Published in final edited form as:

Biometrics. 2017 March ; 73(1): 283–293. doi:10.1111/biom.12560.

Time Series Modeling of Pathogen-Specific Disease Probabilities with Subsampled Data

Leigh Fisher^{1,*}, Jon Wakefield^{1,2}, Cici Bauer³, and Steve Self⁴

¹Department of Biostatistics, University of Washington, Seattle, WA

²Department of Statistics, University of Washington, Seattle, WA

³Department of Biostatistics, Brown University, Providence, Rhode Island, USA

⁴Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA

Summary

Many diseases arise due to exposure to one of multiple possible pathogens. We consider the situation in which disease counts are available over time from a study region, along with a measure of clinical disease severity, for example, mild or severe. In addition, we suppose a subset of the cases are lab tested in order to determine the pathogen responsible for disease. In such a context, we focus interest on modeling the probabilities of disease incidence given pathogen type. The time course of these probabilities is of great interest as is the association with time-varying covariates such as meteorological variables. In this set up, a natural Bayesian approach would be based on imputation of the unsampled pathogen information using Markov Chain Monte Carlo but this is computationally challenging. We describe a practically implementable approach to inference in which we use an empirical Bayes procedure in a first step to estimate summary statistics. We then treat these summary statistics as the observed data and develop a Bayesian generalized additive model. We analyze data on hand, foot and mouth disease (HFMD) in China in which there are two pathogens of primary interest, enterovirus 71 (EV71) and Coxsackie A16 (CA16). We find that both EV71 and CA16 are associated with temperature, relative humidity and wind speed, with reasonably similar functional forms for both pathogens. The important issue of confounding by time is modeled using a penalized B-spline model with a random effects representation. The level of smoothing is addressed by a careful choice of the prior on the tuning variance.

Keywords

Empirical Bayes; generalized additive models; infectious disease modeling; subsampled data

* lhfisher@uw.edu.

7. Supplementary Materials

Web Appendices and Figures referenced in Sections 1, 2, 3.3, 4.2, 4.4, 5.1, and 5.2 are available with this paper at the Biometrics website on Wiley Online Library. An example dataset and sample R code can also be found with the electronic version of this paper.

1. Introduction

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 paper, 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 Coxsackie 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 this infectious disease. 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). More information about these data can be found in Wang et al. (2011).

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 associations differ between the two primary pathogens. If the pathogen responsible for disease were available for all cases, it would be possible to use inferential methods developed for compartmental models, for example susceptible-infectious-recovered (SIR) models in continuous (Ionides et al., 2006) or discrete time (Lekone and Finkenstädt, 2006; Held et al., 2005; McKinley et al., 2009). 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. In this setting, information regarding the responsible pathogen is only available for a subsample of cases, so we have the additional task of estimating the pathogen information for the unsampled cases. For influenza, another multi-pathogen infectious disease with both clinical and laboratory surveillance data, a common approach is to use standard regression methods to estimate the proportion of observed cases attributable to each of the potential components (Bollaerts et al., 2013; Pitman et al., 2007; Dusho et al., 2006).

In this paper we develop a simple method to quickly obtain weekly stratum-specific estimates of the total number of cases attributable to each pathogen, even when there are few cases subsampled. The estimates can then be used for modeling the pathogen-specific temporal dynamics of interest.

The epidemic curves for HFMD epidemics from 2009-2011 for all of China are presented in Figure 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. With each of the three years, there are two epidemic peaks, one in the early summer (June), and the other in the fall (September-October).

For the remainder of the paper, we consider a stratified population partitioned on 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 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.

2. Notation

We consolidate the data into six geographical regions of mainland China for this analysis. Web Figure 1 shows the map of the regions. Within a particular geographic region, we let N_j be the population in the j th combined age-gender stratum, with this population being assumed constant over time. Let Y_{tj}^G be the total number of new 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 in demographic stratum j so that $Y_{tj}^G = Y_{tj}^{GM} + Y_{tj}^{GS}$; these numbers are unobserved. Instead we observe $Y_{tj}^{S\cdot}$ and $Y_{tj}^{M\cdot}$, the total number of severe and mild cases, respectively, in week t and strata j .

