Exploring Cross Immunity of Respiratory Syncytial Virus (RSV) Subtypes

Abstract

RSV is a common respiratory infection with two co-circulating subtypes (A and B) whose cross-immunity is poorly understood. Using a partially observed Markov process (POMP) framework, we fit a two-strain stochastic SEIRS model with seasonal forcing and a cross-immunity parameter q, estimated via iterated filtering and evaluated with profile likelihoods. A negative binomial autoregressive benchmark achieves log-likelihood -1158, while the mechanistic model attains comparable fit (-1159) with well-identified transmission, reporting, and dispersion parameters. However, the initial susceptibility η and cross-immunity parameter q are weakly identified, with flat profile likelihoods. These results indicate that the available data are insufficient to distinguish between competing cross-immunity structures.

1 Introduction

Respiratory Syncytial Virus (RSV) is a highly contagious, seasonal respiratory virus that spreads through respiratory droplets and contaminated surfaces (Mayo Clinic 2024; U.S. FDA 2023). The widespread infection is estimated to cause up to 250,000 hospitalizations among infants and older adults annually, with infections peaking in the winter months (CDC 2023a, 2023b, 2024). Most children will get the infection at least once by the time they are 2 years old, and the mean age of first infection is 6 months old (CDC 2023b). However, unlike other respiratory infections, RSV presents like a common cold in healthy adults whereas young children and immunocompromised older adults suffer more severe symptoms that may ultimately lead them to patient care facilities such as clinics and hospitals.

Testing for RSV is not routinely performed in many hospitalized patients outside of pediatric and other high-risk settings (Prill et al. 2021). Currently, the Center for Disease Control and Prevention (CDC) tracks RSV using reports from participating labs, but healthcare facilities are not required to report positive RSV cases to federal authorities (Prill et al. 2021). Despite low testing rates, cases continue to surge. Like influenza, RSV has two prevalent subtypes, namely, subtypes A and B. However, RSV subtypes are not commonly identified when reported, and most surveillance systems only report aggregate RSV cases. Prior RSV infection with one subtype appears to confer partial but not complete protection against the other, though the degree of cross-immunity remains uncertain (Nuttens et al. 2024).

In this report, the partially observed Markov process (POMP) framework is used to account for underreporting and transmission dynamics between the subtypes. While re-infection is possible, most published compartmental models do not consider whether waning immunity for one strain completely

or partially protects an individual from infection from the other (Nuttens et al. 2024). Thus, the purpose of this project is to assess whether assumptions of full, partial, or no cross-immunity best capture RSV A and B transmission dynamics, and we test these different immunity structures within a POMP framework.

1.1 Data

Subtype-specific RSV time series were obtained from Nextstrain (Nextstrain 2025a, 2025b). Nextstrain is an open-source platform that aggregates publicly available genomic sequence data from the National Center for Biotechnology Information (NCBI), the Global Initiative on Sharing All Influenza Data (GISAID), and the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) at the University of Chicago, among other sources (Hadfield et al. 2018). Each record corresponds to a sequenced RSV sample with an associated collection date and subtype classification (A or B).

Weekly sequence counts from June 1, 2015 to March 15, 2020 (250 weeks) were used to avoid disruptions from COVID-19 interventions. Earlier records were excluded due to low counts. This time frame also precedes changes in RSV reporting and vaccine availability. As shown in Figure 1, both subtypes exhibit annual peaks in December–January, with weekly counts below 100.

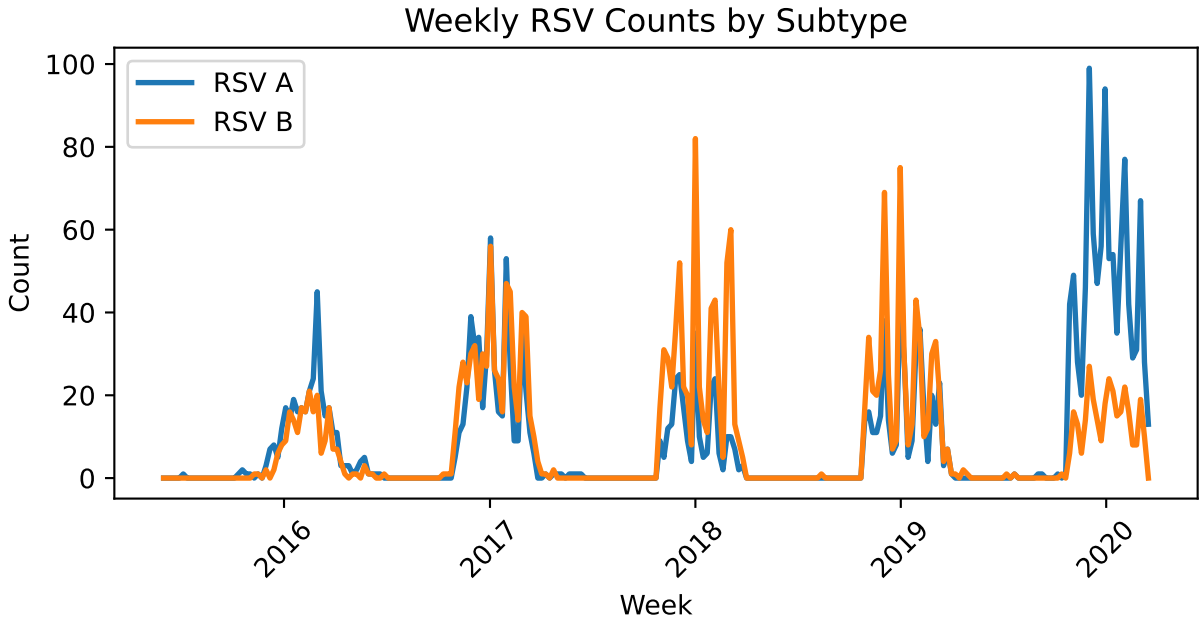


Figure 1: RSV data plot

2 Data limitations

Although RSV surveillance varies across states, national co-circulation patterns of RSV-A and RSV-B are consistent across regions (CDC 2024), and exploratory analysis confirmed minimal differences in seasonal timing across states (more details in supplementary, Figure 7, Figure 8, Figure 9, Figure 10). We therefore treat the United States as a single homogeneous population.

Sequences are primarily derived from clinical testing and likely reflect more severe cases among high-risk populations (infants, older adults, and immunocompromised individuals) (MedlinePlus 2024). We make the simplifying assumption that the sequenced population represents this single high-risk pool of individuals. We also do not stratify by age due to limited metadata and to prevent increased model complexity. Furthermore, while previous studies aimed to identify meaningful differences in disease severity across the subtypes, there is limited evidence that either subtype results in more severe illness (Nuttens et al. 2024). Therefore, we assume that the relative frequency of RSV-A and RSV-B in the sequenced sample reflects the true subtype composition of RSV in the population.

3 Methods

3.1 Statistical Benchmark Models

Initially, we fit a SARIMAX model to serve as a benchmark for comparison to a mechanistic POMP model. However, this initial model highlighted several limitations of the linear-Gaussian framework for count data. In particular, the QQ plot shows heavy right tails, the Jarque-Bera test strongly rejects normality (JB = 755, < 0.001), and the residuals display clear heteroskedasticity, especially during seasonal peaks (see Figure 11 in supplementary). Although a log transformation could potentially address some of these issues, we avoided it to maintain consistency when comparing with our local and global search methods.

Overall, these diagnostics strongly suggest that a count-based model with overdispersion would be more appropriate. To address this, we instead use a negative binomial autoregressive benchmark models fit separately for subtype $i \in \{A, B\}$ with counts Y_{it} satisfying

$$Y_{it} \sim \text{NegBin}(\mu_{it}, r), \quad \mu_{it} = \beta_0 + \beta_1 Y_{i,t-1}, \quad \text{Var}(Y_{it}) = \mu_{it} + \frac{\mu_{it}^2}{r}$$

where $r > 0$ is a dispersion parameter. μ_{it} represents the conditional mean, which is the expected count for subtype i at week t, given the previous week's observed count. $\beta_{0,i}$ expresses a baseline level, and $\beta_{1,i}$ controls how much last week's count carries into this week's prediction.

This benchmark captures overdispersion and short-term dependence in the data without imposing mechanistic assumptions about transmission or immunity. As a result, it serves as a useful point of comparison for the mechanistic POMP models introduced later.

3.2 Mechanistic Transmission Models

To incorporate known biological structure of RSV, such as latent and infected period, we fit stochastic compartmental models for subtypes A and B with a partially observed Markov process (POMP) framework. These models represent latent infection dynamics and allow direct epidemiological interpretation.

