Modeling of Infectious Disease Surveillance Data

Leigh Fisher

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Jon Wakefield, Chair

Vladimir Minin

M. Elizabeth Halloran

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Modeling of Infectious Disease Surveillance Data

Leigh Fisher

Chair of the Supervisory Committee:
Professor Jon Wakefield
Department of Biostatistics & Department of Statistics

A wide range of infectious diseases are monitored at the local, state, and national levels using disease surveillance systems designed to asses current disease burden and to detect emerging outbreaks. Public health officials rely on surveillance data to provide accurate and timely information, which may then be used to inform policy or intervention decisions. For privacy reasons, surveillance data is often aggregated over space and time; hence the data is limited to basic demographic information along with basic spatial or temporal information, such as the county of residence for an infected individual and week of diagnosis. The limited information contained within, and potential bias of, surveillance data can pose serious challenges to valid inference. Nevertheless, for many practical applications, surveillance networks are one of the best sources of data, especially at the state or local level. In this dissertation, we develop models to obtain timely estimates of parameters of interest using infectious disease surveillance data. This work is motivated by surveillance data for hand, foot, and mouth disease (HFMD) in China, influenza in Florida, and measles in Germany. For HFMD, we develop a model to quickly estimate pathogen-specific disease counts, and associations with meteorological variables, when laboratory information is available for only a small subsample of cases. For the flu, we consider approaches to account for potential biases in the data due to disparities in healthcare access. For measles, we develop an ecological model to account for differing levels of vaccination coverage while providing estimates of key epidemic parameters.

TABLE OF CONTENTS

Page

List of Figures v

List of Tables ix

Glossary xi

Chapter 1: Introduction 1
1.1 Motivating examples 1
1.1.1 Hand, foot, and mouth disease in China 1
1.1.2 School located influenza vaccination in Florida 4
1.1.3 Measles epidemics in Germany 6
1.2 Organization of dissertation 7

Chapter 2: Background 8
2.1 Models for infectious diseases 8
2.1.1 Compartmental models 8
2.2 Disease mapping and spatial regression for aggregate data 12
2.3 Bayesian computing 13
2.3.1 Hierarchical Bayesian modeling 13
2.3.2 Markov chain Monte Carlo 14
2.3.3 Integrated nested Laplace approximation 15
2.4 Measures of model fit 17

Chapter 3: Time Series Modeling of Pathogen-Specific Disease Probabilities with Subsampled Data 19
3.1 Introduction 19
3.2 Notation 19
3.3 Model formulation 22

3.3.1 Disease probabilities . . . 22
3.3.2 Laboratory testing subsampling . . . 22
3.3.3 Model development . . . 23
3.3.4 Previous approaches to inference . . . 26
3.4 The hybrid approach . . . 26
3.4.1 Estimation of counts . . . 27
3.4.2 Distribution of the log relative risks . . . 29
3.5 The full probability model . . . 30
3.5.1 The likelihood . . . 31
3.5.2 Prior distributions . . . 33
3.5.3 Posterior distribution . . . 33
3.5.4 Markov Chain Monte Carlo for discrete variables . . . 34
3.5.5 Comparison study of MCMC to the hybrid procedure . . . 36
3.6 Application to HFMD data . . . 39
3.6.1 The model . . . 39
3.6.2 Prior selection for temporal smoothers . . . 44
3.6.3 Simulations . . . 48
3.6.4 Modeling results . . . 50
3.7 Discussion . . . 52

Chapter 4: Negative Controls . . . 55
4.1 Introduction . . . 55
4.2 Previous approach to negative controls . . . 57
4.3 Reframing the negative control approach . . . 63
4.4 Simulations . . . 66
4.4.1 A comparison of negative control modeling approaches . . . 66
4.4.2 Unmeasured confounding and negative control outcomes . . . 70
4.5 Extension to spatially structured data . . . 74
4.5.1 The spatial negative controls model . . . 74
4.5.2 Joint disease models . . . 76
4.5.3 Confounding by location . . . 77
4.5.4 Simulations for spatial negative controls . . . 79
4.6 Extension to stratified populations . . . 87

4.7 Application to the AHCA data . . . 91
4.7.1 Introduction . . . 91
4.7.2 Simple negative control . . . 92
4.7.3 Spatial negative control analysis . . . 94
4.7.4 Evaluating SLIV effect over many seasons . . . 99
4.8 Spatio-temporal negative control analysis . . . 102
4.8.1 Hierarchical spatio-temporal models for ILI . . . 102
4.9 Discussion . . . 107

Chapter 5: Ecological Inference with Surveillance Data . . . 109
5.1 Introduction . . . 109
5.2 Epidemiological parameters of interest . . . 110
5.2.1 Basic reproductive number, R_0 . . . 110
5.2.2 Modeling the force of infection . . . 112
5.3 Aggregate models for infectious disease data . . . 113
5.4 Ecological bias . . . 116
5.5 Ecological models with vaccine coverage . . . 120
5.6 Ecological vaccine model development . . . 121
5.6.1 Introduction and notation . . . 121
5.6.2 All-or-none vaccine ecological model . . . 122
5.6.3 Leaky vaccine ecological model . . . 123
5.6.4 Comments on the ecological vaccine model . . . 125
5.7 Simulations . . . 127
5.7.1 Assesing the simplifying assumptions . . . 127
5.7.2 Assessing the ecological model . . . 129
5.7.3 Asymptotic behavior of the ecological vaccine model . . . 134
5.8 Application to the measles data . . . 136
5.9 Discussion . . . 146

Chapter 6: Discussion and Future Work . . . 147

References . . . 149

Appendix A: Appendix to Chapter 3 158
A.1 Comparison of empirical Bayes and method of moment estimators 158
A.2 Conditional posterior distributions for MCMC 162
A.3 Detailed results of MCMC simulation study 166
A.4 Variance adjustments 181
A.5 Simulations 184
A.5.1 Simulated population setup 185
A.5.2 Estimation of Y_{tj}^{G} 188
A.5.3 Estimation of $\log \theta_t^{\mathrm{G}}$ 198
A.5.4 Modeling $\log \theta_t^{\mathrm{G}}$ 209
A.5.5 Summary of simulations 217
A.6 Sensitivity analysis 218

Appendix B: Appendix to Chapter 4 222
B.1 Parameterizations of the shared component model 222
B.2 ICD-9 Code definitions 225

Appendix C: Appendix to Chapter 5 226
C.1 Additional simulation results assessing simplifying assumptions 226
C.2 Measles analysis with non-informative priors 227
C.3 Measles analysis with informative priors 230

LIST OF FIGURES

Figure Number		Page
1.1	Epidemic curves of the 2009-2011 HFMD epidemics in China.	4
2.1	Diagram of the SIR model	8
3.1	Six geographic regions of mainland China	20
3.2	Data structure for a generic week and demographic stratum	21
3.3	Number of subsampled cases by region and strata	24
3.4	Proportion of subsampled cases by severity, region, and stratum	25
3.5	Estimated log EV71- and CA16-specific relative risk by region by lagged temperature, relative humidity, wind speed, and precipitation	42
3.6	Pairwise relationships between pathogen-specific log relative risks and lagged meteorological variables for the six regions in China	43
3.7	Estimated degrees of freedom for various values of τ_{G}, by region- and pathogen-specific random walks of order 2	47
3.8	Histogram of estimated degrees of freedom for τ_{G} values from Gamma(100, 0.02) distribution for regions in China and by pathogen	48
3.9	Estimated smoothers of meteorological variables for log relative risk of EV71 and CA16	51
3.10	Fitted values for log relative risks of EV71 and CA16	53
4.1	Causal diagram for the negative control outcome	57
4.2	Summary of simulations to compare negative control modeling approaches	68
4.3	Comparison of coverage for negative control modeling approaches	69
4.4	Bias of the unadjusted and adjusted estimates for varying amounts of unmeasured confounding	71
4.5	MSE of the unadjusted and adjusted estimates for varying amounts of unmeasured confounding	72
4.6	Difference in absolute bias and MSE between the adjusted and unadjusted estimates for varying amounts of unmeasured confounding	73

4.7 Coverage of the adjusted and unadjusted estimates for varying amounts of unmeasured confounding 74
4.8 Bivariate ICAR realizations for simulation. 81
4.9 Summary of results for spatial negative controls 85
4.10 Comparison of estimated spatial and independent random effects between including and excluding the Alachua-specific effects 87
4.11 Map of the counties in Florida. Alachua County is highlighted in red, and counties in the Alachua region are in blue. 91
4.12 Summary of AHCA facilities and Florida population in 2010. 93
4.13 Summary of log SMRs for ILI and GI during 2012/2013 influenza season. . . 95
4.14 Scatter plot of log SMRs for ILI against log SMR for GI. 96
4.15 Estimates of the residual relative risk for non-spatial and spatial contributions. Note that the scales differ for the spatial and non-spatial maps. 98
4.16 Raw log SMRs for ILI and GI evaluated at each of the 9 influenza seasons. . 100
4.17 Posterior estimates and 95% CI for the SLIV program effect for 9 seasons. . 101
4.18 log SMRs for influenza-like illness by county and year. 103
4.19 Results from area-specific log linear models. 104

5.1 Summary of measles cases over time and space. 110
5.2 Simulated epidemic curves for three values of R_0. 127
5.3 Simulation results for ecological model simplifying assumptions. 129
5.4 Simulated epidemic curves for partially vaccinated populations. 131
5.5 Estimates when assuming an all-or-none vaccine. 132
5.6 Estimates when assuming a leaky vaccine. 133
5.7 Estimates when assuming an all-or-none vaccine for 10 years worth of data. . 134
5.8 Estimates when assuming leaky vaccine for 10 years worth of data. 135
5.9 Map of estimated vaccine coverage for the MMR vaccine in Germany. 136
5.10 Observed number of cases and prevalence of measles by state and biweek. . . 138
5.11 Histogram of posterior samples of $\hat{\phi}$. 141
5.12 Fitted values for the ecological vaccine model. 142
5.13 Pearson residuals for the ecological vaccine model. 143
5.14 Maps of the random effect estimates for the autoregressive and endemic components in the ecological vaccine model. 144
5.15 Comparison of autoregressive and endemic random effect estimates. 145

A.1 Corrected estimated pathogen-specific disease counts vs uncorrected MoM estimated counts. 161
A.2 Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches in Scenario 1. 167
A.3 Comparison of estimated pathogen-specific disease counts using different estimation approaches in Scenario 1. 168
A.4 Comparison of estimated pathogen-specific log relative risks using different estimation approaches, for Scenario 1. 169
A.5 Comparison of estimated pathogen-specific disease counts using different estimation approaches in Scenario 2. 173
A.6 Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches for Scenario 2. 174
A.7 Comparison of estimated pathogen-specific log relative risks using different estimation approaches in Scenario 2. 175
A.8 Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches, for Scenario 3. 177
A.9 Comparison of estimated pathogen-specific disease counts using different estimation approaches with in Scenario 3. 178
A.10 Comparison of estimated pathogen-specific log relative risks using different estimation approaches, for Scenario 3. 179
A.11 Log relative risks used for simulations when proportionality is valid 186
A.12 Proportions of subsampled cases used for simulations. 187
A.13 Estimated unobserved pathogen-specific disease counts by severity. 190
A.14 Coverage by severity with constant subsampling 191
A.15 Estimates by severity with constant subsampling 194
A.16 Estimated unobserved pathogen-specific disease counts by severity over time 195
A.17 Coverage of the unobserved pathogen-specific disease counts by severity. . . . 196
A.18 Coverage of the unobserved pathogen-specific severe disease counts. 197
A.19 Estimated risk and coverage for simulations with constant subsampling . . . 199
A.20 Estimated risk and coverage for simulations with varying subsampling 201
A.21 Log relative risks used for simulations when proportionality is violated . . . 203
A.22 Coverage when proportionality is invalid and constant subsampling 205
A.23 Coverage when proportionality is invalid and variable subsampling 206
A.24 Comparison of weighted average pathogen-specific log relative risk estimates 208

A.25 Simulation results when estimating temporal effects, under constant subsampling modeling pathogens separately 211
A.26 Simulation results when estimating temporal effects, under constant subsampling modeling pathogens together 212
A.27 Empirical correlations between $\log \widehat{\theta}_t^{\mathrm{E}}$ and $\log \widehat{\theta}_t^{\mathrm{C}}$ over time, as estimated from 500 simulations. 213
A.28 Esimating temporal trends under constant subsampling and modeling pathogens separately 215
A.29 Esimating temporal trends under constant subsampling and modeling pathogens jointly 216
A.30 Sensitivity to temporal smoothing for $\mathrm{E}[\tau_{\mathrm{G}}]$=5,000. 219
A.31 Sensitivity to temporal smoothing for $\mathrm{E}[\tau_{\mathrm{G}}]$=7,500. 220
A.32 Sensitivity to temporal smoothing for $\mathrm{E}[\tau_{\mathrm{G}}]$=10,000. 221

C.1 Pairwise correlation of posterior samples for the ecological vaccine model with noninformative priors. 228
C.2 Fitted values for the ecological vaccine model with noninformative priors. . . 229
C.3 Pairwise correlation of posterior samples for the ecological vaccine model with informative priors. 230

LIST OF TABLES

Table Number Page

1.1 Summary of HFMD cases by severity, year, and strata 3

3.1 Summary of three data generating scenarios. 37

3.2 Summary of estimates from MCMC and hybrid approach. 38

3.3 Comparison of proposal jump sizes, MCMC iterations, and analysis time for the comparison study of the MCMC to the hybrid procedure. 39

4.1 Parameter specifications for spatial simulations 79

4.2 Summary of results for simulations when the negative control outcome matches the spatial variability and and unmeasured confounding in various combinations 86

4.3 Summary of 2012/2013 season. Incidence are per 1,000. 92

4.4 Unadjusted and adjusted estimates of the SLIV program effect by comparison region for the 2012/2013 season. 94

4.5 Unadjusted and adjusted estimates of log SLIV program effect. 99

4.6 Model comparison: the deviance information criteria (DIC) is calculated using p_D is the effective degrees of freedom and the deviance evaluated at the posterior mean, $\bar{D}$ 106

4.7 Summaries of variance components. For the RW2 and ICAR models, the contribution is evaluated empirically, since the variance parameter is conditional rather than marginal. 107

5.1 Summary of the all-or-none and leaky vaccine models and the assumptions for the ecologial model. 126

5.2 Posterior medians and 95% credible intervals for the parameters of the ecological model for the measles data. 140

A.1 Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 1. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12). 170

A.2 Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 2. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12). 172
A.3 Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 3. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12). 180
A.4 Summary of simulation parameters 185
A.5 Scaled MSE for pathogen- and stratum-specific disease counts by severity when the proportion of subsampling is constant over time and strata. 188
A.6 Scaled MSE for pathogen- and stratum-specific disease counts by severity when the proportion of subsampling varies over time and strata. 193
B.1 Simulation results to look into different parameterizations of the shared component model. 224
B.2 ICD-9 codes for ILI and GI 225
C.1 Estimates, standard error, bias, and coverage of the MLE estimates of α_{AR} and α_{EN} for three different simulation scenarios. 226
C.2 Posterior medians and 95% credible intervals for the measles biweekly data with non-informative priors. 227

GLOSSARY

AHCA: Agency for Heath Care Administration

BYM: Besag-York-Mollié

CA16: Coxackie A16

ESSENCE: Early Notification of Community-based Epidemics

EV71: Enterovirus 71

GI: Gastrointestinal illness

HFMD: Hand, foot, and mouth disease

HMC: Hamiltonian Monte Carlo

ICAR: Intrinsic conditional autoregressive model

ILI: Influenza-like illnesss

INLA: Integrated nested Laplace approximation

MCMC: Markov chain Monte Carlo

MMR: Measles, mumps, and rubella

RW2: Random walk of order 2

SCM: Shared component model

SLIV: School-located influenza vaccination program

SMR: Standardized morbidity ratio

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor, Jon Wakefield, for his patience, support, encouragement, and mentorship. His insights into the Seattle music scene, British vocabulary, work-appropriate jokes, and everyday footwear are lessons that are sure to serve me well in all of my future endeavors. I am especially grateful for his compassion and patience when my father died. In a time of great darkness and uncertainty, he provided much needed encouragement and understanding. Jon has been the most metal advisor a girl could ever wish for and I am truly grateful.

I would like to thank my committee members for their helpful feedback and guidance related to my dissertation research. From the University of Washington, I would like to thank Scott Emerson, Nicole Hamblett, Patrick Heagerty, Katie Kerr, Susanne May, Vladimir Minin, Barbara McKnight, Ken Rice, Barbra Richardson, Galen Shorack, Mary Lou Thompson, and Andrew Zhou, for their contributions as instructors, supervisors, and mentors during my time in the department. Additionally, I would like to thank Gitana Garofalo for her administrative and general life support during my time here.

I would also like to thank my fellow students in both the Biostatistics and Statistics departments. And a special thank you to my friends in the Unicorn cohort (honorary members included) for your supportive and fun-filled approach to this really difficult task. Thanks to the Shaggy Muskrats for making sure I got outside once in a while; and thanks to the Kaffeklastch for keeping me caffeinated and part of the community.

I also want to thank two former mentors. First, I am grateful for the enthusiasm and encouragement from Ami Radunskaya, my math advisor and mentor at Pomona College. Our work together was my first exposure to research, and the support from such a bad-ass

woman has left an indelible mark on me. Second, thank you Dan Duxbury, for his continuous encouragement. He has challenged me to do more than I thought I could. His mentorship and friendship prepared me for much more than the race course and the open road.

I am lucky to have incredible friends who have, over the years, provided endless laughs and stuck through the good and the bad. I am grateful for Jennifer Watkins. She first introduced me to biostatistics, when I was unsure about my future path. She is also responsible for me surviving my childhood, as she was the only one brave enough to babysit me as a kid. A special thanks goes to Julia Fathe, my life-long friend. Whether in gymnastics or school, her boldness has always inspired mine. Her decision to return to graduate school planted the idea in my mind. I could not be prouder to call her my friend. Laina Mercer has been a wonderful friend, invaluable teammate, inspirational colleague, and overall go-to resource for all of life's questions. She deserves a special thanks over a large plate of nachos.

My whole family has put up with me through this long process, providing me with love and support along the way. I am grateful for my sister Morgane and her family (Conrad and Leigh) for keeping me honest, being an understanding ear, and sending little reminders of the important things via pictures of my nephew. To Aunt Suzy, Uncle Rick, and Preston, thank you for reminding me to have some fun. Thank you to Earnie for his unwavering love, support, and getting me out of my bubble. He and his funny dogs brought much needed laughs and love, especially during the hard times. Lastly, I must thank my parents, Pat and Mike for their never-ending love and support. I am thankful that they made my education a priority, setting the bar high and encouraging my mathematical inclinations.

It has taken me a long time to get here, and I certainly did not do it on my own. There are so many more people that have contributed to this in more ways than I can include. To have reached this point, with some of my sanity still intact, is truly a tribute to all of the support and love I have in my life.

Thank you all! Now, don't read any more of this and let's get on with the fun!

DEDICATION

In loving memory of my father, Michael Fisher, and
my grandparents Ken and Barbara Clarke.

Chapter 1
INTRODUCTION

Disease surveillance data is easily collected and provides a rich source of information about when and where cases occur in populations over long periods of time, and large areas. Surveillance systems range from daily collection of de-identified electronic medical records to requiring clinicians to report cases of notifiable diseases. The reported data is frequently aggregated over time and space for privacy reasons. As a result, the amount of information about each infected patient is frequently limited to basic demographic information and a general time and a location of diagnosis, aggregated to some administrative set of areas. While surveillance data plays a critical role in public health, there are deficiencies in the data that can make it difficult to obtain unbiased estimates in a timely fashion. In some settings, there is additional information, such as laboratory results, for a small subset of the patients. In other settings, the data is a biased sample of the population of interest, such as electronic medical records from local emergency rooms. In this dissertation, we consider the statistical implications of using surveillance data with such deficiencies and develop methods to obtain reasonable estimates with the available information. This chapter provides a brief introduction to the motivating examples and describes the organization of the rest of this dissertation.

1.1 Motivating examples

1.1.1 Hand, foot, and mouth disease in China

For many infectious diseases, the numbers of new cases are available at regular temporal intervals through surveillance systems, but information on the pathogen responsible is only available on a subset of infected individuals for whom blood samples are obtained for lab

testing. In this thesis, we analyze such data on hand, foot and mouth disease (HFMD), which is an acute contagious viral infection that has caused large-scale outbreaks in Asia during the past decade (Tong and Bible, 2009). HFMD can be caused by a number of different pathogens and often involves mild or moderate symptoms such as fever, oral ulcer or rashes on the hand and foot; in severe cases the disease progresses and causes problems with the nervous system, with symptoms of respiratory and circulatory disturbance. Enterovirus 71 (EV71) and Coxackie A16 (CA16) are the most common pathogens associated with HFMD. Little is known about the etiology of the specific pathogens primarily responsible for HFMD, the factors (for example, meteorological) associated with their spread, or an effective means of public health intervention. Clues to these issues will greatly benefit authorities charged with policy making to control HFMD. In 2003, the Chinese Center for Disease Control and Prevention (CCDC) established a disease surveillance system which regulates the reporting of 39 notifiable infectious diseases including HFMD. The purpose of the surveillance system is to monitor epidemics of infectious diseases, identify areas of high case occurrence, predict and control epidemics, and provide information for formulating policy. Each reported case from the CCDC infectious disease surveillance system consists of the patient's geographical location, gender, age, and the symptom onset date. Additionally, the surveillance data contains the severity of symptoms of each reported case, date of diagnosis, date of death (if applicable), and pathogen responsible for the infection on a subsample of cases. The exact sampling plan varies by region, but in each region a large proportion of severe cases are sampled for virology, with a far smaller proportion of non-severe cases (Table 1.1). More information about these data can be found in Wang et al. (2011).

We consider a stratified population partitioned by both age group and sex. The two age groups are [0, 3) and [3, 100) years of age. Numbers of cases, and severe cases are presented in Table 1.1 by age, sex, year, and pathogen, if available. Between 2009 and 2011, there were 4.7 million cases of HFMD recorded, and less than 2% of those cases were classified as severe. Approximately 63% were seen in children under 3 years of age, and 37% of all cases are female. Only 3% of all HFMD cases recorded between 2009 and 2011 were subsampled

for virology information, while approximately 44% of severe cases have information about the attributable pathogen. Importantly, although 51% of all HFMD cases were found to be caused by EV71, the pathogen was responsible for nearly 82% of all severe HFMD cases. Much of the pathogen-specific research has focused on EV71, since it tends to be the pathogen responsible for the majority of severe cases.

		2009		2010		2011		Total	
	Population	Case	Severe	Case	Severe	Case	Severe	Case	Severe
Under 3 years old									
Female	23,529,240	261,687	3,958	439,789	8,328	380,112	5,152	1,081,588	17,438
Male	25,456,752	455,407	7,343	767,841	16,089	660,534	9,475	1,883,782	32,907
3 years and older									
Female	609,058,569	166,966	888	287,857	2,249	220,454	1,504	675,277	4,641
Male	641,458,527	269,777	1,638	465,551	4,115	360,520	2,651	1,095,848	8,404
Female	632,587,809	428,653	4,846	727,646	10,577	600,566	6,656	1,756,865	22,079
Male	666,915,279	725,184	8,981	1,233,392	20,204	1,021,054	12,126	2,979,630	41,311
Total	1,299,503,088	1,153,837	13,827	1,961,038	30,781	1,621,620	18,782	4,736,495	63,390
Subsampled for Virology									
EV71		11,498	3,334	34,873	12,097	33,552	7,115	79,923	22,546
CA16		7,574	190	18,100	571	17,995	490	43,669	1,251
Other		5,837	612	13,168	2,180	13,994	1,000	32,999	3,792
Total Subsampled		24,909	4,136	66,141	14,848	65,541	8,605	156,591	27,589

Table 1.1: Number of HFMD cases and severe cases by year and demographic stratum for all of China.

We are interested in exploring the temporal dynamics of pathogen-specific (EV71 and CA16) HFMD, as well as the importance of time-varying covariates, for example, meteorological variables, on the risk of EV71- and CA16-related HFMD. We are also interested in how the covariate effects differ between the two primary pathogens. The epidemic curves for HFMD epidemics from 2009 to 2011 for all of China are presented in Figure 1.1. Note that these are presented on the log scale for clarity. The number of subsampled cases with non-missing pathogen information are included in the same figure. The colors represent the available pathogen information by week. Gray bars represent the unsampled HFMD cases, while dark blue, blue, and light blue bars indicate the number of subsampled cases that were found to be caused by EV71, CA16, or Other pathogens, respectively. The vast majority of

cases are missing pathogen information. Within each of the three years, there are two epidemic peaks, one in the early summer (June), and the other in the fall (September-October).

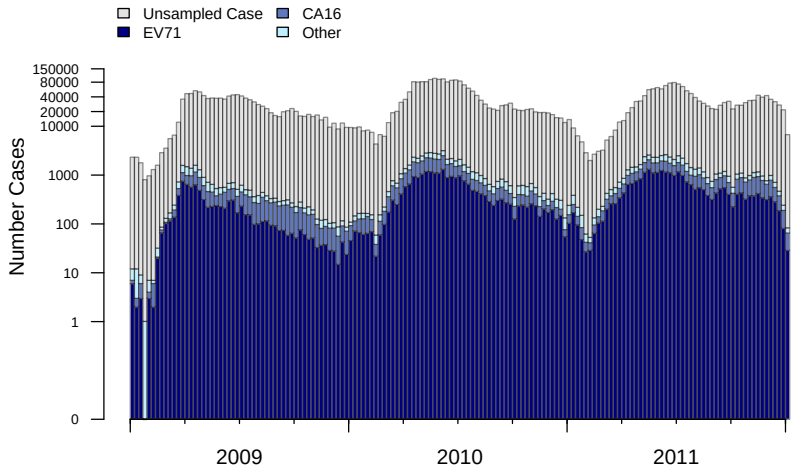


Figure 1.1: Epidemic curves of the 2009-2011 hand, foot, and mouth disease epidemics in China on *log scale*. The majority of cases are missing pathogen information. Few cases each week have specific pathogen information. Cases are aggregated by week.

In Chapter 3, we develop a fast and simple procedure that allows us to impute unobserved pathogen-specific disease counts with corresponding standard errors, which we use to learn about area level covariate effects using standard statistical methods.

1.1.2 School located influenza vaccination in Florida

The Alachua County school-located influenza vaccine (SLIV) program, described in detail in Tran et al. (2010), was fully implemented at the start of the 2009/2010 school year with the goal of vaccinating 70% of pre-K through 8th graders in Alachua county. The program was subsequently expanded to include high-school students. For the SLIV program, there is interest in the direct effects of the seasonal influenza vaccine within the target population, as well as the indirect effects of the program for the greater community. For the program's

target population, 5–17 year olds, the effectiveness is a measure of the overall protection, including both direct and indirect effects. For other age groups, the program's effectiveness measures the indirect effects of the SLIV program. We are interested in understanding the impact the county-wide influenza vaccination program has had on the number of influenza cases in Alachua county.

Tran et al. (2014) evaluated the Alachua SLIV program's effectiveness by estimating age-specific attack rates for Alachua County and comparison regions using data from Florida's Electronic Surveillance System for Early Notification of Community-based Epidemics, abbreviated as ESSENCE. ESSENCE is a statewide surveillance system that combines various data sources in a timely and comprehensive fashion. Participating emergency departments and urgent care centers report daily chief complaint data, along with basic demographic information, on a daily basis. ESSENCE defines cases of composite outcomes like influenza-like illness (ILI) and gastrointestinal illness (GI) based on the chief complaint data.

In 2015, it was discovered that coding changes in Alachua emergency department computers resulted in a severe under-counting of ILI cases. In the spring of 2011, the largest emergency department in Alachua county updated their computer system, which changed the nature of the chief complaint data being sent to ESSENCE. Previously, chief complaints were entered into a free-text box; the new system collected chief complaint data from a series of pull-down menus, which do not include multi-symptom options such as those included in the ILI syndromic case definitions.

Because of the questions about data quality in ESSENCE, especially for Alachua county, we do not use the ESSENCE data for our analyses. Instead we use data from the Florida Agency for Heath Care Administration (AHCA). While the AHCA data is not subject to the same reporting bias as the ESSENCE data, since all emergency departments are required to report by the state, it is still likely to be subject to unobserved confounding like that in ESSENCE. For our purposes, the AHCA data is a suitable replacement for the ESSENCE data.

AHCA provides quarterly data consisting of de-identified electronic medical records from

patient visits to emergency departments throughout Florida. For each visit recorded in the AHCA data set, we know the year and quarter of the visit, the county of the facility, and patient information including sex, age, county of residence, and up to 11 International Classification of Diseases, 9th Revision (ICD-9) diagnosis codes. We consider a patient a case if any of the possible 11 ICD-9 diagnosis codes is included in the syndrome definitions described in Appendix B.2. AHCA has an average of 1,575,000 recorded ED visits each quarter.

In Chapter 4, we explore the AHCA data using a negative control approach to account for the biases arising from differences in healthcare seeking behaviors across counties. We then use negative controls in a disease mapping setting to evaluate the effect of the vaccination program in Alachua county.

1.1.3 Measles epidemics in Germany

Measles is a highly contagious viral infection that can result in death for young or malnourished children. According to the WHO, measles is one of the most contagious diseases known. The average number of secondary infections that arise from a single infective in a completely susceptible population is between 15 and 20 for measles (Sudfeld et al., 2010; Keeling and Rohani, 2008). Although a vaccine is available, measles remains a leading cause of death among young children (Strebel et al., 2013). After a single dose of the MMR vaccine, it is estimated that between 85% and 95% of children will develop immunity, depending on the age of vaccination. A second dose provides nearly 99% vaccine efficacy (Sudfeld et al., 2010). Even with such an effective vaccine, because measles is highly infectious, more than 93% of the population needs to be immune in order to prevent epidemics (CDC, 1998; WHO, 2009). Hence, even in countries with well establish vaccination programs, small outbreaks persist.

We consider data collected on measles outbreaks in Germany from 2005 through 2007. The data come from Germany's national disease surveillance system, the Robert Koch Institute (RKI). This data has been previous used to examine the relationship between vaccination coverage and the size of measles outbreaks; further details about this data and previous anal-

ysis can be found in (Herzog et al., 2011). Weekly disease counts and estimated vaccination coverage data sets are both included in the `surveillance` package for `R` (Held et al., 2005). We are interested in appropriately accounting for vaccination coverage in spatio-temporal models using aggregate data.

In Chapter 5, we develop an ecological framework for infectious disease data. This approach is subsequently used to create an aggregate consistent model for estimating vaccine effects.

1.2 Organization of dissertation

In Chapter 2, we briefly introduce infectious disease and disease mapping models that we build upon throughout the dissertation. We also discuss common computing approaches and methods to asses model fit. In Chapter 3, we develop an approach to obtain fast pathogen-specific disease count estimates that can be used to learn about the pathogen-specific temporal dynamics of HFMD in China. This work has been published in Fisher et al. (2016). In Chapter 4, we formalize via an explicit model the negative controls approach to adjust for confounding and then extend the model to account for spatially structured outcomes. In doing so, we also offer insight into the strong assumptions required for employing negative controls to account for the unmeasured confounding in observational data. In Chapter 5, we develop an ecological model to account for vaccination coverage and investigate how differences in vaccination coverage influence measles outbreaks in Germany. Chapter 6 contains concluding remarks.

Chapter 2

BACKGROUND

2.1 Models for infectious diseases

2.1.1 Compartmental models

Compartmental representations of the disease process are a common approach to modeling infectious disease data. We describe the susceptible-infectious-recovered (SIR) model in both continuous and discrete time. The SIR model describes the progression of a population through three disease states: S denotes the proportion of the population that is susceptible, I represents the proportion of the population that is infected, and R is the proportion recovered. To start, we assume a fixed population of size N, so that $S + I + R = 1$. The SIR model can be depicted pictorially by a flow diagram, as in Figure 2.1.

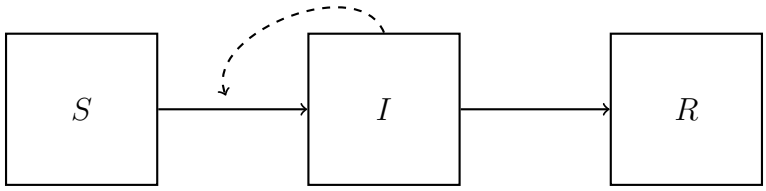


Figure 2.1: Diagram of SIR model. Solid arrows show the movement from S to I to R; the dashed arrow shows how the number of infectious individuals influences the rate at which individuals move from susceptible to infected.

We assume homogeneous mixing, meaning that the probability of contact between any two individuals in the population is the same. The rate of transition from the I to the R compartment is called the removal rate, represented by γ. The reciprocal, $1/\gamma$, corresponds to the average time of infection. The rate of the transition from susceptible to infected (S to

I) is more difficult to specify since there are many contributing factors; the rate of infection will depend on population contact structure, the prevalence of infectious individuals, and the per contact transmission probability. We let β represent the product of the transmission probability and contact rates so that $\lambda = \beta I$ is the force of infection. In Section 5.2.2, we discuss the nature of the transmission term in more detail.

In continuous time, the deterministic movement between the three compartments can be described by a set of differential equations:

$$\frac{dS}{dt} = -\beta SI, \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I. \tag{2.3}$$

The SIR model described in equations (2.1) through (2.3), along with a set of initial conditions, can be solved analytically to arrive at a deterministic solution. The solutions to deterministic models provide insight into epidemic dynamics when a few infectious individuals are introduced into a large, initially naive population. While deterministic models can provide insight into the long-term, large population dynamics of an epidemic, they are not suitable for small populations, or certain scientific questions such as the probability of a major outbreak (Britton, 2010).

An alternative and complement to the deterministic model is the stochastic compartmental model (Andersson and Britton, 2000). In contrast to deterministic models, stochastic models may be defined as a continuous or discrete time Markov chain. An advantage to the stochastic model is that it allows us to model the probability of disease transmission, which is frequently of interest. Additionally, stochastic models can provide measures of uncertainty when estimating quantities of interest.

To define the stochastic SIR model, we change notation slightly. We now let X_t denote the total number of susceptible individuals, Y_t is the total number of infected individuals,

and Z_t the total number recovered. Assuming a closed population of size N, we only have to track the number of susceptibles and infectives. The probability of infection is

$$\Pr(X_{t+h} = i-1, Y_{t+h} = j+1, Z_{t+h} = k | X_t = i, Y_t = j, Z_t = k) = \beta ijh/N + o(h),$$

and the probability of recovery is

$$\Pr(X_{t+h} = i, Y_{t+h} = j-1, Z_{t+h} = k+1 | X_t = i, Y_t = j, Z_t = k) = \gamma jh + o(h),$$

for small h. The length of time an individual spends in a compartment is exponentially distributed, with some compartment-specific rate. For the interval $(\tau_1, T]$, where τ_1 is the time of the first infection, $\boldsymbol{\tau} = (\tau_2, \ldots, \tau_m)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ be vectors of infection and recovery times, respectively. Conditional on $(\boldsymbol{\tau}, \boldsymbol{\omega})$, the likelihood is

$$L(\beta, \gamma, \tau_1 | \boldsymbol{\tau}, \boldsymbol{\omega}) = \prod_{i=1}^{n} \gamma Y_{\omega_i^-} \prod_{j=1}^{m-1} \beta X_{\tau_i^-} Y_{\tau_i^-} \exp\left(- \int_{\tau_1}^{T} (\beta X_t Y_t + \gamma Y_t) dt \right), \qquad (2.4)$$

where $Y_{t^-} = \lim_{t \to t^-} Y_t$. While maximum-likelihood estimates can be obtained from (2.4), we generally do not observe infection and recovery times, making inference difficult for these types of models. A variety of approaches have been developed to address the inference problem for compartmental models. Both Gibson and Renshaw (1998) and O'Neill and Roberts (1999) developed an auxiliary variable approach to inference for the continuous-time SIR model. However, the auxiliary variable approach is only computationally tractable for small populations. When simulation from the model is straightforward, a number of so-called plug-and-play approaches can be considered (He et al., 2010; Ionides et al., 2006; McKinley et al., 2009).

Alternatively, we can consider a discrete approximation to the continuous time series SIR model, following Lekone and Finkenstädt (2006). For the time interval $(t, t+1]$, let B_t denote the number of susceptibles that become infected and C_t the number of infectives who

recover. Then, the discretized SIR model is specified by

$$\begin{aligned} X_{t+1} &= X_t - B_t, \\ Y_{t+1} &= Y_t + B_t - C_t, \\ Z_{t+1} &= Z_t + C_t, \end{aligned}$$

where B_t and C_t are random variables defined by

$$B_{t+1} \sim \text{Binomial}\left(X_t, P_t\right) \qquad \text{and} \qquad C_{t+1} \sim \text{Binomial}\left(Y_t, p_c\right).$$

The transition probabilities are modeled as independent, exponential compartment-specific waiting times for each individual in the population. For a susceptible individual, the time until infection is exponentially distributed with a rate equal to the transmission rate, β. Thus, for an individual who is susceptible at time t, the probability of escaping infection in the next period, $(t, t+1]$, is $\exp(-\beta)$, and

$$\Pr(\text{infected in } (t, t+1] \,|\, \text{susceptible at } t) = 1 - \exp(-\beta).$$

The number of new infections, B_t, follows a binomial distribution, assuming independent times until infection among susceptibles. Hence,

$$B_{t+1}|Y_t \sim \text{Binomial}\big(X_t, 1 - \exp\left(-\beta Y_t/N\right)\big).$$

Analogously, the number of new recoveries, C_t, can be written as

$$C_{t+1} \sim \text{Binomial}\big(Y_t, 1 - \exp\left(-\gamma\right)\big).$$

If we additionally assume that all infected individuals are recovered in the next time period,

i.e. $C_t = Y_t$, then

$$Y_{t+1}|Y_t \sim \text{Binomial}\big(X_t, 1 - \exp\left(-\beta Y_t/N\right)\big),$$

which is a Reed-Frost chain binomial SIR model. The discrete-time SIR model is appealing for surveillance data, which is frequently aggregated into daily or weekly observations.

2.2 Disease mapping and spatial regression for aggregate data

The Besag-York-Mollié (BYM) model is commonly used to analyze area-level disease counts (Besag et al., 1991). Let Y_i and E_i represent the observed and expected disease counts in area i respectively, for $i = 1, \ldots, n$; let θ_i denote the area-specific relative risk. The BYM model takes the form

$$\begin{aligned} Y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \\ \log\theta_i &= \mu + \epsilon_i + S_i, \end{aligned} \tag{2.5}$$

where μ is an overall level; ϵ_i represent unstructured random effects so that $\epsilon_i \sim \text{N}(0, \sigma_\epsilon^2)$ and S_i represent spatially structured random effects, both for $i = 1, \ldots, n$.

The spatially structured random effects $\boldsymbol{S} = (S_1, \ldots, S_n)$ follow an intrinsic conditional autoregressive (ICAR) distribution, where we define area-level effects conditionally on the values of the neighboring areas. Let ∂i denote the set, and n_i the number, of neighbors in area i. Then, conditional on the neighbors, the random effect in area i, S_i is normally distributed such that

$$S_i|\boldsymbol{S}_{-i}, \tau \sim N\left(\frac{1}{n_i}\sum_{j\in\partial i} S_j, \frac{1}{n_i\tau}\right),$$

where $\boldsymbol{S}_{-i} = \{S_j : j \neq i\}$. This form does not produce a proper joint distribution. Rather

the joint distribution takes the form of an intrinsic Gaussian Markov random field (IGMRF)

$$p(\boldsymbol{S}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{j\in\partial i}(S_i - S_i)^2\right),$$
$$= \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{S}^T K \boldsymbol{S}\right),$$

where K is a structure matrix with elements

$$K_{ij} = \begin{cases} n_i & j = i \\ -1 & j \in \partial i \\ 0 & \text{otherwise} \end{cases},$$

when all areas are connected. The structure matrix K has rank $n-1$ as each row of the $n \times n$ matrix sums to zero.

2.3 Bayesian computing

2.3.1 Hierarchical Bayesian modeling

In their simplest form Bayesian hierarchical models are defined by three levels: the data model, the latent model, and the hyperparameter model. Joint densities of known and unknown quantities induced by Bayesian hierarchical models take the form

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.6}$$

where the data y_i for $i = 1, \ldots, n$ are conditionally independent such that $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta})$, the latent variables $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ are hyperparameters with prior density $p(\boldsymbol{\theta})$. Bayes rule gives us the posterior distribution

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

The marginal likelihood $p(\boldsymbol{y})$ is also referred to as the normalizing constant, and is defined as

$$p(\boldsymbol{y}) = \int_{\boldsymbol{x}} \int_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\, p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{x}.$$

For the models developed in this dissertation, we either approximate integrals using the integrated nested Laplace approximation (INLA) method, or simulate samples from the posterior distributions. In this section, we describe the computational approaches we take to Bayesian inference for this dissertation. In Section 2.3.2, we describe the Markov chain Monte Carlo (MCMC) approach to simulating the posterior distribution. In Section 2.3.3, we describe the INLA approach that avoids MCMC methods and instead deterministically approximates the intractable integrals.

2.3.2 Markov chain Monte Carlo

MCMC is a common Bayesian approach to explore the posterior distribution. The idea behind MCMC is to construct a Markov chain, whose stationary distribution is the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$.

The Metropolis-Hastings algorithm is a general approach to MCMC. Starting with an initial value, $\boldsymbol{\theta}^{(0)}$, the algorithm proceeds, for iterations $s = 1, \ldots, S$

1. Sample $\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(s-1)} \sim g(\cdot|\boldsymbol{\theta}^{(s-1)})$, where $g(\cdot|\boldsymbol{\theta}^{(s-1)})$ is the proposal distribution.

2. Compute the acceptance ratio r

$$r = \frac{p(\boldsymbol{\theta}^{\star}|\boldsymbol{y}) \,/\, g(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{y}) \,/\, g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^{\star})}.$$

3. Update such that

$$\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{with probability } \min(1,\, r), \\ \boldsymbol{\theta}^{(s-1)} & \text{otherwise.} \end{cases}$$

MCMC allows us to obtain a large sample from the posterior distribution once the chain has converged. However, MCMC can be very slow for high dimensional problems, making using such an approach impractical for real world applications.

When parameters are highly correlated, more advanced approaches are necessary. Hamiltonian Monte Carlo (HMC) is an MCMC algorithm that can perform better than standard appraoches in complex problems (Duane et al., 1987). HMC introduces auxilliary variables into the standard Metropolis-Hastings algorithm so as to explore the probability space in a more efficient fashion by using gradient information. HMC has recently been implemented in the `rstan` package in `R` by the Stan Development Team (2015a,b).

2.3.3 Integrated nested Laplace approximation

When the latent variables in (2.6) are normally distributed, integrated nested Laplace approximation (INLA) provides a fast alternative to MCMC for estimating the posterior distribution (Rue et al., 2009). INLA has been implemented in the `INLA` package for `R` (Martins et al., 2013). We describe the basic idea behind INLA briefly here.

Suppose we have a hierarchical model, $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and the latent variable $\boldsymbol{x}$ is assumed to be Gaussian, i.e. $p(\boldsymbol{x}|\boldsymbol{\theta}) \sim \text{Normal}(\mu(\boldsymbol{\theta}_1), \boldsymbol{Q}(\boldsymbol{\theta}_2))$. Since $p(\boldsymbol{x}|\boldsymbol{\theta})$ is Gaussian, the posterior distribution has the form

$$p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta}_2)|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta}_2)\boldsymbol{x} + \sum_i \log p(y_i|x_i,\boldsymbol{\theta}_1)\right).$$

In practice, we are interested in learning about the hyperparameters, $p(\boldsymbol{\theta}|\boldsymbol{y})$, or the latent

field, $p(\boldsymbol{x}|\boldsymbol{y})$. The marginal distributions are

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{2.7}$$

and

$$p(\theta_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-i}. \tag{2.8}$$

INLA uses Laplace approximations and numerical integration to approximate the marginal posterior distributions of interest. Let $\tilde{p}(\cdot|\cdot)$ denote an approximate conditional distribution. The posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ is approximated with the Laplace approximation

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) = \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})},$$

where $\tilde{p}_G(\cdot|\cdot)$ is the Gaussian approximation, and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode. Numerical integration is used to then approximate (2.8)

$$\tilde{p}(\theta_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-i}} \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-i}.$$

The posterior distribution for the latent field in (2.7) is approximated via

$$\tilde{p}(x_i|\boldsymbol{y}) = \int \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

While INLA is very fast and accurate, there are some limitations. Importantly, INLA assumes a Gaussian latent field, which can encompass a large variety of models, but is not always applicable.

2.4 Measures of model fit

We briefly describe two common measures of model fit that allow for comparison across a variety of fitted models: the deviance information criteria (DIC) and the Watanabe-Akaike information criteria (WAIC). Both approaches generally try to balance model fit and model complexity.

We consider a generic data model, $p(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters, and define a prior $p(\boldsymbol{\theta})$, so that we have a posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. The deviance, defined as $D = -2\log p(\boldsymbol{y}|\boldsymbol{\theta})$, provides a general measure of lack of model fit. Let $\bar{\boldsymbol{\theta}} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{y}]$ be the posterior mean. DIC is a common approach to compare Bayesian models (Spiegelhalter et al., 2002). We first define the *effective number of parameters*, represented by p_D, to be

$$p_D = \mathrm{E}_{\boldsymbol{\theta}}\left[D|\boldsymbol{y}\right] - D(\bar{\boldsymbol{\theta}}).$$

The effective number of parameters is the posterior mean deviance minus the deviance at the posterior mean. Then, DIC is given by

$$\mathrm{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D = \bar{D} + p_D, \tag{2.9}$$

where $D(\bar{\boldsymbol{\theta}})$ is the deviance evaluated at $\bar{\boldsymbol{\theta}}$, where p_D is the effective number of parameters, and $\bar{D} = \mathrm{E}_{\boldsymbol{\theta}}\left[D|\boldsymbol{y}\right]$ is the posterior mean deviance. In general, lower values of DIC indicate a better model fit. However, there are many known issues with DIC. DIC is not invariant to reparameterization and is not consistency. Most importantly, DIC lacks theoretical justification (Spiegelhalter et al., 2014).

WAIC takes a fully Bayesian approach towards assessing model accuracy and complexity that based on the log pointwise predictive density (Watanabe, 2013; Gelman et al., 2014). The log pointwide predictive density is

$$\sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}) = \sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\theta.$$

And the effective number of parameters is defined as

$$p_W = \sum_{i=1}^{n} \log \mathrm{E}[p(y_i|\boldsymbol{\theta})|\boldsymbol{y}] - \mathrm{E}[\log p(y_i|\boldsymbol{\theta})|\boldsymbol{y}].$$

The WAIC is

$$\mathrm{WAIC} = 2p_W - 2\sum_{i=1}^{n} \log(p_i|\boldsymbol{y}). \tag{2.10}$$

In practice these are computed using posterior samples. Let $\boldsymbol{\theta}^s$ for $s = 1, \ldots, S$ be samples from the posterior distribution. Then

$$\widehat{\mathrm{WAIC}} = 2\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^s)\right) - 2\sum_{i=1}^{n}\left[\log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^s)\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(y_i|\boldsymbol{\theta}^s)\right].$$

From a Bayesian perspective, has the desirable property of utilizing the whole posterior distribution, rather than just the posterior mean, as in DIC. However, since WAIC relies on partitioning the data, spatially or temporally structured data can be problematic.

Chapter 3

TIME SERIES MODELING OF PATHOGEN-SPECIFIC DISEASE PROBABILITIES WITH SUBSAMPLED DATA

3.1 Introduction

In this chapter, we analyze data on hand, foot and mouth disease (HFMD) with the goal of estimating pathogen-specific dynamics when information regarding the responsible pathogen is only available for a subsample of cases. Recall our interest lies in estimating the temporal dynamics of pathogen-specific cases of hand, foot, and mouth disease (HFMD), using data where only a subsample of cases have pathogen-specific information. We are also interested in how the associations differ between the two primary pathogens.

In Section 3.3 we derive a model for the unobserved pathogen-specific cases. In Section 3.4, we develop an approach to inference that uses smoothed estimates of unobserved counts to develop a likelihood for the parameters of interest that reflect this uncertainty. In Section 3.5, we perform simulations to compare our approach to the MCMC for the full model. In Section 3.6, we apply our approach to the HFMD data to model pathogen-specific temporal dynamics.

3.2 Notation

For this analysis, we consolidate the data into six geographical regions of mainland China shown in Figure 3.1. For a single geographic region, we let N_j be the population in the jth age-gender stratum and is assumed constant over time. Let Y_{tj}^{G} be the total number of new HFMD cases in stratum j, week t of pathogen type G, with G=E,C,O, representing EV71, CA16 and Other. EV71 and CA16 represent the majority of the cases, and are of primary interest. We use a second superscript of either M or S to denote mild (non-severe) and severe

cases. Hence, Y_{tj}^{GM} and Y_{tj}^{GS} are the number of mild and severe cases caused by pathogen G in week t and demographic stratum j so that $Y_{tj}^{\mathrm{G}} = Y_{tj}^{\mathrm{GM}} + Y_{tj}^{\mathrm{GS}}$; these numbers are unobserved. Instead we observe $Y_{tj}^{\cdot\mathrm{S}}$ and $Y_{tj}^{\cdot\mathrm{M}}$, the total number of severe and mild cases, respectively, in week t and strata j.

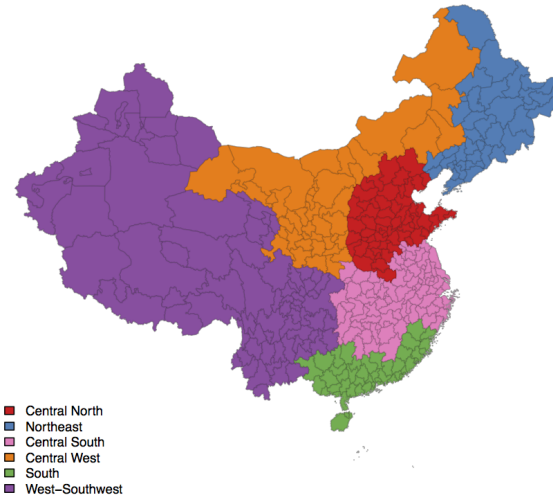


Figure 3.1: Six geographic regions of mainland China

Random samples k_{tj}^{M} and k_{tj}^{S} of mild and severe cases are taken within week t to gain pathogen information. We emphasize that stratified sampling by stratum j is not carried out, but we can distinguish the cases and so split the total samples by stratum, which is useful for building a model. The subsampled cases are lab tested to determine the pathogen responsible for HFMD and we let $Z_{tj}^{\mathrm{GM}}, Z_{tj}^{\mathrm{GS}}$ be the number of mild and severe cases of pathogen type G so that $k_{tj}^{\mathrm{M}} = Z_{tj}^{\mathrm{EM}} + Z_{tj}^{\mathrm{CM}} + Z_{tj}^{\mathrm{OM}}$ and $k_{tj}^{\mathrm{S}} = Z_{tj}^{\mathrm{ES}} + Z_{tj}^{\mathrm{CS}} + Z_{tj}^{\mathrm{OS}}$. Figure 3.2 presents a graph of the data structure for a generic week and demographic stratum, with variables in square boxes being observable and those in circles being unobserved. The arrows encode the conditional dependencies in the graph, which are determined by the model we develop in Section 3.3.

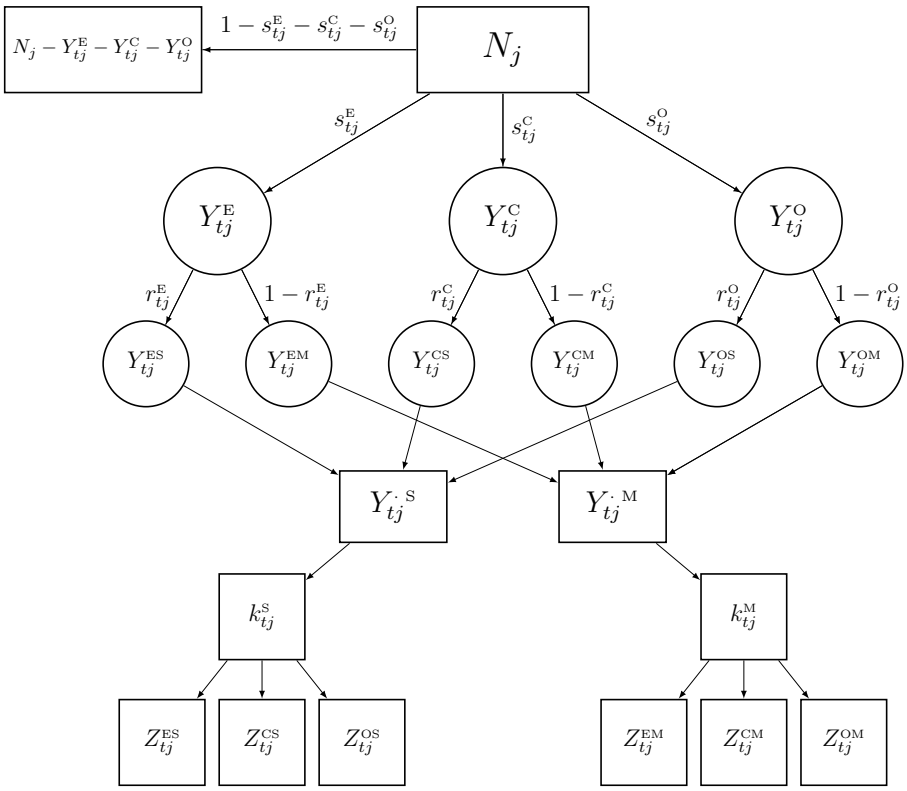


Figure 3.2: Graphical representation of the conditional independencies in a generic week t and a generic age-gender stratum j for the model for the China HFMD. Variables in square boxes are observed, those in circles are unobserved. Generic letters are N for population, Y for cases, k for total subsampled cases and Z for the numbers of cases in each category. The superscripts E, C and O are shorthand for pathogen type EV71, CA16 and Other, and M and S are shorthand for moderate (non-severe) and severe case types.

3.3 Model formulation

We first derive a model where all variables are observed, and then describe how to handle the scenario where pathogen information is measured on only a subsample of all cases.

3.3.1 Disease probabilities

We assume that in week t and stratum j each member of the population can stay uninfected, or be infected by EV71, CA16 or another pathogen, and assign a multinomial distribution to these counts:

$$Y_{tj}^{\mathrm{E}}, Y_{tj}^{\mathrm{C}}, Y_{tj}^{\mathrm{O}}, N_j - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{O}} | \boldsymbol{s}_{tj} \sim \mathrm{Multinomial}(N_j, \boldsymbol{s}_{tj}),$$

where $\boldsymbol{s}_{tj} = [s_{tj}^{\mathrm{E}}, s_{tj}^{\mathrm{C}}, s_{tj}^{\mathrm{O}}]$ and

$$\begin{aligned} s_{tj}^{\mathrm{E}} &= \Pr(\text{EV71 case} \mid \text{week } t, \text{stratum } j), \\ s_{tj}^{\mathrm{C}} &= \Pr(\text{Cox16 case} \mid \text{week } t, \text{stratum } j), \\ s_{tj}^{\mathrm{O}} &= \Pr(\text{Other case} \mid \text{week } t, \text{stratum } j). \end{aligned}$$

If the pathogen responsible for disease were available for all cases, it would be possible to use inferential methods developed for compartmental models, such as described in Section 2.1.1. In such models, the probabilities $s_{tj}^{\mathrm{E}}, s_{tj}^{\mathrm{C}}, s_{tj}^{\mathrm{O}}$ would depend on the number of cases of the respective pathogens in the previous week, but these variables are unobserved.

3.3.2 Laboratory testing subsampling

We observe subsamples of severe and mild cases that are randomly sampled for lab testing, i.e. to determine which pathogen was responsible for disease. A natural model for the numbers falling into the three pathogen classes is the multivariate hypergeometric distribution,

with one distribution for mild and one for severe cases:

$$Z_{tj}^{\mathrm{ES}}, Z_{tj}^{\mathrm{CS}}, Z_{tj}^{\mathrm{OS}} | k_{tj}^{\mathrm{S}}, Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}} \sim \mathrm{MultHyperGeom}(Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}}, k_{tj}^{\mathrm{S}}) \quad (3.1)$$

$$Z_{tj}^{\mathrm{EM}}, Z_{tj}^{\mathrm{CM}}, Z_{tj}^{\mathrm{OM}} | k_{tj}^{\mathrm{M}}, Y_{tj}^{\mathrm{EM}}, Y_{tj}^{\mathrm{CM}}, Y_{tj}^{\mathrm{OM}} \sim \mathrm{MultHyperGeom}(Y_{tj}^{\mathrm{EM}}, Y_{tj}^{\mathrm{CM}}, Y_{tj}^{\mathrm{OM}}, k_{tj}^{\mathrm{M}}). \quad (3.2)$$

3.3.3 Model development

The event of becoming a case in week t is statistically rare for all pathogens and within all stratum j and so we make the approximation

$$Y_{tj}^{\mathrm{G}} | s_{tj}^{\mathrm{G}} \sim \mathrm{Poisson}(N_j s_{tj}^{\mathrm{G}}), \quad (3.3)$$

for pathogens of type $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$. With three pathogens (EV71, CA16, Other), $J = 4$ demographic groups, and 157 weeks of data this leaves us with $3 \times 4 \times 157 = 1,884$ parameters to estimate. Recall that in the three years of data, 3.3% of all HFMD cases were sampled for virology, and only 17.6% of all samples were from severe cases (see Table 1.1). Moreover, even with relatively few ($J = 4$) strata, there are still weeks with no samples for a given stratum. Depending on the region, there are between 18 and 141 weeks where at least one stratum has no subsamples.

As we see in Figure 3.3, the number of HFMD cases subsampled for virology vary by week, region, and demographic stratum. In general, there are more samples for boys under 3 years of age than the other demographic strata. Males account for approximately 60% of the total number of HFMD cases, so it is not surprising to see more samples in the corresponding stratum. Similarly, we see a higher number of cases sampled among the younger patients.

Severe cases are sampled for virology more often than mild cases. Figures 3.4(a) and 3.4(b) show the proportion of mild and severe cases, respectively, that were subsampled by week, region, and demographic strata. Notice that at most 33% of mild cases in a given region, stratum, and week are sampled, but that is very infrequent (three weeks sampled over 25% of mild cases). In contrast, a large proportion of severe cases are sampled for

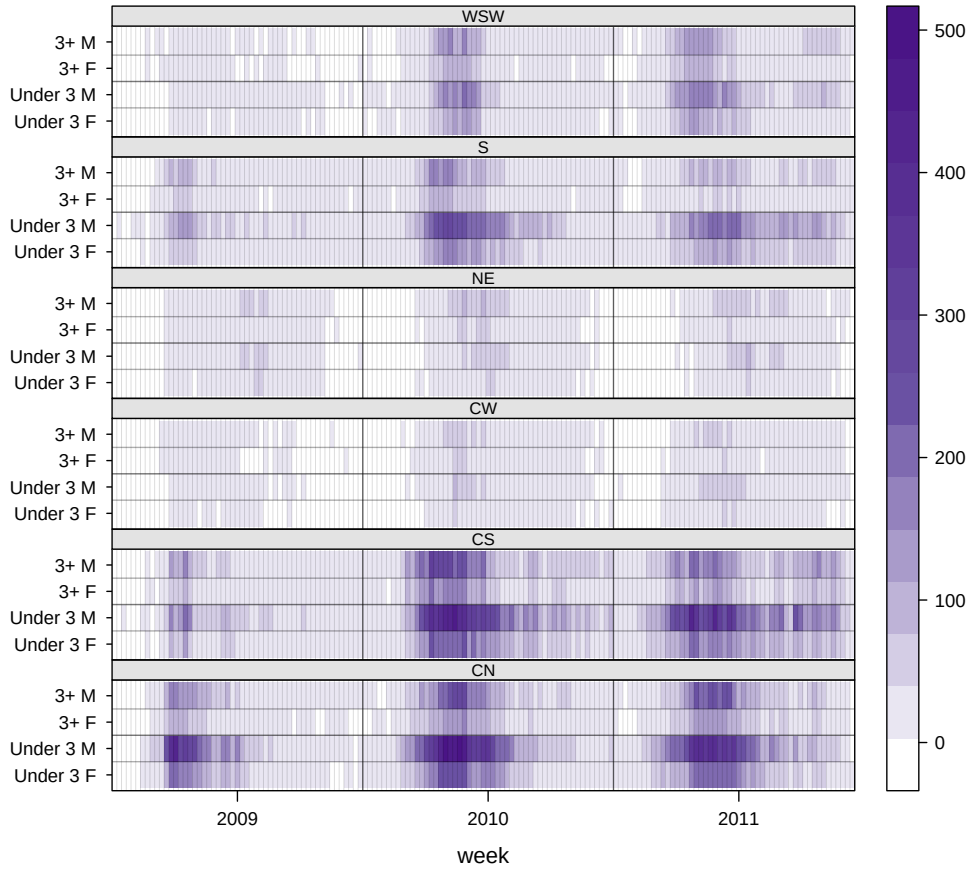


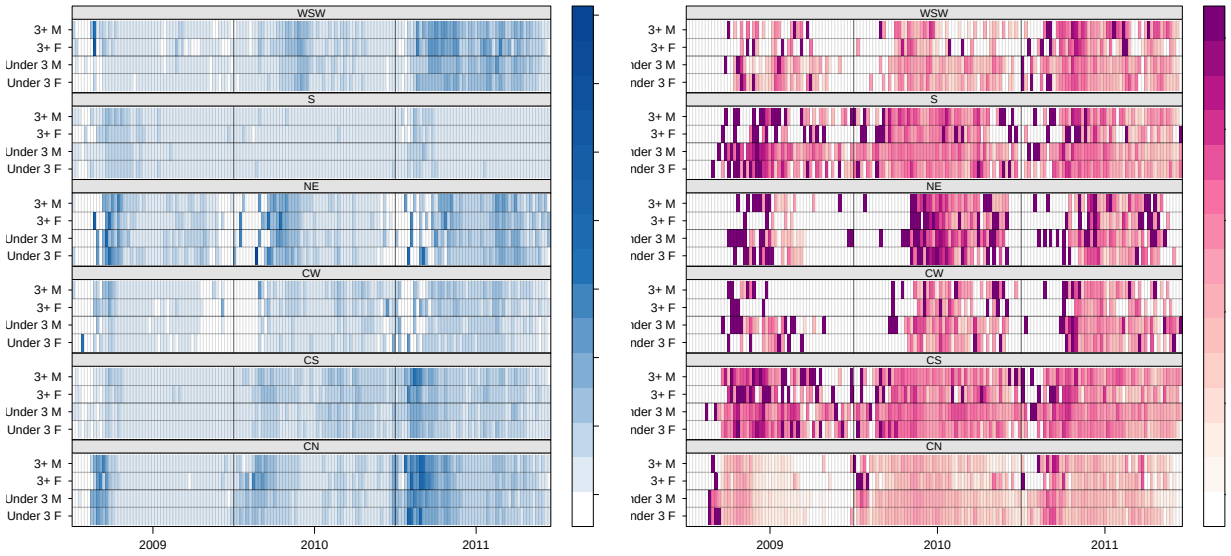
Figure 3.3: Number of HFMD cases (both mild and severe) subsampled for virology by region and demographic strata.

virology, and it is not uncommon for all severe cases in a given stratum and week to be subsampled for virology.

Given the limited data, we cannot estimate distinct probabilities s^{G}_{tj} and so instead we make the proportionality assumption

$$s^{\mathrm{G}}_{tj} = \theta^{\mathrm{G}}_t \times p^{\mathrm{G}}_j, \tag{3.4}$$

where $p^{\mathrm{G}}_j = \Pr(\text{Pathogen G case in a generic week} \mid \text{stratum } j)$ and θ^{G}_t is the relative risk



(a) Mild cases. (b) Severe cases.

Figure 3.4: Proportion of HFMD cases subsampled for virology by case severity, region, and demographic strata.

at time t. Hence, under this model, we assume the effect of time and covariates (which we will model in θ_t^{G}) act equally across all stratum. Analogous to internal standardization, as commonly used in disease mapping (Wakefield et al., 2000), we use the study data to pre-estimate p_j^{G}. Specifically, we begin with the marginal model

$$Y_j^{\mathrm{G}} = \sum_{t=1}^{T} Y_{tj}^{\mathrm{G}} \sim \text{Poisson}(N_j p_j^{\mathrm{G}}),$$

where $j = 1, \ldots, J$ and $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$. The MLE is $\widehat{p}_j^{\mathrm{G}} = Y_j^{\mathrm{G}}/N_j$ and we define the expected numbers as $E^{\mathrm{G}} = \sum_{j=1}^{J} N_j \widehat{p}_j^{\mathrm{G}}$ for $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$. We then obtain

$$Y_t^{\mathrm{G}} = \sum_{j=1}^{J} Y_{tj}^{\mathrm{G}} \sim \text{Poisson}(\theta_t^{\mathrm{G}} \times E^{\mathrm{G}}). \tag{3.5}$$

The usual estimate of the relative risk is

$$\widehat{\theta}_t^{\mathrm{G}} = \frac{Y_t^{\mathrm{G}}}{E^{\mathrm{G}}} = \frac{Y_t^{\mathrm{GM}} + Y_t^{\mathrm{GS}}}{E^{\mathrm{G}}} = \sum_{j=1}^{J} \frac{Y_{tj}^{\mathrm{GS}} + Y_{tj}^{\mathrm{GM}}}{E^{\mathrm{G}}}, \tag{3.6}$$

i.e. the standardized morbidity ratio (SMR). In practice, however, we do not observe the stratum- and pathogen-specific counts of cases, so the above procedure must be modified.

3.3.4 Previous approaches to inference

A natural Bayesian approach to inference for the above model would be to use MCMC methods to impute the unsampled pathogen-specific disease counts, Y_{tj}^{GS}. Bauer (2012) introduced unobserved counts of pathogen-specific severe and mild cases as auxiliary variables for a two-pathogen model of HFMD for the north-west region of China. However, in this setting, where the populations and numbers of cases are large and the number subsampled is small, MCMC methods to impute the unsampled pathogen-specific disease counts are computationally expensive. Particle MCMC methods (Andrieu et al., 2010) have been developed, but are difficult to implement and have not been applied to problems of this size (Dukic et al., 2012; Rasmussen et al., 2011). An alternative that we pursue is to estimate the unobserved pathogen-specific disease counts, and then formulate a likelihood for the estimated log relative risks based on the asymptotic normality of the estimator. Auxiliary variable methods (O'Neill and Roberts, 1999; Gibson and Renshaw, 1998) would be difficult to implement, given the large population sizes and small samples.

3.4 The hybrid approach

In this section, we develop a hybrid approach to inference when the standard approaches are computationally intractable. We refer to the proposed approach as a hybrid because we combine frequentist and Bayesian techniques in order to construct and smooth relevant summary statistics.

3.4.1 Estimation of counts

We propose a procedure to obtain smoothed estimates of the unobserved pathogen-specific disease counts by stratum and severity. Following Hoadley (1969) we place a multivariate Pólya prior on the unknown Y_{tj}^{G}'s; this model is the conjugate prior for a multivariate hypergeometric distribution likelihood. In our case (3.1) and (3.2) are the likelihoods for severe and mild cases, respectively.

