HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2020 January 14.

Published in final edited form as:

J Am Stat Assoc. 2019 ; 114(528): 1561–1573. doi:10.1080/01621459.2019.1585250.

A Spatio-Temporal Modeling Framework for Surveillance Data of Multiple Infectious Pathogens with Small Laboratory Validation Sets

Xueying Tang,
Department of Statistics, University of Florida

Yang Yang*,
Department of Biostatistics and Emerging Pathogens Institute, University of Florida

Hong-Jie Yu,
Chinese Center for Disease Control and Prevention

Qiao-Hong Liao,
Chinese Center for Disease Control and Prevention

Nikolay Bliznyuk
Department of Agricultural and Biological Engineering and Department of Statistics, University of Florida

Abstract

Many surveillance systems of infectious diseases are syndrome-based, capturing patients by clinical manifestation. Only a fraction of patients, mostly severe cases, undergo laboratory validation to identify the underlying pathogen. Motivated by the need to understand transmission dynamics and associate risk factors of enteroviruses causing the hand, foot and mouth disease (HFMD) in China, we developed a Bayesian spatio-temporal modeling framework for surveillance data of infectious diseases with small validation sets. A novel approach was proposed to sample unobserved pathogen-specific patient counts over space and time and was compared to an existing sampling approach. The practical utility of this framework in identifying key parameters was assessed in simulations for a range of realistic sizes of the validation set. Several designs of sampling patients for laboratory validation were compared with and without aggregation of sparse validation data. The methodology was applied to the 2009 HFMD epidemic in southern China to evaluate transmissibility and the effects of climatic conditions for the leading pathogens of the disease, enterovirus 71 and Coxsackie A16.

Keywords

Bayesian Modeling; Infectious Diseases; Missing Data; Hand; Foot and Mouth Disease

*Contact: yangyang@ufl.edu.

Web Appendix
Supplementary technical details, tables and figures are available with this paper at the ASA website.

1 Introduction

More than often, a few genetically and antigenically related infectious pathogens cause similar clinical symptoms in human hosts. Some pathogens may be of more public health importance than others, e.g., associated with higher disease burden, and it is important to understand the epidemiology of these pathogens. For example, a spectrum of enteroviruses (EV) are causative agents for the hand, foot and mouth disease (HFMD), a mild disease commonly seen in children under six years of age. However, neurological complications occasionally occur, in particular in children infected with EV71 (Wang et al., 2011). The surveillance system for HFMD in China, like those for many other infectious diseases, is based on clinical diagnosis, that is, with respect to characteristics of symptoms. This system logged a half million cases in 2008, the first year of its deployment, and 1–2 million cases annually thereafter (Xing et al., 2014). Due to logistic constraints, less than 3% of HFMD cases underwent laboratory confirmation to identify the exact causative virus. It is not clear whether the current sampling practice of validation data for the epidemics of HFMD in China is sufficient to uncover the largely unobserved pathogen-specific spatio-temporal dynamics and to quantify associated epidemiological parameters.

Aggregation of the laboratory data over a greater geographic area or a time period is a common practice for analyzing surveillance data and can be used to improve identifiability of model parameters. When analyzing the HFMD dynamics at the provincial level in China, Takahashi et al. (2016) aggregated the laboratory data by region and month, where each region is composed of typically four provinces, but uncertainty in the laboratory data was ignored. Using the same data but with an aggregation by province, Van Boeckel et al. (2016) modeled each pathogen separately by assuming a binomial distribution for the pathogen-specific case count with the sample proportion in the validation set as the probability. This approach is unlikely to address the uncertainty adequately, as a zero count for a pathogen could occur in a small validation set with a nontrivial probability even if that pathogen contributed substantially to the clinical cases. In addition, it has been shown that joint modeling of multiple pathogens could outperform modeling each pathogen separately (Fisher et al., 2017). Fisher et al. (2017) proposed a multivariate hypergeometric structure to fully account for uncertainty in laboratory confirmation, and analyzed the HFMD surveillance data at the multi-province regional level, where each region was stratified by age group and gender. They used an empirical Bayes approach to estimate the pathogen-specific case counts for each region and week. However, the transmission nature of the disease, i.e., the dependence of the disease risk of susceptible individuals on the number of existing infectious individuals, was ignored.

While aggregation of sparse laboratory data reduces uncertainty in the estimation of pathogen-specific case numbers, it is likely that aggregation across large areas introduces ecological bias in these estimated numbers when the inference is conducted at a finer spatial scale. As a result, a systematic study of the statistical performance of aggregation is needed to ensure proper interpretation of epidemiological parameters estimated from aggregated data. In addition, the sampling of clinical cases for laboratory confirmation is often spatially or temporally imbalanced. Conditioning on the same number of sampled cases, would a homogeneous (random) sampling, where the probability of being sampled is constant over

space and time, be a better alternative? These issues are extremely relevant to the policy of virological surveillance in general.

Motivated by the HFMD surveillance data in China, we propose to study the above issues by designing a general modeling framework for surveillance data of multiple infectious pathogens with a laboratory validation subset. Traditional models for infectious disease transmission either do not consider partially observed infection outcomes (Brix and Diggle, 2001; Paul et al., 2008; Meyer and Held, 2014) or address such issues at the individual rather than the population level (Scharfstein et al., 2006; Yang et al., 2010, 2014). As was mentioned before, existing methods for population level surveillance data of multiple pathogens either inadequately account for uncertainty in laboratory confirmation (Takahashi et al., 2016; Van Boeckel et al., 2016) or do not focus on transmissibility of the disease (Fisher et al., 2017), nor do they account for spatial dependence in pathogen-specific case numbers. The proposed Bayesian framework uses a latent spatio-temporal process model for the unobserved pathogen-specific transmission dynamics, which links the observed non-pathogen-specific case counts and laboratory validation data and thereby accounts for important sources of correlation and uncertainty. The performance of this framework under a variety of sampling schemes for the validation subset is assessed in simulation studies. The method is then used to quantify the transmissibility and environmental risk determinants of EV71 and Coxsackie A16 (CA16), two driving pathogens for the epidemic of the HFMD in southern China during 2009.

2 Data and Notation

The HFMD surveillance data were provided by the Chinese Center for Disease Control and Prevention (CCDC), covering the year of 2009 and all 69 prefectures in five southern provinces (Guangdong, Guangxi, Hunan, Jiangxi and Fujian) with an average population size of 0.41 million per prefecture. Prefecture is an administrative unit between province and county. HFMD epidemics are known to be highly seasonal and affected by climatic conditions (Wang et al., 2011; Xing et al., 2014). Weekly averages of temperature, relative humidity and wind speed during 2009 for each prefecture in this subtropical region were obtained from the National Oceanic and Atmospheric Administration of the United States. Demographic data such as the area and population size of each prefecture were obtained from Chinese Bureau of Statistics. During 2009, a total of 210,628 clinical cases were reported in the study region, of which 4,980 (2.36%) were lab-validated. Fig. S1 in the Web Appendix shows the spatial and temporal distributions of all cases as well as lab-validated cases. The sampling probability for lab-validation among severe cases, 71.5%, is much higher than 2.18% among mild cases, but the number of severe cases itself is small, 571 in total (Web Appendix, Table S1). The clinical definition of mild and severe HFMD cases adopted by CCDC was described elsewhere (Wang et al., 2011). The imbalance in the spatial and temporal distributions of the sampling probability is shown in both Table S1 and Fig. S1, regardless of disease severity. Among lab-validated cases, the distributions of pathogen types also differ considerably in both space and time regardless of disease severity, as shown in Table S2. Such imbalance indicates the potential for biased estimation of pathogen-specific case numbers at the prefecture level and pathogen-specific parameters that are

defined at the individual level, if the laboratory data are aggregated over large space and time domains.

We choose week as the time unit since the incubation period of HFMD is approximately one week (Goh et al., 1982). Without loss of generality, suppose there are R prefectures, T weeks, V pathogens and S disease severity levels. Let $Y_{it}^{(vs)}$ be the number of cases in prefecture i with symptom onset during week t who were infected by pathogen v and had disease severity level s, $i = 1, \ldots, R$, $t = 1, \ldots, T$, $v = 1, \ldots, V$ and $s = 1, \ldots, S$. In reality, we only observe the total number of clinical cases aggregated by severity, $Y_{it}^{(+s)} = \sum_{v=1}^{V} Y_{it}^{(vs)}$, instead of the pathogen-specific numbers. Among the $Y_{it}^{(+s)}$ clinical cases, $Z_{it}^{(+s)} = \sum_{v=1}^{V} Z_{it}^{(vs)}$ patients are lab-validated, where $Z_{it}^{(vs)}$ is the number of lab-validated patients attributed to pathogen v. All the numbers of lab-validated patients are observed. For most diseases, patients with severe symptoms are more likely to be sampled for laboratory validation. It is, however, not unreasonable to assume that the sampling is independent of the underlying pathogen conditioning on the severity level. That is, we assume the $Z_{it}^{(+s)}$ lab-confirmed patients are representative of the $Y_{it}^{(+s)}$ patients in space-time (i,t) and severity level s.

3 Methods

3.1 Transmission

Enteroviruses can sustain in groundwater and soil, forming possible environmental reservoir. They are also known to be passed from person to person by close contact. As a result, we consider both transmission routes, referred to as environment-to-human and human-to-human. For susceptible people in any prefecture i, the transmission rate from environmental reservoir, $E_{it}^{(v)}$, is modeled as

$$\log\left(E_{it}^{(v)}\right) = \log(\gamma_E^{(v)}) + \mathbf{X}'_{E,it}\boldsymbol{\eta}_E^{(v)}, \tag{1}$$

where $\gamma_E^{(v)}$ is the pathogen-specific baseline transmission rate. The coefficient vector $\boldsymbol{\eta}_E^{(v)}$ characterizes the effect of covariates, $\mathbf{X}_{E,it}$, on environment-to-human transmission.