Because RSV infection includes a latent period and temporary immunity, we use an SEIRS-type structure. While many different models exist in current academic literature on RSV, we use a model similar to those shown in Bhattacharyya et al. (2015) and White et al. (2007), but simplified for the purposes of this course project. Individuals transition through susceptible (S), exposed (E),

infectious (I), and recovered (R) compartments. Infection occurs at rate $\lambda(t)$, exposed individuals progress at rate σ, infectious individuals recover at rate γ, and immunity wanes at rate ω.

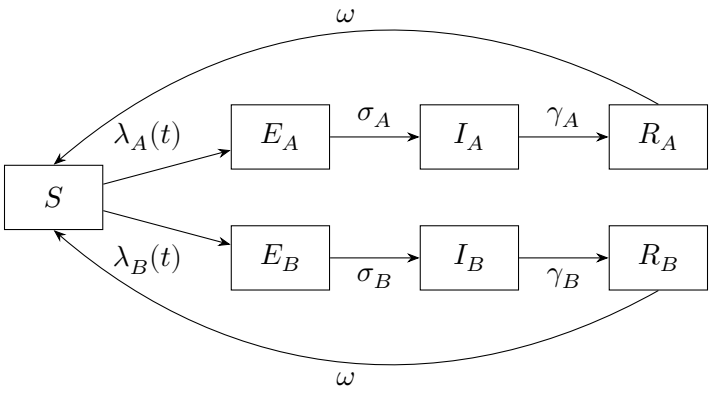


Figure 2: Two-strain SEIRS model for RSV transmission.

Cross-strain interaction is incorporated through the force of infection, with infection rates depending on both within-strain and cross-strain infectious individuals:

$$\lambda_A(t) = \beta_A \frac{I_A(t)+(1-q)I_B(t)}{N}, \quad \lambda_B(t) = \beta_B \frac{I_B(t)+(1-q)I_A(t)}{N}$$

Here, $q \in [0,1]$ represents the degree of cross-immunity between strains, with $q = 0$ corresponding to no cross-immunity and $q = 1$ corresponding to complete cross-immunity. In principle, q can be estimated from the data to capture intermediate levels of interaction. Considering the $q = 0$ case provides a baseline in which strains do not interact through immunity, allowing us to assess whether the data contain evidence for cross-strain effects beyond independent transmission dynamics.

The latent and infectious periods are represented by the transition rates σ and γ. These correspond to average durations of 5 days for the exposure period and 7 days for the infection period. These values were fixed based on well known trends in the epidemiological literature (Negi et al. 2024; Lessler et al. 2009). By fixing σ and γ, the goal is to improve parameter identifiability and stabilize inference for the remaining parameters of interest.

Waning immunity occurs at rate ω, corresponding to an average duration of immunity of six months (Moore et al. 2014), allowing individuals to return to the susceptible class. The parameter N represents an effective population size that scales transmission intensity and does not necessarily correspond to the total population size. The initial fraction susceptible is denoted by η.

3.3 Model Implementation and Comparison

Observed weekly RSV case counts are linked to latent incidence through a negative binomial measurement model. The expected number of reported cases is proportional to the underlying incidence, with reporting governed by a parameter ρ. An adidtional ρ_{slope} term allows the reporting rate to increase linearly by 3% each year to capture gradual improvement in RSV surveillance.

Overdispersion is captured by a size parameter k, allowing the model to accommodate variability beyond the Poisson assumption.

Parameter estimation is performed using likelihood-based methods for β_A, β_B, η, ρ, k, and q. Epidemiological rates σ, γ, and ω were fixed based on values from the literature (Negi et al. 2024; Lessler et al. 2009; Moore et al. 2014). A complete summary of model parameters and their interpretations is provided in Table 2 in the appendix. A local search based on iterated filtering is first used to refine parameter estimates around reasonable starting values and assess convergence behavior. This is followed by a global search over multiple randomized initializations to explore the likelihood surface more thoroughly. Parameter uncertainty and identifiability are assessed using profile likelihoods, constructed by fixing one parameter over a grid while optimizing over the remaining parameters.

We compare the negative binomial benchmark and mechanistic models using likelihood-based criteria. In addition, we evaluate differences in model performance when q is fixed to 0 and when q is allowed to vary to provide insight into the role of cross-strain interaction in shaping RSV transmission dynamics. Likelihood estimates are subject to Monte Carlo variance inherent to particle filter methods; standard errors were estimated by running multiple independent particle filter evaluations at the reported parameter values. We initially explored an extension of the model incorporating full cross-immunity between RSV-A and RSV-B. However, optimization of this specification was highly unstable, with local search trajectories exhibiting large fluctuations in log-likelihood and failing to converge to a consistent region of the parameter space. As a result, we only use the no-immunity model as a simpler and more stable baseline for comparison to the partial immunity model.

4 Results

Starting with the benchmark negative binomial model, the fitted model achieved a log-likelihood of -1158.58 (estimated parameters in Table 3 and diagnostics in Figure 12). Despite not including any mechanistic structure like POMP models, it can still capture key features of the data, including overdispersion and short-term autocorrelation.

We begin by examining the results of the local search procedure. The log-likelihood traces in Figure 3 exhibit substantial variability across replicates, with no clear stabilization by the end of the MIF iterations. This reflects Monte Carlo noise in the likelihood evaluation and the difficulty of optimizing over a complex surface. The estimated standard error of 3.7 for the best log-likelihood indicates that small differences between candidate parameter values are not meaningful.

The best-performing local search run achieves a log-likelihood of $-1{,}164.8$. Despite the overall variability, several parameters (β_A, β_B, k, ρ_0) concentrate within relatively narrow ranges across runs, suggesting that transmission intensity and measurement variability are reasonably well identified. In contrast, η exhibits weaker convergence, indicating limited information about the size of the susceptible population or a poorly chosen initial value. The parameter q shows substantial instability, with estimates spanning nearly the entire parameter space, suggesting that it is not identifiable from the data.

The no immunity model (fixed $q = 0$) local search results are provided in supplementary (Figure 13). The best log likelihood obtained was -1,162.4, which is very similar to results achieved from the

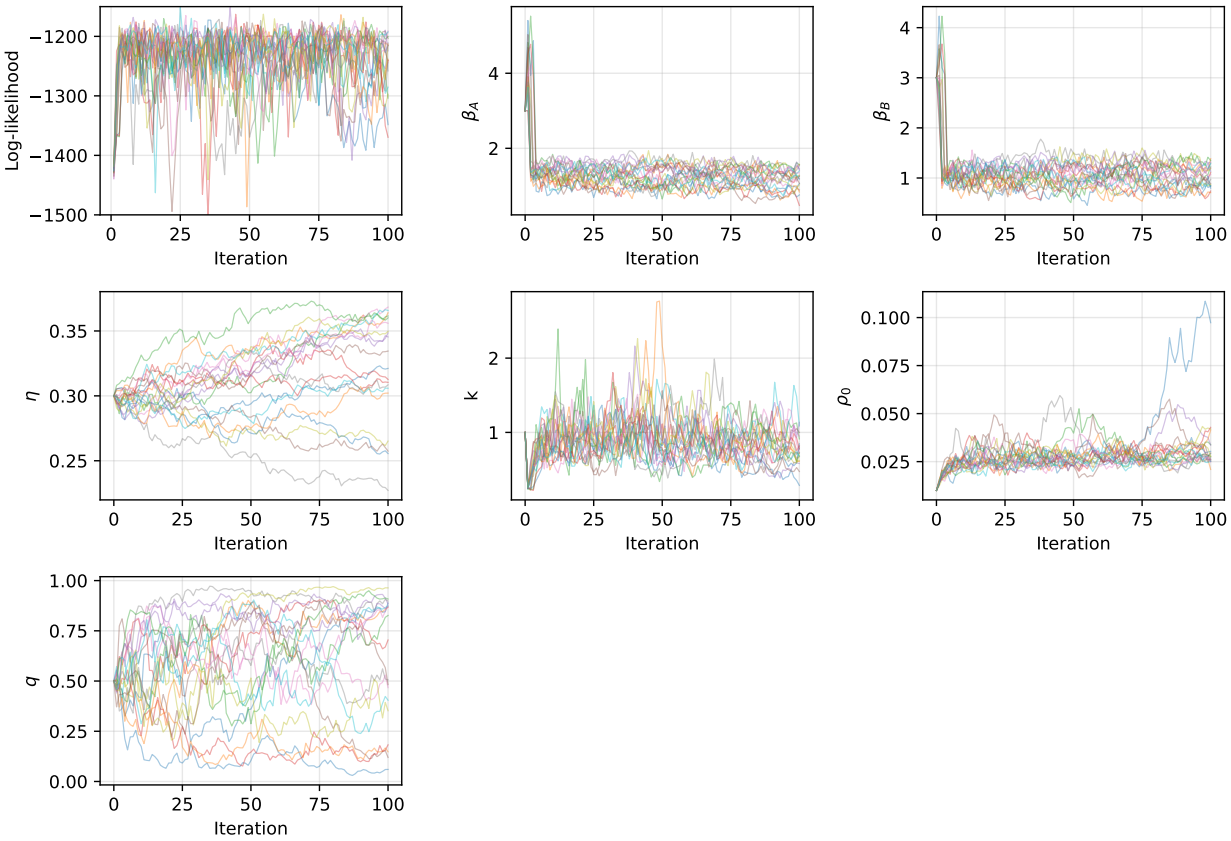


Figure 3: Local search results

partial model. In addition, convergence results are nearly identical to those from the partial immunity model, with the parameter η still showing weak identifiability and the remaining parameters showing convergence around the same narrow ranges.

