Open Forum Infectious Diseases

MAJOR ARTICLE

A Bayesian Spatiotemporal Analysis of Pediatric Group A Streptococcal Infections

Angela Wang,[1] Andrew M. Fine,[2] Erin Buchanan,[3] Mark Janko,[1] Lise E. Nigrovic,[2] and Paul M. Lantos[1]

[1]Duke University, Durham, North Carolina, USA, [2]Boston Children's Hospital, Boston, Massachusetts, USA, and [3]Harrisburg University, Harrisburg, Pennsylvania, USA

Background. Pharyngitis due to group A *Streptococcus* (GAS) is a common pediatric infection. Physicians might diagnose GAS pharyngitis more accurately when given biosurveillance information about GAS activity. The availability of geographic GAS testing data may be able to assist with real-time clinical decision-making for children with throat infections.

Methods. GAS rapid antigen testing data were obtained from the records of 6086 children at Boston Children's Hospital and 8648 children at Duke University Medical Center. Records included children tested in outpatient, primary care settings. We constructed Bayesian generalized additive models, in which the outcome variable was the binary result of GAS testing, and predictor variables included smoothed functions of patient location data and both cyclic and longitudinal time data.

Results. We observed a small degree of geographic heterogeneity, but no convincing clusters of high risk. The probability of a positive test declined during the summer months.

Conclusions. Future work should include geographic data about school catchments to identify whether GAS transmission clusters within schools.

Keywords. Bayesian statistics; epidemiology; geographic information systems; modeling; pediatrics; pharyngitis; *Streptococcus*.

Group A *Streptococcus* (GAS) causes ~600 million annual cases of pharyngitis worldwide [1]. Despite the availability of point-of-care diagnostic testing and clinical risk criteria [2, 3], GAS pharyngitis remains frequently overdiagnosed in both children and adults, leading to unnecessary antibiotic exposure [3–5]. Providing physicians with accurate, up-to-date GAS biosurveillance may improve diagnostic accuracy [6]. GAS pharyngitis cases are spatially and temporally heterogeneous, sometimes occurring sporadically and sometimes in clusters or outbreaks [7–9]. Thus, effective surveillance methods are needed to identify spatial and temporal signals of increased GAS activity in order to inform clinical practice.

We used novel statistical models to identify the spatial and temporal dynamics of GAS diagnoses. We used GAS testing data from the electronic health records of Boston Children's Hospital (Boston, MA, USA) and the Duke University Health System (Durham, NC, USA) with 7 years of pediatric GAS testing data. Our models incorporated individual patient variables and patient location data and evaluated geographic space, time as a longitudinal variable, and time as a cyclic variable to predict the probability that a GAS test will be positive.

METHODS

This study protocol was approved by the Institutional Review Boards of both Boston Children's Hospital and the Duke University Health System with permission for data sharing. Informed consent was waived for this retrospective study.

Study Design

We performed a retrospective study using electronic data from 2 health systems: Boston Children's Hospital (Boston, MA, USA) and Duke University Hospital (Durham, NC, USA).

Study Population

We queried electronic medical records to identify all children who had had a rapid antigen test for GAS between January 1, 2011, and December 31, 2017. The electronic medical records included children seen in primary care settings within each health system, as well as those seen at the hospital for emergency or inpatient care. Tests were included for children who were 5 to 15 years old (inclusive) at the time of testing. As many children were tested more than once during the 7 years of study, we included at most 1 test per 6 months in order to minimize the chance of including multiple tests from the same clinical illness. We excluded children with a primary home address >12 km from the respective hospital. We chose this distance to fit our spatial models to observations within 12 km but predict

Received 13 September 2019; editorial decision 4 December 2019; accepted 9 December 2019.
Correspondence: P. M. Lantos, MD (paul.lantos@duke.edu).

Open Forum Infectious Diseases®

DOI: 10.1093/ofid/ofz524

them within 10 km, thus avoiding predictions to edge areas with sparse data (Figure 1).

Data Collection

We abstracted the results of rapid GAS antigen testing, which were dichotomized as positive or negative. Additionally, we collected each subject's age at the time of testing, gender, race, ethnicity, and residential longitude and latitude.

