BMJ Open

Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014

Xingyu Zhang,[1,2] Fengsu Hou,[3] Zhijiao Qiao,[1] Xiaosong Li,[1] Lijun Zhou,[4] Yuanyuan Liu,[1] Tao Zhang[1]

To cite: 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**:e011038. doi:10.1136/bmjopen-2016-011038

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-011038).

Received 5 January 2016
Revised 11 May 2016
Accepted 13 September 2016

For numbered affiliations see end of article.

Correspondence to
Dr Tao Zhang;
scdxzhangtao@163.com and
Dr Xiaosong Li;
lixiaosong1101@126.com

ABSTRACT

Objectives: Time series models are effective tools for disease forecasting. This study aims to explore the time series behaviour of 11 notifiable diseases in China and to predict their incidence through effective models.

Settings and participants: The Chinese Ministry of Health started to publish class C notifiable diseases in 2009. The monthly reported case time series of 11 infectious diseases from the surveillance system between 2009 and 2014 was collected.

Methods: We performed a descriptive and a time series study using the surveillance data. Decomposition methods were used to explore (1) their seasonality expressed in the form of seasonal indices and (2) their long-term trend in the form of a linear regression model. Autoregressive integrated moving average (ARIMA) models have been established for each disease.

Results: The number of cases and deaths caused by hand, foot and mouth disease ranks number 1 among the detected diseases. It occurred most often in May and July and increased, on average, by 0.14126/100 000 per month. The remaining incidence models show good fit except the influenza and hydatid disease models. Both the hydatid disease and influenza series become white noise after differencing, so no available ARIMA model can be fitted for these two diseases.

Conclusion: Time series analysis of effective surveillance time series is useful for better understanding the occurrence of the 11 types of infectious disease.

Strengths and limitations of this study

- The incidence of 11 notifiable infectious diseases in China from 2009 to 2014 was analysed.
- Decomposition methods were used to explore (1) their seasonality expressed in the form of seasonal indices and (2) their long-term trend in the form of a linear regression model.
- Except for autoregressive integrated moving average (ARIMA) models for influenza and hydatid disease, the incidence models show good fit.
- We could only obtain class C notifiable disease incidence over a period of 6 years because the Chinese Ministry of Health only started to publish these data from 2009. The relatively short length of the series may affect the forecasting efficacy of the time series modelling.

BACKGROUND

Infection surveillance in China has improved since 2003, with a web-based infection surveillance system replacing the previous system over 10 years ago, covering the largest population in the world.[1 2] This web-based surveillance system can report cases of infection, and more infections than previously, in a temporal fashion.[3] It potentially saves lives and maintains the health of the whole population. The quality of the surveillance has greatly improved, with the average omission rate decreased to 13%.[4] This surveillance system currently monitors 39 notifiable infectious diseases, which can be divided into three classes.[5 6] Class A includes plague and cholera, which can cause large epidemics in a very short time.[7] There are no reports on time series of class A notifiable diseases, as only a few cases have been reported over several decades. Class B includes infectious diseases that might cause epidemics such as tuberculosis, syphilis and viral hepatitis.[8] We reported the incidence of class B notifiable diseases in our previous study,[7] with sexual diseases, viral hepatitis and tuberculosis being population health challenges.[7] Class C includes less severe and less infectious diseases such as hand, foot and mouth disease (HFMD), diarrhoea and influenza.

Various methods have been explored for modelling infection surveillance data over the last few decades, and time series models are commonly used.[9 10] Decomposition is a typical time series method, aiming to decompose the infection series into seasonal and long-trend patterns.[9] This method has been used to analyse the seasonality and secular

trend of class B notifiable infectious diseases in China.[7 9] Autoregressive integrated moving average (ARIMA) models are one of the most widely used infection time series models and have been used to fit tuberculosis,[11] typhoid fever,[12] gonorrhoea[13] and hepatitis.[14] ARIMA is composed of a differencing process and an autoregressive and moving average (ARMA) model,[15 16] which views the infection rate at time t as a linear combination of its previous values and the residuals.

The Chinese Ministry of Health has been reporting class C notifiable diseases to the public since 2009. Systemic time series analyses targeting class C notifiable diseases are greatly needed. Therefore, we performed a time series study on the monthly time series data of 11 class C infectious diseases using the decomposition method and the ARIMA model. We described the data's seasonality and long-term trend and established a time series model.

DATA AND METHODS

We collected the available time series data on 11 class C infectious diseases which were reported monthly by the Chinese Center for Disease Prevention and Control from 2009 to 2014. The 11 diseases were HFMD, diarrhoea, influenza, mumps, leprosy, rubella, kala-azar, hydatid disease, typhus disease, conjunctivitis and filariasis. The data were analysed using the decomposition method and the ARIMA model. All analyses were performed using SAS V.9.3.

Decomposition method

The decomposition method was introduced in previous studies.[9] The method breaks the time series into seasonal indices and long-term trend. Let x_{ik} denote the incidence in the k-th month of the i-th year. Then the seasonal index can be calculated in three steps.

1. Calculate the average value in each period

$$\overline{X}_k = \frac{\sum_{i=1}^{k} X_{ik}}{n}, k = 1, 2, \ldots, m.$$

where n is the number of the time points

2. Calculate the overall average value

$$\overline{X} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} x_{ik}}{nm}.$$

3. Calculate the seasonal index

$$S_k = \frac{\overline{X}_k}{\overline{X}}, k = 1, 2, \ldots, m.$$

The 'deseasonalised' series becomes: $SR = x_{ik} - S_k$.