The global search results in Figure 4 reveal a well-defined region of high likelihood, with multiple parameter combinations achieving similar values around -1159.3. These results indicate that the likelihood surface is relatively flat in several dimensions, limiting the ability to distinguish between alternative parameter combinations.

Within this high-likelihood region, the transmission parameters β_A and β_B favor relatively small values (approximately 1–2) and exhibit similar patterns, suggesting that the two strains have comparable transmission dynamics. The initial susceptible fraction η is concentrated in a moderate to high range ($\eta \approx 0.45$–0.85), indicating that a relatively large susceptible pool is required to reproduce the observed epidemic patterns.

In contrast, the reporting rate (ρ_0) and dispersion parameter (k) are tightly concentrated, indicating strong identifiability. The estimates of ρ_0 (approximately 0.024–0.028) suggest a low reporting rate, while the narrow range of k (approximately 0.25–0.29) reflects substantial overdispersion in the observed counts.

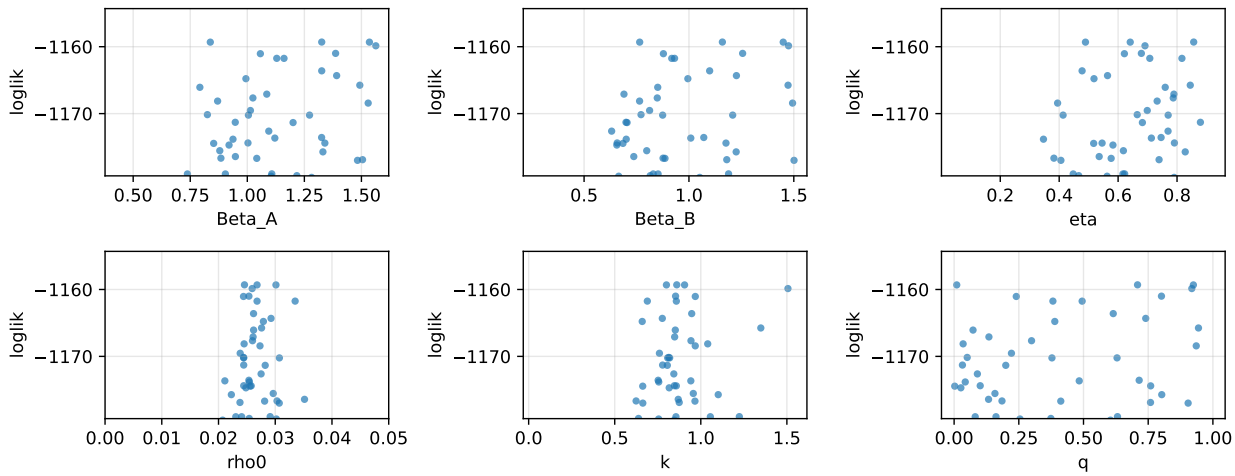


Figure 4: Global search results near the maximum log-likelihood.

Pairwise parameter plots (Supplementary, Figure 14) reveal a clear trade-off between the transmission parameters (β_A, β_B) and the initial susceptible fraction (η), with higher transmission rates corresponding to lower values of η. This indicates that different combinations of transmission intensity and susceptible population size can produce similar epidemic trajectories, limiting the ability to uniquely identify these parameters.

Table 1: Top 10 Largest Log Likelihoods Obtained via Global Search

	loglik	Beta_A	Beta_B	eta	rho0	q	k
59	-1159.304416	1.326057	1.159446	0.856568	0.024575	0.708816	0.860379
82	-1159.317595	0.837899	0.765141	0.641148	0.030124	0.010159	0.905711
186	-1159.317659	1.535012	1.449495	0.489229	0.026795	0.924627	0.800338
73	-1159.862743	1.563039	1.474870	0.691249	0.025933	0.918524	1.506210
4	-1160.997856	1.386863	1.256022	0.678930	0.025359	0.801276	0.853379
188	-1161.044439	1.057258	0.878206	0.621340	0.024404	0.240182	0.967946
176	-1161.738168	1.129022	0.931432	0.707559	0.033506	0.381769	0.689365
15	-1161.744113	1.160809	0.918047	0.816626	0.026794	0.494714	0.857864
161	-1163.606558	1.326069	1.098726	0.477760	0.026174	0.614916	0.947163
157	-1164.320474	1.391450	1.226411	0.564096	0.029260	0.740454	0.776399

Table 1 shows the parameters from the ten best models obtained via the global search. The parameter q exhibits substantial variability across top-performing models, ranging from near 0 to above 0.9 with minimal change in log likelihood. In addition, extremely similar results were observed in the no immunity model (see supplementary Figure 15 and Table 4). The best log-likelihood from the no-immunity global search (-1155.9) exceeds that of the partial immunity model. However, given substantial Monte Carlo uncertainty in likelihood estimates, this difference is not meaningful and further reinforces the conclusion that the two immunity structures are indistinguishable from

the data. Together with the observed trade-offs among other parameters, these results indicate a ridge-like likelihood surface in which multiple parameter combinations yield similarly good fits, while immunity-related parameters remain poorly identified.

The profile likelihood results in Figure 5 provide an assessment of parameter identifiability. Due to Monte Carlo noise in the likelihood estimates, formal 95% profile-likelihood cutoffs were unstable and are not emphasized. The plots are interpreted qualitatively as diagnostics of parameter identifiability. The transmission parameters β_A and β_B are moderately well identified, with clear curvature indicating a preferred range of values. The reporting rate ρ_0 is also well identified, exhibiting a concentrated peak in the profile likelihood.

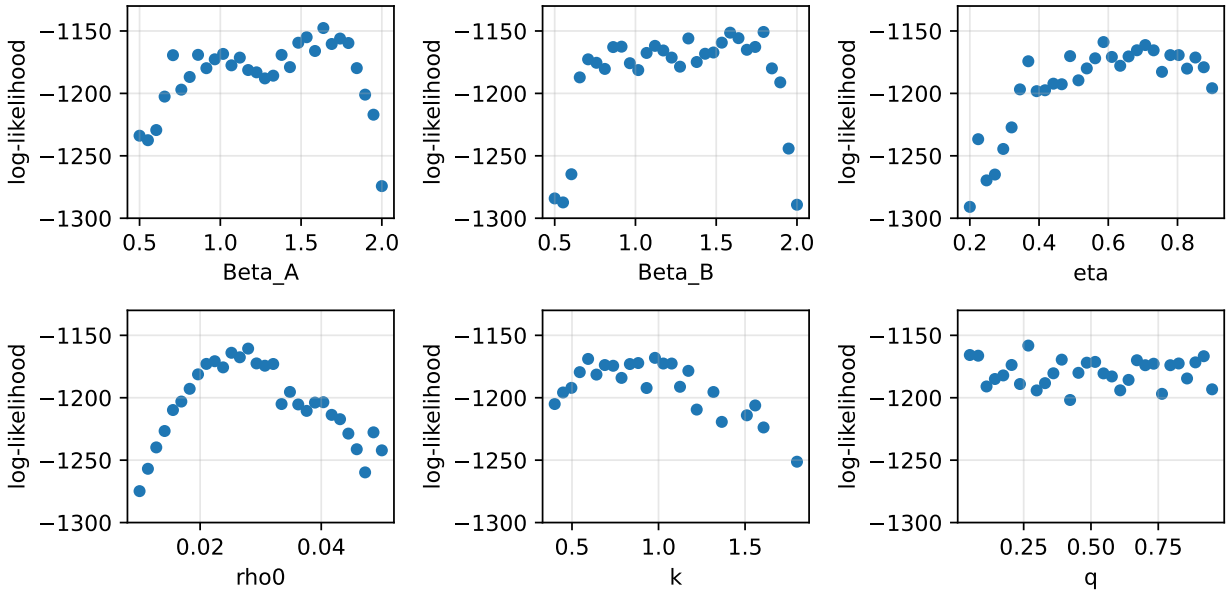


Figure 5: Profile likelihood plots

In contrast, the initial susceptible fraction η displays a flatter profile, with lower values clearly disfavored but a broad range of higher values remaining plausible, indicating weaker identifiability. The dispersion parameter k exhibits a relatively flat and noisy profile, suggesting that it is only weakly identified.

