COMPARISON OF TIME SERIES MODELS PREDICTING TRENDS IN TYPHOID CASES IN NORTHERN INDIA

Kumar Shashvat[1], Rikmantra Basu[1], Amol P Bhondekar[1*], Sanjay Lamba[2*], Karan Verma[2] and Arshpreet Kaur[2]

[1, 2]National Institute of Technology, Delhi, [1*]Central Scientific Instruments Organization, Chandigarh; [2*]Birla Institute of Technology, Pilani, India

Abstract. Time series models have been used with varying degrees of success to predict diseases trends. This study compared and evaluated time series models of typhoid cases and analyzed formulated trends in a northern region of India (Chandigarh) where incidences of typhoid showed an elevated trend annually. The time series analysis was conducted from data collected from the Government of India integrated diseases surveillance programs from 2014 to 2017 using three time series models, namely, autoregressive integrated moving average (ARIMA), support vector machine regression and exponential smoothing. Evaluation of performance in prediction of number of cases is reported as root mean square error and mean absolute error**,** low values indicating superior model**.** Exponential smoothing model outperformed the other two models, with values of root mean square error and mean absolute error for exponential smoothing, support vector regression and ARIMA of 13.30 and 4.51, 16.49 and 10.91 and 17.39 and 11.42, respectively. The results indicate a need of strict actions in the field of sanitation to disrupt the present trend. Future studies will be conducted using more numbers of real time data sets and the ensemble method.

Keywords: autoregressive integrated moving average model, exponential smoothing model, support vector machine regression model, time series model, typhoid, India

INTRODUCTION

Typhoid fever is a noteworthy cause of fatality across the world, especially in developing countries (Dewan *et al*, 2013). A current estimate indicates a mortality of 200,000 per 22 million typhoid cases worldwide, with >90% occurring in Asia (Luby *et al*, 1998). This disease is transmitted by *Salmonella* and the infection in humans is categorized into two types, namely, infection due to low virulence enteric *Salmonella* serotypes, responsible for food poisoning, and that due to enteric *S*. *typhus*, the cause of typhoid, with a group of serovars, *S*. Paratyphi A, B, and C, responsible for paratyphoid (Vollaard *et al*, 2004). Humans can also act as carriers of typhoid.

In general, typhoid is prevalent in impoverished areas of the world, which face challenges in ensuring safe drinking water supply and proper sanitation

Correspondence: Kumar Shashvat, Department of Computer Science and Engineering, National Institute of Technology, New Delhi, Delhi 110040, India.
Tel: +9178 3779 7413; Fax: +9111 2778 7503
E-mail: shashvat.sharma13@gmail.com
*Contributed equally to the work.

(Karkey *et al*, 2010). In developing countries, such as Bangladesh, India and Pakistan, typhoid is the major cause of morbidity and mortality (Ram *et al*, 2007). Risk factors, such as contaminated food (Velema *et al*, 1997; Sharma *et al*, 2009; Wang *et al*, 2012), water (Black *et al*, 1985; Mermin *et al*, 1999) have been acknowledged as the foremost causes for typhoid pervasiveness; other factors, such as close contact with typhoid cases (Tran *et al*, 2005), lack of awareness, living near contaminated water bodies (Sur *et al*, 2007; Kothari *et al*, 2008), flood (Kanungo *et al*, 2008), poor hygiene and standards of living (Ochiai *et al*, 2008; Whitaker *et al*, 2009); and moving into endemic areas (Kelly-Hope *et al* 2007).

Prediction of disease outbreaks can help create a foundation to provide early warning and initiate an advance preparation. In order to achieve these goals, different time series models, such as exponential smoothing and regression methods, have been carried out (Sharma *et al*, 2009). Exponential smoothing model is a time series model that works using a trend projection technique, the prediction of which is considered accurate as it accounts for the difference between projection and what actually has occurred; also it gives more significance to the most recent observations. The regression analysis is a statistical method used to describe relationship among variables to investigate the predictive model; the regression method models the data by forming relationship between the target and independent variable predictor (Agrawal and Adhikari, 2013). Autoregressive integrated moving average model (ARIMA), a serene differencing, autoregressive and moving average model, which inspects infection rate as a linear combination of previous values and residuals, has also been extensively used for prediction of various disease outbreaks (Agrawal and Adhikari, 2013). Recently, machine learning approaches such as the artificial neural network and support vector machines have been used for statistical modelling of the probability of an occurrence of a disease (Zhang *et al*, 2013).

