Final Report (STATS 551)
Snigdha Pakala (vpakala)
December 18, 2024

1. Introduction

Vaccinations have been an integral part of medicine for a long time, dating back to the 15th century when humans attempted to prevent disease through variolation - a process where healthy people were intentionally exposed to smallpox to protect themselves against it. Vaccines have played a critical role in the eradication and reduction of many diseases such as small pox, measles, and polio. However, despite the scientific evidence for their positive impact, much controversy surrounds vaccination. Vaccine hesitancy due to misinformation and distrust in science, have prevented high population vaccination rates.

Vaccination campaigns remain one of the most effective tools for preventing disease outbreaks, but their success depends on addressing these challenges. Understanding how effective vaccination campaigns are is necessary to optimize public health strategies. Recently, the topic of vaccines has gained increased attention and sparked widespread public debate due to the COVID-19 pandemic. This heightened awareness makes a study of vaccination campaign effectiveness particularly timely and relevant.

This study investigates the impact of a pneumococcal vaccination campaign on increasing immunization rates across U.S. communities from 2018 to 2021. We address community-level variability, using a Bayesian hierarchical model to explore how outcomes differ by geographic and demographic factors. Prior research highlights both the potential effectiveness and challenges of vaccination campaigns, but this analysis provides insights through the lens of Bayesian modeling.

2. Data

The dataset used in this analysis comes from the Behavioral Risk Factor Surveillance System (BRFSS), a comprehensive annual survey conducted by the Centers for Disease Control and Prevention (CDC). This data provide vaccination coverage estimates for adults aged 18 years and older who participated in BRFSS interviews. The survey is a critical component of the CDC's AdultVaxView platform, designed to monitor and evaluate adult immunization trends in the United States. The dataset includes vaccination coverage estimates stratified by various demographic and contextual factors, such as geographical regions, age groups, race, and survey year.

It is important to note that data availability is incomplete for all geographic regions. A subset of the data lacks coverage estimates in certain areas or for specific populations due to sample size limitations or other constraints. This variability, along with the presence of many hierarchical models in studying this research question, underscore the importance of robust statistical techniques, such as Bayesian methods, to handle the uncertainty in the dataset.

3. Method

The hierarchical model used in this study is the following:

For the likelihood, it makes sense that the vaccination counts are modeled using a Binomial distribution. Thus, for each observation y_i,

$$y_i \sim \text{Binomial}(\text{total}_i, p_i), \quad \text{logit}(p_i) = \alpha[\text{region}_i] + \beta \cdot \text{time\_period}_i + \gamma \cdot (\text{age\_group}_i - 1).$$

where p_i is the probability of vaccination for observation i.

Regions were created by grouping areas with similar vaccination rates. The four regions in the data are: New England, the most vaccinated region; Middle Atlantic, with distinct vaccination patterns compared to New England but higher than the national average; East North Central, with moderate vaccination rates and similar regional patterns; and other regions with similar rates, which have lower statistical power individually. To incorporate regional effects, each region has its own vaccination rates while sharing information through the global mean and standard deviation. So for each region r,

$$\alpha[r] \sim \text{Normal}(\mu_\alpha, \sigma_\alpha).$$

The informative priors were chosen based on CDC vaccination data to estimate parameters:

$$\mu_\alpha \sim \text{Normal}(\text{logit}(0.65), 0.5), \quad \sigma_\alpha \sim \text{Normal}(0, 0.5),$$

$$\beta \sim \text{Normal}(0, 0.5), \quad \gamma \sim \text{Normal}(0, 0.5).$$

The justification of these is as follows: Since μ_α represents the average vaccination rate across regions, it is centered at 65% with moderate uncertainty to align with the CDC-reported average rate. The logit transofrmation ensures the probabilities remain between $[0, 1]$. This combined with σ_α at Normal(0, 0.5) - which allows for moderate regional variation but penalizes excessively large variability - makes this a weekly informative prior which reflects a reasonable range for variability. These two combined create a standard normal prior for raw region-specific effects, scaled by σ_α, for $\alpha_{raw}[r]$. Next, β captures the time period effect in relation to COVID-19 (pre/post 2020), so it is normally distributed and centered at 0 for no-effect but allowing moderate deviations. Lastly, γ captures the age-group effect, and represented the same as β.

