

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



**Integrating viral RNA sequence and epidemiological data to
define transmission patterns for respiratory syncytial virus**

IVY KADZO KOMBE BSc, MSc

**Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the University of London**

July 2019

Department of Global Health and Development

Faculty of Public Health and Policy

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the Wellcome Trust [107769/Z/10/Z, 102975 and 090853] through the
DELTA Africa Initiative [DEL-15-003].

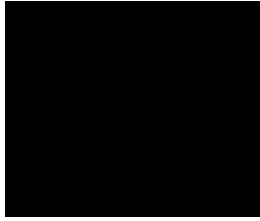
Research group affiliation(s):

Centre for Mathematical Modelling of Infectious Diseases (LSHTM)

Virus Epidemiology and Control (KEMRI-Wellcome Trust)

Declaration

I, Ivy Kadzo Kombe, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.



July 2019

Abstract

The analyses contained herein focus on making comparisons between model inferences obtained using different scales of pathogen identification, with a particular focus on respiratory syncytial virus (RSV). A significant proportion of lower respiratory tract infections in children has been attributed to infection by RSV and as such, there has been global interest in understanding its transmission characteristics in order to plan for effective control. Mathematical models have often been used to explore potential mechanisms that drive the patterns observed in data collected at different scales. Several models have been used to explore how immunity to RSV is acquired and maintained, vaccination strategies and potential drivers of seasonality. However, most of these models do not make a distinction between the two antigenically and genetically distinct RSV groups (RSV A and RSV B), neither do they consider its ecological environment, in particular, potential interactions between RSV and other viral pathogens. This thesis therefore presents work done aimed at understanding the transmission characteristics of viral respiratory pathogens spreading in a group of households using a dynamic model of transmission

The data analysed is cohort data collected between December 2009 and June 2010 from 493 individuals distributed across 47 households from a rural coastal community in Kenya. Individuals in the study had nasopharyngeal swab samples collected twice weekly irrespective of symptom status. Infecting viral pathogens were identified using RT-PCR resulting in the identification of 4 main pathogens: RSV, human coronavirus, rhinovirus and adenovirus. RSV and coronavirus were further classified according to genetically distinct subgroups. Some of the RSV samples were sequenced to obtain whole genome sequences (WGS) and further classified into genetic clades/clusters.

I first conducted a review of methods to identify the best way to integrate social-temporal data and WGS genetic data into a single modelling framework for RSV. Given that the social-temporal data and genetic data were available at different sampling densities, I decided to use a model that focused on the data with the highest density. The results in this thesis are thus presented in three main chapters; the first focuses on analysing social-temporal shedding patterns of RSV identified at the group level (i.e.

distinguish between RSV A and RSV B); the second incorporates the available genetic data into the model used to analyse the social-temporal data (i.e. separating RSV-A into 5 clusters, and RSV-B into 7 clusters); the third is an analysis of the interaction of two pathogens, RSV and coronavirus, identified at two different scales.

One of the main findings in this thesis is that the household setting plays an important role in the spread of RSV, a finding that is made clearer with added detail on pathogen type. In the case of the data analysed here, and the social structuring from which it was collected, RSV clades appeared to mimic household structure as such identification at this level did not drastically change the transmission characteristic observed with identification at the group level. However, the combination of epidemiological and genetic data elucidated transmission chains within the household enabling the identification of the sources of infant RSV infections. For this particular study, it was inferred that the sources of infant RSV infections were both in the same household as the infant and from external sources. Where infant infections occurred in the household, the source of infection was often a child between the ages of 2-13 years. It was inferred that previous infection with one RSV group type reduced susceptibility to re-infection by heterologous group type within the same epidemic. Interactions were also observed between RSV and human coronavirus groups. In particular, previous infection with RSV B was estimated to increase susceptibility to corona OC43 by 81% (95% CrI: 40%, 134%). Detailed data of infection events in individual hosts can provide a wealth of knowledge. The inferences made from this study should be explored at larger spatial and temporal scales to determine the population level impact, and hence public-health significance, of pathogen interactions, whether these interactions are between strains of the same pathogen or between different pathogens. In planning for, and assessing the impact of, an intervention against a particular pathogen, investigators should not ignore the pre-existing ecological balance and should make efforts to understand how this will be disrupted by an intervention against one or more pathogens.

Acknowledgements

I would like to thank my supervisors James Nokes and Graham Medley for their consistent support throughout this PhD project. James, it has taken me over 6 years to get to where I am today, and I would not have managed any of it without your initial guidance. Thank you for giving me the space to explore ideas and grow my networks. Graham, I am grateful for every meeting you made time for, every email and report you responded to with encouragement and diplomatically worded criticism when needed. It was an absolute pleasure working under both yours and James' supervision and I am very excited about the work that has come out of this project. My thanks also to Marc Baguelin for his advice.

I am grateful for the advice and collaboration from Patrick Munywoki, Charles Agoti and George Githinji. Thank you, Benjamin Tsofa and Martha Mwangome, for your mentorship. A special thank you to the training team at KEMRI-Wellcome Trust in Kilifi headed by Sam Kinyanjui, in particular Liz Murabu for your administrative support and being concerned about the welfare of all your students. To my colleagues at KEMRI-Wellcome, thank you for friendship, support and advice. In particular, Alice 'Kare' Kamau, Anne Amulele, James Otieno, Moses Kiti, Patience Kiyuka, Esther Muthumbi and Collins.

To Chee, Vera and Rumbi (my London family), thank you for constantly checking on me and reminding me to take a break and regroup.

I am eternally grateful to my parents and brother for supporting me throughout this journey, believing in me and for celebrating every milestone with me. Thank you for keeping me grounded, helping me to keep my life in order and being there to remind me that this PhD is not just about me, being able to do this is an achievement for the entire Kombe family.

And finally, to Martin 'Teacup' Wanyoike Njoroge, thank you for your love and support and always being proud of how far I have come. Thank you for helping me keep my sanity, for understanding all the times I had to work from home, for being patient and

always being willing to listen to my ranting and ravings about things that did not always make sense to you. I look forward to walking with you through your PhD journey and to finally seeing our sacrifices pay off.

This work was supported through the DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769/Z/10/Z, 102975 and 090853] and the UK government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

Table of contents

Declaration	2
Abstract	3
Acknowledgements	5
Table of contents	7
List of Figures	11
List of Tables	15
Abbreviations	16
1. Introduction	17
1.1. <i>RSV disease burden and epidemiology</i>	17
1.2. <i>Control</i>	19
1.3. <i>Models for improved understanding of disease transmission</i>	21
1.4. <i>Data from the Kilifi household study cohort</i>	22
1.4.1. Description	22
1.4.1. Past work	25
1.5. <i>Motivation for the PhD</i>	26
1.6. <i>Computation</i>	27
1.7. <i>Structure of the thesis</i>	28
1.8. <i>Additional information</i>	29
1.8.1. Ethics statement	29
1.8.2. Training	29
1.9. <i>References:</i>	31
2. Review of models of RSV transmission dynamics and methods for including genetic information	41
2.1. <i>Models of respiratory syncytial virus</i>	41
2.2. <i>Approaches in phylodynamics</i>	52
2.3. <i>References:</i>	56
3. Paper 1: Model based estimates of transmission of respiratory syncytial virus within households.	62
3.1. <i>Overview</i>	62
3.2. <i>Role of candidate</i>	62
3.3. <i>Abstract</i>	66

3.4.	<i>Introduction</i>	67
3.5.	<i>Methods</i>	70
3.6.	<i>Results</i>	83
3.7.	<i>Discussion</i>	93
3.8.	<i>References:</i>	98
4.	Paper 2: Integrating epidemiological and genetic data with different sampling densities into a dynamic model of RSV transmission.	102
4.1.	<i>Overview</i>	102
4.2.	<i>Role of candidate</i>	102
4.3.	<i>Abstract</i>	105
4.4.	<i>Introduction</i>	106
4.5.	<i>Methods</i>	109
4.5.1.	Imputing missing genetic information	111
4.5.2.	Deriving genetic distances between cases	113
4.5.3.	The transmission model	114
4.5.4.	Inference of model parameters and augmented data	121
4.5.5.	Highest probability transmission source	122
4.6.	<i>Results</i>	124
4.6.1.	The data	124
4.6.2.	Inference on model parameters	128
4.6.3.	Highest Probability transmission source	135
4.7.	<i>Discussion</i>	139
4.8.	<i>References:</i>	144
5.	Paper 3: A multi-pathogen model of infection investigating potential interactions between respiratory syncytial virus and coronavirus.	149
5.1.	<i>Overview</i>	149
5.2.	<i>Role of candidate</i>	149
5.3.	<i>Abstract</i>	152
5.4.	<i>Introduction</i>	153
5.5.	<i>Methods</i>	157
5.5.1.	Data	157
5.5.2.	Transmission model	158
5.5.3.	Parameter inference	163
5.6.	<i>Results</i>	165
5.6.1.	Data	165

5.6.2.	Pathogen interactions	167
5.6.3.	Modified pathogen inference	171
5.7.	<i>Discussion</i>	174
5.8.	<i>References</i>	178
6.	Discussion	185
	<i>References:</i>	199
	Appendices	203
	A1: Ethical approval	203
	A2: Supplementary appendix for Paper 1.	206
	A2.1. <i>Imputing shedding durations, symptomatic episodes and viral loads</i>	206
	A2.2. <i>Extra results</i>	211
	A2.3. <i>Modification of the likelihood to establish the most likely infection source for every case.</i>	216
	A2.4. <i>Model validation</i>	218
	A2.5. <i>Sensitivity analysis</i>	225
	A2.6. <i>Checking the contribution of symptomatic and asymptomatic individuals</i>	230
	A2.7. <i>Fitting household size as an ordinal variable</i>	233
	A3: Supplementary appendix for Paper 2.	237
	A3.1. <i>Normalizing the cluster specific background community exposure rate curves</i>	237
	A3.2 <i>Further details on the inference method (RJ-MH-MCMC)</i>	239
	A3.2.1. Metropolis-Hastings Markov Chain Monte Carlo (MH-MCMC) for parameter inference	239
	A3.2.2 Our application of MH-MCMC	240
	A3.3. <i>Establishing the highest probability transmission source (HPTS)</i>	243
	A3.4. <i>Details of the model using pathogen data identified at group resolution</i>	245
	A3.5. <i>Results of the MCMC algorithm</i>	247
	A3.6. <i>Model validation</i>	252
	A4: Supplementary appendix for Paper 3.	258
	A4.1 <i>Extra results</i>	258
	A4.2. <i>Results of the MCMC algorithm</i>	266
	A4.2.1. Multi-pathogen model fit to data with pathogen identification at the group level	266
	A4.2.2. Multi-pathogen model fit to data with pathogen identification at the pathogen level	277

A4.2.3. Single-pathogen model fit to RSV data with pathogen identification at the group level	279
A4.2.4. Single-pathogen model fit to hCoV data with pathogen identification at the group level	282
<i>Appendix References</i>	286

List of Figures

Figure 1. 1: Maps showing the household study site in geographical context as at September 2009. ...	23
Figure 1. 2: Histograms showing the age distribution of the household members.	24
Figure 1. 3: Histograms showing the distribution of the total number of samples collected from the study participants.	24
Figure 1. 4: A flow chart for the SIR model, the model equations and sample deterministic model projections.	42
Figure 1. 5: Possible extensions of the basic SIR model representing different assumptions about the natural history of an infection.	44
Figure 3. 1: Establishing the background community rate function.	75
Figure 3. 2: Shedding patterns for each of the 179 individuals who experienced at least one RSV shedding episode.	84
Figure 3. 3: Comparing the range of within household exposure rate and community exposure rate for a single susceptible individual given different heterogeneities in exposure and infectiousness..	88
Figure 3. 4: A comparison between the simulated data and real epidemics using simulations from 5 different parameter sets estimated from the full model (row 1 to 5).	91
Figure 4. 1: Illustration of how cases become disconnected with added pathogen information	110
Figure 4. 2: Time-resolved maximum likelihood phylogenetic trees for RSV A and RSV B from the <i>Agoti et al</i> ⁴¹ phylogenetic analysis.	111
Figure 4. 3: Distribution of pair-wise nucleotide distances between RSV A sequences.	125
Figure 4. 4: Distribution of pair-wise nucleotide distances between RSV B sequences.	126
Figure 4. 5: Distribution of available sequences across shedding episodes (left) and the results of imputation of cluster durations (right).	127
Figure 4. 6: A comparison of simulated and observed data for RSV A.	132
Figure 4. 7: A comparison of simulated and observed data for RSV B.	133
Figure 4. 8: A comparison of the parameter distributions obtained from the model using different resolutions in pathogen identification.	134
Figure 4. 9: Transmission networks showing the highest probability source of transmission given by our model results.	138
Figure 5. 1: Temporal distribution of cases for the 5 infectious agents clustered by age group.	166
Figure 5. 2: Distribution of shedding episodes for coronavirus and RSV by household and time.	167
Figure 5. 3: Comparing parameter densities obtained from fitting the data at the pathogen level (solid black lines) to the densities obtained from fitting data at the group level (dashed colored lines).	170
Figure 5. 4: Comparison of the RSV specific parameters obtained from fitting a single pathogen model (blue line) to those obtained from fitting a multi-pathogen model (black line).	172

Figure 5. 5: Comparison of the coronavirus specific parameters obtained from fitting a single pathogen model (pink line) to those obtained from fitting a multi-pathogen model (black line).....	173
Figure A2. 1: Distributions of imputed shedding durations for RSV A (left) and RSV (right)	207
Figure A2. 2: Histograms of interpolated viral loads for RSV A (left) and RSV B (right).....	208
Figure A2. 3: Shedding and ARI patterns for each of the 88 individuals who experienced at least one RSV A shedding episode.	209
Figure A2. 4: Shedding and ARI patterns for each of the 113 individuals who experienced at least one RSV B shedding episode.....	210
Figure A2. 5: Trace plots showing convergence for the 15 parameters of interest.....	212
Figure A2. 6: Caterpillar plot of estimated parameters.....	213
Figure A2. 7: Comparing the total household rate of exposure $j \neq i$ HH. Riskh, g, j \rightarrow it between small and large households.	214
Figure A2. 8: Correlation patterns of the different parameters obtained from fitting to the observed data.....	215
Figure A2. 9: Flow chart showing validation process	218
Figure A2. 10: RSV A shedding profiles as observed.	220
Figure A2. 11: RSV B shedding profiles as observed.	221
Figure A2. 12: Histograms of the posterior distributions with vertical lines showing sample sets that were used in simulation.	222
Figure A2. 13: Outcome measures from simulated data when using different sets of parameters drawn from the posterior estimated from the observed data.....	223
Figure A2. 14: Comparing the real (red lines) and simulated(black lines) epidemics	224
Figure A2. 15: Comparing the posterior densities obtained from using the observed data to those from using the simulated data.	225
Figure A2. 16: Using different density functions for the background community rate and comparing results.	226
Figure A2. 17: Using different density functions for the background community rate and comparing results.	227
Figure A2. 18: Using different density functions for the background community rate and comparing results.	228
Figure A2. 19: Frequency distributions of RSV A and RSV B infections by household size	228
Figure A2. 20: Infection patterns in HH5	229
Figure A2. 21: Caterpillar plot showing results obtained when household 5 data was removed from the set.	229
Figure A2. 22: Comparing densities of parameters estimates obtained when using all the data (light red) to densities obtained when using data without household 5 (light blue).	230
	231

Figure A2. 23: Densities comparing the relative total incidence, by RSV group and age group, when the infectiousness of symptomatic individuals is altered or when the infectiousness of asymptomatic individuals is removed.....	231
Figure A2. 24: Caterpillar plot of estimated parameters when only data from symptomatic episodes is used.	232
233	
Figure A2. 25: Comparing densities of parameters estimates obtained when using all the data (light red) to densities obtained when using data from only symptomatic cases (light blue).	233
Figure A2. 26: Caterpillar plot showing the results of estimating a parameter ω (omega) when household size is treated as an ordinal variable.....	234
Figure A2. 27: Caterpillar plot showing the results of estimating a parameter ω (omega) when household size is treated as an ordinal variable.....	235
Figure A2. 28: Caterpillar plot showing the results of estimation when household size is treated as a categorical variable but with the definition of a household changed.	236
Figure A3. 1: The background cluster-specific rate of exposure curves for RSV A.....	238
Figure A3. 2: The background cluster-specific rate of exposure curves for RSV B.....	239
Figure A3. 3: Trace plots of parameters in the cluster model.	248
Figure A3. 4: Trace plots of parameters in the group level data model.	249
Figure A3. 5: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	250
Figure A3. 6: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	251
Figure A3. 7: Violin plots showing the distribution of the total number of people infected in the simulations by RSV group and age.	256
Figure A3. 8: Violin plots showing the distribution of the proportion of cases that had multiple onsets in the simulations by RSV group and age.	256
Figure A3. 9: Violin plots showing the distribution of the number of cases in the first (1 st column) and last (2 nd column) week of the observation/simulation period in the simulations by RSV group.	257
Figure A4. 1: Distributions of shedding durations for the different infectious agents.	258
Figure A4. 2: Distribution of shedding episodes for coronavirus 229E and RSV A by household and time.	262
Figure A4. 3: Distribution of shedding episodes for coronavirus OC43 and RSV B by household and time.	263
Figure A4. 4: Distribution of shedding episodes for RSV A and RSV B by household and time.....	264
Figure A4. 5: Distribution of shedding episodes for coronavirus 229E, NL63 and OC43 by household and time.....	266
Figure A4. 6: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	274

Figure A4. 7: Trace plots of parameters in the multi-pathogen model with pathogen identification at the pathogen level.	277
Figure A4. 8: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	278
Figure A4. 9: Trace plots of parameters in the single-pathogen model for RSV with pathogen identification at the group level.....	280
Figure A4. 10: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	281
Figure A4. 11: Trace plots of parameters in the single-pathogen model for hCoV with pathogen identification at the group level.....	283
Figure A4. 12: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.	284

List of Tables

Table 3. 1: Model Notation	77
Table 3. 2: Summary of shedding episodes	83
Table 3. 4: Results of fitting the transmission model	85
Table 4. 1: Model parameters and their descriptions	118
Table 4. 2: A summary of the distribution of sequences	124
Table 4. 3: Median and 95% credible intervals for parameters estimated using the model with sequence data.	128
Table 4. 4: Characteristics of the transmission chains inferred.	135
Table 4. 5: Age distribution of index cases of household outbreaks.	136
Table 5. 1: Description of variables in the model	160
Table 5. 2: Summary of the data	165
Table 6. 1: A summary of the three variants of the individual level model used to investigate transmission dynamics of RSV.	189
Table A2. 1: Results of fitting a reduced version of the model	215
Table A2. 2: Parameter set used to simulate an epidemic	223
Table A3. 1: The value of the GRB statistic (to 3 significant figures) and the ESS after burn-in are shown for the parameters in the cluster level data model.	250
Table A3. 2: The value of the GRB statistic (to 3 significant figures) and the ESS after burn-in are shown for the parameters in the group level data model.	252
Table A4. 1: Results of parameter estimation using data identified at the pathogen level and group level.	258
Table A.4 1: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the multi pathogen model with pathogen identification at the group level.	275
Table A.4 2: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the multi pathogen model with pathogen identification at the pathogen level.	278
Table A.4. 3: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the single pathogen RSV group model.	281
Table A.4. 4: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the single pathogen hCoV strain model.	285

Abbreviations

ALRI	Acute lower respiratory infection
ESS	Effective sample size
HCoV	Human coronavirus
HMPV	Human metapneumovirus
HPTS	Highest probability transmission source
IPD	Invasive pneumococcal disease
KHDSS	Kilifi Health and Demographic Surveillance System
LRTI	Lower respiratory tract infection
MCMC	Markov Chain Monte Carlo
MMR	Measles mumps and rubella
NPS	Nasopharyngeal swab
ODE	Ordinary differential equations
PCR	Polymerase chain reaction
PCV	Pneumococcal conjugate vaccine
RSV	Respiratory syncytial virus
UK	United Kingdom
URTI	Upper respiratory tract infection
USA	United States of America
WHO	World Health Organization

1. Introduction

1.1. RSV disease burden and epidemiology

The continued identification of respiratory syncytial virus (RSV) as a major cause of acute lower respiratory infection (ALRI) is of global concern. In 2005, an estimated 33.8 million ALRIs in children less than 5 years of age were due to RSV, resulting in 3.4 million hospitalizations¹. Ten years on, and the estimated ALRI burden due to RSV had not changed; 33.1 million cases were estimated to arise from an RSV infection resulting in 118,200 (94,600-149,400) deaths. Over 90% of the estimated RSV burden was carried by developing countries.² A recent study across sites in 7 low-income and low-middle-income countries looking into the aetiology of severe and very severe pneumonia found that RSV has the largest attributable fraction of any single pathogen, including bacterial pathogens³. Infants below 6 months of age experience the most severe disease⁴. Increasingly, RSV is also being identified as a disease causing pathogen in the elderly, with the fraction of disease due to RSV being comparable to that due to non-pandemic influenza⁵. Though studies vary in their definition of lower respiratory illness, this does not alter the fact that RSV has a key role to play.

Individuals are repeatedly infected with RSV throughout their lives, however the risk of disease decreases with age, possibly the result of a combination of physiology (increase in airway size) and immunology (immunological maturity and past exposure)^{6,7}. Immunity to RSV infection is evidently partial and transient^{6,8,9}. Primary infection occurs early in life, and most children will have experienced at least one RSV infection by the age of two years¹⁰⁻¹³. Children are usually born with maternally acquired RSV specific antibodies. Though these wane quickly, high levels have been associated with reduced risk of severe disease in the first 3-6 months of life^{10,14}. However, the protective antibody threshold remains unclear¹⁵ and given that infection still occurs in the 3-6 month age group, the protective effect of maternally acquired antibodies is likely partial and the exact mechanisms of action are yet to be understood¹⁴. Age is not only a factor determining the severity of RSV-related disease, it has also been associated with duration of shedding (younger children have longer durations)¹⁶, and household/ family studies have found that for an infant, having an older sibling of school-going age increases the risk of infection¹⁷⁻²⁰

Respiratory syncytial virus is highly transmissible, evident by its rapid spread in close contact settings²¹⁻²³, but relative to viruses like non-pandemic influenza A, it is less invasive at a cellular level²⁴. A large proportion of RSV infections are asymptomatic²⁵. Mild cases presenting with cold-like symptoms tend to resolve themselves within 2 weeks. Severe cases that require hospitalization receive supportive therapy in the form of administration of supplementary oxygen, mechanical ventilation and fluid replacement²⁶. Though a clear association has been observed between decreasing age and increased disease severity, RSV pathogenicity is likely to be multifactorial, involving a combination of viral and host factors that contribute to a range of infection presentations even within hosts of the same age²⁷.

The RSV genome is about 15000 bases long consisting of 10 genes coding for 11 proteins. Of the three surface proteins, it is only the fusion (F) and attachment (G) glycoprotein that have been found to elicit protective neutralizing antibodies²⁸. Respiratory syncytial virus can be categorized into two antigenically and genetically distinct groups, RSV A and RSV B. The two groups often co-circulate but RSV A has been observed to dominate a majority of outbreaks²⁹⁻³¹. Within each group are genetic subgroups that are continually replaced over time^{29,32,33}. The clustering pattern of RSV sequences in the long term has been found to be more temporal than geographical³⁴⁻³⁶.

The spread of RSV occurs in seasonal patterns. In the temperate regions, seasonality is thought to be driven by low winter temperatures, in the tropics however, the drivers of seasonality are less well defined³⁷⁻³⁹. Given the ubiquitous nature of RSV, it is not uncommon to find other viral pathogens in circulation during an RSV season. Adenovirus and rhinovirus, which tend to be more year-round pathogens than seasonal, are frequently identified either to co-circulate or co-infect with RSV⁴⁰⁻⁴⁵. Influenza, human coronaviruses (HCoVs) and human metapneumovirus (HMPV), all of which have epidemic patterns of spread, have been observed to have overlapping epidemic timings with RSV in some settings^{37,41,46-50}. Influenza is more frequently observed with RSV in temperate regions and less so in the tropics^{37,43,44,51,52}. There is evidence of interactions between RSV and other pathogens. At a cellular level, facilitative interactions have been demonstrated between RSV and bacterial

pathogens⁵³⁻⁵⁶ while a competitive interaction has been demonstrated between RSV and influenza^{24,57}. At a host level, these interactions have sometimes been associated with increased disease severity or longer duration of hospital stay^{45,47,58-64}. In a case-control study that looked at the effect of therapeutic measures against RSV, it was found that there was no significant difference in the rate of occurrence of respiratory illness between the treatment and placebo group, however, within the placebo group, co-infections were more common than RSV infections⁶⁵. This points to a possible competitive interaction between RSV and other viruses that would result in pathogen replacement once an RSV vaccine is in effect. How cellular and host level interaction then scale up to population level dynamics is understudied, a situation which could be remedied by the use of mathematical models informed by experimental and epidemiological studies^{66,67}. Pathogen interactions, whether it is interactions between different strains of the same pathogen or between different pathogens, that have a population level impact on transmission dynamics, could also affect the effectiveness of vaccination strategies. The pneumococcal conjugate vaccine (PCV) has had 3 variants so far, PCV7, PCV10, PCV13 acting against 7, 10 and 13 serotypes of the bacteria *streptococcus pneumoniae*. Though evidence of a reduction in cases of invasive pneumococcal disease (IPD) and pneumonia has been observed^{68,69}, strain replacement which could lead to a mitigation of PCV vaccine efforts is a genuine concern⁷⁰⁻⁷². Though active surveillance is ongoing and several theories behind serotype replacement are being proposed^{71,73}, studies exploring possible multi-strain interactions could further elucidate the mechanism behind replacement. In contrast, evidence of immunomodulation following measles infection that results in a loss of immune memory to other infections has been used to explain the observed reduction in non-measles infectious disease mortality following the introduction of the measles mumps and rubella (MMR) vaccine⁷⁴. Consideration of only the pathogen that is the target of a vaccine without an understanding of its interactions with other pathogens could lead to an under-estimation or over-estimation of vaccine impact at the population level.

1.2. Control

As with most viral infections, there is no specific antiviral treatment for RSV infection. Severe cases requiring hospitalizations receive supportive therapy in the form of

administration of supplementary oxygen, mechanical ventilation and fluid replacement^{3,75}, such facilities are often unavailable in many recourse-poor settings.

Preventive therapy against severe disease in the form of a humanized monoclonal antibody Palivizumab is administered to high-risk infants. Despite Palivizumab being cost effective in preventing RSV disease in high-risk infants, such as those born prematurely or with congenital heart disease, in some high income countries⁷⁶, at approximately 4458 US dollars per child for the recommended 5-month course⁷⁷ it is not affordable for wide scale use in the general population of at-risk infants.

Vaccination of infants <6 months of age is faced by several challenges ranging from interference from maternal antibodies to immunological immaturity of the recipient and risk of enhanced disease upon subsequent natural infection⁷⁸⁻⁸⁰. In recent years however, there has been increased interest in developing a vaccine with three main target groups in mind; infants, pregnant women and the elderly⁸¹. Infants and the elderly would directly benefit from the vaccination while the aim of vaccinating pregnant women would be to provide passive protection to the infant. There are currently over fifty vaccines in different stages of development with the most advanced being a maternal vaccine for which phase III trials were recently completed⁸²⁻⁸⁴. The trial for ResVax™, which enrolled third-trimester pregnant women from countries in the Northern and Southern hemisphere, failed to meet its primary objective of a statistically significant reduction in medically significant RSV-LRTI in the infants born to the vaccinated women. The results did show reasonable vaccine efficacy against RSV LRTI hospitalizations, but timing of the vaccine relative to gestational age was a key determinant of efficacy. Results of timing relative to RSV seasonality were not presented⁸⁴. As with other vaccines in the pipeline ResVax™ was targeted towards the F protein as neutralizing antibodies generated against it have been shown to be protective against severe disease⁸⁵. Of the three RSV surface glycoproteins, the G gene is the most variable and often used for variant typing while the F gene is mostly conserved⁸⁶ and anti-F antibodies have been found to be cross-reactive between RSV A and B⁸⁷, meaning that a successful F-based vaccine should, in theory, work against RSV A and B. Whether a broad-spectrum RSV vaccine will have the intended effect given that interactions between RSV A and B have been shown to contribute to observed seasonal patterns, will be determined once any significant

interaction mechanisms have been considered while making projections of vaccine impact.

1.3. Models for improved understanding of disease transmission

Differences in transmission patterns and host social-demographic factors between locations and settings mean that a vaccine against RSV will have different efficacies and subsequently effectiveness. To gain a better understanding of disease transmission and the effect of an intervention, study investigators often use mathematical models⁸⁸. Models represent a hypothesis of infection transmission, and often, they are compared to data related to the particular disease under study with an aim of estimating model parameters that then allow for inference on transmission dynamics. Some interesting insights gained from modelling include: an analysis of rotavirus that highlighted the role of birth rates in driving the observed seasonal dynamics in the United States of America⁸⁹, parameterization of a model with contact data revealed the importance of contact patterns in identifying at-risk groups⁹⁰, and estimation of the basic reproductive number during an ongoing Ebola virus outbreak highlighted the need for increased bed capacity and case ascertainment if the outbreak was to be controlled⁹¹.

Increasingly, combinations of data streams are being used in models of infectious diseases, perhaps the most popular combination is that of epidemiological and genetic data⁹²⁻⁹⁴. This combination stems from the field of phylodynamics which involves the incorporation of ecological and evolutionary dynamics of a pathogen, based on the assumption that they occur at the same timescale⁹⁵. From a traditional epidemiology view, as opposed to molecular epidemiology, integration of genetic and epidemiological data has been used to infer transmission chains⁹⁶⁻⁹⁸, estimate reproductive numbers⁹⁹⁻¹⁰¹ and other quantities of interest.

A more detailed review of models in the context of RSV and phylodynamics approaches is provided in the next chapter.

Depending on the data available and the level of detail desired in the process represented by the model, mathematical models can vary in complexity and by

association, so can the inference technique. Inference can broadly be categorized as either Frequentist or Bayesian. With a Frequentist approach, point estimates of desired parameters are obtained with confidence intervals based on an empirical distribution of those estimates. With Bayesian, a distribution of parameter estimates is obtained (the posterior) based on the data, the model and prior information. Bayesian inference therefore not only allows for more information on the parameter, it also allows for inferring latent data variables through data augmentation^{102,103}. With increasing model and inference complexity comes increasing computational demands⁸⁸ therefore a balance has to be found that suits the specific study.

1.4. Data from the Kilifi household study cohort

1.4.1. Description

This PhD study was motivated in part by the availability of detailed epidemiological data. During a seasonal RSV outbreak beginning late 2009, members of 47 households in a rural location at the coast of Kenya were intensively followed up for a period spanning 6 months with an aim of recording the incidence of RSV and inferring who infects the household infant¹⁷. A household in this setting is described as composing of members who share a kitchen, in which case a household could be made up of extended family members distributed across several structures on the same compound. The definition of a household in this study is similar to what is used in national surveys in the country¹⁰⁴. In addition to households, a homestead is defined as a group of individuals living in the same compound and may be composed of one or more households. The study was conducted in Kilifi District, an administrative district within the larger Kilifi county. The Kilifi Health and Demographic Surveillance System (KHDSS), highlighted in yellow in the map shown in Figure 1. 1, was set up within the District as a record of births, pregnancies, migration events and deaths. The study area was selected to capture the majority of patients admitted to the main referral District Hospital¹⁰⁵. Matsangoni location, the household study site, as shown in Figure 1. 1, is at the northern tip of the KDHSS. As of 2009, Matsangoni location had a population of 14,998 individuals distributed across 1,835 homesteads. The average number of individuals per homestead was 8.2¹⁰⁶.

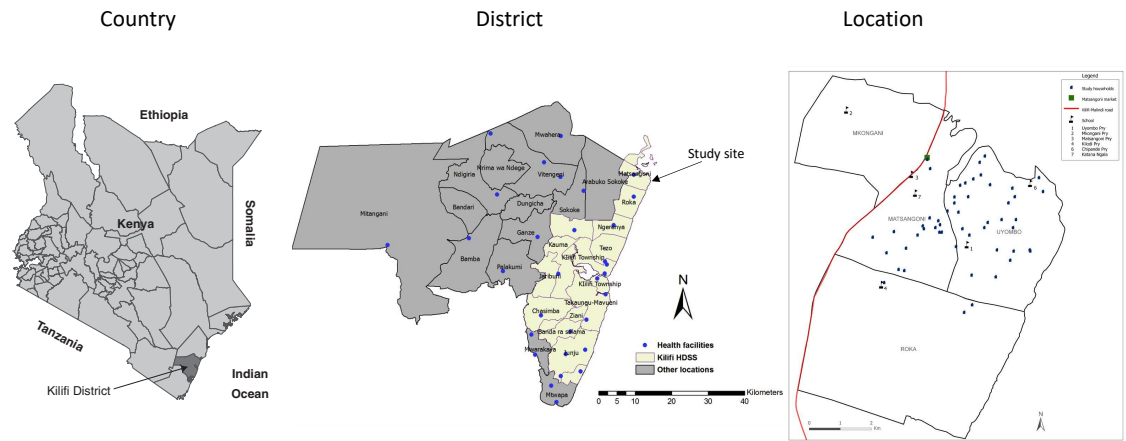


Figure 1. 1: Maps showing the household study site in geographical context as at September 2009.

Left: the map of Kenya shows the position of Kilifi district. Centre: Kilifi district which is further divided into administrative locations. The locations within the KHDSS are highlighted in yellow. Right: Matsangoni location where the household study was carried out¹⁰⁶.

Households were recruited on the basis of having an infant born after the previous RSV epidemic who had at least 1 elder sibling less than 13 years old. Members of the household had nasopharyngeal swab (NPS) samples and clinical data collected every 3-4 days. The samples were tested for RSV and other pathogens using an in-house real-time multiplexed polymerase chain reaction (PCR) assay¹⁰⁷. A sample was considered antigen positive if the PCR cycle threshold value was greater than 0 and less than 35. A Ct value of 0 is interpreted as a lack of genetic signal for the virus of interest while values above the threshold of 35 are interpreted as weak signals which could be due to environmental contamination. Near complete whole genome sequences were obtained for some of the RSV positive samples using the Illumina MisSeq platform^{108,109}.

A total of 47 households, consisting of 493 household members, were successfully followed up. The sizes of the household ranged from 4 to 37 members, with a median of 8. The largest distance between households was 6 km. Of the 493 household members, 272 were female and 221 were male, their age distributions are shown in Figure 1. 2.

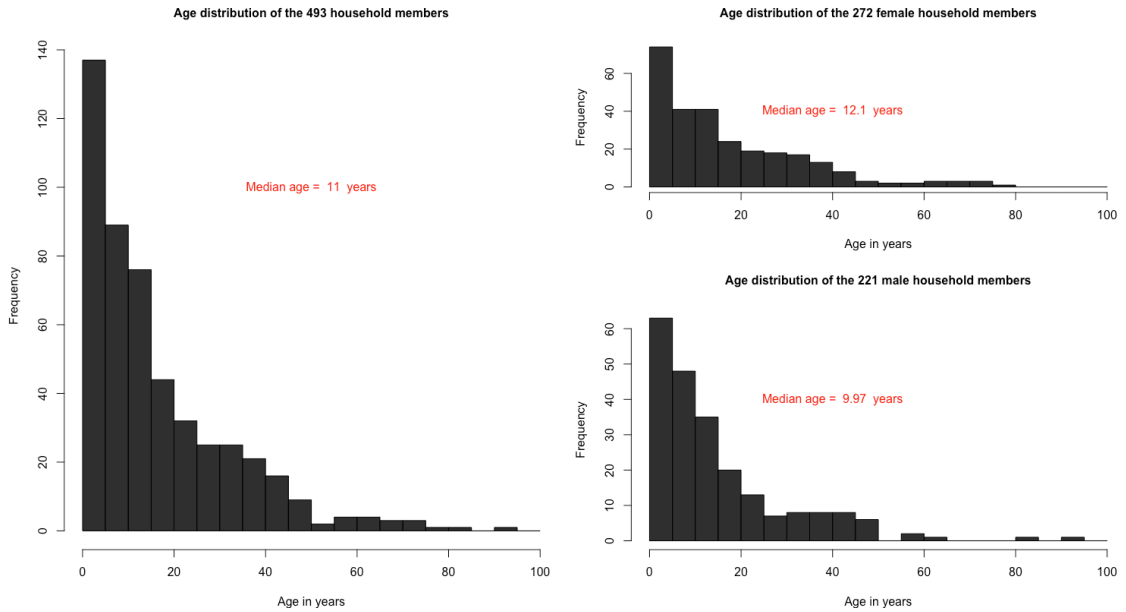


Figure 1. 2: Histograms showing the age distribution of the household members.

A total of 16928 samples were collected from 483 household members. The mean sampling interval was 3.7 days (SD=2.3). The median number of samples collected per participant was 41, the range of samples collected was between 1 and 48. Figure 1. 3 shows the distribution of the number of sampled collected per individual for all the participants, and for the different participant age groups. Of the 16928 samples, 1780 were positive for rhinovirus, 1274 for human coronavirus, 1232 for adenovirus and 537 for RSV^{41,110}.

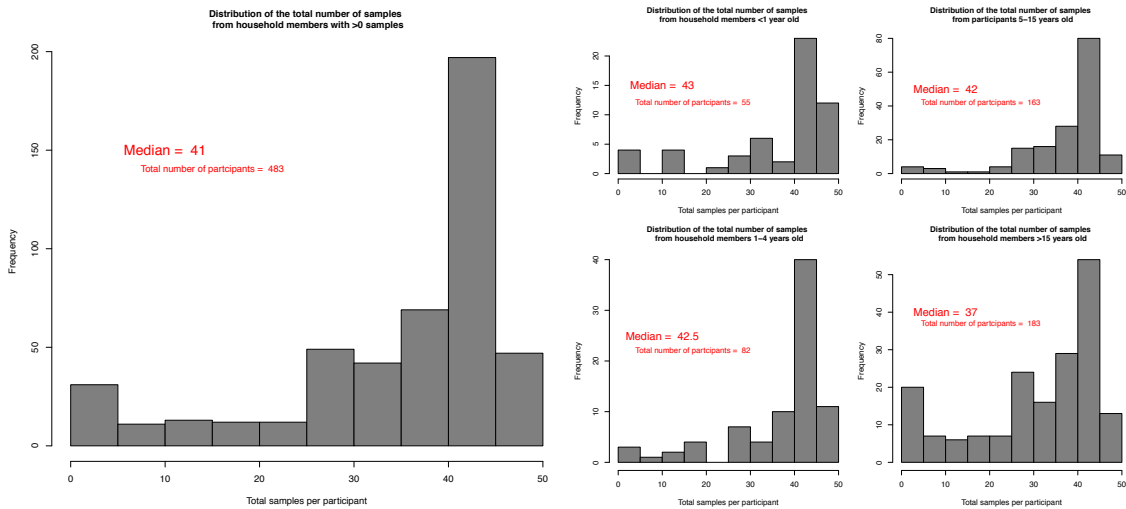


Figure 1. 3: Histograms showing the distribution of the total number of samples collected from the study participants.

1.4.1. Past work

Given that the data has been available for almost a decade, a substantial amount of work has already come out of it revealing several factors about RSV transmission. School-going children were linked to initiating household outbreaks leading to infant infection¹⁷. Bigger household size and infection with RSV group B, higher age, shorter duration of infection, lower peak viral load, absence of concurrent RSV infections within the household, and no prior human rhinovirus infections were found to be independently associated with increased risk of asymptomatic infection²⁵. Shedding durations were found to be longer than previously established, 11.2 days on average relative to a previous range of 3.9-7.4 days. The length of shedding durations was associated with age and severity of disease and reveal potential interactions with other respiratory viruses¹⁶. Individuals experiencing their first infection in an RSV season were found to shed more virus relative to secondary infections; <1 year old, symptomatic shedders and RSV A and B co-infected individuals were identified as most likely to transmit due to their relatively higher viral loads¹¹⁰. In this particular study setting respiratory virus infections and associated illness, are ubiquitous in households. The most frequently detected virus was rhinovirus (10.5% of samples), followed by human coronaviruses (HCoV) (7.5%), adenovirus (7.3%) and RSV (3.2%)¹⁰⁷. Relative to changes observed prior to an upper respiratory tract infection (URTI), the increase in the concentration of *Streptococcus pneumoniae* with RSV or rhinovirus infection was modest. This potentially pointed to the link between viral URIs and pneumococcal disease not being as straightforward as previously thought¹¹¹.

In 2016 the first rhinovirus genomes from Kenya were generated from samples collected by the household study¹¹². A joint epidemiological and phylogenetic analysis of rhinovirus sequences of the VP4/VP2 gene junction from 5 of the 47 households identified 3 species and 26 known subspecies/types in circulation. Repeat infections were common, with up to 8 at an individual level and 13 at a household level in a span of 6 months. Temporal clustering of types was observed within households. Almost all of the reinfections were with heterologous types, indicative of acquisition of immunity against homologous re-infections. Increasing age was associated with decreased infection rate, decreased re-infection rates, decreased duration of shedding and decreased proportion of symptomatic cases. Asymptomatic individuals were not

associated with decreased infectivity and there was evidence of competition between the species¹¹³. A recent phylogenetic analysis of human coronavirus (HCoV) sequences from cases in the household study and cases in an inpatient surveillance study found evidence that changes to the HCoV-NL63 genome are not immune driven¹¹⁴. A phylogenetic analysis of RSV A whole genome sequences from 13 households aimed at inferring transmission chains showed that cases arise more from within household spread rather than multiple introductions¹⁰⁹. A subsequent analysis of RSV A and RSV B whole genome sequences from 20 households found that where transmission pairs could be resolved, the source of infant infection was most likely either a toddler or a school-aged child. However, the conclusion of this study was that there was insufficient diversity in the genomic data for the sequence data alone to be able to fully resolve transmission chains hence they recommended an integrated data analysis combining the genetic data with epidemiological data¹⁰⁸. To somewhat concur with this, an analysis of shared minor variants derived from deep sequencing of some of the RSV samples failed to provide further resolution in the transmission chain beyond that derived from consensus whole genome sequences¹¹⁵.

1.5. Motivation for the PhD

At conceptualization, the aim of this PhD project was to gain a better understanding of RSV transmission dynamics by appropriately analysing a combination of epidemiological and genetic data from a longitudinal household study. The results of this work were intended to inform control strategies and future study designs. The specific objectives were:

- To review current literature in data integration methods and decide on a technique best suited for the data available and analytical objectives.
- To use all available genetic and epidemiological data (including social relationships) to gain a better understanding of the transmission dynamics of RSV in terms of (realized and potential) transmission chains and the factors affecting RSV viral diversity. This will be done in three parts: first using only epidemiological data to infer parameters, second using only epidemiological data to infer transmission chains and finally using epidemiological and genetic data to infer transmission chains.

- To identify the added benefit of viral genetic sequence data in understanding transmission of RSV; and use the methodology to inform on how to efficiently collect data for such analyses and obtain feedback on where we can continue to collect data for further inference.
- To use the integrated data framework developed to explore intervention strategies such as vaccination in terms of the target populations, timings and frequency.

That said, the data available still drove the direction of the analysis which was flexible enough to go in new directions without deviating too far from the initial purpose.

1.6. Computation

Most of the analysis in this thesis was carried out on the R platform¹¹⁶. R is a freely available software with a large community of users and contributors and therefore broad applicability, including analysis of genetic data. All the models used in the analyses presented in this thesis were formulated to suite the household data. Given that the data represented a densely sampled small subset of a community, the models are not overly complicated and it was therefore not necessary to apply complex inference techniques, the use of Metropolis-Hasting MCMC (MH-MCMC) was sufficient¹¹⁷. MH-MCMC was implemented successfully in R, in some instances using pre-existing packages and other times having to write my own functions. However, with the inclusion of sequence data, the model became more complicated including several iterative steps in calculating the likelihood function. In addition, given that sequence data was not available for all the cases I had to extend the MCMC algorithm to include data augmentation. Increased complexity in the model and inference technique meant the analysis was significantly slower to execute and would have taken weeks to run in R. As such, I moved to using the julia platform¹¹⁸. For simplicity I prepared the data in R, saved it at CSV files that were then used in julia. The julia results were then exported as CSV files into R as it has better developed graphics. Even in julia, the analysis still took several days to run, as such I outsourced the computing to a cluster computer based at the KEMRI-Wellcome Trust Research Programme in Kilifi, Kenya. Further details of the methods are provided in subsequent chapters. All the R and julia code used to generate the main analysis is freely available under the

GNU Lesser General Public License v3.0 and can be found at https://github.com/Ikadzo/HH_Transmission_Model.

1.7. Structure of the thesis

Though the PhD is by thesis, not by publication, the main results in this thesis are written in research paper format as some of them have already been published and the rest are intended for publication. Following this introductory chapter are five more chapters:

Review of models of RSV transmission dynamics and methods for including genetic information: This chapter is a review of models that have been used on RSV and approaches in phylodynamics.

Paper 1: Model based estimates of transmission of respiratory syncytial virus within households. Given the choice of approach arrived at after a review of the methods, this chapter is the first stand-alone analysis of the epidemiological data. It is written in paper format and has an abstract, introduction, methods, results, discussion and references section. This analysis has already been published¹¹⁹.

Paper 2: Integrating epidemiological and genetic data with different sampling densities into a dynamic model of RSV transmission. This chapter is an extension of the model presented in the previous chapter with modifications made to allow the use of genetic information. The analysis is presented in paper format; however, this work is yet to be submitted for publication.

Paper 3: A multi-pathogen model of infection investigating potential interactions between respiratory syncytial virus and coronavirus. This chapter is an extension of the model in Paper 1 modified to allow the use of data from multiple pathogens. The analysis is presented in paper format; however, this work is yet to be submitted for publication.

Discussion. Unlike the discussion subsections in the previous three chapters, this is an overall discussion tying together all the conclusions and implications for future work.

Appendices. This section contains supplementary information referenced in different sections of the thesis.

Prior to each chapter written for publication is a copy of a 'research paper cover sheet' signed by one of my supervisors and myself. This is a requirement for this thesis format. Given that the main results have been presented in paper format that are meant to be independently readable, there is some repetition in the content of each paper, particularly the introduction. The references are at the end of each chapter as opposed to being at the end of the thesis, even for the chapters that are not written in paper format.

1.8. Additional information

1.8.1. Ethics statement

For the data collection, informed written consent was obtained from all the study participants or their parents/guardian. The KEMRI-Scientific and Ethical Review Committee in Kenya provided ethical approval. The analysis presented here falls under the expected results from the original data collection study, however, additional ethical approval was obtained from the Observational / Interventions Research Ethics Committee at the London School of Hygiene and Tropical Medicine. The ethical approval letters can be found in appendix section A1: Ethical approval.

1.8.2. Training

To be able to meet the objectives of this PhD I attended several trainings and workshops in order to develop the required skills. I attended a 4-day course from 14-17 June 2016 at the London School of Hygiene and Tropical Medicine on model fitting and inference for infectious disease dynamics ran by Dr Sebastian Funk. The training used the R platform. I attended a 3-month distance-learning course on bioinformatics hosted at KEMRI-Wellcome Trust in Kilifi, Kenya. The course was run and sponsored by the Pan African Bioinformatics Network for H3Africa (H3ABioNet) and it ran from 6th July to 9th October 2016. I attended a one-day phylodynamics workshop on the 15th of February 2018 given by Professor Simon Frost of the Alan Turing Institute in the UK. I was part of a team of three facilitators of a Bayesian statistics workshop from 4-6 June 2018 on the Stan platform ran by Dr Michael Betancourt <https://betanalpha.github.io/>. I attended a 5-day interactive bioinformatics workshop from 20-26 September 2018 sponsored and ran by The Global Initiative for Neuropsychiatric Genetics Education in Research (GINGER). The most recent and final bit of my training was on coding on the julia platform, which I picked up while attending the 3-day EpiRecipes workshop from 1-3 October 2018 at the Alan Turing institute organized by Professor Simon Frost.

1.9. References:

1. Nair, H. *et al.* Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet* **375**, 1545–1555 (2010).
2. Shi, T., McLean, K., Campbell, H. & Nair, H. Aetiological role of common respiratory viruses in acute lower respiratory infections in children under five years: A systematic review and meta-analysis. *J. Glob. Health* **5**, 1–10 (2015).
3. O'Brien, K. L. *et al.* Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *Lancet* **6736**, 1–23 (2019).
4. Prasad, N. *et al.* Interactive effects of age and respiratory virus on severe lower respiratory infection. *Epidemiol. Infect.* **146**, 1861–1869 (2018).
5. Kestler, M., Muñoz, P., Mateos, M., Adrados, D. & Bouza, E. Respiratory syncytial virus burden among adults during flu season: an underestimated pathology. *J. Hosp. Infect.* **100**, 463–468 (2018).
6. Ohuma, E. O. *et al.* The natural history of respiratory syncytial virus in a birth cohort: the influence of age and previous infection on reinfection and disease. *Am. J. Epidemiol.* **176**, 794–802 (2012).
7. PARROTT, R. H. *et al.* Epidemiology of respiratory syncytial virus infection in Washington, DC II. Infection and disease with respect to age, immunologic status, race and sex. *Am. J. Epidemiol.* **98**, 289–300 (1973).
8. Hall, C. B., Walsh, E. E., Long, C. E. & Schnabel, K. C. Immunity to and frequency of reinfection with respiratory syncytial virus. *J. Infect. Dis.* **163**, 693–698 (1991).
9. Glezen, W. P., Taber, L. H., Frank, A. L. & Kasel, J. A. Risk of primary infection and reinfection with respiratory syncytial virus. *Am. J. Dis. Child.* **140**, 543–546 (1986).
10. Ochola, R. *et al.* The level and duration of RSV-specific maternal IgG in infants in Kilifi Kenya. *PLoS One* **4**, e8088 (2009).
11. Nyiro, J. U. *et al.* Defining the vaccination window for respiratory syncytial virus (RSV) using age-seroprevalence data for children in Kilifi, Kenya. *PLoS One* **12**, e0177803 (2017).
12. Cox, M. J., Azevedo, R. S., Cane, P. a, Massad, E. & Medley, G. F.

- Seroepidemiological study of respiratory syncytial virus in São Paulo state, Brazil. *J. Med. Virol.* **55**, 234–9 (1998).
13. Brüssow, H. *et al.* Age-related prevalence of serum antibody to respiratory syncytial virus in ecuadorian and german children. *J. Infect. Dis.* **163**, 680 (1991).
 14. L., L., AM., S., PJ., O. & FJ., C. Immunity to RSV in Early-Life. *Front. Immunol.* **5**, 466 (2014).
 15. Saso, A. & Kampmann, B. Vaccination against respiratory syncytial virus in pregnancy: a suitable tool to combat global infant morbidity and mortality? *The Lancet Infectious Diseases* **16**, e153–e163 (2016).
 16. Munywoki, P. K. *et al.* Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiol. Infect.* **143**, 804–12 (2015).
 17. Munywoki, P. K. *et al.* The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya. *J. Infect. Dis.* **209**, 1685–1692 (2014).
 18. Hall, C. B. *et al.* Respiratory syncytial virus infections within families. *N. Engl. J. Med.* **294**, 414–419 (1976).
 19. Jacoby, P., Glass, K. & Moore, H. C. Characterizing the risk of respiratory syncytial virus in infants with older siblings: A population-based birth cohort study. *Epidemiol. Infect.* **145**, 266–271 (2017).
 20. Heikkinen, T., Valkonen, H., Waris, M. & Ruuskanen, O. Transmission of respiratory syncytial virus infection within families. *Open Forum Infect. Dis.* **2**, ofu118 (2015).
 21. Scott, E. M. *et al.* Risk factors and patterns of household clusters of respiratory viruses in rural Nepal. *Epidemiol. Infect.* **147**, e288 (2019).
 22. French, C. E. *et al.* Risk of nosocomial respiratory syncytial virus infection and effectiveness of control measures to prevent transmission events: a systematic review. *Influenza and other Respiratory Viruses* **10**, 268–290 (2016).
 23. KAPIKIAN, A. Z. *et al.* AN OUTBREAK OF FEBRILE ILLNESS AND PNEUMONIA ASSOCIATED WITH RESPIRATORY SYNCYTIAL VIRUS INFECTION¹. *Am. J. Epidemiol.* **74**, 234–248 (1961).
 24. González-Parra, G. *et al.* A comparison of RSV and influenza in vitro kinetic parameters reveals differences in infecting time. (2018).

doi:10.1371/journal.pone.0192645

25. Munywoki, P. K. *et al.* Frequent Asymptomatic Respiratory Syncytial Virus Infections During an Epidemic in a Rural Kenyan Household Cohort. *J. Infect. Dis.* 1–8 (2015). doi:10.1093/infdis/jiv263
26. Mejías, A., Chávez-Bueno, S. & Sánchez, P. J. Respiratory syncytial virus prophylaxis. *Neoreviews* **6**, e26–e31 (2005).
27. Collins, P. L. & Graham, B. S. Viral and Host Factors in Human Respiratory Syncytial Virus Pathogenesis. *J. Virol.* **82**, 2040–2055 (2008).
28. Cane, P. a. Molecular epidemiology of respiratory syncytial virus. *Rev. Med. Virol.* **11**, 103–16 (2001).
29. Agoti, C. N. *et al.* Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. *J. Virol.* **89**, 11630–11642 (2015).
30. Calderón, A. *et al.* Genetic variability of respiratory syncytial virus A in hospitalized children in the last five consecutive winter seasons in Central Spain. *J. Med. Virol.* **89**, 767–774 (2017).
31. Yu, J. *et al.* Respiratory syncytial virus seasonality, Beijing, China, 2007–2015. *Emerg. Infect. Dis.* **25**, 1127–1135 (2019).
32. van Niekerk, S. & Venter, M. Replacement of previously circulating respiratory syncytial virus subtype B strains with the BA genotype in South Africa. *J. Virol.* **85**, 8789–97 (2011).
33. Cui, G. *et al.* Rapid replacement of prevailing genotype of human respiratory syncytial virus by genotype ON1 in Beijing, 2012–2014. *Infect. Genet. Evol.* **33**, 163–168 (2015).
34. Agoti, C. N. *et al.* Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *J. Virol.* **89**, 3444–3454 (2015).
35. Di Giallonardo, F. *et al.* Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses* **10**, 476 (2018).
36. Tan, L. *et al.* The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. *J. Virol.* **87**, 8213–26 (2013).
37. Li, Y. *et al.* Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: a systematic

- analysis. *Lancet Glob. Heal.* **7**, e1031–e1045 (2019).
38. Rodriguez-Martinez, C. E., Sossa-Briceño, M. P. & Acuña-Cordero, R. Relationship between meteorological conditions and respiratory syncytial virus in a tropical country. *Epidemiol. Infect.* **143**, 2679–2686 (2015).
 39. Obando-Pacheco, P. *et al.* Respiratory syncytial virus seasonality: A global overview. *J. Infect. Dis.* **217**, 1356–1364 (2018).
 40. Yan, X. *et al.* Clinical characteristics and viral load of respiratory syncytial virus and human metapneumovirus in children hospitalized for acute lower respiratory tract infection. *J. Med. Virol.* **89**, 589–597 (2017).
 41. Munywoki, P. K. *et al.* Continuous invasion by respiratory viruses observed in rural households during a respiratory syncytial virus seasonal outbreak in coastal Kenya. *Clin. Infect. Dis.* ciy313–ciy313 (2018).
 42. Taylor, S. *et al.* Respiratory viruses and influenza-like illness: Epidemiology and outcomes in children aged 6 months to 10 years in a multi-country population sample. *J. Infect.* **74**, 29–41 (2017).
 43. Prasad, N. *et al.* Interactive effects of age and respiratory virus on severe lower respiratory infection. *Epidemiol. Infect.* **146**, 1861–1869 (2018).
 44. Douros, K. *et al.* Evidence for respiratory viruses interactions in asymptomatic preschool-aged children. *Allergol. Immunopathol. (Madr)*. **47**, 260–264 (2019).
 45. Mazur, N. I. *et al.* Severity of respiratory syncytial virus lower respiratory tract infection with viral coinfection in HIV-uninfected children. *Clin. Infect. Dis.* **64**, 443–450 (2017).
 46. Friedman, N. *et al.* Human Coronavirus Infections in Israel: Epidemiology, Clinical Symptoms and Summer Seasonality of HCoV-HKU1. *Viruses* **10**, 515 (2018).
 47. da Silva, E. R. *et al.* Severe lower respiratory tract infection in infants and toddlers from a non-affluent population: Viral etiology and co-detection as risk factors. *BMC Infect. Dis.* **13**, 41 (2013).
 48. Pretorius, M. A. *et al.* Respiratory viral coinfections identified by a 10-Plex real-time reverse-transcription polymerase chain reaction assay in patients hospitalized with severe acute respiratory illness-South Africa, 2009-2010. *J. Infect. Dis.* **206**, S159-65 (2012).
 49. Njouom, R. *et al.* Viral etiology of influenza-like illnesses in Cameroon, January-

- December 2009. *J. Infect. Dis.* **206**, S29–S35 (2012).
50. Pilger, D. A., Cantarelli, V. V., Amantea, S. L. & Leistner-Segal, S. Detection of human bocavirus and human metapneumovirus by real-time PCR from patients with respiratory symptoms in Southern Brazil. *Mem. Inst. Oswaldo Cruz* **106**, 56–60 (2011).
 51. Çiçek, C. *et al.* Simultaneous detection of respiratory viruses and influenza A virus subtypes using multiplex PCR. *Mikrobiyol Bul* **48**, 652–660 (2014).
 52. Antalis, E. *et al.* Mixed viral infections of the respiratory tract; an epidemiological study during consecutive winter seasons. *J. Med. Virol.* **90**, 663–670 (2018).
 53. Greenberg, D. *et al.* Nasopharyngeal pneumococcal carriage during childhood community-acquired alveolar pneumonia: Relationship between specific serotypes and coinfecting viruses. in *Journal of Infectious Diseases* **215**, 1111–1116 (2017).
 54. Sande, C. J. *et al.* Airway response to respiratory syncytial virus has incidental antibacterial effects. *Nat. Commun.* **10**, 1–11 (2019).
 55. Avadhanula, V. *et al.* Respiratory Viruses Augment the Adhesion of Bacterial Pathogens to Respiratory Epithelium in a Viral Species-and Cell Type-Dependent Manner Downloaded from. *J. Virol.* **80**, 1629–1636 (2006).
 56. Weinberger, D. M. *et al.* Seasonal drivers of pneumococcal disease incidence: Impact of bacterial carriage and viral activity. *Clin. Infect. Dis.* **58**, 188–194 (2014).
 57. González-Parra, G. & Dobrovolny, H. M. Assessing Uncertainty in A2 Respiratory Syncytial Virus Viral Dynamics. *Comput. Math. Methods Med.* **2015**, 1–9 (2015).
 58. Harada, Y. *et al.* Does respiratory virus coinfection increases the clinical severity of acute respiratory infection among children infected with respiratory syncytial virus? *Pediatr. Infect. Dis. J.* **32**, 441–445 (2013).
 59. Hasegawa, K. *et al.* Multicenter study of viral etiology and relapse in hospitalized children with bronchiolitis. *Pediatr. Infect. Dis. J.* **33**, 809–813 (2014).
 60. Arruda, E. *et al.* The burden of single virus and viral coinfections on severe lower respiratory tract infections among preterm infants a prospective birth cohort study in Brazil. *Pediatr. Infect. Dis. J.* **33**, 997–1003 (2014).
 61. Asner, S. A., Rose, W., Petrich, A., Richardson, S. & Tran, D. J. Is virus coinfection

- a predictor of severity in children with viral respiratory infections? *Clin. Microbiol. Infect.* **21**, 264.e1-264.e6 (2015).
62. Jeannoël, M. *et al.* Microorganisms associated with respiratory syncytial virus pneumonia in the adult population. *Eur. J. Clin. Microbiol. Infect. Dis.* **38**, 157–160 (2019).
 63. Suárez-Arrabal, M. C. *et al.* Nasopharyngeal bacterial burden and antibiotics: Influence on inflammatory markers and disease severity in infants with respiratory syncytial virus bronchiolitis. *J. Infect.* **71**, 458–469 (2015).
 64. Hishiki, H. *et al.* Incidence of bacterial coinfection with respiratory syncytial virus bronchopulmonary infection in pediatric inpatients. *J. Infect. Chemother.* **17**, 87–90 (2011).
 65. Blanken, M. O. *et al.* Respiratory Syncytial Virus and Recurrent Wheeze in Healthy Preterm Infants. *N. Engl. J. Med.* **368**, 1791–1799 (2013).
 66. Opatowski, L., Baguelin, M. & Eggo, R. M. Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling. *PLoS Pathog.* **14**, (2018).
 67. Nikin-Beers, R., Blackwood, J. C., Childs, L. M. & Ciupe, S. M. Unraveling within-host signatures of dengue infection at the population level. *J. Theor. Biol.* **446**, 79–86 (2018).
 68. Silaba, M. *et al.* Effect of 10-valent pneumococcal conjugate vaccine on the incidence of radiologically-confirmed pneumonia and clinically-defined pneumonia in Kenyan children: an interrupted time-series analysis. *Lancet Glob. Heal.* **7**, e337–e346 (2019).
 69. Hammitt, L. L. *et al.* Effect of ten-valent pneumococcal conjugate vaccine on invasive pneumococcal disease and nasopharyngeal carriage in Kenya: a longitudinal surveillance study. *Lancet* **393**, 2146–2154 (2019).
 70. Miller, E., Andrews, N. J., Waight, P. A., Slack, M. P. E. & George, R. C. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: An observational cohort study. *Lancet Infect. Dis.* **11**, 760–768 (2011).
 71. Lewnard, J. A. & Hanage, W. P. Making sense of differences in pneumococcal serotype replacement. *The Lancet Infectious Diseases* **19**, e213–e220 (2019).

72. Kwambana-Adams, B. *et al.* Rapid replacement by non-vaccine pneumococcal serotypes may mitigate the impact of the pneumococcal conjugate vaccine on nasopharyngeal bacterial ecology. *Sci. Rep.* **7**, 1–11 (2017).
73. Lo, S. W. *et al.* Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *Lancet Infect. Dis.* **19**, 759–769 (2019).
74. Mina, M. J., Metcalf, C. J. E., De Swart, R. L., Osterhaus, A. D. M. E. & Grenfell, B. T. Long-term measles-induced immunomodulation increases overall childhood infectious disease mortality. *Science (80-.)*. **348**, 694–699 (2015).
75. Behzadi, M. A. & Leyva-Grado, V. H. Overview of Current Therapeutics and Novel Candidates Against Influenza, Respiratory Syncytial Virus, and Middle East Respiratory Syndrome Coronavirus Infections. *Front. Microbiol.* **10**, 1327 (2019).
76. Resch, B. *et al.* Cost-effectiveness of palivizumab for respiratory syncytial virus infection in high-risk children, based on long-term epidemiologic data from Austria. *Pediatr. Infect. Dis. J.* **31**, e1–e8 (2012).
77. Kamal-Bahl, S., Doshi, J. & Campbell, J. Economic analyses of respiratory syncytial virus immunoprophylaxis in high-risk infants: A systematic review. *Arch. Pediatr. Adolesc. Med.* **156**, 1034–1041 (2002).
78. Siegrist, C.-A. & Lambert, P. H. Maternal immunity and infant responses to immunization: factors influencing infant responses. *Dev. Biol. Stand.* **95**, 133–139 (1997).
79. KAPIKIAN, A. Z., MITCHELL, R. H., CHANOCK, R. M., SHVEDOFF, R. A. & STEWART, C. E. An epidemiologic study of altered clinical reactivity to respiratory syncytial (RS) virus infection in children previously vaccinated with an inactivated RS virus vaccine. *Am. J. Epidemiol.* **89**, 405–421 (1969).
80. Basha, S., Surendran, N. & Pichichero, M. Immune responses in neonates. *Expert Rev. Clin. Immunol.* **10**, 1171–1184 (2014).
81. Anderson, L., Dormitzer, P. & Nokes, D. Strategic priorities for respiratory syncytial virus (RSV) vaccine development. *Vaccine* **31 Suppl 2**, B209-15 (2013).
82. PATH. RSV Vaccine Snapshot - PATH Vaccine Resource Library. (2015). Available at: <http://vaccineresources.org/details.php?i=1562>. (Accessed: 15th July 2019)
83. Novavax. A Study to Determine the Safety and Efficacy of the RSV F Vaccine to

Protect Infants Via Maternal Immunization. *NIH U.S. National Library of Medicine* (2018). Available at:

<https://clinicaltrials.gov/ct2/show/record/NCT02624947>. (Accessed: 16th July 2019)

84. Novavax Inc. Novavax Announces Topline Results from Phase 3 Prepare™ Trial of ResVax™ for Prevention of RSV Disease in Infants via Maternal Immunization. *Globe Newswire* 1–3 (2019).
85. Graham, B. S., Modjarrad, K. & McLellan, J. S. Novel antigens for RSV vaccines. *Current Opinion in Immunology* **35**, 30–38 (2015).
86. Tapia, L. I. *et al.* Gene sequence variability of the three surface proteins of Human Respiratory Syncytial Virus (HRSV) in Texas. *PLoS One* **9**, 90786 (2014).
87. Piedra, P. A., Jewell, A. M., Cron, S. G., Atmar, R. L. & Paul Glezen, W. Correlates of immunity to respiratory syncytial virus (RSV) associated-hospitalization: Establishment of minimum protective threshold levels of serum neutralizing antibodies. in *Vaccine* **21**, 3479–3482 (Elsevier, 2003).
88. Heesterbeek, H. *et al.* Modeling infectious disease dynamics in the complex landscape of global health. *Science (80-.)*. **347**, aaa4339–aaa4339 (2015).
89. Pitzer, V. E. *et al.* Demographic variability, vaccination, and the spatiotemporal dynamics of rotavirus epidemics. *Science* **325**, 290–4 (2009).
90. Mossong, J. *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74 (2008).
91. Lewnard, J. A. *et al.* Dynamics and control of Ebola virus transmission in Montserrado, Liberia: A mathematical modelling analysis. *Lancet Infect. Dis.* **14**, 1189–1195 (2014).
92. Tong, S. Y. C. *et al.* Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res.* **25**, 111–118 (2015).
93. Lau, M. S. Y., Marion, G., Streftaris, G. & Gibson, G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLoS Comput. Biol.* **11**, (2015).
94. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput. Biol.* **10**, e1003457 (2014).
95. Grenfell, B. T. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science (80-.)*. **303**, 327–332 (2004).

96. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput. Biol.* **15**, e1006930 (2019).
97. Campbell, F. *et al.* outbreaker2 : a modular platform for outbreak reconstruction. **19**, (2018).
98. Cori, A. *et al.* A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLoS Comput. Biol.* **14**, 1–22 (2018).
99. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evol.* **34**, 2982–2995 (2017).
100. Gomes Naveca, F. *et al.* Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. (2019).
doi:10.1371/journal.pntd.0007065
101. Kenah, E., Britton, T., Halloran, M. E. & Jr, I. M. L. Molecular Infectious Disease Epidemiology : Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. 1–29 (2016). doi:10.1371/journal.pcbi.1004869
102. Andrews, M. & Baguley, T. Bayesian data analysis. in *The Cambridge Encyclopedia of Child Development* 165–169 (Chapman and Hall/CRC, 2017).
doi:10.1017/9781316216491.030
103. Rao, C. R., Miller, J. P. & Rao, D. C. *Essential Statistical Methods for Medical Statistics. Essential Statistical Methods for Medical Statistics* **27**, (Elsevier, 2011).
104. Kenya, R. of. The 2009 Kenya Population and Housing Census: Volume II, Population and Household Distribution by Socio-Economic Characteristics. (2010).
105. Scott, J. A. G. *et al.* Profile: The Kilifi health and demographic surveillance system (KHDSS). *Int. J. Epidemiol.* **41**, 650–657 (2012).
106. Munywoki, P. K. Transmission of Respiratory Syncytial Virus in Households : Who Acquires Infection From Whom. (Open University UK, 2013).
107. Munywoki, P. K. *et al.* Continuous invasion by respiratory viruses observed in rural households during a respiratory syncytial virus seasonal outbreak in coastal Kenya. *Clin. Infect. Dis.* **67**, 1559–1567 (2018).
108. Agoti, C. N. *et al.* Genomic analysis of respiratory syncytial virus infections in households and utility in inferring who infects the infant. *Sci. Rep.* **9**, 10076

- (2019).
109. Agoti, C. *et al.* Transmission patterns and evolution of RSV in a community outbreak identified by genomic analysis. *Virus Evol.* (2017). doi:In print
 110. Wathuo, M., Medley, G. F., Nokes, D. J. & Munywoki, P. K. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res.* **1**, 27 (2016).
 111. Morpeth, S. C. *et al.* Impact of viral upper respiratory tract infection on the concentration of nasopharyngeal pneumococcal carriage among Kenyan children. *Sci. Rep.* **8**, 11030 (2018).
 112. Agoti, C. N. *et al.* Human Rhinovirus B and C Genomes from Rural Coastal Kenya. *Genome Announc.* **4**, 751–767 (2016).
 113. Kamau, E. *et al.* An intensive, active surveillance reveals continuous invasion and high diversity of rhinovirus in households. *J. Infect. Dis.* **219**, 1049–1057 (2019).
 114. Kiyuka, P. K. *et al.* Human coronavirus NL63 molecular epidemiology and evolutionary patterns in rural coastal Kenya. *J. Infect. Dis.* **217**, 1728–1739 (2018).
 115. Githinji, G. *et al.* Assessing the utility of minority variant composition in elucidating RSV transmission pathways. *bioRxiv* 411512 (2018). doi:10.1101/411512
 116. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
 117. Roberts, G. O. & Rosenthal, J. S. Examples of Adaptive MCMC. *J. Comput. Graph. Stat.* **18**, 349–367 (2009).
 118. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **59**, 65–98 (2017).
 119. Kombe, I. K., Munywoki, P. K., Baguelin, M., Nokes, D. J. & Medley, G. F. Model-based estimates of transmission of respiratory syncytial virus within households. *Epidemics* **27**, 1–11 (2019).

2. Review of models of RSV transmission dynamics and methods for including genetic information

In this chapter I present my review of the methods in two parts: first a review of some of the models that have been used in the context of RSV, second, I review models used in phylodynamics analysis. In the final section I present the logic in deciding which methods to apply to the data available.

2.1. Models of respiratory syncytial virus

Mathematical models of infectious disease transmission (from here on referred to simply as models), as mentioned in the previous chapter, are often used to improve understanding on infection and/or disease dynamics, following from which the same tools can be used to make projections for the future with or without an intervention. Models allow one to represent their assumptions of the natural history of a disease in a manipulatable system of equations, a fundamental element of which is the feedback process between the number of infectious hosts and the risk of infection to the susceptible population. Models at the population level are often compartmental, meaning individuals are grouped into compartments representing their state relative to the infection under study. The most basic is the deterministic, ordinary differential equation (ODE), SIR model. In this model individuals are assumed to be susceptible (S), they get infected at a rate λ and move to the infected (I) class and after a duration of infection $\frac{1}{\gamma}$ they recover into the R class where they have lifelong immunity. The process of transition from one compartment to the next is represented by a system of ODE's. Figure 1. 4 shows a flow chart for the SIR model, the accompany equations and sample model projections.

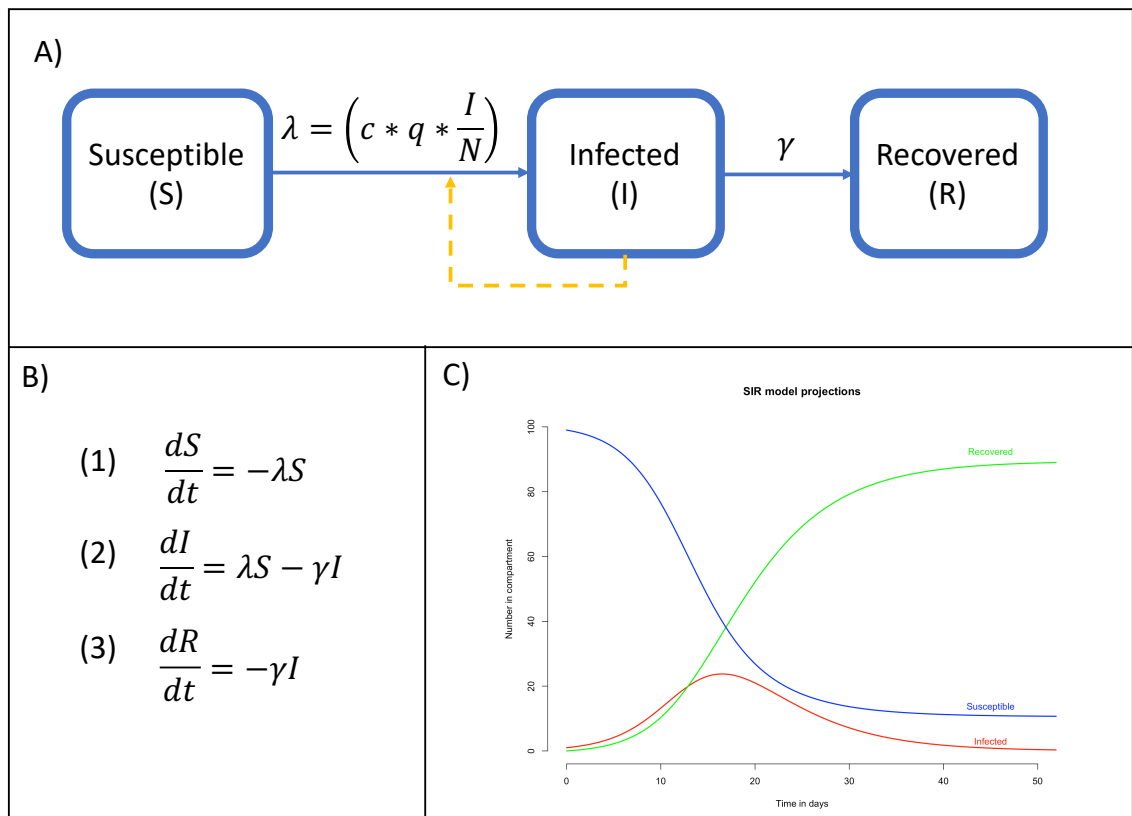


Figure 1. 4: A flow chart for the SIR model, the model equations and sample deterministic model projections.

Panel A): The main disease states in the SIR model are represented by the blue boxes, the transitions are shown by the blue arrow and the rates of transitions between compartments and shown by symbols on top of the blue arrows. The dashed yellow arrow shows the feedback process between the size of the infected compartment and the rate of exposure λ . Panel B): the set of ODE's for the SIR model showing the rates of change in each compartment (dS, dI, dR) per change in time (dt). Panel C): Sample model projections for the SIR model, the values of the parameters used were $c=10/\text{person}/\text{day}$, $q= 0.05$ and $\gamma = 0.2/\text{person}/\text{day}$.