The objectives of the study were to employ ARIMA, exponential smoothing and regression models to predict trends of typhoid cases in a region of India (Chandigarh) during a recent period (2014 - 2017) and to evaluate the model with highest predictive accuracy compared to actual collected data. The results should be of benefit in formulating appropriate prevention measures and policy for control of this disease of public health concern in the country.

MATERIALS AND METHODS

Study region

The study was carried in Chandigarh, a northern region of India located at latitude 76°E and longitude 76°E, covering an area of 114 km^2 of which 4.47 km^2 are rural and 109.53 km^2 urban (Fig 1). Based on the 2011 census, total population is over one million with an average literacy rate of 86.56% (Pradesh, 2011). The annual rainfall is 1,110.7 mm, mainly (80%) during July to October (Kumar *et al*, 2015).

Data collection

Monthly data of typhoid cases were obtained from the Integrated Diseases Surveillance Programme (IDSP), a project of the National Centre for Disease Control, Government of India during 2014 to 2017 (Table 1).

Ethical clearance for the study was approved by the Integrated Diseases Surveillance Programme, Government of India, reference no. E-151913.

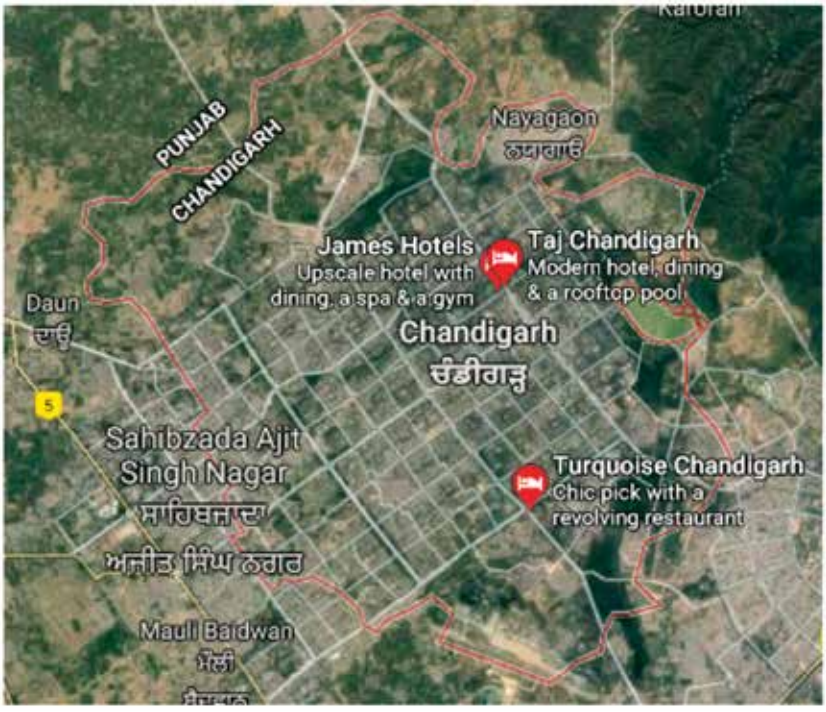

Fig 1-Study region of Chandigarh, India.

Table 1
Monthly typhoid cases in Chandigarh region, India 2014-2017.

Month	2014	2015	2016	2017
January	16	11	15	15
February	24	18	20	28
March	29	25	25	26
April	24	23	21	26
May	23	25	17	42
June	42	35	41	42
July	42	48	50	37
August	51	102	49	37
September	62	71	46	40
October	26	23	30	21
November	16	29	13	43
December	35	22	14	32
Total	390	432	341	389

Data analysis

Typhoid database was analyzed using three time series models, namely, ARIMA, exponential smoothing and support vector machine (SVM) regression. Implementation for this research was performed using R Language Studio version (R 3.0) (R Language Team, Auckland, New Zealand).

Decomposition method. The method decomposes time series into long-term trend and seasonality indices employing the following procedures (Zhang *et al*, 2016):

(1) Compute seasonality index $\overline{Z}_k$ using equation (1):

$$\overline{Z}_k = \frac{\Sigma_{i=1}^{k} Z_{ik}}{x}, \; k = 1,2,....m \qquad (1)$$

where Z_{ik} denote the incidence in the k-th month of the i-th year and x the number of time points.

