JIDC | THE JOURNAL OF INFECTION IN DEVELOPING COUNTRIES

Original Article

A new hybrid model SARIMA-ETS-SVR for seasonal influenza incidence prediction in mainland China

Daren Zhao[1], Ruihua Zhang[2,3]

[1] Department of Medical Administration, Sichuan Provincial Orthopedics Hospital, Chengdu, Sichuan, PR China
[2] School of Management, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, PR China
[3] General Practitioners Training Center of Sichuan Province, Chengdu, Sichuan, PR China

Abstract

Introduction: Seasonal influenza is a serious public health issue in China. This study aimed to develop a new hybrid model for seasonal influenza incidence prediction and provide reference information for early warning management before outbreaks.
Methodology: Data on the monthly incidence of seasonal influenza between 2004 and 2018 were obtained from the China Public Health Science Data Center website. A single seasonal autoregressive integrated moving average (SARIMA) model and a single error trend and seasonality (ETS) model were built. On this basis, we constructed SARIMA, ETS, and support vector regression (SARIMA-ETS-SVR) hybrid model. The prediction performance was determined by comparing mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE), and root mean square error (RMSE) indices.
Results: The optimum SARIMA model was SARIMA (0,1,0) (0,0,1)12. Error trend and seasonality (ETS) (M,A,M) was the SARIMA optimal model. For the fitting performance, the SARIMA-ETS-SVR hybrid model achieved the lowest values of MAE, MSE, and RMSE, in addition to the MAPE. In terms of predictive performance, the SARIMA-ETS-SVR hybrid model had the lowest MAE, MSE, MAPE, and RMSE values among the three models.
Conclusions: The study demonstrated that the SARIMA-ETS-SVR hybrid model provides better generalization ability than a single SARIMA model and a single ETS model, and the predictions will provide a useful tool for preventing this infectious disease.

Key words: SARIMA; ETS; SVR; influenza; infectious disease.

J Infect Dev Ctries 2023; 17(11):1581-1590. doi:10.3855/jidc.18037

(Received 06 February 2023 – Accepted 23 May 2023)

Introduction

Seasonal influenza, caused by the influenza virus, is an acute respiratory disease characterized by the sudden onset of fever, headache, cough, rhinitis, and muscle and joint pain [1,2]. Influenza can be classified into four types: influenza A virus, influenza B virus, influenza C virus, and influenza D virus, among which influenza virus types A and B circulate and cause seasonal influenza [3]. Seasonal influenza has also been shown to contribute significantly to global mortality [4]. The Global Burden of Disease Study (GBD) estimated that 99,000-200,000 deaths could be attributed to seasonal influenza worldwide, accounting for 0.26% of all deaths in 2017 [5]. Additionally, the World Health Organization (WHO) estimates that seasonal influenza leads to approximately 3-5 million cases of severe illness and 290,000-650,000 respiratory deaths annually [3]. Seasonal influenza poses a significant global economic burden. The average annual economic burden of this infectious disease on the healthcare system and society is $11.2 billion in the United States [6]. In Spain, the economic burden of seasonal influenza on primary care, hospitals, and treatment can reach €1 billion annually [7].

In China, seasonal influenza is classified as a Class C infectious disease. Currently, China faces enormous challenges in seasonal influenza control and prevention because of the increased morbidity and mortality associated with this infectious disease [8]. Previous studies have reported that the incidence of seasonal influenza in China increased from 3.51 per 100,000 population in 2005 to 55.09 per 100,000 population in 2018 [8]. The estimated mortality attributable to influenza is an annual average of 88,100 influenza-related deaths in China [9]. Seasonal influenza causes a tremendous disease burden, especially among influenza-associated outpatients, with an average of 2.5 excess influenza-like-illness consultations per 1000 person-years in 30 provinces of China each year between 2006 and 2015 [10]. A recent study revealed that there were 10,025 influenza-related deaths per year, accounting for 5.2% of all deaths in Chongqing

[11]. Therefore, it is crucial to control and prevent seasonal influenza outbreaks in China.

Time-series analysis, a scientific method of quantitative prediction, has been applied to historical data and time variables to predict future developments in infectious diseases [12]. Considerable efforts have been made to develop modeling approaches to explore and understand the regularity of the occurrence of infectious diseases and anticipate outbreaks [13]. Currently, various statistical methods, including traditional mathematical forecasting models and machine-learning-based forecasting models, have been extensively employed in infectious disease forecasting. As for traditional mathematical forecasting models, they include autoregressive integrated moving average (ARIMA) model [14], linear regression [15], grey model first-order one-variable (GM (1,1) model) [16] and exponential smoothing models [17], while machine learning-based forecasting models include artificial neural networks (ANN) [18], support vector regression (SVR) [19] and eXtreme gradient boosting (XGBoost) models [20].

In recent years, the emergence of hybrid methods has provided novel methods for predicting infectious diseases. It has been proven that hybrid methods combine the merits of different methods and may improve the forecast accuracy [21]. However, to date, no studies have been conducted on the use of a hybrid method to predict seasonal influenza epidemic trends in mainland China. In this study, we propose a new seasonal autoregressive integrated moving average (SARIMA), error trend and seasonality (ETS), and support vector regression (SARIMA-ETS-SVR) hybrid model to fit and predict the incidence of seasonal influenza from 2004 to 2018 in mainland China. This study aimed to provide reference information for early warning management and to implement adequate preventive measures before the outbreak of seasonal influenza in mainland China.