Given a population of size N , the rate of exposure λ is determined by an individual's contact rate (c), the probability that a contact is with an infectious person (I/N) and the probability of transmission given an infectious contact (q). Contact rates can be density dependent (increase with increase in population density) or frequency dependent (do not change with population density). A deterministic model does not allow for stochasticity when making predictions, as such given the same value for the

model parameters, the deterministic SIR model will project the same time series of cases an example of which is shown in Figure 1. 4.

If the disease of interest has a latency period, where an individual is infected but not yet infectious, then an exposed compartment **E** is introduced into the SIR structure resulting in an SEIR model. If immunity to infection is not lifelong and individuals can become susceptible again after some time, then a transition is introduced out of the **R** compartment back to the **S** compartment, giving the SIRS model. If the infection does not confer any immunity and individuals are susceptible again as soon as they stop shedding, then the **R** compartment is dropped from the model and from **I**, individuals go back into the **S** compartment, giving the SIS model. If the infection confers partial immunity, i.e. previously infected individuals are less susceptible to future infections, then the **R** compartment is replaced by an **S₂** compartment of reduced susceptibility. These extensions of the SIR are depicted in Figure 1. 5.

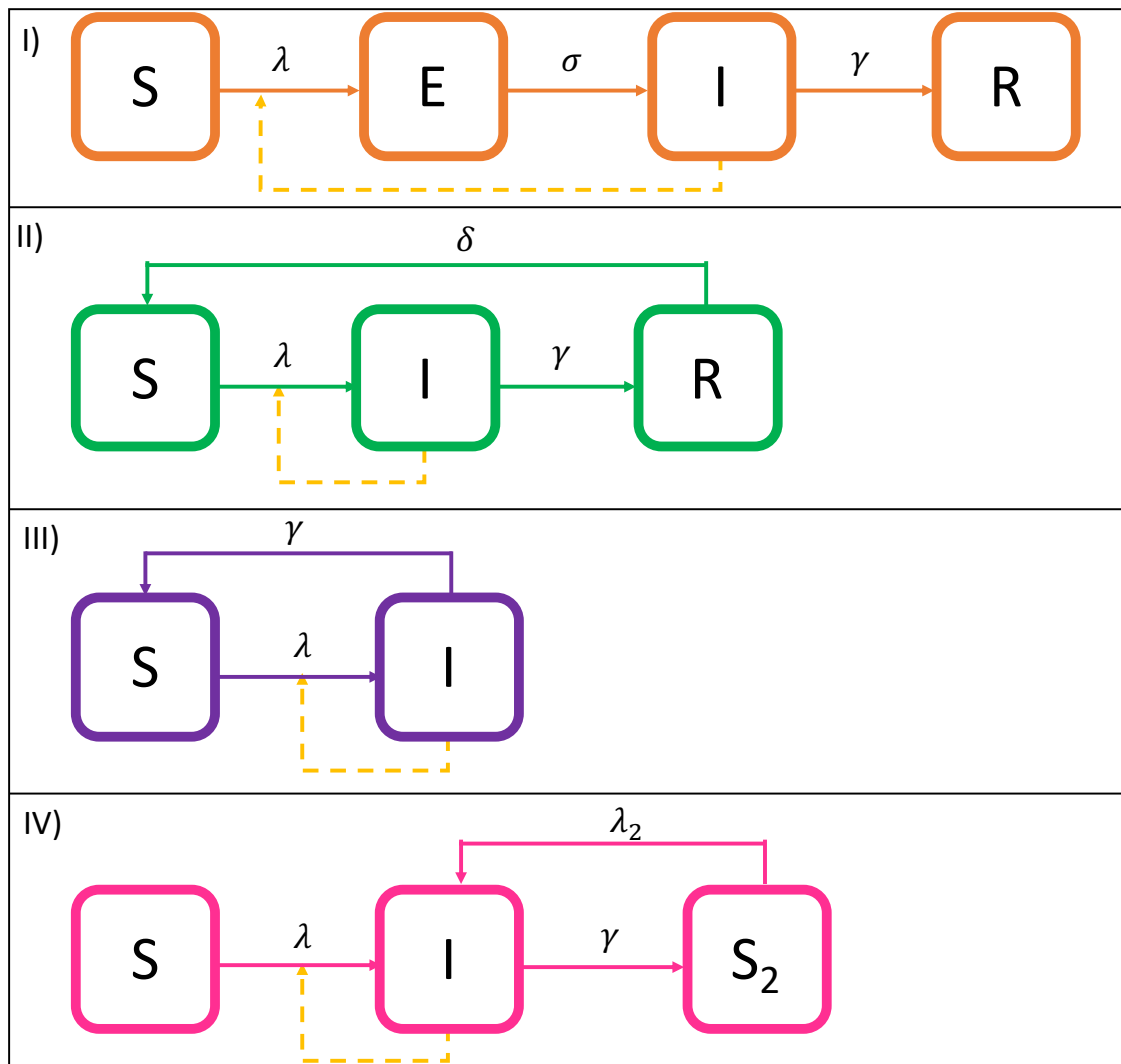


Figure 1. 5: Possible extensions of the basic SIR model representing different assumptions about the natural history of an infection.

Panel I): The SEIR model which assumes a period of latency prior to onset of infectiousness. Panel II): The SIRS model that assumes immunity to infection is transient and is lost after a period = $1/\delta$. Panel III): The SIS model that assumes no immunity following infection. Panel IV): The SIS₂ model that assumes individuals develop partial lifelong immunity following infection.

Numerous other variations of this simple model are possible and have been made to explore a broad range of assumptions^{1,2}. The host population can also be modelled using network models or individual based models, depending on what dynamics are of interest and the nature of the population of interest².

Though epidemiological studies investigating factors associated with RSV transmission have provided some useful insights, it is through mathematical models that the interaction of such factors can be explored, and the overall resultant dynamics analysed. In this section, some of the models that have been used in the context of RSV are reviewed, highlighting the main assumptions that went into them and the main inferences that were drawn. The first set of models to be discussed are those that take an international view of RSV by analysing data from multiple countries. The advantage of such analysis is that they could potentially pick up on the broader RSV specific characteristics and identify differences between countries that could be important for how interventions are planned. RSV occurs in seasonal patterns, but the exact drivers of seasonality are not well defined. It is common practice for mathematical models to use trigonometric functions to force oscillations in model projections (seasonal forcing applied to the rate of exposure, λ), nonetheless, even in doing so a comparative analysis of the forcing functions for different locations can give some insight. One of the earlier studies was conducted by Weber *et al.* using data from the Gambia, USA, Finland and Singapore³. The study explored structural uncertainty by fitting two different compartmental ODE models with seasonal forcing. Their results highlighted the sensitivity of the inferred transmissibility of RSV to the model structure, the model assuming transient full immunity and lifelong partial immunity following primary infection gave higher values of the basic reproductive number compared to the SIRS model of transient full immunity. They also found evidence that different locations have different factors driving seasonality. An attempt to use rainfall and temperature data to explain seasonality in the tropical countries, was unsuccessful, leading the authors to conclude that perhaps it is a combination of meteorological and social factors driving the seasonal patterns. In 2005 White *et al.* analysed data from the UK and Finland using a deterministic compartmental model that distinguished between RSV A and RSV B⁴. They found that in addition to seasonal forcing, the interactions between RSV groups were required to produce the observed seasonal patterns. The group interactions were homogenous across locations, but the seasonality parameters were not. In addition, they found that the data supported the existence of transient partial immunity following infection more so for homologous group reinfection than heterologous. They also estimated that RSV A was about 8% more transmissible than RSV B, perhaps providing an explanation for A being the

dominant group in most epidemics. This study brings out the importance of distinguishing between the RSV groups when investigating transmission determinants, the concluding remarks advocated for longitudinal cohort data in order to obtain a biologically realistic multigroup model for RSV.

Both the Weber and White models were unable to determine the role of young children in RSV epidemiology and they highlighted this as an open question. In a second multi-country analysis using data from 9 locations distributed over 7 country locations White *et al.* used nested deterministic compartmental models to gain a better understanding of RSV natural history. The study found that the data supported the existence of lifelong partial immunity however waning immunity also provided visually good fits to the data. They used a seasonal forcing function, but based on the estimates of the peak timing, they found no indication that temperature had a role to play therefore there was no clear indication of what could be driving the observed differences in seasonality among the countries in the data⁵. The amplitude parameter in the seasonal forcing function not only varied by location but also by model structure. Whereas most of the locations seem to agree on the order of best fitting models, Finland, which has biennial rather than annual epidemic cycles had a different order. It would appear that if the driving forces of seasonality cannot be accounted for, different locations could end up supporting different model structures and hence natural histories in particular, the duration of immunity. The authors hypothesize that there might be similar seasonality drivers across different pathogens e.g. RSV and measles, but the variation in timing could be due to differences in immune mechanisms between the pathogens. Unfortunately, more models looking at multicounty data were not found in the literature, there were however statistical analyses that have found some associations between RSV and meteorological factors. Bloom-Feshbach *et al.* conducted a time series analysis of clinical, geographical and socio-economic data from over 50 countries and compared the variation between RSV and influenza epidemiology⁶. The study found that RSV cases peaked in the winter months in temperate countries but the pattern driving peak incidence in the tropics was less apparent. In another time series analysis of data from two locations in the Philippines and one in Japan, the study concluded that seasonality had more to do with the amplitude and variation in climatic factors than with actual absolute values⁷.

At a narrower geographical scale, several other studies have attempted to understand seasonality drivers and immune mechanisms. In the USA Pitzer *et al*, combined a statistical and dynamic age-structured compartmental model with seasonal forcing and lifelong partial immunity to analyse data from different states. They found that the strongest link to seasonal variation in RSV was potential evapotranspiration, which is a measure of the demand for water from the atmosphere⁸. There was an indication from one of the states that birth rates could also be driving transmission, but this, along with being able to tease apart climatic effects from social patterns, required further studies. Interestingly a regression analysis on data spanning 16 years from a single state in the US found that early epidemic timing was significantly associated with higher population density⁹. Models of RSV transmission fitted to data from Spain have found that most cases occurred in the winter months^{10,11} and through a combination of climatic and clinical data established an association between mean temperature and atmospheric pressure, and RSV activity¹². Studies from Scotland trying to understand drivers of RSV transmission found that RSV transmission was favourable when daily ranges in humidity were narrow¹³. Paynter *et al* fit a compartmental model to data from the Philippines spanning 5 years¹⁴. They found that the peak in the transmissibility parameter in the model preceded the peak in cases and, intriguingly, seasonal malnutrition and rainfall could be driving transmission.

Using an age-structured compartmental model with seasonal forcing and complete but waning immunity calibrated to clinical data from Western Australia spanning a 6-year period, Moore *et al* were able to reproduce the observed biennial seasonal patterns with estimated infectious period ranging from 8-11 days and the duration of immunity being 160 days¹⁵. Given the longer duration of time in-between epidemics, it is perhaps intuitively understandable that a model in such a setting would support the idea of waning complete immunity, however, it is interesting that when White⁵ fit a nested model to Finland's biennial data, an assumption of partial lifelong immunity gave the best fit. This further strengthens the idea that unless seasonality drivers at a given local area are well understood, contrasting natural histories of RSV could be inferred.

If the natural history of RSV inferred from models can be so variable, what impact is this likely to have on vaccination programs? Pan-Ngum *et al* showed that despite structural uncertainty in models, reflecting uncertainty in immunity development and loss following natural infection, vaccines that act to reduce the duration of infection and infectivity are predicted to have the largest impact on cases¹⁶. Surprisingly, maternal vaccination was predicted to have only moderate effects. This is not the only modelling study to suggest that a maternal vaccination might not be the most optimal strategy. A deterministic compartmental model by Kinyanjui *et al* sought to establish the optimal age to vaccinate against RSV, with a particular focus on the inherent mixing assumptions¹⁷. Results using a contact matrix derived from contact diaries were compared to results obtained using a synthetic contact matrix. Though both structures support the vaccination of older infants 5-10 months old, which would result in significant herd immunity, the two different contact matrices predict different mechanisms of vaccine action; the synthetic matrix is such that contacts patterns are dominated by children and so the vaccine works through preventing children from transmitting to the very young, as such it works by reducing secondary infections rather than primary. With the diary-based matrix the force of infection by age shows that primary cases drive transmission as such the vaccination strategy works by impacting primary cases. This work highlights the importance of mixing assumptions and social structure as additional factors that affect model predictions. To further explore the effect of social structure, Poletti *et al.* simulated RSV infection on a synthetic population grouped according to households and schools¹⁸. Given estimated transmission chains, this study found that household transmission was responsible for about 38.3% (35.4,40.9) of infant infection and that school-age children played a key role introducing infection to the household. The impact of vaccination was dependent on the duration of immunity but in general, second to infant vaccination at 3 months of age, annual vaccination of all primary school students (aged 7 to 15 years on average) would result in preventing a significant proportion of infant and community infection. Vaccinating pregnant mothers to protect the infant was effective if it provided an additional 4 months of maternally acquired immunity, beyond the 4 months assumed to occur naturally (i.e. a total of 8 months). Despite infant vaccination being optimal in this study, the risk of maternal antibody interference means 3 months might be too early to vaccinate, in fact Nyiro *et al.* used a catalytic compartmental

model to analyse the seroprevalence profile of children aged between 0 and 145 months and found that targeting vaccination at infants 5 months and older would archive the highest rate of seroconversion¹⁹. Yamin *et al.* used contact data, information on viral load during the course of an infection and data on behaviour change due to RSV symptoms, to parameterize the force of infection in an age structured compartmental model with seasonal forcing, waning immunity and altered infectiousness and disease severity following primary infection²⁰. They calibrated the model to RSV incidence in 5 states in the US and found that vaccinating children <5 years old was the most effective strategy, owing to the fact that they were more infectious (higher viral load and longer durations of infections) and had more frequent contacts. There was, however, geographical variability in predicted vaccine effectiveness across states part of which was attributed to differences in seasonal patterns and population demography. In work that looked at 11 RSV seasons in the USA, Goldstein *et al.* found that children aged between 3 and 6 years old played an important role in propagating the RSV epidemics²¹. These studies suggest that even in the face of uncertainty in how immunity to RSV is built up, a vaccine targeted at the group most likely to infect others would have the biggest impact on overall transmission. It is therefore crucial to establish generalizable transmission chains in a given setting. Several epidemiological studies have found an association between having an older sibling and an increased risk of infant infection, though no direct infection link between the older siblings and the infant was confirmed²²⁻²⁴. These results answer questions raised over 10 years ago by Weber³ and White⁴ on the role of children in RSV epidemiology.

There have also been modelling studies predicting the benefits of a maternal vaccination. Hogan *et al* calibrated an age-structure compartmental model to data from an electronic birth cohort followed up during the period from 1996 to 2012²⁵. In this study they found that a maternal vaccination would lead to a 6-37% reduction in hospitalization in the <3-month-old age group and 30-46% reduction in the 3-5-month-old age group. An analysis by Scheltema *et al.* modelled antibody kinetics starting from trans-placental transfer to waning post-delivery using parameters derived from literature²⁶. They then looked at RSV cases from hospital admission in the UK and the Netherlands and reported deaths in the literature from 20 countries. Based on the age

of the infant at the time of the RSV related outcome (hospitalization or death), and the inferred antibody dynamics from their model, they estimated how many cases would have been averted had a vaccine been administered to the mothers in the third trimester of pregnancy. They found that at least 62% of admissions would have been prevented in the UK and 76% in the Netherlands, while globally, at least 29% of the reported deaths would have been avoided. Similar to the strategy applied to pertussis, Brand *et al.* explored the benefits of a two-vaccine strategy aimed at pregnant women and their household cohabitants. Calibrated to data from a low-income country, the study found that a 50% reduction in RSV hospitalizations is possible if the maternal vaccine effectiveness can achieve 75 days of additional protection for new-borns combined with a 75% coverage of their birth household co-inhabitants (~7.5% population coverage)²⁷.

Other than the study by White *et al.*²⁸, few others have explored interactions between RSV groups. Through estimating group specific reproductive numbers, Otomaru *et al* did find that the range for RSV A was 0.92-1.33 and that for RSV B was 1.04 -1.76, variation being due to epidemic under study and the location. Where time and location results were comparable, RSV A had a slightly higher reproductive number than RSV B, consistent with findings from the White *et al.* study. These estimates were much lower than expected however, the method used to derive them did not include any assumptions of immunity, which from the previous studies, were noted to influence model inference. Going a level beyond looking at RSV group dynamics, Chan *et al.* were interested in understanding drivers of viral diversity and used a compartmental model to establish that viral populations in large cities with dense host populations are more likely to generate new variants²⁹. Comparing RSV to other pathogens Gonzales *et al.* built models looking at RSV at the molecular level and compared *in vitro*³⁰ and *in vivo*³¹ infections of RSV and influenza. The first study found that as a result of RSV having a slower rate of spread from cell to cell, RSV titres increased at a slower rate and reached peak value much later than influenza. The second study found that the infectious cell lifespan was shorter for RSV than influenza. These interactions could shape population level dynamics.

Going beyond using models to infer transmission dynamics and predict vaccine impact, tools for forecasting and tracking ongoing infection trends have been developed in the USA. Using data from 10 RSV seasons, Reis and Shaman built a forecasting tool with 70% accuracy at predicting the peak of an outbreak 4 weeks in advance³². In another first, Oren *et al.* attempted to track the trends in RSV cases using internet search data and found that the regression based method worked fairly well³³.

Taken together, it is clear that RSV has different seasonal transmission dynamics in different climate zones, different countries and even within a country, local areas can have different patterns of transmission. Questions still exist on the exact drivers of transmission, more so from tropical low-income countries from which there is a paucity of data. In the temperate regions, the role of lower temperatures especially in winter months seems to be quite clear. However, climatic factors alone are not enough to explain variations in seasonal patterns and other demographic factors such as birth rates have also been proposed. In a theoretical modelling study, Hogan *et al.* built a model that was able to replicate 4 distinct seasonal patterns that have been observed in real data and identified birth rates as having a key role in shaping some of these patterns³⁴. However even while accounting for difference in birth rates, a seasonal forcing function was still necessary implying that other factors are still influencing seasonality and called for further investigations into the effect of social and climatic factors.

To be able to disentangle and quantify the effect of different factors such as natural history (interactions between the groups and duration of group specific immunity), climatic variables, birth rates, social factors such as crowding behaviours and the role of immunity in driving seasonality, a lot of data is needed. To start with, future work could fit dynamic models to the data from Pacheco *et al.*³⁵ that give a global overview of RSV seasonality. Additional data on the country-level birth rates and average descriptions of climate would also be needed. Quantifying the effect of social factors would be harder to do at a global scale, but it might be possible within country say by grouping locations into rural or urban.

Despite marked difference in seasonality across locations, the benefits of vaccination programs targeting pregnant women, infants or young children have been identified. However, further investigations into the transmission dynamics of RSV in tropical low-income countries are warranted. Countries with functional and consolidated national healthcare registries such as the USA, Australia and Scotland can use electronic records to map transmission, which makes access to data easier. The WHO has recently embarked on a strategy for global RSV surveillance based on the global influenza surveillance and response system, which is promising³⁶. In addition to looking at disease-related factors that influence vaccine effectiveness, studies should also look into how a vaccination program might be impacted by population social-demographic factors.

2.2. Approaches in phylodynamics

Phylodynamics as a field was first formally defined by Grenfell as the unification of immunodynamics, epidemiology and evolutionary biology, processes that potentially simultaneously influence pathogen diversity, in understanding the drivers of observed pathogen phylogenies at different scales³⁷. The main underlying assumption is that the three processes occur at the same timescale. By definition, phylodynamics first came into existence with the aim of understanding observed patterns of genetic sequences data, hence naturally, methods were biased towards detailed models of pathogen evolution and simple birth-death models were used to represent hypotheses of the epidemiological processes³⁸. Increasing complexity in the epidemiological models within a phylodynamics framework made it difficult to infer the epidemiological parameter solely based on observed phylogenies, the use of other complementary data then became useful in distinguishing between competing phylodynamics hypotheses³⁹. From a traditional epidemiology perspective, most of the applications of phylodynamics have been aimed at determining the transmission characteristics such as the reproductive number⁴⁰⁻⁴² or transmission chains⁴³⁻⁴⁵ during an outbreak, more so for viral outbreaks.

This review of methods will focus on phylodynamics approaches that were aimed at inferring epidemiological dynamics from sequence data. The first broad characterization of the methods I will make is grouping them either as methods that

simultaneously infer epidemiological and evolutionary dynamics, or those that apply a two-step inference process beginning with the evolutionary dynamics. Methods such as^{42,46-50} that break the inference process into two parts start by fitting a model of evolution to the sequence data available, resulting in a phylogenetic tree showing relatedness of the sequences based on the inferred model parameters. Following from this, transmission trees⁴⁶ or other epidemiological characteristics of interest such as the basic reproductive number⁴⁷ or hazard ratios⁴² are inferred. Such methods have the advantage of being less computationally intensive than their simultaneous inference counterparts, however, they could potentially result in inconsistencies between the inferred evolutionary and epidemiological dynamics. Methods of simultaneous inference therefore tend to be preferred, in which case the parameters of a model of evolution and that of an epidemiological model of transmission are inferred instantaneously and are therefore allowed to interact^{40,43-45,50-53}. Depending on the complexity of the models of evolution and epidemiology and the mechanism of interaction e.g. through a joint likelihood function, one might then be required to use a sophisticated inference technique such as the methods developed by Lau *et al.*⁵³ and Li *et al.*⁴⁰.

Phylodynamics methods can also be distinguished by the kind of data used. The basic requirement is that for every case under study, at least one genetic sequence and the sampling times are available. The inclusion of other data describing the infection episode and/or the demographics of the host have led to a broad spectrum of methods. To take into account the importance of within host pathogen evolution, methods have been developed that can accommodate more than one sequence per infected host⁵³⁻⁵⁵. Though most of these methods assume that if a host has multiple sequences they are from the same infection episode, a few do allow multiply infected hosts⁵⁵. In addition to sequence sampling times, data giving information on possible exposure times has also been used, particularly for nosocomial infections⁵⁶. Host demographic information such as location^{44,45,52,57} and recent contacts⁴³ has also been used to enrich analyses, with contact information proving highly valuable in clarifying likely infection sources. The use of other data to complement the sequence data has the added advantage of allowing more complex epidemiological models to be used.

Most methods initially assume that every case has a genetic sequence attached to it, however in reality this is seldom the case. Either not all cases have been observed and therefore not all cases have a sequence available for analysis, or some cases have been observed but for one reason or the other, do not have a genetic sequence. Several approaches have been used to tackle the issue of an incomplete observation of the cases in a particular temporal window and geographical region (often considered an outbreak), the most popular of which is to estimate the proportion of the outbreak that is unobserved^{43,51,55}. The method by Didelot *et al.* accounts for missing sequences by allowing additional branches on the transmission tree to be introduced⁴⁶, while the more complicated approach by Lau *et al.* tries to infer the missing sequences in an outbreak⁵³. The first approach is simpler conceptually and computationally. The Lau method requires the use of a sophisticated model of evolution that tries to infer how a genetic sequence might have evolved in a period of time given an initial guess of an introductory sequence, also known as a master sequence. Accounting for unobserved cases can be crucial to an analysis depending on the timeframe under consideration and the pathogen under study. Despite the range of ways to account for missing cases, if a significant fraction of the outbreak is missing data, then no amount of computational suaveness can make up for poor data, as a recent study comparing different methods found⁵⁸.

A transmission bottleneck refers to the limitation of the amount of viral diversity that is passed on from the infecting host to the infected one. A complete bottleneck therefore refers to the situation where only a single strain is passed on. Transmission bottlenecks can also influence the observed phylogeny at a population level, as such, additional assumptions regarding the size of the bottleneck have to be made when analysing population level phylogenies. Most methods assume a complete transmission bottleneck, in that only one lineage is passed on at the time of a transmission event^{46,50,54}. For acute infections, relaxing this assumption may not have a significant impact on inferred dynamics, however, for chronic infections such as HIV where there is a significant amount of within host diversity, one might need to consider an incomplete transmission bottleneck in order to make accurate inference. Volz *et al.* developed a method that looks at both the population and within-host pathogen diversity and allows for an incomplete transmission bottleneck. This method,

surprisingly, does not require infected hosts to have multiple sequences available for analysis⁵⁹.

Several phylodynamics analyses have been conducted for RSV. Tan *et al* analysed 33 RSV A genomes from the Netherlands, Belgium and the USA spanning a period from 2001 to 2011⁶⁰. In their analysis, they found implications that nonselective epidemiological processes, rather than immune pressure, likely play a bigger role in shaping viral diversity observed from the phylogeny. A study carried out using RSV A and RSV B genes of the F protein using samples from Northern Taiwan found that the rate of evolution was dynamic over time, with an increase observed between 2005 and 2010. They did not find evidence of positive selection⁶¹. A positive selection analysis carried out by Do *et al.* using whole genomes for RSV A and RSV B collected over 2 consecutive epidemics found some evidence of positive selection on the G gene both at the population and within host level⁶². In an analysis of 26 sequences obtained over 78 days from chronically infected immune-compromised child, Grad *et al* found some evidence of an adaptive immune response, however further studies are warranted to validate this finding which could be a result of the unique host factors⁶³. As noted by Tan *et al.*⁶⁰, the results of a positive selection analysis could be influenced by the study design, of note is the difference in temporal and geographical scale between the studies that find evidence of positive selection and those that do not. It would appear that in the short term over a local scale (i.e. within a country or an individual), the RSV genome is likely to show evidence of positive selection whereas in the long term, it is not. Analysis of sequences collected at a large geographical scale, whether in the short or long term, could benefit from the inclusion of additional data on the outbreak to aid in distinguishing between competing hypotheses.

2.3. References:

1. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control*. (OUP Oxford, 1992).
2. Keeling, M. J. & Rohani, P. *Modeling infectious diseases in humans and animals*. (Princeton University Press, 2011).
3. Weber, A., Weber, M. & Milligan, P. Modeling epidemics caused by respiratory syncytial virus (RSV). **172**, (2001).
4. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *epidemiology* **2**, 13 (2005).
5. White, L. J. *et al.* Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math. Biosci.* **209**, 222–239 (2007).
6. Bloom-Feshbach, K. *et al.* Latitudinal Variations in Seasonal Activity of Influenza and Respiratory Syncytial Virus (RSV): A Global Comparative Review. *PLoS ONE* **8**, (2013).
7. Kamigaki, T. *et al.* Seasonality of Influenza and Respiratory Syncytial Viruses and the Effect of Climate Factors in Subtropical–Tropical Asia Using Influenza-Like Illness Surveillance Data, 2010 –2012. (2010).
doi:10.1371/journal.pone.0167712
8. Pitzer, V. E. *et al.* Environmental drivers of the spatiotemporal dynamics of respiratory syncytial virus in the United States. *PLoS Pathog.* **11**, e1004591 (2015).
9. Noveroske, D. B., Warren, J. L., Pitzer, V. E. & Weinberger, D. M. Local variations in the timing of RSV epidemics. *BMC Infect. Dis.* **16**, 674 (2016).
10. ACEDO, L., DÍEZ-DOMINGO, J., MORAÑO, J.-A. & VILLANUEVA, R.-J. Mathematical modelling of respiratory syncytial virus (RSV): vaccination strategies and budget applications. *Epidemiol. Infect.* **138**, 853–860 (2010).
11. Jornet-Sanz, M., Corberán-Vallet, A., Santonja, F. J. & Villanueva, R. J. A Bayesian stochastic SIRS model with a vaccination strategy for the analysis of respiratory syncytial virus. *Sort* **41**, 159–176 (2017).

12. Hervás, D., Reina, J. & Hervás, J. A. Meteorologic Conditions and Respiratory Syncytial Virus Activity. *Pediatr. Infect. Dis. J.* **1** (2012).
doi:10.1097/INF.0b013e31825cef14
13. Price, R. H. M., Graham, C. & Ramalingam, S. Association between viral seasonality and meteorological factors. *Sci. Rep.* **9**, (2019).
14. Paynter, S., Yakob, L., Simões, E. A. F., Lucero, M. G. & Tallo, V. Using Mathematical Transmission Modelling to Investigate Drivers of Respiratory Syncytial Virus Seasonality in Children in the Philippines. *PLoS One* **9**, (2014).
15. Moore, H. C., Jacoby, P., Hogan, A. B., Blyth, C. C. & Mercer, G. N. Modelling the seasonal epidemics of respiratory syncytial virus in young children. *PLoS One* **9**, (2014).
16. Pan-Ngum, W. *et al.* Predicting the relative impacts of maternal and neonatal respiratory syncytial virus (RSV) vaccine target product profiles: A consensus modelling approach. *Vaccine* **35**, 403–409 (2017).
17. Kinyanjui, T. M. *et al.* Vaccine Induced Herd Immunity for Control of Respiratory Syncytial Virus Disease in a Low-Income Country Setting. *PLoS One* **10**, e0138018 (2015).
18. Poletti, P. *et al.* Evaluating vaccination strategies for reducing infant respiratory syncytial virus infection in low-income settings. *BMC Med.* **13**, 49 (2015).
19. Nyiro, J. U. *et al.* Defining the vaccination window for respiratory syncytial virus (RSV) using age-seroprevalence data for children in Kilifi, Kenya. *PLoS One* **12**, e0177803 (2017).
20. Yamin, D. *et al.* Vaccination strategies against respiratory syncytial virus. *Proc. Natl. Acad. Sci.* **113**, 201522597 (2016).
21. Goldstein, E. *et al.* On the Relative Role of Different Age Groups during Epidemics Associated with Respiratory Syncytial Virus. *J. Infect. Dis.* **217**, 238–244 (2018).
22. Hardelid, P., Verfuenden, M., Mcmenamin, J., Smyth, R. L. & Gilbert, R. The contribution of child, family and health service factors to respiratory syncytial virus (RSV) hospital admissions in the first 3 years of life: birth cohort study in Scotland, 2009 to 2015. (2019).
23. Otomaru, H. *et al.* Transmission of Respiratory Syncytial Virus Among Children Under 5 Years in Households of Rural Communities, the Philippines. 1–8 (2016).

doi:10.1093/ofid/ofz045

24. Munywoki, P. K. *et al.* The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya. *J. Infect. Dis.* **209**, 1685–1692 (2014).
25. Hogan, A. B. *et al.* Potential impact of a maternal vaccine for RSV : A mathematical modelling study. *Vaccine* **35**, 6172–6179 (2017).
26. Scheltema, N. M. *et al.* Potential impact of maternal vaccination on life-threatening respiratory syncytial virus infection during infancy. *Vaccine* (2018). doi:10.1016/j.vaccine.2018.06.021
27. Brand, S. P. C., Munywoki, P., Walumbe, D., Keeling, M. J. & Nokes, D. J. Reducing RSV hospitalisation in a lower-income country by vaccinating mothers-to-be and their households. *bioRxiv* 1–21 (2019). doi:10.1101/569335
28. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol. Infect.* **133**, 279–289 (2005).
29. Chan, C. H. S., Sanders, L. P. & Tanaka, M. M. Modelling the role of immunity in reversion of viral antigenic sites. *J. Theor. Biol.* **392**, 23–34 (2016).
30. González-Parra, G. *et al.* A comparison of RSV and influenza in vitro kinetic parameters reveals differences in infecting time. (2018). doi:10.1371/journal.pone.0192645
31. González-Parra, G. & Dobrovolny, H. M. Assessing Uncertainty in A2 Respiratory Syncytial Virus Viral Dynamics. *Comput. Math. Methods Med.* **2015**, 1–9 (2015).
32. Reis, J. & Shaman, J. Retrospective Parameter Estimation and Forecast of Respiratory Syncytial Virus in the United States. *PLoS Comput. Biol.* **12**, 1–15 (2016).
33. Oren, E., Frere, J., Yom-Tov, E. & Yom-Tov, E. Respiratory syncytial virus tracking using internet search engine data. doi:10.1186/s12889-018-5367-z
34. Hogan, A. B., Glass, K., Moore, H. C. & Anderssen, R. S. Exploring the dynamics of respiratory syncytial virus (RSV) transmission in children. *Theor. Popul. Biol.* **110**, 78–85 (2016).
35. Obando-Pacheco, P. *et al.* Respiratory syncytial virus seasonality: A global overview. *J. Infect. Dis.* **217**, 1356–1364 (2018).

36. WHO strategy to pilot global respiratory syncytial virus surveillance based on the Global Influenza Surveillance and Response System (GISRS). (2017).
37. Grenfell, B. T. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science (80-.)*. **303**, 327–332 (2004).
38. Baele, G., Suchard, M. A., Rambaut, A. & Lemey, P. Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.* **66**, e47–e65 (2017).
39. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
40. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evol.* **34**, 2982–2995 (2017).
41. Gomes Naveca, F. *et al.* Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. (2019).
doi:10.1371/journal.pntd.0007065
42. Kenah, E., Britton, T., Halloran, M. E. & Jr, I. M. L. Molecular Infectious Disease Epidemiology : Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. 1–29 (2016). doi:10.1371/journal.pcbi.1004869
43. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput. Biol.* **15**, e1006930 (2019).
44. Campbell, F. *et al.* outbreaker2 : a modular platform for outbreak reconstruction. **19**, (2018).
45. Cori, A. *et al.* A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLoS Comput. Biol.* **14**, 1–22 (2018).
46. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
47. Naveca, F. G. *et al.* Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. *PLoS Negl. Trop. Dis.* **13**, e0007065 (2019).
48. Cottam, E. M. *et al.* Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B Biol. Sci.* **275**, 887–895 (2008).

49. Ypma, R. J. F., van Ballegooijen, W. M. & Wallinga, J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062 (2013).
50. Hall, M., Woolhouse, M. & Rambaut, A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Comput. Biol.* **11**, e1004613 (2015).
51. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput. Biol.* **10**, e1003457 (2014).
52. Ypma, R. J. F. *et al.* Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B Biol. Sci.* **279**, 444–450 (2012).
53. Lau, M. S. Y., Marion, G., Streftaris, G. & Gibson, G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLoS Comput. Biol.* **11**, (2015).
54. Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. & Wallinga, J. *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks.* *PLoS Computational Biology* **13**, (2017).
55. De Maio, N., Wu, C. H. & Wilson, D. J. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Comput. Biol.* **12**, 1–23 (2016).
56. Tong, S. Y. C. *et al.* Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res.* **25**, 111–118 (2015).
57. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* **17**, 82 (2016).
58. Firestone, S. M. *et al.* Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Sci. Rep.* **9**, 1–12 (2019).
59. Volz, E. M., Romero-Severson, E. & Leitner, T. Phylodynamic Inference across Epidemic Scales. *Mol. Biol. Evol.* **34**, 1276–1288 (2017).
60. Tan, L. *et al.* Genetic Variability among Complete Human Respiratory Syncytial Virus Subgroup A Genomes: Bridging Molecular Evolutionary Dynamics and Epidemiology. *PLoS One* **7**, 1–15 (2012).
61. Chi, H. *et al.* Molecular Epidemiology and Phylodynamics of the Human

- Respiratory Syncytial Virus Fusion Protein in Northern Taiwan. *PLoS One* **8**, (2013).
62. Do, L. A. H. *et al.* Direct whole-genome deep-sequencing of human respiratory syncytial virus A and B from Vietnamese children identifies distinct patterns of inter- and intra-host evolution. *J. Gen. Virol.* 3470–3483 (2015).
doi:10.1099/jgv.0.000298
63. Grad, Y. H. *et al.* Within-Host Whole-Genome Deep Sequencing and Diversity Analysis of Human Respiratory Syncytial Virus Infection Reveals Dynamics of Genomic Diversity in the Absence and Presence of Immune Pressure. *J. Virol.* **88**, 7286–7293 (2014).

3. Paper 1: Model based estimates of transmission of respiratory syncytial virus within households.

3.1. Overview

This chapter was written in fulfilment of the first part of the second objective. It presents a primary analysis of the social-temporal data from the household cohort study described in Chapter 1. The work in this chapter was published as *Kombe, I. K., Munywoki, P. K., Baguelin, M., Nokes, D. J. & Medley, G. F. **Model-based estimates of transmission of respiratory syncytial virus within households.** *Epidemics* 1–11 (2018).*

3.2. Role of candidate

I formulated the equations and conducted the numerical analysis and wrote the first draft of the paper. Revisions were made with feedback, input, and guidance from my supervisors Graham F. Medley and D. James Nokes, and advisor Marc Baguelin. Patrick K. Munywoki (PKM) was responsible for the original study design and data collection that led to the data used in my analysis, information he provided on the data helped to identify its limitations. Charles Agoti, George Githinji and Sam Brand provided comments on the analysis.

Research paper cover sheet

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	Ish1601271	Title	Ms
First Name(s)	Ivy Kadzo		
Surname/Family Name	Kombe		
Thesis Title	Integrating viral RNA sequence and epidemiological data to define transmission patterns for respiratory syncytial virus		
Primary Supervisor	Professor Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	EPIDEMICS		
When was the work published?	July 2019		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	Choose an item. CC-BY	Was the work subject to academic peer review?	Choose an item. Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

<p>For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)</p>	<p><i>I led the analysis, including writing all the code, wrote the first draft and incorporated co-author comments.</i></p>
---	--

SECTION E

Student Signature	[REDACTED]
Date	<i>25/07/2019</i>

Supervisor Signature	[REDACTED]
Date	<i>25 JULY 2019</i>

Model-based estimates of transmission of respiratory syncytial virus within households

Ivy K. Kombe^{a, b, *}, Patrick K. Munywoki^a, Marc Baguelin^c, D. James Nokes^{a, d}, Graham F. Medley^b[Show more](#)<https://doi.org/10.1016/j.epidem.2018.12.001>

Under a Creative Commons license

[Get rights and content](#)

open access

Abstract

Introduction

Respiratory syncytial virus (RSV) causes a significant respiratory disease burden in the under 5 population. The transmission pathway to young children is not fully quantified in low-income settings, and this information is required to design interventions.

Methods

We used an individual level transmission model to infer transmission parameters using data collected from 493 individuals distributed across 47 households over a period of 6 months spanning the 2009/2010 RSV season. A total of 208 episodes of RSV were observed from 179 individuals. We model competing transmission risk from within household exposure and community exposure while making a distinction between RSV groups A and B.

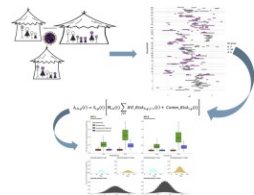
Results

We find that 32–53% of all RSV transmissions are between members of the same household; the rate of pair-wise transmission is 58% (95% CrI: 30–74%) lower in larger households (≥ 8 occupants) than smaller households; symptomatic individuals are 2–7 times more infectious than asymptomatic individuals i.e. 2.48 (95% CrI: 1.22–5.57) among symptomatic individuals with low viral load and 6.7 (95% CrI: 2.56–16) among symptomatic individuals with high viral load; previous infection reduces susceptibility to re-infection within the same epidemic by 47% (95% CrI: 17–68%) for homologous RSV group and 39% (95% CrI: -8%–69%) for heterologous group; RSV B is more frequently introduced into the household, and RSV A is more rapidly transmitted once in the household.

Discussion

Our analysis presents the first transmission modelling of cohort data for RSV and we find that it is important to consider the household social structuring and household size when modelling transmission. The increased infectiousness of symptomatic individuals implies that a vaccine against RSV related disease would also have an impact on infection transmission. Together, the weak cross immunity between RSV groups and the possibility of different transmission niches could form part of the explanation for the group co-existence.

Graphical abstract

[Download high-res image \(110KB\)](#)[Download full-size image](#)

3.3. Abstract

Introduction

Respiratory syncytial virus (RSV) causes a significant respiratory disease burden in the under 5 population. The transmission pathway to young children is not fully quantified in low-income settings, and this information is required to design interventions.

Methods

We used an individual level transmission model to infer transmission parameters using data collected from 493 individuals distributed across 47 households over a period of 6 months spanning the 2009/2010 RSV season. A total of 208 episodes of RSV were observed from 179 individuals. We model competing transmission risk from within household exposure and community exposure while making a distinction between RSV groups A and B.

Results

We find that 32-53% of all RSV transmissions are between members of the same household; the rate of pair-wise transmission is 58% (95% CrI: 30-74%) lower in larger households (≥ 8 occupants) than smaller households; symptomatic individuals are 2-7 times more infectious than asymptomatic individuals i.e. 2.48 (95% CrI: 1.22-5.57) among symptomatic individuals with low viral load and 6.7(95% CrI: 2.56-16) among symptomatic individuals with high viral load; previous infection reduces susceptibility to re-infection within the same epidemic by 47% (95% CrI: 17%-68%) for homologous RSV group and 39% (95%CrI: -8%-69%) for heterologous group; RSV B is more frequently introduced into the household, and RSV A is more rapidly transmitted once in the household.

Discussion

Our analysis presents the first transmission modelling of cohort data for RSV and we find that it is important to consider the household social structuring and household size when modelling transmission. The increased infectiousness of symptomatic individuals implies that a vaccine against RSV related disease would also have an impact on infection transmission. Together, the weak cross immunity between RSV groups and the possibility of different transmission niches could form part of the explanation for the group co-existence.

3.4. Introduction

Respiratory syncytial virus (RSV) is an ubiquitous RNA virus infection that is a major cause of lower respiratory tract disease in children under 5 years of age worldwide ^{1,2}. The estimated global burden of RSV associated acute lower respiratory tract infection (ALRI) in 2015 in under 5 year olds is 33.0 million (21.6-50.3), most of which occurs in developing countries (30.5 million) ³. Of the 3.2 (2.7 -3.8) million hospital admissions associated with RSV in the under 5s, 1.4 (1.2-1.7) million occurred in the 0-5 months age group, and 1.2 (1.0-1.5) million occurred in developing countries.

Despite 50 years of vaccine research none is yet licensed for the prevention of RSV infection or disease. There are currently over fifty vaccines in different stages of development: many with the aim of prevention of early infant RSV disease. While the most advanced (in phase III trials) is a maternal vaccine to boost transplacental antibody transfer ^{4,5}, a variety of product types and range of strategies for protecting young children are under investigation including indirect protection by targeting older infants, elder siblings and family cocooning ⁶⁻⁸.

Prior to vaccine introduction, drivers of transmission need to be well understood in order to predict the potential public health impact of implementation. Investigating outbreaks within the household setting could help to further characterize RSV transmission. The household is an important unit of study for diseases that are transmitted through close contact. The quantitative analysis of household outbreaks has been conducted for influenza ⁹⁻¹⁵. This has led to quantification of transmissibility within the household, improved understanding of the factors that determine level of transmission such as household size and effectiveness of different household level interventions ¹⁶. To date studies of RSV transmission within households or families have been largely observational. One of the earliest is a household cohort study in the USA in which 36 families were followed up for 2 months during the 1974/1975 RSV season ¹⁷. This study found that RSV attack rates in households were high, more so in infants. Older siblings to infants were found to be the most likely index cases in household outbreaks, and illness was found to have an age-related severity. Several other studies over the years across different settings have highlighted the importance

of older children in household outbreaks¹⁸⁻²⁰ which could have implications for control strategies ²¹.

In Kenya, a household cohort study conducted in a rural coastal community during the 2009/2010 RSV epidemic has revealed several patterns. In addition to the importance of older children²⁰, bigger household size and infection with RSV group B, among other factors, were found to be independently associated with increased risk of asymptomatic infection ²²; shedding duration estimates (using molecular diagnostics) were 11.2 days on average, and longer than the previous range reported of 3.9-7.4 days ²³; individuals experiencing the first infection of an RSV season were found to shed more virus relative to secondary infections; children under 1 year old, symptomatic shedders and RSV A and B co-infected individuals were identified as the most likely to transmit due to their relatively higher viral loads ²⁴.

RSV can be categorized into two antigenically and genetically distinct groups, RSV A and RSV B ²⁵. These groups, thought to have diverged about 350 years ago ²⁶, have been observed to co-exist geographically and temporally with most outbreaks being dominated by RSV A and, in some locations, clear patterns of alternating dominance ²⁷. Within the RSV groups are subgroups or genotypes whose frequency changes from season to season, with some genotypes undergoing complete replacement over time ²⁸⁻³³. This pattern of group and genotype replacement is thought to be due to a herd immunity effect ^{25,27,34,35}. A phylogenetic analysis of RSV A sequences from the Kenyan household study showed that most infections arise from a single variant introduction followed by accumulation of household specific variation, i.e. cases arise more from within household spread rather than multiple introductions ³⁶.

However, there is yet to be a mechanistic analysis of RSV household outbreak data that consolidates information on the characteristics of infection episodes and characteristics of the host population into a single dynamic framework. Inference could then be drawn on the competing risks of within household exposure and community (external to household) exposure, in order to quantify the importance of households in RSV transmission. We proposed to use an individual-based approach within a Bayesian framework to analyse the household cohort data from Kenya to

further understand transmission dynamics. We also explore the differences and interactions between RSV groups.

3.5. Methods

Data

The data to be used were collected from a household cohort study conducted in rural coastal Kenya within the Kilifi Health and Demographic Surveillance System (KHDSS) during the 2009/2010 RSV epidemic. Details of the study have been published elsewhere^{20,22,23,37}. In brief, the infant-centric study recruited household members using the criteria that the infant was born after 1 April 2009 (after the previous RSV epidemic) and had at least 1 older sibling less than 13 years old. Deep nasopharyngeal swab (NPS) samples were collected every 3-4 days regardless of symptoms, together with a record of clinical illness. The samples were tested for RSV antigen using an in-house real-time multiplex polymerase chain reaction (PCR) assay. A sample was considered antigen positive if the PCR cycle threshold (Ct) value was 35.0 or below. Positive Ct values were then converted to viral load (\log_{10} RNA equivalent). A household was defined as a group of individuals living in the same compound and who eat together. The data contain information from 493 individuals spread across 47 households whose dates of data collection span 180 days. The household sizes range from 4 to 37 occupants with a median of 8 members.

An RSV A/B shedding episode is defined as a period within which an individual provided PCR positive samples for RSV A/B that were no more than 14 days apart. A shedding episode is referred as symptomatic if within the window of virus shedding, there is at least one day where symptoms were recorded. The symptoms of interest are those of an acute respiratory illness (ARI), which are: cough, or nasal discharge/blockage, or difficulty breathing. Sampling of the study population was done in 3-4 day intervals, as such, complete duration of shedding and ARI episodes had to be imputed, and missing viral loads were linearly interpolated. Shedding durations were imputed first, after which, if there were any days of recorded ARI within shedding episodes, the total duration of the ARI was imputed based on the days of recorded symptoms. As such, the length of an ARI episode within a shedding episode can be \leq length of related shedding episode. The start and end of a shedding and ARI episode were imputed rather than inferred through data augmentation to ensure consistency and hence comparability across studies that have used the same household data^{20,23,24}.

During the sample-collection visits, if a household member was not present, they were recorded as being 'away' on that particular day. As with the shedding information, there was incomplete information on continuous periods of presence or absence from the household. This information was imputed using the same method that was applied to imputing complete shedding durations. There are some instances where an individual was present but not sampled, as such, presence could not purely be identified by the availability of NPS samples. Details of the imputation of shedding, ARI and presence/absence durations and interpolation of viral load can be found in the appendix section A2: Supplementary appendix for Paper 1. For the model, we will assume that all the cases were observed, and ignore the possibility of short duration shedding episodes that could have been missed by the sampling intervals.

We categorized days of shedding according to viral load and symptoms into 4 categories to compare infectiousness: low viral load and asymptomatic, high viral load and asymptomatic, low viral load and symptomatic and, high viral load and symptomatic. High viral load is defined as $>6 \log_{10}$ viral copy number (or a PCR Ct value <23.05).

Transmission model

We built a mechanistic model for RSV that tracks group-specific infection onset at the individual host level. The main aim is to determine the factors that influence infection onset in an individual, and this is the focus of the model formulation. At the start of the outbreak, we assume that everyone is susceptible to RSV infection, but the risk of infection is dependent on age. Once individuals have been exposed to infection, they enter a latency period that ranges between 2 to 5 days after which they become infectious. After the infectious period, individuals become susceptible to infection again, but the risk to subsequent infection is modified, i.e. RSV confers partial transient immunity that lasts as long as the outbreak is ongoing. This partial immunity is assumed to be different from heterologous group re-infection and homologous group re-infection. Having RSV infection risk altered by age and infection history implies the existence of long-term and short-term immunity. This has previously been explored by other modelling studies^{38,39}. Individuals can get heterologous group co-infections, i.e.

we assume infection with RSV A is possible while shedding RSV B, and vice-versa. Per RSV group, our model formulation is similar to the Susceptible(S)-Exposed(E)-Infectious(I)-Susceptible(S₂) type model dynamics.

The main assumptions about transmission are contained in the equation giving the per capita rate of exposure (to infection) per unit time, also known as the infection hazard, denoted $\lambda(t)$. At its base:

$$\begin{aligned} \lambda(t) &= \text{contact rate} * \text{probability of transmission give contact} * \text{number of infectious contacts}(t) \\ &= \text{baseline rate of exposure} * \text{number of infectious contacts}(t) \end{aligned}$$

$$\lambda(t) = \eta * \sum I(t)$$

In our model, a susceptible individual can get infected by someone they share a household with, or from a source outside of the household, splitting λ into two components: a within household exposure component and a community exposure component.

$$\begin{aligned} \lambda(t) &= [\text{baseline household rate of exposure} * \text{number of infectious household contacts}(t)] \\ &+ \\ &[\text{baseline community rate of exposure} * \text{number of infectious community contacts}(t)] \end{aligned}$$

$$\lambda(t) = \left(\eta * \sum_{\text{household}} I(t) \right) + \left(\varepsilon * \sum_{\text{community}} I(t) \right)$$

The number of infectious household contacts is observed in the data. Though there are cases from different households in the data, the sample in the study is small relative to the number of households in the community, as such the true number of infectious community contacts is unknown. We therefore cannot directly infer infectious community contacts and have to use a representative function instead. We do this using a bell-shaped curve that mimics the ongoing outbreak dynamics. We thus have:

$$\lambda(t) = [\text{baseline household exposure rate} * \text{number of infectious household contacts}(t)] \\ + \\ [\text{baseline community exposure rate} * \text{background community function}(t)]$$

$$\lambda(t) = \left(\eta * \sum_{\text{household}} I(t) \right) + (\varepsilon * f(t))$$

We extend this basic formulation to explore if factors such as household size, infectiousness (as determined by viral load and ARI symptoms) and age are determinants of exposure. Further details of each component are provided in the subsequent sections. The rate of exposure to a particular RSV group (index g) is given for a particular individual, (index i) from a given household (index h) at a given day (index t) and is specified by the notation $\lambda_{i,h,g}(t)$.

Within household exposure:

For an individual i , in household h , the rate of exposure at a given time t , is a summation of rates from all the infectious individuals in their household. The rate of exposure from a single infectious housemate (index j) is assumed to depend on the size of the household and the viral load and symptom status. We consider the household size effect as a binary variable where a house with >8 members is considered large. We consider viral load and symptom status as one variable with 4 categories: low viral load and no symptoms, high viral load and no symptoms, low viral load and symptomatic, high viral load and symptomatic. The household rate of exposure from infectious individual j present in the household at time t to i is thus give as:

$$HH_Rate_{h,g,j \rightarrow i}(t) \\ = \eta_g \times \psi_H(\text{Household_size}_i) \times \psi_{I,inf}(\text{Infectivity}_{j,h,g}(t)) \times M_{j,h}(t)$$

η_g is the baseline rate of exposure in the household which is estimated for each of the two RSV groups, RSV A and RSV B. ψ_H is the coefficient modifying exposure in large household relative to small households and $\psi_{I,inf}$ is the coefficient modifying infectiousness based on viral load and symptom status. The within household rate of

exposure only affects susceptible individuals who are present in the household, as such this rate is multiplied by a binary variable $M_{i,h}(t) = 0$ if i is not present in the household at time t and $M_{i,h}(t) = 1$ if i is present.

Community exposure:

For a susceptible individual i , this external to the household source of exposure is assumed to represent both sampled and unsampled cases from other households. Community exposure is assumed to depend on the age of the susceptible individual and time. Age is treated as a categorical variable. The community rate of exposure is thus given as:

$$Comm_{Rate_{i,g}}(t) = \varepsilon_g \times f_g(t) \times \psi_{E,age}(Age_group_{E,i})$$

ε_g is the baseline rate of exposure from the community, which is estimated for each of the two RSV groups. $\psi_{E,age}$ is the coefficient modifying the rate of community exposure by age. For each RSV group, we have $f_g(t)$, a time-unit dependent curve that modifies the community rate of exposure over time, in this case the time period of interest is the duration of the study. We wanted this curve to represent the background epidemic dynamics in the local zone from which the data was collected; as such we proceeded to use the same household dataset to generate it.

The data are calibrated in days and are at the individual level, but to obtain the background community rate, we assumed that this background rate is scalable from the weekly household-level rate of primary incidence, denoted $\lambda_{HH}(t_w)$. The household level rate of primary incidence is the rate at which a household (rather than a single member of a household) acquires the first episode/outbreak in the ongoing RSV season. A household outbreak is a period within which at any given time, at least one household member is shedding RSV. If we treat $\lambda_{HH}(t_w)$ as the hazard rate in a probability distribution, we can estimate it using the following model:

$$I_c(t_w) = N_{HH} \left(1 - \exp^{-\int_0^{t_w} \lambda_{HH}(s)} \right)$$

$$I(t_w) = I_c(t_w) - I_c(t_w - 1)$$

Where

N_{HH} = Total number of households in the study

$I(t_w)$ = Average weekly household-level incidence of primary infection

$I_c(t_w)$ = Weekly cumulative household-level incidence of primary infection

We further assumed that $\lambda_{HH}(t_w) = a_1 \exp\left(-\frac{(t_w - b_1)^2}{c_1}\right)$, giving it a bell-shape, and estimated $\{a_1, b_1, c_1\}$ using maximum likelihood assuming Poisson distributed data.

Once $\lambda_{HH}(t_w)$ was estimated for each RSV group, it was scaled such that it ranges

between 0 and 1 using the formula $X_i^{Scaled} = \frac{X_i - \min(\{X\})}{\max(\{X\}) - \min(\{X\})}$. As such,

$$f_g(t_w) = \frac{\lambda_{HH}(t_w) - \min(\{\lambda_{HH}(1), \lambda_{HH}(2), \dots, \lambda_{HH}(t_w)\})}{\max(\{\lambda_{HH}(1), \lambda_{HH}(2), \dots, \lambda_{HH}(t_w)\}) - \min(\{\lambda_{HH}(1), \lambda_{HH}(2), \dots, \lambda_{HH}(t_w)\})}$$

To turn $f_g(t_w)$ into a daily scale, the value for a given week were assumed to be the values for every day of that week. The resultant background community curves for RSV A and B are shown in Figure 3. 1.

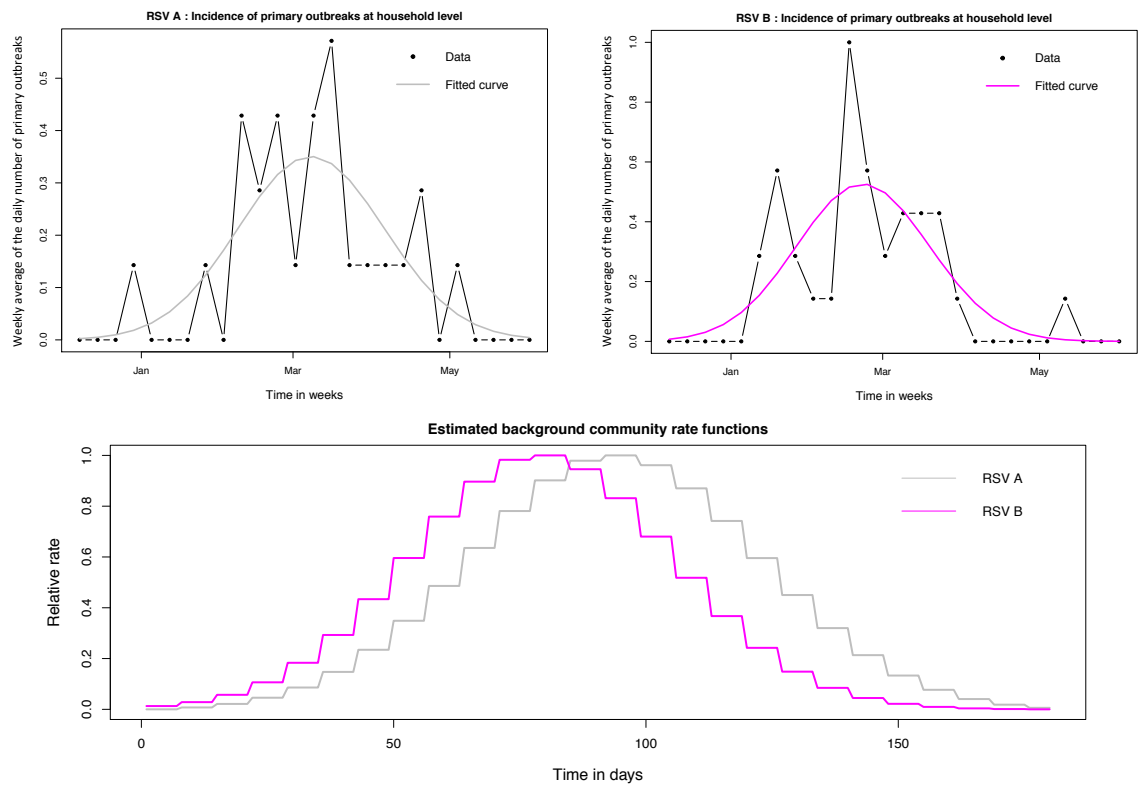


Figure 3. 1: Establishing the background community rate function.

The figures in the top row show a comparison of data and model fit of the weekly household-level rate of primary incidence that was used to derive the background community rate function. Top left: RSV A data and model fit; Top right: RSV B data and model fit; Bottom: Comparing the estimated background community rate function for RSV A and RSV B.

Finally, we assume that susceptibility can be modified according to an individual's infection history within the same epidemic, and their age. These two components are combined into an equation representing relative susceptibility to infection as shown below

$$S_{i,g}(t) = \exp\left(\phi_{Y,hist}(Infection_History_i(t)) + \phi_{X,age}(Age_group_{S,i})\right)$$

$\phi_{X,age}$ is the coefficient modifying susceptibility by age. We categorized infection history into four groups: no previous infection, recovered from an RSV A infection, recovered from an RSV B infection, recovered from both RSV A and B. $\phi_{Y,hist}$ is the coefficient modifying susceptibility to a particular RSV group depending on infection history in the following three ways: by $\exp^{\phi_{Y,hom}}$ if an individual has previously experienced and recovered from infection by the same group (homologous infection), $\exp^{\phi_{Y,het}}$ if the individual has previously experienced and recovered from infection by a different group (heterologous infection) and by $\exp^{(\phi_{Y,hom} + \phi_{Y,het})}$ if an individual has previously experienced and recovered from both RSV A and RSV B infection. This mechanism of interaction between RSV A and B is similar to that applied in a compartmental model used to analyse data from the UK and Finland ²⁷. In combination, all the above assumptions result in the rate of exposure equation shown below

$\lambda_{i,h,g}(t)$ = Rate of exposure of individual i in household h with RSV group g at time t .

$$\lambda_{i,h,g}(t) = S_{i,g}(t) \left[M_{i,h}(t) \sum_{\substack{j \neq i, \\ j \text{ in } i's \\ \text{household}}} HH_{Rate_{h,g,j \rightarrow i}}(t) + Comm_{Rate_{i,g}}(t) \right] \dots (Eq 3.1)$$

The assumption of how age and infection history modify the rate of exposure is similar to the assumptions made in a proportional hazards model.

Additional details on the data variables and parameters are given in Table 3. 1.

Table 3. 1: Model Notation

Symbol	Name	Type	Description
i		Index	Index of individual
h		Index	Index of household
g		Index	Index of RSV group type, either A or B
t		Index	Index of time in days
$I_{j,h,g}(t)$	<i>Infectivity</i>	Data*	Categorical data variable for infectious individuals indicating level of infectivity categorized by viral load and symptom status at time t . The categories are: low viral load and asymptomatic (reference group), high viral load and asymptomatic, low viral load and symptomatic and, high viral load and symptomatic. High viral load is defined as $>6 \log_{10}$ viral copy number.
$Y_i(t)$	<i>Infection_ history</i>	Data	Variable indicating if an individual has experienced and recovered from an infection by a particular RSV group in the current epidemic at time t .
X_i	<i>Age_grou p_s</i>	Data [§]	Categorical data variable indicating the susceptibility age group of an individual. The age groups are <1 year (reference group), 1-4 years, 5-14 years and ≥ 15 years.
$M_{i,h}(t)$		Data	Binary data variable indicating if an individual is present in the household at time t . Absence from the household means that an individual was not present at the point of sample collection and thus in the model they can only get infection from a community

			source and not from an infectious housemate (not sampled and not at household risk). Individuals who were present but not sampled are exposed to both household and community source transmission in the models (not sampled but at household risk).
H_i	Household _size	Data*	Binary data variable indicating whether the individual lives in a large or small household. A small household (reference group) has <8 individuals.
E_i	Age_grou pE	Data [§]	Categorical data variable indicating the community exposure age group of an individual. The age groups are <1 year (reference group), 1-4 years and ≥5 years.
$\phi_{X,age}$	<i>Sus.age.2</i> <i>Sus.age.3</i> <i>Sus.age.4</i>	Parameter	Coefficients modifying susceptibility to RSV depending on age, applied to the age group covariate X_i . <i>Sus.age.2</i> estimates the effect being in age group 1-4 years, <i>Sus.age.3</i> the effect of group 5-15 and <i>Sus.age.4</i> of group ≥15 relative to group <1 year.
$\phi_{Y,hist}$	<i>Prev.hom</i> <i>Prev.het</i>	Parameter	Coefficients modifying susceptibility to infection by a particular RSV group depending on infection history. <i>Prev.hom</i> estimates the effect of a previous homologous group infection, while <i>Prev.het</i> estimates the effect of a previous heterologous group infection. Applied to the categorical covariate $Y_i(t)$.
ψ_H	<i>HH.size</i>	Parameter	Coefficient modifying the amount of within household exposure by household size. <i>HH.size</i> estimates the effect of being in a

			large household relative to a small one. Applied to covariate H_i .
η_g	<i>HH.rsv.a</i> <i>HH.rsv.b</i>	Parameter	Baseline rate of within household exposure by RSV group
$\psi_{I,inf}$	<i>High.Asym</i> <i>Low.Sym</i> <i>High.Sym</i>	Parameter	Coefficients modifying infectiousness by viral load and symptom status. Relative to shedding low viral load and being asymptomatic, <i>High.Asym</i> estimates the effect of shedding high viral load and being asymptomatic, <i>Low.Sym</i> the effect of shedding low viral load and being symptomatic and <i>High.Sym</i> the effect of shedding high viral load and being symptomatic. Applied to the infectivity covariate $I_{j,h,g}(t)$.
$\psi_{E,age}$	<i>Exp.age.2</i> <i>Exp.age.3</i>	Parameter	Coefficients modifying the rate of community exposure by age group. <i>Exp.age.2</i> estimates the effect being in age group 1-4 years and <i>Exp.age.3</i> the effect of group ≥ 5 , relative to the <1-year age group. Applied to the age group covariate E_i
ε_g	<i>Comm.rsv.a</i> <i>Comm.rsv.b</i>	Parameter	Community transmission coefficient by RSV group
$f_g(t)$		Estimated	RSV group specific, time-dependent curve modifying the rate of community exposure.
$U_{i,h,g}$		Data	Set of all days where individual i has an onset of infection with RSV group g . Only includes the first day of shedding for each infection episode.

$A_{i,h,g}$	Data	Set of all the days where individual i is at risk of infection with RSV group g , i.e. they are not currently shedding g .
-------------	------	--

* The choice of cut-off for high viral load and large households was based on initial runs of the inference algorithm that explored different cut-offs for each. The choice of $6 \log_{10}$ copy number for high viral load and 8 persons for large households led to the best convergence. [§]The decision to have different age groups for susceptibility and community exposure was based on initial model runs where the 4th community exposure age group effect (>15 years) was poorly estimated and as such was uninformative. Consequently, this group was merged with the 3rd group.

Following on from the rate of exposure equation are two additional nested equations that make up the model.

$\alpha_{i,h,g}(t)$ = Probability of infection following exposure per day i.e. individual enters the latent phase

$$\alpha_{i,h,g}(t) = (1 - \exp^{-\lambda_{i,h,g}(t)}) \dots \text{(Eq 3.2)}$$

$p_{i,h,g}(t)$ = Probability of starting to shed i.e. individual enters the infectious phase at time t given they did not shed until t .

$$p_{i,h,g}(t) = \sum_{l=0}^L \theta_l \alpha_{i,h,g}(t-l) \dots \text{(Eq 3.3)}$$

Where L is the maximum latent period and θ_l is the probability that the latent period is exactly l days. For $l = \{0,1,2,3,4,5\}$ days, we have the following probabilities $[0,0,4,4,3,1]/12 = [0, 0,0.33,0.33,0.25,0.083]$ ⁴⁰. The same latency distribution is used for RSV A and B.

Since the model is focused on the determinants of infection onset process, the data whose likelihood we are interested in is the individual onset times. As such, we express the likelihood of an individuals observed days of onset as:

$$L_i = \prod_{\text{all RSV groups}} [\text{probability of all onset days and non_onset days with risk of onset} | \text{model}]$$

We assume the data is binomially distributed and write the likelihood as:

$$L_i = \prod_g \left[\prod_{u \in U_{i,h,g}} p_{i,h,g}(u) \prod_{u \in A_{i,h,g}} (1 - p_{i,h,g}(u)) \right]$$

Where $U_{i,h,g}$ is the set of days where individual i had an onset of RSV group g infection and $A_{i,h,g}$ is the set of all days where i did not have an onset but was at risk of infection (i.e. not shedding RSV group g).

The model as presented can be reduced to fit for a single RSV group or for RSV as a single pathogen with no distinction between RSV A and B. Attempts to model household size as a continuous variable were unsuccessful possibly due to our small sample size and hence we modelled transmission within the household as a density dependent process but identified households as either large or small and found that the cut-off between categories of 8 provided the best fit.

Parameter inference

We used Bayesian inference to obtain estimates of the parameters. Adaptive Metropolis Markov Chain Monte Carlo was used as implemented in the R software package *fitR*⁴¹, function *mcmcMH*. The *mcmcMH* function can adapt the size of the proposal distribution, such that the acceptance rate is close to 23.4%, and the shape using the Adaptive metropolis algorithm as in⁴²; the difference in size and shape adaptation being in the scaling factor used. In brief, the method builds a Markov chain which allows us to sample from the posterior distribution $P(\varphi|D)$ of the parameters given the data, where $\varphi = \{\phi_{X,age}, \phi_{Y,hist}, \psi_H, \eta_g, \psi_{I,inf}, \psi_{E,age}, \varepsilon_g\}$. Flat bounded priors were used for all the log of parameters. The limits on the parameters measuring relative effects was -10 to 10, while that on the transmission coefficients was -20 to 0. We initiated 3 chains and set the algorithm to start adapting the size of the proposal distribution after 1000 iterations and the shape after 500 accepted iterations.

Burn-in was assessed visually after which the results of the three concurrent chains were combined to infer the posterior distribution. To obtain fairly accurate values for the 95% credible intervals, we ran the MCMC algorithm until the effective sample size (ESS) was ≥ 4000 ⁴³. The three chains were run for 250,000 iterations each and burn-in for each chain was 80,000, 90,000 and 80,000. After burn-in the reminders of the

three chains were combined into a single chain with an overall acceptance rate of 16.8%. The parameters were estimated on the log scale. All the computation was done using R software package (RStudio version 1.1.383 running R version 3.4.0⁴⁴). The code is freely available under the GNU Lesser General Public License v3.0 and can be found at https://github.com/lkadzo/HH_Transmission_Model.

3.6. Results

Table 3. 2 gives a summary of the shedding episodes in the data. This particular outbreak had more RSV B cases than RSV A, with a significant portion of cases being symptomatic both for RSV A and B. Eighty five percent of the households that were successfully followed up had an introduction of an RSV case. In addition to the information in Table 3. 2; 28 (13.5%) of the total 208 episodes were censored during imputation; of the A and B episodes, 14 (6.7%) were simultaneous RSV A and B shedding episodes, 7 (3.3%) of which had a simultaneous onset; of the 179 individuals who got infected 31 (17.3%) were <1 year old, 41 (22.9%) were 1-4 years, 66 (36.9%) were 5-14 years and 41 (22.9%) ≥15 years old. Of the symptomatic infected individuals, 28 (25.7%) were <1 year old, 35 (32.1%) were 1-4 years, 36 (33%) were 5-14 years and 10 (9.2%) ≥15 years old. A detailed analysis of these shedding patterns has been published elsewhere ²⁴. Figure 3. 2 shows the shedding pattern for all 179 people who had a shedding episode. Figure A2. 3 and Figure A2. 4 in appendix A2 shows the shedding and ARI patterns for RSV A and B respectively.

Table 3. 2: Summary of shedding episodes

	RSV A	RSV B	All RSV
Number of episodes	97	125	208
Number of symptomatic episodes	59	69	119
Number of people infected	88	113	179
Number of people with symptomatic episodes	54	67	109
Number of people with repeat infections	8	12	27
Number of households infected (percentage of total)	25 (53.2%)	34 (72.3%)	40 (85.1%)
Total percentage of household occupants that were infected (total number of occupants) *	30.0% (293)	28.5% (396)	40.5% (442)

* The total number of infected individuals out of the total number of individuals that occupy the infected households.

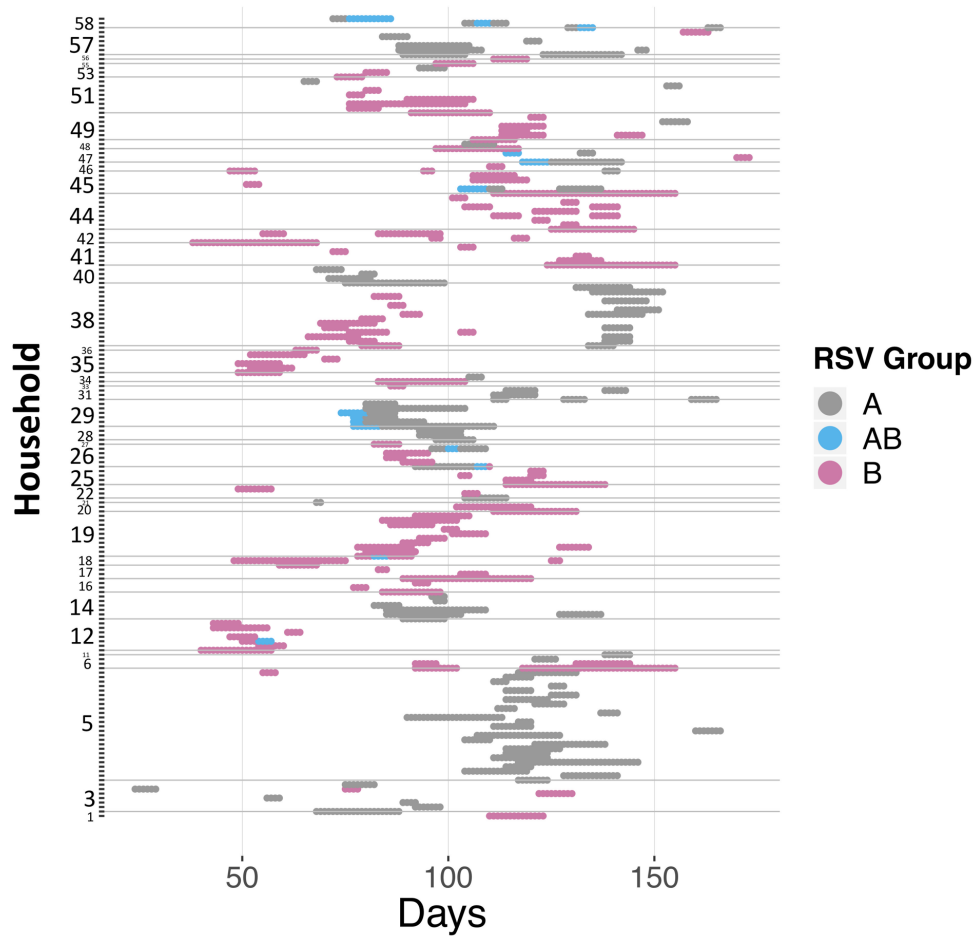


Figure 3. 2: Shedding patterns for each of the 179 individuals who experienced at least one RSV shedding episode.

The y-axis shows the household, time is on the x-axis with zero indicating the day before the first sample was collected. The grey dots show RSV A shedding, dark pink show RSV B and blue shows days of co-shedding. The horizontal grey lines separate the data by household. The study initially recruited 60 households but 13 were lost to follow-up, hence the numbering of the households goes beyond 47.

Transmission model parameter inference

The trace plots used to assess convergence of the three chains are shown in Figure A2. 5 in appendix A2. The resulting parameters estimates are given in **Error! Reference source not found.** and Figure A2. 6.

Table 3. 3: Results of fitting the transmission model.