After the seasonality is removed, a simple linear regression model is established between the deseasonalised incidence and time t, which can be presented in the following formula:

$$SR = \alpha + \beta \cdot t + \varepsilon.$$

The coefficient, constant, R^2 (coefficient of determination) and p values for the regression model are estimated. Changes in incidence can be derived, on average, by month from the coefficient of the regression.

ARIMA model

The ARIMA model is widely used in infectious disease time series modelling. As described in previous studies,[9 12] the model can be formed as ARIMA (p, d, q)×(P, D, Q)s, which can be expressed in the following formula:

$$\Phi(B)U(B^S)\nabla^d\nabla_S^D x_t = V(B^S)\Theta(B)\varepsilon_t,$$

where $\nabla^d = (1-B)^d, \Phi(B) = 1 - \phi_1 B - \cdots - \phi_P B^P, \Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q, \nabla_S^D = (1-B)^{SD}, U(B^S) = 1 - \mu_1 B^S - \cdots - \mu_P B^{PS}, V(B^S) = 1 - \nu_1 B^S - \cdots - \nu_Q B^{QS}$, where B is the backward operator, with $\Phi(B)$, $\Theta(B)$, $U(B^S)$ and $V(B^S)$ being lag polynomials. p and q are non-negative integers that refer to the order of the ARMA parts of the model, respectively, while P and Q represent the order of the seasonal ARMA, respectively. 'd' is the level of integration of the series, 'DS' is the level of seasonal integration, and 'S' is the order of seasonality.

The ARIMA modelling procedure consists of three iterative steps: identification, estimation and diagnostic checking.[17] Several ARIMA models may be identified, and the selection of an optimum model is usually based on the minimum Akaike information criterion (AIC) and Schwartz Bayesian criterion (SBC).[18]

Here, the ARIMA models were established from 2009 to 2013, to test the accuracy for the values of 2014. Several ARIMA models were fitted, and the final ARIMA model was selected on the basis of the minimum AIC and SBC scores for each disease. The mean absolute percentage error (MAPE) and mean square error (MSE) for the forecasting data (2014) were also calculated using the final ARIMA model.

RESULTS

First, the general descriptive analysis is presented, followed by the decomposition method and the ARIMA model results.

Descriptive analysis

Table 1 shows the number of cases and deaths caused by the 11 class C notifiable diseases from 2009 to 2014. The incidence time series of the disease is shown in figure 1. In total, 20 139 572 cases and 3453 deaths were detected in the surveillance system during the 6 years. HFMD ranks first in terms of both reported cases and deaths (figure 2). The proportion of HFMD cases increased from 48% to 68% from 2009 to 2014, and the proportion of deaths was over 80% each year. The number of diarrhoea cases increased from 2009 to 2013, but fell in 2014. Similarly, mumps cases increased from 2010 to 2012, and fell in 2013 and 2014. The

Table 1 Number of cases of class C notifiable diseases in China (ranked by total numbers)

Disease	2009	2010	2011	2012	2013	2014	All
Cases							
HFMD	1 164 784 (48.05)	1 795 336 (55.12)	1 638 743 (52.62)	2 198 442 (58.08)	1 855 849 (54.60)	2 819 581 (67.77)	11 472 735 (56.97)
Diarrhoea	660 410 (27.24)	751 230 (23.07)	841 115 (27.01)	896 808 (23.69)	1 017 512 (29.94)	871 085 (20.94)	5 038 160 (25.02)
Mumps	301 906 (12.45)	299 397 (9.19)	458 232 (14.71)	485 450 (12.83)	332 349 (9.78)	189 469 (4.55)	2 066 803 (10.26)
Influenza	202 667 (8.36)	65 664 (2.02)	66 691 (2.14)	123 491 (3.26)	130 390 (3.84)	218 207 (5.25)	807 110 (4.01)
Conjunctivitis	13 808 (0.57)	292 369 (8.98)	34 570 (1.11)	32 530 (0.86)	36 578 (1.08)	41 741 (1.00)	451 596 (2.24)
Rubella	72 707 (3.00)	44 490 (1.37)	67 887 (2.18)	41 507 (1.10)	18 571 (0.55)	13 305 (0.32)	258 467 (1.28)
Hydatid	3309 (0.14)	4738 (0.15)	3421 (0.11)	3591 (0.09)	4261 (0.13)	4017 (0.10)	23 337 (0.12)
Typhus	2815 (0.12)	2264 (0.07)	2393 (0.08)	2119 (0.06)	2021 (0.06)	1703 (0.04)	13 315 (0.07)
Leprosy	1133 (0.05)	1049 (0.03)	912 (0.03)	975 (0.03)	1113 (0.03)	837 (0.02)	6019 (0.03)
Kala-azar	538 (0.02)	428 (0.01)	346 (0.01)	240 (0.01)	174 (0.01)	301 (0.01)	2027 (0.01)
Filariasis	0 (0.00)	0 (0.00)	1 (0.00)	0 (0.00)	1 (0.00)	1 (0.00)	3 (0.00)
All	2 424 077	3 256 965	3 114 311	3 785 153	3 398 819	4 160 247	20 139 572
Deaths							
HFMD	355 (83.73)	888 (92.31)	506 (91.67)	569 (93.74)	260 (81.50)	508 (86.25)	3086 (89.37)
Diarrhoea	49 (11.56)	52 (5.41)	34 (6.16)	21 (3.46)	31 (9.72)	29 (4.92)	216 (6.26)
Influenza	17 (4.01)	9 (0.94)	3 (0.54)	11 (1.81)	18 (5.64)	46 (7.81)	104 (3.01)
Mumps	2 (0.47)	5 (0.52)	4 (0.72)	1 (0.16)	4 (1.25)	4 (0.68)	20 (0.58)
Leprosy	1 (0.24)	3 (0.31)	1 (0.18)	2 (0.33)	4 (1.25)	0 (0.00)	11 (0.32)
Rubella	0 (0.00)	1 (0.10)	2 (0.36)	1 (0.16)	0 (0.00)	0 (0.00)	4 (0.12)
Kala-azar	0 (0.00)	1 (0.10)	0 (0.00)	1 (0.16)	1 (0.31)	1 (0.17)	4 (0.12)
Hydatid	0 (0.00)	2 (0.21)	0 (0.00)	1 (0.16)	1 (0.31)	0 (0.00)	4 (0.12)
Typhus	0 (0.00)	1 (0.10)	1 (0.18)	0 (0.00)	0 (0.00)	1 (0.17)	3 (0.09)
Conjunctivitis	0 (0.00)	0 (0.00)	1 (0.18)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.03)
Filariasis	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
All	424	962	552	607	319	589	3453