Methodology

Data source

Monthly influenza incidence data from 2004 to 2018 were obtained from the China Public Health Science Data Center website (https://www.phsciencedata.cn/Share/index.jsp, Supplementary File 1). The law of the Peoples Republic of China on the prevention and treatment of infectious diseases requires the inclusion of influenza in the management of category C infectious diseases. If a seasonal influenza case is diagnosed, clinicians must report to the national network reporting system within 24 hours at the local Center for Disease Control and Prevention. In this study, the number of monthly observations of influenza incidence was 180 and data from 2004 to 2018 were used from the database. Data from January 2004 to December 2017 were used to construct the models, and data from January to December 2018 were used to evaluate the predictive performance of each model.

SARIMA model

The ARIMA model is a classical time-series model for infectious disease forecasting [22]. Provided that the seasonality characteristics of the time series are constituted, the model can be recognized as a SARIMA model [23]. In general, the SARIMA model is expressed as SARIMA (p, d, q) (P, D, Q), and its mathematical formula is as follows:

$$(B)\Phi(B^s)(1-B^s)^D(1-B)^d X_t = \theta(B)\Theta(B^s)\varepsilon_t \quad (1)$$

$$\begin{cases} p = 1-\phi_1(B)-\phi_2(B)-\ldots-\phi_q(B)^q \\ d = (1-B)^d \\ q = 1-\theta_1(B)-\theta_2(B)-\ldots-\theta_q(B)^q \\ P = 1-\Phi_1(B^S)-\Phi_2(B^S)^2-\ldots-\Phi_P(B^S)^P \\ D = (1-B^S)^D \\ Q = 1-\Theta_1(B^S)-\Theta_2(B^S)^2-\ldots-\Theta_Q(B^S)^Q \end{cases} \quad (2)$$

where B and ε_t denote the backshift operator and residuals of the seasonal influenza time series, respectively, p is the order of auto-regression, d is the degree of trend difference, q is the order of moving average, P is the seasonal auto regression lag, D is the degree of seasonal difference, Q is the seasonal moving average, and s is the periodicity of the seasonal influenza time series (s = 12) [23].

Several steps are involved in establishing the SARIMA model [23,24]. First, plots of the original seasonal influenza time series or Augmented Dickey-Fuller (ADF) tests were performed to check whether the time series was stationary. If the original seasonal influenza time series is not stationary, differences are used to transform it into a stationary series. Second, the auto-correlation function (ACF) and partial auto-correlation function (PACF) graphics are plotted to verify the identification and estimation of the SARIMA model. Simultaneously, parameters p, q, P, and Q of the SARIMA model can be identified. Third, a Ljung-Box Q test was conducted to perform a white noise test on the seasonal influenza time-series residuals. The independent and normal distributions of seasonal influenza time-series residuals were checked by conducting a normal distribution standardized residual plot or histogram plot. Finally, the lowest values of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were considered for optimal SARIMA models.

ETS model

The ETS model (E, T, S) designates three components, error, trend, and seasonality, which can be combined into different additive or multiplicative combinations to produce the original series [25]. Generally, the ETS model includes three main categories: additive, multiplicative, and mixed models. For detailed analysis, the ETS model was classified into 30 methods [25], as shown in Table 1.

Additive models are expressed as:

$$Y = S + E \tag{3}$$

$$Y = T + S + E \tag{4}$$

Multiplicative models are expressed as:

$$Y = S \times E \tag{5}$$

$$Y = T \times S \times E \tag{6}$$

Mixture models are expressed as:

$$Y = (T \times S) + E \tag{7}$$

$$Y = (T + S) \times (1 + E) \tag{8}$$

The ETS model was built in the R software environment. The optimal ETS prediction model requires Akaike information criterion (AIC), corrected Akaike information criterion (AICc), or Bayesian information criterion (BIC) minima. The Ljung-Box Q-test residuals were also required to be white-noise sequences [26].

SVR model

Support vector regression (SVR) is a machine learning algorithm based on statistical theory and has been applied to regression estimation problems [27]. The basic idea of the SVR model is to train and learn all samples of the research data and distribute them between two straight lines, which requires the total deviation from all points to be the smallest [28]. After the maximum distance between the two lines is obtained, the optimal superposition of the support vector regression is explored. The mathematical formula is as follows [29]:

$$f(x) = w^T \varphi(x) + b \tag{9}$$

Where f(x) represents the prediction values, φ(x) represents the nonlinear mapping, and w and b represent the modifiable coefficients. R(C) is the penalty function, ε is the insensitive loss factor, and ξ_i and ξ_i^* are the relaxation variables.