Median and 95% credible intervals (CrI) are given for the 15 parameters of interest. The posterior distribution for each parameter was obtained by running 3 MCMC chains for 250,000 iterations each. The burn-in for the three chains was 80,000, 90,000 and 80,000 respectively. The reminders of the three chains were combined into a single chain with and overall acceptance rate of 16.8%

Symbol	Description	Name	Median (95% credible interval (CrI))
$\phi_{X,age}$	Coefficients modifying susceptibility to RSV by age. <i>Sus.age.2</i> estimates modification to group 1-4 years, <i>Sus.age.3</i> 5-15 years and <i>Sus.age.4</i> ≥ 15 years relative to group <1 year.	<i>Sus.age.2</i>	0.924 (0.483, 1.87)
		<i>Sus.age.3</i>	0.267 (0.142, 0.537)
		<i>Sus.age.4</i>	0.155 (0.0825, 0.316)
$\phi_{Y,hist}$	Coefficients modifying susceptibility to infection by a particular RSV group depending on infection history. <i>Prev.hom</i> estimates the effect of a previous homologous group infection, and <i>Prev.het</i> the effect of a previous heterologous group infection.	<i>Prev.hom</i>	0.530 (0.316, 0.833)
		<i>Prev.het</i>	0.607 (0.306, 1.08)
ψ_H	Coefficient modifying the amount of within household exposure by household size for households of 8 or more relative to <8.	<i>HH.size</i>	0.424 (0.265, 0.702)
η_g	Baseline rate of within household exposure by RSV group	<i>HH.rsv.a</i>	0.0188 (0.00734, 0.0401)
		<i>HH.rsv.b</i>	0.015 (0.00578, 0.033)
$\psi_{I,inf}$	Coefficients modifying infectiousness by viral load and	<i>High.Asym</i>	0.0704 (0.0000692, 3.15)
		<i>Low.Sym</i>	2.48 (1.22, 5.57)

	symptom status. Relative to shedding low viral load and being asymptomatic, <i>High.Asym</i> estimates the effect of shedding high viral load and being asymptomatic, <i>Low.Sym</i> the effect of shedding low viral load and being symptomatic and <i>High.Sym</i> the effect of shedding high viral load and being symptomatic.	<i>High.Sym</i>	6.7 (2.56, 16.0)
$\psi_{E,age}$	Coefficients modifying the rate of community exposure by age group. <i>Exp.age.2</i> estimates the effect being in age group 1-4 years and <i>Exp.age.3</i> the effect of group ≥ 5 , relative to the <1-year age group.	<i>Exp.age.2</i>	0.563 (0.206, 1.45)
		<i>Exp.age.3</i>	1.87 (0.788, 4.26)
ϵ_g	Community transmission coefficient by RSV group	<i>Comm.rsv.a</i>	0.00338 (0.00203, 0.00530)
		<i>Comm.rsv.b</i>	0.00615 (0.00388, 0.00926)

In short, susceptibility to infection was reduced by previous infection whether these infections were homologous (*Prev.hom* = 0.53 (0.32 - 0.83)) or heterologous (*Prev.het* = 0.61 (0.3 - 1.1)). Increasing age also reduces susceptibility with ages 1-4 years old having an estimated 8% reduction (*Sus.age.2* = 0.92 (0.48 - 1.9)), ages 5-15 years a 73% reduction (*Sus.age.3* = 0.27 (0.14 - 0.53)) and ages ≥ 15 years an 84% reduction (*Sus.age.4* = 0.16 (0.08 - 0.32)). The within household transmission coefficients (*HH.rsv.a* = 0.019 (0.0073 - 0.04) and *HH.rsv.b* = 0.015 (0.0058 - 0.033)) are estimated higher than the community transmission coefficients (*Comm.rsv.a* = 0.0034 (0.002 - 0.0053) and *Comm.rsv.b* = 0.0062 (0.0039 - 0.0093)). The coefficient modifying within household exposure by size (*HH.size* = 0.42 (0.27 - 0.7)) suggests that larger households have less risk of pair-wise within household transmission

($HH.Risk_{h,g,j \rightarrow i}(t)$) than smaller households. However the total risk of household transmission ($\sum_{j \neq i} HH.Risk_{h,g,j \rightarrow i}(t)$) can conceptually be higher than that in smaller households if there are 20 or more infectious household members at a single time point, this is illustrated in Figure A2. 7. However, it should be noted that in this study, the highest number of simultaneously infectious individuals in large households was 14.

Although there is suggestion that pre-school individuals are the least likely to acquire infection from the community, and school-age individuals and older are the most likely to acquire community infection, the evidence is very weak: the relative estimate for age groups 1-4 years is $Exp.age.2 = 0.56$ (0.21 – 1.5) while for age group ≥ 5 years is $Exp.age.3 = 1.9$ (0.78 – 4.2). Symptomatic individuals are more infectious than asymptomatic individuals, more so those with high viral load, the relative estimate for high viral load symptomatic shedders is given as $High.Sym=6.7$ (2.6 – 16). However, there are not enough instances where individuals have high viral load and are asymptomatic to quantify the relative infectiousness of this specific combination, the relative estimate for high viral load asymptomatic shedders, $High.Asym$, has a very wide 95% CrI. Given 71132 person days of observation (493 individuals * 180 days of data, minus days individuals were away), 1021 had RSV A shedding, of which 49 were asymptomatic high viral load shedding days, and 1227 had RSV B shedding with 49 days of asymptomatic high viral load shedding. Given the inability to distinguish between the infectiousness of high versus low viral load asymptomatic shedders, we will not make this distinction in subsequent results and instead just refer to asymptomatic shedders in general.

For a better understanding of the within household and community transmission coefficient parameters, we calculated the different rates of exposure and plotted them as shown in Figure 3. 3.

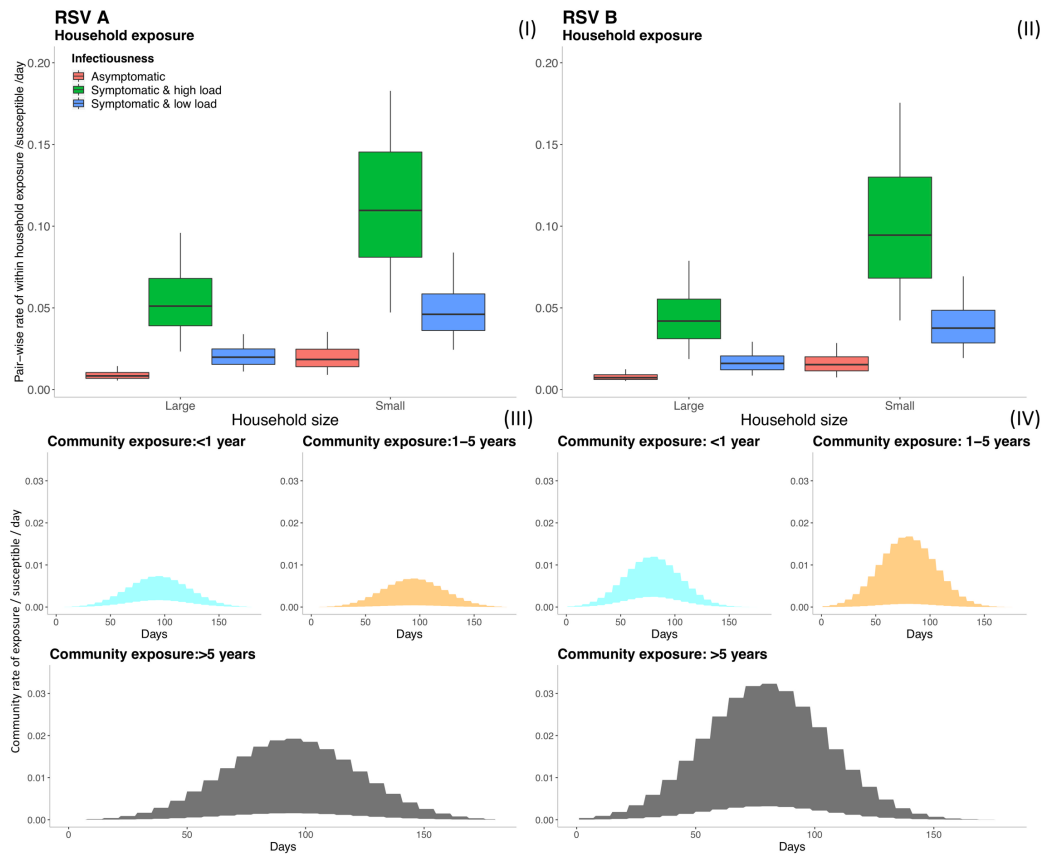


Figure 3. 3: Comparing the range of within household exposure rate and community exposure rate for a single susceptible individual given different heterogeneities in exposure and infectiousness.

Top row: The box plots show the 0.025, 0.25, 0.5, 0.75 and 0.975 percentiles for the rate of exposure per person per day between a single susceptible and a single infectious housemate ($HH_Risk_{h,g,j \rightarrow i}(t)$) for RSV A (I) and RSV B (II). The distributions of rate are categorized by household size and the infectiousness based on viral load and symptom status (see text). Note: outliers have been removed from the box plots for better visualization. Bottom row: The shaded graphs show the range of values over time for the rate of exposure from the community to a single susceptible individual ($Comm_Risk_{i,g}(t)$) for RSV A (III) and RSV B (IV). The graphs are color-coded by the age group of the susceptible individual. The ranges for each age group are determined by the 95% CrI of the parameters that go into the calculations, hence the shaded regions show 95% CrI of the community exposure rate.

Given two competing sources of infection, an infectious housemate and a source outside of the household, a susceptible individual is more likely to get infected within

the household rather than from the community. There is a suggestion that RSV A has a higher transmission potential at the household level relative to RSV B, while the situation is reversed at the community level. However, there is considerable overlap between the distributions of within household transmission coefficient for RSV A and that for RSV B as seen in Figure A2. 6, which shows the distribution of the parameters on the log scale, which is mirrored in the rate of household exposure shown in Figure 3. 3.

We observed some correlations in the estimated parameters. In particular there were strong positive correlations within the relative susceptibility by age parameters. The within household transmission coefficient for RSV A was strongly positively correlated with the within household transmission coefficient for RSV B. The age effects of susceptibility were strongly negatively correlated with the age effects on community exposure. Figure A2. 8 in the supplementary index shows all the pairwise correlation patterns.

Given the posterior densities for the parameters, we calculated the source with the highest likelihood for each infection. While respecting the correlation patterns observed in Figure A2. 8, we sampled 10 different parameter sets and for each, we calculated the proportion of cases whose most likely source was an infectious housemate. The changes made to the likelihood equation to allow for this calculation are described in the appendix A2. For all the infection cases, 32-53% of them were attributed to transmission within the household. For RSV A, this range was 40-59%, while for RSV B it was 26-48%.

To check if any information is lost when we have less data, we refitted the data in three additional ways: RSV A alone, RSV B alone and RSV with no distinction between groups. The results are shown in Table A. 1 in the supplementary index. In reducing the data used to infer parameters we notice that more posterior densities for the relative effect parameters now include 1 in their 95% credible interval, as can be expected. In general, the trends with age, household size and relative infectiousness, as seen in Figure A2. 6, are maintained. However, when RSV is treated as one entity, the protective effect of previous infection is reduced, symptomatic cases are more

infectious, and the estimate of the community transmission coefficient is increased. This suggests that misclassification of viruses disrupts the ability of the model to track transmission patterns, resulting in a greater propensity to account for infections as spontaneous.

Model validation and sensitivity analysis

To validate the model, we checked to see that the range of simulated epidemics contained the real data; then we chose a single simulation with known parameters and re-estimated to see if the posterior distribution contained the known values. Details of this process can be found in the appendix A2, but in general, we were satisfied that the model was working as expected. Figure 3. 4 shows multiple simulated epidemics for different parameter sets relative to the real data. From this we see that as with the real data, the simulations show the RSV B epidemic taking off earlier than the RSV A epidemic. There is a tendency for simulate epidemics to be larger than that observed in terms of total number of cases (Figure A2. 23).

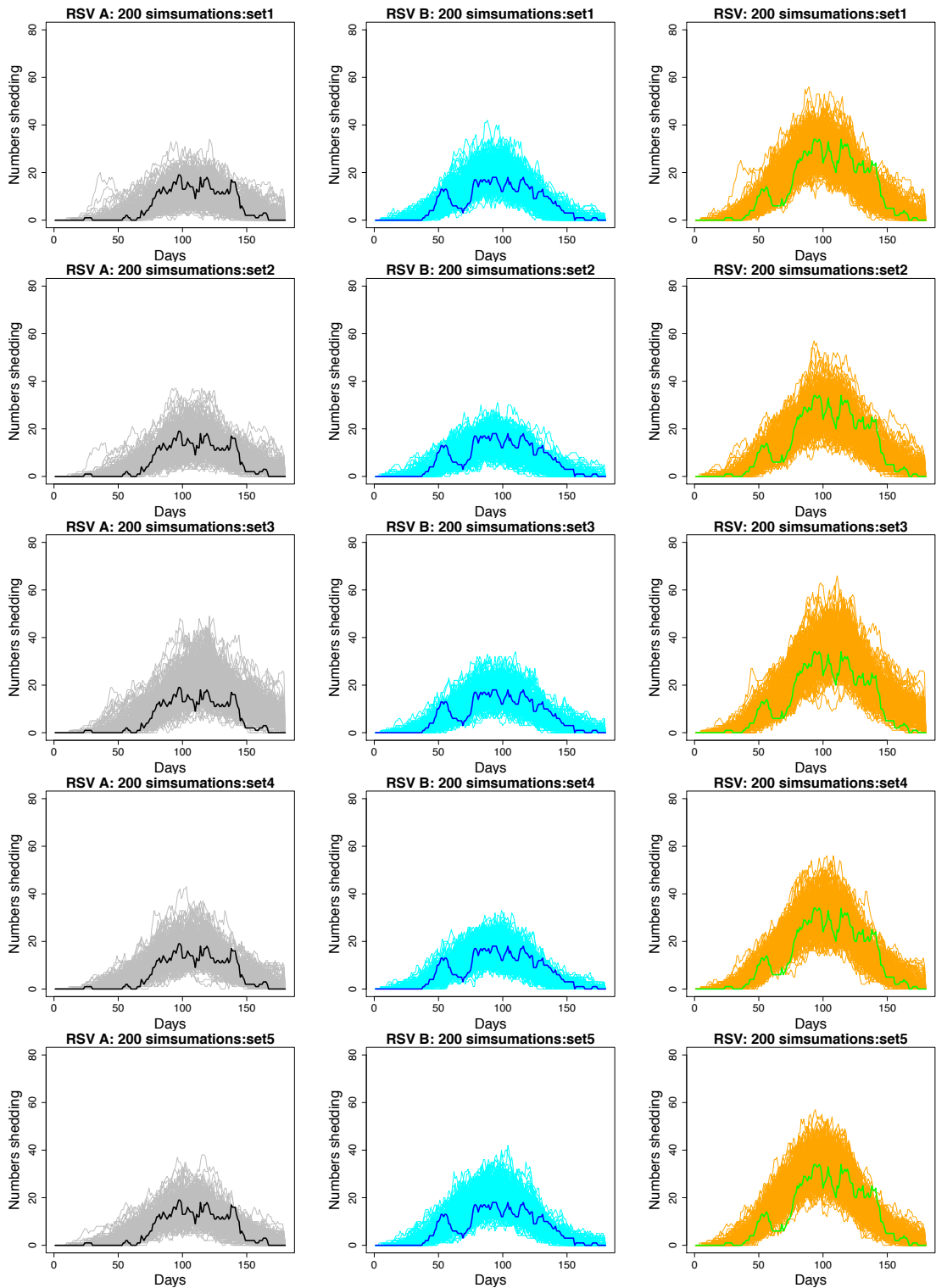


Figure 3. 4: A comparison between the simulated data and real epidemics using simulations from 5 different parameter sets estimated from the full model (row 1 to 5).

First column: RSV A simulated epidemics (grey lines) compared to real data (thick black line). Second column: RSV B simulated epidemics (light blue lines) compared to real

data (thick blue line). Third column: RSV simulated epidemics (orange lines) compared to real data (thick green lines).

We performed a sensitivity analysis to check the robustness of our results to the background community density function. We used 3 additional background functions and found that despite a change in summary values for the parameters, in general the trends were maintained. These results are shown in the appendix A2. They show that the results are robust to the choice in the shape of background community density function.

Finally, we removed the largest household (which had a very large RSV A outbreak but only a single RSV B case) from the data to check if this would change the patterns of the within household transmission coefficients. The results, shown in the appendix A2, were robust to these changes.

Following the validation of the model, we simulated epidemics altering the degree of infectiousness. Initially we reduced the infectiousness of symptomatic individuals to predict the effect of reducing RSV related ARI; then we assumed that asymptomatic individuals are not infectious in order to quantify the contribution of asymptomatic infections to transmission. The results show that reducing infectiousness of symptomatic individuals to the level of asymptomatic individuals lowers the distribution of total number infected. Assuming that asymptomatic individuals are not infectious also tends to decrease the total number infected (see Figure A2. 23 in appendix A2). We also removed the asymptomatic shedding episodes from the data and re-estimated the parameters to check what the effect of only having sampled symptomatic individuals would be. We found that we lose precision in the estimates of the relative infectiousness parameters, previous infection is estimated as being more protective as is being ≥ 15 years old (Figure A2. 24 and Figure A2. 25).

3.7. Discussion

We developed an individual based approach to make Bayesian based inference on transmission parameters using MCMC. We set out to better understand RSV transmission within a household setting using cohort data collected with unprecedented detail during the course of a single RSV epidemic in a rural coastal community in Kenya.

Older individuals are less susceptible to detectable infection, presumably due to immunity acquired in previous epidemics. We found strong evidence of partial immunity to homologous re-infection within the same epidemic for the RSV groups. The effect of previous infections is captured in two different ways in our model. Age (*Sus.age* parameters) captures the combined effect of age and experience of epidemics prior to the one under study, while the estimates for the effect of previous observed infections (*Prev.hom* parameter), captures effect of infections in the current epidemic. It is therefore implicit that immunity to RSV is built up in the long term, from one epidemic to the next and in the short term from one infection to the next. The evidence for cross-immunity between RSV A and B was weaker, which presumably allowed the two virus groups to co-circulate in this epidemic. However, typically, RSV epidemics are dominated by one or other of group A or B and so the particular circumstances of this epidemic might not always hold. It remains to be explored how this individual level parameter estimate is translated into population dynamics.

We found some evidence that individuals aged ≥ 5 years were the most likely to get infection from a community source (less likely to get infected during a household outbreak). This means that given our assumption of latent periods between 2-5 days, which forms the temporal link between cases, individuals ≥ 5 years were the most often identified as index cases in a household outbreak relative to the younger age groups. We have not considered an age-dependent latent period and estimating the latent period from these data is a future goal. The ≥ 5 years age group contains school going children and our result is in line with those of Munywoki et al ²⁰, based on a different analysis of the same study, who found that school-going children were often

initiating household outbreaks. Establishing transmission chains using genomic information could strengthen this result.

We have assumed that the community risk of infection changes smoothly over time and is homogeneous apart from an age effect. These assumptions are necessary as community infections are not completely observed. We are confident that these assumptions do not have significant influence on our estimates of within-household transmission (which is fully observed), but may result in an over-estimate of community exposure, which will be more heterogeneous than we have assumed. Consequently, the simulated epidemics are larger in total numbers than that observed, Figure 3. 4, and our results of up to one half of infections arising from within the household are likely to be a minimum. Data on genetic relatedness between viral isolates will clarify the extent to which individuals are infected from the community during a household outbreak.

By separating RSV A and RSV B we find that RSV B has a higher rate of introduction into the household, and RSV A is more transmissible once in the household, an observation also made by ³⁶ from a phylogenetic analysis of RSV A sequences. This, together with the fact that RSV A had a larger proportion of cases attributed to within household transmission, suggests that there might be some niche separation, explaining how and why these two different groups are able to co-exist and remain separate. It should be noted however that the difference in the distribution of the within household transmission coefficient between the RSV groups is not large, there is a substantial overlap of credible intervals. As such, whatever advantage RSV A might have over RSV B at the household level is small in terms of transmission but might be larger in terms of interaction with other respiratory viruses, and small differences in individual based parameters might translate into large population effects. In the present epidemic, the RSV B epidemic takes off earlier than the RSV A epidemic despite the first case being RSV A (Figure 3. 2). In addition to which, we see that despite RSV B infecting more households than RSV A, RSV A infects a larger proportion of household members (**Error! Reference source not found.**). An examination of the comparative dynamics of RSV A and B within epidemics might be a good way to understand how they interact.

With the definition of a household as a group of individuals living in the same compound and eating food from the same kitchen, we found that the pairwise rate of within household transmission is higher in small households than large ones. The relationship between household size and pair-wise rate of transmission has been observed before for Influenza, ^{11,12,14,15}, however going a step further we show that if households are structured such that they can have at least 20 simultaneously infectious occupants (possible if several members of an extended family live in the same household as is the case in the present study) then larger households will tend to contribute more to transmission than smaller households.

We looked at a combination of presence of symptoms and viral load to infer infectiousness. We found that being symptomatic is of key importance. In general, symptomatic individuals were more infectious, particularly if shedding large amounts of virus. Though this result is not surprising it has an important implication on vaccine effectiveness. If an RSV vaccine works by reducing or preventing disease in the form of an ARI, this will in turn have an impact on transmission potential and we should expect to see reduced morbidity and infection. To check what that potential impact of such a vaccine would be, we simulated epidemics where the infectiousness of symptomatic individuals was equal to that of asymptomatic individuals and we found a significant shift in the overall distribution of simulated case towards smaller total numbers infected. The shift was more for ages between 1 and 15 years, given that this group also had the larger fraction of symptomatic cases, the observation from simulations with reduced infectiousness suggests largely assortative mixing within this group, which in turn means largely assortative transmission. The number of cases in the <1 year age group is not greatly altered by reducing the infectiousness of symptomatic individuals, implying that there are several sources of infection to the infant and reducing or removing only one has little impact Figure A2. 23.

We reduced the model complexity to look at RSV as a single pathogen without distinguishing between groups. This resulted in skewing the parameter estimates away from within household transmission and towards spontaneous infection from external sources, as a result of introductions due to RSV A and RSV B being treated as multiple introduction of the same pathogen thus compounding the effect of community

transmission. This, in addition to the reduced protective effect of previous infection due to misclassification of re-infections, led to the within household transmission parameter being underestimated in order for the model to account for the observed number of infections. In addition, temporally linking RSV A and B cases as a result of misclassification also led to the effect of symptoms on transmission being overestimated. This suggests that the estimates obtained in the present analysis are likely to change if we further classified the cases into RSV subgroups. This goes to illustrate the importance of making distinctions between pathogens in order to obtain accurate estimates of transmission parameters. At any given moment multiple pathogens are co-circulating in a host population, this household study alone had multiple viruses spreading in large numbers during the time of data collection ⁴⁵. How these pathogens interact could have dramatic implications for parameter estimates, and ultimately on how control strategies are implemented. We have seen the effect of the pneumococcal vaccine on the non-vaccine serotypes and how it might mitigate vaccine effectiveness ⁴⁶ and a study on influenza has shown evidence of its controlling effect on other pathogens ⁴⁷. There is an increasing call from such observations to understand how multiple pathogens interact at the host population level.

Our study is not without limitations. The households in the study were selected based on the presence of an infant born after the previous RSV epidemic and older siblings to the infant in order to determine who infects the infant. As such the sample is not random and this might introduce bias in the parameter estimates, the extent of which we are uncertain. Relative to other studies, our sample size in terms of number of households is small. However, the intensive sampling regardless of symptoms means we had less biased observation of infections relative to index-case ascertained household studies that rely on symptom reporting by household contacts. In our study we had 47.2% of RSV A and 40.2% of RSV B positive samples that were symptomatic, 60.8% of RSV A and 55.2% of RSV B episodes were symptomatic. Estimation of parameters only using data from symptomatic episodes shows similar parameter estimates, although with loss of precision, especially in terms of differential infectiousness Figure A2. 24. In addition, sampling was done every 3 or 4 days, which means that short duration infections might have been missed, and we do not have serological data to complement the PCR results. There were 14 instances of RSV A and

RSV B co-infections, 7 of these were apparent co-onset shedding episodes. Our method of imputing the start of a shedding episode is based on the gap between the last negative sample before the first positive sample of the episode, and the first positive sample of the episode. For the co-onset cases, this gap ranged between 3-4 days, and the start of onset was imputed as being halfway between the gap. These may or may not be true co-onset cases, it would require the existence of daily sampling to confirm. Treating the start and end of a shedding episode as augmented data is an alternative to the mid-point estimation, if applied that could lead to different onset days being inferred.

The present analysis could be extended in several ways. We used interpolated shedding durations; it would be an added advantage to use the data to estimate a distribution of shedding durations that could potentially be more generalizable. The inclusion of other sources of information into the analysis could improve parameter inference, as was the case with Li *et al* and the inclusion of genetic data⁴⁸. The inference made on within-household transmission compared to community transmission is based on the latency distribution that links onset of cases. This is a temporal linking of cases that is not always correct. A combination of temporal and genetic distance would allow better inference on linked cases and consequently the competition between within-household and community source transmission. Finally, the RSV A and B model could be used to look at other pathogen interactions and perhaps incorporate more than two pathogens.

In conclusion, our analysis presents the first transmission modelling of cohort data for RSV and we find that it is important to factor in household size and social structuring – such as the tendency for households to contain several members of the extended family – when modelling transmission. It is also important to model competing risks of infection from within the household and the community. There are questions on the mechanisms that allow co-existence of RSV groups temporally and geographically. The weak cross immunity between RSV groups demonstrated by our analysis and the possibility of different transmission niches could form part of the explanation for the co-existence.

3.8. References:

1. Shi, T., McLean, K., Campbell, H. & Nair, H. Aetiological role of common respiratory viruses in acute lower respiratory infections in children under five years: A systematic review and meta-analysis. *J. Glob. Health* **5**, 1–10 (2015).
2. Nair, H. *et al.* Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet* **375**, 1545–1555 (2010).
3. Shi, T. *et al.* Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. **390**, 946–958 (2017).
4. PATH. RSV Vaccine Snapshot - PATH Vaccine Resource Library. (2015). Available at: <http://vaccineresources.org/details.php?i=1562>. (Accessed: 15th July 2019)
5. WHO | WHO vaccine pipeline tracker. *WHO* (2016). Available at: https://www.who.int/immunization/research/vaccine_pipeline_tracker_spreadsheet/en/. (Accessed: 29th July 2019)
6. Kinyanjui, T. M. *et al.* Vaccine Induced Herd Immunity for Control of Respiratory Syncytial Virus Disease in a Low-Income Country Setting. *PLoS One* **10**, e0138018 (2015).
7. Anderson, L., Dormitzer, P. & Nokes, D. Strategic priorities for respiratory syncytial virus (RSV) vaccine development. *Vaccine* **31 Suppl 2**, B209-15 (2013).
8. Poletti, P. *et al.* Evaluating vaccination strategies for reducing infant respiratory syncytial virus infection in low-income settings. *BMC Med.* **13**, 49 (2015).
9. Wu, J. T., Riley, S., Fraser, C. & Leung, G. M. Reducing the Impact of the Next Influenza Pandemic Using Household-Based Public Health Interventions. **3**, (2006).
10. Klick, B. *et al.* Transmissibility of seasonal and pandemic influenza in a cohort of households in Hong Kong in 2009. *Epidemiology* **22**, 793–796 (2011).
11. Lau, M. S. Y., Cowling, B. J., Cook, A. R. & Riley, S. Inferring influenza dynamics and control in households. *Proc. Natl. Acad. Sci.* **112**, 9094–9099 (2015).
12. Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. & Boëlle, P. Y. A Bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data. *Stat. Med.* **23**, 3469–3487 (2004).

13. Yang, Y. *et al.* Household transmissibility of avian influenza a (H7N9) virus, China, february to may 2013 and october 2013 to march 2014. *Eurosurveillance* **20**, 1–11 (2015).
14. Cauchemez, S. *et al.* Household Transmission of 2009 Pandemic Influenza A (H1N1) Virus in the United States. *N Engl J Med* **27361**, 2619–27 (2009).
15. House, T. a. *et al.* Estimation of outbreak severity and transmissibility : influenza A(H1N1)pdm09 in households. *BMC Med.* **10**, 117 (2012).
16. Tsang, T. K., Lau, L. L. H., Cauchemez, S. & Cowling, B. J. Household Transmission of Influenza Virus. *Trends in Microbiology* **24**, 123–133 (2016).
17. Hall, C. B. *et al.* Respiratory syncytial virus infections within families. *N. Engl. J. Med.* **294**, 414–419 (1976).
18. Heikkinen, T., Valkonen, H., Waris, M. & Ruuskanen, O. Transmission of respiratory syncytial virus infection within families. *Open Forum Infect. Dis.* **2**, ofu118 (2015).
19. Jacoby, P., Glass, K. & Moore, H. C. Characterizing the risk of respiratory syncytial virus in infants with older siblings: A population-based birth cohort study. *Epidemiol. Infect.* **145**, 266–271 (2017).
20. Munywoki, P. K. *et al.* The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya. *J. Infect. Dis.* **209**, 1685–1692 (2014).
21. Graham, B. S. Protecting the family to protect the child: Vaccination strategy guided by RSV transmission dynamics. *J. Infect. Dis.* **209**, 1679–1681 (2014).
22. Munywoki, P. K. *et al.* Frequent Asymptomatic Respiratory Syncytial Virus Infections During an Epidemic in a Rural Kenyan Household Cohort. *J. Infect. Dis.* 1–8 (2015). doi:10.1093/infdis/jiv263
23. Munywoki, P. K. *et al.* Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiol. Infect.* **143**, 804–12 (2015).
24. Wathuo, M., Medley, G. F., Nokes, D. J. & Munywoki, P. K. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res.* **1**, 27 (2016).
25. Cane, P. a. Molecular epidemiology of respiratory syncytial virus. *Rev. Med.*

- Viol.* **11**, 103–16 (2001).
26. Zlateva, K. T., Lemey, P., Moe, E., Vandamme, A. & Ranst, M. Van. Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein. **79**, 9157–9167 (2005).
 27. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol. Infect.* **133**, 279–289 (2005).
 28. Agoti, C. N. *et al.* Genetic relatedness of infecting and reinfecting respiratory syncytial virus strains identified in a birth cohort from rural Kenya. *J. Infect. Dis.* **206**, 1532–41 (2012).
 29. Agoti, C. N., Gitahi, C. W., Medley, G. F., Cane, P. a & Nokes, D. J. Identification of group B respiratory syncytial viruses that lack the 60-nucleotide duplication after six consecutive epidemics of total BA dominance at coastal Kenya. *Influenza Other Respi. Viruses* **7**, 1008–12 (2013).
 30. Thongpan, I. *et al.* Respiratory syncytial virus genotypes NA1, ON1, and BA9 are prevalent in Thailand, 2012–2015. *PeerJ* **5**, e3970 (2017).
 31. Song, J. *et al.* Emergence of ON1 genotype of human respiratory syncytial virus subgroup A in China between 2011 and 2015. *Sci. Rep.* **7**, 1–9 (2017).
 32. Rodriguez-Fernandez, R. *et al.* Respiratory Syncytial Virus Genotypes, Host Immune Profiles, and Disease Severity in Young Children Hospitalized With Bronchiolitis. *J. Infect. Dis.* **217**, 24–34 (2017).
 33. Park, E. *et al.* Molecular and clinical characterization of human respiratory syncytial virus in South Korea between 2009 and 2014. *Epidemiol. Infect.* 1–17 (2017). doi:10.1017/S0950268817002230
 34. Botosso, V. F. *et al.* Positive Selection Results in Frequent Reversible Amino Acid Replacements in the G Protein Gene of Human Respiratory Syncytial Virus. *PLoS Pathog.* **5**, e1000254 (2009).
 35. Pretorius, M. A. *et al.* Replacement and positive evolution of subtype A and B respiratory syncytial virus G-protein genotypes from 1997-2012 in South Africa. *J. Infect. Dis.* **208**, 227–237 (2013).
 36. Agoti, C. *et al.* Transmission patterns and evolution of RSV in a community outbreak identified by genomic analysis. *Virus Evol.* (2017). doi:In print

37. Munywoki, P. K. Transmission of Respiratory Syncytial Virus in Households : Who Acquires Infection From Whom. (Open University UK, 2013).
38. White, L. J. *et al.* Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math. Biosci.* **209**, 222–239 (2007).
39. Pan-Ngum, W. *et al.* Predicting the relative impacts of maternal and neonatal respiratory syncytial virus (RSV) vaccine target product profiles: A consensus modelling approach. *Vaccine* **35**, 403–409 (2017).
40. Lee, F. E., Walsh, E. E., Falsey, A. R., Betts, R. F. & Treanor, J. J. Experimental infection of humans with A2 respiratory syncytial virus. *Antivir. Res* **63**, 191–196 (2004).
41. Camacho, A. & Funk, S. fitR: Tool box for fitting dynamic infectious disease models to time series.R package version 0.1. (2017).
42. Roberts, G. O. & Rosenthal, J. S. Examples of Adaptive MCMC. *J. Comput. Graph. Stat.* **18**, 349–367 (2009).
43. Raftery, A. E. & Lewis, S. How many iterations in the Gibbs sampler? *Bayesian Stat.* 763--773 (1992).
44. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
45. Munywoki, P. K. *et al.* Continuous invasion by respiratory viruses observed in rural households during a respiratory syncytial virus seasonal outbreak in coastal Kenya. *Clin. Infect. Dis.* ciy313–ciy313 (2018).
46. Kwambana-Adams, B. *et al.* Rapid replacement by non-vaccine pneumococcal serotypes may mitigate the impact of the pneumococcal conjugate vaccine on nasopharyngeal bacterial ecology. *Sci. Rep.* **7**, 1–11 (2017).
47. Zheng, X., Song, Z., Li, Y., Zhang, J. & Wang, X. L. Possible interference between seasonal epidemics of influenza and other respiratory viruses in Hong Kong, 2014-2017. *BMC Infect. Dis.* **17**, 1–7 (2017).
48. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evol.* **34**, 2982–2995 (2017).

4. Paper 2: Integrating epidemiological and genetic data with different sampling densities into a dynamic model of RSV transmission.

4.1. Overview

This chapter was written in fulfilment of the second part of the second objective, and the third objective. It is an extension of the methods developed in the previous chapter. As with the previous chapter, this chapter is written in the format of a publication and we intend to submit it to a journal with the running title: ***Integrating epidemiological and genetic data with different sampling densities into a dynamic model of RSV transmission.***

4.2. Role of candidate

I formulated the problem, conducted the numerical analysis and wrote the first draft of the chapter. Revisions were made with feedback, input and guidance from my supervisors Graham F. Medley and D. James Nokes, and advisor Marc Baguelin.

Research paper cover sheet



London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	Ish1601271	Title	Ms
First Name(s)	Ivy Kadzo		
Surname/Family Name	Kombe		
Thesis Title	Integrating viral RNA sequence and epidemiological data to define transmission patterns for respiratory syncytial virus		
Primary Supervisor	Professor Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	<i>To be decided</i>
Please list the paper's authors in the intended authorship order:	<i>Kombe I.K., Agoti C.N., Munywoki P.K., Baguelin M., Nokes D.J., Medley G.F.</i>
Stage of publication	Choose an item. <i>Pre-submission draft</i>


SECTION D – Multi-authored work

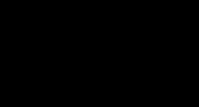
Improving health worldwide

www.lshtm.ac.uk

<p>For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)</p>	<p><i>I led the analysis, including writing all the code, wrote the first draft and incorporated co-author comments.</i></p>
---	--

SECTION E

Student Signature	
Date	<i>26/07/2019</i>

Supervisor Signature	
Date	<i>25 July 2019</i>

4.3. Abstract

Respiratory syncytial virus (RSV) is responsible for a significant burden of respiratory illness in children under 5 years old. A maternal vaccination against RSV disease has recently completed phase III trials where it was reported as being modestly efficacious. Prior to rolling out any vaccination program, a clear understanding of the transmission dynamics is necessary in order to predict which vaccination strategies would be the most effective. We built a dynamic model calibrated at the individual host level that integrated social-temporal data on shedding patterns and genetic clustering patterns derived from a phylogenetic analysis. Through aggregating the genetic information into clusters and the use of data augmentation, we were able to integrate data types of different sampling densities into a single framework. In this study population of 493 individuals with 55 infants under the age of 1 year distributed across 47 households, we found that 52% of RSV B and 60% of RSV A cases arise from infection within the household. Fifty-five percent of infant RSV A infections occur in the household, as do 36% of infant RSV B. Frequently the source of infant infection is a child aged between 2 and 13 years living in the same household as the infant. These results further highlight the importance of school-aged children in RSV transmission, particularly the role they play in directly infecting the infant at the household level. This age group could provide an alternative vaccination target group.

4.4. Introduction

Ever since the term phylodynamics first came into use in 2004¹ there has been an increasing interest to analyse genetic sequences of pathogens while accounting for the epidemiology of the infection: the main question being, how do the pathogen's transmission dynamics shape the observed genetic relationships and conversely, how does the evolution of the pathogen influence how it is transmitted? Developing the methodology that accurately captures both the epidemiological and evolutionary processes, and that is computationally tractable is challenging. The field of phylodynamics has grown to include other multiple data types, more so to determine transmission chains during an outbreak²⁻⁴. A range of methods have been developed with variations observed in the nature of sequence data (single^{2,3,5,6} versus multiple^{7,8} sequences per infected host); complexity of genetic model (coalescent^{7,9,10} versus simple genetic distance models⁵); complexity of transmission model (dynamic transmission models^{2,11,12} versus simple temporal distance models^{4,5}); generalizability across pathogens (often available in packages such as SCOTTI⁷, Outbreaker^{5,6}, PhyDyn¹³, TransPhylo¹⁴); types of data (sequences and collection dates^{5,7,12}; sequences, collection dates and location of host^{3,4}; sequences, collection dates and contact data²) and ability to account for unsampled cases^{5,7,12,14}. A recent attempt has been made to compare the utility of several methods, and though only 9 published models were used, the authors came to the conclusion that "Each model had its own strengths related to the purpose for which it was developed, and limitations related to its assumptions"¹⁵. It is therefore crucial for investigators to bear in mind their specific study design in choosing a method to adopt.

Despite the existence of a wide assortment of methods for integrated epidemiological and evolutionary analyses, most of them use data types that are at the same sampling density, meaning there usually are as many cases as are identified by the generated sequences. To account for there being cases without sequences, some methods estimate or fix the fraction of the outbreak that is unobserved^{5-7,12,14}. Uniquely, the method developed by *Lau et al* can explicitly model confirmed cases that do not have sequence data since the method imputes missing sequences¹². There have been great advancements in the generation of pathogen genetic sequences, however, whole

genome sequencing is not always successful, more so when the pathogen is present at low loads within the host. Given that it might not always be possible to generate genetic sequences for a majority of samples from an outbreak, it would be useful if an analysis technique can simultaneously make use of the sequence data where available, and more readily available spatial or social-temporal shedding patterns to make inferences on transmission characteristics. In this article, we model social-temporal data from an outbreak of respiratory syncytial virus (RSV) and genetic data that covers ~50% of the observed cases to infer the determinants of transmission within a group of households over a 6-month follow-up period.

Respiratory syncytial virus infects all age groups but causes a significant lower respiratory disease burden in children <5 years old, more so in < 6 months old, and the elderly^{16–18}. RSV virus is a negative stranded RNA virus (length ~15,200 bases) that exists in two antigenically and genetically distinct groups estimated to have diverged 350 years ago¹⁹. It spreads in seasonal patterns with most places experiencing annual cycles^{20–23}. Phylogenetic analysis of RSV whole genome sequences from different countries that span several years have estimated mutation rates between 6×10^{-4} and 7×10^{-4} substitutions/site/year^{24–26}. These studies found that the clustering pattern of RSV sequences in the long term is more temporal than geographical. In the short term, changes in the dominant transmitting genotype have been used to understand transmission patterns, as has been the case with studies looking at the distribution of the RSV A ON1 genotype^{27–29}. Genotype replacement from one RSV season to the next is common^{23,30,31}, however short term changes to the genome over the course of a single epidemic could help to determine transmission chains across a limited geographical space. It is such short-term changes observed in the RSV genome that we exploited in the analysis presented here.

The model we use is an extension of our previous work where we successfully used social-temporal patterns of shedding coupled with demographic information on the host to identify symptom status and virus load, household size, age and recent infection history as determinants of transmission at the individual level. In addition, by virtue of RSV A having slightly higher estimates of the parameter quantifying within household transmission relative to RSV B, we hypothesized that RSV A having might

have a transmission niche at the household level³². The question we aim to answer is whether increased resolution in pathogen identification improves inference on transmission characteristics; do weak signals in previously inferred parameters, e.g. a slight difference in the within-household transmission coefficient for RSV A relative to RSV B, become clearer? The easiest place to start would be to build on an existing tool for data integration. The 'Outbreaker' model^{5,6} takes a modular approach by establishing a model for the epidemiological data (epidemiological likelihood) and another for the genetic data (genetic likelihood) and then combines these into a single likelihood for inference on parameters and transmission pairs given genetic sequences and their dates of collection. This modular approach can be used with different data types; temporal (sampling times or exposure data), genetic and spatial^{3,4,12}. An alternative, more classical phylodynamics approach, would be to link the equations of the epidemiological model to the rate of coalescence of the model of evolution and estimate parameters based on the genetic sequences as was the case in the *Li et al* model³³. Joint inference of epidemiological and genetic characteristics would be ideal, however writing down a likelihood and developing an inference technique given different data densities is not straightforward, as demonstrated by *Lau et al*¹² and, depending on what fraction of the outbreak is missing genetic information, an attempt at joint inference can lead to significant inaccuracies even for fairly sophisticated methods¹⁵. Instead, we will take the two-staged approach of first 'learning' from the sequence data and then using the inferred traits within the epidemiological model, similar to^{14,34-37}. The aim of our analysis is therefore to attempt to enrich the densely sampled epidemiological data with the genetic data that is available at a lower sampling density.

4.5. Methods

The model that we extended for this analysis is an individual level transmission model that is calibrated by day and individual host (see Chapter 3). The individuals represented in the data are grouped into households according to the demographic information provided. Individuals can get infected within the household from a sampled infectious individual or from an unsampled infectious individual outside of the household represented by a background community rate. Distinctions are made between RSV A and RSV B group infections and interaction between the two groups is modelled through modified susceptibility to heterologous group reinfection. Further details can be found in ³².

The data used in the present analysis consists of shedding durations imputed from the results of samples collected every 3-4 days, information on symptom status and information on presence or absence from the household. Given the discontinuity in the sampling, complete shedding, ARI and presence/absence durations had to be imputed. This imputation process has been described in detail in A2: Supplementary appendix for Paper 1. In brief, an RSV A/B shedding episode is defined as a period within which an individual provided PCR positive samples for RSV A/B that were no more than 14 days apart. A shedding episode is referred as symptomatic if within the window of virus shedding, there is at least one day where symptoms were recorded. The symptoms of interest are those of an acute respiratory illness (ARI), which are: cough, or nasal discharge/blockage, or difficulty breathing. Individuals are assumed to start shedding halfway between the last negative sample and the first positive sample of the episode, and they stop shedding halfway in between the last positive sample of the episode and the first negative sample. In the same way, complete ARI durations are imputed within shedding episodes and complete presence/absence durations are imputed for all the days of data available for a particular individual. There are some instances where an individual was present but not sampled, as such, presence could not purely be identified by the availability of NPS samples. Imputation was chosen over data augmentation to ensure consistency across studies analysing the same household dataset³⁸⁻⁴⁰.

In the same way that a shedding episode can be identified as RSV A or RSV B, we used the sequence data to further classify shedding episodes into genetic clusters. These genetic clusters are then treated in the same way as genetic groups in the model. Transmission is allowed between members of the same cluster but between clusters transmission is not. Figure 4. 1 illustrates how cases can become disconnected with addition of genetic information. If all the cases are identified at the pathogen level (i.e., all infecting viruses are alike), then the timing of cases is the only thing that informs possible transmission clusters. With information on RSV group, one knows that there are at least 2 transmission clusters since an RSV A case could only have been infected by another RSV A case, and so forth with the genetic clusters.

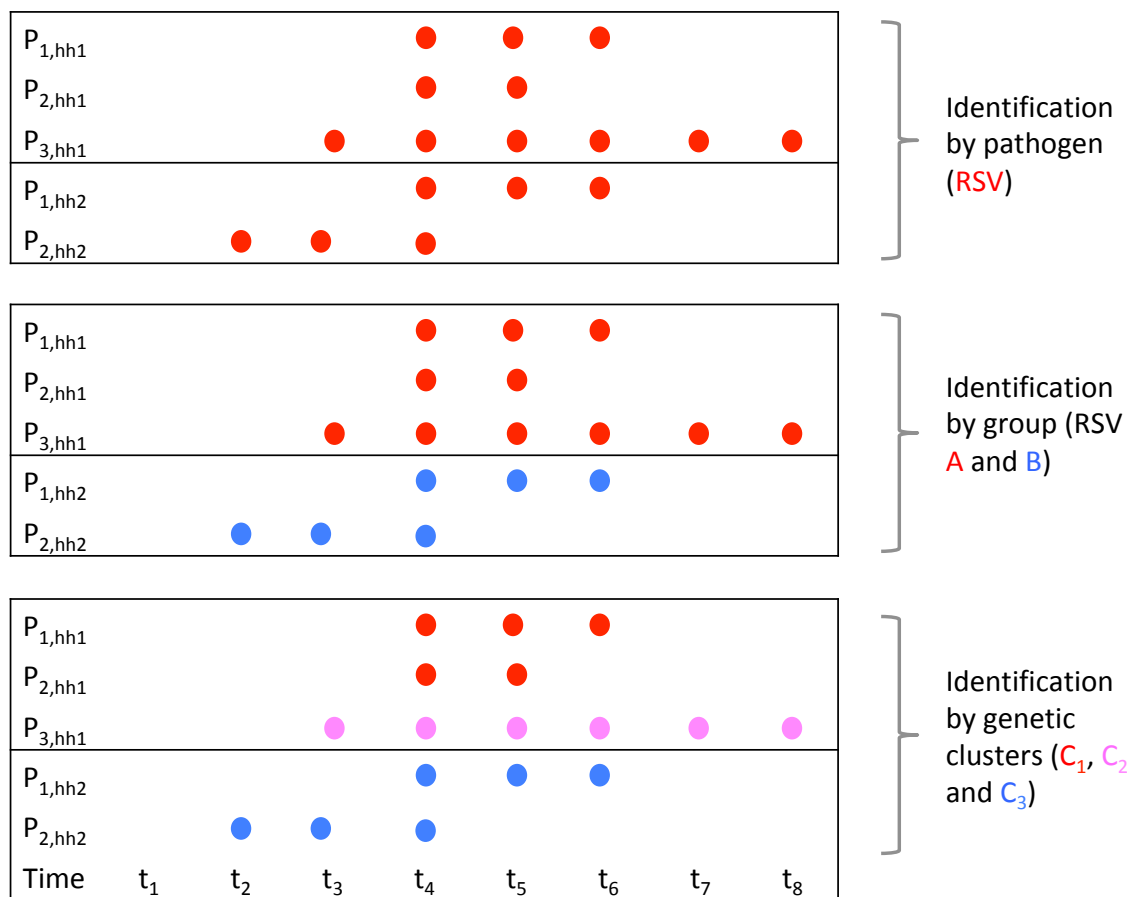


Figure 4. 1: Illustration of how cases become disconnected with added pathogen information

The sequences available from the outbreak are grouped into clusters according to a combination of criteria based on: nucleotide distance cut-off, clustering patterns on the global RSV phylogenetic tree and the inferred date of sequence divergence. Details

of this phylogenetic analysis can be found in ⁴¹. Figure 4. 2, part of Figure 2 in the original paper, shows the time-resolved maximum likelihood phylogenetic trees for RSV A and RSV B showing the estimated node edges and assigned clades and sub-clades. RSV A clustered into one clade with 5 sub-clades while RSV B clustered into 5 clades, two of which had two sub-clades each. For our analysis we do not make a distinction between clades and sub-clades as such we use 5 clusters for RSV A and 7 for RSV B.



Figure 4. 2: Time-resolved maximum likelihood phylogenetic trees for RSV A and RSV B from the *Agoti et al* ⁴¹ phylogenetic analysis.

4.5.1. Imputing missing genetic information

We did not have whole genome sequences for all the positive samples and as such, we needed to impute information where it is lacking. We decided to impute cluster identity rather than sequences, choosing to look at genetic clusters as a way to aggregate genetic information. Augmenting sequences has previously been done by *Lau et al* ⁴², but we proposed to use a simpler approach which does not need any assumptions on sequence evolution. Within a given RSV group, infection by a particular cluster is assumed to be a mutually exclusive process, an individual can only shed one cluster type at a time. The genetic data available is consensus whole genome sequences as such, only one cluster can be identified from a single sample. There are two levels of missing sequence data:

- Partially missing. Where only some of the positive samples in an episode have sequences.

- Completely missing. None of the positive samples in an episode have sequences

Partially missing

If all the sequences within a particular shedding episode belong to the same genetic cluster, then this cluster id is assigned to every day of the shedding episode. If the sequences belong to multiple clusters say C_1 and C_2 , the duration of shedding each is divided such that the first day of shedding up to and including the last day where C_1 appeared are assigned cluster C_1 , subsequent days are assigned C_2 up until the end of shedding, and so forth for >2 cluster identities.

Completely missing

Here we make a further distinction between cases that are part of a household outbreak that has some genetic information, and those that are not. Cases in a household will be assumed to be in the same outbreak if, either there is an overlap in shedding period, or the time between end of shedding of one case and onset in another is ≤ 5 days. Cluster assignment will proceed as follows:

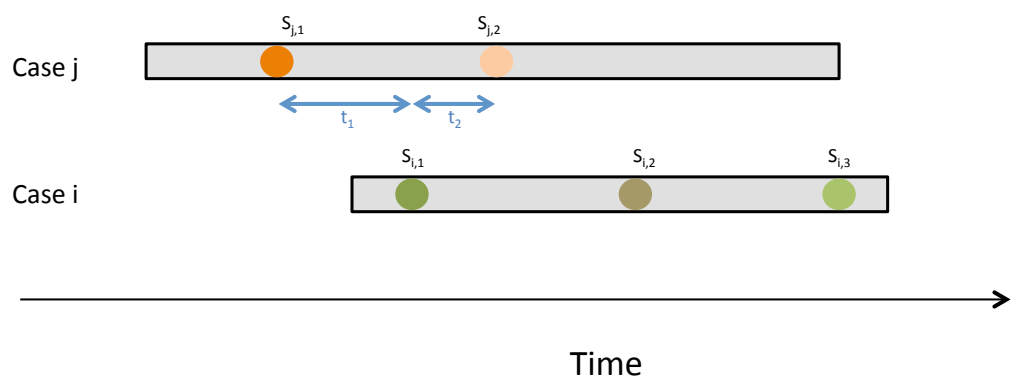
- Cluster assignment for cases that are part of a household outbreak with at least one sequence will depend on the identity of the closest temporal known cluster in the household outbreak. If there is more than one cluster option, this case is left unassigned. The assignment is done sequentially beginning with the case with the earliest onset. Given that this is a deterministic process, these assignments are maintained throughout the fitting process.
- Cluster assignment for cases that are part of a household outbreak with no sequence information. For such cases, the cluster assigned to the entire outbreak is inferred along with model parameters. The option of possible cluster assignment is chosen from the pool of possible clusters. Cases left unassigned by the previous step also have their cluster identity inferred.

A spatial-temporal clustering algorithm was attempted using a distance metric similar to ⁴³. To validate the algorithm, we blinded the algorithm from known clusters, and had it impute them. Over 50% of the time, the algorithm failed.

We also investigated if within cluster variation is informative of transmission events. There is evidence that the clustering pattern mimics closely the household structure⁴¹ and in such a case, the model might not learn much from the genetic clustering alone. We therefore derived genetic distances between cases in the same cluster, which were used to weight the transmission link between said cases.

4.5.2. Deriving genetic distances between cases

Consider a **case i** who had an onset after **case j**, both of whom have sequences. The genetic distance between case **i** and **j** is obtained by comparing the first sequence available from case **i** and any sequence from **j** whose sampling time is closest to the first sequence from **i**. In the illustration below, this would mean comparing sequence $S_{i,1}$ to $S_{j,2}$ to obtain genetic distance $d_{gen}(i,j)$. The phylogenetic analysis of *Agoti et al* ⁴¹ found that long shedding episodes do not have drastically differing genetic sequences (<6 SNPs) as such it should not make a significant difference whether we compare sequences forward ($S_{i,1}$ to $S_{j,2}$) or backward ($S_{i,1}$ to $S_{j,1}$) in time.



If either one or both of the cases do not have sequences, then the genetic distance is obtained from randomly sampling from the set of all pair-wise genetic distances from the specific genetic cluster. For cases with sequence data, $d_{gen}(i,j)$ is fixed, but for cases where one or both is missing sequences, $d_{gen}(i,j)$ changes every time the likelihood is calculated to reflect uncertainty. In this way only pairs of cases with sequence data

contribute definitive genetic information to the parameter inference algorithm while the rest will not. We use nucleotide differences as the distance $d_{gen}(i,j)$.

Once we have $d_{gen}(i,j)$, we then use this to obtain a genetic weight for the probability of a transmission event given by $P_{j \rightarrow i} = \exp^{-d_{gen}(i,j)*\vartheta}$ where ϑ is the rate of exponential decay and is estimated along with other model parameters. This function form results in a negative exponential relationship between the genetic weight and the genetic distance between a pair of cases. As part of uncertainty analysis, we will explore an additional function form where $P_{j \rightarrow i}$ is a step-function such that $P_{j \rightarrow i} = 1$ if $d_{gen}(i,j) \leq \vartheta$, 0 otherwise, where ϑ now becomes a nucleotide distance cut-off for within cluster transmission that is estimated along with other model parameters.

4.5.3. The transmission model

We extend the model in Chapter 3 to track cluster-specific infection onset at the individual host level. We create 3 levels of hierarchy in pathogen identification within the model. At the top level of the structure is identification by pathogen type, at the second level is identification by RSV groups and at the third level is identification by genetic clusters within groups. This hierarchy is in place to allow estimation of some parameters at the pathogen level and others at the group level while identifying the infecting pathogen at the cluster level. The model in Chapter 3 had 2 levels of hierarchy.

Similar to the previous model, everyone is assumed to be susceptible to RSV infection at the start of the outbreak, but the risk of infection is dependent on age. Once individuals have been exposed to infection, they enter a latency period that ranges between 2 to 5 days after which they become infectious. After the infectious period, individuals become susceptible to infection again, but the risk to subsequent infection is modified, i.e. RSV confers partial transient immunity that lasts as long as the outbreak is ongoing. This partial immunity is assumed to be different for heterologous group re-infection and homologous group re-infection. Individuals can get heterologous group co-infections, however, different from the model in the previous chapter, we explore if susceptibility to infection by RSV A is modified if an individual is currently shedding RSV B, and vice-versa.

Let's denote the rate at which individuals get exposed to infection (rate of exposure) as λ . In our model, an individual can get infected by someone they share a household with or from a source outside the household, resulting in a two-component rate of exposure. In a simplified form, we have:

$$\lambda(t) = [\text{baseline household rate of exposure} * \text{number of infectious household contacts}(t)] \\ + \\ [\text{baseline community rate of exposure} * \text{number of infectious community contacts}(t)]$$

$$\lambda(t) = \left(\eta * \sum_{\text{household}} I(t) \right) + \left(\varepsilon * \sum_{\text{community}} I(t) \right)$$

In the previous model, we represented infectious community contacts using a bell-shaped curve that mimicked ongoing transmission dynamics. In this chapter, two changes are made to the community rate of exposure. First, we explore the possibility of transmission between sampled households by introducing a term in the rate of exposure representing risk from sampled neighbours. The risk from sampled neighbours is weighted by a spatial distance kernel which modifies the risk based on the spatial distance between individuals. Second, because we are now modelling the rate of exposure to a specific RSV cluster, deriving a bell-shaped curve to mimic outbreak dynamics for specific clusters is no longer appropriate due to the low number of some cluster specific cases that resulted in unexpected curve shapes. Details of the new formulation of the background cluster-specific community function can be found in the subsequent section. The rate of exposure now takes the form:

$$\lambda(t) = \{\text{baseline household rate of exposure} * \text{number of infectious household contacts}(t)\} \\ + \\ \{\text{baseline community rate of exposure} * [\text{number of infectious neighbour contacts}(t)] + \\ \text{background community function}(t)\}$$

$$\lambda(t) = \left(\eta * \sum_{\text{household contact}} I(t) \right) + \varepsilon * \left(\left(\sum_{\text{sampled neighbour}} I(t) \right) + f(t) \right)$$

In detail, we present the model by specifying the rate of exposure to a particular RSV cluster c acting on a susceptible person i from household h at time t , denoted $\lambda_{i,h,c}(t)$ as:

$$\lambda_{i,h,c}(t) = S_{i,g}(t) \left[M_{i,h}(t) \sum_{j \neq i} HH_Rate_{h,c,j \rightarrow i}(t) + Comm_Rate_{i,c}(t) \right] \dots \text{(Eq 4.1)}$$

Where:

$S_{i,g}(t)$ is the factor modifying exposure by recent group specific infection history, age and group specific shedding status at time t given by:

$$S_{i,g}(t) = \exp \left(\phi_{Y,hist}(Infection_History_i(t)) + \phi_{X,age}(Age_group_{S,i}) + \phi_{W,curr}(Shedding_status_i(t)) \right)$$

$HH_Rate_{h,c,j \rightarrow i}(t)$ is the cluster specific within household exposure rate from infectious individual j present in the household at time t , and is given by:

$$\begin{aligned} HH_Rate_{h,c,j \rightarrow i}(t) &= \eta_g \times \psi_H(Household_size_i) \times \psi_{I,inf}(Infectivity_{j,h,c}(t)) \\ &\times M_{j,h}(t) \end{aligned}$$

$Comm_Rate_{i,c}(t)$ is the cluster specific community (external to the household) exposure rate given by:

$$\begin{aligned}
& Comm_Risk_{i,c}(t) \\
& = \varepsilon_g \\
& \times \psi_{E,age}(Age_group_{E,i}) \left(\left(M_{i,h}(t) \sum_{\substack{j \neq i, j \text{ not in} \\ i's \text{ house}}} Sampled_Neighbour_Risk_{h,c,j \rightarrow i}(t) \right) \right. \\
& \left. + f_c(t) \right)
\end{aligned}$$

Where:

$Sampled_Neighbour_Risk_{h,c,j \rightarrow i}(t)$ is the cluster specific exposure rate from sampled infectious individual j present in a neighbouring household at time t , and is given by:

$$\begin{aligned}
& Sampled_Neighbour_Risk_{h,c,j \rightarrow i}(t) \\
& = \psi_{I,inf}(Infectivity_{j,h,c}(t)) \times K(d_{i,j}, \kappa) \times M_{j,h}(t)
\end{aligned}$$

The parameter κ is the rate of exponential decay for the spatial distance kernel given by $K(d_{i,j}, \kappa) = e^{-\kappa * d_{i,j}}$.

The background community function

We define a background cluster-specific rate of exposure, $f_c(t)$, which affects susceptible individuals outside their household. This background function allows for introduction of new transmission clusters. The function form for a cluster c at time t is given as

$$f_c(t) = \delta + \sum_{\substack{i \text{ shedding} \\ RSV \text{ cluster } c}} e^{(t-\tau_{i,c})\beta}$$

Where δ is the basic risk prior to any observed onsets and β is the rate of exponential decay related to the time since onset of a case shedding cluster type c , β is a measure of the rate at which the cluster might disappear from the community and $\tau_{i,c}$ is the onset time of RSV cluster type c by person i . The parameters δ and β are not cluster or group specific. The sum of the cluster specific curves has to add up to the group specific curve, otherwise using clusters could lead to an over or under representation of the background community exposure rate. To ensure that $\sum f_c(t) = f_g(t)$ we need

to normalize the cluster level curves such that their sum adds up to the group level curve. A description of how this was done can be found in the appendix section A3: Supplementary appendix for Paper 2.

Table 4. 1 lists all the parameters in the model and gives a brief description. Despite identifying the infection pathogen at the cluster level, we do not have any cluster-specific parameters in the model.

Table 4. 1: Model parameters and their descriptions

Parameter (symbol)	Parameter (name)	Description
ϕ_Y	<i>Prev.hom</i> , <i>Prev.het</i>	Coefficients modifying susceptibility to infection by a particular RSV group depending on infection history. <i>Prev.hom</i> estimates the effect of a previous homologous group infection, while <i>Prev.het</i> estimates the effect of a previous heterologous group infection
ϕ_X	<i>Sus.age.2</i> , <i>Sus.age.3</i> , <i>Sus.age.4</i>	Coefficients modifying susceptibility to RSV depending on age. <i>Sus.age.2</i> estimates the effect being in age group 1-4 years, <i>Sus.age.3</i> the effect of group 5-15 and <i>Sus.age.4</i> of group ≥ 15 relative to group < 1 year.
ϕ_W	<i>Curr.het</i>	Coefficient modifying susceptibility to a particular RSV group based on shedding status of the heterologous group type
η_g	<i>HH.rsv.a</i> , <i>HH.rsv.b</i>	Baseline rate of within household exposure by RSV group, per person per day.
ψ_H	<i>HH.size</i>	Coefficient modifying the amount of within household exposure by household size. <i>HH.size</i> estimates the effect of being in a large household (> 8 inhabitants) relative to a small one
ϑ	<i>Gen.rate</i>	For $P_{j \rightarrow i} = \exp^{-d_{gen}(i,j)*\vartheta}$ the genetic distance kernel giving the genetic weight on probability of transmission, <i>Gen.rate</i> is the rate of exponential decay.

ψ_I	<i>Low.Sym</i> <i>High.Sym</i>	Coefficients modifying infectiousness by viral load and symptom status. Relative to being asymptomatic, <i>Low.Sym</i> estimates the effect of shedding low viral load and being symptomatic and <i>High.Sym</i> the effect of shedding high viral load and being symptomatic
ϵ_g	<i>Comm.rsv.a</i> <i>Comm.rsv.b</i>	Baseline rate of community exposure by RSV group, per person per day.
ψ_E	<i>Exp.age.2</i> <i>Exp.age.3</i>	Coefficients modifying the rate of community exposure by age group. <i>Exp.age.2</i> estimates the effect being in age group 1-4 years and <i>Exp.age.3</i> the effect of group ≥ 5 , relative to the <1-year age group
κ	<i>Dist.rate</i>	The rate of exponential decay for the spatial distance kernel given by $K(d_{i,j}, \kappa) = e^{-\kappa*d_{i,j}}$
δ, β	<i>Delta</i> , <i>Beta</i>	For the cluster specific background community function given by <div style="text-align: center;"> $f_c(t) = \delta + \sum_{\substack{i \text{ shedding} \\ \text{RSV cluster } c}} e^{(t-\tau_{i,c})\beta}$ </div> <i>Delta</i> (δ) is the basic risk and <i>Beta</i> (β) is the rate of exponential decay related to the time since onset of a case shedding cluster type <i>c</i> .

Following from the rate of exposure is the probability of exposure to cluster *c* given an exposure event has occurred, expressed as:

Probability of exposure = *prob(any exposure event)* * *prob(exposure to cluster c)*

$$\alpha_{i,h,c}(t) = (1 - \exp^{-\sum_{C'} \lambda_{i,h,c}(t)}) * \left(\frac{\lambda_{i,h,c}(t)}{\sum_{C'} \lambda_{i,h,c}(t)} \right) \quad \dots \quad (Eq 4.2)$$

Where C' is the set of all clusters in a given RSV group.

This formulation factors in the fact that on any given day, an individual can only be shedding virus from a single cluster, in the respective group, this can be seen in the shedding patterns shown in Figure 4. 5. The clusters are therefore competing for

susceptible hosts. Exposure events are mutually exclusive and distributed according to a multinomial distribution. We thus have

$$\left[\text{prob}(\text{no exposure}) + \sum_{\text{All clusters}} \text{prob}(\text{exposure to cluster } c) \right] = 1$$

Assuming that the duration of latency can range from 0 to 5 days with probabilities [0, 0, 0.33, 0.33, 0.25, 0.083]⁴⁴, we then have the following probability of onset at time t given no onsets or shedding until t :

$$p_{i,h,c}(t) = \sum_{l=0}^L \theta_l \alpha_{i,h,c}(t-l)$$

Where L is the maximum latency period and θ_l is the probability that the latency period is exactly l days. In this way, the genetic clusters are used together with the spatial/social clusters (households) and the latency distribution (which implicitly works based on temporal clusters) to make joint inference on transmission parameters.

The likelihood

Since the model is focused on the determinants of infection onset process, the data whose likelihood we are interested in is the onset data. Given the model described, the likelihood of an individual's observed cluster c data is the probability of all the onsets, and days of no onsets where the individual was at risk of infection, i.e. not shedding RSV cluster c . For a particular cluster, this follows a Bernoulli distribution with probability $p_{i,h,c}(u)$.

For i with no onset of type c :

$$L_{i,c} = \prod_{t=1}^T [1 - p_{i,h,c}(t)]$$

Where T is the end of the observation period.

For i with an onset of type c , the likelihood is give as:

$$L_{i,c} = \left[\left(\prod_{u \in \text{Onsets}_{i,h,c}} p_{i,h,c}(u) \right) * \left(\prod_{a \in \text{At}_{risk}_{i,h,c}} (1 - p_{i,h,c}(a)) \right) \right]$$

In this instance, to factor in the genetic data we modify the rate of exposure given in (Eq 4.1) such that:

$$\begin{aligned} HH_Risk_{h,c,j \rightarrow i}(t) &= \eta_g \times \psi_H(\text{Household_size}_i) \times P_{j \rightarrow i} \times \psi_{I,inf}(\text{Infectivity}_{j,h,c}(t)) \\ &\times M_{j,h}(t) \end{aligned}$$

$$\text{Sampled_Neighbour_Risk}_{h,c,j \rightarrow i}(t) = P_{j \rightarrow i} \times \psi_{I,c,j}(t) \times K(d_{i,j}, \kappa) \times M_{j,h}(t)$$

With this formulation, the genetic components of the model are dependent on the epidemiological in that they are not expressed independently in the likelihood function as is the case with modular approaches such as the kind implemented in the Outbreaker package^{5,6}. We introduce $P_{j \rightarrow i}$ into the rate of exposure equation as opposed to directly into the likelihood because for a given case, we are not making direct inference on the source of infection or the exact date of exposure: we consider all likely dates and sources given the latency distribution.

The total likelihood is thus given by the product of $L_{i,c}$ over all the genetic clusters and individuals in the data

$$L = \prod_i \left[\prod_c \left[\left(\prod_{u \in \text{Onsets}_{i,h,c}} p_{i,h,c}(u) \right) * \left(\prod_{a \in \text{At}_{risk}_{i,h,c}} (1 - p_{i,h,c}(a)) \right) \right] \right]$$

4.5.4. Inference of model parameters and augmented data

We used Bayesian inference to obtain estimates of the model parameters $\varphi = \{\text{Prev.hom}, \text{Prev.het}, \text{Sus.age.2}, \text{Sus.age.3}, \text{Sus.age.4}, \text{Curr.het}, \text{HH.rsv.a}, \text{HH.rsv.b}, \text{HH.size}, \text{Gen.rate}, \text{Low.Sym}, \text{High.Sym}, \text{Comm.rsv.a}, \text{Comm.rsv.b}, \text{Exp.age.2}, \text{Exp.age.3}, \text{Dist.rate}, \text{Delta}, \text{Beta}\}$ and the augmented data \mathbf{D}_A given the observed data \mathbf{D} . In brief, Bayesian inference results in an updated distribution of the parameter of interest (posterior distribution) given prior assumptions/knowledge of the parameter (prior distribution) and an expression giving the probability of a parameter value given data (likelihood) i.e. $P(\varphi | \mathbf{D}, \mathbf{D}_A) \propto P(\varphi) \times L(\varphi | \mathbf{D}, \mathbf{D}_A)$.

The adaptive MH-MCMC algorithm is a popular first step for a situation where the target distribution is not simple, and the dimension of the parameters is not small. As we have a total of 19 parameters this seemed a natural starting point. We assume that all the cases were observed but that for some of the cases, there is no information on the cluster id of the shedding episode, as such, the augmented data is the set of all shedding episodes whose cluster id was left unassigned by the imputation process previously described. These include cases that are part of household outbreaks with no genetic information and cases that are part of household outbreaks with more than one possible genetic cluster id. For cases that are part of an outbreak with no genetic information, a single cluster id is inferred for all the cases in the household outbreak. For a brief explanation of our implementation of MH-MCMC, see appendix section A3.

We initiated 3 chains and set the algorithm to start adapting the proposal distribution based on accepted parameters after 10000, 15000 and 10000 iterations respectively. Burn-in was assessed visually after which the results of the three concurrent chains were combined to infer the posterior distribution. The three chains were run for 250,000 iterations each. The parameters were estimated on the log scale. All the computation was done using the *julia* language⁴⁵ (version 1.1)⁴⁶. The code is freely available under the GNU Lesser General Public License v3.0 and can be found at https://github.com/lkadzo/HH_Transmission_Model.

4.5.5. Highest probability transmission source

Following the estimation of the posterior parameter distribution, we randomly selected a subset to determine infection sources for every case. For every case observed in the data we identified the transmission source that had the highest likelihood given the data and a parameter set φ^* sampled from the joint parameter posterior distribution (highest probability transmission source: HPTS). Consider a case i with onset date T_i^O . Given our assumption of a maximum latency duration of 5 days, we define a time window where potential infection could have occurred. For each day in the time window, potential sources of infection are $\{\Omega_i^1, \Omega_i^2 \dots \Omega_i^n\}$. An infection source is assigned if it gives the highest value of i 's likelihood defined as “ the likelihood of i 's onset date, infection date and infection source given sample

parameter set φ^* . Further details of the likelihood function used to identify the HPTS can be found in appendix section A3. A hundred parameter sets were sampled and the HPTS for each case established for each sample. From the distribution of 100 HPTS, the one with the highest frequency was selected as the source of transmission.

4.6. Results

4.6.1. The data

Prior to model fitting, we look at the patterns of the sequence data. The table below quantifies the missing data problem by giving the number of sequences available by RSV group, shedding episode, person and household.

Table 4. 2: A summary of the distribution of sequences

	RSV A	RSV B
No. Samples	250	306
No. Samples with sequences	103 (41.2%)	88 (28.8%)
No. Episodes	97	125
No. Episodes with sequences	54 (55.6%)	54 (43.2%)
No. People infected	88	113
No. People infected with sequences	50 (56.8%)	53 (46.9%)
No. Households	25	34
No. Households with sequences	9 (36%)	15 (44.1%)

Given the genetic clusters imported from the clades and sub-clades of *Agoti et al*⁴¹, Figure 4. 3 and Figure 4. 4 show the distribution of pair-wise nucleotide difference between sequences in the same genetic cluster for RSV A and RSV B respectively. Figure 4. 5 shows the distribution of sequences across the temporal shedding patterns and the results of the cluster duration imputation previously described.

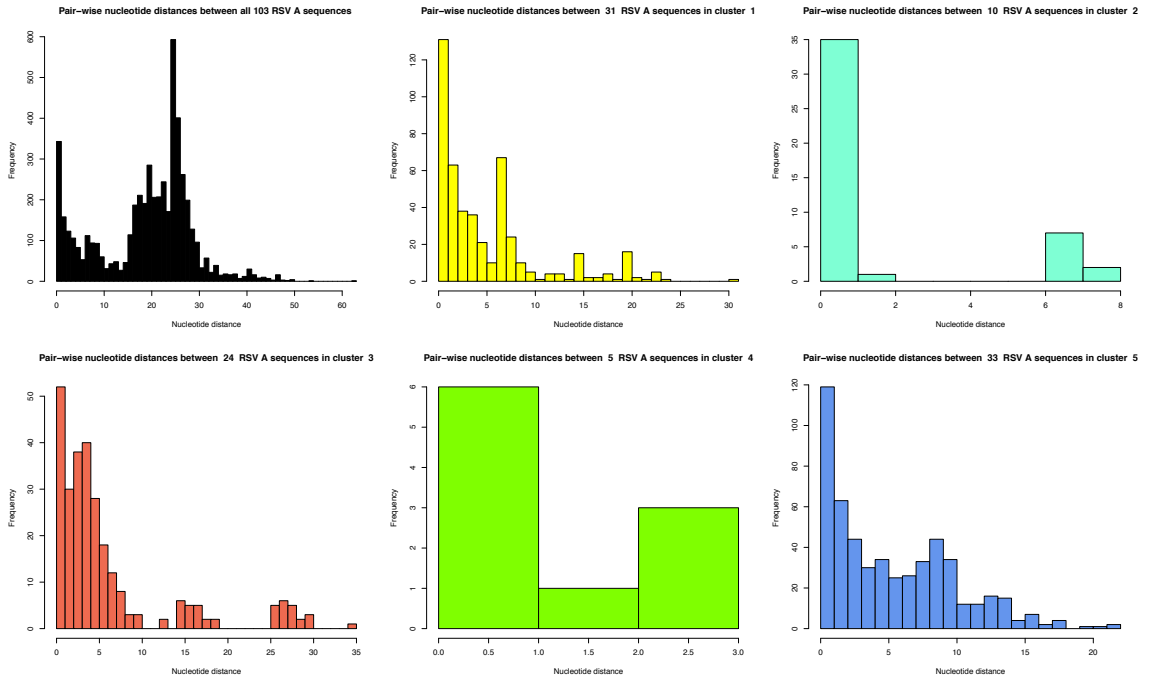


Figure 4. 3: Distribution of pair-wise nucleotide distances between RSV A sequences.

Top row, the first distribution shows all the pair-wise distances, the subsequent figures show the distances by cluster.

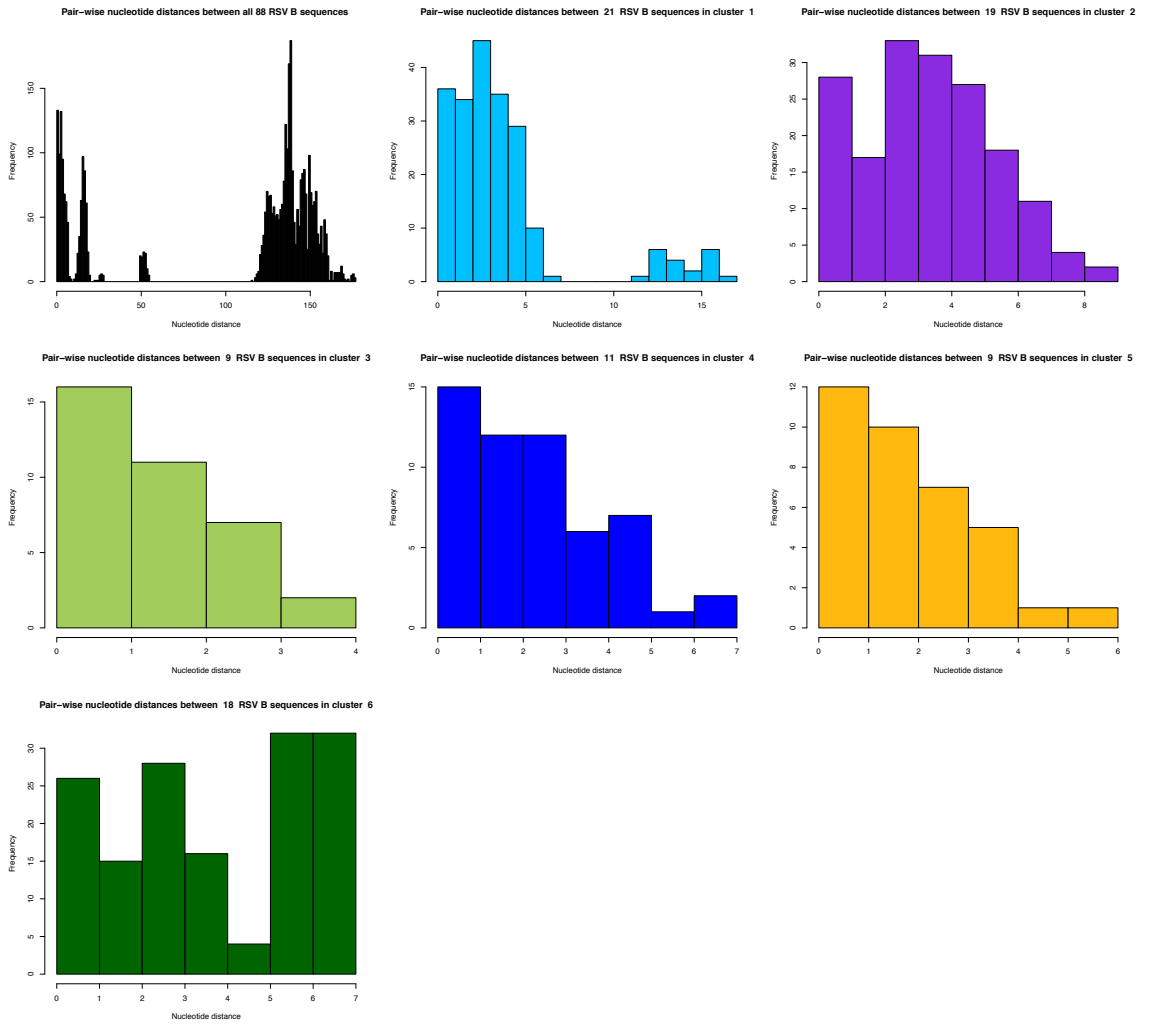


Figure 4. 4: Distribution of pair-wise nucleotide distances between RSV B sequences. Top row, the first distribution shows all the pair-wise distances, the subsequent figures show the distances by cluster.