The parameter q exhibits a nearly flat profile likelihood, confirming that it is not identifiable from the data. This indicates that multiple values of q produce similar likelihoods, consistent with the global search results. The lack of identifiability is likely driven by the limited temporal span of the data and relatively low case counts, which reduce the ability to distinguish between competing cross-immunity structures.

The simulated trajectories in Figure 6 capture the overall seasonal timing and recurrence of RSV outbreaks, but tend to overestimate peak magnitudes and exhibit greater variability than observed data. This suggests that while the model captures the timing of transmission dynamics, it may over-concentrate infections during peak periods. The overestimation of peak magnitudes may partly

reflect uncertainty in the effective population size N. Because the observed counts represent test results submitted by clinical laboratories rather than true case counts, N does not correspond to any well-defined population. Without external information to anchor a choice for N, the model may generate epidemic peaks that are too large relative to the observed sequence counts. This ambiguity is distinct from the reporting rate, ρ, which also absorbs some of this uncertainty but cannot fully resolve the identifiability problem.

The negative binomial benchmark achieved a log-likelihood of -1158, while the POMP model equivalent values around -1159. This suggests that the mechanistic model is able to capture key features of the data and biological reality without sacrificing model fit. However, the POMP model still exhibits some mismatch to the observed patterns, including overestimation of peak magnitudes.

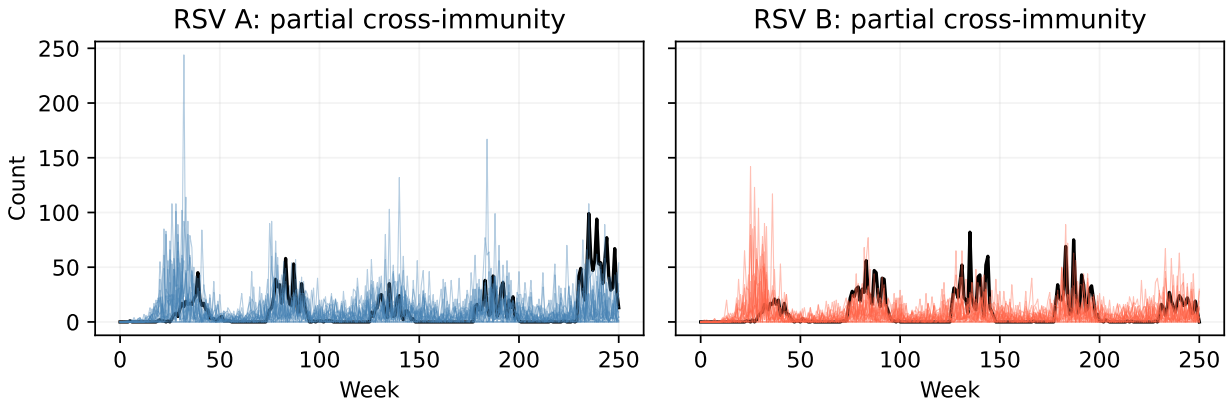


Figure 6: RSV Count Simulations From Best Partial Cross Immunity Model

5 Conclusions

This analysis evaluated a mechanistic model of RSV transmission to assess whether the available data can inform key epidemiological parameters. The results indicate that several parameters, including the transmission rates, reporting rate, and dispersion parameter, are consistently estimated within narrow ranges, suggesting that core aspects of transmission and observation are well identified.

In contrast, parameters governing susceptibility and cross-immunity exhibit substantial uncertainty. In particular, the cross-immunity parameter q displays a flat profile likelihood and fails to converge in local searches, providing strong evidence that it is not identifiable from the available data. The difference in log-likelihood between the partial immunity model (-1159.3) and the no-immunity model (-1162.4) is approximately 3 units, confirming that the data cannot distinguish between these competing immunity structures. Other scientific studies have also failed to identify the degree of cross immunity through models, but biological studies indicate that similarities in the structure of the subtypes could allow for at least partial cross immunity (Nuttens et al. 2024). These results suggest that the dataset lacks sufficient information to distinguish between competing biological assumptions about immunity. Multiple combinations of transmission dynamics and susceptibility

structure can produce similar observed epidemic patterns, particularly in short time series with relatively low counts.

This analysis is subject to several limitations. The relatively short duration of the time series (approximately five years) restricts the ability to estimate parameters governing longer-term dynamics, particularly those related to immunity and reinfection. Additionally, observed case counts are likely subject to underreporting and noise, reducing the signal available for inference. The combination of model complexity and limited data contributes to parameter trade-offs and weak identifiability, particularly for parameters governing cross-immunity. Finally, ikelihood estimates are additionally subject to Monte Carlo variance inherent to particle filter methods, which limits precision when comparing models with similar fits. Future work incorporating longer time series, higher-resolution data, or additional sources of information may improve identifiability and allow for more precise characterization of immunity dynamics in RSV transmission.

6 Scholarship and AI

This project builds on previous STATS 531 infectious diseases projects. We were especially inspired by a previous project on COVID-19 cases in Washtenaw county (*An Analysis of COVID-19 Cases in Washtenaw County* 2021) and another project in Kent county (*Time Series Analysis of COVID-19 Cases in Kent County* 2024). Both used a stochastic SEIR model, particle filtering, global search, profile likelihood, and ARMA likelihood benchmarks. The Washtenaw county project shaped our general workflow and showed the importance of not assuming that a mechanistic model is automatically better; their ARMA benchmark achieved a higher likelihood than the SEIR model, suggesting that the SEIR model missed high-frequency reporting structure such as weekly administrative cycles.

Our project extends this idea from a single-strain epidemic curve to a two-subtype RSV setting. While many prior projects use diseases with similar dynamics (Covid-19, Influenza), none appear to consider interactions between strains of disease subtypes. Our project differs from prior projects as our main scientific goal is to assess whether the data support cross-strain immunity between RSV-A and RSV-B through the parameter q. Like the prior COVID-19 project in Washtenaw County, we first considered simple likelihood benchmarks, including ARMA-type models and negative binomial count models.

We also learned from peer review of prior 531 projects that visual fit alone is not enough evidence for a successful model, and that weakly identified parameters should be investigated rather than only reported as estimates. This motivated our use of profile likelihood for q and our discussion of model limitations, including subtype-specific reporting bias, changing sequencing intensity, age structure, and immune waning.

AI tools, including Claude and ChatGPT were utilized primarily to help debug code, optimize the figure sizes included in the report, and occassionally improve wording in the report. All AI-suggested code was tested to ensure results made sense before being used in the final report. As a team, we relied heavily on course notes and academic papers to inform our model choices, results, citations, and final interpretations. AI was never used as a scientific authority.

Bibliography

An Analysis of COVID-19 Cases in Washtenaw County. 2021. https://ionides.github.io/531w21/final_project/project15/blinded.html.

Bhattacharyya, Samit, Per H. Gesteland, Kent Korgenski, Ottar N. Bjørnstad, and Frederick R. Adler. 2015. "Cross-Immunity Between Strains Explains the Dynamical Pattern of Paramyxoviruses." *Proceedings of the National Academy of Sciences* 112 (43): 13396–400. https://doi.org/10.1073/pnas.1516698112.

CDC. 2023a. *Disease Severity of Respiratory Syncytial Virus Compared with COVID-19 and Influenza Among Hospitalized Adults*. https://www.cdc.gov/mmwr/volumes/72/wr/mm7240a2.htm.

CDC. 2023b. *RSV in Infants and Young Children*. https://www.cdc.gov/rsv/infants-young-children/index.html.

CDC. 2024. *RSV Surveillance Data and Seasonality*. https://www.cdc.gov/rsv/php/surveillance/index.html.

Hadfield, James, Colin Megill, Sidney M Bell, et al. 2018. "Nextstrain: Real-Time Tracking of Pathogen Evolution." *Bioinformatics* 34 (23): 4121–23. https://doi.org/10.1093/bioinformatics/bty407.

Ionides, Edward. 2026. *Notes for STATS 531, Modeling and Analysis of Time Series Data*. https://ionides.github.io/531w26/.

Lessler, Justin, Nicholas G. Reich, Ron Brookmeyer, Trish M. Perl, Kenrad E. Nelson, and Derek A. T. Cummings. 2009. "Incubation Periods of Acute Respiratory Viral Infections: A Systematic Review." *The Lancet Infectious Diseases* 9 (5): 291–300. https://doi.org/10.1016/S1473-3099(09)70069-6.

Mayo Clinic. 2024. *Respiratory Syncytial Virus (RSV) - Symptoms and Causes*. https://www.mayoclinic.org/diseases-conditions/respiratory-syncytial-virus/symptoms-causes/syc-20353098.

MedlinePlus. 2024. *Respiratory Syncytial Virus (RSV) Tests*. https://medlineplus.gov/lab-tests/respiratory-syncytial-virus-rsv-tests/.