The number in parentheses represents the percentage of cases in the particular year.
HFMD, hand, foot and mouth disease.

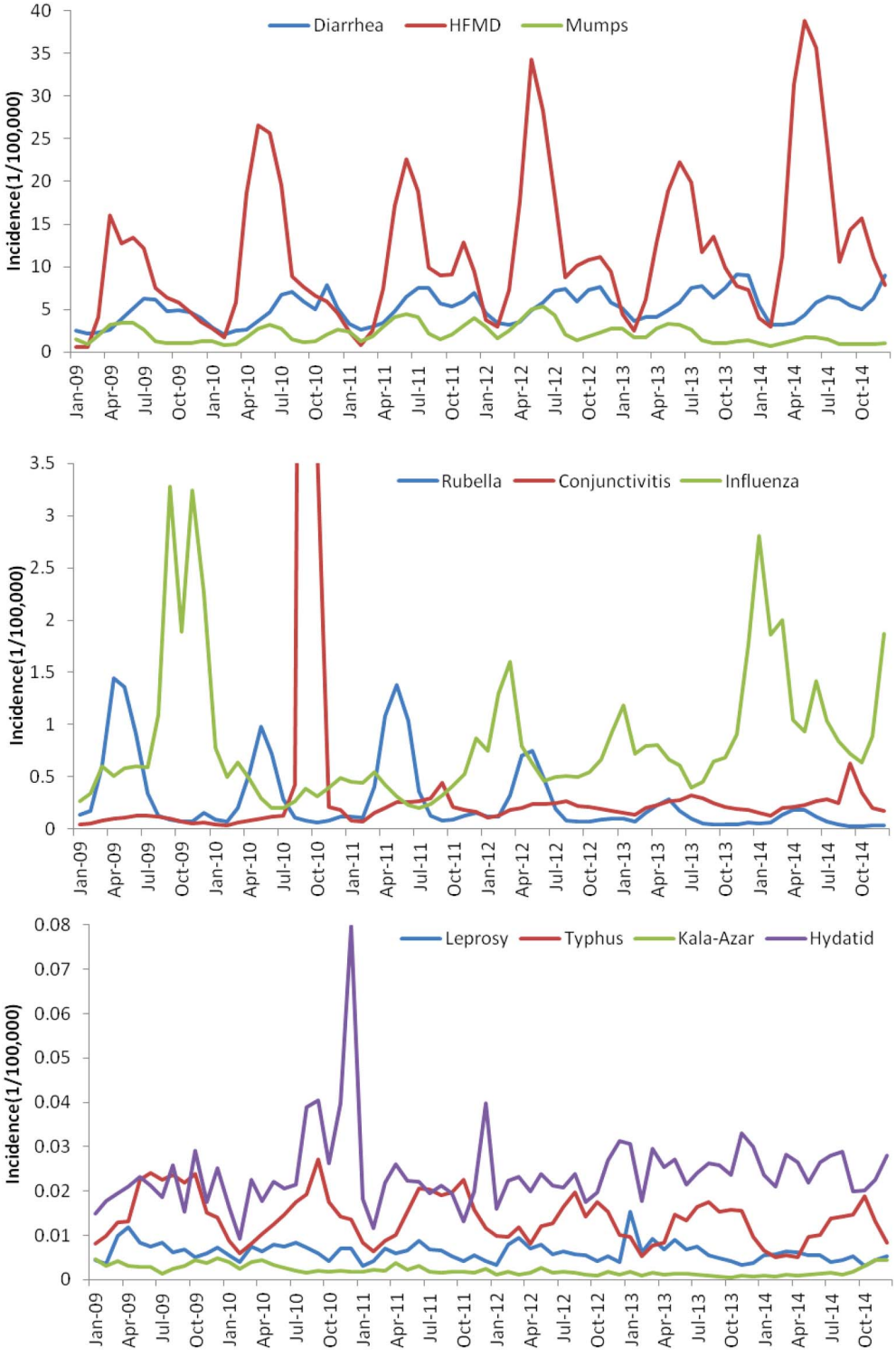

Figure 1 Incidence of the 11 types of class C notifiable disease. HFMD, hand, foot and mouth disease.

number of influenza cases fell from 2009 to 2010, and increased from 2010 to 2014. However, there was a conjunctivitis outbreak in September and October 2010. The incidences for these 2 months were deemed to be outliers, and they were thus replaced by the mean incidence (0.2703/100 000) of September 2009 and September 2010, and the mean incidence (0.1398/100 000) of October 2009 and October 2010. The number of rubella cases, hydatid cases and leprosy cases fluctuated each year. The number of typhus cases fell from 2009 to 2014. The number of kala-azar cases increased from 2009 to 2013 and only fell in 2014. There are only three random filariasis cases reported in the six years.

