Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(529): 66–78. doi:10.1080/01621459.2019.1609480.

Modeling Bronchiolitis Incidence Proportions in the Presence of Spatio-Temporal Uncertainty

Matthew J. Heaton^a, Candace Berrett^a, Sierra Pugh^a, Amber Evans^b, Chantel Sloan^c

^aDepartment of Statistics, Brigham Young University, Provo, UT

^bMPH, Health ResearchTx, LLC, Trevose, PA

^cDepartment of Health Science, Brigham Young University, Provo, UT

Abstract

Bronchiolitis (inflammation of the lower respiratory tract) in infants is primarily due to viral infection and is the single most common cause of infant hospitalization in the United States. To increase epidemiological understanding of bronchiolitis (and, subsequently, develop better prevention strategies), this research analyzes data on infant bronchiolitis cases from the U.S. Military Health System between the years 2003–2013 in Norfolk, Virginia, USA. For privacy reasons, child home addresses, birth dates, and diagnosis dates were randomized (jittered) creating spatio-temporal uncertainty in the geographic location and timing of bronchiolitis incidents. Using spatio-temporal point patterns, we created a modeling strategy that accounts for the jittering to estimate and quantify the uncertainty for the incidence proportion (IP) of bronchiolitis. Additionally, we regress the IP onto key covariates including pollution where we adequately account for uncertainty in the pollution levels (i.e., covariate uncertainty) using a land use regression model. Our analysis results indicate that the IP is positively associated with sulfur dioxide and population density. Further, we demonstrate how scientific conclusions may change if various sources of uncertainty (either spatio-temporal or covariate uncertainty) are not accounted for. Code submitted with this article was checked by an Associate Editor for Reproducibility and is available as an online supplement.

Keywords

Dimension reduction; Measurement error; Respiratory syncytial virus; Spatial statistics

1. Introduction

1.1. Epidemiological Background and Data

Bronchiolitis is a severe acute lower respiratory tract infection (ALRI) that commonly occurs among infants. Most infant bronchiolitis is caused by respiratory syncytial virus

CONTACT Matthew J. Heaton, mheaton@stat.byu.edu, Department of Statistics, Brigham Young University, 223 TMCB, Provo, UT 84602.

Supplementary Materials

Technical details Details regarding the derivation of the likelihood, model fitting and computation. (PDF)

LUR model details Details regarding the LUR model selection and covariates. (PDF)

(RSV) infection during winter season (Shay et al. 2001). Worldwide, there are over 30 million new cases of RSV-associated ALRI each year (Nair et al. 2010). In the United States, the hospitalization rate of infants with RSV was 2345/100,000 person-years from 1993 to 2008 (Zhou et al. 2012). In addition to its acute impact and morbidity, infant bronchiolitis has a strong causal link with the later development of childhood asthma (Heymann et al. 2004; Miller et al. 2013).

Bronchiolitis epidemics emerge from a complex system of interacting components including the poorly understood seasonal circulation of respiratory viruses, individual susceptibility, environmental exposures (such as air pollution), circulating strains of RSV, and climate patterns (e.g., Sirimi et al. 2015; Mehta et al. 2013; Siefker et al. 2016; Walsh et al. 1997; Ye et al. 2016). No effective vaccine has yet been developed, nor do we have a sufficient understanding of epidemic dynamics to allow for their accurate prediction. Currently, monthly RSV immunoprophylaxis is the only available preventive intervention. RSV immunoprophylaxis requires five intramuscular injections and is approved only for those highly susceptible infants with preexisting conditions or those who are born prematurely. The current cost of preventing a single case of bronchiolitis with immunoprophylaxis is estimated at over \$300,000 (Hampp et al. 2011).

To understand drivers of bronchiolitis epidemics, this research seeks to analyze medical claims data between 2003 and 2013 from the Department of Defense (DoD) Military Health System (MHS). The MHS serves all active and retired military personnel and their dependents. While individuals covered by the MHS span the United States, this study focuses on claims occurring in Norfolk, Virginia which is homeport to the Navy's Atlantic Fleet, one of the largest concentrations of naval forces in the United States. The study area lies, approximately, from -76.346° to -76.133° longitude and 36.82° to 36.98° latitude. To focus on bronchiolitis in infants, medical claims from the MHS were isolated for children less than one year of age. The information collected from the MHS database included the birthdate and location of each child and, if the child contracted bronchiolitis, the associated date when they received clinical attention. In total, the MHS data used in this analysis had information on 20,828 susceptible infants of which 2341 contracted bronchiolitis.

Because the MHS database contains highly sensitive protected health information (PHI), the DOD "jittered" all dates (birth and diagnosis dates) to allow us to analyze the data. The amount of jittering is required to be maintained as confidential, though the amount was known to study researchers. Figure 1 shows a kernel density plot of the jittered locations of births in the Norfolk area as well as the density of jittered locations where children contracted bronchiolitis. For further confidentiality reasons, the actual jittered locations are not shown. Comparing the two figures, the locations of those children who contracted bronchiolitis has more spatial variability than all children in general. Hence, we seek here to understand the underlying factors contributing to the larger spread (e.g., environmental exposures, socio-demographics, etc.).

1.2. Research Goals and Contributions

To facilitate epidemiological understanding of the seasonal onsets and drivers of bronchiolitis, this research seeks to estimate, and quantify the uncertainty in, the incidence

proportion (IP) of bronchiolitis using the MHS database described above. The bronchiolitis IP is defined as the number of new bronchiolitis cases per population at risk within a certain spatial and temporal window. Mathematically, for a spatial region $\mathcal{R} \subseteq \mathcal{D} \subset \mathbb{R}^2$ where $\mathcal{D}$ is the entire spatial domain of interest and a time window $\mathcal{W} \subseteq \mathcal{T} \subset \mathbb{Z}^+$ where $\mathcal{T}$ is the total time domain of interest, the IP is defined as,

$$I(\mathcal{R}, \mathcal{W}) = \frac{N_1(\mathcal{R}, \mathcal{W})}{N_0(\mathcal{R}, \mathcal{W})}, \quad (1)$$

where $N_1(\mathcal{R}, \mathcal{W})$ is the number of children who developed bronchiolitis (i.e., “cases”) in $\mathcal{R} \times \mathcal{W}$ while $N_0(\mathcal{R}, \mathcal{W})$ is the number of susceptible children in $\mathcal{R} \times \mathcal{W}$ (i.e., the population at risk). Epidemiological understanding of the seasonal onsets and drivers of bronchiolitis can be determined by regressing $I(\mathcal{R}, \mathcal{W})$ onto appropriate covariates such as environmental exposures, regional demographics, and climate patterns. Thus, our scientific goals can be accomplished through statistical modeling of the IP.

In developing a modeling strategy for $I(\mathcal{R}, \mathcal{W})$ using the MHS data, several issues need to be addressed. First, as mentioned above, the MHS data provides jittered spatial and temporal information leading to spatial and temporal uncertainty in the true location and date of bronchiolitis cases. That is, in (1) neither $N_1(\mathcal{R}, \mathcal{W})$ nor $N_0(\mathcal{R}, \mathcal{W})$ are directly observed for any $\mathcal{R}$ and $\mathcal{W}$ of interest. Of course, the spatio-temporal uncertainty could be ignored but studies by Gabrosek and Cressie (2002) and Cressie and Kornack (2003) suggest that not accounting for this uncertainty may lead to incorrect inference on the main effects of environmental exposures on the IP. Available methods that model spatial uncertainty include Barber, Gelfand, and Silander (2006), Zimmerman (2008), Chakraborty and Gelfand (2010), and Fanshawe and Diggle (2011) but, because of the need to model the “true” spatio-temporal locations, these methods are computationally taxing. In this article, we adopt a strategy similar to Chakraborty and Gelfand (2010) by assuming the true locations and dates follow a nonhomogeneous spatio-temporal point pattern (Baddeley, Rubak, and Turner 2015) with a piecewise constant intensity surface over space and time. We subsequently show that the piece-wise assumption vastly improves computational tractability.

A primary goal of this analysis is to assess the impact of environmental exposures (e.g., pollution) on the associated IP for various spatial regions $\mathcal{R}$ and temporal windows $\mathcal{W}$ of interest. However, pollution information in Norfolk is only available from three pollution monitors; therefore, the pollution concentration throughout all of Norfolk (i.e., the entire spatial domain) is unknown. Additionally, the concentration of pollutants is location-specific (“point referenced” data; see Banerjee, Carlin, and Gelfand 2014) while the IP in (1) is defined areally (i.e., over spatio-temporal regions) leading to a change of support problem when regressing the IP on pollution information (see Gelfand, Zhu, and Carlin 2001; Gotway and Young 2002; Wikle and Berliner 2005, for information and references on the change of support problem).