The prior predictive check shows that the prior predictions capture the general shape with some uncertainty, so the prior is reasonable for our data.

Posterior: Using Bayes' theorem, the posterior distribution is:

$$P(\mu_\alpha, \sigma_\alpha, \boldsymbol{\alpha}, \beta, \gamma \mid \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{p}) P(\mathbf{p} \mid \boldsymbol{\alpha}, \beta, \gamma) P(\boldsymbol{\alpha} \mid \mu_\alpha, \sigma_\alpha) P(\mu_\alpha) P(\sigma_\alpha) P(\beta) P(\gamma)$$

where the posterior distribution reveals the campaign effectiveness on the log-odds scale.

This hierarchical structure is the basis for the Stan model. I set adapt delta at 0.99 to reduce divergent transitions and max tree-depth at 12 to prevent the sampler from getting stuck in very long trajectories and balance computation time with exploration. The model draws 3000 iterations for each of the 4 chains to provide sufficient samples after discarding the first 1500 to allow chains to reach stable stationary distributions. The trace plots, autocorrelation and posterior predictive checks below show stable mixing and good convergence, and the parameters appear well defined. The parameters (β and γ) show great stability with quick decay in autocorrelation. All of the Rhat values being very close to 1 indicate good convergence. Lastly, the effective sample sizes are around 4000 for (β and γ) and between 1378 to 2391 for the α_{raw} parameters, indicating good sampling efficiency for parameters overall.

4. Results

Our results provide many useful implications and understandings of the nuances behind vaccination campaign effectiveness. Firstly, $\beta \approx -0.06$ with 95% credible interval $[-0.08, -0.04]$ is capturing the impact of the COVID-19 time period on vaccination rates. -0.06 suggests that on average, the odds of vaccination during and after the pandemic were approximately 6% lower than pre-COVID times. Since the interval does not include 0, these results are statistically significant. This 6% reduction is a substantial finding when considering a large population. This result highlights the importance of understanding how global events, such as a pandemic, affect public health behaviors. There was also about a 2 percentage point decrease between mean vaccination probability pre-COVID vs. post-COVID, aligning with disruptions caused by the pandemic. Acknowledging this decline allows public health officials to design more targeted interventions aimed at boosting vaccination rates before disastrous crises occur.

Next, regional variations provide much insight on strategization of effective vaccine campaigns. $\sigma_\alpha \approx 0.15$ with 95% credible interval $[0.04, 0.48]$ implies that there are statistically significant regional differences in vaccination rates. This is crucial as it suggests that localized public health interventions such as targeted outreach in under-performing regions might be an effective way to improve vaccination uptake.

Additionally, $\gamma \approx -0.001$ with 95% credible interval $[-0.99, 0.98]$ implies that there were no meaningful differences between the two age groups (18-64 years vs. >= 65 years). However, the uncertainty in this parameter could be consistent with the idea that the effect might vary regionally. This result is also implies vaccine campaigns be created without significant age-based targeting, and instead focus on strategies for particular regions.

Overall, the evidence of regional variation means campaigns can be more specifically designed for under-performing geographical areas. The decline in vaccinations during the pandemic can guide future changes of more frequent targeted campaigns in non-pandemic periods, ensuring that vaccination efforts are not hindered by future crises.

5. Limitations and Next Steps

Despite the model's overall utility, there were also limitations that could have been improved. First, as Figures 9-12 show, the model struggled with extreme values, suggesting that more flexible likelihoods (such as negative binomial) may capture outliers better in regions with high vaccination success. Second, the pre- and post-COVID categories may oversimplify temporal dynamics, potentially missing the nuanced trends visualized in Figure 1. Lastly, Figure 8 reveals correlations between the global mean and the time period effect parameters, indicating further model refinement for increased accuracy.

Along with these changes, future efforts could include incorporating additional predictors such as healthcare infrastructure, demographic details, and vaccine hesitancy, to capture the complex factors influencing vaccine uptake. Also, extending the data after 2021 to assess the robustness of the findings and provide more generalizable insights.

The evidence suggests the pneumococcal vaccination campaign had mixed effectiveness, with an overall decrease in vaccination rates during and after the COVID-19 pandemic, but significant regional variation in outcomes. The model highlights the importance of targeted interventions in under-performing regions and emphasizes the need for future campaigns to address disruptions caused by crises. By refining public health strategies, this work provides valuable insights for improving vaccination outreach and ensuring sustained coverage in the future.

