
## $ government_response_index             <dbl> -89.84, -89.84, -89.84, -89.84, -8~
## $ stringency_index                      <dbl> -100.00, -100.00, -100.00, -100.00~
## $ containment_health_index              <dbl> -91.96, -91.96, -91.96, -91.96, -9~
## $ economic_support_index                <int> -75, -75, -75, -75, -75, -75, -75,~
## $ administrative_area_level             <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ administrative_area_level_1           <chr> "India", "India", "India", "India"~
## $ administrative_area_level_2           <chr> "Dadra And Nagar Haveli", "Dadra A~
## $ administrative_area_level_3           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ latitude                              <dbl> 20.20629, 20.20629, 20.20629, 20.2~
## $ longitude                             <dbl> 73.01859, 73.01859, 73.01859, 73.0~
## $ population                            <int> 586956, 586956, 586956, 586956, 58~
## $ iso_alpha_3                           <chr> "IND", "IND", "IND", "IND", "IND",~
## $ iso_alpha_2                           <chr> "IN", "IN", "IN", "IN", "IN", "IN"~
## $ iso_numeric                           <int> 356, 356, 356, 356, 356, 356, 356,~
## $ iso_currency                          <chr> "INR", "INR", "INR", "INR", "INR",~
## $ key_local                             <chr> "DN", "DN", "DN", "DN", "DN", "DN"~
## $ key_google_mobility                   <chr> "ChIJ-aMVMULL4DsRm4NbagV3U18", "Ch~
## $ key_apple_mobility                    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ key_jhu_csse                          <chr> "INDN", "INDN", "INDN", "INDN", "I~
## $ key_nuts                              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ key_gadm                              <chr> "IND.8_1", "IND.8_1", "IND.8_1", "~
```

Key variables:

- id, date, administrative_area_level_2
- confirmed, deaths, tests
- people_fully_vaccinated, vaccines

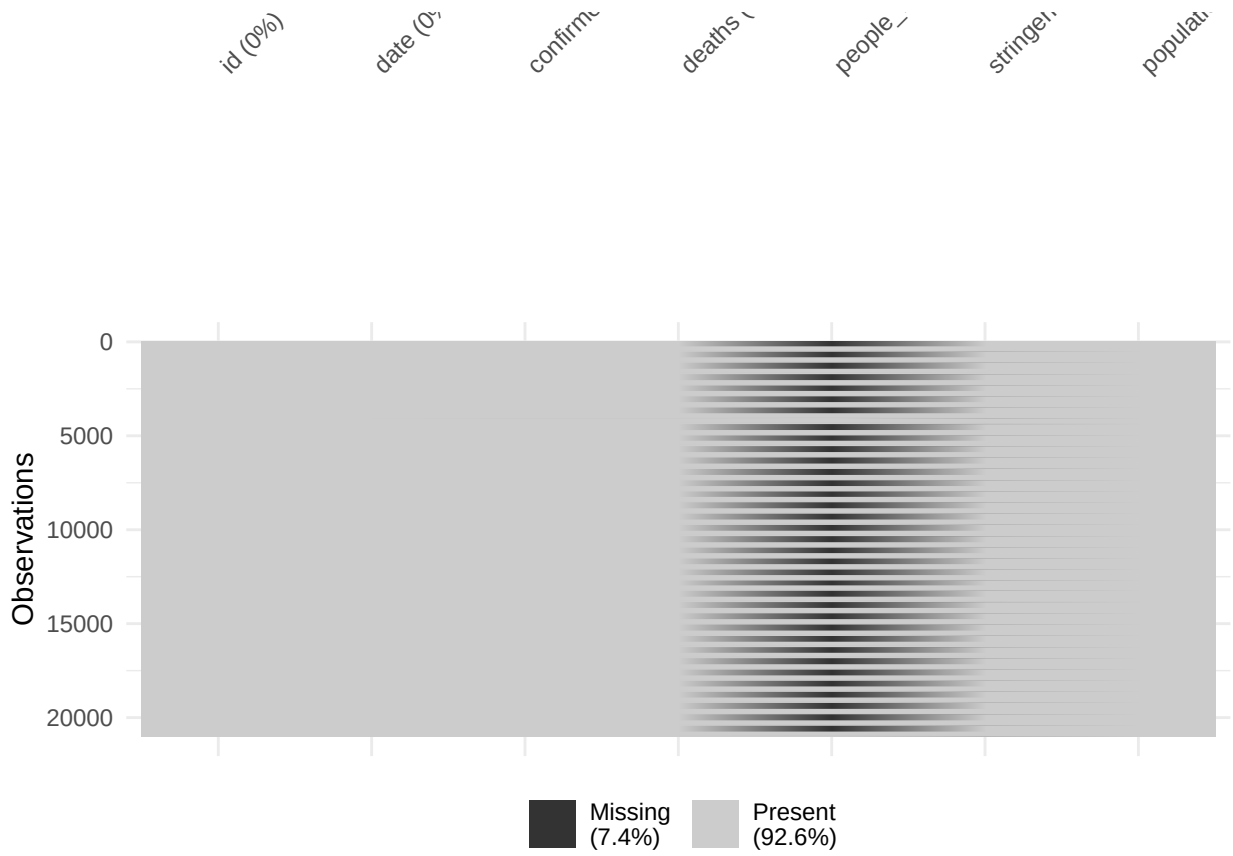
- stringency_index
- population

Missingness diagnosis