$$R(C) = \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{10}$$

$$\text{s.t.} \quad f(x_i) - y_i \le \varepsilon + \xi_i \tag{11}$$

$$y_i - f(x_i) \le \varepsilon + \xi_i^* \tag{12}$$

$$\xi_i \ge 0, \xi_i^* \ge 0, i = 1,2,...m \tag{13}$$

By using Lagrange multiplier, the dual optimization problem can be expressed:

$$\max_{\alpha,\alpha^*} \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)y_i - \sum_{i=1}^{m}(\alpha_j^* + \alpha_j)\varepsilon - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) \tag{14}$$

Here, the α_i , α_i^* , α_j ,and α_j^* are Lagrange multipliers and K(x_i, x_j) is a kernel function. Finally, the SVR model formula is written as:

$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x_i, x_j) + b \tag{15}$$

SARIMA-ETS-SVR hybrid model

In this section, first, the SARIMA and ETS models were constructed respectively, and the predictive values from the SARIMA model(yi) and ETS model(yj) were

Table 1. Trend, seasonality and residuals for different combinations of ETS models.

Additive patterns	**Seasonal Component**		
	N (none)	**A (additive)**	**M (multiplicative)**
N (None)	N, A, N	N, A, A	N, A, M
A (Additive)	A, A, N	A, A, A	A, A, M
AD (Additive damped)	AD, A, N	AD, A, A	AD, A, M
M (multiplicative)	M, A, N	M, A, A	M, A, M
MD (multiplicative damped)	MD, A, N	MD, A, A	MD, A, M
Multiplicative patterns			
N (None)	N, M, N	N, M, A	N, M, M
A (Additive)	A, M, N	A, M, N	A, M, M
AD (Additive damped)	AD, M, N	AD, M, A	AD, N, M
M (multiplicative)	M, M, N	M, M, A	M, M, M
MD (multiplicative damped)	MD, M, N	MD, M, A	MD, M, M

obtained. Subsequently, the predictive values yi and yj were used as input variables, the observed values were used as output values to fit and construct the SVR model, and the SARIMA-ETS-SVR hybrid model and its predictive values were obtained.

Evaluation of prediction performance

In this study, mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE), and root mean square error (RMSE) values were calculated to assess the accuracy of the capability and prediction of each model. The formula can be expressed as [21]:

$$\mathrm{MAE}=\frac{\sum_{t=1}^{n}\left|X_t-\hat{X}_t\right|}{n} \tag{16}$$

$$\mathrm{MSE}=\frac{1}{n}\sqrt{\sum_{t=1}^{n}(X_t-\hat{X}_t)^2} \tag{17}$$

$$\mathrm{MAPE}=\frac{\sum_{t=1}^{n}\left|\frac{X_t-\hat{X}_t}{X_t}\right|\times 100\%}{n} \tag{18}$$

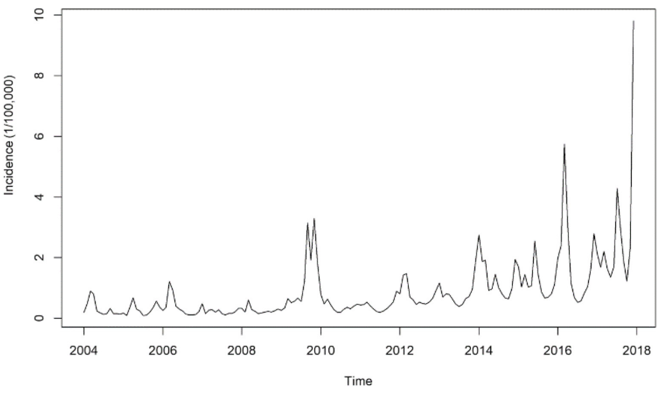
$$\mathrm{RMSE}=\sqrt{\frac{\sum_{t=1}^{n}(X_t-\hat{X}_t)^2}{n}} \tag{19}$$

where, $\hat{X}_t$ is the predicted value, X_t is the observed value, and n is the sequence sample size.

Data analysis

The R software version 4.1.1 was applied to construct the SARIMA, ETS, and SVR models, among which the "forecast," "zoo" and "tseries" packages were used in the construction of the SARIMA and ETS models, and the "e1071," "caret" and "tidyverse" packages were used to construct the SVR model. The level of significance was set at $p < 0.05$.

Figure 1. The original seasonal influenza time series between 2004 and 2018 in mainland China.



Results

SARIMA model

The original seasonal influenza time series from 2004 to 2018 in mainland China is shown in Figure 1. The monthly seasonal influenza time series showed a long-term fluctuating trend, indicating that it was not a stationary time series. As shown in Figure 2, the monthly seasonal influenza time series showed apparent seasonality, trends, periodicity, and randomness between 2004 and 2018 in mainland China. Therefore, a trend difference was carried out to eliminate the data instabilities. After a trend difference of the original seasonal monthly influenza time series (Figure 3A), the differenced time series became stationary (ADF test, t=-5.368, $p < 0.05$), and the parameters of d and D of the SARIMA model were 1 and 0, respectively.

For the SARIMA seasonal part, the ACF plot with the differenced time series showed a significant spike at lags 12 or 21 (Figure 3B), and the PACF plot with the differenced time series showed a significant spike at lag 12 (Figure 3C).

Figure 2. Decomposition of the original seasonal influenza time series between 2004 and 2018 in mainland China. **A:** Observed value plot; **B:** Trend plot; **C:** Seasonal plot; **D:** Random plot.