The individual categories of race and ethnicity differed between Boston and Durham. Thus, for simplicity, racial categories were consolidated to "black," "white," and "other or unavailable." The latter category included several self-reported racial categories that were represented in small numbers, categories such as "other," "multiracial," and "2 or more races," as well as individuals who declined to provide a race. Exploratory modeling did not show any statistical disadvantage to this consolidated categorization. Reference values were set for categorical variables as follows: "female" for gender, "unavailable or other" for race, and "unavailable" for ethnicity. Age was centered on 0 by subtracting the mean and standardized by dividing by 2 standard deviations [10]. For temporal modeling, we determined the week of the year (from 1 to 53) and the cumulative week (from 1 to 371) for the date on which a test was performed.

Statistical Analyses

Our primary model was a logistic generalized additive model (GAM). GAMs are regression models that use nonparametric functions to model nonlinear relationships between independent variables and an outcome variable of interest [11]. We used the statistical programming language R (www.r-project.org) and the brms and mgcv packages [11–13]. Mgcv is a comprehensive package for the specification of GAMs. Brms, through its dependency on mgcv, allows the construction of Bayesian GAMs that are then sent to the program Stan (www.mc-stan.org) for sampling of the posterior probability distribution.

The response variable in our models was the binary result of streptococcal testing (negative vs positive), and our fixed linear predictors were age, race, and ethnicity. We used three spline functions to incorporate space and time: (1) an isotropic 2-dimensional spline of longitude and latitude to model geographic heterogeneity; (2) a spline to model temporal variability over the length of our study period; and (3) a cyclic spline to model seasonal variability observed cyclically over the years of study. Approaches to modeling time, space, and seasonality using splines are supported in mgcv and brms [11–14]. We chose normally distributed priors with mean 0 and standard deviation