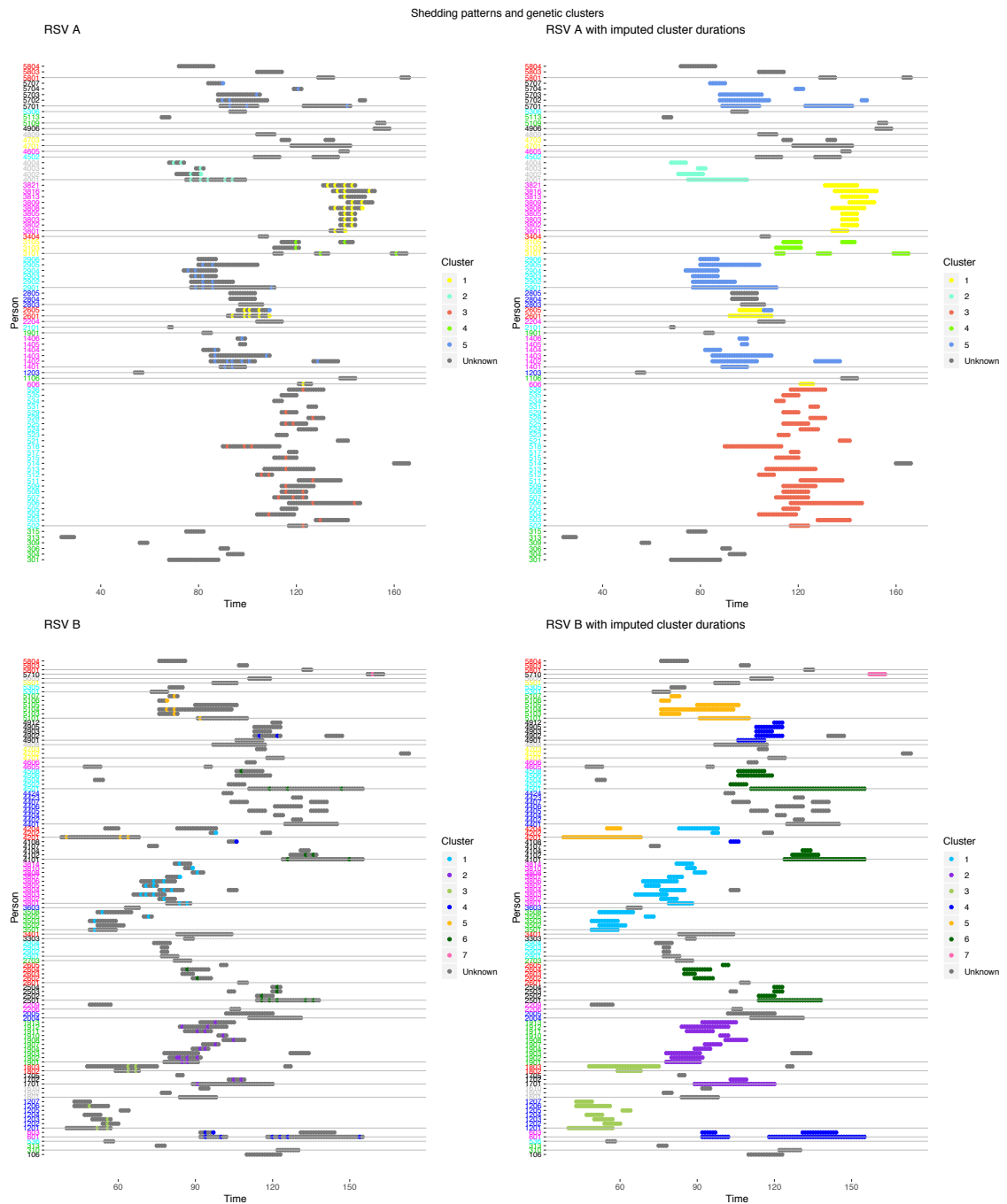


Figure 4. 5: Distribution of available sequences across shedding episodes (left) and the results of imputation of cluster durations (right).

RSV A data is shown in the top row shows and RSV B is shown at the bottom.

Imputation of cluster shedding durations was done for episodes that had at least one sequence, and for episodes with no sequences but that were part of a household outbreak with at least one sequence.

4.6.2. Inference on model parameters

The median and 95% credible intervals of the 19 parameters inferred using the model and data with three levels of hierarchy in pathogen identification are shown in Table 4.

3. The trace plots showing the results of the MCMC algorithm are given in A3:

Supplementary appendix for Paper 2. Convergence was assessed visually and confirmed using the Gelman-Rubin-Brooks (GRB) statistic⁴⁷.

Table 4. 3: Median and 95% credible intervals for parameters estimated using the model with sequence data.

Symbol	Description	Name	Median (95% Credible interval)
ϕ_Y	Coefficients modifying susceptibility to infection by a particular RSV group depending on infection history. <i>Prev.hom</i> estimates the effect of a previous homologous group infection, and	<i>Prev.hom</i>	0.4328 (0.2665, 0.6727)
	<i>Prev.het</i> the effect of a previous heterologous infection	<i>Prev.het</i>	0.5126 (0.2601, 0.8985)
ϕ_W	Coefficient modifying susceptibility to a particular RSV group based on	<i>Curr.het</i>	0.9520 (0.2494, 2.262)

	shedding status of the heterologous group type		
ϕ_X	Coefficients modifying susceptibility to RSV by age.	<i>Sus.age.2</i>	0.8804 (0.4997, 1.616)
		<i>Sus.age.3</i>	0.2741 (0.1591, 0.4946)
	<i>Sus.age.2</i> estimates	<i>Sus.age.4</i>	0.1562 (0.08867, 0.2852)
	modification to group 1-4 years, <i>Sus.age.3</i> 5-15 years and <i>Sus.age.4</i> ≥ 15 years relative to group <1 year.		
η_g	Baseline rate of within household exposure by RSV group, per person per day.	<i>HH.rsv.a</i>	0.02360 (0.0119, 0.04361)
		<i>HH.rsv.b</i>	0.02272 (0.01120, 0.04196)
ψ_H	Coefficient modifying the amount of within household exposure by household size for households of 8 or more relative to <8.	<i>HH.size</i>	0.4457 (0.2892, 0.6843)
ψ_I	Coefficients modifying	<i>Low.Sym</i>	2.1 (1.214, 3.67)
		<i>High.Sym</i>	4.437 (1.8, 8.959)

infectiousness by viral load and symptom status. Relative to being asymptomatic, *Low.Sym* estimates the effect of shedding low viral load and being symptomatic and *High.Sym* the effect of shedding high viral load and being symptomatic

κ	The rate of exponential decay on the spatial distance kernel	<i>Dist.rate</i>	207.7 (7.819, 169100)
ϑ	The rate of exponential decay on the genetic weight function.	<i>Gen.rate*</i>	0.0002631 (0.000001027, 0.003817)
ε_g	Baseline rate of community exposure by RSV group, per person per day.	<i>Comm.rsv.a</i> <i>Comm.rsv.b</i>	0.0003091 (0.0001198, 0.0008682) 0.0003849 (0.0001525, 0.001072)
ψ_E	Coefficients modifying the rate of community exposure by age	<i>Exp.age.2</i> <i>Exp.age.3</i>	0.5311 (0.2179, 1.221) 1.64 (0.7705, 3.386)

	group. <i>Exp.age.2</i> for 1-4 years and <i>Exp.age.3</i> for ≥ 5 years, relative < 1 year		
δ, β	Parameters for the cluster specific background community function.	<i>Delta</i> <i>Beta</i>	1.58 (0.5466, 4.693) 0.1929 (0.08315, 0.7321)

* Here **Gen.rate** is the rate of exponential decay.

Previous infection reduces the risk of re-infection in the same outbreak by ~50% based on the estimates of *Prev.hom* and *Prev.het* parameters which measure the relative reduction in susceptibility to infection by a particular RSV group given previous homologous or heterologous group infection, respectively. Estimates of *Sus.age.2*, *Sus.age.3* and *Sus.age.4* imply an inverse relationship between age and susceptibility to infection. Households of 8 or more individuals have ~55% reduction in pair-wise rate of exposure within the household relative to smaller households, *HH.size* = 0.4457 (0.2892, 0.6843). Symptomatic cases are 2-4 times more infectious than asymptomatic cases, *Low.Sym* = 2.1 (1.214, 3.67) and *High.Sym* = 4.437 (1.8, 8.959). The high estimate for the rate of exponential decrease in probability of transmission with increasing distance between households, *Dist.rate* = 207.7 (7.819, 169100), means that transmission between household for this study population is unlikely to have occurred, although the large credibility interval suggests that there is limited information for this parameter. Estimates of the rate of exponential decrease in transmission probability with increasing genetic distance (*Gen.rate*) parameter imply that within cluster transmission was nearly 100% likely regardless of the pair-wise nucleotide distances. The uncertainty analysis where a genetic cut-off for transmission was estimated rather than a rate of exponential decay resulted in the same outcome (results not shown). Age is unlikely to affect the rate of exposure to infection from

sources outside of the household, *Exp.age.2* and *Exp.age.3*, were estimated with credible intervals including 1.

To validate the model, we simulated multiple epidemics and checked to see if the observed epidemic was captured by the range of simulated dynamics. Details of the simulation algorithm can be found in A3: Supplementary appendix for Paper 2. We sampled 12 sets of parameters from the posterior distribution, and for each set, simulated 100 epidemics. The results of these simulations are shown in Figure 4. 6 for RSV A and Figure 4. 7 for RSV B. In addition to comparing the time course of cases, we also looked at the total number of cases in an epidemic, the proportion of individuals with multiple onsets and the number of cases in the first and last week of the time period. These values from the data were compared to the range of simulated values to check that key aspects of the epidemic were being reproduced by the simulations. These results are shown in A3: Supplementary appendix for Paper 2. From these results, we concluded that the model sufficiently captured key aspects of the epidemic.

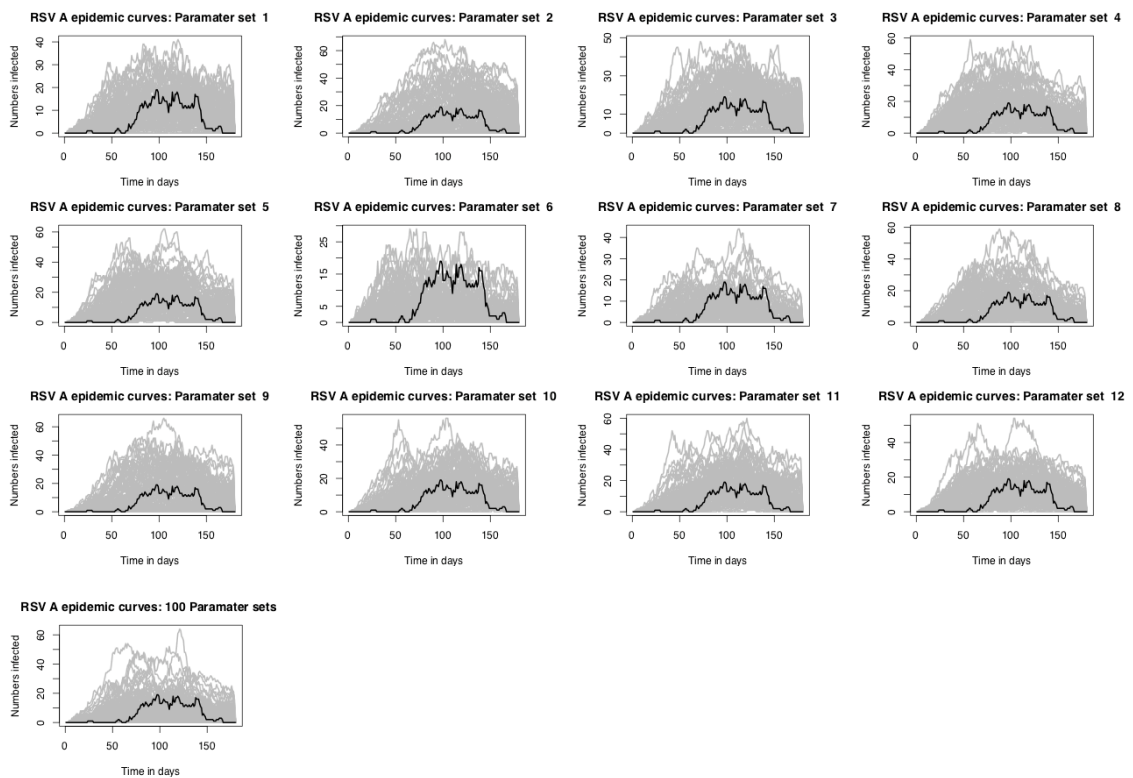


Figure 4. 6: A comparison of simulated and observed data for RSV A.

Each panel shows the results of 100 simulations from a single parameter set. The grey lines show the simulated data while the black lines show the observed data. Time is shown on the x-axis while the y-axis shows the total number of people who are shedding at a given point in time.

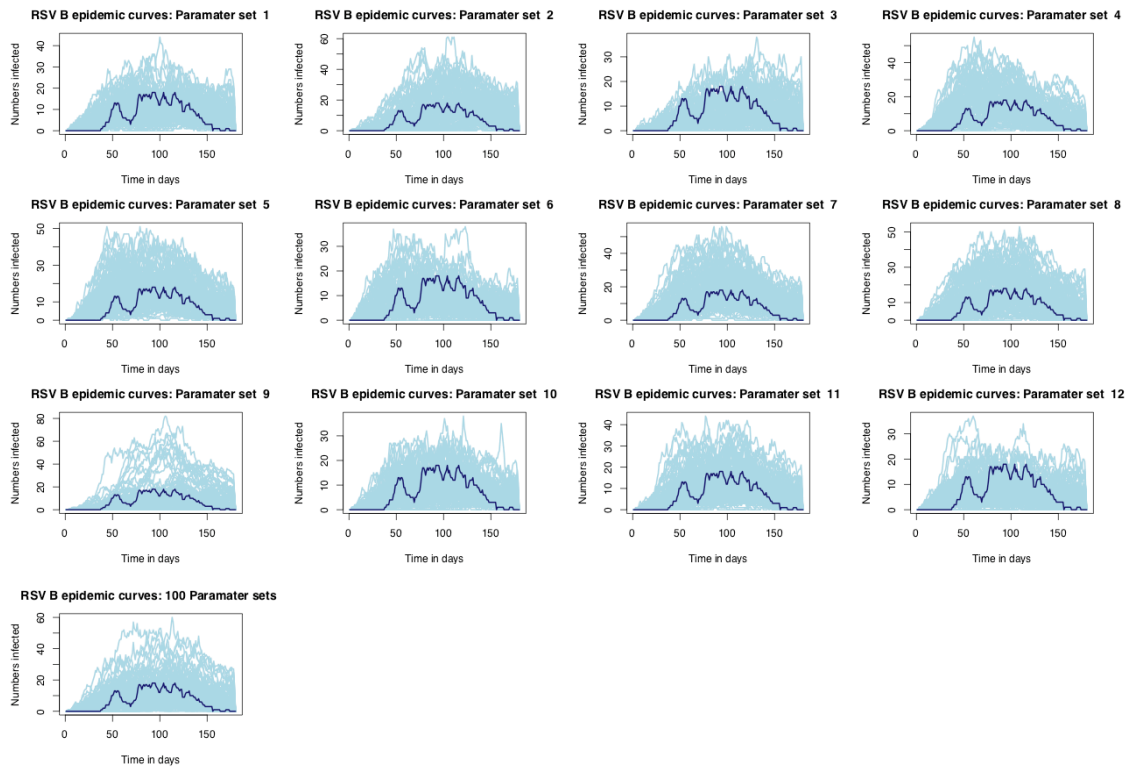


Figure 4. 7: A comparison of simulated and observed data for RSV B.

Each panel shows the results of 100 simulations from a single parameter set. The light blue lines show the simulated epidemic data while the dark blue lines show the observed data. Time is shown on the x-axis while the y-axis shows the total number of people who are shedding at a given point in time.

To assess the impact of increased resolution in pathogen identification on estimated parameters we compared the distributions of parameters estimated using RSV cases identified at the pathogen level, group level and cluster level. Figure 4. 8 shows the density plots comparing these distributions, details of the model modifications to allow fitting of group level data are given in appendix section A3. This figure shows 17 of the 19 parameters in the model with genetic clusters, parameters *Dist.rate* and *Gen.rate* are not included. For parameters that are present in the model with group and cluster level identification but not in the model with pathogen level identification,

e.g. *Prev.het*, we used the corresponding parameter assumption, i.e. *Prev.het* = *Prev.hom*.

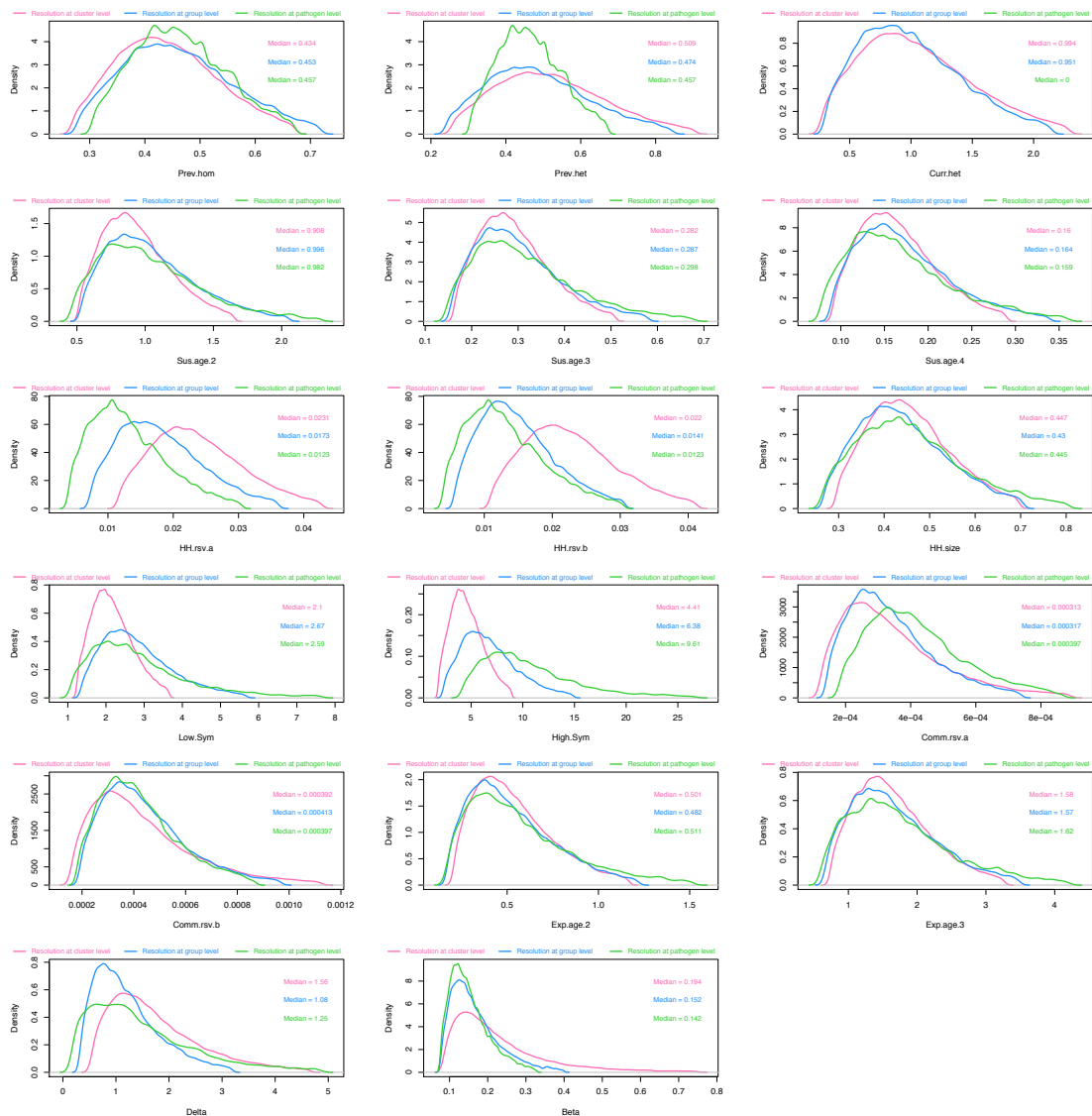


Figure 4. 8: A comparison of the parameter distributions obtained from the model using different resolutions in pathogen identification.

The green curves show the results using data at the pathogen level, the blue curves shows the group level and the pink curves show the cluster level. Each panel shows 1 one of 17 shared parameters.

The results show that for most of the parameters, the estimated distributions do not differ by the resolution in pathogen identification. The parameters measuring the effect of viral load and symptoms on infectiousness (*Low.Sym* and *High.Sym*) are estimated with increased precision when pathogen resolution is increased. The

distribution of the within household transmission coefficients shift slightly towards higher values with increased resolution both for RSV A and B (*HH.rsv.a* and *HH.rsv.b*) while the community transmission coefficient for RSV A (*Comm.rsv.a*) has a slight shift towards lower values.

4.6.3. Highest Probability transmission source

For each case in the data, we established the HPTS given a particular set of parameters and matching augmented data. For a particular case, the frequency of each HPTS across the sample was recorded and only the most frequent HPTS is show in the transmission networks in Figure 4. 9. Table 4. 4 gives additional characteristics of the transmission networks.

Table 4. 4: Characteristics of the transmission chains inferred.

	RSV A	RSV B
Number of cases	97	125
Number of introductions into households (index cases)	39	60
Number of introductions leading to onward transmission	13	23
Number of infant cases	20	22
Number of non-index infant cases	11	8
Number of household outbreaks initiated by an infant	3	9

Thirty-nine out of ninety-seven (40%) of the RSV A cases were from sources outside of the household, while for RSV B 60 (48%) cases were are result of non-household exposures; 33% (13/39) of RSV A introductions into the household led to infection of other household members, as did 38% (23/60) of RSV B introductions; 55% (11/20) infant, children <1 year old, RSV A infections were acquired within the household as were 36% (8/22) infant RSV B infections. Of the 11 infant RSV cases that were infected within the household, 8 were infected by children aged between 2 and 13 years, 1 was infected by another infant , 1 by a 16-year old and 1 by a 37-year old adult. Five out of

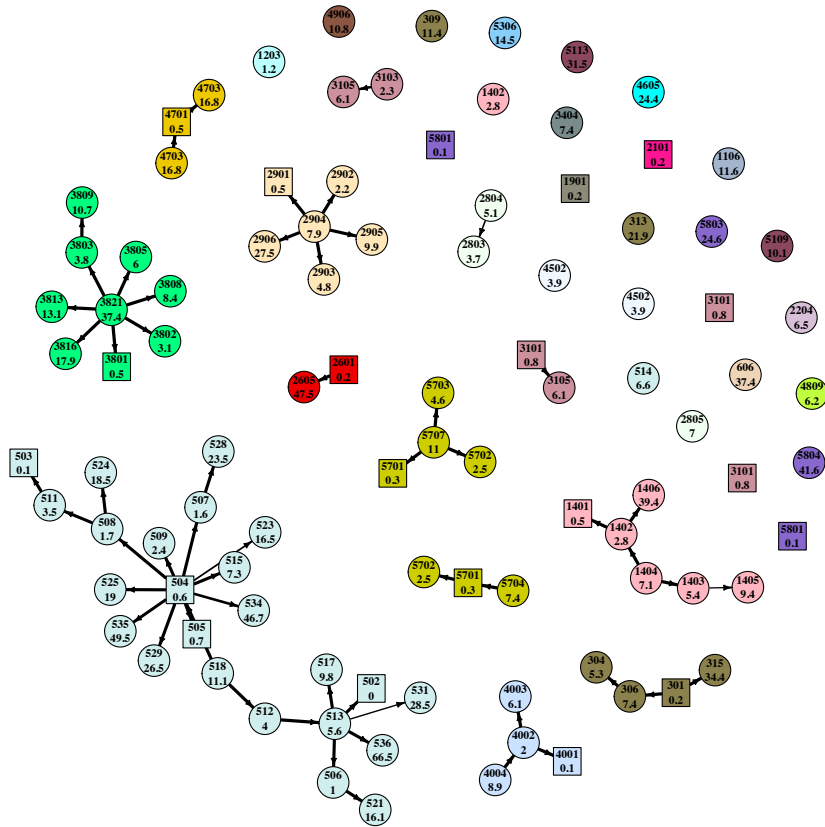
8 of the infant RSV B cases infected within the household were infected by children between 2 and 13 years, 2 were infected by a 16 and 18-year-old while one was most likely infected by a 49 year old. Table 4. 5 gives the age distribution of index cases that led to other infections in the household (HH outbreaks) compared to the age distribution of index cases that did not. Household outbreaks were, more often than not, initiated by children below 13 years old (31 out of 36 index cases).

Table 4. 5: Age distribution of index cases of household outbreaks.

Index cases are clustered into 3 age groups and according to whether they led to onward transmission in the household or not.

<i>Age Group</i>	No. index cases leading to onward transmission		No. index cases NOT leading to onward transmission	
	RSV A	RSV B	RSV A	RSV B
< 1	3	9	6	5
1 – 13	8	11	13	19
≥ 13	2	3	7	13
<i>Total</i>	13	23	26	37

RSV A



RSV B

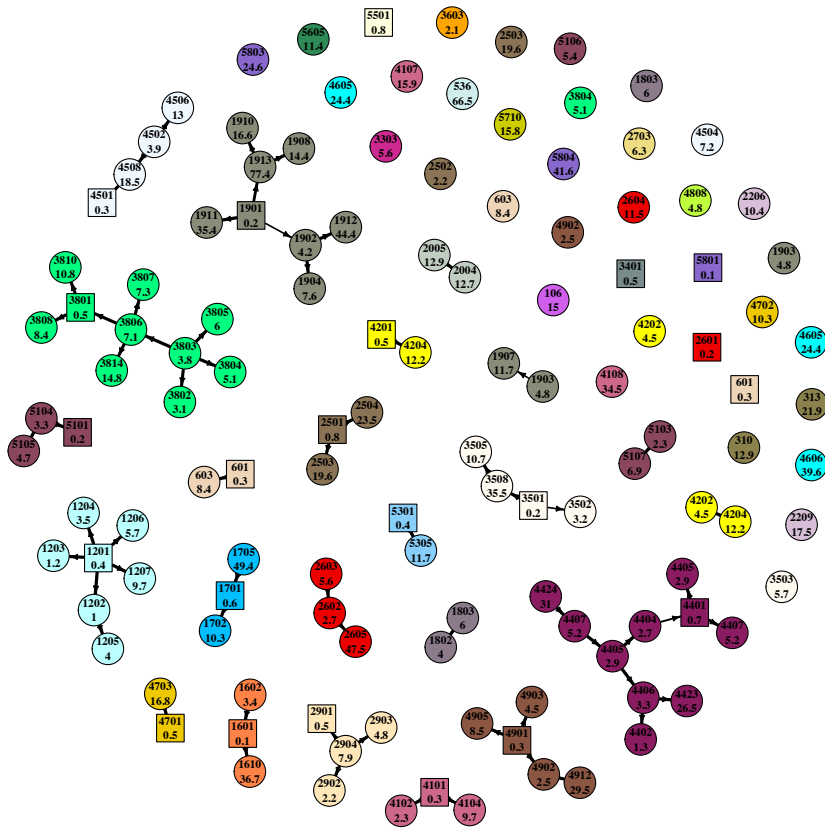


Figure 4. 9: Transmission networks showing the highest probability source of transmission given by our model results.

Each vertex is an RSV case labelled by individual study number (top) and age in years (bottom) and color-coded by household. Cases that are <1 year old are represented by square shaped vertices. The width of the connecting edge is proportional to the frequency at which the particular source was identified as the HPTS given different parameter set values.

4.7. Discussion

We carried out an analysis of data on the social-temporal and genetic pattern of spread of RSV in a group of households in rural Kenya followed up during a six-month study period covering the time frame of an entire RSV epidemic in the local area. Through systematically integrating all the available information of the infection episodes and host demographics, we were able to infer sources of infant infections. Fifty five percent of infant RSV A infections were acquired within the household, compared to 36% of infant RSV B infections. There were 8% more RSV B introductions into the household than RSV A, and a 5% difference in the proportion of introductions that led to onward within household transmission between the RSV groups. In this study population, there is evidence of differences in transmission dynamics between the two RSV groups, parts of which could be due to RSV B dominating in this particular outbreak. However, despite the seemingly slight transmission advantage of RSV B, a larger fraction of infant RSV A infections was acquired within the household. This points to the household not only being an important environment for RSV transmission in general, but possibly, more specifically, for infant RSV A transmission.

This work is an extension of a previous analysis on social-temporal data³² that now incorporates the output of a phylogenetic analysis⁴¹ with the aim of utilizing all the available data from an outbreak to define transmission dynamics. In doing so, we also assessed the difference in model inference when different data resolutions were used for pathogen identification; resolution at the pathogen level (RSV), resolution at the group level (RSV A and RSV B), and resolution at the genetic cluster level (5 RSV A clusters and 7 RSV B clusters). We found that increased resolution did not dramatically change the distribution of estimated parameters. With resolution at the group level we had previously inferred possible niche separations between RSV A and RSV B based on the distribution of the transmission coefficients. The evidence of this was not overwhelming to begin with, and the slight change in parameter distributions as a result of increased resolution resulted in this line of evidence being lost. However, in both the present and previous analysis, a larger fraction of RSV A cases were acquired in the household relative to RSV B. In the present analysis, 60% of RSV A and 52% of RSV B infections occurred in the household. In the previous analysis, 40-59% of RSV A

and 26-48% of RSV B infection occurred in the household. As previously established in the social-temporal data analysis, households with less than 8 occupants had a higher pair-wise risk of exposure compared to larger household, increasing age had a protective effect on transmission as did previous infection in the same outbreak. Symptomatic cases with high viral load were more infectious than asymptomatic cases, the effect of which was inferred more precisely with the inclusion of genetic data. We found that transmission between the households in this study was unlikely to have occurred, which is in line with the results of the phylogenetic analysis of Agoti *et al*⁴¹. Different resolutions of the data had different ways of suggesting a difference in transmission niche between RSV A and RSV B; the group data inferred overlapping but slightly different values for the transmission coefficients, the cluster resolution data inferred almost similar distributions for the transmission coefficients between RSV A and B, but RSV A was better transmitted to infants within the household. Respiratory syncytial virus is an important pathogen to the under 5 years olds, with <6 month olds experiencing the most severe disease burdens⁴⁸. It is a ubiquitous pathogen that circulates in seasons which are not only characterized by a change in the dominant group type, but also changes to the genotype composition^{23,29}. The slower mutation rates of RSV A⁴⁹ could account for its niche being in young infection-naïve infants. In accord with this, White *et al* found evidence that RSV A is slightly more transmissible than RSV B⁵⁰. Their study used a compartmental multi-strain model to fit data from the UK and Finland. From the household study that we analysed, we cannot state with certainty that there is a difference in transmission niche between the two groups, a study that incorporates information from different potential transmission hubs such as households, schools and workplaces would be better placed to do so.

Increasing the resolution in pathogen identification did not have a drastic impact on estimated parameters; this could be due to the study design. Nasopharyngeal swab (NPS) samples to test for the presence of infection were collected twice a week every week for 6 months from all the participants present in the households at the time of the sample collection visits. This resulted in densely sampled detailed data that left little room for uncertainty in when individuals got infected. In addition, information on the social structuring of the population in the form of households provides information on some of the most frequent contacts each participant had. This level of detail is

likely why the addition of genetic information, whose clustering mimicked the household structure, did not lead to much further resolution on who might have acquired infection from whom, hence no significant changes in most of the inferred parameter distribution. This result should not have been surprising; Campbell *et al* in integrating genetic, temporal and contact data found that the contact data could replace the genetic data in a model trying to infer the transmission chains². In their work, Kinyanjui *et al* highlighted the importance of mixing assumptions and social structure in models of RSV transmission⁵¹. This implies that good quality data on timing of cases and their most frequent contacts is key to be able to determine transmission characteristics of an infection. However, this could be limited to the type of infection under study and it should be borne in mind that contact data can be difficult to gather in the heat of an ongoing outbreak. In place of a detailed epidemiological study with dense sampling, integrating temporal and genetic data is the next best thing, particularly if the priority is transmission chain inference. A possible further analysis of these data would be to determine to what extent the genetic information can recapture the household clustering, i.e. to fit a model which does not include the household information.

Through combining epidemiological and phylogenetic inference, our method was able to determine transmission chains within households with greater certainty than a preceding phylogenetic analysis by Agoti *et al*⁵². In general, the networks inferred from the present analysis did not contradict any of the inference from the phylogenetic analysis. However, for one of the infected infants the inferred source of transmission differed. We assigned individual 3806 as the source of infant 3801's RSV B infection while Agoti assigned 3805. Both 3806 and 3805 were children of school going age and both had sequences that were 3 nucleotides apart from the closest temporal sequence from 3801. In addition to considering the social grouping, infection window and genetic cluster, our approach also considers the infectiousness of a potential source. In this case, 3806 had symptoms and a high viral load in the three days preceding shedding onset in 3801, while 3805 did not. Our model assigned 3806 as the infection source due to their higher infectiousness relative to 3805. Such an example highlights the strength in our technique in being able to incorporate all possible determinants of a transmission event. Despite the marked improvement in transmission chain

inference when combining an epidemiological and phylogenetic analysis, we did notice that our method has a propensity to infer super-spreaders, examples are infant 504 who was implicated in infecting 10 household cohabitants and 3821 who seeded 7 other infections. Though there might be some truth to these dynamics, based on the roles of the different members of the household, the model arrives at these networks based on the patterns in the available data. If all the criteria for a transmission event have been met, i.e. in the same genetic cluster, within a reasonable infection window, in the same household, a highly infectious potential source, then the model will create a link between cases. To tease apart true super-spread events from “convenience” networks, additional data on within household contacts would be needed to inform the model, data such as the kind collected by Kiti and colleagues⁵³.

This study is not without its limitations. Similar to previous work^{14,35,37}, we used a two-step approach in our application of phylodynamics. This has the potential to lead to inconsistencies that would otherwise not occur with simultaneous inference of the evolutionary and epidemiological dynamics. However, given that we only used aggregated results of the phylogenetic analysis, in the form of clusters, and raw nucleotide distances as opposed to phylogenetic tree distances, we do not heavily rely on the exact results of the independent phylogenetic analysis. Using genetic clusters provides the advantage of being able to identify obvious separate introductions, a characteristic that can be difficult to account for in the models of simultaneous inference. In addition, given that the genetic clusters were generated using a combination of criteria makes it less likely that the wrong clustering pattern was inferred. As with previous work¹⁴, our two-step approach is more computationally tractable than a simultaneous-inference version of it would have been. We were able to include data from individuals who did not have genetic sequences and use a non-trivial epidemiological model.

Despite the fact that sequence data did not make a significant change in understanding overall transmission patterns for RSV in this study, we believe that there is still a lot of potential for phylodynamics in RSV. At a larger geographical scale, say country level, sequences collected over several months coupled with a stochastic transmission model of RSV could be used to determine patterns of spread within the country; answering

questions such as where are new strains of RSV introduced, how quickly do they spread across the country and what are the drivers of the spread. Sequences collected over several years coupled with a transmission model and social-demographic data on the host population could help determine what drives the replacement of RSV genotypes and what allows the co-existence of RSV groups; potentially giving a more definitive answer as to whether immune pressure plays a role in changes to the RSV genome in the short term. RSV has already been shown to have geographically different transmission patterns and potential drivers of seasonality^{21,54}, if a phylodynamics analysis reveals that there are also geographically different drivers of genotype replacement, this could have significant implications to vaccine development and effectiveness. Finally, improved surveillance is needed in order to get better data on RSV outbreaks. There is currently a spatial bias in the RSV sequences available in GenBank as Giallonardo *et al* reported²⁵, however, there is an ongoing WHO effort to develop a global RSV surveillance strategy⁵⁵.

In conclusion, we were able to integrate the results of a phylogenetic analysis with epidemiological data to infer that nearly half of the RSV infections in this study were acquired within the household. A significant portion of infant RSV infections occur in the household, more so for RSV A than RSV B, and a majority of these are a result of transmission from children aged between 2 and 13 years old. Vaccination of this age group would therefore provide indirect protection to the infant.

4.8. References:

1. Grenfell, B. T. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science (80-.)*. **303**, 327–332 (2004).
2. Campbell, F., Cori, A., Ferguson, N. & Jombart, T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput. Biol.* **15**, e1006930 (2019).
3. Cori, A. *et al.* A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLoS Comput. Biol.* **14**, 1–22 (2018).
4. Ypma, R. J. F. *et al.* Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B Biol. Sci.* **279**, 444–450 (2012).
5. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput. Biol.* **10**, e1003457 (2014).
6. Campbell, F. *et al.* outbreaker2 : a modular platform for outbreak reconstruction. **19**, (2018).
7. De Maio, N., Wu, C. H. & Wilson, D. J. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Comput. Biol.* **12**, 1–23 (2016).
8. Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. & Wallinga, J. *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks.* *PLoS Computational Biology* **13**, (2017).
9. Rasmussen, D. A., Boni, M. F. & Koelle, K. Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam. *Mol. Biol. Evol.* **31**, 258–271 (2014).
10. Rasmussen, D. A., Volz, E. M. & Koelle, K. Phylodynamic inference for structured epidemiological models. *PLoS Comput. Biol.* **10**, e1003570 (2014).
11. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
12. Lau, M. S. Y., Marion, G., Streftaris, G. & Gibson, G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLoS Comput. Biol.* **11**, (2015).
13. Volz, E. M. & Siveroni, I. Bayesian phylodynamic inference with complex models. *PLoS Comput. Biol.* **14**, e1006546 (2018).

14. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
15. Firestone, S. M. *et al.* Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Sci. Rep.* **9**, 1–12 (2019).
16. Kestler, M., Muñoz, P., Mateos, M., Adrados, D. & Bouza, E. Respiratory syncytial virus burden among adults during flu season: an underestimated pathology. *J. Hosp. Infect.* **100**, 463–468 (2018).
17. Shi, T. *et al.* Global , regional , and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015 : a systematic review and modelling study. **390**, 946–958 (2017).
18. Troeger, C. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect. Dis.* **18**, 1191–1210 (2018).
19. Zlateva, K. T., Lemey, P., Moe, E., Vandamme, A. & Ranst, M. Van. Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein. **79**, 9157–9167 (2005).
20. Jepsen, M. T. *et al.* Incidence and seasonality of respiratory syncytial virus hospitalisations in young children in Denmark , 2010 to 2015. 1–8 (2018).
21. Obando-Pacheco, P. *et al.* Respiratory syncytial virus seasonality: A global overview. *J. Infect. Dis.* **217**, 1356–1364 (2018).
22. Broberg, E. K. *et al.* Seasonality and geographical spread of respiratory syncytial virus epidemics in 15 European countries, 2010 to 2016. *Eurosurveillance* **23**, 17–00284 (2018).
23. Agoti, C. N. *et al.* Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. *J. Virol.* **89**, 11630–11642 (2015).
24. Agoti, C. N. *et al.* Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *J. Virol.* **89**, 3444–3454 (2015).
25. Di Giallonardo, F. *et al.* Evolution of Human Respiratory Syncytial Virus (RSV) over Multiple Seasons in New South Wales, Australia. *Viruses* **10**, 476 (2018).
26. Tan, L. *et al.* The comparative genomics of human respiratory syncytial virus

- subgroups A and B: genetic variability and molecular evolutionary dynamics. *J. Virol.* **87**, 8213–26 (2013).
27. Agoti, C. N., Otieno, J. R., Gitahi, C. W., Cane, P. A. & Nokes, D. J. Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. *Emerg. Infect. Dis.* **20**, 950–9 (2014).
 28. Duvvuri, V. R. *et al.* Genetic diversity and evolutionary insights of respiratory syncytial virus A ON1 genotype: global and local transmission dynamics OPEN. (2015). doi:10.1038/srep14268
 29. Otieno, J. R. *et al.* Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evol.* **4**, vey027 (2018).
 30. van Niekerk, S. & Venter, M. Replacement of previously circulating respiratory syncytial virus subtype B strains with the BA genotype in South Africa. *J. Virol.* **85**, 8789–97 (2011).
 31. Cui, G. *et al.* Rapid replacement of prevailing genotype of human respiratory syncytial virus by genotype ON1 in Beijing, 2012–2014. *Infect. Genet. Evol.* **33**, 163–168 (2015).
 32. Kombe, I. K., Munywoki, P. K., Baguelin, M., Nokes, D. J. & Medley, G. F. Model-based estimates of transmission of respiratory syncytial virus within households. *Epidemics* **27**, 1–11 (2019).
 33. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Mol. Biol. Evol.* **34**, 2982–2995 (2017).
 34. Didelot, X., Gardy, J. & Colijn, C. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. (2014). doi:10.1093/molbev/msu121
 35. Cottam, E. M. *et al.* Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B Biol. Sci.* **275**, 887–895 (2008).
 36. Kenah, E., Britton, T., Halloran, M. E. & Jr, I. M. L. Molecular Infectious Disease Epidemiology : Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. 1–29 (2016). doi:10.1371/journal.pcbi.1004869
 37. Naveca, F. G. *et al.* Genomic, epidemiological and digital surveillance of

- Chikungunya virus in the Brazilian Amazon. *PLoS Negl. Trop. Dis.* **13**, e0007065 (2019).
38. Munywoki, P. K. *et al.* The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya. *J. Infect. Dis.* **209**, 1685–1692 (2014).
 39. Munywoki, P. K. *et al.* Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiol. Infect.* **143**, 804–12 (2015).
 40. Wathuo, M., Medley, G. F., Nokes, D. J. & Munywoki, P. K. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res.* **1**, 27 (2016).
 41. Agoti, C. N. Genomic analysis of respiratory syncytial virus infections in households and utility in inferring who infects the infant. *Sci. Rep.* (2019).
 42. Lau, M. S. Y., Marion, G., Streftaris, G. & Gibson, G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. 1–27 (2015).
doi:10.1371/journal.pcbi.1004633
 43. Ypma, R. J. F., Donker, T., van Ballegooijen, W. M. & Wallinga, J. Finding evidence for local transmission of contagious disease in molecular epidemiological datasets. *PLoS One* **8**, e69875 (2013).
 44. Lee, F. E., Walsh, E. E., Falsey, A. R., Betts, R. F. & Treanor, J. J. Experimental infection of humans with A2 respiratory syncytial virus. *Antivir. Res* **63**, 191–196 (2004).
 45. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **59**, 65–98 (2017).
 46. Edelman, A. The Julia Language. 1–51 (2013).
 47. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations)? *J. Comput. Graph. Stat.* **7**, 434–455 (1998).
 48. Shi, T., McLean, K., Campbell, H. & Nair, H. Aetiological role of common respiratory viruses in acute lower respiratory infections in children under five years: A systematic review and meta-analysis. *J. Glob. Health* **5**, 1–10 (2015).
 49. Bose, M. E. *et al.* Sequencing and analysis of globally obtained human respiratory syncytial virus a and B genomes. *PLoS One* **10**, (2015).

50. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *epidemiology* **2**, 13 (2005).
51. Kinyanjui, T. M. *et al.* Vaccine Induced Herd Immunity for Control of Respiratory Syncytial Virus Disease in a Low-Income Country Setting. *PLoS One* **10**, e0138018 (2015).
52. Agoti, C. N. *et al.* Genomic analysis of respiratory syncytial virus infections in households and utility in inferring who infects the infant. *Sci. Rep.* **9**, 10076 (2019).
53. Kiti, M. C. *et al.* Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. *EPJ Data Sci.* **5**, 21 (2016).
54. White, L. J. *et al.* Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math. Biosci.* **209**, 222–239 (2007).
55. *WHO strategy to pilot global respiratory syncytial virus surveillance based on the Global Influenza Surveillance and Response System (GISRS).* (2017).

5. Paper 3: A multi-pathogen model of infection investigating potential interactions between respiratory syncytial virus and coronavirus.

5.1. Overview

This chapter presents an analysis based on an extension of the model first introduced in Chapter 3. As with the previous chapter, this chapter is written in the format of a publication and we intend to submit it to a journal with the running title: ***A multi-pathogen model of infection investigating potential interactions between respiratory syncytial virus and coronavirus.***

5.2. Role of candidate

I formulated the problem, conducted the numerical analysis and wrote the first draft of the chapter. Revisions were made with feedback, input and guidance from my supervisors Graham F. Medley and D. James Nokes and advisor Marc Baguelin.

Research paper cover sheet



London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	Ish1601271	Title	Ms
First Name(s)	Ivy Kadzo		
Surname/Family Name	Kombe		
Thesis Title	Integrating viral RNA sequence and epidemiological data to define transmission patterns for respiratory syncytial virus		
Primary Supervisor	Professor Graham Medley		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	<i>To be decided</i>
Please list the paper's authors in the intended authorship order:	<i>Kombe I. K., Munywoki P.K., Baqelro M. Nokes D.J., Medley G.F.</i>
Stage of publication	Choose an item. <i>Pre-submission draft</i>

SECTION D – Multi-authored work

Improving health worldwide

www.lshtm.ac.uk

<p>For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)</p>	<p><i>I led the analysis, wrote the first draft and incorporated co-author comments. I also wrote all the code for the analysis.</i></p>
---	--

SECTION E

Student Signature	[Redacted]
Date	<i>25/07/2019</i>

Supervisor Signature	[Redacted]
Date	<i>2019 July 25</i>

5.3. Abstract

Respiratory syncytial virus (RSV) is a common viral pathogen that causes a significant burden of respiratory disease in children. RSV circulates in seasonal patterns and has often been observed to co-circulate with other viral pathogens such as influenza, human coronavirus, rhinovirus etc. Often, viral pathogens are assumed to be circulating independently, ignoring any possible interactions between pathogens in the same ecological system. This could lead to a miss-representation of the true disease burden attributed to a particular pathogen and ill-informed projections of the effect of an intervention targeted at a single pathogen. In light of this, we extended a previously developed multi-strain model of RSV to include data on RSV and human coronavirus in order to investigate potential interactions at the individual host level and extend these to infer transmission dynamics of the two pathogens in a small population of hosts. We found that interactions between the two pathogens are specific to particular groups. RSV B interacted with coronavirus OC43 through increased susceptibility to heterologous pathogen infection, where the susceptibility to corona OC43 was increased by about 81% (95% CrI: 40%, 134%) following an RSV B infection. Though the results of this study are based on a small population of hosts, the inferred interactions imply that a vaccine that reduces the transmission of RSV would also reduce the transmission of coronavirus OC43 and its associated disease burden. Further studies are warranted to explore these and other interactions between RSV and other pathogens at a larger geographical and temporal scale.

5.4. Introduction

Respiratory syncytial virus (RSV) is recognized as a major cause of respiratory disease in children less than 5 years old¹⁻³. RSV has clear epidemic patterns that coincide with winter seasons in temperate countries, but has less definitive correlates of seasonality in the tropics⁴. Co-circulation of RSV with other pathogens is common especially adenovirus and rhinovirus, both of which tend to be year-round as opposed to seasonal⁵⁻¹⁰. However, RSV and influenza have been shown to have similar epidemic timings during winter in temperate regions^{8,9,11-13}. In the tropics, it is not as clear which pathogens share similar epidemic timings with RSV, but observations have been made on co-circulation with human coronaviruses (HCoVs) and human metapneumovirus (HMPV)^{6,14-18}.

Given the ubiquitous nature of RSV and other respiratory pathogens, co-infections are common. In studies looking at the distribution of pathogens present in cases of respiratory illness, RSV-virus or RSV-bacteria co-infected samples represent a significant fraction of the total RSV samples^{5,7,10,15,19,20}. The effects of viral co-infections in general are not clear. Some studies report co-infections being associated with increased disease risk^{10,21,22}, others do not find any associations^{23,24}, while some have found an association with decreased disease risk²⁵. Viral-bacterial co-infections do have a clearer pattern of increased disease risk or severity²⁶. Cases of respiratory illness co-infected with RSV and a bacterial pathogen have been associated with increase disease severity²⁷⁻²⁹. More specifically, when looking at *Streptococcus pneumoniae*, Greenberg *et al* found that RSV more commonly occurs with non-invasive serotypes than invasive serotypes, hypothesizing that non-invasive serotypes do not typically cause disease unless there is a viral co-infection³⁰. Further evidence of a facilitative relationship between RSV and bacterial pathogens has been found³¹⁻³³. RSV-virus co-infections have been associated with increased disease risk or longer duration of hospital stay in some instances^{10,15,21,34-36}, which could also be indicative of facilitation, but further evidence is warranted.

Competitive relationships between RSV and other viruses have been proposed as explanations for the observed associations in data. Greer *et al* looked at data from

cases presenting with ARTI at hospitals and found an association between detection of RSV, rhino and HMPV and decreased detection of other viruses, implying competition³⁷. The cases from this study were mostly children. Martin *et al* followed children in day-care for two years and characterized respiratory illness. Though adenovirus, human bocavirus (HBoV), HCoVs, HMPV and rhinovirus often occurred together, RSV and rhinovirus occurred together less frequently than would happen by chance, a signal of competition between the two³⁸. Bhattacharyya *et al* found evidence of cross-immunity between paramyxoviruses, more so an immunizing effect of RSV, the strength of which increased with decreasing phylogenetic distance between viruses³⁹. At a cellular level, Shinjoh *et al* showed competition between RSV and influenza³⁸.

Facilitative or competitive pathogen interactions can occur at a cellular and/or host and/or population level. Despite the biological and epidemiological evidence of RSV interacting with other pathogens, many mathematical models of RSV do not account for it. In fact, pathogen interactions are rarely accounted for in most studies of viral transmission dynamics, yet if there are important pathogen interactions that affect population level dynamics, models that do not take this into account could be erroneous in their analysis of individual viruses. Mechanistic mathematical models are powerful tools for gaining a better understanding of disease transmission. They can be used not only to investigate and quantify mechanisms of pathogen interactions, but also to predict the population level impact of an intervention against one pathogen that interacts with other pathogens in the same host population⁴⁰. Multiple-pathogen models, in general, are not common. Statistical models that take into account co-circulation of pathogens have been used to assign causality to cases of respiratory disease^{3,41}. Asten *et al* looked at data on influenza like illness (ILI) spanning 10 years from the Netherlands and were able to establish, through regression, that a change in shift in the influenza A epidemic resulted in changes to the epidemics of other pathogens that usually circulate around the same time. When influenza A outbreaks occurred earlier, RSV outbreaks were delayed, and coronavirus outbreaks were intensified. RSV outbreaks in this dataset tended to start earlier than influenza outbreaks, as such when the influenza outbreaks were early and the RSV outbreaks late, they overlapped more⁴². In a recent systematic analysis of data from multiple

sites distributed across the globe, Li *et al* found that the global seasonal trend of RSV and influenza differed by both timing and length and despite RSV and HMPV having similar epidemic durations, they did not co-circulate in most countries, an observation which led to the hypothesis of a potential competitive interaction between RSV and HMPV. This analysis had the strong advantage of more study sites, hence more geographically representative data¹¹. Merler *et al* used a simple compartmental model with homogenous mixing to explore the hypothesis that co-infection with an acute respiratory infection increased the transmissibility of pandemic influenza, leading to multiple waves of cases during an outbreak. Their model, which had no seasonal forcing, was able to produce output that was in agreement with data showing multiple waves of the 1918 Spanish flu⁴³. Velasco-Hernández *et al* used a deterministic compartmental model to explore a hypothesis of a competitive interaction between RSV and influenza through super-infection, where influenza was treated as the superior pathogen capable of infecting hosts already infected with RSV. Their hypothesis was partially validated, using data from children <5 years old seeking treatment at a hospital in Mexico, despite that fact that their model needed further complexities to be more realistic⁴⁴. Pinky *et al* built an ODE model to explore RSV and influenza viral kinetics at a cellular level. In the model, within a co-infected host, RSV and influenza were competing to infect cells in the respiratory tract. The model was fit to data of *in vitro* co-infection and then used to determine that the virus with the highest growth-rate will outcompete other co-infecting viruses and infect more cells, however, this competitive advantage could be surpassed if the slower virus had a higher initial inoculum or an earlier infection time. Single pathogen *in vitro* and *in vivo* models comparing influenza and RSV found that indeed RSV had a slower rate of spread from cell to cell and hence viral titres increased at a slower rate. In addition, the infectious cell lifespan was shorter for RSV than influenza^{45,46}. Multi-strain models are much more common with a huge volume of literature around models of influenza⁴⁷. Models that look at the interactions between RSV groups are few and far between⁴⁸. In a recent review, Opatowskia *et al* highlighted the importance of multi-pathogen mechanistic models. They argue that once pathogen interactions have been adequately accounted for, then a more accurate picture of disease burden can be established and intervention programs can be better optimized⁴⁰.

In studying the interaction between species, the level of the data is critical. Most of the previous studies have either been at the population level (looking at population level patterns) or the within-individual level. The data from the HH study enable us to look in detail at the individual level, i.e. the simultaneous exposure and transmission patterns in co-circulating viruses. In this study, we propose to use a mechanistic mathematical model that tracks infection at the individual host level to investigate interactions between different strains (groups) of RSV and endemic strains of human coronavirus (HCoV). This choice of pathogens was based on the fact that RSV and HCoV were the only pathogens whose identification was also at a group level for all the observed cases. Rhinovirus was typed in only 5 of the 47 households, while no typing was carried out for adenovirus. The most frequently circulating strains of HCoV are HCoV-OC43, HCoV-NL63, HCoV-229E and HCoV-HKU1⁴⁹. The four groups cause mild to severe disease and have often been observed to co-circulate^{6,14,50-52}. As with RSV, children and the elderly are at an increased risk of symptomatic infection^{53,54}. There is evidence that the strains differ by host age group⁵⁵, symptoms⁵⁶ and seasonality⁵⁶. Repeat infections with HCoV are common⁶ and a recent phylogenetic analysis found evidence that changes to the HCoV-NL63 genome are not immune driven⁵⁷. Similar to RSV, HCoV-NL63 has been shown to have a facilitative interaction with *Streptococcus pneumoniae* through enhancing its cell adherence⁵⁸.

5.5. Methods

5.5.1. Data

We use data on RSV and human coronavirus (hCoV) shedding patterns collected from a household cohort study conducted in rural coastal Kenya within the Kilifi Health and Demographic Surveillance System (KHDSS) during the 2009/2010 RSV epidemic. A household was defined as a group of individuals living in the same compound who share a kitchen. Details of the study have been published elsewhere^{6,59–62}. In brief, the infant-centric study recruited household members using the criteria that the infant was born after 1 April 2009 (after the previous RSV epidemic) and had at least 1 older sibling less than 13 years old. Deep nasopharyngeal swab (NPS) samples were collected every 3-4 days regardless of symptoms, together with a record of clinical illness. The focus of the study was to investigate who infects the infant with RSV, however, three other pathogens were identified as frequently circulating in the study participants: rhinoviruses, adenoviruses and human coronaviruses with coronavirus further classified by group into 229E, OC43 and NL63. We did not include adenovirus and rhinovirus in the present analysis due to the lack of information on infecting virus species. This resulted in the observation of some shedding durations being as long as >60 days, which were probably re-infections by different species. Test runs were conducted with adenovirus and rhinovirus data in the model, but each time the inference algorithm failed to converge. Details of all the pathogens identified from the household study can be found in⁶. The data contain information from 493 individuals spread across 47 households whose dates of data collection span 180 days.

In addition to the data on shedding and symptom status, there is information on presence or absence from the household. Given the discontinuity in the sampling, complete shedding, and presence/absence durations had to be imputed. This imputation process has been described in detail in A2: Supplementary appendix for Paper 1. In brief, a virus shedding episode is defined as a period within which an individual provided PCR positive samples for the virus that were no more than 14 days apart. Individuals are assumed to start shedding halfway between the last negative sample and the first positive sample of the episode, and they stop shedding halfway in between the last positive sample of the episode and the first negative sample. In the

same way, complete presence/absence durations are imputed for all the days of data available for a particular individual. There are some instances where an individual was present but not sampled, as such, presence could not purely be identified by the availability of NPS samples. Imputation was chosen over data augmentation to ensure consistency across studies analysing the same household^{59,61,63}.

5.5.2. Transmission model

To interrogate the data on any possible interactions between pathogens, we built a model that tracks infection with RSV and coronavirus at the individual host level. This model is a modified version of an earlier model that was used to analyse RSV A and RSV B data⁶⁴. We extend the logic applied to modelling multiple groups/strains of RSV to model multiple groups/strains of multiple pathogens interacting through modified susceptibility. Every group within a pathogen is treated independently, so for RSV with two groups, we treat RSV A and RSV B as distinct infectious agents and look for an interaction. In the case where we have RSV and coronaviruses, we have a total of 5 infectious agents: RSV A, RSV B, corona 229E, corona OC43 and corona NL63. An individual is either susceptible to, or infected with, a particular infectious agent. An individual who is currently not shedding any of the infectious agents is considered susceptible to all of them, if they are shedding one, say RSV A, then they are susceptible to the other 4, RSV B, corona 229E, corona OC43 and corona NL63. For a single infectious agent, we assume SEIS₂ type dynamics where an individual is initially susceptible, they get exposed and go through a period of latency prior to onset of infectiousness after which they become susceptible again, but the susceptibility is modified as a result of having experienced an infection. Unlike in the previous models in Chapter 3 and 4. We do not assume an age effect on susceptibility, this is purely for computational reasons in order to reduce the dimension of the parameter vector. Also different from the models in the previous chapter is that fact that we do not fix the latency distribution. We assume that every pathogen has a gamma latency distribution with a specific mean and standard deviation (SD). The pathogen-specific mean and SD for these distributions are estimated along with other model parameters.

We model the rate at which an individual i , is getting exposed to infection by infectious agent v at time t , denoted $\lambda_{i,h,v}(t)$. Individuals can get infected by someone they share

a household with or from a source outside of the household, resulting, as in previous model versions, in a two-part rate of exposure equation.

$$\lambda(t) = [\text{baseline household exposure rate} * \text{number of infectious household contacts}(t)] \\ + \\ [\text{baseline community exposure rate} * \text{background community function}(t)]$$

$$\lambda(t) = \left(\eta * \sum_{\text{household}} I(t) \right) + (\varepsilon * f(t))$$

The number of infectious household contacts is observed in the data while infectious contacts from the community are represented using a derived function. We use the same function form that was used in Chapter 4, We define an infectious-agent specific background rate of exposure, $f_v(t)$ given as

$$f_v(t) = \delta + \sum_{\substack{i \text{ shedding} \\ \text{infection } v}} e^{(t-\tau_{i,v})\beta}$$

is the basic risk prior to any observed onsets and β is the rate of exponential decay related to the time since onset of a case shedding infectious agent v , β is a measure of the rate at which the infectious agent might disappear from the community and τ_i is the onset time by person i . The parameters δ and β are not infectious-agent or pathogen specific.

Where δ is the basic risk prior to any observed onsets and β is the rate of exponential decay related to the time since onset of a case shedding infectious agent v , β is a measure of the rate at which the infectious agent might disappear from the community and $\tau_{i,v}$ is the onset time of infectious agent v by person i .

Unlike in the previous model iterations, we do not include the effect of household size, viral load and ARI, or age in the rate of exposure. As with the removal of age effects on susceptibility, this was also done to reduce the dimension of the parameter vector. The main aim of this version of the model is to investigate possible pathogen interactions.

We model pathogen interactions through parameters that modify susceptibility based on infection history and current infection status. The interactions are investigated in a pair-wise manner. Say we have three infectious agents in the data, V_1 , V_2 and V_3 . Susceptibility to V_1 is modified based on previous infection with: V_1 indicated by the parameter ($risk.V_1.prev.V_1$), V_2 indicated by the parameter ($risk.V_1.prev.V_2$) and

V_3 indicated by the parameter ($risk.V_1.prev.V_3$). The first modification is due to previous homologous infection while the second and third are due to previous heterologous infection. In addition, susceptibility to V_1 is also modified if the individual is currently shedding V_2 indicated by the parameter ($risk.V_1.curr.V_2$) and V_3 indicated by the parameter ($risk.V_1.curr.V_3$). The same logic applies to modification of susceptibility to V_2 , V_3 and V_4 . We assume that the modification to risk of V_1 given previous or current infection with V_2 = modification of risk to V_2 given previous or current infection with V_1 , i.e. ($risk.V_1.prev.V_2$) = ($risk.V_2.prev.V_1$) and ($risk.V_1.curr.V_2$) = ($risk.V_2.curr.V_1$). This greatly reduces the number of interaction parameters to be estimated. The effect of multiple infections by different infectious agents is cumulative, if at time t an individual has experienced and recovered from infection by V_2 and V_3 then their susceptibility to V_2 is modified by a factor = $e^{(risk.V_2.prev.V_2 + risk.V_2.prev.V_3)}$. The rate of exposure to a particular infectious agent (index v) is given for a particular individual, (index i) from a given household (index h) at a given day (index t) and is specified by the notation $\lambda_{i,h,v}(t)$. The rate of exposure is given in equation Eq.5.1 and the variables are described in Table 5. 1.

$$\lambda_{i,h,v}(t) = \exp\left(\phi_Y(t) \times Y_{i,h,v}(t) + \phi_S \times S_{i,h,v}(t)\right) \left[\left(M_{i,h}(t) \times \eta_v \sum_{\substack{j \neq i, \\ j \text{ in} \\ i's \text{ household}}} (S_{j,h,v}(t) \times M_{j,h}(t)) \right) + (\varepsilon_v \times f_v(t)) \right] \dots \text{ (Eq 5.1)}$$

Table 5. 1: Description of variables in the model

Symbol	Type	Description
i	Index	Index of individual
h	Index	Index of household
v	Index	Index of the type of infectious agent
t	Index	Index of time in days
p	Index	Index of the type of pathogen

$S_{i,h,v}(t)$	Data	Binary data variable indicating if an individual i is shedding infectious agent v at time t
$Y_{i,h,v}(t)$	Data	Binary variable keeping track of an individual's infection history with respect to infectious agent v by time t .
$M_{i,h}(t)$	Data	Binary data variable indicating if an individual is present in the household at time t . Absence from the household means that an individual was not present at the point of sample collection and thus in the model, they can only get infection from a community source and not from an infectious housemate (not sampled and not at household risk). Individuals who were present but not sampled are exposed to both household and community source transmission in the models (not sampled but at household risk).
ϕ_Y	Parameter	Coefficients modifying susceptibility to infection by a particular infectious agent depending on infection history. The estimated effect could be due to previous homologous or previous heterologous infection. Applied to the categorical covariate $Y_{i,h,v}(t)$. The parameter name is <i>risk.V1.prev.V2</i> .
ϕ_S	Parameter	Coefficients modifying susceptibility to infection by a particular infectious agent depending on shedding status. The estimated effect is due to heterologous infection. Applied to the categorical covariate $S_{i,h,v}(t)$. The parameter name is <i>risk.V1.curr.V2</i> .
η_v	Parameter	Baseline rate of within household exposure specific to the infectious agent. The parameter name is <i>HH.V</i> .
ε_v	Parameter	Community transmission coefficient specific to the infectious agent. The parameter name is <i>Comm.V</i> .
δ, β	Parameters	For the infectious-agent specific background community function given by

$$f_v(t) = \delta + \sum_{\substack{i \text{ shedding} \\ \text{infection } v}} e^{(t-\tau_{i,v})\beta}$$

Delta (δ) is the basic risk and *Beta* (β) is the rate of exponential decay related to the time since onset of a case shedding infectious agent v . The parameter names are *Delta* and *Beta* and they are not pathogen or infectious-agent specific.

μ_p, σ_p Parameters The mean μ_p and standard deviation σ_p of a pathogen specific gamma distribution used to approximate the distribution of latency durations. Different groups/species in a single pathogen are assumed to have the same latency distribution. Latency durations are used in calculating the probability of onset given exposure.

$U_{i,h,v}$ Data Set of all days where individual i has an onset of infection with infectious-agent v . Only includes the first day of shedding for each infection episode.

$A_{i,h,g}$ Data Set of all the days where individual i is at risk of infection with infectious-agent v , i.e. they are not currently shedding v .

Following on from the rate of exposure equation are two additional nested equations that make up the model.

$\alpha_{i,h,v}(t)$ = Probability of infection following exposure per day i.e. individual enters the latent phase

$$\alpha_{i,h,v}(t) = (1 - \exp^{-\lambda_{i,h,v}(t)}) \dots \text{ (Eq 5.2)}$$

$P_{i,h,v}(t)$ = Probability of starting to shed i.e. individual enters the infectious phase at time t given they did not shed until t .

$$P_{i,h,v}(t) = \sum_{l=0}^L \theta_{l,p} \alpha_{i,h,v}(t-l) \dots \text{ (Eq 5.3)}$$

Where L is the maximum latent period and $\theta_{l,p}$ is the probability that the latent period is exactly l days. We assumed that the latency durations follow a

discretized gamma distribution truncated at 7 days. Given that incubation periods have been estimated to range from 2-5 days⁶⁵, we chose 0 to 7 days for possible latency durations. The mean and standard deviation of the gamma distribution are estimated for every pathogen p .

Since the model is focused on investigating if pathogen interactions determine the infection onset process, the data whose likelihood we are interested in is the onset data for the different infectious agents. As such, we express the likelihood of an individual's observed days of onset for all infectious agents as:

$$L_i = \prod_v \left[\prod_{u \in U_{i,h,v}} P_{i,h,v}(u) \prod_{a \in A_{i,h,v}} (1 - P_{i,h,v}(a)) \right]$$

Where $U_{i,h,v}$ is the set of days where individual i had an onset of infectious agent v and $A_{i,h,v}$ is the set of all days where i did not have an onset but was at risk of infection (i.e. not shedding infectious agent v). As with the previous iterations of the model, we assumed binomially distributed data.

While fitting the data with the pathogens identified at the group level, the model has 37 parameters, however, reducing the model to fit the data identified at the pathogen level results in 14 parameters. The parameters for the latency distribution μ_p, σ_p were estimated once using the data identified at the pathogen level. The inferred values were then fixed for the model with pathogen identification at the group level.

5.5.3. Parameter inference

We used Bayesian inference to obtain estimates of the parameters. Adaptive Metropolis Markov Chain Monte Carlo was used to explore the parameter space⁶⁶. In brief, the method builds a Markov chain which allows us to sample from the posterior distribution $P(\varphi/D)$ of the parameters given the data, where $\varphi = \{ \phi_Y, \phi_S, \eta_v, \varepsilon_v, \delta, \beta, \mu_p, \sigma_p \}$. Normal distributions with large standard deviations were used as weakly informative priors for the log of all the parameters. The algorithm is initiated with narrow standard deviations in the joint parameter proposal distribution, which are adjusted after a specified number of accepted proposals.

We initiated 3 chains and set the algorithm to start adapting the proposal distribution based on accepted parameters after 25,000 iterations. Burn-in was assessed visually after which the results of the three concurrent chains were combined to infer the posterior distribution. The three chains were run for 250,000 iterations each. The parameters were estimated on the log scale All the computation was done using R software package (RStudio version 1.1.383 running R version 3.4.0⁶⁷). The code is freely available under the GNU Lesser General Public License v3.0 and can be found at [https://github.com/Ikadzo/HH Transmission Model](https://github.com/Ikadzo/HH_Transmission_Model).

5.6. Results

5.6.1. Data

A summary of the distribution of cases by infectious agent is given in Table 5. 2. A majority of the onsets were due to coronaviruses with Corona OC43 having the highest number of onsets and Corona 229E the lowest. All the households in the study experienced at least one coronavirus infection, while 7 households did not get any RSV infections. There were some cases that experienced re-infections in each of the 5 infectious agents considered. RSV A and RSV B had similar proportions of onsets that were re-infections (10.2% and 10.6% respectively), while of the coronaviruses, NL63 had the highest proportion of re-infections (32.5%). RSV A had the highest proportion of onsets accompanied by an acute respiratory illness while HCoV 229E had the lowest, 61% and 26% respectively.

Table 5. 2: Summary of the data

Infectious agent	Number of onsets	Number of onsets with an ARI	Number of people infected	Number of repeat infections	Number of households infected
RSV A	97	59	88	9	25
RSV B	125	69	113	12	34
RSV	208	119	179	29	40
Corona-229E	133	34	119	14	30
Corona-NL63	216	85	163	53	33
Corona-OC43	260	118	215	45	44
Corona	565	228	346	219	47

Figure 5. 1 shows the temporal distribution of cases clustered by age group for ages <1 year, 1-5 years, 5-15 years and >15 years. The number of individuals in the study in each age group increased by age, from 55 <1 year olds to 191 >15 year olds, however the number of onsets for each pathogen was highest in the 5-15 age group. Figure 5. 2

shows the distribution of shedding episodes for each individual that had a recorded onset. Despite the fact that there were hundreds of coronavirus cases, there were very few RSV-coronavirus co-infections. Figure A4. 1 in A4: Supplementary appendix for Paper 3. shows the distribution of shedding durations by infectious agent and pathogen.

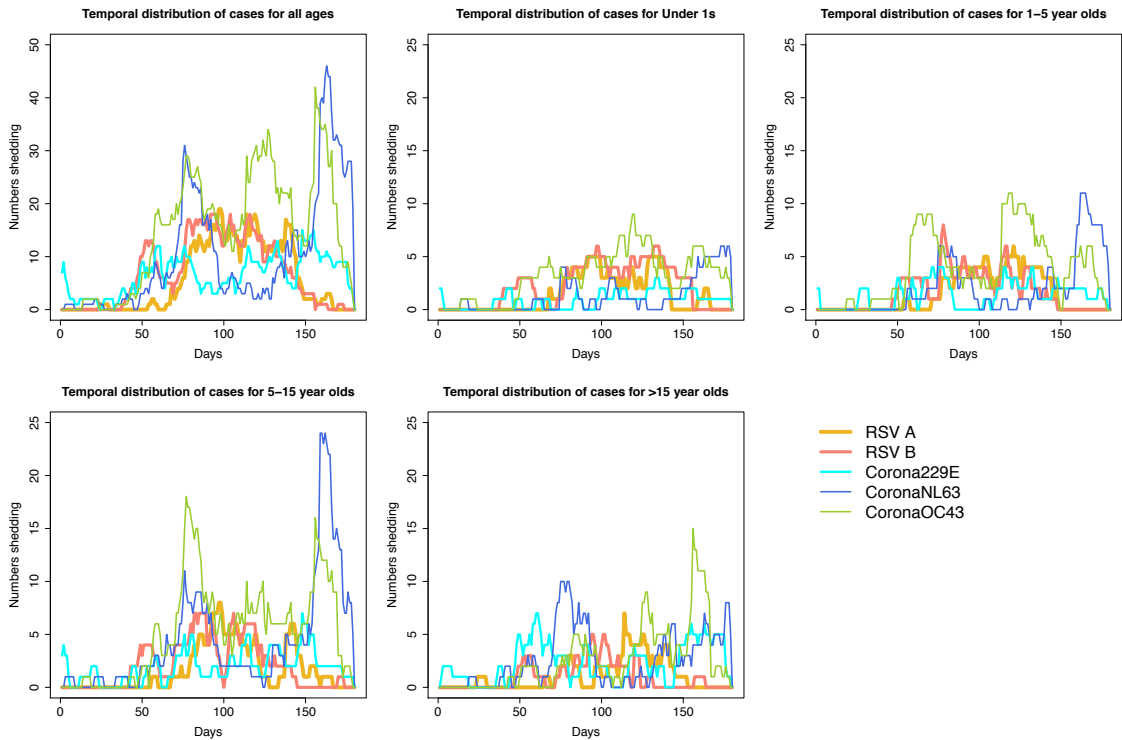


Figure 5. 1: Temporal distribution of cases for the 5 infectious agents clustered by age group.

In each panel, the x-axis shows time in days while the y-axis shows the total number of infectious people. Top-left: The temporal distribution of all the cases in the data; Top-centre: temporal distribution for all the cases <1 year old; Top-right: 1-5 year olds; Bottom-left: 5-15 year olds and Bottom-centre: > 15 year olds.

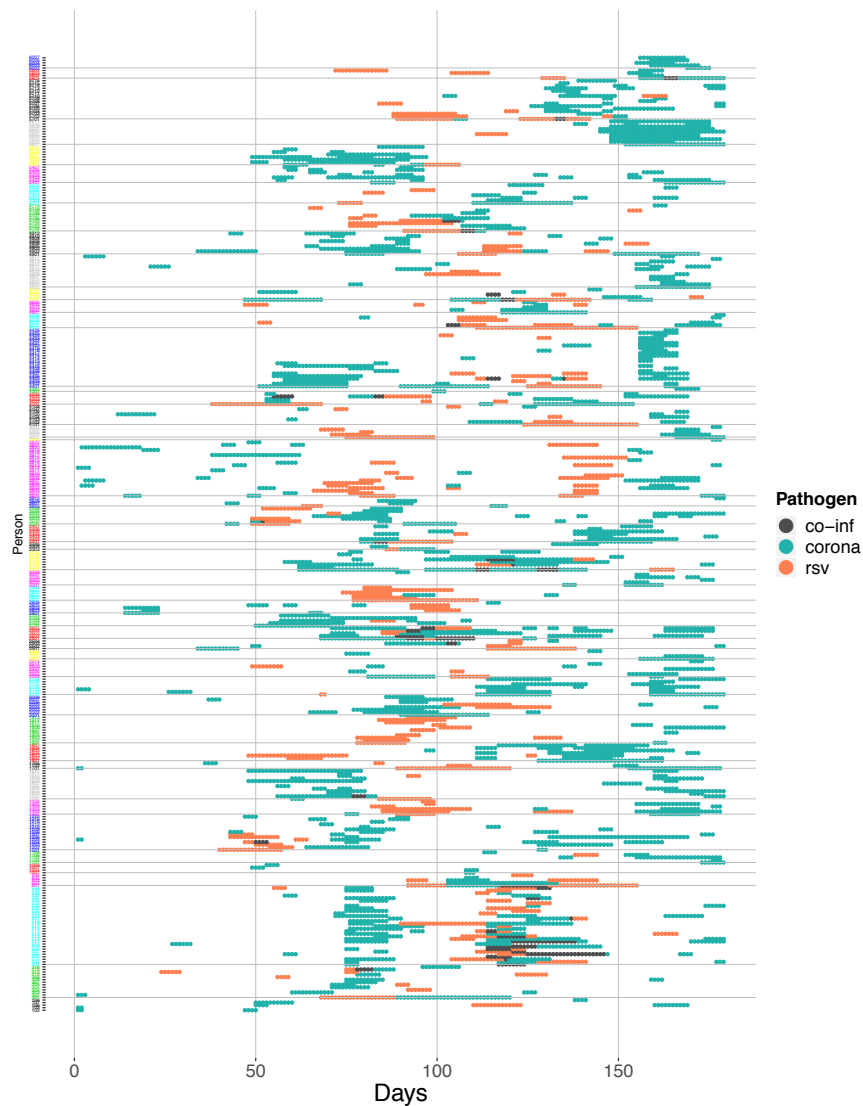


Figure 5. 2: Distribution of shedding episodes for coronavirus and RSV by household and time.