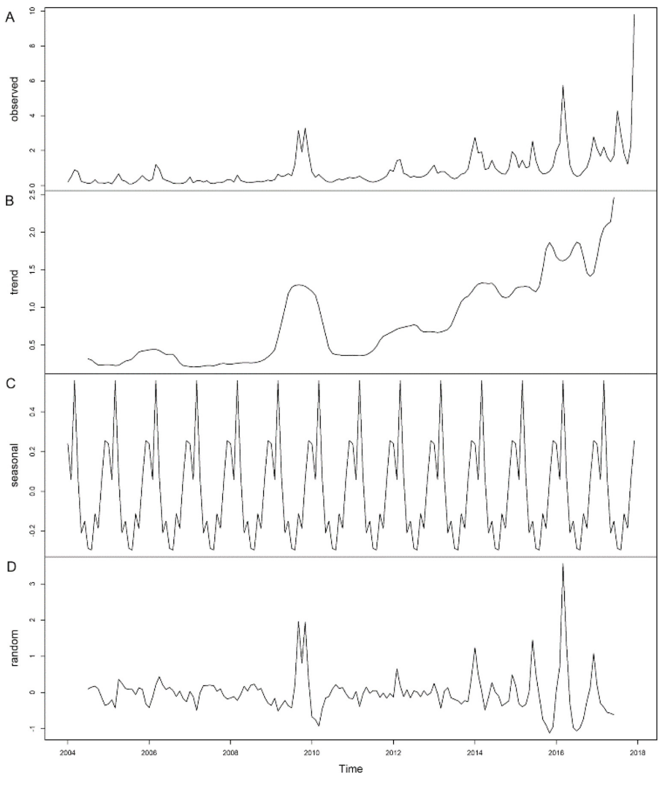


Figure 3. Time series plot of after a trend difference and the ACF and PACF plots of after a trend difference of the original seasonal monthly influenza. **A:** After a trend difference plot; **B:** ACF plot; **C:** PACF plot.

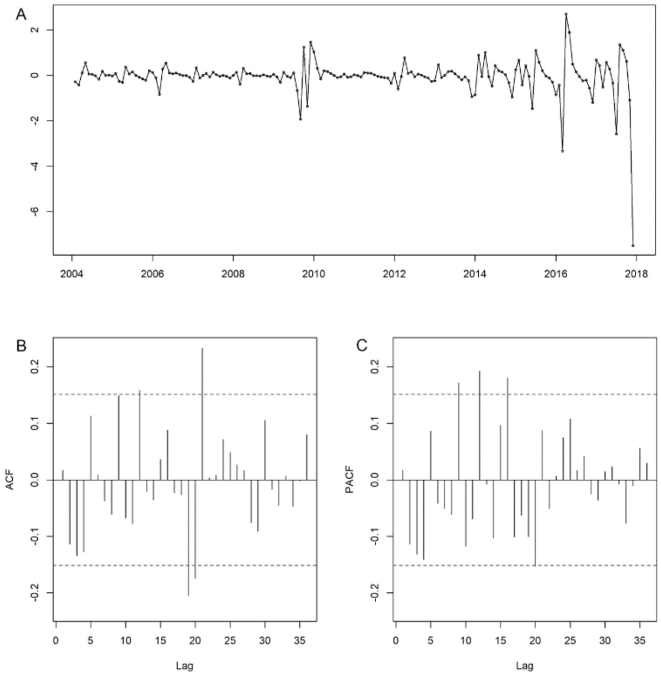


Figure 4. SARIMA (0,1,0) (0,0,1)12 model's residual. **A:** Standardized residuals plot; **B:** ACF of residuals plot; **C:** p values for Ljung-Box statistic.

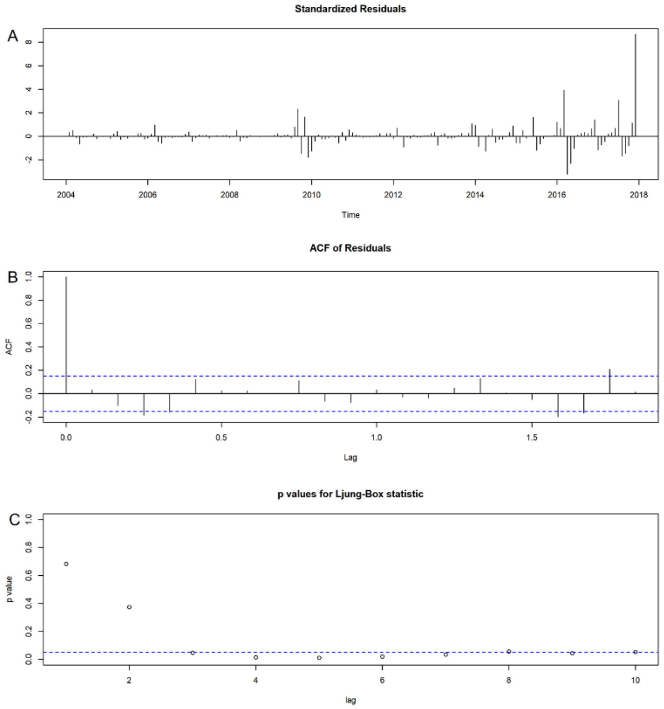


Table 2. The candidate SARIMA models and Ljung-Box Q test.