```
missing_prop <- sort(colMeans(is.na(india_raw)), decreasing = TRUE)

missing_prop[1:15]
```

```
##                         hosp                          icu
##                   1.00000000                   1.00000000
##                         vent administrative_area_level_3
##                   1.00000000                   1.00000000
##                     key_nuts           key_apple_mobility
##                   1.00000000                   0.97099724
##                     vaccines            people_vaccinated
##                   0.50795314                   0.50795314
##     people_fully_vaccinated            transport_closing
##                   0.50795314                   0.03428898
##                        tests          key_google_mobility
##                   0.03257453                   0.02876464
##                     key_gadm            workplace_closing
##                   0.02876464                   0.01200114
##                cancel_events
##                   0.01200114
```

```
india_raw %>%
  dplyr::select(
    id,
    date,
    confirmed,
    deaths,
    people_fully_vaccinated,
    stringency_index,
    population
  ) %>%
  visdat::vis_miss()
```



Findings:

- Vaccination variables are missing in early 2020–early 2021 (before rollout).
- stringency_index has gaps for some state-days.
- The population is valid for most states; a small number of rows require filtering.
- Cumulative counts show occasional downward corrections → we must construct daily increments.

Cleaning and feature engineering

Filter valid rows and time window

```
india_clean <- india_raw %>%
  filter(!is.na(population), population > 0) %>%
  filter(date >= as.Date("2020-03-01"),
         date <= as.Date("2022-12-31"))
```

Construct daily increments from cumulative records

```
india_daily <- india_clean %>%
  group_by(id) %>%
  arrange(date, .by_group = TRUE) %>%
  mutate(
    daily_cases  = confirmed - lag(confirmed),
    daily_deaths = deaths    - lag(deaths),
```

```
    daily_tests  = tests     - lag(tests),
    daily_cases  = ifelse(daily_cases  < 0, 0, daily_cases),
    daily_deaths = ifelse(daily_deaths < 0, 0, daily_deaths),
    daily_tests  = ifelse(daily_tests  < 0, 0, daily_tests)
  ) %>%
  ungroup()
```

Per-capita measures

```
india_daily <- india_daily %>%
  mutate(
    cases_per_100k   = 1e5 * daily_cases  / population,
    deaths_per_100k  = 1e5 * daily_deaths / population,
    tests_per_100k   = 1e5 * daily_tests  / population,
    vax_full_per_100 = 100 * people_fully_vaccinated / population
  )
```

Impute vaccination and stringency, and aggregate weekly

```
india_daily <- india_daily %>%
  group_by(id) %>%
  arrange(date, .by_group = TRUE) %>%
  tidyr::fill(vax_full_per_100, stringency_index, .direction = "down") %>%
  ungroup()