Accounting for missing covariate data is a challenge on its own. Several approaches have been suggested to accommodate missing covariate values, for example, multiple imputation,

maximum likelihood, fully Bayesian, and weighted estimating equations (see Ibrahim et al. 2005, for an overview of these methods). These same methods have been applied to the case where the covariates may be spatially dependent (e.g., Baker, White, and Mengersen 2014; Panzera, Bendetti, and Postiglione 2016; Rathbun 2013; Waagepetersen 2008; Zidek, Sun, and Le 2000). The fully Bayesian approach is to place a prior distribution on the missing covariate values and estimate their posterior distributions within the Bayesian framework. Due to the large number of locations needing to be imputed (recall we only have three pollution monitors within our region of interest), this is infeasible in our example. One simple solution to this issue is to separately model the pollution information and spatio-temporally kriging the monitoring station data across the spatial domain thereby treating the kriged values as the “true” values. However, such an approach ignores the spatio-temporal uncertainty in the covariates. In this research, we account for this additional uncertainty by building a land-use regression (LUR) model for the monitoring station data and sampling covariate values across the spatio-temporal domain from the modeled predictive distribution.

Due to the short spatial range of pollution concentrations, any impact of pollutants on bronchiolitis IP is likely to be highly localized. Hence, to assess their impacts, modeling for the IP must also be highly localized. Fine scale spatio-temporal modeling using, say, Gaussian processes, however, often leads to computational issues (see Sun, Li, and Genton 2012). Hence, in this research, our modeling strategy is fine-scaled (i.e., high resolution) yet we utilize low-rank approaches to facilitate the spatial modeling.

In conclusion, this research seeks to estimate and quantify the uncertainty for the bronchiolitis IP while accounting for various aspects of spatio-temporal uncertainty and computational complexity. Interestingly, while our model fully accounts for the spatio-temporal uncertainty, we also compare these results to alternative model specifications that fail to do so thus quantifying the potential folly in not accounting for such sources of uncertainty. Epidemiologically, this research furthers understanding of the drivers of the onset and peak rates of bronchiolitis uncertainty on a fine spatial scale.

2. A Point Process Model for Incidence Proportions

2.1. Notation and Model Overview

For purposes of this analysis, let $\mathcal{D} \subset \mathbb{R}^2$ be the continuous spatial domain of interest and $\mathcal{T} \subset \mathbb{Z}^+$ be a discrete subset of the positive integers that corresponds to the time periods of interest. Because the IP is defined for specific regions and time windows, we assume that $\mathcal{D} = \bigcup_r \mathcal{R}_r$ where $\mathcal{R}_1, \dots, \mathcal{R}_R$ is a partition of the spatial domain (i.e., $\mathcal{R}_k \cap \mathcal{R}_j = \emptyset$ for $k \neq j$) into R regions and $\mathcal{T} = \bigcup_w \mathcal{W}_w$ where $\mathcal{W}_1, \dots, \mathcal{W}_W$ are W disjoint subsets of the temporal domain. We assume that $\{\mathcal{R}_r\}$ and $\{\mathcal{W}_w\}$ are carefully chosen for the specific analysis such that inference is desired for $I(\mathcal{R}_r, \mathcal{W}_w)$ for all $r = 1, \dots, R$ and $w = 1, \dots, W$. We further note that the choice of $\{\mathcal{R}_r\}$ and $\{\mathcal{W}_w\}$ may be data dependent. That is, the choice of such regions may depend on the amount of data available.

Notationally, let $s_i^\star \in \mathcal{D}$ denote the recorded birth location (with jitter) of the i th child that is associated with a true birth location s_i where $i = 1, \dots, N$ such that N denotes the total number of children in the dataset. Likewise, let $b_i^\star$ denote the recorded (again, with jitter) birthday of the i th child that is associated with a true (unjittered) birthday b_i . Of the N observed children, we observe that $N_1 < N$ of these children contracted bronchiolitis and define $t_j^\star$ as the recorded time (with jitter) in which bronchiolitis was contracted with an associated true date of t_j for $j = 1, \dots, N_1$.

To properly assess the uncertainty associated with $I(\mathcal{R}_r, \mathcal{W}_w)$ for each r, w in Equation (1), we must specify a model for both $N_1(\mathcal{R}_r, \mathcal{W}_w)$ (the number of cases in $\mathcal{R}_r \times \mathcal{W}_w$) and $N_0(\mathcal{R}_r, \mathcal{W}_w)$, the corresponding number of susceptible children (a child is classified as “susceptible” here if they are less than one year of age). Using the above notation,

$$N_1(\mathcal{R}_r, \mathcal{W}_w) = \sum_{j=1}^{N_1} \mathbb{1}_{\{t_j \in \mathcal{W}_w\}} \mathbb{1}_{\{s_j \in \mathcal{R}_r\}} \mathbb{1}_{\{b_j \in \mathcal{W}_w^\star\}} \text{ and}$$

$$N_0(\mathcal{R}_r, \mathcal{W}_w) = \sum_{i=1}^N \mathbb{1}_{\{s_i \in \mathcal{R}_r\}} \mathbb{1}_{\{b_i \in \mathcal{W}_w^\star\}} \text{ where } \mathbb{1}_{\mathcal{A}} \text{ is an indicator for the set } \mathcal{A} \text{ and } \mathcal{W}_w^\star$$

denotes the union of $\mathcal{W}_w$ and all time periods less than a year before $\min(\mathcal{W}_w)$. The challenge in this application, however, is that we only observe $s_i^\star \neq s_i$, $t_i^\star \neq t_i$ and $b_i^\star \neq b_i$. Thus, to specify statistical models for $N_1(\mathcal{R}_r, \mathcal{W}_w)$ and $N_0(\mathcal{R}_r, \mathcal{W}_w)$ we must hierarchically associate the jittered information to the actual information. As a final layer of model complexity, the total number of children N should also be modeled accordingly to completely quantify the uncertainty associated with the IP.

Based on the arguments above, our goal is to specify a model for $\{s_i^\star, s_i, t_i^\star, t_i, b_i^\star, b_i\}$ and N . As a general strategy, we factor the joint distribution of $(\{s_i^\star, s_i, t_i^\star, t_i, b_i^\star, b_i\}, N)$ as

$$\begin{aligned} [\{s_i^\star, s_i, t_i^\star, t_i, b_i^\star, b_i\}, N] &= [\{t_j^\star\} \mid \{t_j\}] \times [\{t_j\} \mid \{s_i, b_i\}, N] \times \dots \\ &\quad [\{s_i^\star\} \mid \{s_i\}, N] \times [\{b_i^\star\} \mid \{b_i\}, N] \times [\{s_i, b_i\}, N], \end{aligned} \quad (2)$$

where $[.]$ denotes a general distribution. We note that the factorization in (2) is conditional on a set of parameters which we suppressed from the notation for brevity. The following subsections describe each term in (2) in more detail and describe the associated parameters.

Intuitively, the factorization in (2) considers the data in the following sequence. First, N children are born at locations $\{s_i\}$ and dates $\{b_i\}$ which are then jittered according to $[s_i^\star \mid s_i]$ and $[b_i^\star \mid b_i]$. Second, given the number of susceptible children, N_1 of these children contract bronchiolitis at times distributed according to $[\{t_j\} \mid \{s_i, b_i\}, N]$ (i.e., the times which children contract bronchiolitis depends on their spatial location and birth date). And, third, the dates $\{t_j\}$ are subsequently jittered according to $[t_j^\star \mid t_j]$.

2.2. Point Pattern Model for Births

Consider, first, the point pattern of births given by the last term in Equation (2). We assume that the true locations $\{s_i, b_i\}$ are a realization of a Poisson point process with intensity $\Lambda(\mathbf{u}, v)$ for any $\mathbf{u} \in \mathcal{D}$ and $v \in \mathcal{T}$. Using this point process approach,

$$N \sim \mathcal{P}(\Lambda(\mathcal{D}, \mathcal{T})),$$

where $\mathcal{P}(\mu)$ denotes the Poisson distribution with mean μ and, for notational simplicity, $\Lambda(\mathcal{D}, \mathcal{T}) = \sum_{v \in \mathcal{T}} \int_{\mathcal{D}} \Lambda(\mathbf{u}, v) d\mathbf{u}$. Additionally, $(s_i, b_i) \sim \lambda(\mathbf{u}, v)$ independently where $\lambda(\mathbf{u}, v)$ is the normalized intensity (i.e., $\lambda(\mathbf{u}, v) = \Lambda(\mathbf{u}, v) / \Lambda(\mathcal{D}, \mathcal{T})$).