6. Supplemental Details and Materials

0.0.1 Figures

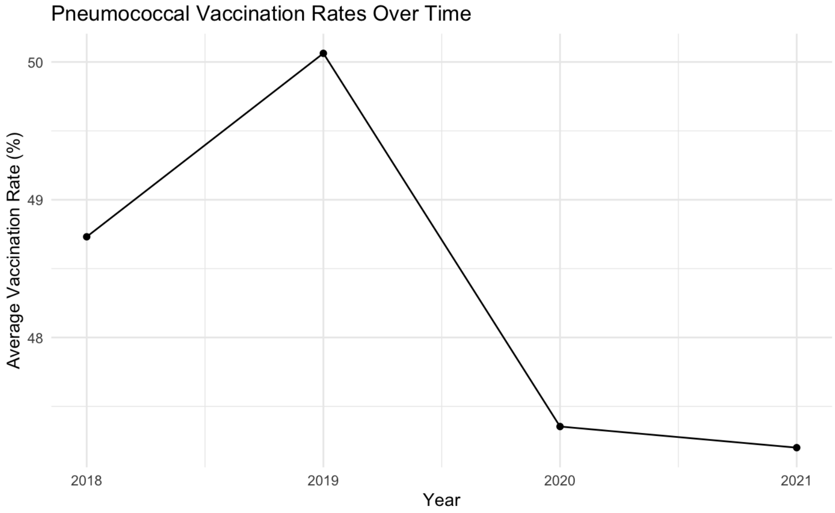


Figure 1: Trend of Pneumococcal Vaccination 2018-2021

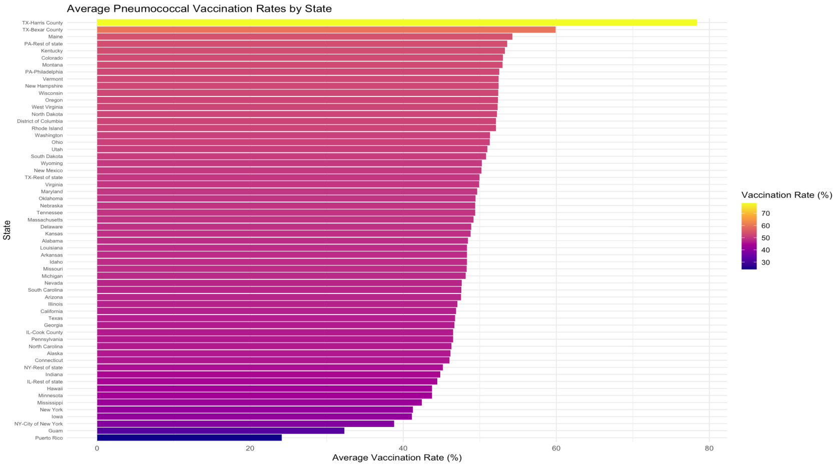


Figure 2: Pneumococcal Vaccination Rates by State

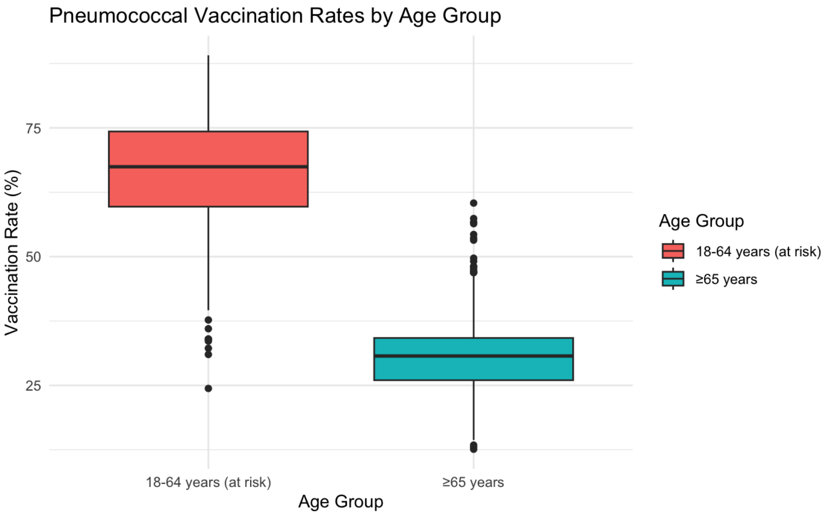