(2) Compute overall average value $\overline{Z}$ using equation (1.1):

$$\overline{Z} = \frac{\Sigma_{i=1}^{x} \Sigma_{k=1}^{m} x_{ik}}{xm} \qquad (1.1)$$

where x_{ik} denotes the cases in the k-th month of i-th year, x the number of time points and m the total number of time points as specified in equation 1.

(3) Compute seasonal index S_k using equation (1.2):

$$S_k = \frac{\overline{Z}_k}{\overline{\overline{Z}}}, \; k = 1,2,... \; ... \; m \qquad (1.2)$$

where k denotes month of i-th year.

The decomposed series is then:

$$SR = Z_{ik} - S_k \qquad (1.3)$$

The decomposition method is used to decompose the time series into seasonal indices and long term trend. The seasonal index is calculated by using the above three steps. Firstly, it calculates the seasonal index by determining the average value in each period of equation (1). Secondly, it calculates the overall average value as described in equation (1.1). Thirdly, it calculates the seasonal index and decompose the series using equation (1.2).

ARIMA model. The ARIMA model converts a non-stationary time series into stationary data by handling the level of differences in the data points using the following equations (Agrawal and Adhikari, 2013):

$$\gamma(L)(1 - L)^d x_t = \theta(L)\varepsilon_t \text{ i.e.} \qquad (2)$$

$$\left(1 - \sum_{i=1}^{p} \gamma_i L^i\right)(1 - L)^d x_t = \left(1 + \sum_{k=1}^{q} \theta_k L^k\right)\varepsilon_t \qquad (2.1)$$

where p, d and q are integers ≥0 and consigned to the order of the autoregressive, integrated and moving average parts of the model, and d denotes significance of level of difference. In general, d = 1 is sufficient in most cases.

Exponential smoothing model. The exponential smoothing model assigns more weight to recent observations and exponentially decreases the weight of past observations over time using the following equations:

$$S_0 = y_0 \qquad (3)$$

$$S_t = \alpha\, y_{t-1} + (1 - \alpha)\, S_{t-1}, \; t > 0 \qquad (3.1)$$

where α is smoothing factor and S_t is output of the exponential smoothing model.

SVM regression model. SVM is applied to binary classification problems to find a canonical hyper plane, which maximally separates two given classes of training samples. For two sets of linearly separable training data points that are classified into one of two classes, B1 and B2 (say), using linear hyper planes (*ie* straight lines) from an infinite number of separating hyper planes the one with maximum margin is selected as having the best classification and generalization (Fig 2) (Agrawal and Adhikari, 2013). In most applications data

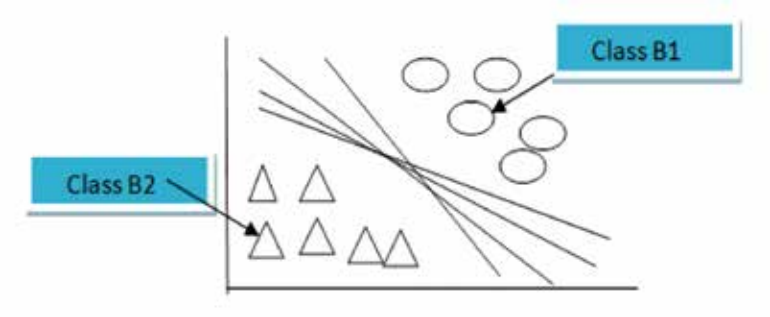

Fig 2-Support vector machine applied to find a canonical hyper plane with maximal separation between two classes of training samples. For two sets of linearly separable training data points that are classified into one of two classes, B1 and B2, using linear hyper planes (*ie*, straight lines) from an infinite number of separating hyper planes the one with maximum margin is selected as having the best classification and generalization (Agrawal and Adhikari, 2013).

points are not linearly separable and in such cases, a soft margin hyper plane classifier is constructed as follows:

$X \varepsilon R^n$: $w^T x + b = 0$, where $w \varepsilon R^n$, $b \varepsilon R$ (4)

Minimization:

$$K(w, \varepsilon) = \frac{1}{2} \| w \|^2 + c(\Sigma_{i=1}^{n} \varepsilon_i) \quad (4.1)$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i \; \forall_i = 1,2 \ldots N \; \varepsilon_i \geq 0 \quad (4.2)$$

In SVM regression model, as in the classification problem, if y_i belongs to, for example, A1and A2, $y_i \varepsilon$ □ where □ is a real number and other variables are same as for the classification problem.