The x-axis shows the time in days while the y-axis shows the individuals, where each notch is a single individual. The horizontal lines demarcate the different households. The labels on the y-axis are color-coded to separate the different households.

5.6.2. Pathogen interactions

This section of results aims to answer two questions: Are there interactions between RSV and coronavirus that can be detected from the data? Are these interactions detected at different scales of pathogen identification? To answer these questions, the data was fitted in two ways, first with identification of coronavirus and RSV at the

pathogen level, then at the group level. The comparable parameter densities that came out of this analysis are shown in Figure 5. 3 and Table A4. 1 in appendix A4. The trace plots showing the results of the MCMC algorithm are given in appendix A4. Convergence was assessed visually and confirmed using the Gelman-Rubin-Brooks (GRB) statistic ⁶⁸.

Starting from panel *a*) in the top row in Figure 5. 3, the risk of RSV infection given an individual has already experienced at least one other RSV infection in the present outbreak is reduced by about 40% compared to an individual who has no previous RSV infection. This reduction in susceptibility remains approximately unchanged whether infection is identified at a pathogen (solid black line) or group level (dashed coloured lines). However, the 95% Credible interval for the parameters estimated with the group level data now includes 1 on the fringes of the interval. The first interaction parameter in panel *b*) measures the reduce susceptibility to reinfection by a heterologous pathogen, i.e. risk of RSV given previous coronavirus and vice versa. Here there is a noticeable change in distribution when the pathogens are identified at a finer scale. When identification is at a pathogen level, the modified susceptibility is 1.33 (95% CrI: 1.14, 1.57), which indicates an increased risk of infection. However, with identification at the group level, this effect is increased for an (RSV B - OC43) interaction to 1.81 (1.4, 2.34) and reduced for an (RSV A - 229E) interaction to 0.698 (0.383, 1.17). These parameters are estimated such that the interactions are assumed to be symmetric, i.e. previous infection with RSV B increases the risk to subsequent infection with coronavirus OC43 and vice versa. However, the data shows that in 66% (40 out of 61) of the individuals who had an RSV B and OC43 infection, the RSV B infection preceded the OC43 infection. In individuals who had an RSV A and 229E infection, 59% (10 out of 17) of them had the coronavirus infection prior to the RSV infection. These shedding patterns are shown in Figure 5. 2 and Figure A4. 2, Figure A4. 3, Figure A4. 4 and Figure A4. 5 in appendix A4. The other group level interactions have their distributions centred closer to 1, which implies no effect. The risk of infection with HCoV is modified given previous infection within the same epidemic, however, unlike with RSV, the direction of effect is not as clear, panel *c*). Previous infection reduces susceptibility to homologous group infection for OC43 ($risk.oc43.prev.oc43=0.58$ (0.413, 0.79)) and NL63 ($risk.nl63.prev.nl63=0.617$ (0.438,

0.844)). The interaction between NL63 and the other HCoV groups appears to be of increased susceptibility, as the medians of the distributions are above 1, however, the 95% Credible intervals do include one so the effect cannot be stated with significance ($risk.229e.prev.nl63=1.1$ (0.806, 1.46) and $risk.nl63.prev.oc43=1.16$ (0.914, 1.48)).

The other interaction between RSV and coronavirus measured by the susceptibility to co-infection ($risk.rsv.curr.corona$) shown in panel *d*) does not have a strong signal regardless of the level of pathogen identification, the distribution of the estimated parameters are either centred around 1 or have a wide credible interval. The values of the within household transmission coefficients are dependent on the level of pathogen identification. For RSV ($HHrsv$) shown in panel *e*), when identification is at the pathogen level, the value is 0.0038 (0.00291, 0.00508) while identification at the group level increases the within household coefficient for RSV A to 0.00544 (0.00379, 0.00758). The change in distributions with increased pathogen resolution is also observed when looking at the coronavirus within household transmission coefficient ($HH.corona$) shown in panel *f*). In general, coronaviruses had higher values for the within household transmission coefficient than RSV ($HH.229e=0.00795$ (0.00577, 0.0108), $HH.nl63 = 0.0117$ (0.00939, 0.0145) and $HH.oc43 = 0.00547$ (0.00428, 0.00681), compared to $HH.rsva = 0.00544$ (0.00379, 0.00758) and $HH.rsvb = 0.00408$ (0.00282, 0.00555)). The distributions of the community transmission coefficients have a narrower credible interval when identification of the pathogen is at the group level; however, the distributions of the individual groups are not too different from each other whether looking at RSV ($Comm.rsva = 0.000186$ (0.000101, 0.000317) and $Comm.rsvb = 0.000217$ (0.00012, 0.000357)) shown in panel *g*), or coronavirus ($Comm.229e = 0.000242$ (0.000132, 0.000398), $Comm.nl63 = 0.000181$ (0.0000996, 0.000297) and $Comm.oc43 = 0.000297$ (0.000167, 0.000485)) shown in panel *h*). The parameters for the background community function are not significantly altered by the resolution of pathogen identification (Δ and β) shown in panel *i*) and *j*) respectively.

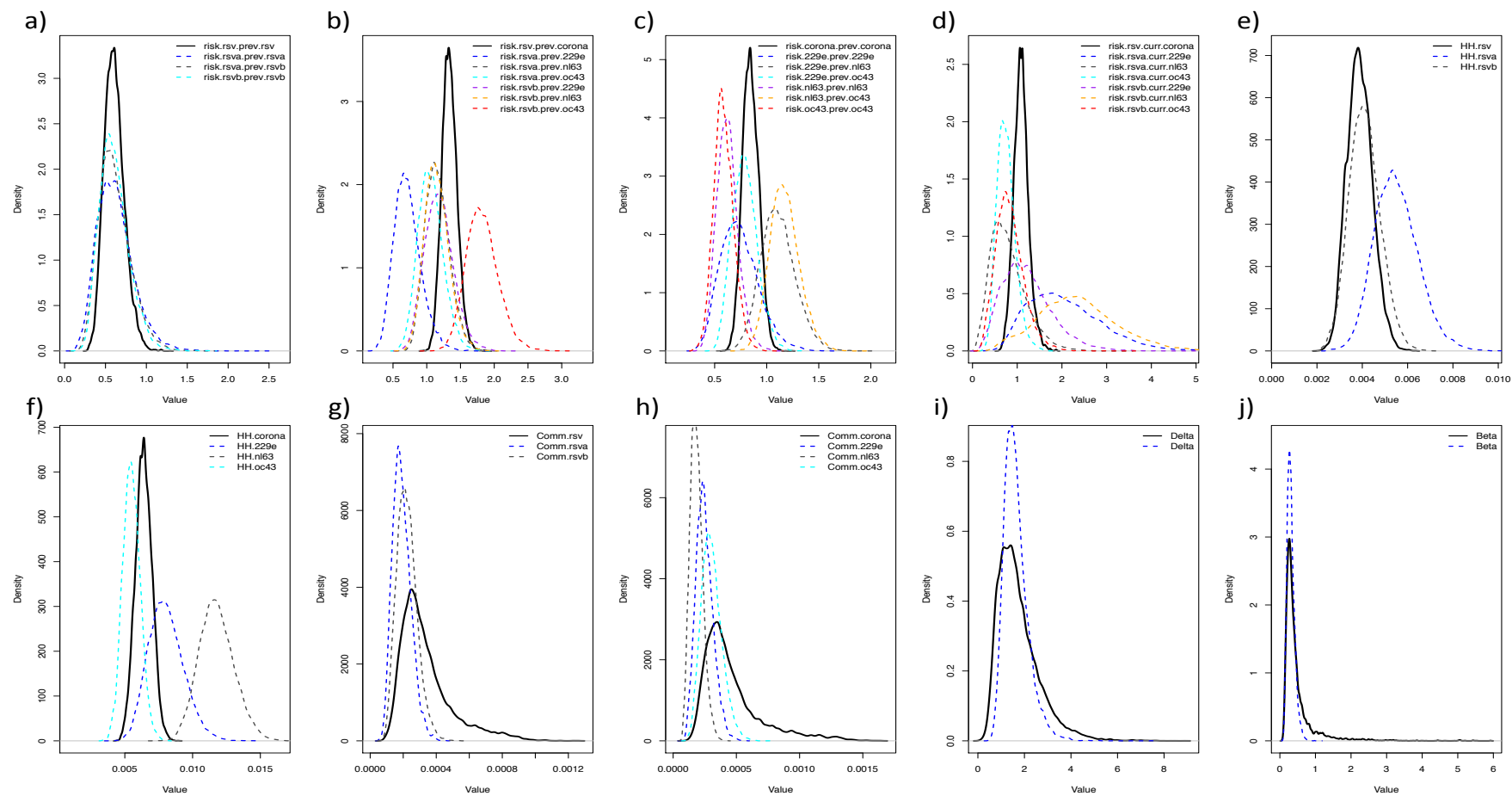


Figure 5. 3: Comparing parameter densities obtained from fitting the data at the pathogen level (solid black lines) to the densities obtained from fitting data at the group level (dashed coloured lines).

Estimates of the mean and standard deviation of latency distribution were derived for RSV and coronavirus. RSV had a mean latency distribution of 3.05 days (2.19, 3.7) and standard deviation of 0.683 (0.323, 1.21); coronavirus had a mean of 2.95 days (2.4, 3.62) and standard deviation of 0.712 (0.518, 0.938).

5.6.3. Modified pathogen inference

This section of results compares the pathogen specific parameters estimated when independently fitting the pathogen (either RSV or coronavirus) to parameters estimated when fitting multiple pathogens (RSV and coronavirus). The results are shown in Figure 5. 4 for RSV and **Error! Reference source not found.** for coronavirus. The trace plots showing the results of the MCMC algorithm are given in appendix A4. For RSV, there are slight shifts in the distributions of the parameters for RSV B within household transmission coefficients (towards smaller values), and RSV A community transmission coefficients (towards larger values) when going from a single-pathogen fit to a multi-pathogen fit. The shift observed in the parameters of the background community function (*Delta* and *Beta*) imply that when considering multiple interacting pathogens in a model, the basic risk prior to any observed onsets is higher and the rate of exponential decay of risk following observed onsets is much faster. The rates of exposure from within the household and from the community need to balance out in a way that explains the timing of the cases. Since *Delta* and *Beta* are not pathogen specific, the shift was probably necessary in order to explain the number of observed RSV and HCoV index cases in the data and subsequent cases in the same time window that were likely part of the same transmission chain as the index cases.

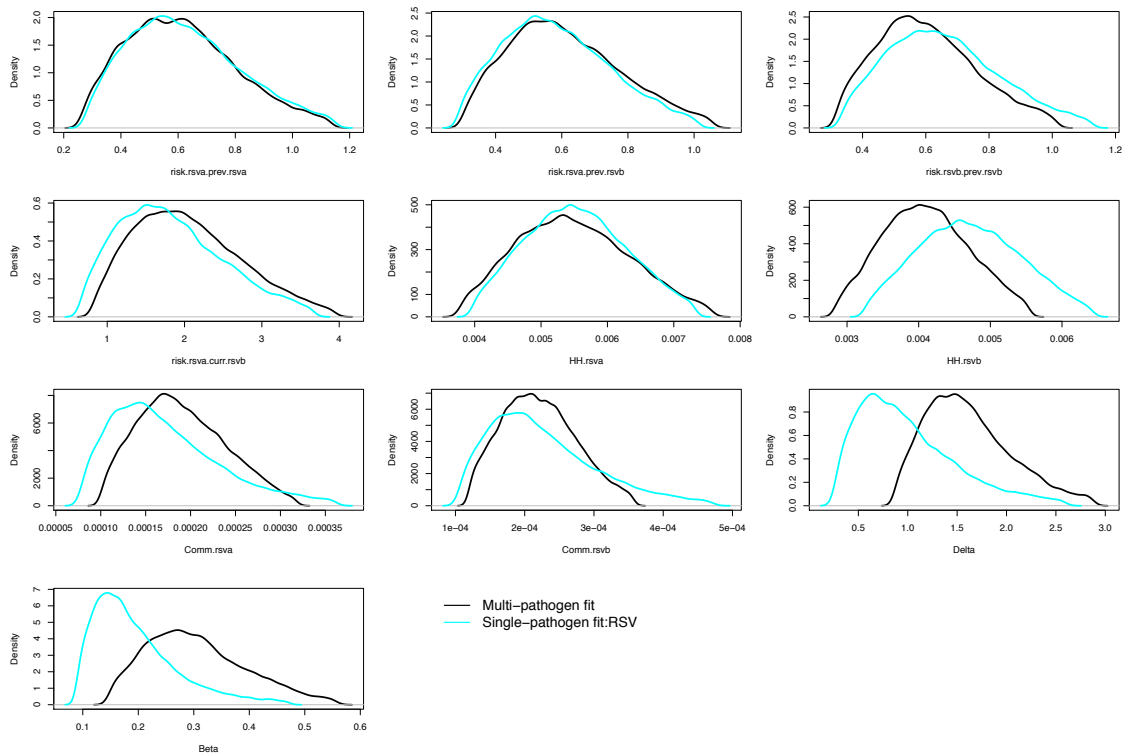


Figure 5. 4: Comparison of the RSV specific parameters obtained from fitting a single pathogen model (blue line) to those obtained from fitting a multi-pathogen model (black line).

For coronavirus, fitting of multiple interacting pathogens results in slight shifts in the distributions of some of the parameters, while the parameters that define the background community rate of exposure (*Comm.229e*, *Comm.nl63*, *Comm.oc43*, *Beta*) become better defined. Noticeably, when fitting HCoV as a single pathogen the distributions for *Comm.229e*, *Comm.nl63*, *Comm.oc43* and *Beta* appear to be bimodal. It is worth noting however that the model with HCoV data took a lot longer to converge relative to the version with RSV data. Increasing the number of iterations to 400,000 resulted in an increase in the effective sample sizes, however it did not resolve the bimodal distributions observed for these parameters as seen in Figure 5. 5.

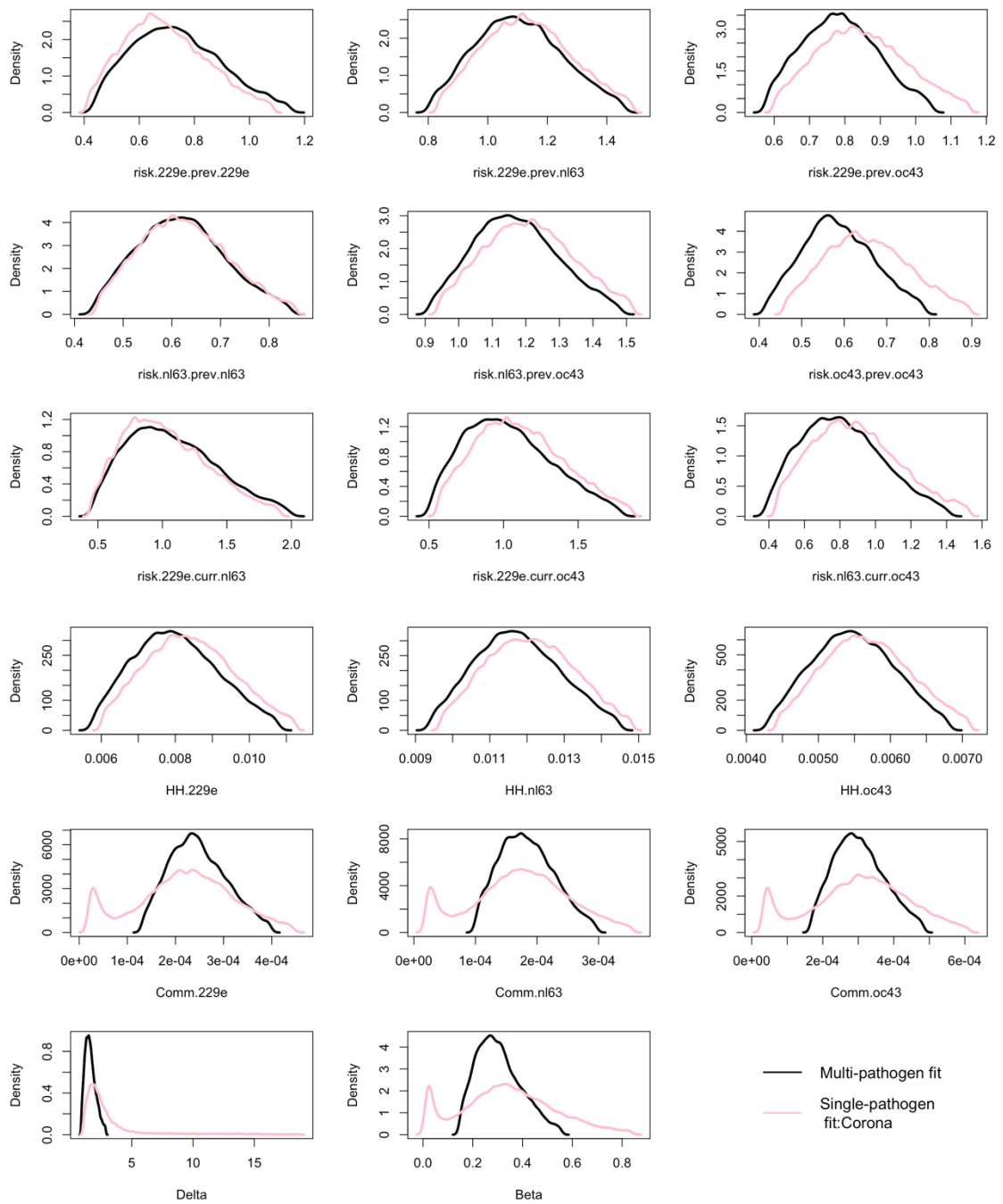


Figure 5. 5: Comparison of the coronavirus specific parameters obtained from fitting a single pathogen model (pink line) to those obtained from fitting a multi-pathogen model (black line).

5.7. Discussion

We extended a multi-strain individual level transmission model of RSV to be able to fit data from multiple pathogens and investigated the interactions between RSV and coronavirus. The two pathogens interacted through modified susceptibility and we found evidence that RSV and HCoV interacted through increased susceptibility to heterologous pathogen re-infection. With increased resolution at the group level, this effect was only significant for the interaction between RSV B and OC43 where previous infection with RSV B increased an individual's susceptibility to coronavirus OC43 by about 81% (95% CrI: 40%, 134%). Though the modification to susceptibility was assumed to be symmetric, i.e. modified susceptibility to RSV B given previous OC43 equals modified susceptibility to OC43 given previous RSV B, the pattern in the data was such that RSV B infections preceded OC43 infections 66% of the time. The exact mechanism of this facilitative interaction between RSV B and HCoV OC43 is unknown; differences in target host cells, immune response to infection or modification in host behaviour following infection could play a role in the observed dynamics. *In vitro* and *in vivo* infection studies coupled with models of viral kinetics would provide more data at different levels of interaction^{45,46}. In addition to investigating interactions, we compared inference made when fitting data from a single pathogen to fitting data from multiple pathogens and found that though there were no drastic changes to the pathogen specific parameters, fitting more data did lead to better resolution in some of the parameter distributions, more so for HCoV.

The data used in this study showed that 26-45% of coronavirus onsets were accompanied by an acute respiratory illness, with OC43 having the higher case-disease ratio. Though this is not as high as for RSV, 55% for RSV B and 61% for RSV A, it is still significant evidence of the importance of coronaviruses as contributors to respiratory illness. Given our inferred interaction between RSV B and coronavirus OC43, it follows that if a vaccine against RSV was introduced leading to a reduction in RSV transmission, it could also lead to a reduction in coronavirus OC43 transmission and associated disease. Since there is no clear evidence of competition between the coronavirus groups, a reduction in OC43 transmission is unlikely to result in its replacement in dominance by the other coronavirus groups. In this case, the inferred pathogen interaction would lead to a positive unintended effect of RSV vaccination against

coronavirus. Such a scenario would not be novel. It has been shown that the effect of the measles mumps and rubella (MMR) vaccine has had an impact on measles related cases and non-measles related deaths. This is as a result of preventing measles infections which lead to immunomodulation making the host more susceptible to infections for up to 3 years after the measles infection⁶⁹. This is one significant example of how complex pathogen interactions can be, necessitating the need to consider viruses as potentially being linked to each other in order to fully understand their epidemiology.

Previously, a different version of the model used in the present analysis was used to fit RSV data identified up to the level of genetic clusters (see Chapter 4). This analysis found evidence of an interaction between RSV A and RSV B where previous infection with either RSV A or RSV B reduced the risk of heterologous group re-infection by 49%(95% CrI: 10%-74%). In the present analysis, this effect was not as clear as the estimates for the modified susceptibility included one. This could be a result of a difference in model assumptions or due to the increased resolution in pathogen identification. Even with such detailed data, the more pathogens that are considered together in the same system, the more interaction parameters are needed leading to a decrease in statistical power to make inference and an increase in computational demands. In moving from the model in Chapter 4 to the model presented here, parameters quantifying the effect of age, household size and viral load were forgone in order to estimate parameters that would allow inference on potential interactions.

The results of this analysis must be considered along with its limitations. Interaction between the infectious agents was through modified susceptibility to heterologous re-infections or co-infections. Other potential mechanisms of interaction such as modified infectiousness, modified duration of infection or ecological interference by way of modified behaviour following infection, were not explored. Ecological interference is unlikely to have a significant effect when considering transmission events within households⁷⁰. Modified susceptibility is a common place to begin when investigating pathogen interactions^{39,48,71}. While it is plausible that a combination of modifications could be contributing to observed pathogen dynamics, including all possible interactions could quickly lead to an intractable model, as such choices have to be

made on the system being represented by the model and conclusions should be drawn in light of the simplifying assumptions⁷². There is however evidence from this data set that pathogens could be interacting through modified duration of infection. Previous infection with other viruses was found to be associated with shorter RSV shedding episodes while co-infections were associated with longer shedding episodes⁶¹. It was assumed that the interactions between the pathogens were symmetric, effectively ignoring the order of infection events within a single individual. Though this is also assumed when using compartmental models that are fitted to data that span larger geographical and temporal scales^{39,48,71}, it is often done due to a lack of suitable data to determine the order of infection with multiple pathogens. That is not the case with the household data used in the present analysis, as a first step to extend this analysis in preparation for publication, the symmetry assumption will be relaxed. It was also assumed that immunity following exposure to multiple pathogens is built up geometrically, i.e. if previous infection with pathogen X reduces susceptibility to infection by Y by a factor of 0.7 and previous infection with Z reduces susceptibility to Y by a factor of 0.5, then previous infection by X and Z reduced susceptibility to Y by a factor of 0.35. Making such an assumption for pathogens that appear to co-circulate is reasonable since the data has a record of infection history during the epidemic period. However, were it clear that one of the pathogens is significantly out of sync with the rest, then this assumption would not be appropriate. Instead, the model could be formulated such that only the most recent infection contributes to modified susceptibility⁷².

The data spanned a relatively short temporal period of 6 months bringing into question whether the observations are generalizable across epidemics. However, the detailed nature of the data that captured repeat infections with different pathogens is a great advantage relative to cross-sectional studies that mostly capture single onsets making it difficult to explore pathogen relationships at the host level. Cross-sectional studies often have the advantage of capturing transmission dynamics at a larger temporal and geographical scale, such data is available from the same local area as the household data and has previously been used to fit a compartmental model⁷³. A multi-scale model that combines the individual level dynamics and extrapolates these to the population level would elucidate if the pathogen interactions observed at the

individual host level significantly impact transmission at the population level. The model was only fit to data from RSV and coronavirus onsets but there are other viral pathogens in circulation, most notably rhinovirus and adenovirus, both of which could have potential interactions with RSV, as well as unobserved bacterial infections. The choice in which pathogen to fit was based on the availability of pathogen identification at a finer scale. Rhinovirus has over 100 serotypes as such, treating it as a single homogenous pathogen is likely to lead to incorrect inference. All the households in the data experience at least one rhinovirus introduction, but only 5 of the 47 households in the data had the rhinoviruses typed⁷⁴. In preparing this analysis for publication, a sub-analysis of 5 households will be included where three pathogens with identification at the group level are used to fit the model, RSV, HCoV and rhinovirus. Due to the relatively short study period, the circulation of other pathogens might have been missed, in particular the HMPV epidemic. Susceptibility to HMPV has been shown to be altered by an RSV infection³⁹ which could explain why the epidemics were not observed to overlap. Such potentially competitive interactions warrant further exploration and extrapolation to a larger population of hosts.

In conclusion, this study, to the best of our knowledge, presents the first dynamic multi-pathogen model of RSV and coronavirus group specific data. We show that interactions between the two pathogens are group specific, which could explain the contradictory observations from previous studies on the effect of RSV infection on other respiratory pathogens. Quantifying the level of interaction between pathogens and understanding how this influences the transmission dynamics of each could help to design optimized control strategies. Future studies should look at pathogen interactions at multiple levels such as cellular, individual host, etc., identifying parameters that could then be used in mechanistic population models of transmission. Most models assume viruses are independent, a factor that is unlikely to be true. There are challenges in determining the level and extent of interactions; however, pathogen interactions could shed light on long standing issues such as drivers of seasonality and pathogenicity.

5.8. References

1. Nair, H. *et al.* Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet* **375**, 1545–1555 (2010).
2. Shi, T., McLean, K., Campbell, H. & Nair, H. Aetiological role of common respiratory viruses in acute lower respiratory infections in children under five years: A systematic review and meta-analysis. *J. Glob. Health* **5**, 1–10 (2015).
3. O'Brien, K. L. *et al.* Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *Lancet* **6736**, 1–23 (2019).
4. Obando-Pacheco, P. *et al.* Respiratory syncytial virus seasonality: A global overview. *J. Infect. Dis.* **217**, 1356–1364 (2018).
5. Yan, X. *et al.* Clinical characteristics and viral load of respiratory syncytial virus and human metapneumovirus in children hospitalized for acute lower respiratory tract infection. *J. Med. Virol.* **89**, 589–597 (2017).
6. Munywoki, P. K. *et al.* Continuous invasion by respiratory viruses observed in rural households during a respiratory syncytial virus seasonal outbreak in coastal Kenya. *Clin. Infect. Dis.* **ciy313–ciy313** (2018).
7. Taylor, S. *et al.* Respiratory viruses and influenza-like illness: Epidemiology and outcomes in children aged 6 months to 10 years in a multi-country population sample. *J. Infect.* **74**, 29–41 (2017).
8. Prasad, N. *et al.* Interactive effects of age and respiratory virus on severe lower respiratory infection. *Epidemiol. Infect.* **146**, 1861–1869 (2018).
9. Douros, K. *et al.* Evidence for respiratory viruses interactions in asymptomatic preschool-aged children. *Allergol. Immunopathol. (Madr)*. **47**, 260–264 (2019).
10. Mazur, N. I. *et al.* Severity of respiratory syncytial virus lower respiratory tract infection with viral coinfection in HIV-uninfected children. *Clin. Infect. Dis.* **64**, 443–450 (2017).
11. Li, Y. *et al.* Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: a systematic analysis. *Lancet Glob. Heal.* **7**, e1031–e1045 (2019).
12. Ciçek, C. *et al.* Simultaneous detection of respiratory viruses and influenza A virus subtypes using multiplex PCR. *Mikrobiyol Bul* **48**, 652–660 (2014).

13. Antalis, E. *et al.* Mixed viral infections of the respiratory tract; an epidemiological study during consecutive winter seasons. *J. Med. Virol.* **90**, 663–670 (2018).
14. Friedman, N. *et al.* Human Coronavirus Infections in Israel: Epidemiology, Clinical Symptoms and Summer Seasonality of HCoV-HKU1. *Viruses* **10**, 515 (2018).
15. da Silva, E. R. *et al.* Severe lower respiratory tract infection in infants and toddlers from a non-affluent population: Viral etiology and co-detection as risk factors. *BMC Infect. Dis.* **13**, 41 (2013).
16. Pretorius, M. A. *et al.* Respiratory viral coinfections identified by a 10-Plex real-time reverse-transcription polymerase chain reaction assay in patients hospitalized with severe acute respiratory illness-South Africa, 2009-2010. *J. Infect. Dis.* **206**, S159-65 (2012).
17. Njouom, R. *et al.* Viral etiology of influenza-like illnesses in Cameroon, January-December 2009. *J. Infect. Dis.* **206**, S29–S35 (2012).
18. Pilger, D. A., Cantarelli, V. V., Amantea, S. L. & Leistner-Segal, S. Detection of human bocavirus and human metapneumovirus by real-time PCR from patients with respiratory symptoms in Southern Brazil. *Mem. Inst. Oswaldo Cruz* **106**, 56–60 (2011).
19. Appak, Ö., Duman, M., Belet, N. & Sayiner, A. A. Viral respiratory infections diagnosed by multiplex polymerase chain reaction in pediatric patients. *J. Med. Virol.* **91**, 731–737 (2019).
20. Jiang, W. *et al.* Etiologic spectrum and occurrence of coinfections in children hospitalized with community-acquired pneumonia. *BMC Infect. Dis.* **17**, 787 (2017).
21. Asner, S. A., Rose, W., Petrich, A., Richardson, S. & Tran, D. J. Is virus coinfection a predictor of severity in children with viral respiratory infections? *Clin. Microbiol. Infect.* **21**, 264.e1-264.e6 (2015).
22. Marcone, D. N. *et al.* Viral etiology of acute respiratory infections in hospitalized and outpatient children in Buenos Aires, Argentina. *Pediatr. Infect. Dis. J.* **32**, e105-10 (2013).
23. Petrarca, L. *et al.* Acute bronchiolitis: Influence of viral co-infection in infants hospitalized over 12 consecutive epidemic seasons. *J. Med. Virol.* **90**, 631–638

- (2018).
24. Ljubin-Sternak, S. *et al.* Etiology and Clinical Characteristics of Single and Multiple Respiratory Virus Infections Diagnosed in Croatian Children in Two Respiratory Seasons. *J. Pathog.* **2016**, 1–8 (2016).
 25. Martin, E. T., Kuypers, J., Wald, A. & Englund, J. A. Multiple versus single virus respiratory infections: Viral load and clinical disease severity in hospitalized children. *Influenza Other Respi. Viruses* **6**, 71–77 (2012).
 26. Beadling, C. & Slifka, M. K. How do viral infections predispose patients to bacterial infections? [Miscellaneous Article]. *Curr. Opin. Infect. Dis.* **17**, 185–191 (2004).
 27. Jeannoël, M. *et al.* Microorganisms associated with respiratory syncytial virus pneumonia in the adult population. *Eur. J. Clin. Microbiol. Infect. Dis.* **38**, 157–160 (2019).
 28. Suárez-Arrabal, M. C. *et al.* Nasopharyngeal bacterial burden and antibiotics: Influence on inflammatory markers and disease severity in infants with respiratory syncytial virus bronchiolitis. *J. Infect.* **71**, 458–469 (2015).
 29. Hishiki, H. *et al.* Incidence of bacterial coinfection with respiratory syncytial virus bronchopulmonary infection in pediatric inpatients. *J. Infect. Chemother.* **17**, 87–90 (2011).
 30. Greenberg, D. *et al.* Nasopharyngeal pneumococcal carriage during childhood community-acquired alveolar pneumonia: Relationship between specific serotypes and coinfecting viruses. in *Journal of Infectious Diseases* **215**, 1111–1116 (2017).
 31. Sande, C. J. *et al.* Airway response to respiratory syncytial virus has incidental antibacterial effects. *Nat. Commun.* **10**, 1–11 (2019).
 32. Avadhanula, V. *et al.* Respiratory Viruses Augment the Adhesion of Bacterial Pathogens to Respiratory Epithelium in a Viral Species-and Cell Type-Dependent Manner Downloaded from. *J. Virol.* **80**, 1629–1636 (2006).
 33. Weinberger, D. M. *et al.* Seasonal drivers of pneumococcal disease incidence: Impact of bacterial carriage and viral activity. *Clin. Infect. Dis.* **58**, 188–194 (2014).
 34. Harada, Y. *et al.* Does respiratory virus coinfection increases the clinical severity of acute respiratory infection among children infected with respiratory syncytial

- virus? *Pediatr. Infect. Dis. J.* **32**, 441–445 (2013).
35. Hasegawa, K. *et al.* Multicenter study of viral etiology and relapse in hospitalized children with bronchiolitis. *Pediatr. Infect. Dis. J.* **33**, 809–813 (2014).
 36. Arruda, E. *et al.* The burden of single virus and viral coinfections on severe lower respiratory tract infections among preterm infants a prospective birth cohort study in Brazil. *Pediatr. Infect. Dis. J.* **33**, 997–1003 (2014).
 37. Greer, R. M. *et al.* Do rhinoviruses reduce the probability of viral co-detection during acute respiratory tract infections? *J. Clin. Virol.* **45**, 10–15 (2009).
 38. Martin, E. T., Fairchok, M. P., Stednick, Z. J., Kuypers, J. & Englund, J. A. Epidemiology of multiple respiratory viruses in childcare attendees. *J. Infect. Dis.* **207**, 982–989 (2013).
 39. Bhattacharyya, S., Gesteland, P. H., Korgenski, K., Bjørnstad, O. N. & Adler, F. R. Cross-immunity between strains explains the dynamical pattern of paramyxoviruses. *Proc. Natl. Acad. Sci.* **112**, 13396–13400 (2015).
 40. Opatowski, L., Baguelin, M. & Eggo, R. M. Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling. *PLoS Pathog.* **14**, (2018).
 41. Upshur, R. E. G., Moineddin, R., Crighton, E. J. & Mamdani, M. Interactions of viral pathogens on hospital admissions for pneumonia, croup and chronic obstructive pulmonary diseases: results of a multivariate time-series analysis. *Epidemiol. Infect.* **134**, 1174–1178 (2006).
 42. van Asten, L. *et al.* Early occurrence of influenza A epidemics coincided with changes in occurrence of other respiratory virus infections. *Influenza Other Respi. Viruses* **10**, 14–26 (2016).
 43. Merler, S., Poletti, P., Ajelli, M., Caprile, B. & Manfredi, P. Coinfection can trigger multiple pandemic waves. *J. Theor. Biol.* **254**, 499–507 (2008).
 44. Velasco-Hernández, J. X., Núñez-López, M., Comas-García, A., Cherpitel, D. E. N. & Ocampo, M. C. Superinfection between Influenza and RSV alternating patterns in San Luis Potosí State, México. *PLoS One* **10**, (2015).
 45. González-Parra, G. *et al.* A comparison of RSV and influenza in vitro kinetic parameters reveals differences in infecting time. (2018).
doi:10.1371/journal.pone.0192645

46. González-Parra, G. & Dobrovolny, H. M. Assessing Uncertainty in A2 Respiratory Syncytial Virus Viral Dynamics. *Comput. Math. Methods Med.* **2015**, 1–9 (2015).
47. Kucharski, A. J., Andreasen, V. & Gog, J. R. Capturing the dynamics of pathogens with many strains. *J. Math. Biol.* 1–24 (2015). doi:10.1007/s00285-015-0873-4
48. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *epidemiology* **2**, 13 (2005).
49. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and sources of endemic human coronaviruses. in *Advances in virus research* **100**, 163–188 (Elsevier, 2018).
50. Nascimento-Carvalho, A. C. *et al.* Respiratory viruses among children with non-severe community-acquired pneumonia: A prospective cohort study. *J. Clin. Virol.* **105**, 77–83 (2018).
51. Davis, B. M. *et al.* Human coronaviruses and other respiratory infections in young adults on a university campus: prevalence, symptoms, and shedding. *Influenza Other Respi. Viruses* **12**, 582–590 (2018).
52. Varghese, L. *et al.* Epidemiology and clinical features of human coronaviruses in the pediatric population. *J. Pediatric Infect. Dis. Soc.* **7**, 151–158 (2017).
53. Gorse, G. J., Donovan, M. M., Patel, G. B., Balasubramanian, S. & Lusk, R. H. Coronavirus and other respiratory illnesses comparing older with young adults. *Am. J. Med.* **128**, 1251-e11 (2015).
54. Zhang, S. *et al.* Epidemiology characteristics of human coronaviruses in patients with respiratory infection symptoms and phylogenetic analysis of HCoV-OC43 during 2010-2015 in Guangzhou. *PLoS One* **13**, e0191789 (2018).
55. Killerby, M. E. *et al.* Human coronavirus circulation in the United States 2014–2017. *J. Clin. Virol.* **101**, 52–56 (2018).
56. Zeng, Z.-Q. *et al.* Epidemiology and clinical characteristics of human coronaviruses OC43, 229E, NL63, and HKU1: a study of hospitalized children with acute respiratory tract infection in Guangzhou, China. *Eur. J. Clin. Microbiol. Infect. Dis.* **37**, 363–369 (2018).
57. Kiyuka, P. K. *et al.* Human coronavirus NL63 molecular epidemiology and evolutionary patterns in rural coastal Kenya. *J. Infect. Dis.* **217**, 1728–1739

- (2018).
58. Golda, A. *et al.* Infection with human coronavirus NL63 enhances streptococcal adherence to epithelial cells. *J. Gen. Virol.* **92**, 1358–1368 (2011).
 59. Munywoki, P. K. *et al.* The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya. *J. Infect. Dis.* **209**, 1685–1692 (2014).
 60. Munywoki, P. K. Transmission of Respiratory Syncytial Virus in Households : Who Acquires Infection From Whom. (Open University UK, 2013).
 61. Munywoki, P. K. *et al.* Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiol. Infect.* **143**, 804–12 (2015).
 62. Munywoki, P. K. *et al.* Frequent Asymptomatic Respiratory Syncytial Virus Infections During an Epidemic in a Rural Kenyan Household Cohort. *J. Infect. Dis.* 1–8 (2015). doi:10.1093/infdis/jiv263
 63. Wathuo, M., Medley, G. F., Nokes, D. J. & Munywoki, P. K. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res.* **1**, 27 (2016).
 64. Kombe, I. K., Munywoki, P. K., Baguelin, M., Nokes, D. J. & Medley, G. F. Model-based estimates of transmission of respiratory syncytial virus within households. *Epidemics* **27**, 1–11 (2019).
 65. Lessler, J. *et al.* Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet Infectious Diseases* **9**, 291–300 (2009).
 66. Roberts, G. O. & Rosenthal, J. S. Examples of Adaptive MCMC. *J. Comput. Graph. Stat.* **18**, 349–367 (2009).
 67. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
 68. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations)? *J. Comput. Graph. Stat.* **7**, 434–455 (1998).
 69. Mina, M. J., Metcalf, C. J. E., De Swart, R. L., Osterhaus, A. D. M. E. & Grenfell, B. T. Long-term measles-induced immunomodulation increases overall childhood infectious disease mortality. *Science (80-.)*. **348**, 694–699 (2015).
 70. Rohani, P., Green, C. J., Mantilla-Beniers, N. B. & Grenfell, B. T. Ecological interference between fatal diseases. *Nature* **422**, 885–888 (2003).

71. Pitzer, V. E. *et al.* Modeling rotavirus strain dynamics in developed countries to understand the potential impact of vaccination on genotype distributions. *Proc. Natl. Acad. Sci.* **108**, 19353–19358 (2011).
72. Wikramaratna, P. S. *et al.* Five challenges in modelling interacting strain dynamics. *Epidemics* **10**, 31–34 (2015).
73. Kinyanjui, T. M. *et al.* Vaccine Induced Herd Immunity for Control of Respiratory Syncytial Virus Disease in a Low-Income Country Setting. *PLoS One* **10**, e0138018 (2015).
74. Kamau, E. *et al.* An intensive, active surveillance reveals continuous invasion and high diversity of rhinovirus in households. *J. Infect. Dis.* **219**, 1049–1057 (2019).

6. Discussion

Recognizing the importance of understanding the transmission dynamics of RSV, this PhD project was conceptualized with an aim to use an integrated data analysis to identify transmission chains within a household setting. The idea was to use all the data available on the shedding episodes and host social-demographic factors within a single modelling framework. A key determinant of the methodology that could be applied was the availability of genetic data. Naturally, this called for an investigation into the field of phylodynamics^{1,2}. A review of the methods available revealed that applying any of the pre-existing methods that simultaneously infer ecological and evolutionary dynamics would be challenging due to the nature of the data. The genetic sequences were available at about half the sampling density of the pathogen-positive samples. It was therefore decided that a two-part analysis of the data would be more suitable beginning with the epidemiological data, then extending the model to include genetic information, similar to previous work^{3,4}.

The epidemiological data, consisting of timings of positive samples, viral load and symptom status, was modelled using a dynamic transmission model calibrated at the individual host level and time in days. RSV cases were identified by group either as RSV A or RSV B. The results of this primary analysis revealed that during the course of a single epidemic, individuals acquire partial immunity that is stronger against homologous group re-infection than heterologous. The existence of re-infections within the same epidemic is evidence that even in the short-term, immunity to RSV infection is not complete and this incompleteness was quantified as a 47% (95% CrI: 17%-68%) reduction in susceptibility to homologous re-infection and 39% (95%CrI: -8%-69%) reduction to heterologous. An effect of increasing age on susceptibility was also inferred. Older individuals were less susceptible to RSV infection. This could be indicative of a lifelong partial immunity that builds up with repeated exposure, a mechanism previously explored by Weber *et al*⁵ and Kinyanjui *et al*⁶. However, it was assumed that exposure and hence infectious contacts occurred homogeneously with respect to age. This is not necessarily the case, as a recent study on household contact revealed⁷, as such the estimated age effects on susceptibility could include effects of age related contacts within the household. Nonetheless, these results imply that there are short-term and long-term immune dynamics against RSV. If vaccinations are timed

according to the seasonal patterns, models assessing the impact of such strategies will need to account for both the short and long-term dynamics. Individuals who were shedding large quantities of virus and had ARI symptoms, characteristics which have been found to be correlated⁸, were estimated as being more infectious than their asymptomatic low viral load counterparts. Simulations were run to assess the impact of a vaccine that worked by eliminating symptoms. These showed that such a vaccine would reduce the projected number of cases significantly. It has previously been shown that a vaccine that worked by reducing infectiousness and duration of shedding would have the highest impact⁹. Two unique findings came out of this analysis: firstly, households of less than 8 occupants were inferred as having an increased pair-wise risk of transmission; secondly, there was evidence that RSV A had a slight transmission advantage over RSV B in the household. The first observation could be the result of smaller households, as defined in this setting, being more likely to have fewer structures and as such members come into contact more freely, and therefore are more likely to infect each other. If so, then this means that social structuring needs to be considered when modelling transmission, or at a minimum, population density. High population density has already been linked to the generation of new RSV viral variants¹⁰ and household structure has been included in a dynamic model used to explore a joint maternal and cocoon vaccination strategy¹¹. The second observation is not completely new; other studies have found evidence that RSV A is more transmissible than RSV B^{12,13}, perhaps providing an explanation for RSV A dominating most outbreaks. Finding some evidence that RSV A and RSV B might have different transmission niches could point more to the ecology of the two groups and the mechanisms by which they manage to co-exist.

The genetic information was included into the model by allowing further classification of the infecting virus into genetic clusters. These clusters were derived from a separate phylogenetic analysis¹⁴ and were used to create mutually exclusive pathogen identities. The model and inference technique were then adapted to include the genetic cluster information where available and infer it where it was not. Increased resolution in pathogen identification did not result in significant shifts in the distribution of estimated parameters, however, the weak signal of there being separate transmission niches for RSV A and B transmission inferred through differences

in the distribution of the group-specific transmission coefficients was lost. In inferring the transmission networks, which were refined by the pathogen resolution, I did find that despite RSV B being the dominant pathogen in circulation, RSV A was more successfully transmitted to infants in the household. It was also shown that when infants were infected within the household, it was by a child under the age of 13 years which is supported by a household contact study that showed that children 6-14 years old frequently contacted children 0-5 years old in the same household⁷. Infants were also frequently found to be index cases in household outbreaks, this result taken together with observations from the previously mentioned household contact study that found children 0-5 years old were frequently contacted by adults outside of their households⁷, could point to childcare practices that are important in determining the source of infant infections. Since the analysis by White *et al*¹⁵, there aren't other publications that look into quantifying the interaction between RSV groups and inferring what such interactions mean for transmission dynamics. Such studies could aid in gaining a better understanding of why the RSV groups manage to co-exist with RSV A being the more dominant pathogen. Why does RSV A not replace RSV B? Interactions between the RSV groups could be ecological and/or immunological, and disentangling these effects would result in a better understanding of transmission drivers thereby allow more effective control¹⁶.

Following the results above, there was increased interest in identifying if RSV interactions within the household also occur with other pathogens. The model was adapted to fit data from two different pathogens, RSV and human coronavirus (HCoV), each identified at the group level. Coronavirus was classified as corona-229E, corona-NL63 or corona-OC43. Given how frequently individuals in the study were infected with HCoVs, it was surprising that the frequency of RSV-HCoV co-infections was not more common than what was observed. Whether this frequency was less than what would be expected to occur if the pathogens were independent was investigated by estimating if susceptibility to RSV infection is modified if one is currently shedding HCoV, and *vice versa*. Whether susceptibility was altered depended on the level of pathogen identification. If all RSV shedding episodes were treated homogeneously and all the HCoV episodes were also treated homogeneously, it was inferred that susceptibility was not altered; implying the lack of frequent co-infections was to be

expected. However, with pathogen identification at the group level, there was a combination of effects, with some pathogen group pairs showing reduced susceptibility and others showing increased. A similar effect was observed when attempts to infer if susceptibility to one pathogen was altered by previous infection to another. Interactions with RSV A seemed to result in reduced susceptibility while interactions with RSV B resulted in increased susceptibility. These observations could form part of the explanation as to why epidemiological studies investigating the effect of viral co-infections find conflicting interactions, with some reporting increased disease risk¹⁷⁻¹⁹ and other not²⁰⁻²². However, these results should be interpreted with caution. A strong assumption was made by treating the interactions as symmetric, i.e. the effect of say previous RSV A infection on susceptibility to HCoV-OC43 was the same as the effect of previous HCoV-OC43 infection on susceptibility to RSV A. This assumption was made to reduce the dimensions of the parameters being estimated for computational efficiency and ease of interpretation. However, the results suggest that this should be challenged, and asymmetric relationships explored. Asymmetric fitting is beyond the scope of this thesis, however since the results of the multi-pathogen fitting are intended for publication, this will be done prior to submitting to a journal. The analysis we presented in Chapter 5 only begins to scratch the surface of possible interaction mechanisms that could be driving patterns of pathogen transmission observed at the host level. A majority of models treat pathogens as existing independently, an assumption which has already been challenged by studies showing RSV replacement as a disease causing agent in the face of prophylactic treatment²³, competitive interactions between RSV and influenza^{24,25}, competitive interactions between RSV and human metapneumovirus (HMPV)^{16,26}, and facilitative interactions between viruses and bacteria²⁷⁻²⁹. It is increasingly crucial for investigators to begin to consider multi-pathogen interactions even if the focus is just on one particular pathogen³⁰. The challenge for such studies will be in determining the level and extent of these interactions. Models would become increasingly intractable with increase in the number of pathogens being considered. The resolution with which these pathogens are identified will also play a role in the consistency of inferred interactions. All these factors mean that a lot of data is required at different potential levels of interactions, within the host, individual host level and between host level.

The table below gives a summary of the three versions of the individual level dynamic model used in Chapter 3, 4 and 5.

Table 6. 1: A summary of the three variants of the individual level model used to investigate transmission dynamics of RSV.

The models are named according to the chapter in which they were presented.

	Chapter 3	Chapter 4	Chapter 5
Data	<ul style="list-style-type: none"> • Social-temporal RSV shedding patterns where shedding episodes are identified by RSV group. • Individual’s demographics e.g. age 	<ul style="list-style-type: none"> • Social-temporal RSV shedding patterns. • Individual’s demographics e.g. age • Viral genetic sequence data in the form of genetic clusters used to further classify shedding episodes 	Social-temporal RSV and hCoV shedding patterns
Analysis objectives	To define transmission patterns for RSV	<ul style="list-style-type: none"> • To define transmission patterns for RSV • To identify transmission chains and source of infant infections 	To investigate possible pathogen interactions between RSV and hCoV at the individual host level

Assumptions

- Susceptible-Exposed-Infected-Susceptible (SEIS₂) RSV natural history
- Infection from household and external unknown sources (community exposure)
- Group-specific community exposure to infection can be represented by a bell-shaped curve estimated from household level incidence
- Age, household size, viral load and symptom might affect transmission
- Latency period ranges between 2-5 days
- Transmission between households in the study not explicitly modelled

- Susceptible-Exposed-Infected-Susceptible (SEIS₂) RSV natural history
- Infection from household and external unknown sources (community exposure)
- Cluster-specific community exposure to infection can be represented by adding up an exponential function relating the rate of exposure to time since onset in every case.
- Age, household size, viral load and symptom might affect transmission
- Latency period ranges between 2-5 days
- Possible transmission between households in the study

- Susceptible-Exposed-Infected-Susceptible (SEIS₂) RSV and hCoV natural history
- Infection from household and external unknown sources (community exposure)
- pathogen-specific community exposure to infection can be represented by adding up an exponential function relating the rate of exposure to time since onset in every case.
- Pathogen-specific latency durations gamma distributed with mean and SD estimated from the data
- Transmission between households in the study not explicitly modelled

Findings

- Smaller households have a higher pairwise rate of exposure.
 - Increasing age estimated to reduce susceptibility to infection.
 - RSV confers partial short-term immunity more so against homologous group re-infections.
 - A vaccine that works to eliminate symptoms would have an impact on overall transmission
 - Estimates of the baseline rates of exposure within the household and at the community level for RSV A and B suggest a possible transmission niche for RSV A within the household.
 - Estimated 40-59% of RSV A and 26-48% of RSV B cases occurred in the HH
- Smaller households have a higher pairwise rate of exposure.
 - As with model in Chapter 3, age estimated to affect susceptibility and RSV infection estimated to confer short-term immunity
 - Inclusion of genetic data in the model resulted in slight shifts in the distributions of the baseline rates of exposure for RSV A and B, resulting in the evidence of a transmission niche from the previous model being lost
 - Increased precision in infection source attribution, estimated 60% of RSV A cases from the HH, while 52% of RSV B were from the HH.
- Pathogen interactions become cleared with increased resolution of pathogen identification, which could explain conflicting evidence of how RSV interacts with other pathogens from other studies.
 - RSV B and hCoV OC43 estimated to have a facilitative interaction where previous infection with one increases susceptibility to the other.

		<ul style="list-style-type: none"> • Over half of infant RSV A infections contracted within the household, less so for RSV B • Where infant infections occurred in the household, often the source of infection was a child between the ages of 2 and 13. • Transmission between households in the study unlikely to have occurred. 	
Limitations	<ul style="list-style-type: none"> • Small sample size • Sampling frequency means short duration episodes might have been missed 	Used two-step approach in data integration which could introduce inconsistency in inferred dynamics	<ul style="list-style-type: none"> • Assumed symmetry in pathogen interactions • Only used data from 2 pathogens • Data represents short temporal window
Recommendations	Inclusion of other data types such as genetic data to further elucidate	<ul style="list-style-type: none"> • Targeting school-aged children for vaccination would result in an 	<ul style="list-style-type: none"> • Pathogen interactions should not be ignored if we are to fully

	<p>transmission chains and clarify hypothesis of transmission niche.</p>	<p>indirect protective effect on the infant</p> <ul style="list-style-type: none">• Though genetic data did not lead to a drastic change in the inferred transmission dynamics, its utility should not be ruled out in future studies conducted at a broader temporal and geographical scale• Inferred interactions between RSV A and B, and differences in transmission such as the observation that more RSV A infections occur in the household relative to RSV, warrant further investigations.	<p>understand their pathogen transmission dynamics.</p> <ul style="list-style-type: none">• Evidence of an interaction between RSV and hCoV warrants further investigation
--	--	--	--

In all the three main analyses presented in this thesis, it was considered that infection can occur within a household, through contact with an infectious household cohabitant, or outside the household (community level). The possibility of between household transmission as a source of community infection was considered in the model in Chapter 4, however there were numerous households in the study area that were not recruited, making the possibility of direct transmission occurring between the few that were sampled unlikely. It was therefore necessary to find a way to account for community exposure in order to allow introductions into the households. Two different approaches to account for this were implemented. In Chapter 3, a bell-shaped incidence curve was fitted to primary onsets of household outbreaks and used as a proxy for the background community rate. This background function was derived at the RSV group level. In Chapter 4, it was no longer feasible to use this function with the identification of genetic clusters. This was because some clusters had such a low representation in the population of infected households that the background curves derived in this way took unexpected shapes. As such, a new function form was adopted that was based on the timings of cases infected with a particular cluster. The 'signal' of the cluster in the broader community was assumed to wane exponential from the time of onset in an individual. Signals from all the individuals with onsets to a particular cluster were added up to give the total background community function. Though this formulation worked, it requires further validation. Accounting for exposure from sources outside of the household is crucial, if one community exposure can lead to a household member getting infected, it would be wrong to assume that this external exposure is not competing with exposure at the household level once infection is introduced. However, throughout all the analyses, the inferred pair-wise rate of within household exposure was much higher than the community rate of exposure.

Depending on the definition of RSV disease, the efficacy and effectiveness of preventive measures may vary and be affected by population characteristics (genetic or otherwise) and circulation patterns^{31,32}. This necessitates the understanding of RSV transmission dynamics at geographical scales that have generalizable seasonality, population demographics and infrastructure for implementation of vaccine policy. Kenya is a lower middle-income country in the tropical eastern coast of the African

continent. As with other tropical locations, the seasonality drivers of RSV are not well understood, however, the burden of disease due to RSV has been shown to be significant³³⁻³⁸, making a vaccine against RSV a subject of national interest. Though the results of this study are based on a small number of individuals, the characterization of the ecology of RSV calls for further investigations to determine the role of pathogen interactions and social-demographic characteristics in driving the observed viral seasonality patterns. Would an RSV vaccine need to be such that it prevents disease without disrupting the viral ecology so as not to elicit pathogen replacement?

All the inference made in this work is based on data collected from a small fraction of the population. Due to the intense sampling of the individuals in the study, it was not feasible to extend follow-up to large groups of individuals due to constraints, the least of which is logistics. The results presented here must therefore be taken with the knowledge that it was not possible to characterize what was not observed, and transmission in a majority of the local area was largely unobserved. In addition to the data being limited in geographical scale, there are also temporal limits to bear in mind. Data collection only covered six months of the year, meaning seasonality of viruses such as HMPV were missed, (which could also be another indicator of competition between RSV and HMPV). The strengths of such study designs are being able to observe infection dynamics at the individual host level, picking up on repeat infections, co-infections, asymptomatic infections and variations in the infecting pathogen. On the opposite end of the spectrum are studies such as those conducted by Li and colleagues that collated data from different sources to come up with a global picture of pathogen dynamics spanning several years²⁶. Such studies are useful in being able to compare and contrast seasonality and therefore infer potential drivers, however, it is too coarse to infer factors such as the role of repeat infections. Depending on the purpose of a study, a balance must be found between the number of samples and the information content of each sample.

In the analyses presented in this thesis, individual level dynamics were used to predict population level transmission dynamics, albeit in a small population. Due to the size of the population, it was possible to use individual-level mathematical models. Such models can become increasingly complicated and computationally intensive when they

are used to represent large heterogeneous populations. At which point a popular alternative is compartmental models that group hosts into several states. As a middle ground, multi-scale models are used to link individual level dynamics such as variation in infectiousness based on age of infection, to populations level transmission dynamics³⁹. As more data on the characteristics of individual infections are becoming available, especially genetic data, such multi-scale models should concurrently increase in frequency. This naturally might mean an increase in complexity of the inference technique, requiring the use of advanced techniques such as particle filter MCMC⁴⁰.

This PhD project was conceptualized with a broad aim of gaining a better understanding of RSV transmission dynamics by interrogating different data types collected from a longitudinal household study. Specifically, I had four main objectives. The first was simply to conduct a review of the literature and identify the best way to integrate different data types into a single modelling framework. This review, presented in Chapter 2, was successful in identifying that the choice of method should be data driven, thus given the unique nature of the data, it was decided that the model will be focused on representing short-term (6 months) infection dynamics at the individual host. The model also focused on the most abundant type of data, social-temporal shedding patterns, and used any other data types as enhancements. The second objective was to use all the available genetic data and epidemiological data to infer transmission dynamics and transmission chains within and possibly between households. The model was built-up in stages, first using the epidemiological (social-temporal) data to infer transmission characteristics and then extending the model to include genetic data as a way to further clarify transmission clusters. These two versions of the model are presented in Chapter 3 and 4 respectively. Including genetic data in the model did not result in a drastic change in the inferred dynamics, possibly due to the study design, as discussed in Chapter 4. Despite the hypothesis of a difference in transmission niche between RSV A and B inferred in Chapter 3 from the distribution of parameters being nullified in Chapter 4, there was consistency in the fact that more of the RSV A cases were attributed to within household transmission than RSV B. The evidence for the existence of a transmission niche might not be clear, but the differences inferred between the two groups warrant further investigation.

The third objective was to identify the added benefit of viral genetic sequence data and provide advice on data collection for future studies. The use of genetic clusters in the model increased the certainty with which transmission clusters were inferred, therefore allowing more precision in some of the epidemiological parameters. However, given the relatively smaller sampling density of the genetic data compared to the social-temporal data, the utility of the former was not greatly observed in the model inference. However, I do not dismiss how informative viral genetic sequences can be and argue that at a larger temporal and geographical scale the insights gained from integrating genetic data with epidemiological data will be much more impactful. Finally, the integrated data framework was to be used to explore vaccination strategies and give information on target population, timing and frequency. Though I did not explicitly model vaccination, I did infer that targeting school going children would lead to indirect protection of the infant and that a vaccine would have an overall effect on transmission even if it only worked to eliminate symptomatic infections. The inference from the analyses presented in this thesis could be used in multi-scale model of vaccination that aims to translate individual level dynamics onto population level effects. Though it was not part of the main objectives, Chapter 5 presents the results of extending the model in Chapter 3 to fit data from multiple pathogens. The need to do so arose after the observations of the interactions between RSV A and B. Though the analysis had numerous simplifying assumptions, it was able to highlight the need to not only consider group/species interactions of the same pathogen, but also between pathogen interactions as they could have an impact on how effective and intervention will be.

In conclusion, this thesis presents the progression of a data driven analysis that began with an aim of simply inferring transmission dynamics through integrating genetic sequence data and epidemiological data. Though this target was met, inference on the ecological dynamics of RSV groups and RSV with other pathogens was made. Evidently, there is an interaction between the two RSV groups and possible differences in transmission propensity within the household that require further investigation. Signals of possible multi-pathogen interactions also warrant further investigations and should serve as a precaution for future studies that treat RSV as a homogenous independent pathogen. During vaccine trials, samples should be taken to consider the

potential enhanced or reduced impact of a reduction in RSV transmission. Extensions of this analysis are possible on two fronts. First, integration of epidemiological and genetic data could be done at larger temporal and geographical scales to reveal interactions between ecological and evolutionary dynamics. The approach taken here was two-staged, avoiding simultaneous inference of the epidemiological and genetic models for computational reasons and due to restrictions in the data brought about by differences in sampling densities. However, much insight can be gained through simultaneous phylodynamic inference. Second, future models of RSV could aim to be a combination of inference on the longitudinal short-term host dynamics, such as the kind inferred here in Chapter 4, and population level long-term dynamics such as the kind inferred by Kinyanjui et al⁶. Such models would not only be able to make better inference on interactions between RSV groups, but also interactions between different pathogens, and in doing so, be better placed to make predictions on the impact of an intervention strategy against one or more pathogens³⁰.

References:

1. Grenfell, B. T. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science (80-.)*. **303**, 327–332 (2004).
2. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
3. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
4. Cottam, E. M. *et al.* Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B Biol. Sci.* **275**, 887–895 (2008).
5. Weber, A., Weber, M. & Milligan, P. Modeling epidemics caused by respiratory syncytial virus (RSV). **172**, (2001).
6. Kinyanjui, T. M. *et al.* Vaccine Induced Herd Immunity for Control of Respiratory Syncytial Virus Disease in a Low-Income Country Setting. *PLoS One* **10**, e0138018 (2015).
7. Kiti, M. C. *et al.* Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. *EPJ Data Sci.* **5**, 21 (2016).
8. Wathuo, M., Medley, G. F., Nokes, D. J. & Munywoki, P. K. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res.* **1**, 27 (2017).
9. Pan-Ngum, W. *et al.* Predicting the relative impacts of maternal and neonatal respiratory syncytial virus (RSV) vaccine target product profiles: A consensus modelling approach. *Vaccine* **35**, 403–409 (2017).
10. Chan, C. H. S., Sanders, L. P. & Tanaka, M. M. Modelling the role of immunity in reversion of viral antigenic sites. *J. Theor. Biol.* **392**, 23–34 (2016).
11. Brand, S. P. C., Munywoki, P., Walumbe, D., Keeling, M. J. & Nokes, D. J. Reducing RSV hospitalisation in a lower-income country by vaccinating mothers-to-be and their households. *bioRxiv* 1–21 (2019). doi:10.1101/569335
12. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol. Infect.*

- 133**, 279–289 (2005).
13. Otomaru, H. *et al.* Transmission of Respiratory Syncytial Virus Among Children Under 5 Years in Households of Rural Communities , the Philippines. 1–8 (2016). doi:10.1093/ofid/ofz045
 14. Agoti, C. N. *et al.* Genomic analysis of respiratory syncytial virus infections in households and utility in inferring who infects the infant. *Sci. Rep.* **9**, 10076 (2019).
 15. White, L. J., Waris, M., Cane, P. A., Nokes, D. J. & Medley, G. F. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *epidemiology* **2**, 13 (2005).
 16. Bhattacharyya, S., Gesteland, P. H., Korgenski, K., Bjørnstad, O. N. & Adler, F. R. Cross-immunity between strains explains the dynamical pattern of paramyxoviruses. *Proc. Natl. Acad. Sci.* **112**, 13396–13400 (2015).
 17. Mazur, N. I. *et al.* Severity of respiratory syncytial virus lower respiratory tract infection with viral coinfection in HIV-uninfected children. *Clin. Infect. Dis.* **64**, 443–450 (2017).
 18. Asner, S. A., Rose, W., Petrich, A., Richardson, S. & Tran, D. J. Is virus coinfection a predictor of severity in children with viral respiratory infections? *Clin. Microbiol. Infect.* **21**, 264.e1-264.e6 (2015).
 19. Marcone, D. N. *et al.* Viral etiology of acute respiratory infections in hospitalized and outpatient children in Buenos Aires, Argentina. *Pediatr. Infect. Dis. J.* **32**, e105-10 (2013).
 20. Petrarca, L. *et al.* Acute bronchiolitis: Influence of viral co-infection in infants hospitalized over 12 consecutive epidemic seasons. *J. Med. Virol.* **90**, 631–638 (2018).
 21. Ljubin-Sternak, S. *et al.* Etiology and Clinical Characteristics of Single and Multiple Respiratory Virus Infections Diagnosed in Croatian Children in Two Respiratory Seasons. *J. Pathog.* **2016**, 1–8 (2016).
 22. Martin, E. T., Kuypers, J., Wald, A. & Englund, J. A. Multiple versus single virus respiratory infections: Viral load and clinical disease severity in hospitalized children. *Influenza Other Respi. Viruses* **6**, 71–77 (2012).
 23. Blanken, M. O. *et al.* Respiratory Syncytial Virus and Recurrent Wheeze in

- Healthy Preterm Infants. *N. Engl. J. Med.* **368**, 1791–1799 (2013).
24. González-Parra, G. *et al.* A comparison of RSV and influenza in vitro kinetic parameters reveals differences in infecting time. (2018).
doi:10.1371/journal.pone.0192645
 25. González-Parra, G. & Dobrovolny, H. M. Assessing Uncertainty in A2 Respiratory Syncytial Virus Viral Dynamics. *Comput. Math. Methods Med.* **2015**, 1–9 (2015).
 26. Li, Y. *et al.* Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: a systematic analysis. *Lancet Glob. Heal.* **7**, e1031–e1045 (2019).
 27. Sande, C. J. *et al.* Airway response to respiratory syncytial virus has incidental antibacterial effects. *Nat. Commun.* **10**, 1–11 (2019).
 28. Avadhanula, V. *et al.* Respiratory Viruses Augment the Adhesion of Bacterial Pathogens to Respiratory Epithelium in a Viral Species-and Cell Type-Dependent Manner Downloaded from. *J. Virol.* **80**, 1629–1636 (2006).
 29. Weinberger, D. M. *et al.* Seasonal drivers of pneumococcal disease incidence: Impact of bacterial carriage and viral activity. *Clin. Infect. Dis.* **58**, 188–194 (2014).
 30. Opatowski, L., Baguelin, M. & Eggo, R. M. Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling. *PLoS Pathog.* **14**, 1–28 (2018).
 31. Aranda, S. S. & Polack, F. P. Prevention of Pediatric Respiratory Syncytial Virus Lower Respiratory Tract Illness: Perspectives for the Next Decade. *Front. Immunol.* **10**, 1006 (2019).
 32. Janet, S., Broad, J. & Snape, M. D. Respiratory syncytial virus seasonality and its implications on prevention strategies. *Human Vaccines and Immunotherapeutics* **14**, 234–244 (2018).
 33. Nyawanda, B. O. *et al.* Evaluation of case definitions to detect respiratory syncytial virus infection in hospitalized children below 5 years in Rural Western Kenya, 2009–2013. *BMC Infect. Dis.* **16**, 218 (2016).
 34. Emukule, G. O. *et al.* The burden of influenza and RSV among inpatients and outpatients in rural western Kenya, 2009–2012. *PLoS One* **9**, e105543 (2014).
 35. Breiman, R. F. *et al.* Severe acute respiratory infection in children in a densely

- populated urban slum in Kenya, 2007–2011. *BMC Infect. Dis.* **15**, 95 (2015).
36. Okiro, E. A., Ngama, M., Bett, A. & Nokes, D. J. The incidence and clinical burden of respiratory syncytial virus disease identified through hospital outpatient presentations in Kenyan children. *PLoS One* **7**, e52520 (2012).
 37. Nokes, D. J. *et al.* Respiratory syncytial virus infection and disease in infants and young children observed from birth in Kilifi District, Kenya. *Clin. Infect. Dis.* **46**, 50–57 (2008).
 38. Nokes, D. J. *et al.* Incidence and severity of respiratory syncytial virus pneumonia in rural Kenyan children identified through hospital surveillance. *Clin. Infect. Dis.* **49**, 1341–1349 (2009).
 39. Nikin-Beers, R., Blackwood, J. C., Childs, L. M. & Ciupe, S. M. Unraveling within-host signatures of dengue infection at the population level. *J. Theor. Biol.* **446**, 79–86 (2018).
 40. Andrieu, C., Doucet, A. & Holenstein, R. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 269–342 (2010).

Appendices

A1: Ethical approval

1. KEMRI (SERU) ethical approval



KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel: (254) (020) 2722541, 2713349, 0722-205901, 0733-400003, Fax: (254) (020) 2720030
E-mail: director@kemri.org, info@kemri.org, Website. www.kemri.org

KEMRI/RES/7/3/1

November 7, 2016

TO: DR. PATRICK MUNYWOKI,
PRINCIPAL INVESTIGATOR

THROUGH: DR. BENJAMIN TSOFA,
DIRECTOR, CGMR-C,
KILIFI

Dear Sir,

RE: **SSC PROTOCOL NO. 1651 (RESUBMITTED REQUEST FOR AMENDMENT 2):
HOUSEHOLD TRANSMISSION OF RESPIRATORY SYNCYTIAL VIRUS (RSV): WHO
ACQUIRES INFECTION FROM WHOM?**

Reference is made to your letter dated October 25, 2016. KEMRI/Scientific and Ethics Review Unit (SERU) acknowledges receipt of the revised document on October 27, 2016.

The Committee acknowledges receipt of the following documents:

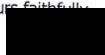
1. Clean copy of the amended version of the protocol (version 2.1_170816)
2. SERU letter dated October 17, 2016

This is to inform you that the Committee determines that the issues raised at the 256th Committee A, meeting of the KEMRI/Ethics Review Committee held on October 11, 2016 are adequately addressed. You are therefore **authorized** to implement the Amendments accordingly:

1. Administrative changes on co-investigators listing:
 - a. An update of Dr. Paul Kellam's institutional affiliation to include Imperial College, UK.
 - b. That, two investigators, Drs. Clayton Onyango and Nelson Onyango have stepped down as co-investigators from the study.
2. In page 10 of the protocol, insertion of the phrase "or samples will be sequenced as a service provision at one of our collaborating sites in the UK, Universities of Warwick or Imperial Health Protection Agency".
3. Page 14; the timelines for data analysis and final report writing have been revised to: "August 2010 to December 2020."

Please note that you are responsible for submitting any further changes to the approved version of the study protocol to SERU for review and the changes should not be initiated until written approval from the SERU is received.

Yours faithfully,


FOR: DR. EVANS AMUKOYE,
ACTING HEAD,
KEMRI SCIENTIFIC AND ETHICS REVIEW UNIT



In Search of Better Health

2. LSHTM ethical approval

London School of Hygiene & Tropical Medicine

Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk



Observational / Interventions Research Ethics Committee

Ms. Ivy Kombe
LSHTM

11 September 2017

Dear Ivy,

Study Title: Integrating viral RNA sequence and epidemiological data to define transmission patterns for respiratory syncytial virus (RSV)

LSHTM ethics ref: 14209

Thank you for your application for the above research, which has now been considered by the Observational Committee.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Protocol / Proposal	Proposal	28/06/2017	1
Investigator CV	CV-Ivy_K_Kombe	05/07/2017	1
Investigator CV	CV_James_Nokes	05/07/2017	1
Investigator CV	CV_Marc_Baguelin	05/07/2017	1
Local Approval	Amended_protocol_2016	05/07/2017	1
Local Approval	BREC_Approval_2009	05/07/2017	1
Local Approval	ERC_Approval_2009	05/07/2017	1
Local Approval	SERU_Approval_2015	05/07/2017	1
Local Approval	SERU_Approval_2016	05/07/2017	1
Local Approval	Original_protocol_2009	06/07/2017	1
Investigator CV	CV_Graham_Medley	17/07/2017	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: <http://leo.lshtm.ac.uk>

Additional information is available at: www.lshtm.ac.uk/ethics

Yours sincerely,



Professor John DH Porter
Chair

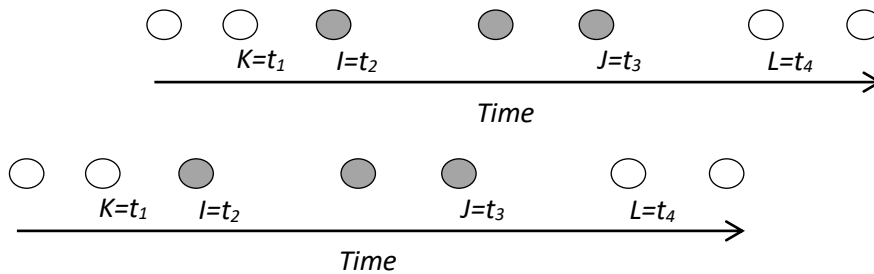
ethics@lshtm.ac.uk
<http://www.lshtm.ac.uk/ethics/>

Improving health worldwide

A2: Supplementary appendix for Paper 1.

A2.1. Imputing shedding durations, symptomatic episodes and viral loads

An RSV A/B shedding episode is defined as a period within which an individual provided PCR positive samples for RSV A/B that were no more than 14 days apart. Sampling of the study population was done in intervals, as such, complete shedding episodes had to be imputed using the mid-point method described. Shedding was assumed to start mid-way between the last negative sample and the first positive sample, and it ended midway between the last positive sample and the first negative sample of an episode. This is illustrated below:



Filled circles are positive samples in a single episode, empty circle are negative. t_1 , t_2 , t_3 and t_4 are dates of sample collection.

For $(t_4 - t_3)$ and $(t_2 - t_1) \leq 7$ days

$$\text{Duration} = \left[t_3 + \left(\frac{t_4 - t_3}{2} \right) \right] - \left[t_2 - \left(\frac{t_2 - t_1}{2} \right) \right]$$

For $(t_4 - t_3) > 7$

$$\text{Duration} = \left[t_3 + \left(\frac{x}{2} \right) \right] - \left[t_2 - \left(\frac{t_2 - t_1}{2} \right) \right] : \text{Right censoring}$$

For $(t_2 - t_1) > 7$

$$\text{Duration} = \left[t_3 + \left(\frac{t_4 - t_3}{2} \right) \right] - \left[t_2 + \left(\frac{x}{2} \right) \right] : \text{Left censoring}$$

Where x = mean of sampling intervals for samples in an episode, which was found to be 3.45 days.

Any negative samples ($C_t > 35$ or $C_t = 0$) in between a shedding episode were ignored, i.e. were not treated like true end of shedding. Figure A2. 1 shows the distribution of imputed shedding durations for RSV A and RSV B episodes.

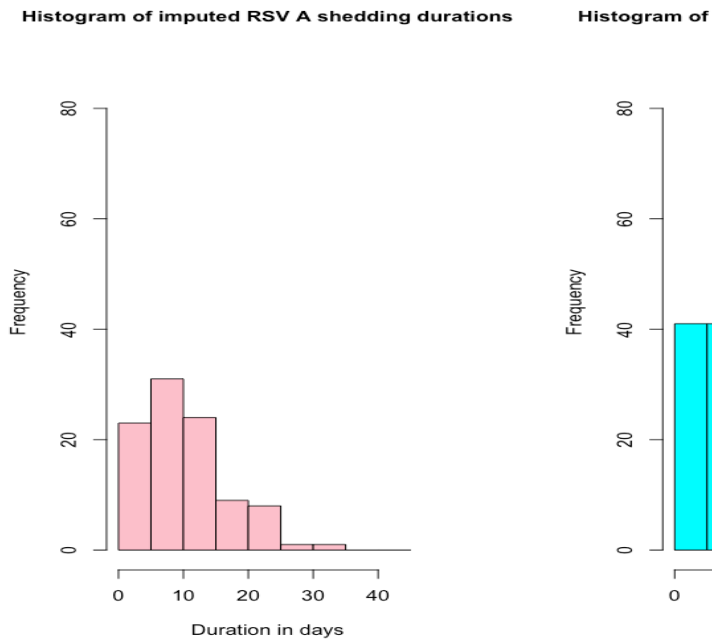


Figure A2. 1: Distributions of imputed shedding durations for RSV A (left) and RSV (right)

In order to include information of the amount of virus shed by an infected person into the transmission model, the Ct value need to be converted to \log_{10} RNA copy number which is a more direct measure of viral load. The formula used to convert Ct values to their \log_{10} RNA equivalent was $y = -3.308x + 42.9$, where $y = \text{Ct values}$ and $x = \log_{10}$ RNA copy number[1,2].