Figure 3: Age Group Variation in Vaccination Rates

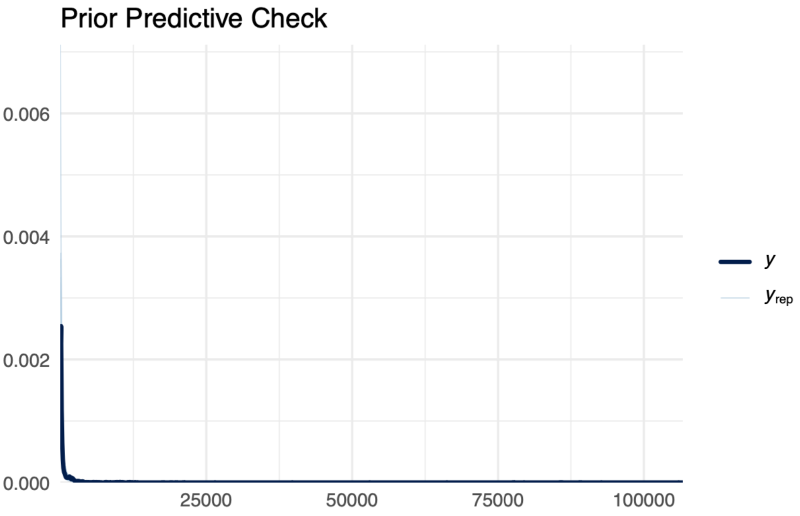


Figure 4: Prior Predictive Checking

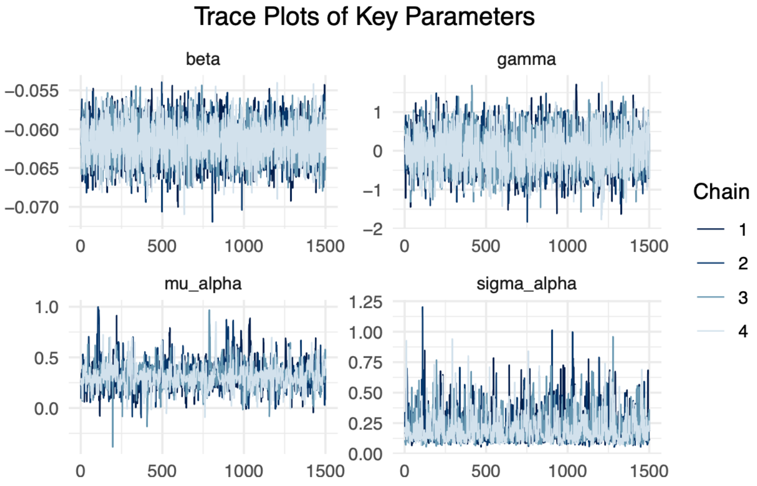


Figure 5: Trace Plots for Parameters

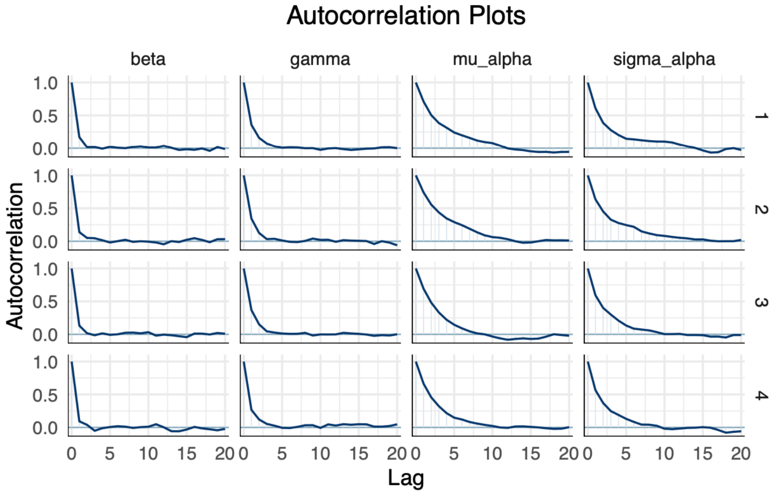