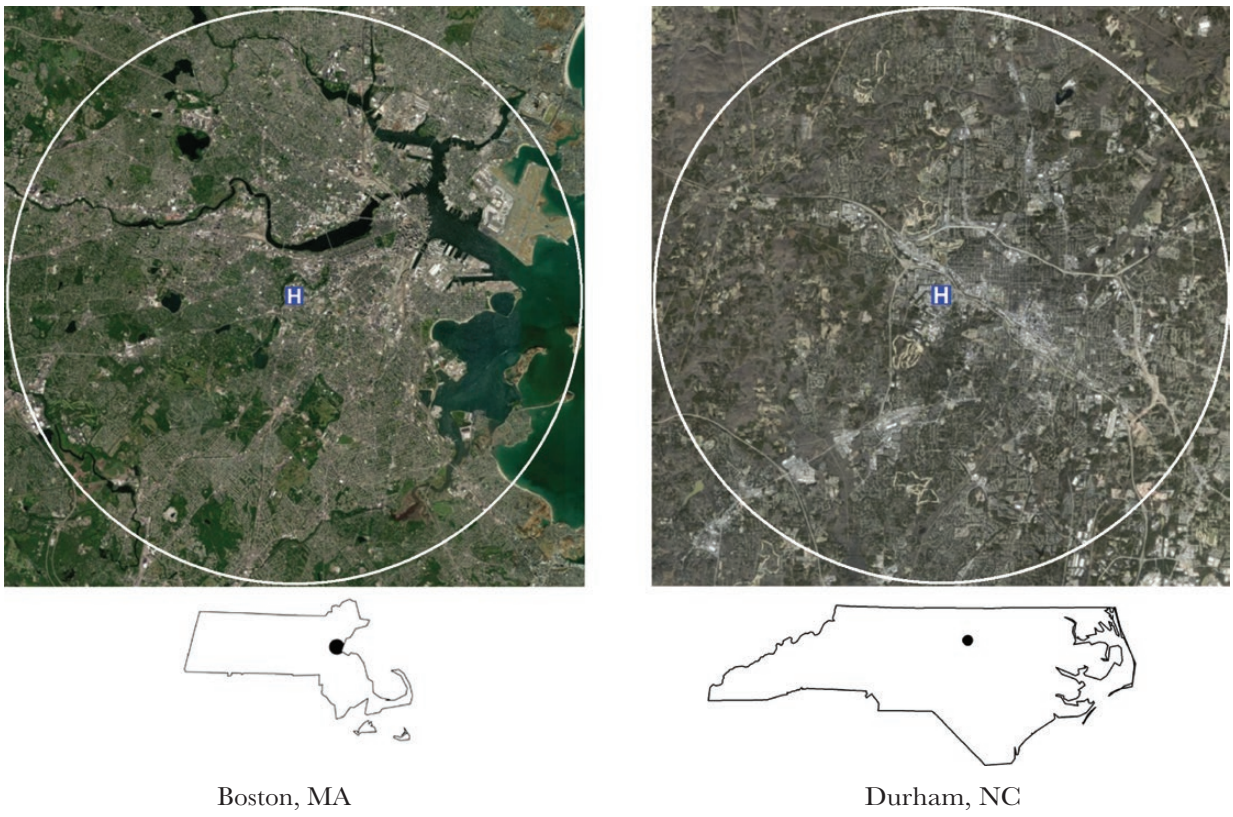

Figure 1. Study locations. These aerial images illustrate 10-km radius circles around Boston Children's Hospital (Boston, MA, USA) and Duke University Hospital (Durham, NC, USA). Children whose testing data were used to populate our models had home addresses within a 12-km radius of their respective hospital. After fitting our models, we predicted the probability of group A *Streptococcus* pharyngitis in the 10-km radius circles illustrated here. Aerial imagery was provided by ESRI through its ArcGIS basemaps (ESRI, Redlands, CA, USA).

1 for the odds ratio (OR) of fixed effects and for the log odds of the models' intercepts. Default priors were accepted for smoothed terms, which were a minimally informative Student *t* distribution.

We then constructed grids onto which we could predict our models. The grids were composed of dense longitude–latitude coordinate pairs covering a 10-km radius circle around each hospital. Using loops, we predicted the probability of a positive GAS test for each of 371 consecutive weeks and for the week of the year. We used contours to circumscribe areas where there was a 90%, 95%, or 99% probability that the local odds differed from the average odds. Each prediction was saved as an image, after which the images were joined sequentially to create an animation. Results are expressed in probability and in 95% uncertainty intervals, which represent the values bounding the 95% uncertainty interval.

RESULTS

We fit our Boston model using data from 7169 GAS tests in 6086 children, of which 1567 (21.9%) were positive. Our Durham model included 13 129 tests in 8648 children, 2421 of which were positive (18.4%). The demographic characteristics of our subjects can be found in Table 1.

The impact of individual covariates in our models is presented in Figure 2. Neither race nor ethnicity was associated with the probability of a positive GAS test in either site. For both sites, younger age was the most important individual predictor of a positive test; an increase in age of 6 years was associated with an OR of 0.49 in Durham (95% uncertainty interval [UI], 0.44–0.54) and 0.68 in Boston (95% UI, 0.60–0.77). The odds of a positive GAS test were higher for males in Durham (OR, 1.13; 95% UI, 1.03–1.23) and trended similarly in Boston (OR, 1.08; 95% UI, 0.96–1.21).