We note that the above point pattern model is in terms of births of children but this model induces a distribution for the number of susceptible children. That is, the implied distribution for $N_0(\mathcal{R}_r, \mathcal{W}_w)$ for any r, w is $N_0(\mathcal{R}_r, \mathcal{W}_w) \sim \mathcal{P}(\Lambda(\mathcal{R}_r, \mathcal{W}_w^\star))$ where, as above, $\mathcal{W}_w^\star$ denotes the union of $\mathcal{W}_w$ and all time periods less than a year before $\min(\mathcal{W})$. We point out that $I(\mathcal{R}_r, \mathcal{W}_w)$ in (1) is undefined with positive probability because $\text{Prob}(N_0(\mathcal{R}_r, \mathcal{W}_w) = 0) > 0$ (division by zero). Note that this is not an issue related to our modeling approach but rather an issue associated with the IP in general. That is, for any spatial region $\mathcal{R}$ where $N_0(\mathcal{R}, \mathcal{W}) = 0$ (i.e., there are no susceptible persons), the IP is undefined even when not treating IP as a random variable. To overcome this issue, we define an approximate expected incidence ratio (aEIR) as

$$\tilde{I}(\mathcal{R}_r, \mathcal{W}_w) = \frac{\mathbb{E}(N_1(\mathcal{R}_r, \mathcal{W}_w))}{\mathbb{E}(N_0(\mathcal{R}_r, \mathcal{W}_w))}, \quad (3)$$

where $\mathbb{E}(\cdot)$ denotes the expectation. In contrast to the IP, the aEIR in (3) is well defined for any measurable sets $\mathcal{R}_r$ and $\mathcal{W}_w$. The remainder of this article will focus on estimating aEIR rather than IP.

Conditional on the true birth locations s_i and birthdays b_i , the recorded locations and birthdays, as described in Section 1, are then uniformly jittered away from the true locations (days). That is, for the term $\{s_i^\star\} \mid \{s_i\}$ in (2) we assume $s_i^\star \mid s_i \sim \mathcal{U}(\mathcal{C}_\ell(s_i))$ independently where $\mathcal{U}(\cdot)$ denotes the uniform distribution and $\mathcal{C}_\ell(c)$ denotes a two-dimensional cube with center $c \in \mathbb{R}^2$ and side length ℓ such that $\ell/2$ is the spatial jittering length in each dimension. Note that, in this application and in most applications, the maximum jittering distance $\ell/2$ is known (but, again, not shared here for confidentiality reasons). Likewise, for the term $\{b_i^\star\} \mid \{b_i\}$ in (2), we assume $b_i^\star \mid b_i \sim \mathcal{J}_d(b_i)$ where $\mathcal{J}_d(t) = \{t-d, t-(d-1), \dots, t, t+1, \dots, t+d\}$ represents the temporal jittering set of length $d \in \mathbb{Z}^+$ days where, for this application, d is the known temporal jittering length. We note that the uniform distribution is appropriate in this analysis because, in discussion with the owners of the data, this was the actual distribution used to perform the jittering. However, other jittering distributions (e.g., Gaussian as in Chakraborty and Gelfand 2010 or Fanshawe and Diggle 2011) would be appropriate in other settings and such a choice of

jittering distribution would impact statistical inference. Hence, because the jittering distribution is known for this application, we do not explore the sensitivity of our results to different jittering distributions but leave this as an interesting path for future research.

Using the derivation of the likelihood discussed in the supplementary materials, the resulting log-likelihood for $\Lambda(\cdot)$ given the recorded births $\{s_i^\star\}$ and $\{b_i^\star\}$ is given by

$$-\Lambda(\mathcal{D}, \mathcal{T}) + N \log(\Lambda(\mathcal{D}, \mathcal{T})) + \sum_{i=1}^N \log \left(\sum_{b_i \in \mathcal{T}} \int_{\mathcal{D}} \lambda(s_i, b_i) \mathbb{1}_{\{s_i^\star \in \mathcal{C}_l(s_i)\}} \mathbb{1}_{\{b_i^\star \in \mathcal{J}_d(b_i)\}} ds_i \right) \quad (4)$$

where the indicator functions $\mathbb{1}_{\mathcal{A}}$ arise from the assumption that $s_i^\star \mid s_i \sim \mathcal{U}(\mathcal{C}_l(s_i))$ and $b_i^\star \mid b_i \sim \mathcal{U}(\mathcal{J}_d(b_i))$ as above. The form of the likelihood in (4) is troublesome to use for estimation due to the presence of the integrals which are not available in closed form. Taking a Bayesian perspective (as is the approach here) and specifying a prior for the intensity surface (e.g., a Gaussian process prior), the resulting integrals are random integrals with a complex dependency structure (see Heaton and Gelfand 2012). One common solution to this issue is to assume $\Lambda(\mathbf{u}, \mathbf{v})$ is proportional to closed-form density functions (see Kottas and Sansó 2007; Chakraborty and Gelfand 2010) so that the integrals can be easily computed but, with this approach, specifying/estimating a sufficiently flexible density can be cumbersome.

Due to the complex nature of the likelihood in (4), we need a parameterization of $\Lambda(\mathbf{u}, \mathbf{v})$ that results in computationally feasible methods for inference. In this article, similar to the approach of Liang, Carlin, and Gelfand (2009) who approximated the integrals using a discrete Riemann sum, we rely on the fact that our quantity of interest $\tilde{I}(\mathcal{R}_r, \mathcal{W}_w)$ is only defined for prespecified regions and temporal windows. Hence, we define a “regionally constant” model for $\Lambda(\mathbf{u}, \mathbf{v})$ such that

$$\Lambda(\mathbf{u}, \mathbf{v}) = \delta \sum_{w=1}^W \sum_{r=1}^R \frac{\lambda_{rw}}{|\mathcal{R}_r| |\mathcal{W}_w|} \mathbb{1}_{\{\mathbf{u} \in \mathcal{R}_r\}} \mathbb{1}_{\{\mathbf{v} \in \mathcal{W}_w\}}, \quad (5)$$

where $\lambda_{rw} \in [0, 1]$ with the constraint $\sum_{w=1}^W \sum_{r=1}^R \lambda_{rw} = 1$, $|\mathcal{R}_r|$ is the area of region r and $|\mathcal{W}_w|$ is the number of elements in the temporal window w . Intuitively, we assume that the intensity surface for births $\Lambda(\mathbf{u}, \mathbf{v})$ is constant over each spatial region $\mathcal{R}_r$ and temporal window $\mathcal{W}_w$ of interest. For this regionally constant specification, it can be easily shown (see the supplemental material) that $\Lambda(\mathcal{D}, \mathcal{T}) = \delta$ such that δ is the total number of births that are expected over the entire spatial domain and time period of this study. Likewise, using the arguments presented in the supplemental material, $\Lambda(\mathcal{R}_r, \mathcal{W}_w) = \delta \lambda_{rw}$ such that λ_{rw} can be interpreted as the fraction of expected births that occur within $\mathcal{R}_r \times \mathcal{W}_w$.

Beyond the intuitive interpretation of δ and λ_{rw} in (5), under this regionally constant specification the likelihood in (4) can be rewritten as

$$\begin{aligned}
& -\delta + N \log(\delta) \\
& + \sum_{i=1}^N \log \left(\sum_{w=1}^W \sum_{r=1}^R \lambda_{rw} \frac{|\mathcal{C}_l(s_j^*) \cap \mathcal{R}_r|}{|\mathcal{R}_r|} \frac{|\mathcal{J}_h(b_j^*) \cap \mathcal{W}_w|}{|\mathcal{W}_w|} \right), \tag{6}
\end{aligned}$$

where $|\mathcal{C}_l(s_j^*) \cap \mathcal{R}_r| / |\mathcal{R}_r|$ is the fraction of $\mathcal{R}_r$ that overlaps with the two-dimensional cube $\mathcal{C}_l(s_j^*)$ and $|\mathcal{J}_h(b_j^*) \cap \mathcal{W}_w| / |\mathcal{W}_w|$ is the fraction of the temporal window $\mathcal{W}_w$ also contained in the jittering set $\mathcal{J}_h(b_j^*)$ (see the supplemental material for a derivation). From (6), note that each of the integrals in (4) are replaced by summations. Finally, because we know how the data were jittered (uniformly with known spatial and temporal distance), the terms $|\mathcal{C}_l(s_j^*) \cap \mathcal{R}_r| / |\mathcal{R}_r|$ and $|\mathcal{J}_h(b_j^*) \cap \mathcal{W}_w| / |\mathcal{W}_w|$ in (6) need only be calculated once rather than iteratively in an MCMC or optimization algorithm. This fact, in conjunction with summations rather than integrals, renders the likelihood computationally feasible for Bayesian inference.

The specification in $\Lambda(\mathbf{u}, \mathbf{v})$ in (5) assumes that the intensity is constant within each $\mathcal{R}_r \times \mathcal{W}_w$. The validity of the assumption relies on the fact that $\mathcal{R}_1, \dots, \mathcal{R}_R$ and $\mathcal{W}_1, \dots, \mathcal{W}_W$ are of high enough resolution to capture the majority of spatio-temporal variability. We find this to be a sufficiently safe assumption given the fact that each $\mathcal{R}_r$ and $\mathcal{W}_w$ were defined according to the needs of the analysis and that finer resolutions are not of direct scientific interest. Note, however, that using (5) does not constrain the calculation of the aEIR to $\mathcal{R}_r \times \mathcal{W}_w$ only. Rather, (5) can be used to calculate $\mathbb{E}(N_0(\mathcal{R}, \mathcal{W}))$ for any $\mathcal{R} \subseteq \mathcal{D}$ and $\mathcal{W} \subseteq \mathcal{T}$.