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}^{GM}, Z_{tj}^{GS} be the number of mild and severe cases of pathogen type G so that $k_{tj}^M = Z_{tj}^{EM} + Z_{tj}^{CM} + Z_{tj}^{OM}$ and $k_{tj}^S = Z_{tj}^{ES} + Z_{tj}^{CS} + Z_{tj}^{OS}$. Figure 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. Model Formulation

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

3.1 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:

$$Y_{tj}^E, Y_{tj}^C, Y_{tj}^O, N_j - Y_{tj}^E - Y_{tj}^C - Y_{tj}^O | s_{tj} \sim \text{Multinomial}(N_j, s_{tj}),$$

where $s_{tj} = [s_{tj}^E, s_{tj}^C, s_{tj}^O]$ and

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

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:

$$Z_{tj}^{ES}, Z_{tj}^{CS}, Z_{tj}^{OS} | k_{tj}^S, Y_{tj}^{ES}, Y_{tj}^{CS}, Y_{tj}^{OS} \sim \text{MultHyperGeom}(Y_{tj}^{ES}, Y_{tj}^{CS}, Y_{tj}^{OS}, k_{tj}^S) \quad (1)$$

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

(2)

3.3 Model Development

If we were to model the data via a discrete time SIR model the probabilities $s_{tj}^E, s_{tj}^C, s_{tj}^O$ would depend on the number of cases of the respective pathogens in the previous week, but these variables are unobserved and to introduce these as auxiliary variables in the case of large populations has a substantial computational burden. (See Web Appendix F for simulations to support this.)

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

$$Y_{tj}^G | s_{tj}^G \sim \text{Poisson} \left(N_j s_{tj}^G \right), \quad (3)$$

for pathogens of type $G = E, C, O$. 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). 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. See Web Appendix A.2 for further details.

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

$$s_{tj}^G = \theta_t^G \times p_j^G, \quad (4)$$

where $p_j^G = \Pr(\text{Pathogen } G \text{ case in a generic week} \mid \text{stratum } j)$ and θ_t^G is the relative risk 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

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

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

$$Y_t^G = \sum_{j=1}^J Y_{tj}^G \sim \text{Poisson} \left(\theta_t^G \times E^G \right). \quad (5)$$

The usual estimate of the relative risk is

$$\widehat{\theta}_t^G = \frac{Y_t^G}{E^G} = \frac{Y_t^{GM} + Y_t^{GS}}{E^G} = \sum_{j=1}^J \frac{Y_{tj}^{GS} + Y_{tj}^{GM}}{E^G}, \quad (6)$$

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.

4. The Likelihood

4.1 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.

4.2 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 (1) and (2) are the likelihoods for severe and mild cases, respectively.

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

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

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

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

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, $\mathbf{Y}_{tj}^S = (Y_{tj}^{ES}, Y_{tj}^{CS}, Y_{tj}^{OS})$ be the unobserved cases we wish to estimate and $\mathbf{Z}_{tj}^S = (Z_{tj}^{ES}, Z_{tj}^{CS}, Z_{tj}^{OS})$ 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,

$$\begin{aligned} \mathbf{Z}_{tj}^S | \mathbf{Y}_{tj}^S, k_{tj}^S &\sim \text{MultHyperGeom}(\mathbf{Y}_{tj}^{ES}, \mathbf{Y}_{tj}^{CS}, \mathbf{Y}_{tj}^{OS}, k_{tj}^S), \\ \mathbf{Y}_{tj}^S | \boldsymbol{\alpha}_j^S &\sim \text{MultPolya}(\mathbf{Y}_{tj}^S; \boldsymbol{\alpha}_j^S), \end{aligned}$$

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

$$\mathbf{Y}_{tj}^S - \mathbf{Z}_{tj}^S | \mathbf{Z}_{tj}^S, \boldsymbol{\alpha}_j^S, k_{tj}^S \sim \text{MultPolya}((\mathbf{Y}_{tj}^S - k_{tj}^S); \boldsymbol{\alpha}_j^S + \mathbf{Z}_{tj}^S). \quad (7)$$

For $G = E, C, O$

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

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

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

The posterior distribution in (7) allows for computation of measures of uncertainty about the estimated counts of pathogen-specific severe cases. In particular, the posterior variance is

$$\begin{aligned} &Var[Y_{tj}^{GS} | Z_{tj}^{GS}, Y_{tj}^S, \boldsymbol{\alpha}_j^S] \\ &= \frac{(Y_{tj}^S - k_{tj}^S) (\alpha_j^{GS} + Z_{tj}^{GS}) (Y_{tj}^S + \alpha_j^S) (\alpha_j^S + k_{tj}^S - \alpha_j^{GS} - Z_{tj}^{GS})}{(\alpha_j^S + k_{tj}^S)^2 (\alpha_j^S + k_{tj}^S + 1)}. \end{aligned} \quad (8)$$

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

$$\begin{aligned}
& Cov \left[Y_{tj}^{ES}, Y_{tj}^{CS} | \mathbf{Z}_{tj}^S, Y_{tj}^S, \boldsymbol{\alpha}_j^S \right] \\
&= - \frac{\left(Y_{tj}^S - k_{tj}^S \right) \left(\alpha_j^{ES} + Z_{tj}^{ES} \right) \left(\alpha_j^{CS} + Z_{tj}^{CS} \right) \left(Y_{tj}^S + \alpha_j^S \right)}{\left(\alpha_j^S + k_{tj}^S \right)^2 \left(\alpha_j^S + k_{tj}^S + 1 \right)}. \quad (9)
\end{aligned}$$