Our spatiotemporal models (Figure 3, Videos 1 and 2) demonstrate a pronounced cyclical trend in the probability of a positive test. The probability drops markedly each year during the summer months in both Boston and Durham. By contrast, the probability varied far less significantly throughout the remainder of the year. The overall longitudinal trend over the years of study was fairly constant in both sites, with probabilities in the peak and nadir periods differing by only about 5%. In both sites, areas of higher probability appeared to migrate slowly across the study area over time. We did not, however, confidently resolve spatial clusters where the probability of a positive test clearly differed from the surroundings.

Table 1. Demographic Characteristics of the Study Population

	Boston	Durham
Age, y	8.5 (6.5–11.4)	9.2 (7.0–12.1)
Gender		
Female	3745 (52.2)	6979 (53.2)
Male	3424 (47.8)	6150 (46.8)
Race		
Black	2220 (31.0)	5047 (38.4)
White	1035 (14.4)	5282 (40.2)
Other or unavailable	3914 (54.6)	2800 (21.3)
Ethnicity		
Hispanic	3781 (52.7)	1704 (13.0)
Non-Hispanic	3102 (43.2)	10 735 (81.8)
Unavailable	286 (4.0)	690 (5.3)

Age is presented as median with interquartile range. Gender, race, and ethnicity are presented as number and percentage.

DISCUSSION

We have described the spatiotemporal dynamics of positive GAS tests over 7 study years using clinical data from 2 metropolitan areas. In both study sites, the probability of a positive GAS test migrated spatially over time. This spatial heterogeneity may reflect local outbreaks of GAS pharyngitis in the high-probability areas, whereas low-probability areas may represent outbreaks of viruses or other pathogens in which many children are tested for pharyngitis but GAS is found less commonly. Overall, however, the geographic heterogeneity of each study site was of low amplitude and uncertain significance. Generally there was a <10% difference in the probability of a positive GAS test between the local maxima and minima at any given time. We did not convincingly resolve spatial clusters in either site where there was a persistently high probability of a positive test.