Candidate models	**AIC**	**BIC**	**LL**	**L-BQS**	***p***
SARIMA (0,1,0) (0,0,1)$_{12}$	414.21	420.446	-205.11	0.1674	0.6824
SARIMA (1,1,0) (1,0,0)$_{12}$	414.32	423.676	-204.16	0.0052	0.9424
SARIMA (1,1,0) (0,0,1)$_{12}$	415.90	425.251	-204.95	0.0047	0.9450
SARIMA (1,1,1) (0,0,1)$_{12}$	415.03	427.499	-203.51	0.0227	0.8801
SARIMA (0,1,0) (0,0,2)$_{12}$	415.08	424.439	-204.54	0.1827	0.6690
SARIMA (0,1,1) (0,0,2)$_{12}$	416.60	429.069	-204.30	0.0059	0.9386

SARIMA: seasonal autoregressive integrated moving average; AIC: Akaike information criterion; BIC: Bayesian Schwarz information criterion; LL: log likelihood; L-BQS: Ljung-Box Q statistics; *p:* p-values.

Table 3. The predictive values of three models 2018 in mainland China.

Date	**Observed value**	**SARIMA**		**ETS**		**SARIMA-ETS-SVR**	
		Forecasted value	**Absolute error**	**Forecasted value**	**Absolute error**	**Forecasted value**	**Absolute error**
January	19.4372	9.5542	9.8830	8.6710	10.7662	3.6476	15.7896
February	9.533	9.3930	0.1400	7.9802	1.5528	3.4313	6.1017
March	4.8001	9.2950	4.4949	11.7584	6.9583	4.6550	0.1451
April	1.9192	9.3368	7.4176	8.6507	6.7315	3.7544	1.8352
May	1.5839	9.4035	7.8196	4.9052	3.3213	1.6817	0.0978
June	1.1623	9.5598	8.3975	4.2836	3.1213	1.1679	0.0056
July	1.0131	10.2551	9.2420	4.4883	3.4752	0.8671	0.1460
August	0.8517	9.8746	9.0229	4.3950	3.5433	1.0488	0.1971
September	0.955	9.5512	8.5962	5.6280	4.6730	2.0563	1.1013
October	1.1024	9.3698	8.2674	4.8000	3.6976	1.6322	0.5298
November	2.0909	9.6279	7.5370	5.9985	3.9076	2.2367	0.1458
December	10.6363	11.5859	0.9496	10.4981	0.1382	3.1426	7.4937

SARIMA: seasonal autoregressive integrated moving average; ETS: error trend and seasonality; SVR: Support vector regression.

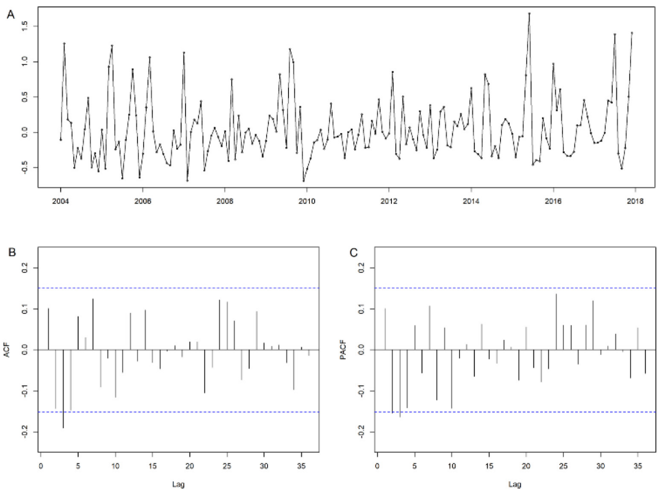
Therefore, parameter P was 0 or 1, and Q was 0, 1, or 2. For the SARIMA non-seasonal part of the first cycle, both the ACF and PACF plots with the differenced time series showed a significant spike at lags 9 or 12. Therefore, the parameters p and q are either 0 or 1. The candidate SARIMA models are listed in Table 2. These candidate SARIMA models residual all passed the Ljung-Box Q test, indicating that the residual series were white noise time series.

The optimum SARIMA model was SARIMA (0,1,0) (0,0,1) 12; its estimates of seasonal moving average at lag one (SMA1) were 0.2729, standard error (SE) was 0.0967, and it had the lowest values of AIC and BIC. As illustrated in Figure 4, the residual from SARIMA (0,1,0) (0,0,1) 12 passed the Ljung-Box Q-test (χ^2= 0.1674, test statistic p > 0.05). Finally, the SARIMA (0,1,0) (0,0,1) 12 model was used to predict seasonal influenza time series from January to December 2018 in mainland China (Table 3).