Performance evaluation criteria

Root mean square (RMSE) error and mean absolute error (MAE) were used to evaluate the performance of each model (Zhang *et al*, 2014). These performance values were calculated using the following equations:

$$RMSE = \frac{1}{N} \sqrt{\sum_{i=1}^{n} (P_n - Z_n)^2} \quad (5)$$

$$RMSE = \frac{1}{N} \sum_{i=1}^{n} | P_n - Z_n | \quad (5.1)$$

RESULTS

Monthly indices of typhoid cases in Chandigarh region, India each year from 2014 to 2017 showed an increase in the number of cases from June till September (Table 2). The indices for these months were higher compared to other months in each year surveyed.

The prediction from ARIMA model that included a non-seasonal MA(1) term and a seasonal MA(1) term with no differentiation among seasonal periods (S =12) compared to actual results is illustrated in Fig 3, that from exponential smoothing model in Fig 4 and from SVM regression model in Fig 5. A comparison of performance among the three prediction models indicated the exponential smoothing model was the best as it had the lowest RMSE and MAE values (Table 3). Prediction of the total number of typhoid cases in Chandigarh region from January to October 2018 indicated that the number of cases from exponential smoothing model was lower than that from ARIMA and higher than support vector regression models (Table 4).

Table 2
Monthly indices of typhoid incidence in Chandigarh region, India 2014-2017.

Month	2014	2015	2016	2017
January	0.49	0.30	0.52	0.46
February	0.73	0.5	0.70	0.86
March	0.89	0.69	0.88	0.80
April	0.73	0.63	0.73	0.80
May	0.70	0.69	0.59	1.29
June	1.29	0.97	1.44	1.29
July	1.29	1.33	1.76	1.14
August	1.56	2.83	1.72	1.14
September	1.90	1.97	1.61	1.23
October	0.8	0.63	1.05	0.6
November	0.49	0.80	0.45	1.32
December	1.07	0.61	0.49	0.98

Monthly index = Average monthly case / Average number of cases for all months.

DISCUSSION

Typhoid is one of the most neglected tropical diseases having the most number of incidences in northern parts of India as well as in other parts of the country (Kelly-Hope *et al*, 2007). Poor sanitation plays an important role in the spread of this disease but is not the only contributing factor. Surveillance of an infectious disease constitutes an important aspect in disease management and prevention. In this study, we compared surveillance data of typhoid cases in Chandigarh region from 2014-2017 with three time series prediction models and found the exponential smoothing model gave the optimal accurate prediction compared to the actual collected data. It was also noted typhoid cases in this northern of the country peaked region annually between June to September.

Zhang *et al* (2013) reported a comprehensive study of different time series prediction methods on the monthly incidence of typhoid incidence involving ARIMA, backpropagation neural network, radial basis neural network, and Elman recurrent neural network models. Performances of these models evaluated against data obtained from the Chinese Center for Diseases Control and Prevention 2005-2009 indicated radial basis function network has the best performance. Subsequently, Zhang *et al* (2014) described a study using data of nine types of infectious diseases data obtained from China public health surveillance system 2005-2011 for comparison that demonstrated prediction performances of ARIMA, exponential smoothing and SVM regression models perform equally well. More recently, Zhang *et al* (2016) reported the application of ARIMA model to predict monthly cases of 11 notifiable diseases in China. It is worth noting parameters used to evaluate time series prediction models are different among reported studies

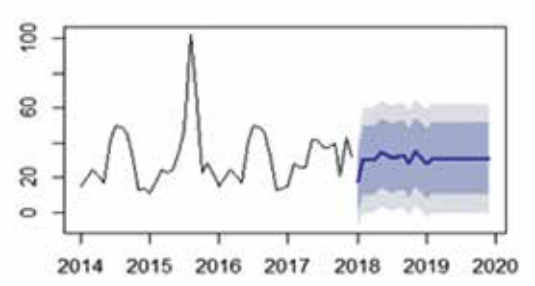

Fig 3-Prediction from autoregressive integrated moving average of monthly typhoid cases in Chandigarh region, India. Y-axis shows predicted monthly number of typhoid cases and x-axis shows year. Blue curve highlighted in grey indicates period with no actual collected data.