The dominant temporal trend we observed was the low probability of a positive test during summer months, which was observed annually throughout the study period. This phenomenon was of greater amplitude than any longitudinal trend or spatial pattern. The most likely explanation is that transmission of GAS occurs more widely during months when schools are in session. In addition to the above spatiotemporal observations, we also found that younger children had the highest probability of a positive test. These findings suggest that the most important epidemiologic signals of GAS activity will be found in elementary schools.

The temporal pattern we have identified would certainly have been demonstrable even without the inclusion of spatial data in our models. However, this would leave untested the question of whether a seasonal trend is global or local. By including geographic coordinates in our models, we have demonstrated that the seasonality of GAS is a global phenomenon and not (clearly) due to local recurrences. Furthermore, adding geographic space to our models can also be seen as a covariate adjustment like our other independent variables: The seasonal variability in GAS was prominent even after adjusting for geography.

There is relatively little published literature about the spatial epidemiology of GAS. In Kenya, the incidence of rheumatic heart disease, an important complication of GAS pharyngitis, was found to be spatially heterogeneous [15]. In recent years, China has experienced an increasing incidence of scarlet fever

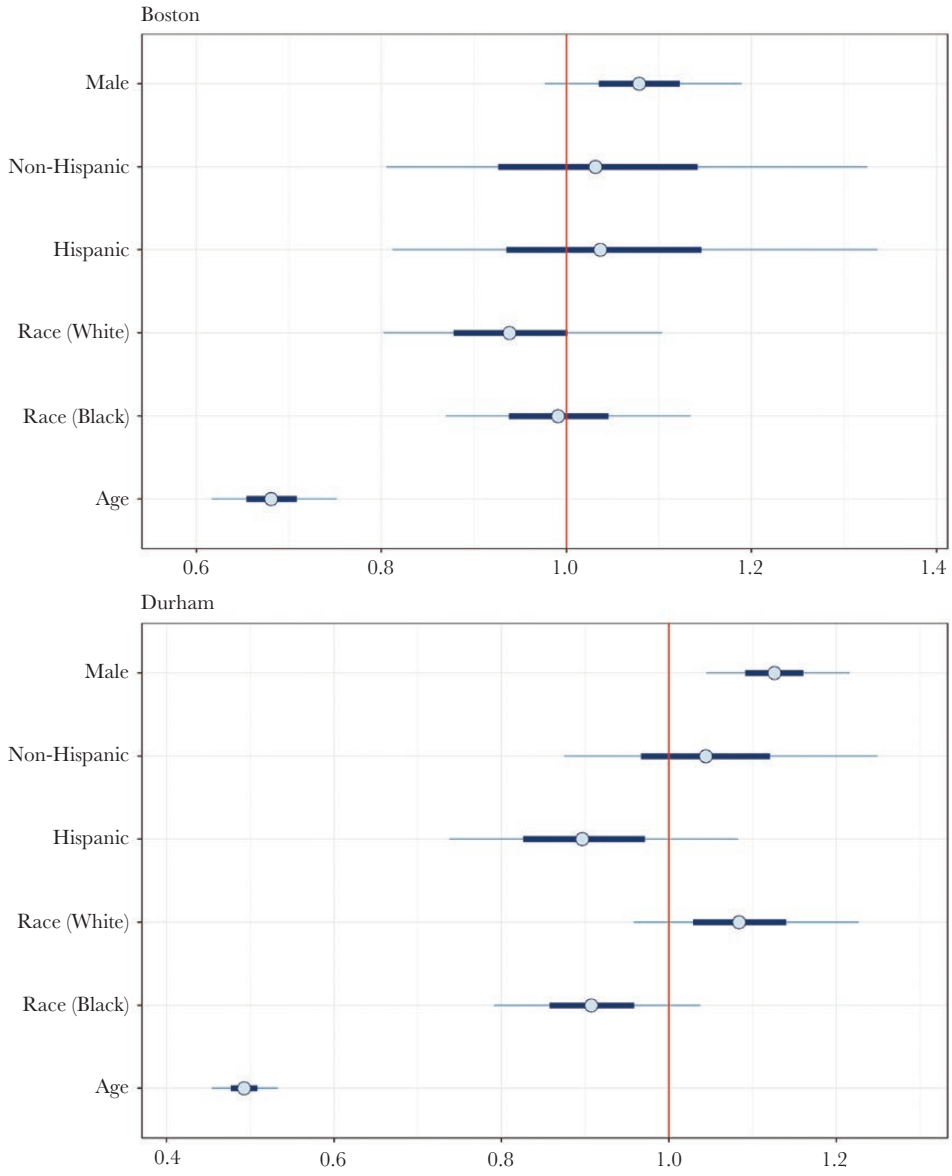

Figure 2. Influence of covariates on group A *Streptococcus* (GAS) pharyngitis. This figure illustrates the odds ratio (OR) of a positive GAS test. For a given covariate, the x-axis value represents the OR of a positive GAS test when all other variables are held at their mean value. The circles, thick bars, and thin bars represent the mean, 50% highest posterior density, and 95% highest posterior density, respectively, for each covariate. If a covariate's posterior distribution does not include 1, it can be assumed highly probable that it influences the OR of a positive GAS test. In both Boston and Durham, higher age was associated with a lower likelihood of a positive GAS test, as the posterior distribution is well to the left of 1. In Durham, male gender was associated with higher odds of a positive GAS test, something also observed but with less confidence in Boston. Race and ethnicity were not associated with GAS test results.

[16], a cutaneous manifestation of GAS pharyngitis. This has inspired an intriguing body of studies, which have identified spatial heterogeneity of scarlet fever in some sites, as well as statistical associations between scarlet fever and various meteorologic variables and pollutants. These studies have been conducted in several Chinese cities and provinces and have employed different modeling approaches, such as regression models (including GAMs) and discontinuous cluster statistics [17–22]. It is important to consider that both rheumatic fever and scarlet fever are a subset of total GAS cases and are associated with particular M-protein types in the former case and toxin elaboration in the latter. It may be that any spatial or spatiotemporal variability in scarlet fever or rheumatic fever is primarily due to heterogeneity in the circulating GAS strains. If so, that may explain why our study, which looked globally at GAS testing, did not replicate the spatial heterogeneity reported elsewhere.

Our study is limited by a number of factors inherent to retrospective research, including inability to specify and standardize subject recruitment and data collection in advance. However, manual chart review for the >20 000 included cultures would not have been feasible. We did not collect the results of GAS