The multivariate Pólya distribution is defined as a multinomial averaged over a Dirichlet($\boldsymbol{a}$) distribution. Let $\boldsymbol{X}$ be a generic random variable, with $\boldsymbol{X} = (X_1, \ldots, X_K)$. We denote the multivariate Pólya by $\boldsymbol{X} \sim \mathrm{MultPolya}(n, \boldsymbol{a})$ with

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!\Gamma(a_+)}{\Gamma(n + a_+)} \prod_{k=1}^{K} \frac{\Gamma(x_k + a_k)}{x_k!\Gamma(a_k)},$$

and where $\boldsymbol{a} = (a_1, \ldots, a_K)$ are specified a priori and $a_+ = \sum_{k=1}^{K} a_k$. The mean of this distribution is $\mathrm{E}[X_k] = na_k/a_+$, with variances and covariances (Hoadley, 1969, Section 4)

$$\mathrm{Var}[X_k] = \frac{na_k(n + a_+)(a_+ - a_k)}{a_+^2(1 + a_+)}, \qquad \mathrm{Cov}[X_k, X_{k'}] = -\frac{na_{k'}a_k(n + a_+)}{a_+^2(1 + a_+)}.$$

To illustrate the steps of the Bayesian procedure, we consider the problem of estimating severe cases for a given week t and strata, j. Let, $\boldsymbol{Y}_{tj}^{\mathrm{S}} = (Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}})$ be the unobserved cases we wish to estimate and $\boldsymbol{Z}_{tj}^{\mathrm{S}} = (Z_{tj}^{\mathrm{ES}}, Z_{tj}^{\mathrm{CS}}, Z_{tj}^{\mathrm{OS}})$ be the observed data. Recall, that conditional on the number of pathogen-specific severe cases, subsampled severe cases follow a multivariate hypergeometric distribution. Including the multivariate Pólya prior on the total number of pathogen-specific severe cases results in the following model,

$$\boldsymbol{Z}_{tj}^{\mathrm{S}} \,\big|\, \boldsymbol{Y}_{tj}^{\mathrm{S}}, k_{tj}^{\mathrm{S}} \sim \mathrm{MultHyperGeom}\left(Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}}; k_{tj}^{\mathrm{S}}\right),$$

$$\boldsymbol{Y}_{tj}^{\mathrm{S}} \,\big|\, \boldsymbol{\alpha}_j^{\mathrm{S}} \sim \mathrm{MultPolya}\left(Y_{tj}^{\cdot\mathrm{S}}; \boldsymbol{\alpha}_j^{\mathrm{S}}\right),$$

where $\boldsymbol{\alpha}_j^{\mathrm{S}} = \left(\alpha_j^{\mathrm{ES}}, \alpha_j^{\mathrm{CS}}, \alpha_j^{\mathrm{OS}}\right)$. This yields a posterior distribution for the unsampled severe cases of the form,

$$\boldsymbol{Y}_{tj}^{\mathrm{S}} - \boldsymbol{Z}_{tj}^{\mathrm{S}} \left| \boldsymbol{Z}_{tj}^{\mathrm{S}}, \boldsymbol{\alpha}_j^{\mathrm{S}}, k_{tj}^{\mathrm{S}} \right. \sim \mathrm{MultPolya}\Big(\left(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}}\right); \boldsymbol{\alpha}_j^{\mathrm{S}} + \boldsymbol{Z}_{tj}^{\mathrm{S}} \Big). \tag{3.7}$$

For $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$

$$\mathrm{E}\left[Y_{tj}^{\mathrm{GS}} - Z_{tj}^{\mathrm{GS}} \middle| Z_{tj}^{\mathrm{GS}}, Y_{tj}^{\cdot\mathrm{S}}, \boldsymbol{\alpha}_j^{\mathrm{S}} \right] = \left(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}}\right) \left(\frac{\alpha_j^{\mathrm{GS}} + Z_{tj}^{\mathrm{GS}}}{\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}}} \right),$$

where $\alpha_j^{\cdot\mathrm{S}} = \alpha_j^{\mathrm{ES}} + \alpha_j^{\mathrm{CS}} + \alpha_j^{\mathrm{OS}}$. Thus, we can derive a posterior estimate of the number of severe cases for a given pathogen as

$$\widetilde{Y_{tj}^{\mathrm{GS}}} = Z_{tj}^{\mathrm{GS}} + \left(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}}\right) \left(\frac{\alpha_j^{\mathrm{GS}} + Z_{tj}^{\mathrm{GS}}}{\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}}} \right).$$

The posterior distribution in (3.7) allows for computation of measures of uncertainty about the estimated counts of pathogen-specific severe cases. In particular, the posterior variance is

$$\mathrm{Var}\left[Y_{tj}^{\mathrm{GS}} \middle| Z_{tj}^{\mathrm{GS}}, Y_{tj}^{\cdot\mathrm{S}}, \boldsymbol{\alpha}_j^{\mathrm{S}} \right] = \frac{\left(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}}\right)\left(\alpha_j^{\mathrm{GS}} + Z_{tj}^{\mathrm{GS}}\right)\left(Y_{tj}^{\cdot\mathrm{S}} + \alpha_j^{\cdot\mathrm{S}}\right)\left(\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}} - \alpha_j^{\mathrm{GS}} - Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}}\right)^2 \left(\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}} + 1\right)}. \tag{3.8}$$

Similarly, we are able to estimate the covariance for the estimated counts of two pathogens. For example, for severe EV71 and CA16 cases, the posterior covariance is

$$\mathrm{Cov}\left[Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}} \middle| \boldsymbol{Z}_{tj}^{\mathrm{S}}, Y_{tj}^{\cdot\mathrm{S}}, \boldsymbol{\alpha}_j^{\mathrm{S}} \right] = -\frac{(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}})(\alpha_j^{\mathrm{ES}} + Z_{tj}^{\mathrm{ES}})(\alpha_j^{\mathrm{CS}} + Z_{tj}^{\mathrm{CS}})(Y_{tj}^{\cdot\mathrm{S}} + \alpha_j^{\cdot\mathrm{S}})}{(\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}})^2(\alpha_j^{\cdot\mathrm{S}} + k_{tj}^{\mathrm{S}} + 1)}. \tag{3.9}$$

A simple method of moments (MoM) estimator corresponds to $\alpha_j^{\mathrm{GS}} = 0$. A comparison of the two estimators appears in Appendix A.1.

We take an empirical Bayes approach to choosing α_j^{GS} and α_j^{GM} for $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$. We use

the totality of lab data to estimate the proportions of severe and mild samples falling into the three pathogen types (E, C, O) in stratum j and then pick $\alpha_j^{\mathrm{ES}}, \alpha_j^{\mathrm{CS}}$, and α_j^{OS} in the same proportions, with the constraint that the prior sample size ($\sum_{j=1}^{J} \alpha_j^{\mathrm{GS}}$) is one. Hence, the priors we obtain are

$$\alpha_j^{\mathrm{GS}} = \sum_{t=1}^{T} Z_{tj}^{\mathrm{GS}} \Bigg/ \sum_{j=1}^{J} \sum_{t=1}^{T} Z_{tj}^{\mathrm{GS}} \qquad \text{and} \qquad \alpha_j^{\mathrm{GM}} = \sum_{t=1}^{T} Z_{tj}^{\mathrm{GM}} \Bigg/ \sum_{j=1}^{J} \sum_{t=1}^{T} Z_{tj}^{\mathrm{GM}}.$$

3.4.2 Distribution of the log relative risks

As likelihood, we take the asymptotic distribution of the log relative risks, with the estimates and variances being based on the posterior means and posterior covariances derived above. To estimate the pathogen-specific relative risk of disease, the parameter we are interested in, we collapse over strata to obtain

$$\widehat{\theta_t^{\mathrm{G}}} = \frac{1}{E^{\mathrm{G}}} \left(\sum_{j=1}^{J} \widetilde{Y_{tj}^{\mathrm{GS}}} + \widetilde{Y_{tj}^{\mathrm{GM}}} \right).$$

We estimate the corresponding variance using the results of equation (3.8) which simplifies, as a result of independence between strata and severity, to

$$\mathrm{Var}\left[\widehat{\theta_t^{\mathrm{G}}}\right] = \left(\frac{1}{E^{\mathrm{G}}}\right)^2 \mathrm{Var}\left[\sum_{j=1}^{J} \widetilde{Y_{tj}^{\mathrm{GS}}} + \widetilde{Y_{tj}^{\mathrm{GM}}}\right] = \left(\frac{1}{E^{\mathrm{G}}}\right)^2 \sum_{j=1}^{J} \mathrm{Var}\left[\widetilde{Y_{tj}^{\mathrm{GS}}}\right] + \mathrm{Var}\left[\widetilde{Y_{tj}^{\mathrm{GM}}}\right].$$

Similarly, we can estimate the covariance between two pathogen-specific relative risk estimates. For example,

$$\mathrm{Cov}\left[\widehat{\theta_t^{\mathrm{E}}}, \widehat{\theta_t^{\mathrm{C}}}\right] = \frac{1}{E^{\mathrm{E}} E^{\mathrm{C}}} \sum_{j=1}^{J} \mathrm{Cov}\left[\widetilde{Y_{tj}^{\mathrm{ES}}}, \widetilde{Y_{tj}^{\mathrm{CS}}}\right] + \mathrm{Cov}\left[\widetilde{Y_{tj}^{\mathrm{EM}}}, \widetilde{Y_{tj}^{\mathrm{CM}}}\right].$$

Again, this is simplified tremendously as a result of modeling disease severity and demographic strata independently.

To model these parameters on the log scale, standard delta method computations provide asymptotic variance and covariances. For week t,

$$\text{Var}\left[\log \widehat{\theta_t^{\text{G}}}\right] = \left(\sum_{j=1}^{J} \text{Var}\left[\widetilde{Y_{tj}^{\text{GS}}}\right] + \text{Var}\left[\widetilde{Y_{tj}^{\text{GM}}}\right]\right)\left(\sum_{j=1}^{J} \widetilde{Y_{tj}^{\text{GS}}} + \widetilde{Y_{tj}^{\text{GM}}}\right)^{-2}.$$

We now index by region r, and let $\boldsymbol{W}_{rt} = \left(\log \widehat{\theta_{rt}^{\text{E}}}, \log \widehat{\theta_{rt}^{\text{C}}}\right)$ be the log pathogen-specific relative risks and $\boldsymbol{\lambda}_{rt} = \left(\text{E}[\log \widehat{\theta_{rt}^{\text{E}}}], \text{E}[\log \widehat{\theta_{rt}^{\text{C}}}]\right)$ be the means in region r and in week t. We model the sampling distribution of the estimators $\boldsymbol{W}_{rt}$ as follows. For region r and week t, we model

$$\boldsymbol{W}_{rt}|\boldsymbol{\lambda}_{rt} \sim N_2\left(\boldsymbol{\lambda}_{rt}, \mathbf{V}_{rt}\right), \tag{3.10}$$

where $\mathbf{V}_{rt}$ is the known covariance matrix obtained from the estimating procedure.

In summary, in the hybrid procedure is an easily implementable approach to inference in which we use an empirical Bayes procedure to obtain smoothed estimates of unobserved disease counts. We then take the asymptotic distribution as the observed data to model the parameters of interest.

3.5 The full probability model

We compare the hybrid method to a MCMC analysis of the full probability model in the small population situation, when this approach is computationally feasible. We find that the hybrid procedure and MCMC procedures produce similar estimates, with comparable coverage properties and mean squared error (MSE). When true disease counts are small, the hybrid procedure tends to produce estimate slightly larger than those obtained via MCMC, however coverage probabilities are close. The most dramatic difference between the two procedures is the time it takes to complete the analysis. The MCMC is orders of magnitude longer, even for relatively small populations. We spend the remainder of this section describing the details.

3.5.1 The likelihood

The observed random variables are the numbers of severe and mild cases by week and stratum and the subsampled cases by severity. Let $\boldsymbol{Y}^{\mathrm{S}}_{tj} = (Y^{\mathrm{ES}}_{tj}, Y^{\mathrm{CS}}_{tj}, Y^{\mathrm{OS}}_{tj})$, $\boldsymbol{Z}^{\mathrm{S}}_{tj} = (Z^{\mathrm{ES}}_{tj}, Z^{\mathrm{CS}}_{tj}, Z^{\mathrm{OS}}_{tj})$, and similarly define $\boldsymbol{Y}^{\mathrm{M}}_{tj}$ and $\boldsymbol{Z}^{\mathrm{M}}_{tj}$ for mild cases. Define $\boldsymbol{Y}_{tj} = (Y^{\mathrm{E}}_{tj}, Y^{\mathrm{C}}_{tj}, Y^{\mathrm{O}}_{tj})$, $\boldsymbol{Z}_{tj} = (\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj})$, $\boldsymbol{k}_{tj} = (k^{\mathrm{S}}_{tj}, k^{\mathrm{M}}_{tj})$, $\boldsymbol{\theta}_t = (\theta^{\mathrm{E}}_t, \theta^{\mathrm{C}}_t, \theta^{\mathrm{O}}_t)$, and $\boldsymbol{r}_{tj} = (r^{\mathrm{E}}_{tj}, r^{\mathrm{C}}_{tj}, r^{\mathrm{O}}_{tj})$. Lastly, denote the total number of cases for a given stratum and week by Y^{+}_{tj}. We are interested in the posterior distributions of $\boldsymbol{Y}^{\mathrm{S}}_{tj}$, $\boldsymbol{Y}_{tj}$, and $\boldsymbol{\theta}_t$.

We have for a single week t and stratum j, the likelihood:

$$
\begin{aligned}
p(\boldsymbol{Y}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Z}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{\cdot\mathrm{M}}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{p}_j, \boldsymbol{r}_{tj}, \boldsymbol{k}_{tj}, N_j) = & \underbrace{\Pr(\boldsymbol{Z}^{\mathrm{S}}_{tj} | \boldsymbol{Y}^{\mathrm{S}}_{tj}, k^{\mathrm{S}}_{tj}) \Pr(\boldsymbol{Z}^{\mathrm{M}}_{tj} | \boldsymbol{Y}^{\mathrm{M}}_{tj}, k^{\mathrm{M}}_{tj})}_{MultHyperGeoms} \\
& \times \quad \underbrace{\Pr(Y^{\cdot\mathrm{S}}_{tj} | \boldsymbol{Y}^{\mathrm{S}}_{tj}) \Pr(Y^{\cdot\mathrm{M}}_{tj} | \boldsymbol{Y}^{\mathrm{M}}_{tj})}_{Deterministic} \\
& \times \quad \underbrace{\Pr(Y^{\mathrm{ES}}_{tj} | Y^{\mathrm{E}}_{tj}, r^{\mathrm{E}}_{tj}) \Pr(Y^{\mathrm{CS}}_{tj} | Y^{\mathrm{C}}_{tj}, r^{\mathrm{C}}_{tj}) \Pr(Y^{\mathrm{OS}}_{tj} | Y^{\mathrm{O}}_{tj}, r^{\mathrm{O}}_{tj})}_{Binomials} \\
& \times \quad \underbrace{\Pr(Y^{\mathrm{E}}_{tj} | \theta^{\mathrm{E}}_t, p^{\mathrm{E}}_j) \Pr(Y^{\mathrm{C}}_{tj} | \theta^{\mathrm{C}}_t, p^{\mathrm{C}}_j) \Pr(Y^{\mathrm{O}}_{tj} | \theta^{\mathrm{O}}_t, p^{\mathrm{O}}_j)}_{Poissons}.
\end{aligned}
$$

As previously specified in Section 3.3, for a single week t and stratum j, we have a likelihood for each component as follows:

1. We approximate the Multinomial distribution with the product of Poisson distributions. For G=E, C, O,

$$Y^{\mathrm{G}}_{tj} | s^{\mathrm{G}}_{tj} \sim \text{Poisson}\left(N_j s^{\mathrm{G}}_{tj}\right).$$

   And as before (see Section 3.3.3), we make the proportionality assumption, $s^{\mathrm{G}}_{tj} = \theta^{\mathrm{G}}_t \times p^{\mathrm{G}}_j$ so that for G=E, C, O,

$$Y^{\mathrm{G}}_{tj} | \theta^{\mathrm{G}}_t \sim \text{Poisson}\left(N_j \theta^{\mathrm{G}}_t p^{\mathrm{G}}_j\right).$$

   Analogous to the hybrid procedure, we use the study data to pre-estimate p^{G}_j and compute expected numbers. The MLE is $\widehat{p}^{\mathrm{G}}_j = Y^{\mathrm{G}}_j / TN_j$ and we define the expected

numbers as $E^{\mathrm{G}} = \sum_{j=1}^{J} N_j \widehat{p}_j^{\mathrm{G}}$ for $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$. We then obtain

$$Y_t^{\mathrm{G}} = \sum_{j=1}^{J} Y_{tj}^{\mathrm{G}} \sim \text{Poisson}(\theta_t^{\mathrm{G}} \times E^{\mathrm{G}}). \tag{3.11}$$

The contribution to the likelihood is, for G=E, C, O

$$p(Y_t^{\mathrm{G}} | \theta_t^{\mathrm{G}}) = \frac{(\theta_t^{\mathrm{G}} E^{\mathrm{G}})^{Y_t^{\mathrm{G}}} e^{-\theta_t^{\mathrm{G}} E^{\mathrm{G}}}}{Y_t^{\mathrm{G}}!}.$$

In practice, we estimate Y_{tj}^{G} by $\widehat{Y_{tj}^{\mathrm{G}}}$ to obtain $\widehat{E^{\mathrm{G}}}$ for substitution into (3.11), using the procedure outlined in the paper. We then treat the $\widehat{E^{\mathrm{G}}}$'s as known and proceed with estimation. In the MCMC, we will use the current estimates of Y_{tj}^{G} and form expected numbers that change at each iteration.

2. Independent $Y_{tj}^{\mathrm{GS}} | Y_{tj}^{\mathrm{G}}, r_{tj}^{\mathrm{G}} \sim \text{Binomial}(Y_{tj}^{\mathrm{G}}, r_{tj}^{\mathrm{G}})$ for G=E, C, O,

$$p(Y_{tj}^{\mathrm{GS}} | Y_{tj}^{\mathrm{G}}, r_{tj}^{\mathrm{G}}) = \binom{Y_{tj}^{\mathrm{G}}}{Y_{tj}^{\mathrm{GS}}} \left(r_{tj}^{\mathrm{G}}\right)^{Y_{tj}^{\mathrm{GS}}} \left(1 - r_{tj}^{\mathrm{G}}\right)^{Y_{tj}^{\mathrm{G}} - Y_{tj}^{\mathrm{GS}}}.$$

3. Independent multivariate hypergeometric distributions for the observed subsamples by disease severity. We let $\boldsymbol{Z}_{tj}^{\mathrm{S}} | k_{tj}^{\mathrm{S}}, \boldsymbol{Y}_{tj}^{\mathrm{S}} \sim \text{MultHyperGeom}(Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}}, k_{tj}^{\mathrm{S}})$ and $\boldsymbol{Z}_{tj}^{\mathrm{M}} | k_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{\mathrm{M}} \sim \text{MultHyperGeom}(Y_{tj}^{\mathrm{EM}}, Y_{tj}^{\mathrm{CM}}, Y_{tj}^{\mathrm{OM}}, k_{tj}^{\mathrm{M}})$, which contribute to the likelihood as

$$p(\boldsymbol{Z}_{tj}^{\mathrm{S}} | k_{tj}^{\mathrm{S}}, \boldsymbol{Y}_{tj}^{\mathrm{S}}) = \binom{Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{ES}}} \binom{Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CS}}} \binom{Y_{tj}^{\mathrm{OS}}}{Z_{tj}^{\mathrm{OS}}} \Big/ \binom{Y_{tj}^{\cdot \mathrm{S}}}{k_{tj}^{\mathrm{S}}},$$

and

$$p(\boldsymbol{Z}_{tj}^{\mathrm{M}} | k_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{\mathrm{M}}) = \binom{Y_{tj}^{\mathrm{EM}}}{Z_{tj}^{\mathrm{EM}}} \binom{Y_{tj}^{\mathrm{CM}}}{Z_{tj}^{\mathrm{CM}}} \binom{Y_{tj}^{\mathrm{OM}}}{Z_{tj}^{\mathrm{OM}}} \Big/ \binom{Y_{tj}^{\cdot \mathrm{M}}}{k_{tj}^{\mathrm{M}}}.$$

3.5.2 Prior distributions

In a full Bayesian analysis, we need priors for the model parameters. For G=E, C, O, and for $t = 1, \ldots, T$,

$$\theta_t^{\mathrm{G}} \sim \mathrm{Gamma}(\alpha^{\mathrm{G}}, \beta^{\mathrm{G}}),$$

for fixed α^{G} and β^{G}. For G=E, C, O, $j = 1, \ldots, J$, and for $t = 1, \ldots, T$,

$$r_{tj}^{\mathrm{G}} \sim \mathrm{Beta}(a^{\mathrm{G}}, b^{\mathrm{G}}),$$

for fixed a^{G} and b^{G}.

3.5.3 Posterior distribution

With a likelihood and prior, we can determine the posterior up to proportionality for the unobserved disease counts and parameters of interest. We manipulate the unnormalized posterior to obtain the form of conditional densities. For a single week t and stratum j, we have the posterior

$$\begin{aligned}
&p(\boldsymbol{Y}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj} | \boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj}, N_j) \\
&\quad \propto p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj} | N_j) \\
&\quad = p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) \\
&\quad = p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj} | Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj} | \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) \\
&\qquad\qquad \times p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) \\
&\quad = p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj} | Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) \\
&\quad = p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj} | Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, Y^{\cdot\mathrm{S}}_{tj} | \boldsymbol{Y}_{tj}, Y^{+}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{Y}_{tj}, Y^{+}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}) \\
&\quad = p(\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj} | Y^{\cdot\mathrm{S}}_{tj}, Y^{+}_{tj}, \boldsymbol{k}_{tj}, \boldsymbol{Y}^{\mathrm{S}}_{tj}, \boldsymbol{Y}_{tj}, N_j) p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, Y^{\cdot\mathrm{S}}_{tj} | \boldsymbol{Y}_{tj}, Y^{+}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) \\
&\qquad\qquad \times p(\boldsymbol{Y}_{tj} | Y^{+}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(Y^{+}_{tj} | \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) p(\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}).
\end{aligned}$$

The joint distribution $p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}|\boldsymbol{Y}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j)$ is

$$\begin{aligned}
p(\boldsymbol{Y}^{\mathrm{S}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}|\boldsymbol{Y}_{tj}, Y^{+}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, N_j) = &\binom{Y^{\mathrm{E}}_{tj}}{Y^{\mathrm{ES}}_{tj}} \left(r^{\mathrm{E}}_{tj}\right)^{Y^{\mathrm{ES}}_{tj}} \left(1-r^{\mathrm{E}}_{tj}\right)^{Y^{\mathrm{E}}_{tj}-Y^{\mathrm{ES}}_{tj}} \\
&\times \binom{Y^{\mathrm{C}}_{tj}}{Y^{\mathrm{CS}}_{tj}} \left(r^{\mathrm{C}}_{tj}\right)^{Y^{\mathrm{CS}}_{tj}} \left(1-r^{\mathrm{C}}_{tj}\right)^{Y^{\mathrm{C}}_{tj}-Y^{\mathrm{CS}}_{tj}} \\
&\times \binom{Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj}}{Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj}} \left(r^{\mathrm{O}}_{tj}\right)^{(Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj})} \\
&\times \left(1-r^{\mathrm{O}}_{tj}\right)^{(Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj})-(Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj})}.
\end{aligned}$$

Therefore the posterior distribution for a single week t and stratum j is

$$\begin{aligned}
&p(\boldsymbol{Y}_{tj}, \boldsymbol{Y}^{\mathrm{S}}{}_{tj}, \boldsymbol{\theta}_t, \boldsymbol{r}_{tj}|\boldsymbol{Z}^{\mathrm{S}}_{tj}, \boldsymbol{Z}^{\mathrm{M}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, Y^{\cdot\mathrm{M}}_{tj}, \boldsymbol{k}_{tj}, N_j) \\
&\quad= \binom{Y^{\mathrm{ES}}_{tj}}{Z^{\mathrm{ES}}_{tj}} \binom{Y^{\mathrm{CS}}_{tj}}{Z^{\mathrm{CS}}_{tj}} \binom{Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj}}{Z^{\mathrm{OS}}_{tj}} \Big/ \binom{Y^{\cdot\mathrm{S}}_{tj}}{k^{\mathrm{S}}_{tj}} \\
&\qquad\times \binom{Y^{\mathrm{E}}_{tj}-Y^{\mathrm{ES}}_{tj}}{Z^{\mathrm{EM}}_{tj}} \binom{Y^{\mathrm{C}}_{tj}-Y^{\mathrm{CS}}_{tj}}{Z^{\mathrm{CM}}_{tj}} \binom{Y^{+}_{tj}-Y^{\cdot\mathrm{S}}_{tj}-(Y^{\mathrm{E}}_{tj}-Y^{\mathrm{ES}}_{tj})-(Y^{\mathrm{C}}_{tj}-Y^{\mathrm{CS}}_{tj})}{Z^{\mathrm{OM}}_{tj}} \Big/ \binom{Y^{\cdot\mathrm{M}}_{tj}}{k^{\mathrm{M}}_{tj}} \\
&\qquad\times \binom{Y^{\mathrm{E}}_{tj}}{Y^{\mathrm{ES}}_{tj}} \left(r^{\mathrm{E}}_{tj}\right)^{Y^{\mathrm{ES}}_{tj}} \left(1-r^{\mathrm{E}}_{tj}\right)^{Y^{\mathrm{E}}_{tj}-Y^{\mathrm{ES}}_{tj}} \binom{Y^{\mathrm{C}}_{tj}}{Y^{\mathrm{CS}}_{tj}} \left(r^{\mathrm{C}}_{tj}\right)^{Y^{\mathrm{CS}}_{tj}} \left(1-r^{\mathrm{C}}_{tj}\right)^{Y^{\mathrm{C}}_{tj}-Y^{\mathrm{CS}}_{tj}} \\
&\qquad\times \binom{Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj}}{Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj}} \left(r^{\mathrm{O}}_{tj}\right)^{(Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj})} \left(1-r^{\mathrm{O}}_{tj}\right)^{(Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj})-(Y^{\cdot\mathrm{S}}_{tj}-Y^{\mathrm{ES}}_{tj}-Y^{\mathrm{CS}}_{tj})} \\
&\qquad\times \frac{\left(\theta^{\mathrm{E}}_t E^{\mathrm{E}}\right)^{Y^{\mathrm{E}}_{tj}} e^{-\theta^{\mathrm{E}}_t E^{\mathrm{E}}}}{Y^{\mathrm{E}}_{tj}!} \times \frac{\left(\theta^{\mathrm{C}}_t E^{\mathrm{C}}\right)^{Y^{\mathrm{C}}_{tj}} e^{-\theta^{\mathrm{C}}_t E^{\mathrm{C}}}}{Y^{\mathrm{C}}_{tj}!} \times \frac{\left(\theta^{\mathrm{O}}_t E^{\mathrm{O}}\right)^{(Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj})} e^{-\theta^{\mathrm{O}}_t E^{\mathrm{O}}}}{(Y^{+}_{tj}-Y^{\mathrm{E}}_{tj}-Y^{\mathrm{C}}_{tj})!} \\
&\qquad\times \pi(\theta^{\mathrm{E}}_t)\pi(\theta^{\mathrm{C}}_t)\pi(\theta^{\mathrm{O}}_t)\pi(r^{\mathrm{E}}_{tj})\pi(r^{\mathrm{C}}_{tj})\pi(r^{\mathrm{O}}_{tj}).
\end{aligned}$$

Conditional posterior distributions, along with their support can be found in Appendix A.2.

3.5.4 Markov Chain Monte Carlo for discrete variables

In this section, we describe the MCMC algorithm to obtain posterior estimates. To normalize the discrete distributions of Y^{GS}_{tj} and Y^{G}_{tj} is computationally expensive for large populations. As an alternative, we implement a Metropolis-Hastings algorithm for sampling discrete vari-

ables, as described by Wakefield et al. (2011). This scheme is based on the use of Markov bases (Diaconis and Sturmfels, 1998). The MCMC proceeds as follows:

1. Set initial values for $Y_{tj}^{\mathrm{ES}\,(0)}$, $Y_{tj}^{\mathrm{CS}\,(0)}$, $Y_{tj}^{\mathrm{E}\,(0)}$, $Y_{tj}^{\mathrm{C}\,(0)}$, and $\theta_t^{\mathrm{G}\,(0)}$, $r_{tj}^{\mathrm{G}\,(0)}$ for G=E, C, O. We define initial values as follows: For G=E, C,

$$Y_{tj}^{\mathrm{GS}\,(0)} = Y_{tj}^{\cdot\mathrm{S}} \times \frac{Z_{tj}^{\mathrm{GS}}}{k_{tj}^{\mathrm{S}}} \qquad \text{and} \qquad Y_{tj}^{\mathrm{G}\,(0)} = Y_{tj}^{\cdot\mathrm{S}} \times \frac{Z_{tj}^{\mathrm{GS}}}{k_{tj}^{\mathrm{S}}} + Y_{tj}^{\cdot\mathrm{M}} \times \frac{Z_{tj}^{\mathrm{GM}}}{k_{tj}^{\mathrm{M}}}.$$

   $Y_{tj}^{\mathrm{OS}\,(0)}$ and $Y_{tj}^{\mathrm{O}\,(0)}$ are deterministically defined. That is, $Y_{tj}^{\mathrm{OS}\,(0)} = Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}\,(0)} - Y_{tj}^{\mathrm{CS}\,(0)}$ and $Y_{tj}^{\mathrm{O}\,(0)} = Y_{tj}^{\cdot\mathrm{S}} + Y_{tj}^{\cdot\mathrm{M}} - Y_{tj}^{\mathrm{E}\,(0)} - Y_{tj}^{\mathrm{C}\,(0)}$. Lastly, for G=E, C, and O, $\log \theta_t^{\mathrm{G}\,(0)} = 0$, $r_{tj}^{\mathrm{G}\,(0)} = 0.1$.

2. Update Y_{tj}^{ES} via the following Metropolis-Hastings procedure. Given the current value $Y_{tj}^{\mathrm{ES}\,(\mathrm{s})}$, we propose a new point, $Y_{tj}^{\mathrm{ES}\,(*)} = Y_{tj}^{\mathrm{ES}\,(\mathrm{s})} + d^{\mathrm{ES}}$, where d^{ES} is drawn uniformly from $\{-D^{\mathrm{ES}}, \ldots, -1, 1, \ldots, D^{\mathrm{ES}}\}$, for a fixed $D^{\mathrm{ES}} > 0$. We confirm that the proposed point is valid with respect to its range. If $Y_{tj}^{\mathrm{ES}\,(*)}$ is a valid, we move to the new point with probability $\min\{q, 1\}$. The acceptance ratio q is defined for updating Y_{tj}^{ES} as:

$$q = \frac{p\left(Y_{tj}^{\mathrm{ES}\,(*)} | \boldsymbol{r}_{tj}^{(\mathrm{s}\text{-}1)}, \boldsymbol{Z}_{tj}^{\mathrm{S}}, \boldsymbol{Z}_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{(\mathrm{s}\text{-}1)}, Y_{tj}^{\mathrm{CS}\,(\mathrm{s}\text{-}1)}, Y_{tj}^{\cdot\mathrm{S}}, Y_{tj}^{+}\right)}{p\left(Y_{tj}^{\mathrm{ES}\,(\mathrm{s})} | \boldsymbol{r}_{tj}^{(\mathrm{s}\text{-}1)}, \boldsymbol{Z}_{tj}^{\mathrm{S}}, \boldsymbol{Z}_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{(\mathrm{s}\text{-}1)}, Y_{tj}^{\mathrm{CS}\,(\mathrm{s}\text{-}1)}, Y_{tj}^{\cdot\mathrm{S}}, Y_{tj}^{+}\right)}.$$

   If $Y_{tj}^{\mathrm{GS}\,(*)}$ is not within the correct range, we remain at the current value.

3. Update Y_{tj}^{CS} using an analogous Metropolis-Hastings procedure as Y_{tj}^{ES}.

4. Update Y_{tj}^{OS}, which is defined as $Y_{tj}^{\mathrm{OS}\,(\mathrm{s})} = Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}\,(\mathrm{s})} - Y_{tj}^{\mathrm{CS}\,(\mathrm{s})}$.

5. Compute E^{G} for G=E, C, O, based on the previous iteration's estimates: $E^{\mathrm{G}\,(\mathrm{s})} = \sum_{j=1}^{J} N_j \widehat{p_j^{\mathrm{G}}}$.

6. Update Y_{tj}^{E} and Y_{tj}^{C} from their respective conditional distributions using the Metropolis-Hastings procedure to that described in step 2.

7. Update Y_t^{O}, defined by $Y_{tj}^{\mathrm{O}\,(\mathrm{s})} = Y_{tj}^{\cdot\mathrm{S}} + Y_{tj}^{\cdot\mathrm{M}} - Y_{tj}^{\mathrm{E}\,(\mathrm{s})} - Y_{tj}^{\mathrm{C}\,(\mathrm{s})}$.

8. Update θ_t^{G} and r_{tj}^{G} for G=E, C, O.

9. Repeat steps 2 through 7 for the desired number of iterations.

The choice of the proposal jump sizes (the D^{GS}'s and D^{G}'s) can dramatically impact how quickly the MCMC chains will converge. Jumps that are too small will result in a high acceptance probability, but will be slow to explore the parameter space. If the proposed jumps are too large, the acceptance of new points will be low, again slowing the time to convergence. We set jump sizes to obtain an acceptance rate around 30% (Roberts et al., 1997).

3.5.5 Comparison study of MCMC to the hybrid procedure

The primary purpose of this analysis is to show that the estimates obtained by the hybrid procedure are close to those obtained via the MCMC described in Section 3.5.4. In each case, the population consists of a single stratum with six months worth of weekly data. We consider three data generating scenarios:

- Scenario 1: small population and higher disease and severity probabilities than observed in the HFMD data.

- Scenario 2: medium population and disease and severity probabilities close to what we observed in the HFMD data.

- Scenario 3: large population and disease and severity probabilities close to what we observed in the HFMD data.

We also assume constant subsampling over time; 45% of severe cases and 10% of mild cases are subsampled for virology. The other parameters used to generate the data for analysis are summarized in Table 3.1. For each scenario, we compare the estimated unobserved

pathogen-specific disease counts, as well as the estimated pathogen-specific log relative risk estimates, obtained via both procedures. We also compare the time each procedure takes to obtain estimates.

Table 3.1: Summary of three data generating scenarios.

		Disease Probabilities			Severity Probabilities		
Scenario	N	p^{E}	p^{C}	p^{O}	r^{E}	r^{C}	r^{O}
Scenario 1	5,000	0.180	0.10440	0.07560	0.3	0.135	0.150
Scenario 2	50,000	0.018	0.01044	0.00756	0.3	0.030	0.015
Scenario 3	500,000	0.018	0.01044	0.00756	0.3	0.030	0.015

We evaluate both the mean square error (MSE), the scaled MSE, and the proportion of credible intervals that cover the true value. As in the previous simulation studies, we scale the MSE because the true unobserved disease counts differ over time. The scaled MSE for Y_{tj}^{ES}, for example, is computed as

$$\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\widetilde{Y_{tj}^{\mathrm{ES}}}-Y_{tj}^{\mathrm{ES}}\right)^2/Y_{tj}^{\mathrm{ES}}. \tag{3.12}$$

Credible intervals for the hybrid estimates of the unobserved disease counts are obtained from the posterior estimates of Y_t^{GS}, as described in Section 3.4.1. Recall, hybrid estimates of the pathogen-specific log relative risk are obtained via the delta method for the posterior estimates of the pathogen-specific relative risk. As point estimates from the MCMC procedure, we have posterior medians, with central 95% credible intervals. We emphasize that the models used in the two analyses are not exactly the same. In particular, the priors are defined differently between the two analyses, with an empirical Bayes prior for the hybrid model.

We summarize the simulation results here, while complete results for these simulations are presented in Appendix A.3. Table 3.2 summarizes the MSE for both the hybrid and MCMC approaches in the three scenarios. For pathogen-specific disease counts, the scaled

MSE of the estimated counts for the two approaches are comparable across the simulation scenarios.

	Scenario 1		Scenario 2		Scenario 3	
Scaled MSE	Hybrid	MCMC	Hybrid	MCMC	Hybrid	MCMC
Y_{tj}^{E}	3.59	3.71	4.75	5.11	4.19	4.42
Y_{tj}^{C}	6.38	6.70	6.79	7.24	5.51	5.71
Y_{tj}^{O}	7.60	7.85	6.52	6.73	5.58	5.72
Y_{tj}^{ES}	0.53	0.54	0.13	0.12	0.06	0.07
Y_{tj}^{CS}	1.19	1.24	1.36	1.29	1.25	1.30
Y_{tj}^{OS}	0.80	0.77	1.80	1.78	1.05	1.09
MSE ($\times 10^3$)	Hybrid	MCMC	Hybrid	MCMC	Hybrid	MCMC
$\log \theta_t^{\mathrm{E}}$	5.44	5.30	10.54	10.56	1.38	1.40
$\log \theta_t^{\mathrm{C}}$	16.85	17.42	14.82	15.80	0.91	0.92
$\log \theta_t^{\mathrm{O}}$	21.58	22.91	25.59	27.12	1.85	1.87

Table 3.2: Scaled MSE for the estimated unobserved pathogen-specific disease counts and MSE for the pathogen-specific log relative risk estimates obtained by MCMC and the hybrid method in three simulation scenarios. Scaled MSE is the average of squared errors divided by the true value, as in (3.12).

When true disease counts are small, the hybrid procedure tends to produce estimate slightly larger than those obtained via MCMC, however coverage over the single simulation are close. The most dramatic difference between the two procedures is the time it takes to complete the analysis (summarized in Table 3.3). The hybrid procedure takes less than a second to obtain reasonable estimates of the pathogen-specific disease counts and log-relative risks while the MCMC takes approximately 40 minutes. The choice of proposal jump sizes will greatly influence how quickly the MCMC chain converges. We summarize the values used for reasonable convergence in the MCMC analyses in Table 3.3.

While the time to complete the MCMC analysis seems reasonable in these scenarios, we should not expect similar results for more complex scenarios, like that for the HFMD analysis. Additional strata will require more unobserved disease count estimation, and we would expect this to increase the time it will take for the MCMC chain to converge. Similarly,

modeling the pathogen-specific log relative risk would require more iterations of the MCMC. Therefore, we expect the hybrid procedure to continue to produce reasonable estimates of the unobserved disease counts and the pathogen-specific log relative risks when the MCMC analysis is no longer feasible. We now turn our attention to modeling the estimates obtained in the hybrid method for the HFMD data.

	Proposal Jump Sizes				MCMC Specifics		Time Comparisons	
Scenario	D^{ES}	D^{CS}	D^{E}	D^{C}	Iterations	Burn-in	Hybrid	MCMC
1	20	15	75	50	250,000	200,000	0.082 seconds	40.2 min
2	7	5	50	60	250,000	200,000	0.100 seconds	38.7 min
3	20	20	175	175	250,000	200,000	0.059 seconds	39.0 min

Table 3.3: Comparison of proposal jump sizes, MCMC iterations, and analysis time for the comparison study of the MCMC to the hybrid procedure.

3.6 Application to HFMD data

3.6.1 The model

We are interested in better understanding the temporal dynamics of EV71 and CA16, as well as the importance of area level characteristics on disease. Following the setup in (3.10), we model the means as

$$\lambda_{rt}^{\text{G}} | \, \boldsymbol{\beta}^{\text{G}}, \boldsymbol{\delta}^{\text{G}}, \boldsymbol{\gamma}^{\text{G}} = \beta_{0r}^{\text{G}} + \boldsymbol{\gamma}^{\text{G}} \boldsymbol{x}_{rt} + g_r(t, \boldsymbol{\delta}^{\text{G}}) + \sum_{k=1}^{K} f_k(z_k, \boldsymbol{\beta}_k^{\text{G}}), \tag{3.13}$$

where β_{0r}^{G} is a region- and pathogen-specific intercept, $\boldsymbol{x}_{rt}$ are covariates we model parametrically, g_r is a smooth function of time in region r, and f_k are smooth functions of (meteorological) covariates z_k not including time for $k = 1, \ldots, K$ covariates. Hence, we have chosen to have a common regression model for the meteorological variables but different temporal smoothers for each region. In the interest of parsimony we would like a simple model, but fits from an initial model with a single time smoother for all regions was

inadequate (when residuals were examined). As we will see the temporal pattern is complex and it is not surprising that the confounder by time model needs to be region-specific.

Smoothers were fit to each of four area-level meteorological covariates, temperature, relative humidity, wind speed, and log precipitation at a one week lag, so that $K = 4$. Average weekly meteorological data was obtained from averaging daily weather data from multiple weather stations within each region. From an epidemiological standpoint, there is good reason to include lagged covariates in our models; the estimated incubation period of HFMD is between 3 and 7 days (Huang et al., 2013; Wu et al., 2014; Chang et al., 2012). Hence, if there is an association, it is likely that the meteorological conditions one week prior would be related to the number of new cases in a given week.

Temperature, relative humidity, wind speed, and log precipitation were modeled using cubic regression splines with 3 degrees of freedom (and a B-spline representation), and with knots at the tertiles of the observed values for each region. Splines have previously been used to smooth meteorological covariates in models of total cases of HFMD (Wu et al., 2014; Onozuka and Hashizume, 2011). In this exploratory analysis, the number of knots was chosen as a combination of what has been previously used in the literature and empirical investigation.

We also included an indicator for the time when school is in session versus closed (Jan 15-Feb 15 and July 1-Aug 31). We would expect to see the risk of disease decrease when schools are closed; we expect the transmission rates to decline when school is on break since there are fewer contacts. While the actual periods of school closures differed across prefectures and regions, we had no further information and therefore simply modeled school closing with the indicator. This follows the approach used in Wang et al. (2011).

The effects we are interested in are on a short time scale. That is, we are interested in the short-term effects of meteorological conditions on the number of HFMD cases in a given week. To this end, we want to adjust for the large-scale or long-term temporal effects. We have strong confounding by time and so we will adjust for long-term seasonality trends with, as discussed above, region-specific effects.

Figures 3.5(a) and 3.5(b) show the empirical estimates of the pathogen-specific log relative risks by lagged temperature, relative humidity, wind speed, and precipitation, by region. A lowess smoothed line shows the general trend. The two pictures are very similar, supporting the decision to use the same smoothing model for both EV71 and CA16 trends with different parameters. Furthermore, most trends are not too non-linear, which supports the use of a relatively small number of knots (three) in the smoothers.

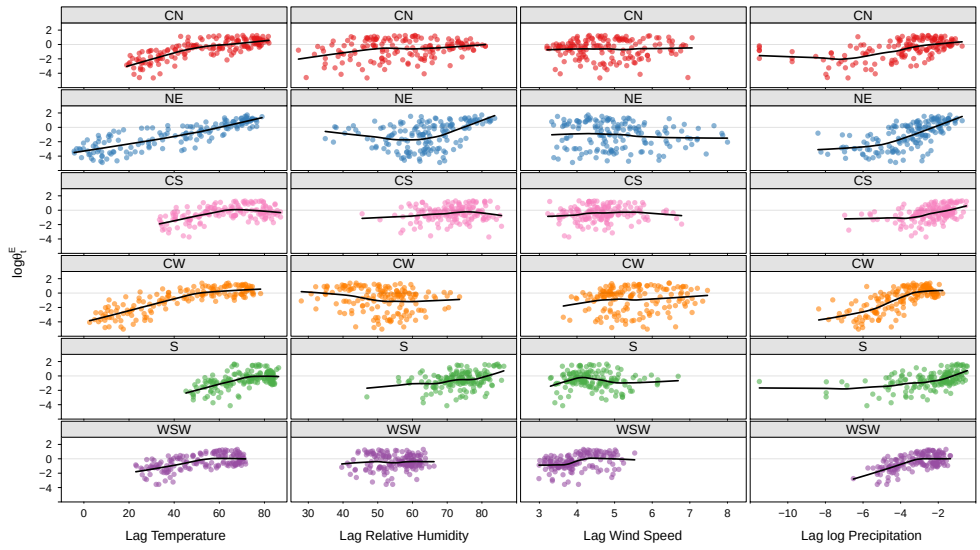
Figure 3.6 shows the pairwise relationships between pathogen-specific empirical estimates of log relative risk and the meteorological covariates of interest. Again, the potential for confounding by time is evident.

There has been significant work into understanding how to appropriately adjust for temporal trends in the air pollution literature (Peng et al., 2006). The risk of biased results can be high if care is not paid to these subtleties, especially when the magnitude of the effects of interest is small (Dominici et al., 2004). The amount of smoothing in the model can influence both the bias and standard error of the estimates associated with the meteorological covariates.

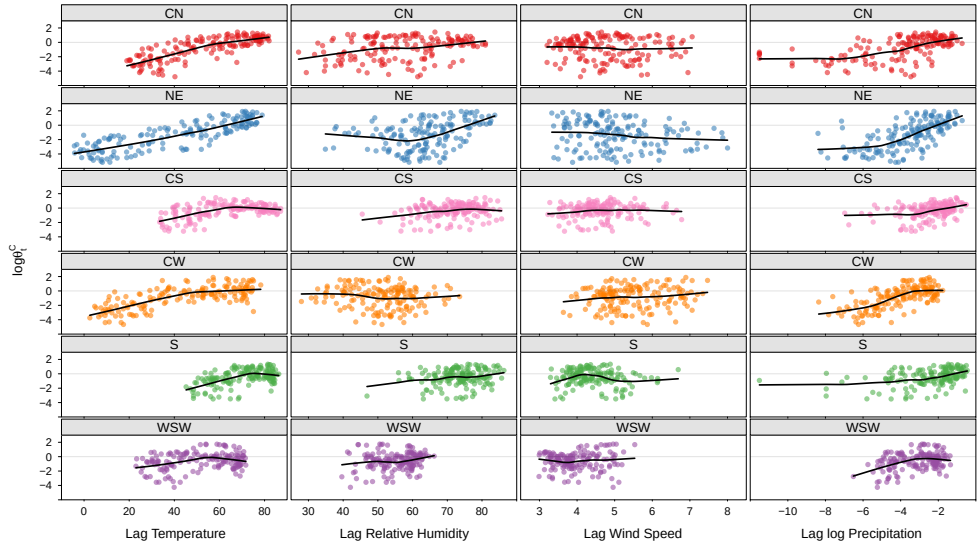
We take a Bayesian approach to inference and model time as a second-order random walk (RW2), i.e. $g_r(t, \delta_t^{\mathrm{G}}) = \delta_t^{\mathrm{G}}$, where δ_t^{G} is specified below. We place independent priors on the remaining coefficients. The model for region r and week t is

$$
\begin{aligned}
\boldsymbol{W}_{rt} | \boldsymbol{\lambda}_{rt} &\sim N_2\left(\boldsymbol{\lambda}_{rt}, \mathbf{V}_{rt}\right), \\
\lambda_{rt}^{\mathrm{G}} &= \beta_{0r}^{\mathrm{G}} + \beta_{1r}^{\mathrm{G}} t + \boldsymbol{\gamma}^{\mathrm{G}} \boldsymbol{x}_{rt} + \sum_{k=1}^{K} \mathbf{B}(\boldsymbol{z}_{rtk}) \boldsymbol{\beta}_k^{\mathrm{G}} + \delta_t^{\mathrm{G}}, \\
\delta_t^{\mathrm{G}} - 2\delta_{t-1}^{\mathrm{G}} + \delta_{t-2}^{\mathrm{G}} | \tau_{\mathrm{G}} &\sim N(0, \tau_{\mathrm{G}}^{-1}), \\
\tau_{\mathrm{G}} &\sim \mathrm{Gamma}(a, b), \quad a = 100, b = 0.02, \\
\boldsymbol{\gamma}^{\mathrm{G}} &\sim N(\mathbf{0}, \sigma_{\gamma^{\mathrm{G}}}^2 I), \qquad \sigma_\gamma^2 = 100, \\
\boldsymbol{\beta}^{\mathrm{G}} &\sim N(\mathbf{0}, \sigma_\beta^2 I), \qquad \sigma_\beta^2 = 100,
\end{aligned}
\tag{3.14}
$$

where $\boldsymbol{x}_{rt}$ are covariates modeled parametrically, in our case an indicator for school closure,



(a) EV71



(b) CA16

Figure 3.5: Estimated log EV71- (left) and CA16- (right) specific relative risk by region by lagged temperature, relative humidity, wind speed, and precipitation. A lowess smoothed line is included.

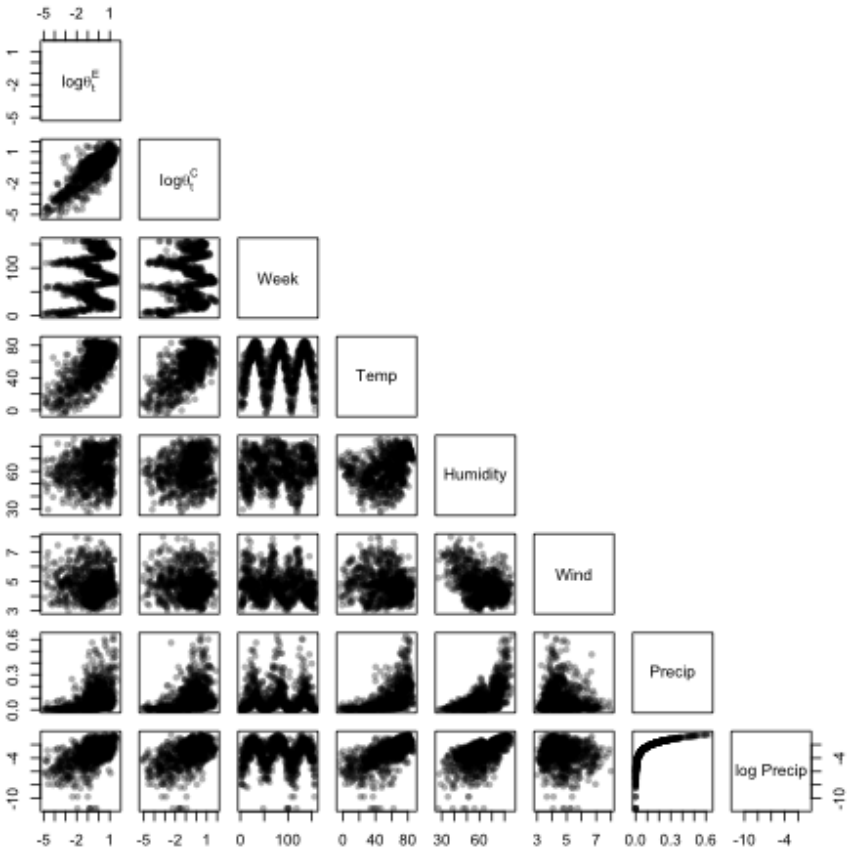


Figure 3.6: Pairwise relationships between pathogen-specific log relative risks and lagged meteorological variables for the six regions in China.

and $\mathbf{B}(z)$ is the vector of B-spline basis functions evaluated at z. In the next section, we discuss prior selection.

We fit the model using the integrated nested Laplace approximation (INLA) approach (see Section 2.3.3), to quickly fit such hierarchical Bayesian models. Fitting the above model to data across all of China and over 3 years takes approximately 5.5 minutes on a 2.6GHz Intel Core i5 and 16GB memory computer using OSX 10.9.5.

3.6.2 Prior selection for temporal smoothers

RW2 models as linear smoothers

The second order random walk (RW2) in time has improper "density"

$$\pi(\boldsymbol{\delta}^{\mathrm{G}}|\tau_{\mathrm{G}}) \propto \tau_{\mathrm{G}}^{(T-2)/2} \exp\left\{-\frac{\tau_{\mathrm{G}}}{2}(\boldsymbol{\delta}^{\mathrm{G}})^{\mathrm{T}}\boldsymbol{R}\boldsymbol{\delta}^{\mathrm{G}}\right\} \tag{3.15}$$

where the "precision" matrix is $\tau_{\mathrm{G}}\boldsymbol{R}$, where $\boldsymbol{R}$ is the $T \times T$ structure matrix, defined as

$$\boldsymbol{R} = \begin{bmatrix} -1 & 2 & -1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & -1 & 2 & -1 \end{bmatrix}.$$

Notice that $\boldsymbol{R}$ has rank $T-2$, and therefore (3.15) is improper. To adjust for temporal smoothing in the HFMD model (3.14), we place a gamma prior on the precision, $\pi(\tau_{\mathrm{G}}|a,b) \sim \text{Gamma}(a,b)$, where $\mathrm{E}[\tau_{\mathrm{G}}] = a/b$ and $\mathrm{Var}[\tau_{\mathrm{G}}] = a/b^2$.

We know that small precision corresponds to little temporal smoothing since the random effects $\boldsymbol{\delta}^{\mathrm{G}}$ are allowed to vary relatively freely. In contrast, large precisions yield smoother curves, since the random effects are strongly encouraged to be similar. But translating these observations on the smoothing into the parameters of a gamma distribution is not easy. A more intuitive approach is to consider the number degrees of freedom per year. For example, 6 degrees of freedom roughly corresponds to extracting structure beyond two months in time, so that the covariates associations are being estimated on time scales of 2 months or less.

When the precision τ_{G} is known, we have the following model for the data $\boldsymbol{W}^{\mathrm{G}}$

$$\begin{aligned}\boldsymbol{W}^{\mathrm{G}}|\boldsymbol{\delta}^{\mathrm{G}} &\sim N(\boldsymbol{\delta}^{\mathrm{G}}, \boldsymbol{V}^{\mathrm{G}})\\ \boldsymbol{\delta}^{\mathrm{G}}|\tau_{\mathrm{G}} &\sim N(\mathbf{0}, \tau_{\mathrm{G}}^{-1}\boldsymbol{R}^{-1})\end{aligned}$$

where $\boldsymbol{V}^{\mathrm{G}}$ is the known vector of measurement errors for $\boldsymbol{W}^{\mathrm{G}}$, as described in the main paper. This yields a posterior of

$$\boldsymbol{\delta}^{\mathrm{G}}|\boldsymbol{W}^{\mathrm{G}} \sim N\big((\boldsymbol{\Lambda}^{\mathrm{G}} + \tau_{\mathrm{G}}\boldsymbol{R})^{-1}\boldsymbol{\Lambda}^{\mathrm{G}}\boldsymbol{W}^{\mathrm{G}},\, (\boldsymbol{\Lambda}^{\mathrm{G}} + \tau_{\mathrm{G}}\boldsymbol{R})^{-1}\big) \tag{3.16}$$

where $\boldsymbol{\Lambda}^{\mathrm{G}} = (\boldsymbol{V}^{\mathrm{G}})^{-1}$. Thus we obtain posterior mean estimates (fitted values) of

$$\widehat{\boldsymbol{W}^{\mathrm{G}}} = \mathrm{E}[\boldsymbol{\delta}^{\mathrm{G}}|\boldsymbol{W}^{\mathrm{G}}] = (\boldsymbol{\Lambda}^{\mathrm{G}} + \tau_{\mathrm{G}}\boldsymbol{R})^{-1}\,\boldsymbol{\Lambda}^{\mathrm{G}}\boldsymbol{W}^{\mathrm{G}} = \boldsymbol{S}^{\mathrm{G}}\boldsymbol{W}^{\mathrm{G}}, \tag{3.17}$$

where $\boldsymbol{S}^{\mathrm{G}}$ is the linear smoother matrix. Note that we could write $\boldsymbol{S}^{\mathrm{G}} = \boldsymbol{S}^{\mathrm{G}}(\tau_{\mathrm{G}})$ to emphasize the role of τ_{G}. And the estimated degrees of freedom for this model is found by

$$\mathrm{df} = \mathrm{tr}\left((\boldsymbol{\Lambda}^{\mathrm{G}} + \tau_{\mathrm{G}}\boldsymbol{R})^{-1}\,\boldsymbol{\Lambda}^{\mathrm{G}}\right) = \mathrm{tr}(\boldsymbol{S}^{\mathrm{G}}),$$

i.e., the trace of the hat matrix, see for example Ruppert et al. (2003, Sec 3.13).

This gives us a straightforward way to estimate the degrees of freedom associated with a fixed value for the precision. To implement this approach to prior selection, we start

with a desired range for the degrees of freedom and then pick the values a and b of the gamma distribution so that the distribution on τ_{G} has the required range of degrees of freedom. Since gamma distributions are defined by two parameters, choosing one of the parameters will determine the second for a given mean value (where we center the distribution about a particular number of degrees of freedom). The coefficient of variation for a gamma distribution is $1/\sqrt{a}$ and so we can choose a values that are commensurate with the amount of relative variability. Given the mean, the choice of a determines b (since we know the required mean), and thus the prior distribution. In the next section we pick values for our HFMD study.

Choice of HFMD priors

The choice of degrees of freedom per year has been discussed extensively in the air pollution literature. For example, Dominici et al. (2003) use smooth functions of time to adjust for large-scale fluctuations in mortality over time. They choose a smoothing parameter "equal to seven degrees of freedom per year of data so that little information from time scales longer than approximately two months in included when estimating β." Since we are also interested in removing long term trends, we consider between 4 and 8 degrees of freedom per year to be an appropriate amount of smoothing. This range corresponds to, effectively, allowing 3 to 1.5 months of data to be used to estimate the meteorological associations.

Figure 3.7 plots the estimated number of degrees of freedom for each region- and pathogen-specific random walk. We see that for τ_{G} between 5,000 and 10,000, all pathogen- and region-specific models have degrees of freedom within the acceptable range of 4 to 8 df/year. We choose a mean value for τ_{G} of 5,000, which corresponds to the maximum amount of wiggliness in the temporal smoothers, while maintaining an acceptable amount of smoothing. We fix a coefficient of variation ($= 1/\sqrt{a}$) at 10%, which gives $a = 100$. Thus, with $\tau_{\text{G}} = 5,000$, and $a = 100$, we use a Gamma(100, 0.02) prior for the temporal smoothing in our main analysis.

While on average, this prior should provide the appropriate amount of smoothing, we can examine the distribution of degrees of freedom. Figure 3.8 shows histograms for each

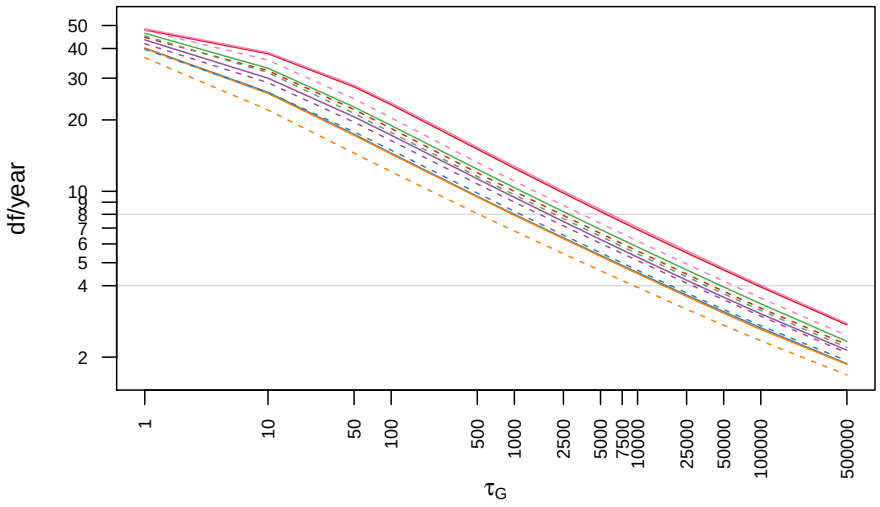


Figure 3.7: Estimated degrees of freedom for various values of τ_G, by region- and pathogen-specific random walks of order 2. Solid lines are for EV71 and dashed for CA16. Colors correspond to regions as defined on the map (Figure 3.1). The two grey horizontal lines show the range of degrees of freedom we would like our priors to concentrate on.

pathogen- and region-specific smoother where the degree of freedom was estimated using random draws from the specified prior. These plots show a reasonable spread, and these are the priors we use in our main analysis. In the next section, we examine the sensitivity of our analysis to this prior selection.

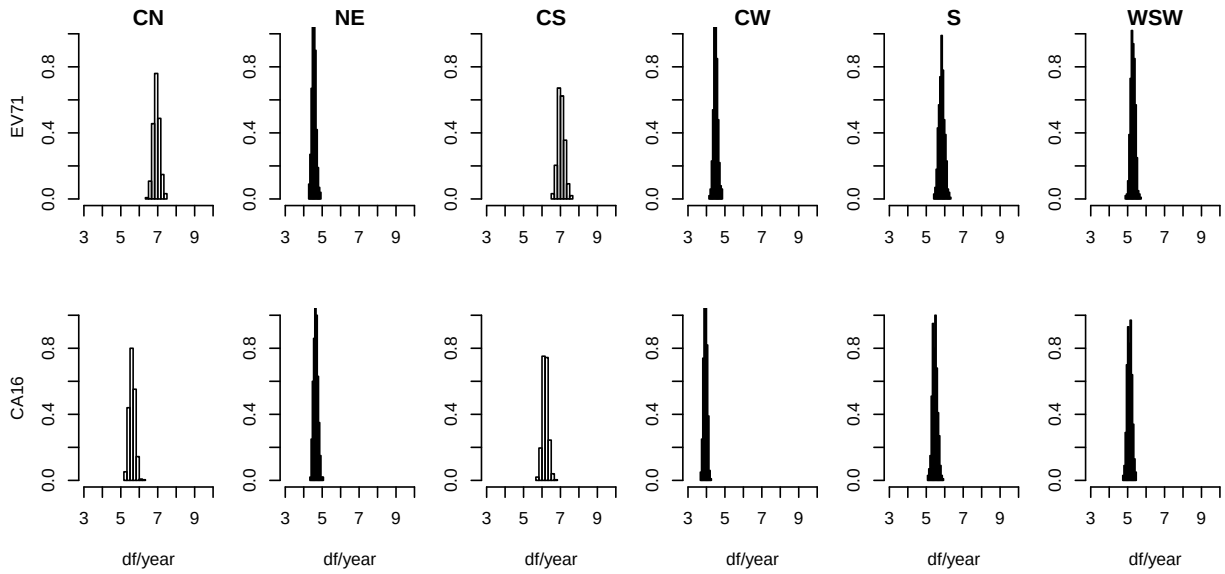


Figure 3.8: Histogram of estimated degrees of freedom for τ_{G} values from Gamma(100, 0.02) distribution for regions in China and by pathogen.

3.6.3 Simulations

We conducted simulations studies to investigate three aspects of the proposed modeling approach. We summarize the results here, and present complete results in Appendix A.5. The first study examined the pre-processing procedure used to obtain unobserved disease counts. We consider two subsampling scenarios for this study: when the proportion of cases subsampled is constant across time and stratum, and when the proportion of subsampled cases varies across time and stratum. We compare our results to those that would be obtained

in the complete case scenario (i.e. when 100% of cases are subsampled for virology). When the amount of subsampling is constant across strata and time, we obtain good estimates for the two primary pathogens of interest. The corresponding confidence intervals have nearly 95% coverage, except when the true number of cases is very small (fewer than 5 cases). When the amount of subsampling varies, the resulting estimates perform somewhat worse than those obtained from constant subsampling. In particular, when there are no samples for a given stratum or week, the resulting estimates have poor coverage. Therefore, this procedure would not be ideal for estimating exact numbers in these cases. Nevertheless, as subsequent simulation demonstrates, the pre-processing procedure allows us to obtain reasonable estimates of the pathogen-specific log relative risks. The details of this simulation study are presented in Appendix A.5.2.

The second simulation study investigated the performance of the pathogen-specific log relative risks estimates, and their sensitivity to the modeling assumptions. In particular, we were interested in the impact of the proportionality assumption on the performance of the estimated pathogen-specific log relative risk. We compare both constant and variable subsampling schemes to the complete case scenario. When proportionality holds, we tend to estimate the true pathogen-specific log relative risk for the two primary pathogens of interest well, and corresponding standard error estimates yield intervals with nearly 95% coverage. When subsampling is variable, these estimates are slightly worse, but coverage is still very good. When the proportionality assumption is not valid, the proposed method yields good estimates of the weighted average pathogen-specific log relative risk. The details of this simulation study are presented in Appendix A.5.3.

The last simulation study investigated the impact of modeling choices in the ability to estimate underlying temporal and covariate trends. In the sensitivity analyzes for the primary analysis, we saw that that our results were relatively stable over a range of reasonable priors (See Section A.6 for further details). Therefore we do not consider the effect of prior selection in these simulations. We consider the impact of modeling the two pathogens jointly. We found that accounting for the correlation induced by subsampling between the

two pathogens by modeling them jointly yielded better estimates of the true effects, as well as more consistent estimation of the underlying temporal effects. When consider estimating the covariate effect, both methods produce slightly biased estimates of the true effect. However, the estimates obtained when modeling the two pathogens together yield covariate effect estimates with a smaller MSE. The details of this simulation study are presented in Appendix A.5.4.

Overall, simulations studies show that the methods proposed in this paper result in reasonable estimates of the unobserved disease counts, the underlying pathogen-specific log relative risk, and estimated effects of covariates on the risk. We see that even when the proportionality assumption fails, we still obtain good estimates of relevant parameters.

3.6.4 Modeling results

We fit the model developed in Section 3.6.1 to three years of HFMD data for the six geographic regions. In the appendix, Figures A.30-A.32 show the effects of changing the amount of temporal smoothing on the estimated associations of interest. There is reasonable stability of the associations in the meteorological variables over a range of levels of temporal smoothing as dictated though our prior on τ_{G}. Here we present results for the model where $\tau_{\text{G}} \sim \text{Gamma}(100, 0.02)$.

Figure 3.9 shows the estimated pathogen-specific smoothers for both the common meteorological covariates and the region- and pathogen-specific temporal smoothers. The posterior mean curve is the solid line and approximate 95% pointwise credibility intervals are also included as shaded regions. We see a positive effect of temperature on both EV71- and CA16-specific HFMD log relative risk, although to varying degrees. For both pathogens, we see an increasing trend below 45°F, and a flattening above that. CA16 seems to also show an increasing trend above 70°F, though the data are more sparse for these temperatures. Similar trends have been found in Wu et al. (2014), although for all HFMD and not pathogen specific cases. EV71- and CA16-specific log relative risks increase with increasing relative humidity, though the association is steeper for EV71. A similar relationship was found for

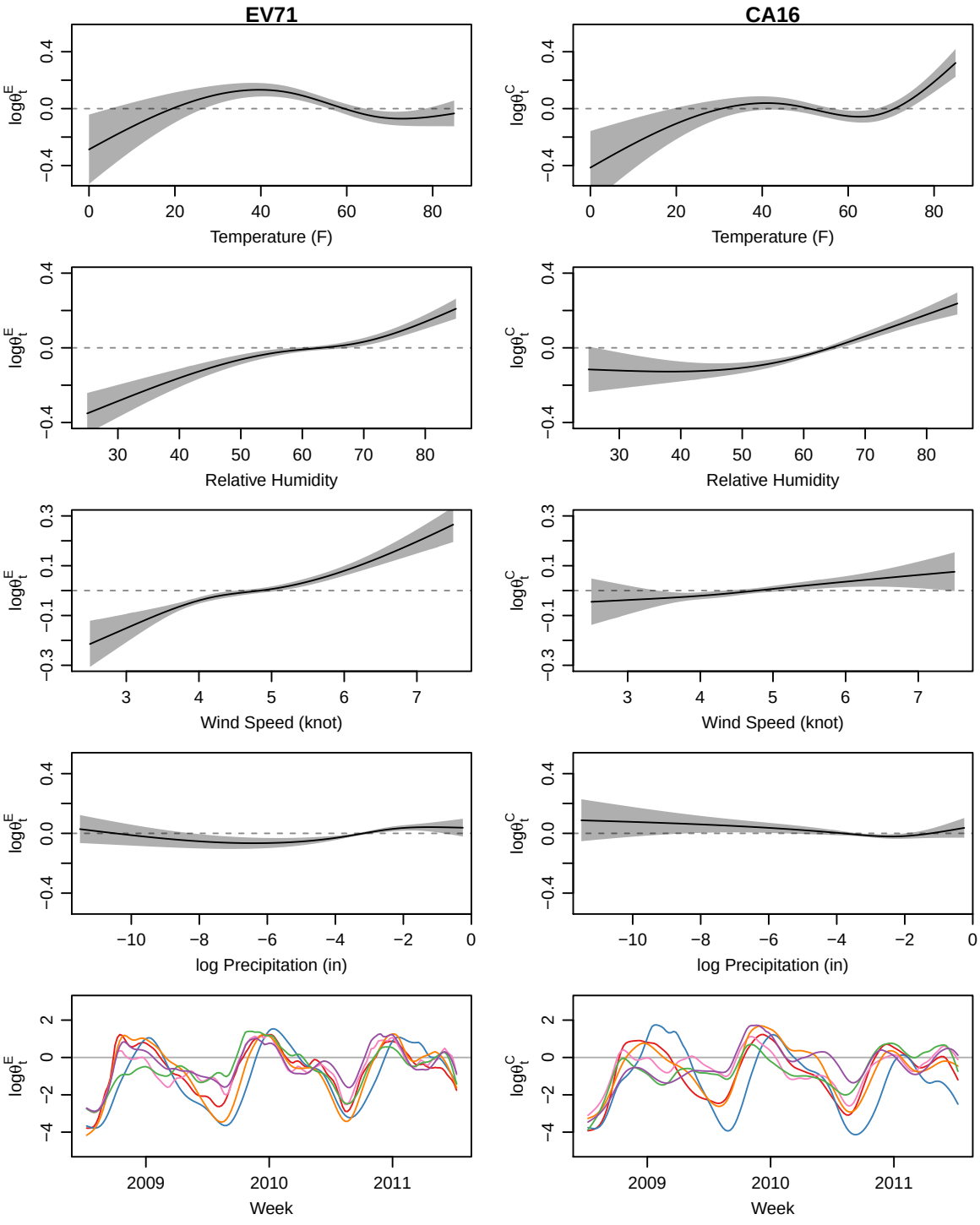


Figure 3.9: Estimated smoothers of meteorological variables for log relative risk of EV71 (left) and CA16 (right). The solid line is the posterior median and the shaded regions are 95% pointwise credible intervals. For temporal soothers in the bottom panels, the color corresponds to the region. This figure appears in color in the electronic version of this article.

all HFMD cases in Wang et al. (2011).

For both EV71- and CA16-specific log relative risk, there seems to be little change associated with small amounts of average log precipitation. There is evidence of an increase in EV71-specific log relative risk for average log rainfall, with a flattening out above 0.1 inches of lagged average precipitation (or -2.3 for log precipitation). In contrast, the effect of precipitation on the CA16-specific log relative risk seems to decline slightly, and then increase for lagged weekly precipitation above 0.1 inches. We estimate a 6.3% higher risk of EV71-specific HFMD when school is in session (95% CI: 2.4-10.4% higher). Likewise, when school is in session, we estimate a 14% higher relative risk of CA16-specific HFMD cases (95% CI: 9.6-19). Similar school effects have been reported in other studies (Wang et al., 2011). Figure 3.10 shows the observed and fitted SIRs in each of the six regions and for each of EV71 and CA16. We see that the fit is reasonable in each case.

3.7 Discussion

While there have been a number of studies examining the relationship between HFMD and meteorological covariates in a variety of areas, very few have looked at these trends for specific pathogens. Moreover, most pathogen-specific studies have focused on EV71 (not CA16), due to its association with severe disease. Nevertheless, our results are consistent with the current understanding in the literature. In general, we see that warmer temperatures (up to a point) and elevated humidity are associated with an increased rate of HFMD infections, as well as EV71-specific HFMD, consistent with other studies (Chang et al., 2012; Ma et al., 2010; Onozuka and Hashizume, 2011; Wang et al., 2011; Wu et al., 2014).