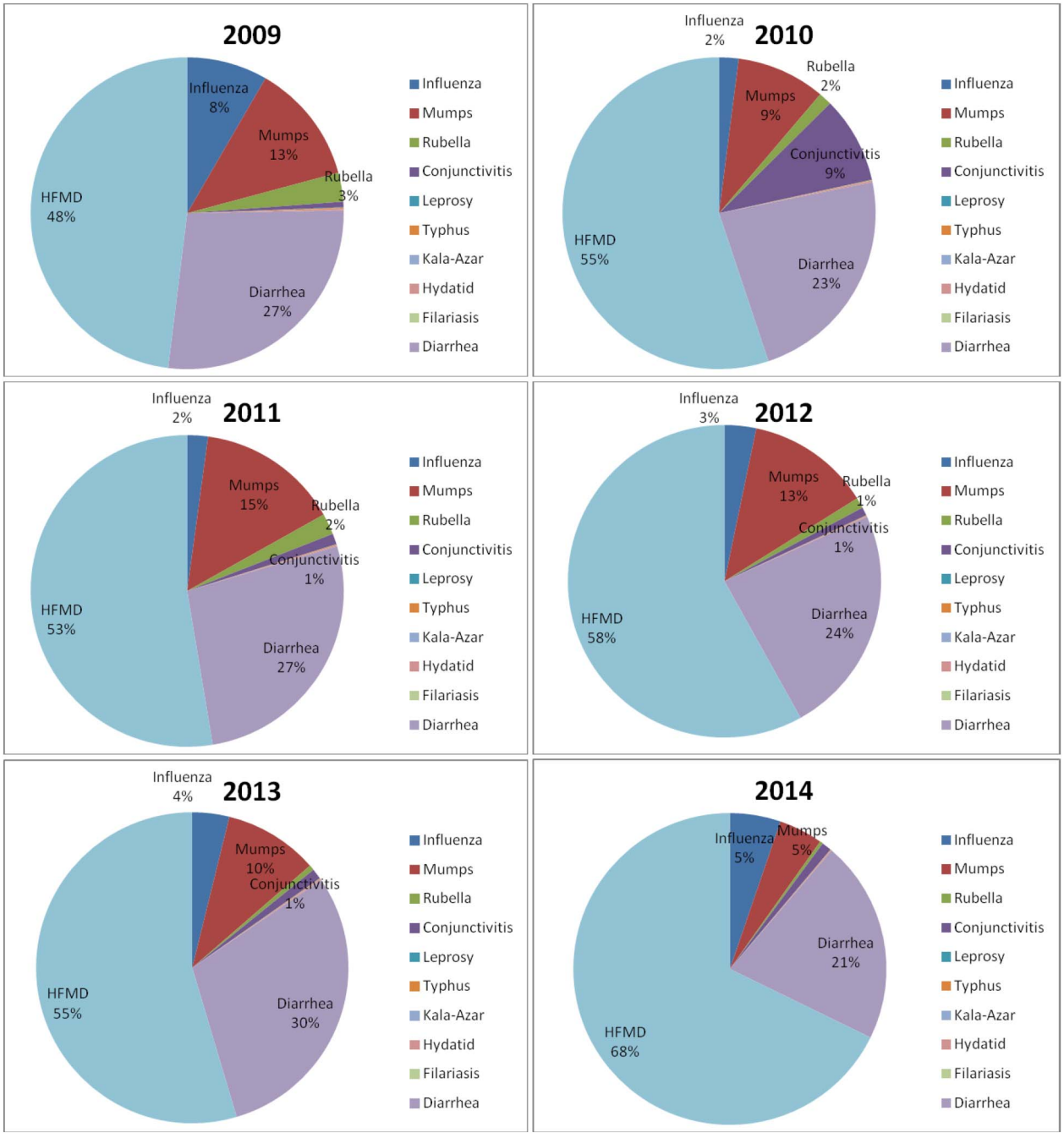

Figure 2 Change in proportion of cases of the different diseases from 2009 to 2014. HFMD, hand, foot and mouth disease.

Decomposition

Table 2 and figure 3 present the seasonal indices for each disease. The ranges of seasonal indices of rubella, HFMD, conjunctivitis, mumps and influenza were >1. In general, the occurrence of each disease was greatest during specific months as follows: rubella, April to June (peaked in May); HFMD, April to July (peaked in May); conjunctivitis, July to September (peaked in September). Typhus and hydatid disease did not peak during a specific month but occurred most often from August to October and December, respectively. Diarrhoea and leprosy on the other hand only occurred/peaked in August and May, respectively. There was no fixed seasonality for the incidence of influenza: it occurred most often from September to January (autumn and winter in China) in 2009, whereas from 2011 to 2014 it occurred most often from December to April (winter and spring). There was no obvious seasonality for kala-azar disease.