Figure 6: Autocorrelation Plots for Parameters

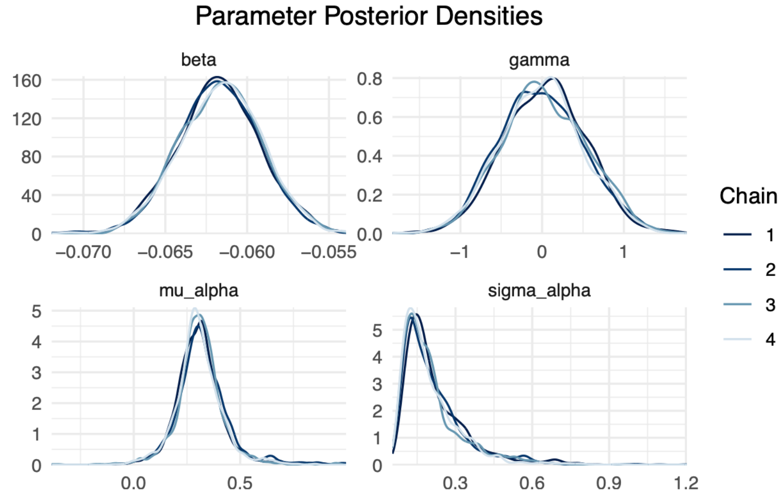


Figure 7: Parameter Posterior Densities

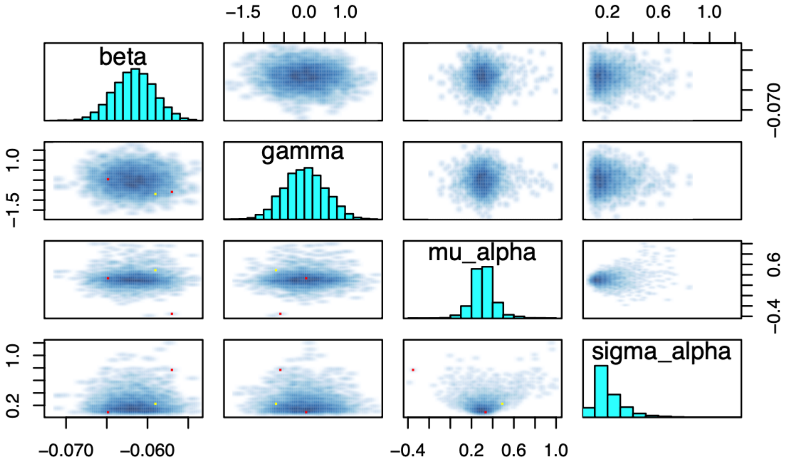


Figure 8: Parameter Pairs Plots

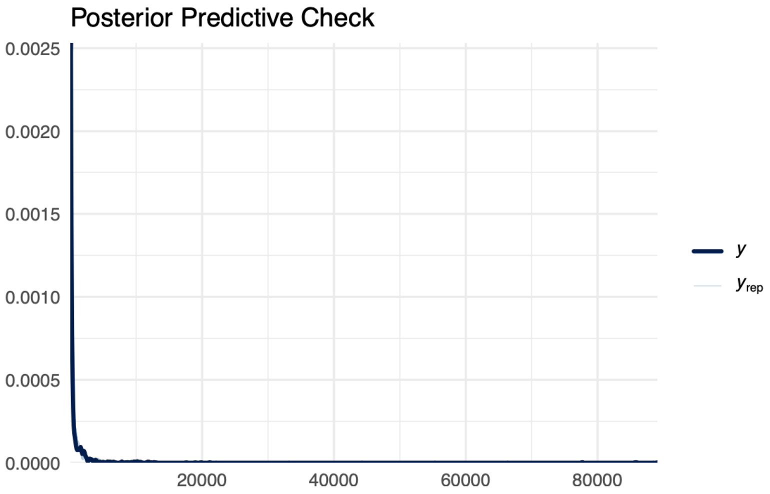


Figure 9: Posterior Predictive Check