For the purposes of this paper, we only considered those meteorological covariates that had been previously shown to have a strong relationship with the number of overall HFMD cases. The approach could be used to investigate the relationship between pathogen-specific HFMD and other meteorological covariates. A limitation of the method and the data analysis is that we do not account for co-infections. While some have found that approximately 10% of HFMD cases showed signs of EV71 and CA16 co-infections (Liu et al., 2011), co-infection

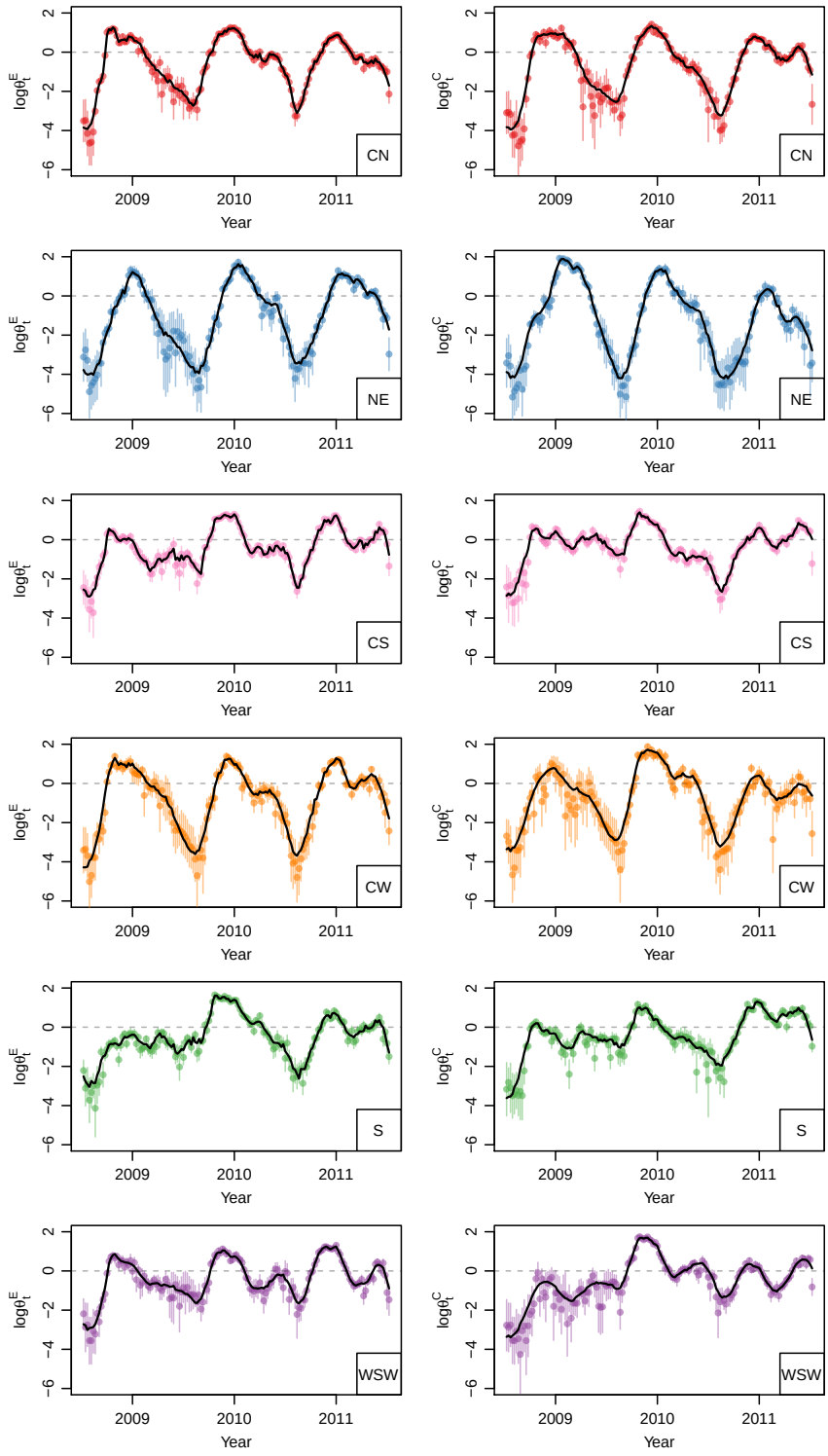


Figure 3.10: Fitted values for log relative risks of EV71 (left) and CA16 (right). The dots are the estimated log relative risks, and the vertical lines are the 95% intervals based on the known standard errors. The color corresponds to the region in China. This figure appears in color in the electronic version of this article.

data was not available to us.

We have proposed a fast and simple procedure that allows us to easily impute unobserved pathogen-specific disease counts with corresponding standard errors. Once pathogen-specific disease counts are estimated, we can use these to learn about area level covariate effects using standard statistical methods which provide results commensurate with those previously published. Simulation studies show that this procedure yields good estimates and that the subsequent modeling approach was fairly robust to modeling assumptions.

The key to our approach was a first step in which we constructed relevant summary statistics to model. We believe this general approach is applicable in many situations with complex sampling schemes, such as in disease surveillance systems. The asymptotic distribution of the summary statistics is then used with hierarchical smoothing to help with small sample problems, and semi-parametric regression reveals non-linear associations.

Chapter 4

NEGATIVE CONTROLS

4.1 Introduction

In this chapter, we are interested in evaluating the effectiveness of the school-located influenza vaccine (SLIV) program in reducing county-wide risk of influenza and influenza-like illness (ILI) using emergency department (ED) data from AHCA. The quarterly ED data consists of patient zip code, basic demographics, and ICD-9 codes. Details about this study and the data were introduced in Section 1.1.2.

While passively collected disease surveillance data, such as AHCA, are inexpensive and easily collected, they also carry a high risk for biased (or erroneous) inference, depending on the scientific question. If, for example, we were interested in the incidence of broken legs across counties in Florida, we may be less concerned about the risk of bias using the AHCA data, since most people will seek emergency care for such a serious injury. In our setting, we are interested in estimating the effect of the SLIV program on county-wide ILI cases. Hence, our target of inference is the relative risk of ILI for Alachua county versus all other non-Alachua counties of Florida over a single influenza season. Since we are interested in county-level ILI risk, we are concerned with potential bias due to differences in ascertainment of ILI-associated ED visits between Alachua county and other counties in Florida. Unlike a broken leg, most individuals with ILI will not seek medical care at all, and among those that do, even fewer will seek care at an ED. Severity of illness, insurance status, proximity, and a variety of socio-economic factors will influence the population that arrives at an emergency department for ILI. For unbiased estimates of population-level incidence, we would need to adjust for these factors.

However, surveillance data rarely includes patient information beyond basic demographics

and key disease indicators. As a result, we are concerned that patients seeking healthcare for ILI at an ED are not a representative sample of the county-wide ILI population. With such data limitations, one option is to simply limit the scope of conclusions to patients seeking health care at an ED.

As an alternative, we examine the use of negative controls in the disease surveillance setting. We acknowledge that surveillance data provides little information about the healthcare seeking behavior of sick individuals and instead consider the healthcare seeking behavior to be a significant source of unmeasured confounding when making inference about disease risk. For simplicity, we assume that healthcare seeking behavior of sick individuals is the only source of unmeasured confounding. In practice, there are many sources of confounding, some of which we can control for in the model (such as sex), but we also accept that there is likely remaining confounding. In controlled experiments, we may randomize to reduce the impact of such confounding, but in observational studies, we must accept that we have controlled for the largest sources of confounding and hope that any remaining unmeasured confounding is minimal.

Lab scientists frequently use a negative control to check that the experiment, in the absence of the exposure, proceeded as expected. A negative control outcome is subject to the same kinds of unmeasured confounding as the outcome of interest, but is *not* in the causal pathway of interest and not affected by the exposure of interest or the treatment. Let X be the exposure of interest, Y the outcome of interest, U be unobserved confounding, and Z be the negative control outcome. Figure 4.1 depicts the causal diagram of unobserved confounding with a negative control outcome.

Lipsitch et al. (2010) proposed using negative controls as a way to detect and adjust for the biases that can invalidate causal inference in epidemiological studies. Subsequently, the negative control framework has been applied to a variety of settings (Richardson et al., 2014; Flanders et al., 2011). Richardson et al. (2015) formalized the use of negative controls outcomes to adjust SMR estimates, as well as the conditions under which the adjustment will reduce bias in the estimates.

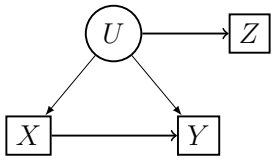


Figure 4.1: Diagram depicting unmeasured confounding. X is the exposure of interest, Y the outcome of interest, U unobserved confounding, and Z the negative control outcome. Squares indicate observed variables; circles are unobserved. Arrows denote causal associations.

4.2 Previous approach to negative controls

We describe the negative control approach in the context of the Florida influenza study. We assume a single treatment region, and note that this approach can extend to model multiple treatment areas. For $n+1$ independent study counties, let $i = 1, \ldots, n$ denote control regions, and $i = n + 1$ the treatment region. Let Y_i be the number of cases in county i who sought care at an ED for ILI during the influenza season. Let N_i be the total population in county i, $i = 1, \ldots, n + 1$. We assume the true data generating model for seeking care for ILI at an ED during a single influenza season (a rare event) is

$$Y_i \,|\, \theta_i \sim \text{Poisson}\left(N_i r_i \theta_i\right), \tag{4.1}$$

where $r_i = \Pr(\text{ED} \,|\, \text{ILI}, i)$ is the probability of going to the ED given ILI in area i, and $\theta_i = \Pr(\text{ILI} \,|\, i)$ is the probability of being sick with ILI in area i. Hence, $r_i\theta_i = \Pr(\text{ED and ILI} \,|\, i)$ is the probability of seeking healthcare for ILI at an ED given the patient is a resident of county i, $i = 1, \ldots, n + 1$.

Writing the probability as the product of conditional probabilities allows us to distinguish the disease process (θ_i) from the health care seeking process (r_i). While the data we collect is influenced by both the risk of disease and the healthcare seeking behavior of sick individuals, the ILI data alone has little information about the healthcare seeking behavior of sick individuals. As a result, in this kind of disease surveillance data, the healthcare seeking behavior (r_i) confounds inference about the risk of disease (θ_i).

In our setting, we are interested in estimating the effects of the SLIV program in a single county. If we are unaware of unmeasured confounding, a naïve approach would be to simply fit the log linear model

$$Y_i \,|\, \mu, \alpha \sim \text{Poisson}\left(N_i e^{\mu+\alpha x_i}\right), \tag{4.2}$$

where μ is the average log risk of ILI in the non-intervention counties; x_i is an indicator variable equal to one in the treatment area and zero otherwise; and $\exp\{\alpha\}$ is the relative risk associated with the SLIV program, with $\exp\{\alpha\} < 1$ corresponding to a reduction in risk in the intervention county. The MLE for the estimated relative risk associated with the intervention from (4.2) is

$$e^{\hat{\alpha}} = \frac{Y_{n+1}/N_{n+1}}{(\sum_{i=1}^{n} Y_i)/(\sum_{i=1}^{n} N_i)}, \tag{4.3}$$

which we call the unadjusted estimate. The expectation of the unadjusted estimate, under the true data generating model in (4.1) is

$$\begin{aligned}
\mathrm{E}\left[e^{\hat{\alpha}}\right] =& \mathrm{E}\left[\frac{Y_{n+1}/N_{n+1}}{(\sum_{i=1}^{n} Y_i)/(\sum_{i=1}^{n} N_i)}\right] \\
=& \left(\frac{\sum_{i=1}^{n} N_i}{N_{n+1}}\right) \mathrm{E}\left[\frac{Y_{n+1}}{(\sum_{i=1}^{n} Y_i)}\right] \\
\approx& \left(\frac{\sum_{i=1}^{n} N_i}{N_{n+1}}\right) \frac{\mathrm{E}\left[Y_{n+1}\right]}{(\sum_{i=1}^{n} \mathrm{E}\left[Y_i\right])} \qquad \text{by Slutsky's Theorem} \\
=& \left(\frac{\sum_{i=1}^{n} N_i}{N_{n+1}}\right) \frac{N_{n+1} r_{n+1} e^{\mu+\alpha}}{(\sum_{i=1}^{n} N_i r_i e^{\mu}))} = e^{\alpha} \left(\frac{r_{n+1}}{\bar{r}}\right),
\end{aligned}$$

where $\bar{r} = \sum_{i=1}^{n} N_i r_i / \sum_{i=1}^{n} N_i$ is the weighted average of the area-specific healthcare seeking

behavior probabilities for ILI. Hence, the unadjusted estimate has a bias of

$$\begin{aligned} \mathrm{E}\left[e^{\hat{\alpha}}\right] - e^{\alpha} &\approx \frac{\mathrm{E}[Y_{n+1}]/N_{n+1}}{\left(\sum_{i=1}^{n} \mathrm{E}[Y_i]\right) / \left(\sum_{i=1}^{n} N_i\right)} - e^{\alpha} \\ &= e^{\alpha}\left[\frac{r_{n+1}\left(\sum_{i=1}^{n} N_i\right)}{\left(\sum_{i=1}^{n} N_i r_i\right)} - 1\right] = e^{\alpha}\left(\frac{r_{n+1}}{\bar{r}} - 1\right). \end{aligned} \tag{4.4}$$

Notice that the bias is zero when $r_{n+1} = \bar{r}$, i.e. when the healthcare seeking behavior probability in the intervention area is equal to the weighted average of healthcare seeking behavior probabilities in the control areas. To understand when the naïve model will yield a biased estimate, we consider the simplest case of two regions; let $n = 1$, so that $i = 2$ is the treatment region. The usual relative risk estimate then for two regions has bias

$$\mathrm{E}[e^{\hat{\alpha}}] - e^{\alpha} \approx e^{\alpha}\left[\frac{r_2}{r_1} - 1\right].$$

The bias in the unadjusted estimate is determined by the difference in unmeasured area-level confounding, i.e. when $r_2 \neq r_1$, and the direction of the bias depends on how the two regions differ.

To correct for this bias, we consider a negative control outcome. A good negative control would be a condition similar in nature to ILI, but would not be expected to be influenced by vaccination. We consider the AHCA data on gastrointestinal illness (GI) to be a suitable negative control since we do not expect GI cases to be influenced by a flu vaccination program, and we might speculate that the healthcare seeking behavior for GI complaints in an ED is similar to that for ILI in that ED.

Let Z_i denote the number of GI cases in county i that sought care in an ED. We suppose the true data generating model for GI cases in the ED during a single influenza season is

$$Z_i \,|\, s_i, q_i \sim \text{Poisson}\left(N_i s_i q_i\right), \tag{4.5}$$

where $s_i = \Pr(\text{ED} \,|\, \text{GI}, i)$ is the risk of going to the ED given GI in area i; and $q_i = \Pr(\text{GI} \,|\, i)$

is the risk of GI in area i. As before, we decompose the risk of being a GI case in the ED into the risk of being a case, and the probability of seeking healthcare at the ED conditional on being ill with GI. Although we choose the negative control outcome so that we expect no treatment effect, we can fit the same model we used for ILI to the negative control GI:

$$Z_i \,|\, \nu, \beta \sim \text{Poisson}(N_i e^{\nu+\beta x_i}), \tag{4.6}$$

where $\exp\{\nu\}$ is the average risk of GI in the non-intervention areas, x_i is a treatment area indicator, and $\exp\{\beta\}$ is the log relative risk associated with the intervention. Note that we expect $\beta = 0$ since the negative control was chosen such that it is unaffected by the treatment. Hence, a non-zero estimate of β reflects residual bias not accounted for in the model. In the negative control model, the MLE for the "treatment" effect is

$$e^{\hat{\beta}} = \frac{Z_{n+1}/N_{n+1}}{\left(\sum_{i=1}^{n} Z_i\right) / \left(\sum_{i=1}^{n} N_i\right)}.$$

Analogous to the ILI estimate in (4.4), the expected "treatment" effect for the negative control, under the true data generating model (4.5), is

$$\mathrm{E}\left[e^{\hat{\beta}}\right] \approx \frac{\mathrm{E}[Z_{n+1}]/N_{n+1}}{\left(\sum_{i=1}^{n} \mathrm{E}[Z_i]\right) / \left(\sum_{i=1}^{n} N_i\right)} = \left(\frac{s_{n+1} \sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} N_i s_i}\right) = \left(\frac{s_{n+1}}{\bar{s}}\right),$$

when we assume $\beta = 0$, and where $\bar{s} = \sum_{i=1}^{n} N_i s_i / \sum_{i=1}^{n} N_i$ is the weighted average of the GI-specific risk of seeking care at an ED in the control areas.

When the healthcare seeking behavior for GI differs in the treatment and control areas (i.e. $s_{n+1} \neq \bar{s}$), the difference is reflected in our estimate of β. In this way, the negative control outcome can be used to detect unmeasured confounding in the AHCA data.

Assuming that the healthcare seeking behavior is identical for ILI and GI and assuming there is no treatment effect for GI, we can use the estimated "treatment" effect from the negative control model to adjust for the same bias in our estimated treatment effect for ILI.

Mathematically,

$$\mathrm{E}\left[e^{\hat{\alpha}-\hat{\beta}}\right] \approx e^{\alpha}\left(\frac{r_{n+1}}{\bar{r}}\right)\left(\frac{\bar{s}}{s_{n+1}}\right),$$

which equals e^{α} when $r_{n+1}/\bar{r} = s_{n+1}/\bar{s}$, since $\beta = 0$. To formalize this, define $\gamma = \alpha - \beta$ to be the adjusted treatment effect. Therefore, the estimate of the adjusted treatment effect

$$e^{\hat{\gamma}} = e^{\hat{\alpha}-\hat{\beta}} = \frac{Y_{n+1}/\left(\sum_{i=1}^{n} Y_i\right)}{Z_{n+1}/\left(\sum_{i=1}^{n} Z_i\right)}, \tag{4.7}$$

which is the ratio of the unadjusted estimate of the treatment effect from (4.3) and the estimated "treatment" effect for the negative control outcome. The bias of the adjusted estimate of the treatment effect is approximately

$$\mathrm{E}\left[e^{\hat{\gamma}}\right] - e^{\alpha} \approx e^{\alpha}\left[\frac{r_{n+1}}{s_{n+1}}\frac{\bar{s}}{\bar{r}} - 1\right]. \tag{4.8}$$

Therefore, the adjusted estimate of the treatment effect is less biased than the unadjusted estimate when

$$\left|\frac{r_{n+1}}{s_{n+1}}\frac{\bar{s}}{\bar{r}} - 1\right| < \left|\frac{r_{n+1}}{\bar{r}} - 1\right|. \tag{4.9}$$

To better understand when the negative control approach will lead to a reduction in bias, we consider the simplest scenario of two regions. Since the negative control was chosen such that it is not affected by the treatment (i.e. $\beta = 0$), the estimated effect of the intervention on the ED visits for GI is

$$\mathrm{E}\left[e^{\hat{\beta}}\right] \approx \frac{\mathrm{E}[Z_2]/N_2}{\mathrm{E}[Z_1]/N_1} = \frac{s_2 e^{\nu+\beta}}{s_1 e^{\nu}} = e^{\beta}\frac{s_2}{s_1} = \frac{s_2}{s_1}.$$

Hence, the "treatment" effect for the negative control is equal to the relative risk of seeking healthcare for GI in an ED in the treatment area compared to the control area.

If the nature of the unobserved confounding is similar for both outcomes, we would want to use the estimated relative risk for healthcare seeking behavior from the negative control outcome to reduce the bias in the estimated treatment effect. Since

$$\mathrm{E}\left[e^{\hat{\alpha}-\hat{\beta}}\right] = \frac{(\mathrm{E}[Y_2]/N_2)/(\mathrm{E}[Y_1]/N_1)}{(\mathrm{E}[Z_2]/N_2)/(\mathrm{E}[Z_1]/N_1)} = e^{\alpha} \times \frac{r_2}{s_2} \times \frac{s_1}{r_1},$$

we will have less bias when $r_1/s_1 \approx r_2/s_2$. That is, if the ratio of unmeasured confounding for the two outcomes is similar across areas, we will expect a reduction in bias.

Standard error estimates for the adjusted treatment effect, $\hat{\gamma}$, can be obtained from the result of fitting both models,

$$\mathrm{Var}\left[\hat{\gamma}\right] = \mathrm{Var}\left[\hat{\alpha} - \hat{\beta}\right] = \mathrm{Var}\left[\hat{\alpha}\right] + \mathrm{Var}\left[\hat{\beta}\right].$$

In practice, those that use the negative control approach do not typically adjust the standard error estimates. More often, confidence intervals around the adjusted estimate are formed using the standard error estimates obtained in the unadjusted analysis (Richardson et al., 2015).

Thus far, we have described the previous approaches to adjusting for negative controls. While the approach is a straightforward way to both detect and adjust for unmeasured confounding, standard error estimates tend to be too small because they do not account for the variation in the negative control outcomes. In the next section, we reframe the negative control approach to directly model the adjusted treatment effect. A major benefit of our proposed approach is that by modeling the adjusted estimate directly, it automatically accounts for the variability in both outcomes. Additionally, our proposed model can also be extended to more complex situations. In Section 4.4.1, we compare the two approaches via simulation.

4.3 Reframing the negative control approach

We reframe the standard negative control framework to automatically obtain more appropriate standard error estimates and better accommodate model extensions. The key to our approach to the negative control framework is to recognize that the adjusted estimate in (4.7) is simply a ratio of risk ratios (or relative risks), which is the usual parameter associated with an interaction term in a log-linear regression model. Hence, we reframe the negative control approach as an interaction model. Following the previous section, we work through the results of an interaction model to obtain the same adjusted estimate. The advantage to modeling the adjusted treatment effect directly is that we automatically obtain appropriate standard error estimates.

The interaction model requires a slight change in notation. We now let Y_{ci} be the number of cases of condition c in area i, where $c = 1$ denote the outcome of interest (ILI) and $c = 2$ the negative control (GI). As before, we assume the true data generating model for a single influenza season is

$$Y_{1i} \,|\, r_i, \mu_1, \alpha \sim \text{Poisson}\left(N_i r_i \exp\{\mu_1 + \alpha x_i\}\right), \tag{4.10}$$

$$Y_{2i} \,|\, s_i, \mu_2 \sim \text{Poisson}\left(N_i s_i \exp\{\mu_2\}\right), \tag{4.11}$$

where μ_1 is the average log risk for the outcome of interest in the non-intervention area; μ_2 is the average risk for the negative control outcome; and α is the additional log risk associated with the treatment in the outcome of interest. As in (4.1) and (4.5), the area- and condition-specific heathcare seeking behaviors $r_i = \Pr(\text{ED} \,|\, \text{ILI}, i)$ and $s_i = \Pr(\text{ED} \,|\, \text{GI}, i)$ represent unmeasured confounding.

As before, if we naïvely fit $Y_{1i} \,|\, \mu, \alpha \sim \text{Poisson}\left(N_i \exp\{\mu + \alpha x_i\}\right)$, the MLE is a biased estimate of the treatment effect

$$\text{E}\left[\hat{\alpha}\right] = \text{E}\left[\log\left(\frac{Y_{1,n+1}/N_{n+1}}{(\sum_{i=1}^{n} Y_{1i})/(\sum_{i=1}^{n} N_i)}\right)\right] \approx \alpha + \log\left(\frac{r_{n+1}}{\bar{r}}\right).$$

To reduce the bias, we jointly model the outcome of interest and the negative control and model the adjusted estimate directly as an interaction in the model

$$Y_{ci} \,|\, \mu_1, \mu_2, \beta, \gamma \sim \text{Poisson}\left(N_i \exp\{\mu_1 1_{[c=1]} + \mu_2 1_{[c=2]} + \beta x_i + \gamma x_i 1_{[c=1]}\}\right), \tag{4.12}$$

where μ_1 is the average log relative risk for the outcome of interest; μ_2 is the log relative risk associated with the negative control outcome compared to the outcome of interest; β is the log relative risk of the treatment for the negative control outcome; and γ is the adjusted treatment effect of interest ($\gamma = \alpha - \beta$). The adjusted MLE of the treatment effect is

$$e^{\hat{\gamma}} = \frac{Y_{1,n+1} / \sum_{i=1}^{n} Y_{1i}}{Y_{2,n+1} / \sum_{i=1}^{n} Y_{2i}};$$

which is identical to the adjusted treatment effect estimate in equation (4.7). Under the true data generating model, the adjusted treatment effect is consistent for

$$\text{E}\left[\hat{\gamma}\right] = \text{E}\left[\log\left(\frac{Y_{1,n+1} / \sum_{i=1}^{n} Y_{1i}}{Y_{2,n+1} / \sum_{i=1}^{n} Y_{2i}}\right)\right] \approx \alpha + \log\left(\frac{r_{n+1}}{\bar{r}}\right) - \log\left(\frac{s_{n+1}}{\bar{s}}\right).$$

Therefore, the adjusted model will result in a less biased estimate of the treatment effect, α, as compared to the unadjusted estimate when

$$\left|\log\left(\frac{r_{n+1}}{\bar{r}}\right) - \log\left(\frac{s_{n+1}}{\bar{s}}\right)\right| < \left|\log\left(\frac{r_{n+1}}{\bar{r}}\right)\right|. \tag{4.13}$$

Note that this is equivalent to the conditions for bias reduction in equation (4.9) in the previous section. When the probability of seeking healthcare at an ED for ILI is the same in the two areas, i.e. $r_{n+1} = \bar{r}$, the unadjusted estimate is unbiased. In practice, if we believed the unmeasured confounding was identical across study areas, we would correctly fit the unadjusted model and obtain an unbiased estimate of the treatment effect. However, employing the negative control framework will yield unbiased estimates only if the unmeasured confounding for the negative control outcome is also identical in the two areas, i.e. $s_{n+1} = \bar{s}$.

When the unmeasured confounding is not identical across study areas, there is the opportunity to reduce the bias in the estimates with the use of negative controls. In general, the adjusted estimates bias will depend on how close the unobserved confounding for the negative control outcome is to that of the outcome of interest across all study areas. In particular, when $r_{n+1} > \bar{r}$, bias is reduced when

$$1 < \frac{s_{n+1}}{\bar{s}} < \left(\frac{r_{n+1}}{\bar{r}}\right)^2.$$

And when $r_{n+1} < \bar{r}$, bias is reduced when

$$\left(\frac{r_{n+1}}{\bar{r}}\right)^2 < \frac{s_{n+1}}{\bar{s}} < 1.$$

There are two scenarios when the the unadjusted and adjusted estimates are equally biased. First, when there is no difference in the amount of unmeasured confounding for the negative control outcome in the two areas i.e. $s_{n+1} = \bar{s}$, the adjusted estimate will have the same bias as the unadjusted estimate. Second, the adjusted estimate will be equally biased when $(r_{n+1}/\bar{r})^2 = s_{n+1}/\bar{s}$.

In the next section, we conduct a simulation study to demonstrate the differences in modeling approaches when using the negative control outcome to adjust for unmeasured confounding and to understand when the use of negative controls is beneficial.

4.4 Simulations

4.4.1 A comparison of negative control modeling approaches

We first consider simulations that compare different approaches to estimating a treatment effect when there is unmeasured confounding. We simulate two regions each with a population of 50,000. While ILI and GI are not rare diseases, people rarely arrive at the ED for such conditions. For all simulations, the overall disease risks, i.e. Pr(ILI) and Pr(GI), are 0.18 and 0.20 for the outcome of interest and the negative control outcome respectively. The log treatment effect is fixed at $\log(0.75) \approx -0.288$, corresponding to a 25% reduction in disease risk in the treatment area. We initially assume a perfect negative control outcome, so that the unmeasured confounding is identical for each condition; for the treatment area, $r_2 = s_2 = 0.55$ and for the reference area $r_1 = s_1 = 0.45$. For each of the 10,000 simulations, we compare the estimates obtained from three different models:

1. The *unadjusted* model fits the naïve log-linear model (4.2)

$$Y_{1i}|\beta_0, \beta_1 \sim \text{Poisson}\left(N_i \exp\{\beta_0 + \beta_1 x_i\}\right)$$

2. The *adjusted* model fits both outcomes separately and adjusts the estimated treatment effect, but does not adjust the standard error estimates. This is the standard approach described in Section 4.2. For each outcome $c = 1, 2$, we fit

$$Y_{ci} \,|\, \beta_{c0}, \beta_{c1} \sim \text{Poisson}\left(N_i \exp\{\beta_{c0} + \beta_{c1} x_i\}\right),$$

   and the adjusted estimate is $\exp\{\widehat{\beta_{11}} - \widehat{\beta_{21}}\}$.

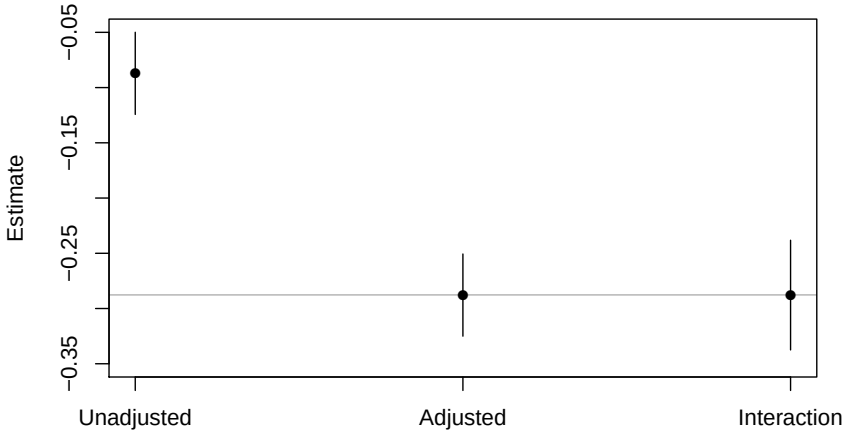
3. The *interaction* model fits both outcomes together and estimates the adjusted treatment effect directly, as in equation (4.12)

$$Y_{ci} \,|\, \mu_1, \mu_2, \beta, \gamma \sim \text{Poisson}\left(N_i \exp\{\mu_1 1_{[c=1]} + \mu_2 1_{[c=2]} + \beta x_i + \gamma x_i 1_{[c=1]}\}\right)$$

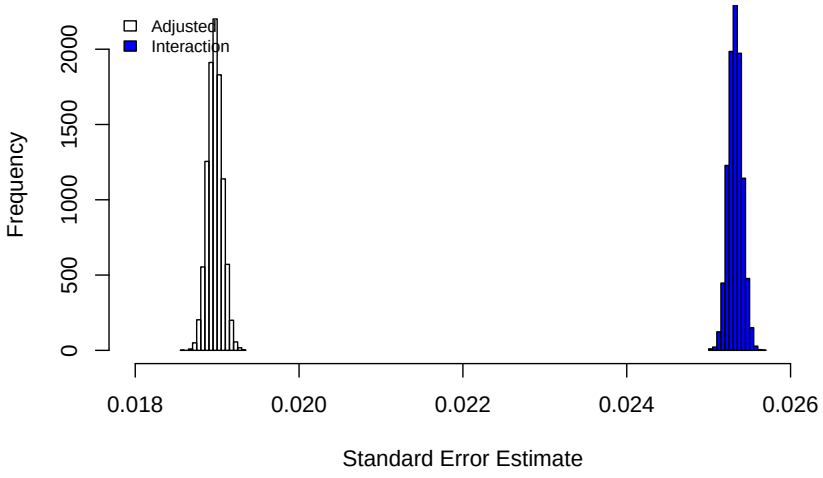
We expect that the adjusted models, which account for the unmeasured confounding via the negative control, will produce unbiased estimates. For now, we only consider the ideal negative control outcome. In subsequent simulations, we explore how the differences in unmeasured confounding will effect estimation.

Figure 4.2 compares the estimates and standard errors of the log relative risk associated with treatment in the three models. Figure 4.2(a) plots the average log relative risk associated with treatment obtained from each of the three models. We see that the unadjusted estimates are larger than the treatment effect, while the estimates from the adjusted and interaction models are correctly estimating the true treatment effect. Recall that the adjusted model estimates this bias explicitly. In fact, the unadjusted estimate is biased by the log ratio of unmeasured confounding in the two areas, i.e. $\log(r_2/r_1) = \log(0.55/0.45) \approx 0.2$. In the adjusted model, the bias that arises from the unmeasured confounding is estimated by analysis of the negative control outcome; the interaction model estimates this bias with the main treatment effect. Both approaches produce identical estimates of 0.2 (95% CI: 0.167-0.233). Notice that if the unmeasured confounding had been reversed for the two areas, we would see that the unadjusted estimate would be approximately 0.2 lower than the true value.

In Figure 4.2(b) we plot histograms of the standard error estimates obtained from the adjusted and interaction models. Recall that in these simulations, the adjusted model fits separate models for each outcome and obtains an adjusted estimate after, but fails to correct the standard error estimate to account for the variability in estimate from negative control estimate. As such the standard errors for the adjusted model are exactly those from the unadjusted model. We see that the adjusted model that fails to correct the standard errors yields standard error estimates that are much smaller than those obtained from the interaction model.



(a) Estimates of log relative risk.



(b) Histograms of standard error estimates.

Figure 4.2: On the top, we plot estimated log relative risk associated with treatment from the unadjusted, adjusted, and interaction models. The horizontal line indicates the true treatment effect used in simulations. On the bottom, we plot histograms of standard error estimates from the adjusted (white) and interaction (blue) models.

In Figure 4.3, we plot the coverage estimates for the three models. As expected, the bias of the unadjusted model results in poor coverage of the true treatment effect. The adjusted model, although unbiased, fails to account for the variability in the negative control outcomes, and thus yields 95% confidence intervals that are too narrow. In contrast, the interaction model has good coverage properties with approximately 95% coverage. By modeling the adjusted treatment effect directly, as in the interaction model, we obtain standard errors that more appropriately reflect the uncertainty.

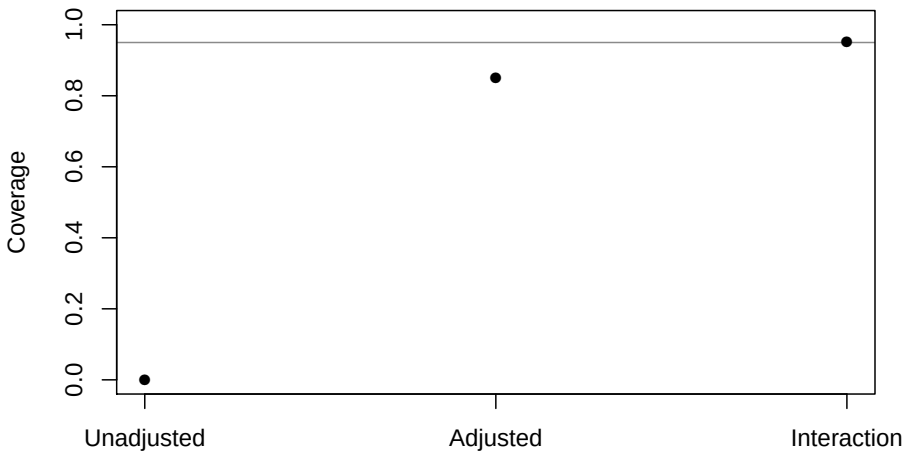


Figure 4.3: Coverage for the unadjusted, adjusted, and interaction models. The horizontal line indicates 95% coverage.

As simulations in the simplest scenario demonstrate, the interaction model produces less biased estimates, along with appropriate standard errors. These simulations have demonstrated the utility of the negative control outcome framework, when the nature of the unmeasured confounding is the same across the outcomes, i.e. in ideal circumstances. These simulations also demonstrate the utility in directly modeling the adjusted treatment effect using the interaction model.

In the next section, we conduct a simulation study to illuminate how the nature of the unmeasured confounding impacts the results of both the unadjusted and interaction models. In subsequent simulations, we only consider the interaction model when estimating the adjusted effect.

4.4.2 Unmeasured confounding and negative control outcomes

We conduct a series of simulations to examine when adjusting for unmeasured confounding via a negative control outcome yields bias reduction. We simulate observed ILI and GI cases following equation (4.10),

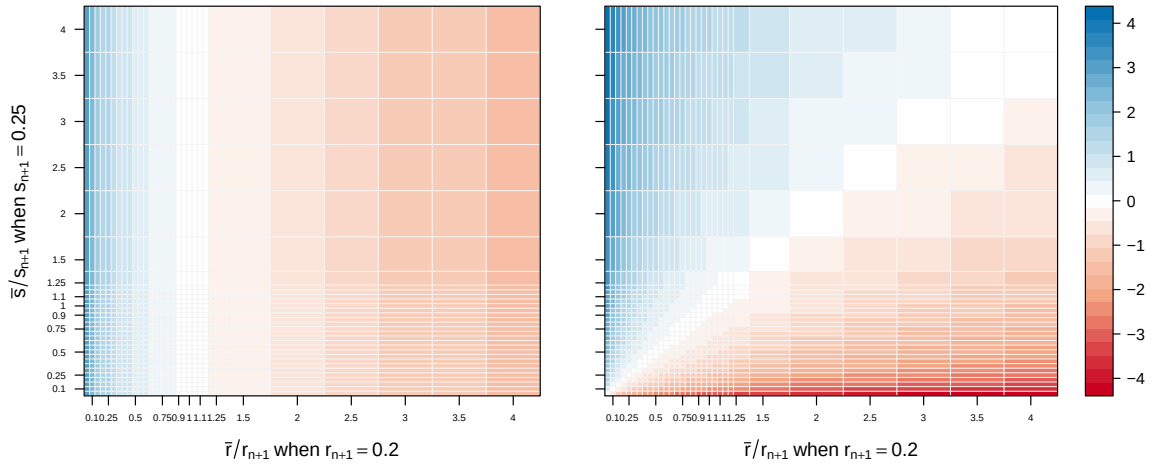
$$Y_{1i} \,|\, \mu_1, r_i, \alpha \sim \text{Poisson}\left(N_i \, r_i \, e^{\mu_1 + \alpha x_i}\right),$$
$$Y_{2i} \,|\, \mu_2, s_i \sim \text{Poisson}\left(N_i \, s_i \, e^{\mu_2}\right),$$

where $N_i = 50,000$ for all i, $\mu_1 = \log(0.18)$, $\mu_2 = \log(0.20)$, $\alpha = \log(0.75)$, and we specify r_i and s_i under a variety of scenarios. Since we can collapse into treatment and control areas, we let $i = 1, 2$. We fit both the unadjusted and the adjusted model and compare the estimates of the treatment effect. We refer to the estimates obtained using only ILI data as the unadjusted estimates; adjusted estimates are those from the interaction model in equation (4.12).

In Figure 4.4, we plot the bias of the unadjusted and adjusted estimates for varying amounts of unobserved confounding in the control areas for both outcomes. For the unadjusted estimates, we see that the bias only depends on how much the unmeasured confounding differs between the treatment region and the average of the control regions for the outcome of interest; this is expected since the unadjusted estimate ignores the negative control outcome. The unadjusted estimate has good properties when the amount of unobserved confounding is approximately the same in the treatment and average control regions for the outcome of interest. In contrast, the adjusted estimate will have small bias when the ratio of unmeasured confounders in the treatment and average control regions are similar for both the outcome

of interest and the negative control outcome (along the diagonal).

While the adjusted estimate performs well when the negative control outcomes are subject to similar unmeasured confounding as the outcome of interest, we see that the adjusted estimates can be severely biased if the negative control outcome is not carefully selected. For example, when the unobserved confounding for the outcome of interest in the average control region is four times that in the treatment region, but the opposite is true for the negative control outcome, the adjusted estimate is more biased than the unadjusted.

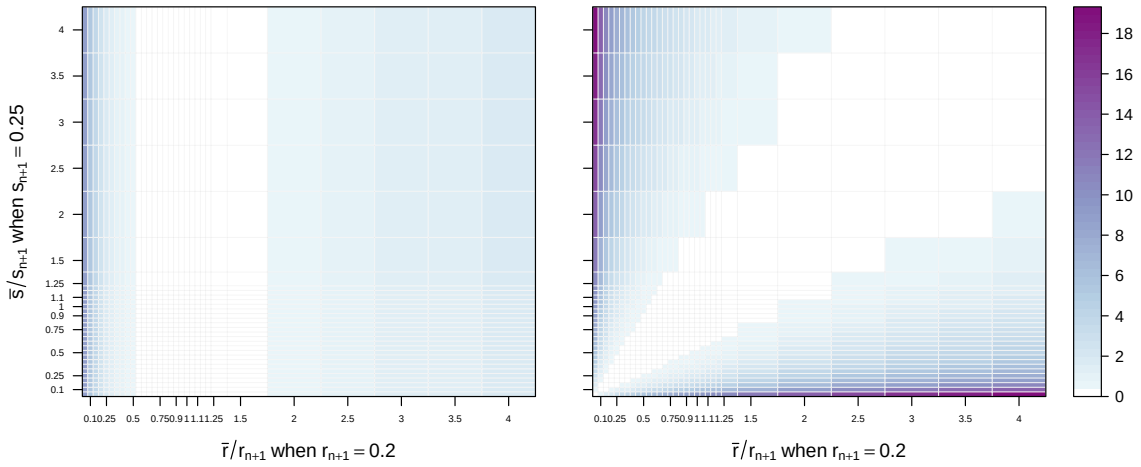


(a) Unadjusted estimates.

(b) Adjusted estimates.

Figure 4.4: Bias of the unadjusted (left) and adjusted (right) estimates for varying amounts of unmeasured confounding in the control areas. In the treatment region, $r_{n+1} = 0.20$ and $s_{n+1} = 0.25$.

A comparison of the mean squared error (MSE) of the unadjusted estimate to that of the adjusted as the amount of unmeasured confounding varies is presented in Figure 4.5. Similar to the bias, we see that when the amount of unobserved confounding across regions is similar for the outcome of interest, the unadjusted estimates have a small MSE. Moreover, when the unobserved confounding is similar for both outcomes the adjusted estimates have a low MSE. A poorly matched negative control can also produce an adjusted estimate with a very large MSE compared to that of the unadjusted estimate.



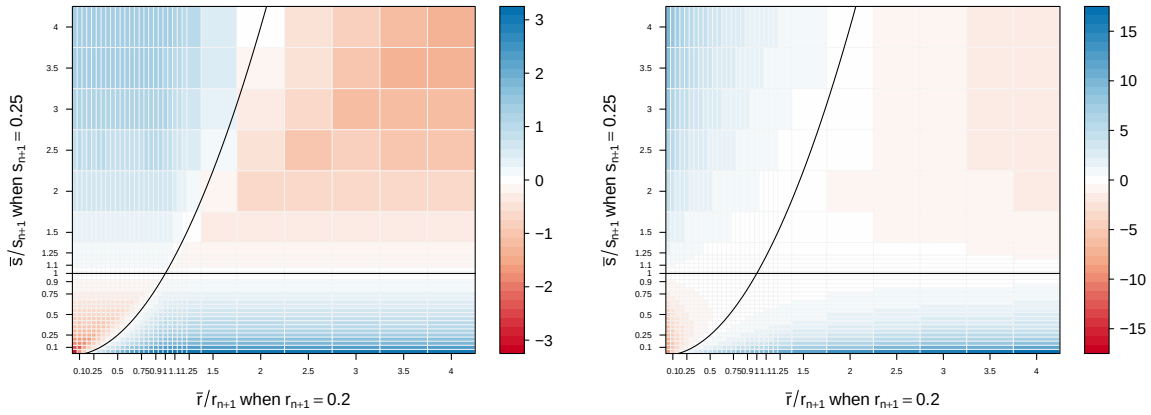
(a) Unadjusted estimates

(b) Adjusted estimates

Figure 4.5: Mean squared error (MSE) of the unadjusted (left) and adjusted (right) estimates for varying amounts of unmeasured confounding in the control areas. In the treatment region, $r_{n+1} = 0.20$ and $s_{n+1} = 0.25$.

In Figure 4.6, we plot the difference in absolute bias and MSE between the adjusted and unadjusted estimates for varying amounts of unmeasured confounding in the control regions areas. The red indicates a reduction in the absolute bias of the adjusted estimates compared to the absolute bias of the unadjusted estimates; blue indicates an increase in the absolute bias of the adjusted estimates compared to the absolute bias of the unadjusted estimates. The black lines indicate where there is no reduction in bias, when $s/s_{n+1} = 1$ and when $(\bar{s}/s_{n+1}) = (\bar{r}/r_{n+1})^2$. The absolute bias (and MSE) is reduced when the condition in (4.13) holds, as expected. Again, a poorly chosen negative control outcome will result in a worse estimate of the true treatment effect.

Coverage for both the unadjusted and adjusted estimates are presented in Figure 4.7 for various amounts of unobserved confounding. As expected, when the confounding is similar in the treatment and control regions, the unadjusted estimates have good coverage properties; when the negative control outcome is subject to similar confounding as the outcome of interest, the adjusted estimate has the expected coverage properties. For both estimates,



(a) Difference in absolute bias.

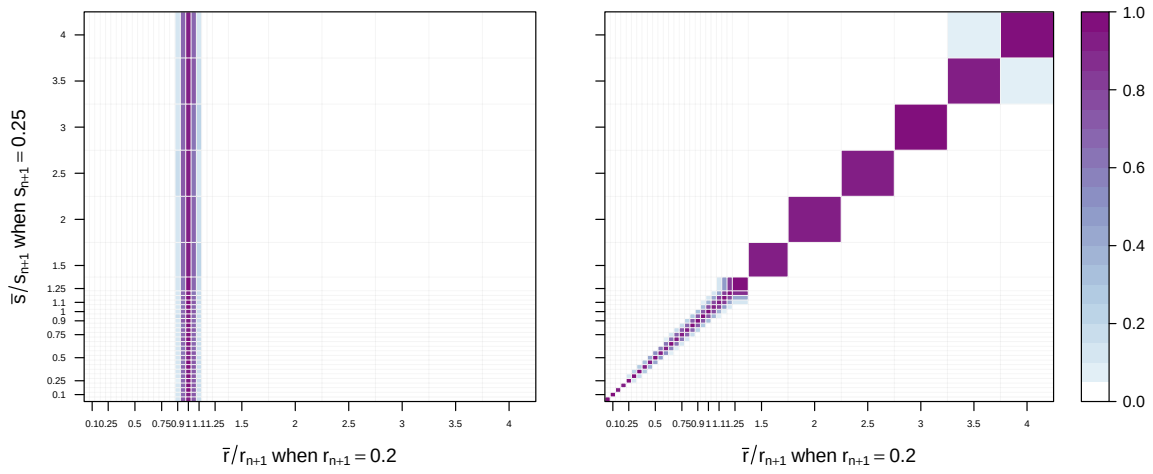
(b) Difference in MSE.

Figure 4.6: Difference in absolute bias (left) and MSE (right) between the adjusted and unadjusted estimates for varying amounts of unmeasured confounding in the control regions. In the treatment region, $r_{n+1} = 0.20$ and $s_{n+1} = 0.25$. The black lines indicate where there is no reduction in bias, when $s/s_{n+1} = 1$ and when $(\bar{s}/s_{n+1}) = (\bar{r}/r_{n+1})^2$. The red (blue) indicates a reduction (increase) in the absolute bias of the adjusted estimates compared to the absolute bias of the unadjusted estimates.

coverage quickly decreases when the confounding differs in the two areas and across outcomes.

It is clear from these simulations that utilizing negative controls can be beneficial at reducing bias, under certain circumstances. However, these results also emphasize the importance of a well selected negative control outcome. When the direction of the unmeasured confounding is dissimilar across outcomes, adjusted estimates can be much worse than the unadjusted estimates.

In practice, the data cannot be used to determine if we have selected a good negative control outcome. Thus, negative controls should be selected using additional subject matter and institutional knowledge along with external studies. The validity of the results from a negative control analysis rests on the untestable assumption that the two outcomes are subject to similar biases. Thus results from *both* the unadjusted and adjusted analyses should be presented, so that the reader can see how results differ under of the negative control assumptions.



(a) (b)

Figure 4.7: Coverage of the unadjusted (left) and adjusted (right) estimates for varying amounts of unmeasured confounding in the control regions. In the treatment region, $r_{n+1} = 0.20$ and $s_{n+1} = 0.25$.

4.5 Extension to spatially structured data

4.5.1 The spatial negative controls model

With surveillance data, we often have observations not just across multiple areas but also over multiple time points. Although the data is still subject to unmeasured confounding, the nature of repeated observations frequently eases some of the concerns. For example, with one study area, pre-intervention observations can serve as controls for the post-intervention observations assuming that the nature of the unmeasured confounding has remained stable over time, which is more likely if time periods are close together. If it is reasonable to assume the amount of unmeasured confounding is stable across the multiple time points, it would not be necessary to adjust estimates using a negative control outcome. In fact, using a negative control outcome to adjust the original estimates in this scenario could result in a more biased estimate, as we saw in the simulations of Section 4.4.2.

In this section, we consider the problem of estimating the treatment effect for a single

observation with spatially structured observations. Let $i = 1, \ldots, n$ denote the control areas, and $i = n + 1$ represents the treatment area. As before, we let Y_{ci} be the number of cases of condition c in county i that sought care at an ED; let N_i be the total population in county i; and the data generating model is

$$Y_{1i} \,|\, r_i, \theta_i \sim \text{Poisson}\left(N_i r_i \theta_i\right),$$
$$Y_{2i} \,|\, s_i, q_i \sim \text{Poisson}\left(N_i s_i q_i\right),$$

where we decompose the area- and condition-specific risk of seeking healthcare at the ED into the product of the area-specific disease risk and the conditional probability of seeking healthcare given a specific condition and area. The negative control model for spatially structured data is $Y_{ci} \,|\, \theta_{ci} \sim \text{Poisson}(N_i\, \theta_{ci})$, with

$$\log \theta_{ci} = \beta_0 1_{[c=1]} + \beta_1 1_{[c=2]} + \beta_2 x_i + \gamma x_i 1_{[c=1]} + \epsilon_{ci} + S_{ci}, \tag{4.14}$$

where β_0 is an overall level for the outcome of interest; β_1 is the overall level for the negative control outcome; β_2 is the treatment effect; γ is the adjusted treatment effect; $\boldsymbol{\epsilon}_c = (\epsilon_{c1}, \ldots, \epsilon_{c,n+1})$ are independent condition- and area-specific random effects; and $\boldsymbol{S}_c = (S_{c1}, \ldots, S_{c,n+1})$ are condition-specific spatially structured random effects. We discuss the forms of the spatially structured random effects in the next section.

Implementing the seemingly simple form of (4.14) requires serious care and consideration for two reasons. First, to adjust for a negative control in a spatially structured setting requires the joint modeling of two diseases. Second, when the exposure of interest (here a treatment indicator) also has spatial structure, spatial random effects may misattribute some of the exposure signal to the underlying spatial surface. In Section 4.5.2, we describe an approach to modeling disease outcomes jointly. In Section 4.5.3, we consider the risk for confounding by location in our application and discuss modeling approaches to handle it. We then return our attention to the spatial negative control model. In Section 4.5.4

we conduct simulations to examine how our choice of joint disease model and accounting for spatial confounding affect the estimated treatment effects when adjusted for a negative control.

4.5.2 Joint disease models

We consider two different approaches to specific spatial structure in the spatial negative control model. The first approach is to simply model each condition with independent ICAR distributions, so that in (4.14), $\boldsymbol{S}_1 \sim \text{ICAR}(\tau_1)$ and $\boldsymbol{S}_2 \sim \text{ICAR}(\tau_2)$. However, as many diseases share common risk factors it may be possible to leverage information to obtain better estimates of the parameters of interest. The shared component model (SCM) considers both a shared and disease-specific spatial structure. While the shared component model was originally developed using a cluster model (Knorr-Held and Best, 2001), it has since been extended to map multiple diseases jointly. We describe a common parameterization of the SCM, following Dabney and Wakefield (2005).

For ease, we describe the SCM for two diseases, indexed with $c = 1, 2$. Let Y_{ci} and E_{ci} be the observed and expected counts for disease c in area i. The first stage of the SCM assumes cases follow a Poisson distribution, $Y_{ci} \,|\, \theta_{ci} \sim \text{Poisson}(E_{ci}\theta_{ci})$, where θ_{ci} is the disease- and area-specific relative risk, and disease-specific relative risks take the form

$$\log \theta_{1i} = \beta_{10} + \eta_{1i} + \delta\phi_i,$$
$$\log \theta_{2i} = \beta_{20} + \eta_{2i} + \phi_i/\delta,$$

where β_{10} and β_{20} are disease-specific intercepts, ϕ_i is the shared component, δ scales the shared component, and η_{1i} and η_{2i} are disease specific random effects. For identifiability, proper priors must be assigned to η_{ci} and ϕ_i. The SCM can also accommodate area-level covariates for spatial regression, but as we will see in subsequent sections, care should be taken to avoid spatial confounding.

4.5.3 Confounding by location

In the disease mapping literature, it has been observed that including spatial random effects can alter the estimated fixed effects of interest (Clayton et al., 1993; Wakefield, 2007; Hodges and Reich, 2010). Including spatially structured random effects can be particularly problematic when covariates in the model are also spatially structured. In our setting, the covariate we include is binary indicating and non-zero for the single county where the school-located vaccination program was implemented. The SLIV program analysis is an extreme example of confounding by location.

To see why this is problematic, consider the univariate spatial regression model to examine the effect of the intervention, where x_i, is the binary treatment indicator and non-zero in a single area $i = n + 1$.

$$
\begin{aligned}
Y_i \,|\, \theta_i &\sim \text{Poisson}(N_i\theta_i), \\
\log \theta_i &= \beta_0 + \beta_1 x_i + \epsilon_i + S_i,
\end{aligned}
$$

where β_0 is an overall level, β_1 is the treatment effect, $\epsilon_i \sim \text{N}(0, \sigma^2_\epsilon)$ are unstructured random effects, and $\boldsymbol{S} = (S_1, \ldots, S_{n+1})$ are spatially structured random effects. However, with a single intervention area, and data from a single time only, it is impossible to distinguish the area-level effects, S_{n+1} and ϵ_{n+1} from the intervention effect β_1. That is, the fixed effect of interest is totally confounded by space.

Hughes and Haran (2013) proposed fitting spatial random effects that are orthogonal to the explanatory variables included in the regression. By re-parameterizing the standard convolution model, Hughes and Haran (2013) remove the spatial confounding, as well as reduce the dimensionality of the spatial random effects. In our situation, we define $\boldsymbol{X}$ to be the fixed effect design matrix and $\boldsymbol{A}$ the adjacency matrix for the counties of Florida. The Moran operator matrix is defined as $\boldsymbol{P}^\perp \boldsymbol{A} \boldsymbol{P}^\perp$, where $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ is the usual projection matrix onto the null space, and $\boldsymbol{M}$ contains the first h eigenvectors of the Moran

operator. The Hughes and Haran model then fits

$$\begin{aligned} Y_{1i} \,|\, \theta_i &\sim \text{Poisson}(N_i\theta_i), \\ \log \theta_i &= \beta_0 + \beta_1 x_i + \boldsymbol{M}_i\boldsymbol{\delta}_s, \\ p(\delta_s \,|\, \tau) &\propto \tau^{h/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\delta}_s^{'}\boldsymbol{Q}_s\boldsymbol{\delta}_s\right), \end{aligned}$$

where $\boldsymbol{Q}_s = \boldsymbol{M}^{'}\boldsymbol{Q}\boldsymbol{M}$, and $\boldsymbol{Q}$ is the full structure matrix from the ICAR model. While the orthogonalization approach successfully removes the spatial confounding in the estimated fixed effects, in general, we lose the nice interpretation of the spatial random effects and therefore do not consider it in our application.

A simple approach, when the only covariate is the intervention indicator, would be to fit the BYM model, but remove the random effects in the treatment region. In essence this would model the fixed effects in terms of the intervention region, and allow variability around that estimate in the reference regions. We refer to this as the modified Besag-York-Mollie (mBYM). Mathematically, the unadjusted mBYM model is

$$\begin{aligned} Y_{1i} \,|\, \theta_i &\sim \text{Poisson}(N_i\theta_i), \\ \log \theta_i &= \beta_0 + \beta_1 x_i + \epsilon_i + S_i, \\ \epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, n, \\ \boldsymbol{S} &= (S_1, \ldots, S_n) \sim \text{ICAR}(\tau_s), \end{aligned}$$

where ϵ_i are independent random effects and $\boldsymbol{S}$ are spatially structured random effects, both of which are defined only for non-treatment areas, (i.e. $i = 1, \ldots, n$). However, as we will see in the simulations in Section 4.5.4, this approach leads to standard error estimates that are too small.

4.5.4 Simulations for spatial negative controls

We simulate two outcomes for the counties of Florida in the following way. For $i = 1, \ldots, 67$, and $c = 1, 2$, we let

$$Y_{ic} \,|\, r_{ic}, \mu_c, S_{ic}, \alpha \sim \text{Poisson}(N_i r_{ic} \exp\{\mu_c + S_{ic} + \alpha 1_{[c=1]} 1_{[i=67]}\}),$$

where N_i is the population of county i, r_{ic} is the unmeasured confounding for area i and outcome c, μ_c is the condition specific average log risk, S_{ic} is the spatial disease component, and α is the log relative risk associated with the intervention in Alachua ($i = 67$). For all simulations, we fix $\mu_1 = \log(0.18) \approx -1.715$, $\mu_2 = \log(0.20) \approx -1.609$, and $\alpha = \log(0.75) \approx -0.288$. We describe how r_{ic} and $\boldsymbol{S}$ will vary over simulations in Table 4.1.

Outcomes Matching		Bivariate ICAR			Alachua		Rest of FL	
Spatially	Confounding	τ_1	τ_2	ρ	$r_{30,1}$	$r_{30,2}$	r_{i1}	r_{i2}
Similar	Good	9	9	0.99	0.5	0.45	0.75	0.7
Similar	Bad	9	9	0.99	0.5	0.75	0.75	0.5
Moderately similar	Good	9	9	0.5	0.5	0.45	0.75	0.7
Moderately similar	Bad	9	9	0.5	0.5	0.75	0.75	0.5
Dissimilar	Good	9	9	0	0.5	0.45	0.75	0.7
Dissimilar	Bad	9	9	0	0.5	0.75	0.75	0.5

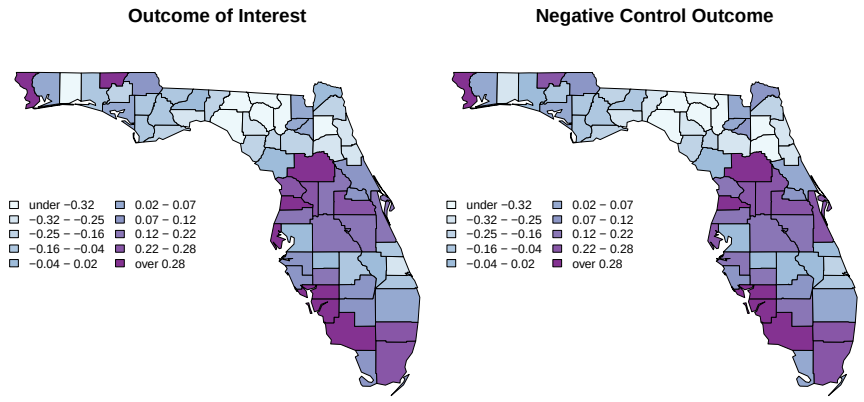
Table 4.1: Parameter specifications for spatial simulations

The outcomes are considered to be well matched in terms of confounding when the relationship between the treatment and control areas is similar across the two outcomes. Note that for the outcome of interest, the unmeasured confounding is constant across simulations.

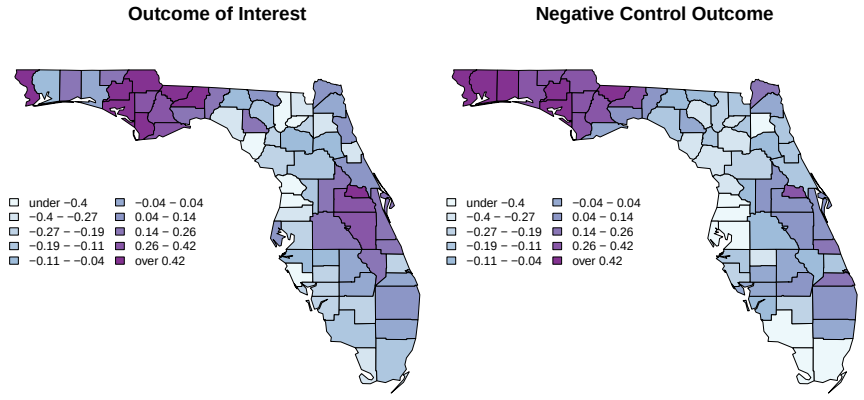
We generate $\boldsymbol{S}$ from a bivariate ICAR distribution, following Martinez-Beneito (2013). To do so, let ϕ_{ij} be the spatial effect in area i for condition j, where $i = 1, \ldots, n$ and $j = 1, 2$. Let $\boldsymbol{\phi}_i = (S_{i1}, S_{i2})$ be the area-level spatial effects for the two conditions of interest, and let $\boldsymbol{\phi}_j$ be the vector of random effects for the jth condition. We simulate independent realizations from the ICAR distribution, following Algorithm 3.1 described by Rue and Held

(2005), and then collect these into a $n \times 2$ matrix, where $\Phi = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$. Let $\boldsymbol{\Lambda}$ be the 2×2 correlation matrix for the two spatial random effects within a given area, and let $\tilde{\boldsymbol{\Lambda}}$ denote the upper-triangular matrix of the Cholesky decomposition. Martinez-Beneito (2013) showed that bivariate ICAR realizations, denoted $\boldsymbol{S}_c$ for condition c, are obtained by transforming independent ICAR variables as follows: $(\boldsymbol{S}_1 \ \boldsymbol{S}_2) = \Phi\tilde{\boldsymbol{\Lambda}}$. That is, we induce correlation between two independent ICAR variables via the upper-triangular matrix from the Cholesky decomposition of the between-disease correlation matrix. Further details can be found in Martinez-Beneito (2013) and Botella-Rocamora et al. (2015).

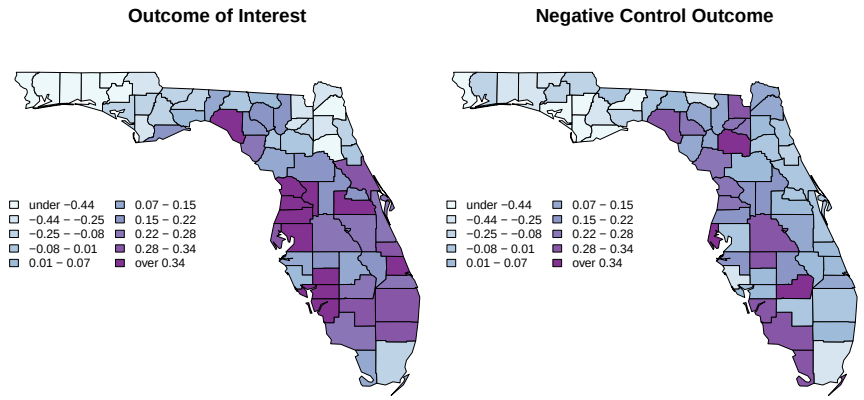
In simulations, we examine three possible underlying spatial structures for the two outcomes: when outcomes are highly correlated (and spatially similar); when outcomes are moderately correlated; and when outcomes each have spatial structure, but are uncorrelated (and spatially dissimilar). In Figure 4.8, we plot examples of the spatial structure for each of the outcomes in the three scenarios. Adjusting for a negative control outcome requires modeling two outcomes, each of which may have spatial structure. We therefore consider different approaches to modeling two disease outcomes. In Section 4.5.2, we discussed possible approaches to modeling multiple diseases.



(a) When the two outcomes are spatially similar, $\rho = 0.99$.



(b) When the two outcomes are moderately similar, $\rho = 0.5$.



(c) When the two outcomes are spatially dissimilar, $\rho = 0$.

Figure 4.8: Example of realizations from bivariate ICAR models used to simulate the true underlying spatial structure for outcomes that are well matched spatially (top), outcomes are poorly matched (bottom), and in between (middle).

For these simulations, we fit and compare three unadjusted and five adjusted models. All models are implemented in INLA. For all models, x_i is the treatment indicator, equal to one only for the outcome of interest in the intervention region ($c = 1$ and $i = 67$):

1. Unadjusted GLM:

$$\begin{aligned} Y_{1i} \,|\, \theta_i &\sim \text{Quasi-Poisson}(N_i\theta_i) \\ \log \theta_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

2. Unadjusted BYM model:

$$\begin{aligned} Y_{1i} \,|\, \theta_i &\sim \text{Poisson}(N_i\theta_i) \\ \log \theta_i &= \beta_0 + \beta_1 x_i + \epsilon_i + S_i \\ \epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \quad \text{for } i = 1, \ldots, 67 \\ \boldsymbol{S} &= (S_1, \ldots, S_{67}) \sim \text{ICAR}(\tau_s) \end{aligned}$$

3. Unadjusted modified BYM model:

$$\begin{aligned} Y_{1i} \,|\, \theta_i &\sim \text{Poisson}(N_i\theta_i) \\ \log \theta_i &= \beta_0 + \beta_1 x_i + \epsilon_i + S_i \\ \epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \quad \text{for } i = 1, \ldots, 66 \\ \boldsymbol{S} &= (S_1, \ldots, S_{66}) \sim \text{ICAR}(\tau_s) \end{aligned}$$

4. Adjusted GLM:

$$\begin{aligned} Y_{ci} \,|\, \theta_{ci} &\sim \text{Quasi-Poisson}(N_i\theta_{ci}) \\ \log \theta_{ci} &= \beta_0 + \beta_1 x_i + \beta_2 1_{[c=1]} + \beta_3 x_i 1_{[c=1]} \end{aligned}$$

5. Adjusted BYM model:

$$
\begin{aligned}
Y_{ci} \,|\, \theta_{ci} &\sim \text{Poisson}(N_i \theta_{ci}) \\
\log \theta_{ci} &= \beta_0 + \beta_1 x_i + \beta_2 1_{[c=1]} + \beta_3 x_i 1_{[c=1]} + \epsilon_i + S_i \\
\epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \quad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{S} &= (S_1, \ldots, S_{67}) \sim \text{ICAR}(\tau_s)
\end{aligned}
$$

6. Adjusted modified BYM:

$$
\begin{aligned}
Y_{ci} \,|\, \theta_{ci} &\sim \text{Poisson}(N_i \theta_{ci}) \\
\log \theta_{ci} &= \beta_0 + \beta_1 x_i + \beta_2 1_{[c=1]} + \beta_3 x_i 1_{[c=1]} + \epsilon_i + S_i \\
\epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \quad \text{for } i = 1, \ldots, 66 \\
\boldsymbol{S} &= (S_1, \ldots, S_{66}) \sim \text{ICAR}(\tau_s)
\end{aligned}
$$

7. Adjusted condition-specific BYM (cBYM):

$$
\begin{aligned}
Y_{ci} \,|\, \theta_{ci} &\sim \text{Poisson}(N_i \theta_{ci}) \\
\log \theta_{ci} &= \beta_0 + \beta_1 x_i + \beta_2 1_{[c=1]} + \beta_3 x_i 1_{[c=1]} + \epsilon_{ci} + S_{ci} \\
\epsilon_{1i} &\sim \text{N}(0, \sigma_1^2) \quad \text{for } i = 1, \ldots, 67 \\
\epsilon_{2i} &\sim \text{N}(0, \sigma_2^2) \quad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{S}_1 &= (S_{1,1}, \ldots, S_{1,67}) \sim \text{ICAR}(\tau_1) \\
\boldsymbol{S}_2 &= (S_{2,1}, \ldots, S_{2,67}) \sim \text{ICAR}(\tau_2)
\end{aligned}
$$

8. Shared components model (SCM):

$$
\begin{aligned}
Y_{ci} \,|\, \theta_{ci} &\sim \text{Poisson}(N_i \theta_{ci}) \\
\log \theta_{ci} &= \beta_0 + \beta_1 x_i + \beta_2 1_{[c=1]} + \beta_3 x_i 1_{[c=1]} + \epsilon_{ci} + S_i 1_{[c=1]} + \delta S_i 1_{[c=2]} \\
\boldsymbol{S} &= (S_1, \ldots, S_{67}) \sim \text{ICAR}(\tau) \\
\epsilon_{1i} &\sim \text{N}(0, \sigma_1^2) \quad \text{for } i = 1, \ldots, 67 \\
\epsilon_{2i} &\sim \text{N}(0, \sigma_2^2) \quad \text{for } i = 1, \ldots, 67
\end{aligned}
$$

We summarize the estimates and 95% confidence or credible intervals for the various models over 250 simulations in Figure 4.9 and Table 4.2. As expected, the unadjusted models produce similar estimates regardless of the nature of the negative control outcome. All unadjusted models produce similarly biased estimates, while the standard errors vary across models. Adjusted models are close to the true value only when the negative control outcome has unmeasured confounding similar to that of the outcome of interest. Poorly matched confounding leads to severely biased estimates. Models that do not include condition-specific random effects (such as adjBYM) have very narrow credible intervals. When the spatial structures are dissimilar, the shared component model has much wider intervals. As before, we see that adjusting for the negative control outcome is only beneficial when the nature of the unmeasured confounding is similar across outcomes. Failing to account for the spatial structure, as in both the unadjusted and adjusted GLM models, results in very large standard error estimates, and wide intervals.

In many simulation settings, the adjusted models that model common spatial and independent random effects across outcomes (Adjusted BYM and Adjusted mBYM) perform identically in terms of estimating the adjusted treatment effect. While these two models produce different estimates for the other fixed effects, they perform similarly in terms of estimating the parameters of interest. In Figure 4.10, we plot the random effect estimates obtained from the Adjusted mBYM model against those from the Adjusted BYM. Removing the spatial random effect for Alachua does not change the estimated spatial random effects

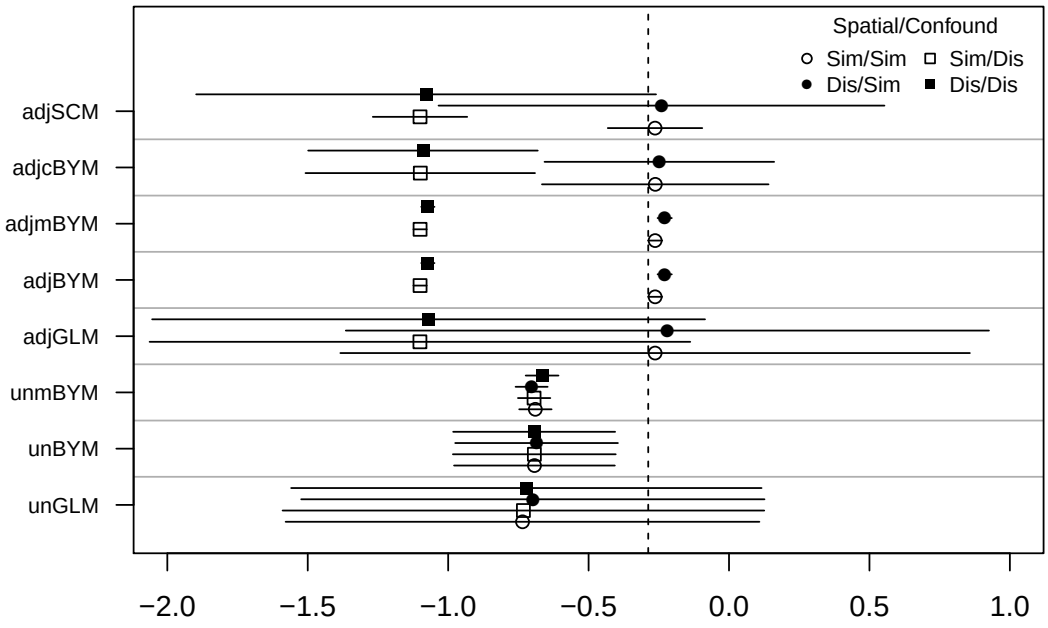


Figure 4.9: Simulations with well matched confounding are denoted by circles; squares correspond to poorly matched confounding; open symbols indicate similar spatial patterns; and filled symbols are simulations with spatially dissimilar outcomes. The dashed, vertical line denotes the truth.

for the other regions, while the independent random effect for Alachua county is estimated to be nearly zero.

