Using big data to predict pertussis infections in Jinan city, China: a time series analysis

Yuzhou Zhang\textsuperscript{1} · Hilary Bambrick\textsuperscript{1} · Kerrie Mengersen\textsuperscript{2} · Shilu Tong\textsuperscript{1,3,4} · Lei Feng\textsuperscript{5} · Li Zhang\textsuperscript{5} · Guifang Liu\textsuperscript{5} · Aiqiang Xu\textsuperscript{5} · Wenbiao Hu\textsuperscript{1}

Received: 14 February 2019 / Revised: 6 July 2019 / Accepted: 27 August 2019 / Published online: 2 September 2019
© ISB 2019

Abstract

This study aims to use big data (climate data, internet query data and school calendar patterns (SCP)) to improve pertussis surveillance and prediction, and develop an early warning model for pertussis epidemics. We collected weekly pertussis notifications, SCP, climate and internet search query data (Baidu index (BI)) in Jinan, China between 2013 and 2017. Time series decomposition and temporal risk assessment were used for examining the epidemic features in pertussis infections. A seasonal autoregressive integrated moving average (SARIMA) model and regression tree model were developed to predict pertussis occurrence using identified predictors. Our study demonstrates clear seasonal patterns in pertussis epidemics, and pertussis activity was most significantly associated with BI at 2-week lag ($r_{BI} = 0.73$, $p < 0.05$), temperature at 1-week lag ($r_{temp} = 0.19$, $p < 0.05$) and rainfall at 2-week lag ($r_{rainfall} = 0.27$, $p < 0.05$). No obvious relationship between pertussis peaks and school attendance was found in the study. Pertussis cases were more likely to be temporally concentrated throughout the epidemics during the study period. SARIMA models with 2-week-lagged BI and 1-week-lagged temperature had better predictive performance ($\beta_{search\ query} = 0.06$, $p = 0.02$; $\beta_{temp} = 0.16$, $p = 0.03$) with large correlation coefficients ($r = 0.67$, $p < 0.01$) and low root mean squared error (RMSE) value ($r = 3.59$). The regression tree model identified threshold values of potential predictors (search query, climate and SCP) for pertussis epidemics. Our results showed that internet query in conjunction with social and climatic data can predict pertussis epidemics, which is a foundation of using such data to develop early warning systems.

Keywords Climate · Social factor · Search terms · Pertussis · Prediction

Introduction

Pertussis, also known as whooping cough is one of the most risky infectious diseases for children, which caused by \textit{Bordetella pertussis}, leading to the annual cases of 143,963 in 2017 (WHO 2018), and around 63,000 deaths in children aged $<$ 5 years in 2013 worldwide (WHO 2015). Despite extensive immunizations, the disease has resurged in many countries over the last decade, such as the USA, Japan and Australia (Kamiya et al. 2012; Octavia et al. 2012; Schmidtke et al. 2012). Moreover, 95 % of cases occur in developing countries with a high case-fatality rate of up to 4 % (WHO 2003).
A vaccination programme against pertussis was commenced from the early 1960s in China, which involved giving three doses at 3, 4 and 5-months old. Another booster dose at 18–24 months was added in 1978 (Guo et al. 2013). The vaccine coverage increased from 58 to 99 % in the target populations between 1983 and 2018 in China (World Health Organization 2018). China has experienced a rapid rise in pertussis infections (Zhang et al. 2019), with a 66 % increase in the reported cases between 2005 and 2015 (Zeng et al. 2016).

The effects of climate factors on the transmission of infectious diseases has raised significant concerns of public health authorities and government (Haines et al. 2014), as they can affect both the survival of the pathogen in the environment and human population behaviours (Grassly and Fraser 2006). Current knowledge of the impacts of climate variables on pertussis transmission is limited. Previous studies reported that the rainy season, temperature and vapour pressure may contribute to the peak of transmission in pertussis infection, and may also be correlated to, school calendar pattern and socioeconomic index (Blackwood et al. 2012; Huang et al. 2017a). However, a substantial discrepancy was observed between countries with similar socioeconomic and vaccination conditions (Jackson and Rohani 2014). This indicates that the complex associations between pertussis with social and environmental factors need to be quantified for different areas to imply targeted preventive measures.