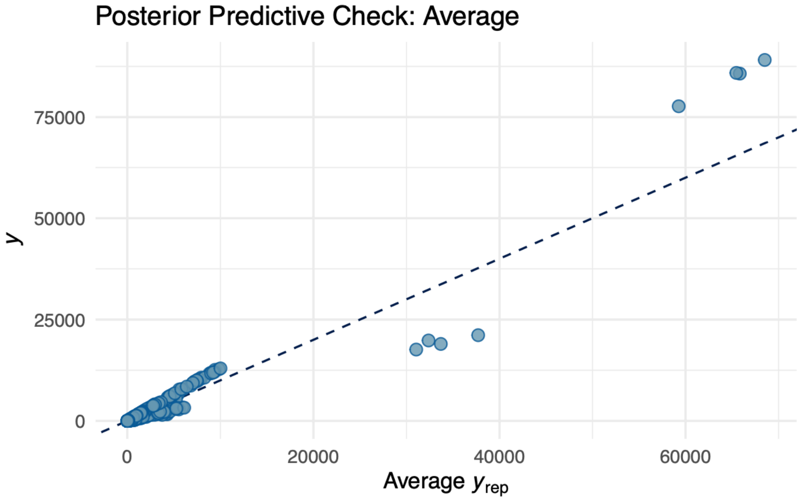


Figure 10: Posterior Predictive Check (Average)

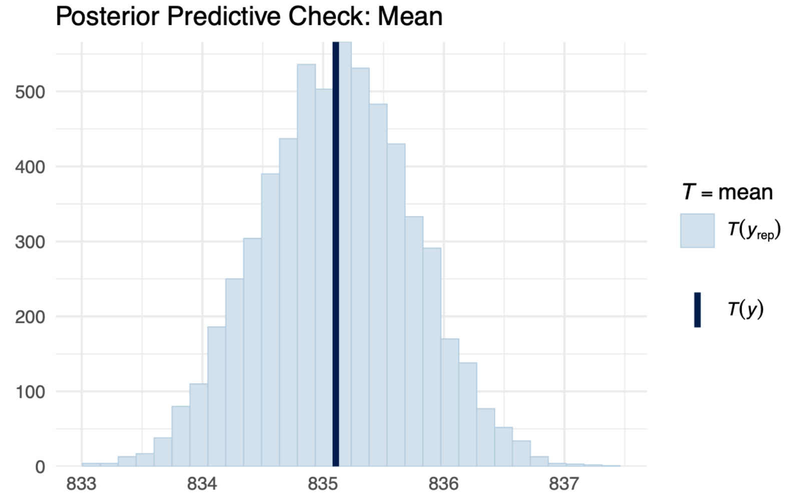


Figure 11: Posterior Predictive Check (Tstat vs. Predictive)

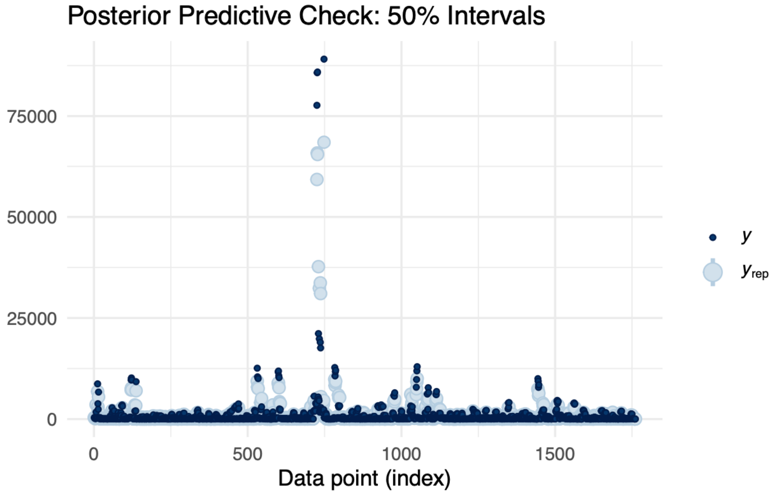


Figure 12: Posterior Predictive Check (50% Intervals)