Following conversion of the PCR Ct values to viral load, we proceeded to interpolate the viral loads for days in an episode that did not have data. Linear interpolation was used for all the shedding episodes. It was assumed that the starting and ending sample, if data was missing, had a viral load of 2.388 \log_{10} RNA (baseline positive Ct value converted to viral load). For two samples of viral load V_a and V_b at times t_a and t_b , $t_b > t_a$, the gap in between is filled out as follows:

For $t_b - t_a = n$, viral load V_j at time point t_j for $j=1 \dots (n-1)$ is given by

$$V_j = V_a + \frac{j(V_b - V_a)}{n}$$

Viral loads lower than 2.388 \log_{10} RNA in between an episode were not included in the interpolation. Figure A2. 2 shows histograms of interpolated viral loads for RSV A and RSV B.

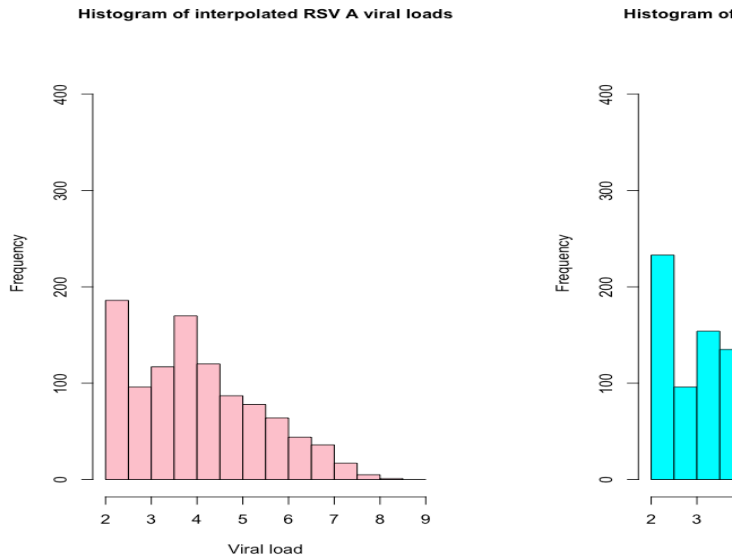
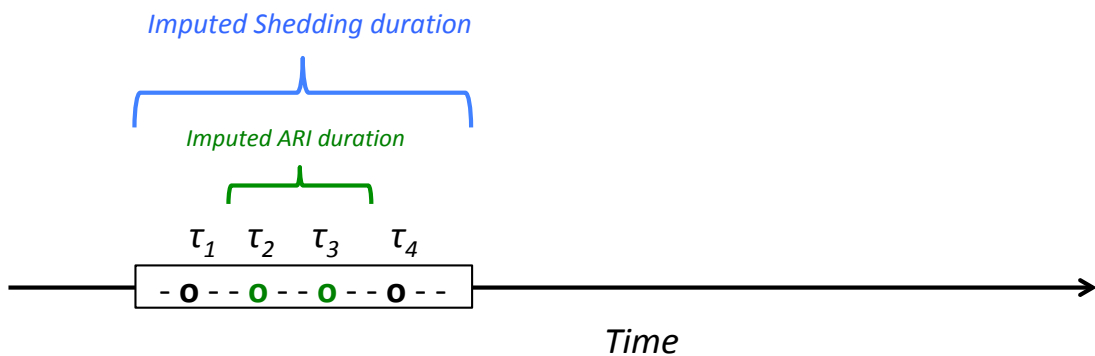


Figure A2. 2: Histograms of interpolated viral loads for RSV A (left) and RSV B (right).

We define symptomatic as having an acute respiratory illness (ARI), which is defined as having at least one of three traits: cough or nasal discharge/ blockage or difficulty breathing. Within virus shedding episodes, we imputed complete ARI episodes from intervals of recorded ARI. A virus shedding episode that had no day where an ARI was reported was assumed to be asymptomatic. For a virus shedding episode with at least one day of recorded ARI, the duration of symptoms was imputed using the midpoint method described for shedding episodes. This is illustrated below:



Green open circles are reported ARI symptoms (ARI positive) within the shedding episode and black open circles are confirmed absence of ARI (ARI negative). τ_1 , τ_2 , τ_3 and τ_4 are days within the shedding episode where information on symptoms was collected.

In this case, the mean sampling interval for ARI ‘samples’ within an episode was 3.78 days. This was obtained from all ARI episodes not just the ones within shedding

episodes. Figure A2. 3 and Figure A2. 4 show the shedding patterns by RSV group and ARI status.



Figure A2. 3: Shedding and ARI patterns for each of the 88 individuals who experienced at least one RSV A shedding episode.

The y-axis shows the individuals with labels color-coded by household, time is on the x-axis with zero indicating the day before the first sample was collected. The green dots show virus shedding and orange dots show the virus shedding days that were accompanied by an ARI.

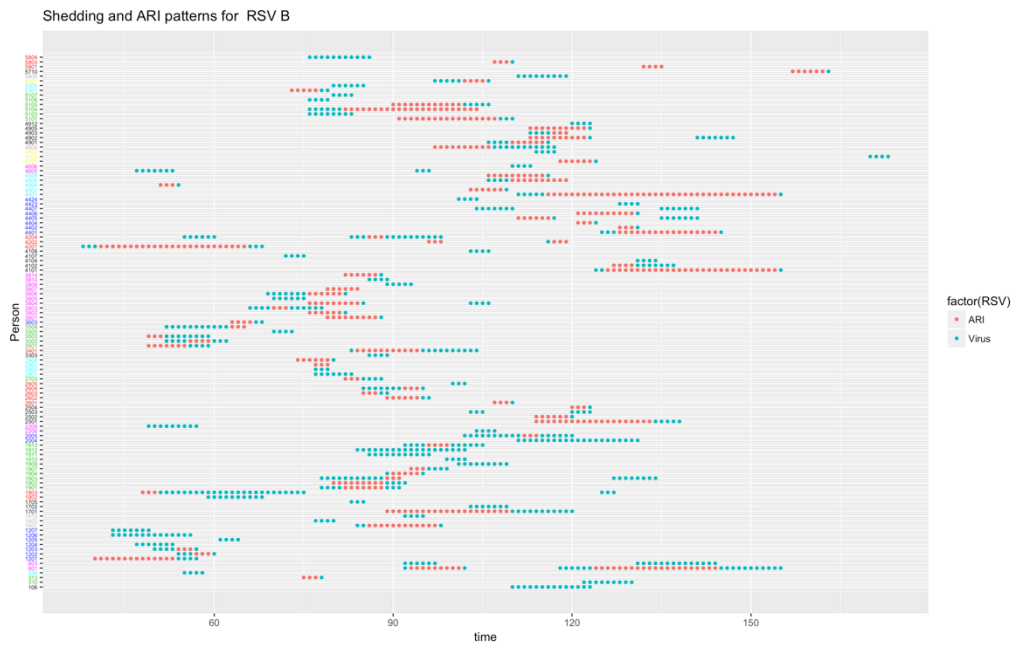
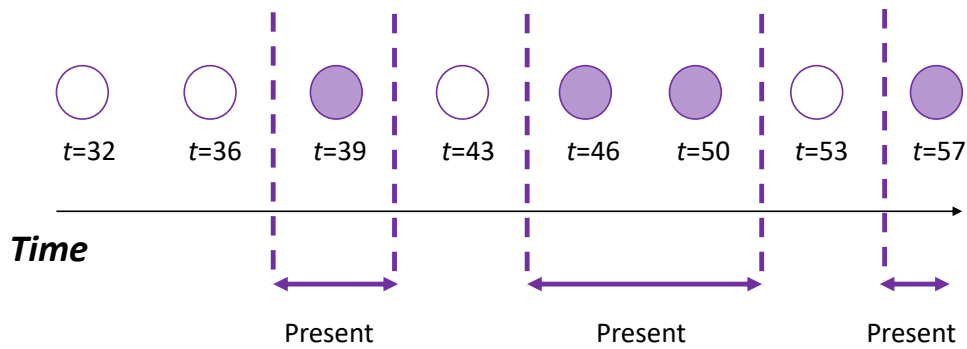


Figure A2. 4: Shedding and ARI patterns for each of the 113 individuals who experienced at least one RSV B shedding episode.

The y-axis shows the individuals with labels color-coded by household, time is on the x-axis with zero indicating the day before the first sample was collected. The green dots show virus shedding and orange dots show the virus shedding days that were accompanied by an ARI.

The imputation of continuous periods of presence or absence from the household was done similar to the imputation of shedding durations, however, there was no left or right censoring. Each participant had a set of days of recorded data, these days were either marked as 'away' or 'present' in the household, e.g. a participant might have data on days {32, 36, 39, 43, 46, 50, 53, 57} with status {away, away, present, away, present, present, away, present}. Since no data is available for this individual before day 32 and after day 57, no imputation is done outside this time window. For the days within the window, imputation is done as illustrated below:



Filled circles are days when the participant was recorded as being present while open ones is when they were away. The present period starts halfway between the last 'away' and first 'present' and ends halfway between the last 'present' and first 'away'.

A2.2. Extra results

This section shows some additional results that are mentioned in the main text. Three chains with different starting points were used to generate the parameter estimates. The trace plots are shown in Figure A.5. Chain 3 was run in three parts each with a length of 50000, 100000 and 100000 respectively. The starting point of the second part was the end point of the first part, and so on for the third part. This was done in an attempt to reduce total computation time. The model runs were implemented on a cluster computer that appeared to be slowing down tasks that were taking up a lot of time and resources, as such, to try and work around this, the long chains were split up to give the impression of a new task. The Final results given after a burn-in of 80000 iterations exclude the re-start period seen between iteration 150000 and 175000. However, including it does not make a significant difference to the inferred posterior distributions.

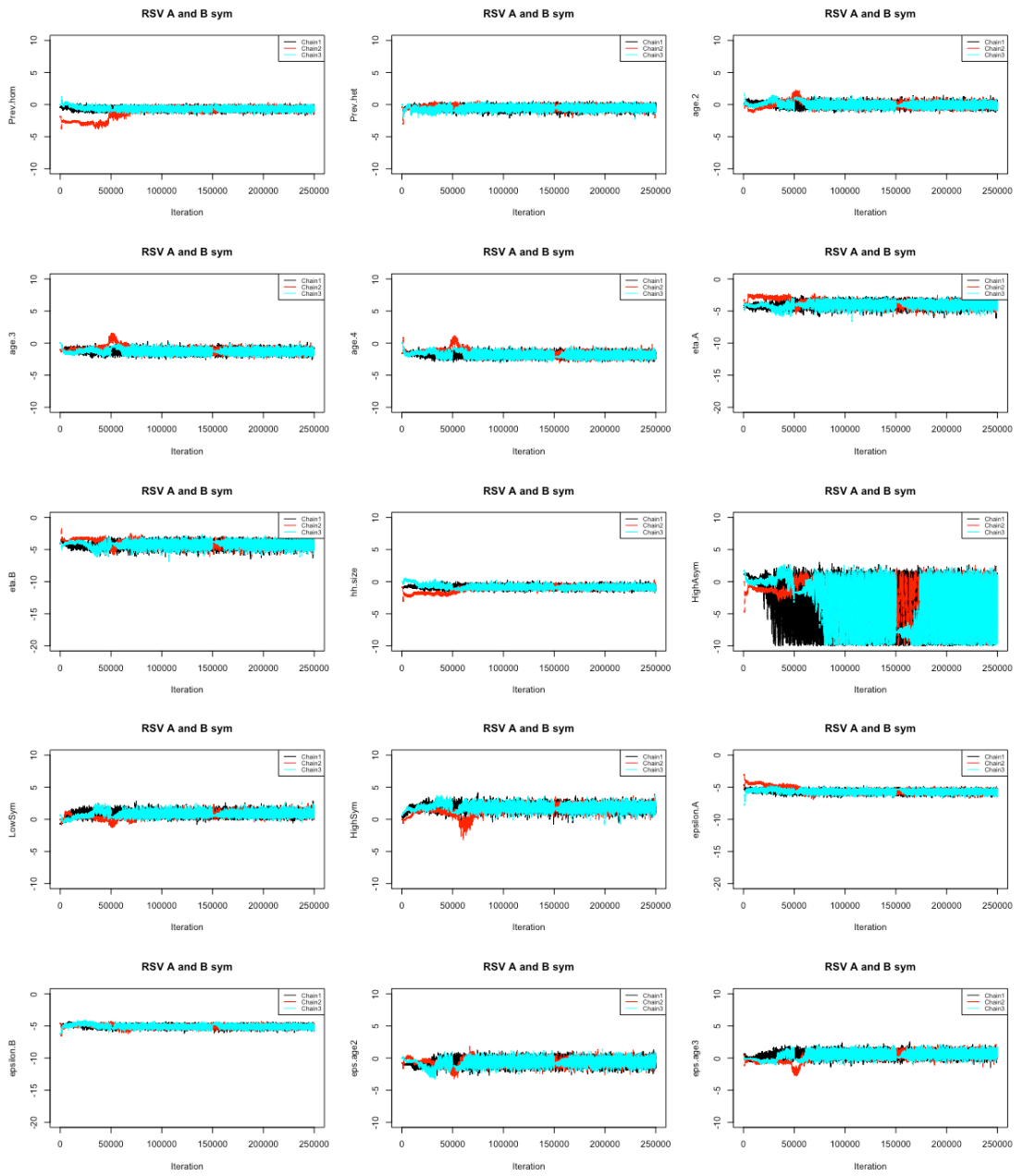


Figure A2. 5: Trace plots showing convergence for the 15 parameters of interest.

Three chains with different starting points were used.

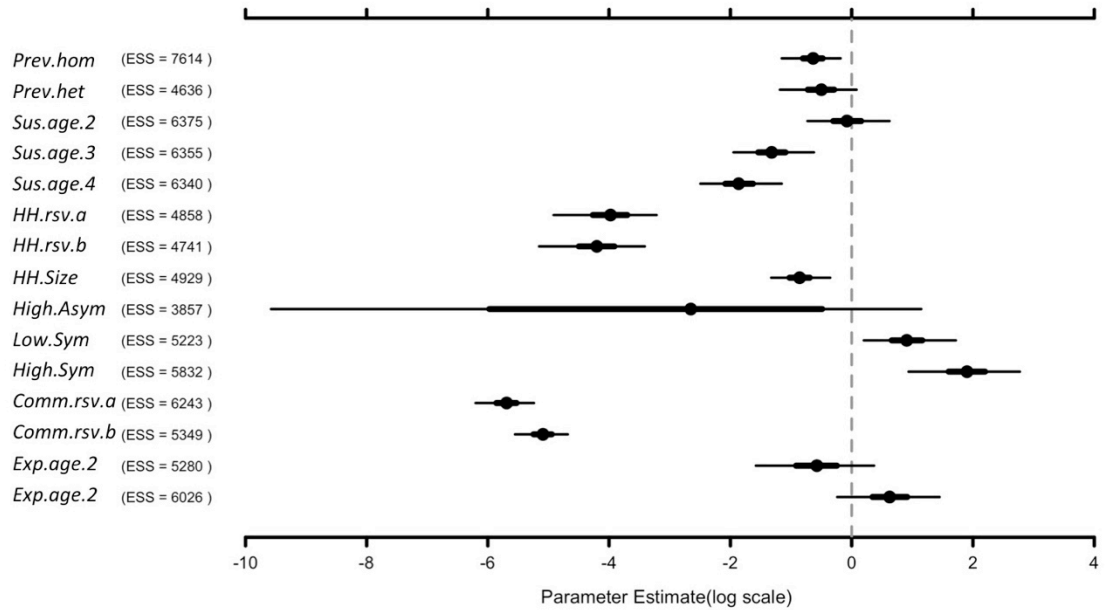


Figure A2. 6: Caterpillar plot of estimated parameters.

The 15 parameters estimated, and their respective effective sample sizes are shown. Points represent posterior medians, the thick lines represent 50% credible region and the thin lines represent 95% credible region. Except η_A and η_B (within household transmission coefficients) ε_A , and ε_B (community transmission coefficients respectively) all the other parameters represent relative effects where a reference group exists. If a relative effect parameter is equal to 1(0 on the log scale) then the group it represents, and the reference group are not different. ESS is the effective sample size

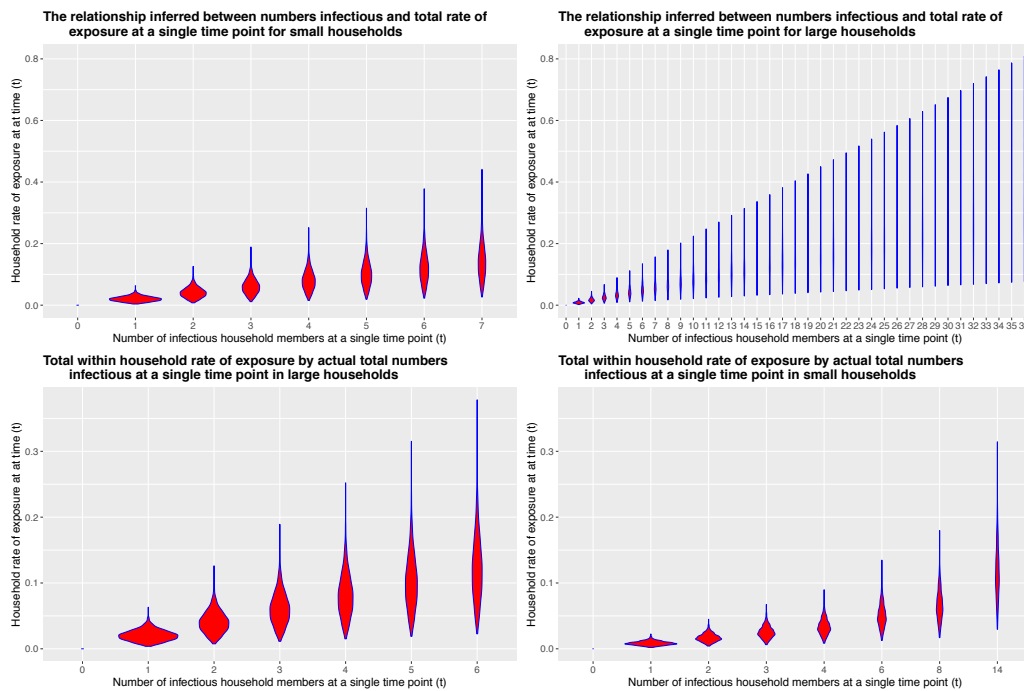


Figure A2. 7: Comparing the total household rate of exposure
 $(\sum_{j \neq i} HH.Risk_{h,g,j \rightarrow i}(t))$ between small and large households.

Each panel shows violin plots (combination of box plots and density plots) giving the distribution of the total household rate of exposure by total number of people infectious in the household at a given time point. The x-axis shows the total number infectious and the y-axis shows the value of the total household rate of exposure. The top row shows the linear relationship between total number of infectious individuals and total rate of within household exposure inferred from the parameter estimates for small households (left panel) and large household (right panel). The bottom row shows the same linear relationship, but with actual observed number of infectious household members. In this data set, small households had at most 6 simultaneously infectious household members while large household had 14.

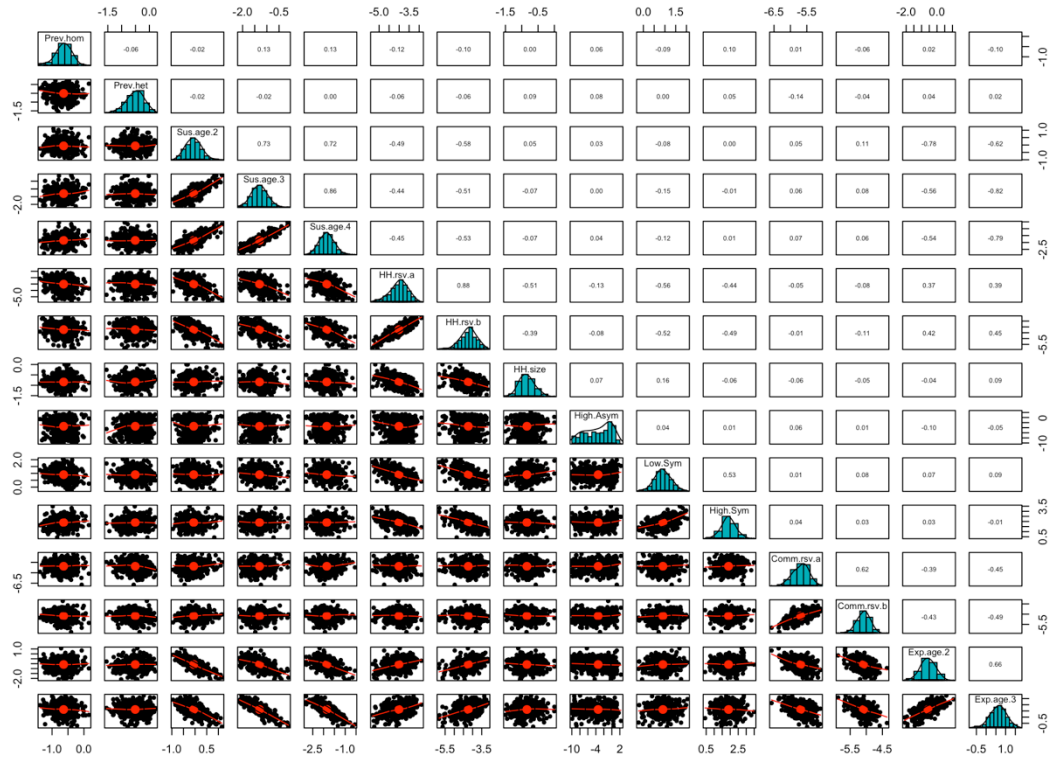


Figure A2. 8: Correlation patterns of the different parameters obtained from fitting to the observed data.

Table A2. 1: Results of fitting a reduced version of the model.

We reduced the model such that there are no interactions between different RSV groups and refit the data in three additional ways: RSV A alone, RSV B alone and RSV with no distinction between groups.

Parameter symbol	Parameter name	RSV A	RSV B	RSV
$\phi_{Y,hom}$	Prev.hom	0.444 (0.0194, 0.963)	0.547 (0.276, 0.983)	0.643 (0.423, 0.978)
ϕ_{X2}	Sus.age.2	1.01 (0.436, 3.16)	0.773 (0.263, 2.99)	0.919 (0.456, 2)
ϕ_{X3}	Sus.age.3	0.293 (0.134, 0.821)	0.234 (0.0866, 0.859)	0.294 (0.154, 0.63)
ϕ_{X4}	Sus.age.4	0.187 (0.0829, 0.534)	0.129 (0.047, 0.494)	0.159 (0.0811, 0.343)

η	HH.rsv	0.021 (0.00544, 0.0561)	0.0101 (0.000453, 0.0343)	0.0133 (0.00457, 0.0311)
ϕ_H	HH.size	0.337 (0.184, 0.659)	0.606 (0.285, 1.52)	0.467 (0.277, 0.847)
ϕ_{I2}	High.Asym	0.0401 (0.0000626, 1.72)	0.243 (0.0000662, 18.1)	1.03 (0.000126, 6.36)
ϕ_{I3}	Low.Sym	1.91 (0.711, 6.36)	3.39 (1.20, 71.1)	2.17 (0.931, 5.93)
ϕ_{I4}	High.Sym	7.28 (0.701, 25.4)	6.31 (0.885, 158)	8.76 (3.72, 23.5)
ε	Comm.rsv	0.00328 (0.00159, 0.00594)	0.006 (0.00339, 0.0096)	0.00939 (0.00588, 0.014)
ϕ_{E2}	Exp.age.2	0.335 (0.0745, 1.46)	0.815 (0.16, 3.44)	0.574 (0.187, 1.65)
ϕ_{E3}	Exp.age.3	1.73 (0.548, 5.26)	2.13 (0.484, 7.6)	1.81 (0.712, 4.4)

A2.3. Modification of the likelihood to establish the most likely infection source for every case.

The rate of exposure in the model is give as:

$$\lambda_{i,h,g}(t) = S_{i,g}(t) \left[M_{i,h}(t) \sum_{j \neq i} HH_Risk_{h,g,j \rightarrow i}(t) + Comm_Risk_{i,g}(t) \right]$$

This can be expanded to show all the variables and parameters as shown:

$$\lambda_{i,h,g}(t) = \exp(\phi_{Y,hist}(t) + \phi_{X,age}) \left[M_{i,h}(t) \sum_{j \neq i} (\eta_g * \psi_H * \psi_{I,inf}) + (\varepsilon_g * f_g(t) * \psi_{E,age}) \right] \quad (1)$$

For a given case i , in order to be able to calculate the likelihood of infection from a particular source Ω_i , either a sampled housemate or an unknown community source, we need to formulate the probability of transmission from said source at time t . This is given by:

$$Pr_{\Omega_i \rightarrow i, h, g}(t) = \frac{\lambda_{\Omega_i \rightarrow i, h, g}(t)}{\lambda_{i, h, g}(t)} \quad (2)$$

For Ω_i in the same household as i , the rate of exposure is given by

$$\lambda_{\Omega_i \rightarrow i, h, g}(t) = \exp\left(\phi_{Y, hist}(t) + \phi_{X, age}\right) \left[M_{i, h}(t) \eta_g \psi_{i, H} \psi_{\Omega_i, l, inf}(t) M_{\Omega_i, h}(t) \right]$$

For Ω_i an unknown source external to the household, the rate of exposure is given by

$$\lambda_{\Omega_i \rightarrow i, h, g}(t) = \exp\left(\phi_{Y, hist}(t) + \phi_{X, age}\right) \left[\varepsilon_g * f_g(t) * \psi_{E, age} \right]$$

The likelihood function

The probability given in (2) is calculated for a time point $t =$ exposure time of individual i , t_i^E . This is not observed in the data, however, given our assumption on the latency duration, we can define a 6-day window of possibility. If case i had a shedding onset at time T_i^O , then the window for transmission is from day $(T_i^O - 5)$ to $(T_i^O - 0)$. For each day in the window, potential sources are identified based on shedding status and for each combination of infection source Ω_i and exposure date t_i^E , the likelihood is calculated using the formula below:

$$\begin{aligned} L(\varphi | \{T_i^O, t_i^E, \Omega_i\}) &= \left(1 - e^{-\lambda_{i, g, h}(t_i^E)}\right) * \left(\prod_{t_i \neq t_i^E} e^{-\lambda_{i, g, h}(t_i)}\right) * \left(\theta_l(T_i^O - t_i^E)\right) \\ &* \left(\frac{\lambda_{\Omega_i \rightarrow i, h, g}(t_i^E)}{\lambda_{i, h, g}(t_i^E)}\right) \end{aligned}$$

The first part of the product is the probability of infection at time t_i^E , the second part is the probability of escaping infection at any time $t_i \neq t_i^E$, the third is the probability of

a latency duration of length $(T_i^o - t_i^E)$ and the last term is the probability of transmission from source Ω_i to i .

Given the likelihood, the highest-probability-source is chosen as the infection source that give the highest value of the likelihood.

A2.4. Model validation

This is a two part process: first we check if the parameters estimated can reproduce the results (or something similar) by simulation; then we check if given simulated data, we can re-estimate parameters that are similar to the ones used to simulate the data. This process is illustrated in the flow chart below.

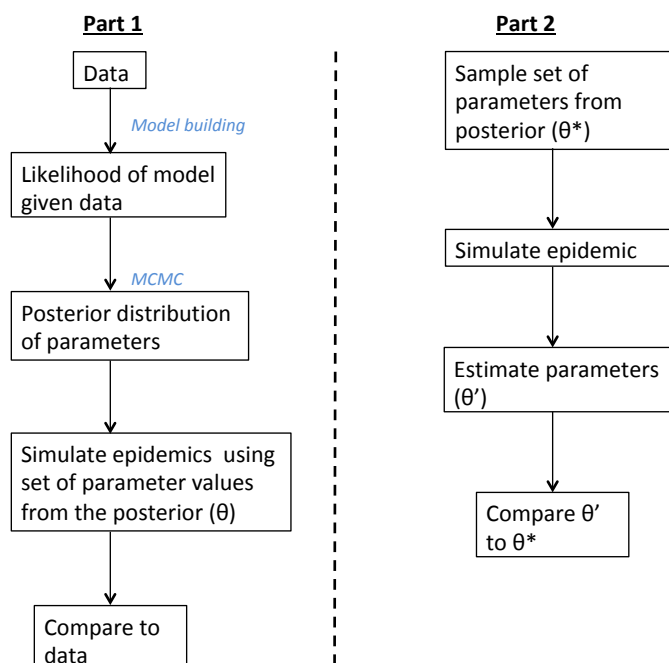


Figure A2. 9: Flow chart showing validation process

Given a set of parameter values, the simulation pseudo code per simulation is as follows:

1. Initiate system such that everyone one is susceptible to RSV A and RSV B.
2. At every time step keep track of:
 - a. Susceptibility status of every individual
 - b. Exposure status
 - c. Infectious status (viral load and infectivity group)

d. Infection history

3. At every time step:

- a. Determine number of transmission events using

$$\Delta E = \text{Poisson} \left(\sum_{i \in C_E} (1 - e^{-rate_{E,i}}) \right)$$

ΔE = number of events of type E at a given time point

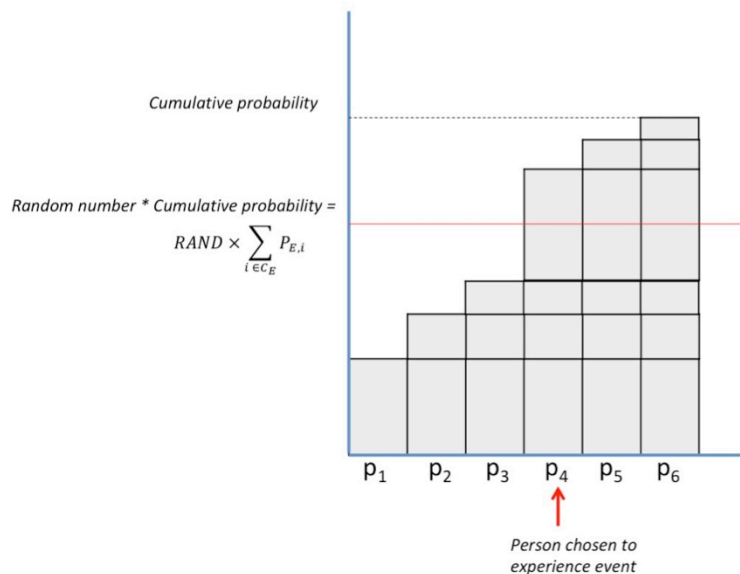
C_E = set of all individuals capable of experiencing event E .

$rate_{E,i}$ = rate of occurrence of event E on person i .

- b. Determine who experiences each event. For a given event, order individuals capable of experiencing the event. For a given person p to experience the event, the following inequality has to be satisfied.

$$\sum_{i=1}^{i \leq p-1} P_{E,i} < \left(\text{RAND} \times \sum_{i \in C_E} P_{E,i} \right) \leq \sum_{i=1}^{i \leq p} P_{E,i}$$

Where $P_{E,i} = 1 - e^{-rate_{E,i}}$ = probability of person i experiencing event E . RAND = a random number between (but not including) 0 and 1. This is illustrated in the figure below.



Repeat this until the required number of events

- c. For each individual experiencing a transmission event, assign a latency duration and shedding profile by sampling from the relevant empirical

distributions. The empirical latency distribution is the same as was used in estimating the parameters and is homogeneous for every individual. The shedding profiles are grouped by age in the following 4 groups <1, 1-5, 5-15 and ≥15 years (see Figure A2. 10 and Figure A2. 11 for age grouped shedding profiles). An assigned shedding profile is a combination of duration of shedding, viral loads and symptom status. Once latency durations and shedding profiles have been assigned, the state variables for each individual are updated accordingly.

d. Update rate of exposure.

The rate of exposure/transmission for susceptible individuals changes according to

$$\lambda_{ihg}(t) = \exp(\phi_X X_i + \phi_{Yg} Y_{ig}(t)) \left[M_{ih}(t) \eta_g \phi_H H_i \sum_{j \neq i} \phi_I I_{jhg}(t) + \phi_E E_i \varepsilon_g f_g(t) \right]$$

Figure A2. 10 and Figure A2. 11 show the shedding profiles as observed from the data for RSV A and B, clustered by age and symptom status.

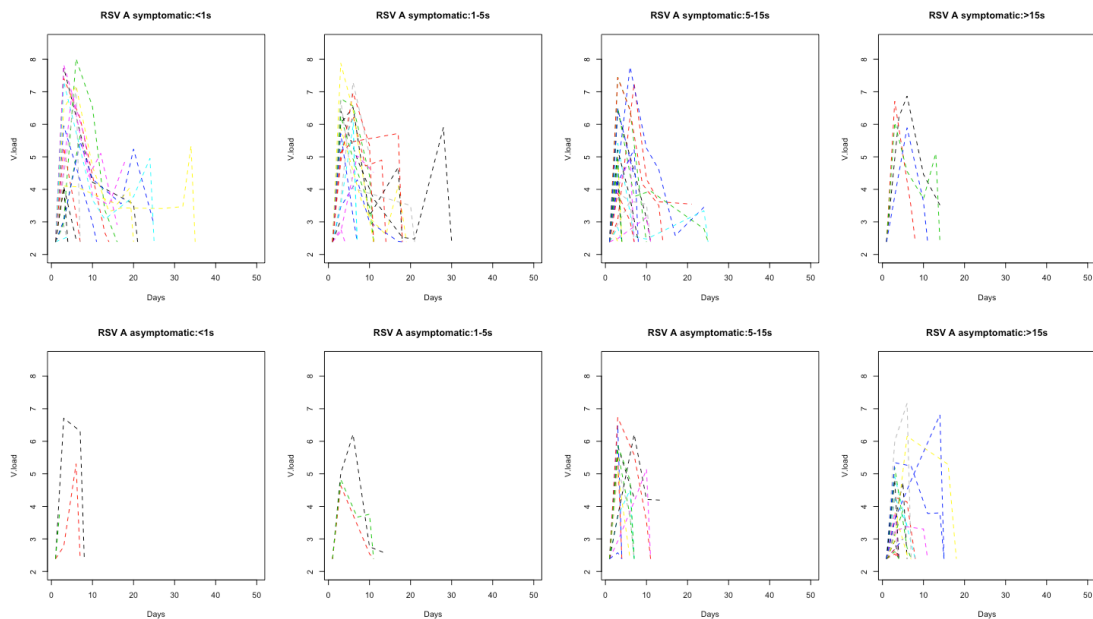


Figure A2. 10: RSV A shedding profiles as observed.

Each figure shows the viral loads on different days of shedding for each infection episode observed. The top row shows profiles for symptomatic RSV A shedding by age group in years, the bottom row shows profiles for asymptomatic RSV A shedding by age group in years.

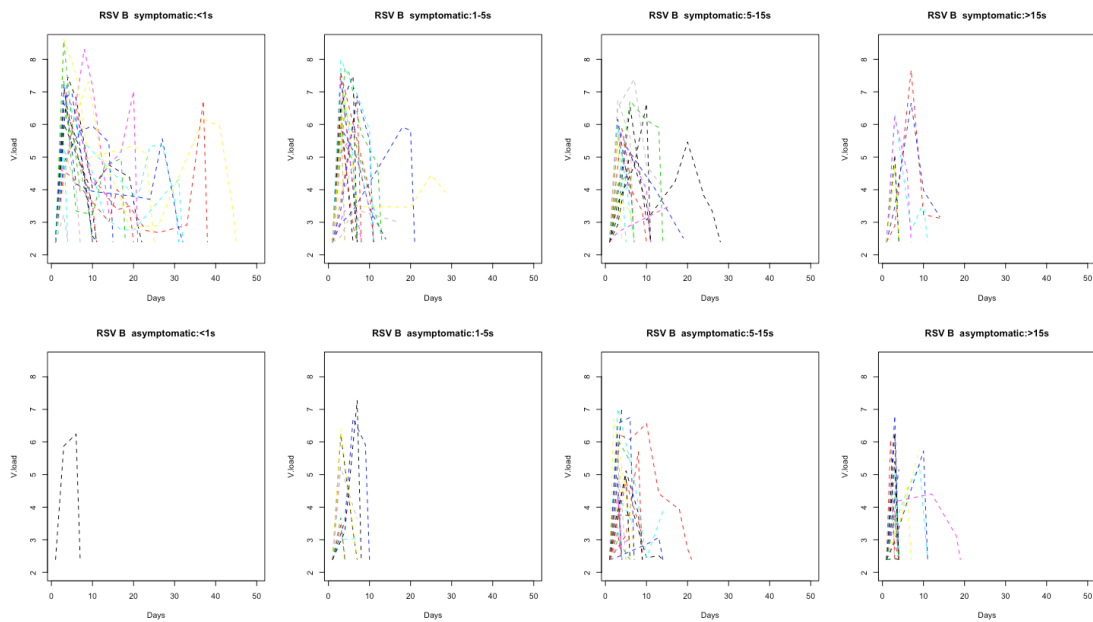


Figure A2. 11: RSV B shedding profiles as observed.

Each figure shows the viral loads on different days of shedding for each infection episode observed. The top row shows profiles for symptomatic RSV B shedding by age group in years, the bottom row shows profiles for asymptomatic RSV B shedding by age group in years.

We sampled 5 sets of parameters (dependent sampling to maintain the correlations observed) and for each set simulated 200 epidemics to compare to the data. The sampled parameters relative to the posterior distribution are shown in Figure A2. 12. In addition to looking at the projected epidemics, we also look at the following outcome measures to make comparisons:

- Total number of individuals infected
- Total number of households infected
- Proportion of individuals with repeat infections
- Timing of epidemic peak

Figure A2. 13 and Figure 3. 4 in the main text show the results of the simulations relative to the observed data.

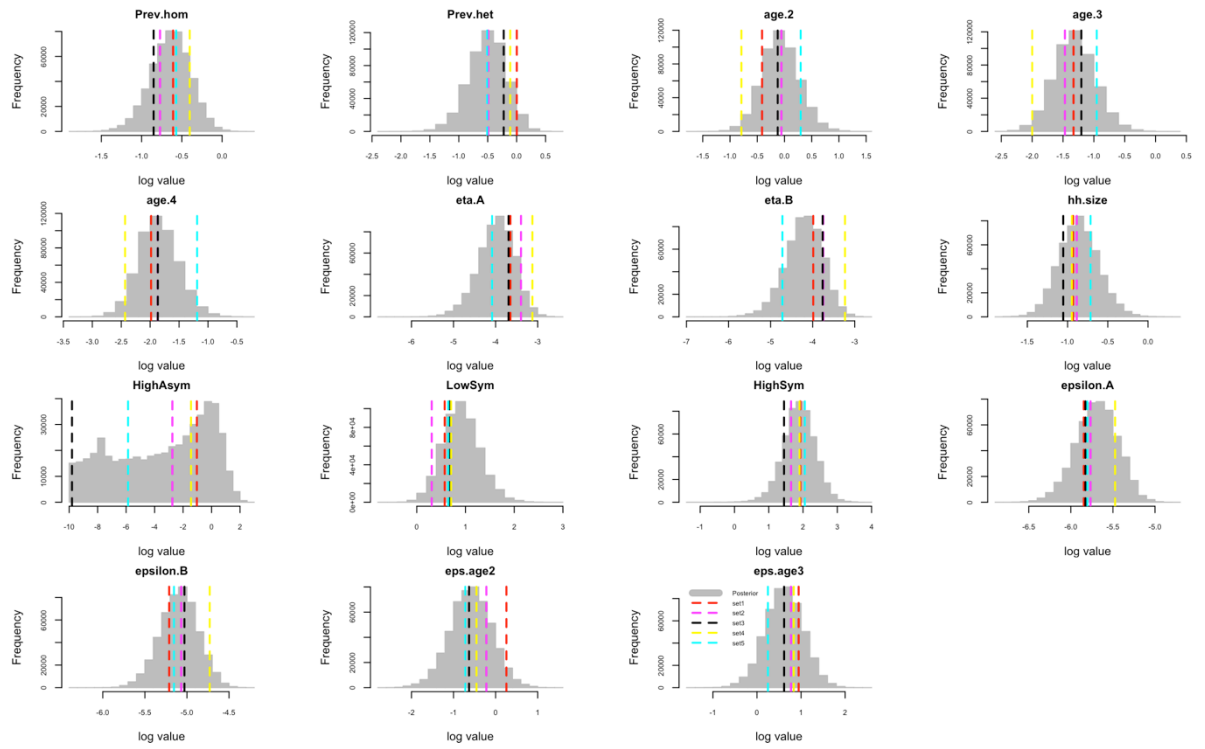


Figure A2. 12: Histograms of the posterior distributions with vertical lines showing sample sets that were used in simulation.

Each panel shows histograms of different parameters in grey. Red dashed lines show the value of the parameter in set1, dark pink shows set2, black set3, yellow set4 and blue set5.

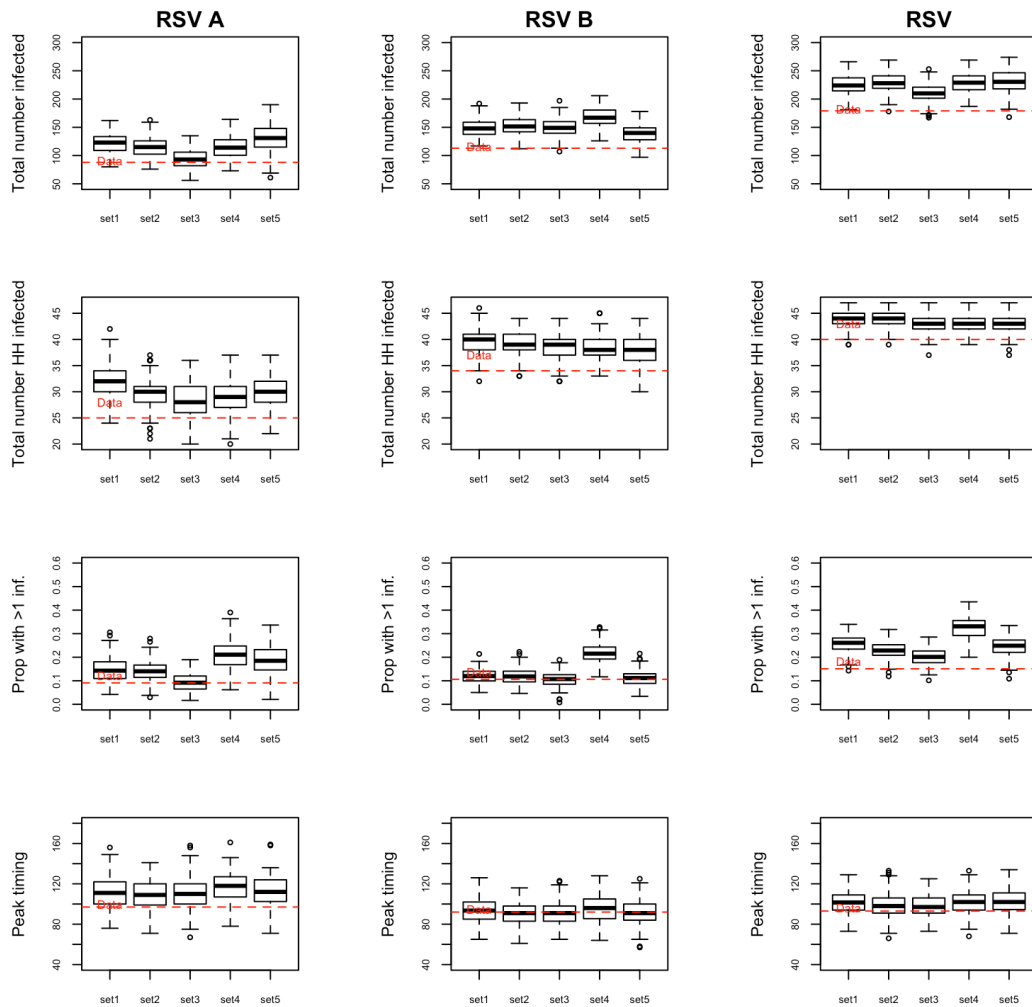


Figure A2. 13: Outcome measures from simulated data when using different sets of parameters drawn from the posterior estimated from the observed data.

Each box and whisker plot is the distribution of the specific outcome measure from 200 simulations set run from a single sampled data set.

To check if the model can re-estimate known parameter values, we simulated an epidemic and compared the re-estimated densities to the densities given by using the observed data. Table A2. 2 gives the values of the parameters used to simulate the epidemic, Figure A2. 14 compares the real and simulated epidemics and Figure A2. 15 compares the original and re-estimated parameter densities.

Table A2. 2: Parameter set used to simulate an epidemic

Parameter symbol	Parameter name	Set1
$\phi_{Y,hom}$	Prev.hom	0.544

$\phi_{Y,het}$	Prev.het	1
ϕ_{X2}	Age.2	0.662
ϕ_{X3}	Age.3	0.265
ϕ_{X4}	Age.4	0.138
η_A	Eta.A	0.026
η_B	Eta.B	0.0186
ψ_H	hh.size	0.394
ψ_{I2}	HighAsym	0.356
ψ_{I3}	LowSym	1.77
ψ_{I4}	HighSym	6.85
ϵ_A	Epsilon.a	0.00289
ϵ_B	Epsilon.b	0.00545
ψ_{E2}	Eps.age2	1.29
ψ_{E3}	Eps.age3	2.57

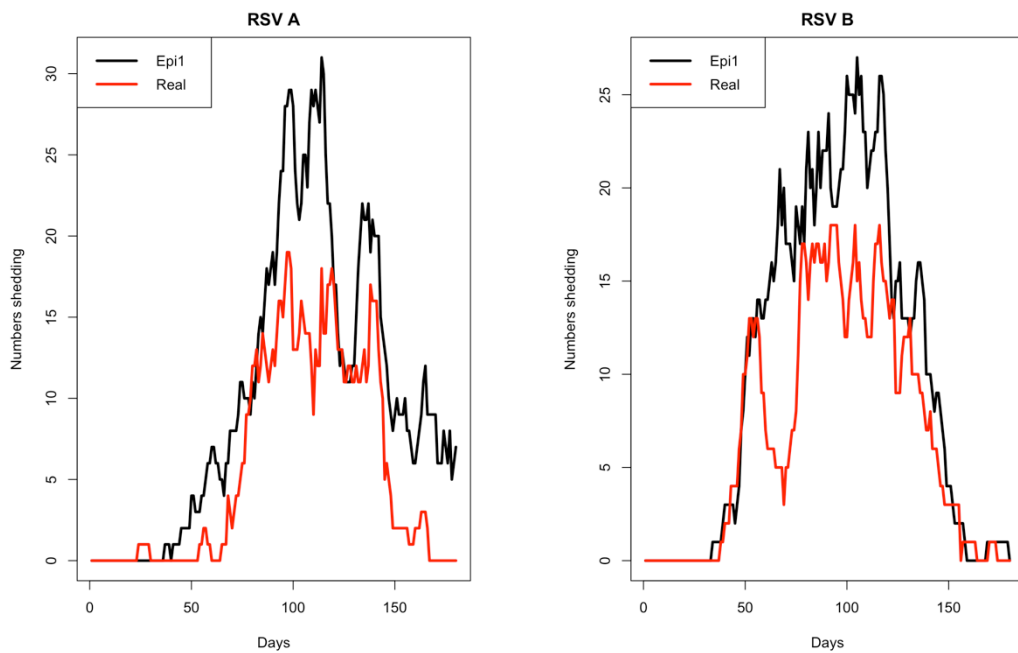


Figure A2. 14: Comparing the real (red lines) and simulated(black lines) epidemics

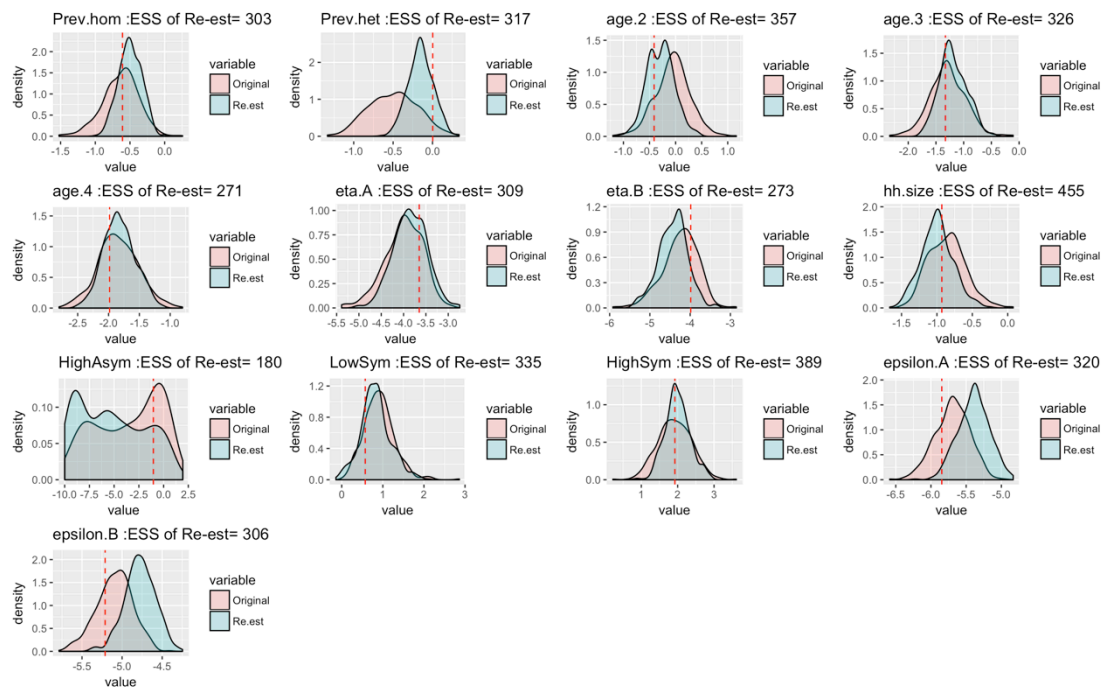


Figure A2. 15: Comparing the posterior densities obtained from using the observed data to those from using the simulated data.

The posterior densities from using real data are shown in red while the ones from simulated data are shown in blue. The dashed red line shows the value of the parameter used to simulate the epidemic. ESS is the effective sample size.

The re-estimated distributions capture the parameter used to simulate the epidemic and in general fall within the ranges of the original distributions obtained from the real data.

A2.5. Sensitivity analysis

We check if our results were sensitive to the background community density function by exploring 3 additional function forms. The results are presented in the following three figures. Option1 shows the density curves used in the main analysis, Options 2,3 and 4 show the curves used in the sensitivity analysis. The first additional function form was sampling switching the RSV A curve with the RSV B curve to check for sensitivity to peak epidemic timing. The second function, option 3, is curves generated from RSV A and RSV B hospital incidence from the same sampling period. The data on RSV admissions was obtained from the main referral hospital in the area. The curve in

option 4 is a reverse of the curves in option 3, i.e. the RSV A curve was swapped with the RSV B curve.

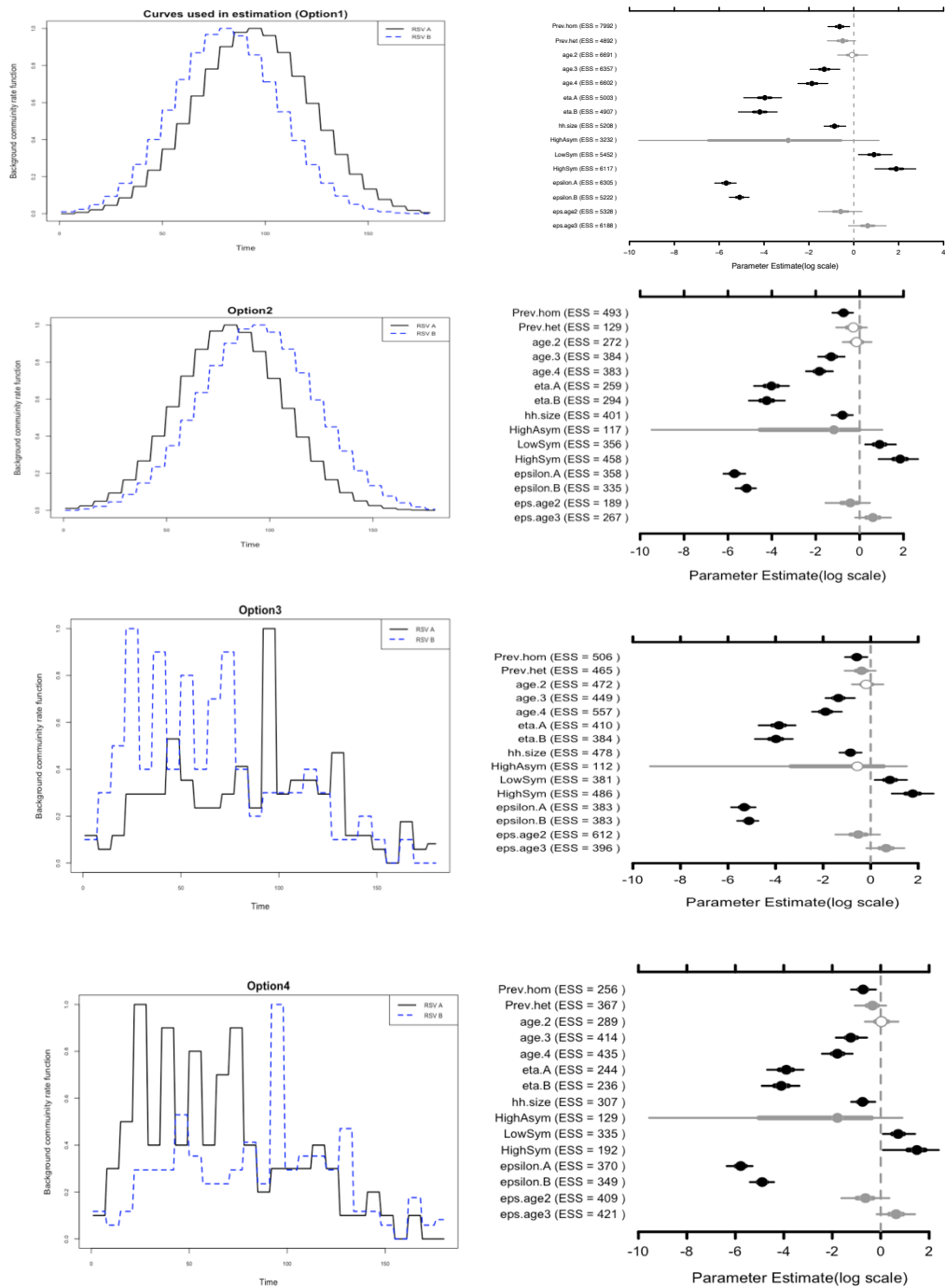


Figure A2. 16: Using different density functions for the background community rate and comparing results.

The left side shows the density functions for RSV A and RSV B, the right side shows the re-estimated parameters after 20,000 iterations.

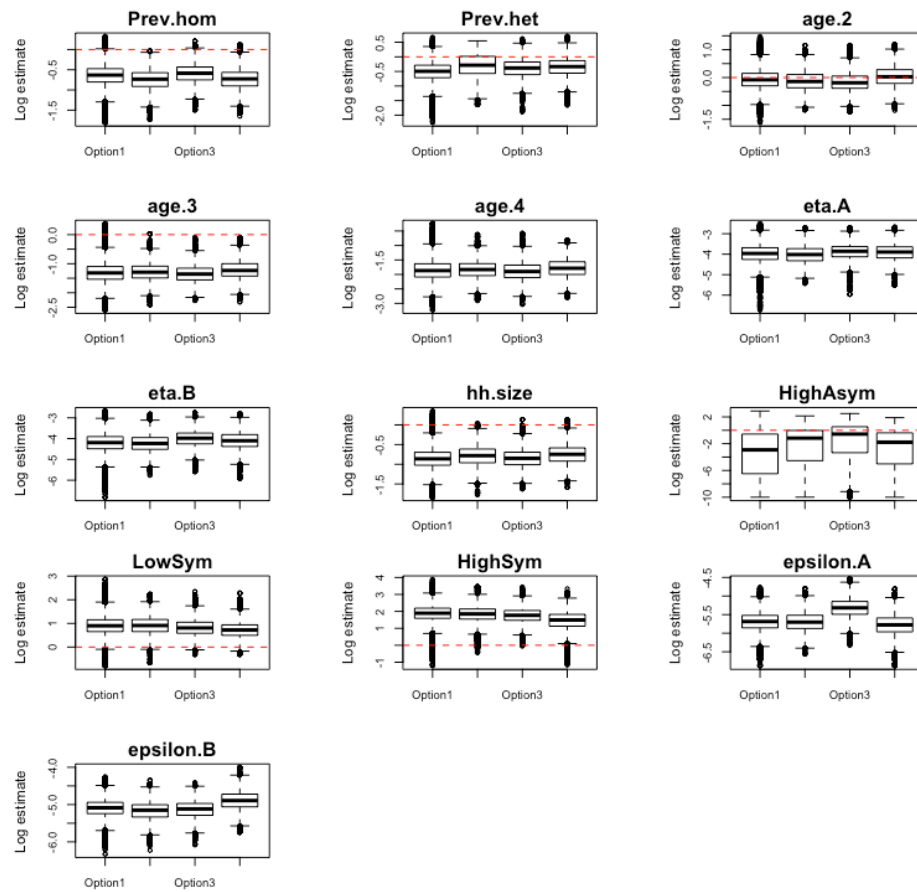


Figure A2. 17: Using different density functions for the background community rate and comparing results.

Box plots comparing the estimated parameters.

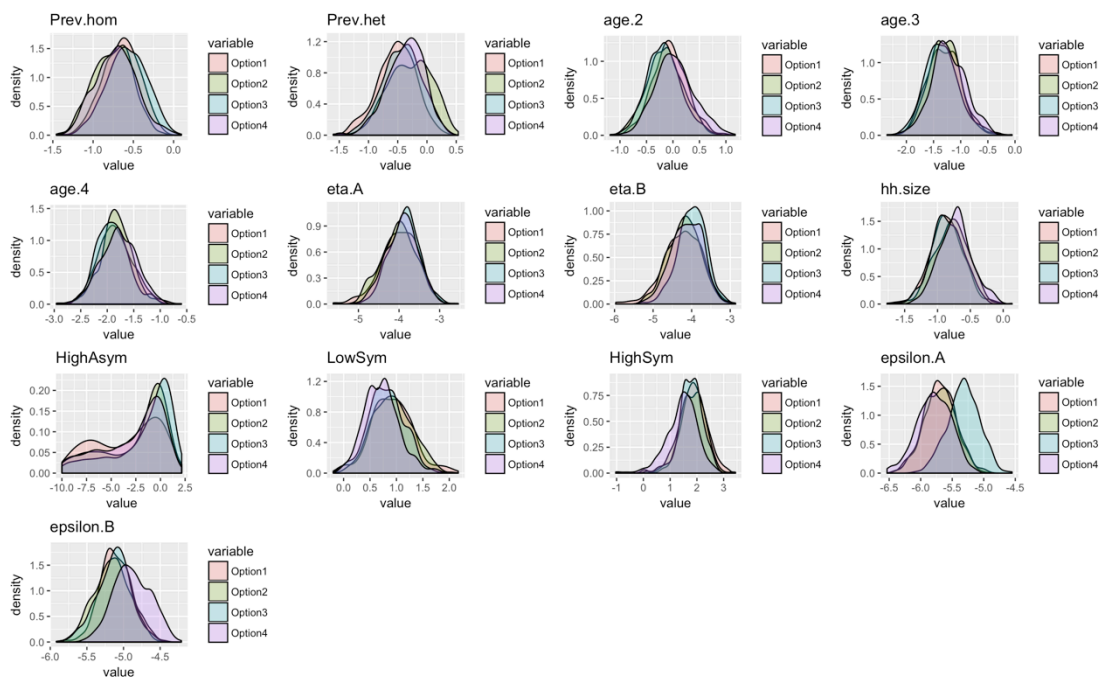


Figure A2. 18: Using different density functions for the background community rate and comparing results.

Density plots comparing the estimated parameters.

We also looked at the distribution of cases by household size. In Figure A2. 19 we see that RSV A got into the largest household and infected significantly more people than RSV B. Looking at Figure A2. 20, it seems that all but one RSV A case were probably part of a single outbreak (based on perceived temporal distance). To check if this could be the reason for the difference in within household transmission coefficient estimated, we removed data from the largest household (HH5) and re-estimated the parameters. The results of this are shown in Figure A2. 21 and Figure A2. 22. The slight difference between the RSV groups in the within household transmission parameter is still present.

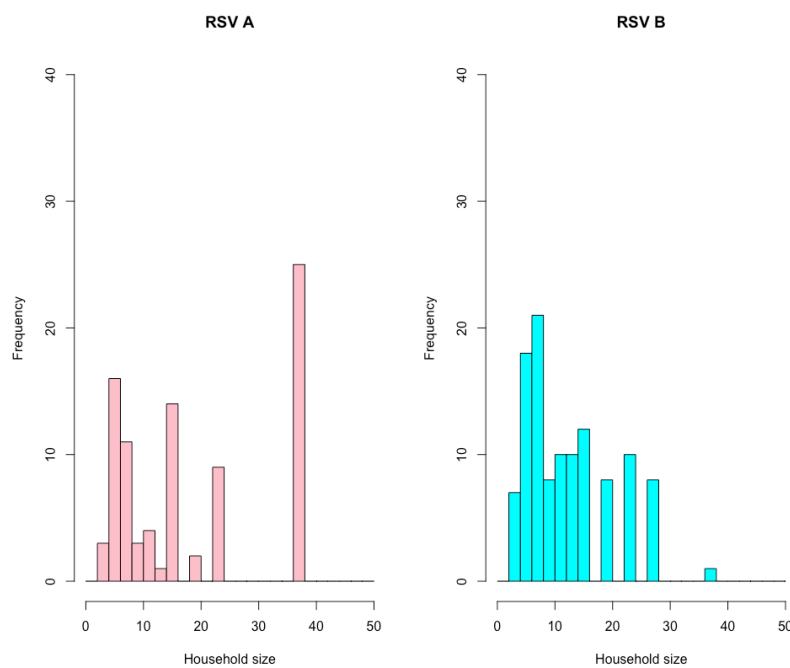


Figure A2. 19: Frequency distributions of RSV A and RSV B infections by household size

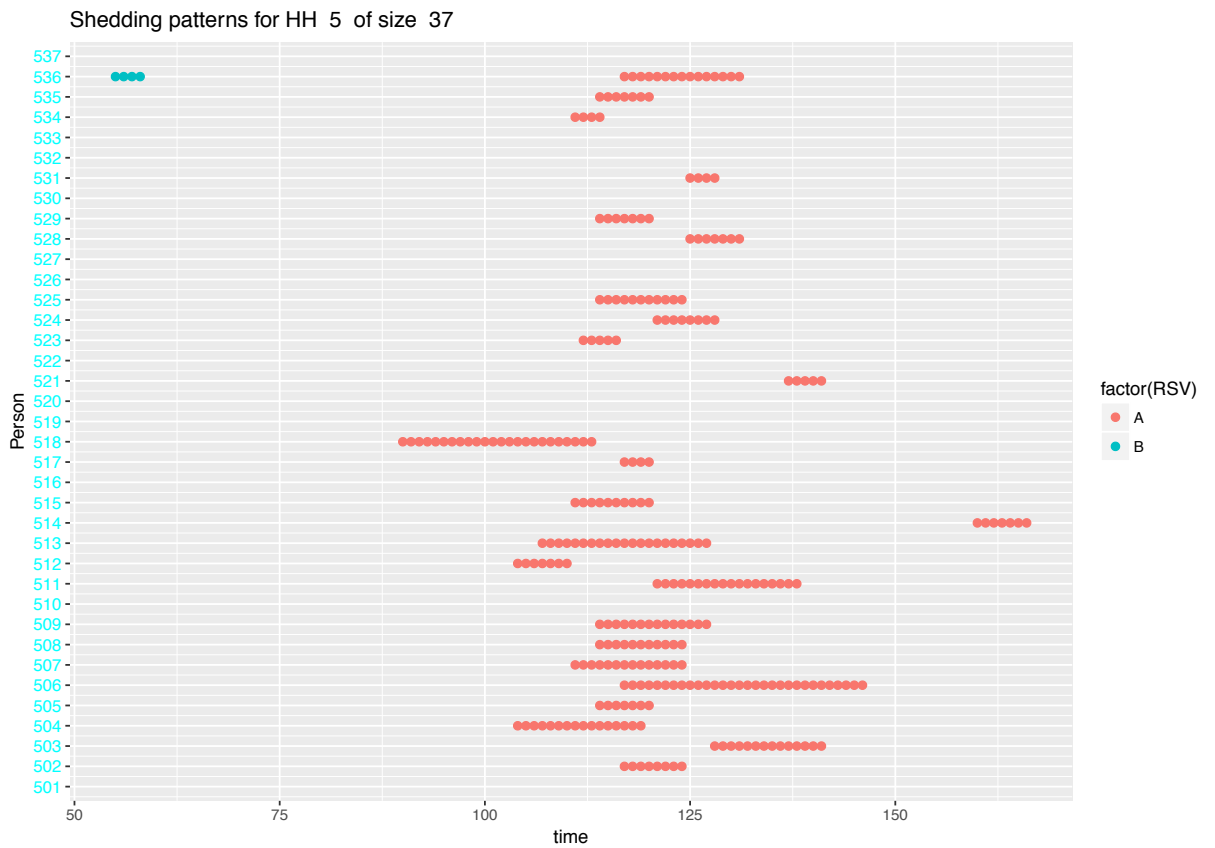


Figure A2. 20: Infection patterns in HH5

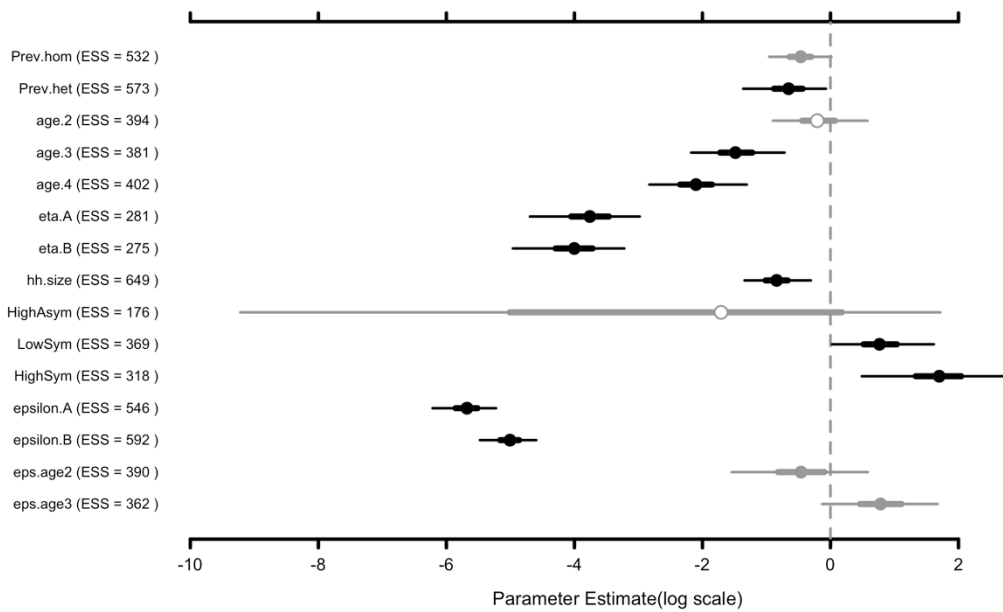


Figure A2. 21: Caterpillar plot showing results obtained when household 5 data was removed from the set.

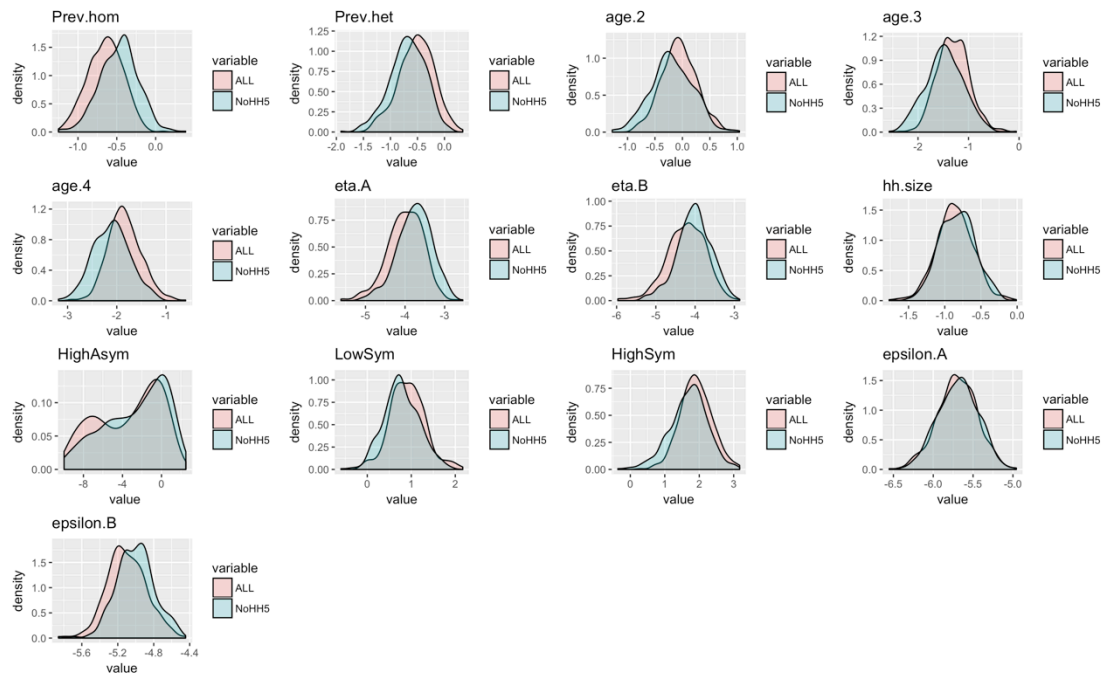


Figure A2. 22: Comparing densities of parameters estimates obtained when using all the data (light red) to densities obtained when using data without household 5 (light blue).

A2.6. Checking the contribution of symptomatic and asymptomatic individuals

In this section we show the results of simulations where infectiousness was altered, Figure A2. 23, and parameter estimation where only a subset of the data was used Figure A2. 24. For the simulation we compare three scenarios: Infectiousness of symptomatics and asymptomatics as given in the model parameters presented in Table 3; Infectiousness of the symptomatic individuals is reduced to match that of asymptomatic individuals (this is done so as to get an idea of what the effect of a vaccine that reduces symptoms would be); Infectiousness of asymptomatic individuals is assumed to be 0 such that they cannot transmit (this is done so as to get an idea of the contribution of asymptomatic infections to transmission). For each scenario, 10000 simulations were used based on sampling 100 different parameter (and making the modifications necessary for scenario 2 and 3) sets and for each set simulation 100 epidemics.

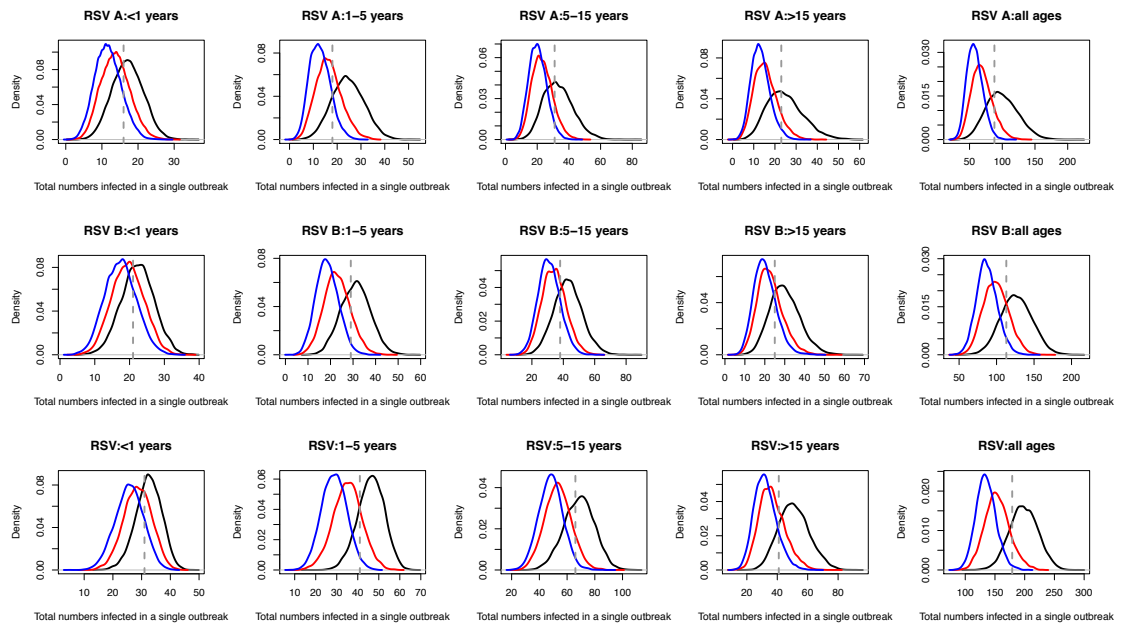


Figure A2. 23: Densities comparing the relative total incidence, by RSV group and age group, when the infectiousness of symptomatic individuals is altered or when the infectiousness of asymptomatic individuals is removed.

The black line shows the distribution of total number of people infected from 10000 simulations for estimated (unaltered) parameters scenario where symptomatic individuals are more infectious than asymptomatic. The red line shows the case when where the parameters used in simulation have been altered to force symptomatic individuals to be as infectious as asymptomatic individual (i.e. reduced infectiousness). The blue line shows when asymptomatic individuals are assumed to not be infectious at all.

From the figure above we notice that the greater shift in the distribution of cases when infectiousness of symptomatics is reduced occurs in the 1-15 year old age group. The reduction in the <1 year age group is not huge, presumably because transmission to this age group is from several sources as such reducing the infectiousness of symptomatics has little impact on the total numbers infected during an outbreak. We also notice that assuming asymptomatic cases are not infectious leads to far less number than were actually observed. This highlights the importance of asymptomatic individuals in transmission.

Following on from the simulation, we used a subset of the data that had only symptomatic episodes to re-estimate the model parameters (this is done to give an

idea of how much information would be missed if the sampling had only been of individuals who showed symptoms). In this case, we still had days in the data with shedding, but no symptoms and ARI episodes are not necessarily as long as the entire virus shedding episode.