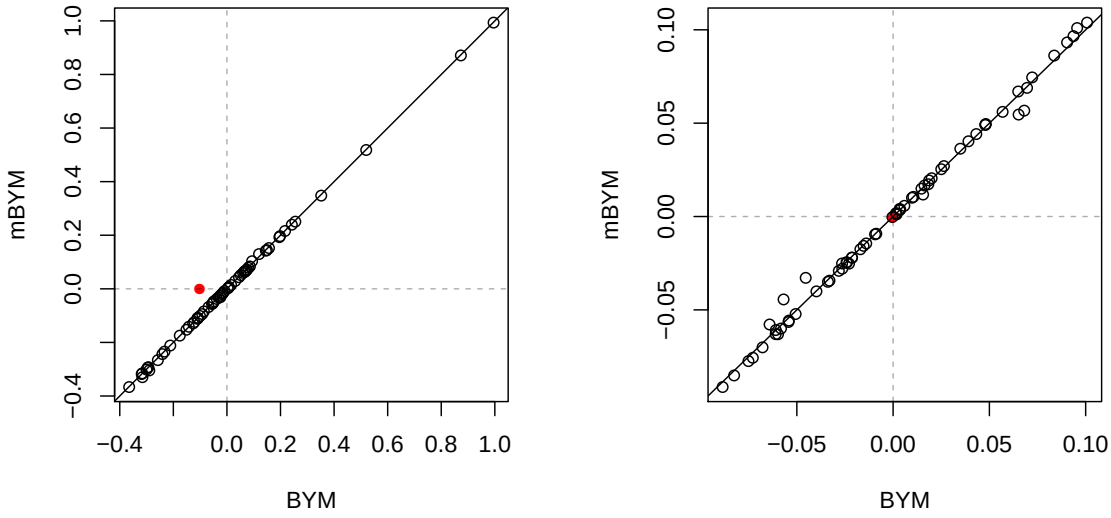
As we have seen in previous simulations, when the unmeasured confounding is similar between the treatment and control areas, it is advantageous to adjust for the negative control outcome. When both the spatial structure and the unmeasured confounding are also similar, the shared component model will provide the least biased estimate, and the corresponding standard error estimates appropriately reflect the uncertainty. When the spatial structures differ between the two diseases, but the negative control is well matched, we see that modeling the spatial structure separately (cBYM) yields estimates with low MSE and standard error

		Unmeasured Confounding Well Matched							Unmeasured Confounding Poorly Matched						
		Est	SD	95% CI		Cover	Bias	MSE	Est	SD	95% CI		Cover	Bias	MSE
$\rho = 0.99$	Unadjusted														
	GLM	-0.712	0.439	-1.573	0.148	0.973	-0.425	0.244	-0.733	0.455	-1.625	0.158	0.967	-0.446	0.263
	BYM	-0.674	0.172	-1.013	-0.336	0.353	-0.387	0.175	**-0.696**	**0.168**	**-1.028**	**-0.364**	**0.273**	**-0.408**	**0.191**
	mBYM	-0.668	0.019	-0.707	-0.630	0.040	-0.381	0.190	-0.695	0.019	-0.732	-0.657	0.020	-0.407	0.210
	Adjusted models														
	GLM	-0.217	0.595	-1.383	0.949	1.000	0.070	0.024	-1.094	0.516	-2.105	-0.082	0.673	-0.806	0.673
	BYM	-0.220	0.013	-0.245	-0.194	0.107	0.068	0.022	-1.095	0.012	-1.118	-1.071	0.000	-0.807	0.674
	cBYM	-0.221	0.242	-0.698	0.255	1.000	0.066	0.024	-1.095	0.238	-1.563	-0.628	0.020	-0.807	0.676
	mBYM	-0.220	0.013	-0.245	-0.194	0.107	0.068	0.022	-1.095	0.012	-1.118	-1.071	0.000	-0.807	0.674
	SCM	**-0.219**	**0.157**	**-0.528**	**0.089**	**0.973**	**0.068**	**0.022**	-1.096	0.156	-1.403	-0.790	0.000	-0.809	0.676
		Est	SD	95% CI		Cover	Bias	MSE	Est	SD	95% CI		Cover	Bias	MSE
$\rho = 0.5$	Unadjusted														
	GLM	-0.739	0.481	-1.683	0.204	0.973	-0.451	0.271	-0.759	0.455	-1.651	0.134	0.967	-0.471	0.285
	BYM	-0.689	0.166	-1.017	-0.362	0.327	-0.402	0.183	**-0.705**	**0.168**	**-1.036**	**-0.374**	**0.287**	**-0.417**	**0.198**
	mBYM	-0.701	0.018	-0.738	-0.664	0.007	-0.413	0.215	-0.717	0.019	-0.754	-0.679	0.013	-0.429	0.234
	Adjusted models														
	GLM	-0.259	0.652	-1.536	1.019	1.000	0.029	0.071	-1.102	0.522	-2.125	-0.080	0.673	-0.815	0.729
	BYM	-0.252	0.013	-0.278	-0.227	0.100	0.036	0.060	-1.105	0.012	-1.129	-1.081	0.000	-0.818	0.724
	cBYM	**-0.244**	**0.237**	**-0.711**	**0.222**	**0.967**	**0.043**	**0.036**	-1.110	0.238	-1.578	-0.643	0.020	-0.822	0.708
	mBYM	-0.252	0.013	-0.278	-0.227	0.100	0.036	0.060	-1.105	0.012	-1.129	-1.081	0.000	-0.818	0.724
	SCM	-0.251	0.270	-0.782	0.279	0.973	0.037	0.047	-1.109	0.263	-1.626	-0.592	0.113	-0.821	0.720
		Est	SD	95% CI		Cover	Bias	MSE	Est	SD	95% CI		Cover	Bias	MSE
$\rho = 0$	Unadjusted														
	GLM	-0.705	0.432	-1.551	0.141	0.967	-0.417	0.230	-0.724	0.434	-1.574	0.126	0.953	-0.436	0.245
	BYM	-0.682	0.169	-1.016	-0.348	0.327	-0.394	0.179	**-0.697**	**0.169**	**-1.030**	**-0.364**	**0.313**	**-0.410**	**0.192**
	mBYM	-0.662	0.019	-0.699	-0.624	0.013	-0.374	0.180	-0.679	0.019	-0.717	-0.642	0.020	-0.392	0.191
	Adjusted models														
	GLM	-0.219	0.593	-1.381	0.943	1.000	0.069	0.115	-1.108	0.504	-2.096	-0.120	0.607	-0.820	0.783
	BYM	-0.222	0.013	-0.248	-0.197	0.060	0.065	0.096	-1.105	0.012	-1.129	-1.082	0.000	-0.817	0.762
	cBYM	**-0.235**	**0.240**	**-0.707**	**0.238**	**0.967**	**0.053**	**0.044**	-1.115	0.239	-1.585	-0.646	0.047	-0.828	0.728
	mBYM	-0.222	0.013	-0.248	-0.197	0.060	0.065	0.096	-1.105	0.012	-1.129	-1.082	0.000	-0.817	0.762
	SCM	-0.245	0.315	-0.865	0.375	0.980	0.043	0.058	-1.122	0.318	-1.747	-0.497	0.193	-0.834	0.760

Table 4.2: Summary of results for simulations when the negative control outcome matches the spatial variability and and unmeasured confounding in various combinations. The true treatment effect is $\log(0.75) \approx -0.288$.

estimates that are appropriate. When the unmeasured confounding differs between the two outcomes, the unadjusted analyses yield less biased results, and including spatial structure in the usual way is the best option. Although simulations suggest that the unadjusted GLM has good coverage properties, this will not generally be the case. The unadjusted GLM estimate will be biased, and ignoring the spatial structure results in very wide intervals.

These simulations show that while it is important to understand the nature of the condition-specific spatial structures before modeling the outcomes jointly, it is much more important to have a negative control that properly adjusts for the unmeasured confounding. Nevertheless, if the negative control approach is to be used with spatially structured



(a) Spatial random effects.

(b) Independent random effects.

Figure 4.10: Comparison of estimated spatial (left) and independent (right) random effects when excluding the Alachua-specific effects against those from the full BYM model. Estimates corresponding to Alachua county (the treatment area) are in red and equal to zero when not modeled.

outcomes, it is critical to perform exploratory analyses in order to best account for the underlying condition-specific spatial structure by comparing the log SMR estimates for each outcome. Moreover, it is important to include the exploratory results along with those from the unadjusted and adjusted analyses. In Section 4.7.3, where we apply these methods to the AHCA data, we use exploratory analyses to learn about the underlying spatial structures and inform our modeling decisions.

4.6 Extension to stratified populations

We briefly consider stratification in the negative control framework. Stratification is a frequently employed technique to control for known (*a priori*) confounders. In epidemiological studies, it is common to stratify the population on demographic characteristics such as age and sex. We first outline an approach in the absence of unmeasured confounding, then

examine how the the presence of unmeasured confounding affects the analysis.

In general, we assume $i = 1, \ldots, n+1$ study regions, and $j = 1, \ldots, J$ strata. We denote the number of cases in area i and stratum j by Y_{ij}, and let N_{ij} be the population of the jth stratum in area i. For rare disease, we assume

$$Y_{ij} \,|\, p_{ij} \sim \text{Poisson}\left(N_{ij} p_{ij}\right), \tag{4.15}$$

where p_{ij} is an area- and stratum-specific risk. As before, we assume that $i = n+1$ is the intervention (or treatment) region, while $i = 1, \ldots, n$ correspond to the control regions. The model described by equation (4.15) has $(n+1) \times J$ parameters to estimate. Frequently this exceeds the number of disease events, and hence is unestimable. When this is the case, it is common to make a proportionality assumption to reduce the dimensionality of the problem. In particular, we assume that the area- and stratum-specific risk is proportional to a stratum-specific reference risk and write $p_{ij} = p_j \theta_i$, where $\theta_i = p_{ij}/p_j$ is the relative risk of disease associated with area i and p_j is a stratum-specific reference rate. Internally standardized reference rates are defined as:

$$\hat{p}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{\sum_{i=1}^{n} N_{ij}},$$

for $j = 1, \ldots, J$. Stratum-specific reference rates allow us to define area-specific expected disease counts as $E_i = \sum_{j=1}^{J} N_{ij} \hat{p}_j$. The proportionality assumption implies the area-level model

$$Y_i = \sum_{j=1}^{J} Y_{ij} \sim \text{Poisson}\left(\theta_i \sum_{j=1}^{J} N_{ij} p_j\right) = \text{Poisson}(E_i \theta_i), \tag{4.16}$$

and in our situation, we consider modeling $\log \theta_i = \mu + \alpha x_i$, were x_i is a treatment indicator.

To examine how the internal standardization procedure performs in the presence of unmeasured confounding, we let Y_{ij} denote the number of ILI cases in county i and stratum j who sought care at the ED; and N_{ij} is the total population of stratum j in county i. As

before, the data generating model is assumed to be

$$Y_{ij} \sim \text{Poisson}\left(N_{ij} p_{ij}\right), \tag{4.17}$$

where $p_{ij} = \Pr\left(\text{ED and ILI} \,|\, i, j\right)$, and we can decompose the risk in (4.17) such that

$$p_{ij} = \Pr\left(\text{ED and ILI} \,|\, i, j\right) = \Pr\left(\text{ED} \,|\, \text{ILI}, i, j\right) \Pr\left(\text{ILI} \,|\, i, j\right) = q_{ij}\, \phi_{ij}, \tag{4.18}$$

where ϕ_{ij} is the stratum- and area-specific disease risk and q_{ij} is the stratum- and area-specific risk of seeking healthcare given ILI. An extension of the usual proportionality assumption implies

$$q_{ij} = \Pr\left(\text{ED} \,|\, \text{ILI}, i, j\right) = \Pr\left(\text{ED} \,|\, \text{ILI}, j\right)\, r_i = d_j\, r_i, \tag{4.19}$$

$$\phi_{ij} = \Pr\left(\text{ILI} \,|\, i, j\right) = \Pr\left(\text{ILI} \,|\, j\right) \theta_i = p_j\, \theta_i, \tag{4.20}$$

where the area-level parameters

$$\theta_i = \frac{\Pr\left(\text{ILI} \,|\, i, j\right)}{\Pr\left(\text{ILI} \,|\, j\right)} \qquad \text{and} \qquad r_i = \frac{\Pr\left(\text{ED} \,|\, \text{ILI}, i, j\right)}{\Pr\left(\text{ED} \,|\, \text{ILI}, j\right)},$$

are each relative risks. Moreover,

$$r_i \theta_i = \frac{\Pr\left(\text{ED} \,|\, \text{ILI}, i, j\right) \Pr\left(\text{ILI} \,|\, i, j\right)}{\Pr\left(\text{ED} \,|\, \text{ILI}, j\right) \Pr\left(\text{ILI} \,|\, j\right)} = \frac{\Pr\left(\text{ED and ILI} \,|\, i, j\right)}{\Pr\left(\text{ED and ILI} \,|\, j\right)},$$

is also a relative risk. When the very strong assumption of proportionality is made for for the disease and healthcare-seeking components, as above, the stratum-specific probabilities can be combined so that

$$d_j\, p_j = \Pr\left(\text{ED} \,|\, \text{ILI}, j\right) \Pr\left(\text{ILI} \,|\, j\right) = \Pr\left(\text{ED and ILI} \,|\, j\right) = \lambda_j.$$

Hence, the standard proportionality assumption results in the unadjusted area- and stratum-

specific model

$$Y_{ij} \,|\, r_i, \theta_i \sim \text{Poisson}\left(N_{ij}\lambda_j r_i \theta_i\right), \tag{4.21}$$

where λ_j is the stratum-specific risk of seeking healthcare at an ED and having ILI; r_i is the area-specific relative risk of seeking healthcare given ILI; and θ_i is the area-specific relative risk of disease. If we then follow the usual internal standardization procedure, we obtain stratum-specific reference rates

$$\hat{\lambda}_j = \sum_{i=1}^{n} Y_{ij} \Big/ \sum_{i=1}^{n} N_{ij},$$

which account for stratum-specific differences in healthcare seeking behavior. Area-specific expected counts are then $E_i = \sum_{j=1}^{J} N_{ij}\hat{\lambda}_j$, and the resulting analysis fits

$$Y_i \sim \text{Poisson}\left(E_i r_i \theta_i\right), \tag{4.22}$$

where r_i and θ_i are relative risks. Any stratum-level confounding is adjusted for via internal standardization, and the only remaining unmeasured confounding is the area-specific healthcare seeking behaviors (the r_i's). Thus, when the proportionality assumption is appropriate for the underlying data generating model, the usual internal standardization approach will result in the model analogous to those previously developed.

When the proportionality assumptions for both (4.19) and (4.20) are too strong, it would also be possible to assume proportionality in the only disease risk, say, so that $p_{ij} = q_{ij}\,\theta_i\,p_j$. In the settings where the proportionality assumption is too strong for both the heathcare seeking behavior and disease risk, it is still possible to decompose the risk as in (4.18).

4.7 Application to the AHCA data

4.7.1 Introduction

Recall, we are interested in learning about the effects of the school-located influenza vaccination program that was implemented at the start of the 2009/2010 school year in Alachua county, Florida. Alachua County is located in northwest Florida, see Figure 4.11, and is neighbored by seven other Florida counties; we define the Alachua region, in blue, to be the surrounding 23 counties. In 2005, Florida had a population of 17.84 million people, which

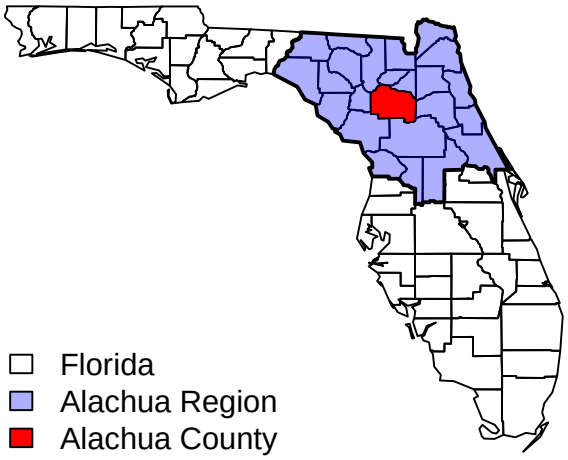


Figure 4.11: Map of the counties in Florida. Alachua County is highlighted in red, and counties in the Alachua region are in blue.

has grown to nearly 20 million in 2014, according to the US Census. Alachua County's population has seen less growth over the course of the study period, compared to the rest of the state; Alachua has had between 235,000 and 250,000 residents over the course of the study period.

In Figure 4.12, we summarize the locations of AHCA facilities (EDs) in relation to the Florida population for 2010. The number of AHCA facilities is relatively stable over time, as is the density of the Florida population over the course of the study period. Not surprisingly, more facilities are located in counties with larger populations, such as Miami-Dade county in

southeast Florida. In more rural areas, such as the northwestern panhandle, a single facility may serve multiple counties.

The influenza seasons typically begin in October and run through May (CDC, 2016). We define the influenza season as consecutive fourth and first quarters; that is, the fourth quarter of 2011 and the first quarter of 2012 make up the 2011/2012 flu season. AHCA provides quarterly data from 2005 through 2014, which makes nine complete influenza seasons available for analysis. We drop the first quarter of 2005 and the fourth quarter of 2014, since each are only half of a influenza season. We stratify county populations into five age groups: $[0, 5), [5, 20), [20, 45), [45, 65), 65+$.

4.7.2 Simple negative control

We examine the effect of the SLIV program in Alachua county for the 2012/2013 influenza season and consider using a negative control to obtain less biased estimates. As always, the result of adjusting for a negative control depends upon an untestable assumption: that the nature of the unmeasured confounding is similar across treatment and control areas.

A simple first analysis compares Alachua county to the surrounding region and to the rest of Florida by collapsing across the counties in the two possible reference regions. In Figure 4.11, we highlight the 23 counties which define the Alachua Region. Table 4.3 summarizes the ILI and GI counts for Alachua, the Alachua region, and the rest of Florida during the two quarters that make up the 2012/2013 influenza season.

		Total Cases		Incidence	
	Population	ILI	GI	ILI	GI
Alachua	252,299	7,136	4,170	28.2	16.5
Alachua Region	3,240,610	141,005	65,147	43.5	20.1
Florida	19,201,766	710,599	375,367	37.0	19.5

Table 4.3: Summary of 2012/2013 season. Incidence are per 1,000.

We present the unadjusted and adjusted estimate of the vaccine program effect for two

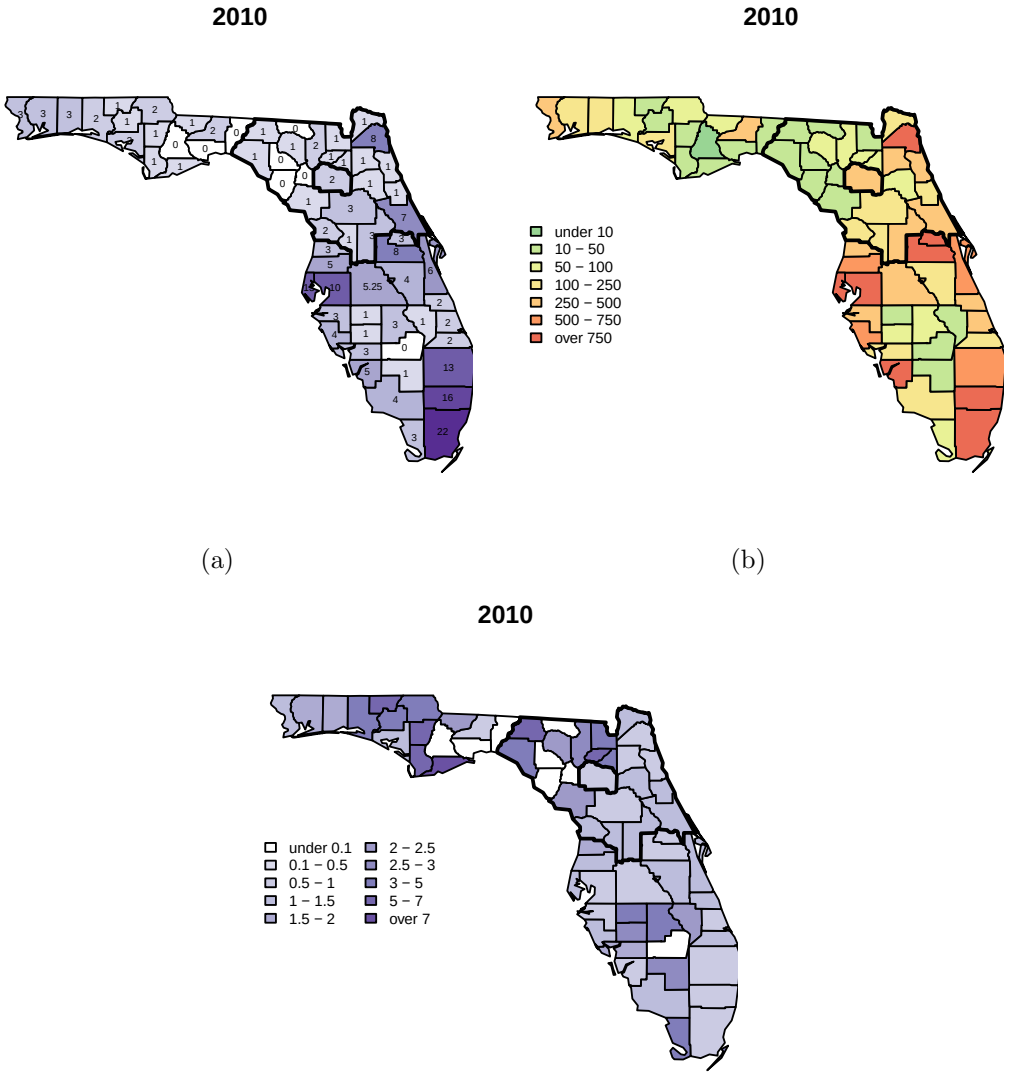


Figure 4.12: Summary of AHCA facilities and Florida population in 2010. The number of AHCA facilities in each county are shown in (a); the 2010 population density per square mile is shown in (b); and the number of AHCA facilities per 100,000 residents is shown in (c).

comparison regions in Table 4.4. All comparisons suggest a potential program effect for the 2012/2013 influenza season, however the size of the effect and the corresponding uncertainty varies across the comparisons. We estimate the vaccination program is associated with a 40% lower rate of ILI cases in Alachua compared to the surrounding counties (95% CI: 41%–39%). However, by adjusting for unmeasured confounding via a negative control, we estimate an 18% lower rate associated with the vaccination program compared to the surrounding counties (95% CI: 29%–5%). Compared to the rest of Florida, the vaccination program is associated with a 27% lower rate of ILI in Alachua (95% CI: 29%–25%); adjustment for the negative control results in an estimated 5% lower rate of cases in Alachua county, as compared to the rest of Florida (95% CI: 21% lower to 14% higher).

	log Scale						Original Scale			
	Unadjusted			Adjusted			Unadjusted		Adjusted	
Comparison	Est	SE	95% CI	Est	SE	95% CI	Est	95% CI	Est	95% CI
Alachua Region	-0.51	0.01	(-0.53, -0.49)	-0.19	0.07	(-0.34, -0.05)	0.60	(0.59, 0.61)	0.82	(0.71, 0.95)
Florida	-0.32	0.01	(-0.34, -0.29)	-0.05	0.09	(-0.23, 0.13)	0.73	(0.71, 0.75)	0.95	(0.79, 1.14)

Table 4.4: Unadjusted and adjusted estimates of the SLIV program effect by comparison region for the 2012/2013 season.

As we see from the very simple analysis, there is some suggestion of a reduction of ILI-related ED visits associated with the SLIV program. However, collapsing the counties into treatment and control regions in this way is certainly a reduction in information. In the next section, we do not collapse across counties and instead estimate the single season SLIV program effect using the spatial negative control analysis.

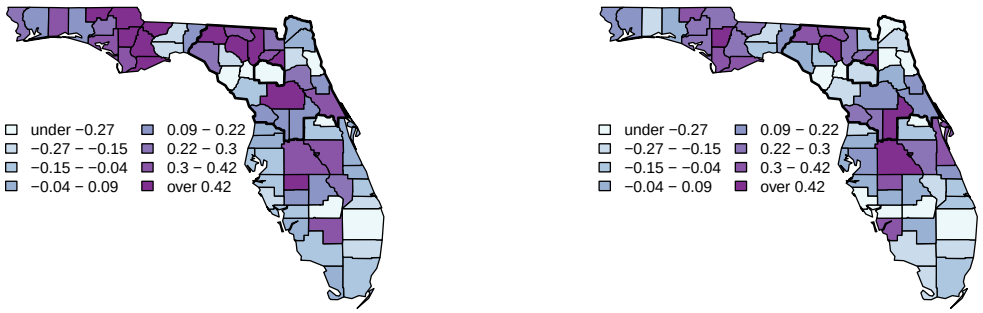
4.7.3 Spatial negative control analysis

We now consider estimating the SLIV effect, taking into account the spatial nature of our data. To account for differences in the county-level age structures, we standardize counts via

internal standardization, as described in Section 4.6. The spatial negative control model is

$$Y_{ci}|\theta_{ci} \sim \text{Poisson}(E_{ci}\,\theta_{ci})$$
$$\log \theta_{ci} = \beta_0 1_{[c=1]} + \beta_1 1_{[c=2]} + \beta_2 x_i + \gamma x_i 1_{[c=1]} + \epsilon_{ci} + S_{ci},$$

where E_{ci} is the expected count for county i and condition c, and θ_{ci} is the relative risk of condition c in area i. As we saw in Section 4.5.4, it is important to understand the nature of the underlying spatial structure for both outcomes before fitting the spatial negative control model. To this end, as an exploratory first step for the spatial analysis, we plot the log standardized morbidity ratios (SMRs) of both ILI and GI for the 2012/2013 flu season in Figure 4.13. Broadly, the patterns of the condition-specific log SMRs appear similar.



(a) log SMR for ILI in 2012/2013 season.

(b) log SMR for GI in 2012/2013 season.

Figure 4.13: Summary of log SMRs for ILI and GI during 2012/2013 influenza season.

To better understand the spatial similarities, we plot the raw log SMRs for ILI against those of GI in Figure 4.14. As expected, we see a strong association and find that the log SMRs have a correlation of 0.81. For the 2012/2013 influenza season, the similarity in spatial structure of the two conditions is clear from our initial examination of the raw log SMRs. However, in other settings, we can imagine the similarity in spatial structure is less obvious, especially in areas with small populations, where SMRs will be unstable.

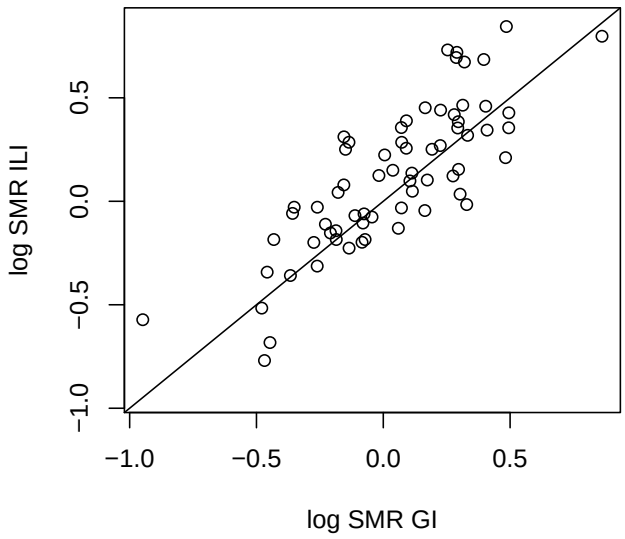


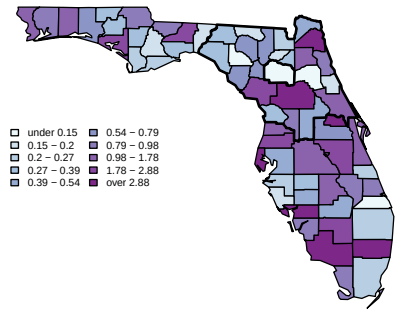
Figure 4.14: Scatter plot of log SMRs for ILI against log SMR for GI.

When the spatial similarity is less obvious, we can take a modeling approach to examine the similarity in spatial structure for the two outcomes. Specifically, for each outcome, $c = 1, 2$, we fit the following model to obtain smoothed estimates of the condition-specific relative risk.

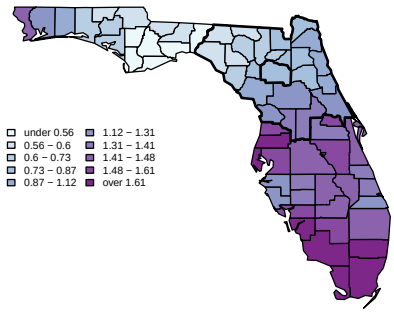
$$
\begin{aligned}
Y_{ci}|\theta_{ci} &\sim \text{Poisson}(N_i\theta_{ci}) \\
\log\theta_{ci} &= \beta_0 + \epsilon_i + S_i \\
\epsilon_i &\sim \text{N}(0, \sigma^2_\epsilon) \quad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{S} &= (S_1, \ldots, S_{67}) \sim \text{ICAR}(\tau_s).
\end{aligned}
$$

In Figure 4.15 we plot the residual relative risk estimates for the non-spatial and spatial components of the separate smooth models. The estimated spatial residual relative risks for each outcome in Figures 4.15(b) and 4.15(d) are very similar. However, for both outcomes, the scale of non-spatial residual relative risks is much wider than the spatial residual relative risks, suggesting that the spatial structure can only explain a small portion of the residual risk for these data. Using empirical estimates of the marginal variances for the spatially structured random effects, we can estimate the proportion of the total variance that is explained by the spatially structured random effects. For ILI, we estimate that approximately 18% of the total variation can be explained by the spatially structured random effects; for GI, spatial random effects explain 16% of the total variation.

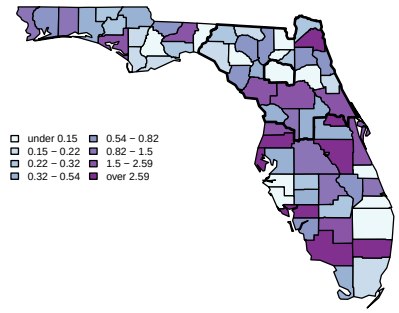
With a good understanding of the spatial nature of the two conditions, we can now turn our attention towards estimating the SLIV program effect. As we saw in simulations, the unadjusted model is preferable to an adjusted model using a poorly chosen negative control outcome. With spatial data, the standard BYM spatial regression model is the unadjusted model that best accommodates the spatial structure. Assuming that GI is a good negative control outcome for ILI, then adjusted models yield less biased estimates. The simulations in Section 4.5.4 showed that when there is strong spatial similarity in the outcomes, the shared component model yields less biased estimates with appropriate standard errors.



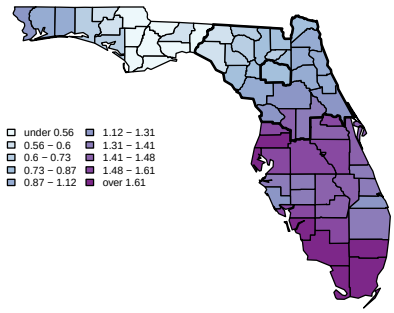
(a) Non-spatial residual relative risk for ILI.



(b) Spatial residual relative risk for ILI.



(c) Non-spatial residual relative risk for GI.



(d) Spatial residual relative risk for GI.

Figure 4.15: Estimates of the residual relative risk for non-spatial and spatial contributions. Note that the scales differ for the spatial and non-spatial maps.

Since it is impossible to determine the quality of GI as a negative control GI for ILI, we present the results from both the unadjusted and the adjusted model. Estimated log program effects from both models are summarized in Table 4.5. We estimate that the SLIV program is associated with a 34% lower rate of ILI cases for the 2012/2013 influenza season, with a 95% credible interval ranging from a 65% lower to 23% higher rate of cases. Using the adjusted model, we see the estimated SLIV program effect to be less dramatic. In particular, we now estimate the SLIV program is associated with a 13% lower rate of ILI cases, with a 95% credible interval ranging from 42% lower to 32% higher. We should be cautious when

	log Scale			Original Scale	
	Est	SD	95% CI	Est	95% CI
Unadjusted (BYM)	-0.421	0.318	(-1.047, 0.204)	0.656	(0.351, 1.227)
Adjusted (SCM)	-0.133	0.210	(-0.546, 0.279)	0.875	(0.579, 1.322)

Table 4.5: Unadjusted and adjusted estimates of log SLIV program effect.

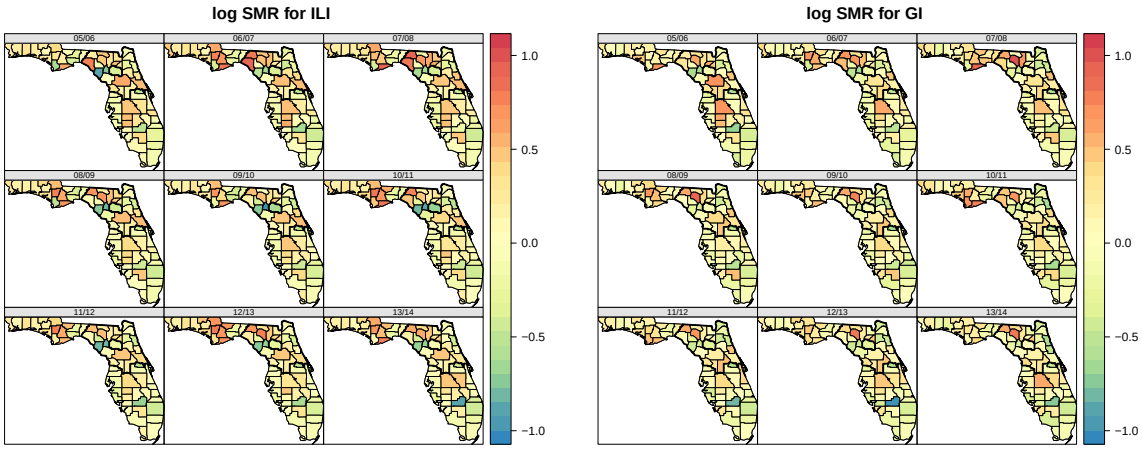
interpreting the results of these models. It seems unlikely that GI is a perfect negative control for ILI. While healthcare seeking behavior for GI may be very similar to ILI, it is unlikely that other sources of unmeasured confounding are equally similar across the two outcomes. This analysis does not account for background vaccination levels, which are likely to differ dramatically across counties. By standardizing the rates before the analysis, we have accounted for differences in age groups, however there are likely additional regional differences that are unmeasured and cannot be adjusted using GI as a negative control.

As a result, it is likely that both the unadjusted and adjusted estimates are biased. Nevertheless, both analyses suggest that the SLIV program is associated with a lower rate of ILI. While neither analysis yields statistically significant credible intervals, there is little power to detect a reduction with one treatment area and data aggregated to influenza seasons and counties.

While the estimated effect of the SLIV program for a single season is of interest, the AHCA data set has a total of 10 years of quarterly data, half of which is after the initiation of the SLIV program in Alachua. In the next section, we explore estimating the program effect changes over time.

4.7.4 Evaluating SLIV effect over many seasons

As a first look into how the estimated SLIV program effect changes over time, we examine each influenza season separately, following the single season analysis in the previous section. In Figure 4.16 we plot the raw log SMRs for both ILI and GI. We see that in general, the spatial patterns are similar over time. In fact, the correlations between seasonal raw log



(a) log SMRs for ILI.

(b) log SMRs for GI.

Figure 4.16: Raw log SMRs for ILI and GI evaluated at each of the 9 influenza seasons.

SMRs range from 0.75 to 0.87, with an average of approximately 0.80.

For each of the nine seasons, we fit three models: the unadjusted quasi-Poisson GLM, the unadjusted spatial BYM, and the adjusted SCM. In Figure 4.17, we plot the posterior estimated log SLIV program effect for the nine seasons. For pre-SLIV program seasons, the unadjusted estimates are relatively stable, with wide credible intervals and in general, the adjusted estimates are closer to zero. Note that we expect the estimated effect to be zero in pre-intervention periods. However, the difference between the unadjusted and adjusted estimates changes over the pre-intervention seasons. This suggests that the nature of unmeasured confounding in ILI and GI changes over time. During the first two years of the SLIV program, we see some evidence of a positive effect, with credible intervals excluding zero for all three models in the 2010/2011 influenza season. The results for the three most recent seasons are more difficult to interpret as there is not a consistent trend in the estimates.

We expect the SLIV program effect to vary in each season for a number of reasons. Part of the effect will depend on the efficacy of the influenza vaccine in use each season. Another

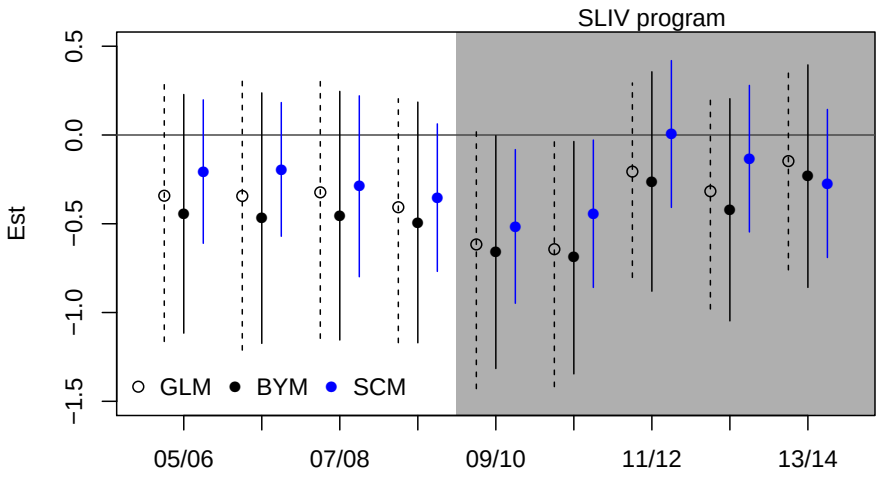


Figure 4.17: Posterior estimates and 95% CI for the SLIV program effect for 9 seasons. Dashed lines indicate estimates from the unadjusted quasi-Poisson model; credible intervals from unadjusted and adjusted spatial models are depicted in solid black and blue, respectively.

part of the effect will depend on vaccine coverage. In particular, there are three types of coverage to consider: coverage of the vaccination program itself, among school-aged children in Alachua county, coverage in the other age groups in Alachua county, and coverage levels in the other counties in Flordia. Each of these aspects is likely to vary over time, suggesting that a more complex spatio-temporal model for the program effect is necessary.

Extending the negative control framework to include spatially structured random effects required extensive simulations to understand how the choice of modeling in the two disease context affects the adjusted estimate. In the next section, we explore the underlying spatio-temporal structure of the two outcomes.

4.8 Spatio-temporal negative control analysis

4.8.1 Hierarchical spatio-temporal models for ILI

As we saw with space, extending the negative control framework for more complex spatio-temporal models requires careful consideration of the underlying patterns to ensure the effects of the program are not mis-attributed to spatial, temporal, or spatio-temporal effects. We first examine the spatio-temporal patterns of the AHCA data to determine appropriate models for ILI and GI over 9 years.

In Section 4.7.4, we only considered data from the two quarters that defined a influenza season. To model temporal trends we consolidate quarters for yearly observations, where we define years starting in the second quarter and through the first quarter of the subsequent year and collapse over the four quarters. For example, the 07/08 year consists of the second, third, and fourth quarter of 2007, as well as the first quarter of 2008. In this way, we are allowing influenza seasons to be analyzed together in a single year. In Figure 4.18 we map the log SMRs for ILI by county and years. We see that the maps tend to get redder (higher SMR's) over time, suggesting a temporal trend. Spatial trends are harder to distinguish, and will therefore require further investigation.

As a first step towards understanding the spatio-temporal patterns in the AHCA yearly data, we fit separate quasi-Poisson models to each area, where $\log \theta_{it} = \beta_{0i} + \beta_{1i} t$. We compare estimates across the counties to gain a sense into the nature and variability of the area-specific estimates. In Figure 4.19, we summarize the results from these models. Figure 4.19(a) shows a histogram of the area-specific slopes, Figure 4.19(b) plots the fitted area-specific slopes. Most counties have a gradual increasing trend, and Alachua (in red) looks like most of the region, as well as the rest of Florida. It does appear as though Alachua tended to have a lower log relative risk of ILI compared to the rest of Florida, however. Figure 4.19(c) maps the exponentiated slopes, suggesting some potential spatial structure in the estimated slopes.

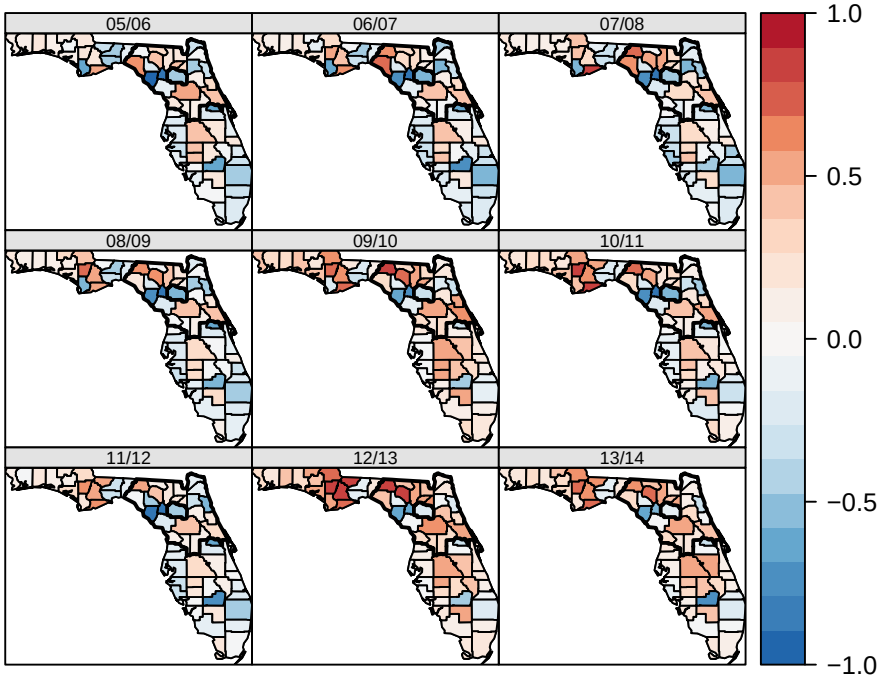
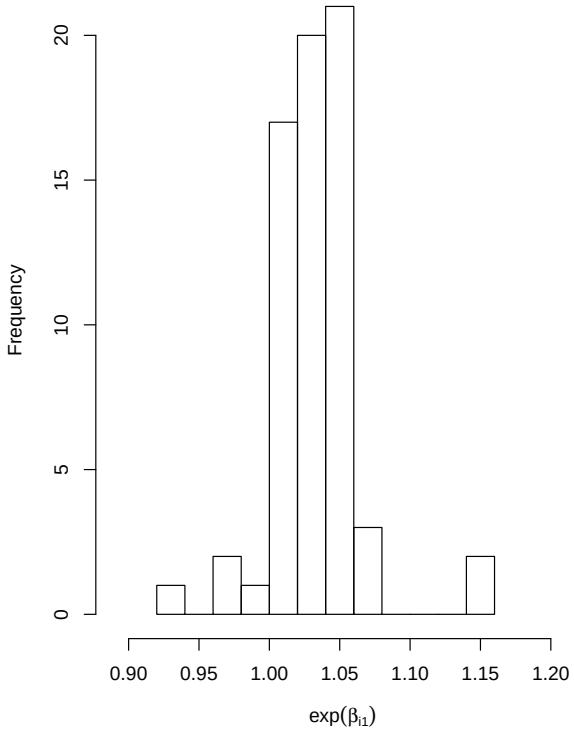
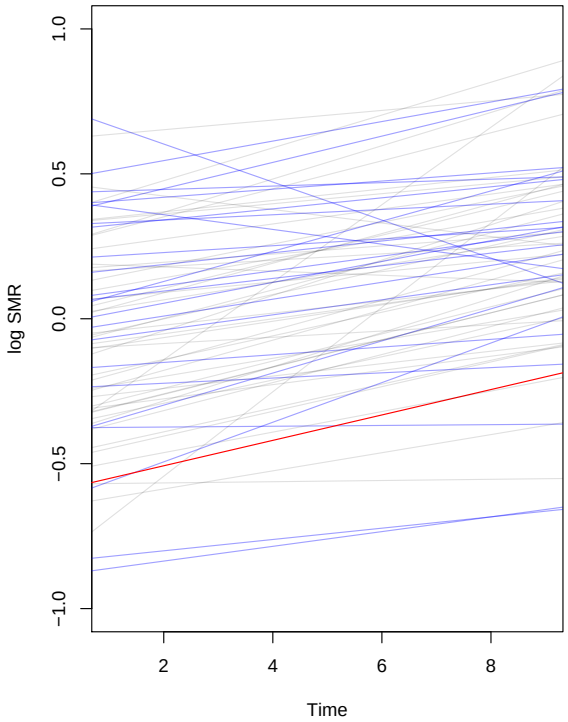


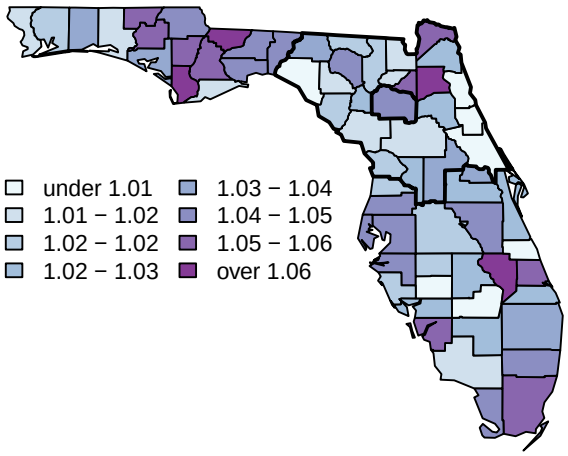
Figure 4.18: log SMRs for influenza-like illness by county and year.



(a) Histogram of exponentiated slopes.



(b) Fitted slopes on the log relative risk scale.



(c) Map of exponentiated slopes.

Figure 4.19: Results from area-specific log linear models. In (a), we summarize the area-specific slopes. In (b), the red line indicates Alachua county, regional fits are in blue, and the remainder of Florida in gray. (c) maps the exponentiated slopes.

For all hierarchical models we consider the first stage model for the number of cases in area i at time t as

$$Y_{it}|\theta_{it} \sim \text{Poisson}\left(E_{it}\theta_{it}\right),$$

where E_{it} are the expected counts, and θ_{it} is the relative risk of disease. A common Bayesian approach to modeling spatio-temporal data is to model the spatial effects as ICAR and independent of the non-parametric temporal effects (Knorr-Held and Besag, 1998). In what follows, we refer to this model, which does not include a space-time interaction, as the "main effects" model. The second stage of the main effects model fits

$$\log \theta_{it} = \beta_0 + \beta_1 t + \epsilon_i + S_i + \gamma_t, \tag{4.23}$$

where $\epsilon_i \sim \text{N}(0, \sigma_v^2)$ are spatially unstructured (iid) random effects, S_i are spatially structured ICAR random effects with precision τ_s, and $\gamma_t \sim \text{N}(0, \sigma_\gamma^2)$ are temporally unstructured random effects.

However, the results from simple space-time models suggest that we may need to include a spatio-temporal interaction term. We model the log relative risk as

$$\log \theta_{it} = \beta_0 + \epsilon_i + S_i + \gamma_t + \tau_t + \delta_{it}, \tag{4.24}$$

where $\epsilon_i \sim \text{N}(0, \sigma_v^2)$ are spatially unstructured (iid) random effects; S_i are spatially structured ICAR random effects with precision $1/\sigma_s^2$; $\gamma_t \sim \text{N}(0, \sigma_\gamma^2)$ are temporally unstructured random effects; τ_t are temporally structured second order random walk (RW2) with precision $1/\sigma_\tau^2$; and δ_{it} is a space-time interaction random effect. Following Knorr-Held (2000), we consider four priors for δ_{it}, which correspond to four kinds of possible structures for the space-time interaction:

1. Type I: To model independent space-time effects, we assume each δ_{it} is independent and identically distributed Gaussian with mean 0 and variance σ_δ^2.

2. Type II: The interaction between the temporally structured main effect and the unstructured spatial effects.

3. Type III: Assuming spatial trends differ over time, with no temporal structure, requires independent ICAR-type priors over time.

4. Type IV: The interaction between spatially and temporally structured main effects assumes that spatial patterns are similar in neighboring years.

We compare the main effects model to the four interaction models using the deviance information criteria, or DIC, the log normalizing constant, and WAIC (see Section 2.4 for details).

Table 4.6 summarizes the various methods of model comparisons. The main effects model has very large DIC and WAIC values, especially compared to models with spatio-temporal random effects. Including a space-time interaction dramatically improves the model fit, and an independent space-time interaction seems to provide the best fit, based on these measures. We see a similar trend for the WAIC when modeling simply the space-time effects.

	$\log p(y)$	p_D	$\bar{D}$	DIC	WAIC
Main Effects	-17,891.29	153.80	29,006.43	29,160.23	36,572.26
Type I	-4,918.70	585.53	6,943.31	**7,528.84**	**7,373.93**
Type II	-4,792.55	555.93	6,989.09	7,545.02	7,468.05
Type III	-4,917.26	579.84	6,950.29	7,530.12	7,388.33
Type IV	-4,705.82	544.04	7,001.98	7,546.02	7,490.29

Table 4.6: Model comparison: the deviance information criteria (DIC) is calculated using p_D is the effective degrees of freedom and the deviance evaluated at the posterior mean, $\bar{D}$

We also consider the contribution of the variance components for the various models. For the RW2 and ICAR models, the variance contribution is evaluated empirically, since the variance parameter is conditional rather than marginal. We summarize the variance components from the considered models in Table 4.7 and see that structured and unstructured

temporal effects explain at least 50% of the variation for all of the models, the structured spatial effects tend to explain a large proportion of the variation, too.

	RW2	ICAR	iid time	iid space	interaction
Main Effects	0.36	0.32	0.29	0.03	
Type I	0.31	0.27	0.24	0.02	0.17
Type II	0.29	0.18	0.24	0.17	0.12
Type III	0.26	0.39	0.24	0.03	0.07
Type IV	0.33	0.19	0.26	0.17	0.06

Table 4.7: Summaries of variance components. For the RW2 and ICAR models, the contribution is evaluated empirically, since the variance parameter is conditional rather than marginal.

Here, we have only explored the spatio-temporal structure for ILI. As we can see, modeling the SLIV program over the 9 years of data is a complex problem. Adjusting for the negative control in models with complex space-time interactions would require considerable care, and extensive simulations to understand the situations in which such an approach is beneficial. With the available AHCA data, complex spatio-temporal analyses are not likely to offer much insight into the nature of the SLIV program. We leave as future work the task of understanding how to best model spatial, temporal, and spatio-temporal components for two outcomes while still obtaining reasonable estimates.

4.9 Discussion

In the presence of unmeasured confounding, unadjusted estimates can be biased; the direction of this bias depends on the direction of differences in unmeasured confounding across study populations. As we have seen, the negative control framework is a potentially powerful tool to detect and adjust for unmeasured confounding. However, the performance of the adjusted models depends heavily on the choice of negative control. If both the outcome of interest and the negative control are subject to a similar magnitude of confounding, then adjusting for the negative control will result in less biased estimates. Unbiased estimates are possible

when the unmeasured confounding is equal across outcomes. However, when the unmeasured confounding in the negative control differs from that in the outcome of interest, introducing the negative control can produce severely biased estimates more extreme than those from an unadjusted analysis. The direction of the bias in the adjustment will again depend on the nature of the unmeasured confounding across areas and outcomes.

Of course, there is nothing in the data that can help us determine the quality of the negative control outcome. The confounding we are adjusting for is unmeasured and cannot be assessed or accounted for in the models. Thus, the negative control framework relies on the untestable assumption that the nature of the unmeasured confounding is similar in magnitude across the study populations and the two outcomes.

In practice, perfect negative controls will generally be hard to find, as there can be a variety of sources of unmeasured confounding, especially with surveillance data. Nevertheless, if there is a known significant source of unmeasured confounding, a good negative control outcome could remove a major source of bias. For the AHCA analysis, we assume that differences in health care seeking behavior given ILI is the primary source of unmeasured confounding. However, it is possible that there are other sources of bias remaining in our analyses that we have not accounted for, such as differences in diagnosis across for ILI-specific ICD-9 codes. While we find it reasonable to assume that GI suffers from the same sources of confounding as ILI in the AHCA data, another reader may disagree. Both the unadjusted and adjusted results should be presented in any analysis that employs the negative control outcome.

It is possible that using additional information from an external data source could help avoid the strong assumptions of the negative control approach. For example, area-level data about the factors that influence healthcare seeking behavior, such as socio-economic indicators, could be included in standard models to adjust for ascertainment differences across counties. Individual level data, or further information about the healthcare seeking behavior of cases (or a subsample of cases) may also allow us to more directly control for the bias in the AHCA data for studying influenza.

Chapter 5

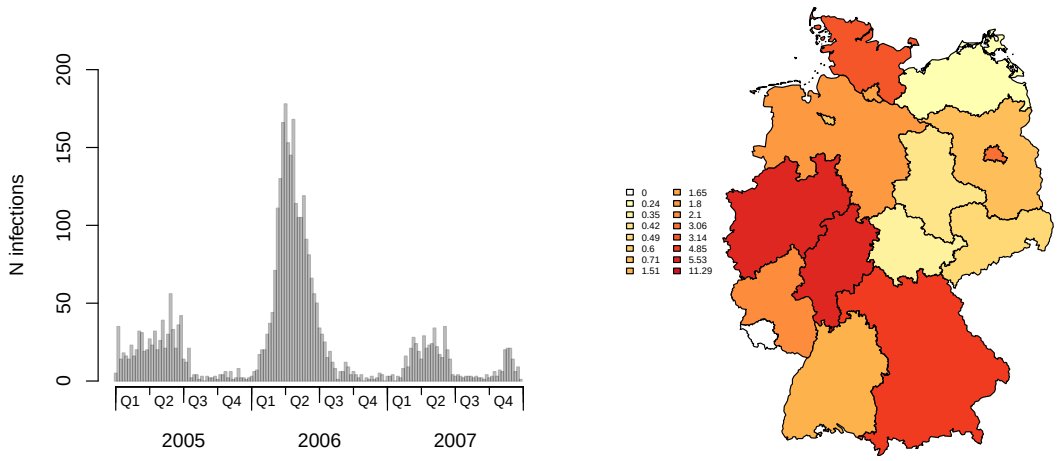
ECOLOGICAL INFERENCE WITH SURVEILLANCE DATA

5.1 Introduction

Surveillance data is commonly aggregated over space and time. When there is interest in studying covariate effects on the spread of infectious disease with aggregated data, disease mapping approaches, such as ecological regression, ignore the dependent nature of infectious diseases and are typically not considered. Current models for aggregated infectious disease data are difficult to interpret and prone to ecological bias. In this chapter, we develop an ecological infectious disease model for a partially vaccinated population that provides estimates for familiar epidemiological parameters by starting with an individual-level model to derive an ecologically consistent model for infectious diseases in partially vaccinated populations.

We consider surveillance data about measles outbreaks in Germany in partially vaccinated populations, which was previously introduced in Section 1.1.3. In Figure 5.1(a), we plot weekly counts of measles cases for all of Germany. In 2006, there seems to be a large increase in the total number of weekly measles cases early in the year. Overall, the total number of cases in a given week is fairly small, especially in a population of over 82 million people. In Figure 5.1(b), we plot the number of cases per 100,000 residents for the three years of study data. The distribution of cases is not uniform across the 16 states of Germany. In general, states that formally made up East Germany have a lower incidence of measles compared to the states of West Germany.

The remainder of this chapter is organized as follows: in Section 5.2, we introduce the key epidemiological parameters we are interested in estimating. We then discuss aggregate models for infectious disease data in Section 5.3. In Section 5.4, we discuss the risk of ecological bias in aggregated infectious disease models. In Section 5.5, we describe a previous



(a) Weekly measles cases over time.

(b) Measles cases per 100,000 residents.

Figure 5.1: Summary of measles cases over time and space.

approach for including vaccination in the aggregate infectious disease model. In Section 5.6, we develop an individual-level model for infectious diseases and consider the implied ecological models in partially vaccinated populations. We examine, via simulation, the assumptions and behavior of the ecological vaccine models in Section 5.7. Finally, we apply these methods to the measles data in Section 5.8 and compare the results to current approaches.

5.2 Epidemiological parameters of interest

5.2.1 Basic reproductive number, R_0

A key parameter for quantifying infectious diseases is the basic reproductive number, represented by R_0 (Keeling and Rohani, 2008). This quantity is defined as the average number of individuals a typical infectious individual would infect in a completely susceptible population. When R_0 is greater than one, a single infectious individual will, on average, infect at least one other individual before recovering resulting (in expectation) in a major outbreak. Due to stochastic effects, only minor outbreaks can also occur when R_0 is greater than one. When R_0 is less than one, meaning that on average a single infected individual will infect

fewer than one person, minor outbreaks are possible. Importantly, R_0 is not only determined by the disease, but is also influenced by population characteristics such as density, and mobility. Therefore, R_0 is a parameter defined with respect to the disease in a specific population at some point in time. When a portion of the population is immune, either because of vaccination or previous infection, the average number of new infections caused by a single infectious is called the *effective reproductive number*, represented by R. When x is the proportion of the population that is immune to infection, $R = (1 - x)R_0$. As with R_0, when the effective reproductive number R is less than 1, a major outbreak can be prevented.

The *critical vaccination coverage*, denoted v_c, is the minimum fraction of the population that needs to be vaccinated to prevent a major outbreak. With a perfect vaccine, where all vaccine recipients have 100% immunity, $v_c = 1 - 1/R_0$, and vaccination coverage at least as large as v_c will provide protection (in expectation) from a major outbreak in the population. If the vaccine is not perfect and the risk of infection for a vaccine recipient is $(1 - \phi)$, then the critical vaccination coverage is

$$v_c = \left(1 - \frac{1}{R_0}\right)\frac{1}{\phi}.$$

Major epidemics can be avoided with high probability if more than v_c of the population is vaccinated. In other words, an entire population can be protected from a major outbreak when at least v_c of the population is vaccinated (Britton, 2010).

This phenomenon is referred to as herd or community immunity, and it reflects the indirect effects of vaccination (Halloran et al., 2010). Specifically, when portions of a community are vaccinated and the number of susceptible individuals are reduced, disease transmission is also reduced. As a result, an unvaccinated individual in the partially vaccinated population gains some level of protection even though they have no individual-level protection from infection.

5.2.2 Modeling the force of infection

Following Halloran et al. (2010), the force of infection (hazard), for a susceptible individual is

$$\lambda_t^\dagger = c(N) \times \frac{y_{t-1}}{N} \times p, \tag{5.1}$$

where $c(N)$ is the contact rate for individuals in a population of size N, the prevalence of infectives in the population is y_{t-1}/N, and p is the per-contact probability of infection, or the transmission probability. Hence, the hazard rate $\lambda_t^\dagger$ determines the time until infection for an individual who is susceptible at time $t-1$ in a population with y_{t-1} infectives.

The contact rate $c(N)$ is frequently modeled as density dependent or frequency dependent (Keeling and Rohani, 2008). Frequency dependent transmission, or mass action, assumes that the contact rate is independent of the population size, i.e. $c(N) = c^{\text{FD}}$. In contrast, density dependent transmission assumes that the contact rate is proportional to the size of the population, so that $c(N) = c^{\text{DD}}N$.

The hazard in (5.1) only reflects the risk for an individual becoming infected from direct contact with an infected individual in that population. For many diseases, there are multiple ways to become infected, however. When there are competing sources of transmission, we consider additive hazards,

$$\lambda_t^\dagger = \lambda_t^{\text{AR}} + \lambda_t^{\text{NE}} + \lambda_t^{\text{EN}},$$

where λ_t^{AR}, λ_t^{NE}, and λ_t^{EN} correspond to risk from autoregressive, neighborhood, and endemic sources, respectively. The models described in Section 5.3 are of this form.

5.3 Aggregate models for infectious disease data

Held et al. (2005) proposed an epidemic-endemic framework to model aggregated disease counts over time and space. The flexible framework has been extended by Paul et al. (2008), Paul and Held (2011), and Meyer and Held (2014), and models are implemented in the `surveillance` package in `R` (Meyer et al., 2016). The Held framework is motivated by spatial branching processes, and is closely related to standard SIR and multivariate time series SIR models (Xia et al., 2004). Bauer and Wakefield (2016) extend the Held framework to handle a stratified population, and in doing so provide a novel derivation to the model by aggregating the individual level model. We briefly describe the derivation from the individual-level to the aggregated model before discussing the specifics of the Held framework.

Let Y_{it} be the number of cases in area i and time t, and N_{it} be the population size for area i and time t. Time here is relative to the disease, meaning that we are assuming the sum of incubation and infectious times is approximately that of the observation times. For the individual level model, the probability of a susceptible individual in area i and time $t-1$ becoming infected by time t is determined by the hazard rate $\lambda_{it}^{\dagger}$. Assuming a constant hazard between time steps, for a susceptible in area i at time $t-1$, the probability of remaining susceptible at time t is given by the survivor function

$$\Pr\left(\text{escaping infection in } (t-1,t] \,|\, \text{no infection by } t-1,\ \text{area } i\right) = e^{-\lambda_{it}^{\dagger}}.$$

The probability an individual who is susceptible at time $t-1$ is infected by time t is

$$\Pr\left(\text{infection in } (t-1,t] \,|\, \text{no infection by } t-1,\ \text{area } i\right) = 1 - e^{-\lambda_{it}^{\dagger}}.$$

Assuming the time until infection is independent for all susceptible individuals, we get a Reed-Frost chain binomial SIR, and the number of new infectives in area i at time t is

$$Y_{it} | \lambda_{it}^{\dagger} \sim \text{Binomial}\left(S_{i,t-1}, 1 - e^{-\lambda_{it}^{\dagger}}\right).$$

When $\lambda_{it}^{\dagger}$ is small, the Taylor expansion

$$1 - \exp(-\lambda_{it}^{\dagger}) \approx \lambda_{it}^{\dagger},$$

simplifies the form of the probability of infection. When the number of susceptibles, $S_{i,t-1}$ is large, and the probability of infection is small, the binomial distribution can be approximated by a Poisson distribution so that

$$Y_{it}|\lambda_{it}^{\dagger} \sim \text{Poisson}\left(S_{i,t-1}\lambda_{it}^{\dagger}\right).$$

Additionally, when the number of new infections is small, the number of susceptibles can be approximated by the initial number of susceptibles. In a completely unvaccinated population, the initial number of susceptible individuals is the total population.

The Held framework assumes that for rare diseases, the number of cases in area i and time t is modeled by a Poisson random variable. Specifically, Held models $Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it})$, where the mean, μ_{it}, can be decomposed into three components: autoregressive (AR), neighborhood (NE), and endemic (EN). Mathematically, the mean has the following form

$$\mu_{it} = \underbrace{\lambda_{it}^{\text{AR}}\, y_{i,t-1}}_{\text{Autoregressive}} + \underbrace{\lambda_{it}^{\text{NE}} \sum_{j \neq i} w_{ji}\, y_{j,t-1}}_{\text{Neighborhood}} + \underbrace{\lambda_{it}^{\text{EN}} N_{it}}_{\text{Endemic}}. \tag{5.2}$$

To allow for over-dispersion, cases can also be modeled with the negative binomial distribution although we do not pursue that here. The autoregressive component accounts for the disease risk from the number of counts in the previous time period in the same area. The neighborhood component reflects the additional risk from infected individuals in neighboring areas, where the neighbors are defined though the weights w_{ji}. The parameters λ_{it}^{AR} and λ_{it}^{NE} are rates and determine the relative contributions of cases in the same and in neighboring regions. The endemic component describes additional risk that is not accounted for in the autoregressive or neighborhood components. Each component can be modeled with a log-

linear model to include covariates as well as fixed and random effects. For example, the autoregressive component may take the form

$$\log \lambda_{it}^{\mathrm{AR}} = \alpha_{\mathrm{AR}} + a_i + \boldsymbol{\beta}^{\mathrm{AR}} \boldsymbol{x}_{it}, \tag{5.3}$$

where α_{AR} is an overall log-risk, a_i are area-specific fixed (or random) effects, $\boldsymbol{x}_{it}$ are area- and time-specific covariates, and $\boldsymbol{\beta}^{\mathrm{AR}}$ are the associated covariate effects. The neighborhood and endemic components can be modeled in a similar fashion. In practice, the endemic component tends to account for environmental reservoirs that contribute to the risk of infection. For seasonal diseases, the endemic component is modeled to account for the seasonality via

$$\log \lambda_{it}^{\mathrm{EN}} = \alpha_{\mathrm{EN}} + \alpha_i + \sum_{s=1}^{S} \left[\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)\right],$$

where S is the number of pairs of sines and cosines to include and ω_s are Fourier frequencies. For weekly data, $\omega_s = 2\pi s/52$. In the `surveillance` package, parameter estimates are quickly obtained for the Held models via penalized maximum likelihood estimation. While these models are relatively easy to implement, parameter interpretation is cumbersome. In (5.3), for a single covariate x_{it}, the corresponding $e^{\beta^{\mathrm{AR}}}$ is interpreted as the ratio of autoregressive rates that corresponds to a one-unit increase in x_{it}.

In the subsequent sections, we develop an ecological model that accounts for the binary vaccination status covariate that provides familiar epidemiological parameters associated with infectious disease models. We start with an individual level model for a partially vaccinated population and derive an ecological model that can be easily fit with surveillance data and provides parameters which are easy to interpret.

5.4 Ecological bias

Before turning our attention to modeling a vaccination coverage in infectious disease models for aggregated data, we first consider the risk of ecological bias when modeling aggregated infectious disease data. In the social sciences and non-infectious disease epidemiology, the risk of drawing erroneous individual-level conclusions from group-level data has been well characterized (Selvin, 1958; Robinson, 1950; Greenland, 1992; Greenland and Robins, 1994; Richardson and Monfort, 2000; Wakefield, 2008). This phenomenon is referred to as ecological bias and occurs as a result of the within-area variability of exposures and confounders. When the individual-level risk of disease is nonlinear, the form of the marginal aggregate risk changes and results in so-called pure specification bias. As we have seen from our discussion of compartmental models in Section 2.1.1, infectious disease risks are very nonlinear and are susceptible to ecological bias. Aside from Koopman and Longini (1994), there has been little discussion of ecological bias for aggregate infectious disease models.

To better understand ecological bias in the infectious disease setting, we consider a simple autoregressive model and start with an individual level disease model. Let Y_{itj} be the disease indicator for susceptible individual j in week t and area i and x_{itj} a covariate of interest. Then, under the approximations discussed in Section 5.3,

$$Y_{itj}|x_{itj} \sim \text{Bernoulli}(\lambda^{\dagger}_{itj}).$$

Recall, the force of infection for the purely autoregressive model can be written as the product of three quantities

$$\lambda^{\dagger}_{itj} = c_{itj}(N)p_{ij}Y_{i,t-1}/N_i,$$

where $c_{itj}(N)$ is the individual contact rate, p_{ij} is the individual's per-contact transmission probability, and $Y_{i,t-1}/N_i$ is the within-area disease prevalence. From this point onward, we assume a frequency dependent contact rate, so that $c_{itj}(N) = c^{\text{FD}}_{itj}$ and we let $\lambda^{\text{AR}}_{itj} = c_{itj}(N)p_{ij}$

denote the individual-level portion of the force of infection, so that

$$Y_{itj}|x_{itj} \sim \text{Bernoulli}(\lambda_{itj}^{\text{AR}} Y_{i,t-1}/N_i).$$

For an infectious disease, we can imagine that exposure to certain variables will influence an individual's contact rate, or transmission probability. For example, cold weather may reduce an individual's contact rate, or an underlying illness may increase the individual's transmission rate. In other words, for some univariate exposure of interest x_{itj}, we could re-write the individual's risk of infection as

$$\lambda_{itj}^{\text{AR}} = e^{\alpha_0} f(\alpha_1, x_{itj}), \tag{5.4}$$

where $f(\alpha_1, x)$ describes the relationship between the variable x and the product of the contact rate and transmission probability. In some ecological settings we only observe the area-level aggregate data, such as total disease counts,

$$Y_{it} = \sum_{j=1}^{N_i} Y_{itj},$$

and average exposure,

$$\bar{x}_{it} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{itj}.$$

We assume that exposures within the same area follow an area-level distribution, $g_{it}(x|\boldsymbol{\omega}_{it})$, where $\boldsymbol{\omega}_{it}$ are area- and week-level parameters for that distribution. Then the individual-level model in (5.4) implies an aggregate risk for week t and area i of

$$\overline{\lambda_{it}^{\text{AR}}} = e^{\alpha_0} \int_{A_i} f(\alpha_1, x) g_{it}(x|\boldsymbol{\omega}_{it}) dx, \tag{5.5}$$

assuming a continuous exposure, and where A_i represents area i. For concreteness, suppose

that $f(\alpha_1, x) = \exp(\alpha_1 x)$, so that

$$\lambda_{itj}^{\text{AR}} = \exp\left(\alpha_0 + \alpha_1 x_{itj}\right). \tag{5.6}$$

In this scenario, if we assume that within each area, the distribution of the exposures is normally distributed, i.e. $x_{itj}|\bar{x}_{it},\, \sigma_{it}^2 \sim \text{Normal}\left(\bar{x}_{it}, \sigma_{it}^2\right)$, then the marginal area-level risk is

$$\text{E}\left[Y_{it}|Y_{i,t-1}, \bar{x}_{it}\right] = \exp\left(\alpha_0 + \alpha_1 \bar{x}_{it} + \alpha_1^2 \sigma_{it}^2/2\right)\, Y_{i,t-1}. \tag{5.7}$$

Thus the implied aggregate risk is a function of both the average exposure and the variability of the exposure within an area. For further details in a non-infectious disease setting see Richardson et al. (1987); Plummer and Clayton (1996). Moreover, the individual-level association α_1 could be distorted in either direction, depending on the relationship between $\bar{x}_{it}$ and σ_{it}^2.