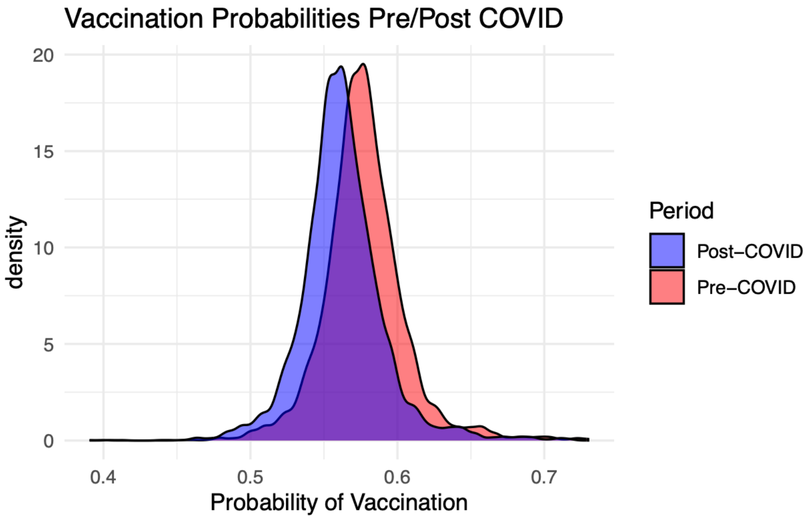


Figure 13: Campaign Effectiveness (Pre/Post COVID-19)

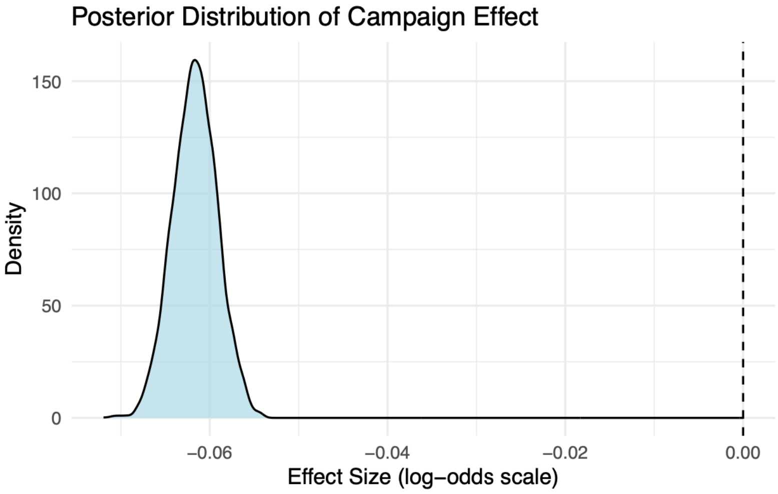
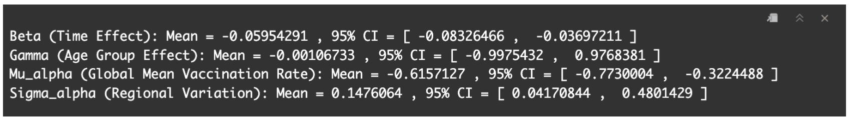


Figure 14: Campaign Effectiveness (Posterior)



```
Beta (Time Effect): Mean = -0.05954291 , 95% CI = [ -0.08326466 ,  -0.03697211 ]
Gamma (Age Group Effect): Mean = -0.00106733 , 95% CI = [ -0.9975432 ,  0.9768381 ]
Mu_alpha (Global Mean Vaccination Rate): Mean = -0.6157127 , 95% CI = [ -0.7730004 ,  -0.3224488 ]
Sigma_alpha (Regional Variation): Mean = 0.1476064 , 95% CI = [ 0.04170844 ,  0.4801429 ]
```

Figure 15: 95% Credible Intervals for Parameters

0.0.2 References

1. World Health Organization. "A Brief History of Vaccination." WHO, https://www.who.int/news-room/spotlight/history-of-vaccination/a-brief-history-of-vaccination.

2. Poland, Gregory A., et al. "The Age-Old Struggle against the Antivaccinationists." New England Journal of Medicine, vol. 364, no. 2, 2011, pp. 97–99. PubMed, https://pubmed.ncbi.nlm.nih.gov/23444591/

3. Centers for Disease Control and Prevention. "AdultVaxView: General Population." CDC, https://www.cdc.gov/adultvaxview/about/general-population.html.

4. R Core Team. "Tutorial 100: Posterior Predictive Checks." isotracer: R Package Documentation, CRAN, https://cran.r-project.org/web/packages/isotracer/vignettes/tutorial-100-posterior-predictive-checks.html

5. R Core Team. "Logistic." R Documentation, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Logistic.

6. UCLA Statistical Consulting Group. "Logistic Regression in R." UCLA Institute for Digital Research and Education, https://stats.oarc.ucla.edu/r/dae/logit-regression/.