Estimations of the coefficient, constant, R^2 and p values for the regression model are shown in table 3. The regression models for influenza, mumps and hydatid disease showed no significance ($p>0.05$), and R^2 for the leprosy model was low ($R^2=0.055$). Of the class C notifiable diseases, HFMD incidence increased most

Table 2 Seasonal index of each type of class C infectious disease

Disease	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Range	STD
HFMD	0.25	0.16	0.53	1.46	2.09	2.08	1.59	0.81	0.86	0.82	0.75	0.59	1.93	0.63
Diarrhoea	0.77	0.55	0.59	0.63	0.85	1.09	1.34	1.35	1.10	1.12	1.33	1.28	0.80	0.29
Mumps	0.94	0.55	0.79	1.23	1.60	1.68	1.39	0.75	0.56	0.64	0.84	1.03	1.12	0.37
Influenza	1.25	1.03	1.24	0.81	0.68	0.70	0.58	0.68	1.17	0.9	1.32	1.63	1.05	0.31
Conjunctivitis	0.52	0.49	0.78	0.90	1.05	1.13	1.21	1.45	1.68	1.05	0.91	0.83	1.19	0.33
Rubella	0.38	0.38	1.13	2.59	3.06	2.13	0.84	0.34	0.25	0.23	0.28	0.39	2.83	0.97
Hydatid	0.83	0.69	1.00	0.94	0.97	0.92	0.91	1.14	0.96	0.91	1.11	1.62	0.93	0.22
Typhus	0.62	0.51	0.66	0.67	1.05	1.16	1.29	1.37	1.37	1.40	1.08	0.82	0.89	0.32
Leprosy	1.00	0.85	1.32	1.19	1.21	1.12	1.10	0.95	0.90	0.67	0.85	0.84	0.66	0.18
Kala-azar	1.12	0.87	1.10	1.17	1.05	1.02	0.74	0.73	0.85	0.98	1.15	1.21	0.48	0.16

HFMD, hand foot and mouth disease.

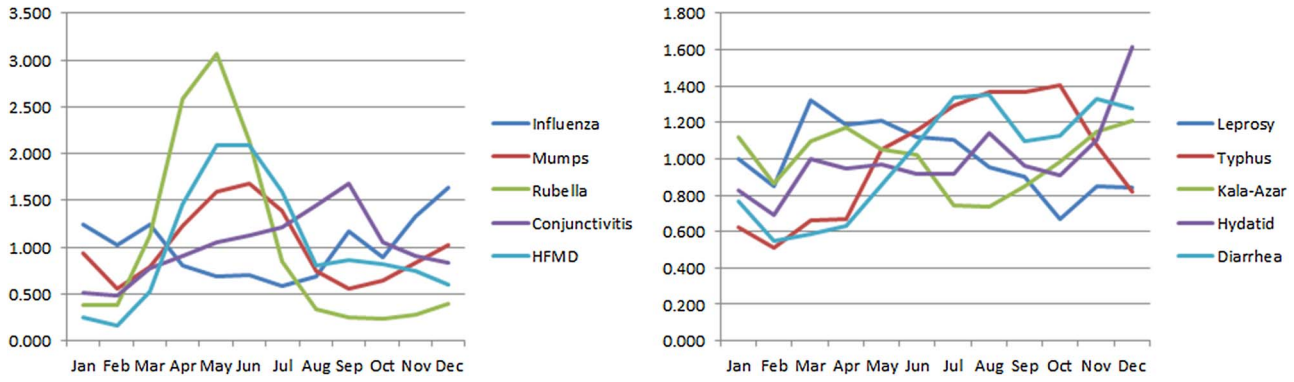

Figure 3 Seasonal index of each type of infectious disease. HFMD, hand, foot and mouth disease.

Table 3 Regression results of each series with seasonality removed

Disease	Constant	Constant 95% CI		Coefficient	Coefficient 95% CI		p Value	R^2
HFMD	6.67593	5.25891	8.09296	0.14126	0.10752	0.17499	<0.001	0.49905
Diarrhoea	4.17851	3.8286	4.52841	0.02801	0.01968	0.03634	<0.001	0.3912
Mumps	2.34317	2.00014	2.68619	−0.00566	−0.01382	0.00251	0.171	0.02654
Influenza	0.64218	0.37914	0.90522	0.00523	−0.00103	0.01149	0.101	0.03815
Conjunctivitis	0.09454	0.07167	0.11741	0.00258	0.00204	0.00313	<0.001	0.56109
Rubella	0.43629	0.39905	0.47353	−0.00461	−0.0055	−0.00372	<0.001	0.6056
Hydatid	0.02253	0.0196	0.02547	0.00004	−0.00003	0.00011	0.22	0.02124
Typhus	0.01745	0.01638	0.01851	−0.0001	−0.00013	−0.00008	<0.001	0.47228
Leprosy	0.00694	0.00613	0.00775	−0.00002	−0.00004	0.00000	0.046	0.0556
Kala-azar	0.0032	0.00281	0.00358	−0.00003	−0.00004	−0.00002	<0.001	0.37894

HFMD, hand, foot and mouth disease.

rapidly, by an average of 0.14126/100 000 per month. The incidence of rubella, typhus and kala-azar decreased after removal of seasonality.

ARIMA model

The results of the ARIMA estimations, MAPE and MSE for each disease time series are shown in table 4. The final selected ARIMA model (based on minimum AIC and SBC scores) is shown in bold font. The fitting and forecasting performance of each model is shown in figure 4. The incidence of hydatid disease and influenza after differencing was random (white noise), so no available ARIMA model could be fitted for these two diseases. The other disease series were well fitted. MAPEs for conjunctivitis, typhus and HFMD were as expected (0.1638, 0.1819 and 0.3497, respectively). Those for mumps, rubella, kala-azar and diarrhoea were slightly high (0.5417, 0.6948, 0.6837 and 0.6838, respectively).