Traditional disease surveillance takes up to 2 weeks from the commencement to receipt of the notification by health authorities (Chan et al. 2010). This delay weakens the ability of the monitoring system in responding to possible outbreaks and provide timely epidemiological intelligence (Project 2011). Collectively, these challenges call for the development of surveillance systems to detect emerging epidemics. Internet search query data has been seen as a new tool of diseases surveillance and early warning in the high-risk area or the place where has limited access to traditional disease surveillance (Chretien et al. 2008). There is an increasing interest in achieving near real-time monitoring and prediction of infectious disease using internet data (Cho et al. 2013; Husnayain et al. 2019; Kang et al. 2013; Kapitány-Fővény et al. 2019; Seo et al. 2014; Shin et al. 2016). Currently, there is an increasing interest in performing internet search query to detect and even forecast pertussis outbreaks (Nagel et al. 2013; Pollett et al. 2015; Zhang et al. 2017).

Currently, little is known about the effects of climate and social factors on pertussis infections, especially in China. Furthermore, the value of combining internet search query, climate variables and social factors in tracking pertussis epidemics in China has not yet been examined. The aims of the study are identifying potential risk factors of pertussis infection and developing an early warning model using identified risk factors and internet query data.

### Methods

#### Study setting and data collection

Jinan is located in the east of China, it is the capital city of Shandong province (Fig. 1). Shandong is the second-largest province by population in China (National Statistics Bureau of China 2010), and Jinan’s total population is 7.32 million and landscape is 7998 km² as of 2016 (Statistics 2016). Jinan located in temperate climate with four distinct seasons, lying in the transition between the humid subtropical and humid continental zones (Zhang et al. 2015).

Weekly data on the total number of clinic- and laboratory-confirmed pertussis cases in Jinan, China from January, 2013 to December, 2017 were obtained from the Chinese National Notifiable Disease Reporting System (CNNDRS), this database has been widely used in previous studies of pertussis and other infectious diseases (Wang et al. 2018; Zeng et al. 2016). The detailed diagnosis confirmation of pertussis (Weisheng (WS, hygiene in Chinese) 274–2007) was issued by the Chinese Ministry of Health on 17 April 2007, and all laboratory-confirmed and clinically diagnosed cases must be uploaded to CNNDRS within 24 h after diagnosis (National Health Commission of the PRC 2007).

Climate data, including daily maximum temperature (°C) and rainfall (mm) data were obtained for Jinan from the National Oceanic and Atmospheric Administration (NOAA). The school calendar pattern (SCP) data were extracted from the Jinan Education Department. Weekly search metrics data for Jinan in the same period were obtained from the Baidu index (BI). The BI was produced by Baidu Company, which dominates the Chinese internet search market with 72 % search engine market share (Statcounter 2018). The BI is based on the search frequency at city level of keywords from internet users who used the Baidu search engine, and there is an increasing trend in using the BI in infectious diseases detection in China (Li et al. 2017; Liu et al. 2016).

We selected ten top search terms which were highly correlated with the term “pertussis” to try our best to capture different internet-seeking behaviours among populations (Milinovich et al. 2014) (Table S1); these ten pertussis-related search terms were officially provided by BI. Then, we used the search query which included the ten terms to download the data.

We transformed the daily maximum temperature and rainfall to weekly mean maximum temperature and cumulative rainfall for the analysis for consistency with the other time series data. The data analysis was completed via R software, version 3.5.2 and SPSS software, version 25 (SPSS Inc.; Chicago, IL, USA) (Liu et al. 2016).
Data analysis

Descriptive analysis of pertussis notifications, climate and search data

Time series decomposition analysis was performed to discover the systematic seasonality of pertussis infections and climate data (Zhang et al. 2017). This method decomposes time series data into three components, including trend, systematic seasonal factors and residual through performing a sequence of smoothing operations (Ke et al. 2016; Lee et al. 2017; Zhang et al. 2017). The statistical equation underlying this analysis is given in the Supplemental Material. Seasonal patterns of pertussis infections were also compared to SCP data (school attending periods).

Time series cross-correlation analysis

Associations between pertussis occurrence with BI and climate factors were assessed between weekly pertussis notifications with weekly BI, weekly mean maximum temperature and weekly cumulative rainfall applying time series cross-correlation analysis. As the factors are highly correlated to pertussis occurrence by different lags in time, we used the lagged variables with maximum correlation coefficient to develop the models in this paper (Sang et al. 2015).

Temporal risk assessment

The temporal patterns of an epidemic can be considered as a temporally sporadic or clustered occurrence. Several indicators were employed to quantify the magnitude and severity of disease outbreaks, and incidence rate is the most common indicator for measuring the temporal feature of an outbreak (Dunn et al. 2001). For the purposes of this study, an epidemic of pertussis was defined when the number of cases successively exceeded the mean weekly notifications of each calendar year for at least 2 weeks (Neuzil et al. 2000). Three temporal risk indices were employed to assess the severity and magnitude of a pertussis epidemic, namely (1) the frequency of occurrence of an epidemic, (2) the duration of an epidemic and (3) the intensity of an epidemic (Wen et al. 2006).