However, if we started with the aggregate-level data, we would likely fit the naïve ecological model

$$\text{E}\left[Y_{it}|Y_{i,t-1}, \bar{x}_{it}\right] = \exp(\beta_0 + \beta_1 \bar{x}_i) Y_{i,t-1}, \tag{5.8}$$

where $\exp(\beta_1)$ is the relative risk associated with a one unit increase in $\bar{x}_i$. The correct interpretation of $\exp(\beta_1)$ is estimating the *contextual* effect, that is the effect of the group average on individual risk. When either the mean and variance are independent or when there is no within-area variability of exposures, $\sigma_{it}^2 = 0$ for all areas i and weeks t, the ecological model of (5.8) is identical to the implied aggregate model of (5.7). Models for the within-area distribution of the x_{itj}'s that extend beyond normality require all of the higher moments to be independent of the mean. However, without additional information about the within-area variability of the exposures, the ecological model will not be estimating individual-level parameters.

For a discrete exposure x_k, with K levels, the aggregate risk is

$$\overline{\lambda_{it}^{\text{AR}}} = e^{\alpha_0} \sum_{k=1}^{K} f(\alpha_1, x_k) g_{it}(x_k|\boldsymbol{\omega}_{it}).$$

We again let $f(\alpha_1, x) = \exp(\alpha_1 x)$, and now assume a binary exposure, x_{itj}. Let $N_{it1} = \sum_{A_i} x_{itj}$ be the number exposed in area i and time t. With a binary exposure, the implied aggregate risk is

$$\overline{\lambda_{it}^{\text{AR}}} = e^{\alpha_0} \left[\left(1 - \frac{N_{it1}}{N_i} \right) + \frac{N_{it1}}{N_i} e^{\alpha_1} \right] N_i.$$

Then, the marginal area-level model is

$$\text{E}\left[Y_{it}|Y_{i,t-1}, \bar{x}_{it}\right] = \exp\left(\alpha_0\right) \left[1 - \bar{x}_{it} + \bar{x}_{it} \exp(\alpha_1)\right] Y_{i,t-1}, \tag{5.9}$$

where $\bar{x}_{it} = N_{it1}/N_i$ is the proportion of exposed individuals. Again, we see that the implied aggregated model of (5.9), which is estimating individual level effects, differs from the naïve ecological model in (5.8). The two approaches are estimating different parameters. The naïve ecological model estimates the risk associated with the average exposure, while the implied aggregate model estimates the average of individual risks.

In the non-infectious disease setting, it is well understood that when data are aggregated to the group level, individual-level associations can become distorted, leading to ecological bias. In the infectious disease setting the individual-level risk of disease is very non-linear, and therefore we expect naïve aggregate models to suffer from ecological bias. We now turn our attention towards modeling aggregated infectious diseases with vaccination coverage.

5.5 Ecological models with vaccine coverage

Herzog et al. (2011) considered models to estimate the multiplicative difference in incidence due to difference in vaccination coverage. The area-level vaccination coverage was included in either the endemic or auto-regressive part of the model. In their analysis of measles in Germany, Herzog et al. (2011) found that including the log proportion of unvaccinated individuals in the epidemic component gives the best fit, in terms of AIC, compared to including the coverage covariate in the endemic component. Specifically, for area-level coverage estimates x_i, the Herzog et al. (2011) model with the best fit is specified as follows:

$$\begin{aligned} Y_{it}|\mu_{it} &\sim \text{Poisson}(\mu_{it}), \\ \mu_{it} &= e^{\alpha_0}\,(1-x_i)^{\alpha_1}\,y_{i,t-1} + N_{it}\nu_{it}, \\ \log \nu_{it} &= \beta_0 + \gamma \sin(2\pi t/26) + \delta \cos(2\pi t/26). \end{aligned}$$

The log proportion of the unvaccinated individuals can be thought of as a proxy for the number of susceptibles in the population. As such, α_1 cam be thought of as a flexibility parameter to improve model fit. Herzog et al. (2011) do not interpret the parameter associated with vaccination coverage.

Meyer et al. (2016) applied the same approach as Herzog et al. (2011) to including vaccination coverage in the model for a different measles example dataset, and provide more discussion of the parameter associated with vaccination coverage. In their review article, Meyer et al. (2016) determined that the model which included vaccination coverage in the endemic component gave the best fit, in terms of AIC, for their example. Specifically,

$$\begin{aligned} Y_{it}|\mu_{it} &\sim \text{Poisson}(\mu_{it}), \\ \mu_{it} &= e^{\alpha_0}\,y_{i,t-1} + N_{it}\,(1-x_i)^{\alpha_1}\nu_{it}, \\ \log \nu_{it} &= \beta_0 + \alpha_1 \log(1-x_i) + \gamma \sin(2\pi t/26) + \delta \cos(2\pi t/26), \end{aligned}$$

provided the best model fit. Meyer et al. (2016) interpret α_1, the parameter associated with

$\log(1-x_i)$ in the endemic component, as follows: a doubling of the proportion of susceptibles in area i corresponds to an expected multiplicative change in the endemic incidence of 2^{α_1}. While the approach taken by Herzog et al. (2011) and Meyer et al. (2016) perform reasonably in terms of model fit, the parameter interpretation is unclear, making them unsuitable models for inference.

5.6 Ecological vaccine model development

5.6.1 Introduction and notation

To properly model the effects of vaccination, at the population level, it is important to consider how the vaccine reduces an individual's risk of infection. We consider the aggregate models for two modes of vaccine action: leaky and all-or-none. Leaky vaccines are assumed to reduce the risk of infection by a constant proportion for all vaccinated individuals; in contrast, "all-or-none" vaccines provide full protection from infection to vaccinated individuals when successful, but fail to provide protection with some probability (Halloran et al., 2010, Chapter 7). In other words, leaky vaccines reduce the per-exposure risk of infection, while an all-or-none vaccine's protection is independent of the number of contacts made.

We assume a single area with a fixed population of size N. Let S_t denote the number of susceptibles at time t, and Y_t the total number of infectives at time t. We let x be the proportion of the population who are vaccinated, which is assumed to be constant over time. We let $\lambda_t^\dagger$ denote the force of infection, or the risk of an individual who was susceptible at time $t-1$ becoming infected by time t. We further assume that all infectives at a given time are recovered at the subsequent time point. For clarity, we develop the ecological model with a generic force of infection. In Section 5.2.2 we discussed different approaches to modeling the force of infection. We assume that the vaccine only affects an individual's susceptibility to infection (and not infectiousness or disease progression) and that vaccination provides lifetime immunity. And we let ϕ be the reduction in a vaccine recipient's risk of infection (the vaccine effect) after a single vaccination.

5.6.2 All-or-none vaccine ecological model

We consider the group level model and the resulting ecological model with an all-or-none vaccine. For an all-or-none vaccine, a vaccinated individual remains susceptible to infection because the vaccine fails with probability $(1-\phi)$ (Halloran et al., 2010, Chapter 7). When the vaccine fails to take, we would expect the vaccinated susceptibles' risk of infection to be the same as an unvaccinated susceptible. In a partially vaccinated population of size N, the $(1-x)N$ unvaccinated individuals will certainly remain susceptible to infection. Assuming an all-or-non vaccine means that $(1-\phi)xN$ of the vaccinated individuals are still at risk of infection due to vaccine failure. We denote the number of susceptibles at time t by $S_t(\phi)$ to emphasize that the number of susceptibles is a function of the vaccine effect. Hence, the number of susceptibles at time $t=0$ is

$$S_0(\phi) = (1-x)N + (1-\phi)xN = (1-\phi x)N.$$

The number of new infections at time $t+1$ can be modeled as

$$Y_{t+1} \,|\, \lambda_t^\dagger \sim \text{Binomial}\left(S_t(\phi),\, 1-\exp(-\lambda_t^\dagger)\right), \tag{5.10}$$

where $S_t(\phi) = S_{t-1}(\phi) - Y_t$, $t = 1, \ldots, T$, and $\lambda_t^\dagger$ is a generic force of infection. In the rare disease setting, the binomial can be approximated by a Poisson and when $\lambda_t^\dagger$ is small, a Taylor expansion approximates $1-\exp(-\lambda_t^\dagger) \approx \lambda_t^\dagger$ so that

$$Y_{t+1} \,|\, \lambda_t^\dagger \sim \text{Poisson}\left(S_t(\phi)\,\lambda_t^\dagger\right), \tag{5.11}$$

where $S_t(\phi) = (1-\phi x)N - \sum_{k=1}^{t} Y_k$. When the susceptible population is sufficiently large, and the number of cases is small, the number of susceptibles can be approximated such that $S_t(\phi) \approx (1-\phi x)N$. The ecological model in (5.10) can be approximated by

$$Y_{t+1} \,|\, \lambda_t^\dagger \sim \text{Poisson}\left(\lambda_t^\dagger(1-\phi x)N\right), \tag{5.12}$$

when the approximations are valid. In Section 5.7.1, we consider the conditions under which these modeling assumptions can be made.

5.6.3 Leaky vaccine ecological model

A bit more notation is needed in the leaky vaccine setting, since a leaky vaccine reduces the risk of infection for all recipients. We let S_{ut} and S_{vt} denote the number of unvaccinated and vaccinated susceptibles at time t, respectively; and let Y_{ut} and Y_{vt} be the total number of unvaccinated and vaccinated infectives at time t, such that $Y_t = Y_{ut} + Y_{vt}$. Since the entire population is susceptible to infection, we can assume that $S_{u0} = (1-x)N$ and $S_{v0} = xN$. The assumption of a leaky vaccine allows us to write the hazard for the vaccinated population as a function of that of the unvaccinated population and the vaccine effect:

$$\lambda_{vt}^{\dagger} = (1-\phi)\lambda_{ut}^{\dagger}. \tag{5.13}$$

Then, the number of new infections at time $t+1$ can be modeled as

$$Y_{u,t+1} \,|\, \lambda_{ut}^{\dagger} \sim \text{Binomial}\Big(S_{ut}, 1-\exp(-\lambda_{ut}^{\dagger})\Big), \tag{5.14}$$

$$Y_{v,t+1} \,|\, \lambda_{vt}^{\dagger} \sim \text{Binomial}\Big(S_{vt}, 1-\exp(-\lambda_{vt}^{\dagger})\Big), \tag{5.15}$$

where $\lambda_{ut}^{\dagger}$ is the force of infection for an unvaccinated susceptible at time t, and $\lambda_{vt}^{\dagger}$ is defined in (5.13); the number of susceptibles at time $t+1$ are

$$S_{u,t+1} = S_{u,t} - Y_{u,t+1} \qquad \text{and} \qquad S_{v,t+1} = S_{v,t} - Y_{v,t+1}.$$

The resulting aggregate model is a convolution of binomials, where

$$\Pr\left(Y_t = y \,\middle|\, \lambda_{ut}^{\dagger},\, \lambda_{vt}^{\dagger}\right) = \sum_{z=0}^{y} \Pr\left(Y_{ut} = z \,\middle|\, \lambda_{ut}^{\dagger}\right) \Pr\left(Y_{vt} = y - z \,\middle|\, \lambda_{vt}^{\dagger}\right). \tag{5.16}$$

In settings where populations or disease counts are large, the aggregate model will be computationally expensive and potentially intractable.

When $\lambda_{ut}^{\dagger}$ is small, the Taylor approximation

$$1 - \exp(-\lambda_{ut}^{\dagger}) \approx \lambda_{ut}^{\dagger},$$

simplifies the probability of infection in equation (5.14). An analogous approximation can be made for $1-\exp(-\lambda_{vt}^{\dagger})$. Moreover, when cases are rare, the binomial distributions can be approximated by Poissons. Hence, when risk of infection is small for both the unvaccinated and vaccinated populations, the number of new infections in each group is approximately

$$Y_{u,t+1}|\lambda_{ut}^{\dagger} \sim \text{Poisson}\Big(S_{ut}\,\lambda_{ut}^{\dagger}\Big), \tag{5.17}$$

$$Y_{v,t+1}|\lambda_{ut}^{\dagger},\,\phi \sim \text{Poisson}\Big(S_{vt}\,(1-\phi)\lambda_{ut}^{\dagger}\Big). \tag{5.18}$$

Since the sum of Poisson random variables is also Poisson, the resulting aggregate model, when the risk is small for both vaccinated and unvaccinated groups is

$$Y_{t+1} \,\big|\, \lambda_{ut}^{\dagger},\,\phi \sim \text{Poisson}\left(\big(S_{ut} + S_{vt}(1-\phi)\big)\lambda_{ut}^{\dagger}\right), \tag{5.19}$$

is certainly a tractable likelihood, especially compared to the convolution model of (5.16). However, this model still requires knowing the number of susceptibles by vaccination status. If it is reasonable to assume that the number of infectives is negligible, i.e. $S_{ut} \approx S_{u0}$ and $S_{vt} \approx S_{v0}$, the ecological model for a partially vaccinated population is approximately

$$Y_{t+1} \,\big|\, \lambda_{ut}^{\dagger},\,\phi \sim \text{Poisson}\left(\lambda_{ut}^{\dagger}(1-\phi x)N\right), \tag{5.20}$$

which is identical to (5.12), the ecological model derived assuming an all-or-none vaccine.

5.6.4 Comments on the ecological vaccine model

We summarize the development of the ecological vaccine model starting from the all-or-none and leaky vaccine assumptions, as well as the simplifying assumptions that result in the ecological vaccine model in Table 5.1. Both the all-or-none and leaky vaccine models can be approximated by the ecological vaccine model when the following simplifying assumptions can be made:

1. Poisson approximation to the binomial distribution

2. Force of infection approximation: $1 - e^{-\lambda_t^\dagger} \approx \lambda_t^\dagger$

3. Negligible number of infections: $S_{ut} \approx S_{u0}$ for unvaccinated individuals, and $S_{vt} \approx S_{v0}$ for vaccinated individuals. Note that the number of susceptibles may also be a function of the vaccine effect.

This list of assumptions helps illuminate when the ecological vaccine model we have developed is appropriate to use. When there is a major outbreak and the progression of cases is limited by the number of remaining susceptibles, the number of infections will not be negligible; we would not expect the ecological vaccine model to perform well in this scenario.

The ecological models developed in the previous sections were done using a generic force of infection, $\lambda_t^\dagger$, but in practice we will consider an additive hazard model, like that described in Section 5.2.2.

In the next section, we conduct simulations to better understand when the approximations are appropriate in the infectious disease setting and in the absence of vaccination. We first use maximum likelihood estimates (MLE's) to examine the basic attributes of these models in simple scenarios. We then develop a hierarchical model to accommodate more complex forms of the hazard components. We use Stan and Hamiltonian Monte Carlo (HMC) to fit these more complex ecological models (see Section 2.3.2 for details).

	All-or-none	**Leaky**
Initial susceptible population		
	$S_{u0}(\phi) = (1-x)N$	$S_{u0} = (1-x)N$
	$S_{v0}(\phi) = (1-\phi)xN$	$S_{v0} = xN$
Force of infection		
	$\lambda^\dagger_{ut} = \lambda^\dagger_t, \quad \lambda^\dagger_{vt} = \lambda^\dagger_t$	$\lambda^\dagger_{ut} = \lambda^\dagger_t\ , \quad \lambda^\dagger_{vt} = (1-\phi)\lambda^\dagger_t$
Progression		
$Y_{u,t+1} \mid \lambda^\dagger_{ut}$	$\text{Bin}\Big(S_{ut}(\phi), 1-e^{-\lambda^\dagger_t}\Big)$	$\text{Bin}\Big(S_{ut}, 1-e^{-\lambda^\dagger_t}\Big)$
$Y_{v,t+1} \mid \lambda^\dagger_{vt}$	$\text{Bin}\Big(S_{vt}(\phi), 1-e^{-\lambda^\dagger_t}\Big)$	$\text{Bin}\Big(S_{vt}, 1-e^{-(1-\phi)\lambda^\dagger_t}\Big)$
Implied aggregate model		
$Y_{t+1} \mid \lambda^\dagger_t$	$\text{Bin}\Big(S_\text{t}(\phi), 1-e^{-\lambda^\dagger_t}\Big)$	Convolution of binomials
Simplifying assumptions		
Poissons approximate binomials		
	$\text{Poi}\Big(S_t(\phi)\big(1-e^{-\lambda^\dagger_t}\big)\Big)$	$\text{Poi}\Big(S_{ut}\big(1-e^{-\lambda^\dagger_t}\big) + S_{vt}\big(1-e^{-(1-\phi)\lambda^\dagger_t}\big)\Big)$
Taylor approximation		
$1-\exp(-\lambda^\dagger_t) \approx \lambda^\dagger_t$	$\text{Poi}\Big(S_t(\phi)\lambda^\dagger_t\Big)$	$\text{Poi}\Big(\big(S_{ut} + (1-\phi)S_{vt}\big)\lambda^\dagger_t\Big)$
Negligible number of infections		
	$S_t(\phi) \approx (1-\phi x)N$	$S_{ut} \approx (1-x)N, \quad S_{vt} \approx xN$
Ecological vaccine model		
	$Y_{t+1} \mid \lambda^\dagger_t, \phi \sim \text{Poisson}\Big(\lambda^\dagger_t(1-\phi x)N\Big)$	

Table 5.1: Summary of the all-or-none and leaky vaccine models and the assumptions for the ecological vaccine model. N is the total population; x denotes the proportion of the population vaccinated (assumed constant over time); ϕ is the vaccine effect on susceptibility; S_{ut} and S_{vt} denote the number of unvaccinated and vaccinated susceptibles at time t, respectively; Y_{ut} and Y_{vt} denote new cases in time t among unvaccinated and vaccinated, respectively; and $\lambda^\dagger_t$ is a generic force of infection.

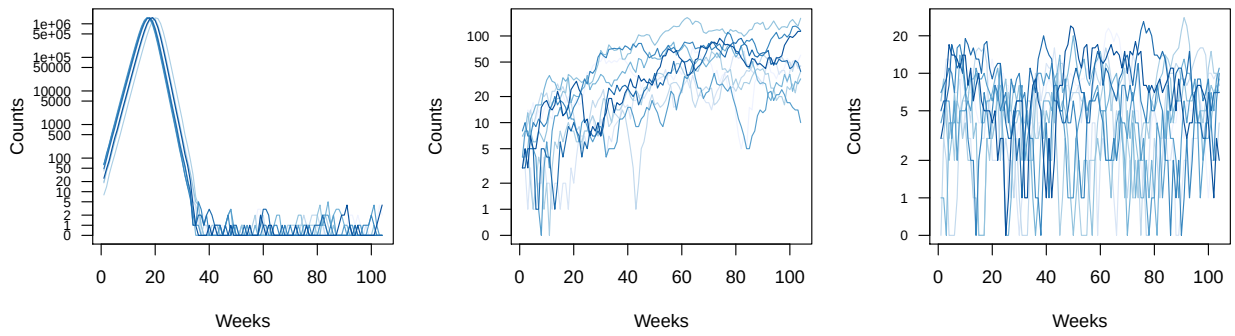
5.7 Simulations

5.7.1 Assesing the simplifying assumptions

In the previous sections, we started from two different modes of vaccine action and arrived at the same ecological model under certain simplifying assumptions. In this section, we conduct simulations to examine when those simplifications are appropriate, in the absence of vaccination. We consider a totally unvaccinated population of 10 million and two years of weekly disease count data. We start each simulation with 1 infected individual and allow the number of cases to grow stochastically for 4 weeks before collecting the two years of weekly observations. We simulate epidemics for a high, medium, and low values of R_0 following the binomial model

$$
\begin{aligned}
Y_{t+1}|\lambda_t^\dagger &\sim \text{Binomial}\left(S_t, 1 - e^{-\lambda_t^\dagger}\right), \\
\lambda_t^\dagger &= e^{\alpha_{\text{AR}}} Y_t/N + e^{\alpha_{\text{EN}}}, \\
S_t &= N - \sum_{k=1}^{t} Y_k.
\end{aligned}
$$

In Figure 5.2 we present example realizations of the disease counts over time for $R_0 = 2, 1.15$, and 0.85.



(a) $R_0 = 2$ (b) $R_0 = 1$ (c) $R_0 = 0.85$

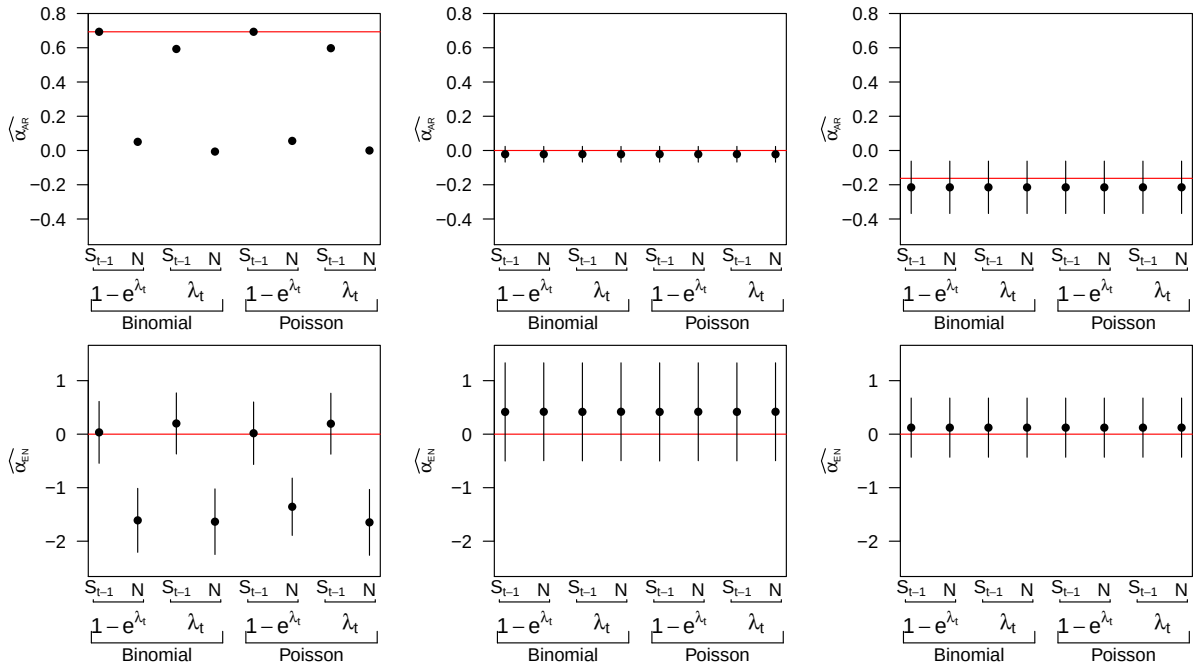
Figure 5.2: Ten simulated epidemic curves for three values of R_0 in the absence of vaccination.

The three simplifying assumptions necessary for the ecological vaccine model give us eight possible models to compare for each simulated epidemic. We model the force of infection as $\lambda_t^\dagger = e^{\alpha_{\text{AR}}} Y_t/N + e^{\alpha_{\text{EN}}}$, for all models. For each simulated epidemic, we compare the MLEs from the following eight models:

1. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(S_t, 1 - e^{-\lambda_t^\dagger}\right)$.
2. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(N, 1 - e^{-\lambda_t^\dagger}\right)$.
3. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(S_t, \lambda_t^\dagger\right)$.
4. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(N, \lambda_t^\dagger\right)$.
5. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(S_t\,(1 - e^{-\lambda_t^\dagger})\right)$.
6. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(N\,(1 - e^{-\lambda_t^\dagger})\right)$.
7. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(S_t\,\lambda_t^\dagger\right)$.
8. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(N\,\lambda_t^\dagger\right)$.

In Figure 5.3, we summarize the results of the simulations examining the simplifying assumptions of the ecological model in a totally unvaccinated population. When $R_0 = 2$, the epidemic is limited by the number of susceptibles, and eventually dies off when there are few remaining susceptible individuals. In this scenario it is inappropriate to approximate the number of susceptibles with the initial number of susceptibles. Moreover, when R_0 is large, the estimates obtained using the Taylor approximation differ from the estimates using the true risk. In Figure 5.3(a), we plot the estimates from the eight models when $R_0 = 2$ and see that those models that approximate the number of susceptibles with the initial number of susceptibles perform poorly compared to those that account for the number of susceptibles. Figures 5.3(b) and 5.3(c), correspond to when $R_0 = 1$ and $R_0 = 0.85$, respectively. For

the smaller values of R_0, the models that approximate the number of susceptibles produces similar estimates to those which use the true number of susceptible individuals. In fact, for the smaller values of R_0, we see that all eight models produce similar estimates, with some slight underestimation of the autoregressive term and slight overestimation of the endemic term. In Appendix C, we include a table with the estimates, bias, and coverage for these simulations. When $R_0 = 0.85$ all eight models produce identical estimates up to three significant digits.



(a) $R_0 = 2$ (b) $R_0 = 1$ (c) $R_0 = 0.85$

Figure 5.3: Simulation results for ecological model simplifying assumptions. Rows correspond to the parameter, columns to the true values of R_0. The first row shows estimates of α_{AR}; the second row depicts estimates of α_{EN}. True parameter values are denoted by red lines.

5.7.2 Assessing the ecological model

We now consider the performance of the ecological model within a partially vaccinated population. For identifiablity, we need multiple populations with varying levels of vaccine

coverage or time-varying vaccine coverage. We simulate data for five areas with varying levels of coverage as follows assuming either an all-or-none or leaky vaccine as

$$Y_{u,t+1} \mid \lambda^{\dagger}_{ut} \sim \text{Binomial}\Big(S_{ut}, 1 - \exp(-\lambda^{\dagger}_{ut})\Big), \tag{5.21}$$

$$Y_{v,t+1} \mid \lambda^{\dagger}_{vt} \sim \text{Binomial}\Big(S_{vt}, 1 - \exp(-\lambda^{\dagger}_{vt})\Big), \tag{5.22}$$

where we let $Y_{u0} = 1$ and $Y_{v0=0}$; the initial number of susceptibles by vaccination status (S_{ut} and S_{vt}) is determined by the assumed vaccine mode of action, as will the specific forms for the force of infection by vaccination status ($\lambda^{\dagger}_{ut}$ and $\lambda^{\dagger}_{vt}$). For each model, we let $\lambda^{\dagger}_{it} = \exp(\alpha_{\text{AR}})\, Y_{it}/N_i + \exp(\alpha_{\text{EN}})$. We assume there are no infections from other areas, i.e. no neighborhood component. We compare model fits from the following models:

1. Fully observed all-or-nothing model:

$$\begin{aligned}
Y_{ui,t+1} \mid \lambda^{\dagger}_{it}, \phi &\sim \text{Binomial}\Big(S_{uit}(\phi), 1 - e^{-\lambda^{\dagger}_{it}}\Big), \\
Y_{vi,t+1} \mid \lambda^{\dagger}_{it}, \phi &\sim \text{Binomial}\Big(S_{vit}(\phi), 1 - e^{-\lambda^{\dagger}_{it}}\Big), \\
S_{uit}(\phi) &= (1 - x_i)N_i - \sum_{k=1}^{t} Y_{uik}, \\
S_{vit}(\phi) &= (1 - \phi)x_i N_i - \sum_{k=1}^{t} Y_{vik}.
\end{aligned}$$

2. Fully observed leaky model:

$$\begin{aligned}
Y_{ui,t+1} \mid \lambda^{\dagger}_{it}, \phi &\sim \text{Binomial}\Big(S_{uit}, 1 - e^{-\lambda^{\dagger}_{it}}\Big), \\
Y_{vi,t+1} \mid \lambda^{\dagger}_{it}, \phi &\sim \text{Binomial}\Big(S_{vit}, 1 - e^{-(1-\phi)\lambda^{\dagger}_{it}}\Big), \\
S_{uit} &= (1 - x_i)N_i - \sum_{k=1}^{t} Y_{uik}, \\
S_{vit} &= x_i N_i - \sum_{k=1}^{t} Y_{vik}.
\end{aligned}$$

3. Ecological vaccine model:

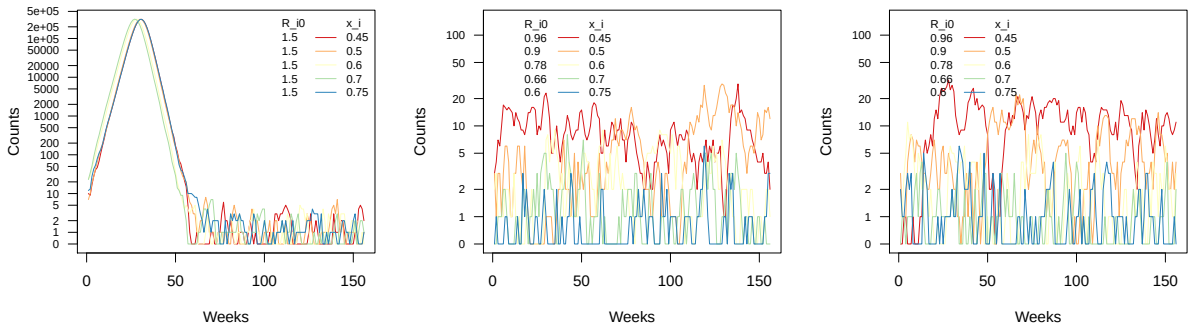
$$Y_{i,t+1}|\lambda_{it}^{\dagger} \sim \text{Poisson}\left(N_i(1-\phi x_i)\lambda_{it}^{\dagger}\right).$$

4. Herzog model:

$$\begin{aligned} Y_{i,t+1}|\mu_{it} &\sim \text{Poisson}(\mu_{it}), \\ \mu_{it} &= \exp(\alpha_0)\,(1-x_i)^{\alpha_1}\,Y_{it} + N_i \exp(\beta_0). \end{aligned}$$

We have parameterized $\lambda_{it}^{\dagger}$ so that α_0 and β_0 in the Herzog model are comparable to α_{AR} and α_{EN}, respectively, in the other models. However, the parameter associated with vaccine coverage in the Herzog model, α_1, is not directly comparable to the vaccine effect ϕ of the other models.

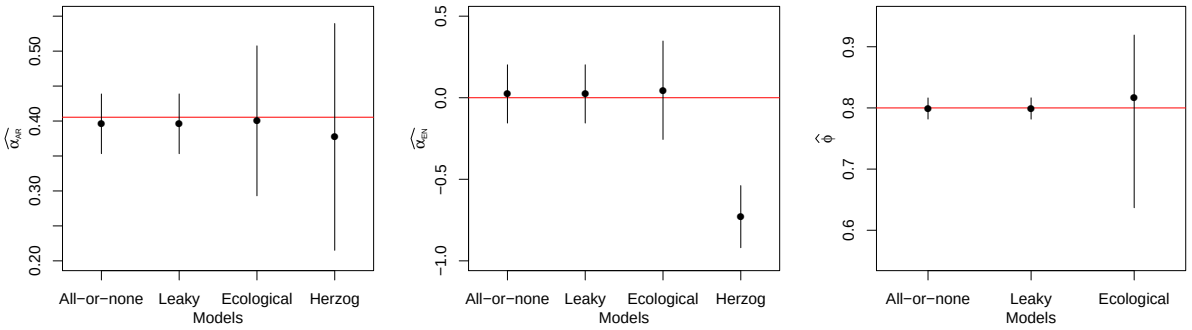
We consider the scenario when it is reasonable to assume that the ecological model is a good approximation to the true data generating model. We consider five equally sized populations, In Figure 5.4 we present an example of realizations for the five populations under the assumption of no vaccine effect, an all-or-none vaccine, and a leaky vaccine.



(a) No vaccine effect ($\phi = 0$) (b) All-or-none vaccine ($\phi = 0.8$) (c) Leaky vaccine ($\phi = 0.8$)

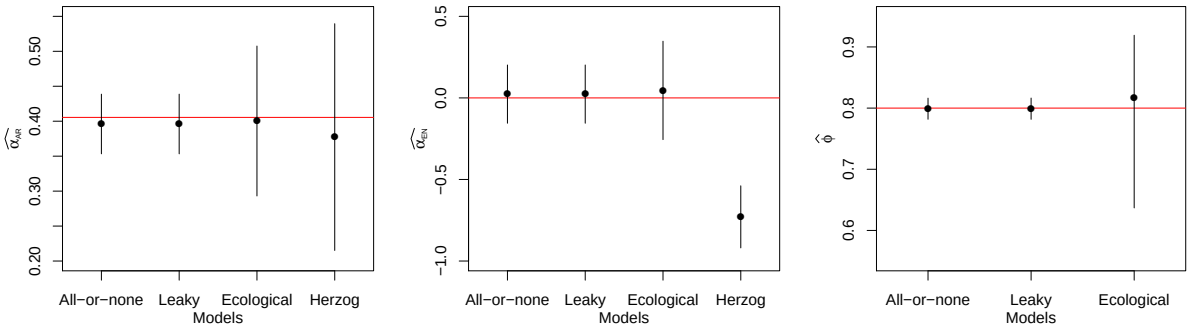
Figure 5.4: Simulated epidemic curves for five populations, when there is no vaccine effect, an effective all-or-none vaccine, and an effective leaky vaccine. Red and orange curves have lower vaccination coverage compared to the green and blue curves.

In Figure 5.5 we present the estimates obtained under the fully observed models, the ecological vaccine model, and the Herzog model when the data were simulated assuming an all-or-none vaccine. We present similar estimates for simulations assuming a leaky vaccine in Figure 5.6. Recall, both the all-or-none and the leaky models need the number of cases by vaccination status, which is not the case for the ecological and Herzog models. The ecological vaccine model has wider intervals than the all-or-none and leaky vaccine models, which appropriately reflect the lost information as a result of the aggregation. Nevertheless, the estimates from the ecological model are comparable to those obtained using completely observed data. In contrast, the Herzog models yield estimates that are very different from the true autoregressive and endemic parameter values. We do not include the Herzog estimates in the pictures for the estimates of the vaccine effect, ϕ, since the Herzog parameter is not comparable to the parameters in the other models.



(a) $\widehat{\alpha_{\mathrm{AR}}}$ (b) $\widehat{\alpha_{\mathrm{EN}}}$ (c) $\hat{\phi}$

Figure 5.5: Estimates and 95% confidence intervals of (a) α_{AR}, (b) α_{EN}, and (c) ϕ for the fully observed all-or-none and leaky models, the ecological vaccine model, and the Herzog model for data simulated assuming an *all-or-none* vaccine. Red horizontal lines denote the true parameter values.

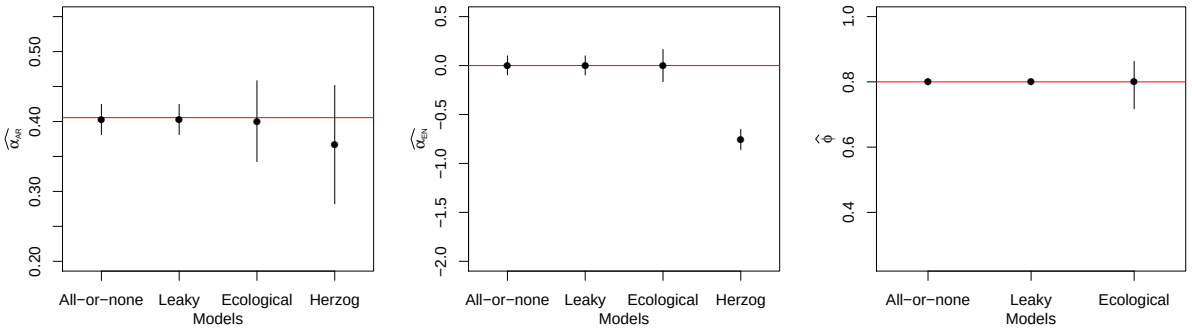


(a) $\widehat{\alpha_{\mathrm{AR}}}$ (b) $\widehat{\alpha_{\mathrm{EN}}}$ (c) $\hat{\phi}$

Figure 5.6: Estimates and 95% confidence intervals of (a) α_{AR}, (b) α_{EN}, and (c) ϕ for the fully observed all-or-none and leaky models, the ecological vaccine model, and the Herzog model for data simulated assuming a *leaky* vaccine. Red horizontal lines denote the true parameter values.

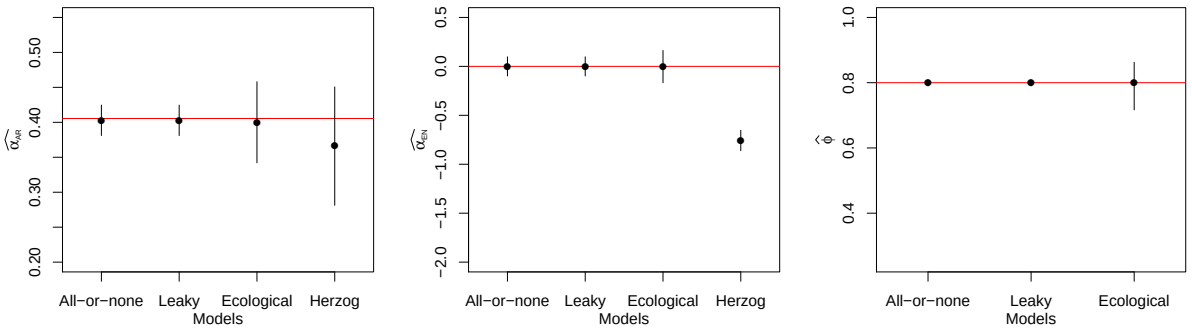
5.7.3 Asymptotic behavior of the ecological vaccine model

Under the same conditions as the simulations in Section 5.7.2, we consider the results for ten years worth of data. In Figures 5.7 and 5.8 we present estimates from the fully observed all-or-none and leaky models, along with estimates from the ecological vaccine model and the Herzog model. We see that the estimates for the fully observed models, as well as the ecological vaccine models are much closer to the true parameter values compared to the previous simulations, which used only 3 years of weekly data. With long time series, the ecological vaccine model provides unbiased estimates for all model parameters.



(a) $\widehat{\alpha_{\mathrm{AR}}}$ (b) $\widehat{\alpha_{\mathrm{EN}}}$ (c) $\hat{\phi}$

Figure 5.7: Estimates and 95% confidence intervals for the fully observed all-or-none and leaky models, the ecological vaccine model, and the Herzog model for 10 years worth of weekly data simulated assuming an *all-or-none* vaccine. Red horizontal lines denote the true parameter values.

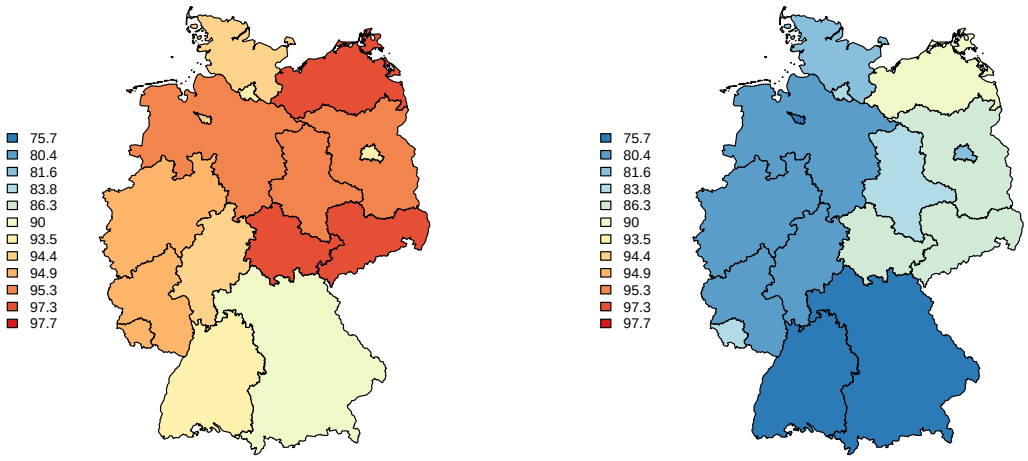


(a) $\widehat{\alpha_{\mathrm{AR}}}$ (b) $\widehat{\alpha_{\mathrm{EN}}}$ (c) $\hat{\phi}$

Figure 5.8: Estimates and 95% confidence intervals for the fully observed all-or-none and leaky models, the ecological vaccine model, and the Herzog model for 10 years worth of weekly data simulated assuming a *leaky* vaccine. Red horizontal lines denote the true parameter values.

5.8 Application to the measles data

We now apply these methods to data on measles outbreaks in Germany described in Section 1.1.3. Estimated measles, mumps, and rubella (MMR) vaccination coverage is based on the number of students presenting vaccination cards at the required medical exam for school entry (Herzog et al., 2011). Between 87% and 95% of students brought vaccination cards to the entry exam preceding the start of the 2006–2007 school year. In Figure 5.9 we plot the percentage of children who brought vaccination documentation to the medical exam and had received one or more MMR vaccines (left) and at least two vaccines (right).



(a) At least 1 MMR vaccination.

(b) At least 2 MMR vaccinations.

Figure 5.9: Percentage of children with vaccination cards who had at least 1 MMR vaccination (left) and at least 2 MMR vaccinations (right) in 2006.

For our analysis, we estimate the coverage for at least one MMR vaccine by assuming that the coverage in the population that did not bring the vaccination cards is half that of those who did have vaccination cards. The available vaccination data is for children starting primary school, typically between 4 and 7 years of age. We assume that the MMR vaccination coverage for the whole population is the same as the estimated vaccination coverage for children starting primary school. Herzog et al. (2011) also assumed constant coverage across

the population, and made the same assumption about the vaccination rate for those without a vaccination card. As Herzog et al. (2011) discussed, the estimated coverage is likely to be an overestimate, as those who show up for the annual medical exam and bring vaccination cards are more likely to have more complete medical records. In their analysis, Herzog et al. (2011) conduct sensitivity analyses to examine the assumption of vaccine coverage among those with no vaccination card.

We assume a two week time step, based on the approximate generation time for measles (Bjørnstad et al., 2002; Herzog et al., 2011). As such, data is aggregated into biweekly data. In Figure 5.10, we plot the total number of observed measles cases and prevalence per 100,000 people, by state and biweek for the sixteen states in Germany. The left axis indicates the total number of cases; the right axis indicates the prevalence per 100,000 people. We estimate the coverage for at least one MMR vaccine, and include the estimated vaccine coverage in the upper right corner of each frame. Vaccine coverage estimates range from 88% to 95% for each of the states. Saarland had no reported cases of measles over the three years, while the North Rhine-Westphalia state had nearly 300 cases in a single biweek. In general, most states have under 50 cases in a given biweek.

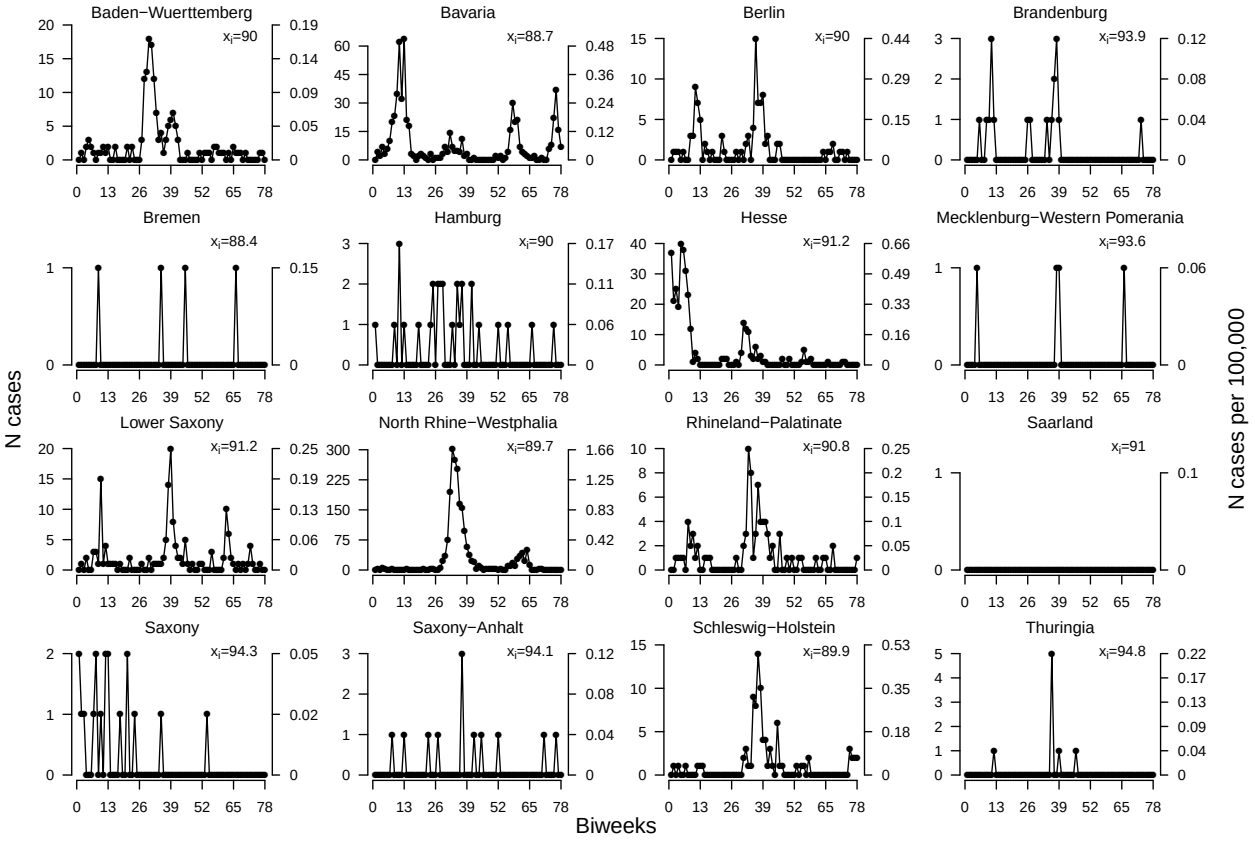


Figure 5.10: Number of measles cases and prevalence by state and biweek from 2005 through 2007. The left axis indicates the total number of cases; the right axis indicates the prevalence per 100,000 people. Estimated MMR vaccine coverage is included in the upper right corner of each plot.

We incorporate our previous knowledge about the MMR vaccine effectiveness by placing a beta prior on ϕ, where 90% of the mass is between 0.6 and 0.99. During the study period, Germany had a population of over 80 million people, and it is reasonable to assume that the number of cases is negligible. We fit the following ecological model to the measles data:

$$\begin{aligned}
Y_{i,t+1} \,|\, \mu_{it}, \phi &\sim \text{Poisson}(N_i(1-\phi x_i)\mu_{it}), \\
\mu_{it} &= \lambda_i \frac{y_{it}}{N_i} + \nu_{it}, \\
\log \lambda_i &= \alpha_{\text{AR}} + a_i, \\
\log \nu_{it} &= \alpha_{\text{EN}} + b_i + \gamma \sin(2\pi t/26) + \delta \cos(2\pi t/26), \\
a_i &\sim \text{N}(0, \sigma^2_{\text{AR}}), \\
b_i &\sim \text{N}(0, \sigma^2_{\text{EN}}), \\
\phi &\sim \text{Beta}(10, 2.5),
\end{aligned}$$

where x_i is the estimated vaccine coverage in area i; component-specific random effects a_i and b_i are assumed independent.

We summarize the fixed effects estimates in Table 5.2. We find that the posterior estimate of R_0 is 2.38, with a 95% interval between 0.74 and 5.22. This estimate is much smaller than the known R_0 value for measles, which is between 15 and 20 (Sudfeld et al., 2010; Keeling and Rohani, 2008). There are many possible sources of this underestimation, for example, under-reporting of measles cases, poor estimation of vaccine coverage and the discretization of time. It is also possible that the underestimating is due to the assumption of a frequency dependent contact rate. However, Bjørnstad et al. (2002) found no evidence of a relationship between measles transmission and population size in England and Wales from 1944 to 1964, suggesting that the frequency dependent assumption is reasonable.

The estimated vaccine effect is 0.91, with a posterior credible interval from 0.64 to 0.99, which is commensurate with the known vaccine efficacy for the MMR vaccine. In Figure 5.11 we plot a histogram of posterior samples of $\hat{\phi}$ along with the prior Beta(10, 2.5) curve.

	Median	2.5%	97.5%
α_{AR}	0.87	-0.30	1.65
ϕ	0.91	0.64	0.99
α_{EN}	3.52	2.53	4.17
γ	0.71	0.55	0.86
δ	-0.20	-0.36	-0.04
σ_{AR}	0.70	0.28	1.66
σ_{EN}	0.53	0.27	0.96
R_0	2.38	0.74	5.22

Table 5.2: Posterior medians and 95% credible intervals for the parameters of the ecological model for the measles data.

The posterior is similar to the prior, suggesting that there is little information about the vaccine effect in this data. As a sensitivity analysis, we fit the same hierarchical model with a non-informative prior for ϕ. The results for this analysis are presented in Appendix C.2. The non-informative prior on ϕ results in slightly higher estimates for both α_{AR} and ϕ, but each have substantially wider credible intervals. The prior choice for ϕ had little effect on the posterior estimates of the parameters in the endemic component of the model.

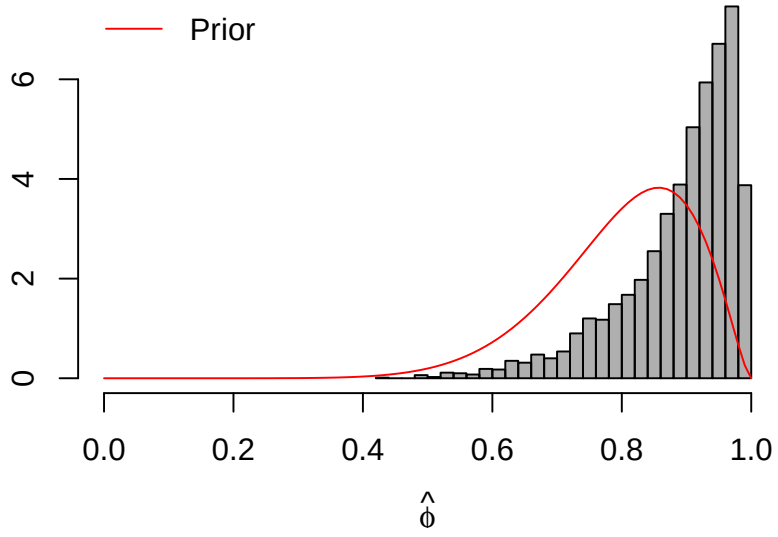


Figure 5.11: Histogram of posterior samples of $\hat{\phi}$. The red curve is the prior distribution, Beta(10, 2.5).

In Figure 5.12 we compare the fit of the ecological model with the observed data. Fitted values are computed as

$$\hat{Y}_{it} = (1 - \hat{\phi} x_i) \exp\left(\widehat{\alpha_{\text{AR}}} + \widehat{a}_i\right) Y_{i,t-1} + N_i \exp\left(\widehat{\alpha_{\text{EN}}} + \widehat{b}_i + \hat{\gamma}\sin(\omega_t) + \hat{\delta}\cos(\omega_t)\right), \quad (5.23)$$

where $Y_{i,t-1}$ is the observed number of counts for area i and week $t-1$, and $\omega_t = 2\pi t/26$. In general, the ecological vaccine model provides good estimates for the number of cases. In the upper right corner of each state-specific plot, we include the estimated effective reproductive number, $\hat{R} = (1 - x_i)\exp(\widehat{\alpha_{\text{AR}}} + \hat{a}_i)$.

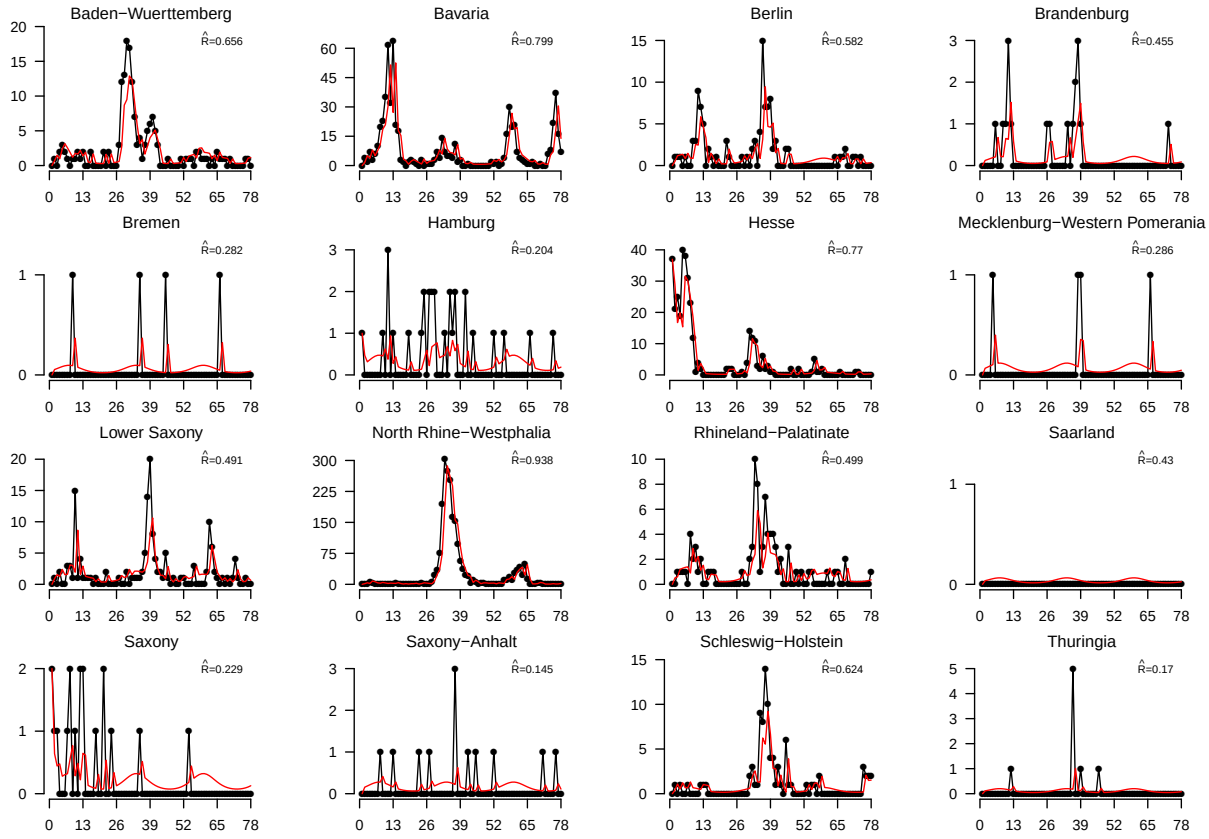


Figure 5.12: Fitted cases for ecological model that includes random effects in both the endemic and epidemic components. Black points denote observed data, red lines are fitted values.

In Figure 5.13 we plot the Pearson residuals, $r_{it} = \left(Y_{it} - \hat{Y}_{it}\right) / \sqrt{\hat{Y}_{it}}$, where $\hat{y}_{it}$ is the fitted value for area i and biweek t which is estimated following (5.23). In general, it appears that this model provides a reasonable fit to the data.

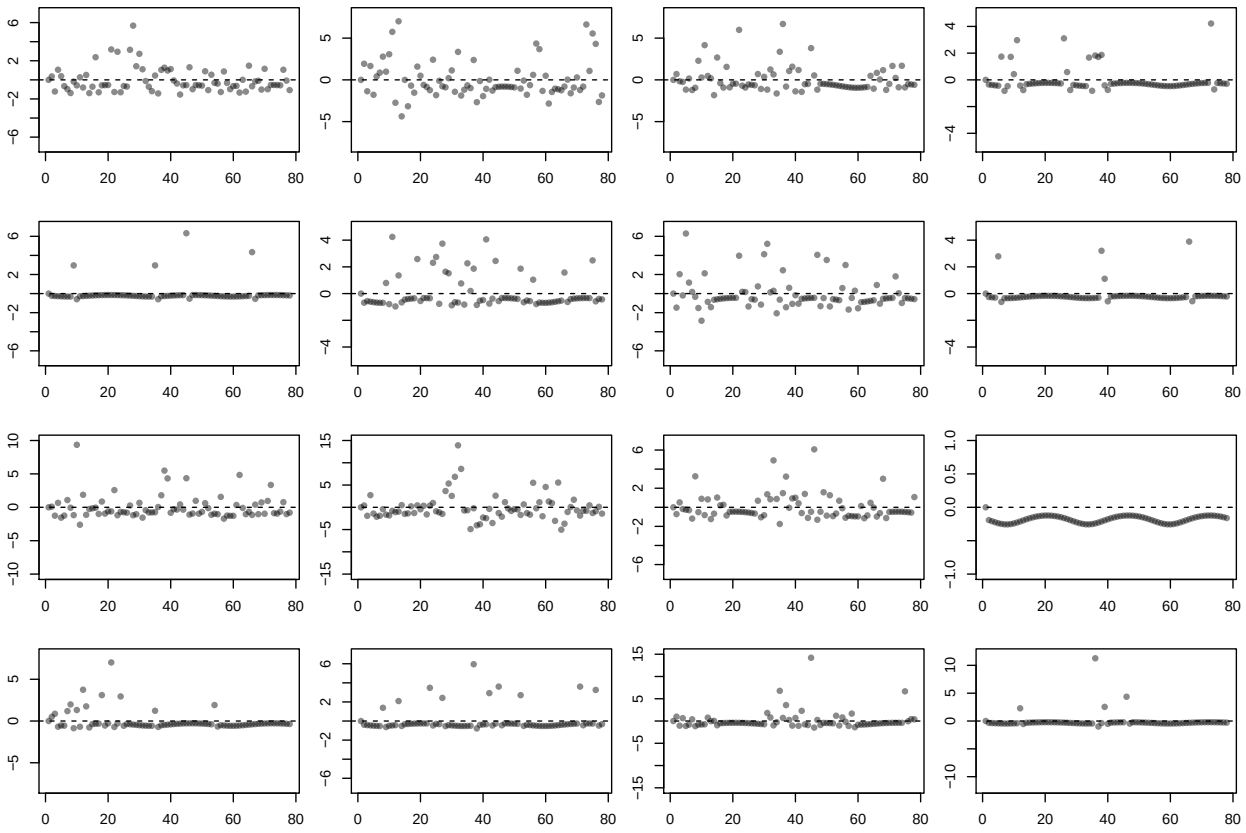
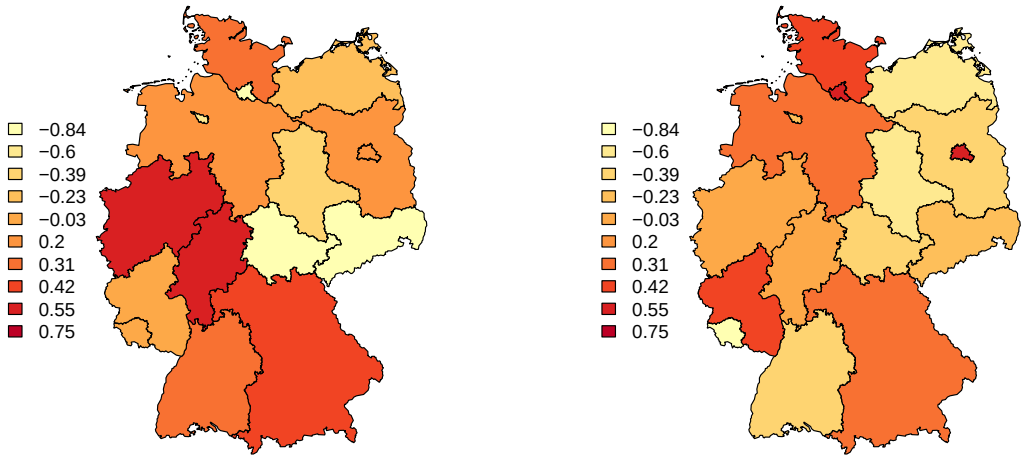


Figure 5.13: Pearson residuals for the ecological vaccine model.

In Figure 5.14 we plot the area-specific random effects for the autoregressive and endemic components. The states with the highest prevalence have higher autoregressive random effects. The endemic random effects do not appear to have a similar spatial structure as the autoregressive random effects.



(a) AR random effects.

(b) EN random effects.

Figure 5.14: Maps of the random effect estimates for the autoregressive and endemic components in the ecological vaccine model.

When the autoregressive random effects are plotted against the endemic random effects, as in Figure 5.15, there is no evidence of a strong correlation between the two components. This supports our decision to model the component-specific random effects separately. However, in other settings, we may want to consider more complex forms of random effects. For example, if there were strong correlations between the component-specific random effects, bivariate random effects may be more appropriate.

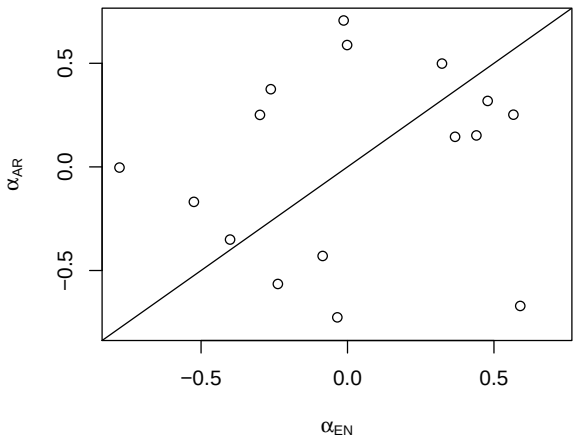


Figure 5.15: Comparison of autoregressive and endemic random effect estimates.

5.9 Discussion

We derived an ecological model for infectious disease data that accounts for vaccination coverage that is consistent for the individual-level vaccine effects. We considered different modes of action for the vaccine, which lead to the same ecological vaccine model under common simplifying assumptions. Simulations showed that the ecological vaccine model performs reasonably well in many practical scenarios and illuminated when certain assumptions are inappropriate.

A key benefit to our approach is that we obtain estimates of familiar epidemiological parameters which are easy to interpret. However, the ecological vaccine model proposed makes many assumptions that may not be appropriate for many infectious diseases. For example, it would be beneficial to extend the ecological vaccine model to account for a non-constant and perhaps longer infectiousness period. The work presented in this chapter would benefit from additional investigations. It may be interesting to consider bivariate random effects, or spatially structured random effects in the autoregressive and/or endemic components. Future work will be focused on extending the method to account for stratified population structures and including neighborhood effects in the ecological vaccine model.

In summary, we developed an ecological model that accounts for area-level vaccination coverage and provides estimates of familiar epidemiological parameters that are easy to interpret. Our approach started with an individual-level model and made several simplifying assumptions to arrive at an aggregate consistent model.

Chapter 6

DISCUSSION AND FUTURE WORK

Infectious disease surveillance systems provide a rich source of data regarding infectious diseases at the local, state, and national level. Public health officials use this data in a variety of ways, including to assess the current disease burden, detect emerging outbreaks, and inform policy decisions. While surveillance data plays a critical role in public health, there are deficiencies in the data that can make it difficult to obtain unbiased estimates of quantities of interest in a timely fashion. In this dissertation, we addressed three distinct methodological challenges in an effort to obtain timely and unbiased estimates of meaningful infectious disease parameters using surveillance data.

In Chapter 3, we developed a hybrid approach to quickly obtain estimates of the unobserved pathogen-specific disease counts with corresponding standard errors when the large number of cases and small number of subsamples with virological information made typical MCMC approaches intractable. These estimates were then used to model pathogen-specific disease dynamics and covariate effects. This work has been published in Fisher et al. (2016).

In Chapter 4, we sought to evaluate the effect of a school-located influenza vaccine program on community-wide influenza like illness using passively collected emergency department data across the state of Florida. In this setting, the surveillance data was not a representative sample of influenza cases in a county since unmeasured socio-economic factors influenced which individuals sought care for ILI at their local emergency department. The unmeasured differences in ascertainment due to differences in healthcare seeking behavior counfounds results and yields biased estimates. We re-framed the negative control framework and extended it to account for spatial data, highlighting the strong assumptions needed to adjust for unmeasured confounding.

In Chapter 5, we developed an ecological infectious disease model for a partially vaccinated population that provides estimates of familiar epidemiological parameters. Current approaches to ecological regression for infectious diseases in partially vaccinated populations are computationally expensive, lack familiar interpretation, and are susceptible to ecological bias. As an alternative, we start with an individual-level model with familiar parameters and derive an ecologically consistent model for infectious diseases in partially vaccinated populations.

The methods described in this dissertation were developed to provide timely, unbiased estimates of relevant parameters using infectious disease surveillance data. Infectious disease surveillance systems are an important and affordable source of information. We explored three instances where common deficiencies in the surveillance data made answering specific scientific questions more challenging. It is important to note that the deficiencies of interest in our settings may not be problematic in answering a different scientific question or using a different source of data. Moreover, there are other common deficiencies in surveillance data, such as under-reporting, which we did not address in this dissertation. The specific deficiencies that make inference challenging will depend on both the nature of the surveillance data collected and the scientific question of interest.

BIBLIOGRAPHY

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, volume 4. Springer New York.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:269–342.

Bauer, C. and Wakefield, J. (2016). Stratified space-time infectious disease modeling. *Submitted*.

Bauer, C. X. C. (2012). *Bayesian Modeling of Health Data in Space and Time*. PhD thesis, University of Washington.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43:1–59.

Betancourt, J. A., Hakre, S., Polyak, C. S., and Pavlin, J. A. (2007). Evaluation of ICD-9 codes for syndromic surveillance in the electronic surveillance system for the early notification of community-based epidemics. *Military medicine*, 172:346–352.

Bjørnstad, O. N., Finkenstädt, B. F., and Grenfell, B. T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. *Ecological Monographs*, 72:169–184.

Botella-Rocamora, P., Martinez-Beneito, M. A., and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in medicine*, 34:1548–1559.

Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225:24 – 35.

CDC (1998). Measles, mumps, and rubella – vaccine use and strategies for elimination of measles, rubella, and congenital rubella syndrome and control of mumps: Recommendations of the advisory committee on immunization practices (ACIP). http://www.cdc.gov/mmwr/preview/mmwrhtml/00053391.htm.

CDC (2011). *Assessment of ESSENCE performance for influenza-like illness surveillance after an influenza outbreak–US Air Force Academy, Colorado, 2009.* Centers for Disease Control and Prevention.

CDC (2016). The flu season. http://www.cdc.gov/flu/about/season/flu-season.htm.

Chang, H.-L., Chio, C.-P., Su, H.-J., Liao, C.-M., Lin, C.-Y., Shau, W.-Y., Chi, Y.-C., Cheng, Y.-T., Chou, Y.-L., Li, C.-Y., et al. (2012). The association between enterovirus 71 infections and meteorological parameters in Taiwan. *PLoS One*, 7:e46845.

Clayton, D., Bernardinelli, L., and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22:1193–1202.

Dabney, A. R. and Wakefield, J. C. (2005). Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, 14:83–112.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26:363–397.

Dominici, F., McDermott, A., and Hastie, T. (2004). Improved semi-parametric time series models of air pollution and mortality. *Journal of the American Statistical Association*, 468:938–948.

Dominici, F., Sheppard, L., and Clyde, M. (2003). Health effects of air pollution: A statistical review. *International Statistical Review*, 71:243–276.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, 195(2):216 – 222.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association*, 107(500):1410–1426.

Fisher, L., Wakefield, J., Bauer, C., and Self, S. (2016). Time series modeling of pathogen-specific disease probabilities with subsampled data. *Biometrics*. http://dx.doi.org/10.1111/biom.12560.

Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., and Tolbert, P. E. (2011). A method to detect residual confounding in spatial and other observational studies. *Epidemiology (Cambridge, Mass.)*, 22:823.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11:397–412.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24:997–1016.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15:19–40.

Greenland, S. (1992). Divergent biases in ecologic and individual level studies. *Statistics in Medicine*, 11:1209–1223.

Greenland, S. and Robins, J. (1994). Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology*, 139:747–760.

Halloran, M., I.M. Longini, J., and Struchiner, C. (2010). *Design and Analysis of Vaccine Studies*. Springer, New York.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*, 7:271–283.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5:187–199.

Herzog, S., Paul, M., and Held, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in german surveillance data. *Epidemiology and Infection*, 139:505–515.

Hoadley, B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*, 64:216–229.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64:325–334.

Huang, Y., Deng, T., Yu, S., Gu, J., Huang, C., Xiao, G., Hao, Y., et al. (2013). Effect of meteorological variables on the incidence of hand, foot, and mouth disease in children: a time-series analysis in Guangzhou, China. *BMC Infectious Diseases*, 13:134.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, 75:131–159.

Ionides, E., Bretó, C., and King, A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103:18438–18443.

Keeling, M. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton and Oxford.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, 17:2045–60.

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164:73–85.

Koopman, J. S. and Longini, I. M. (1994). The ecological effects of individual exposures and nonlinear disease dynamics in populations. *American Journal of Public Health*, 84:836–842.

Lekone, P. and Finkenstädt, B. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62:1170–1177.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21:383.

Liu, M. Y., Liu, W., Luo, J., Liu, Y., Zhu, Y., Berman, H., and Wu, J. (2011). Characterization of an outbreak of hand, foot, and mouth disease in Nanchang, China in 2010. *PLoS One*, 6:e25287.

Ma, E., Lam, T., Wong, C., and Chuang, S. (2010). Is hand, foot and mouth disease associated with meteorological parameters? *Epidemiology and Infection*, 138:1779–1788.

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100:539–553. doi: 10.1093/biomet/ast023.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67:68–83.

McKinley, T., Cook, A., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5:24.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8:1612–1639.

Meyer, S., Held, L., and Höhle, M. (2016). Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *Journal of Statistical Software*. Preprint available at `http://arxiv.org/abs/1411.0416`.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162:121–129.

Onozuka, D. and Hashizume, M. (2011). The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan. *Science of The Total Environment*, 410–411:119 – 125.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, 30:1118–1136.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27:6250–6267.

Peng, R. D., Dominici, F., and Louis, T. A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169:179–203.

Plummer, M. and Clayton, D. (1996). Estimation of population exposure. *Journal of the Royal Statistical Society, Series B*, 58:113–126.

Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol*, 7:e1002136.

Richardson, D. B., Keil, A. P., Tchetgen Tchetgen, E., and Cooper, G. (2015). Negative

control outcomes and the analysis of standardized mortality ratios. *Epidemiology*, 26:727–732.

Richardson, D. B., Laurier, D., Schubauer-Berigan, M. K., Tchetgen, E. T., and Cole, S. R. (2014). Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *American Journal of Epidemiology*, 180:933–940.