ETS model

The ets ( ) function in the R software forecast package was used to fit the ETS model. ETS (M, A, M) was the optimal model, with the lowest values of AIC (439.436), AICc (443.516), and BIC (492.543). The smoothing parameters alpha, beta, and gamma of the ETS (M,A,M) model are 0.999, 0.003, and 0.0004, respectively. Residual series from ETS (M, A, M) between 2004 and 2018 was shown in Figures 5A. As illustrated in Figures 5B and C, the residual series from ETS (M, A, M) of the ACF and PACF were all within their two standard error bounds, and the residual from ETS (M, A, M) passed the white noise test (Ljung-Box Q Statistics χ^2= 1.5121, p > 0.05). Based on the above residual test results, we believe that the residual series from ETS (M, A, M) is a white noise time series. Finally, the ETS (M, A, M) model was used to predict the seasonal influenza time series from January to December 2018 in mainland China (Table 3).

Figure 5. Time series plot of residual from ETS (M, A, M). **A:** Residual series from ETS (M, A, M) between 2004 and 2018; **B:** Residual series from ETS (M, A, M) of ACF plot; **C:** Residual series from ETS (M, A, M) of PACF plot.



SARIMA-ETS-SVR hybrid model

Owing to the trend difference in the original seasonal influenza time series, the 13-month values were lost in the SARIMA modeling process. Therefore, 155 observed values were considered as the database to construct the SARIMA-ETS-SVR hybrid model. Initially, SARIMA and ETS models were constructed, respectively. The predictive values from the SARIMA and ETS models were used as the input data, and the observed values were used as the output values to construct the SVR model. For the SVR modeling process, the grid search optimization method was used to determine the parameters C, γ, and ε. Subsequently, the optimal residual SVR model was selected using the function tune.svm ( ) of R software, and parameters C, γ, and ε were set to 100, 0.01, and 0.1, respectively. Finally, the SARIMA-ETS-SVR hybrid model and its predictive values were obtained. The SARIMA-ETS-SVR hybrid model was adopted to predict seasonal influenza time series from January to December 2018 in mainland China (Table 3).

Comparison of three models

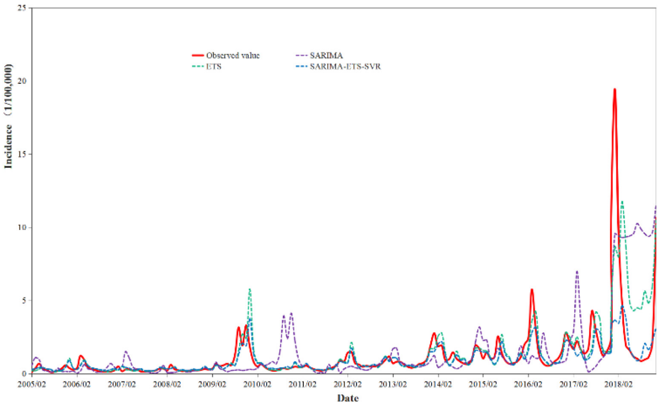
In this section, a performance assessment of the forecasts is conducted by comparing the MAE, MSE, MAPE, and RMSE indices. The incidence of influenza in January 2018 was excluded from the performance assessment of the forecasts because it was an outlier (19.4372 per 100,000 population). For the fitting performance part, the values of MAE, MSE, and RMSE of the SARIMA model were larger than those of the ETS and SARIMA-ETS-SVR hybrid models, and the ETS and SARIMA-ETS-SVR hybrid model indices did not differ significantly (Table 4). However, for the forecasting performance, the SARIMA-ETS-SVR hybrid model's values of MAE, MSE, MAPE, and RMSE were the lowest among the three models (Table 4). As shown in Figure 6, the predicted values fitted by the SARIMA-ETS-SVR model can simulate the trend of the observed values better than those of a single SARIMA model and a single ETS model.

Discussion

To the best of our knowledge, this is the first study to develop a SARIMA-ETS-SVR hybrid model in detail to forecast the incidence of seasonal influenza from 2004 to 2018 in mainland China. In this study, firstly, a single SARIMA model and a single ETS model were built respectively; On this basis, the predictive values from SARIMA and ETS models were obtained, which were used as input variables to fit and construct the SARIMA-ETS-SVR hybrid model. Subsequently, the three models were used to predict the seasonal influenza incidence, and their prediction performance was determined by comparing the MAE, MSE, MAPE, and RMSE indices.

Undoubtedly, scientific and reliable forecasting of infectious disease incidence is essential for timely implementation of precautionary measures [30]. As each predictive method has advantages and disadvantages, choosing an appropriate forecasting method based on data characteristics and sample size played a very important role in the prediction of infectious diseases [31]. The SARIMA model, a classical time-series model, is widely used to predict infectious diseases [12]. It contains a seasonality component and is applied in the field of infectious disease prediction, because it considers factors such as periodicity, seasonality, and randomness in the construction of the model [32]. The SARIMA model has the potential to eliminate time-series instability and is regarded as a practical forecasting tool for early warning and effective preventive measures against infectious diseases [33]. The ETS model can not only capture the dynamic relationship between internal regulations and external results, but can also describe the internal regulations of the time series with the current and historical minimum information [25]. Compared to the ARIMA model, the ETS model has a higher capacity to capture the dynamic dependence structures of the time series [25]. The SVR model, proposed by Vapnik, is a machine learning algorithm based on statistical theory and has been adopted in numerous fields in practice [34]. It specializes in processing nonlinear problems [28]. A high-dimensional space using structural risk minimization and a small or large sample size to build the model are advantages of this model [27].