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

We take an empirical Bayes approach to choosing α_j^{GS} and α_j^{GM} for $G = E, C, O$. 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 α_j^{ES} , α_j^{CS} , and α_j^{OS} in the same proportions, with the constraint that the prior sample size $\left(\sum_{j=1}^J \alpha_j^{GS} \right)$ is one. Hence, the priors we obtain are

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

4.3 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

$$\widehat{\theta}_t^G = \frac{1}{E^G} \left(\sum_{j=1}^J \widetilde{Y}_{tj}^{GS} + \widetilde{Y}_{tj}^{GM} \right).$$

We estimate the corresponding variance using the results of equation (8) which simplifies, as a result of independent strata and severity, to

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

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

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

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 ,

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

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

$$\mathbf{W}_{rt} | \boldsymbol{\lambda}_{rt} \sim N_2 (\boldsymbol{\lambda}_{rt}, \mathbf{V}_{rt}), \quad (10)$$

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

4.4 Simulation study

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 can be found in Web Appendix F.

We also compare our procedure to a MCMC analysis of the full probability model in the small population situation, when this approach is computationally feasible. The details of this analysis can be found in Web Appendix G. We find that the hybrid and MCMC

procedures produce similar estimates, with comparable coverage properties and 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.

5. Application to HFMD Data

5.1 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 (10), we model the means as

$$\lambda_{rt}^G | \beta^G, \delta^G, \gamma^G = \beta_{0r}^G + \gamma^G \mathbf{x}_{rt} + g_r(t, \delta^G) + \sum_{k=1}^K f_k(z_k, \beta_k^G), \quad (11)$$

where β_{0r}^G is a region- and pathogen-specific intercept, $\mathbf{x}_{rt}$ 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 $k = 1, \dots, K$ 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. See Web Appendix C for further details.

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. $g_r(t, \delta_t^G) = \delta_t^G$. We place independent priors on the remaining coefficients.

The model for region r and week t is

$$\begin{aligned} W_{rt} | \lambda_{rt} &\sim N_2(\lambda_{rt}, \mathbf{V}_{rt}), \\ \lambda_{rt}^G &= \beta_{0r}^G + \beta_{1r}^G t + \gamma^G x_{rt} + \sum_{k=1}^K \mathbf{B}(z_{rtk}) \beta_k^G + \delta_t^G, \\ \delta_t^G &= -2\delta_{t-1}^G + \delta_{t-2}^G | \tau_G \sim N(0, \tau_G^{-1}), \\ \tau_G &\sim \text{Gamma}(a, b), \quad a=100, b=0.02, \\ \gamma^G &\sim N(\mathbf{0}, \sigma_{\gamma^G}^2 \mathbf{I}), \quad \sigma_{\gamma^G}^2=100, \\ \beta^G &\sim N(\mathbf{0}, \sigma_{\beta^G}^2 \mathbf{I}), \quad \sigma_{\beta^G}^2=100, \end{aligned}$$

where x_{rt} are covariates modeled parametrically, in our case an indicator for school closure, and $\mathbf{B}(z)$ is the vector of B-spline basis functions evaluated at z .

The specification of the gamma prior on the RW2 precision determines the smoothness of the estimated temporal trend and we examine the sensitivity to this key choice. The RW2 model gives fitted values that are linear in the data, i.e. $\hat{\mathbf{W}} = \mathbf{S}(\tau_G) \mathbf{W}$, with the smoother matrix $\mathbf{S}(\tau_G)$ being a function of the precision of the RW2 model. The trace of $\mathbf{S}(\tau_G)$ gives the effective number of degrees of freedom of the RW2 smoother (Ruppert et al. 2003, Section 3.13) and so we place a gamma prior on τ_G which reflects the time scale of the long term temporal confounding we wish to adjust for. In our case we specify a prior with 4–8 degrees of freedom per year. Details about the prior selection can be found in Web Appendix D.

We fit the model using the integrated nested Laplace approximation (INLA) approach (Rue et al., 2009), which has an implementation within the R programming environment. INLA is able to quickly fit such hierarchical Bayesian models by combining numerical integration

and Laplace approximations in an efficient manner. 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.

5.2 Results

Web Figures 11-13 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 $\tau_G \sim \text{Gamma}(100, 0.02)$.

Figure 3 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 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 4 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.