2.3. Binomial Model for Bronchiolitis Cases

Having built a model for $[\{s_j, b_j\}, N]$ as well as $[\{s_i^* \mid s_i\}]$ and $[\{b_i^* \mid b_i\}]$ in the previous subsection, consider now the model for the times children contract bronchiolitis (the term $[\{t_j\} \mid \{s_j, b_j\}, N]$ in (2)). Again using the regional nature of the aEIR, we define a regionally constant probability model for the bronchiolitis dates. That is, conditional on the spatial location s_j and birthdate b_j , we assume that,

$$\mathbb{P}(\text{Prob}(t_j = t \mid s_j, b_j) = \sum_{r,w} \frac{\theta_{rw}}{|\mathcal{W}_w|} \mathbb{1}_{\{t \in \mathcal{W}_w\}} \mathbb{1}_{\{s_j \in \mathcal{R}_r\}} \mathbb{1}_{\{b_j \in \mathcal{W}_w^*\}}. \tag{7}$$

Intuitively, (7) states that the probability of contracting bronchiolitis changes across temporal windows $\mathcal{W}_w$ and spatial regions $\mathcal{R}_r$ but is constant within each $\mathcal{R}_r \times \mathcal{W}_w$. Notice that $\theta_{rw} = \mathbb{P}(\text{Prob}(t_j \in \mathcal{W}_w \mid s_j \in \mathcal{R}_r, b_j \in \mathcal{W}_w^*))$; that is, the probability that a bronchiolitis case occurs in $\mathcal{W}_w$ given that the child was born within a year of $\mathcal{W}_w(b_j \in \mathcal{W}_w^*)$ and resides in spatial region $\mathcal{R}_r$. Subsequently, we assume that the jittered bronchiolitis dates $t_j^* \sim \mathcal{U}(\mathcal{J}_d(t_j))$ where, as above, $\mathcal{J}_d(t_j) = \{t_j - d, \dots, t_j + d\}$ is the temporal jittering set with known jittering distance.

The model in (7) implicitly assumes independence between the children who eventually contract bronchiolitis. Because bronchiolitis is a contagious disease, this assumption may be violated resulting in a clustering of the locations and dates of observed bronchiolitis cases. To check this assumption a priori, we calculated L -functions (see Baddeley, Rubak, and Turner 2015, sec. 7.3) within each time window and found that within the majority of time windows, independence is supported (see the supplemental materials, e.g., L -functions). Additionally, as described below, assessment of model fit using posterior predictive checks further justifies that the assumption of independence is justified for this dataset.

Assuming independence among the children, the implied distribution of $N_1(\mathcal{R}_r, \mathcal{W}_w)$ (the number of bronchiolitis cases in $\mathcal{R}_r \times \mathcal{W}_w$) under (7) is,

$$N_1(\mathcal{R}_r, \mathcal{W}_w) \mid N_0(\mathcal{R}_r, \mathcal{W}_w) \sim \mathcal{B}(N_0(\mathcal{R}_r, \mathcal{W}_w), \theta_{rw}), \quad (8)$$

where $\mathcal{B}(n, \theta)$ denotes the binomial distribution with n trials and success probability θ and $N_0(\mathcal{R}_r, \mathcal{W}_w)$ is the number of susceptible children. Under the binomial formulation in (8), the aEIR in (3) simplifies to,

$$\tilde{I}(\mathcal{R}_r, \mathcal{W}_w) = \frac{\mathbb{E}(N_1(\mathcal{R}_r, \mathcal{W}_w))}{\mathbb{E}(N_0(\mathcal{R}_r, \mathcal{W}_w))} = \frac{\theta_{rw} \mathbb{E}(N_0(\mathcal{R}_r, \mathcal{W}_w))}{\mathbb{E}(N_0(\mathcal{R}_r, \mathcal{W}_w))} = \theta_{rw} \quad (9)$$

via properties of conditional expectation. Hence, for purposes of this research θ_{rw} for all $r = 1, \dots, R$ and $w = 1, \dots, W$ are the primary parameters of interest.

A primary goal of this study is, then, to associate θ_{rw} with pollution and socio-demographic covariates. Hence, hierarchically, we assume the spatio-temporal generalized linear mixed model

$$\log\left(\frac{\theta_{rw}}{1 - \theta_{rw}}\right) = \alpha + \mathbf{x}'_{rw}\boldsymbol{\beta} + \eta_w + \psi_r, \quad (10)$$

where α is an intercept, $\mathbf{x}'_{rw}$ is a vector of covariates describing the areal unit $\mathcal{R}_r \times \mathcal{W}_w$ with associated coefficients $\boldsymbol{\beta}$, η_w is a zero-mean, temporal random effect and ψ_r is a zero mean, spatial random effect. In terms of epidemiological research, the $\boldsymbol{\beta}$ coefficients are of scientific interest because these quantify the extent to which pollution and demographics impact the aEIR for bronchiolitis.

The spatio-temporal model in (10) makes a strong assumption of separability between the spatial (ψ_r) and temporal effects (η_w). In early phases of this work, more complex modeling structures were considered (i.e., nonseparable effects). However, the MHS data being analyzed here was found to be too sparse to consistently estimate a nonseparable space-time structure. Notably, according to the jittered locations, the median number of children who contracted bronchiolitis in a time window was 6 with a maximum of 33. Likewise, the 82.4% of the spatial regions used in this study did not see a single case of bronchiolitis. Hence, given the sparsity of the data, it was determined to use a separable space-time structure.

Let $\boldsymbol{\eta} = (\eta_1, \dots, \eta_W)$ and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_R)'$ denote the vector of temporal and spatial random effects across each of the W temporal windows and R spatial regions. Because a primary research question investigated here is the potential effect of pollution on the aEIR, setting W and R to be large is necessary so as to define the model at a sufficiently high resolution to capture the space-time variability of pollution. However, setting R to be large sacrifices computational tractability in spatial analyses (see Sun, Li, and Genton 2012). Hence, here we use a common approach in spatial statistics by using a basis function expansion for $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$ (see, e.g., Higdon 1998; Banerjee et al. 2008; Cressie and Johannesson 2008). Specifically, we let $\boldsymbol{\psi} = \mathbf{M}_\psi \boldsymbol{\psi}_0$ and $\boldsymbol{\eta} = \mathbf{M}_\eta \boldsymbol{\eta}_0$ where $\mathbf{M}_\psi$ and $\mathbf{M}_\eta$ are matrices of Moran basis functions following Hughes and Haran (2013) with unknown coefficients $\boldsymbol{\psi}_0$ and $\boldsymbol{\eta}_0$, respectively.

2.4. Hyperprior Distributions

To complete the model specification, consider now prior assumptions for the probabilities of birth at certain locations and dates given by $\boldsymbol{\lambda} = (\lambda_{11}, \dots, \lambda_{RW})'$, the expected number of children δ and the parameters in (10) associated with the space-time dynamics of the probability of bronchiolitis (specifically, α , β , $\boldsymbol{\eta}_0$, and $\boldsymbol{\psi}_0$). Because we used the Moran basis function expansions for $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$, Hughes and Haran (2013) show that the associated prior distribution for $\boldsymbol{\psi}_0$ (and, equivalently, $\boldsymbol{\eta}_0$) is proportional to

$$[\boldsymbol{\psi}^*] \propto \tau_\psi^q / 2 \exp \left\{ -\frac{\tau_\psi}{2} \boldsymbol{\psi}_0' \mathbf{M}_\psi' (\mathbf{W}_+ - \mathbf{W}) \mathbf{M}_\psi \boldsymbol{\psi}_0 \right\},$$

where τ_ψ is a prior precision, $\mathbf{W}$ is a $R \times R$ spatial adjacency matrix and $\mathbf{W}_+ = \text{diag}(\mathbf{W}\mathbf{1})$ with $\mathbf{1}$ denoting a column vector of ones.

As discussed in Section 2.2, $\boldsymbol{\lambda}$ represents the probabilities of being born in each spatial region and time window and is constrained such that $\sum_{t,w} \lambda_{tw} = 1$. For this reason, we assume that $\boldsymbol{\lambda}$ a priori follows a Dirichlet distribution with parameter vector $\mathbf{1}$ to portray our prior knowledge that a birth is equally likely across all time periods and spatial regions. The parameter $\delta > 0$ is the expected number of births that occur across all regions and time windows $\mathcal{D} \times \mathcal{T}$. Given the form of the likelihood in (6), we assume that the prior distribution for δ is the Jeffreys' prior for Poisson means and proportional to $\delta^{-1/2}$.

3. Accounting for Space-Time Uncertainty in the Covariates

The regression equation in (10) includes effects for various pollutants on the aEIR. Specifically, we are interested in the effect of fine particulate matter (PM_{2.5}) and sulfur dioxide (SO₂) on the aEIR. However, as described in Section 1, these pollutants are only observed at a few locations in our region of interest. To capture this relationship via Equation (10), we need pollution measurements for each spatial region $\mathcal{R}_r$ and temporal window $\mathcal{W}_w$ of interest. With such sparse measurements, a simple solution is to predict covariate values at unknown locations using a spatial kriging approach and use these as the measured covariate values for all regions of interest. However, these predictions will have inherent uncertainty which may impact the estimated coefficients if ignored. Hence, we

propose to account for uncertainty in the kriged estimates by integrating over the uncertainty in the point estimates.