Figure 6. Comparison of forecasts performance of the three models.



Given this background, this study focused on the construction of the SARIMA-ETS-SVR hybrid model and applied it to predict influenza incidence from 2004 to 2018 in mainland China. The reasons include the following: first, the SARIMA and ETS models require a sample size of at least 30 [35], whereas the SVR model requires a small or large sample size [27]. From the point of view of analyzing the sample size and data characteristics, 180 months data of influenza incidence were collected, which meets the data needs of the SARIMA, ETS, and SVR models. Second, from a prediction approach choice analysis perspective, the prediction approach we selected was reasonable and scientific. In the SARIMA-ETS-SVR hybrid modeling process, the SARIMA and ETS models were used to predict the influenza incidence, and their predictive values were considered as input variables to construct the SVR model, which could fully utilize the advantages of the three models. The SARIMA and ETS models were specialized in extracting the linear information of the influenza time series, whereas the SVR model had excellent performance in addressing the nonlinear information of the influenza time series. Moreover, the influenza incidence prediction issue was

Table 4. Performances assessment of three models.

Evaluating indicator	Fitting performance			Forecasting performance		
	SARIMA	ETS	SARIMA-ETS-SVR	SARIMA	ETS	SARIMA-ETS-SVR
MAE	0.6558	0.3461	0.3219	6.5350	3.7382	1.6181
MSE	7.4480	4.6684	4.6519	11.9759	6.9344	4.9588
MAPE	0.8868	0.3437	0.3641	5.1717	2.5111	0.4070
RMSE	10.5331	6.6021	6.5788	16.9365	9.8068	7.0128

SARIMA: seasonal autoregressive integrated moving average; ETS: error trend and seasonality; SVR: support vector regression; MAE: mean absolute error; MSE: mean squared error; MAPE: mean absolute percentage error; RMSE: root mean square error.

converted into a high-dimensional feature space through a nonlinear transformation to create an SVR model with a good generalization ability [29]. Generally, the incidence of infectious diseases has both linear and nonlinear characteristics in real-world studies [36]. Therefore, in our study, the SARIMA-ETS-SVR hybrid model can better extract linear and nonlinear information on influenza incidence. Third, the predictive performance was verified by comparison with MAE, MSE, MAPE, and RMSE evaluation indices. In the fitting performance part, the predictive performance of the SARIMA-ETS-SVR hybrid model was slightly better than that of a single SARIMA model and a single ETS model, while in the forecasting performance part, the predictive performance of the SARIMA-ETS-SVR hybrid model was significantly better than that of a single SARIMA model and a single ETS model. It was suggested that the SARIMA-ETS-SVR hybrid model provided more generalization ability than a single SARIMA model and a single ETS model.

For these reasons, we proposed a new hybrid SARIMA-ETS-SVR model for the prediction of influenza incidence between 2004 and 2018 in mainland China. Despite the fact that the SARIMA-SVR hybrid model achieves better performance, there are certain limitations to this study. First, seasonal influenza outbreaks are subject to many factors, such as meteorological factors [37], the level of healthcare, and residents' awareness and behavioral level of influenza [38]. However, as these factors have been excluded from the modeling process of the SARIMA-ETS-SVR hybrid model, the forecast results cannot fully interpret the practical situation of seasonal influenza. Second, each of the predictive methods has advantages and disadvantages [36], the SARIMA-ETS-SVR hybrid model is no exception. If we do not update the data in time, the predictive results of the SARIMA-ETS-SVR hybrid will not be accurately simulated. Therefore, future work should continuously update influenza data in the modeling process of the SARIMA-ETS-SVR hybrid model and obtain more accurate predictive results.

Conclusions

In this study, we collected monthly data on influenza incidence from 2004 to 2018 from the website of the data-center of China Public Health Science and proposed the SARIMA-ETS-SVR hybrid model to predict seasonal influenza incidence. The results suggest that the SARIMA-ETS-SVR hybrid model is highly capable of simulating real-world situations of the changes and trends in influenza incidence, which will provide a useful information for preventing this infectious disease. As a result, the government and relevant ministries need to strengthen influenza surveillance and prediction, and formulate corresponding preventive measures to reduce the spread of influenza. However, there are still some weaknesses in this study that affect the accuracy of the influenza prediction results. Since data on influenza may be under-reported or misreported, the accuracy of the prediction results is to some extent affected. Besides, the SARIMA-ETS-SVR hybrid model suffers from the lack of a large sample to validate its predictive performance. Therefore, in future work, we intend to collect sufficiently detailed data jointly with the Center for Disease Control and Prevention and use the SARIMA-ETS-SVR hybrid model for validation on a large sample to continuously improve and optimize the model to provide an effective tool for influenza surveillance and early warning.

Acknowledgements

We thank the Sichuan Provincial Primary Health Service Development Research Center (Grant no. SWFZ21-Q-59), Sichuan Provincial Orthopedics Hospital (Grant No. 2021GL01) for funding this study.

Authors' contributions

DZ, and RHZ: project design, data collection, data analysis, and manuscript editing. All authors contributed to the article and approved the submitted version.