6. 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 (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 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. Example code is included in Web Appendix H.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The first author was supported by 2R01CA095994-05A1 and R21 AI119773 from the National Institutes of Health. The second author was supported by 2R01CA095994-05A1 from the National Institutes of Health. We would also like to thank Håvard Rue for help with implementing these models in INLA, Gavin Shaddick for useful conversations, and China CDC for making the data available. The authors would like to thank an associate editor for constructive comments that helped improve the manuscript.

References

- Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72:269–342.
- Bauer, CXC. PhD thesis. University of Washington; 2012. Bayesian Modeling of Health Data in Space and Time..
- Bollaerts K, Antoine J, Van Casteren V, Ducoffre G, Hens N, Quoilin S. Contribution of respiratory pathogens to influenza-like illness consultations. *Epidemiology & Infection*. 2013; 141:2196–2204. [PubMed: 23217849]
- 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. The association between enterovirus 71 infections and meteorological parameters in Taiwan. *PLoS One*. 2012; 7:e46845. [PubMed: 23071650]
- Dominici F, McDermott, Hastie T. Improved semi-parametric time series models of air pollution and mortality. *Journal of the American Statistical Association*. 2004; 468:938–948.
- Dukic V, Lopes HF, Polson NG. Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association*. 2012; 107:1410–1426.

- Dushoff J, Plotkin JB, Viboud C, Earn DJ, Simonsen L. Mortality due to influenza in the United States —an annualized regression approach using multiple-cause mortality data. *American Journal of Epidemiology*. 2006; 163:181–187. [PubMed: 16319291]
- Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*. 1998; 15:19–40.
- Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*. 2005; 5:187–199.
- Hoadley B. The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*. 1969; 64:216–229.
- Huang Y, Deng T, Yu S, Gu J, Huang C, Xiao G, Hao Y, et al. 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*. 2013; 13:134. [PubMed: 23497074]
- Ionides E, Bretó C, King A. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*. 2006; 103:18438–18443.
- Lekone P, Finkenstädt B. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*. 2006; 62:1170–1177. [PubMed: 17156292]
- Liu MY, Liu W, Luo J, Liu Y, Zhu Y, Berman H, Wu J. Characterization of an outbreak of hand, foot, and mouth disease in Nanchang, China in 2010. *PLoS One*. 2011; 6:e25287. [PubMed: 21980416]
- Ma E, Lam T, Wong C, Chuang S. Is hand, foot and mouth disease associated with meteorological parameters? *Epidemiology and Infection*. 2010; 138:1779–1788. [PubMed: 20875200]
- McKinley T, Cook A, Deardon R. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*. 2009; 5:24.
- O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 1999; 162:121–129.
- Onozuka D, Hashizume M. The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan. *Science of The Total Environment*. 2011; 410–411:119–125.
- Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2006; 169:179–203.
- Pitman R, Melegaro A, Gelb D, Siddiqui M, Gay N, Edmunds W. Assessing the burden of influenza and other respiratory infections in England and Wales. *Journal of Infection*. 2007; 54:530–538. [PubMed: 17097147]
- Rasmussen DA, Ratmann O, Koelle K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol*. 2011; 7:e1002136. [PubMed: 21901082]
- Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2009; 71:319–392.
- Ruppert, D., Wand, M., Carroll, R. *Semiparametric Regression*. Cambridge University Press; New York: 2003.
- Tong CW, Bible JM. Global epidemiology of enterovirus 71. *Future Virology*. 2009; 4:501–510.
- Wakefield, JC., Best, NG., Waller, LA. Bayesian approaches to disease mapping.. In: Elliott, P.Wakefield, JC.Best, NG., Briggs, D., editors. *Spatial Epidemiology: Methods and Applications*. Oxford University Press; Oxford: 2000. p. 104–27.
- Wang Y, Feng Z, Yang Y, Self S, Gao Y, Longini IM, Wakefield J, Zhang J, Wang L, Chen X, et al. Hand, foot and mouth disease in China: Patterns of spread and transmissibility during 2008–2009. *Epidemiology (Cambridge, Mass.)*. 2011; 22:781.
- Wu H, Wang H, Wang Q, Xin Q, Lin H. The effect of meteorological factors on adolescent hand, foot, and mouth disease and associated effect modifiers. *Global Health Action*. 2014; 7:24664. [PubMed: 25098727]

Figure 1.

Epidemic curves of the 2009-2011 hand, foot, and mouth disease epidemics in, China on *log scale*. The majority of cases are cases missing pathogen information. Few cases each week have specific pathogen information. Cases are aggregated by week. This figure appears in color in the electronic version of this article.

Figure 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.

Figure 3.

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.

Figure 4.

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.

Table 1

Number of HFMD Cases & Severe Cases by Year and Demographic Stratum for all of China.

	Population	2009		2010		2011		Total	
		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