We make use of a LUR model to spatially and temporally kriging the observed pollution measurements. Given the strong positive association of pollutants with urban areas, a LUR model is explicitly built for exploiting this relationship to perform prediction. Additionally, Mercer et al. (2011) illustrated the utility of accounting for spatial dependence in the LUR model. Therefore, to spatially predict pollution measurements at the unobserved locations (and times), we use a spatio-temporal LUR model, using the approach proposed by Szpiro et al. (2010) and Lindstrom et al. (2014).

For a given pollutant, let x_{rw} be the observed average pollutant measurement in spatial region $\mathcal{R}_r$ and across time window $\mathcal{W}_w$ and let n_x represent the number of locations with observed measurements of the pollutant. Following Szpiro et al. (2010), we write the model using only the observed locations for simplicity, but the joint model of both the unobserved and observed pollution measurements follows directly. In the LUR model, the pollution measurements are written as

$$x_{rw} = \xi_{rw} + v_{rw},$$

where ξ_{rw} is the spatio-temporal trend modeled with land-use covariates and v_{rw} is the spatio-temporal residual. The trend is broken into a linear combination of M basis functions, $f_1(w), \dots, f_M(w)$ such that

$$\xi_{rw} = \alpha_{0r} + \sum_{m=1}^M f_m(w) \alpha_{mr},$$

where $\mathbf{a}_m = (\alpha_{m1}, \dots, \alpha_{mR})'$ is the vector of basis coefficients for basis $f_m(\cdot)$. The basis functions $\{f_m(\cdot)\}$ are determined using singular value decomposition (SVD) on the (standardized) observations (an expectation-maximization algorithm is used for the SVD to accommodate missing values) and then smoothed using smoothing splines. For more details on this particular choice of basis functions, see Szpiro et al. (2010) and Fuentes, Guttorp, and Sampson (2006). The coefficient vectors $\mathbf{a}_m$ are modeled using a spatially dependent regression model,

$$\alpha_m \sim \mathcal{N}(\mathbf{B}_m \boldsymbol{\zeta}_m, \boldsymbol{\Sigma}(\boldsymbol{\phi}_m)),$$

where $\mathbf{B}_m$ is a matrix of land-use covariates, $\boldsymbol{\zeta}_m$ is a vector of coefficients, and $\boldsymbol{\Sigma}_a(\boldsymbol{\phi}_m)$ is a spatial covariance matrix dependent on parameters $\boldsymbol{\phi}_m$. Here, we use an exponential covariance structure for $\boldsymbol{\Sigma}(\boldsymbol{\phi}_m)$ with spatial decay $\boldsymbol{\phi}_m = \boldsymbol{\phi}_m$. Finally, the residual, v_{rw} , is assumed to be a zero-mean Gaussian process with spatial dependence $\boldsymbol{\Sigma}(\boldsymbol{\phi}_v)$, where again we use an exponential covariance structure. We fit the model using the R package SpatioTemporal (Lindstrom et al. 2013).

For pollution measurements, we used pollution monitors that were within 3 latitudes north or south and 1 longitude east or west of our region of interest. Figure 2 contains a map showing the location of the monitors for each pollutant along with the region of interest. Admittedly, this captures a very large area around our region of interest. That said, Figure 3 shows that the pollutant measurements across monitors (monitors located in the region of interest are red) follow similar trends. We examine a cross-validation study in Section 4.2 to determine the efficacy of the LUR model in this instance. Additionally, because we also account for the uncertainty of the predicted covariate values in the full bronchiolitis model (see Section 4.1), we expect that this uncertainty is reflected in the analysis results.

One common and valid concern about pollution measurements is the inherent bias that may be present due to the placement of pollution monitors (see, e.g., Gelfand, Sahu, and Holland 2012; Zidek, Shaddick, and Taylor 2014). Jerrett et al. (2005) showed that for $PM_{2.5}$, road networks tend to matter most, while for SO_2 , industrial areas tend to matter. To account for this, we use both land cover variables and distance to nearest major road, described more below, in the LUR model. While we expect the land cover variable to capture most of that impact on SO_2 measurements, we also examined the location of each of the 29 SO_2 monitors. By looking at maps of each of the monitor locations, we found that four were possibly industrial (outside of typical developed areas) and 25 were not. We expect that distance to road will capture the impact of monitor placement on $PM_{2.5}$ measurements.

To determine the land cover, we used the 2011 National Land Cover Database 30m resolution data product (Homer et al. 2015) and used the most common land cover within the region. Because of the small number of observed locations, we aggregated to four land use categories: high or medium developed (DevHigh); low or open space developed (DevLow); wetlands, forest, shrubs, or grasslands (Green), and cultivated or pastures (Crop). To determine the distance to a major road, we used the Natural Earth North America Roads Database version 2.0.0 10m resolution data product (Earth 2012) and calculated the distance to nearest major freeway, tollway or primary road (DistRoad). Figure 4 shows both of these variables for our study area. We used both the land cover and the distance to major road as covariates ($\mathbf{B}_m$) for each basis coefficient, α_m .

By using this model, we not only obtain spatially smoothed predictions for all $\mathcal{R}_r$ and $\mathcal{W}_w$ by meaningfully relating pollution levels to land cover but, importantly for this research, we obtain standard errors for our predictions which we use when fitting the model in Equation (10) to the MHS database. That is, as discussed more in Section 4.1, we treat the pollution covariates as “unknown” and iteratively sample the covariates within the Markov chain Monte Carlo (MCMC) algorithm.

4. Application to the MHS Database

4.1. MCMC Algorithm Overview

Bayesian model fitting for the model proposed in Sections 2 and 3 proceeds by MCMC sampling of the nonjittered spatio-temporal information $\{s_i\}$, $\{b_i\}$, and $\{t_i\}$, model parameters $\boldsymbol{\lambda} = (\lambda_{11}, \dots, \lambda_{RW})'$, δ , α , $\boldsymbol{\beta}$, $\boldsymbol{\eta}_0$, and $\boldsymbol{\psi}_0$ as well as the pollution covariate values

in $\mathbf{x}_{rw}$. While details for each complete conditional distribution is available in the supplementary materials, we briefly outline the MCMC algorithm here.

First, under the model formulation in Section 2.2, sufficient statistics for $\{s_i\}$, $\{b_i\}$, and $\{t_j\}$ are simply the index of the spatial region and temporal windows to which s_i , b_i , and t_j belong. Hence, to simplify computation and because s_i , b_i , and t_j are not of direct interest, we choose to sample the indices of spatial regions and temporal windows directly from their complete conditional distributions rather than the true spatio-temporal locations. Second, the form of the complete conditional distributions for δ and λ are gamma and Dirichlet, respectively, and we update each by drawing from these distributions. Third, due to a lack of a closed form complete conditional distribution, the parameters α , β , η_0 , and ψ_0 are updated jointly via adaptive Metropolis steps (Haario, Saksman, and Tamminen 2001). Finally, we drew the values of the pollution covariates by drawing from the associated predictive distribution of the LUR model described in Section 3.

We ran the algorithm for 2,000,000 iterations, using the first 1,000,000 iterations as burn-in. These values for burn and total iterations were determined by assessing chain mixing properties via trace plots, effective sample size and Monte Carlo standard error (Jones et al. 2006; Gong and Flegal 2016). Example trace plots and further discussion of convergence are included in the supplementary materials.

To assess model fit, we employed posterior predictive checks advocated by Leininger and Gelfand (2017). Namely, for each draw from the posterior distribution of model parameters, we simulated a realization of a point pattern and compared simulated pairwise distances to observed pairwise distances using $H(r) = (n(n-1))^{-1} \sum_{i \neq j} \mathbb{1}\{d_{ij} \leq r\}$ where r is a given distance and d_{ij} is the distance between case i and case j . Figures comparing the posterior predictive distribution of $H(r)$ versus the observed $H(r)$ are shown in the supplementary materials. To summarize, we found that pairwise distances as simulated from our model closely follow the observed pairwise distances suggesting a good model fit and justifying our assumption of independence between bronchiolitis cases.

4.2. Results When Accounting for Uncertainty

For the MHS data, we divided up the spatial region shown in Figure 1 into $R = 6514$ spatial regions $\mathcal{R}_r$ that corresponded to 200m grid cells. Furthermore, we partitioned the temporal domain of 2004–2013 into $W = 208$ two week intervals totaling 6514×208 total spatio-temporal domains on which we wish to estimate and quantify the uncertainty regarding the aEIR. In the aEIR model, we used 100 Moran basis functions for $\mathbf{M}_\psi$ and 75 Moran basis functions for $\mathbf{M}_\eta$ which represent approximately 25% and 50% of the total Moran bases associated with positive eigenvalues. This number of basis functions was chosen by considering increments of 25 basis functions up to a maximum of 200 (beyond which the computation became burdensome) and choosing the number that minimized the deviance information criterion (DIC; Spiegelhalter et al. 2002, 2014). Beyond the PM_{2.5} and SO₂ pollutant covariates, we also included an indicator for summer (defined as May–August), the interaction between summer and both pollutants as well as socio-demographic covariates for

percent with some college, percent single parent households, percent below poverty, percent African American, percent Hispanic, and population density.