india_weekly <- india_daily %>%
  mutate(week = floor_date(date, unit = "week")) %>%
  group_by(id, week) %>%
  summarize(
    weekly_cases_per_100k   = sum(cases_per_100k,   na.rm = TRUE),
    weekly_deaths_per_100k  = sum(deaths_per_100k,  na.rm = TRUE),
    weekly_tests_per_100k   = sum(tests_per_100k,   na.rm = TRUE),
    mean_stringency         = mean(stringency_index, na.rm = TRUE),
    end_vax_full_per_100    = dplyr::last(na.omit(vax_full_per_100)),
    population              = dplyr::first(population),
    state_name              = dplyr::first(administrative_area_level_2),
    .groups = "drop"
  ) %>%
  filter(
    !is.na(weekly_deaths_per_100k),
    !is.na(mean_stringency)
  )
```

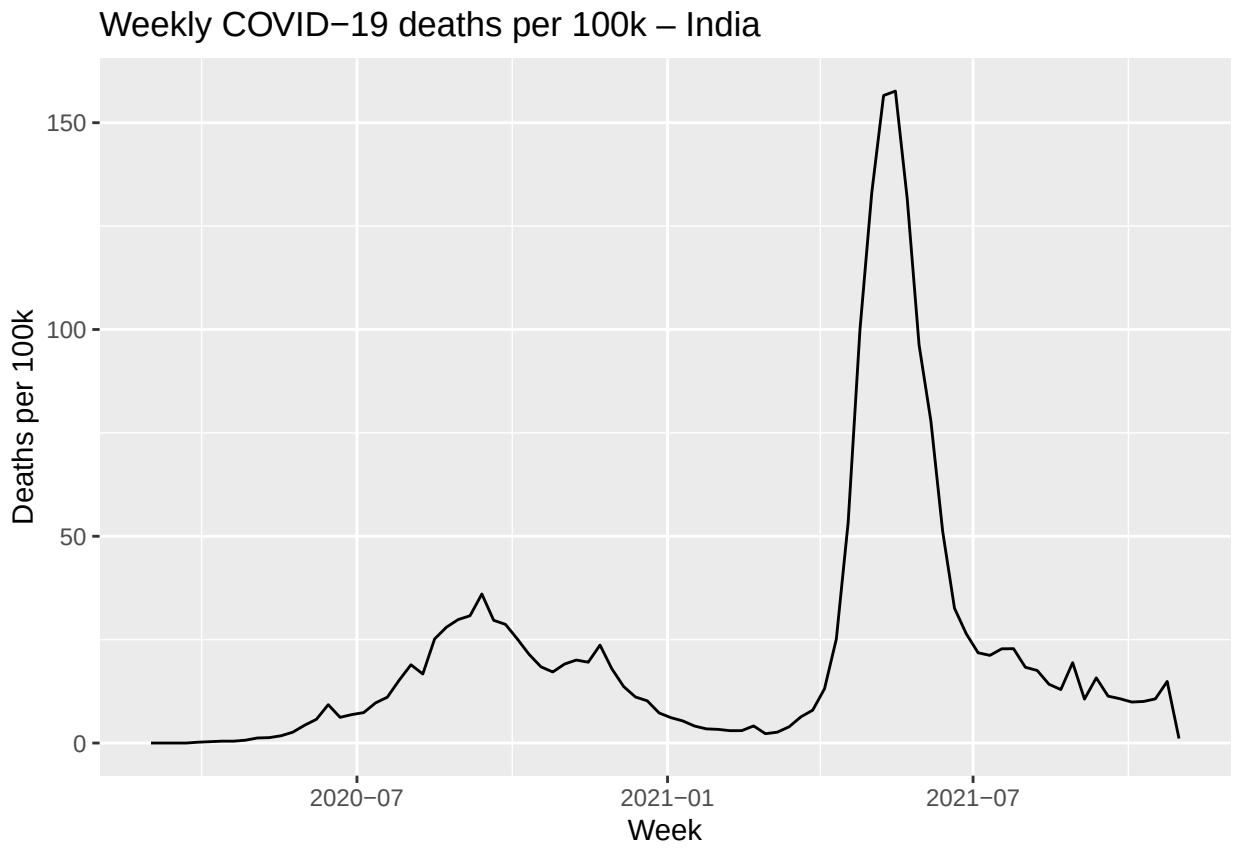
```
sort(colMeans(is.na(india_weekly)), decreasing = TRUE)
```

```
##   end_vax_full_per_100                     id                   week
##              0.4914586              0.0000000              0.0000000
##  weekly_cases_per_100k weekly_deaths_per_100k  weekly_tests_per_100k
##              0.0000000              0.0000000              0.0000000
##        mean_stringency             population             state_name
##              0.0000000              0.0000000              0.0000000
```

EDA

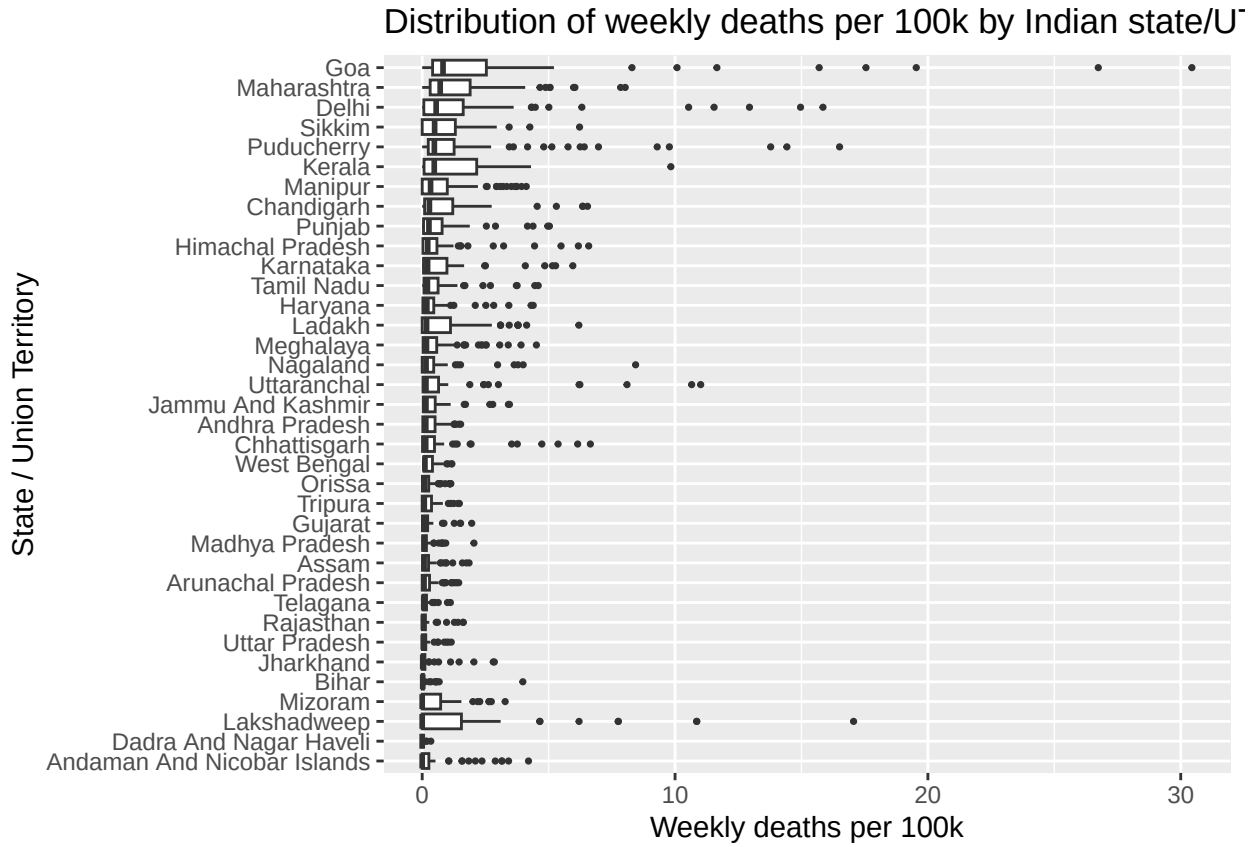
National weekly mortality and vaccination