For human sources of infection, we consider local transmissions within each prefecture and between neighboring prefectures, but ignore long-distance transmission between non-adjacent prefectures partly because children, the majority of HFMD cases, usually do not travel far. Denote the set of neighboring prefectures of prefecture i by Ω_i and the number of neighbors by ω_i, and let $\bar{\Omega}_i = \Omega_i \cup \{i\}$. Let $\mathbf{1}_{(\cdot)}$ be the indicator function. The human-to-human transmission rate from prefecture j to prefecture i, $H_{ij,t}^{(v)}$, is modeled as

$$\log\left(H_{ij,t}^{(v)}\right) = \mathbf{1}_{(j=i)}\log(\gamma_{H1}^{(v)}) + \mathbf{1}_{(j\in\Omega_i)}\log(\gamma_{H2}^{(v)}) + \mathbf{X}'_{H,it}\boldsymbol{\eta}_H^{(v)} + \alpha_i^{(v)} + \beta_t^{(v)}, \tag{2}$$

where $\gamma_{H1}^{(v)}$ and $\gamma_{H2}^{(v)}$ are baseline human-to-human transmission rates within the same prefecture and between adjacent prefectures, respectively. For notational simplicity, we assume the two types of human-to-human transmission share the same effect $\boldsymbol{\eta}_H^{(v)}$ of the same collection of covariates $\mathbf{X}'_{H,it}$. To account for both spatial heterogeneity and dependence, we include a spatial random effect $\alpha_i^{(v)}$, for which we assume an intrinsic conditional autoregressive (ICAR) structure with variance $\sigma^2_{\alpha^{(v)}}$:

$$\alpha_i^{(v)} \mid \alpha_j^{(v)}, j \in \Omega_i \sim \mathcal{N}\left(\frac{1}{\omega_i}\sum_{j \in \Omega_i}\alpha_j^{(v)}, \frac{\sigma^2_{\alpha^{(v)}}}{\omega_i}\right), \tag{3}$$

subject to the constraint $\sum_{i=1}^{R}\alpha_i^{(v)} = 0$ for each pathogen type v, as all prefectures constitute a connected graph (Gelfand et al., 2010). The temporal variation $\beta_t^{(v)}$ is modeled using regression splines as

$$\beta_t^{(v)} = \mathbf{b}^{\star}(t)'\boldsymbol{\eta}_B^{(v)} = \sum_{k=1}^{K}\eta_{Bk}^{(v)}b_k^{\star}(t), \tag{4}$$

where $\boldsymbol{\eta}_B^{(v)} = (\eta_{B1}^{(v)}, \ldots, \eta_{BK}^{(v)})'$ are the coefficients, and $\mathbf{b}^{\star}(t) = \left(b_1^{\star}(t), \ldots, b_K^{\star}(t)\right)'$ are the centered cubic B-spline basis functions, i.e., $b_k^{*}(x) = b_k(x) - \frac{1}{T}\int_0^T b_k(y)dy$ based on the regular B-spline basis functions $b_k(x)$, $k = 1, \ldots, K$. To better capture the curvature of the temporal trend, we place three inner knots at weeks 20, 30 and 40 near the two peaks and the valley of the epidemic curve, and two external knots at weeks 1 and 53 (Fig. S1(A)). The temporal term in our model satisfies the constraint $\int_0^T \beta_t^{(v)}dt = 0$, so that the baseline transmission rates are identifiable and can be interpreted as the mean rates over time. There are $K = 6$ basis functions in total.

To calculate the incidence rate in each space-time unit, we make the following assumptions about the natural history of disease for HFMD. First, the HFMD has an incubation period (time from infection to symptom onset) of about one week, i.e., cases with symptom onset in week t were infected in week $t - 1$. Second, we assume a case is infectious for one week, i.e., during the symptom onset week. This assumption is reasonable because (i) symptoms of the HFMD usually resolve in a week; (ii) children diagnosed with HFMD might have been home-quarantined or hospitalized if requested by physicians according to the prevention and control guidelines issued by CCDC since 2009 (Chan et al., 2017); and (iii) the infectiousness level via human-to-human contact during the second week was likely much lower than the first week with symptom onset (Wang et al., 2011). Under these assumptions, the disease incidence rate at the individual level is given by

$$\lambda_{it}^{(v)} = E_{i,t}^{(v)} + \sum_{j \in \Omega_i} H_{ij,t}^{(v)} Y_{j(t-1)}^{(v+)}. \tag{5}$$

We adjust $\lambda_{it}^{(v)}$ for covariates at the time of disease onset (t), but one can also adjust it for covariates at the time of infection ($t-1$). As the prefecture-level population size is large, a common practice is to assume that the number of new cases in each space-time unit follows a Poisson distribution, i.e.,

$$Y_{it}^{(v+)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}\left(S_{i(t-1)}^{(v)} \lambda_{it}^{(v)}\right), \tag{6}$$

where $\mathbf{Y}_{it-}^{(v+)} = \{Y_{j(t-1)}^{(v+)} : j \in \bar{\Omega}_i\}$ represents the set of historical cases who contributed to the generation of $Y_{it}^{(v+)}$, and $S_{it}^{(v)}$ is the number of people susceptible to pathogen v at t. In practice, $S_{it}^{(v)}$ is often approximated by the population size N_i when the majority of the population is susceptible (as in the case of HFMD in China during 2009) or the disease is endemic ($S_{it}^{(v)}$ is stable over time). A previous investigation on the same epidemic found an alternative assumption, $Y_{it}^{(v+)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}\left(\lambda_{it}^{(v)}\right)$, to provide more satisfactory fit to the data than (6), though the interpretation of $\lambda_{it}^{(v)}$ would change from the individual level to the population level (Wang et al., 2011). Consequently, we assume $Y_{it}^{(v+)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}\left(\lambda_{it}^{(v)}\right)$ throughout this investigation.

3.2 Pathogenicity

For the HFMD surveillance data, it is sufficient to consider two severity levels ($S = 2$): mild ($s = 1$) and severe ($s = 2$). Given the pathogen-specific case count $Y_{it}^{(v+)}$, we assume that the number of mild cases follows a binomial distribution

$$Y_{it}^{(v1)} \mid Y_{it}^{(v+)} \sim \text{Binomial}\left(Y_{it}^{(v+)}, p_{it}^{(v)}\right), \tag{7}$$

where $p_{it}^{(v)}$ is the probability of developing mild disease for a person in region i who was infected with pathogen v and had symptom onset during week t. Pathogenicity is adjusted for covariates $\mathbf{X}_{P,it}$ via the regressions

$$g(p_{it}^{(v)}) = g(p_0^{(v)}) + \mathbf{X}'_{P,it} \boldsymbol{\eta}_P^{(v)}, \tag{8}$$

where $g(p)$ is an appropriate link function, $p_0^{(v)}$ is the baseline probability of being mild given infection, and $\boldsymbol{\eta}_P^{(v)}$ is the covariate coefficients specific to pathogen v. As the majority of HFMD cases are mild ($p_0^{(v)} \approx 1$), the complementary log-log link function, $g(p) = \text{cloglog}(p) = \log(-\log(1-p))$, is a reasonable choice to better differentiates covariate effects. When $p_0^{(v)}$ is not close to 1, the logit transformation may be appropriate. It is convenient to combine

infection and pathogenicity into a single Poisson structure, conditioning on historic cases and relevant parameters:

$$Y_{it}^{(v1)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}\left(\lambda_{it}^{(v)} p_{it}^{(v)}\right) \quad \text{and} \quad Y_{it}^{(v2)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}\left(\lambda_{it}^{(v)}(1-p_{it}^{(v)})\right). \tag{9}$$

3.3 Laboratory Validation

To link the unknown number of cases, $Y_{it}^{(vs)}$, to the observed number of lab-validated cases, $Z_{it}^{(vs)}$, we assume that, conditioning on severity level, the sampling of cases for lab-validation is random and thus independent of the underlying pathogen. As a result, the sampling process itself is ignorable (Daniels and Hogan, 2008). Given the total number of validated cases, $Z_{it}^{(+s)}$, and the pathogen-specific case numbers, $Y_{it}^{(1s)}, \ldots, Y_{it}^{(Vs)}$, of severity level s in prefecture i and week t, we assume that the numbers of pathogen-specific lab-validated cases, $Z_{it}^{(1s)}, \ldots, Z_{it}^{(Vs)}$, follow a multivariate hypergeometric distribution (Fisher et al., 2017):

$$\Pr\left(Z_{it}^{(1s)}, \ldots, Z_{it}^{(Vs)} \mid Z_{it}^{(+s)}, Y_{it}^{(1s)}, \ldots, Y_{it}^{(Vs)}\right) = \frac{\prod_{v=1}^{V} \binom{Y_{it}^{(vs)}}{Z_{it}^{(vs)}}}{\binom{Y_{it}^{(+s)}}{Z_{it}^{(+s)}}}. \tag{10}$$

To investigate how inference is influenced by aggregation of laboratory validation data, we explore two aggregation schemes: (1) aggregation by neighborhood and (2) aggregation by province. The latter was used by Van Boeckel et al. (2016) and is more aggressive than the former, as the number of prefectures ranges 10–20 in a province and 3–8 in a neighborhood $\bar{\Omega}_i$. Aggregation by region (4–5 provinces) in previous analyses is not considered here as our study area contains only five provinces. Let

$$\mathscr{A}_{it}^{(s)} = \{(j,\tau) : j \in \bar{\Omega}_i \text{ and } |t-\tau| \leq 2 \text{ and } Z_{j\tau}^{(+s)} > 0\}$$

be the collection of prefecture i and its neighbors and weeks close to t, where lab-validation was performed on cases of severity category s. Analogously, define the provincial-level collection

$$\mathscr{B}_{it}^{(s)} = \{(j,\tau) : j \text{ in the same province of } i \text{ and } |t-\tau| \leq 2 \text{ and } Z_{j\tau}^{(+s)} > 0\}.$$

Let $\rho(\mathscr{A}_{it}^{(s)}) = \sum_{(j,\tau)\in\mathscr{A}_{it}^{(s)}} Z_{j\tau}^{(+s)} / \sum_{(j,\tau)\in\mathscr{A}_{it}^{(s)}} Y_{j\tau}^{(+s)}$ be the validation proportion aggregated by neighborhood, and let $\xi_v(\mathscr{A}_{it}^{(s)}) = \sum_{(j,\tau)\in\mathscr{A}_{it}^{(s)}} Z_{j\tau}^{(vs)} / \sum_{(j,\tau)\in\mathscr{A}_{it}^{(s)}} Z_{j\tau}^{(+s)}$ be the proportion

of pathogen v in the aggregated validation set. similarly define $\rho(\mathscr{B}_{it}^{(s)})$ and $\xi_v(\mathscr{B}_{it}^{(s)})$ for the aggregation by province. For any prefecture i and week t, if $Y_{it}^{(+s)} > 0$, the aggregation schemes are:

- By neighborhood: If $\rho(\mathscr{A}_{it}^{(s)}) > 0$ and $\frac{Z_{it}^{(+s)}}{Y_{it}^{(+s)}} < \rho(\mathscr{A}_{it}^{(s)})$, set $\hat{Z}_{it}^{(+s)} = Y_{it}^{(+s)}\rho(\mathscr{A}_{it}^{(s)})$ and $\hat{Z}_{it}^{(vs)} = \hat{Z}_{it}^{(+s)}\xi_v(\mathscr{A}_{it}^{(s)})$, If $\rho(\mathscr{A}_{it}^{(s)}) = 0$, use aggregation by province.
- By province: If $\frac{Z_{it}^{(+s)}}{Y_{it}^{(+s)}} < \rho(\mathscr{B}_{it}^{(s)})$, set $\hat{Z}_{it}^{(+s)} = Y_{it}^{(+s)}\rho(\mathscr{B}_{it}^{(s)})$ and $\hat{Z}_{it}^{(vs)} = \hat{Z}_{it}^{(+s)}\xi_v(\mathscr{B}_{it}^{(s)})$.