For the LUR model, we used three SVD basis functions for the trend with the associated land use covariates given in Figure 4. We used cross-validation to select the number of bases, M , as well as to examine the model efficacy. Specifically, after removing data from two monitors in the Norfolk, VA study region, we fit the LUR model using $M = 2, \dots, 5$ bases for each pollutant. For $PM_{2.5}$, the number of bases used made little difference as all predicted quite well relative to the range in measurements (combined MSE for both locations was 6.462, 6.469, 6.430, and 6.494, for 2, 3, 4, and 5 bases, respectively). For SO_2 , the model with 2 bases fit rather poorly (MSE of 5.169), while the models with 3, 4, and 5 bases fit similarly (MSEs of 1.971, 3.770, and 3.213, respectively). For this reason, we chose to use the model with 3 bases. Importantly for this analysis, our 90% prediction intervals also captured approximately the true predicted value 90% of the time suggesting our measures of uncertainty, not just point predictions, are also adequate. The supplementary files contain figures showing the observed, predicted and 90% confidence bands for each pollutant using a model with the different number of bases as well as covariance parameter estimates.

The columns of Table 1 under the “full model” correspond to posterior summaries of α and β when accounting for uncertainty in the spatio-temporal locations of children (i.e., accounts for uncertainty in $\{s_i\}$, $\{b_i\}$, $\{t_i\}$, and $\mathbf{x}_{rw}$). At the 95% level, increased levels of SO_2 positively affect the aEIR in the winter while $PM_{2.5}$ has no effect on aEIR in the winter. Counter-intuitively, in the summer, $PM_{2.5}$ negatively affects the aEIR while SO_2 has no effect. Further investigation revealed that Norfolk experiences a spike in $PM_{2.5}$ in the summer (likely due to increased population from tourism) while bronchiolitis typically peaks in the winter months. Additional reasons for the counterintuitive sign could be due to multicollinearity (correlations between $PM_{2.5}$ and SO_2 measurements ranged from -0.62 to 0.78 with most correlations falling between 0 and 0.35, depending on the locations considered), or measurement error (measurements used in this analysis were averaged across all observed daily measurements in the two-week time window; therefore, if some windows contained only a few daily observations, these could be erroneously high or low). Hence, for Norfolk, this negative effect of $PM_{2.5}$ in the summer is partially understandable but should not be considered generalizable.

Interestingly, the only demographic variable that seems to impact the aEIR is population density. Respiratory viruses are considered density-dependent, in that their transmission increases with increased population density. However, the lack of significant socio-demographic indicators is of interest, as these are known to be strongly associated with RSV hospitalization in other studies as well as with influenza and other respiratory viruses (Sloan et al. 2015; Sangaré, Curtis, and Ahmad 2006). Further studies in other areas are needed to determine if this lack of association is unique to Norfolk or is consistent nationwide.

The left panel of Figure 5 displays the spatial random effects ψ while the right panel displays the posterior probability $\mathbb{P}(\text{Prob}(\psi_r > 0))$ for those locations where this probability exceeds 0.5 (indicating locations where the aEIR is elevated). Spatially, the largest random

effects occur directly over the Naval base (approximately -76.30° longitude and 36.94° latitude) but are generally widespread throughout Norfolk. This pattern was also seen in Figure 1 which showed the spatial distribution of bronchiolitis cases was more widespread than the density of births. Interestingly, the spatial patterns in ψ_r are not attributable to patterns in socio-demographics (recall the lack of significance of these variables in Table 1). As noted by, Reich, Hodges, and Zadnik (2006) and Hughes and Haran (2013), including spatial random effects may confound the main effects. However, we followed the approach of Hughes and Haran (2013) by using the Moran basis functions which constrains the spatial random effects to be orthogonal to the main effects. Hence, we suggest that further research is needed to investigate the driving factors for the spatial structure seen here.

Figure 6 displays the temporal random effects η . Intuitively, the temporal patterns of bronchiolitis aEIR is largely cyclical with peaks occurring in the winter. In general, the peaks in the later years (2010–2013) seem to be the highest observed peaks in the aEIR over the temporal domain (09/2004–04/2013). Not only do these years have the highest rates but they also have the longest durations and thus higher overall disease burdens. However, further research is needed to see if this increasing trend in time and duration is unique to Norfolk or commonly found across the United States.

As discussed briefly in Section 1, immunoprophylaxis is effective at preventing bronchiolitis in susceptible infants but requires monthly intramuscular injections to be effective. Hence, administering immunoprophylaxis prior to the onset of the bronchiolitis season, and continuing the injections throughout the season, is imperative. Using the posterior draws of η we calculate the “onset” of the bronchiolitis season to be the first window $\mathcal{W}_w$ in each year such that $\mathbb{P}(\eta_w > 0 \mid \text{data}) > 0.5$. Likewise, we calculate the end of the season as the first window where $\mathbb{P}(\eta_w > 0 \mid \text{data}) > 0.5$. Using this definition, the onset of the bronchiolitis season occurred as early as the first two weeks in November (which occurred in 2006, 2010, and 2012) or as late as the last few weeks in December (which occurred in 2005). Bronchiolitis seasons ended as late as the first two weeks in April (in 2011 and 2012) but most commonly ended by March 1st. Hence, administering immunoprophylaxis starting in October and continuing through February may be the best preventive measure.

4.3. Comparison to Results When Ignoring Uncertainty

To assess the impact of the spatio-temporal uncertainty in the location, birth date, and date diagnosed with bronchiolitis as well as the uncertainty in the pollution covariates, we compared the results from our full model to two alternative model specifications. First, we considered ignoring the uncertainty in the pollution covariates and treated the predicted values from the LUR model as the “true” values of the pollution across all $\mathcal{R}_r$ and $\mathcal{W}_w$ (we call this the “partial model”). And, second, we considered a model that ignored all sources of uncertainty and treated not only the pollution values as known but also made the assumption that the recorded birth locations, birth dates and dates diagnosed with bronchiolitis were, in fact, the true spatio-temporal locations of the children in this study.

Table 1 displays the posterior summaries for β under these two alternative model specifications. First, comparing β using the full model versus the model that ignores all

sources of uncertainty, a quick initial inspection of the corresponding credible intervals suggests that both models reach the same conclusions regarding the effect of each covariate (i.e., only SO₂ in the winter and population density are “significant”). However, a closer inspection reveals that credible intervals are wider for the full model than for the model without spatio-temporal uncertainty. The corresponding widths of the credible intervals, however, can lead to different scientific conclusions between the two models in the following manner. Consider, for example, the coefficient for percent of single parent households. The posterior probability $\mathbb{P}(\text{Prob}(\beta_{\text{Single}} > 0 \mid \text{data}) \approx 0.55$ for the full model which is nearly equivalent evidence for a positive effect as a negative effect. However, $\mathbb{P}(\text{Prob}(\beta_{\text{Single}} > 0 \mid \text{data}) \approx 0.81$ using the model that ignores uncertainty suggesting strong evidence (approximately 4 to 1 odds) for a positive effect for single parent households.

The middle columns of Table 1 labeled “partial model” displays 95% central credible intervals when we account for uncertainty in the spatio-temporal locations but ignore covariate uncertainty (i.e., treat the kriged pollution values as “true” pollution levels). The uncertainty expressed in the pollution effects for this partial model align with that expressed when ignoring both sources of uncertainty suggesting that ignoring uncertainty in covariates overstates the certainty of scientific conclusions. Interestingly, many of the 95% credible intervals widths for socio-demographic effects under the partial model are in between the corresponding widths for the full model and the no uncertainty model. This seemingly suggests that ignoring one source of uncertainty may have broader impacts on the analysis than the portion of the model directly associated with this uncertainty.

5. Conclusions

In this research, we sought to investigate the spatio-temporal patterns in the aEIR for bronchiolitis using a dataset obtained from the MHS. To preserve patient confidentiality, however, the spatial and temporal information for children in the MHS dataset was “jittered” in that the spatial locations in the dataset were randomized to within a known distance of the true location and the dates were randomized to within a known number of days of the true date. To account for this spatio-temporal uncertainty, we developed a modeling framework using spatio-temporal point patterns to estimate the aEIR for bronchiolitis and regress this aEIR onto covariates of interest. Furthermore, due to limited information on pollution levels over the spatial domain of interest, we also accounted for uncertainty in the pollution covariates values through the use of a LUR model that not only predicts pollutions levels to all locations and time periods of interest but also gives a measure of uncertainty in the predicted pollution value.