Richardson, S. and Monfort, C. (2000). Ecological correlation studies. In Elliott, P., Wakefield, J. C., Best, N. G., and Briggs, D. J., editors, *Spatial Epidemiology: Methods and Application*, chapter 11. Oxford University Press, Oxford.

Richardson, S., Stucker, I., and Hémon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, 16:111–20.

Roberts, G., Gelman, A., and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–57.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 71:319–392.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Selvin, H. (1958). Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology*, 63:607–619.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64:583–639.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B*, 64:485–493.

Stan Development Team (2015a). Stan: A c++ library for probability and sampling, version 2.8.0.

Stan Development Team (2015b). *Stan Modeling Language User's Guide and Reference Manual, Version 2.8.0*.

Strebel, P. M., Papania, M. J., Fiebelkorn, A. P., and Halsey, N. A. (2013). 20 - measles vaccine. In Plotkin, S. A., Orenstein, W. A., and Offit, P. A., editors, *Vaccines (Sixth Edition)*, pages 352 – 387. W.B. Saunders, London, sixth edition edition.

Sudfeld, C. R., Navar, A. M., and Halsey, N. A. (2010). Effectiveness of measles vaccination and vitamin a treatment. *International Journal of Epidemiology*, 39:i48–i55.

Tong, C. W. and Bible, J. M. (2009). Global epidemiology of enterovirus 71. *Future Virology*, 4:501–510.

Tran, C. H., McElrath, J., Hughes, P., Ryan, K., Munden, J., Castleman, J. B., Johnson, J., Doty, R., McKay, D. R., Stringfellow, J., et al. (2010). Implementing a community-supported school-based influenza immunization program. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 8:331–341.

Tran, C. H., Sugimoto, J. D., Pulliam, J. R., Ryan, K. A., Myers, P. D., Castleman, J. B., Doty, R., Johnson, J., Stringfellow, J., Kovacevich, N., et al. (2014). School-located influenza vaccination reduces community risk for influenza and influenza-like illness emergency care visits. *PLoS One*, 9:e114479.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8:158–183.

Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health*, 29:75–90.

Wakefield, J., Haneuse, S., Dobra, A., and Teeple, E. (2011). Bayes computation for ecological inference. *Statistics in medicine*, 30:1381–1396.

Wakefield, J. C., Best, N. G., and Waller, L. A. (2000). Bayesian approaches to disease mapping. In Elliott, P., Wakefield, J. C., Best, N. G., and Briggs, D., editors, *Spatial Epidemiology: Methods and Applications*, pages 104–27. Oxford University Press, Oxford.

Wang, Y., Feng, Z., Yang, Y., Self, S., Gao, Y., Longini, I. M., Wakefield, J., Zhang, J., Wang, L., Chen, X., et al. (2011). Hand, foot and mouth disease in China: Patterns of spread and transmissibility during 2008-2009. *Epidemiology (Cambridge, Mass.)*, 22:781.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897.

WHO (2009). Measles vaccines: WHO position paper. *Weekly Epidemiological Record*, 84:349–60. http://www.who.int/wer/2009/wer8435/en/.

Wu, H., Wang, H., Wang, Q., Xin, Q., and Lin, H. (2014). The effect of meteorological factors on adolescent hand, foot, and mouth disease and associated effect modifiers. *Global Health Action*, 7:24664.

Xia, Y., Bjørnstad, O. N., and Grenfell, B. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164:267–281.

Appendix A

APPENDIX TO CHAPTER 3

A.1 Comparison of empirical Bayes and method of moment estimators

A method of moment (MoM) approach to estimate the unobserved Y_{tj}^{GS} for $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$ yields

$$\widehat{Y_{tj}^{\mathrm{GS}}} = Y_{tj}^{\cdot\mathrm{S}} \left(\frac{Z_{tj}^{\mathrm{GS}}}{k_{tj}^{\mathrm{S}}} \right) \quad \text{and} \quad \widehat{Y_{tj}^{\mathrm{GM}}} = Y_{tj}^{\cdot\mathrm{M}} \left(\frac{Z_{tj}^{\mathrm{GM}}}{k_{tj}^{\mathrm{M}}} \right). \tag{A.1}$$

The MoM method provides an intuitive estimate of the unobserved counts; it simply scales the total number of severe (or mild) cases by the proportion of subsampled cases that were caused by a specific pathogen. Derived standard error estimates (using the moments of a multivariate hypergeometric distribution) provide a measure of the uncertainty in these estimates,

$$\begin{aligned} \mathrm{Var}\left[\widehat{Y_{tj}^{\mathrm{GS}}}\right] &= \frac{Z_{tj}^{\mathrm{GS}}(k_{tj}^{\mathrm{S}} - Z_{tj}^{\mathrm{GS}})(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}})(Y_{tj}^{\cdot\mathrm{S}})^2}{(Y_{tj}^{\cdot\mathrm{S}} - 1)(k_{tj}^{\mathrm{S}})^3} \\ &= \frac{Y_{tj}^{\cdot\mathrm{S}}}{k_{tj}^{\mathrm{S}}(Y_{tj}^{\cdot\mathrm{S}} - 1)} \left(\frac{Z_{tj}^{\mathrm{GS}}(k_{tj}^{\mathrm{S}} - Z_{tj}^{\mathrm{GS}})(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}})Y_{tj}^{\cdot\mathrm{S}}}{(k_{tj}^{\mathrm{S}})^2} \right), \end{aligned} \tag{A.2}$$

as well as covariance estimates within stratum and severity, for example

$$\mathrm{Cov}\left[\widehat{Y_{tj}^{\mathrm{ES}}}, \widehat{Y_{tj}^{\mathrm{CS}}}\right] = -\frac{Z_{tj}^{\mathrm{ES}} Z_{tj}^{\mathrm{CS}}(Y_{tj}^{\cdot\mathrm{S}} - k_{tj}^{\mathrm{S}})(Y_{tj}^{\cdot\mathrm{S}})^2}{(Y_{tj}^{\cdot\mathrm{S}} - 1)(k_{tj}^{\mathrm{S}})^3}. \tag{A.3}$$

While these estimators are intuitive and easy to compute without using MCMC, there are some practical details that require thought.

In the case of HFMD, the number of cases (both severe and mild) that are subsampled

for virology varies greatly, ranging from 0 to 483 samples for a given strata and week (see Figures 3.3 and 3.4). When no subsamples are taken for virology in week t and strata j (e.g. when $k^{\mathrm{S}}_{tj} = 0$), we cannot estimate pathogen-specific disease counts. In the HFMD data, approximately 36% of the strata-specific samples have no severe cases sampled. Samples are generally not large enough to obtain at least one sample from each of the three pathogen types, resulting in an estimated pathogen-specific disease count of zero for the unrepresented pathogen.

Moreover, standard error estimates of zero are obtained when there are no observed cases attributable to a particular pathogen. Although, it is likely that all three pathogens are responsible for some number of cases in a given week, this is not reflected in the estimates, and corresponding standard error estimates, produced by the simple MoM procedure.

In the chapter, we describe an empirical Bayesian procedure to obtain estimates of Y^{G}_{tj}. Specifically, we found

$$\widetilde{Y^{\mathrm{GS}}_{tj}} = Z^{\mathrm{GS}}_{tj} + \left(Y^{\cdot\mathrm{S}}_{tj} - k^{\mathrm{S}}_{tj}\right)\left(\frac{\alpha^{\mathrm{GS}}_{j} + Z^{\mathrm{GS}}_{tj}}{\alpha^{\cdot\mathrm{S}}_{j} + k^{\mathrm{S}}_{tj}}\right).$$

Notice that when $\alpha^{\mathrm{GS}}_{j} = 0$,

$$\widetilde{Y^{\mathrm{GS}}_{tj}} = Z^{\mathrm{GS}}_{tj} + \left(Y^{\cdot\mathrm{S}}_{tj} - k^{\mathrm{S}}_{tj}\right)\left(\frac{Z^{\mathrm{GS}}_{tj}}{k^{\mathrm{S}}_{tj}}\right) = \frac{Z^{\mathrm{GS}}_{tj}k^{\mathrm{S}}_{tj} + \left(Y^{\cdot\mathrm{S}}_{tj} - k^{\mathrm{S}}_{tj}\right)Z^{\mathrm{GS}}_{tj}}{k^{\mathrm{S}}_{tj}} = \frac{Z^{\mathrm{GS}}_{tj}Y^{\cdot\mathrm{S}}_{tj}}{k^{\mathrm{S}}_{tj}}$$

which is the same estimates as we derived previously (A.1). Furthermore, the estimated variance when $\alpha^{\mathrm{GS}}_{j} = 0$ for all $\mathrm{G} = \mathrm{E}, \mathrm{C}, \mathrm{O}$,

$$\mathrm{Var}\left[Y^{\mathrm{GS}}_{tj}\middle| Z^{\mathrm{GS}}_{tj}, Y^{\cdot\mathrm{S}}_{tj}, \boldsymbol{\alpha}^{\mathrm{S}}_{j}\right] = \frac{1}{k^{\mathrm{S}}_{tj}+1}\left(\frac{Z^{\mathrm{GS}}_{tj}(k^{\mathrm{S}}_{tj} - Z^{\mathrm{GS}}_{tj})(Y^{\cdot\mathrm{S}}_{tj} - k^{\mathrm{S}}_{tj})Y^{\cdot\mathrm{S}}_{tj}}{(k^{\mathrm{S}}_{tj})^2}\right),$$

differs from the variance in Equation (A.2) by only a factor of $k^{\mathrm{S}}_{tj}(Y^{\cdot\mathrm{S}}_{tj} - 1)/(k^{\mathrm{S}}_{tj} + 1)Y^{\cdot\mathrm{S}}_{tj} \approx 1$. When both the number of severe cases and the number of severe cases subsampled within a stratum are large, these variances are nearly indistinguishable. The covariance estimates

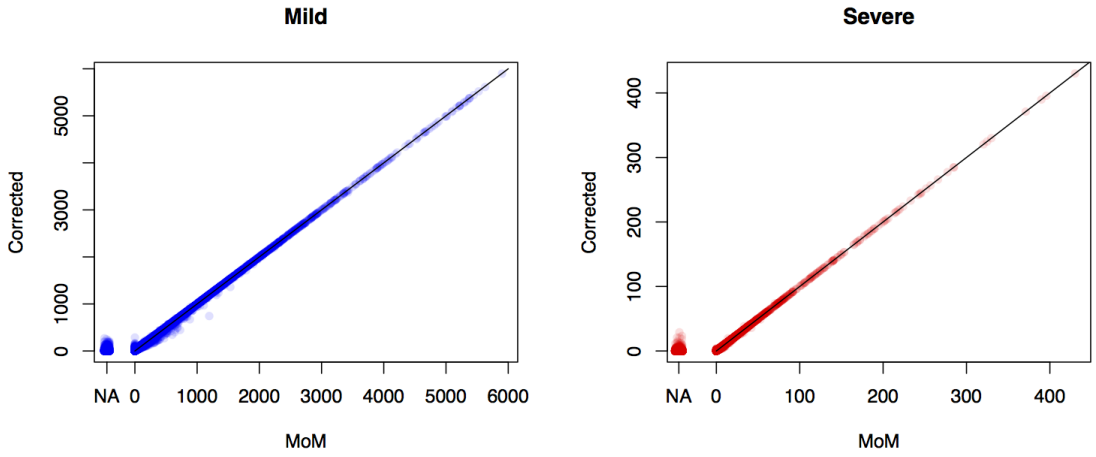
differ by the same factor.

Figure A.1(a) plots the corrected estimated counts computed using our procedure against the uncorrected (MoM) estimates. In general, the corrected estimates are close to those from the MoM method. The most noticeable differences in the estimated counts between the two methods tends to be close to zero, as we would expect. Even when none of the subsampled cases are found to be caused by a given pathogen, our procedure will estimate that the number of cases that are caused by that pathogen as non-zero. This is desirable since it is likely that there are cases (both severe and mild) caused by all three pathogens, even if we do not see any cases in our subsample. Note that, in general, our Bayes procedure will lead to non-integer estimates of the number of cases. However, non-integer disease counts is not an issue when modeling the log relative risk.

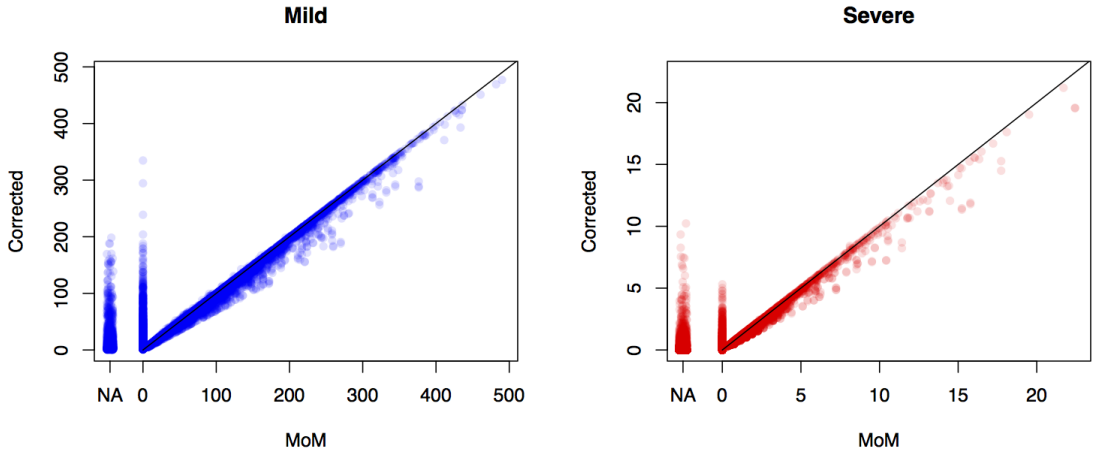
In Figure A.1(b), we plot the posterior standard deviations against the standard errors obtained by the simple method of moments estimates. We see that, in general for values greater than zero our method provides smaller standard error estimates. Again, this is partially due to the fact that we have chosen small prior sample sizes. We find that when the MoM estimates standard errors of zero, our procedure generally yields non-zero standard error estimates.

Even with the correction, there are cases when estimates of stratum- and pathogen-specific counts are zero. It is also still possible to still obtain estimated standard errors of zero. Adjusted estimates of counts will be zero when there are no observed cases in a particular stratum and week (i.e. $Y_{tj}^{\cdot \mathrm{S}} = 0$ and $Y_{tj}^{\cdot \mathrm{M}} = 0$). When there are no observed cases of HFMD (severe or mild) for a particular stratum and week, there are also no cases to sample from, and thus will estimate no pathogen-specific cases, and that estimate will have a standard error of zero. When all of the observed cases for a given stratum and week are sampled for virology, we will again obtain estimate the standard error to be zero. In both of these circumstances, we know the number of pathogen-specific cases within a given stratum and week with total certainty.

Importantly, these zeros are not problematic for inference. Recall, that our estimate of



(a) Comparison of estimated counts.



(b) Comparison of standard error estimates.

Figure A.1: Corrected estimated pathogen-specific disease counts vs uncorrected MoM estimated counts (a) and standard error estimates (b). Mild disease counts are on the left. Counts that could not be estimated under MoM are presented with a jitter.

the pathogen-specific relative risk is summed across stratum. There are no weeks where there are no cases of HFMD for all strata, hence we will always have non-zero estimates of the relative risk, θ_t^{G}. The same is true of the estimated standard error.

A.2 Conditional posterior distributions for MCMC

The conditional posterior distributions required for an MCMC algorithm are:

- Conditional posterior distribution for Y_{tj}^{ES}:

$$
\begin{aligned}
&p(Y_{tj}^{\mathrm{ES}}|\boldsymbol{r}_{tj},\boldsymbol{Z}_{tj}^{\mathrm{S}},\boldsymbol{Z}_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},Y_{tj}^{\mathrm{CS}},Y_{tj}^{\cdot\mathrm{S}},Y_{tj}^{+},\boldsymbol{k}_{tj},N_j)\\
&\qquad\propto \binom{Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{ES}}}\binom{Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{OS}}}\binom{Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{EM}}}\binom{Y_{tj}^{+}-Y_{tj}^{\cdot\mathrm{S}}-(Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}})-(Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}}\\
&\qquad\qquad\times\binom{Y_{tj}^{\mathrm{E}}}{Y_{tj}^{\mathrm{ES}}}\left(r_{tj}^{\mathrm{E}}\right)^{Y_{tj}^{\mathrm{ES}}}\left(1-r_{tj}^{\mathrm{E}}\right)^{Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}}\\
&\qquad\qquad\times\binom{Y_{tj}^{+}-Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{C}}}{Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}}\left(r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}})}\left(1-r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{+}-Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{C}})-(Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}})}
\end{aligned}
$$

  with $\max\left\{Z_{tj}^{\mathrm{ES}},Z_{tj}^{\mathrm{OM}}-Y_{tj}^{\cdot\mathrm{M}}+Y_{tj}^{\mathrm{E}}+Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}\right\}\leq Y_{tj}^{\mathrm{ES}}\leq\min\left\{Y_{tj}^{\mathrm{E}}-Z_{tj}^{\mathrm{EM}},Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{CS}}-Z_{tj}^{\mathrm{OS}}\right\}$

- Conditional posterior distribution for Y_{tj}^{CS}:

$$
\begin{aligned}
&p(Y_{tj}^{\mathrm{CS}}|\boldsymbol{r}_{tj},\boldsymbol{Z}_{tj}^{\mathrm{S}},\boldsymbol{Z}_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},Y_{tj}^{\mathrm{ES}},Y_{tj}^{\cdot\mathrm{S}},Y_{tj}^{+},\boldsymbol{k}_{tj},N_j)\\
&\qquad\propto \binom{Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CS}}}\binom{Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{OS}}}\binom{Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CM}}}\binom{Y_{tj}^{+}-Y_{tj}^{\cdot\mathrm{S}}-(Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}})-(Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}}\\
&\qquad\qquad\times\binom{Y_{tj}^{\mathrm{C}}}{Y_{tj}^{\mathrm{CS}}}\left(r_{tj}^{\mathrm{C}}\right)^{Y_{tj}^{\mathrm{CS}}}\left(1-r_{tj}^{\mathrm{C}}\right)^{Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}}\\
&\qquad\qquad\times\binom{(Y_{tj}^{+}-Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{C}})}{Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}}\left(r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}})}\left(1-r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{+}-Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{C}})-(Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}})}
\end{aligned}
$$

  where

$$\max\left\{Z_{tj}^{\mathrm{CS}},Z_{tj}^{\mathrm{OM}}-Y_{tj}^{\cdot\mathrm{M}}+Y_{tj}^{\mathrm{C}}+Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}\right\}\leq Y_{tj}^{\mathrm{CS}}\leq\min\left\{Y_{tj}^{\mathrm{C}}-Z_{tj}^{\mathrm{CM}},Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Z_{tj}^{\mathrm{OS}}\right\}$$

- Y_{tj}^{OS} is deterministic: $Y_{tj}^{\mathrm{OS}}=Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}$

- Conditional posterior distribution for Y_{tj}^{E}:

$$
\begin{aligned}
&p(Y_{tj}^{\mathrm{E}}|\boldsymbol{\theta}_t, \boldsymbol{r}_{tj}, \boldsymbol{Z}_{tj}^{\mathrm{S}}, \boldsymbol{Z}_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{\mathrm{S}}, Y_{tj}^{\mathrm{C}}, Y_{tj}^{\cdot\mathrm{S}}, Y_{tj}^{+}, \boldsymbol{k}_{tj}, N_j) \\
&\qquad \propto \binom{Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{EM}}} \binom{Y_{tj}^{+} - Y_{tj}^{\cdot\mathrm{S}} - (Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}) - (Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}} \\
&\qquad\qquad \times \binom{Y_{tj}^{\mathrm{E}}}{Y_{tj}^{\mathrm{ES}}} \left(1 - r_{tj}^{\mathrm{E}}\right)^{Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}} \\
&\qquad\qquad \times \binom{Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}}}{Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}}} \left(r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}})} \left(1 - r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}}) - (Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}})} \\
&\qquad\qquad \times \frac{(\theta_t^{\mathrm{E}} E^{\mathrm{E}})^{Y_{tj}^{\mathrm{E}}} e^{-\theta_t^{\mathrm{E}} E^{\mathrm{E}}}}{Y_{tj}^{\mathrm{E}}!} \frac{(\theta_t^{\mathrm{O}} E^{\mathrm{O}})^{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}})} e^{-\theta_t^{\mathrm{O}} E^{\mathrm{O}}}}{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}})!}
\end{aligned}
$$

- Conditional posterior distribution for Y_{tj}^{C}:

$$
\begin{aligned}
&p(Y_{tj}^{\mathrm{C}}|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{p}\boldsymbol{Z}_{tj}^{\mathrm{S}}, \boldsymbol{Z}_{tj}^{\mathrm{M}}, \boldsymbol{Y}_{tj}^{\mathrm{S}}, Y_{tj}^{\mathrm{E}}, Y_{tj}^{\cdot\mathrm{S}}, Y_{tj}^{\cdot\mathrm{M}}, \boldsymbol{k}_{tj}, N_j) \\
&\qquad \propto \binom{Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CM}}} \binom{Y_{tj}^{+} - Y_{tj}^{\cdot\mathrm{S}} - (Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}) - (Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}} \\
&\qquad\qquad \times \binom{Y_{tj}^{\mathrm{C}}}{Y_{tj}^{\mathrm{CS}}} \left(1 - r_{tj}^{\mathrm{C}}\right)^{Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}}} \\
&\qquad\qquad \times \binom{Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}}}{Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}}} \left(r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}})} \left(1 - r_{tj}^{\mathrm{O}}\right)^{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}}) - (Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}})} \\
&\qquad\qquad \times \frac{(N_j \theta_t^{\mathrm{C}} p_j^{\mathrm{C}})^{Y_{tj}^{\mathrm{C}}} e^{-N_j \theta_t^{\mathrm{C}} p_j^{\mathrm{C}}}}{Y_{tj}^{\mathrm{C}}!} \frac{(\theta_t^{\mathrm{O}} E^{\mathrm{O}})^{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}})} e^{-\theta_t^{\mathrm{O}} E^{\mathrm{O}}}}{(Y_{tj}^{+} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}})!}
\end{aligned}
$$

- Y_{tj}^{O} is deterministic: $Y_{tj}^{\mathrm{O}} = Y_{tj}^{\cdot\mathrm{S}} + Y_{tj}^{\cdot\mathrm{M}} - Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{C}}$

- The conditional posterior distributions for severity probabilities r_{tj}^{G}'s are defined for

G=E, C, O as

$$\begin{aligned} p(r_{tj}^{\mathrm{G}}|\boldsymbol{Y}_{tj},\boldsymbol{Y}_{tj}^{\mathrm{S}}) &\propto \left(r_{tj}^{\mathrm{G}}\right)^{Y_{tj}^{\mathrm{GS}}}\left(1-r_{tj}^{\mathrm{G}}\right)^{Y_{tj}^{\mathrm{G}}-Y_{tj}^{\mathrm{GS}}}\left(r_{tj}^{\mathrm{G}}\right)^{a^{\mathrm{G}}-1}\left(1-r_{tj}^{\mathrm{G}}\right)^{b^{\mathrm{G}}} \\ &= \mathrm{Beta}\left(Y_{tj}^{\mathrm{GS}}+a^{\mathrm{G}}, Y_{tj}^{\mathrm{G}}-Y_{tj}^{\mathrm{GS}}+b^{\mathrm{G}}\right) \end{aligned}$$

- The conditional posterior distributions for weekly pathogen-specific relative risks θ_t^{G}'s are defined for G=E, C, O as

$$\begin{aligned} p(\theta_t^{\mathrm{G}}|\boldsymbol{p},\boldsymbol{Y},\boldsymbol{N}) &\propto (\theta_t^{\mathrm{G}})^{\alpha_t^{\mathrm{G}}-1}\, e^{-\beta_t^{\mathrm{G}}\theta_t^{\mathrm{G}}}\, (\theta_t^{\mathrm{G}})^{\sum_{j=1}^{J} Y_{tj}^{\mathrm{G}}}\, e^{-\theta_t^{\mathrm{G}}\sum_{j=1}^{J} N_j p_j^{\mathrm{G}}} \\ &= \mathrm{Gamma}\left(\alpha_t^{\mathrm{G}}+Y_t^{\mathrm{G}},\ \beta_t^{\mathrm{G}}+E^{\mathrm{G}}\right) \end{aligned}$$

To determine the support of the conditional distributions, we notice that $Y_{tj}^{\mathrm{ES}}+Y_{tj}^{\mathrm{CS}}+Y_{tj}^{\mathrm{OS}}=Y_{tj}^{\cdot\mathrm{S}}$ and $Y_{tj}^{\mathrm{GM}}=Y_{tj}^{\mathrm{G}}-Y_{tj}^{\mathrm{GS}}$, so that we can rewrite $p(\boldsymbol{Z}_{tj}^{\mathrm{M}}|k_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj}^{\mathrm{M}})$ in terms of $\boldsymbol{Y}_{tj}$ and $\boldsymbol{Y}_{tj}^{\mathrm{S}}$. In particular,

$$p(\boldsymbol{Z}_{tj}^{\mathrm{M}}|k_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},\boldsymbol{Y}_{tj}^{\mathrm{S}}) = \binom{Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{EM}}}\binom{Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CM}}}\binom{Y_{tj}^{\mathrm{O}}-Y_{tj}^{\mathrm{OS}}}{Z_{tj}^{\mathrm{OM}}}\Big/\binom{Y_{tj}^{\cdot\mathrm{M}}}{k_{tj}^{\mathrm{M}}}.$$

Let $Y_{tj}^{+}=Y_{tj}^{\mathrm{E}}+Y_{tj}^{\mathrm{C}}+Y_{tj}^{\mathrm{O}}$. Then,

$$p(\boldsymbol{Z}_{tj}^{\mathrm{M}}|k_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},\boldsymbol{Y}_{tj}^{\mathrm{S}}) = \binom{Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{EM}}}\binom{Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CM}}}\binom{Y_{tj}^{+}-Y_{tj}^{\cdot\mathrm{S}}-(Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}})-(Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}}\Big/\binom{Y_{tj}^{\cdot\mathrm{M}}}{k_{tj}^{\mathrm{M}}}.$$

Therefore,

$$\begin{aligned} p(\boldsymbol{Z}_{tj}^{\mathrm{S}},&\boldsymbol{Z}_{tj}^{\mathrm{M}}|k_{tj}^{\mathrm{S}},k_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},\boldsymbol{Y}_{tj}^{\mathrm{S}}) = p(\boldsymbol{Z}_{tj}^{\mathrm{S}}|k_{tj}^{\mathrm{S}},\boldsymbol{Y}_{tj}^{\mathrm{S}})p(\boldsymbol{Z}_{tj}^{\mathrm{M}}|k_{tj}^{\mathrm{M}},\boldsymbol{Y}_{tj},\boldsymbol{Y}_{tj}^{\mathrm{S}}) \\ &= \binom{Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{ES}}}\binom{Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CS}}}\binom{Y_{tj}^{\cdot\mathrm{S}}-Y_{tj}^{\mathrm{ES}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{OS}}}\Big/\binom{Y_{tj}^{\cdot\mathrm{S}}}{k_{tj}^{\mathrm{S}}} \\ &\times \binom{Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}}}{Z_{tj}^{\mathrm{EM}}}\binom{Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}}}{Z_{tj}^{\mathrm{CM}}}\binom{Y_{tj}^{\cdot\mathrm{M}}-(Y_{tj}^{\mathrm{E}}-Y_{tj}^{\mathrm{ES}})-(Y_{tj}^{\mathrm{C}}-Y_{tj}^{\mathrm{CS}})}{Z_{tj}^{\mathrm{OM}}}\Big/\binom{Y_{tj}^{\cdot\mathrm{M}}}{k_{tj}^{\mathrm{M}}}. \end{aligned}$$

From these two multivariate hypergeometric distributions, we can find the domain for Y_{tj}^{ES}

and Y_{tj}^{CS}. For example, we can find the domain for Y_{tj}^{ES} from the following limits

$$\begin{aligned}
Z_{tj}^{\mathrm{ES}} &\leq Y_{tj}^{\mathrm{ES}} \\
Z_{tj}^{\mathrm{EM}} &\leq Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}} \\
Z_{tj}^{\mathrm{OS}} &\leq Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}} \\
Z_{tj}^{\mathrm{OM}} &\leq Y_{tj}^{\cdot\mathrm{M}} - (Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}) - (Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}}),
\end{aligned}$$

which implies

$$\max\left\{Z_{tj}^{\mathrm{ES}}, Z_{tj}^{\mathrm{OM}} - Y_{tj}^{\cdot\mathrm{M}} + Y_{tj}^{\mathrm{E}} + Y_{tj}^{\mathrm{C}} - Y_{tj}^{\mathrm{CS}}\right\} \leq Y_{tj}^{\mathrm{ES}} \leq \min\left\{Y_{tj}^{\mathrm{E}} - Z_{tj}^{\mathrm{EM}}, Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{CS}} - Z_{tj}^{\mathrm{OS}}\right\}.$$

Similarly, for Y_{tj}^{CS}, we find

$$\max\left\{Z_{tj}^{\mathrm{CS}}, Z_{tj}^{\mathrm{OM}} - Y_{tj}^{\cdot\mathrm{M}} + Y_{tj}^{\mathrm{C}} + Y_{tj}^{\mathrm{E}} - Y_{tj}^{\mathrm{ES}}\right\} \leq Y_{tj}^{\mathrm{CS}} \leq \min\left\{Y_{tj}^{\mathrm{C}} - Z_{tj}^{\mathrm{CM}}, Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Z_{tj}^{\mathrm{OS}}\right\}.$$

And while Y_{tj}^{OS} can be determined deterministically, i.e. $Y_{tj}^{\mathrm{OS}} = Y_{tj}^{\cdot\mathrm{S}} - Y_{tj}^{\mathrm{ES}} - Y_{tj}^{\mathrm{CS}}$, it must also fall within similar bounds. The multivariate hypergeometric distributions also imposes bounds for Y_{tj}^{E} and Y_{tj}^{C}:

$$\max\left\{0, Z_{tj}^{\mathrm{EM}} + Y_{tj}^{\mathrm{ES}}\right\} \leq Y_{tj}^{\mathrm{E}} \leq Y_{tj}^{\cdot\mathrm{M}} - Y_{tj}^{\mathrm{CM}} - Z_{tj}^{\mathrm{OM}} + Y_{tj}^{\mathrm{ES}},$$

and

$$\max\left\{0, Z_{tj}^{\mathrm{CM}} + Y_{tj}^{\mathrm{CS}}\right\} \leq Y_{tj}^{\mathrm{C}} \leq Y_{tj}^{\cdot\mathrm{M}} - Y_{tj}^{\mathrm{EM}} - Z_{tj}^{\mathrm{OM}} + Y_{tj}^{\mathrm{CS}}.$$

A.3 Detailed results of MCMC simulation study

Scenario 1

In Scenario 1, we have a population of 5,000 and the disease and severity probabilities are higher than what was observed in the China HFMD data set. Here, we expect the MCMC to run quickly and perform well because we have a large number of observed cases, in a relatively small population.

We implemented the MCMC model in `R` with 250,000 iterations being carried out. We discarded the first 200,000 iterations as burn-in, and used the remaining MCMC samples for inference. Trace plots showed convergence of the MCMC chains. The MCMC procedure took approximately 40 minutes, while the hybrid procedure produced estimates in 0.08 seconds. In Figure A.2, we plot the estimates and 95% credible intervals over time; true disease count are in black, MCMC estimates are in blue, and hybrid estimates in red. We see that the credible intervals tend to be very similar in length. Figure A.3 compares the estimated unobserved pathogen-specific disease counts obtained using the two procedures; we plot the hybrid estimates against the MCMC estimates and see that the two procedures produce very similar estimates.

We also consider how well each method estimates the pathogen-specific log relative risks ($\log \theta_t^G$). We summarize these results in Figure A.4, where we see that overall, the hybrid approach produces results that are comparable to those obtained via MCMC. In general, the estimates obtained using the MCMC procedure produced credible intervals that are slightly narrower than the intervals obtained using the hybrid procedure. As a result, the coverage for hybrid intervals of the unobserved disease counts tends to be slightly higher than the coverage using MCMC intervals. However, when estimating pathogen-specific log relative risks ($\log \theta_t^G$'s), the hybrid procedure uses a normal approximation, which results in slightly worse coverage for these intervals compared to the MCMC credible intervals. Note that coverage probabilities are over a single simulation consisting of six months of weekly data. These results are summarized in Table A.1.

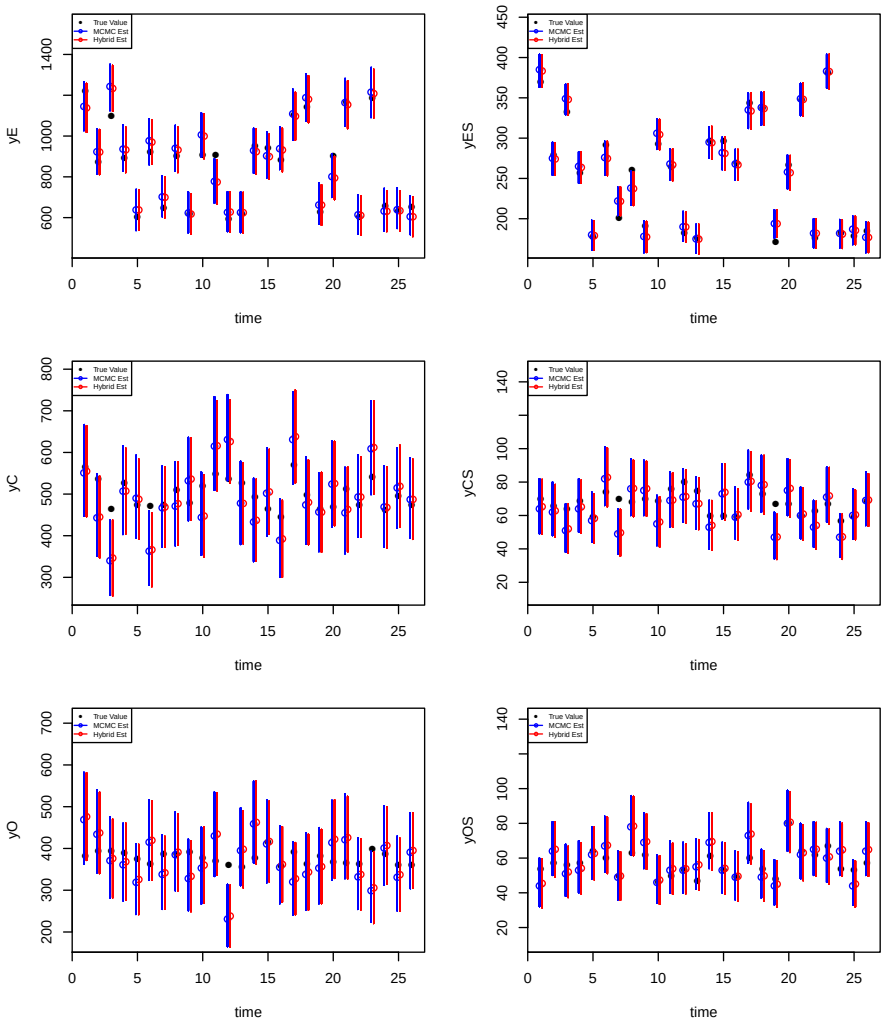


Figure A.2: Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches in Scenario 1.

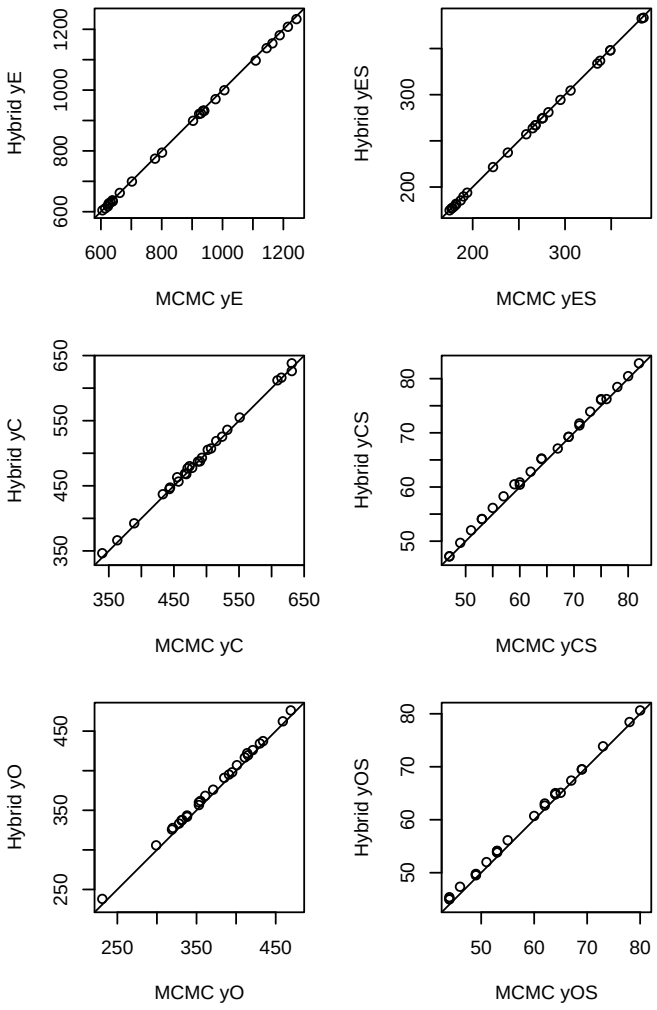


Figure A.3: Comparison of estimated pathogen-specific disease counts using different estimation approaches in Scenario 1.

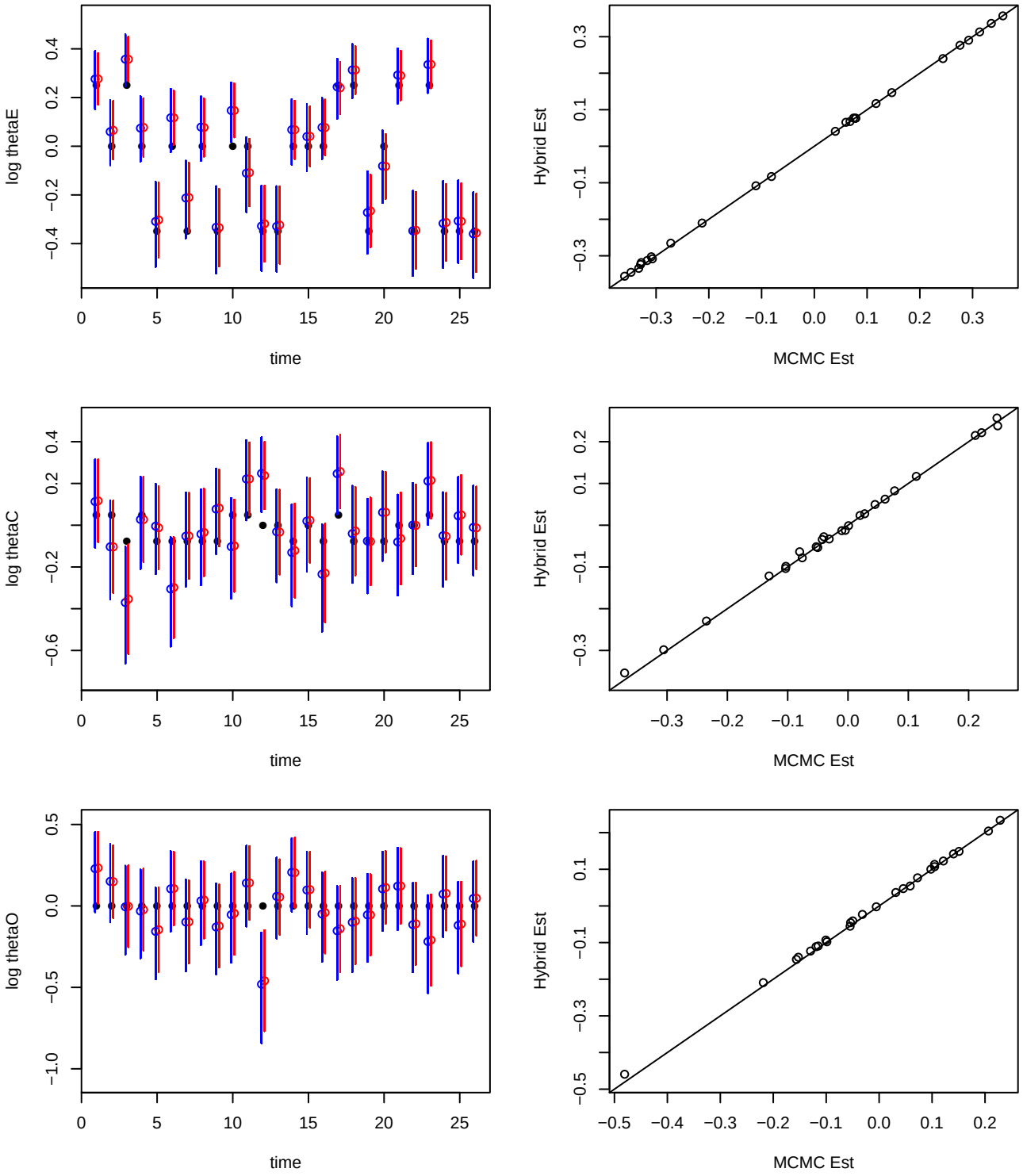


Figure A.4: Comparison of estimated pathogen-specific log relative risks using different estimation approaches, for Scenario 1.

	MSE		Scaled MSE		Prop of CI's cover	
	Hybrid	MCMC	Hybrid	MCMC	Hybrid	MCMC
Y_{tj}^{E}	3311.31	3433.46	3.59	3.71	0.885	0.923
Y_{tj}^{C}	3198.55	3353.92	6.38	6.70	0.923	0.923
Y_{tj}^{O}	2855.22	2950.15	7.60	7.85	0.923	0.923
Y_{tj}^{ES}	125.06	126.08	0.53	0.54	0.885	0.885
Y_{tj}^{CS}	80.22	83.62	1.19	1.24	0.923	0.923
Y_{tj}^{OS}	46.13	44.50	0.80	0.77	1.000	1.000
	MSE ($\times 10^3$)				Prop of CI's cover	
	Hybrid	MCMC			Hybrid	MCMC
$\log \theta_t^{\mathrm{E}}$	5.44	5.30			0.885	0.962
$\log \theta_t^{\mathrm{C}}$	16.85	17.42			0.885	0.885
$\log \theta_t^{\mathrm{O}}$	21.58	22.91			0.923	0.962

Table A.1: Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 1. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12).

Scenario 2

In this scenario, we compare the two approaches with a slightly larger population (50,000), and now with disease and severity probabilities more like what we observe in the China HFMD data. The MCMC procedure requires 250,000 iterations to converge and takes 38.7 minutes; the hybrid procedure takes 0.1 seconds to estimate both the unobserved disease counts and the pathogen-specific log relative risks. Table A.2 summarizes the MSE, scaled MSE, and proportion of covering intervals for both methods. The MSE's are very similar, as are the coverages over a single dataset, with the hybrid procedure being slightly favored.

Figure A.5 compares the estimated unobserved pathogen-specific disease counts obtained using the two procedures; we plot the hybrid estimates against the MCMC estimates and see that the two procedures produce very similar estimates. In Figure A.6, we plot the estimates and 95% credible intervals over time; true disease count are in black, MCMC estimates are in blue, and hybrid estimates in red. We compare the estimates obtained via both procedures in Figure A.5. When counts are small, the hybrid procedure tends to somewhat over estimate the unobserved counts compared to the MCMC estimates.

	MSE		Scaled MSE		Prop of CI's cover	
	Hybrid	MCMC	Hybrid	MCMC	Hybrid	MCMC
Y_{tj}^{E}	3867.01	4142.88	4.75	5.11	0.923	0.923
Y_{tj}^{C}	3489.15	3724.23	6.79	7.24	0.923	0.885
Y_{tj}^{O}	2492.85	2561.27	6.52	6.73	0.962	0.962
Y_{tj}^{ES}	32.16	30.00	0.13	0.12	0.962	0.885
Y_{tj}^{CS}	21.84	20.81	1.36	1.29	0.923	0.885
Y_{tj}^{OS}	10.50	10.38	1.80	1.78	0.885	0.808
	MSE ($\times 10^3$)				Prop of CI's cover	
	Hybrid	MCMC			Hybrid	MCMC
$\log \theta_t^{\mathrm{E}}$	10.54	10.56			0.769	0.885
$\log \theta_t^{\mathrm{C}}$	14.82	15.80			0.923	0.923
$\log \theta_t^{\mathrm{O}}$	25.59	27.12			0.923	0.923

Table A.2: Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 2. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12).

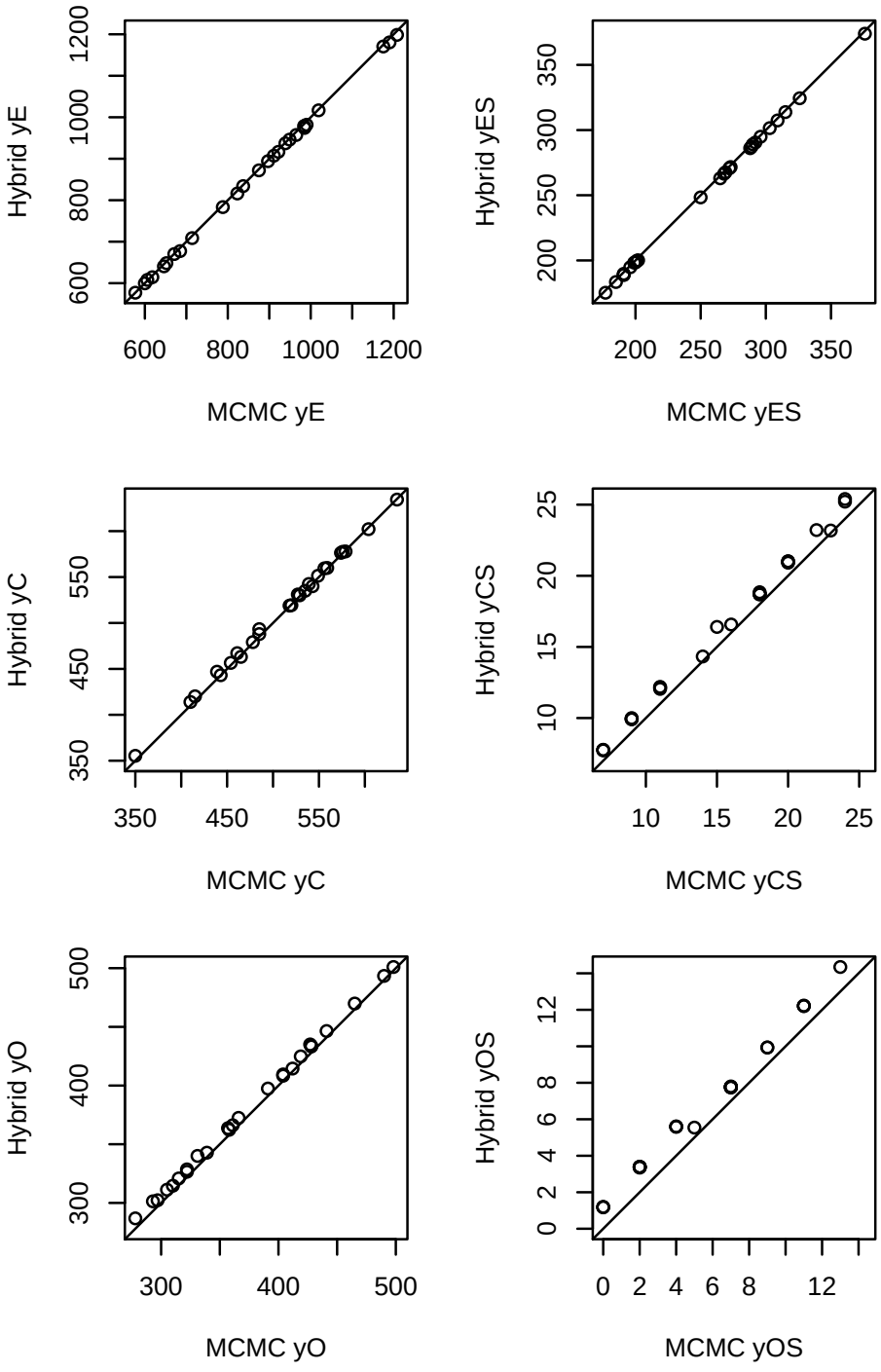


Figure A.5: Comparison of estimated pathogen-specific disease counts using different estimation approaches in Scenario 2.

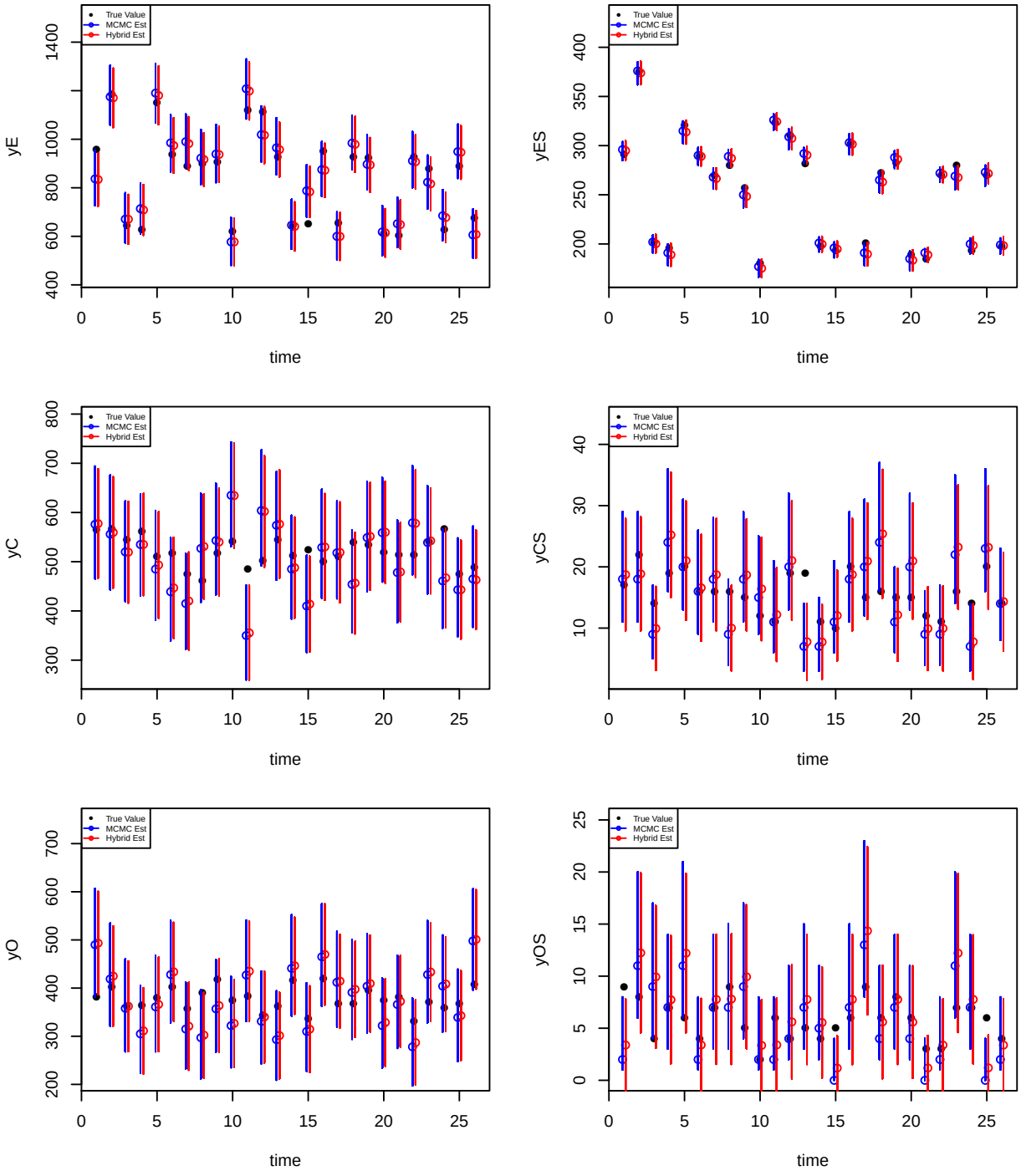


Figure A.6: Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches for Scenario 2.

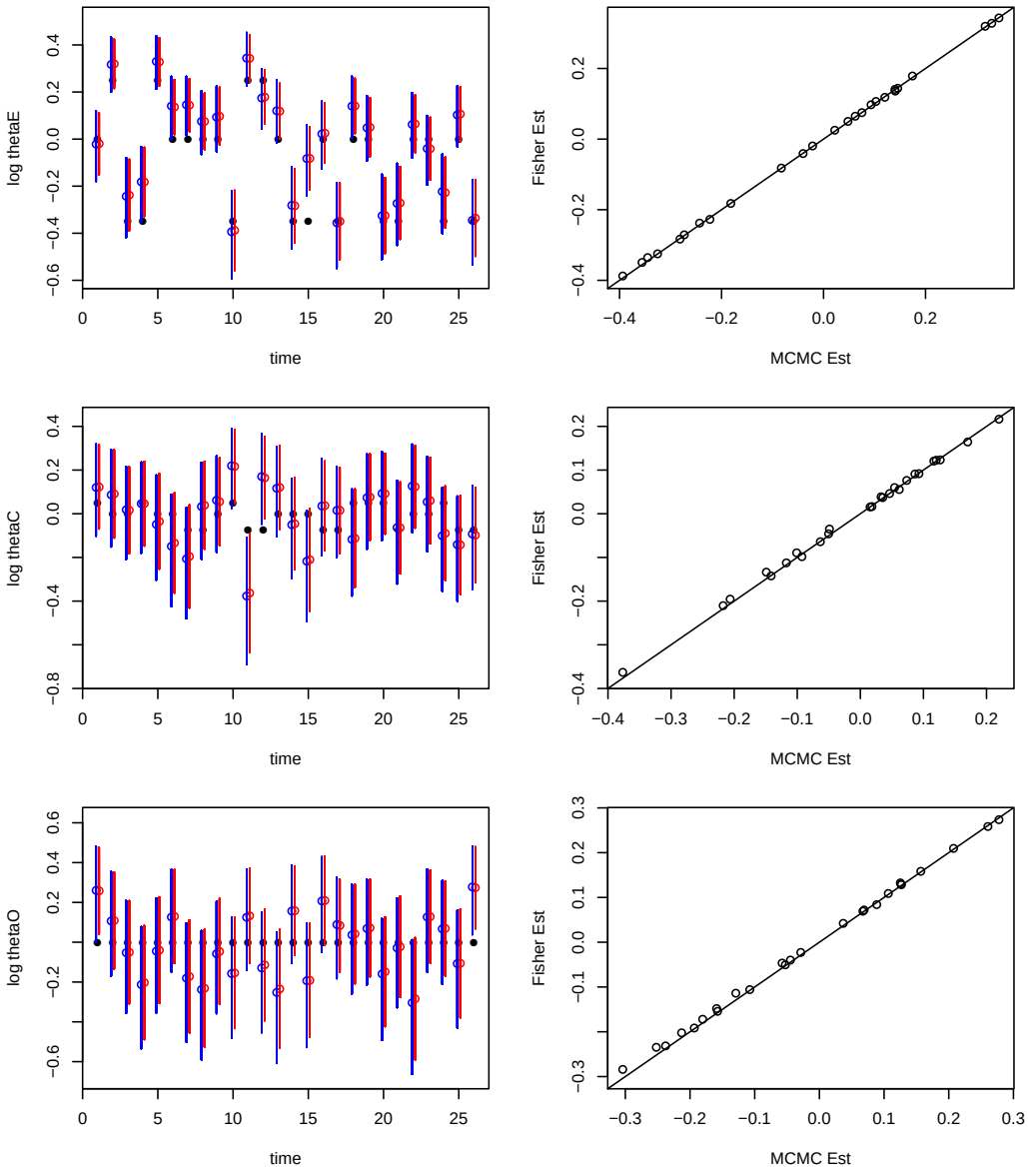


Figure A.7: Comparison of estimated pathogen-specific log relative risks using different estimation approaches in Scenario 2.

Scenario 3

For this scenario, we consider a population of 500,000, and use disease and severity probabilities that are closer to those observed in the HFMD data. Here, 250,000 MCMC iterations took 39 minutes, and the final 50,000 iterations were used for posterior estimates. In contrast, the hybrid procedure takes 0.06 seconds to estimate both the unobserved disease counts and the pathogen-specific log relative risks. Table A.3 summarizes the MSE, scaled MSE, and proportion of covering intervals for both methods. Again, the hybrid procedure produces estimates that are very similar to those obtained by the MCMC.

In Figure A.8, we plot the estimates and 95% credible intervals for the unobserved pathogen-specific disease counts over time; true counts are in black, MCMC estimates are in blue, and hybrid estimates in red. Figure A.9 compares the estimated unobserved pathogen-specific disease counts obtained using the two procedures; we plot the hybrid estimates against the MCMC estimates and see that the two procedures produce very similar estimates when disease counts are large. When counts are small, the hybrid procedure tends to somewhat over estimate the unobserved counts. Lastly, in Figure A.10, we compare the estimates of the pathogen-specific log relative risks. We see that the two approaches yield very similar estimates, in terms of MSE and coverage only slightly favors the MCMC procedure for the single simulated population.

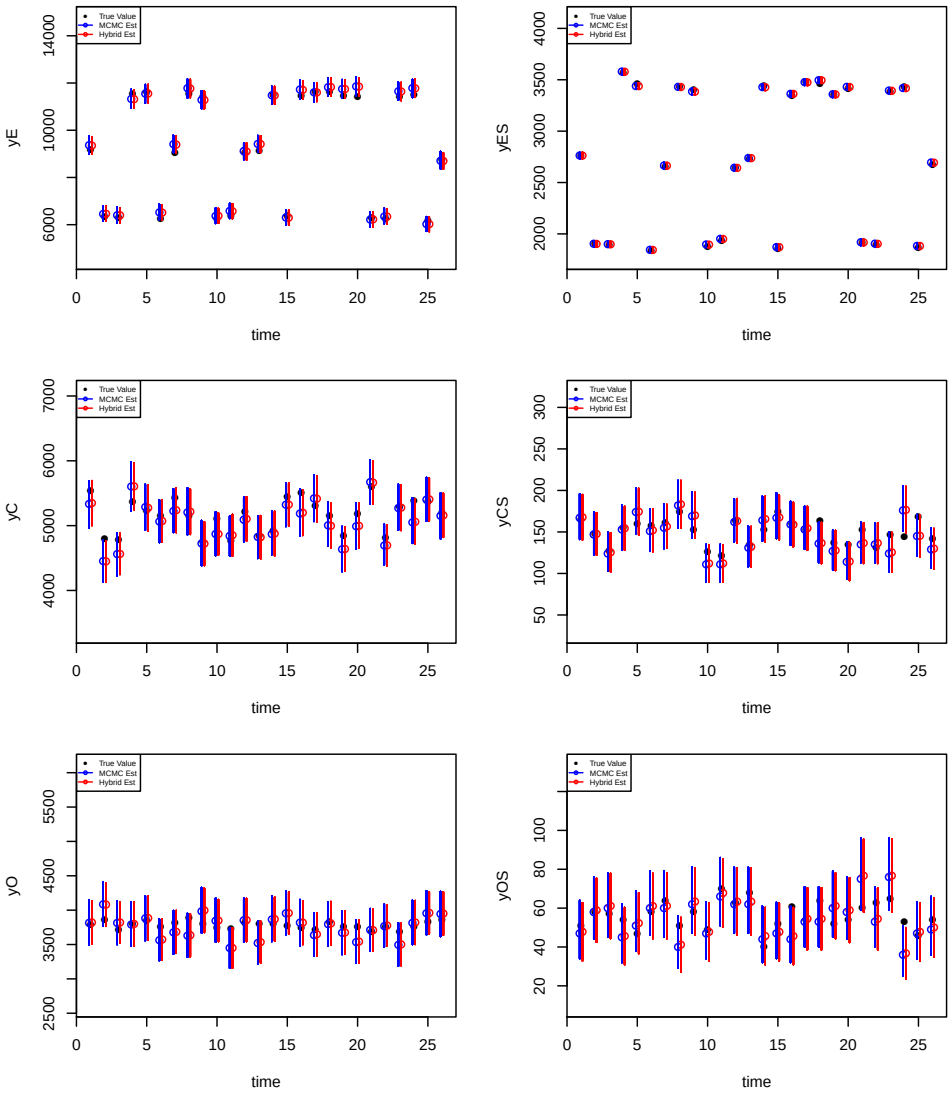


Figure A.8: Comparison of estimated pathogen-specific disease counts over time, using different estimation approaches, for Scenario 3.

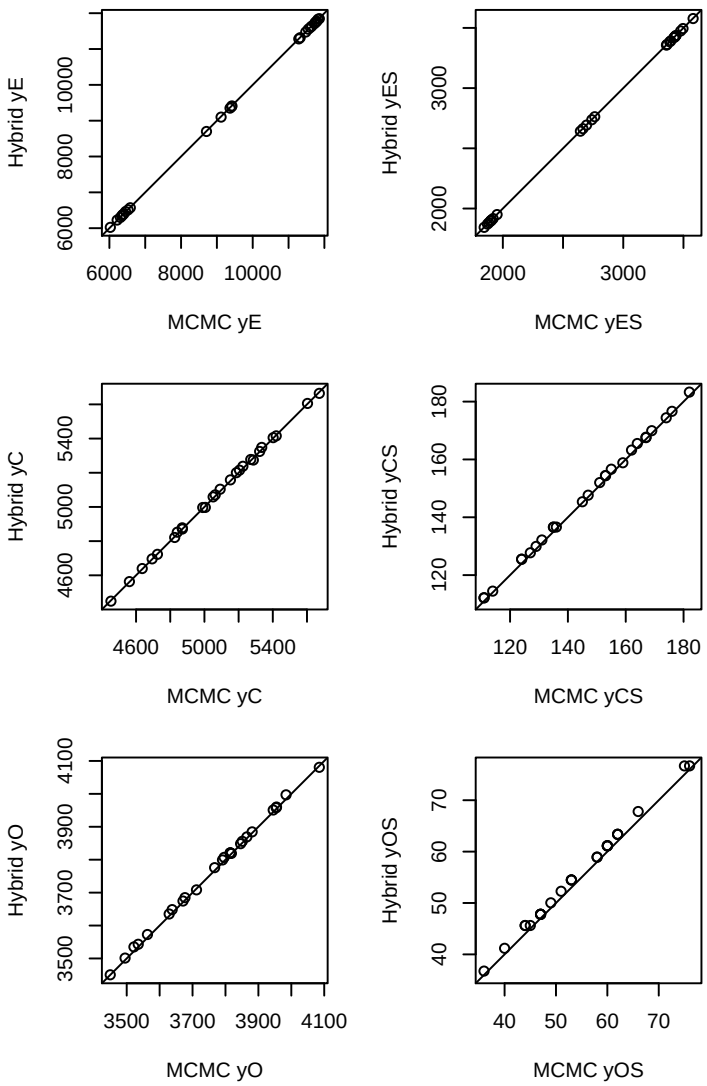


Figure A.9: Comparison of estimated pathogen-specific disease counts using different estimation approaches with in Scenario 3.

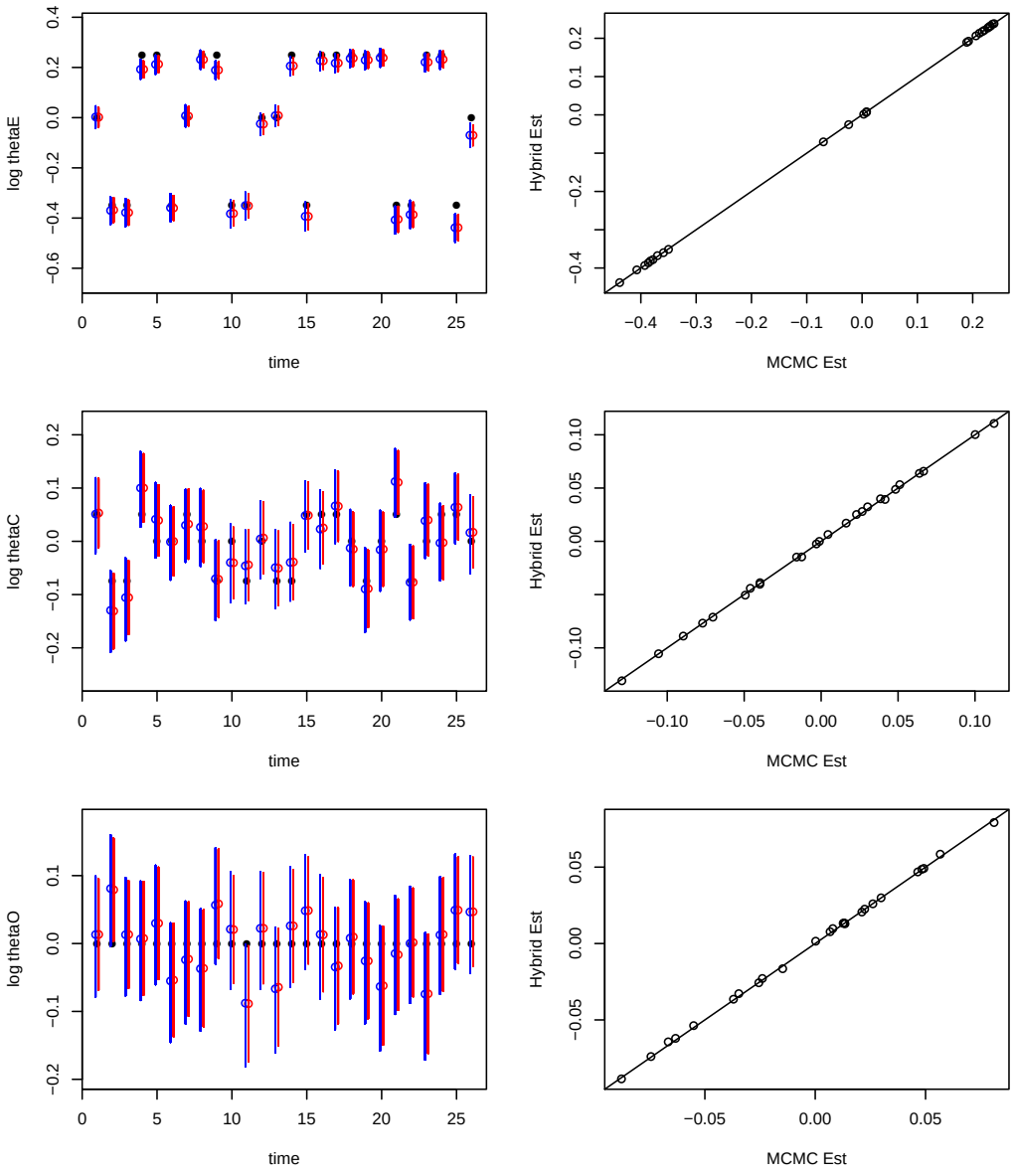


Figure A.10: Comparison of estimated pathogen-specific log relative risks using different estimation approaches, for Scenario 3.

	MSE		Scaled MSE		Prop of CI's cover	
	Hybrid	MCMC	Hybrid	MCMC	Hybrid	MCMC
Y_{tj}^{E}	39143.05	41337.31	4.19	4.42	0.962	0.962
Y_{tj}^{C}	28479.67	29594.69	5.51	5.71	0.962	0.962
Y_{tj}^{O}	21145.40	21692.35	5.58	5.72	1.000	1.000
Y_{tj}^{ES}	184.58	199.23	0.06	0.07	0.962	0.962
Y_{tj}^{CS}	185.92	193.12	1.25	1.30	0.923	0.923
Y_{tj}^{OS}	59.70	62.65	1.05	1.09	0.923	0.923
	MSE ($\times 10^3$)				Prop of CI's cover	
	Hybrid	MCMC			Hybrid	MCMC
$\log \theta_t^{\mathrm{E}}$	1.38	1.40			0.731	0.731
$\log \theta_t^{\mathrm{C}}$	0.91	0.92			0.962	1.000
$\log \theta_t^{\mathrm{O}}$	1.85	1.87			0.923	0.962

Table A.3: Comparison of MSE, scaled MSE, and proportion of covering intervals for estimates obtained using different methods in Scenario 3. Coverage probabilities are over a single simulated population. Scaled MSE is the average of squared errors divided by the true value, as in (3.12).