For either scheme, in equation (10), replace $Z_{it}^{(+s)}$ with $\hat{Z}_{it}^{(+s)}$ and $Z_{it}^{(vs)}$ with $\hat{Z}_{it}^{(vs)}$. If $\rho(\mathscr{B}_{it}^{(s)}) = 0$, which is very rare, no further aggregation is pursued for prefecture-week (i,t) and its lab-validation component does not contribute to the likelihood.

3.4 Likelihood

Define $\mathbf{Y} = \{Y_{it}^{(vs)}, i = 1, \ldots, R,\ \ s = 1, 2,\ \ t = 1, \ldots, T,\ \ v = 1, \ldots, V\}$ and $\mathbf{Y}_+ = \{Y_{it}^{(+s)}, i = 1, \ldots, R,\ \ s = 1, 2,\ \ t = 1, \ldots, T\}$, and similarly define $\mathbf{Z}$ and $\mathbf{Z}_+$. Let $\boldsymbol{\theta}^{(v)} = (\gamma_E^{(v)}, \gamma_{H1}^{(v)}, \gamma_{H2}^{(v)}, \sigma^2_{\alpha^{(v)}}, p_0^{(v)}, \boldsymbol{\eta}_E^{(v)}, \boldsymbol{\eta}_H^{(v)}, \boldsymbol{\eta}_P^{(v)}, \boldsymbol{\eta}_B^{(v)})'$ and $\boldsymbol{\alpha}^{(v)} = (\alpha_1^{(v)}, \ldots, \alpha_R^{(v)})'$. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(v)} : v = 1, \ldots, V\}$ and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^{(v)} : v = 1, \ldots, V\}$. The joint probability density of complete data is given by

$$f(\mathbf{Z}, \mathbf{Z}_+, \mathbf{Y}, \boldsymbol{\alpha} \mid \boldsymbol{\theta}) = f(\mathbf{Z} \mid \mathbf{Y}, \mathbf{Z}_+) f(\mathbf{Z}_+ \mid \mathbf{Y}) f(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) f(\boldsymbol{\alpha} \mid \boldsymbol{\theta}), \tag{11}$$

where

$$\begin{aligned} f(\mathbf{Z} \mid \mathbf{Y}, \mathbf{Z}_+) &= \prod_{i=1}^{R}\prod_{t=1}^{T}\prod_{s=1}^{2}\binom{Y_{it}^{(+s)}}{Z_{it}^{(+s)}}^{-1}\prod_{v=1}^{V}\binom{Y_{it}^{(vs)}}{Z_{it}^{(vs)}}, \\ f(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) &\propto \prod_{i=1}^{R}\prod_{t=1}^{T}\prod_{v=1}^{V}\exp(-\lambda_{it}^{(v)})\left\{\frac{(\lambda_{it}^{(v)}p_{it}^{(v)})^{Y_{it}^{(v1)}}}{Y_{it}^{(v1)}!}\frac{\left(\lambda_{it}^{(v)}(1-p_{it}^{(v)})\right)^{Y_{it}^{(v2)}}}{Y_{it}^{(v2)}!}\right\}, \end{aligned} \tag{12}$$

and

$$f(\boldsymbol{\alpha} \mid \boldsymbol{\theta}) \propto \prod_{v=1}^{V}(\sigma^2_{\alpha^{(v)}})^{-\frac{R-G}{2}}\exp\left[-\frac{1}{4\sigma^2_{\alpha^{(v)}}}\sum_{i=1}^{R}\sum_{j\in\Omega_i}(\alpha_i^{(v)} - \alpha_j^{(v)})^2\right]. \tag{13}$$

The component $f(\mathbf{Z}_+ \mid \mathbf{Y})$ in (11), which is equivalent to $f(\mathbf{Z}_+ \mid \mathbf{Y}_+)$, does not involve unknown quantities related to transmission or pathogenicity and can be omitted in posterior inference. In (13), G is the number of isolated sets of connected prefectures (Gelfand et al., 2010). For the surveillance data we consider, $G = 1$. The multiplier of the variance in the exponential term is 4 because each pair of neighboring $\alpha_i^{(v)}$ and $\alpha_j^{(v)}$ appears twice in the numerator. A schematic plot of the primary model structure is shown in Fig. S2.

3.5 Priors

We assume $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(V)}$ are mutually independent a priori. For a given pathogen v, the elements of $\boldsymbol{\theta}^{(v)}$ are also assumed to be independent. To avoid impropriety of the posterior, we put proper but weakly informative priors on these parameters. A gamma prior is used for each of the environment-to-human and human-to-human transmission rates, but the shape and rate differ slightly between simulations and the case study, (0.1, 0.1) in the former and (1.0, 0.5) in the latter. For both simulations and the case study, an inverse gamma prior with a shape of 2.1 and a rate of 1.0 is assumed for $\sigma^2_{\alpha^{(v)}}$, and a normal prior with a zero mean and a variance of 1000 is assumed for the coefficients, $(\boldsymbol{\eta}_E^{(v)}, \boldsymbol{\eta}_H^{(v)}, \boldsymbol{\eta}_P^{(v)}, \boldsymbol{\eta}_B^{(v)})$, and the baseline pathogenicity parameters, $\text{cloglog}(\mathbf{p}_0^{(v)})$. We compare the posteriors to the priors in simulations and examine the sensitivity of the case study results to the hyper-parameters of the priors.

3.6 Posteriors and Inference

Let $\mathscr{C}$ denote the set of values of $\mathbf{Y}$ satisfying two constraints:

$$\sum_{v=1}^{V} Y_{it}^{(vs)} = Y_{it}^{(+s)},\ i = 1, \ldots, R,\ t = 1, \ldots, T,\ s = 1, 2, \tag{14}$$

and

$$Y_{it}^{(vs)} \geq Z_{it}^{(vs)},\ i = 1, \ldots, R,\ t = 1, \ldots, T,\ s = 1, 2,\ v = 1, \ldots, V\,. \tag{15}$$

The joint posterior density of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ conditional on the data $\mathbf{Z}$ and $\mathbf{Y}_+$ is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\alpha} \mid \mathbf{Z}, \mathbf{Y}_+) \propto \sum_{\mathbf{Y} \in \mathscr{C}} f(\mathbf{Z} \mid \mathbf{Y}, \mathbf{Z}_+) f(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) f(\boldsymbol{\alpha} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})\,. \tag{16}$$

Since the posterior density is complex and cannot be sampled from exactly, we use Monte Carlo Markov chain (MCMC) methods such as a Gibbs sampler (Gelfand and Smith, 1990) to obtain posterior samples. However, the size of $\mathscr{C}$ is excessively large, making the summation in (16) impractical. To circumvent this difficulty, we consider sampling from the joint posterior density

$$\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{Y} \mid \mathbf{Z}, \mathbf{Y}_+) \propto f(\mathbf{Z} \mid \mathbf{Y}, \mathbf{Z}_+) f(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) f(\boldsymbol{\alpha} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathbf{1}_{(\mathbf{Y} \in \mathscr{C})}. \quad (17)$$

Except for the variances $\sigma^2_{\alpha^{(v)}}$, $v = 1, \ldots, V$, which can be drawn directly from an inverse gamma distribution, the full conditional distributions of all other quantities in $\boldsymbol{\theta}$, as well as $\boldsymbol{\alpha}$ and $\mathbf{Y}$, are not in standard forms. We use a Metropolis step to sample $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. Wakefield et al. (2011) used a Markov basis sampling (MBS) method to sample each individual $Y_{it}^{(vs)}$ in a model for pathogenicity but without transmission. However, in preliminary analyses of the HFMD data, we encountered convergence issues with this method. We propose to use a multinomial proposal for sampling $\mathbf{Y}_{it}^{(s)} = (Y_{it}^{(1s)}, \ldots, Y_{it}^{(Vs)})'$ for given i, t and s. According to the joint probability (11), the full conditional distribution of $\mathbf{Y}_{it}^{(s)}$, conditioning on all other random quantities as well as their sum $Y_{it}^{(+s)}$, can be rewritten as

$$l(\mathbf{Y}_{it}^{(s)}) \prod_{v=1}^{V} \frac{(\lambda_{it}^{(v)} \tilde{p}_{it}^{(vs)})^{Y_{it}^{(vs)}}}{Y_{it}^{(vs)}!} \binom{Y_{it}^{(vs)}}{Z_{it}^{(vs)}} \propto l(\mathbf{Y}_{it}^{(s)}) \prod_{v=1}^{V} \frac{(\lambda_{it}^{(v)} \tilde{p}_{it}^{(vs)})^{Y_{it}^{(vs)} - Z_{it}^{(vs)}}}{(Y_{it}^{(vs)} - Z_{it}^{(vs)})!}, \quad (18)$$

where $\tilde{p}_{it}^{(vs)} = \mathbf{1}_{(s=1)} p_{it}^{(v)} + \mathbf{1}_{(s=2)}(1 - p_{it}^{(v)})$, and $l(\mathbf{Y}_{it}^{(s)}) = \prod_{j \in \bar{\Omega}_i} \prod_{v=1}^{V} \exp(-\lambda_{j(t+1)}^{(v)}) (\lambda_{j(t+1)}^{(v)})^{Y_{j(t+1)}^{(v+)}}$ contains all transmission risks imposed by the $\mathbf{Y}_{it}^{(s)}$ cases of prefecture i through $\lambda_{j(t+1)}^{(v)}$ on the neighboring prefectures during their infectious week $t+1$. Let $\mathbf{Z}_{it}^{(s)} = (Z_{it}^{(1s)}, \ldots, Z_{it}^{(Vs)})'$. This expression suggests we can sample $\mathbf{Y}_{it}^{(s)} - \mathbf{Z}_{it}^{(s)}$ from a multinomial distribution with size $Y_{it}^{(+s)} - Z_{it}^{(+s)}$ and a probability vector of normalized $(\lambda_{it}^{(1)} \tilde{p}_{it}^{(1s)}, \ldots, \lambda_{it}^{(V)} \tilde{p}_{it}^{(Vs)})'$. Denoting the new sample by $\mathbf{Y}_{it}^{(\star s)}$, we then accept it with probability $l(\mathbf{Y}_{it}^{(\star s)}) / l(\mathbf{Y}_{it}^{(s)})$. This Metropolized independence sampling (MIS) approach works well when the proposal provides a reasonable coverage of the domain of $\mathbf{Y}_{it}^{(\star s)}$ (Chib and Greenberg, 1995; Liu, 1996). The sampling approaches for spatial random effects $\boldsymbol{\alpha}$ and parameters $\boldsymbol{\theta}$ are described in Section 1 of the Web Appendix.