```
india_nat_weekly <- india_weekly %>%
  group_by(week) %>%
  summarize(
    deaths_per_100k  = sum(weekly_deaths_per_100k),
    cases_per_100k   = sum(weekly_cases_per_100k),
    tests_per_100k   = sum(weekly_tests_per_100k),
    mean_vax_per_100 = weighted.mean(end_vax_full_per_100,
                                     population, na.rm = TRUE),
    mean_stringency  = weighted.mean(mean_stringency,
                                     population, na.rm = TRUE)
  )

ggplot(india_nat_weekly, aes(week, deaths_per_100k)) +
  geom_line() +
  labs(
    title = "Weekly COVID-19 deaths per 100k - India",
    x = "Week", y = "Deaths per 100k"
  )
```



State-level variation (boxplots / heatmap)

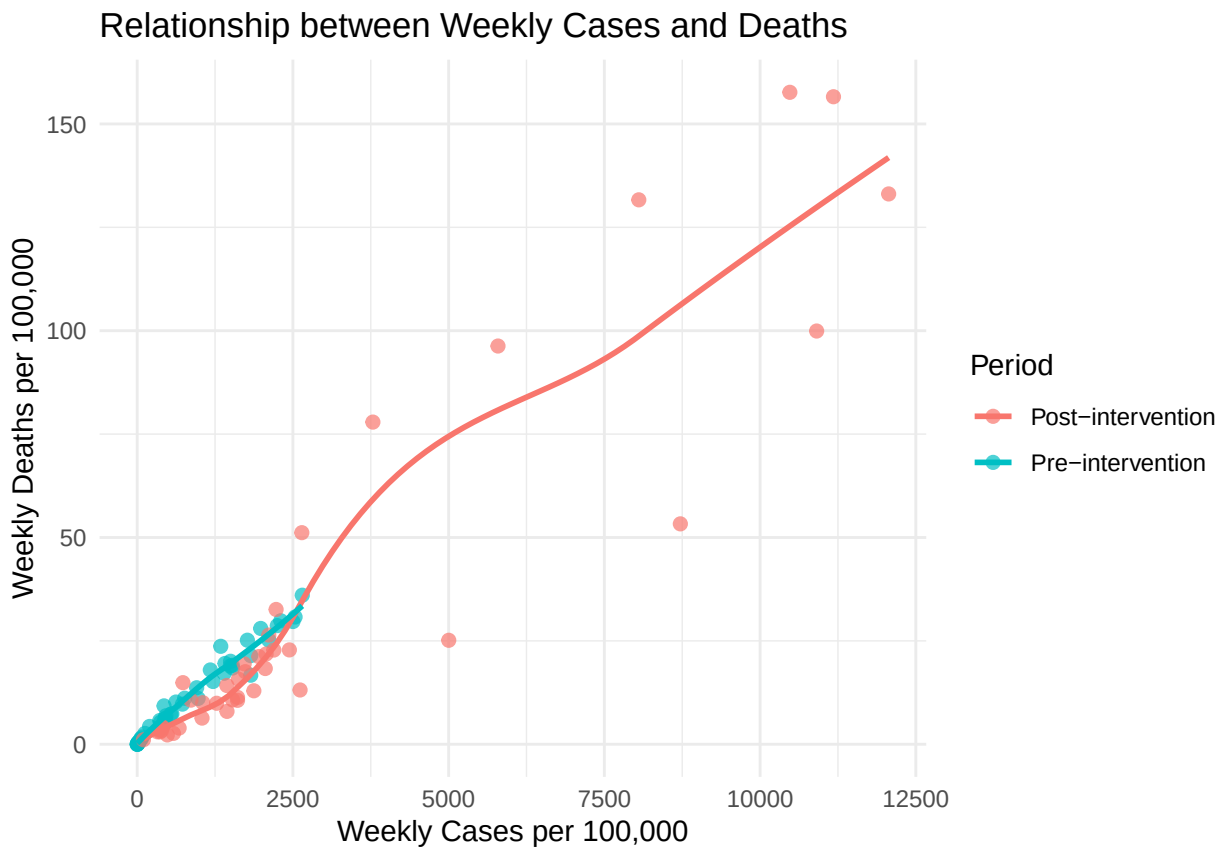
```
ggplot(india_weekly,
       aes(x = reorder(state_name, weekly_deaths_per_100k, median),
           y = weekly_deaths_per_100k)) +
  geom_boxplot(outlier.size = 0.6) +
  coord_flip() +
  labs(
    title = "Distribution of weekly deaths per 100k by Indian state/UT",
    x = "State / Union Territory",
    y = "Weekly deaths per 100k"
  )
```