DISCUSSION

Infection surveillance is important in infectious disease management and prevention. In this paper, we use the infection surveillance data to show the infection

Table 4 ARIMA models for the infectious diseases

Disease	Identification	AIC	BIC	MAPE	MSE
HFMD	ARIMA $(0,0,0)\times(1,1,0)_{12}$	260.17	262.02	0.3497	8.8720
	ARIMA $(0,0,0)\times(1,1,0)_{12}$	260.81	264.51		
	ARIMA $(0,0,0)\times(1,1,0)_{12}$	267.87	269.72		
	ARIMA $(0,0,0)\times(1,1,0)_{12}$	264.79	266.65		
Diarrhoea	**ARIMA $(1,0,1)\times(0,1,1)_{12}$**	113.73	119.28	0.6838	3.4459
	ARIMA $(1,0,0)\times(0,1,0)_{12}$	137.59	139.44		
	ARIMA $(0,0,1)\times(0,1,0)_{12}$	128.35	130.20		
	ARIMA $(1,0,0)\times(1,1,0)_{12}$	120.50	124.21		
	ARIMA $(0,0,1)\times(1,1,0)_{12}$	117.07	120.79		
Mumps	**ARIMA $(0,0,1)\times(0,1,1)_{12}$**	54.13	57.83	0.5417	0.6300
	ARIMA $(0,0,1)\times(0,1,1)_{12}$	64.16	66.01		
	ARIMA $(0,0,2)\times(0,1,0)_{12}$	58.00	60.00		
	ARIMA $(0,0,2)\times(0,1,1)_{12}$	53.23	56.93		
Influenza	**Become white noise after differencing**				
Conjunctivitis	**ARIMA $(0,0,1)\times(0,1,1)_{12}$**	−120.65	−116.95	0.1638	0.1091
	ARIMA $(0,0,1)\times(0,1,0)_{12}$	−103.13	−101.28		
	ARIMA $(1,0,0)\times(0,1,0)_{12}$	−94.49	−92.64		
	ARIMA $(1,0,0)\times(1,1,0)_{12}$	−102.63	−98.93		
	ARIMA $(2,0,0)\times(0,1,0)_{12}$	−99.17	−95.47		
Rubella	**ARIMA $(0,0,1)\times(0,1,1)_{12}$**	−34.53	−29.68	0.6948	0.1215
	ARIMA $(0,0,1)\times(0,1,0)_{12}$	−31.54	−29.68		
	ARIMA $(0,0,1)\times(1,1,0)_{12}$	−33.75	−30.05		
Hydatid	**Become white noise after differencing**				
Typhus	**ARIMA $(0,0,1)\times(0,1,1)_{12}$**	−401.92	−398.22	0.1819	0.0021
	ARIMA $(1,0,0)\times(0,1,0)_{12}$	−384.83	−382.83		
	ARIMA $(0,0,1)\times(0,1,0)_{12}$	−391.05	−389.20		
	ARIMA $(1,0,0)\times(1,1,0)_{12}$	−393.26	−389.57		
	ARIMA $(1,0,0)\times(0,1,1)_{12}$	−397.08	−393.38		
	ARIMA $(0,0,1)\times(1,1,0)_{12}$	−400.224	−396.524		
Leprosy	**ARIMA $(0,0,1)\times(0,1,0)_{12}$**	−429.84	−427.99	0.4767	0.0035
	ARIMA $(1,0,0)\times(0,1,0)_{12}$	−414.98	−413.14		
Kala-azar	**ARIMA $(1,0,0)\times(0,1,1)_{12}$**	−531.63	−527.93	0.6837	0.0021
	ARIMA $(1,0,0)\times(0,1,0)_{12}$	−530.26	−528.40		
	ARIMA $(0,0,1)\times(0,1,0)_{12}$	−525.32	−523.47		
	ARIMA $(0,0,1)\times(0,1,1)_{12}$	−527.16	−523.46		

The final ARIMA model selected is highlighted in bold.
AIC, Akaike information criterion; ARIMA, autoregressive integrated moving average; HFMD, hand, foot and mouth disease; MAPE, mean absolute percentage error; MSE, mean square error.

characteristics of the 11 class C notifiable diseases in China. Of these diseases, HFMD is the most serious in terms of both incidence and death rate, which agrees with previous studies.[19 20] The disease is caused by enterovirus and coxsackievirus, which are very prevalent in children under the age of five. HFMD can cause herpes in the hands, feet and mouth, as well as other complications such as myocarditis, pulmonary oedema and aseptic meningoencephalitis.[21] Some severely affected patients may die because of the rapid progress of the disease. From 2009 to 2014, more than 11 million HFMD cases were detected leading to 3086 fatalities. HFMD appears to occur most often from April to July (peaking in May), and increased, on average, by 0.14126/100 000 per month with seasonality removed. Strategies for the control and prevention of HFMD include promoting healthcare education, improving hygiene conditions in hospitals and schools, and strengthening the control of cross-infection.[21]

Seasonal patterns are a major cornerstone in understanding subtle but drastic effects of climate change on disease dynamics.[7 22] From the present analysis of surveillance data on China's population, we conclude that rubella, HFMD and diarrhoea most frequently occur in summer, whereas conjunctivitis and typhus are most prevalent during summer and autumn, and hydatid disease incidence peaks in winter. There is no fixed seasonality for influenza incidence, and no obvious seasonality was detected for kala-azar.