Moore, Hannah C., Peter Jacoby, Alexandra B. Hogan, Christopher C. Blyth, and Geoffry N. Mercer. 2014. "Modelling the Seasonal Epidemics of Respiratory Syncytial Virus in Young Children." *PLOS ONE* 9 (6): e100422. https://doi.org/10.1371/journal.pone.0100422.

Negi, Ruchita, Jonathan Shabab, and Frank Esper. 2024. "The RSV Roadmap in Children: Lessons Learned, Paths Forward." *Cleveland Clinic Journal of Medicine* 91 (9 suppl 1): S19–25. https://doi.org/10.3949/ccjm.91.s1.04.

Nextstrain. 2025a. *RSV-a Phylogeny*. https://nextstrain.org/rsv/a/genome/6y.

Nextstrain. 2025b. *RSV-b Phylogeny*. https://nextstrain.org/rsv/b/genome/6y.

Nuttens, Charles, Juliette Moyersoen, Daniel Curcio, et al. 2024. "Differences Between RSV a and RSV b Subgroups and Implications for Pharmaceutical Preventive Measures." *Infectious Disease and Therapy*, ahead of print. https://doi.org/10.1007/s40121-024-01012-2.

Prill, Mila M., Gayle E. Langley, Amber Winn, and Susan I. Gerber. 2021. "Respiratory Syncytial Virus-Associated Deaths in the United States According to Death Certificate Data, 2005 to 2016." *Health Science Reports* 4 (4): e428. https://doi.org/https://doi.org/10.1002/hsr2.428.

Time Series Analysis of COVID-19 Cases in Kent County. 2024. https://ionides.github.io/531w24/final_project/project12/blinded.html.

U.S. FDA. 2023. *FDA Approves First RSV Vaccine*. https://www.fda.gov/news-events/press-announcements/fda-approves-first-respiratory-syncytial-virus-rsv-vaccine.

White, L. J., J. N. Mandl, M. G. M. Gomes, et al. 2007. "Understanding the Transmission Dynamics of Respiratory Syncytial Virus Using Multiple Time Series and Nested Models." *Mathematical Biosciences* 209 (1): 222–39. https://doi.org/10.1016/j.mbs.2006.08.018.

7 Supplementary material

7.1 Reproducibility

To implement the methods described above, code from lectures 13, 15, and 16 (Ionides 2026) was used as a base and modified for use with the new two strain SEIRS model described in greater detail in the main report. To ensure the cached results load without issues, use version 0.4.3 of pypomp and version 3.0.2 of pandas.

7.2 Selected Exploratory Data Analysis

Sequences in this dataset were contributed by laboratories across the United States, with the largest contributions from New York, Texas, and Washington. Although RSV surveillance is inconsistent across states, most studies have documented consistent national co-circulation patterns of RSV-A and RSV-B, with both subtypes following similar seasonal cycles across regions (CDC 2024). Shifts of this typical seasonal pattern have occurred in Florida, Alaska, Hawaii, and other US territories (CDC 2024). However, initial exploratory data analysis of RSV counts by state shown in the plots below (Figure 7, Figure 8, Figure 9, Figure 10) suggest minimal differences in seasonal timing across these states and New York's typical seasonal pattern. Thus, counts for all US states and territories were included to conserve the number of observations. Under this assumption, we treat the United States as a single homogeneous population and do not model spatial differences.

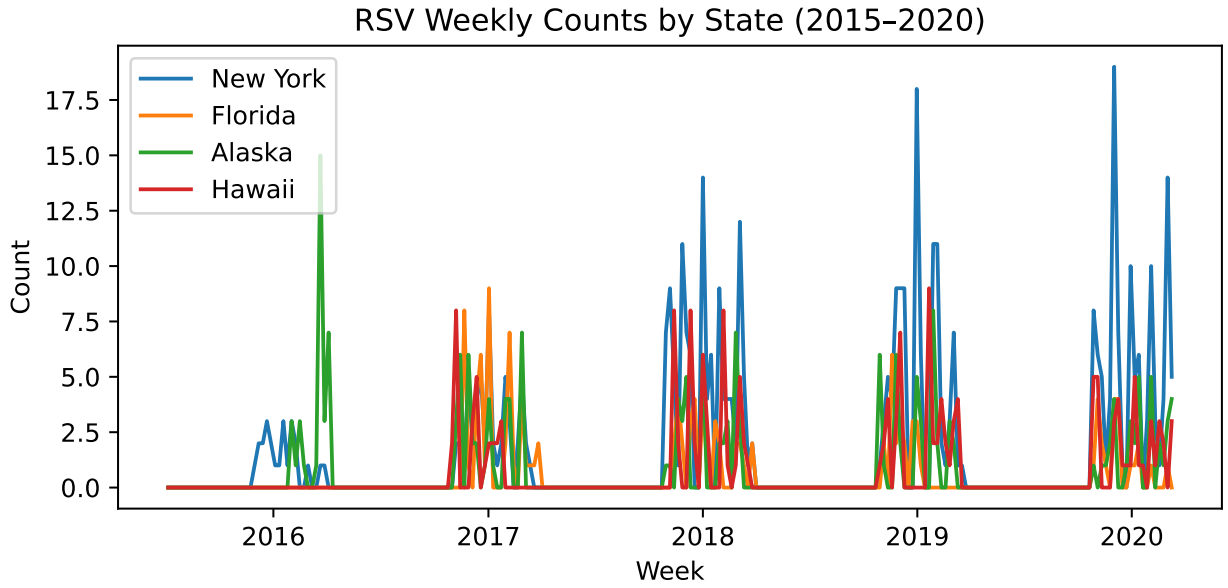


Figure 7: Comparison of Counts for Selected States (2015–2020)

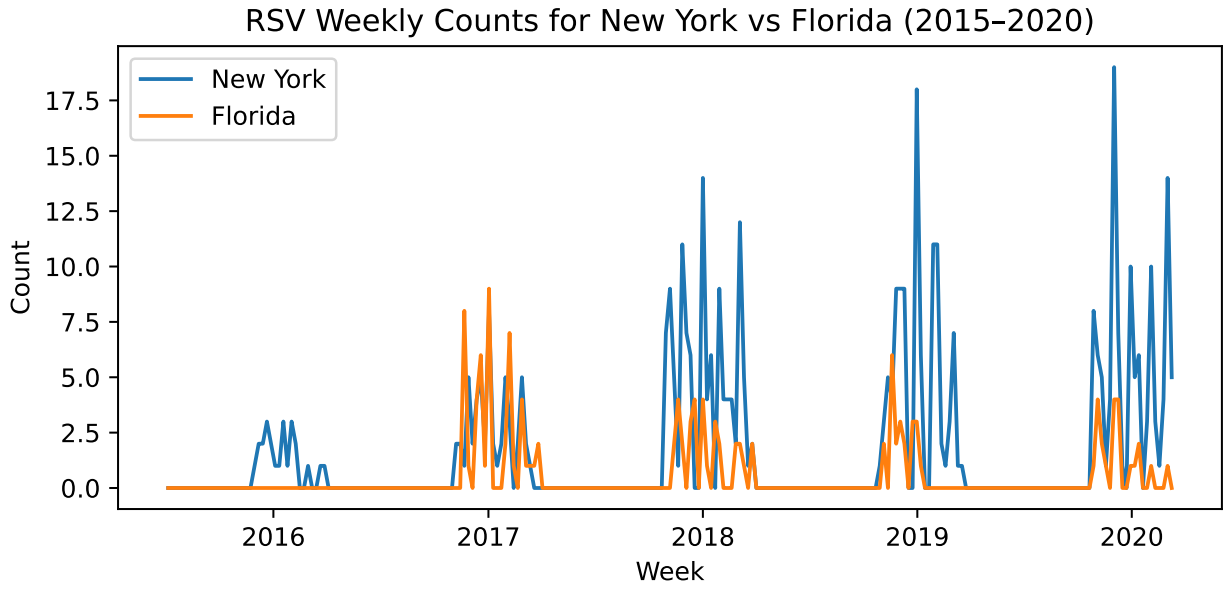


Figure 8: Comparison of Counts for New York and Florida (2015–2020)