To evaluate the performance of a given simulation setting, we examine the distributions of a mixing statistic and the posterior mean squared error (PMSE) calculated for each parameter over simulated epidemics. For a parameter with true value θ_0 and posterior samples $\{x_{ij} : i = 1, \ldots, m; j = 1, \ldots, n\}$ where i and j index chain and iteration respectively, the PMSE is defined as $\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \theta_0)^2$, and the mixing statistic is calculated as $\frac{m(n-1)}{m-1} \frac{\sum_{i=1}^{m} (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{\cdot\cdot})^2}$, where $\bar{x}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$ and $\bar{x}_{\cdot\cdot} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}$. This statistic plus 1 approximates the Gelman-Rubin statistic (Gelman et al., 1995).

3.7 Extension to multiple severity categories and a longer infectious period

When $S > 2$, one can use a nested binomial structure given by

$$Y_{it}^{(vs)} \mid \sum_{k=s}^{S} Y_{it}^{(vk)} \sim \text{Binomial}\left(\sum_{k=s}^{S} Y_{it}^{(vk)}, p_{it}^{(vs)}\right), \quad s = 1, \ldots, S-1, \tag{19}$$

where $p_{it}^{(vs)}$ is the probability of falling into severity category s given that the severity is of level s or higher for a person in region i who was infected with pathogen v and had symptom onset during week t. For example, if $S = 3$ and $s = 1, 2, 3$ represent mild, severe outpatient and severe hospitalized respectively, $p_{it}^{(v1)}$ is the probability of the disease being mild, and $p_{it}^{(v2)}$ is the probability of being outpatient given that the disease is severe. Similar to the two-category scenario, pathogenicity is adjusted for covariates via $g(p_{it}^{(vs)}) = g(p_0^{(vs)}) + \mathbf{X}'_{P,it}\boldsymbol{\eta}_P^{(vs)}$, where the baseline probabilities, $p_0^{(vs)}$, and the covariate coefficients, $\boldsymbol{\eta}_P^{(vs)}$, are now specific to severity category s, $s = 1, \ldots, S-1$. The nested binomial parameterization uniquely determines a multinomial structure via the transformation $\tilde{p}_{it}^{(v1)} = p_{it}^{(v1)}, \tilde{p}_{it}^{(vS)} = \prod_{k=1}^{S-1}(1 - p_{it}^{(vk)})$ and $\tilde{p}_{it}^{(vs)} = p_{it}^{(vs)}\prod_{k=1}^{s-1}(1 - p_{it}^{(vk)})$, $s = 2, \ldots, S-1$, such that

$$(Y_{it}^{(v1)}, \ldots, Y_{it}^{(vS)}) \sim \text{Multinomial}(Y_{it}^{(v+)}, \tilde{\mathbf{p}}_{it}^{(v)}), \tag{20}$$

where $\tilde{\mathbf{p}}_{it}^{(v)} = \tilde{p}_{it}^{(v1)}, \ldots, \tilde{p}_{it}^{(vS)})'$. In the nested binomial parameterization, $\{p_0^{(vs)}, \boldsymbol{\eta}_P^{(vs)} : s = 1, \ldots, S-1\}$ can be independently sampled without violating the intrinsic constraint $\sum_{s=1}^{S} \tilde{p}_{it}^{(vs)} = 1$. Combining infection and pathogenicity into a single Poisson structure, we have $Y_{it}^{(vs)} \mid \mathbf{Y}_{it-}^{(v+)} \sim \text{Poisson}(\lambda_{it}^{(v)}\tilde{p}_{it}^{(vs)})$, $s = 1, \ldots, S$, and (12) becomes

$$f(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) \propto \prod_{i=1}^{R}\prod_{t=1}^{T}\prod_{v=1}^{V} \exp(-\lambda_{it}^{(v)}) \left\{\prod_{s=1}^{S} \frac{(\lambda_{it}^{(v)}\tilde{p}_{it}^{(vs)})^{Y_{it}^{(vs)}}}{Y_{it}^{(vs)}!}\right\}. \tag{21}$$

To extend the model to an infectious period of multiple weeks, let d be the duration of the infectious period, and assume the infectiousness level decays exponentially with a weekly reduction rate of ρ, where $0 < \rho < 1$. Under these assumptions, the disease incidence rate in (5) now takes the form

$$\lambda_{it}^{(v)} = E_{i,t}^{(v)} + \sum_{\tau=t-d}^{t-1} \sum_{j \in \Omega_i} \rho^{t-1-\tau} H_{ij,t}^{(v)} Y_{j\tau}^{(v+)}. \tag{22}$$

and $\mathbf{Y}_{it-}^{(v+)}$ in (6) need to be redefined as $\{Y_{j\tau}^{(v+)}: j \in \bar{\Omega}_i, \tau = t-d, \ldots, t-1\}$. In addition, $l(\mathbf{Y}_{it}^{(s)})$ in (18) should be rewritten as $\prod_{j \in \bar{\Omega}_i} \prod_{\tau=t}^{t+d-1} \prod_{v=1}^{V} \exp(-\lambda_{j\tau}^{(v)}) \left(\lambda_{j\tau}^{(v)}\right)^{Y_{j\tau}^{(v+)}}$ to account for multiple infectious weeks of the $\mathbf{Y}_{it}^{(s)}$ cases. The additional parameter, ρ, may not be identifiable in all settings.

4 Simulation Studies

Using the 69 prefectures in southern China as a template, we simulated epidemics of three pathogens with two severity categories, mild and severe, based on the proposed model, keeping the total number of cases comparable to the real epidemic in 2009 in the same area. The epidemic in each prefecture started with either one case or none, each with probability 0.5, in the first week. For simplicity, we let the infectious period to be one week ($d = 1$), and only temperature was considered as a linear covariate for both transmission and pathogenicity. A total of 100 epidemics were generated for each scenario discussed below. Each epidemic was analyzed with five parallel chains.

We first assessed identifiability of parameters for three sampling proportions of cases for lab-validation: 2%, 5% and 10%, coupled with two sampling designs: (1) balanced design, where all prefecture-weeks share a common validation proportion; and (2) imbalanced design, where a subset of prefecture-weeks was chosen and the validation proportions in these units were scaled to reach an overall validation proportion similar to the balanced design. The sampling proportion among severe cases was set to 70% uniformly for all prefecture-weeks and simulation settings. As the majority of cases were mild, the overall sampling proportion among mild cases was close to that among all cases. For the imbalanced design, the selection of prefecture-weeks was based on fitted probabilities derived from a logistic regression of presence of laboratory validation on prefecture-specific temporal smoothing splines. This was to mimic the spatio-temporal imbalance in the presence/absence of lab-validation in reality. To achieve an overall proportion of 2% for the imbalanced design, prefecture-weeks with fitted probabilities above 40% for the presence of lab-validation were selected and the validation proportions among mild cases in these prefecture-weeks were set to 13.3%. When the overall proportion was increased to 5% and 10%, we explored two variations of the imbalanced design: (I) increasing the validation proportion while fixing the number of lab-validated prefecture-weeks; or (II) increasing the number of prefecture-weeks with lab-validation while keeping the validation proportion unchanged.

The sample mean and standard deviation of the posterior means over 100 simulated epidemics for each parameter are presented in Table 1 for 2% vs. 10% and in Table S4 (Web Appendix) for 2% vs. 5%. Under the balanced design with a 2% lab-validation, most parameters, except for the covariate effects associated with environment-to-person transmission ($\eta_E^{(v)}$, $v = 1, \ldots, 3$), were reasonably identified with no or moderate bias and small standard deviations (SD). The difficulty in estimating $\eta_E^{(v)}$s is likely due to the fact that the magnitude of the environment-to-person transmission rate is relatively small such that

the environmental exposure accounts for much fewer infections than infectious human cases for most of the time and places (see Section 2 and Table S3 in the Web Appendix for a simplified simulation study showing such a possibility). Interestingly, the identifiability of the same parameter varied between pathogens, e.g., the estimates of $\gamma_E^{(3)}$, $\gamma_{H2}^{(3)}$ and $\eta_P^{(3)}$ for pathogen 3 seemed more biased than those for the other two pathogens. Such heterogeneity resulted from the fact that different parameter values between the pathogens yielded different epidemic scales that contained different amounts of information for inference. Increasing the lab-validation proportion from 2% to 10% did in general reduce both biases and SDs. The greatest reduction in SD occurred for $\gamma_{H1}^{(v)}$, $\gamma_{H2}^{(v)}$ and $\eta_E^{(v)}$, $v = 1, \ldots, 3$. Bias decreased substantially for $\eta_E^{(v)}$ and also notably for $\gamma_E^{(v)}$ ($v = 1, 3$), $\gamma_{H1}^{(3)}$ and $\eta_P^{(3)}$. The SDs for pathogenicity parameters, $p_0^{(v)}$'s and $\eta_P^{(v)}$'s, were not affected much by the increase in lab-validation proportion, because the increase mainly occurred among mild cases but not among severe cases, as in reality the latter was already densely sampled. Consequently, pathogenicity parameters benefit much less from increasing lab-validation, and stochastic fluctuation dominated the SDs of their estimators.