The relative frequency of epidemic ($\alpha$) was defined as the total number of weeks in which epidemic occurred (OW) out of the weeks’ number within a surveillance period (TW):

$$\alpha = \frac{OW}{TW}$$

Here, the surveillance period was taken to be a calendar year, so that TW = 52.

An epidemic duration index ($\beta$) was defined as the average number of weeks of an epidemic in the surveillance period:

$$\beta = \frac{OW}{ON}$$

where OW is as mentioned above and ON means the total number of epidemics for a calendar year. This index may reflect the effectiveness of the preventive measures used within the epidemic period. Moreover, a bigger coefficient might indicate infected cases take longer to be resolved and hence increase the risk of virus mutation.

An intensity index ($\gamma$) was defined as the increase in the rate of new cases within an epidemic. Unlike the incidence rate, which gives an overall estimate for the surveillance period, the

Fig. 1 The location of Jinan (green area) in Shandong province, China
intensity index aims to assess an epidemic during successive weeks where infections have continually occurred (Wen et al. 2006). The equation of intensity index is as follows:

$$\gamma = \frac{IN}{ON}$$

where IN represents the total number of pertussis notifications during the defined epidemics for each year; ON is as described above. The $\gamma$ value can be larger if most infections are temporally clustered during the epidemics. A small number of $\gamma$ indicates that the cases are more temporally dispersed, as more epidemics occurred in a calendar year.

**Seasonal autoregressive integrated moving average (SARIMA) model with climate and search data**

The SARIMA model is one of most common model to predict the infections of infectious diseases (Midekisa et al. 2012; Ren et al. 2013; Wongkoon et al. 2012). We developed SARIMA models to predict pertussis epidemics using climate and BI data. The predictors with the lag which has the highest cross-correlation were selected for developing the models. Three important components are included in the model, which are autoregressive (AR), differencing and moving average (MA). Usually, we choose three parameters for developing the model, $(p, d, q)$ $(P, D, Q)$, where $p$ is the order of the AR, $d$ is the order of the differencing, $q$ is the order of the MA, $P$ is the order of the seasonal AR, $D$ is the order of the seasonal differencing and $Q$ is the order of the seasonal MA (Box et al. 2015).

The formula of the developed SARIMA model in our study is as follows:

$$y_t = q(A) q(A^s)a_t \rho(A^s) / \rho(A) \rho(A)(1-A)^d (1-A^s)^D + BI + Tem + Rain$$

where $\Phi_p(A')$ is seasonal autoregressive operator, $\Theta_q(A)$ is autoregressive operator, $\Theta_q(A)$ is the operator of moving averages, $a_t$ is white noise, $y_t$ is predicted pertussis notifications, $BI$, $Tem$ and $Rain$ are the regressive coefficients of BI and climate factors.

The goodness-of-fit of our model was evaluated using partial autocorrelation of residuals function (PACF) and autocorrelation function (ACF) analyse. Furthermore, the Bayesian information criterion (BIC), stationary $R$ squared ($R^2$), maximum absolute percent error (MAPE) and root mean squared error (RMSE) values were applied for assessing the goodness-of-fit of our model. The dataset from January 2013 to December 2016 was used to train the SARIMA model, and data from January 2017 to December 2017 were used to validate the predicted values of the model. The predictive capacity of SARIMA models, which included and excluded BI and climate data was also examined, with preference given to the model with the bigger $R^2$ value and smaller BIC, RMSE and MAPE values. The model has better performance was selected in our final prediction and two metrics were used to quantify its predictive performance, namely Pearson correlation and RMSE.

**Regression tree analysis**

The regression tree model is a flexible, robust and non-parametric model which is widely used in infectious disease study (De'ath and Fabricius 2000; Liu et al. 2016; Zhang et al. 2018). A regression tree model was developed by dividing SCP, weekly climate and search data to subsets, and assess the levels of the correlations to weekly pertussis epidemics. The predictors with the biggest cross-correlation values were involved for developing our model. We used the data of SCP with two categories, attending and closure periods. The resultant binary SCP variable was then defined as a factor variable in the model construction. We also used the lagged pertussis data corresponding to the maximum autoregression (AR) coefficient in the model construction (Dugas et al. 2013). To select the best tree size, cross-validation was applied and estimated the prediction errors. The best tree was defined when the tree had the lowest estimated error rate within one standard error of the minimum with the smallest tree size (Breiman et al. 1984).