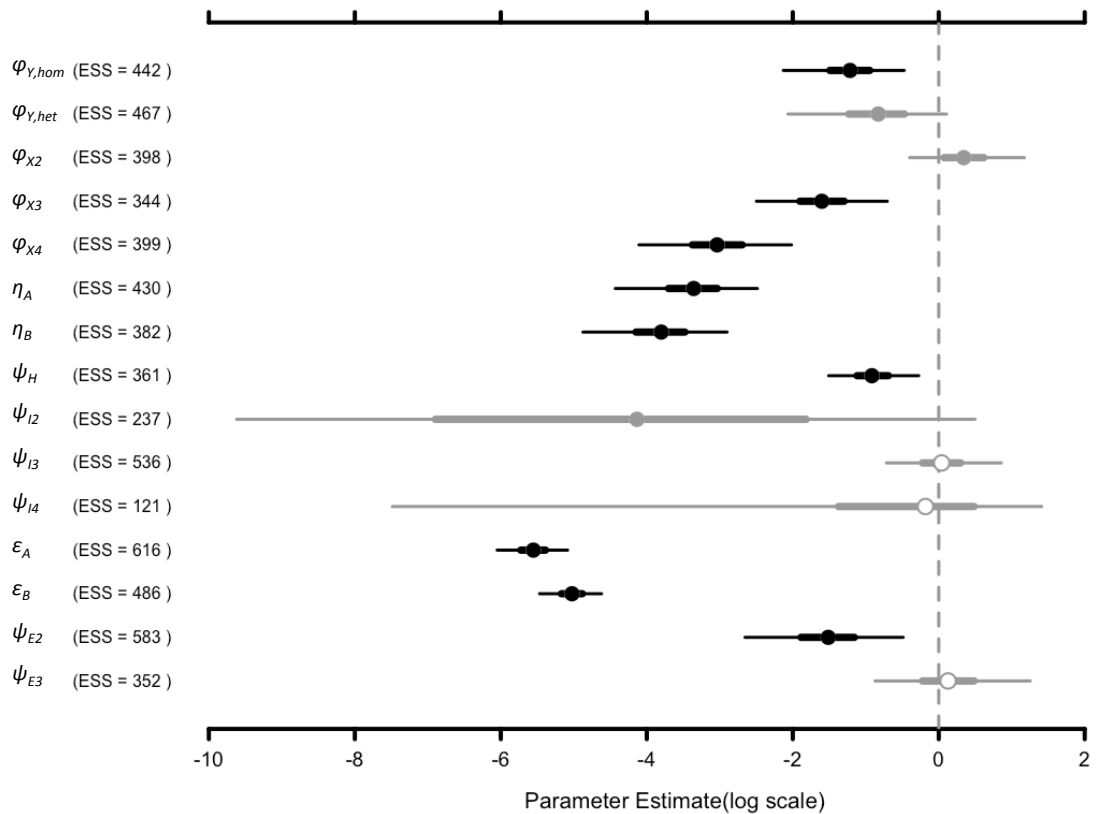


Figure A2. 24: Caterpillar plot of estimated parameters when only data from symptomatic episodes is used.

The 15 parameters estimated, and their respective effective sample sizes are shown. Points represent posterior medians, the thick lines represent 50% credible region and the thin lines represent 95% credible region. Except η_A and η_B (within household transmission coefficients) ϵ_A , and ϵ_B (community transmission coefficients) all the other parameters represent relative effects where a reference group exists. If a relative effect parameter is equal to 1(0 on the log scale) then the group it represents and the reference group are not different. Parameters where 50% credible interval overlaps with 0(dashed vertical line) are shown by open grey circles, where the 50% credible intervals do not overlap with 0 but the 95% credible interval does, filled grey

circles show these parameters. If there is not overlap with 0, the circles are black and filled.

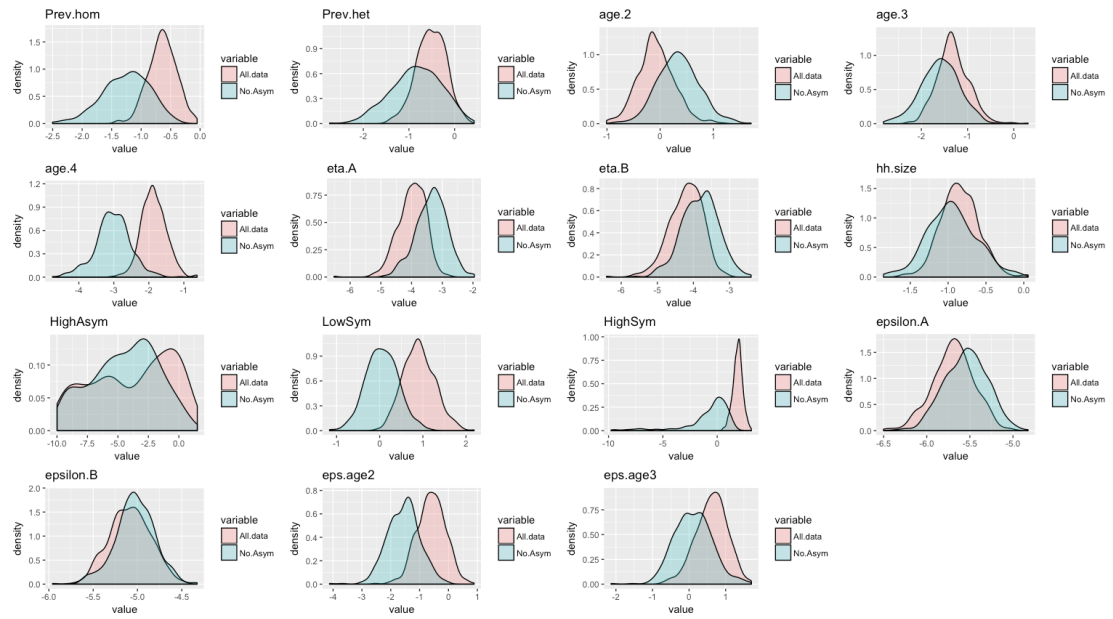


Figure A2. 25: Comparing densities of parameters estimates obtained when using all the data (light red) to densities obtained when using data from only symptomatic cases (light blue).

A2.7. Fitting household size as an ordinal variable

As the model was built up in stages, this section was done prior to the inclusion of symptom data; instead only viral load was used as a proxy to infectivity. To fit household size as an ordinal variable, the rate of exposure equation is as below

$$\lambda_{ihg}(t) = \exp(\phi_X X_i + \phi_{Yg} Y_{ig}(t)) \left[M_{ih}(t) \eta_g (N_i - 1)^{-\omega} \sum_{j \neq i} \phi_I I_{jhg}(t) + \phi_E E_i \varepsilon_g f_g(t) \right]$$

The factor $(N_{ih} - 1)^{-\omega}$ modifies the within household transmission coefficient, where N_i is the household size for susceptible i and ω determines that kind of transmission. If $\omega \rightarrow 0$, it points to density dependent transmission, $\omega=1$ implies frequency dependence. The estimation of ω was done using the entire data set and again using a subset where the definition of a household was changed such that a household is defined as individuals who share a building unit. The results of this are shown below.

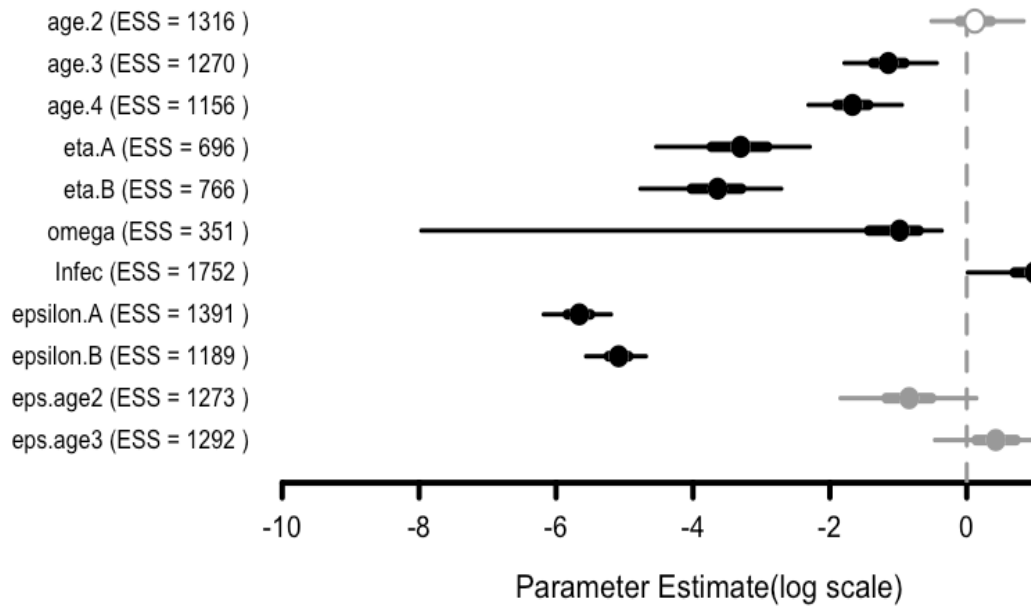


Figure A2. 26: Caterpillar plot showing the results of estimating a parameter ω (omega) when household size is treated as an ordinal variable.

These results were obtained when fitting was done using all the data available.

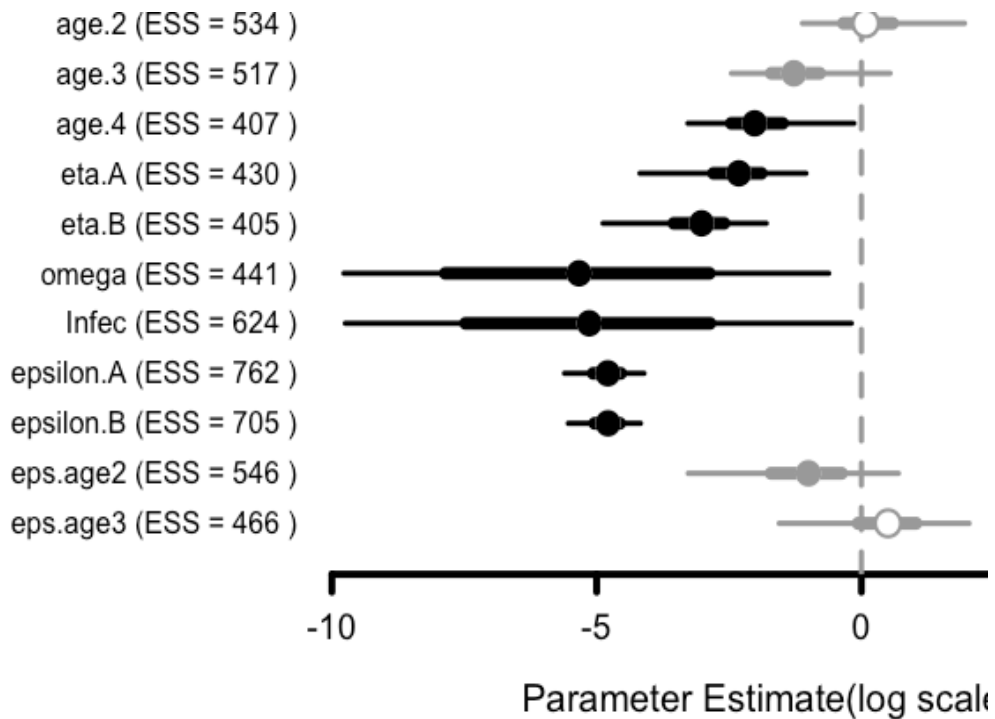


Figure A2. 27: Caterpillar plot showing the results of estimating a parameter ω (omega) when household size is treated as an ordinal variable.

These results were obtained when fitting was done using a subset of the data that had complete information on building units and hence a household could be redefined as a building unit.

Neither the entire data set nor the subset with redefined households seems to be able to give proper estimates of ω (omega). The distribution for this parameter is wide, but it should be noted that it does not include 1 (0 on the log scale) as such, the transmission is not frequency dependent in the usual notation. We also used the subset with redefined households to fit for a categorical effect of household size, the results of which are shown in Figure A2. 28. The subset does not have enough information in it to narrow down on the effect of categorical household size, the effect of previous heterologous infection and the effect of high viral load. In fact, the latter distribution seems to have a reversed direction from previous results, implying high viral load reduces transmission. This is a curious result that perhaps further highlights the need to also use information on symptoms.

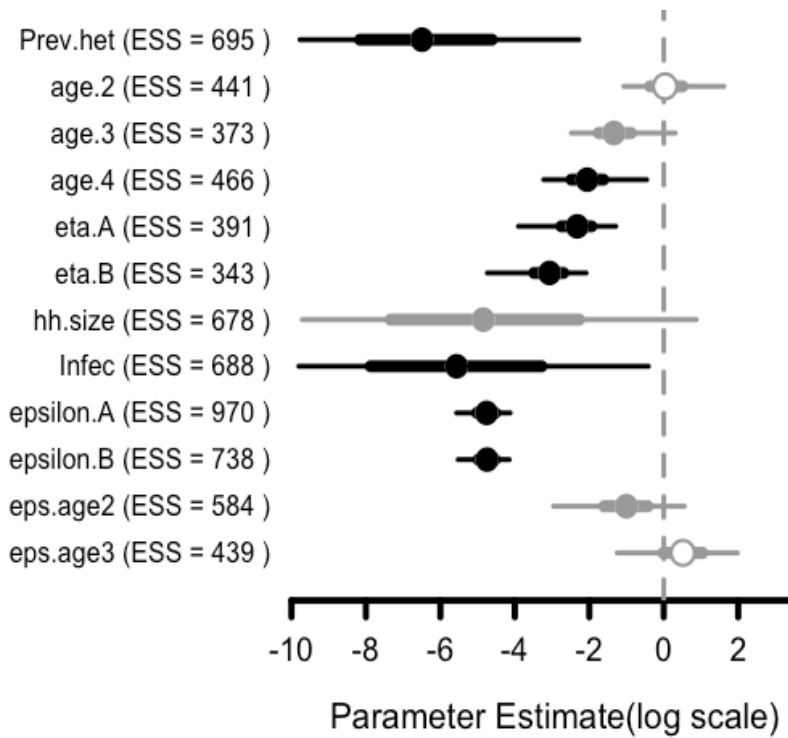


Figure A2. 28: Caterpillar plot showing the results of estimation when household size is treated as a categorical variable but with the definition of a household changed.

These results were obtained when fitting was done using a subset of the data that had complete information on building units and hence a household could be redefined as a building unit.

A3: Supplementary appendix for Paper 2.

A3.1. Normalizing the cluster specific background community exposure rate curves

We define a background cluster-specific rate of exposure, $f_c(t)$, for a cluster c at time t as

$$f_c(t) = \delta + \sum_{\substack{i \text{ shedding} \\ \text{RSV cluster } c}} e^{(t-\tau_{i,c})\beta}$$

Where δ is the basic risk and β is the rate of exponential decay related to the time since onset of a case shedding cluster type c . β is a measure of the rate at which the cluster might disappear from the community. τ_i is the onset time by person i .

To ensure that $\sum_{c'} f_c(t) = f_g(t)$ we need to normalize the cluster level curves such that their sum adds up to the group level curve. We describe how to do this using the illustration below:

P ₁		x	x	x	x	x			
P ₂				x	x	x	x	x	
P ₃			x	x	x	x	x	x	x
P ₄					x	x	x	x	x
	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	T ₈	T ₉

Say person P₁, P₂, P₃ and P₄ all have RSV A but there are 6 clusters in this particular configuration. The issue with multiple clusters showing up like above is that we get onsets where we previously had none which leads to $\sum f_c(t) > f_g(t)$. By looking at RSV, we have 4 onsets at T₂, T₃, T₄ and T₅ respectively. By cluster we have 6 onsets, with an extra one at T₅ and T₈. Up until T₄ $\sum f_c(t) = f_g(t)$, at T₅ we can normalize such that $f_g(t_5)$ is divided proportionally among the 4 cluster. In normalizing, the absolute value is reduced, but the clusters are weighed appropriately. At T₈, by looking at RSV we see no new onset, but by cluster we have one and so

$$f_g(t_8) = \delta + (e^{(t_8-t_2)\beta} + e^{(t_8-t_3)\beta} + e^{(t_8-t_4)\beta} + e^{(t_8-t_5)\beta})$$

since the most recent onset was at t_5 . At weighing, this value is divided proportionally among the red ($f_{red}(t_8) = \delta + e^{(t_8-t_2)\beta}$), black ($f_{black}(t_8) = \delta + e^{(t_8-t_4)\beta}$), green ($green(t_8) = \delta + e^{(t_8-t_5)\beta}$), dark blue ($f_{D.blue}(t_8) = \delta + e^{(t_8-t_3)\beta}$), purple $f_{purple}(t_8) = \delta + e^{(t_8-t_5)\beta}$ and light blue ($f_{L.blue}(t_8) = \delta + 1$) clusters. The equation for the normalized function $\hat{f}_c(t)$ is given as:

$$\hat{f}_c(t) = \left(\delta + \sum_{\substack{i \text{ shedding} \\ RSV \text{ cluster } c}} e^{(t-\tau_{i,c})\beta} \right) \times \left(\sum_{c \in C'} \left(\delta + \sum_{\substack{i \text{ shedding} \\ RSV \text{ cluster } c}} e^{(t-\tau_{i,c})\beta} \right) \right)$$

An example of the shapes of the background community rate of exposure curves is shown in Figure A3.1 for the 5 clusters in RSV A and FigureA3.2 for the 7 clusters in RSV B.

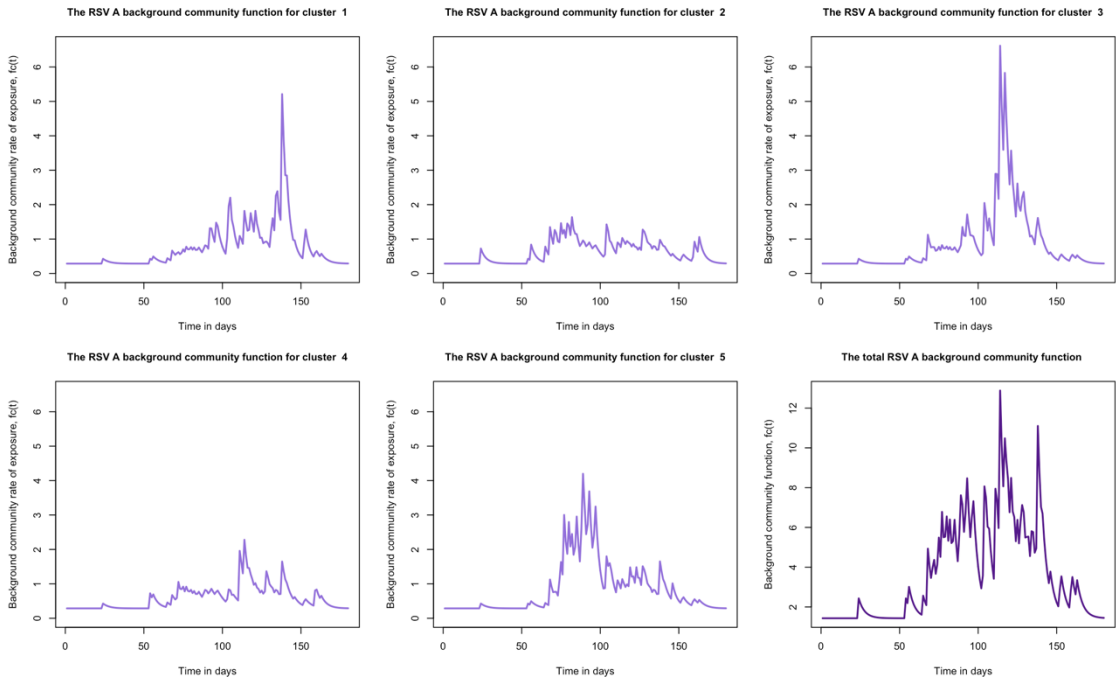


Figure A3. 1: The background cluster-specific rate of exposure curves for RSV A.
The normalized $f_c(t)$ curves are shown for the 5 different clusters and the group.

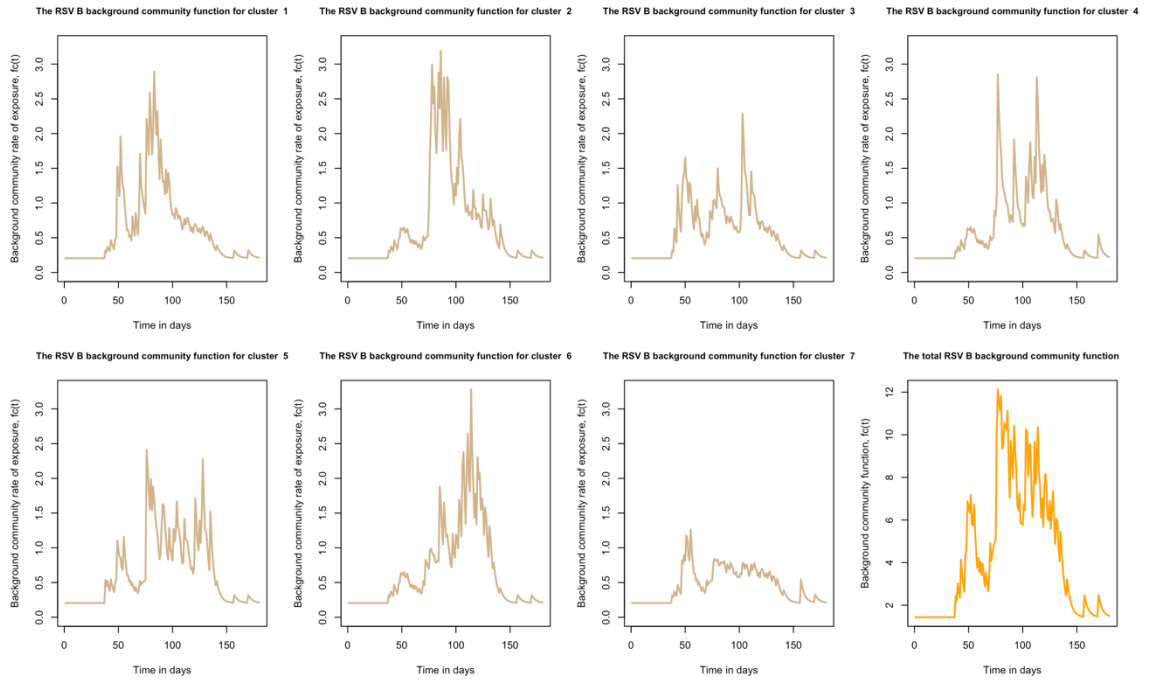


Figure A3. 2: The background cluster-specific rate of exposure curves for RSV B.
The normalized $f_C(t)$ curves are shown for the 7 different clusters and the group.

A3.2 Further details on the inference method (MH-MCMC)

A3.2.1. Metropolis-Hastings Markov Chain Monte Carlo (MH-MCMC) for parameter inference

The MH-MCMC algorithm is a popular first step for a situation where the target distribution is not simple and the dimension of the parameters is not small. As we have a total of 19 parameters this seemed like a natural starting point. For a given intractable target distribution, the MH-MCMC algorithm creates a chain of auto-correlated samples, for each desired parameter, whose equilibrium distribution is drawn from the desired target density. The samples that form part of the chain are proposed from a distribution q and are either accepted or rejected based on an acceptance probability ρ . A generic MH-MCMC algorithm is as follows:

For a desired target density $\pi(x)$, where x is the set of parameters, given x^n (the set of parameters at iteration n of the chain):

1. Generate $Y_n \sim q(y|x^n)$
2. Take the next set of parameters:

$$X^{n+1} = Y_n \text{ with acceptance probability, } \rho = \min \left\{ 1, \frac{\pi(Y_n)q(X_n|Y_n)}{\pi(X_n)q(Y_n|X_n)} \right\}$$

x_n otherwise

For a symmetric proposal distribution where $q(y/x) = q(x/y)$ the acceptance ratio

$\frac{\pi(Y_n)q(X_n|Y_n)}{\pi(X_n)q(Y_n|X_n)}$ reduces to $\frac{\pi(Y_n)}{\pi(X_n)}$ making the acceptance probability independent of $q(\cdot)$.

However the choice of $q(\cdot)$ does determine the performance of the algorithm as such $q(\cdot)$ has to be carefully chosen. The conditional probability $q(y/x)$ means that the samples in the MH-MCMC chain are dependent. Variations of the algorithm can use an independent proposal g such that $q(y/x)=g(y)$. The construction of an appropriate proposal distribution can be difficult as such, an alternative to doing this is to slowly approach the target distribution by exploring the parameter space close to current values of the MH-MCMC chain. This is what the random walk MH-MCMC does. The algorithm for this is:

1. Generate $Y_n = x_n + \varepsilon_n$, where $\varepsilon_n \sim g(\cdot)$.
2. Take the next set of parameters:

$$X^{n+1} = Y_n \text{ with acceptance probability, } \rho = \min \left\{ 1, \frac{\pi(Y_n)}{\pi(X_n)} \right\}$$

x_n otherwise

If $g(\cdot)$ is a uniform distribution then $Y_n \sim U(x^n - \delta, x^n + \delta)$, for $g(\cdot)$ a Normal distribution $Y_n \sim N(x_n, \sigma^2)$. For a pair (x_n, y_n) the acceptance ratio will be the same whether y_n came from a Uniform or Normal proposal distribution. However the choice of $g(\cdot)$ does determine the range of proposed values as such must be made such that the boundaries of the target distribution $\pi(x)$ are explored[3]. However, in practice an additional condition to accepting a proposed value is used to make sure that even low probability regions of the parameter space are explored and thus represented in the final equilibrium distribution. If $\rho \neq 1$, generate $r \sim Uniform(0,1)$, if $\rho > r$ then the proposed value is accepted.

A3.2.2 Our application of MH-MCMC

We denote the observed data as D , the augmented data as D_A and the set of parameters as φ . The target distribution is given as $p(\varphi|D, D_A) = P(D|D_A)L(\varphi|D, D_A)P(\varphi)$; $P(D|D_A)$ = probability of the observed data give the augmented data; $L(\varphi|D, D_A)$ = the likelihood of the parameters given the observed and augmented data; $P(\varphi)$ = the prior probability of the data. The augmented and observed data are independent and we have no information to inform what the

missing cluster ids could be, making every combination of D and D_A equally likely. Consequently, we did not include $P(D|A)$ when calculating the posterior probability.

The parameters will be updated first, followed by an update of the augmented data. We will assume weakly informative priors in the form of a normal distribution with mean 0 and a standard deviation of ~ 3 for the log of parameters. There is only one move to update the data with a probability of occurrence =1, i.e. the updates to D_A are carried out at every iteration. Given the significant number of uninformed outbreaks, for the same set of parameter values, the likelihood value (and subsequently the posterior value) can vary drastically with new configurations of the missing cluster ids. This is very likely to lead to the proposed change in D_A being rejected and if it is accepted subsequent updates to the parameter values might get rejected even when the standard deviation for the proposal distribution is small. As such, to mimic a gradual change in cluster configurations, at every iteration of the MCMC algorithm, the random allocation of cluster ids will be done for one household outbreak at a time.

A3.2.3.1. Choice of proposal distributions for the parameters

For the parameter set φ we will use a multivariate normal distribution as the proposal distribution. For iteration n in the chain a new set φ^* will be proposed such that $\varphi^* \sim Normal(\varphi^{n-1} | \Sigma)$. The choice of the variance-covariance matrix Σ will determine the size of the space that is explored and how fast the MCMC chain converges. This can be fixed at the start of the algorithm and regular manual checks conducted to make sure the chain is progressing well and modifying Σ if it is not, e.g. by making sure the acceptance rate is not too high (implying the standard deviation is too low and thus only the very close neighbours of a current value are being explored, leading to the acceptance ratio being high most of the time and hence more accepted values) or too low (implying the inverse problem). Alternatively the modification of Σ can be automated through an adaptive random walk MH-MCMC algorithm. There are several adaptation algorithms [4], we will chose one that learns from the empirical distribution of values up to the $(n-1)^{th}$ iteration to modify the Σ at iteration n . For samples $\{\varphi_1, \varphi_2, \varphi_3, \dots \varphi_{n-1}\}$ in the MCMC chain so far, at iteration n the proposal density $g(\cdot)$ is given by

$$g_n(\cdot) = (1 - \varepsilon)N(\varphi^{n-1} | 2.38^2 \Sigma_{n-1} / d) + \varepsilon N(\varphi^{n-1} | 0.1^2 \Sigma_0 / d)$$

Where:

ε = A small positive constant, chosen to be 0.05 as in [4].

Σ_{n-1} = The empirical variance-covariance matrix derived from samples
 $\{\varphi_1, \varphi_2, \varphi_3, \dots, \varphi_{n-1}\}$

d = The dimension of the parameter set

Σ_0 = The initial guess of the parameter variance-covariance matrix. This is usually a diagonal matrix of variances.

This notation means for a fraction of the time $(1 - \varepsilon)$, the proposal distribution will be $N(\varphi^{n-1} | 2.38^2 \Sigma_{n-1} / d)$ and the rest of the time it will be $N(\varphi^{n-1} | 0.1^2 \Sigma_0 / d)$. Prior to adaptation beginning at iteration n , the proposal distribution at iteration k is given by $g_k(.) = N(\varphi^{k-1} | 0.1^2 \Sigma_0 / d)$

A3.2.3.2 Pseudo algorithm for our implementation of MH-MCMC

For each MCMC chain

1. Set initial values for the parameters and assign cluster ids at random for the outbreaks with no sequence information (uninformed outbreaks).
2. For every iteration n
 - a. Update parameter values
 - i. Propose a new set of parameters by sampling from the proposal distribution: $\varphi^* \sim Normal(\varphi^{n-1} | \Sigma)$
 - ii. Calculate the acceptance probability $\rho(\varphi^{n-1}, \varphi^*) = \min \left\{ 1, \frac{p(\varphi^* | D, D_A^{n-1})}{p(\varphi^{n-1} | D, D_A^{n-1})} \right\}$
 - iii. If $\rho(\varphi^{n-1}, \varphi^*) > r \sim Uniform(0,1)$ update $\varphi^n = \varphi^*$ otherwise $\varphi^n = \varphi^{n-1}$
 - b. Update cluster id for a single uniformed outbreak
 - i. Randomly select an uniformed outbreak from the set of uninformed outbreaks, all with the same probability of being selected.

- ii. Given the present cluster id for the chosen outbreak C_r , randomly select a new cluster id from the set of all possible clusters excluding C_r .
- iii. With C_s as the proposed cluster id, the proposed change to the augmented data is accepted with probability

$$\rho'(D_A^{n-1}, D_A^*) = \min \left\{ 1, \frac{p(\varphi^n | D, D_A^*)}{p(\varphi^n | D, D_A^{n-1})} \frac{|C_r|}{|C_s| + 1} \right\}$$

Where $|C_r|$ is the number of household outbreaks in C_r in the present

permutation of the augmented data D_A^{n-1} and $|C_s|$ is the number of household outbreaks in C_s .

- iv. If $\rho'(D_A^{n-1}, D_A^*) > r' \sim Uniform(0,1)$ update D_A^n, D_A^* otherwise D_A^n, D_A^{n-1}

The correction factor $\frac{|C_r|}{|C_s|+1}$ is introduced into the acceptance ratio for a proposed change in cluster id because the proposal distributions are not symmetric. For an update of cluster id from C_s to C_r , the proposed change is uniformly distributed over the set of all household outbreaks/cases in cluster C_s that are part of the augmented dataset. Conversely the reverse move of a change of cluster id from C_r to C_s is uniformly distributed over the set of all household outbreaks/cases in cluster C_r that are part of the augmented dataset. As such, the proposal distributions are dependent on the number of uniformed household outbreaks in each cluster.

A3.3. Establishing the highest probability transmission source (HPTS)

We modified the likelihood to establish the most likely infection source (HPTS) for every case. For a given case i infected with RSV cluster c within group g , there are three possible sources of infection (Ω_i), either a sampled housemate, a sampled neighbour or an unknown community source. The total rate of exposure is given as:

$$\lambda_{i,h,c}(t) = S_{i,g}(t) \left[M_{i,h}(t) \sum_{j \neq i} HH_{Rate_{h,c,j \rightarrow i}}(t) + Comm_Rate_{i,c}(t) \right] \quad (1)$$

Where (as in the main text):

$S_{i,g}(t)$ is the factor modifying exposure by recent group specific infection history, age and group specific shedding status at time t

$Comm_Rate_{i,c}(t)$ is the cluster specific community (external to the household) exposure rate.

The probability of exposure is = $prob(\text{any exposure event}) * prob(\text{exposure to cluster } c)$

$$\alpha_{i,h,c}(t) = (1 - \exp^{-\sum_{c'} \lambda_{i,h,c}(t)}) * \left(\frac{\lambda_{i,h,c}(t)}{\sum_{c'} \lambda_{i,h,c}(t)} \right) \quad (2)$$

For a given source of infection Ω_i in the same household as i , the rate of exposure is given by:

$$\lambda_{\Omega_i \rightarrow i,h,c}(t) = S_{i,g}(t) [M_{i,h}(t) \times P_{\Omega_i \rightarrow i} \times \eta_g \times \psi_H(\text{Household_size}_i) \times \psi_{I,inf}(\text{Infectivity}_{\Omega_i,h,c}(t)) \times M_{\Omega_i,h}(t)]$$

For Ω_i not in the same household as i but among the sampled individuals, the rate of exposure is given by:

$$\lambda_{\Omega_i \rightarrow i,h,c}(t) = S_{i,g}(t) [\varepsilon_g \times \psi_{E,age}(\text{Age}_{group_{E,i}}) \times M_{i,h}(t) \times P_{\Omega_i \rightarrow i} \times \psi_{I,inf}(\text{Infectivity}_{\Omega_i,h,c}(t)) \times K(d_{i,\Omega_i}, \kappa) \times M_{\Omega_i,h}(t)]$$

For Ω_i an unknown source external to the household, the rate of exposure is given by:

$$\lambda_{\Omega_i \rightarrow i,h,c}(t) = S_{i,g}(t) [\varepsilon_g \times \psi_{E,age}(\text{Age}_{group_{E,i}}) \times f_c(t)]$$

The probability of transmission from a single source Ω_i at time t thus becomes:

$$Pr_{\Omega_i \rightarrow i,h,c}(t) = \frac{\lambda_{\Omega_i \rightarrow i,h,c}(t)}{\lambda_{i,h,c}(t)} \quad (I)$$

The likelihood function

The probability given in (I) is calculated for a time point $t = \text{exposure time of individual } i, t_i^E$. This is not observed in the data, however, given our assumption on the latency

duration, we can define a 6-day window of possibility. If case i had a shedding onset at time T_i^O , then the window for transmission is from day $(T_i^O - 5)$ to $(T_i^O - 0)$. For each day in the window, potential sources are identified based on shedding status and for each combination of infection source Ω_i and exposure date t_i^E , the likelihood is calculated using the formula below:

$$L(\varphi|\{T_i^O, t_i^E, \Omega_i\}) = \alpha_{i,h,c}(t) * \left(\prod_{t_i \neq t_i^E} (1 - \alpha_{i,h,c}(t)) \right) * (\theta_i(T_i^O - t_i^E)) * \left(\frac{\lambda_{\Omega_i \rightarrow i,h,c}(t_i^E)}{\lambda_{i,h,c}(t_i^E)} \right)$$

The first part of the product is the probability of infection with cluster c at time t_i^E , the second part is the probability of escaping infection at any time $t_i \neq t_i^E$, the third is the probability of a latency duration of length $(T_i^O - t_i^E)$ and the last term is the probability of transmission from source Ω_i to i .

Given the likelihood, the highest-probability-source is chosen as the infection source that gives the highest value of the likelihood.

A3.4. Details of the model using pathogen data identified at group resolution

The null model is similar in structure to the model of sequence data presented in the main text, however, there is no identification of the infecting pathogen at the cluster level, only at the group level. The rate of exposure to a particular RSV cluster g acting on a susceptible person i from household h at time t :

$$\lambda_{i,h,g}(t) = S_{i,g}(t) \left[M_{i,h}(t) \sum_{j \neq i} HH_{Rate}_{h,g,j \rightarrow i}(t) + Comm_{Rate}_{i,g}(t) \right] \quad (1)$$

Where:

$S_{i,g}(t)$ is the factor modifying exposure by recent group specific infection history, age and group specific shedding status at time t given by:

$$S_{i,g}(t) = \exp\left(\phi_{Y,hist}(Infection_History_i(t)) + \phi_{X,age}(Age_group_{S,i}) + \phi_{W,curr}(Shedding_status_i(t))\right)$$

$HH_Rate_{h,g,j \rightarrow i}(t)$ is the group specific within household exposure rate given by:

$$\begin{aligned} HH_Rate_{h,g,j \rightarrow i}(t) &= \eta_g \times \psi_H(Household_size_i) \times \psi_{I,inf}(Infectivity_{j,h,g}(t)) \\ &\times M_{j,h}(t) \end{aligned}$$

$Comm_Rate_{i,g}(t)$ is the cluster specific community (external to the household) exposure rate given by:

$$Comm_Rate_{i,g}(t)$$

$$= \varepsilon_g$$

$$\times \psi_{E,age}(Age_group_{E,i}) \left(\left(M_{i,h}(t) \sum_{\substack{j \neq i, \\ j \text{ not in} \\ i's \text{ house}}} Sampled_Neighbour_Rate_{h,g,j \rightarrow i}(t) \right) + f_g(t) \right)$$

Where:

$$Sampled_Neighbour_Rate_{h,g,j \rightarrow i}(t) = \psi_{I,g,j}(t) \times K(d_{i,j}, \kappa) \times M_{j,h}(t)$$

The background function $f_g(t)$ is derived the same way $f_c(t)$ is, as described in the main text. Since we do not use genetic distances in this version of the model, we do not estimate ϑ for $P_{j \rightarrow i} = \exp^{-d_{gen}(i,j)*\vartheta}$ or $P_{j \rightarrow i} = 1$ if $d_{gen}(i,j) \leq \vartheta, 0$ otherwise, making the total number of parameters 17.

Following from the rate of exposure is the probability of exposure given by:

$$\alpha_{i,h,g}(t) = (1 - \exp^{-\lambda_{i,h,g}(t)}) \quad (2)$$

The probability of onset is given as:

$$p_{i,h,g}(t) = \sum_{l=0}^L \theta_l \alpha_{i,h,g}(t-l)$$

Where L is the maximum latency period and θ_l is the probability that the latency period is exactly l days.

The likelihood for individual i 's data is given as:

$$L_i = \prod_g \left[\prod_{u \in U_{i,h,g}} p_{i,h,g}(u) \prod_{a \in A_{i,h,g}} (1 - p_{i,h,g}(a)) \right]$$

The total likelihood is thus given by the product of L_i over all the individuals in the data

$$L = \prod_i \left[\prod_g \left[\prod_{u \in U_{i,h,g}} p_{i,h,g}(u) \prod_{a \in A_{i,h,g}} (1 - p_{i,h,g}(a)) \right] \right]$$

A3.5. Results of the MCMC algorithm

The figures below show the evolution of the parameter value with increasing number of iterations for the model with pathogen identification at the genetic cluster level (cluster model) and at the group level (group model).

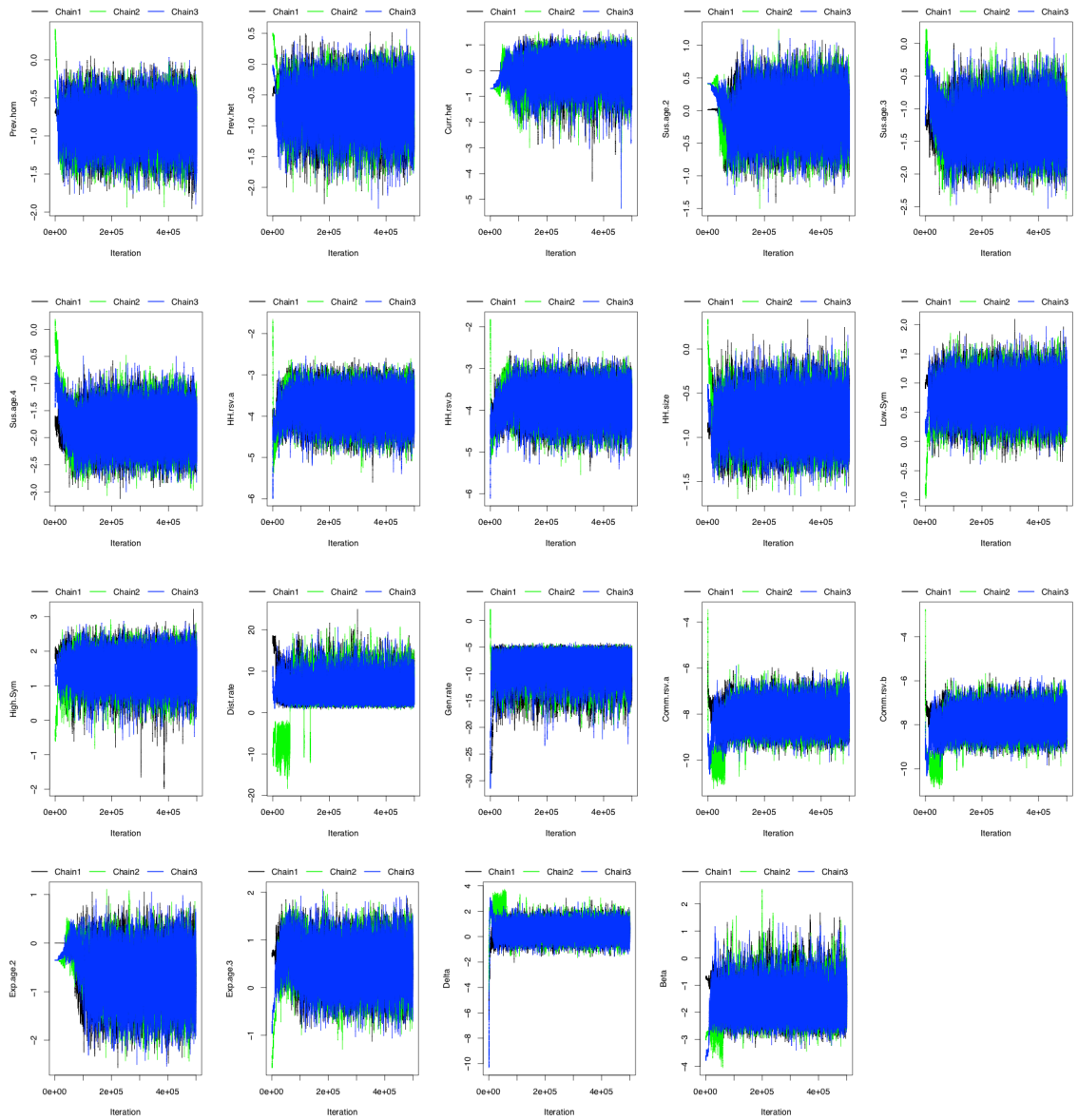


Figure A3. 3: Trace plots of parameters in the cluster model.

Three chains were initiated at different parameter values and these are shown in black (Chain 1), green (Chain 2) and blue (Chain 3) lines. The x-axis shows the iteration number, while the y-axis shows the log parameter value.

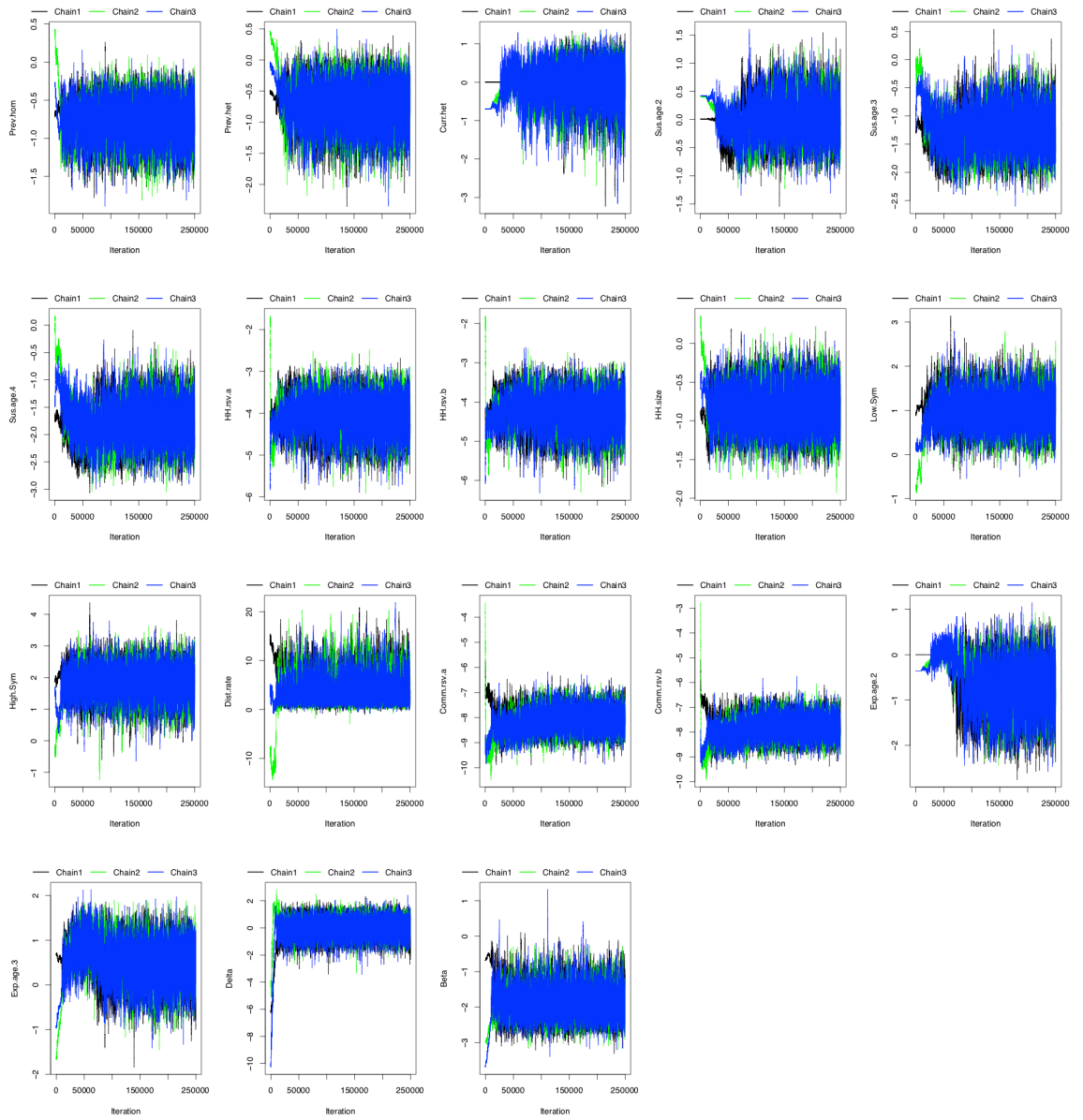


Figure A3. 4: Trace plots of parameters in the group level data model.

Three chains were initiated at different parameter values and these are shown in black (Chain 1), green (Chain 2) and blue (Chain 3) lines. The x-axis shows the iteration number, while the y-axis shows the log parameter value.

To confirm convergence observed in the trace plots, we calculated the Gelman-Rubin-Brooks statistic and the effective sample size. When using the GRB statistic, convergence is said to have occurred if the ratio of pooled/within chain variance is close to 1. The GRB statistic assumes that the target distribution is Normal. The plot below shows the value of the GRB statistic as the number of iterations increases for each parameter. This is to check whether a value close to one was reached by chance or if the trend line had truly stabilized close to 1.

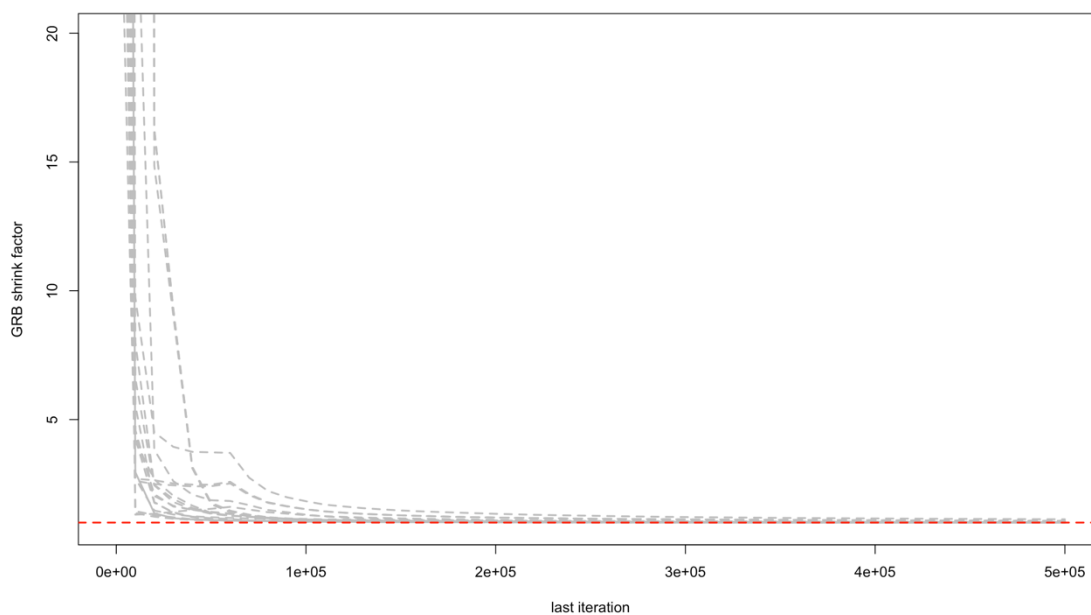


Figure A3. 5: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the cluster level data model and the dashed red line shows the value 1.

The point estimated of the GRB and the values of the ESS after burn in are given in the table below.

Table A3. 1: The value of the GRB statistic (to 3 significant figures) and the ESS after burn-in are shown for the parameters in the cluster level data model.

Parameter	Point estimate GRB statistic	ESS
Prev.hom	1	10607
Prev.het	1	10073
Curr.het	1.01	7131
Sus.age.2	1.01	9154
Sus.age.3	1.02	9771
Sus.age.4	1.02	10384
HH.rsv.a	1	9476

HH.rsv.b	1.01	9765
HH.size	1	10147
Low.Sym	1.02	9987
High.Sym	1.01	9774
Dist.rate	1.16	10455
Gen.rate	1.04	10436
Comm.rsv.a	1.09	7847
Comm.rsv.b	1.09	7823
Exp.age.2	1	8432
Exp.age.3	1.01	9863
Delta	1.04	7908
Beta	1.03	6678

The mGRB is 1.07 and the mESS is 10008.

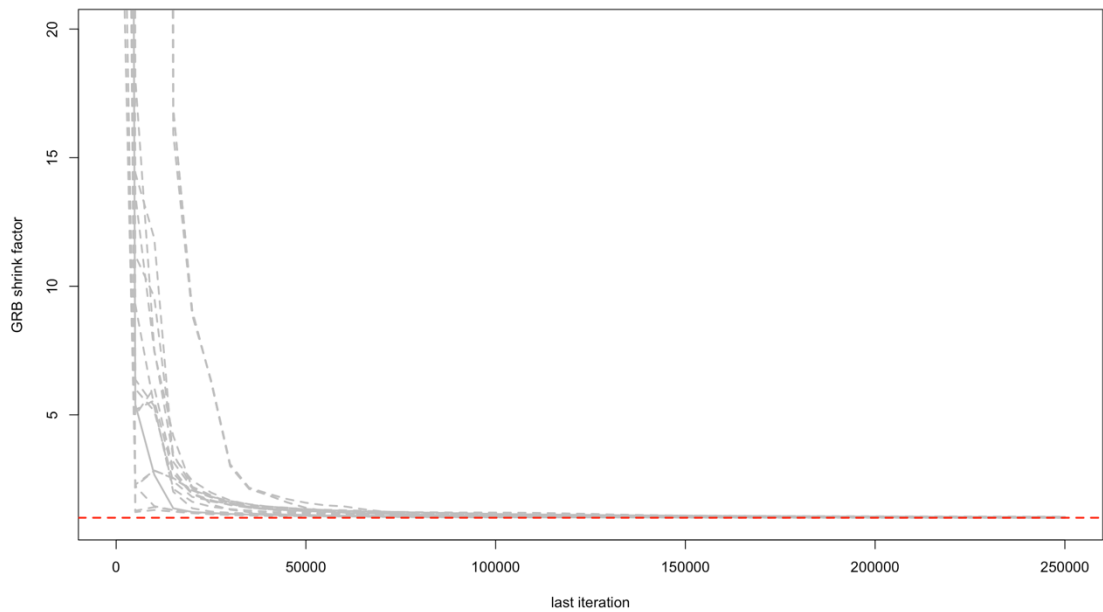


Figure A3. 6: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the group level data model and the dashed red line shows the value 1.

Table A3. 2: The value of the GRB statistic (to 3 significant figures) and the ESS after burn-in are shown for the parameters in the group level data model.

Parameter	Point estimate GRB statistic	ESS
Prev.hom	1.01	3713
Prev.het	1.02	3978
Curr.het	1.07	2309
Sus.age.2	1.02	2998
Sus.age.3	1.03	3617
Sus.age.4	1.04	3694
HH.rsv.a	1.01	3426
HH.rsv.b	1.01	3361
HH.size	1.02	3673
Low.Sym	1.04	3957
High.Sym	1.03	3744
Dist.rate	1.07	3374
Comm.rsv.a	1.05	4069
Comm.rsv.b	1.05	4093
Exp.age.2	1.02	2858
Exp.age.3	1.02	3476
Delta	1.04	5331
Beta	1.04	3873

The mGRB is 1.09 and the mESS is 4146.

As a rule of thumb, a GRB of <1.1 is generally considered good, as such, it is safe to conclude that there was convergence.

A3.6. Model validation

The results of the model fitting are the posterior parameter distribution and corresponding augmented data for the cluster ids of cases with no genetic information. A simulation based on a set of parameter values will also be based on the

corresponding augmented data which will be used to derive a complete set of shedding profiles from the observed data. A single shedding profile is a combination of duration of shedding, viral loads and symptom status, and genetic cluster. The simulation pseudo code per simulation is as follows:

1. Initiate system such that everyone one is susceptible to RSV.
2. At every time step keep track of the following variables:
 - a. Exposure status (by RSV cluster)
 - b. Shedding status by group
 - c. Shedding status by genetic cluster
 - d. Infectiousness status (combination of viral load and symptom status)
 - e. Infection history (by RSV group)
 - f. The background rate of exposure from the community
3. At every time step:
 - a. Update the background community function to reflect any new shedding onsets
 - b. Calculate the cluster specific rate of exposure, $\lambda_{i,h,c}(t)$, as defined in the main text.
 - c. Determine the number of group specific transmission events E_g where

$$E_g = \text{Poisson} \left(\sum_{i \in S_{E_g}} P_{E_g,i} \right)$$

S_{E_g} = set of all individuals susceptible to infection event E_g .

$P_{E_g,i}$ = probability of person i experiencing event E_g

$$P_{E_g,i} = \sum_{\substack{c = \text{clusters} \\ \text{in } g}} \left((1 - \exp^{-\sum_{c'} \lambda_{i,h,c}(t)}) * \left(\frac{\lambda_{i,h,c}(t)}{\sum_{c'} \lambda_{i,h,c}(t)} \right) \right)$$

Where $\lambda_{i,h,c}(t)$ = rate at which person i is exposed to infection of cluster type C .

- d. Given the number of group specific transmission events, determine the cluster id of each through weighted sampling. E.g. if $E_g = 4$ and $\mathbf{c} = \{1,2,3\}$ are the cluster ids in the group, the probability of a case being any one if the three clusters is:

$$\left\{ \frac{\lambda_{h,1}(t)}{\sum_{c'} \lambda_{h,c}(t)}, \frac{\lambda_{h,2}(t)}{\sum_{c'} \lambda_{h,c}(t)}, \frac{\lambda_{h,3}(t)}{\sum_{c'} \lambda_{h,c}(t)} \right\}, \text{ for } \lambda_{h,1}(t) = \sum_i \lambda_{i,h,c}(t)$$

- e. Determine who experiences each cluster specific transmission event. For a given event, order individuals capable of experiencing the event. For a given person p to experience the event, the following inequality has to be satisfied.

$$\sum_{i=1}^{i \leq p-1} P_{E_c,i} < \left(\text{RAND} \times \sum_{i \in S_{E_c}} P_{E_c,i} \right) \leq \sum_{i=1}^{i \leq p} P_{E_c,i}$$

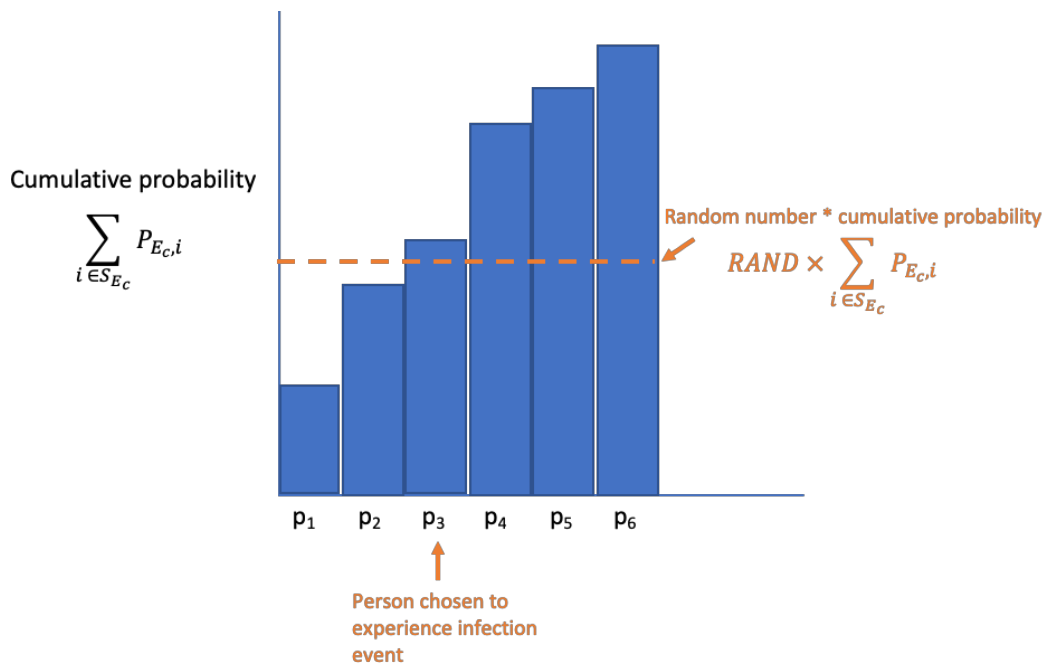
Where:

$$P_{E_c,i} = (1 - \exp^{-\sum_{c'} \lambda_{i,h,c}(t)}) * \left(\frac{\lambda_{i,h,c}(t)}{\sum_{c'} \lambda_{i,h,c}(t)} \right)$$

S_{E_c} = all individuals susceptible to infection of cluster type c .

RAND = a random number between (but not including) 0 and 1.

This is illustrated in the figure below.



Repeat this until the required number of events

- f. For each individual experiencing a transmission event, assign a latency duration and shedding profile by sampling from the relevant empirical distributions. The empirical latency distribution is the same as was used in estimating the parameters and is homogeneous for every individual. Shedding profiles are derived from the observed data and a combination of duration of shedding, viral loads and symptom status, and genetic cluster. The shedding profiles are grouped by age in the following 4 groups <1,1-5, 5-15 and ≥ 15 years. Once latency durations and shedding profiles have been assigned, the state variables for each individual are updated accordingly.

To explore how much variation there can be in the simulations from a single parameter set, a set of 12 parameter set samples were used, and for each set, 100 simulations were run, giving a total of 1200 simulations. We then sampled 100 parameter sets and run single simulations from each to explore between-parameter-set variation. The results of the simulations are presented in the form of epidemic curves and summary measures that are used to compare the main features of the outbreak. The summary measures shown in the subsequent figures are: total number of people infected, the proportion of cases that had multiple onsets and the number of cases in the first and last week of the observation/simulation period.

The results of the simulation are shown in Figure 4. 7 and Figure 4. 8 in the main text and Figure A3. 7, Figure A3. 8 and Figure A3. 9.

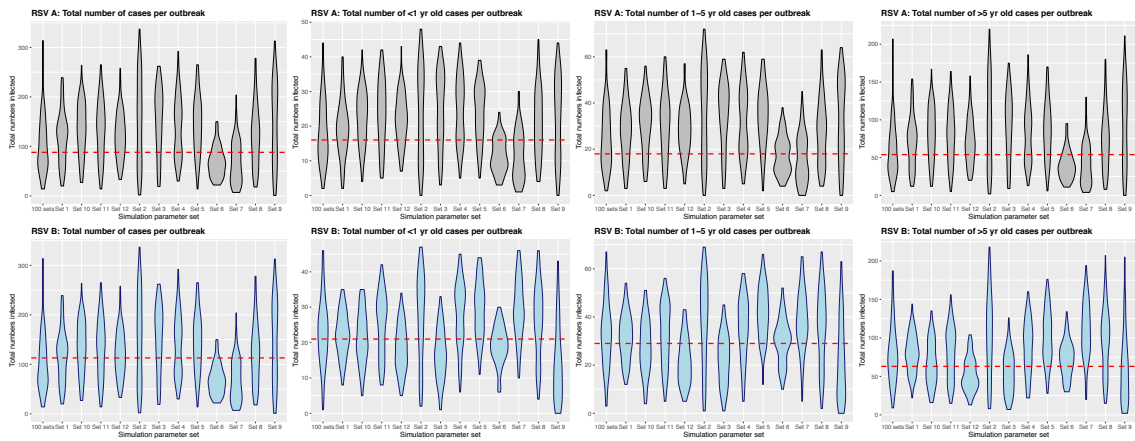


Figure A3. 7: Violin plots showing the distribution of the total number of people infected in the simulations by RSV group and age.

Each panel shows the distribution of the total numbers infected in the simulations run using 12 different parameter sets (violin plots) compared to the total number from the observed data (dashed red line). The y-axis shows the total number and the x-axis is labelled by parameter set used. Top row: RSV A results for all the cases (1st column), cases < 1 year old (2nd column), cases between 1-5 years old (3rd column) and cases > 5 years old (4th column). Bottom row: RSV B results. Violin plots are a combination of box plots and density distributions, the shapes should therefore be interpreted as density plots while the ranges should be interpreted as the tips of whiskers in a box and whisker plots.

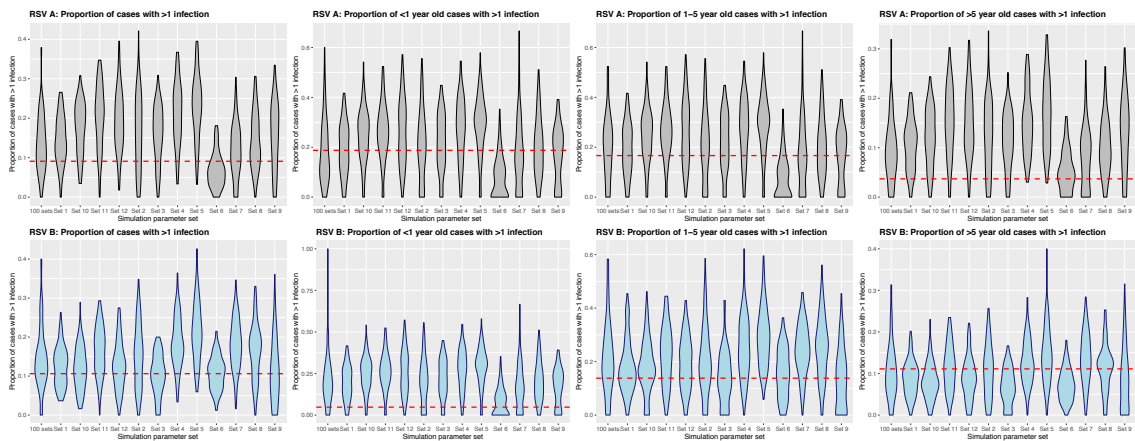


Figure A3. 8: Violin plots showing the distribution of the proportion of cases that had multiple onsets in the simulations by RSV group and age.

Each panel shows the distribution of the proportion of cases that had multiple onsets in the simulations run using 12 different parameter sets (violin plots) compared to the proportion from the observed data (dashed red line). The y-axis shows the proportion

and the x-axis is labelled by parameter set used. Top row: RSV A results for all the cases (1st column), cases < 1 year old (2nd column), cases between 1-5 years old (3rd column) and cases > 5 years old (4th column). Bottom row: RSV B results.

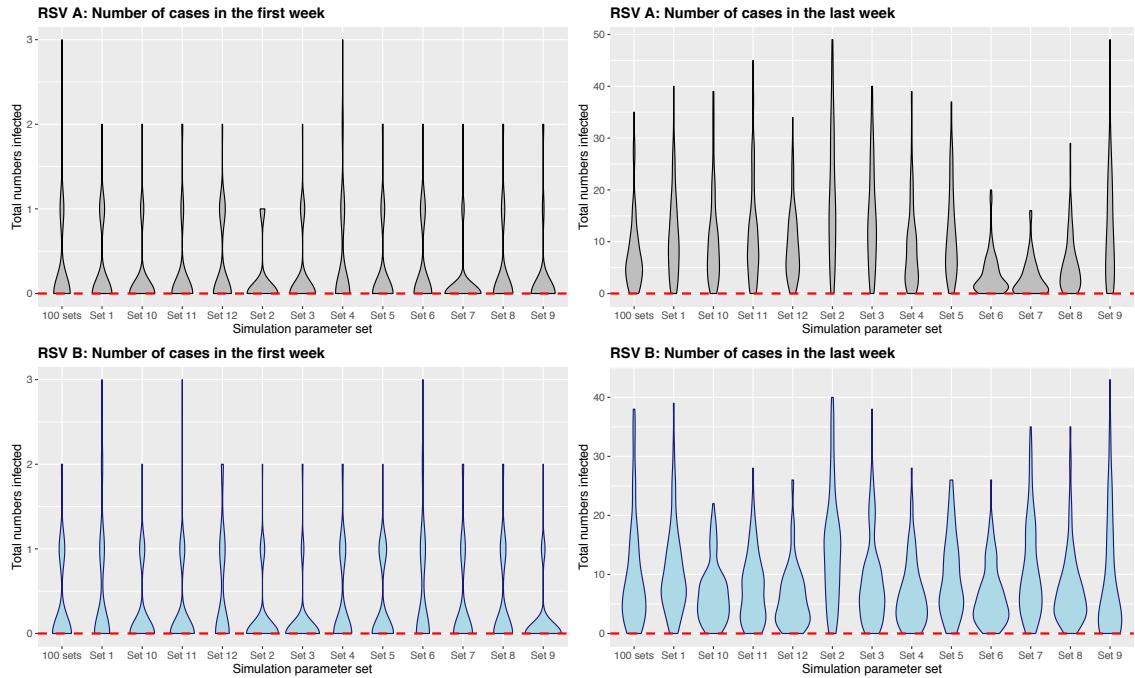


Figure A3. 9: Violin plots showing the distribution of the number of cases in the first (1st column) and last (2nd column) week of the observation/simulation period in the simulations by RSV group.

The y-axis shows the total number of people infected and the x-axis is labelled by parameter set used. The dashed red line shows what was observed in the data, i.e. there were no cases observed in the first and last week of the 180-day observation period.

A4: Supplementary appendix for Paper 3.

A4.1 Extra results

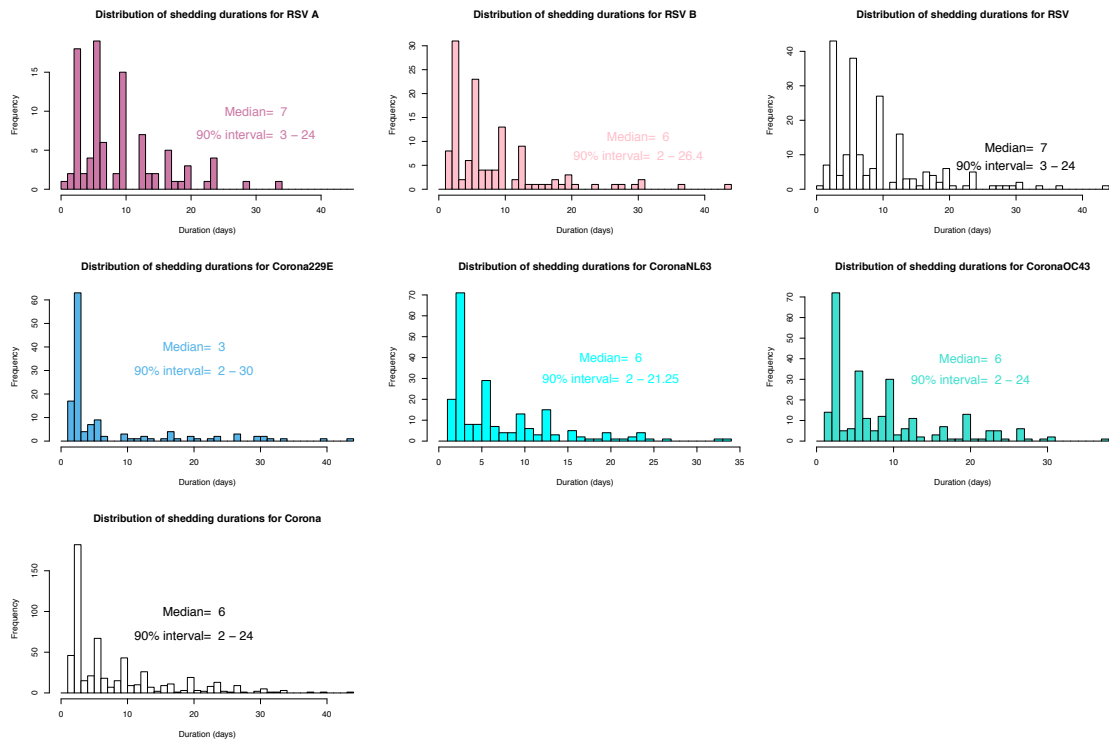


Figure A4. 1: Distributions of shedding durations for the different infectious agents.

Each panel shows data from the 5 different infectious agents and the white histograms show the pathogen level distribution of durations. The medium and 90% interval are given in text in each panel.

Table A4. 1: Results of parameter estimation using data identified at the pathogen level and group level.

Pathogen level		Group level	
Parameter	Median (95% CrI)	Parameter	Median (95% CrI)
risk.rsv.prev.rsv	0.599 (0.395, 0.878)	risk.rsva.prev.rsva	0.598 (0.265, 1.13)
		risk.rsvb.prev.rsvb	0.589 (0.317, 1.01)

		risk.rsva.prev.rsvb	0.591 (0.302, 1.06)
risk.rsv.prev.corona	1.33 (1.14, 1.57)	risk.rsva.prev.229e	0.698 (0.383, 1.17)
		risk.rsva.prev.nl63	1.15 (0.846, 1.53)
		risk.rsva.prev.oc43	1.05 (0.728, 1.45)
		risk.rsvb.prev.229e	1.17 (0.811, 1.64)
		risk.rsvb.prev.nl63	1.14 (0.828, 1.54)
		risk.rsvb.prev.oc43	1.81 (1.4, 2.34)
risk.corona.prev.corona	0.843 (0.706, 1)	risk.229e.prev.229e	0.724 (0.435, 1.15)
		risk.229e.prev.nl63	1.1 (0.806, 1.46)
		risk.229e.prev.oc43	0.784 (0.576, 1.05)
		risk.nl63.prev.nl63	0.617 (0.438, 0.844)
		risk.nl63.prev.oc43	1.16 (0.914, 1.48)
		risk.oc43.prev.oc43	0.58 (0.413, 0.79)
risk.rsv.curr.corona	1.09 (0.786, 1.46)	risk.rsva.curr.229e	1.996 (0.841, 3.95)
		risk.rsva.curr.nl63	0.741 (0.281, 1.76)
		risk.rsva.curr.oc43	0.733 (0.418, 1.21)
		risk.rsvb.curr.229e	1.15 (0.375, 2.53)
		risk.rsvb.curr.nl63	2.3 (0.774, 4.31)
		risk.rsvb.curr.oc43	0.804 (0.321, 1.57)
		risk.rsva.curr.rsvb	1.98 (0.831, 3.95)
		risk.229e.curr.nl63	1.04 (0.465, 1.99)
		risk.229e.curr.oc43	1 (0.508, 1.79)

		risk.nl63.curr.oc43	0.799 (0.387, 1.41)
HH.rsv	0.00387 (0.00291, 0.00508)	HH.rsva	0.00544 (0.00379, 0.00758)
		HH.rsvb	0.00408 (0.00282, 0.00555)
HH.corona	0.00636 (0.00518, 0.00755)	HH.229e	0.00795 (0.00577, 0.0108)
		HH.nl63	0.0117 (0.00939, 0.0145)
		HH.oc43	0.00547 (0.00428, 0.00681)
Comm.rsv	0.000296 (0.000146, 0.000798)	Comm.rsva	0.000186 (0.000101, 0.000317)
		Comm.rsvb	0.000217 (0.00012, 0.000357)
Comm.corona	0.000395 (0.000199, 0.00119)	Comm.229e	0.000242 (0.000132, 0.000398)
		Comm.nl63	0.000181 (0.0000996, 0.000297)
		Comm.oc43	0.000297 (0.000167, 0.000485)
Delta	1.55 (0.567, 3.94)	Delta	1.55 ()
Beta	0.338 (0.148, 2.19)	Beta	0.294 (0.148, 0.557)

mu.rsv	3.05 (2.19, 3.7)		
sigma.rsv	0.683 (0.323, 1.21)		
mu.corona	2.95 (2.4, 3.62)		
sigma.corona	0.712 (0.518, 0.938)		

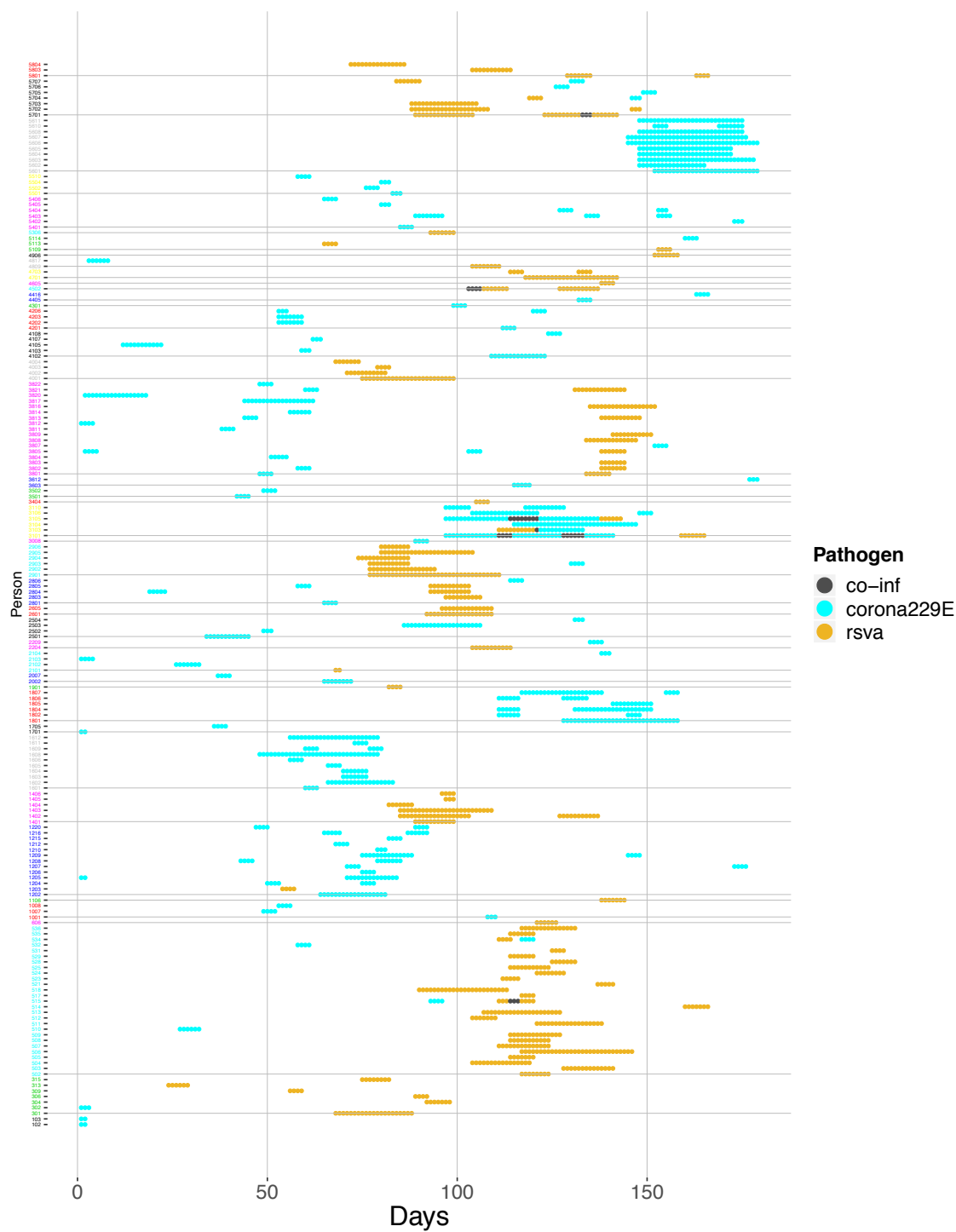


Figure A4. 2: Distribution of shedding episodes for coronavirus 229E and RSV A by household and time.