In general, the imbalanced designs were associated with slightly larger biases and moderately larger SDs for transmission-related parameters, but performed similarly for pathogenicity-related parameters (except for $\eta_p^{(3)}$), compared to the balanced design under the same lab-validation proportion (Table 1). With 10% lab-validation, imbalanced design II outperformed design I in terms of smaller bias and variation for most parameters, suggesting that, given limited resources, it may be more fruitful to cover a wider range of locations than to sample more patients at a limited number of locations. The distributions of the mixing statistic and the PMSE over the 100 simulated epidemics are shown for the transmission rates and $\text{cloglog}(p_0^{(v)})$ in Fig. 1, covariate effects in Fig. S3, and spatial effects $\boldsymbol{\alpha}^{(v)}$ and temporal effects $\boldsymbol{\eta}_B^{(v)}$ in Fig. S4, with all parameters averaged over the three pathogens. These figures confirm the superiority of the balanced design over imbalanced designs, as well as sampling more locations with fixed lab-validation proportions over sampling more patients at fixed locations, in terms of both mixing and PMSE for estimating transmission rates and the associated covariate coefficients. The superiority of imbalanced design II over design I was more evident when the validation proportion was high, i.e., at 10% vs. at 5%, as shown by PMSE in these figures. The results for the validation proportion of 2% provide no information for comparing the two imbalanced designs, as they are exactly the same in our simulation setup. The pathogenicity-related parameters, $p_0^{(v)}$ (Fig. 1) and $\eta_P^{(v)}$ (Fig. S3), had better mixing under the balanced design than under the imbalanced designs when the validation proportion was low (2–5%), and under imbalanced design II than under design I; however, PMSE was similar across designs and even across validation proportions, for the same reason mentioned above. At the validation proportion of 10%, the balanced design and imbalanced design II performed similarly in terms of both mixing and PMSE for the pathogenicity-related parameters. The differences between validation proportions and sampling designs in estimating the overall temporal trends, $\beta_t^{(v)}$, are less obvious than but

largely consistent with those seen on the temporal effects $\boldsymbol{\eta}_B^{(v)}$ (Figs. S5–S7). The temporal trends were captured with reasonable accuracy, though more so for pathogens 1 and 2 than for pathogen 3. We also compared the performance between an empirical approach and the proposed MCMC model for estimating the pathogen-specific case numbers, $Y_{it}^{(vs)}$ (Web Appendix, Section 3 and Fig. S8).

For all the sampling designs and lab-validation proportions, the posteriors of 50 randomly selected simulated epidemics are compared to the priors for key parameters of pathogen $v = 1$ in Figs. S9–S16, where the posterior density estimates were smoothed with a normal kennel and a Silverman's rule-of-thumb bandwidth (Silverman, 1986). Generally, the posteriors are distinct from the priors and the posterior modes are centered near the true values, suggesting that these parameters are well identified even with a 2% lab-validation. It is also clear that the higher the lab-validation proportion, the more clustered the posterior modes, and the trend is more notable for the balanced design, e.g., for $\boldsymbol{\eta}_S^{(1)}$ in Fig. S14.

We then examined the performance of lab data aggregation for both balanced and imbalanced designs II, fixing the overall lab-validation proportion at 2%. Aggregation improved the mixing of the Markov chains for all parameters, as shown in Figs. 2, S17 and S18. Compared to aggregation by province, aggregation by neighborhood showed slightly better mixing under the balanced design for the transmission rates, spatial effects $\boldsymbol{\alpha}^{(v)}$, and temporal effects $\boldsymbol{\eta}_B^{(v)}$. No qualitative difference in mixing was found between the two aggregation schemes under the imbalanced sampling design, except for the slight advantage of aggregation by neighborhood in sampling $\boldsymbol{\alpha}^{(v)}$. Under the balanced sampling, aggregation by neighborhood reduced notably the PMSE for $\gamma_{H1}^{(v)}$ (Fig. 2) and $\boldsymbol{\eta}_B^{(v)}$ (Fig. S18), inflated the PMSE for pathogenicity-related parameters $p_0^{(v)}$, (Fig. 2) and $\eta_P^{(v)}$ (Fig. S17), and provided comparable estimates for other parameters, in comparison to no aggregation. The performance of aggregation by neighborhood under the imbalanced sampling design was similar to that under the balanced design, except that the PMSE was increased for two more parameters, $\eta_H^{(v)}$ (Fig. S17) and $\boldsymbol{\alpha}^{(v)}$ (Fig. S18). Aggregation by neighborhood gave notably smaller PMSE than aggregation by province for $\gamma_E^{(v)}$, $\gamma_{H1}^{(v)}$ (Fig. 2) and $\boldsymbol{\eta}_B^{(v)}$ (Fig. S18) under the balanced sampling design and for $\boldsymbol{\alpha}^{(v)}$ (Fig. S18) under the imbalanced design. On the other hand, aggregation by province offered slightly better PMSE for $\gamma_{H2}^{(v)}$ (Fig. 2) and $\eta_E^{(v)}$ (Fig. S17) under the balanced design.

While aggregation led to notable inflation in the PMSE for pathogenicity-related parameters ($p_0^{(v)}$ and $\eta_P^{(3)}$) regardless of the sampling design, the actual biases and SDs of the estimates associated with aggregation were still reasonably small compared to the scale of these parameters (Table S5). The only considerable bias was in the estimates for $\eta_P^{(3)}$, perhaps mainly because of the small scale we assumed for this parameter. Further examination of

pathogen-specific $p_0^{(v)}$ and $\eta_P^{(v)}$ confirmed that parameters associated with pathogen 3, in particular $\eta_P^{(3)}$, had much larger PMSE compared to the other two pathogens (Fig. S19).

Finally, we compared the proposed MIS approach to the existing MBS approach for sampling the unobserved pathogen-specific patient counts, $Y_{it}^{(vs)}$. The lab-validation proportions across prefecture-weeks exactly followed the 2009 HFMD data in southern China to best mimic the spatio-temporal imbalance in reality. The overall lab-validation proportions in these simulated epidemics were about 1.7–1.8%. No aggregation was performed. We found no qualitative difference in terms of PMSE between the two approaches in simulations, but the MIS approach is associated with notably better mixing statistics for $\text{cloglog}(p_0^{(v)})$, $\eta_P^{(v)}$ and $\boldsymbol{\alpha}^{(v)}$ (Figs. S20–S22).

5 Case Study

We aim to estimate the transmissibility and effects of climatic factors of EV71 and CA16 in the southern provinces of China during 2009. All other minor HFMD-related enteroviruses were grouped into a single category “Other”. Each patient in the surveillance database is marked as either “mild” or “severe”. Consequently, $V = 3$ and $S = 2$ in this case study. In view of the extremely low and spatio-temporally imbalanced lab-validation proportions in the surveillance data (Fig. S1), we simplified the proposed model by assuming (1) the spatial random effects, $\boldsymbol{\alpha}^{(v)}$, were specific to each province rather than to each prefecture; and (2) the infectious period is one week. Laboratory data were aggregated by neighborhood in the data analysis. Convergence problems (different posterior modes across chains with different initial values) would appear for some parameters if either the model simplifications or the aggregation were not implemented. The simplified and aggregated model is also robust to moderate perturbation of the priors; for example, changing the prior for transmission rates from Gamma(1.5, 0.5) to Gamma(0.1, 0.1) makes little changes to the posteriors. As climatic conditions and population density are relevant to both environmental exposure and human-to-human contact, all transmission rates were adjusted for temperature, relative humidity, wind speed and the logarithm of population density, each as a cubic polynomial. In addition, as pathogenicity is modeled mainly to account for the different sampling probabilities between mild and severe cases, we do not adjust pathogenicity for covariates in this analysis. The MCMC was implemented with ten parallel chains.

Estimates for the transmission rates and pathogenicities are given in Table 2. CA16 was substantially less transmissible from environment to human and from human to human locally (i.e., within prefecture) than EV71 and other enteroviruses. In particular, the environmental transmissibility of CA16 was much lower than the other two pathogens. However, CA16 appeared more transmissible from human to human across neighboring prefectures, about five times that of EV71 and twice that of other enteroviruses. The transmission rates across neighboring prefectures, 0.002–0.01, were almost negligible as compared to the rates within prefectures, 0.58–0.90. However, the role of cross-prefecture transmission in the spatial diffusion of the disease should not be undervalued. Take CA16 for example, the infection risk imposed by 100 infectious cases in neighboring prefectures

(100 × 0.01) is much higher than the environmental risk (0.11). As expected, EV71 showed the highest pathogenicity, 0.58% (95% credible interval: 0.52%, 0.64%), triple that of enteroviruses in the "other" category and over ten times more pathogenic than CA16. Mild disease is usually associated with better mobility and could partially account for the relatively strong cross-prefecture transmissibility of CA16.

Covariate effects on the environmental exposure level were distinct between EV71 and CA16 (Fig. 3). The environmental exposure level to EV71 was not much affected by wind speed or temperature, whereas CA16 preferred less windy conditions and temperature near 25°C. EV71 was most active with relative humidity in the range of 75–85%, whereas, for CA16, the higher relative humidity the better. Both sparsely and densely populated areas were associated with elevated environmental risk of EV71, but the risk of CA16 increased only in densely populated areas. The distinction in covariate effects on human-to-human transmission between EV71 and CA16 was also notable (Fig. 4). In the most common range of wind speed, 1–3 meters per second, EV71 was associated with lower, while CA16 was associated with higher, risk of human-to-human transmission as the wind speed increases. Extremely windy conditions further decreased the risk of CA16. Temperatures above 20°C were slightly more suitable for human-to-human transmission of EV71; in contrast, human to human transmission of CA16 was facilitated by either relatively cold (near 5°C) or very hot (near 30°C) conditions. In the most common ranges of relative humidity (60–85%) and population density (10^2–10^3 people/km^2), there were positive associations for both pathogens.

Spatial heterogeneity in baseline human-to-human transmissibility unexplained by covariates was shown in Fig. S23 by province-level random effects. Fujian Province, the east tip of southern China, had the highest baseline transmissibility for both EV71 and CA16, followed by Hunan for EV71 and Guangdong for CA16. The lowest baseline transmissibility was observed in Guangxi for EV71 and in Jiangxi for CA16. Posterior means of covariate-adjusted effective incidence rates, $\lambda_{it}^{(v)}$s, were averaged over the year and mapped in Fig. S24 (right) at the prefecture level, together with annual averages of posterior means of $Y_{it}^{(v+)}$s (middle) and empirically imputed $Y_{it}^{(v+)}$s (left). The empirical imputation of $Y_{it}^{(vs)}$ was implemented by multiplying $Y_{it}^{(+s)}$ with the observed proportion of pathogen v among all lab-tested cases of severity category s, with the same aggregation as used by the model. As expected, the posterior means of $\lambda_{it}^{(v)}$s and $Y_{it}^{(v+)}$s are very close to each other, and both are similar to the empirically imputed $Y_{it}^{(v+)}$s.