**Results**

**Descriptive analysis**

Similar patterns between the reported pertussis notifications and the BI data were found during the study period (Fig. 2). Moreover, there was an increasing trend in pertussis notifications of Jinan over the study period. Analyses of systematic seasonal variations for the pertussis notifications, temperature and rainfall showed clear seasonal patterns over the study period (Fig. 2). The pertussis notifications were peaked in summer (July–September) and winter (November–January). However, temperature and rainfall were observed to peak only in summer. The systematic seasonal patterns indicated that there is a clear semi-annual epidemic in pertussis infections with a higher peak in summer months, relative to the winter peak. Furthermore, the winter peaks of pertussis notifications occurred in the attending period of SCP, but the summer peaks occurred in the school closure period at the commencement of the summer holidays (Fig. 3).

**Time series cross-correlation analysis**

As depicted in Fig. 4, the time series cross-correlation analysis demonstrated that weekly pertussis notifications were strongly correlated with weekly BI metrics at the time lags of 2 weeks.
correlation coefficient of 0.73). Moreover, the strongest associations between weekly pertussis occurrence with weekly mean maximum temperature and cumulative rainfall were observed at 1-week and 2-week lags (correlation coefficients of 0.19 and 0.27, respectively). See Table S2 for details.

**Temporal risk analysis**

The epidemics number (ON) and frequency index (α) were relatively stable during the study period. However, there were increasing trends in notifications number (IN), the epidemic duration index (β) and the epidemic intensity index (γ). The largest IN, β and γ figures were observed in 2017 with values of 586, 11.0 and 293.0 respectively (Table 1).

Pairwise correlations among the total number of pertussis notifications during epidemics (IN), the epidemic duration index (β) and the epidemic intensity index (γ) are shown in Table S3. It was found that both β and γ are highly correlated with IN.

**SARIMA model**

BI, mean maximum temperature and cumulative rainfall with maximum cross-correlation coefficient were included in the study to develop the model. We used 2-weeks lag of BI, 1-week lag of mean maximum temperature and 2-weeks lag of rainfall in the development of the models. The SARIMA model (1,0,2) (1,0,0) provided the best fit to the data. The goodness-of-fit analysis indicated that our SARIMA models...
fitted the data well, which the PACF and ACF fluctuated randomly close to zero in the plots (Figure S1). The results of the SARIMA models were listed in Table S4. The model with BI and temperature data was selected based on its performance metrics, as the model has larger $R^2$ and smaller MAPE, RMSE and BIC coefficients. This model was used to fit and predict pertussis notifications 1-week ahead in 2017, and was validated by pertussis surveillance data (Figure S2) (Fig. 5).

The assessment of the predictive capacity of our model was shown in Table S5. The table indicates that the predictive performance of our model was high overall based on the Pearson correlation of 0.67 ($p < 0.01$) and the relatively small RMSE of 3.59.

**Regression tree analysis**

The BI and climate data with the same lag in the SARIMA development were included in the construction of regression tree models to explain the variation in weekly pertussis notifications. The model also included SCP as a factor variable and AR variable at lags of 1 week. The 1-week AR term for pertussis was found to be the most important factor among the candidate predictors (Fig. 6). Based on the model, the mean weekly pertussis notifications in Jinan increased by 4.2-fold (33.8/8.0) when weekly BI at 2-week lag exceeded 115 and 1-week-lagged weekly mean maximum temperature exceeded 30.1 °C. Furthermore, the mean weekly number of pertussis cases increased by 2.5-fold (20.3/8.0) when weekly BI at 2-week lag was < 115 and children were attending school.

**Discussion**

Pertussis is recognised as a resurgent disease in many countries, such as Australia, the USA and the UK (Cherry 2012; Mooi et al. 2014; Zhang et al. 2017). However, there has been limited research on the occurrence of pertussis in China. The
increased notifications of pertussis may partly result from the transition of vaccine-type from whole-cell pertussis vaccine to acellular vaccine, which has a lower protection level than that of whole-cell vaccine (Gambhir et al. 2015). Additionally, the appearance of erythromycin-resistant *Bordetella pertussis* should be alarming in China, as it is has been the traditional antimicrobial drug of choice for post-exposure prophylaxis of pertussis (Wang et al. 2014; Zhang et al. 2013).