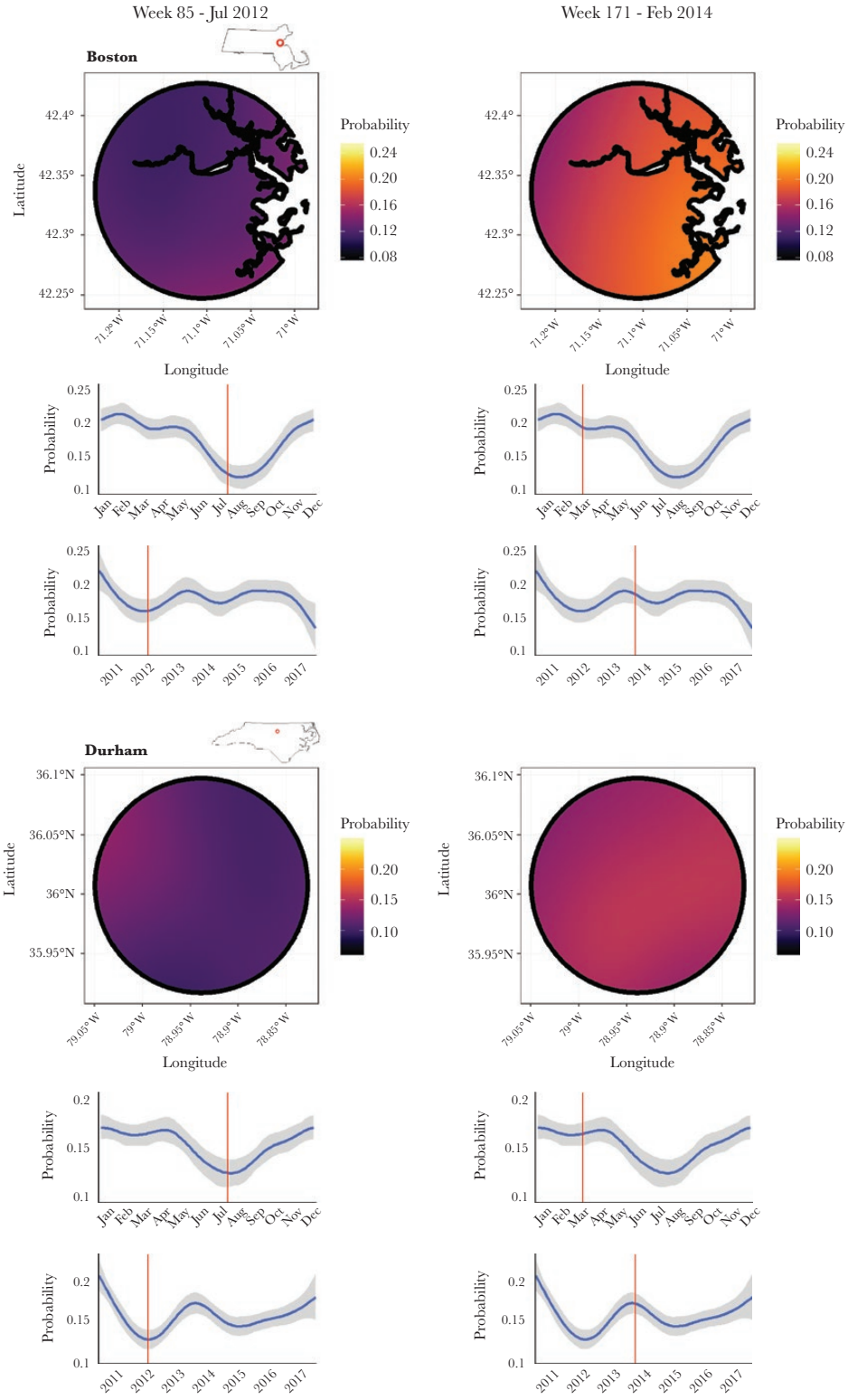

Figure 3. Videos 1 and 2: Spatiotemporal distribution of group A *Streptococcus* (GAS) testing data. We have selected 2 time points in Figure 3 to illustrate the probabilty of positive GAS tests in space at different time points. The supplementary videos show this evolution over our entire 371-week study period. The map figure encompasses a 10-km radius around Boston Children's Hospital and Duke University Hospital, onto which we have predicted the probability of a positive test for each of our 371 weeks. Contours that briefly appear in the animations show regions where the local probability of a positive GAS test differs from the mean with 90% (dotted), 95% (dashed), and 99% (solid) probability. Blue contours represent lower-than-average probability, and red represents higher than average. Below the map are 2 temporal smooths, showing the cyclic probability (top) and the longitudinal probability (bottom) of a positive GAS test. These animations show the spatial evolution of the odds of GAS pharyngitis, but no clear or sustained clusters of high or low probability. Although there is some temporal variability in both sites, the dominant finding is a decrease in probability during the summer months in both Boston and Durham.

culture testing. Some children with a negative GAS rapid test may later grow GAS from throat culture. However, we wanted to examine the potential for rapid test results to inform real-time clinical decision-making. In addition, not all clinical labs will perform culture for negative GAS results, which will be positive in a small minority of cases. With retrospective geospatial analyses, we are forced to accept a certain amount of error and uncertainty in location data. For instance, the address location we recorded represents each child's most recent address, but may not be where the child lived at the time of their illness. Even when home address data are accurate, disease exposure may have happened elsewhere. We have to assume, albeit with caution, that areas of potential disease around most children's addresses follow logic and probability; for instance, within a neighborhood with a sample of individual addresses, the probability of exposure is highest at shared school locations and other gathering places. Consequently, large data sets such as ours may illuminate real spatial trends. Most importantly, we did not have information for the school each child attended. In a further examination of our data, we will incorporate individual school catchment boundaries using a multilevel modeling approach to identify clustering of risk.

In summary, we have conducted a novel geostatistical analysis of GAS testing data. We constructed models that evaluated 2-dimensional geographic space, time as both a longitudinal variable and a cyclic (ie, seasonal) variable, using smoothing splines and Bayesian inference. This modeling approach may be an efficient, robust approach to infectious disease surveillance using electronic health data. In our study, although we did not identify compelling spatial trends in GAS risk, we did identify temporal patterns that suggest clustering among young children during the school year. Further development of this research should take into account school enrollment and evaluate the influence of environmental exposures.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

Financial support. None.