The temporal trends in baseline human-to-human transmissibility unexplained by co-variates were shown in Fig. S25 for each type of pathogen, where a bimodal seasonality was shared by all pathogens. The first peak occurred near weeks 11–12, corresponding to early March when the empirically imputed pathogen-specific epidemic curves grew exponentially. The second peak of human-to-human transmissibility appeared near week 45, corresponding to early November, which was actually after the second peak in the empirical epidemic curves. This gap is likely a result of synergy of both local variation in the epidemics and covariate

adjustment. Taking CA16 as an example, Fujian province had its second epidemic peak at week 50, much delayed than week 40 in Guangdong (Fig. S26). Meanwhile, the interquartile range of temperatures in early November (14 – 23°C) were less suitable for transmission than that (24.5 – 29°C) in mid-September (week 37-38), according to Figs. 3 and 4. Together, they suggest a late fall peak in the temporal effect best explains the winter peak of CA16 cases (Fig. S25). The posterior means of $\lambda_{it}^{(v)}$s were averaged over all prefectures and plotted over time for each province and each pathogen (blue) in Fig. S26, together with averages of posterior means of $Y_{it}^{(v+)}$s (gray) and empirically imputed $Y_{it}^{(v+)}$s (red). As with the spatial patterns, consistency was observed between model-based $\lambda_{it}^{(v)}$s and $Y_{it}^{(v+)}$s. Empirically imputed $Y_{it}^{(v+)}$s followed similar patterns with occasional departures from the smoother model-fitted patterns at some weeks, indicating that the current model fits the data reasonably well.

6 Discussion

We proposed a Bayesian framework for analyzing surveillance data of infectious diseases with sparse laboratory validation set and assessed the identifiability of key parameters related to transmissibility and pathogenicity. Even under a very low overall lab-validation proportion of 2%, pathogen-specific human-to-human transmission rates, probabilities of severe disease, and associated covariate effects can be accurately estimated in the simulations. In our application, covariate effects on environment-to-human transmission risk are the least identifiable. For most infectious pathogens with both human-to-human and environment-mediated transmission routes, the latter is usually relatively low and dominated by the former during epidemics, leading to the difficulty in estimating the latter as has been previously recognized by Eisenberg et al. (2013) and also shown in our simulation studies. Explicit measurement of environmental exposure levels could mitigate this difficulty but is often not feasible due to detection limit or logistic constraints. In such cases, a simplified environmental transmission component is modeled to account for data variation unexplained by human-to-human transmission rather than to draw reliable inference on risk factors associated with environment-to-human transmission.

For some parameters, the improvement in estimation does not seem commensurate with a five-fold increase in the lab-validation proportion (from 2% to 10%), in particular for the imbalanced designs, e.g., $\gamma_{H2}^{(3)}$ in Table 1. This is perhaps related to the complexity of the model and the nature of the data. The surveillance data entail two layers of competing risks: (a) competition among pathogens and (b) competition among three types of infectious sources (environment, cases in local prefectures, and cases in neighbor prefectures). Layer (a) falls in the standard multi-cause survival setting with an independent structure among the latent pathogen-specific times to infection, but is subject to missingness due to limited lab-testing. The identifiability of cause-specific hazards and regressor effects for full data under mild conditions were shown in Heckman and Honore (1989). Layer (b) differs slightly from the standard setting in that it is rarely observed which type of source infects first. The parameters are identifiable largely because the risk levels vary differently over time between

source types. Identifiability of source-specific parameters has been shown empirically in similar settings (Longini and Koopman, 1982; Yang et al., 2010; Meyer and Held, 2014). The random effects $\boldsymbol{a}^{(v)}$ in (2) further complicate identifiability of source-specific parameters when laboratory data are sparse, as seen in our data analysis where prefecture-level random effects led to convergence problems. If the interest is to detect a few outlying prefectures with unusual transmission risk, one could consider shrinkage approaches used for variable selection, e.g., a spike-and-slab prior for the $\alpha_i^{(v)}$'s (Ishwaran and Rao, 2005).

The most computationally challenging component in our framework is the sampling of **Y**. The discrete nature of these variables precludes the use of efficient sampling methods designed for continuous variables, e.g., the Hamiltonian dynamics (Neal, 2010). The high dimension of **Y** also makes it difficult to use partition-based algorithms such as the stochastic approximation Monte Carlo algorithm (Liang et al., 2007). In addition, the formulation of $\lambda_{it}^{(v)}$ (the mean of $Y_{it}^{(v+)}$) in (5), which is a widely accepted choice in the context of disease transmission (Meyer and Held, 2014; Malesios et al., 2017), makes it difficult to find a link function that is linear in all the regressors. In the absence of a generalized linear mixed model representation, efficient analytic techniques such as the integrated nested Laplace approximation are not applicable (Rue and Martino, 2009). On the other hand, the computationally efficient two-step approach suggested by Fisher et al. (2017), i.e., first estimate **Y** solely from the case and laboratory data and then use the asymptotic or posterior distribution of the estimated **Y** for the inference on other parameters, could be adapted to the transmission setting. We proposed the MIS approach with a multinomial proposal to sampling $\{Y_{it}^{(vs)} : v = 1, \ldots, V\}$ simultaneously. The MIS approach is comparable to the MBS approach in terms of PMSE but gives better mixing results for some parameters. When applied to the real data, however, the MBS approach showed less satisfactory mixing than the MIS approach for $\gamma_E^{(v)}$s (Fig. S27) and $\gamma_{H2}^{(v)}$s (Fig. S28), suggesting the new approach is likely more favorable. The performance of the MIS approach could be further improved (Web Appendix, Section 5).

In both simulations and the case study, the infectious period was assumed to be one week. In an additional simulation study, a one-week infectious period was associated with slightly better mixing behavior of the MCMC for most parameters and smaller PMSE for some parameters, compared to a two-week infectious period (Figs. S29–S31). A possible reason is that a longer infectious period induces extra dependence among the elements of **Y** and thereby makes the sampling more challenging. It should be noted that ρ in (22) is not identifiable even with 10% sampling probability for lab-validation and was hence assumed known. It was previously found that a value of 0.2 for ρ provided a better fit to the HFMD epidemic data in 2009 than larger values (Wang et al., 2011). Consequently, $\rho = 0.2$ was assumed for the results presented in Figs. S29–S31. Another caveat of the case study is that children who might have been clinically misdiagnosed as HFMD were not excluded. Some clinical cases sampled for laboratory validation might have been tested negative. The number of test-negative children were not reported to CCDC and are thus not available to infer how many non-tested cases were actually misdiagnosed. Finally, our model does not address underreporting of cases. For instance, some mild cases might not seek medical

assistance at surveillance-covered hospitals, leading to possible underestimation of $p_0^{(v)}$. However, other unknown but systematic underreporting mechanisms might also exist, and how they affect the inference is generally unpredictable. Extension of our model to adjust for misdiagnosis and underreporting is open to future investigation.

Based on the simulation results, we recommend the balanced design for virological surveillance of infectious diseases whenever feasible. If a balanced design is not achievable, it is better to sample more locations for lab-validation rather than to sample more cases from limited locations. To analyze surveillance data with sparse laboratory validation, aggregation could improve inference on pathogen-specific human-to-human transmission rates and their temporal trend, but may compromise the inference on other parameters especially when the laboratory validation is imbalanced. Inferential gain or loss is not always monotonic in the level of aggregation. More importantly, our study builds a foundation for studying immunological interactions among antigenically related pathogens at the population level for a wide spectrum of infectious diseases such as influenza and dengue, which will inform future field and modeling studies on forecasting of epidemics and optimization of control and prevention strategies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Dr. Zijian Feng and Dr. Jing Zhang at Chinese CDC for their support in the early phase of this project. This work was supported by the National Institute of Allergy and Infectious Diseases grant R21-AI119773 and the National Institute of General Medicine grant U54 GM111274.

References

Brix A and Diggle P (2001). Spatiotemporal prediction for log-gaussian cox processes. Journal of the Royal Statistical Society B 63, 823–841.

Chan J, Law C, Hamblion E, Fung H, and Rudge J (2017). Best practices to prevent transmission and control outbreaks of hand, foot, and mouth disease in childcare facilities: a systematic review. Hongkong Medical Journal 23, 177–190.

Chib SC and Greenberg E (1995). Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Daniels MJ and Hogan JW (2008). Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. New York: Wiley.

Eisenberg M, Robertson S, and Tien J (2013). Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. J Theor Biol 324, 84–102. [PubMed: 23333764]

Fisher L, Wakefield J, Bauer C, and Self S (2017). Time series modeling of pathogen-specific disease probabilities with subsampled data. Biometrics 73, 283–293. [PubMed: 27378138]

Gelfand A, Diggle P, Fuentes M, and Guttorp P (2010). Handbook of spatial statistics. Boca Raton: Chapman & Hall/CRC.

Gelfand A and Smith A (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Gelman A, Carlin J, Stern H, and Rubin D (1995). Bayesian Data Analysis. London: Chapman & Hall.

Goh K, Doraisingham S, Tan J, Lim G, and Chew S (1982). An outbreak of hand, foot and mouth disease in singapore. Bulletin of the World Health Organization 60, 965–969. [PubMed: 6297819]

Heckman J and Honore B (1989). The identifiability of the competing risks model. Biometrika 76, 325–330.

Ishwaran H and Rao JS (2005). Spike and slab variable selection: frequentist and bayesian strategies. The Annals of Statistics 33, 730–773.

Liang F, Liu C, and Carroll R (2007). Stochastic approximation in Monte Carlo computation. Journal of the American Statistical Association 102, 305–320.

Liu JS (1996). Metropolized independence sampling with comparison to rejection sampling and importance sampling. Statistics and Computing 6, 113–119.

Longini I and Koopman J (1982). Household and community transmission parameters from final distributions of infections in households. Biometrics 38, 115–126. [PubMed: 7082755]

Malesios C, Demiris N, Kalogeropoulos K, and N. I. (2017). Bayesian epidemic models for spatially aggregated count data. Statistics in Medicine 36, 3216–3230. [PubMed: 28608436]

Meyer S and Held L (2014). Power-law models for infectious disease spread. The Annals of Applied Statistics 8, 1612–1639.

Neal RM (2010). MCMC using Hamiltonian dynamics, Chapter 5 Boca Raton: Chapman & Hall/ CRC Press.

Paul M, Held L, and Toschke A (2008). Multivariate modelling of infectious disease surveillance data. Statistics in Medicine 27, 6250–6267. [PubMed: 18800337]

Rue H and Martino S (2009). Apprixmate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society B 71, 319–392.

Scharfstein DO, Halloran ME, Chu H, , and Daniels MJ (2006). On estimation of vaccine efficacy using validation samples with selection bias. Biostatistics 7, 615–629. [PubMed: 16556610]

Silverman BW (Ed.) (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/ CRC Press.