7. OpenAI. "ChatGPT." ChatGPT by OpenAI, OpenAI, chat.openai.com (general sentence refinement conciseness to fit report in the 2-3 page limit)

0.0.3 Code

0.1 Appendix

1. Likelihood (Observation Level)

For each observation i:

$$P(y_i \mid p_i, \text{total}_i) = \text{Binomial}(y_i \mid \text{total}_i, p_i),$$

where the logit of p_i is modeled as:

$$\text{logit}(p_i) = \alpha[\text{region}_i] + \beta \cdot \text{time\_period}_i + \gamma \cdot (\text{age\_group}_i - 1).$$

Thus, the probability of y_i depends on:

- $\alpha[\text{region}_i]$: Region-specific baseline vaccination rate.
- β: Effect of time period.
- γ: Effect of age group.

2. Region-Specific Effects

For each region $r = 1, \ldots, R$:

$$P(\alpha[r] \mid \mu_\alpha, \sigma_\alpha) = \text{Normal}(\alpha[r] \mid \mu_\alpha, \sigma_\alpha).$$

Here, $\alpha[r]$ depends on:

- μ_α: Global mean baseline vaccination rate.
- σ_α: Standard deviation of vaccination rates across regions.

3. Priors on Global Parameters

The global parameters have the following priors:

$$P(\mu_\alpha) = \text{Normal}(\mu_\alpha \mid \text{logit}(0.65), 0.5),$$

$$P(\sigma_\alpha) = \text{Normal}(\sigma_\alpha \mid 0, 0.5),$$
$$P(\beta) = \text{Normal}(\beta \mid 0, 0.5),$$
$$P(\gamma) = \text{Normal}(\gamma \mid 0, 0.5).$$

4. Full Conditional Model Specification

The joint probability of the data $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ and parameters is:

$$P(\mathbf{y}, \mathbf{p}, \boldsymbol{\alpha}, \mu_\alpha, \sigma_\alpha, \beta, \gamma) = P(\mathbf{y} \mid \mathbf{p}) \cdot P(\mathbf{p} \mid \boldsymbol{\alpha}, \beta, \gamma) \cdot P(\boldsymbol{\alpha} \mid \mu_\alpha, \sigma_\alpha) \cdot P(\mu_\alpha) \cdot P(\sigma_\alpha) \cdot P(\beta) \cdot P(\gamma),$$

where:

- **Data likelihood:**
$$P(\mathbf{y} \mid \mathbf{p}) = \prod_{i=1}^{N} \text{Binomial}(y_i \mid \text{total}_i, p_i).$$
- **Probability model for p_i:**
$$P(\mathbf{p} \mid \boldsymbol{\alpha}, \beta, \gamma) = \prod_{i=1}^{N} \delta \left(p_i - \text{logit}^{-1}(\alpha[\text{region}_i] + \beta \cdot \text{time\_period}_i + \gamma \cdot (\text{age\_group}_i - 1)) \right).$$
- **Region-specific effects:**
$$P(\boldsymbol{\alpha} \mid \mu_\alpha, \sigma_\alpha) = \prod_{r=1}^{R} \text{Normal}(\alpha[r] \mid \mu_\alpha, \sigma_\alpha).$$
- **Global priors:**
$$P(\mu_\alpha) \cdot P(\sigma_\alpha) \cdot P(\beta) \cdot P(\gamma).$$

5. Posterior Distribution

Using Bayes' theorem, the posterior distribution is:

$$P(\mu_\alpha, \sigma_\alpha, \boldsymbol{\alpha}, \beta, \gamma \mid \mathbf{y}) \propto P(\mathbf{y} \mid \mu_\alpha, \sigma_\alpha, \boldsymbol{\alpha}, \beta, \gamma) \cdot P(\mu_\alpha) \cdot P(\sigma_\alpha) \cdot P(\beta) \cdot P(\gamma),$$

where the likelihood is:

$$P(\mathbf{y} \mid \mu_\alpha, \sigma_\alpha, \boldsymbol{\alpha}, \beta, \gamma) = \prod_{i=1}^{N} \text{Binomial} \left(y_i \mid \text{total}_i, \text{logit}^{-1}(\alpha[\text{region}_i] + \beta \cdot \text{time\_period}_i + \gamma \cdot (\text{age\_group}_i - 1)\right.$$