When there is substantial heterogeneity among different years, then conclusions on seasonal patterns based solely on seasonal indices may not be reliable. This may be the case for conjunctivitis, hydatid and influenza disease, as outbreaks in some years may have not been

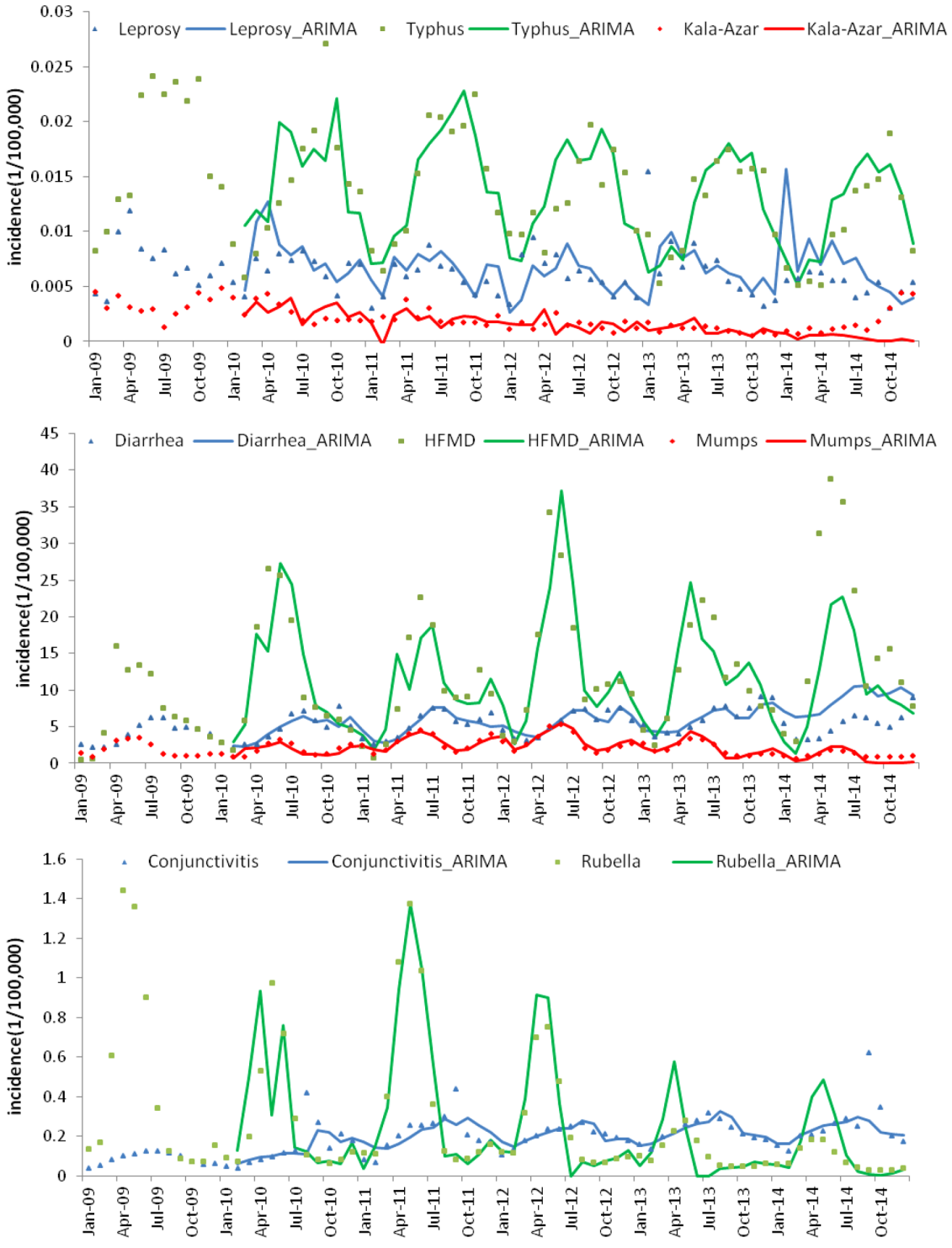

Figure 4 Incidence and fitting values predicted by autoregressive integrated moving average (ARIMA) models. HFMD, hand, foot and mouth disease.

related to seasonal effects. We calculated the seasonal indices for each year as a comparison (see online supplementary appendix figure A1) by using the incidences divided by the average incidence for the corresponding year. The seasonal patterns for influenza in 2009 were slightly different from other years. The incidence of influenza generally peaked in December, January and March, but in 2009 it peaked in September and November. The incidence of conjunctivitis peaked in August and September in 2010, 2011 and 2014, but peaked in June to August in the other years. Hydatid disease showed strong seasonality with consistent peaks in December for every year analysed.

The surveillance data used in this study covered different climate zones and different provinces with diversified urbanisation levels. The heterogeneity of

seasonality not only exists in different years, but also occurs in different geographic regions. To support the conclusion, we take influenza data as an example and calculate the regional seasonal indices using the monthly data (http://www.phsciencedata.cn/Share/en/index.jsp) of 31 Chinese provinces from 2009 to 2012 (see online supplementary appendix table A1). The seasonal patterns are slightly different among the different provinces.

The long-term patterns of the 11 types of class C infectious disease have also been shown with a linear regression model between the deseasonalised series and time t. The model shows that rubella, typhus and kala-azar decreased after removal of seasonality with the improvement of public health surveillance and management. The regression model is useful for understanding long-term epidemic trends, which can be applied to forecast future incidence, greatly facilitating management of public health resources such as vaccine preparation.[7]

An ARIMA model has been established for each disease. All of the incidence models show good fitting performance except those for influenza and hydatid disease. Influenza is a well-known typical infectious disease with seasonal trend[23] (range 1.05 in table 2); however, the ARIMA model cannot be applied to it. Possible reasons are the heterogeneity among the different epidemic periods and climate zones mentioned above. The incidence series of hydatid disease becomes white noise after differencing, suggesting that the occurrence of the disease is random without seasonal impact.