Takahashi S, Liao Q, Van Boeckel T, and Others (2016). Hand, foot, and mouth disease in china: Modeling epidemic dynamics of enterovirus serotypes and implications for vaccination. PLoS Medicine 13, e1001958. [PubMed: 26882540]

Van Boeckel T, Takahashi S, Liao Q, Xing W, Lai S, Hsiao V, Liu F, Zheng Y, Chang Z, Yuan C, Metcalf C, Yu H, and Grenfell B (2016). Hand, foot, and mouth disease in china: Critical community size and spatial vaccination strategies. Scientific Reports 6, 25248. [PubMed: 27125917]

Wakefield J, Haneuse S, Dobra A, and Teeple B (2011). Bayes computation for ecological inference. Statistics in Medicine 30, 1381–1396. [PubMed: 21341304]

Wang Y, Feng Z, Yang Y, Self S, Gao Y, Wakefield J, Wang L, Zhang J, Chen X, Yao L, Stanaway J, Wang Z, and Yang W (2011). Hand, foot, and mouth disease in china: Patterns of spread and transmissibility. Epidemiology 22, 781–792. [PubMed: 21968769]

Xing W, Liao Q, Viboud C, Zhang J, Sun J, Wu J, Chang Z, Liu F, Fang V, Zheng Y, Cowling B, Varma J, Farrar J, Leung G, and Yu H (2014). Hand, foot, and mouth disease in china, 2008-12: an epidemiological study. Lancet Infectious Diseases 14, 308–318. [PubMed: 24485991]

Yang Y, Halloran ME, Chen Y, and Kenah E (2014). A pathway em-algorithm for estimating vaccine efficacy with a non-monotone validation set. Biometrics 70, 568–578. [PubMed: 24766139]

Yang Y, Halloran ME, Daniels M, Longini IM, B. D. S., and Cummings D (2010). Modeling competing infectious pathogens from a Bayesian perspective: application to influenza studies with incomplete laboratory results. Journal of the American Statistical Association 105, 1310–1322. [PubMed: 21472041]

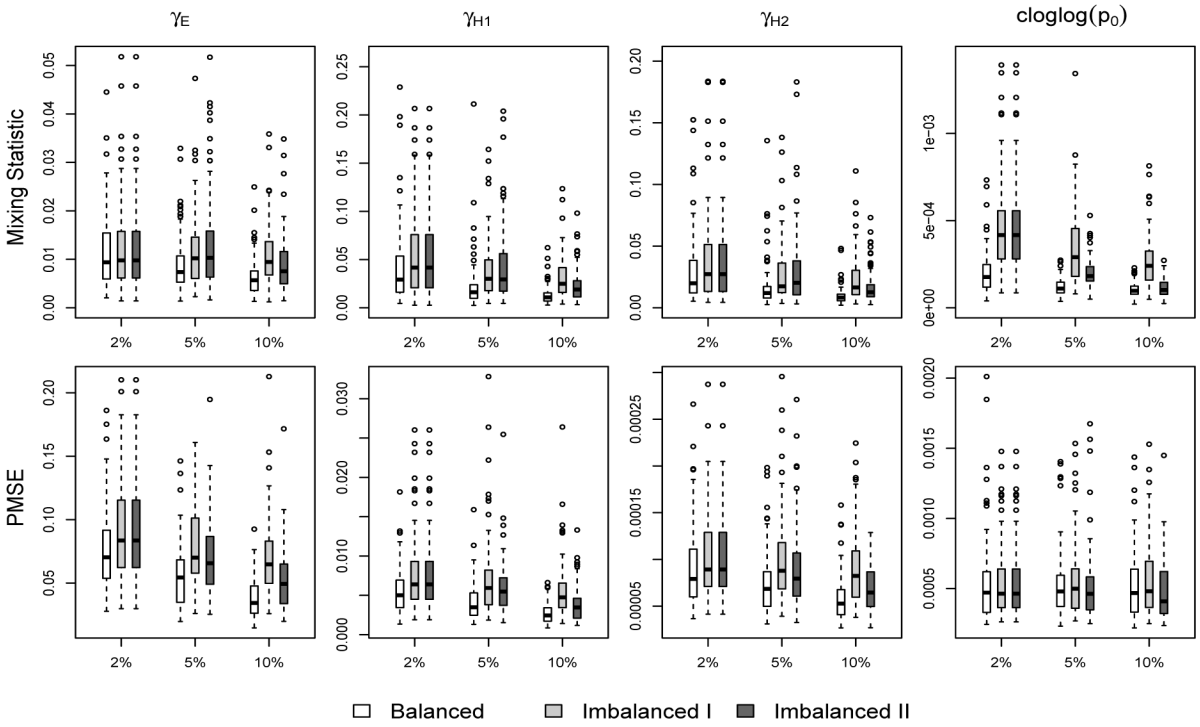

Figure 1:

Simulation study for parameter identifiability: box plots of the mixing statistic and the posterior mean square error (PMSE) over the 100 epidemics for baseline transmission rates ($\gamma_E^{(v)}, \gamma_{H1}^{(v)}, \gamma_{H2}^{(v)}$) and baseline pathogenicity ($\text{cloglog}(p_0^{(v)})$). Presented statistics are averaged over all three pathogens. The box plots are stratified by overall lab-validation proportion (2%, 5% and 10%) and sampling design (white: balanced, light grey: imbalanced design I, dark grey: imbalanced design II). At 2% lab-validation, the two imbalanced designs I and II are exactly the same design and thus the boxes are duplicates of each other.

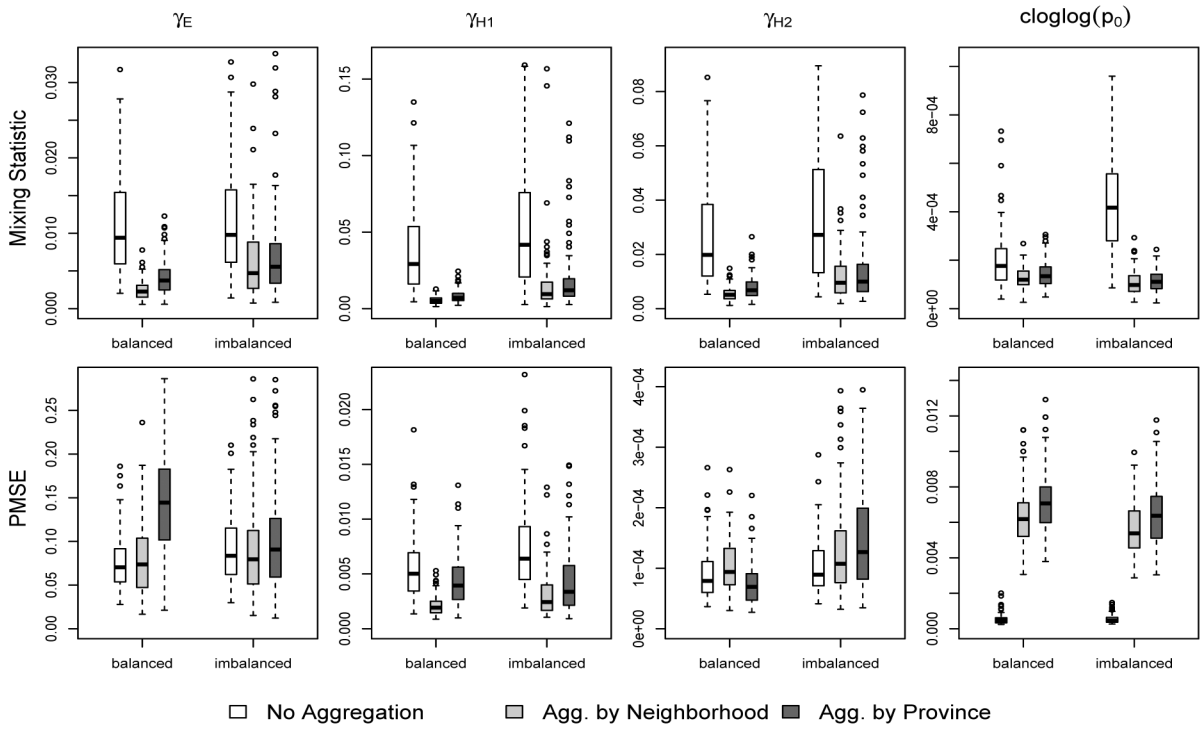

Figure 2:

Simulation study for aggregation of laboratory data: box plots of the mixing statistic and the posterior mean square error (PMSE) over the 100 epidemics for baseline transmission rates $(\gamma_E^{(v)}, \gamma_{H1}^{(v)}, \gamma_{H2}^{(v)})$ and baseline pathogenicity $(\text{cloglog}(p_0^{(v)}))$. Presented statistics are averaged over all three pathogens. The box plots are stratified by lab-sampling design and aggregation scheme (white: no aggregation, light grey: aggregation by neighborhood, dark grey: aggregation by province). The overall lab-validation proportion was set to 2%.