```
intervention_date <- as.Date("2021-01-16")
india_nat_weekly$period <- ifelse(
  india_nat_weekly$week <= intervention_date,
  "Pre-intervention",
  "Post-intervention"
)

ggplot(india_nat_weekly,
                    aes(x = cases_per_100k,
                        y = deaths_per_100k,
                        color = period)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(
```

```
    title = "Relationship between Weekly Cases and Deaths",
    x = "Weekly Cases per 100,000",
    y = "Weekly Deaths per 100,000",
    color = "Period"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Constructing the causal-impact dataset (national level)

```
# Choose key covariates: cases and stringency (and optionally tests)
ci_data <- india_nat_weekly %>%
  dplyr::select(
    week,
    deaths_per_100k,
    cases_per_100k,
    tests_per_100k,
    mean_stringency
  ) %>%
  arrange(week)

head(ci_data)
```

```
## # A tibble: 6 x 5
```

```
##   week       deaths_per_100k cases_per_100k tests_per_100k mean_stringency
##   <date>               <dbl>          <dbl>          <dbl>           <dbl>
## 1 2020-03-01         0               0.0139          0              -22.2
## 2 2020-03-08         0.00164         0.471           0              -29.5
## 3 2020-03-15         0               4.35            0              -58.9
## 4 2020-03-22         0.00327         3.97            0              -96.2
## 5 2020-03-29         0.211           8.84            5.74          -100
## 6 2020-04-05         0.330          12.1           242.            -100
```

Define intervention date and periods

India's national vaccination rollout began mid-January 2021. We treat that as the **intervention start**, and choose:

- **pre-period:** from the start of the dataset until just before rollout
- **post-period:** from rollout onward

Note: adjust exact boundaries if needed for your analysis.

```
intervention_date <- as.Date("2021-01-16")

pre.period <- c(min(ci_data$week),
                intervention_date - 7)    # week ending before rollout

post.period <- c(intervention_date,
                 max(ci_data$week))
pre.period
```

```
## [1] "2020-03-01" "2021-01-09"
```

```
post.period
```

```
## [1] "2021-01-16" "2021-10-31"
```

Prepare matrix for CausalImpact

`CausalImpact` expects a matrix/data frame where:

- Column 1 = response (deaths_per_100k)
- Columns 2+ = covariates

```
ci_mat <- ci_data %>%
  dplyr::select(
    deaths_per_100k,
    cases_per_100k,
    tests_per_100k,
    mean_stringency
  )

ci_mat <- as.matrix(ci_mat)
```