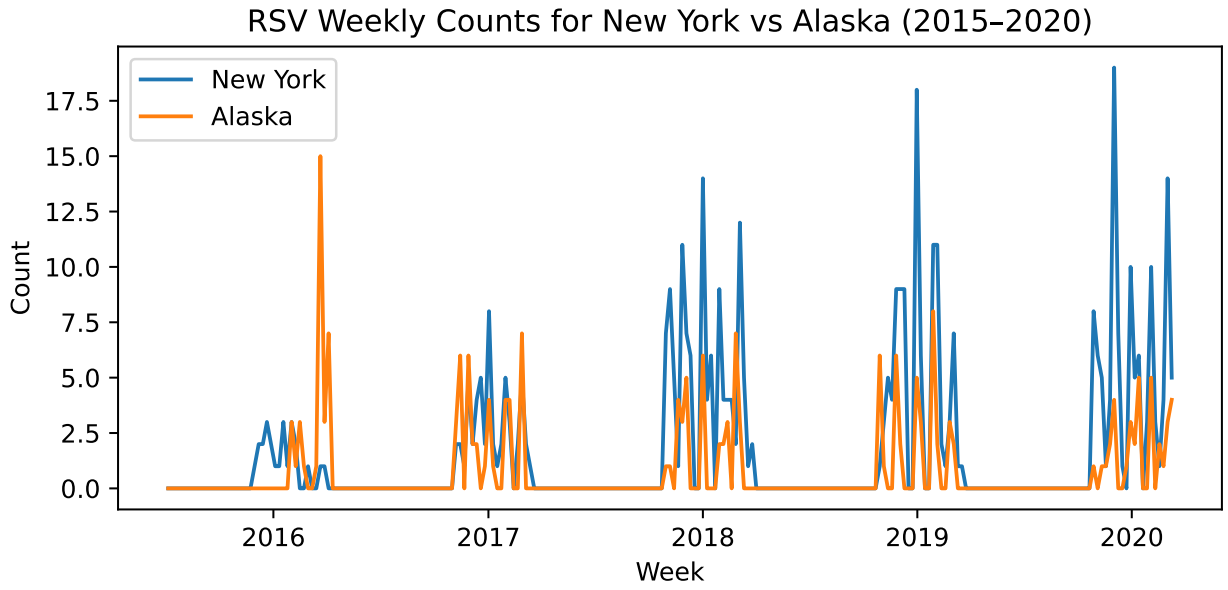


Figure 9: Comparison of Counts for New York and Alaska (2015–2020)

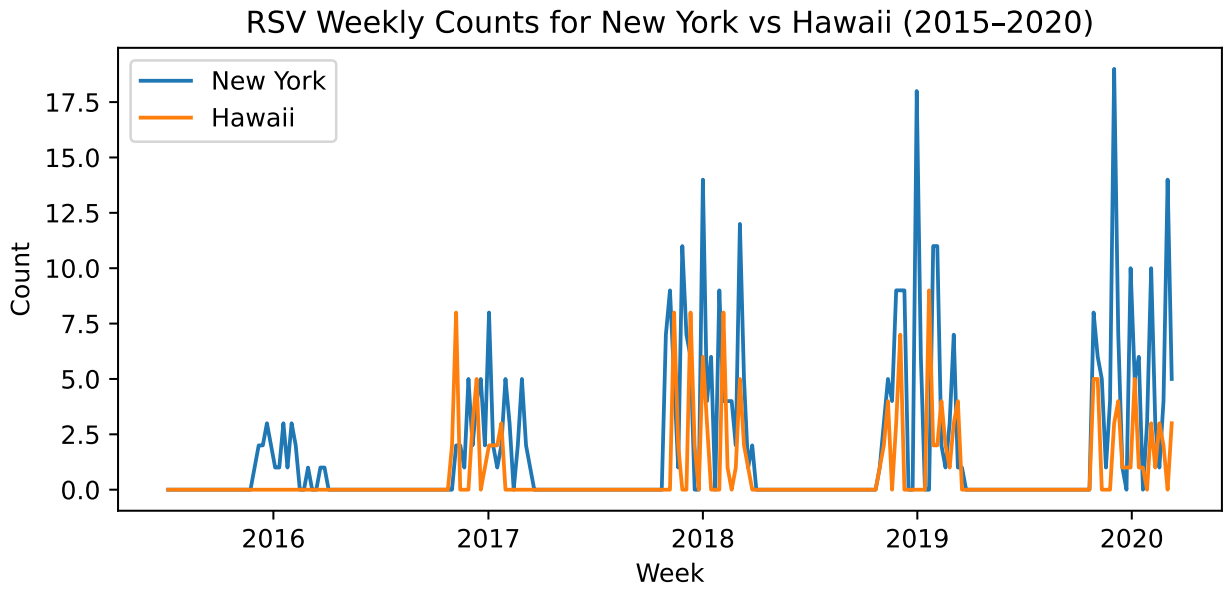


Figure 10: Comparison of Counts for New York and Hawaii (2015–2020)

7.3 Methods Details

7.3.1 SARIMAX Model Diagnostics

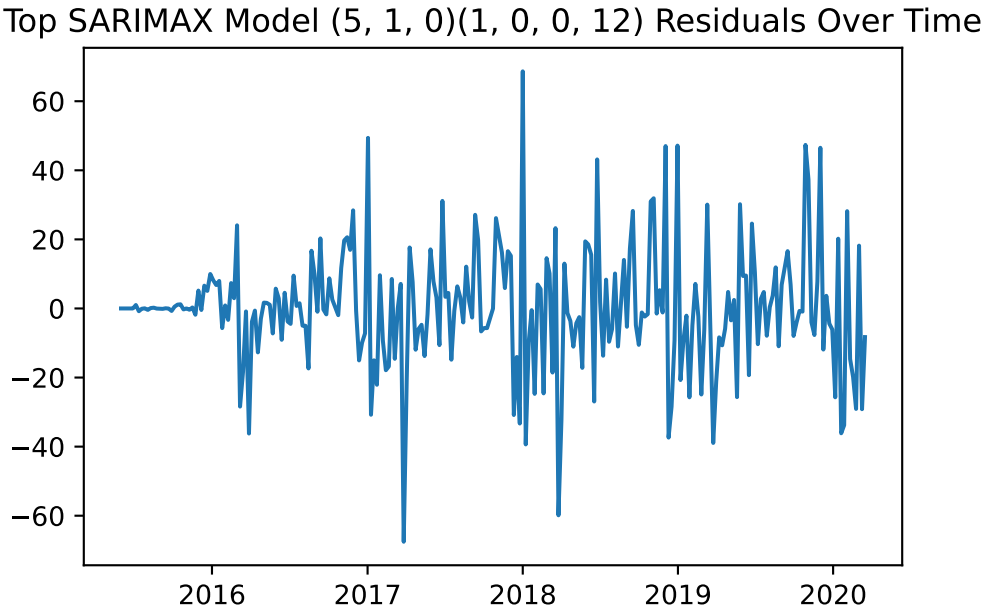


Figure 11: Residuals of the SARIMAX Model over Time

Figure 11 shows the residuals of the top-performing SARIMAX model over time. The residuals display clear heteroskedasticity, with substantially larger variance during seasonal peaks (around weeks 300–350 and 600–650) compared to off-peak periods. This pattern, combined with the Jarque-Bera test result (JB = 755, $p < 0.001$) reported in the main text, confirms that the linear-Gaussian assumptions of SARIMAX are violated and motivates the use of a count-based model.

7.3.2 Model Parameters

Table 2: Model parameters for the two-strain SEIRS model

Parameter	Description	Role	Notes
β_A, β_B	Transmission rates for RSV-A and RSV-B	Estimated	Seasonally varying
σ_A, σ_B	Progression rate from exposed to infectious	Fixed	$1/\sigma$ = latent period
γ_A, γ_B	Recovery rate from infectious to recovered	Fixed	$1/\gamma$ = infectious period
ω	Waning immunity rate	Fixed	$1/\omega$ = duration of immunity
η	Initial susceptible fraction	Estimated	Determines initial conditions

Parameter	Description	Role	Notes
ρ_0	Initial Reporting rate	Estimated	Links incidence to observed cases at initialization
ρ_{slope}	Reporting rate increase over time	Time varying	Links incidence to observed cases changing over time due to increased reporting
k	Dispersion parameter (NegBin)	Estimated	Smaller k = more overdispersion
q	Cross-immunity parameter	Estimated	$q = 0$ none, $q = 1$ full
N	Effective population size	Fixed	Scales transmission
t_{peak}	Seasonal peak timing	Fixed	Controls timing of epidemics
$A(t)$	Seasonal amplitude	Stochastic	Varies annually
a, b	Shape parameters for $A(t)$	Fixed	Beta distribution
$A_{\max}$	Maximum seasonal amplitude	Fixed	Scales seasonal variation

Stochastic Transition Equations

The strain-specific forces of infection are given by:

$$\lambda_A(t) = \beta_A(t)\frac{I_A(t) + (1-q)I_B(t)}{N}, \qquad \lambda_B(t) = \beta_B(t)\frac{I_B(t) + (1-q)I_A(t)}{N}.$$

Seasonal forcing is incorporated through:

$$\beta_A(t) = \beta_{0,A}\left[1 + A(t)\cos\left(\frac{2\pi(t - t_{\text{peak}})}{52}\right)\right], \qquad \beta_B(t) = \beta_{0,B}\left[1 + A(t)\cos\left(\frac{2\pi(t - t_{\text{peak}})}{52}\right)\right].$$

Transitions between compartments use a binomial approximation with exponential transition probability:

$$N_{SE,A}(t) \sim \text{Binomial}\left(S(t), 1 - e^{-\lambda_A(t)\delta}\right)$$

This was also used for all other transitions (E to I, I to R, R to S) for both strains.