There were similar trends between weekly pertussis notifications and BI during the study period. The findings indicate that the search query data, BI has a great potential in monitoring pertussis notifications in Jinan city. Seasonal decomposition analysis showed clear seasonal patterns of pertussis epidemics and climate data in the study period. They were observed to peak in summer months. However, pertussis infection was also observed to peak in winter months, which was a trough for temperature and rainfall. The results are in agreement with a previous study in the Netherlands, which observed that the annual peak of pertussis activity occurs in summer (de Greeff et al. 2009). This may partially result from the similar summer climate conditions between the two regions, where both of them has temperature climate with mean temperatures of 24 and 23 °C in July in Shandong and the Netherlands, respectively (Royal Netherlands Meteorological Institute 2018; Zhang et al. 2015). Moreover, the precipitation mostly occurs in summer in both of the two areas (MENG and Xu 2007; Royal Netherlands Meteorological Institute 2018). A high temperature might impact the survival of the pathogen and host susceptibility in the transmission of infectious diseases (Dowell 2001). Additionally, a previous laboratory study showed that the activity of pertussis toxin was promoted at a higher temperature (Murayama et al. 1994). In terms of rainfall, a study in Thailand reported that the periods of the peak of pertussis activity and heavy rainfall were both observed between May and June, which indicated that rainfall may contribute to transmission in pertussis infections (Blackwood et al. 2012). On the other hand, indoor crowding in rainy weather may also increase the risk of pertussis transmission (Huang et al. 2017b).

Interestingly, the peak season of pertussis is variable in different countries. The peaks of pertussis activity have been observed in autumn and winter months in Australia (Zhang et al. 2017), and in spring months in Canada (Skowronski et al. 2002). Moreover, the seasonal patterns of pertussis epidemics can change in the same area for different time periods (FINE and Clarkson 1986; Skowronski et al. 2002). The complex seasonal patterns of pertussis activity may reveal significant factors in the natural history of pertussis (FINE and Clarkson 1986).

Several studies have suggested that there might be a relationship between an increasing activity of pertussis and the opening period of school (Anderson et al. 1984; FINE and Clarkson 1986; Grenfell 1989), as schoolmates play a significant role in the transmission of pertussis during school outbreaks periods (Brennan et al. 2000; De Serres et al. 2000). However, our study showed that the summer peaks occurred in the school closure period (summer holidays) in the study. Interestingly, a study in Canada among schoolchildren found a significant rising of pertussis notifications in June, while the typical peaks months were August and September for infants and pre-schoolchildren (Skowronski et al. 2002). The results may suggest that schoolchildren become infected after an academic year and transmit pertussis to their family members in the following summer holidays (de Greeff et al. 2009). Our results suggest that further study is needed to offer insight into the complex pertussis transmission dynamics with socio-environmental factors.

Strong positive associations were observed at 2-weeks lag for BI, 1 week lag for temperature and 2 weeks lag for rainfall. The results suggest that an early warning system for pertussis would use both BI and climate data, as the findings reveal how much faster search query and climate data can facilitate pertussis epidemic intelligence. Thus, using search and climate data for monitoring pertussis can assist government and public health authorities to implement cost-effective control activity with a longer time window.

Importantly, a positive value for the duration index ($\beta$) is important for public health officers since it indicates the potential contributions of the preventive measures in the identified period of an epidemic (Wen et al. 2006). A bigger coefficient of $\beta$ reflects the occurrence has less opportunity to disappear when it occurs, which has more opportunity of the mutation for virus (Wen et al. 2006). It should be noted that there were increasing trends in the epidemic weeks (OW), the total number of notifications (IN), the duration index ($\beta$) and the intensity index ($\gamma$) in the study period, although the epidemics number (ON) and frequency index ($\alpha$) were relatively stable. This indicates that pertussis cases were increasingly

### Table 1

Statistics of temporal risk indices of pertussis in Jinan from 2013 to 2017

| Year | OW | ON | IN | Frequency index ($\alpha$) | Duration index ($\beta$) | Intensity index ($\gamma$) |
|------|----|----|----|--------------------------|------------------------|--------------------------|
| 2013 | 10 | 3  | 15 | 0.2                      | 3.3                     | 5.0                      |
| 2014 | 19 | 4  | 141| 0.4                      | 4.6                     | 35.3                     |
| 2015 | 15 | 2  | 180| 0.3                      | 7.5                     | 90.0                     |
| 2016 | 16 | 3  | 288| 0.3                      | 5.3                     | 96.0                     |
| 2017 | 22 | 2  | 586| 0.4                      | 11.0                    | 293.0                    |
likely to be temporally concentrated throughout the epidemics, and more likely to continue once they occur over the study period. Additionally, strong correlations between IN with the duration index ($\beta$) and the intensity index ($\gamma$) for pertussis were observed for this study in Jinan. This indicates that a larger size of pertussis epidemic occurred when cases were more temporally concentrated and last for a longer time throughout a calendar year.