Bayesian Causal Impact analysis

Fit the model

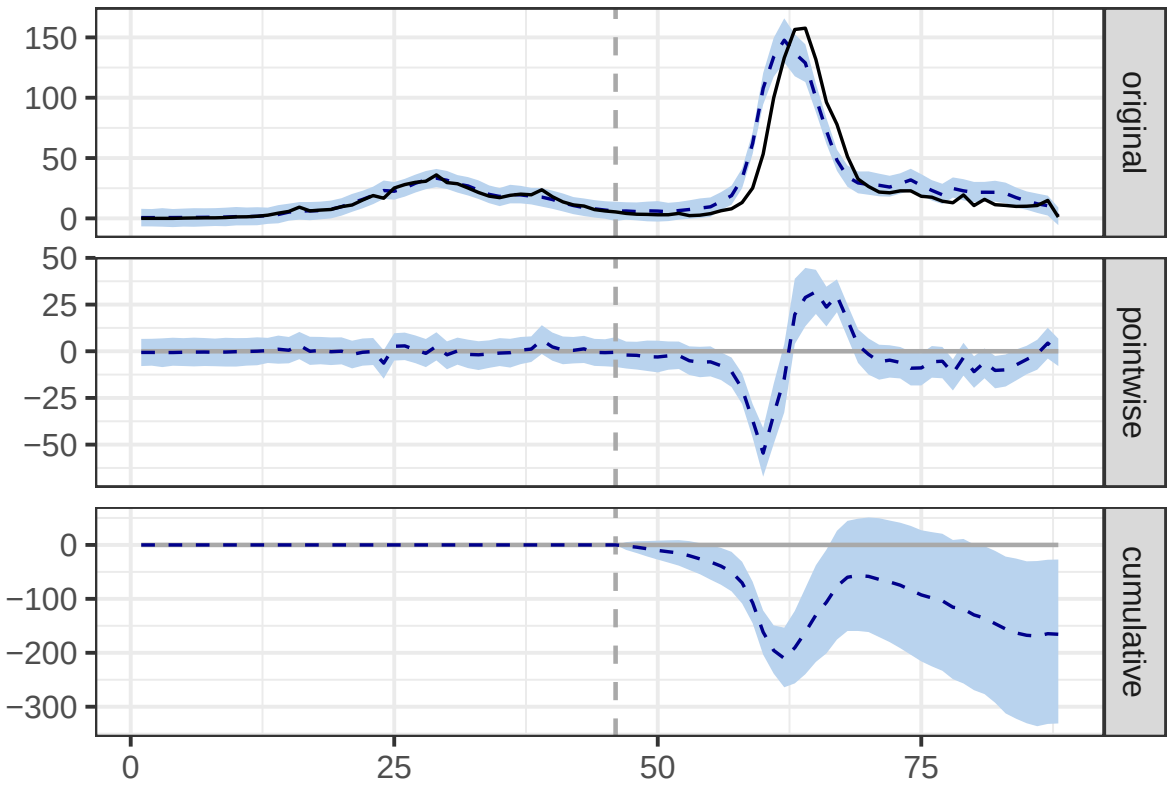
```
intervention_idx <- which.min(abs(ci_data$week - intervention_date))
pre.period  <- c(1, intervention_idx - 1)
post.period <- c(intervention_idx, nrow(ci_data))
impact <- CausalImpact(ci_mat, pre.period, post.period)

# Overall numeric summary
summary(impact)
```

```
## Posterior inference {CausalImpact}
##
##                          Average         Cumulative
## Actual                   33              1365
## Prediction (s.d.)        36 (1.9)        1531 (78.0)
## 95% CI                   [33, 40]        [1392, 1696]
##
## Absolute effect (s.d.)   -3.9 (1.9)      -165.6 (78.0)
## 95% CI                   [-7.9, -0.64]   [-331.1, -27.02]
##
## Relative effect (s.d.)   -11% (4.5%)     -11% (4.5%)
## 95% CI                   [-20%, -1.9%]   [-20%, -1.9%]
##
## Posterior tail-area probability p:   0.00901
## Posterior probability of an effect:  99.0991%
##
## For more details, type: summary(impact, "report")
```

```
# More detailed explanation-style summary
summary(impact, "report")
```

```
## Analysis report {CausalImpact}
##
##
## During the post-intervention period, the response variable had an average value of approx. 32.50. By
##
## Summing up the individual data points during the post-intervention period (which can only sometimes
##
## The above results are given in terms of absolute numbers. In relative terms, the response variable sh
##
## This means that the negative effect observed during the intervention period is statistically signifi
##
## The probability of obtaining this effect by chance is very small (Bayesian one-sided tail-area probab
```

```
plot(impact)
```



Extracting posterior effect and credible intervals

```
summary(impact, "report")
```

```
## Analysis report {CausalImpact}
##
##
## During the post-intervention period, the response variable had an average value of approx. 32.50. By
##
## Summing up the individual data points during the post-intervention period (which can only sometimes
##
## The above results are given in terms of absolute numbers. In relative terms, the response variable sh
##
## This means that the negative effect observed during the intervention period is statistically signifi
##
## The probability of obtaining this effect by chance is very small (Bayesian one-sided tail-area probal
```