A.4 Variance adjustments

The variance and covariances presented in the chapter were conditional on both the observed subsampled counts (the Z_{tj}^{GS}'s) and the total number of severe (or mild) cases observed ($Y_{tj}^{\cdot\mathrm{S}}$). However, these fail to take into account the variability of the total number of severe cases for a given stratum and week. Therefore, we derive the unconditional variances to be used.

$$\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right] = \sum_{\mathrm{G}} \mathrm{E}\left[\mathrm{E}\left[Y_{tj}^{\mathrm{GS}}|Y_{tj}^{\mathrm{G}}\right]\right] = \sum_{\mathrm{G}} \mathrm{E}\left[Y_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}\right] = \sum_{\mathrm{G}} N_j s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}} \tag{A.4}$$

Since we do not know the truth, we estimate

$$\widehat{s_{tj}^{\mathrm{G}}} = Y_{tj}^{\mathrm{G}}/N_j \qquad \text{and} \qquad \widehat{r_{tj}^{\mathrm{G}}} = Y_{tj}^{\mathrm{GS}}/Y_{tj}^{\mathrm{G}}.$$

As expected, we estimate $\widehat{\mathrm{E}}\left[Y_{tj}^{\cdot\mathrm{S}}\right] = Y_{tj}^{\cdot\mathrm{S}}$. We can similarly estimate the unconditional variance of the total number of severe cases as follows.

$$\begin{aligned}
\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right] &= \mathrm{Var}\left[\sum_{\mathrm{G}} Y_{tj}^{\mathrm{GS}}\right] \\
&= \sum_{\mathrm{G}} \mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}\right] + 2\sum_{\mathrm{G}\neq\mathrm{G}'} \mathrm{Cov}\left[Y_{tj}^{\mathrm{GS}}, Y_{tj}^{\mathrm{G}'\mathrm{S}}\right] \\
&= \sum_{\mathrm{G}} \mathrm{Var}\left[\mathrm{E}\left[Y_{tj}^{\mathrm{GS}}|Y_{tj}^{\mathrm{G}}\right]\right] + \mathrm{E}\left[\mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}|Y_{tj}^{\mathrm{G}}\right]\right] + 2\sum_{\mathrm{G}\neq\mathrm{G}'} \mathrm{Cov}\left[\mathrm{E}\left[Y_{tj}^{\mathrm{GS}}|Y_{tj}^{\mathrm{G}}\right], \mathrm{E}\left[Y_{tj}^{\mathrm{G}'\mathrm{S}}|Y_{tj}^{\mathrm{G}'}\right]\right] \\
&= \sum_{\mathrm{G}} \mathrm{Var}\left[Y_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}\right] + \mathrm{E}\left[Y_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}(1 - r_{tj}^{\mathrm{G}})\right] + 2\sum_{\mathrm{G}\neq\mathrm{G}'} \mathrm{Cov}\left[Y_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}, Y_{tj}^{\mathrm{G}'} r_{tj}^{\mathrm{G}'}\right] \\
&= \sum_{\mathrm{G}} N_j s_{tj}^{\mathrm{G}}(1 - s_{tj}^{\mathrm{G}})(r_{tj}^{\mathrm{G}})^2 + N_j s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}(1 - r_{tj}^{\mathrm{G}}) - 2\sum_{\mathrm{G}\neq\mathrm{G}'} N_j s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}} s_{tj}^{\mathrm{G}'} r_{tj}^{\mathrm{G}'} \\
&= \sum_{\mathrm{G}} N_j s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}(1 - s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}}) - 2\sum_{\mathrm{G}\neq\mathrm{G}'} N_j s_{tj}^{\mathrm{G}} r_{tj}^{\mathrm{G}} s_{tj}^{\mathrm{G}'} r_{tj}^{\mathrm{G}'}
\end{aligned}$$

Again, we estimate the variance with

$$\widehat{\mathrm{Var}}\left[Y_{tj}^{\cdot\mathrm{S}}\right] = \sum_{\mathrm{G}} Y_{tj}^{\mathrm{GS}}\left(1 - \frac{Y_{tj}^{\mathrm{GS}}}{N_j}\right) - \frac{2}{N_j}\sum_{\mathrm{G}\neq\mathrm{G}'} Y_{tj}^{\mathrm{GS}} Y_{tj}^{\mathrm{G}'\mathrm{S}}$$

A similar procedure produces estimates for the total number of mild cases. We can then use these estimates to find the marginal variance of our estimated pathogen- and severity-specific disease counts. In particular,

$$\mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]=\mathrm{E}\left[\mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},Y_{tj}^{\cdot\mathrm{S}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]+\mathrm{Var}\left[\mathrm{E}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},Y_{tj}^{\cdot\mathrm{S}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right] \tag{A.5}$$

For ease, we estimate these two parts separately, and then re-combine them at the end.

$$\begin{aligned}
&\mathrm{E}\left[\ \mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},Y_{tj}^{\cdot\mathrm{S}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\\
&\quad=\mathrm{E}\left[\frac{\left(Y_{tj}^{\cdot\mathrm{S}}-k_{tj}^{\mathrm{S}}\right)\left(\alpha_{j}^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(Y_{tj}^{\cdot\mathrm{S}}+\alpha_{j}^{\cdot\mathrm{S}}\right)\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}-\alpha_{j}^{\mathrm{GS}}-Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)^{2}\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\\
&\quad=\frac{\left(\alpha_{j}^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}-\alpha_{j}^{\mathrm{GS}}-Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)^{2}\left(\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\left(\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right]+\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]^{2}-\left(k_{tj}^{\mathrm{S}}-\alpha_{j}^{\cdot\mathrm{S}}\right)\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]-k_{tj}^{\mathrm{S}}\alpha_{j}^{\cdot\mathrm{S}}\right)
\end{aligned}$$

The second part of (A.5) is

$$\begin{aligned}
\mathrm{Var}\left[\mathrm{E}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},Y_{tj}^{\cdot\mathrm{S}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]&=\mathrm{Var}\left[Z_{tj}^{\mathrm{GS}}+\left(Y_{tj}^{\cdot\mathrm{S}}-k_{tj}^{\mathrm{S}}\right)\left(\frac{\alpha_{j}^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}}{\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}}\right)\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\\
&=\left(\frac{\alpha_{j}^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}}{\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}}\right)^{2}\mathrm{Var}\left[\left(Y_{tj}^{\cdot\mathrm{S}}-k_{tj}^{\mathrm{S}}\right)\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_{j}^{\mathrm{S}}\right]\\
&=\left(\frac{\alpha_{j}^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}}{\alpha_{j}^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}}\right)^{2}\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right]
\end{aligned}$$

Recombining these two parts, we continue with (A.5), to find

$$\begin{aligned}
\mathrm{Var}\left[Y_{tj}^{\mathrm{GS}}\middle|Z_{tj}^{\mathrm{GS}},\boldsymbol{\alpha}_j^{\mathrm{S}}\right] &= \frac{\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}-\alpha_j^{\mathrm{GS}}-Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)^2\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\left(\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]-k_{tj}^{\mathrm{S}}\right)\left(\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]+\alpha_j^{\cdot\mathrm{S}}\right) \\
&\quad+\frac{\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}-\alpha_j^{\mathrm{GS}}-Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)^2\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right]+\left(\frac{\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}}{\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}}\right)^2\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right] \\
&= \frac{\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}-\alpha_j^{\mathrm{GS}}-Z_{tj}^{\mathrm{GS}}\right)}{\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)^2\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\left(\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]-k_{tj}^{\mathrm{S}}\right)\left(\mathrm{E}\left[Y_{tj}^{\cdot\mathrm{S}}\right]+\alpha_j^{\cdot\mathrm{S}}\right) \\
&\quad+\frac{\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}+1\right)}{\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right] \\
&= \mathrm{Var}[\widetilde{Y_{tj}^{\mathrm{GS}}}|Z_{tj}^{\mathrm{GS}}]+\frac{\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}\right)\left(\alpha_j^{\mathrm{GS}}+Z_{tj}^{\mathrm{GS}}+1\right)}{\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}\right)\left(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1\right)}\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right]
\end{aligned}$$

The same can be done for mild cases. We also derive a similar adjustment for covariances. For example

$$\mathrm{Cov}\left[Y_{tj}^{\mathrm{ES}},Y_{tj}^{\mathrm{CS}}|\boldsymbol{Z}_{tj}^{\mathrm{S}}\right]=\mathrm{Cov}\left[Y_{tj}^{\mathrm{ES}},Y_{tj}^{\mathrm{CS}}|Y_{tj}^{\cdot\mathrm{S}},\boldsymbol{Z}_{tj}^{\mathrm{S}}\right]+\frac{(\alpha_j^{\mathrm{ES}}+Z_{tj}^{\mathrm{ES}})(\alpha_j^{\mathrm{CS}}+Z_{tj}^{\mathrm{CS}})}{(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}})(\alpha_j^{\cdot\mathrm{S}}+k_{tj}^{\mathrm{S}}+1)}\mathrm{Var}\left[Y_{tj}^{\cdot\mathrm{S}}\right].$$

A.5 Simulations

We conduct simulation studies to investigate the performance of the proposed methods, which we refer to as the hybrid method, and the impact of the various modeling assumptions. We investigate the performance of unobserved disease counts under various subsampling procedures. We find that estimates from the hybrid method had approximately 95% coverage, even when few cases are subsampled for virology. Coverage is somewhat worse when the true number of cases is very small. Not surprisingly, when there are no subsamples for a given stratum or week, the resulting estimates have poor coverage.

The hybrid method estimates the true pathogen-specific log relative risk for the two primary pathogens of interest well, and corresponding standard error estimates yield intervals with close to 95% coverage. When the proportionality assumption is not valid, the hybrid method yields good estimates of the weighted average pathogen-specific log relative risk. We consider the impact of modeling the two pathogens jointly. We find that accounting for the correlation induced by subsampling between the two pathogens by modeling them jointly yielded better estimates of the true effects, as well as more consistent estimation of the underlying temporal effects. When we consider estimates of the covariate effect, we obtain slightly biased estimates of the true effect. However, the estimates obtained when modeling the two pathogens together yield covariate effect estimates with a smaller mean squared error (MSE) compared to the MSE from modeling the two pathogens separately. Details of these analyses are described below.

A.5.1 Simulated population setup

We simulate 3 years of disease surveillance for a population with 4 strata, where the strata sizes are similar to those in China. We first simulate the true number of pathogen-specific disease counts, followed by the number of severe cases for the entire population.

For each stratum and week, we generate pathogen- and stratum-specific disease counts by severity as follows:

$$Y_{tj}^{\mathrm{G}} | s_{tj}^{\mathrm{G}} \sim \text{Poisson}(N_j s_{tj}^{\mathrm{G}}),$$
$$Y_{tj}^{\mathrm{GS}} | Y_{tj}^{\mathrm{G}}, r_j^{\mathrm{G}} \sim \text{Binomial}(Y_{tj}^{\mathrm{G}}, r_j^{\mathrm{G}}),$$

where N_j is the stratum-specific population size s_{tj}^{G} is the pathogen- and stratum-specific disease probability, and r_j^{G} is the pathogen- and stratum-specific severe disease probability. We start by assuming the pathogen-specific disease probabilities can be written as $s_{tj}^{\mathrm{G}} = \theta_t^{\mathrm{G}} \times p_j^{\mathrm{G}}$, although this will change later on. Table A.4 presents the specific pathogen- and stratum- specific values used in simulations based on the HFMD data. For each stratum, EV71 is simulated to account for the the majority of cases observed at any time, while CA16 and Other pathogens each make up a smaller proportion, with CA16 having a slightly higher probability.

Stratum	N_j/N	p_j^{E}	p_j^{C}	p_j^{O}	r_{tj}^{E}	r_{tj}^{C}	r_{tj}^{O}
1	0.018	0.01800	0.010440	0.007560	0.30	0.030	0.0150
2	0.019	0.02000	0.011600	0.008400	0.30	0.030	0.0150
3	0.469	0.00050	0.000290	0.000210	0.27	0.027	0.0135
4	0.494	0.00075	0.000435	0.000315	0.27	0.027	0.0135

Table A.4: Summary of stratum-specific parameter values for simulation

Pathogen-specific relative risks for three years of weekly data are generated as follows:

$$\log \theta_t^{\mathrm{G}} = \sum_{k=1}^{2} b_k^{\mathrm{G}} \sin(\omega_k t - \pi/4) + c_k^{\mathrm{G}} \cos(\omega_k t - \pi/4) - d^{\mathrm{G}}$$

where $\omega_k = 2k\pi/52.25$, both b_k^{G} and c_k^{G} are pathogen-specific constants, and d^{G} is a constant to ensure $\sum_T \theta_t^{\mathrm{G}} = T$. Figure A.11 shows the true pathogen-specific log relative risks over time, when pathogen-specific disease probabilities can be written as $s_{tj}^{\mathrm{G}} = \theta_t^{\mathrm{G}} \times p_j^{\mathrm{G}}$. Subsamples

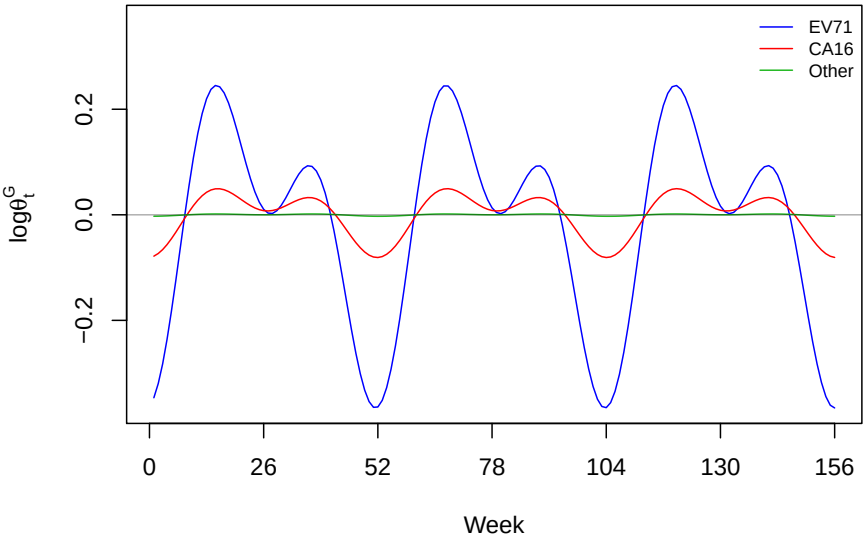


Figure A.11: Pathogen-specific log relative risks over time used for simulations, when the proportionality assumption is valid.

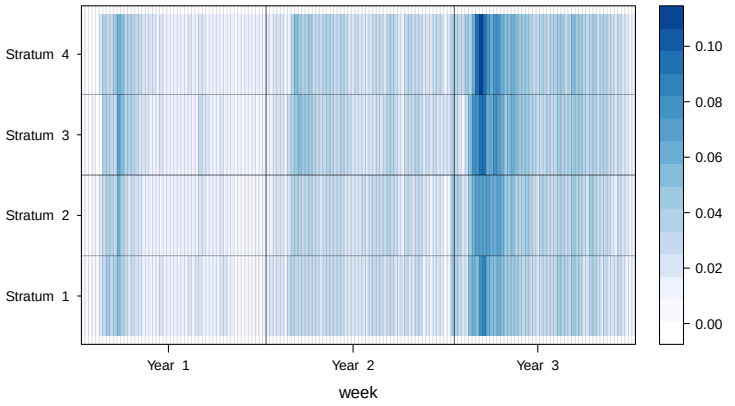
are drawn from the total severe and mild cases separately,

$$Z_{tj}^{\mathrm{ES}}, Z_{tj}^{\mathrm{CS}}, Z_{tj}^{\mathrm{OS}} | k_{tj}^{\mathrm{S}}, Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}} \sim \mathrm{MultHyperGeom}(Y_{tj}^{\mathrm{ES}}, Y_{tj}^{\mathrm{CS}}, Y_{tj}^{\mathrm{OS}}, k_{tj}^{\mathrm{S}})$$
$$Z_{tj}^{\mathrm{EM}}, Z_{tj}^{\mathrm{CM}}, Z_{tj}^{\mathrm{OM}} | k_{tj}^{\mathrm{M}}, Y_{tj}^{\mathrm{EM}}, Y_{tj}^{\mathrm{CM}}, Y_{tj}^{\mathrm{OM}} \sim \mathrm{MultHyperGeom}(Y_{tj}^{\mathrm{EM}}, Y_{tj}^{\mathrm{CM}}, Y_{tj}^{\mathrm{OM}}, k_{tj}^{\mathrm{M}}),$$

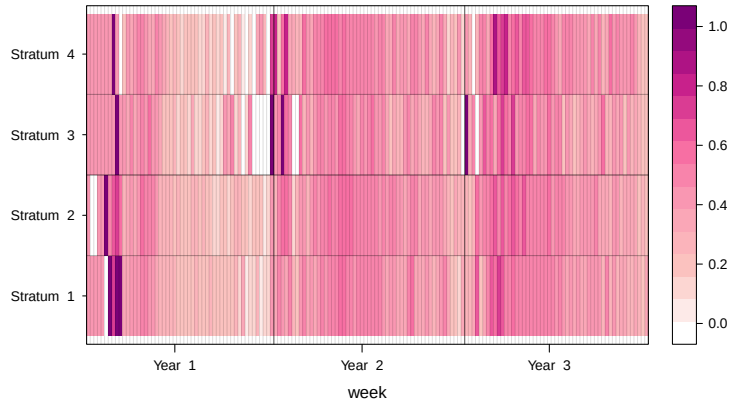
where k_{tj}^{S} and k_{tj}^{M} are defined to be some proportion of the total number of severe or mild cases, respectively.

We consider two different scenarios for the proportion of cases subsampled for virology: when the proportion is constant over time and strata, and when the proportion of cases

subsampled varies over time and strata. Both scenarios assume that the proportion of cases subsampled differs by severity. For simulations assuming constant sampling over time and strata, we suppose that for all strata and weeks, 45% of all severe cases and 10% of mild cases were subsampled. For simulations with variable amounts of subsampling, we use proportions comparable to those observed in the HFMD data. Specific proportions by strata, week, and severity are presented in Figure A.12.



(a) Proportion of mild cases subsampled



(b) Proportion of severe cases subsampled

Figure A.12: Proportions of mild and severe cases subsampled for simulations when subsampling varies by stratum and over time.

A.5.2 Estimation of Y_{tj}^{G}

Simulation specifics

In the simulation study we examine how well we estimate the unobserved pathogen- and stratum-specific disease counts over repeated subsamples from the true population under different amounts of subsampling. We evaluate the performance of our estimates by examining the bias, the scaled mean squared error (MSE), and the coverage of the true values. Since the magnitude of the various unobserved counts differs dramatically across pathogen and stratum, we compare a scaled version of the MSE. For example, to evaluate our estimates of Y_{tj}^{ES}, we compute

$$\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\widetilde{Y_{tj}^{\text{ES}}} - Y_{tj}^{\text{ES}}\right)^2 / Y_{tj}^{\text{ES}}. \tag{A.6}$$

Constant subsampling

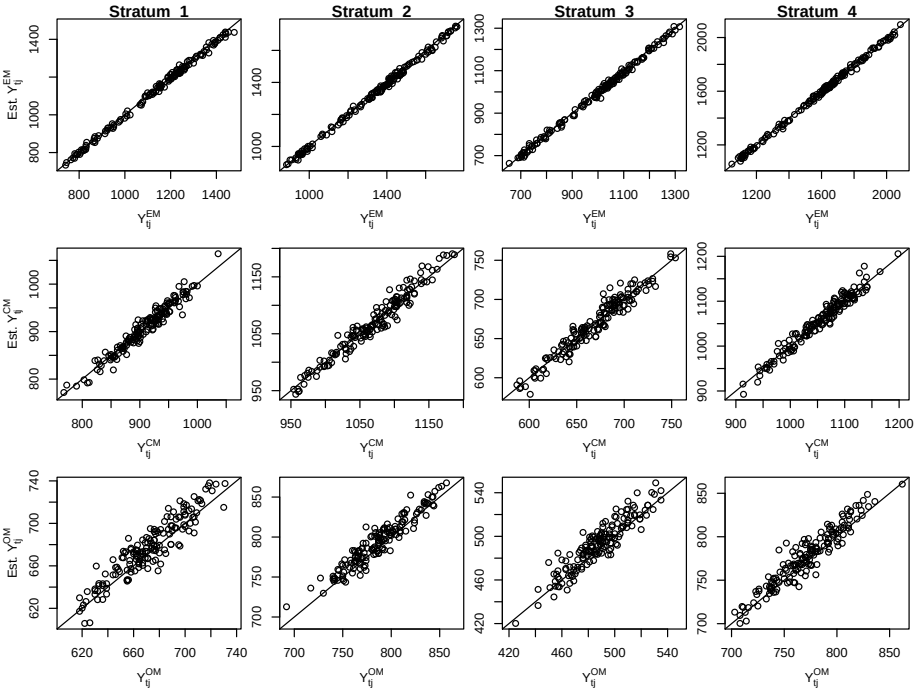
Pathogen- and stratum-specific estimated counts against the true number of cases for both mild and severe cases are plotted in Figure A.13. We see that when the number of disease cases is large, our estimates tend to do well. However, when the true number of counts are small or zero, we tend to overestimate the number of cases.

Table A.5 summarizes the scaled mean squared error (MSE) for pathogen- and stratum-specific disease counts by severity. We see that, in general, when pathogen- and stratum-specific disease counts are large, the scaled MSE is relatively small.

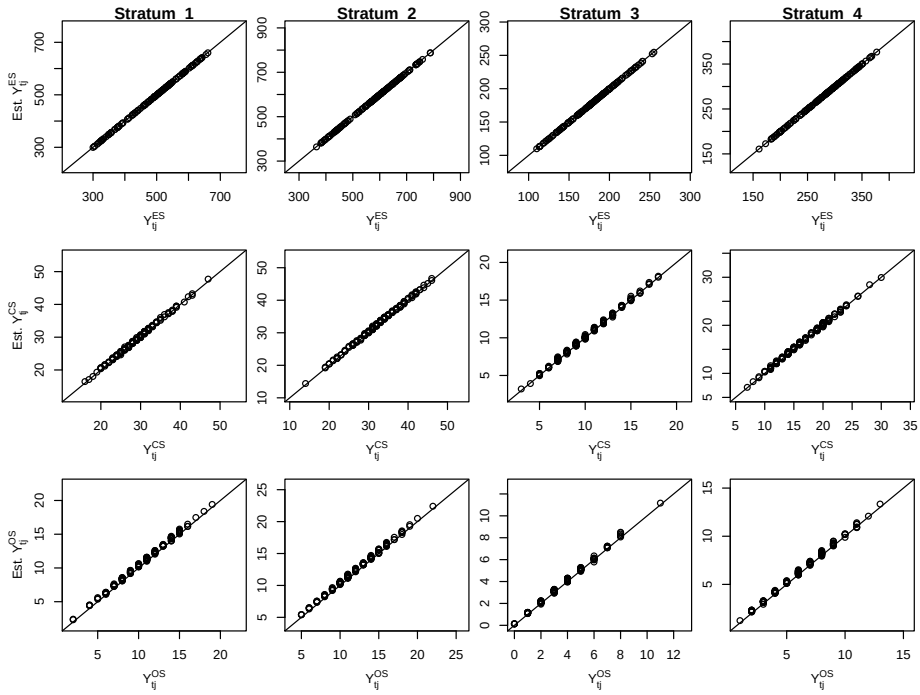
Strata	yES	yCS	yOS	yEM	yCM	yOM
1	0.095	1.16	1.21	54.96	62.17	70.54
2	0.091	1.15	1.21	55.10	62.06	70.52
3	0.090	1.16	1.20	50.21	64.11	72.64
4	0.094	1.16	1.20	50.86	65.29	72.80

Table A.5: Scaled MSE for pathogen- and stratum-specific disease counts by severity when the proportion of subsampling is constant over time and strata.

In Figure A.14, we present coverage of the pathogen- and stratum-specific disease counts by severity when the proportion of subsampling is constant over time and strata. For mild cases (see Figure A.14(a)), when the number of cases for each pathogen and stratum are large, we see minimal differences between the performance of unadjusted and adjusted variance estimates. Posterior estimates for mild cases yield credible intervals that are somewhat too small (leading to under-coverage); for mild cases, coverage is between 90 and 95% for all pathogens and strata. Coverage for severe cases is presented in Figure A.14(a). We see that the variance adjustment may make a difference, but generally will lead to over-coverage. Unadjusted estimates yield credible intervals that have good coverage for larger numbers of true cases. However, when the true number of cases is very small, but not zero (less than 5), we see worse coverage.

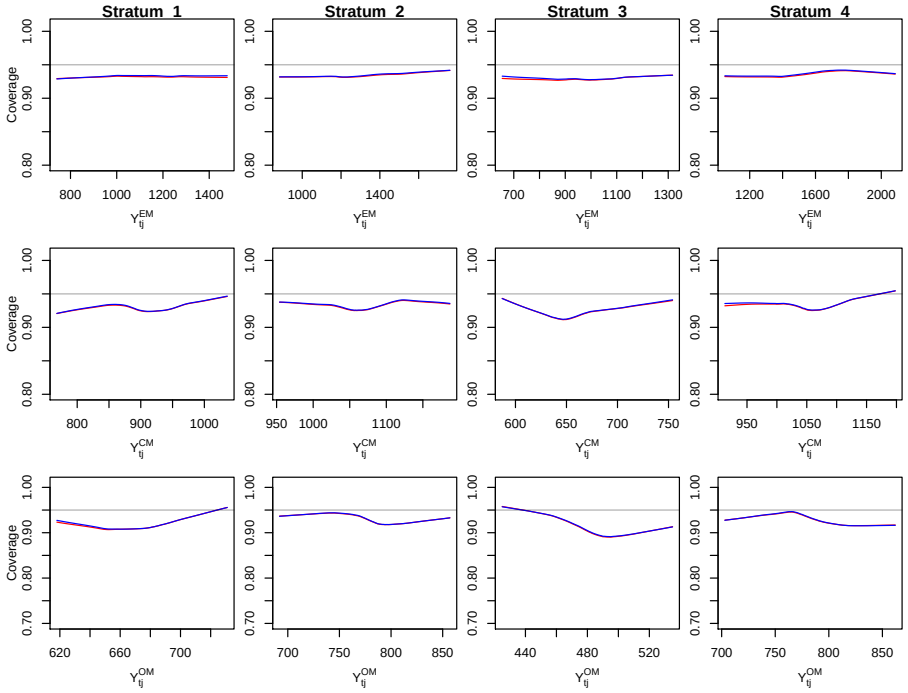


(a) Mild cases

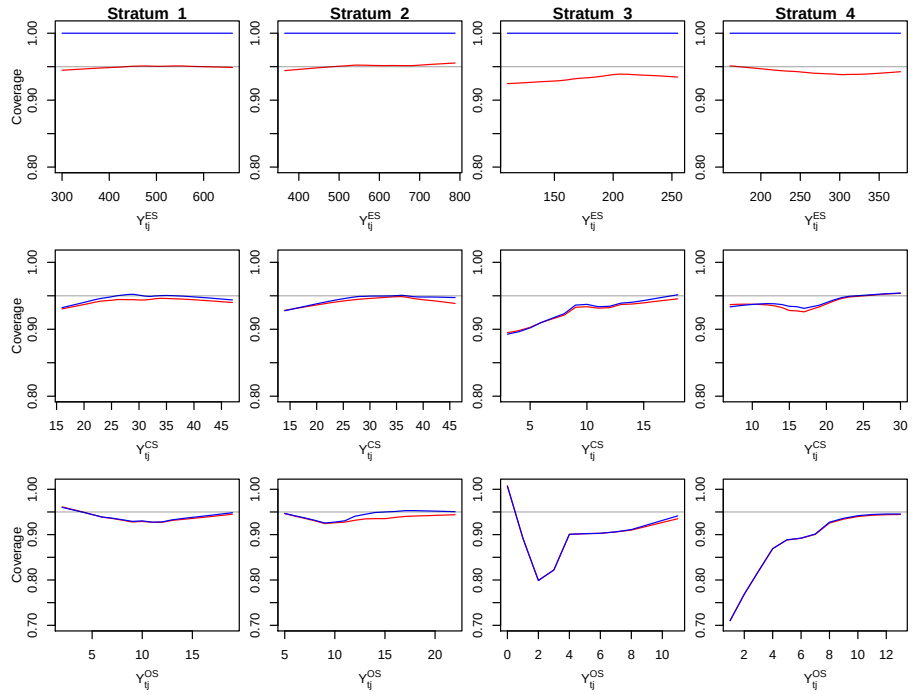


(b) Severe cases

Figure A.13: Estimated unobserved pathogen-specific disease counts by severity. Simulation results for 500 simulations, when the proportion of severe and mild cases are constant over time and stratum.



(a) Coverage for mild cases



(b) Coverage for severe cases

Figure A.14: Coverage of the unobserved pathogen-specific disease counts by severity. Red lines are coverage estimates using the unadjusted variance estimates; blue lines plot the coverage using the adjusted variance estimates. Results are based on 500 simulations, when the proportion of severe and mild cases are constant over time and stratum.

Variable subsampling

Pathogen- and stratum-specific estimated counts against the true number of cases for both mild and severe cases are plotted in Figure A.15. We see that, on average, we tend to estimate the pathogen- and stratum-specific disease counts by severity well. However, there are estimates that are very different from the true number of cases, especially in the estimates of severe cases.

To better understand why we consistently under- (or over-) estimate certain counts, we plot both the estimated and true pathogen- and stratum-specific counts by stratum over time in Figure A.16. We see that even with a small amount of subsampling, we obtain reasonable estimates for the unobserved pathogen-specific disease counts. However, when there are no cases subsampled for virology, we tend to dramatically over- or under-estimate the true number of cases. In these instances, the pathogen-specific estimates for a given stratum and week sum to the total severe (or mild) cases observed. In fact, when there are no subsamples, the posterior estimates are $\widetilde{Y_{tj}^{\text{GS}}} = Y_{tj}^{\cdot\text{S}}\left(\alpha_j^{\text{GS}}/\alpha_j^{\cdot\text{S}}\right)$.

Notice that the estimates we obtain, are an improvement over the standard MoM estimates described in Section A.1, where we cannot form estimates when there were no cases subsampled. It would be possible to perform additional smoothing in time to obtain better estimates when no cases are subsampled, but we did not pursue this further. In subsequent simulations, we examine how variable subsampling (and sometimes very biased estimates) effects the estimates of pathogen-specific log relative risks.

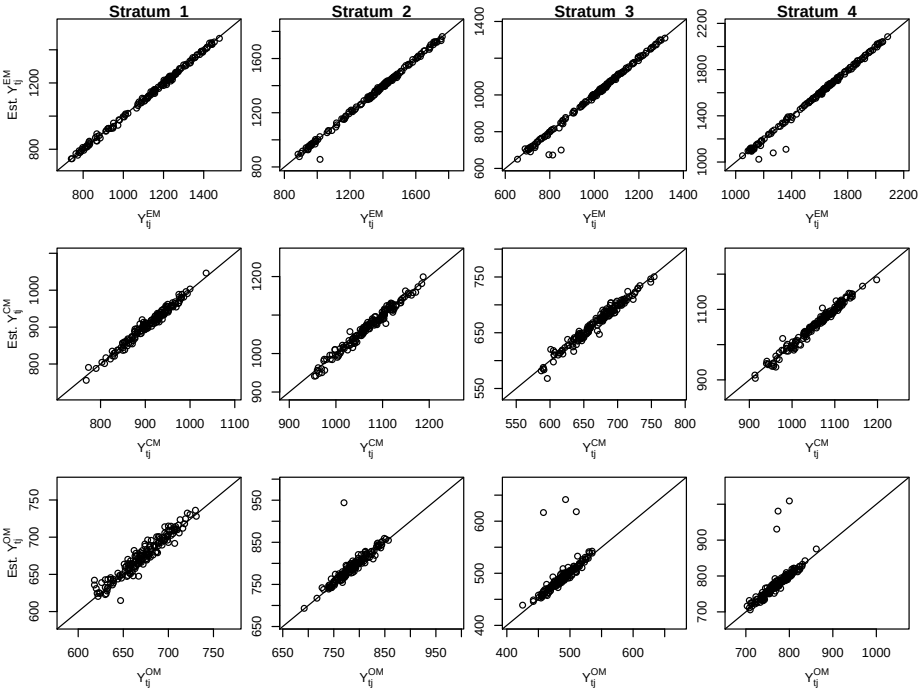
Table A.6 summarizes the scaled MSE for pathogen- and stratum-specific disease counts by severity averaged over time. As in the constant proportion of subsampling scenario, we tend to have a small scaled MSE. However, we now see that when total number of cases is small, the scaled MSE is much larger than before. We expect the scaled MSE to be sensitive to bias when the true number of counts is small, and the estimates to be somewhat worse with the variable subsampling. Thus, we are not surprised by the dramatically larger scaled MSEs, especially for the counts of severe cases attributable to Other pathogens for strata 3.

Coverage of the pathogen- and stratum-specific disease counts by severity when subsampling varies by stratum and week are presented in Figure A.18. As before, we see that posterior credible intervals yield nearly 95% coverage when estimating mild cases, even with variable amounts of subsampling. However, for severe cases, the effects of variable subsampling are clear. When there is a large number of pathogen-specific severe cases for a given stratum, we still tend to obtain reasonable coverage from the posterior credible intervals. However, as we know, when no cases are subsampled, we obtain very biased estimates, which result in poor coverage of the credible intervals.

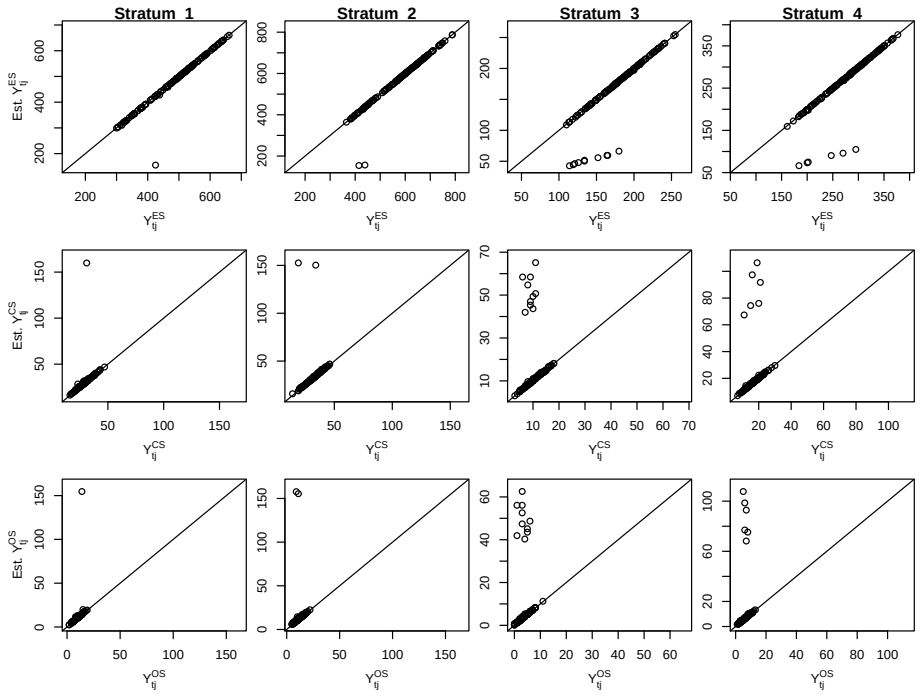
Strata	yES	yCS	yOS	yEM	yCM	yOM
1	1.27	5.46	11.22	32.58	35.73	40.15
2	2.37	10.42	29.93	28.78	31.80	36.01
3	3.71	15.72	65.26	24.47	29.84	34.09
4	3.80	12.99	43.80	23.86	29.22	33.66

Table A.6: Scaled MSE for pathogen- and stratum-specific disease counts by severity when the proportion of subsampling varies over time and strata.

To clarify, we plot the coverage estimates for pathogen- and stratum-specific severe disease counts over time in Figure A.18. We see that, for a given strata, when there are no subsamples, credible intervals have very low coverage for all pathogens. In these cases, the variance adjustment tends to improve the coverage dramatically.

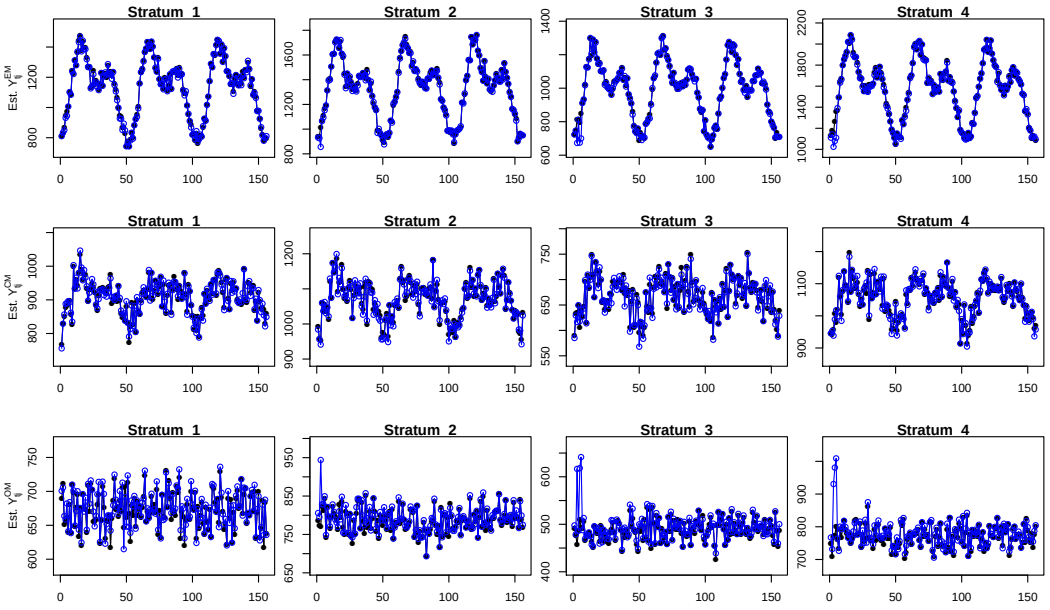


(a) Mild cases

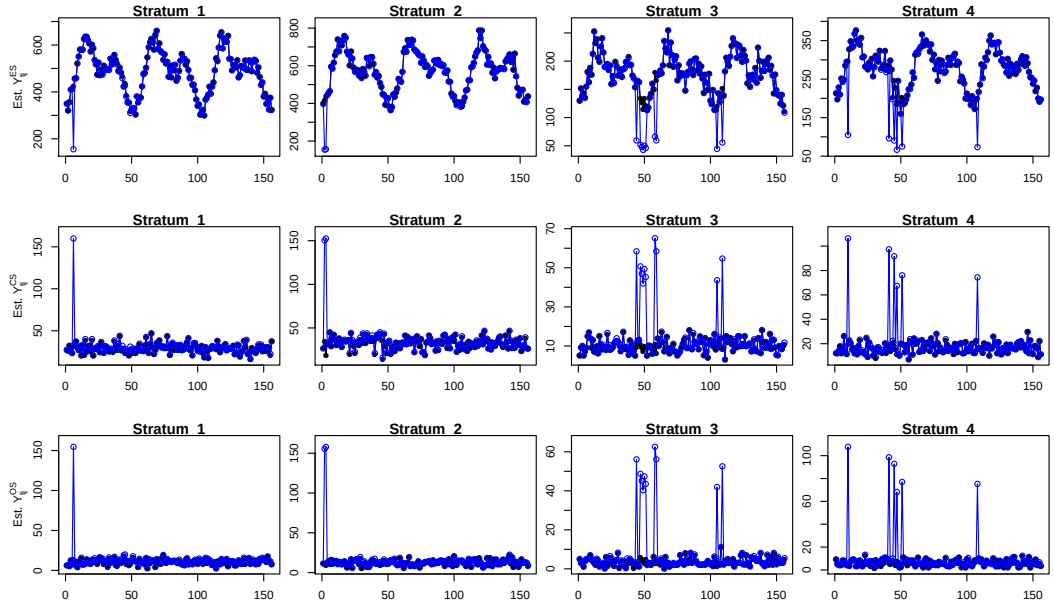


(b) Severe cases

Figure A.15: Estimated unobserved pathogen-specific disease counts by severity. Simulation results for 500 simulations, when the proportion of severe and mild cases varies over time and stratum.

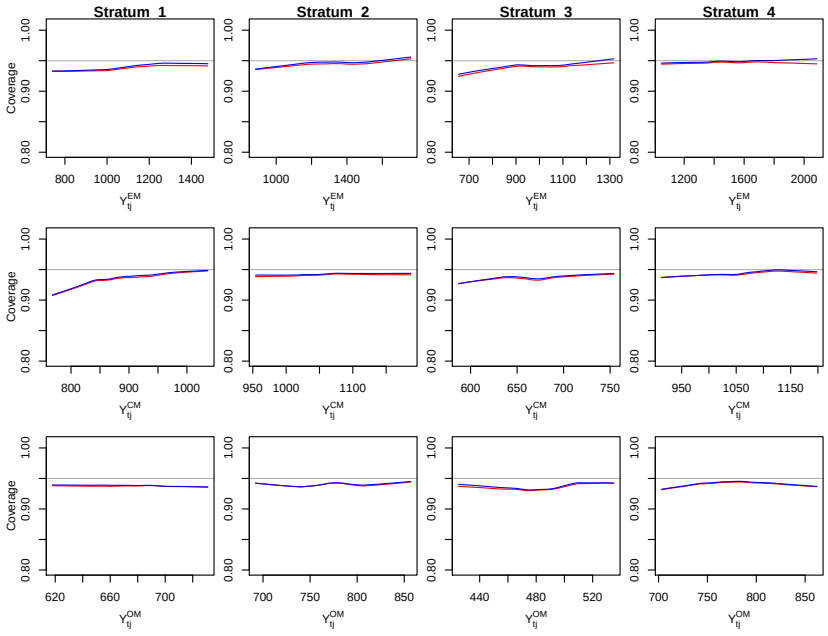


(a) Mild cases

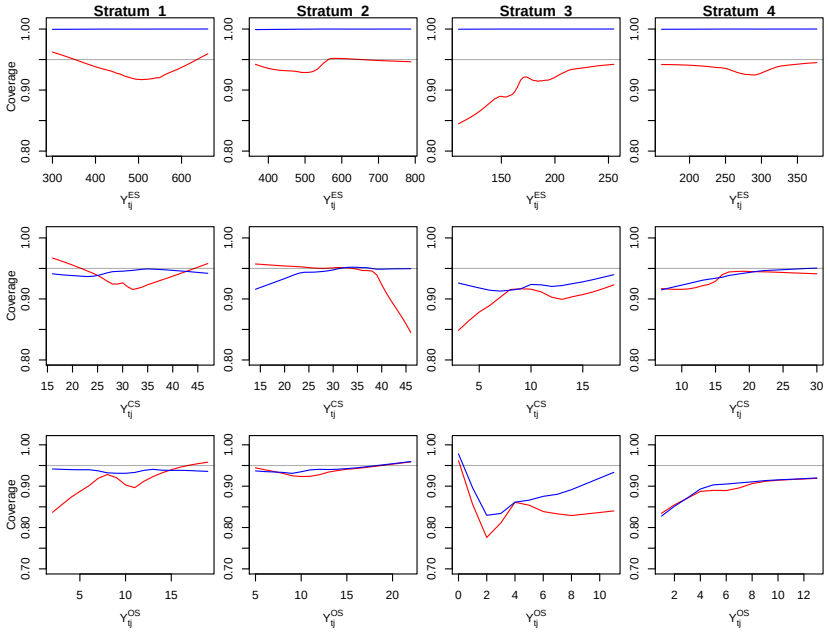


(b) Severe cases

Figure A.16: Estimated unobserved pathogen-specific disease counts by severity over time. Black dots are the true values, and blue circles are the average estimates. Simulation results for 500 simulations, when the proportion of severe and mild cases varies over time and stratum.



(a) Coverage for mild cases by value



(b) Coverage for severe cases by value

Figure A.17: Coverage of the unobserved pathogen-specific disease counts by severity. Red lines are lowess curves for coverage estimates using the unadjusted variance estimates; blue lines plot the lowess curve for coverage using the adjusted variance estimates. Results are based on 500 simulations, *when the proportion of severe and mild cases varies over time and stratum.*

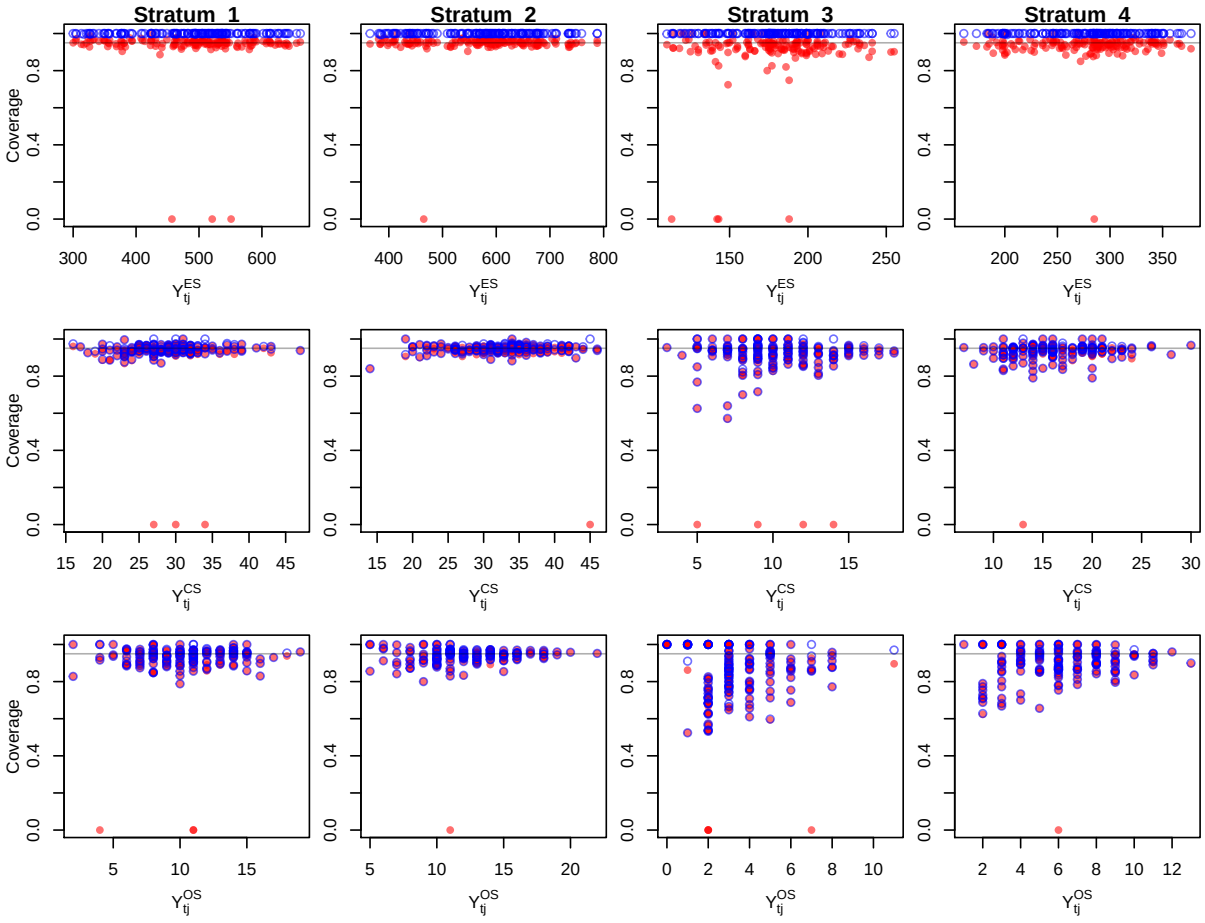


Figure A.18: Coverage of the unobserved pathogen-specific severe disease counts. Red points are coverage estimates using the unadjusted variance estimates; blue points plot the coverage using the adjusted variance estimates. Results are based on 500 simulations, when the proportion of severe and mild cases varies over time and stratum.

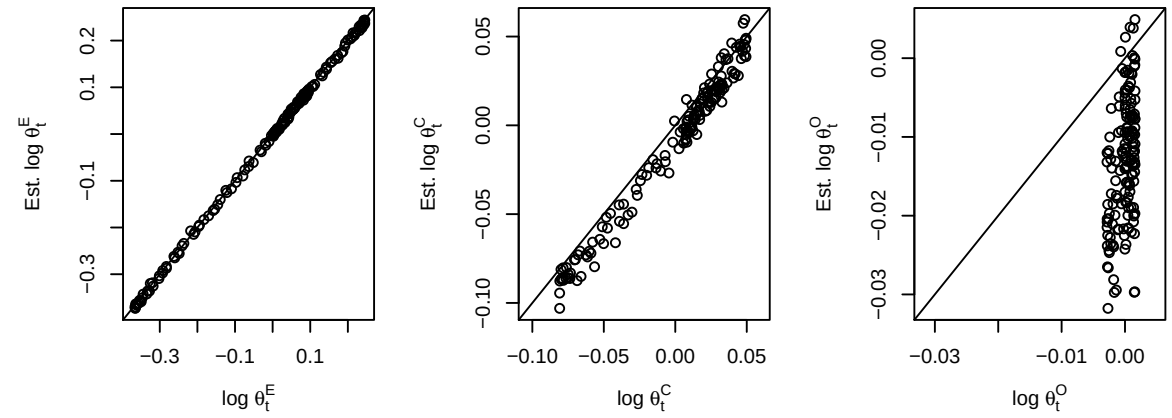
A.5.3 Estimation of $\log \theta_t^G$

Setup

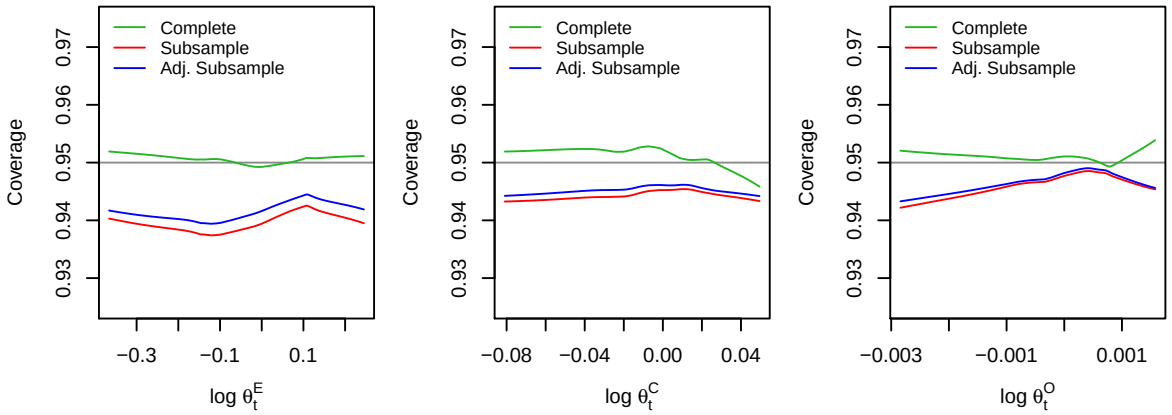
To evaluate how well we estimate $\log \theta_t^{\mathrm{G}}$, we compare the performance of our estimators, both with and without the variance adjustments, to those that would be obtained if the complete data had been observed as well as the true values. When the totality of cases have been observed, there is no need to estimate pathogen- and stratum-specific disease counts by severity and pathogen-specific log relative risk estimates are obtained in the standard way, $\log \widehat{\theta_t^{\mathrm{G}}} = \log\left(\sum_{j=1}^{J} Y_{tj}^{\mathrm{GS}} + Y_{tj}^{\mathrm{GM}}\right) - \log(E^{\mathrm{G}})$, with $\mathrm{Var}\left[\log \widehat{\theta_t^{\mathrm{G}}}\right] = 1/(E^{\mathrm{G}}\widehat{\theta_t^{\mathrm{G}}})$. We consider the effects of constant and variable subsampling, and then consider the case when pathogen-specific log relative risks are not proportionate across strata.

Constant proportion of subsampling

Figure A.19(a) plots the average estimated pathogen-specific log relative risk estimates for 500 simulations against the true values. Estimates of $\log \theta_t^{\mathrm{E}}$ tend to be very good, while $\log \theta_t^{\mathrm{C}}$ estimates tend to underestimate the true value. The estimates for $\log \theta_t^{\mathrm{O}}$ tend to also underestimate the true value, but are more variable. Moreover, $\log \theta_t^{\mathrm{O}}$ is very close to 0, so the bias is very small. Figure A.19(b) shows lowess curves of the point-wise coverage of the pathogen-specific log relative risk estimates obtained in three ways: using the complete data and using subsampled data with and without the variance adjustments discussed in Section A.4. The estimates obtained from totally observed data have approximately 95% coverage for each of the pathogens and true log-relative risk values. When only a subsample of the data is observed, the credible intervals have slightly lower coverage of the true value. When we take into account the additional variability in the number of severe and mild cases, the coverage improves some.



(a) Average estimates of pathogen-specific log relative risk, by pathogen.



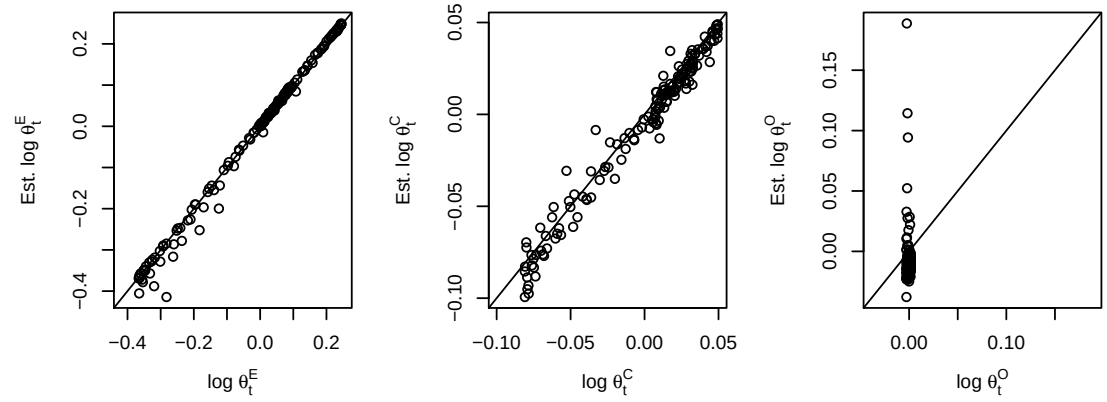
(b) Point-wise coverage for pathogen-specific log relative risk, by pathogen. The green line is the lowess for coverage of estimates using the complete data, while the blue and red lines are lowess curves based on subsampled data with and without the variance adjustments, respectively.

Figure A.19: Estimated pathogen-specific log relative risks and pointwise coverage by true value. The green line is a lowess curve for the coverage of estimates using the complete data, while the blue and red lines are lowess curves based on subsampled data with and without the variance adjustments, respectively. Simulation results for 500 simulations, where 45% of all severe cases and 10% of mild cases were subsampled.

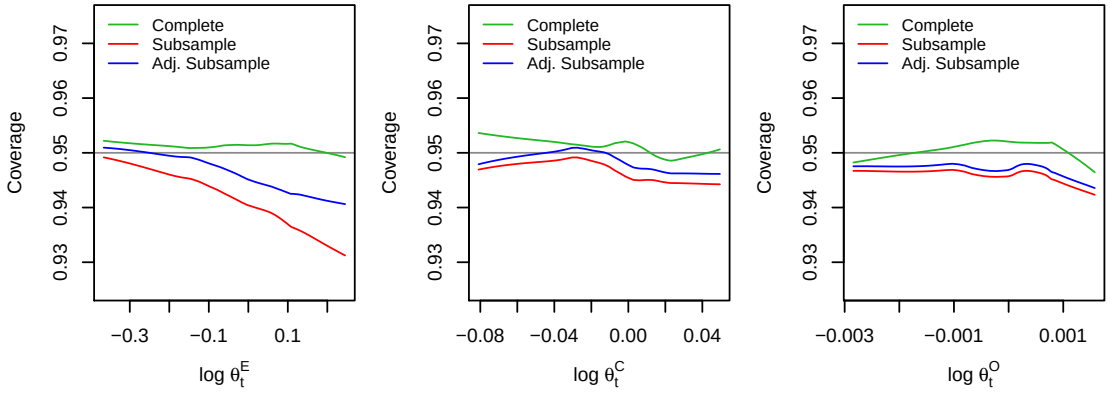
Variable amounts of subsampling

Figure A.20(a) plots the average estimates of pathogen-specific log relative risk against the true values, by pathogen. We see that they tend to be good for the dominant pathogen, but

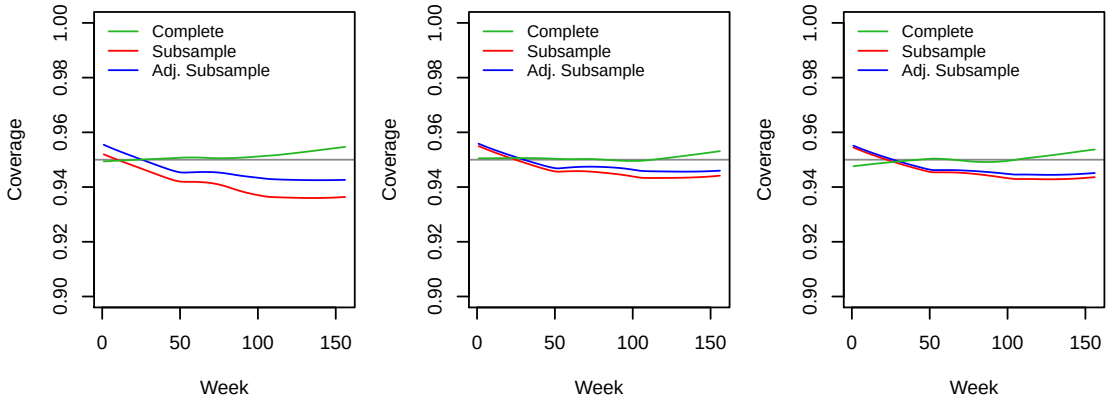
can underestimate small effects, especially for the rarest pathogen. Figure A.20(b) shows the point-wise coverage of the pathogen-specific log relative risk. We examine the coverage of pathogen-specific log relative risks over time in Figure A.20(c). Although we tended to poorly estimate the unobserved pathogen- and stratum-specific disease counts when no cases were subsampled, our estimates of pathogen-specific log relative risks have good coverage.



(a) Average estimates of pathogen-specific log relative risk, by pathogen.



(b) Point-wise coverage for pathogen-specific log relative risk, by pathogen. The green line is coverage of estimates using the complete data, while the blue and red lines are based on subsampled data with and without the variance adjustments, respectively.



(c) Point-wise coverage for pathogen-specific log relative risk over time.

Figure A.20: Estimated pathogen-specific log relative risks, and pointwise coverage by value and over time. The green line is a lowess curve for the coverage of estimates using the complete data, while the blue and red lines are lowess curves based on subsampled data with and without the variance adjustments, respectively. Simulation results for 500 simulations, when the proportion of severe and mild cases varies over time and by stratum.

Proportionality assumption

Recall, that since we cannot estimate distinct probabilities, we assume that the pathogen-specific disease probabilities can be written as a product of independent temporal and stratum-specific parameters. That is, $s_{tj}^{\mathrm{G}} = \theta_t^{\mathrm{G}} \times p_j^{\mathrm{G}}$.

We examine how our procedure performs when the proportionality assumption is violated. In particular, we examine the case when

$$s_{tj}^{\mathrm{G}} = \theta_{tj}^{\mathrm{G}} \times p_j^{\mathrm{G}},$$

where θ_{tj}^{G} varies both over time and by strata and is generated by

$$\log \theta_{tj}^{\mathrm{G}} = \sum_{k=1}^{2} b_{jk}^{\mathrm{G}} \sin(\omega_k t - \pi/4) + c_{jk}^{\mathrm{G}} \cos(\omega_k t - \pi/4) - d_j^{\mathrm{G}}$$

where $\omega_k = 2k\pi/52.25$, both b_{jk}^{G} and c_{jk}^{G} are pathogen-specific constants, and d_j^{G} is a constant to ensure $\sum_T \theta_{tj}^{\mathrm{G}} = T$. Pathogen- and stratum-specific log relative risks over time are plotted in Figure A.21. For each pathogen, the stratum-specific log relative risks differ. For EV71, the log relative risks vary dramatically across strata; the CA16 stratum-specific log relative risks are similar in shape, but all have similar magnitudes; the stratum-specific log relative risk for Other pathogens varies across strata more than CA16, but less than EV71.

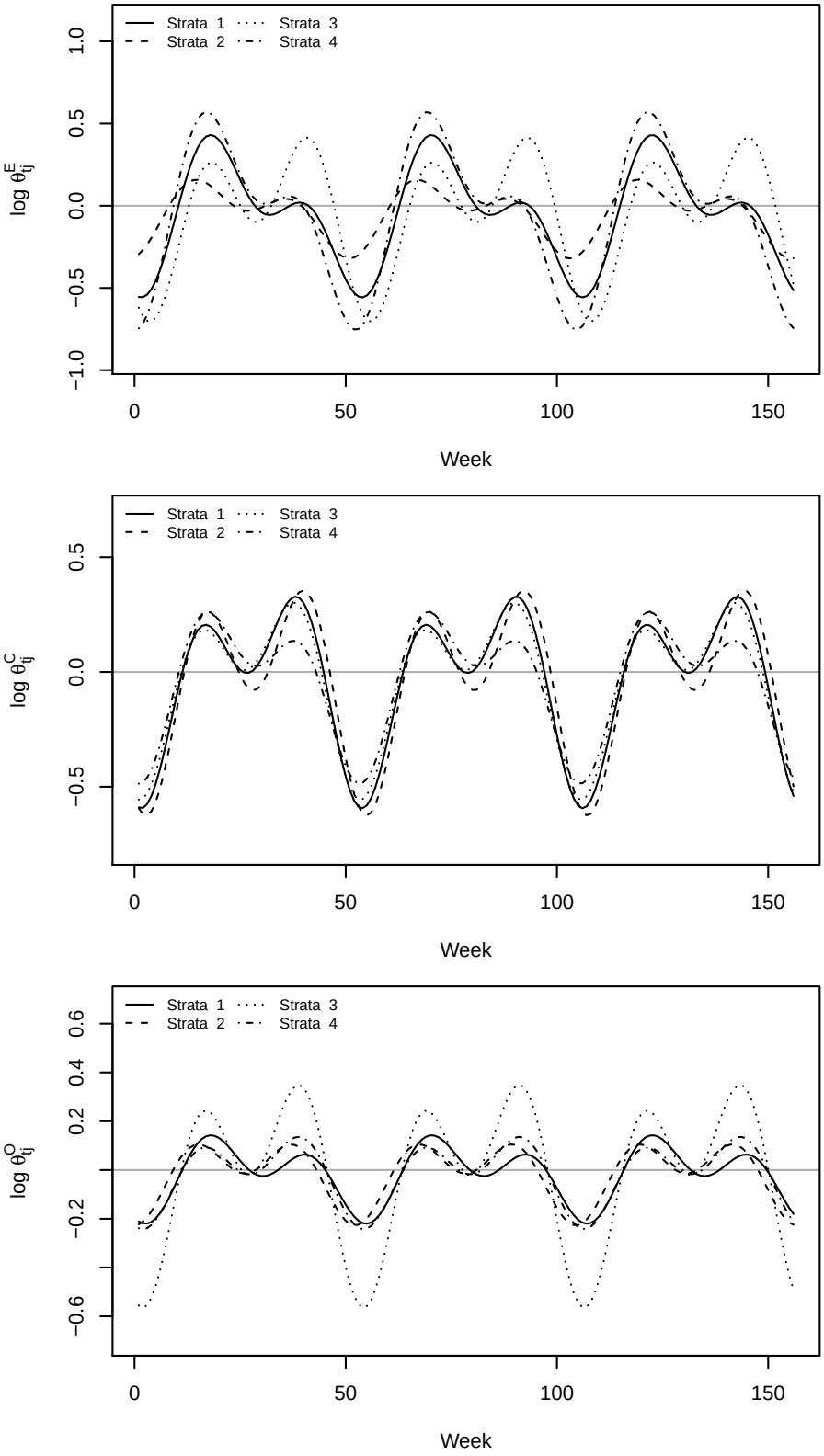


Figure A.21: Pathogen-specific log relative risks over time used for simulations when the proportionality assumption is clearly violated.

When the proportionality assumption is violated, the pathogen-specific relative risk that we estimate is a weighted average of the pathogen- and stratum-specific relative risks. Specifically, $Y_{tj}^{\mathrm{G}} \sim \text{Poisson}\left(N_j \theta_{tj}^{\mathrm{G}} p_j^{\mathrm{G}}\right)$ implies

$$Y_t^{\mathrm{G}} = \sum_{j=1}^{J} Y_{tj}^{\mathrm{G}} \sim \text{Poisson}\left(\sum_{j=1}^{J} N_j \theta_{tj}^{\mathrm{G}} p_j^{\mathrm{G}}\right).$$

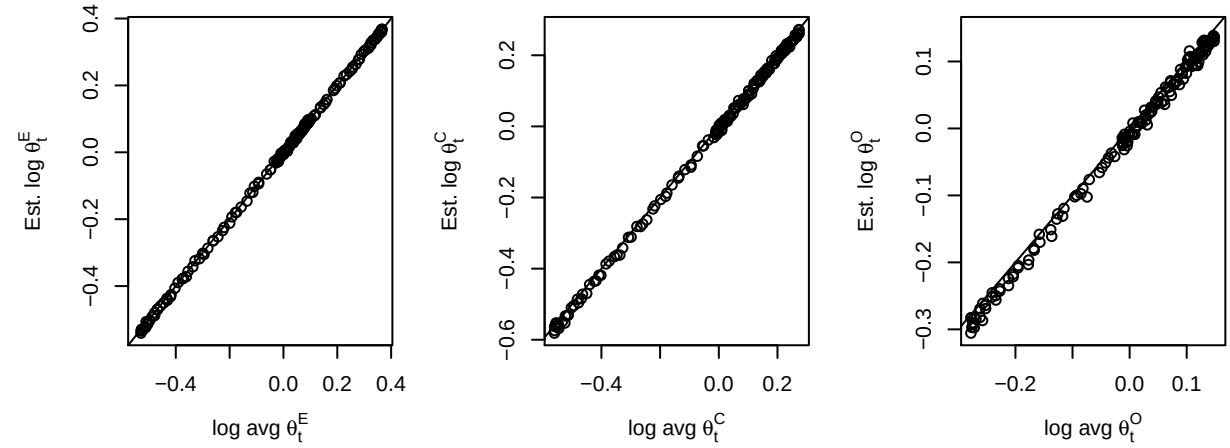
We can find the expectation of our usual estimate of θ_t^{G}:

$$\mathrm{E}[\widehat{\theta_t^{\mathrm{G}}}] = \frac{\mathrm{E}[Y_t^{\mathrm{G}}]}{E^{\mathrm{G}}} = \frac{\sum_{j=1}^{J} N_j \theta_{tj}^{\mathrm{G}} p_j^{\mathrm{G}}}{\sum_{j=1}^{J} N_j p_j^{\mathrm{G}}} = \sum_{j=1}^{J} \theta_{tj}^{\mathrm{G}} \frac{N_j p_j^{\mathrm{G}}}{\sum_{j=1}^{J} N_j p_j^{\mathrm{G}}},$$

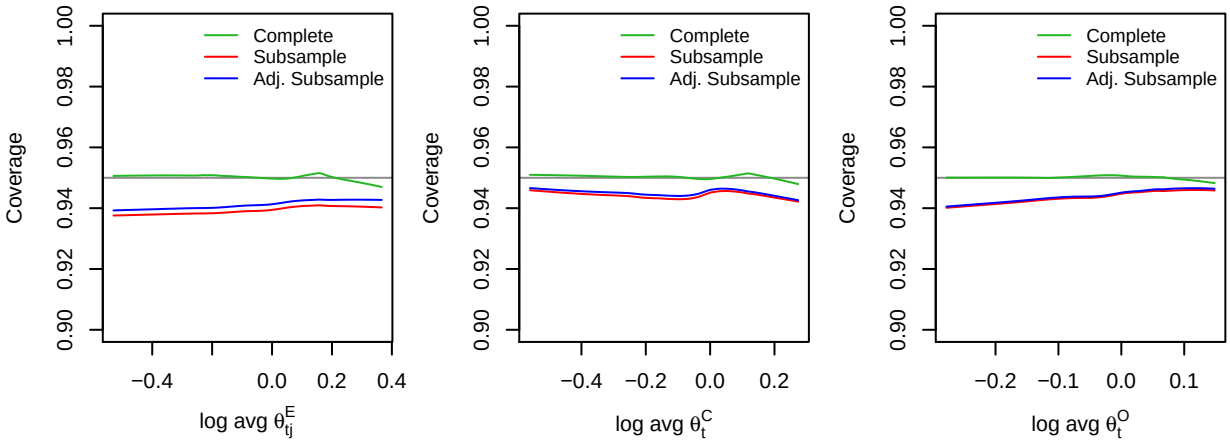
the weighted average of the pathogen- and stratum-specific relative risks.

In Figure A.22(b) we plot the point-wise coverage of the estimated pathogen-specific log relative risk against the true average pathogen-specific log relative risk. Estimates are obtained assuming proportionality when the truth violates this assumption. We see that for all pathogens, the estimates obtained using the completely observed data have good coverage of the true average log relative risk. Although the estimates seem to be reasonable (see Figure A.24(a)), variance estimates are too small, thus resulting in severe under coverage. In contrast, estimates obtained from subsampled data still have relatively good coverage properties, even when the proportionality assumption is violated.

We also consider the impact on variable amounts of subsampling has on estimates when the proportionality assumption is violated; we present the results in Figure A.23. As we have seen before, compared to when the proportion of subsampled cases is constant, variable amounts of subsampling can lead to slightly worse coverage in the estimates of the average pathogen-specific log relative risk. When we examine the coverage over time, in Figure A.23(c), we see that when the proportion of subsamples is small (or zero) for many stratum-severity combinations, we have nearly 100% coverage.