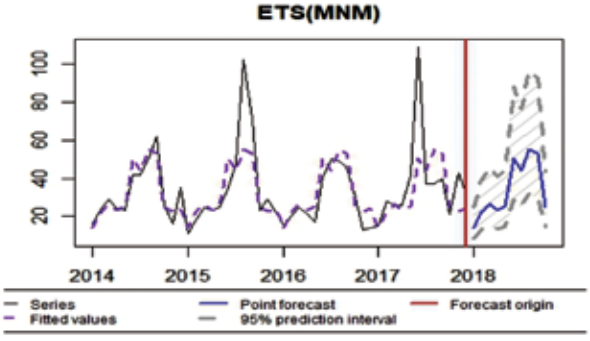

Fig 4-Prediction from exponential smoothing model of monthly typhoid cases in Chandigarh region, India. Y-axis shows predicted monthly number of typhoid cases and x-axis shows year. Blue curve indicates period with no actual collected data. ETS(MNM), exponential smoothing with multiplicative seasonality.

Table 3
Comparison among three time series prediction models to that of monthly collected typhoid incidences in Chandigarh region, India 2014-2017.

Model	Root mean square error	Mean absolute error
Autoregressive integrated moving average	16.50	10.91
Support vector machine regression	17.39	11.42
Exponential smoothing	13.30	7.91

(Wang *et al*, 2012) and this could impact upon the performance of a particular test model.

In summary, this study finds among three time series prediction models, namely, autoregressive integrated moving

Table 4
Predicted total number of typhoid cases in Chandigarh region, India from January to October 2018 from three time series prediction models.

Model	Predicted typhoid cases
Exponential smoothing	339
Autoregressive integrated moving average	361
Support vector machine regression	237

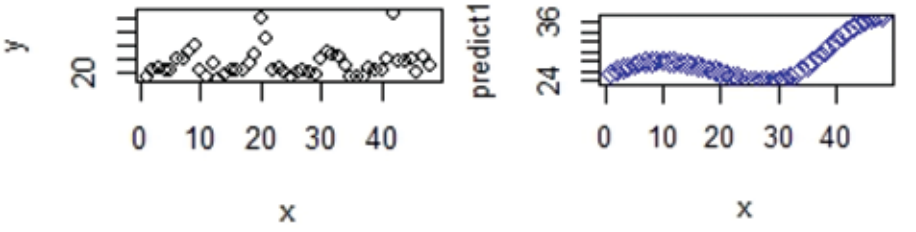

Fig 5-Prediction from support vector machine regression model of monthly typhoid cases in Chandigarh region, India. Y-axis shows the predicted values and x-axis shows the total time period of data collected. Left panel, actual data; right panel, predicted values.

Table 5
Research gap analysis.

Model	Capability	Limitation
Autoregressive integrated moving average	Very popular method Suitable for all type of time series data	If there are insufficient data, prediction is not as accurate as exponential smoothing Requires a minimum of 30-50 observations
Support vector machine regression	Provides optimal solution Allows mapping	When data set is large, requires a large number of tuning parameters Tuning parameters increases complexity
Exponential smoothing	Produces an accurate prediction Gives more significance to recent observations	Cannot handle trend well Produces predictions that lag behind actual trend

average, exponential smoothing and support vector machine regression, exponential smoothing model produced the most accurate predicted monthly typhoid cases in Chandigarh region from 2014 to 2017 compared to actual collected data. However, it is worth bearing in mind the advantages and limitations of

the three time series models (Table 5), and the optimal prediction model should be evaluated on a case-by-case basis. The limitations of this study was the limited time duration of data collection. Future studies should be able to correct this problem. The implication of this research on public health policy is to highlight the importance of such models in predicting the number of future typhoid cases, which should assist in informing the general public regarding the particular season of the year when special caution and preventive measures should be undertaken against contract typhoid. Furthermore, this study should stimulate other developing countries to carry out tests on the most appropriate time series prediction model that allows alertness beforehand of impending diseases so that the required medical facilities will be in place.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support from Integrated Diseases Surveillance Programme, Chandigarh for providing the database, and thank Dr G Dewan, Director of Health and Services, Chandigarh for giving approval for collection of data of Chandigarh region.

REFERENCES

Agrawal RK, Adhikari R. An introductory study on time series modeling and forecasting. *arXiv Preprint arXiv:1302.6613*, *1302.6613* 2013: 1-68.