For this particular application, we showed that ignoring the spatio-temporal uncertainty versus modeling the uncertainty can overstate the certainty in scientific conclusions. Specifically, 95% credible intervals for the covariate effects were more narrow when ignoring the uncertainty in the covariates and spatio-temporal locations. While, for this particular application, conclusions about the “significance” of coefficients based on 95% credible intervals were the same when modeling versus ignoring sources of uncertainty, the narrower posterior distributions greatly increased the evidence of a significant covariate

effect. Hence, we conclude that ignoring uncertainty can potentially lead to “false positive” scientific conclusions.

Statistically, our primary contributions have been (i) a computationally feasible framework that can be used to account for various forms of spatio-temporal uncertainty and (ii) assessing the impact of this spatio-temporal uncertainty of scientific conclusions. First, we emphasize that, while we used specific assumptions about the nature of the spatio-temporal uncertainty that adheres to our MHS application, such assumptions could be easily altered to account for different forms of uncertainty in different datasets (e.g., a Gaussian versus uniform spatial jittering distribution or a continuous rather than discrete jittering distribution in time).

While our modeling strategy was effective in assessing factors that impact the aEIR, our modeling strategy is not without its limitations. For example, our application was for a small spatial region (Norfolk, VA) and extending this modeling strategy across a larger spatial region (say, the contiguous United States) would be infeasible for computational reasons (i.e., a 200 m grid across the contiguous United States would yield too many spatial regions to handle).

From an epidemiological perspective, our analysis of the MHS database reached interesting conclusions regarding the drivers of bronchiolitis epidemics while, in other respects, left several unanswered questions. For example, our results indicated that, aside from population density, none of the other socio-demographic variables strongly impact the aEIR of bronchiolitis. While this is an interesting result of itself, much additional epidemiological work is needed to investigate what variables, if any, contribute to increases or decreases in the aEIR of bronchiolitis in other locations. Additionally, our analysis in Norfolk found a positive association between the aEIR and SO_2 in the winter but a negative association between $\text{PM}_{2.5}$ and the aEIR in the summer. This negative association goes against intuition, and, as described may not be generalizable. We found elevated levels of $\text{PM}_{2.5}$ during the summer in Norfolk which led to a negative association with the aEIR which peaked in the winter.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank Pingsheng Wu, Tebeb Gebretsadik, Nicholas Sicignano, Rees Lee, and Tina Hartert for their valuable insight into the project.

Funding

Research reported in this publication was supported by the National Institute of Environmental Health Science of the National Institutes of Health under award number R03ES025295. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Navy, Department of Defense of the U.S. Government. Procedures for analysis and visualization of the data were reviewed and approved by the IRB at Brigham Young University.

References

- Baddeley A, Rubak E, and Turner R (2015), *Spatial Point Patterns: Methodology and Applications With R*, Boca Raton, FL: CRC Press. [2,5]
- Baker J, White N, and Mengersen K (2014), “Missing in Space: An Evaluation of Imputation Methods for Missing Data in Spatial Analysis of Risk Factors for Type II Diabetes,” *International Journal of Health Geographics*, 13, 47. [3] [PubMed: 25410053]
- Banerjee S, Carlin BP, and Gelfand AE (2014), *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed.), Boca Raton, FL: CRC Press. [2]
- Banerjee S, Gelfand AE, Finley AO, and Sang H (2008), “Gaussian Predictive Process Models for Large Spatial Datasets,” *Journal of the Royal Statistical Society, Series B*, 70, 825–848. [5]
- Barber JJ, Gelfand AE, and Silander JA (2006), “Modelling Map Positional Error to Infer True Feature Location,” *The Canadian Journal of Statistics*, 34, 659–676. [2]
- Chakraborty A, and Gelfand AE (2010), “Analyzing Spatial Point Patterns Subject to Measurement Error,” *Bayesian Analysis*, 5, 97–122. [2,4]
- Cressie N, and Johannesson G (2008), “Fixed Rank Kriging for Very Large Spatial Data Sets,” *Journal of the Royal Statistical Society, Series B*, 70, 209–226. [5]
- Cressie N, and Kornack J (2003), “Spatial Statistics in the Presence of Location Error With an Application to Remote Sensing of the Environment,” *Statistical Science*, 18, 436–456. [2]
- Earth N (2012), “Natural Earth North America Roads Database Version 2.0.0,” available at <http://www.naturalearthdata.com/downloads/10m-cultural-vectors/roads/>. [6]
- Fanshawe TR, and Diggle PJ (2011), “Spatial Prediction in the Presence of Positional Error,” *Environmetrics*, 22, 109–122. [2,4]
- Fuentes M, Guttorp P, and Sampson PD (2006), “Using Transforms to Analyze Space-Time Processes,” in *Statistical Methods for Spatio-Temporal Systems*, eds. Finkenstadt B, Held L, and Isham V, Boca Raton FL: CRC/Chapman and Hall, pp. 77–150. [6]
- Gabrosek J, and Cressie N (2002), “The Effect on Attribute Prediction of Location Uncertainty in Spatial Data,” *Geographical Analysis*, 34, 262–285. [2]
- Gelfand AE, Sahu SK, and Holland DM (2012), “On the Effect of Preferential Sampling in Spatial Prediction,” *Environmetrics*, 23, 565–578. [6] [PubMed: 24077640]
- Gelfand AE, Zhu L, and Carlin BP (2001), “On the Change of Support Problem for Spatio-Temporal Data,” *Biostatistics*, 2, 31–45. [2] [PubMed: 12933555]
- Gong L, and Flegal JM (2016), “A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo,” *Journal of Computational and Graphical Statistics*, 25, 684–700. [8]
- Gotway CA, and Young LJ (2002), “Combining Incompatible Spatial Data,” *Journal of the American Statistical Association*, 97, 632–648. [2]
- Haario H, Saksman E, and Tamminen J (2001), “An Adaptive Metropolis Algorithm,” *Bernoulli*, 2, 223–242. [8]
- Hampp C, Kauf TL, Saidi AS, and Winterstein AG (2011), “Cost-Effectiveness of Respiratory Syncytial Virus Prophylaxis in Various Indications,” *Archives of Pediatrics & Adolescent Medicine*, 165, 498–505. [1] [PubMed: 21300647]
- Heaton MJ, and Gelfand AE (2012), “Kernel Averaged Predictors for Spatio-Temporal Regression Models,” *Spatial Statistics*, 2, 15–32. [4] [PubMed: 24010051]
- Heymann PW, Carper HT, Murphy DD, Platts-Mills TA, Patrie J, McLaughlin AP, Erwin EA, Shaker MS, Hellems M, Peerzada J, and Hayden FG (2004), “Viral Infections in Relation to Age, Atopy, and Season of Admission Among Children Hospitalized for Wheezing,” *Journal of Allergy and Clinical Immunology*, 114, 239–247. [1] [PubMed: 15316497]
- Higdon D (1998), “A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean,” *Environmental and Ecological Statistics*, 5, 173–190. [5]
- Homer CG, Dewitz JA, Yang L, Jin S, Danielson P, Xian G, Coulston J, Herold ND, Wickham JD, and Megown K (2015), “Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change Information,” *Photogrammetric Engineering and Remote Sensing*, 81, 345–354. [6]