The findings suggest that SARIMA models that include pertussis notifications data with internet search and climate data may afford more accurate forecasting at 1 week ahead. The findings suggest that a SARIMA model that includes internet search query data and climate variables provides a good fit to pertussis surveillance data, and has a good 1 week ahead forecasting result. The incorporation of such a model, which combines internet search and climate data with conventional surveillance system may greatly improve pertussis epidemics tracking by limiting the impact of reporting lag on the system.

The results of the regression tree analysis showed that the 1-week-lagged AR of pertussis was the most important driver of the occurrence of pertussis epidemics in Jinan. The results revealed that our models based on AR of pertussis, BI, SCP and climate data can identify the threshold value of the included risk factors for pertussis infections, where use the pertussis surveillance data.

We acknowledge the following limitations inherent in using the data. First, CNNDRS only includes the reported

---

**Fig. 5** Observed and 1-week ahead predicted pertussis notifications during week 1–52, 2017 based on the SARIMA model. (LCL, lower 95% confidence limits; UCL, upper 95% confidence limits).

---

**Fig. 6** A regression tree model of the hierarchical relationship between weekly pertussis notifications with BI, climate and SCP in Jinan between 2013 and 2017. (The regression trees show the threshold values of the predictor variables at the splits and the mean weekly pertussis notifications in the node boxes; $N$ is the total number of weeks of occurrence of pertussis outbreaks in the subgroup of observations indicated by the path leading to that node; AR, autoregression.)
number of cases, thus, the patients of pertussis infection but do not seek health care or have been misdiagnosed by a clinic and laboratory are exclude. Second, the reported notifications may be impacted by the increasing reporting awareness among the physicians in medical institutes (Zhang et al. 2019). Third, the search query data of BI may be more accurate when more search terms included in our future study. Fourth, previous studies showed that different internet-seeking behaviours, as well as over media and self-reporting of disease-related information may also influence the results (Milinovich et al. 2014). This motivates further research into ways in which these types of bias can be detected and mitigated.

Conclusion

In conclusion, pertussis has recently resurged and remains epidemic globally and a substantial health burden. This disease re-emerged as a public health concern in China in the last decade, despite an extensive national immunization programme. The epidemic of pertussis had a larger peak in summer months, relative to the winter peak. The winter peak occurs following a return to school, but the summer peak occurs in the school holiday. This study suggests that internet search query data combined with climate and social factors have great potential to more accurately forecast pertussis epidemics, and this can be seen as a foundation of developing an early warning system of pertussis occurrence. Such systems may facilitate public health, as well as government to forecast and respond to the future pertussis outbreaks.

Acknowledgements

We would like to express our gratitude to the Shandong Center for Disease Control and Prevention for providing pertussis surveillance data. Y.Z. was supported by the China Scholarship Council Postgraduate Scholarship and the Queensland University of Technology Higher Degree Research Tuition Fee Sponsorship. L. F. was supported by Shandong Medical and Health Science and Technology Development Programs (award no. FT140101216). K. M. was supported by ARC Laureate Fellowship (award no. FL150100150) and an ARC Centre of Excellence in Mathematical and Statistical Frontiers (award no. CE140100049).

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Anderson R, Grenfell B, May R (1984) Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. Epidemiol Infect 93:587–608

Blackwood J, Cummings D, Brouth H, Iamsirithaworn S, Rohani P (2012) The population ecology of infectious diseases: pertussis in Thailand as a case study. Parasitology 139:1888–1898

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks Monterey, CA

Brennan M et al (2000) Evidence for transmission of pertussis in schools, Massachusetts, 1996: epidemiologic data supported by pulsed-field gel electrophoresis studies. J Infect Dis 181:210–215

Chan EH et al (2010) Global capacity for emerging infectious disease detection. Proc Natl Acad Sci 107:21701–21706

Cherry JD (2012) Epidemic pertussis in 2012—the resurgence of a vaccine-preventable disease. N Engl J Med 367:785–787

Cho S et al (2013) Correlation between national influenza surveillance data and google trends in South Korea. PLoS One 8:e81422

Chretien J-P et al (2008) Syndromic surveillance: adapting innovations to developing settings. PLoS Med 5:e72

de Greeff SC, Dekkers AL, Teunis P, Rahamat-Langendonc JC, Mooi FR, de Melker HE (2009) Seasonal patterns in time series of pertussis. Epidemiol Infect 137:1388–1395

De Serres G et al (2000) Morbidity of pertussis in adolescents and adults. J Infect Dis 182:174–179

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192

Dowell SF (2001) Seasonal variation in host susceptibility and cycles of certain infectious diseases. Emerg Infect Dis 7:369

Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, Rothman RE (2013) Influenza forecasting with Google flu trends. PLoS One 8:e56176

Dunn CE et al (2001) Analysing spatially referenced public health data: a comparison of three methodological approaches. Health Place 7:1–12