Incidence Process

Incidence is defined as the number of new infections entering the exposed compartments.

$$H_A(t+\delta) = H_A(t) + dN_{SE,A}(t), \qquad H_B(t+\delta) = H_B(t) + dN_{SE,B}(t).$$

Initial Conditions

Initial conditions are defined as:

$$E_A(0) = 3, \quad I_A(0) = 1, \quad E_B(0) = 3, \quad I_B(0) = 1,$$

The seasonal amplitude state is initialized as:

$$A(0) = A_{\text{mean}}.$$

Measurement Model

Observed case counts are modeled using a negative binomial distribution:

$$Y_{A,t} \sim \text{NegBin}(\mu_{A,t}, k), \qquad Y_{B,t} \sim \text{NegBin}(\mu_{B,t}, k),$$

with

$$\mu_{A,t} = \rho_t H_A(t), \qquad \mu_{B,t} = \rho_t H_B(t),$$

where ρ_t is the reporting rate and k is the dispersion parameter.

Seasonal Amplitude Process

Seasonal variation in transmission is governed by a stochastic annual amplitude:

$$A_{\text{year}} \sim A_{\max} \cdot \text{Beta}(a, b).$$

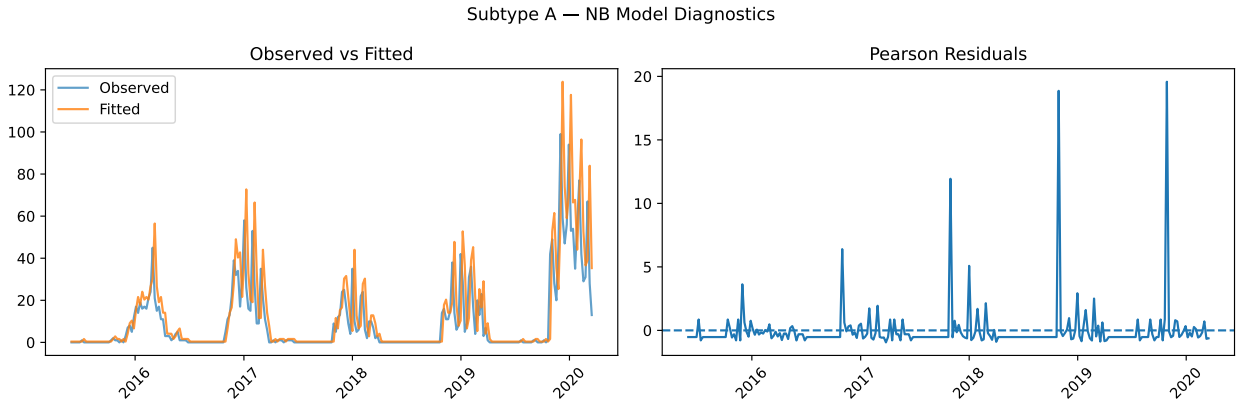
The sampled amplitude is held constant within each year and reset at annual boundaries.

7.4 Benchmark Negative Binomial Model Parameter Estimates

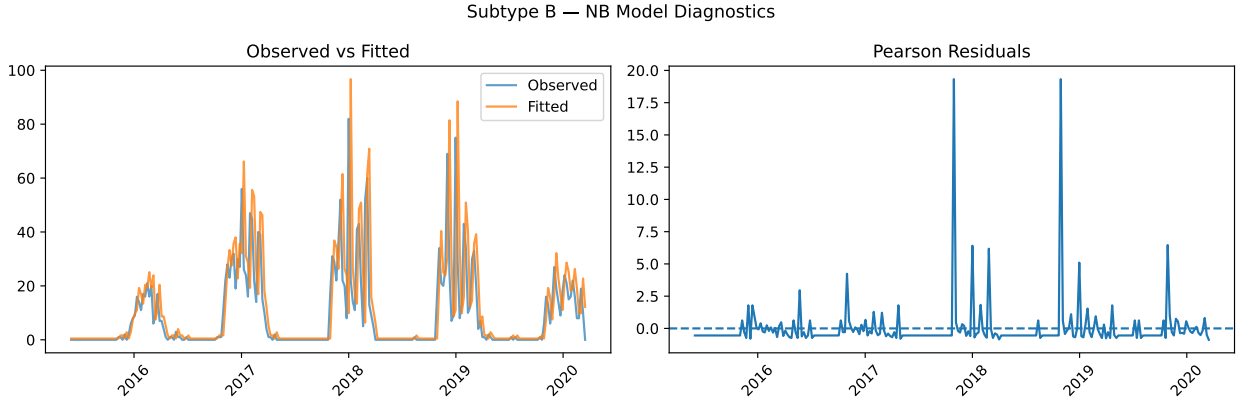
```
Subtype A: n_obs = 251, log-likelihood = -580.51
Subtype B: n_obs = 251, log-likelihood = -578.07

Combined log-likelihood: -1158.58
```

Table 3: Negative Binomial Benchmark Model Parameter Estimates

	Subtype	mu_intercept	beta1	size (k)	log_likelihood	n_obs
0	A	0.378847	1.247385	1.007119	-580.509132	251
1	B	0.468271	1.173637	0.831954	-578.073809	251



(a) Negative Binomial Benchmark Diagnostics



(b)

Figure 12

7.5 No Immunity Model Local Search Results

```
Local search runtime (RL=2): 353.24 seconds
```

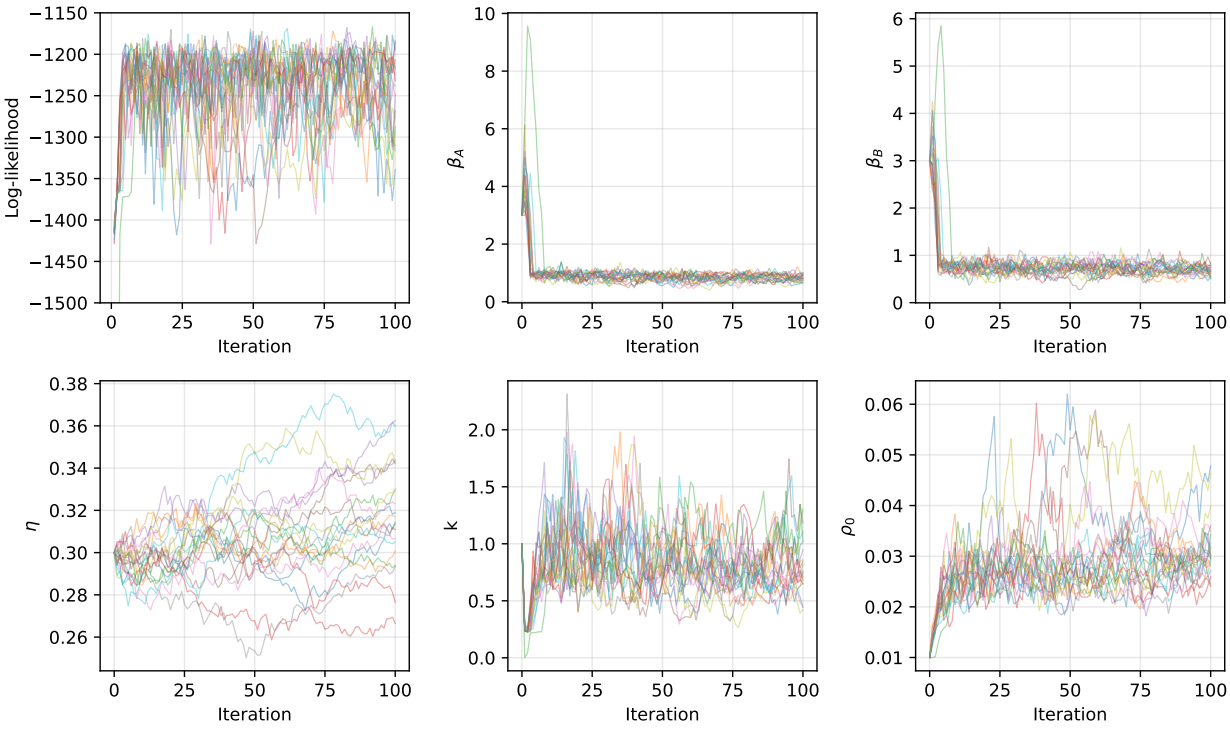


Figure 13: Local Search Results for the No Immunity Model

```
Best log-likelihood: -1162.4
(SE: 16.8)
```