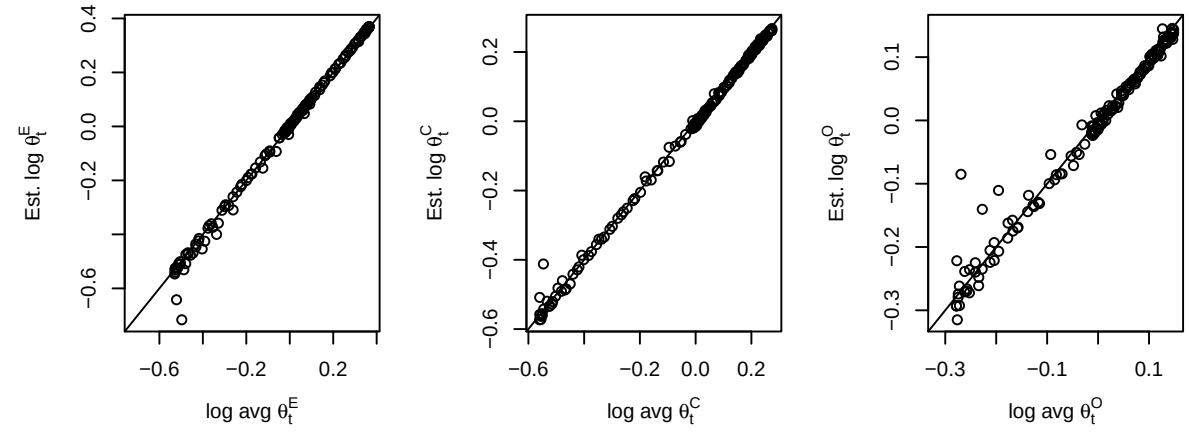


(a) Estimates

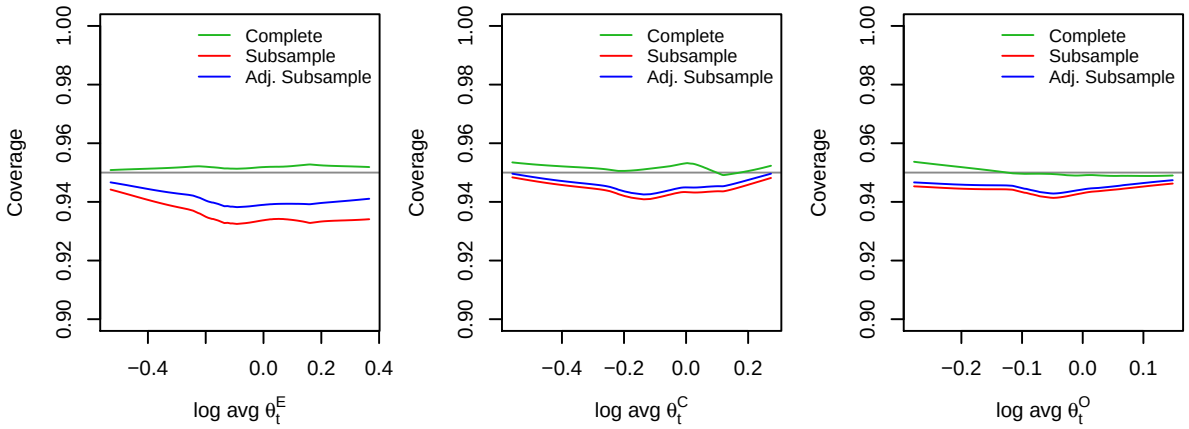


(b) Coverage

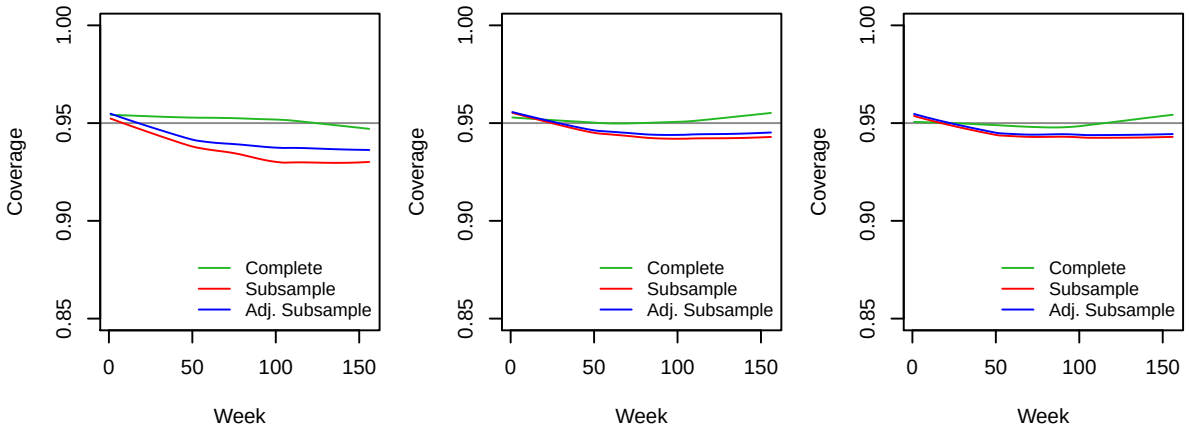
Figure A.22: Coverage of the average pathogen-specific log relative risk over time using estimates of the pathogen-specific log relative risk assuming proportionality. The green line is a lowess curve for the coverage of estimates using complete data, while the blue and red lines are lowess curves based on subsampled data with and without the variance adjustments, respectively. Simulation results for 500 simulations, when the proportion of severe and mild cases are constant over time and stratum.



(a) Estimates



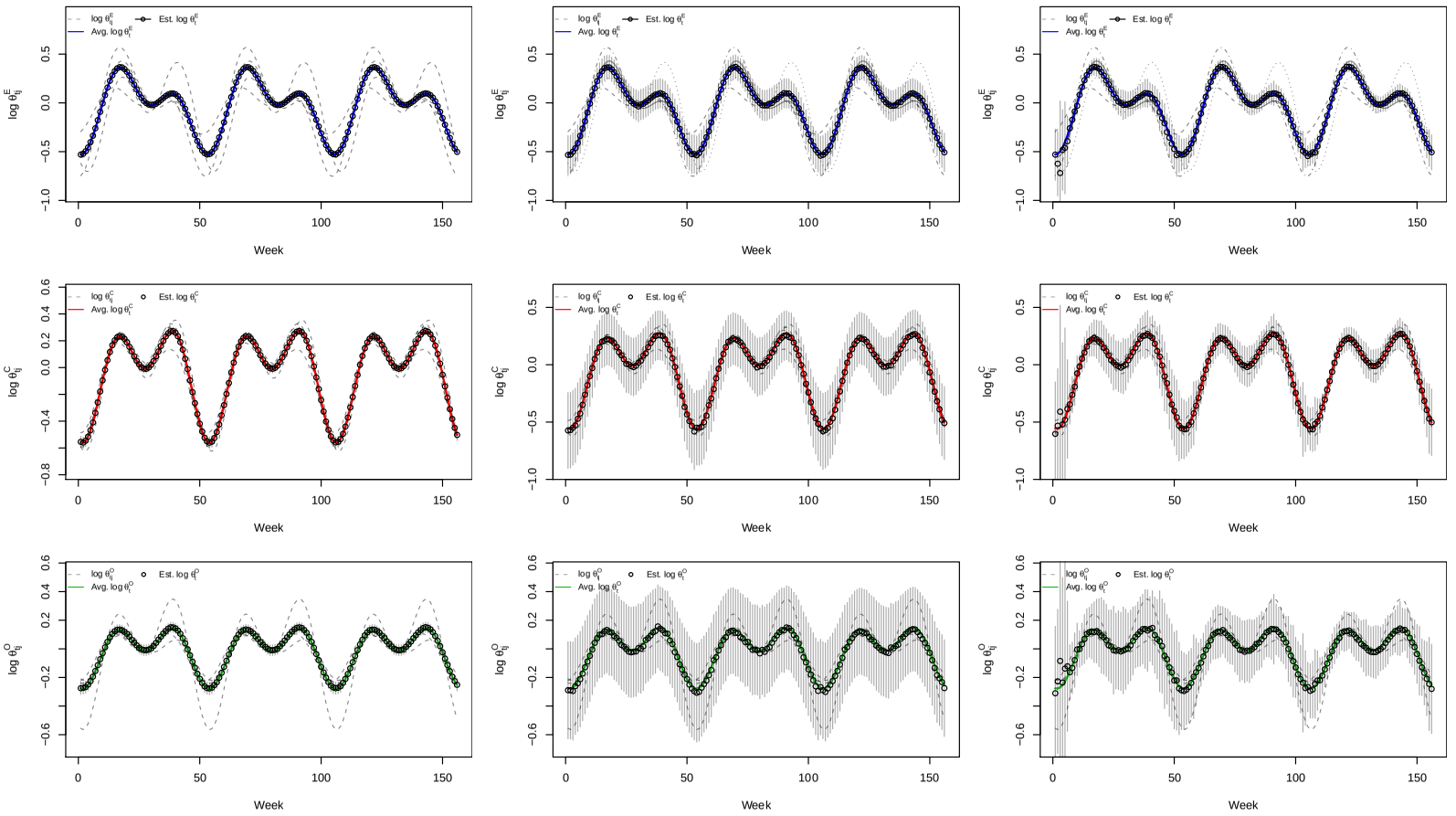
(b) Coverage by average value



(c) Coverage over time

Figure A.23: Coverage of the average pathogen-specific log relative risk over time using estimates of the pathogen-specific log relative risk assuming proportionality is not true. The green line is a lowess curve for the coverage of estimates using complete data, while the blue and red lines are lowess curves based on subsampled data with and without the variance adjustments, respectively. Simulation results for 500 simulations, when the proportion of severe and mild cases varies over time and by stratum.

As a summary, we compare the weighted average pathogen-specific log relative risk estimates obtained from the complete data, and the two subsampling schemes in Figure A.24. All three scenarios result in good estimates of the average pathogen-specific log relative risk, but tend to perform somewhat worse along with the amount of the data used. We see that when we have complete data, as in Figure A.24(a) , we estimate the weighted average pathogen-specific log relative risk well and have narrow point-wise confidence intervals. When using data obtained from the constant subsampling scenario, as shown in Figure A.24(b), we still obtain good estimates of the average pathogen-specific log relative risk, but the confidence intervals are much wider than those using complete data. As expected, the wider confidence intervals reflect the increase in uncertainty in our estimates that comes from using only a subsample of the true data, see Figure A.24(c). When there is a very small proportion of cases subsampled for many strata, the estimates are not as good and the resulting confidence intervals are very wide. Nevertheless, the procedure yields good estimates of the true weighted average pathogen-specific log relative risk, and provides corresponding measurements of uncertainty that reflect the uncertainty of our estimates based on the data being used.



(a) Complete data (b) Constant subsampled data (c) Variable subsampled data

Figure A.24: Dashed lines are the pathogen- and stratum-specific log relative risks over time, the solid colored line is the average pathogen-specific log relative risk over time, and the points are the estimates of the pathogen-specific log relative risk assuming proportionality. These results are based on 500 simulations, when the proportion of severe and mild cases subsampled are constant over time and stratum.

A.5.4 Modeling $\log \theta_t^G$

Simulation setup

We also perform simulations to examine some of the modeling decisions we used to analyze the HFMD data; specifically we examine the decision to model the pathogens jointly. We examine the effect of modeling $\log \widehat{\theta}_t^{\mathrm{E}}$ and $\log \widehat{\theta}_t^{\mathrm{C}}$ jointly, and the role of the prior in the performance of our estimates.

As before, we denoted $\boldsymbol{W}_t = (W_t^{\mathrm{E}}, W_t^{\mathrm{C}}) = \left(\log \widehat{\theta}_t^{\mathrm{E}}, \log \widehat{\theta}_t^{\mathrm{C}}\right)$, $\boldsymbol{\lambda}_t = \left(\mathrm{E}[\log \widehat{\theta_{rt}^{\mathrm{E}}}], \mathrm{E}[\log \widehat{\theta_{rt}^{\mathrm{C}}}]\right)$, and $\mathbf{V}_t$ is the known covariance matrix,

$$\mathbf{V}_t = \begin{bmatrix} \mathrm{Var}\left[W_t^{\mathrm{E}}\right] & \mathrm{Cov}\left[W_t^{\mathrm{E}}, W_t^{\mathrm{C}}\right] \\ & \mathrm{Var}\left[W_t^{\mathrm{C}}\right] \end{bmatrix}.$$

Both models will take as likelihood the asymptotic sampling distribution of the log relative risks, and model time as a second-order random walk (RW2). Specifically, to model the pathogen-specific log relative risks separately, we fit $W_t^{\mathrm{G}}|\lambda_t^{\mathrm{G}} \sim N(\lambda_t^{\mathrm{G}}, V_t^{\mathrm{G}})$, for G=E and C, where $V_t^{\mathrm{G}} = \mathrm{Var}\left[W_t^{\mathrm{E}}\right]$ is the known variance estimate. The joint model assumes $\boldsymbol{W}_{rt}|\boldsymbol{\lambda}_{rt} \sim N_2\left(\boldsymbol{\lambda}_{rt}, \mathbf{V}_{rt}\right)$, where $\mathbf{V}_{rt}$ is the known covariance matrix.

For simulations that consider only the temporal effects, we model the means as

$$\begin{aligned} \lambda_t^{\mathrm{G}} &= \beta_0^{\mathrm{G}} + \beta_1^{\mathrm{G}} t + \delta_t^{\mathrm{G}} \\ \delta_t^{\mathrm{G}} - 2\delta_{t-1}^{\mathrm{G}} + \delta_{t-2}^{\mathrm{G}} | \tau_{\mathrm{G}} &\sim N(0, \tau_{\mathrm{G}}^{-1}), \\ \tau_{\mathrm{G}} &\sim \mathrm{Gamma}(a, b), \\ \boldsymbol{\beta}^{\mathrm{G}} &\sim N(\mathbf{0}, \sigma_\beta^2 I), \qquad \sigma_\beta^2 = 100, \end{aligned}$$

where priors for τ_{G} are chosen in the same manner as described in Section 3.6.2. All models are fit using integrated nested Laplace approximation (INLA) implemented within the `R` programming environment (Rue et al., 2009).

Modeling only temporal effects

We use pathogen-specific log relative risks that are proportional across strata and examine fitted values and the smoothed random walk curves. We compare results from the separate and joint models, each fit using for priors corresponding to little and more appropriate amounts of temporal smoothing. For simulations with little temporal smoothing, we let $a^{\mathrm{G}} = 10$ and $b^{\mathrm{G}} = 1/100$ for G=E and C. This corresponds to the case when $\mathrm{E}[\tau_{\mathrm{G}}] = 1,000$, and $a = 10$.

Figure A.25 summarizes the results from simulations where we fit the second-order random walk model to each pathogen separately. On the left, we plot the estimated pathogen-specific log relative risks against the true values; red lowess smoothers are superimposed. On the right, we plot the estimated temporal trends for each simulation; the solid blue (red) curve corresponds to the true EV71 (CA16) temporal effect used to simulate the data. We see that when we model each pathogen separately, we risk estimating a temporal trend that differs dramatically from the true underlying curve.

We summarize the results from simulations where we fit the second-order random walk model to both pathogen simultaneously in Figure A.26. When we model the two pathogens jointly, the estimates and models perform much better. The estimated pathogen-specific log relative risks are closer to the true values, as are the estimated smoothers.

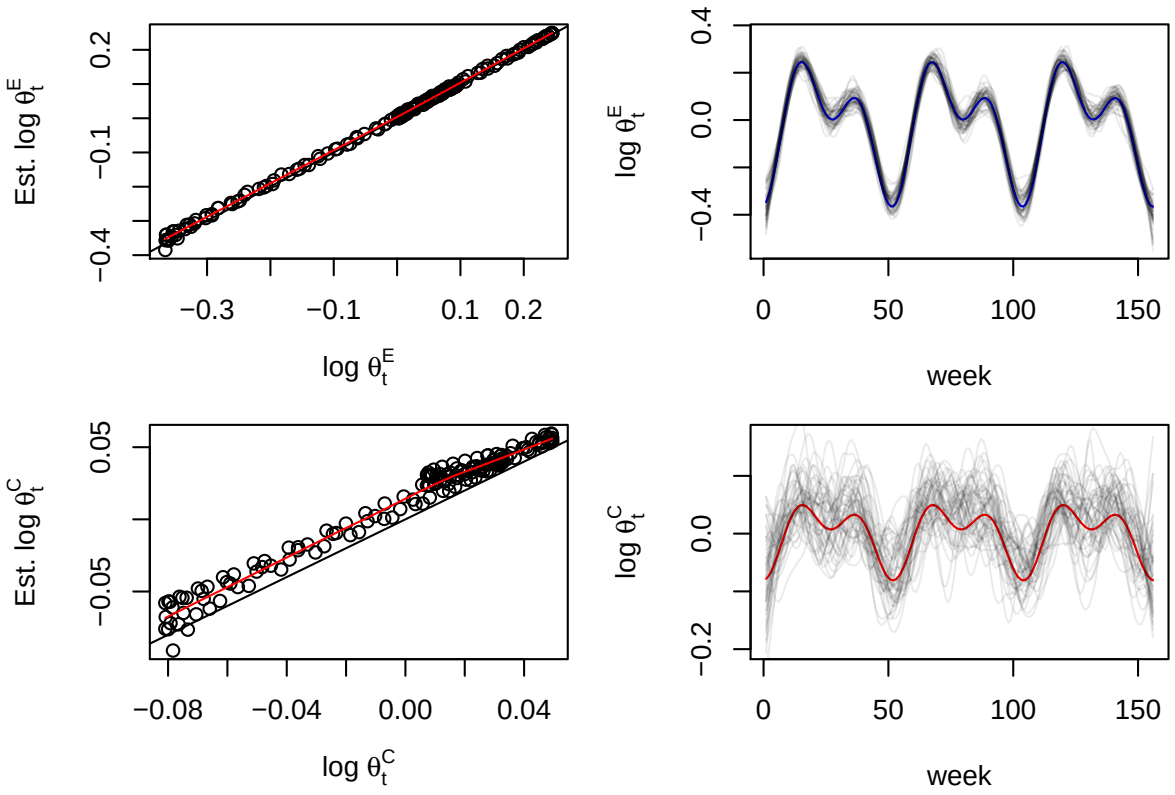


Figure A.25: Simulation results for 100 simulations, when the proportion of severe and mild cases are constant over time and stratum. Pathogen-specific models are fit *separately*, with only the second-order random walk fit to time. On the left, we plot the estimated pathogen-specific log relative risks against the true values; red lowess smoothers are superimposed. On the right, we plot the estimated temporal trends for each simulation; the solid blue (red) curve corresponds to the true EV71 (CA16) temporal effect used to simulate the data.

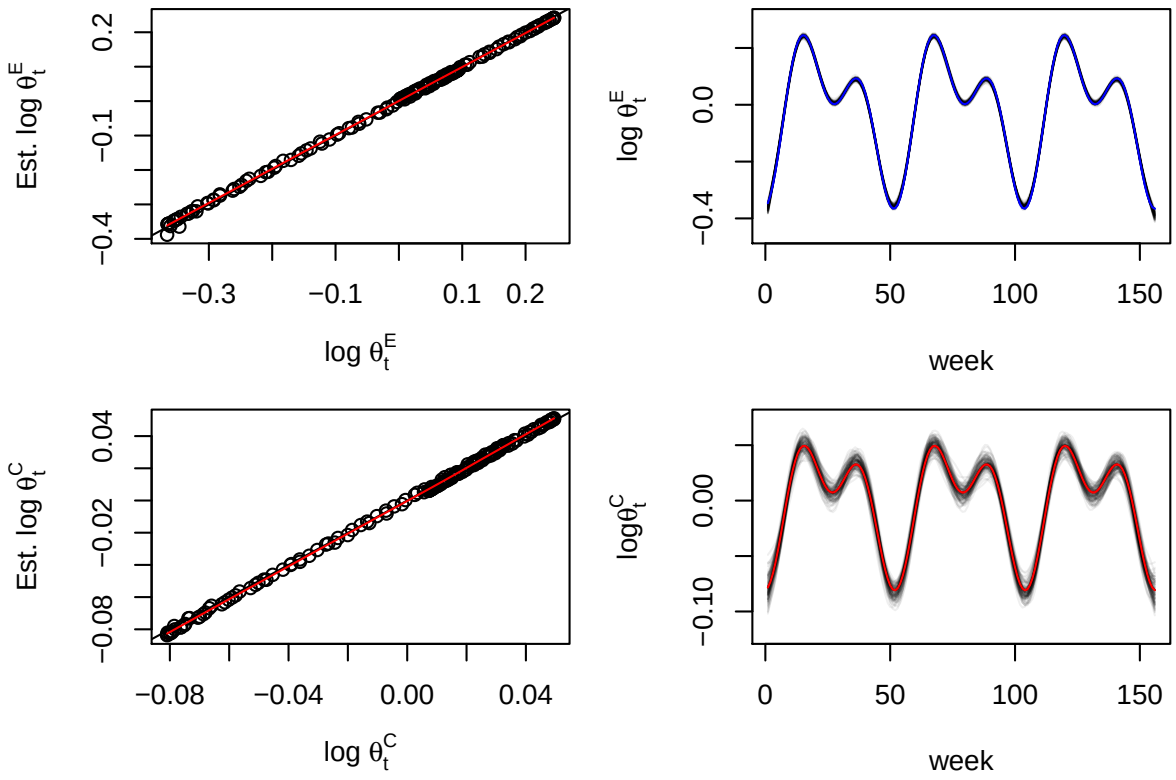


Figure A.26: Simulation results for 100 simulations, when the proportion of severe and mild cases are constant over time and stratum. Pathogen-specific models are fit simultaneously, with only the second-order random walk fit to time. On the left, we plot the estimated pathogen-specific log relative risks against the true values; red lowess smoothers are superimposed. On the right, we plot the estimated temporal trends for each simulation; the solid blue (red) curve corresponds to the true EV71 (CA16) temporal effect used to simulate the data.

The observed improvement from simultaneously modeling both pathogens should not come as a surprise. We expect the subsampled counts to be correlated, and therefore using additional information should lead to better model performance. In Figure A.27, we plot the empirical correlation between the two pathogen-specific log relative risk estimates; we see that the two pathogens have a strong (negative) correlation.

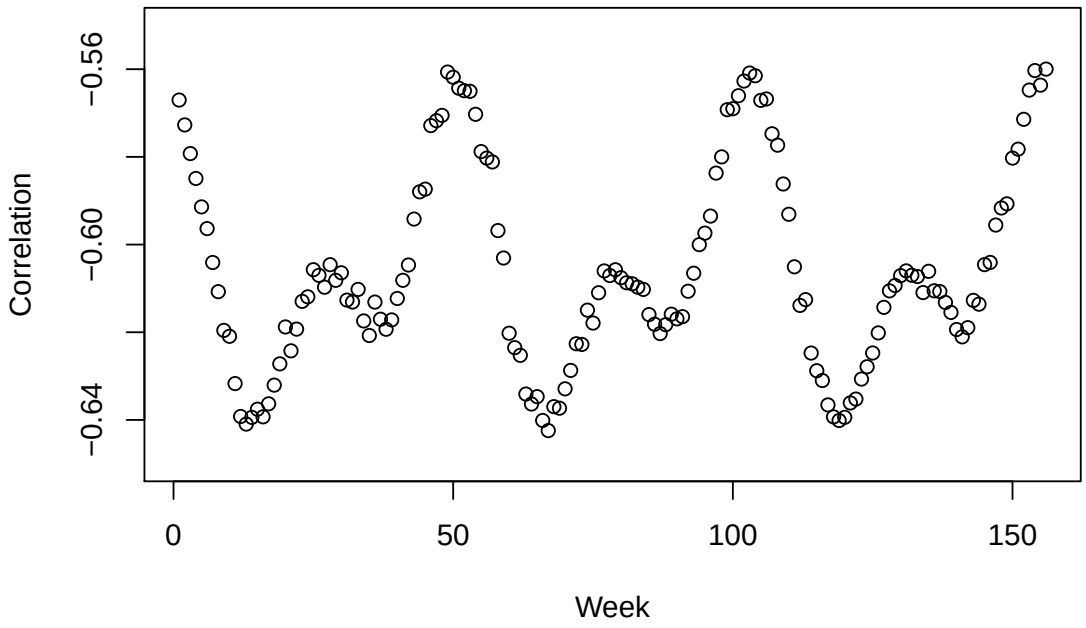


Figure A.27: Empirical correlations between $\log \widehat{\theta}_t^{\mathrm{E}}$ and $\log \widehat{\theta}_t^{\mathrm{C}}$ over time, as estimated from 500 simulations.

Estimating a linear trend in temperature

We now turn our attention towards the performance of the model when estimating a linear trend in temperature. For simulations, we used the average lagged temperature from the Northeastern region of China to include a linear trend in temperature and simulated the true pathogen specific log relative risk as

$$\log \theta_t^{\mathrm{G}} = a^{\mathrm{G}} x_t + \sum_{k=1}^{2} b_k^{\mathrm{G}} \sin(\omega_k t - \pi/4) + c_k^{\mathrm{G}} \cos(\omega_k t - \pi/4) - d^{\mathrm{G}},$$

where x_t is the temperature at time t; a^{G} is the pathogen-specific effect we are interested in estimating.

We compare the estimates obtained from modeling the two pathogen-specific log relative risks jointly and separately. Figure A.28 summarizes the results from simulations where we model EV71 and CA16 separately. While both models tend to produce estimates that are, on average, closer to the true temperature effect, the variability of these estimates across simulations is much greater than when we model the pathogen-specific relative risks jointly. Moreover, the pathogen-specific temporal trends for each pathogen are much more variable. Note that the priors are fairly flexible and identical for both models.

Figure A.29 summarizes the results from simulations where we fit the EV71-CA16 model jointly. Histograms of estimated pathogen-specific temperature effects show overestimation in the EV71-specific temperature effects, while the CA16-specific temperature effects tend to be closer to the true value. The smoothers in time in the right panels show consistent underestimation for the EV71-specific trend in time, while the time trends for CA16 seem to be more variable around the truth. As a result of strong confounding by time, especially for EV71, we tend to slightly over-estimate the effect of temperature and underestimate the effect of time.

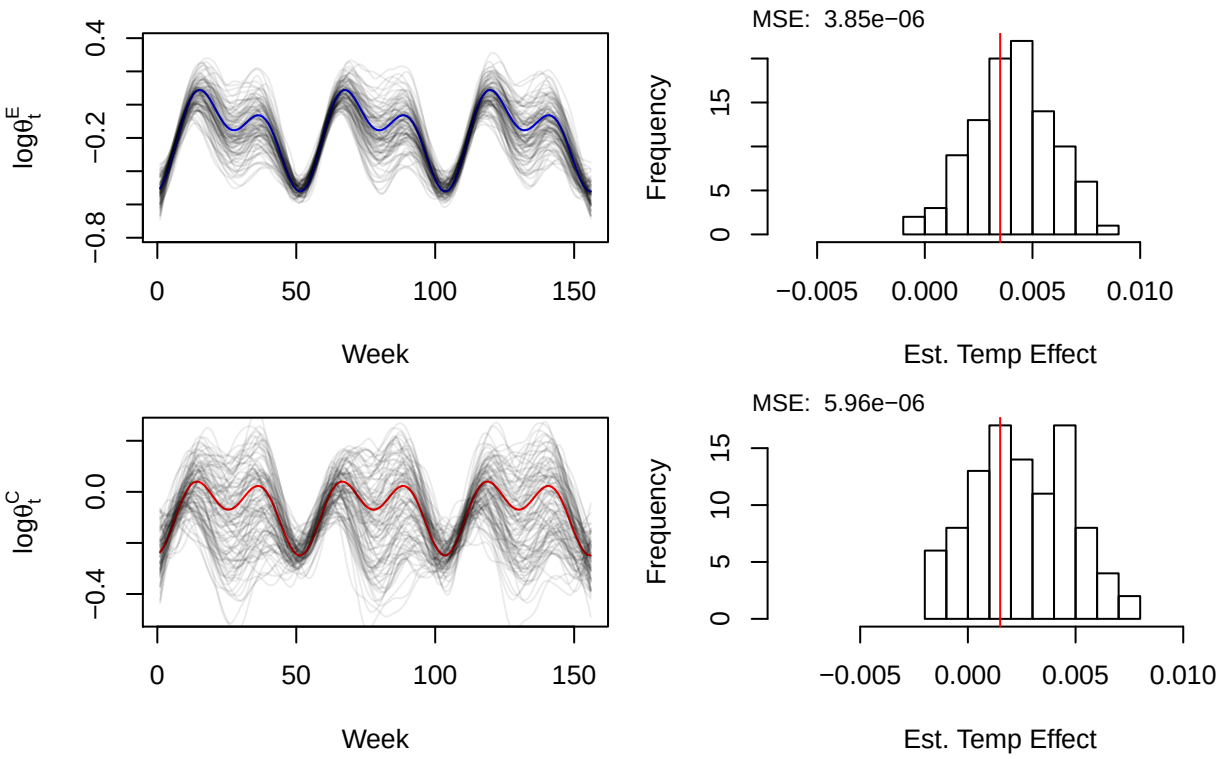


Figure A.28: Simulation results for 100 simulations where EV71 and CA16 estimated *separately*, and the proportion of severe and mild cases are constant over time and stratum. Right panels show histograms of estimated pathogen-specific temperature effects, with the red line at the true value. Estimated pathogen-specific smoothers in time are shown in the left panels.

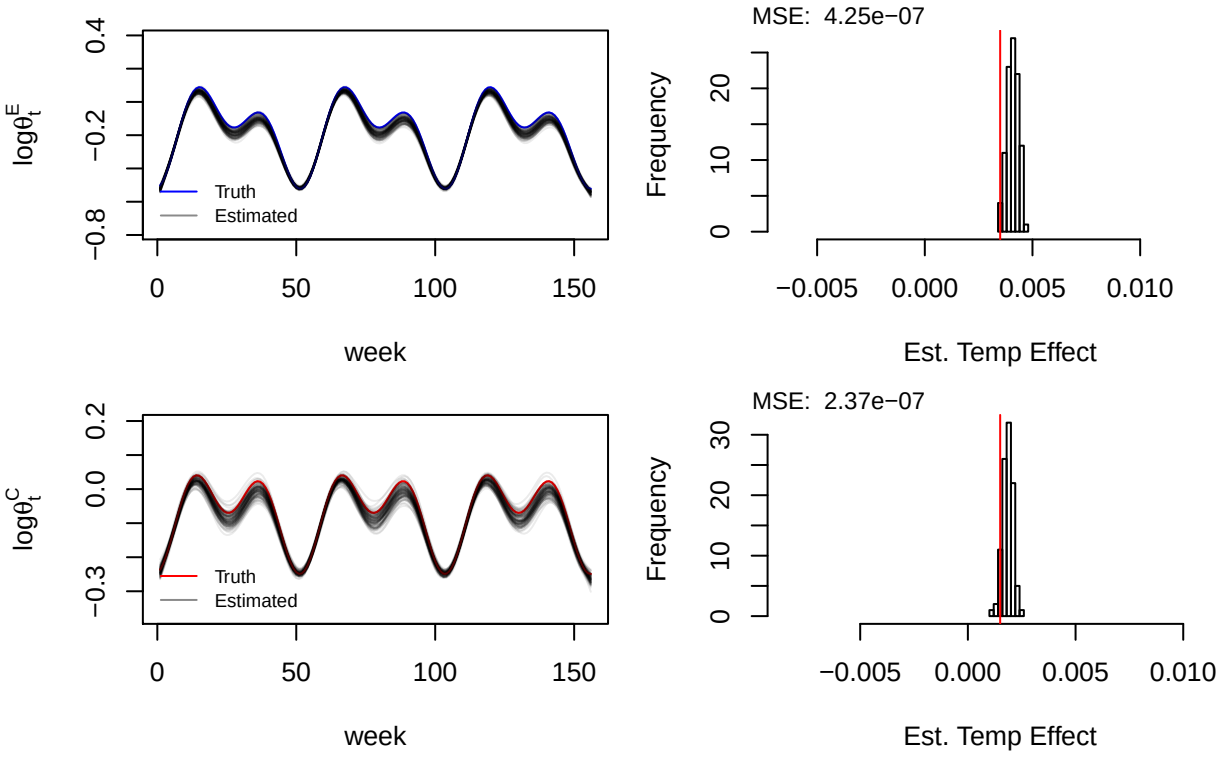


Figure A.29: Simulation results for 100 simulations where EV71 and CA16 estimated *jointly*, and the proportion of severe and mild cases are constant over time and stratum. Right panels show histograms of estimated pathogen-specific temperature effects, with the red line at the true value. Estimated pathogen-specific smoothers in time are shown in the left panels.

A.5.5 Summary of simulations

We conducted simulations studies to investigate three important aspects of the proposed methodology. Overall, simulations studies show that the methods proposed in this paper result in reasonable estimates of the unobserved disease counts, the underlying pathogen-specific log relative risk, and estimated effects of covariates on the risk. We further see that even when the proportionality assumption fails, we still obtain good estimates of relevant parameters.

The first study examined the pre-processing procedure used to obtain estimate unobserved disease counts. We consider two subsampling scenarios for this study: when the proportion of cases subsampled is constant across time and stratum, and when the proportion of subsampled cases varies across time and stratum. We compare our results to those that would be obtained in the complete case scenario (i.e. when 100% of cases are subsampled for virology). When the amount of subsampling is constant across strata and time, we obtain good estimates for the two primary pathogens of interest. The corresponding confidence intervals have nearly 95% coverage, except when the true number of cases is very small (fewer than 5 cases). When the amount of subsampling varies, the resulting estimates perform somewhat worse than those obtained from constant subsampling. In particular, when there are no samples for a given stratum or week, the resulting estimates have poor coverage. Therefore, this procedure is would not be ideal for estimating exact numbers in these cases. Nevertheless, as subsequent simulation demonstrates, the pre-processing procedure allows us to obtain reasonable estimates of the pathogen-specific log relative risks. The details of this simulation study are presented in Section A.5.2.

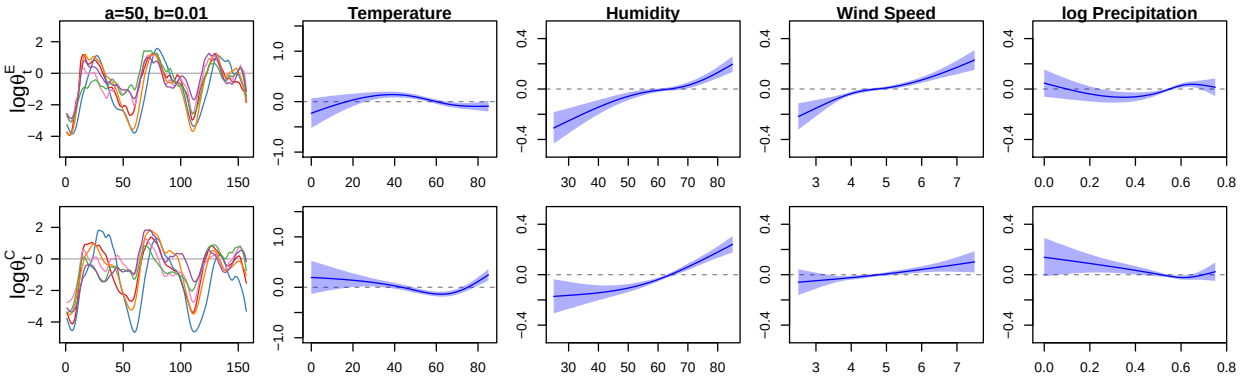
The second simulation study investigated the performance of the pathogen-specific log relative risks estimates, and their sensitivity to the modeling assumptions. In particular, we were interested in the impact of the proportionality assumption on the performance of the estimated pathogen-specific log relative risk. We compare both constant and variable subsampling schemes to the complete case scenario. When proportionality holds, we tend to

estimate the true pathogen-specific log relative risk for the two primary pathogens of interest well, and corresponding standard error estimates yield intervals with nearly 95% coverage. When subsampling is variable, these estimates are slightly worse, but coverage is still very good. When the proportionality assumption is not valid, the proposed method yields good estimates of the weighted average pathogen-specific log relative risk. The details of this simulation study are presented in Section A.5.3.

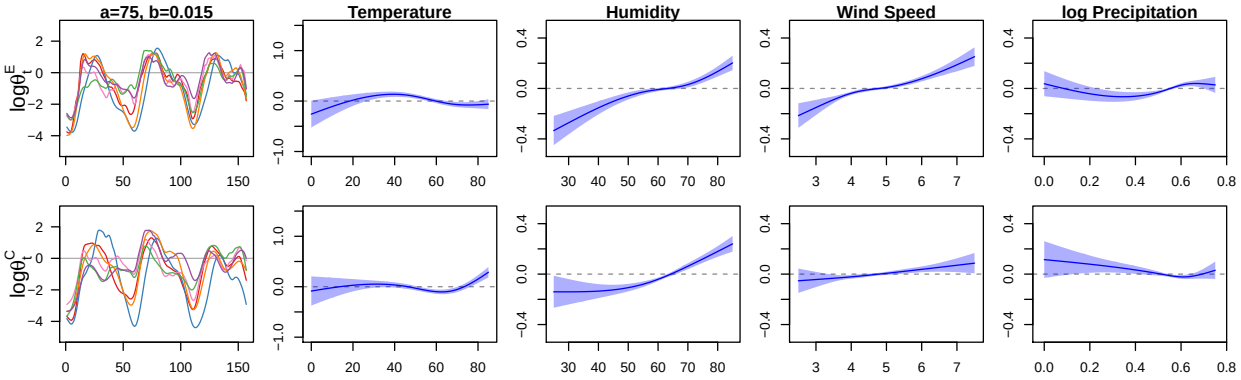
The last simulation study investigated the impact of modeling choices in the ability to estimate underlying temporal and covariate trends. In the sensitivity analyzes for the primary analysis, we saw that that our results were relatively stable over a range of reasonable priors (See Subsection A.6 for further details). Therefore we do not consider the effect of prior selection in these simulations. We consider the impact of modeling the two pathogens jointly. We found that accounting for the correlation induced by subsampling between the two pathogens by modeling them jointly yielded better estimates of the true effects, as well as more consistent estimation of the underlying temporal effects. When consider estimating the covariate effect, both methods produce slightly biased estimates of the true effect. However, the estimates obtained when modeling the two pathogens together yield covariate effect estimates with a smaller MSE. The details of this simulation study are presented in Section A.5.4.

A.6 Sensitivity analysis

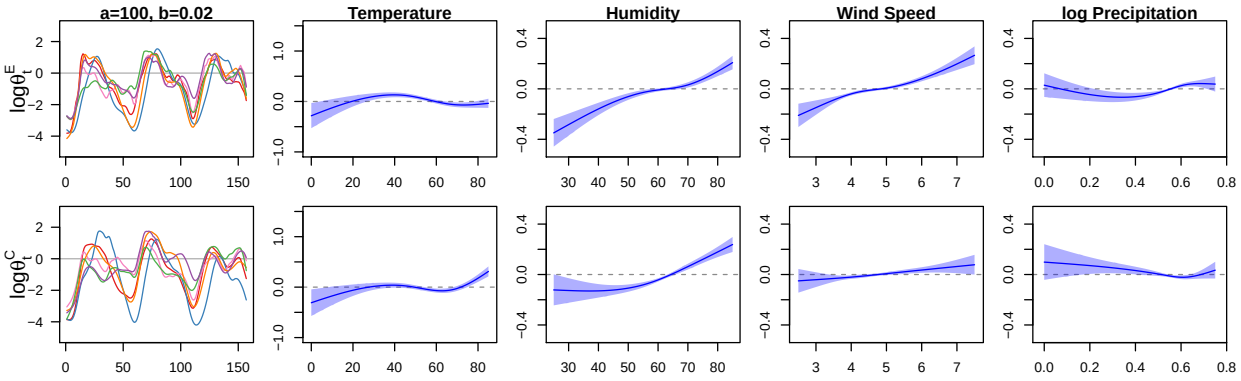
To examine the sensitivity of our results to the priors, we vary both the mean of τ_{G} and a. We consider three values for $\text{E}[\tau_{\text{G}}]$: 5,000, 7,500, and 10,000. Values of a between 50 and 100 yield coefficients of variation between 0.14 and 0.10, respectively. We examine the sensitivity to the a parameter by considering three values: 50, 75, and 100. Figures A.30, A.31, and A.32 present results for fixed values of the mean of τ_{G} and with varying a values. The results appear to be robust to prior specification. The CA16-specific log relative risk seems to be the most sensitive to prior specification for low temperatures.



(a) $\mathrm{E}[\tau_{\mathrm{G}}] = 5,000$ and $a = 50$

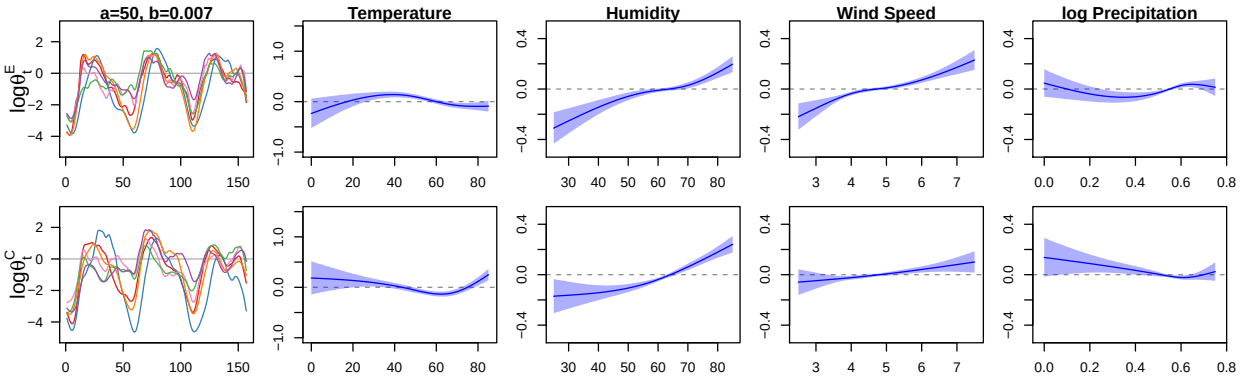


(b) $\mathrm{E}[\tau_{\mathrm{G}}] = 5,000$ and $a = 75$

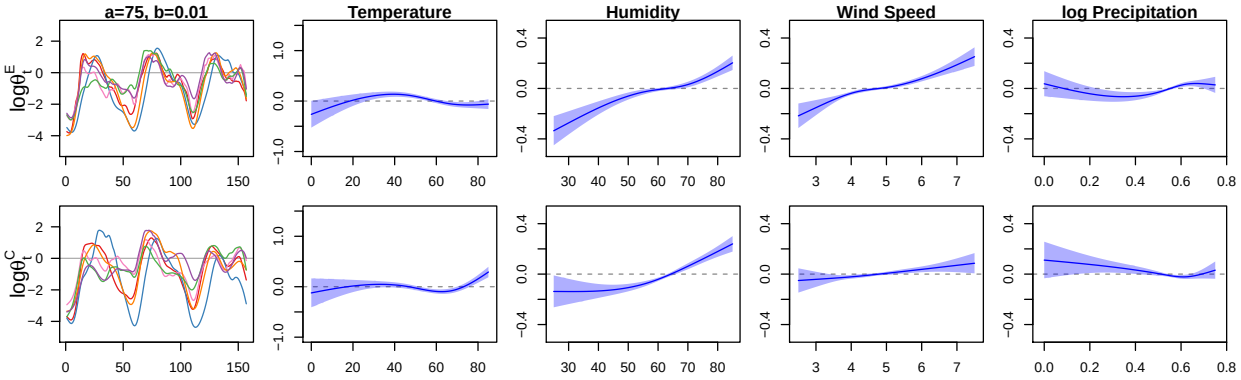


(c) $\mathrm{E}[\tau_{\mathrm{G}}] = 5,000$ and $a = 100$

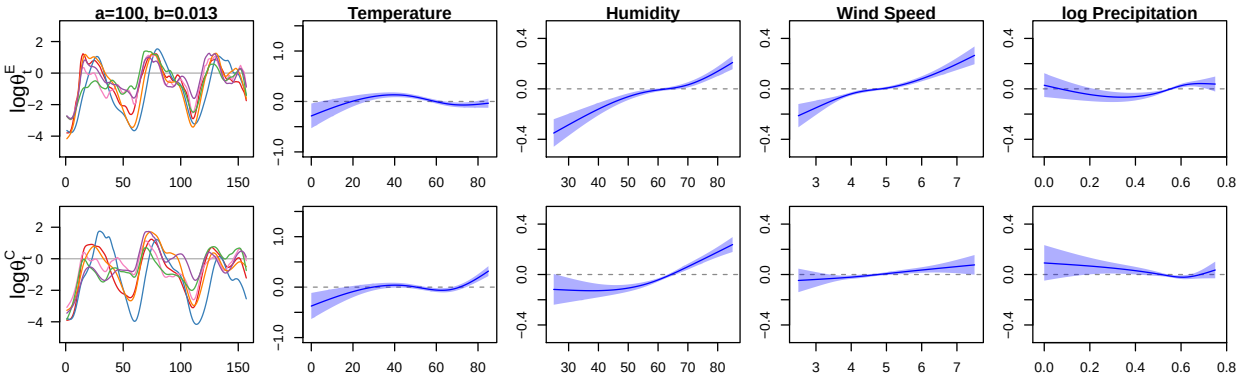
Figure A.30: Results for $\mathrm{E}[\tau_{\mathrm{G}}]$=5,000 and three values of a. For each analysis, estimated smoothers of meteorological variables for log relative risk of EV71 (top) and CA16 (bottom). Temporal soothers in the left panels, the color corresponds to the region. The solid line is the posterior median and the shaded regions are 95% pointwise credible intervals.



(a) $\mathrm{E}[\tau_{\mathrm{G}}] = 7,500$ and $a = 50$

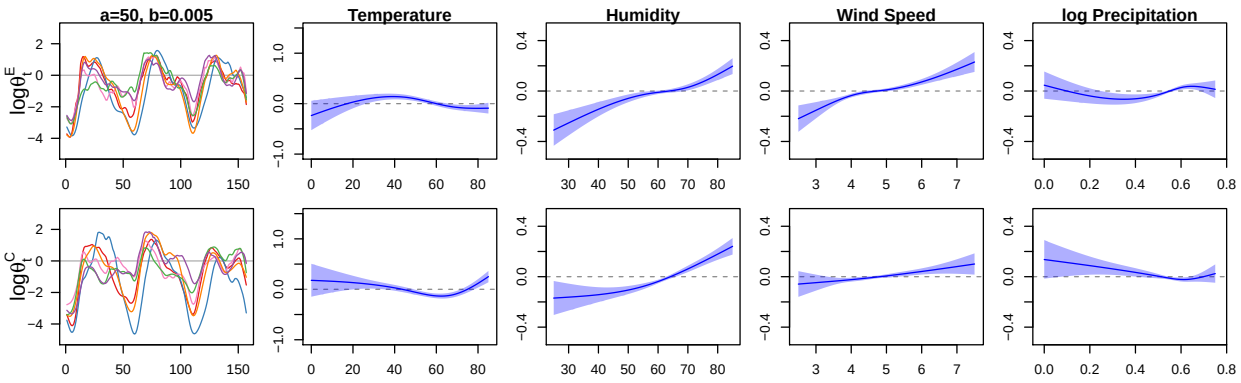


(b) $\mathrm{E}[\tau_{\mathrm{G}}] = 7,500$ and $a = 75$

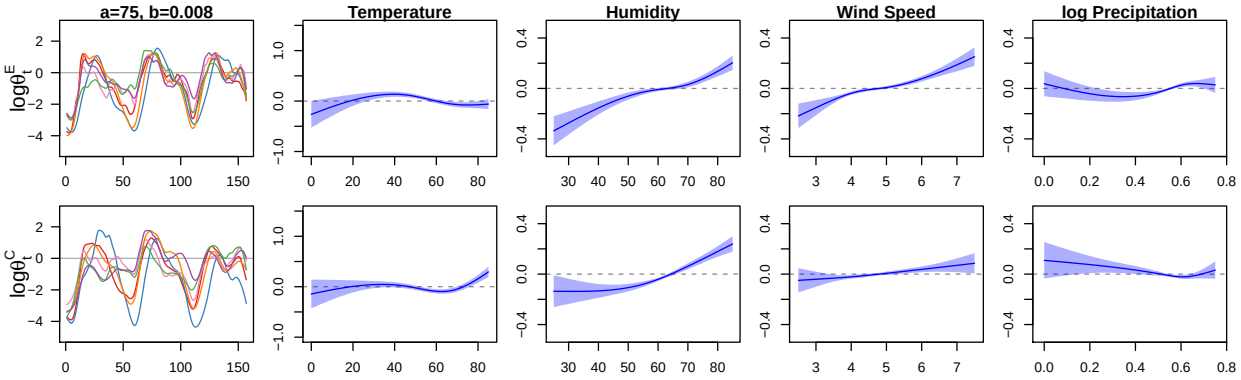


(c) $\mathrm{E}[\tau_{\mathrm{G}}] = 7,500$ and $a = 100$

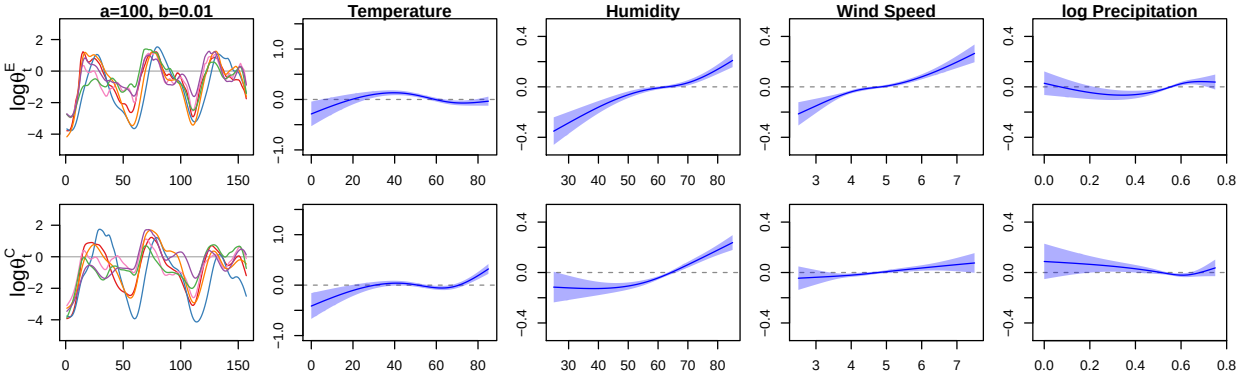
Figure A.31: Results for $\mathrm{E}[\tau_{\mathrm{G}}]$=7,500 and three values of a. For each analysis, estimated smoothers of meteorological variables for log relative risk of EV71 (top) and CA16 (bottom). Temporal soothers in the left panels, the color corresponds to the region. The solid line is the posterior median and the shaded regions are 95% pointwise credible intervals.



(a) $\mathrm{E}[\tau_{\mathrm{G}}] = 10,000$ and $a = 50$



(b) $\mathrm{E}[\tau_{\mathrm{G}}] = 10,000$ and $a = 75$



(c) $\mathrm{E}[\tau_{\mathrm{G}}] = 10,000$ and $a = 100$

Figure A.32: Results for $\mathrm{E}[\tau_{\mathrm{G}}]$=10,000 and three values of a. For each analysis, estimated smoothers of meteorological variables for log relative risk of EV71 (top) and CA16 (bottom). Temporal soothers in the left panels, the color corresponds to the region. The solid line is the posterior median and the shaded regions are 95% pointwise credible intervals.

Appendix B

APPENDIX TO CHAPTER 4

B.1 Parameterizations of the shared component model

While `INLA` has a built-in latent model for the shared component model, we briefly explore various parameterizations of the shared component model via simulations. For σ_η^{-2} and σ_ϵ^{-2}, we assume that for each unstructured random effect, the residual relative risks lie between 0.1 and 10 with probability 0.9, and for τ we assume that the residuals lie between 0.5 and 2 with probability 0.95 (Fong et al., 2010). For all models we have the same hyper-priors:

$$\begin{aligned}
\sigma_\eta^{-2} &\sim \text{Gamma}(0.5, 0.0164), \\
\sigma_\epsilon^{-2} &\sim \text{Gamma}(0.5, 0.0164), \\
\tau &\sim \text{Gamma}(0.5, 0.001488).
\end{aligned}$$

1. INLA built-in approach (`besag2`):

$$\begin{aligned}
\log \theta_{1i} &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \eta_i + \delta\phi_i \\
\log \theta_{2i} &= \beta_0 + \beta_1 x_i + \epsilon_i + \phi_i/\delta \\
\eta_i &\sim \text{N}(0, \sigma_\eta^2) \qquad \text{for } i = 1, \ldots, 67 \\
\epsilon_i &\sim \text{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{\phi} &= (\phi_1, \ldots, \phi_{67}) \sim \text{ICAR}(\tau) \\
\delta &\sim \text{N}(0, 0.4106^2)
\end{aligned}$$

2. By hand (using `copy`):

$$\begin{aligned}
\log \theta_{1i} &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \eta_i + \phi_i \\
\log \theta_{2i} &= \beta_0 + \beta_1 x_i + \epsilon_i + \delta\phi_i \\
\eta_i &\sim \mathrm{N}(0, \sigma_\eta^2) \qquad \text{for } i = 1, \ldots, 67 \\
\epsilon_i &\sim \mathrm{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{\phi} &= (\phi_1, \ldots, \phi_{67}) \sim \mathrm{ICAR}(\tau)
\end{aligned}$$

3. By hand without Alachua:

$$\begin{aligned}
\log \theta_{1i} &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \eta_i + \delta\phi_i \\
\log \theta_{2i} &= \beta_0 + \beta_1 x_i + \epsilon + \phi_i/\delta \\
\eta_i &\sim \mathrm{N}(0, \sigma_\eta^2) \qquad \text{for } i = 1, \ldots, 66 \\
\epsilon_i &\sim \mathrm{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, 66 \\
\boldsymbol{\phi} &= (\phi_1, \ldots, \phi_{66}) \sim \mathrm{ICAR}(\tau)
\end{aligned}$$

4. Asymmetric by hand (OI)

$$\begin{aligned}
\log \theta_{1i} &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \phi_i \\
\log \theta_{2i} &= \beta_0 + \beta_1 x_i + \epsilon_i + \delta\phi_i \\
\epsilon_i &\sim \mathrm{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, 67 \\
\boldsymbol{\phi} &= (\phi_1, \ldots, \phi_{67}) \sim \mathrm{ICAR}(\tau)
\end{aligned}$$

5. Asymmetric by hand (NC)

$$\log \theta_{1i} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \eta_i + \delta\phi_i$$
$$\log \theta_{2i} = \beta_0 + \beta_1 x_i + \phi_i$$
$$\eta_i \sim \mathrm{N}(0, \sigma_\epsilon^2) \qquad \text{for } i = 1, \ldots, 67$$
$$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{67}) \sim \mathrm{ICAR}(\tau)$$

We compare the results when there is a strong spatial correlation (i.e. when we expect the shared component model to perform well) in order to decide what model is appropriate for the simulation study. The results are summarized in Table B.1. We see that for our purposes,

Model	Est	SD	95% CI	Coverage	Bias	MSE
1	-0.264	0.904	(-2.148, 1.617)	0.990	0.023	0.035
2	-0.255	0.171	(-0.591, 0.082)	0.960	0.033	0.026
3	-0.253	0.025	(-0.301, -0.204)	0.240	0.035	0.026
4	-0.251	0.164	(-0.573, 0.071)	0.950	0.037	0.026
5	-0.257	0.162	(-0.576, 0.062)	0.950	0.031	0.026

Table B.1: Simulation results to look into different parameterizations of the shared component model.

the `besag2` parameterization (Model 1) produces credible intervals around the estimated treatment effect that are slightly too large, and the mean squared error is larger than the other models. In contrast, our hand-coded approach in Model 2 provides good estimates with reasonable standard error estimates. As in other simulations, omitting the random effects for the treatment area (as in Model 3) result in credible intervals that are too narrow. We also considered two asymmetric parameterizations of the shared component model. We see that the asymmetric approaches seem to perform reasonably well in simulations, with only minor differences in the two. In practice, however, it seems difficult to justify using the asymmetric approach. In the simulations, we use the hand-coded symmetric model, as described in Model 2.

B.2 ICD-9 Code definitions

We define influenza-like illness (ILI) and gastrointestinal illness (GI) by the ICD-9 codes listed in Table B.2. The ILI and GI ICD-9 code definition were adapted from Betancourt et al. (2007). The definition for ILI was additionally informed by CDC (2011).

ILI	GI
079.99	005.9
382.9	008.69
460	008.8
461.9	009.0
465.8	009.1
465.9	009.2
488.0	009.3
488.01	535.00
488.8	535.40
488.81	535.50
488.02	536.2
488.82	555.9
488.09	558.9
488.1	578.0
488.19	787.01
488.89	787.02
487.0	787.03
487.1	787.30
488.11	787.91
488.12	787.99
487.8	
486	
466.0	
490	
786.2	
780.6	
780.60	

Table B.2: ICD-9 codes for ILI and GI

Appendix C

APPENDIX TO CHAPTER 5

C.1 Additional simulation results assessing simplifying assumptions

	$\widehat{\alpha_{\text{AR}}}$				$\widehat{\alpha_{\text{EN}}}$			
Model	Est.	Std. Err	Bias	Cover	Est.	Std. Err	Bias	Cover
$R_0 = 2$								
$\text{Bin}(S_{t-1},\, 1 - e^{-\lambda_t^\dagger})$	0.693	3.5e-04	0.000	0.958	-0.009	0.301	-0.009	0.962
$\text{Bin}(N,\, 1 - e^{-\lambda_t^\dagger})$	0.051	3.5e-04	-0.643	0.00	-1.660	0.313	-1.660	0.006
$\text{Bin}(S_{t-1},\, \lambda_t^\dagger)$	0.593	3.2e-04	-0.100	0.00	0.153	0.299	0.153	0.858
$\text{Bin}(N,\, \lambda_t^\dagger)$	-0.006	3.4e-04	-0.700	0.00	-1.674	0.320	-1.674	0.004
$\text{Poi}(S_{t-1}\,(1 - e^{-\lambda_t^\dagger}))$	0.693	3.9e-04	0.000	0.970	-0.032	0.305	-0.032	0.958
$\text{Poi}(N\,(1 - e^{-\lambda_t^\dagger}))$	0.056	3.7e-04	-0.637	0.00	-1.443	0.285	-1.443	0.008
$\text{Poi}(S_{t-1}\,\lambda_t^\dagger)$	0.597	3.5e-04	-0.096	0.00	0.145	0.298	0.145	0.860
$\text{Poi}(N\,\lambda_t^\dagger)$	0.000	3.5e-04	-0.693	0.00	-1.688	0.321	-1.688	0.000
$R_0 = 1$								
$\text{Bin}(S_{t-1},\, 1 - e^{-\lambda_t^\dagger})$	-0.023	0.0232	-0.023	0.896	0.467	0.458	0.467	0.728
$\text{Bin}(N,\, 1 - e^{-\lambda_t^\dagger})$	-0.024	0.0233	-0.024	0.896	0.471	0.456	0.471	0.726
$\text{Bin}(S_{t-1},\, \lambda_t^\dagger)$	-0.024	0.0233	-0.024	0.896	0.468	0.457	0.468	0.728
$\text{Bin}(N,\, \lambda_t^\dagger)$	-0.024	0.0233	-0.024	0.896	0.470	0.456	0.470	0.726
$\text{Poi}(S_{t-1}\,(1 - e^{-\lambda_t^\dagger}))$	-0.023	0.0233	-0.023	0.896	0.467	0.458	0.467	0.728
$\text{Poi}(N\,(1 - e^{-\lambda_t^\dagger}))$	-0.024	0.0233	-0.024	0.896	0.471	0.456	0.471	0.726
$\text{Poi}(S_{t-1}\,\lambda_t^\dagger)$	-0.024	0.0233	-0.024	0.896	0.468	0.457	0.468	0.728
$\text{Poi}(N\,\lambda_t^\dagger)$	-0.024	0.0233	-0.024	0.896	0.470	0.456	0.470	0.726
$R_0 = 0.85$								
$\text{Bin}(S_{t-1},\, 1 - e^{-\lambda_t^\dagger})$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Bin}(N,\, 1 - e^{-\lambda_t^\dagger})$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Bin}(S_{t-1},\, \lambda_t^\dagger)$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Bin}(N,\, \lambda_t^\dagger)$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Poi}(S_{t-1}\,(1 - e^{-\lambda_t^\dagger}))$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Poi}(N\,(1 - e^{-\lambda_t^\dagger}))$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Poi}(S_{t-1}\,\lambda_t^\dagger)$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892
$\text{Poi}(N\,\lambda_t^\dagger)$	-0.207	0.076	-0.045	0.958	0.126	0.284	0.126	0.892

Table C.1: Estimates, standard error, bias, and coverage of the MLE estimates of α_{AR} and α_{EN} for three different simulation scenarios.

C.2 Measles analysis with non-informative priors

We also consider the results of the model which does not have a strong prior for the vaccine effect, ϕ. Posterior median and 95% credible intervals are presented in Table C.2. Compared with the results presented in the chapter, the estimates for the vaccine effect have a much wider credible interval, although the median is slightly higher in this setting. The estimate α_{AR} are slightly larger in the non-informative setting. The α_{AR}, ϕ and α_{EN} are highly

	Median	2.5%	97.5%
$\widehat{\alpha_{\text{AR}}}$	1.29	-0.83	1.86
$\hat{\phi}$	0.97	0.18	1.00
$\widehat{\alpha_{\text{EN}}}$	3.88	1.85	4.34
$\hat{\gamma}$	0.71	0.55	0.87
$\hat{\delta}$	-0.19	-0.35	-0.04
$\hat{R}_0$	3.63	0.43	6.40

Table C.2: Posterior medians and 95% credible intervals for the measles biweekly data with non-informative priors.

correlated, as we see in Figure C.1.

In Figure C.2 we compare the fit of the ecological model with the observed data. Compared to the analysis with stronger priors that was presented in the chapter, the the fitted values for the analysis that has non-informative priors does not appear to fit the data as well as the model that include stronger priors.

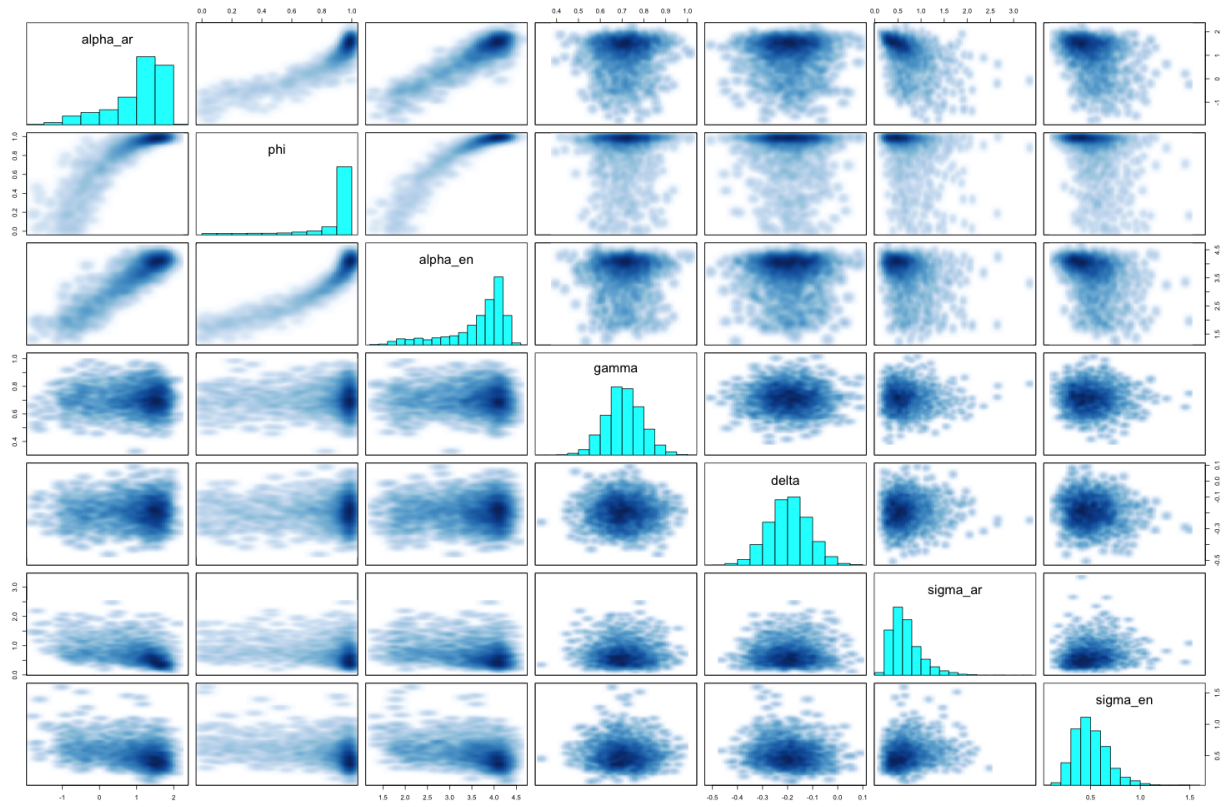


Figure C.1: Pairwise correlation of posterior samples for ecological model with noninformative priors.

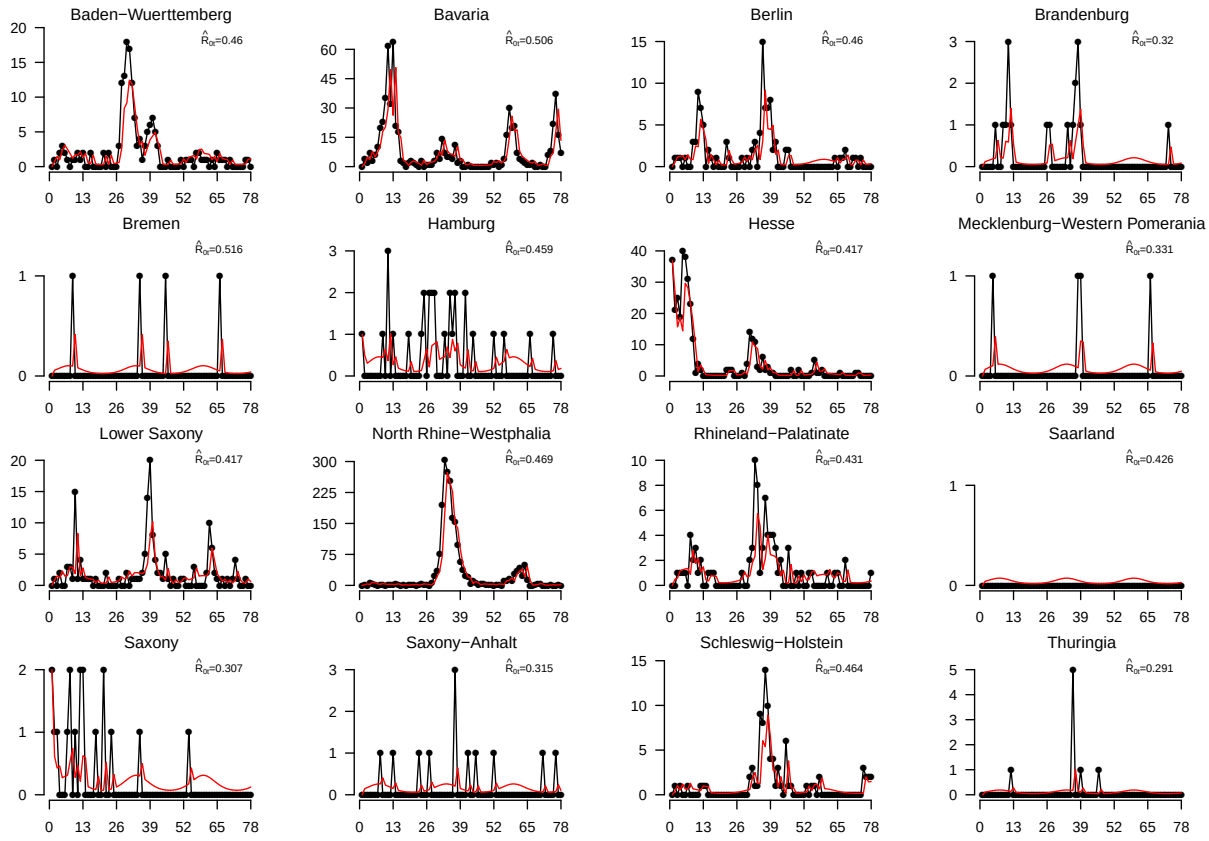


Figure C.2: Fitted cases for ecological model with noninformative priors. Black points denote observed data, red lines are fitted values.

C.3 Measles analysis with informative priors

In Figure C.3 we plot the pairwise correlation of the posterior samples for the ecological vaccine model with informative priors.

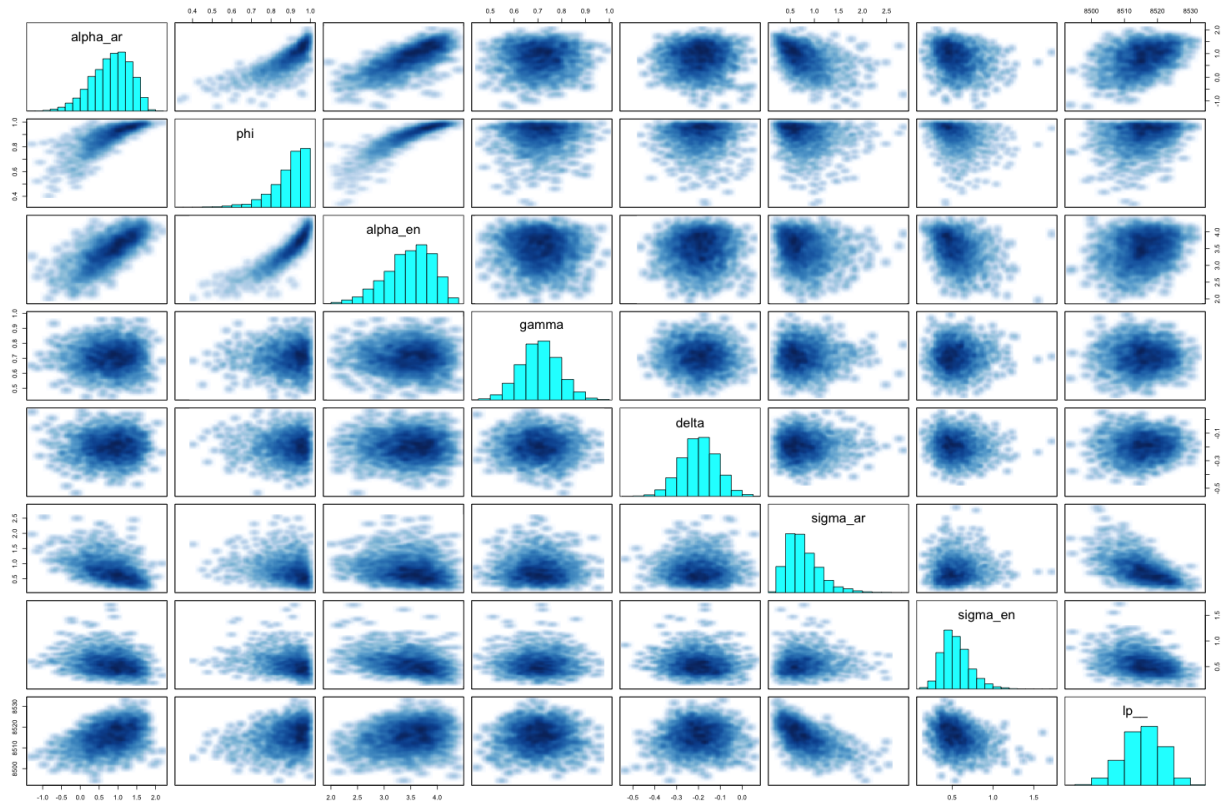


Figure C.3: Pairwise correlation of posterior samples for ecological model with informative priors.

VITA

Leigh Fisher was born a third generation Arizonan to Patricia and Michael Fisher in Phoenix, Arizona. After attending boarding school at The Taft School in Watertown, Connecticut, she decided to return to her southwestern roots for college. She graduated in 2005 from Pomona College in Claremont, California with degrees in Mathematics and Philosophy. After college, she worked as a financial advisor and as an assistant coach to a competitive junior rowing program. In 2014 she earned her M.S. in Biostatistics from the University of Washington in Seattle, Washington. In 2016, under the supervision of Dr. Jon Wakefield, she earned her Ph.D. in Biostatistics from the University of Washington.