Fine PE, Clarkson JA (1986) Seasonal influences on pertussis. Int J Epidemiol 15:237–247

Gamblir M, Clark TA, Cauchemez S, Tartof SY, Swerdlow DL, Ferguson NM (2015) A change in vaccine efficacy and duration of protection explains recent rises in pertussis incidence in the United States. PLoS Comput Biol 11:e1004138

Grassly NC, Fraser C (2006) Seasonal infectious disease epidemiology. Proc R Soc Lond B Biol Sci 273:2541–2550

Grenfell B (1989) Pertussis in England and Wales: an investigation of transmission dynamics and control by mass vaccination. Proc R Soc Lond B 236:213–252

Guo B, Page A, Wang H, Taylor R, McIntyre P (2013) Systematic review of reporting rates of adverse events following immunization: an international comparison of post-marketing surveillance programs with reference to China. Vaccine 31:603–617

Haines A, Ebi KL, Smith KR, Woodward A (2014) Health risks of climate change: act now or pay later. Lancet 384:1073–1075

Huang X et al (2017a) Assessing the social and environmental determinants of pertussis epidemics in Queensland, Australia: a Bayesian spatio-temporal analysis. Epidemiol Infect 145:1221–1230

Huang X, Mengersen K, Milinovich G, Hu W (2017b) Effect of weather variability on seasonal influenza among different age groups in Queensland, Australia: a Bayesian spatiotemporal analysis. J Infect Dis 215:1695–1701

Husnayain A, Fuad A, Lazuardi L (2019) Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action 12:1552652

Jackson D, Rohani P (2014) Perplexities of pertussis: recent global epidemiological trends and their potential causes. Epidemiol Infect 142:672–684

Kamiya H, Otsuka N, Ando Y, Odaira F, Yoshino S, Kawano K, Takahashi H, Nishida T, Hidaka Y, Toyozumi-Ajisaka H,
Shibayama K, Kamachi K, Sunagawa T, Taniguchi K, Okabe N (2012) Transmission of *Bordetella holmesii* during pertussis outbreak, Japan. Emerg Infect Dis 18:1166

Kang M, Zhong H, He J, Rutherford S, Yang F (2013) Using google trends for influenza surveillance in South China. PLoS One 8: e55205

Kaptitány-Fövény M, Ferenci T, Sulyok Z, Kegele J, Richter H, Vályi-Nagy I, Sulyok M (2019) Can Google Trends data improve forecasting of Lyme disease incidence? Zoonoses Public Health 66:101–107

Ke G et al (2016) Epidemiological analysis of hemorrhagic fever with renal syndrome in China with the seasonal-trend decomposition method and the exponential smoothing model. Sci Rep 6:39350

Lee HS, Nguyen-Viet H, Nam VS, Lee M, Won S, Duc PF, Grace D (2017) Seasonal patterns of dengue fever and associated climate factors in 4 provinces in Vietnam from 1994 to 2013. BMC Infect Dis 17:218

Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, Deng A, Peng Z, Xiao J, Rutherford S, Xie R, Zeng W, Li X, Ma W (2017) Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China. PLoS Negl Trop Dis 11: e0005354

Liu K et al (2016) Using Baidu search index to predict Dengue outbreak in China. Sci Rep 6:38040

Meng C, Xu Z (2007) Relation between ENSO and Pprecipitation in Shandong [J]. Yellow River 1:014

Midekisa A, Senay G, Henebry GM, Semunigse P, Wimberly MC (2012) Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. Malar J 11:165

Milinovich GJ, Williams GM, Clements AC, Hu W (2014) Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 14:160–168

Mooi FR, Van Der Maas NA, De Melker HE (2014) Pertussis resurgence: waning immunity and pathogen adaptation—two sides of the same coin. Epidemiol Infect 142:685–694

Murayama T, Hewlett EL, Maloney NJ, Justice JM, Moss J (1994) Effect of temperature and host factors on the activities of pertussis toxin and Bordetella adenylate cyclase. Biochemistry 33:15293–15297

Nagel AC et al. (2013) The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. J Med Internet Res 15

National Health Commission of the PRC (2007) Pertussis diagnostic criteria. http://www.nhfpc.gov.cn/zwjgzt/s9491/201410/52040bc16d3b4ecae56e2883358666.shtml

National Statistics Bureau of China (2010) The Sixth National Population Census data. http://data.stats.gov.cn/

Neuzil KM, Wright PF, Mitchell EF Jr, Griffin MR (2000) The burden of influenza illness in children with asthma and other chronic medical conditions. J Pediatr 137:856–864