7.6 Partial Immunity Model Full Global Search Plots

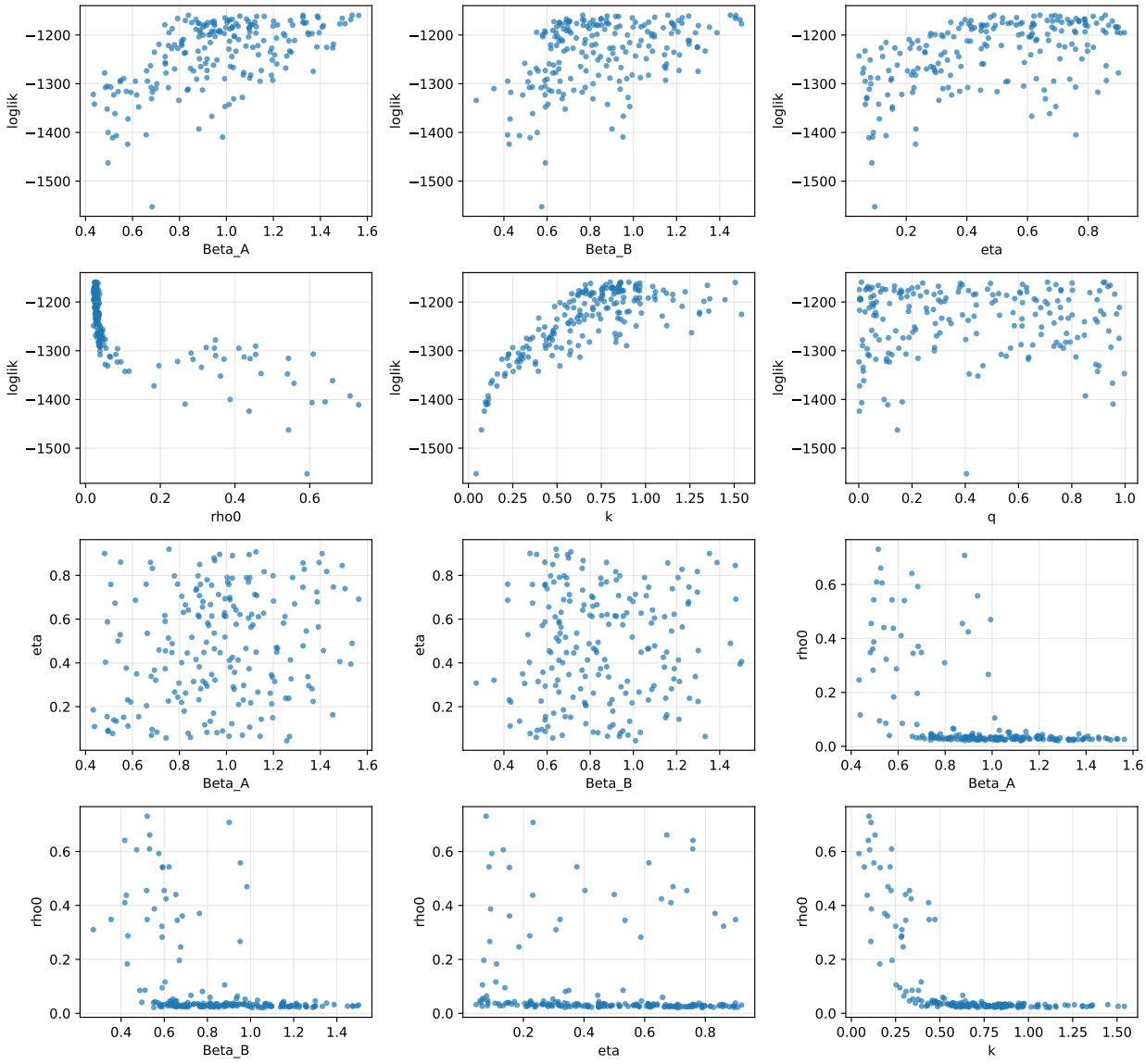


Figure 14: Full Global Search Results for the Partial Immunity Model

7.7 No Immunity Model Full Global Search Plots and Best Model Parameters

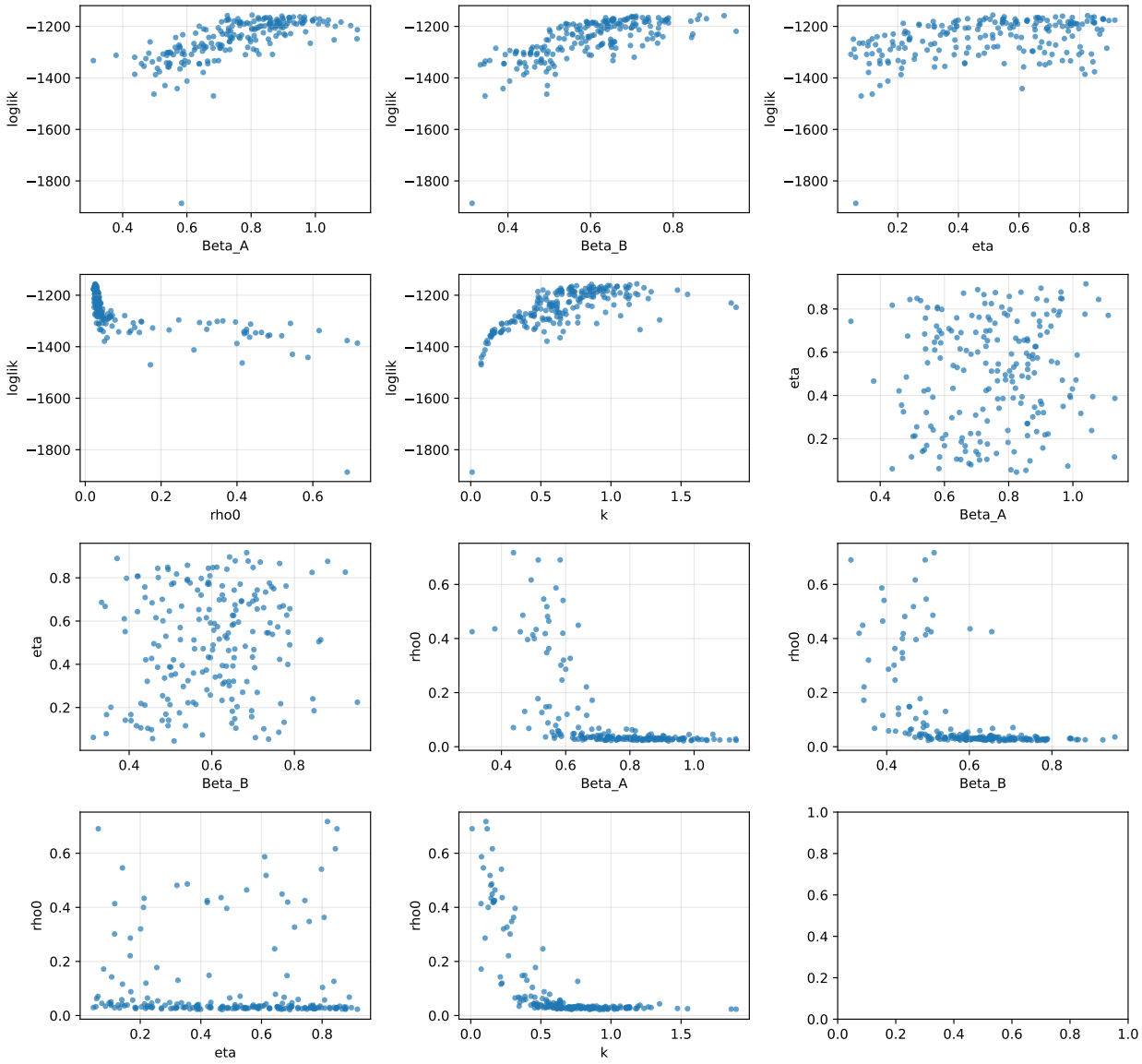


Figure 15: Global Search Results for the No Immunity Model

Table 4: Top 10 Largest Log Likelihoods Obtained via Global Search for No Immunity Model

	loglik	Beta_A	Beta_B	eta	rho0	k
155	-1155.914877	0.802926	0.863725	0.513512	0.026269	1.184851
78	-1158.221267	0.725922	0.923268	0.826030	0.025261	0.860685
3	-1160.035600	0.895631	0.709932	0.743530	0.027503	1.134499
145	-1162.601639	0.951855	0.699039	0.551306	0.026519	0.978710

Table 4: Top 10 Largest Log Likelihoods Obtained via Global Search for No Immunity Model

	loglik	Beta_A	Beta_B	eta	rho0	k
18	-1163.223981	0.966280	0.700041	0.847905	0.029079	0.992684
148	-1163.400122	0.931638	0.703633	0.720617	0.025569	0.754709
39	-1165.441350	0.918677	0.739536	0.775570	0.025433	1.045218
20	-1165.844128	0.857828	0.725514	0.661758	0.028548	1.105108
187	-1166.197225	0.917865	0.636509	0.839819	0.030294	0.718879
142	-1166.271321	0.822007	0.764327	0.664813	0.023403	0.930681