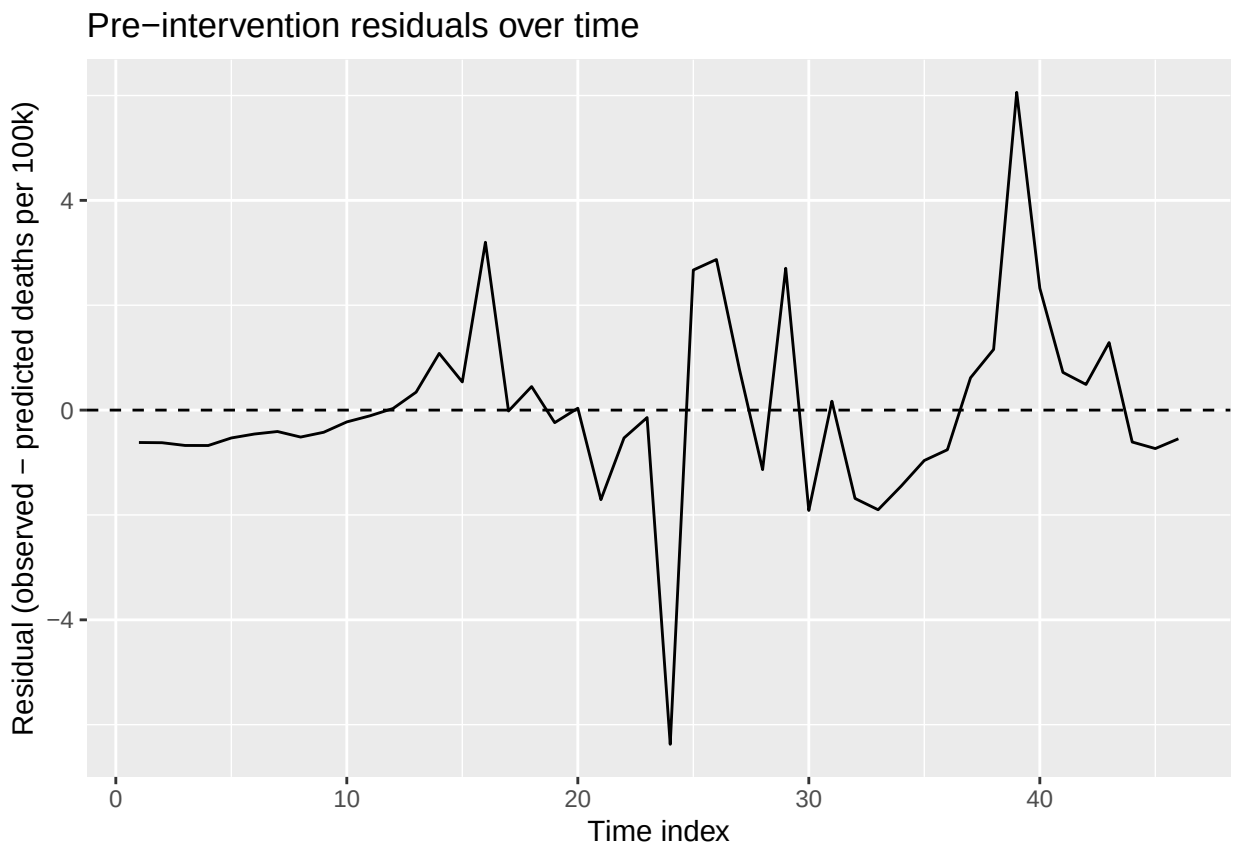
```
impact_table <- impact$summary
impact_table
```

```
##                Actual       Pred Pred.lower Pred.upper   Pred.sd   AbsEffect
## Average      32.50042   36.44437   33.14386   40.38365  1.856095   -3.943948
## Cumulative 1365.01769 1530.66350 1392.04209 1696.11338 77.955993 -165.645815
##            AbsEffect.lower AbsEffect.upper AbsEffect.sd  RelEffect
## Average          -7.883231      -0.6434382     1.856095 -0.1057794
```

```
## Cumulative     -331.095689     -27.0244037    77.955993 -0.1057794
##             RelEffect.lower RelEffect.upper RelEffect.sd alpha           p
## Average          -0.1952084     -0.01941345   0.04515646  0.05 0.009009009
## Cumulative       -0.1952084     -0.01941345   0.04515646  0.05 0.009009009
```

Diagnostics

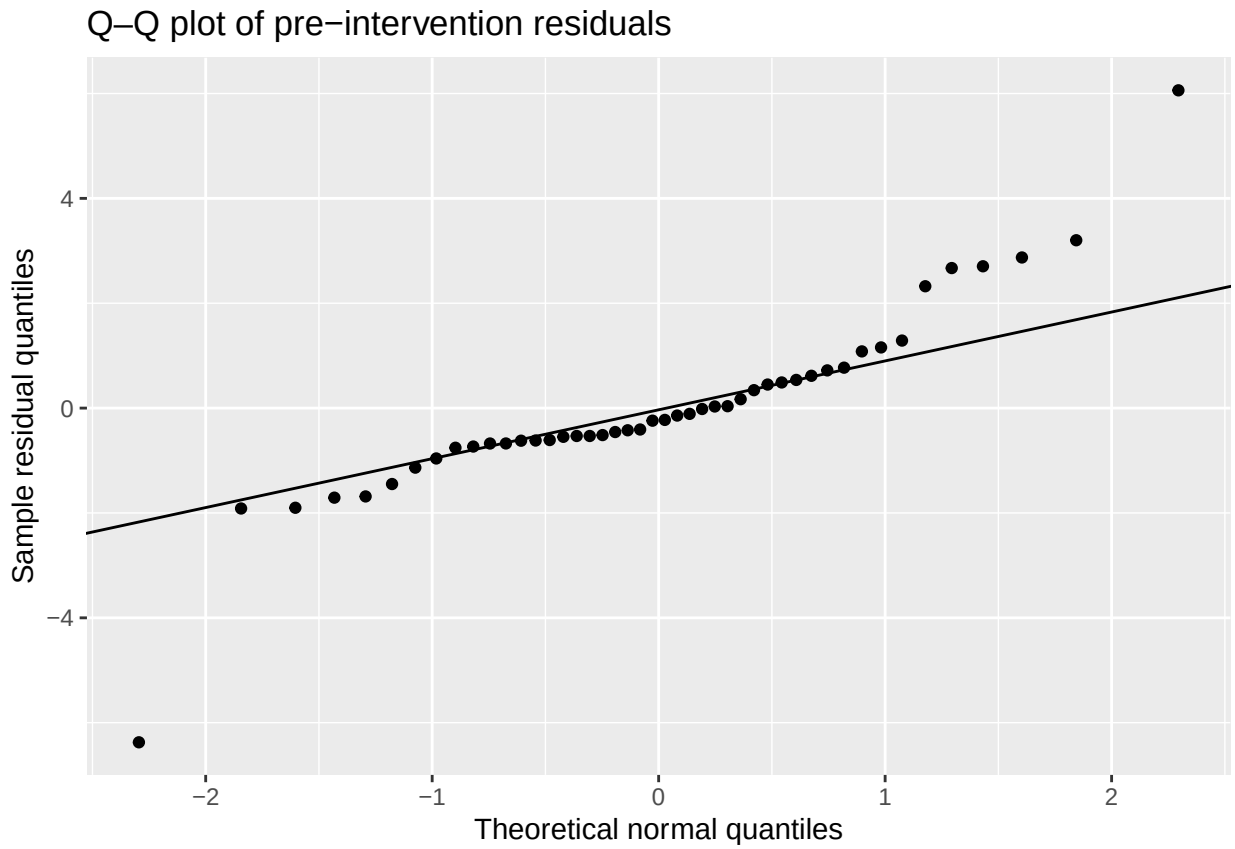
```
# Turn the series into a data frame
impact_df <- as.data.frame(impact$series)

# Add an integer time index
impact_df$time_idx <- seq_len(nrow(impact_df))

# Compute residuals: observed - predicted
impact_df$resid <- impact_df$response - impact_df$point.pred

# Pre-period ends at pre.period[2] (since you used integer indices)
pre_df <- impact_df[1:pre.period[2], ]

# Plot residuals over time in the pre period
ggplot(pre_df, aes(x = time_idx, y = resid)) +
  geom_line() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    title = "Pre-intervention residuals over time",
    x = "Time index",
    y = "Residual (observed - predicted deaths per 100k)"
  )
```