Octavia S, Smitchenko V, Gilbert GL, Lawrence A, Keil AD, Hogg G, Lan R (2012) Newly emerging clones of Bordetella pertussis carrying pm2 and pttxP3 alleles implicated in Australian pertussis epidemic in 2008–2010. J Infect Dis 205:1220–1224

Pollett S et al. (2015) Validating the use of Google trends to enhance pertussis surveillance in California PLoS Curr 7

Project TS (2011) Assessment of syndromic surveillance in Europe. Lancet 378:1833–1834

Ren H, Li J, Yuan Z-A, Hu J-Y, Yu Y, Lu Y-H (2013) The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. BMC Infect Dis 13:421

Royal Netherlands Meteorological Institute (2018) Monthly overview of the weather in the Netherlands. https://www.knmi.nl/nederland-nl/klimatologie/gegevens/mnow

Sang S et al (2015) Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. PLoS Negl Trop Dis 9:e0003808

Schmidtke AJ, Boney KO, Martin SW, Skoff TH, Tondella ML, Tatti KM (2012) Population diversity among Bordetella pertussis isolates, United States, 1935–2009. Emerg Infect Dis 18:1248

Seo D-W et al (2014) Cumulative query method for influenza surveillance using search engine data. J Med Internet Res 16:e289

Shin S-Y et al (2016) Correlation between national influenza surveillance data and search queries from mobile devices and desktops in South Korea. PLoS One 11:e0158539

Skowronski DM, de Serres G, MacDonald D, Wu W, Shaw C, Macnabb J, Champagne S, Patrick DM, Halperin SA (2002) The changing age and seasonal profile of pertussis in Canada. J Infect Dis 185:1448–1453

Statcounter (2018) Search engine market share China. http://gs.statcounter.com/search-engine-market-share/all/china

Statistics SPBo (2016) Shandong statistical yearbook. http://www.stats- sd.gov.cn/gtb/index.jsp?url=http%3A%2F%2Fwww.stats-sd.gov.cn%2Fart%2F2018%2F4%2F2%art_6131_404684.html

Wang Z, Cui Z, Li Y, Hou T, Liu X, Yi Y, Liu L, He Q (2014) High prevalence of erythromycin-resistant Bordetella pertussis in Xi’an, China. Clin Microbiol Infect 20:O825–O830

Wang Y, Xu C, Wang Z, Zhang S, Zhu Y, Yuan J (2018) Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. PLoS One 13: e0208404. https://doi.org/10.1371/journal.pone.0208404

West H, Lin NH, Lin C-H, King C-C, Su M-D (2006) Spatial mapping of temporal risk characteristics to improve environmental health risk identification: a case study of a dengue epidemic in Taiwan. Sci Total Environ 367:631–640

WHO (2003) WHO-recommended surveillance standard of pertussis. http://www.who.int/immunization.monitoring_surveillance/burden/vpd/surveillance_type/passive/pertussis_standards/en/

WHO (2015) Weekly epidemiological record. https://www.who.int/wer/2015/wer9035.pdf?ua=1

WHO (2018) Pertussis. https://www.who.int/immunization_monitoring_surveillance/burden/vpd/surveillance_type/passive/pertussis_standards/en/

Wongkoon S, Jaroensutasinee M, Jaroensutasinee K (2012) Assessing the temporal modelling for prediction of dengue infection in northern and northeastern, Thailand. Trop Biomed 29:339–348

World Health Organization (2018) WHO vaccine-preventable diseases: monitoring system. 2018 global summary, http://apps.who.int/immunization_monitoring/globalsummary/countries?countrycriteria%5Bcountry%5D=CN&commit=OK

Zeng Q et al (2016) Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016. Sci Rep 6:32367

Zhang Q, Li M, Wang L, Xin T, He Q (2013) High-resolution melting analysis for the detection of two erythromycin-resistant Bordetella pertussis strains carried by healthy schoolchildren in China. Clin Microbiol Infect 19:E260–E262

Zhang J, Su Y, Wu J, Liang H (2015) GIS based land suitability assessment for tobacco production using AHP and fuzzy set in Shandong province of China. Comput Electron Agric 114:202–211

Zhang Y, Milinovich G, Xu Z, Bambrick H, Mengersen K, Tong S, Hu W (2017) Monitoring pertussis infections using internet search queries. Sci Rep 7:10437

Zhang Y, Bambrick H, Mengersen K, Tong S, Hu W (2018) Using Google Trends and ambient temperature to predict seasonal influenza outbreaks. Environ Int 117:284–291

Zhang Y et al. (2019) Resurgence of pertussis infections in Shandong, China: space-time cluster and trend analysis. Am J Trop Med Hyg:tpmd190013