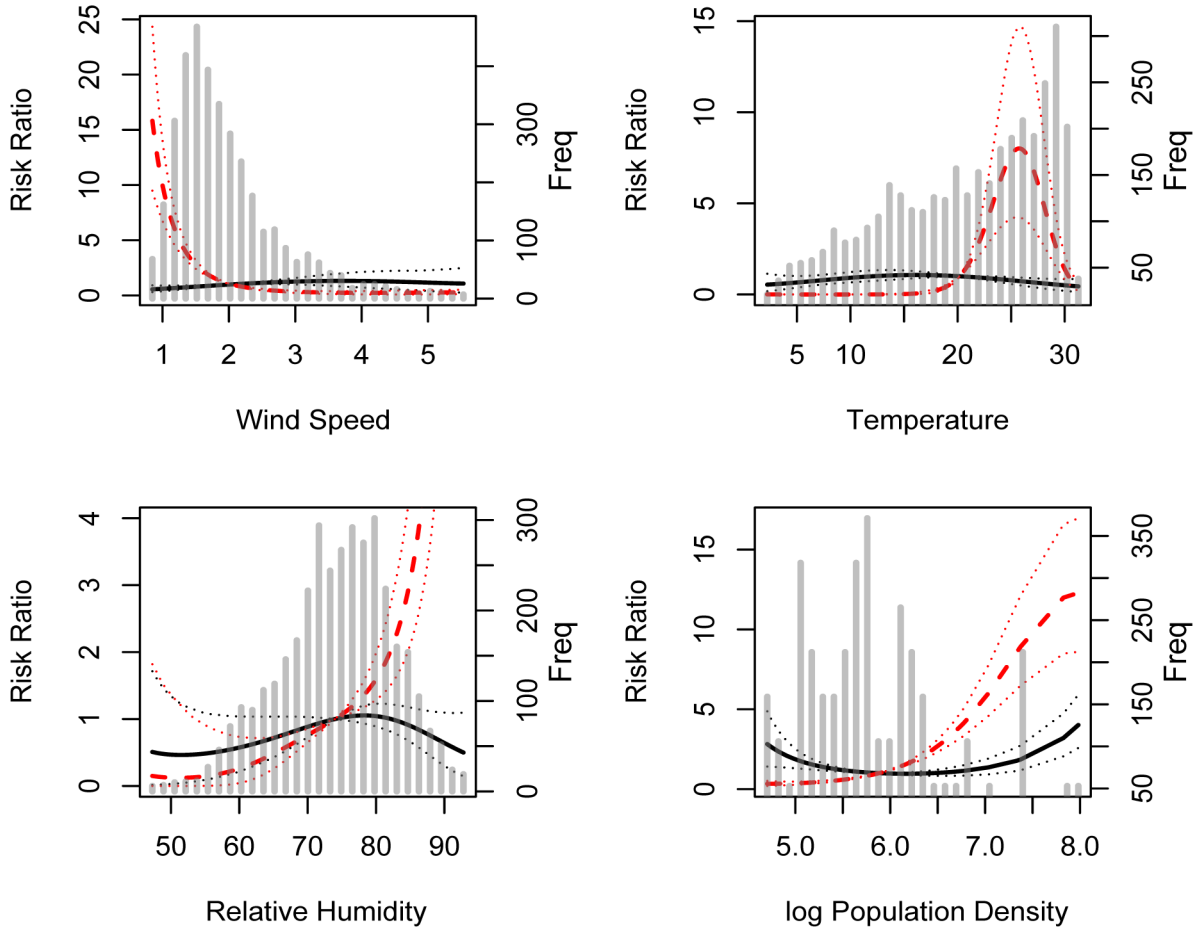

Figure 3:

Covariate effects on baseline environment-to-human transmission rates ($\gamma_E^{(v)}$) for EV71 (solid) and CA16 (dashed). The background histograms in gray represent distributions of the corresponding covariates. Presented are posterior means (solid or dashed) and 95% credible intervals (dotted) of risk ratios ($e^{\eta_E^{(v)}}$).

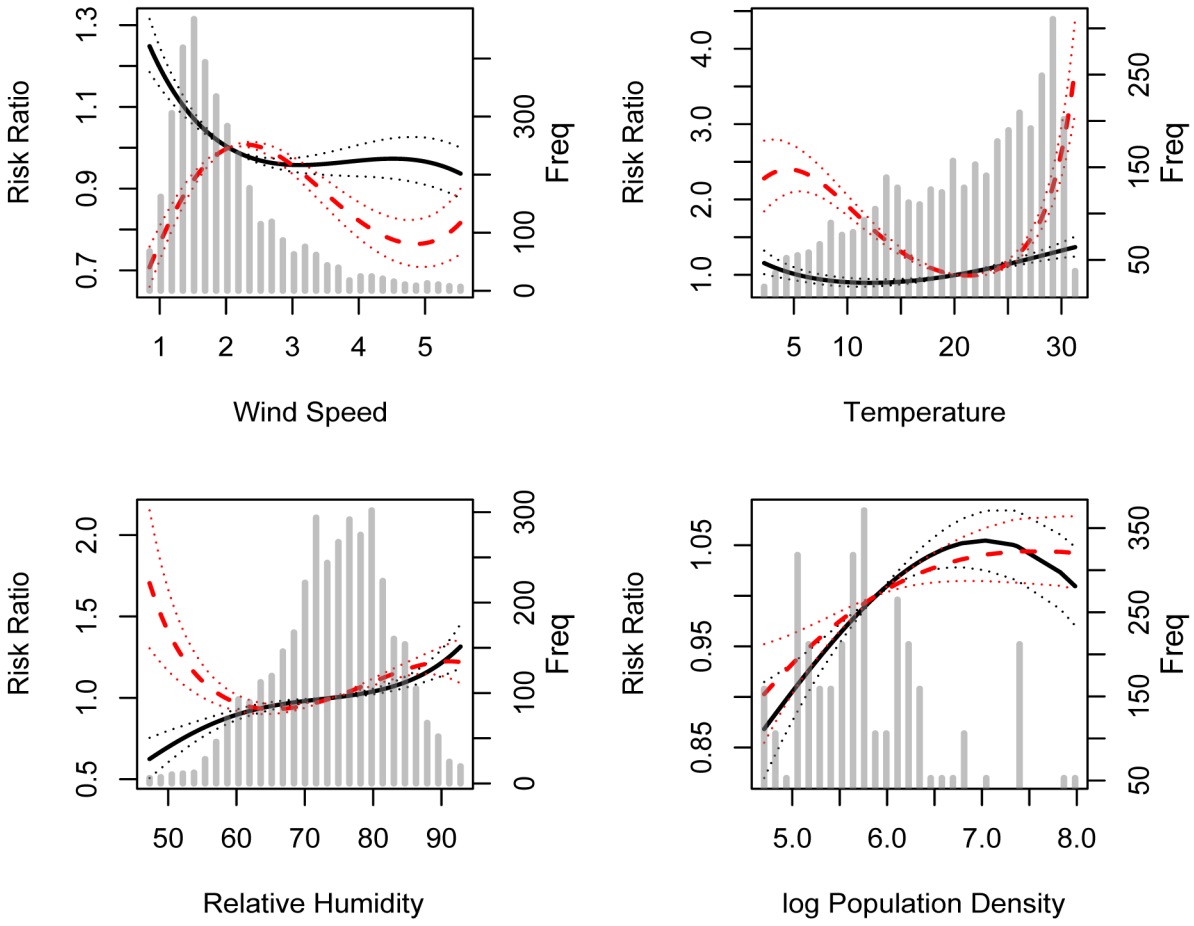

Figure 4:

Covariate effects on baseline human-to-human transmission rates ($\gamma_{H1}^{(v)}$ and $\gamma_{H2}^{(v)}$) for EV71 (solid) and CA16 (dashed). The background histograms in gray represent distributions of the corresponding covariates. Presented are posterior means (solid or dashed) and 95% credible intervals (dotted) of risk ratios ($e^{\boldsymbol{\eta}_H^{(v)}}$).

Table 1:

Average posterior means of pathogen-specific transmission and pathogenicity parameters over 100 simulated epidemics, stratified by sampling design (balanced, imbalanced I and imbalanced II) and lab-validation proportion (2% and 10%).

Parameter	True Value	Average Posterior Mean					SD of Posterior Means				
		Balanced		Imbalanced			Balanced		Imbalanced		
					10%					10%	
		2%	10%	2%	I†	II‡	2%	10%	2%	I†	II‡
$\gamma_E^{(1)}$	1	0.90	0.99	0.89	0.92	0.93	0.130	0.113	0.160	0.146	0.124
$\gamma_E^{(2)}$	1	1.06	1.06	1.04	1.05	1.07	0.165	0.137	0.170	0.164	0.147
$\gamma_E^{(3)}$	1	1.29	1.17	1.32	1.26	1.21	0.130	0.092	0.139	0.125	0.118
$\gamma_{H1}^{(1)}$	0.5	0.51	0.49	0.52	0.51	0.51	0.038	0.031	0.039	0.036	0.027
$\gamma_{H1}^{(2)}$	0.5	0.46	0.46	0.46	0.46	0.46	0.042	0.030	0.049	0.041	0.032
$\gamma_{H1}^{(3)}$	0.5	0.43	0.46	0.41	0.42	0.45	0.037	0.028	0.043	0.038	0.037
$\gamma_{H2}^{(1)}$	0.05	0.053	0.052	0.053	0.054	0.053	0.0048	0.0043	0.0057	0.0055	0.0045
$\gamma_{H2}^{(2)}$	0.05	0.049	0.049	0.048	0.048	0.049	0.0056	0.0043	0.0064	0.0056	0.0049
$\gamma_{H2}^{(3)}$	0.05	0.041	0.043	0.040	0.041	0.042	0.0048	0.0041	0.0048	0.0043	0.0040
$\text{cloglog}(p_0^{(1)})$	1.61	1.61	1.61	1.61	1.62	1.61	0.014	0.014	0.014	0.015	0.013
$\text{cloglog}(p_0^{(2)})$	1.79	1.80	1.79	1.80	1.80	1.79	0.020	0.020	0.019	0.018	0.017
$\text{cloglog}(p_0^{(3)})$	1.71	1.71	1.71	1.71	1.70	1.71	0.014	0.014	0.014	0.016	0.014
$\eta_P^{(1)}$	−0.15	−0.16	−0.15	−0.16	−0.16	−0.16	0.015	0.015	0.015	0.016	0.014
$\eta_P^{(2)}$	−0.22	−0.23	−0.22	−0.23	−0.23	−0.22	0.020	0.020	0.019	0.018	0.018
$\eta_P^{(3)}$	0.06	0.068	0.061	0.070	0.069	0.065	0.016	0.016	0.017	0.016	0.015
$\eta_E^{(1)}$	−0.05	−0.11	−0.09	−0.11	−0.09	−0.09	0.078	0.058	0.091	0.089	0.069
$\eta_E^{(2)}$	0.15	−0.083	0.011	−0.133	−0.098	−0.043	0.097	0.080	0.095	0.087	0.084
$\eta_E^{(3)}$	0.12	0.19	0.13	0.23	0.19	0.15	0.068	0.058	0.078	0.067	0.075
$\eta_H^{(1)}$	0.18	0.18	0.19	0.18	0.18	0.18	0.029	0.023	0.033	0.030	0.024
$\eta_H^{(2)}$	0.35	0.34	0.35	0.34	0.34	0.35	0.016	0.014	0.018	0.017	0.014
$\eta_H^{(3)}$	−0.18	−0.18	−0.18	−0.19	−0.18	−0.18	0.027	0.020	0.030	0.026	0.023

†: Increase the lab-validation proportion in fixed number of prefectures

‡: Increase the number of sampled prefectures but fix the lab-validation proportion

Table 2:

Posterior means and 95% credible intervals (CI) of pathogen-specific transmission rates and pathogenicity for the hand, foot and mouth disease epidemic during 2009 in southern five provinces of China.

Parameter	EV71		CA16		Other	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
$\gamma_E^{(v)}$	0.69	(0.51, 0.91)	0.11	(0.061, 0.18)	0.47	(0.32, 0.70)
$\gamma_{H1}^{(v)}$	0.81	(0.78, 0.84)	0.58	(0.55, 0.61)	0.90	(0.86, 0.94)
$\gamma_{H2}^{(v)}$	0.002	(0.0009, 0.003)	0.010	(0.009, 0.012)	0.005	(0.003, 0.007)
$1-p_0^{(v)}$	0.58%	(0.52%, 0.64%)	0.044%	(0.030%, 0.060%)	0.20%	(0.17%, 0.24%)