The forecasting accuracy is not ideal compared with some other diseases such as typhoid fever[12] and syphilis.[24] This may be because the relatively short length of the time series reduced the predictive power of the model. This is one of the limitations of the study. More time series data need to be collected for future exploration. Besides collecting more data, relevant secondary variables, such as average monthly temperatures, would provide information about the underlying reasons for seasonal patterns of outbreaks and may enhance future predictions. We collected the average monthly temperature time series data from the National Bureau of Statistics of China and calculated the time series correlation coefficients[24] between influenza incidence and average temperature. As this paper mainly focuses on univariate time series analysis, we have placed these results in online supplementary appendix figure A2 as support information. Certain correlations (0.33) were observed between the average temperature and the disease series. In a future study, we will collect data on more environment and weather variables to enhance the infection predictions.

Author affiliations
[1]Department of Epidemiology & Health Statistics, West China School of Public Health, Sichuan University, Chengdu, Sichuan, China
[2]Department of Anatomy with Radiology, University of Auckland, Auckland, New Zealand
[3]Sun Yat-sen Global Health Institute, School of Public Health, Sun Yat-sen University, Guangzhou, Guangdong, China
[4]Department of Respiratory, Sichuan Centre for Disease Control and Prevention, Sichuan, China

Acknowledgements XZ gratefully acknowledges financial support from the China Scholarship Council. We thank Susann Beier for careful proofreading of the manuscript.

Contributors XZ, TZ, LZ and FH conceived and designed the experiments. XZ and TZ collected the data and performed the statistical analysis. XZ, FH, ZQ, XL, LZ, YL and TZ participated in drafting the manuscript including data analysis and interpretation of results. All authors read and approved the final manuscript.

Funding TZ was supported by Sichuan University grant 'the Fundamental Research Funds for the Central Universities' (grant number 2016SCU11006) and the National Natural Science Foundation of China (grant no.81602935). The research is funded by the National Science and Technology Major Special Project 'Data mining and analysis of the surveillance data of five syndrome pathogens (grant number 2012ZX10004201-006)'. XZ was supported financially by the China Scholarship Council for his doctoral studies.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The dataset is available from the corresponding author at scdxzhangtao@163.com.

REFERENCES

1. Ortiz E, Clancy CM. Use of information technology to improve the quality of health care in the United States. *Health Serv Res* 2003;38:xi–xxii.
2. Wang L, Wang Y, Jin S, *et al*. Emergence and control of infectious diseases in China. *Lancet* 2008;372:1598–605.
3. Kass-Hout T, Zhang X. *Biosurveillance: methods and case studies*. CRC Press, 2010.
4. Qing G, Chun-yi Z, Yi-bing J, *et al*. Investigation of infectious disease direct reporting network management in Chinese medical institutions. *Disease Surveillance* 2010;25:410–13.
5. Zhang L, Wilson DP. Trends in notifiable infectious diseases in China: implications for surveillance and population health policy. *PLoS ONE* 2012;7:e31076.
6. Liang S, Yang C, Zhong B, *et al*. Surveillance systems for neglected tropical diseases: global lessons from China's evolving schistosomiasis reporting systems, 1949–2014. *Emerg Themes Epidemiol* 2014;11:19.
7. Zhang X, Hou F, Li X, *et al*. Study of surveillance data for class B notifiable disease in China from 2005 to 2014. *Int J Infect Dis* 2016;48:7–13.
8. Yang W, Li Z, Lan Y, *et al*. A nationwide web-based automated system for outbreak early detection and rapid response in China. *Western Pac Surveill Response J* 2011;2:10–15.
9. Zhang X, Zhang T, Young AA, *et al*. Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS ONE* 2014;9:e88075.
10. Nobre FF, Monteiro ABS, Telles PR, *et al*. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Stat Med* 2001;20:3051–69.
11. Rios M, Garcia JM, Sanchez JA, *et al*. A statistical analysis of the seasonality in pulmonary tuberculosis. *Eur J Epidemiol* 2000;16:483–8.
12. Zhang X, Liu Y, Yang M, *et al*. Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PLoS ONE* 2013;8:e63116.
13. Dowell D, Tian LH, Stover JA, *et al*. Changes in fluoroquinolone use for gonorrhea following publication of revised treatment guidelines. *Am J Public Health* 2012;102:148–55.

14. Ture M, Kurt I. Comparison of four different time series methods to forecast hepatitis A virus infection. *Expert Syst Appl* 2006;31:41–6.
15. Zaidi AA, Schnell DJ, Reynolds GH. Time series analysis of syphilis surveillance data. *Stat Med* 1989;8:353–62.
16. Pai P-F, Lin C-S. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 2005;33:497–505.
17. Ho SL, Xie M, Goh TN. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. *Comput Ind Eng* 2002;42:371–5.
18. Koehler AB, Murphree ES. A comparison of the Akaike and Schwarz criteria for selecting model order. *Appl Stat* 1988;37:187–95.
19. Liu Y, Wang X, Pang C, *et al*. Spatio-temporal analysis of the relationship between climate and hand, foot, and mouth disease in Shandong province, China, 2008–2012. *BMC Infect Dis* 2015;15:146.
20. Zhu Q, Hao Y, Ma J, *et al*. Surveillance of hand, foot, and mouth disease in mainland China (2008–2009). *Biomed Environ Sci* 2011;24:349–56.
21. Li Y, Zhang J, Zhang X. Modeling and preventive measures of hand, foot and mouth disease (HFMD) in China. *Int J Environ Res Public Health* 2014;11:3108.
22. Pascual M, Dobson A. Seasonal patterns of infectious diseases. *PLoS Med* 2005;2:e5.
23. Lofgren E, Fefferman NH, Naumov YN, *et al*. Influenza seasonality: underlying causes and modeling theories. *J Virol* 2007;81:5429–36.
24. Zhang X, Zhang T, Pei J, *et al*. Time series modelling of syphilis incidence in China from 2005 to 2012. *PLoS ONE* 2016;11: e0149401.