These diagnostics help argue that:

- The model fits well in the pre-period (good predictive power without vaccines).
- Deviations in the post-period can reasonably be attributed to the intervention, given the assumptions.

```
ggplot(pre_df, aes(sample = resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(
    title = "Q-Q plot of pre-intervention residuals",
    x = "Theoretical normal quantiles",
    y = "Sample residual quantiles"
  )
```



Strengths:

- Uses a principled **Bayesian structural time series** approach.
- Adjusts for other time-varying drivers (cases per 100k, testing intensity, stringency index).
- Quantifies uncertainty via **posterior credible intervals**.
- Fits well in the pre-period (residual diagnostics and Q–Q plot)

Key assumptions/limitations:

- The **relationship between covariates and deaths** learned in the pre-vaccination period would have persisted in the post-period in the absence of the rollout.
- No major unmeasured intervention occurs exactly at the rollout time that changes mortality dynamics in a way not captured by covariates (e.g., a huge change in clinical treatment protocols).
- Vaccination is modeled as a **single national intervention**; heterogeneity across states is not explicitly modeled in the causal component (though it informed cleaning and EDA).
- As with all observational causal analyses, **confounding** cannot be completely ruled out.

Conclusion:

Given model fit and posterior estimates, the analysis suggests that India's vaccination rollout **causally reduced weekly COVID-19 mortality per 100,000** beyond what would be expected based on infection dynamics and policy stringency alone. The project illustrates how a messy, high-dimensional public-health dataset can be turned into a clear **causal story with uncertainty** using an end-to-end data science pipeline.

References

(Guidotti and Ardia 2020; Guidotti 2022; Hale et al. 2021; Brodersen et al. 2015; Makridakis 2018; Choudhary, Choudhary, and Singh 2021)

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. "Inferring causal impact using Bayesian structural time-series models." *The Annals of Applied Statistics* 9 (1): 247–74. https://doi.org/10.1214/14-AOAS788.

Choudhary, Om Prakash, Priyanka Choudhary, and Indraj Singh. 2021. "India's COVID-19 Vaccination Drive: Key Challenges and Resolutions." *Lancet Infect Dis* 21 (11): 1483–84.

Guidotti, Emanuele. 2022. "A Worldwide Epidemiological Database for COVID-19 at Fine-Grained Spatial Resolution." *Scientific Data* 9 (1): 112. https://doi.org/10.1038/s41597-022-01245-1.

Guidotti, Emanuele, and David Ardia. 2020. "COVID-19 Data Hub." *Journal of Open Source Software* 5 (51): 2376. https://doi.org/10.21105/joss.02376.

Hale, Thomas, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, et al. 2021. "A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker)." *Nature Human Behaviour* 5 (4): 529–38. https://doi.org/10.1038/s41562-021-01079-8.

Makridakis, Evangelos AND Assimakopoulos, Spyros AND Spiliotis. 2018. "Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward." *PLOS ONE* 13 (3): 1–26. https://doi.org/10.1371/journal.pone.0194889.