- Hughes J, and Haran M (2013), “Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models,” *Journal of the Royal Statistical Society, Series B*, 75, 139–159. [5,9]
- Ibrahim JG, Chen M, Lipsitz SR, and Herring AH (2005), “Missing-Data Methods for Generalized Linear Models: A Comparative Review,” *Journal of the American Statistical Association*, 100, 332–346, DOI: 10.1198/016214504000001844. [2]
- Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, and Giovis C (2005), “A Review and Evaluation of Intraurban Air Pollution Exposure Models,” *Journal of Exposure Science and Environmental Epidemiology*, 15, 185. [6]
- Jones GL, Haran M, Caffo BS, and Neath R (2006), “Fixed-Width Output Analysis for Markov Chain Monte Carlo,” *Journal of the American Statistical Association*, 101, 1537–1547. [8]
- Kottas A, and Sansó B (2007), “Bayesian Mixture Modeling for Spatial Poisson Process Intensities, With Applications to Extreme Value Analysis,” *Journal of Statistical Planning and Inference*, 137, 3151–3163. [4]
- Leininger TJ, and Gelfand AE (2017), “Bayesian Inference and Model Assessment for Spatial Point Patterns Using Posterior Predictive Samples,” *Bayesian Analysis*, 12, 1–30. [8]
- Liang S, Carlin BP, and Gelfand AE (2009), “Analysis of Minnesota Colon and Rectum Cancer Point Patterns With Spatial and Nonspatial Covariate Information,” *Annals of Applied Statistics*, 3, 943–962. [4]
- Lindstrom J, Szpiro A, Sampson PD, Bergen S, and Oron AP (2013), *SpatioTemporal: Spatio-Temporal Model Estimation*, R Package Version 1.1.7. [6]
- Lindstrom J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson TV, and Sheppard L (2014), “A Flexible Spatio-Temporal Model for Air Pollution With Spatial and Spatio-Temporal Covariates,” *Environmental and Ecological Statistics*, 21, 411–433, DOI: 10.1007/s10651-013-0261-4. [6] [PubMed: 25264424]
- Mehta S, Shin H, Burnett R, North T, and Cohen AJ (2013), “Ambient Particulate Air Pollution and Acute Lower Respiratory Infections: A Systematic Review and Implications for Estimating the Global Burden of Disease,” *Air Quality, Atmosphere & Health*, 6, 69–83. [1]
- Mercer LD, Szpiro AA, Sheppard L, Lindström J, Adar SD, Allen RW, Avol EL, Oron AP, Larson T, Liu L-JS, and Kaufman JD (2011), “Comparing Universal Kriging and Land-Use Regression for Predicting Concentrations of Gaseous Oxides of Nitrogen (NO_x) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air),” *Atmospheric Environment*, 45, 4412–4420, DOI: 10.1016/j.atmosenv.2011.05.043. [6] [PubMed: 21808599]
- Miller EK, Gebretsadik T, Carroll KN, Dupont WD, Mohamed YA, Morin L-L, Heil L, Minton PA, Woodward K, Liu Z, and Hartert TV (2013), “Viral Etiologies of Infant Bronchiolitis, Croup, and Upper Respiratory Illness During Four Consecutive Years,” *The Pediatric Infectious Disease Journal*, 32, 950–955. [1] [PubMed: 23694832]
- Nair H, Nokes DJ, Gessner BD, Dherani M, Madhi SA, Singleton RJ, O’Brien KL, Roca A, Wright PF, Bruce N, and Chandran A (2010), “Global Burden of Acute Lower Respiratory Infections due to Respiratory Syncytial Virus in Young Children: A Systematic Review and Meta-Analysis,” *The Lancet*, 375, 1545–1555. [1]
- Panzer D, Bendetti R, and Postiglione P (2016), “A Bayesian Approach to Parameter Estimation in the Presence of Spatial Missing Data,” *Spatial Economic Analysis*, 11, 201–218. [3]
- Rathbun SL (2013), “Optimal Estimation of Poisson Intensity With Partially Observed Covariates,” *Biometrika*, 100, 277–281. [3]
- Reich BJ, Hodges JS, and Zadnik V (2006), “Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models,” *Biometrics*, 62, 1197–1206. [9] [PubMed: 17156295]
- Sangaré L, Curtis MP, and Ahmad S (2006), “Hospitalization for Respiratory Syncytial Virus Among California Infants: Disparities Related to Race, Insurance, and Geography,” *The Journal of Pediatrics*, 149, 373–377. [9] [PubMed: 16939750]
- Shay DK, Holman RC, Roosevelt GE, Clarke MJ, and Anderson J (2001), “Bronchiolitis-Associated Mortality and Estimates of Respiratory Syncytial Virus? Associated Deaths Among US Children, 1979–1997,” *Journal of Infectious Diseases*, 183, 16–22. [1] [PubMed: 11076709]

- Siefker D, Williams RW, Lu L, You D, and Cormier S (2016), “The Influence of Genetic Diversity on Neonatal Respiratory Syncytial Virus Disease,” *The Journal of Immunology*, 196, 80–89. [1] [PubMed: 26608911]
- Sirimi N, Miligkos M, Koutouzi F, Petridou E, Siahianidou T, and Michos A (2015), “Respiratory Syncytial Virus Activity and Climate Parameters During a 12-Year Period,” *Journal of Medical Virology*, 88, 931–937. [1] [PubMed: 26575521]
- Sloan C, Chandrasekhar R, Mitchel E, Schaffner W, and Lindegren ML (2015), “Socioeconomic Disparities and Influenza Hospitalizations, Tennessee, USA,” *Emerging Infectious Diseases*, 21, 1602. [9] [PubMed: 26292106]
- Spiegelhalter DJ, Best NG, Carlin BP, and van der Linde A (2014), “The Deviance Information Criterion: 12 Years on,” *Journal of the Royal Statistical Society, Series B*, 76, 485–493. [8]
- (2002), “Bayesian Measures of Model Complexity and Fit” (with discussion), *Journal of the Royal Statistical Society, Series B*, 64, 583–639. [8]
- Sun Y, Li B, and Genton MG (2012), “Geostatistics for Large Datasets,” in *Advances and Challenges in Space-Time Modelling of Natural Events*, Berlin, Heidelberg: Springer, pp. 55–77. [3,5]
- Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, and Kaufman JD (2010), “Predicting Intra-urban Variation in Air Pollution Concentrations With Complex Spatio-Temporal Dependencies,” *Environmetrics*, 21, 606–631, DOI: 10.1002/env.1014. [6]
- Waagepetersen R (2008), “Estimating Functions for Inhomogeneous Spatial Point Processes With Incomplete Covariate Data,” *Biometrika*, 95, 351–363. [3]
- Walsh EE, McConnochie KM, Long CE, and Hall CB (1997), “Severity of Respiratory Syncytial Virus Infection Is Related to Virus Strain,” *Journal of Infectious Diseases*, 175, 814–820. [1] [PubMed: 9086135]
- Wikle CK, and Berliner LM (2005), “Combining Information Across Spatial Scales,” *Technometrics*, 47, 80–91. [2]
- Ye Q, Fu J-F, Mao J-H, and Shang S-Q (2016), “Haze Is a Risk Factor Contributing to the Rapid Spread of Respiratory Syncytial Virus in Children,” *Environmental Science and Pollution Research*, 23, 20178–20185. [1] [PubMed: 27439752]
- Zhou H, Thompson WW, Viboud CG, Ringholz CM, Cheng P-Y, Steiner C, Abedi GR, Anderson LJ, Brammer L, and Shay DK (2012), “Hospitalizations Associated With Influenza and Respiratory Syncytial Virus in the United States, 1993–2008,” *Clinical Infectious Diseases*, 54, 1427–1436. [1] [PubMed: 22495079]
- Zidek JV, Shaddick G, and Taylor CG (2014), “Reducing Estimation Bias in Adaptively Changing Monitoring Networks With Preferential Site Selection,” *The Annals of Applied Statistics*, 8, 1640–1670. [6]
- Zidek JV, Sun W, and Le ND (2000), “Designing and Integrating Composite Networks for Monitoring Multivariate Gaussian Pollution Fields,” *Journal of the Royal Statistical Society, Series C*, 49, 63–79. [3]
- Zimmerman DL (2008), “Estimating the Intensity of a Spatial Point Process From Locations Coarsened by Incomplete Geocoding,” *Biometrics*, 64, 262–270. [2] [PubMed: 17680833]

Figure 1. Kernel density plots of recorded (a) birth location of each child and (b) birth location of children who contracted bronchiolitis in Norfolk, Virginia.

Figure 2.
Locations of the observations used in the LUR model for (a) PM_{2.5} and (b) SO₂.

Figure 3. PM_{2.5} (top) and SO₂ (bottom) measurements from stations used in the analysis. The red lines are the measurements from monitors within the Norfolk study region.

Figure 4.
Images of the two land use variables, land cover (left) and distance to major road (right),
used in the LUR model to predict pollution within our study region.

Figure 5. Posterior means of (a) the spatial random effects Ψ and (b) $\mathbb{P}(\Psi_r > 0 \mid \text{data})$ for locations where $\mathbb{P}(\Psi_r > 0 \mid \text{data}) > 0.5$. In both panels, locations that are predominantly water are omitted.

Figure 6.
Posterior mean of the temporal random effects η .

Table 1.

Posterior medians and 95% credible intervals for α and β according to a model that fully accounts for all sources of uncertainty (full model), a model that ignores uncertainty in the pollution covariates but accounts for uncertainty in the recorded dates and locations (partial model), and a model that ignores covariate and spatio-temporal location uncertainty (no uncertainty).

	Full model			Partial model			No uncertainty		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
Intercept	-5.31	-5.07	-4.88	-5.33	-5.05	-4.90	-5.53	-5.16	-5.00
SO ₂	0.02	0.04	0.06	0.03	0.04	0.05	0.03	0.04	0.05
PM _{2.5}	-0.01	0.00	0.01	-0.04	-0.03	-0.02	-0.04	-0.03	-0.02
Summer	-0.01	0.00	0.02	-0.01	0.00	0.02	-0.01	0.00	0.01
Summer*SO ₂	-0.03	0.00	0.03	-0.02	0.00	0.02	-0.02	0.00	0.02
Summer*PM _{2.5}	-0.03	-0.02	-0.01	-0.01	-0.02	-0.01	-0.02	-0.01	-0.01
Education	-0.12	0.12	0.29	-0.12	0.05	0.30	-0.09	0.14	0.24
Single	-0.21	0.04	0.24	-0.08	0.05	0.19	-0.04	0.07	0.16
Poverty	-0.44	-0.07	0.02	-0.18	-0.05	0.02	-0.06	-0.02	0.02
%Af Am	-0.11	0.05	0.39	-0.06	0.14	0.40	-0.07	0.01	0.12
%Hispanic	-0.74	-0.21	0.11	-0.18	-0.07	0.10	-0.20	-0.05	0.10
PopDens	0.02	0.12	0.24	0.01	0.06	0.10	0.02	0.05	0.09

NOTE: Quantiles are subject to rounding error.