The x-axis shows the time in days while the y-axis shows the individuals, where each notch is a single individual. The horizontal lines demarcate the different households.

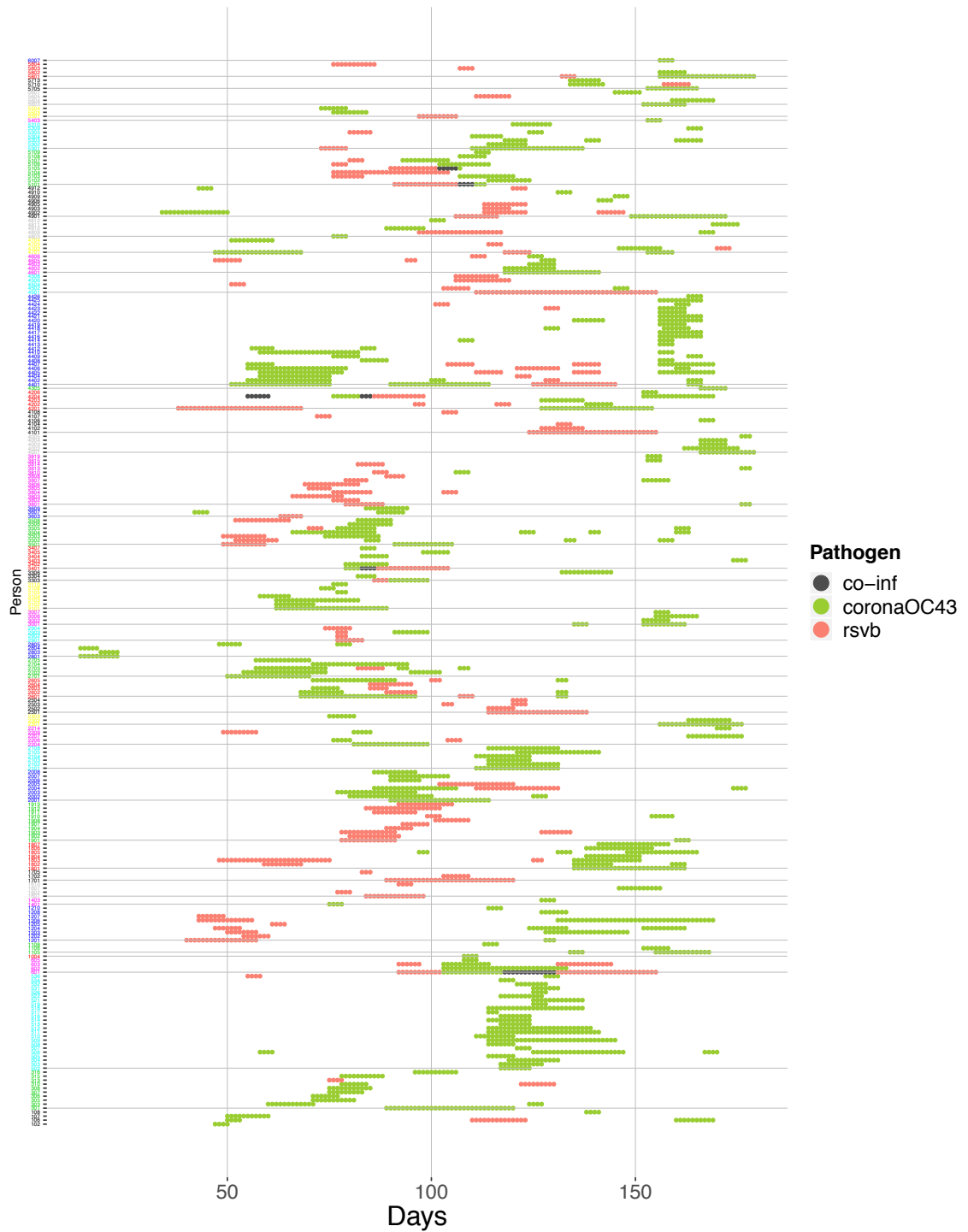


Figure A4. 3: Distribution of shedding episodes for coronavirus OC43 and RSV B by household and time.

The x-axis shows the time in days while the y-axis shows the individuals, where each notch is a single individual. The horizontal lines demarcate the different households.

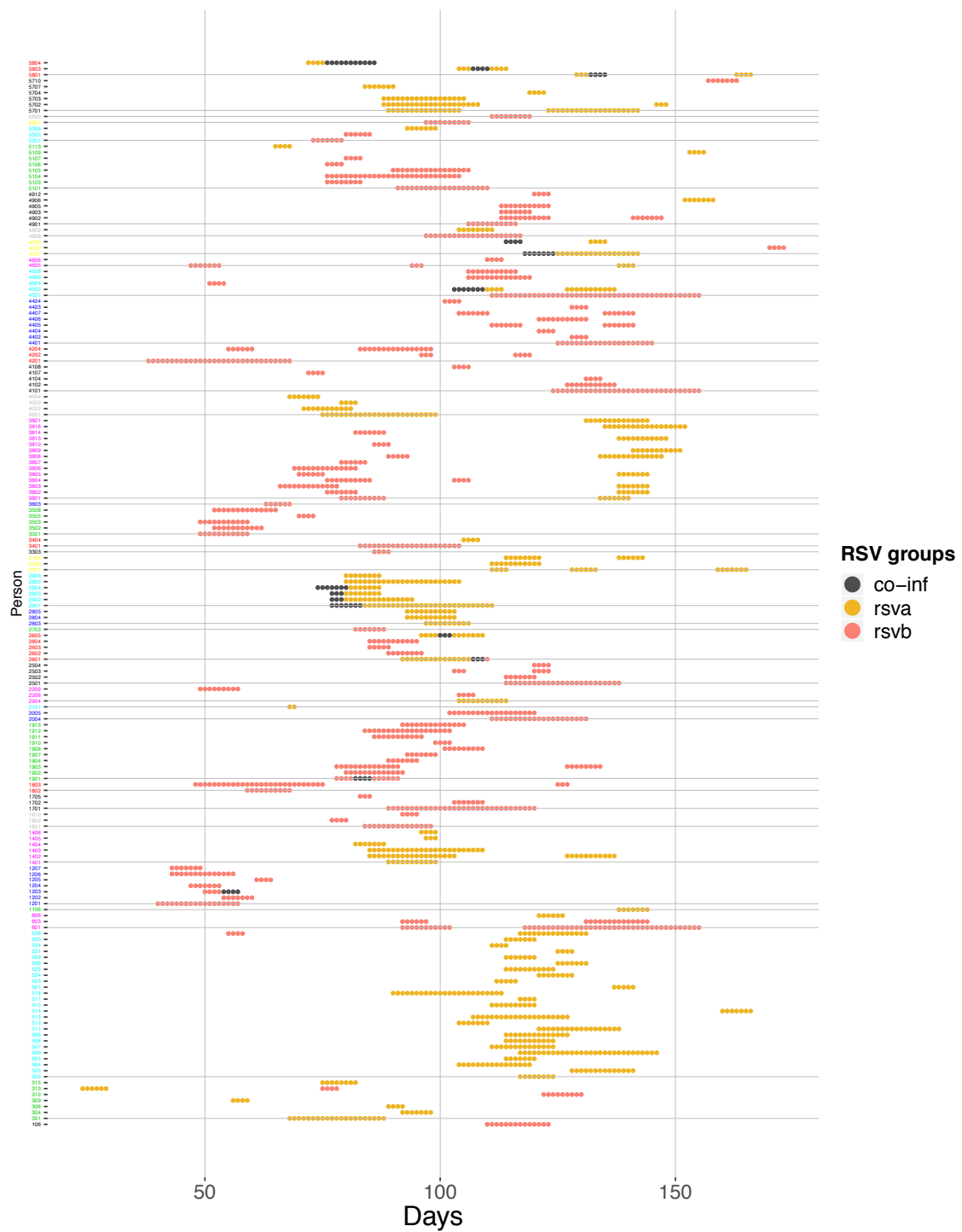


Figure A4. 4: Distribution of shedding episodes for RSV A and RSV B by household and time.

The x-axis shows the time in days while the y-axis shows the individuals, where each notch is a single individual. The horizontal lines demarcate the different households.

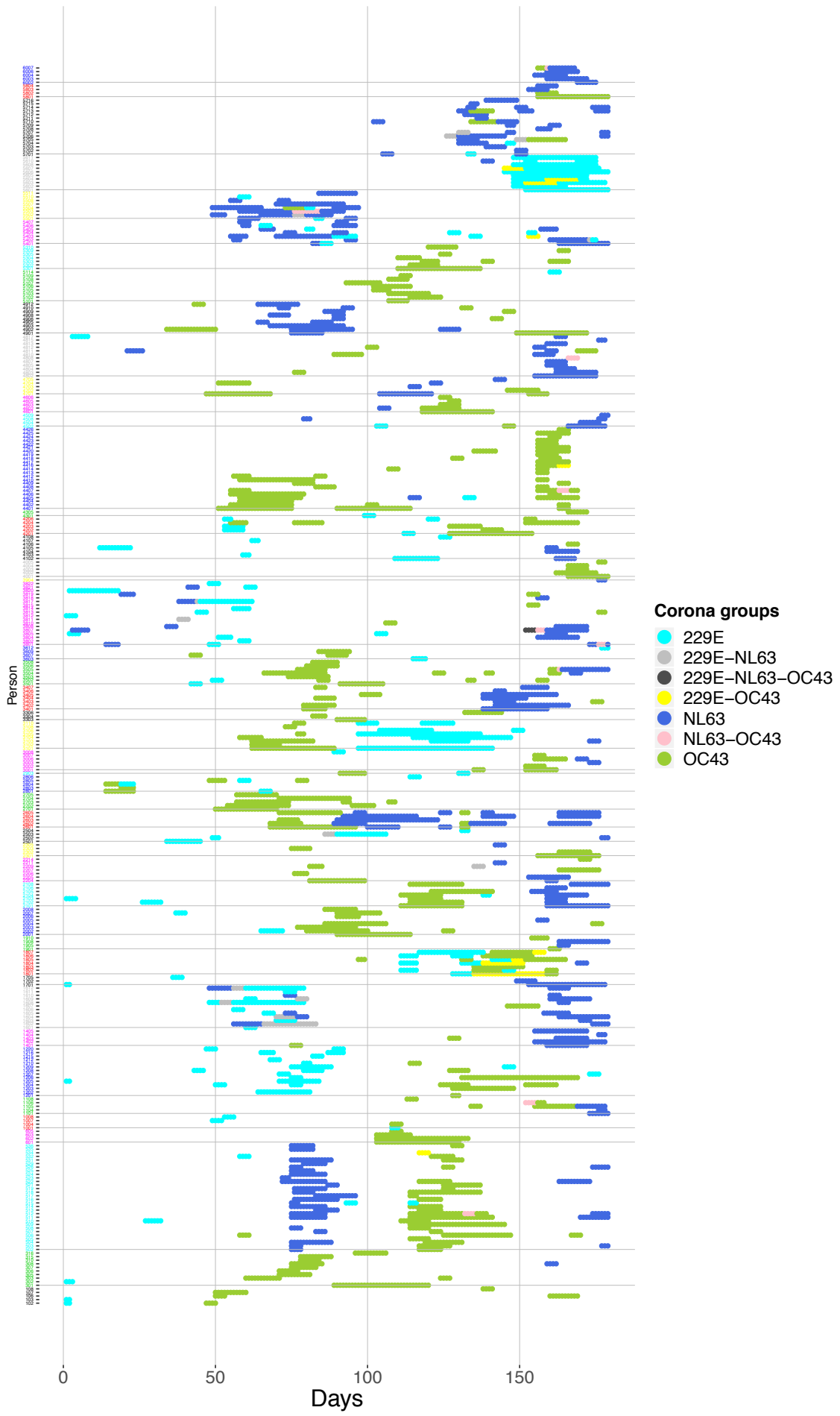


Figure A4. 5: Distribution of shedding episodes for coronavirus 229E, NL63 and OC43 by household and time.

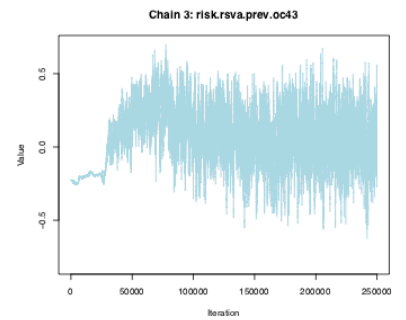
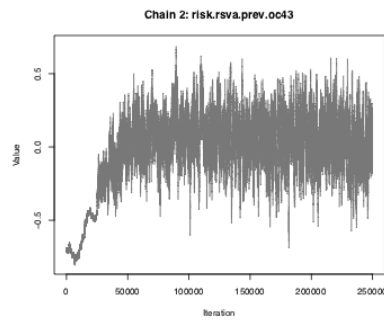
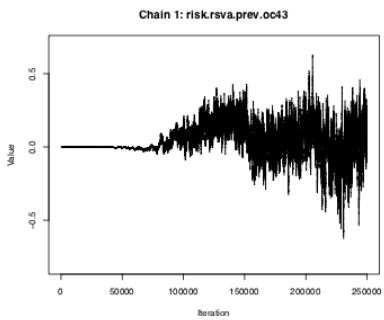
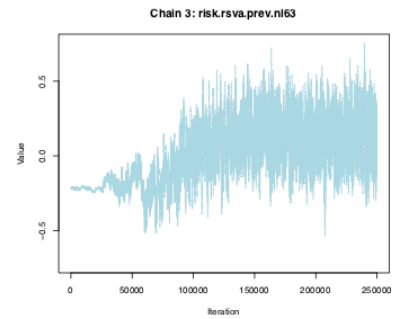
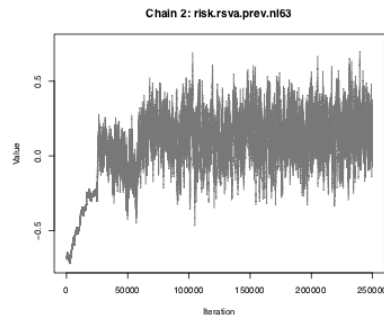
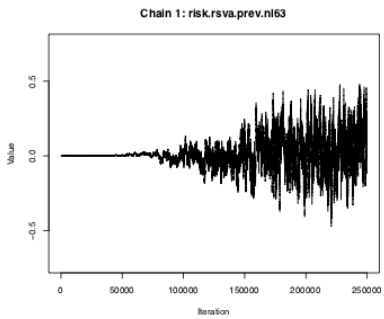
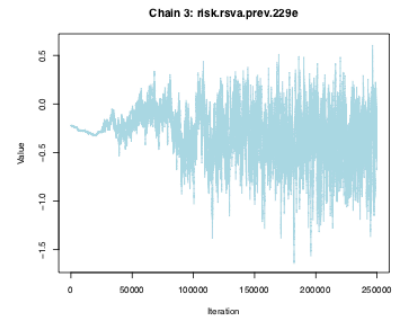
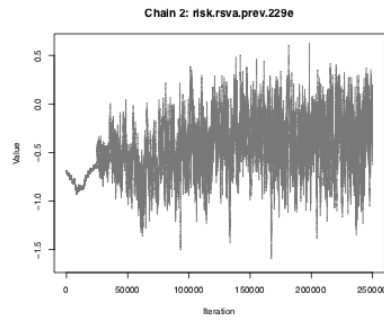
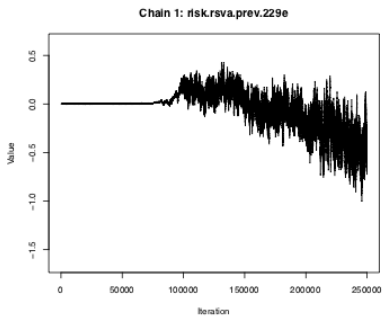
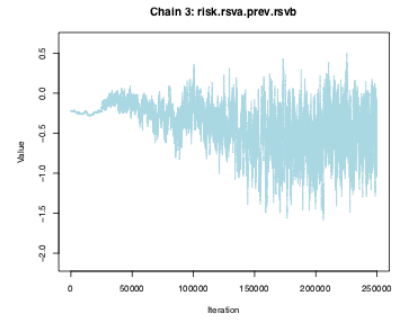
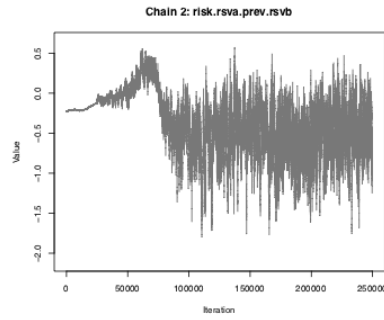
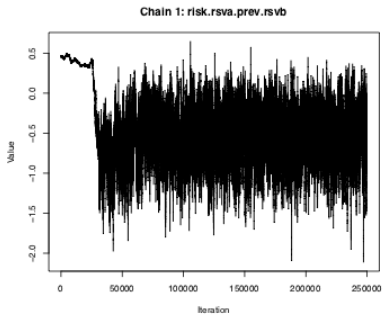
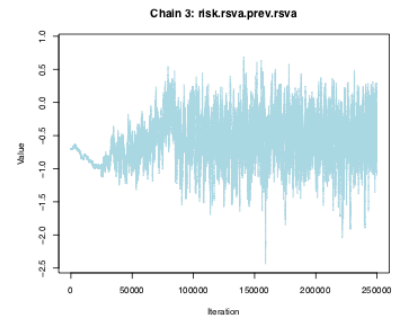
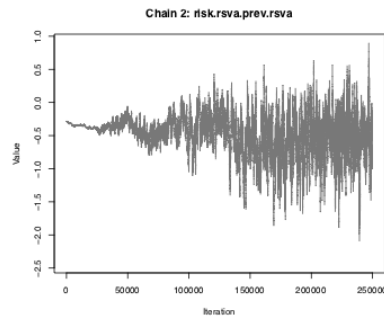
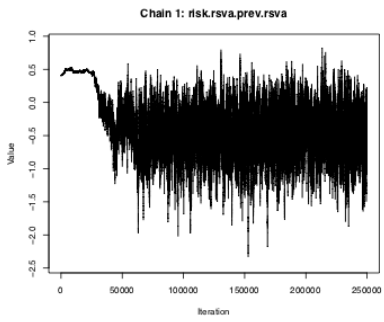
The x-axis shows the time in days while the y-axis shows the individuals, where each notch is a single individual. The horizontal lines demarcate the different households.

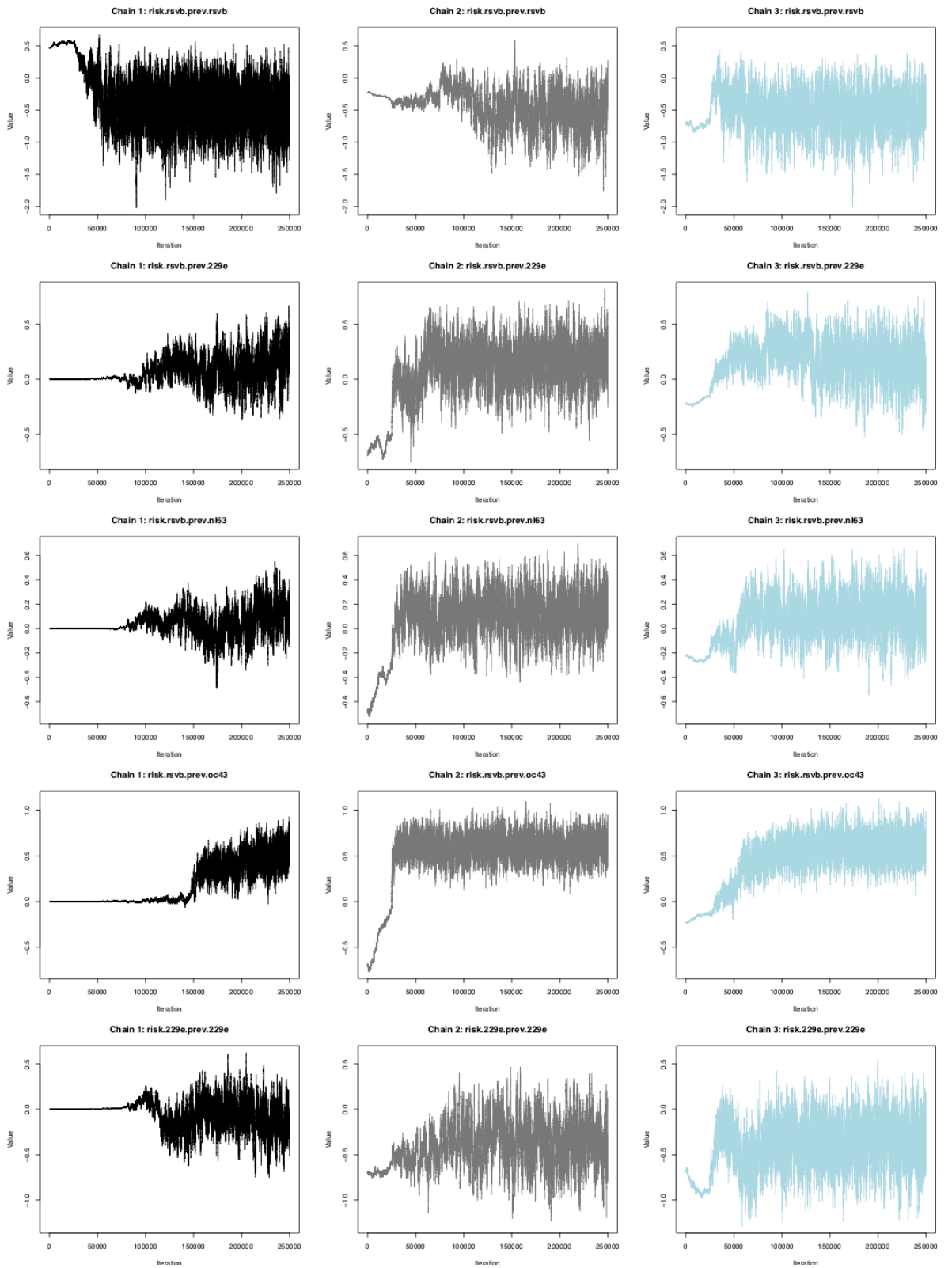
A4.2. Results of the MCMC algorithm

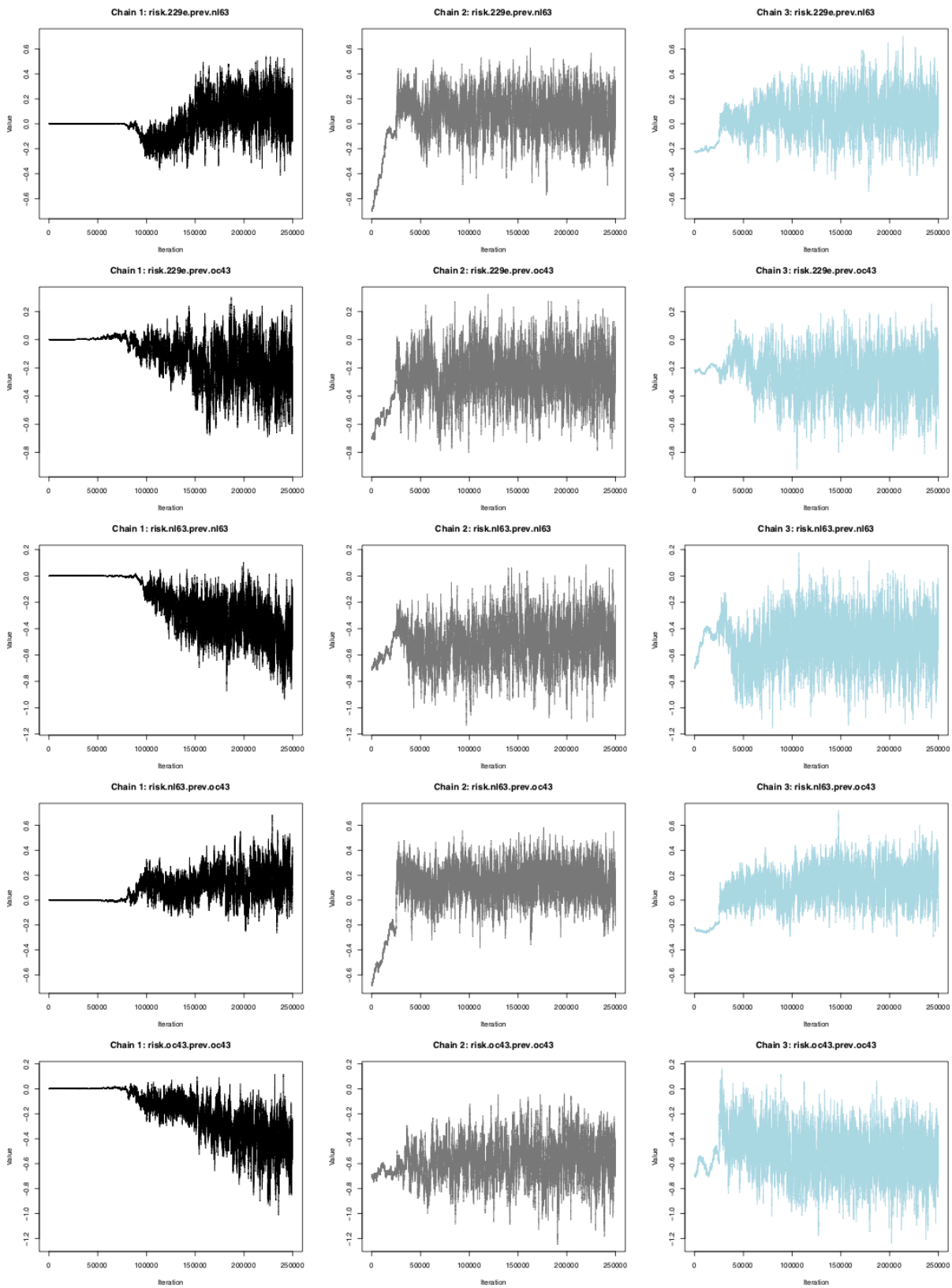
In Chapter 4, there were 4 different model fits; a fit of the multi-pathogen model to data with pathogen identification at the group level, a fit of the multi-pathogen model to data with pathogen identification at the pathogen level, a fit of the single-pathogen model to RSV data with pathogen identification at the group level, a fit of the single-pathogen model to hCoV data with pathogen identification at the group level. The following sections show the parameter trace plots, GRB statistic values and ESS values for each of the 4 fits.

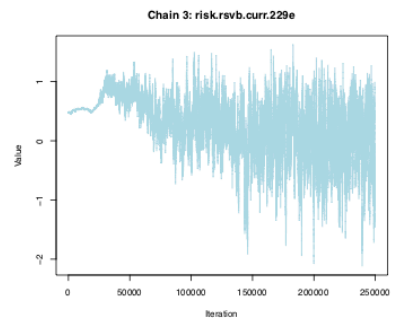
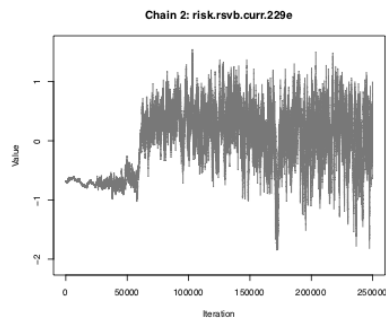
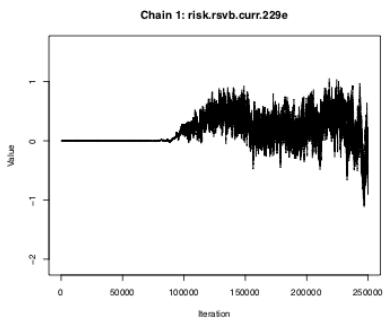
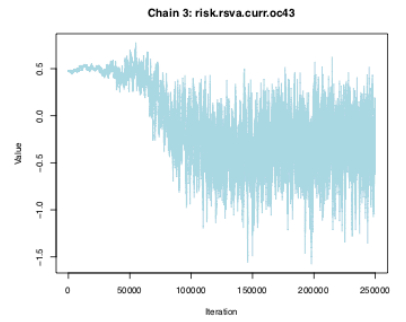
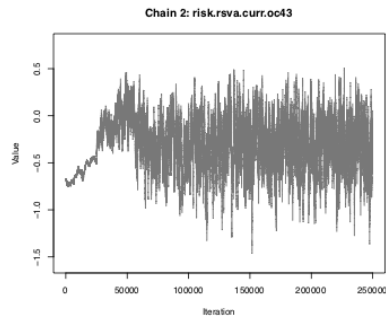
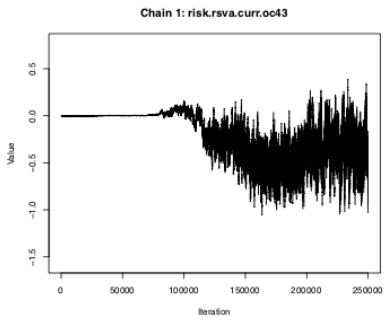
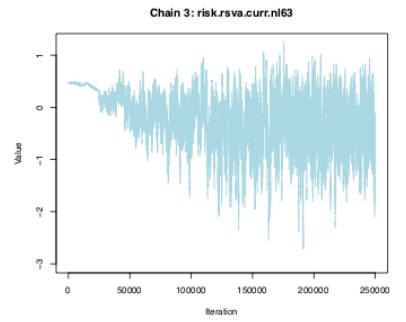
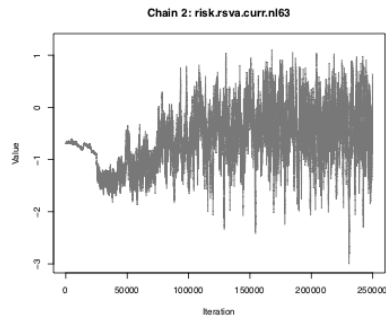
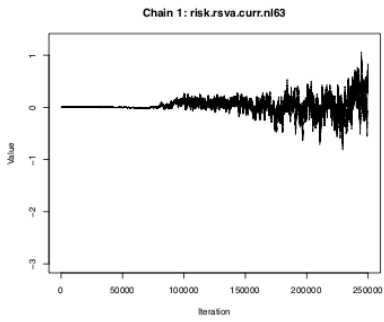
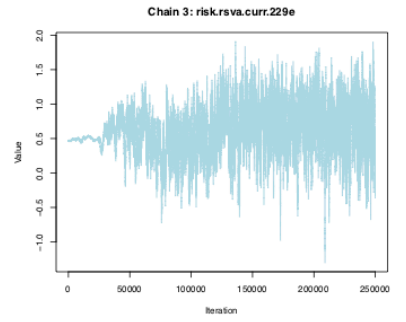
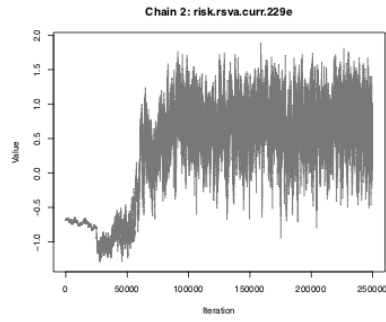
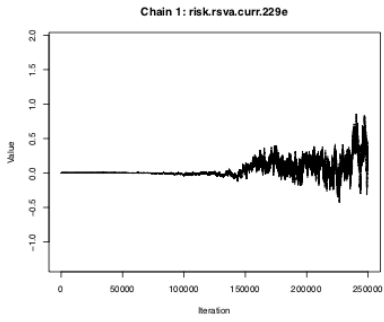
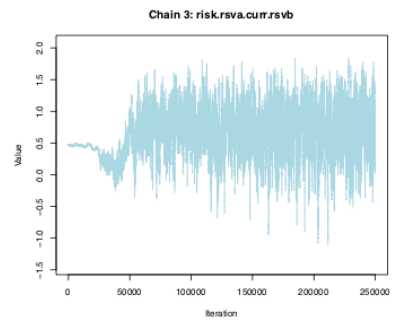
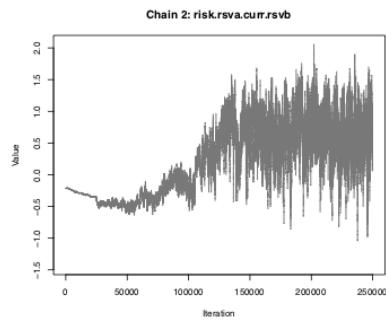
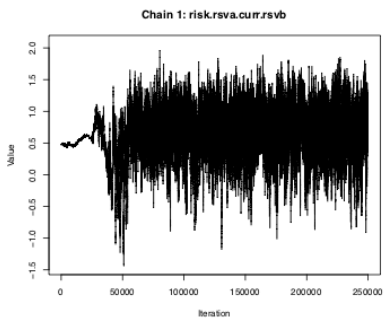
A4.2.1. Multi-pathogen model fit to data with pathogen identification at the group level

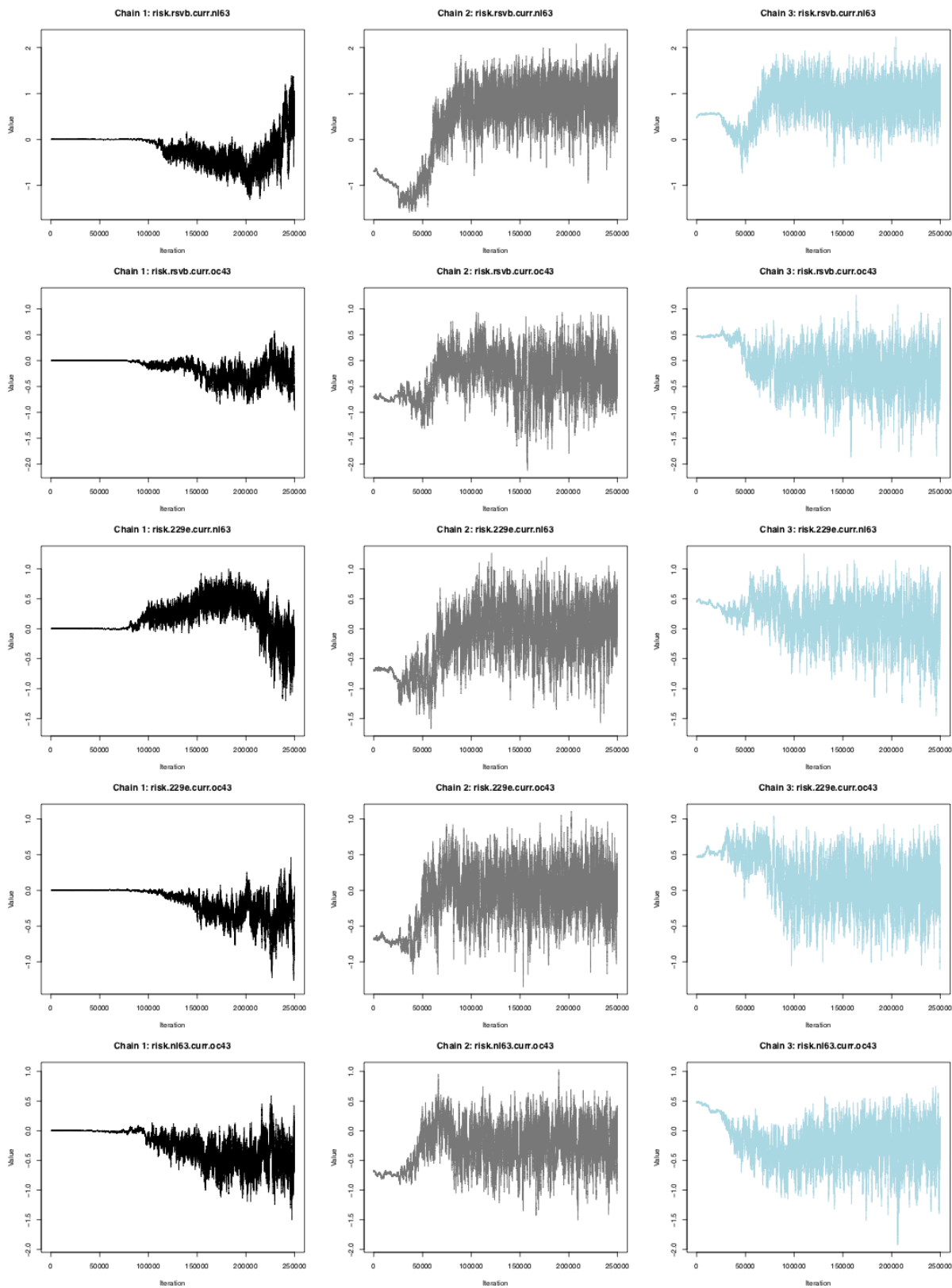
The following are trace plots of the 37 parameters in the multi-pathogen model with pathogen identification at the group level.

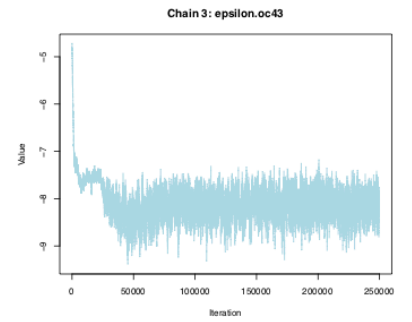
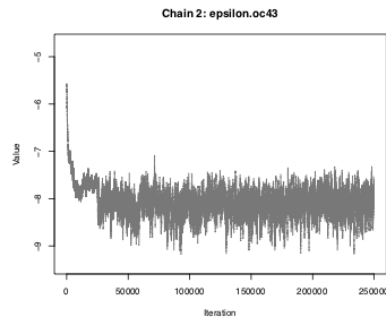
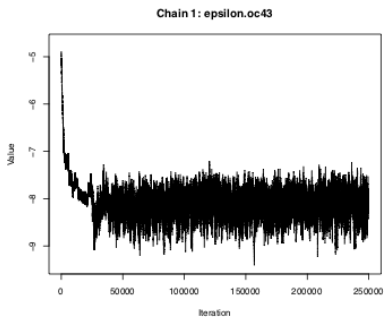
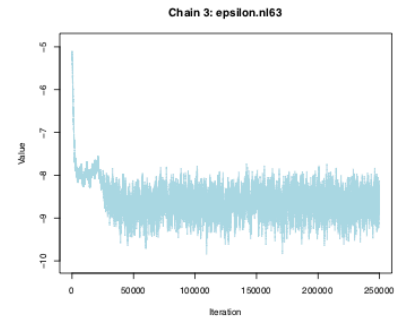
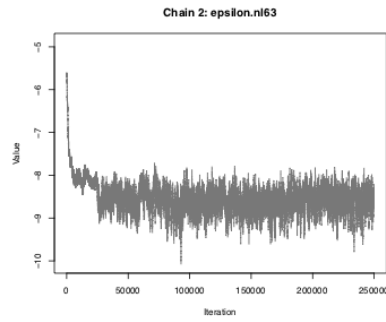
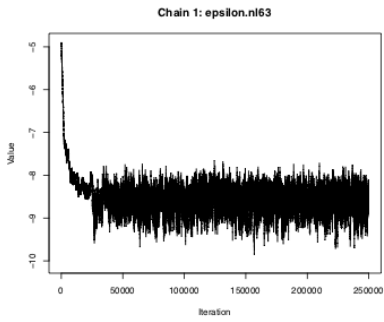
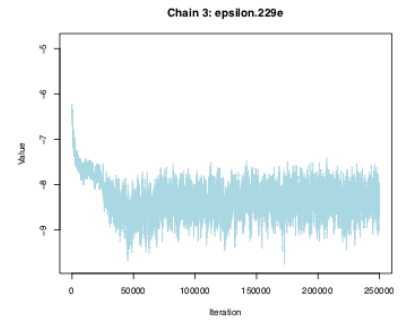
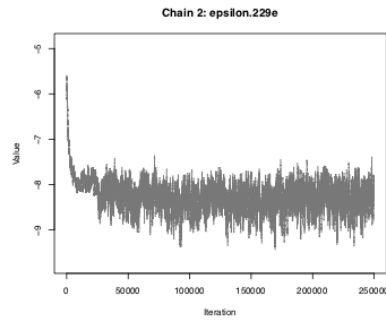
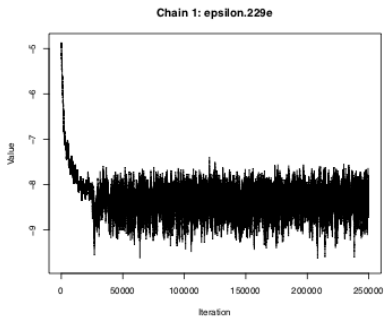
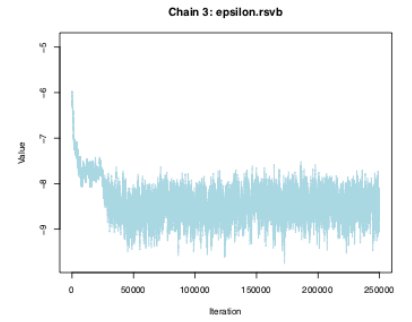
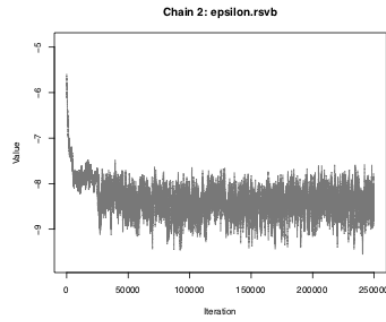
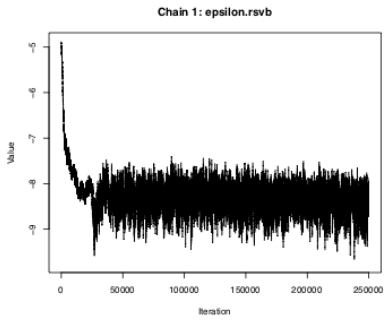
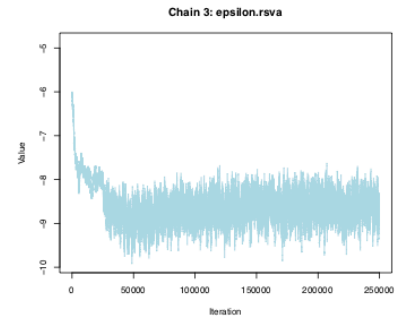
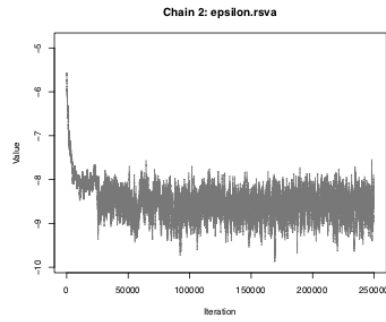
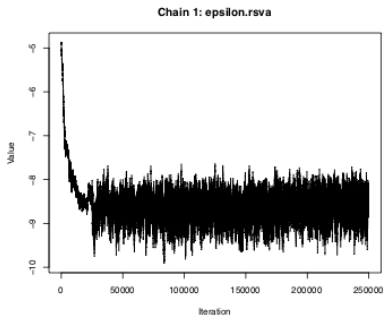












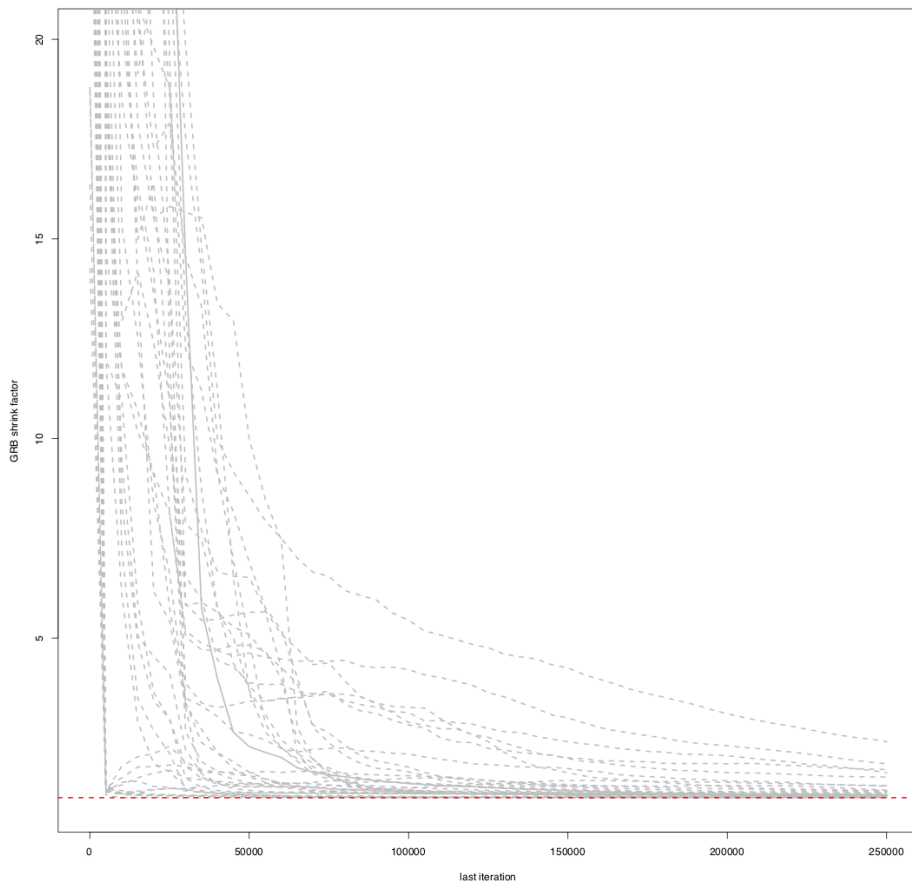
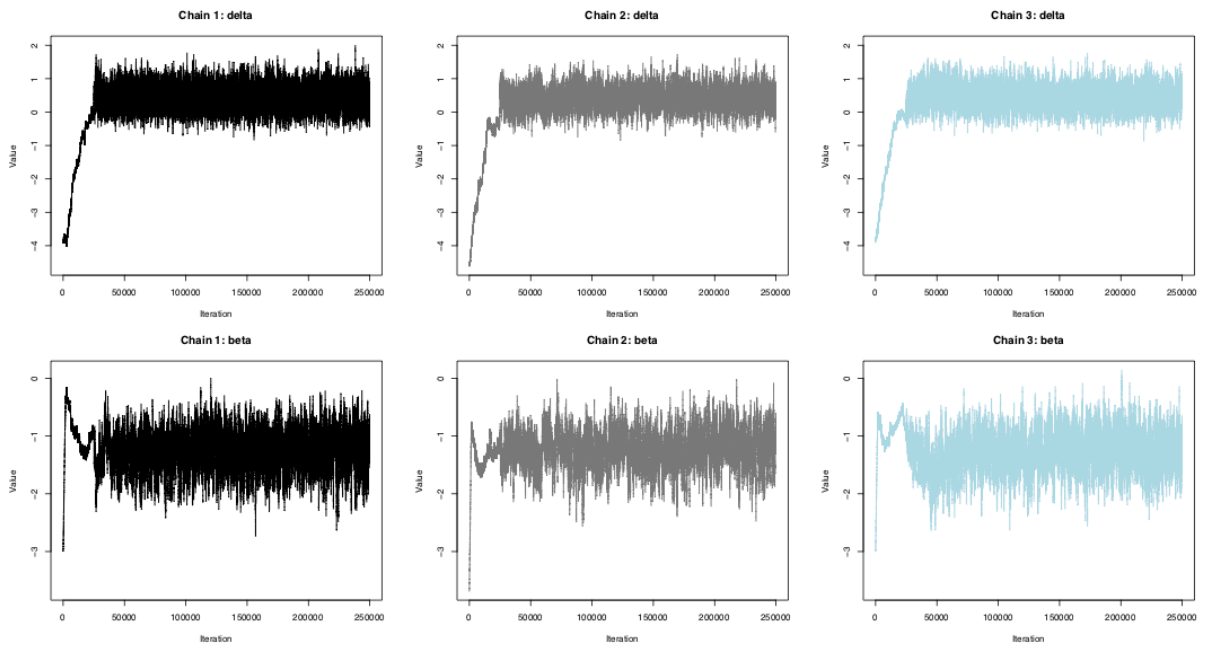


Figure A4. 6: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the multi-pathogen model with pathogen identification at the group level and the dashed red line shows the value 1.

The burn-in point was chosen as 225,000 for chain 1, 125,000 for chain 2 and 150,000 for chain 3.

Table A.4 1: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the multi pathogen model with pathogen identification at the group level.

Parameter	Point estimate GRB statistic	ESS
risk.rsva.prev.rsva	1.25	616
risk.rsva.prev.rsvb	1.08	865
risk.rsva.prev.229e	2.73	736
risk.rsva.prev.nl63	1.1	929
risk.rsva.prev.oc43	1.13	877
risk.rsvb.prev.rsvb	1.26	823
risk.rsvb.prev.229e	1.1	758
risk.rsvb.prev.nl63	1.1	894
risk.rsvb.prev.oc43	1.34	1010
risk.229e.prev.229e	2.51	717
risk.229e.prev.nl63	1.1	787
risk.229e.prev.oc43	1.83	848
risk.nl63.prev.nl63	3.63	864
risk.nl63.prev.oc43	1.09	921
risk.oc43.prev.oc43	4.76	828
risk.rsva.curr.rsvb	2.41	886
risk.rsva.curr.229e	1.11	601
risk.rsva.curr.nl63	2.23	545
risk.rsva.curr.oc43	1.43	776
risk.rsvb.curr.229e	1.13	521
risk.rsvb.curr.nl63	1.11	375
risk.rsvb.curr.oc43	1.33	580
risk.229e.curr.nl63	1.36	600

risk.229e.curr.oc43	1.18	631
risk.nl63.curr.oc43	1.28	635
eta.rsva	1.1	1060
eta.rsvb	1.01	1250
eta.229e	1.12	1290
eta.nl63	1.43	1190
eta.oc43	1.16	1070
epsilon.rsva	1.03	1060
epsilon.rsvb	1.01	1180
epsilon.229e	1.03	935
epsilon.nl63	1.01	1090
epsilon.oc43	1.01	1110
delta	1	1900
beta	1.01	1090

The mGRB is 7.43 and the mESS is 1099.

Visually the chains look like they do converge, but the values of the GRB and ESS suggest that longer runs are needed.

A4.2.2. Multi-pathogen model fit to data with pathogen identification at the pathogen level

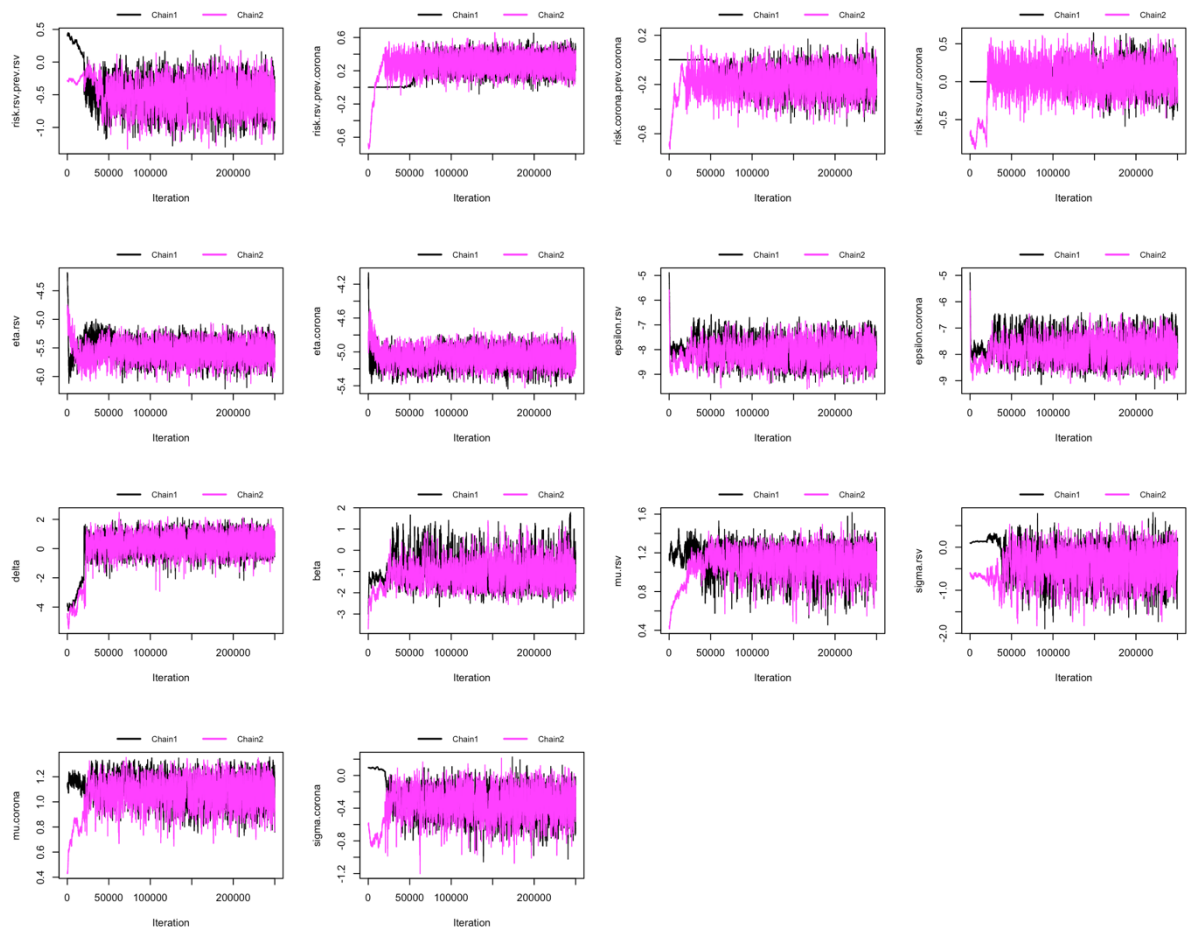


Figure A4. 7: Trace plots of parameters in the multi-pathogen model with pathogen identification at the pathogen level.

Two chains were initiated at different parameter values and these are shown in black (Chain 1) and pink (Chain 2). The x-axis shows the iteration number, while the y-axis shows the log parameter value.

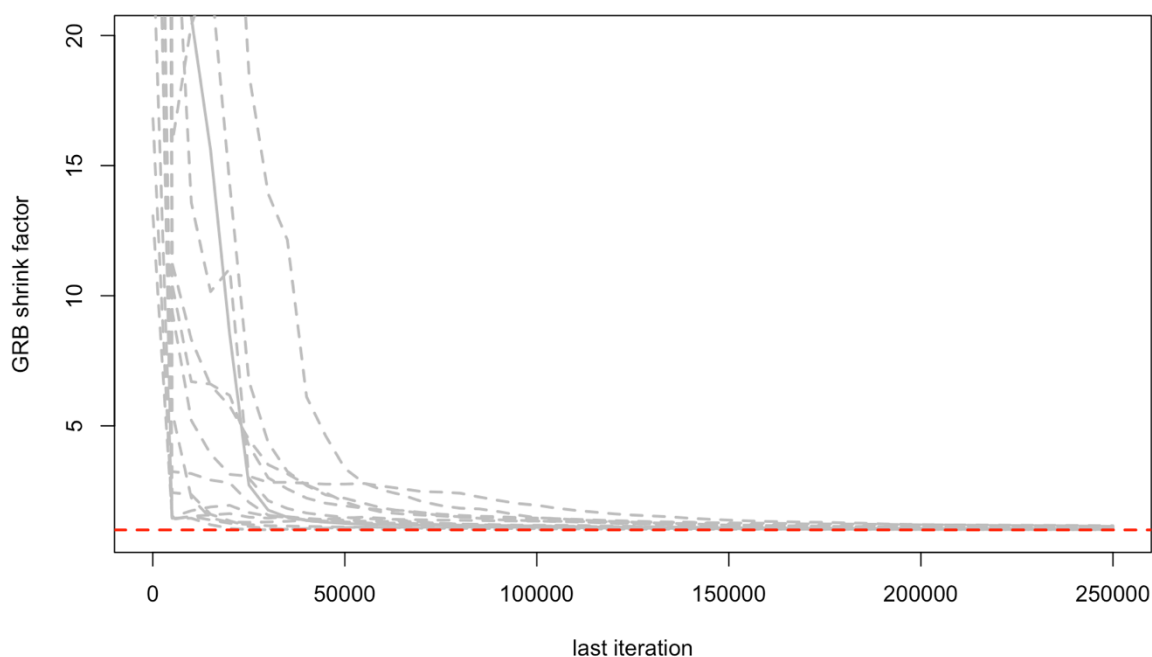


Figure A4. 8: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the multi-pathogen model with pathogen identification at the pathogen level and the dashed red line shows the value 1.

Burn-in point was chosen as 75000.

Table A.4 2: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the multi pathogen model with pathogen identification at the pathogen level.

Parameter	Point estimate GRB statistic	ESS
risk.rsv.prev.rsv	1.07	1840
risk.rsv.prev.corona	1.05	2940
risk.corona.prev.corona	1.29	1560
risk.rsv.curr.corona	1.23	1160
eta.rsv	1.02	2160

eta.corona	1.04	1770
epsilon.rsv	1.04	1250
epsilon.corona	1.04	1110
delta	1.01	2050
beta	1.04	981
mu.rsv	1.13	1220
sigma.rsv	1.14	1570
mu.corona	1.16	1130
Sigma.corona	1.13	1460

The mGRB is 1.23 and the mESS is 1893.

A4.2.3. Single-pathogen model fit to RSV data with pathogen identification at the group level

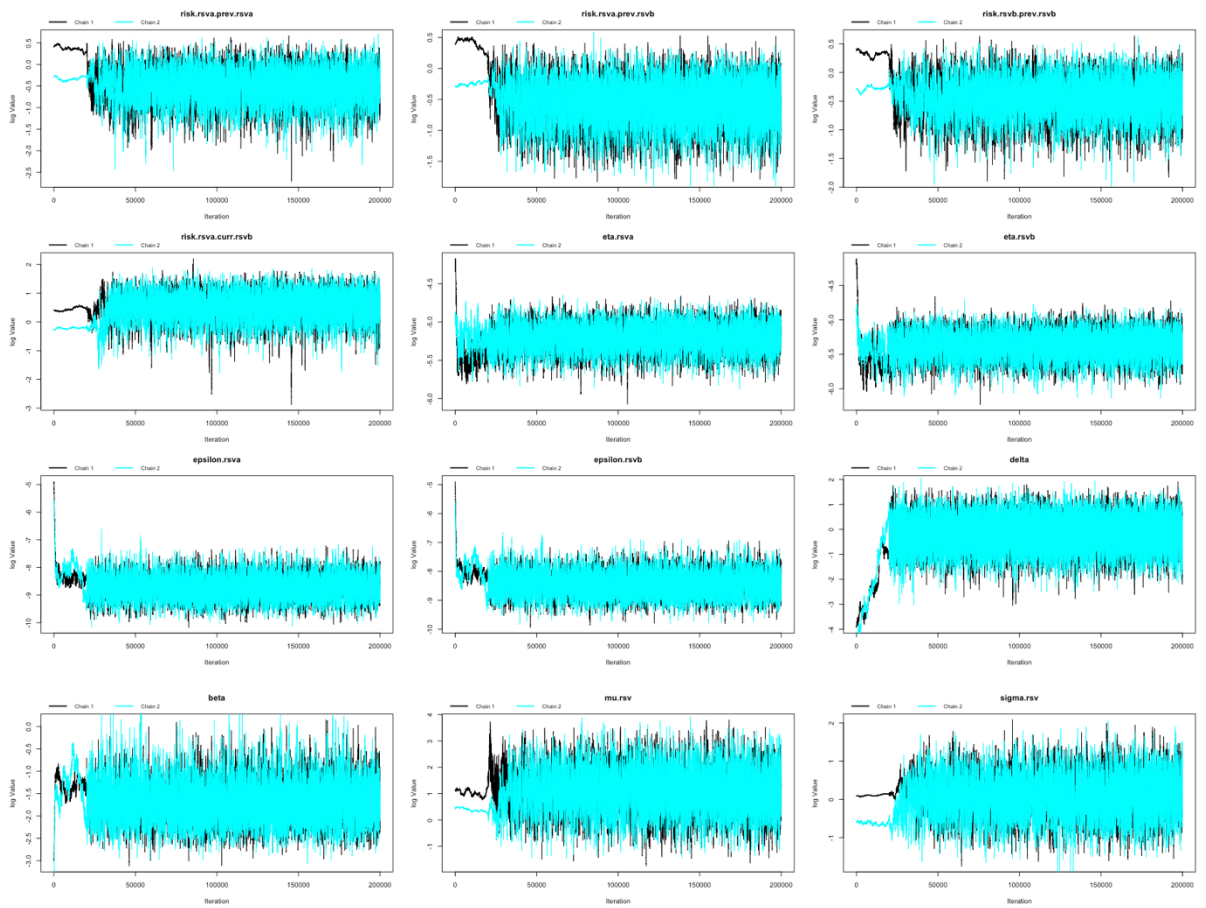


Figure A4. 9: Trace plots of parameters in the single-pathogen model for RSV with pathogen identification at the group level.

Two chains were initiated at different parameter values and these are shown in black (Chain 1) and blue (Chain 2). The x-axis shows the iteration number, while the y-axis shows the log parameter value.

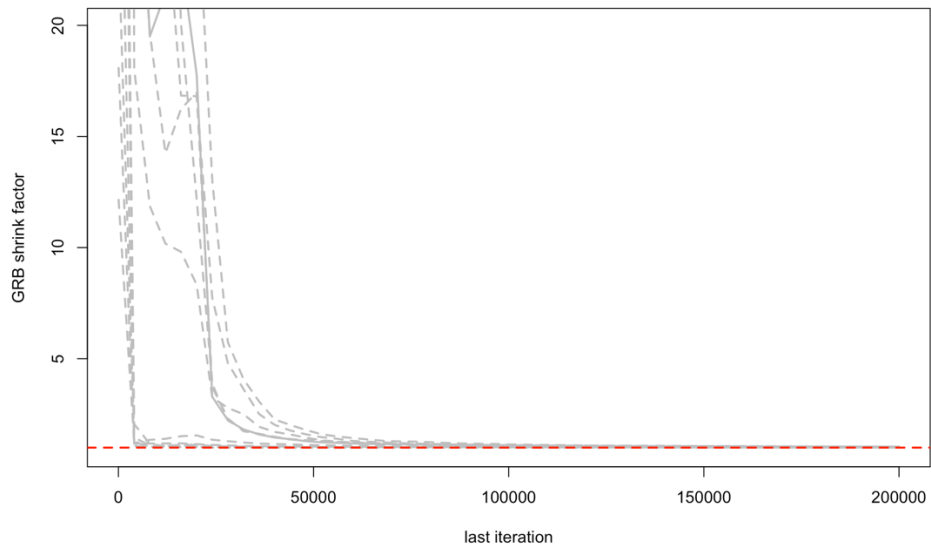


Figure A4. 10: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the single-pathogen model for RSV with pathogen identification at the group level and the dashed red line shows the value 1.

Burn-in was set at 40,000 for each chain.

Table A.4. 3: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the single pathogen RSV group model.

Parameter	Point estimate GRB statistic	ESS
risk.rsva.prev.rsva	1.05	3970
risk.rsva.prev.rsvb	1.06	4290
risk.rsvb.prev.rsvb	1.05	3920
risk.rsva.curr.rsvb	1.04	4250
eta.rsva	1.03	4820
eta.rsvb	1.01	5180
epsilon.rsva	1.00	4990
epsilon.rsvb	1.00	4580

delta	1.00	7580
beta	1.01	4180
mu.rsv	1.02	3720
sigma.rsv	1.07	3770

The mGRB is 1.08 and the mESS is 5239.

A4.2.4. Single-pathogen model fit to hCoV data with pathogen identification at the group level

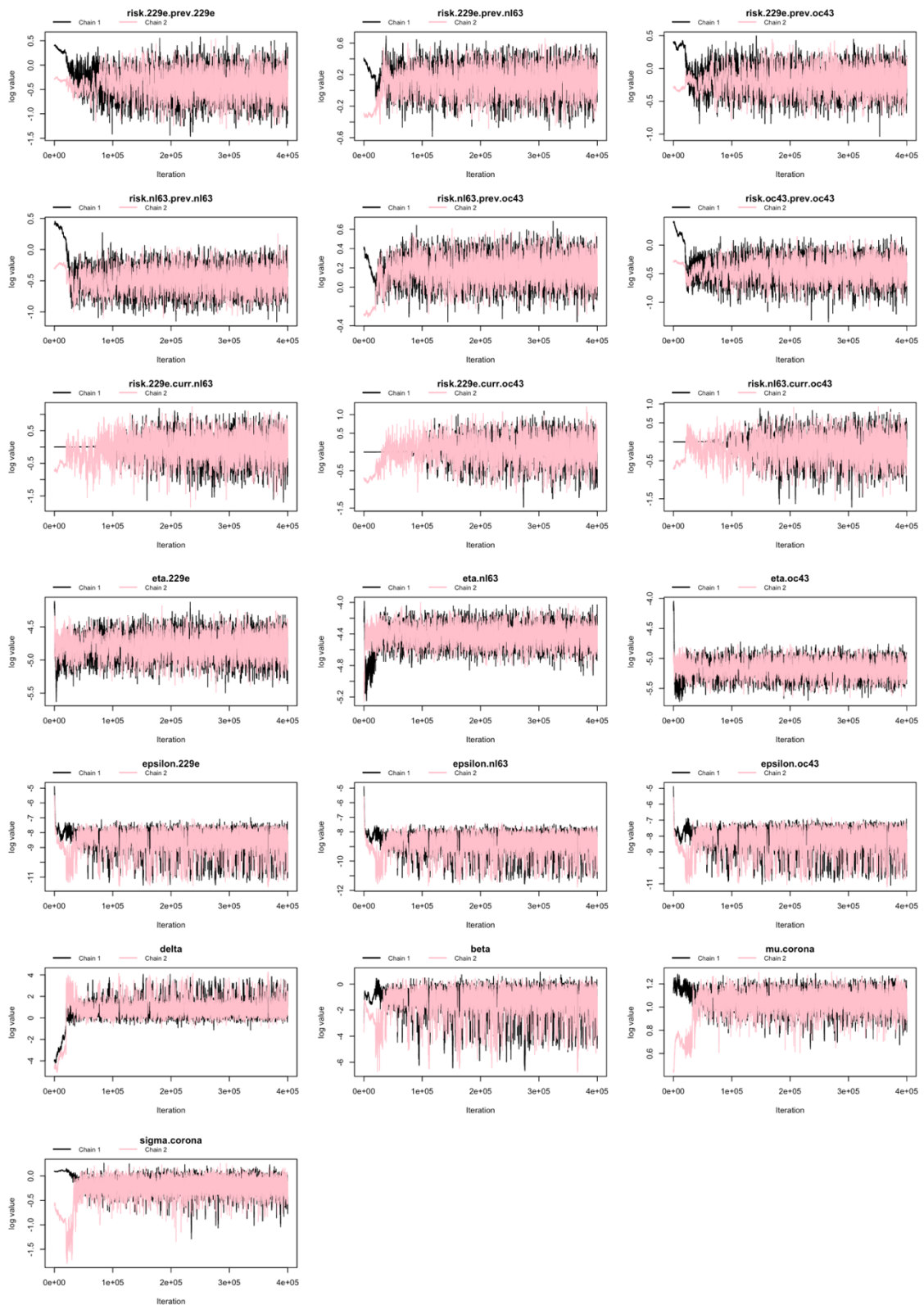


Figure A4. 11: Trace plots of parameters in the single-pathogen model for hCoV with pathogen identification at the group level.

Two chains were initiated at different parameter values and these are shown in black (Chain 1) and pink (Chain 2). The x-axis shows the iteration number, while the y-axis shows the log parameter value.

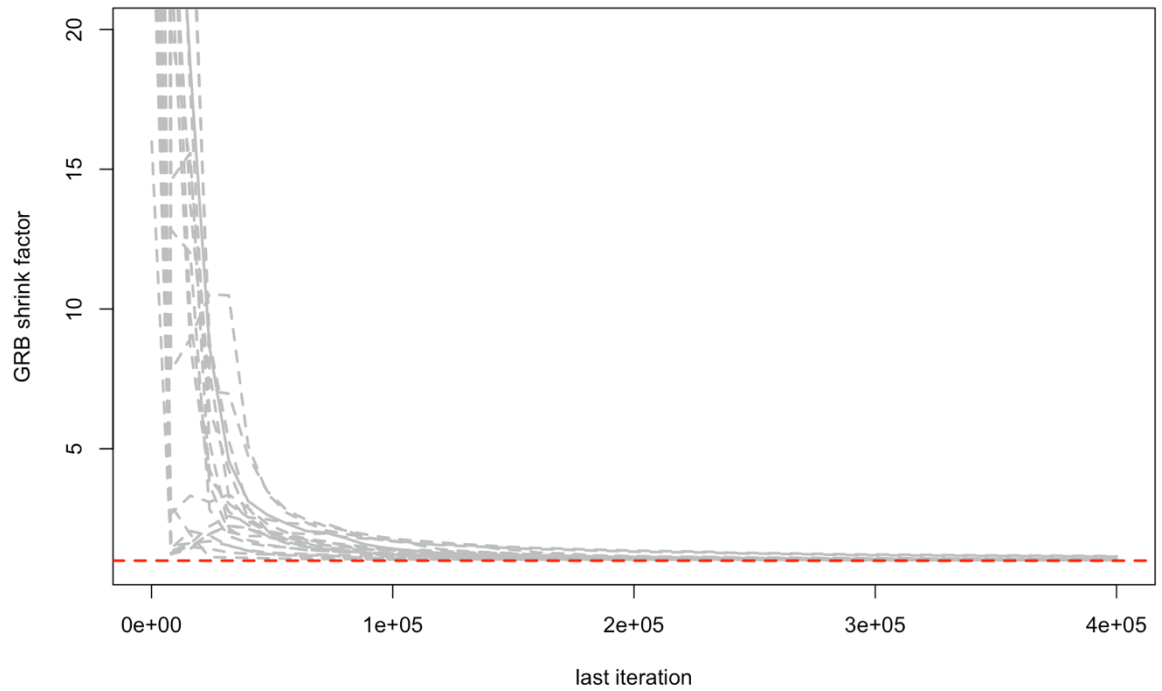


Figure A4. 12: The evolution of the Gelman-Rubin-Brooks (GRB) statistic (shrink factor) as the number of iterations increases.

Each grey line represents a model parameter in the single-pathogen model for hCoV with pathogen identification at the group level and the dashed red line shows the value 1.

Chose a burn-off of 100,000 for each chain.

Table A.4. 4: The values of the GRB statistic (to 3 significant figures) and the effective sample size are shown for all the parameters in the single pathogen hCoV strain model.

Parameter	Point estimate GRB statistic	ESS
risk.229e.prev.229e	1.05	2660
risk.229e.prev.nl63	1.06	2730
risk.229e.prev.oc43	1.07	2640
risk.nl63.prev.nl63	1.09	2970
risk.nl63.prev.oc43	1.06	2720
risk.oc43.prev.oc43	1.08	2500
risk.229e.curr.nl63	1.04	2000
risk.229e.curr.oc43	1.06	1950
risk.nl63.curr.oc43	1.04	2070
eta.229e	1.00	2920
eta.nl63	1.04	3120
eta.oc43	1.02	3190
epsilon.229e	1.03	1290
epsilon.nl63	1.03	1360
epsilon.oc43	1.03	1320
delta	1.02	1570
beta	1.04	1160
mu.corona	1.19	3070
sigma.corona	1.22	3850

The mGRB is 1.11 and the mESS is 3097.

Appendix References

1. Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. *Nat Protoc. Nature Research*; **2006**; 1(3):1559–1582.
2. Wathuo M, Medley GF, Nokes DJ, Munywoki PK. Quantification and determinants of the amount of respiratory syncytial virus (RSV) shed using real time PCR data from a longitudinal household study. *Wellcome Open Res* [Internet]. **2017**; 1(0):27. Available from: <https://wellcomeopenresearch.org/articles/1-27/v2>
3. Robert CP, Casella G. Metropolis–Hastings Algorithms BT - Introducing Monte Carlo Methods with R. In: Robert C, Casella G, editors. New York, NY: Springer New York; 2010. p. 167–197. Available from: https://doi.org/10.1007/978-1-4419-1576-4_6
4. Roberts GO, Rosenthal JS. Examples of Adaptive MCMC. *J Comput Graph Stat* [Internet]. **2009**; 18(2):349–367. Available from: <http://www.tandfonline.com/doi/abs/10.1198/jcgs.2009.06134>