Potential conflicts of interest. All authors: no reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis **2005**; 5:685–94.
2. Centor RM, Witherspoon JM, Dalton HP, et al. The diagnosis of strep throat in adults in the emergency room. Med Decis Making **1981**; 1:239–46.
3. McIsaac WJ, Kellner JD, Aufricht P, et al. Empirical validation of guidelines for the management of pharyngitis in children and adults. JAMA **2004**; 291:1587–95.
4. Harrist A, Van Houten C, Shulman ST, et al. Notes from the field: group A streptococcal pharyngitis misdiagnoses at a rural urgent-care clinic—Wyoming, March 2015. MMWR Morb Mortal Wkly Rep **2016**; 64:1383–5.
5. Dooling KL, Shapiro DJ, Van Beneden C, et al. Overprescribing and inappropriate antibiotic selection for children with pharyngitis in the United States, 1997-2010. JAMA Pediatr **2014**; 168:1073–4.
6. Fine AM, Nizet V, Mandl KD. Improved diagnostic accuracy of group A streptococcal pharyngitis with use of real-time biosurveillance. Ann Intern Med **2011**; 155:345–52.
7. Asteberg I, Andersson Y, Dotevall L, et al. A food-borne streptococcal sore throat outbreak in a small community. Scand J Infect Dis **2006**; 38:988–94.
8. Danchin MH, Rogers S, Kelpie L, et al. Burden of acute sore throat and group A streptococcal pharyngitis in school-aged children and their families in Australia. Pediatrics **2007**; 120:950–7.
9. Kaplan EL, Wotton JT, Johnson DR. Dynamic epidemiology of group A streptococcal serotypes associated with pharyngitis. Lancet **2001**; 358:1334–7.
10. Gelman A. Scaling regression inputs by dividing by two standard deviations. Stat Med **2008**; 27:2865–73.
11. Wood SN. *Generalized Additive Models: An Introduction With R*. 2nd ed. Boca Raton, FL: CRC Press/Taylor & Francis Group; **2017**.
12. Bürkner P-C. brms: an r package for Bayesian multilevel models using Stan. J Stat Softw **2017**; 80:28.
13. Wood S. Package "mgcv": mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. Available at: https://cran.r-project.org/web/packages/mgcv/mgcv.pdf. Accessed 29 May 2017.
14. Simpson G. Modelling seasonal data with GAMs. Available at: https://www.fromthebottomoftheheap.net/2014/05/09/modelling-seasonal-data-with-gam/. Accessed 8 November 2019.
15. Lumsden RH, Akwanalo C, Chepkwony S, et al. Clinical and geographic patterns of rheumatic heart disease in outpatients attending cardiology clinic in Western Kenya. Int J Cardiol **2016**; 223:228–35.
16. Liu Y, Chan TC, Yap LW, et al. Resurgence of scarlet fever in China: a 13-year population-based surveillance study. Lancet Infect Dis **2018**; 18:903–12.
17. Duan Y, Yang LJ, Zhang YJ, et al. Effects of meteorological factors on incidence of scarlet fever during different periods in different districts of China. Sci Total Environ **2017**; 581–582:19–24.
18. Lu JY, Chen ZQ, Liu YH, et al. Effect of meteorological factors on scarlet fever incidence in Guangzhou City, Southern China, 2006–2017. Sci Total Environ **2019**; 663:227–35.
19. Mahara G, Wang C, Yang K, et al. The association between environmental factors and scarlet fever incidence in Beijing region: using GIS and spatial regression models. Int J Environ Res Public Health **2016**; 13(11):1083.
20. Duan Y, Huang XL, Wang YJ, et al. Impact of meteorological changes on the incidence of scarlet fever in Hefei City, China. Int J Biometeorol **2016**; 60:1543–50.
21. Tang JH, Tseng TJ, Chan TC. Detecting spatio-temporal hotspots of scarlet fever in Taiwan with spatio-temporal Gi* statistic. PLoS One **2019**; 14:e0215434.
22. Zhang Q, Liu W, Ma W, et al. Spatiotemporal epidemiology of scarlet fever in Jiangsu Province, China, 2005–2015. BMC Infect Dis **2017**; 17:596.