Black RE, Cisneros L, Levine MM, Banfi A, Lobos H, Rodriguez H. Case-control study to identify risk factors for pediatric endemic typhoid fever in Santiago, Chile. *Bull World Health Organ* 1985; 5: 899-904.

Dewan AM, Corner R, Hashizume M, Ongee, ET. Typhoid fever and its association with environmental factors in the Dhaka metropolitan area of Bangladesh: a spatial and time-series approach. *PLOS Negl Trop Dis* 2013; 7: 24998.

Kanungo S, Dutta S, Sur D. Epidemiology of typhoid and paratyphoid fever in India. *J Infect Dev Ctries* 2008; 2: 454-60.

Karkey A, Arjyal A, Anders KL, *et al*. The burden and characteristics of enteric fever at a healthcare facility in a densely populated area of Kathmandu. *PLOS One* 2010; 5(11): e13988.

Kelly-Hope LA, AlonsoWJ, Thiem VD, *et al*. Geographical distribution and risk factors associated with enteric diseases in Vietnam. *Am J Trop Med Hyg* 2007; 4: 706-12.

Kothari A, Pruthi A, Chugh TD. The burden of enteric fever. *J Infect Dev Ctries* 2008; 4: 253-9.

Kumar VP, Rao VUM, Sarath Chandran MA, Subba Rao AVM, Bapuji Rao. Annual Progress Report-AICRPAM, 2015.

Luby SP, Faizan MK, Fisher-Hoch SP, *et al*. Risk factors for typhoid fever in an endemic setting, Karachi, Pakistan. *Epidemiol Infect* 1998; 2: 129-38.

Mermin JH, Villar R, Carpenter J, *et al*. A massive epidemic of multidrug resistant typhoid fever in Tajikistan associated with consumption of municipal water. *J Infect Dis* 1999; 6:1416-22.

Pradesh A. Literate *vs* illiterate. Census of India 2011. [Cited 2018 May 19]. Available from: http://www.censusindia.gov.in/2011-prov-results/paper2-vol2/data_files/AP/Chapter_VI.pdf, 201; 2.

Ochiai RL, Acosta CJ, Danovaro-Holliday MC, *et al*. A study of typhoid fever in five Asian countries: disease burden and implications for controls. *Bull World Health Organ* 2008; 4: 260-8.

Ram PK, Naheed A, Brooks WA, *et al*. Risk factors for typhoid fever in a slum in Dhaka, Bangladesh. *Epidemiol Infect* 2007; 3: 458-65.

Sharma PK, Ramakrishnan R, Hutin Y, Manickam P, Gupte MD. Risk factors for typhoid in Darjeeling, West Bengal, India:

evidence for practical action. *Trop Med Int Health* 2009; 6: 696-702.

Sur D, Ali M, Von Seidlein L, *et al*. Comparisons of predictors for typhoid and paratyphoid fever in Kolkata, India. *BMC Public Health* 2007; *7*: 1-10.

Tran HH, Bjune G, Nguyen BM, Rottingen JA, Grais RF, Guerin PJ. Risk factors associated with typhoid fever in Son La province, northern Vietnam. *Trans R Soc Trop Med Hyg* 2005; 11: 819-26.

Velema JP, van Wijnen G, Bult P, van Naerssen T, Jota S. Typhoid fever in Ujung Pandang, Indonesia--high-risk groups and high-risk behaviours. *Trop Med Int Health* 1997; 11:1088-94.

Vollaard AM, Ali S, van Asten HAGH, *et al*. Risk factors for typhoid and paratyphoid fever in Jakarta, Indonesia. *JAMA.* 2004; 21: 2607-15.

Wang L-X, Li X, Fang L-Q, Wang D-C, Cao W-C, Kan B. Association between the incidence of typhoid and paratyphoid fever and meteorological variables in Guizhou, China. *Chin Med J* 2012; 125: 455-60.

Whitaker JA, Franco-paredes C, Rio C, Edupuganti S. Rethinking typhoid fever vaccines: implications for travelers and people living in highly endemic areas. *J Travel Med* 2009; 16: 46-52.

Zhang X, Hou F, Qiao Z, *et al*. Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014. *BMJ Open* 2016; 6:10.

Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X. Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PLOS One* 2013; 8:5.

Zhang X, Zhang T, Young AA, Li X. Applications and comparisons of four time series models in epidemiological surveillance data. *PLOS One*, 2014; 2:1-16.