References

1. Zhu X, Fu B, Yang Y, Ma Y, Hao J, Chen S, Liu S, Li T, Liu S, Guo W, Liao Z (2019) Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinformatics 20: 575. doi: 10.1186/s12859-019-3131-8.
2. Murayama T, Shimizu N, Fujita S, Wakamiya S, Aramaki E (2020) Robust two-stage influenza prediction model considering regular and irregular trends. PLoS One 15: e0233126. doi: 10.1371/journal.pone.0233126.
3. World Health Organization (2022) Influenza (seasonal) Available: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: 2 March 2022.
4. Cozza V, Campbell H, Chang HH, Iuliano AD, Paget J, Patel NN, Reiner RC, Troeger C, Viboud C, Bresee JS, Fitzner J (2021) Global seasonal influenza mortality estimates: a comparison of 3 different approaches. Am J Epidemiol 190: 718-727. doi: 10.1093/aje/kwaa196.
5. GBD 2017 Influenza Collaborators (2019) Mortality, morbidity, and hospitalisations due to influenza lower respiratory tract infections, 2017: an analysis for the global burden of disease study 2017. Lancet Respir Med 7: 69-89. doi: 10.1016/S2213-2600(18)30496-X.
6. Putri WCWS, Muscatello DJ, Stockwell MS, Newall AT (2018) Economic burden of seasonal influenza in the United

States. Vaccine 36: 3960-3966. doi: 10.1016/j.vaccine.2018.05.057.

7. Pérez-Rubio A, Platero L, Eiros Bouza JM (2019) Seasonal influenza in Spain: clinical and economic burden and vaccination programmes. Med Clin (Barc) 153: 16-27. doi: 10.1016/j.medcli.2018.11.014.
8. Zhang Y, Wang X, Li Y, Ma J (2019) Spatiotemporal analysis of influenza in China, 2005-2018. Sci Rep 9: 19650. doi: 10.1038/s41598-019-56104-8.
9. Li L, Liu Y, Wu P, Peng Z, Wang X, Chen T, Wong JYT, Yang J, Bond HS, Wang L, Lau YC, Zheng J, Feng S, Qin Y, Fang VJ, Jiang H, Lau EHY, Liu S, Qi J, Zhang J, Yang J, He Y, Zhou M, Cowling BJ, Feng L, Yu H (2019) Influenza-associated excess respiratory mortality in China, 2010-15: a population-based study. Lancet Public Health 4: e473-e481. doi: 10.1016/S2468-2667(19)30163-X.
10. Feng L, Feng S, Chen T, Yang J, Lau YC, Peng Z, Li L, Wang X, Wong JYT, Qin Y, Bond HS, Zhang J, Fang VJ, Zheng J, Yang J, Wu P, Jiang H, He Y, Cowling BJ, Yu H, Shu Y, Lau EHY (2020) Burden of influenza-associated outpatient influenza-like illness consultations in China, 2006-2015: a population-based study. Influenza Other Respir Viruses 14: 162-172. doi: 10.1111/irv.12711.
11. Qi L, Li Q, Ding XB, Gao Y, Ling H, Liu T, Xiong Y, Su K, Tang WG, Feng LZ, Liu QY (2020) Mortality burden from seasonal influenza in Chongqing, China, 2012-2018. Hum Vaccin Immunother 16: 1668-1674. doi: 10.1080/21645515.2019.1693721.
12. Qiu H, Zhao H, Xiang H, Ou R, Yi J, Hu L, Zhu H, Ye M (2021) Forecasting the incidence of mumps in Chongqing based on a SARIMA model. BMC Public Health 21: 373. doi: 10.1186/s12889-021-10383-x.
13. He F, Hu ZJ, Zhang WC, Cai L, Cai GX, Aoyagi K (2017) Construction and evaluation of two computational models for predicting the incidence of influenza in Nagasaki Prefecture, Japan. Sci Rep 7: 7192. doi: 10.1038/s41598-017-07475-3.
14. Ala'raj M, Majdalawieh M, Nizamuddin N (2021) Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. Infect Dis Model 6: 98-111. doi: 10.1016/j.idm.2020.11.007.
15. Rath S, Tripathy A, Tripathy AR (2020) Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes Metab Syndr 14: 1467-1474. doi: 10.1016/j.dsx.2020.07.045.
16. Li H, Zeng B, Wang J, Wu H (2021) Forecasting the number of new coronavirus infections using an improved grey prediction model. Iran J Public Health 50: 1842-1853. doi: 10.18502/ijph.v50i9.7057.
17. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, Xu JY, Shi J, Sun DY (2020) Analysis and predication of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty 9: 123. doi: 10.1186/s40249-020-00742-y.
18. Mohammadinia A, Saeidian B, Pradhan B, Ghaemi Z (2019) Prediction mapping of human leptospirosis using ANN, GWR, SVM and GLM approaches. BMC Infect Dis 19: 971. doi: 10.1186/s12879-019-4580-4.
19. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma W (2017) Developing a dengue forecast model using machine learning: a case study in China. PLoS Negl Trop Dis 11: e0005973. doi: 10.1371/journal.pntd.0005973.
20. Lv CX, An SY, Qiao BJ, Wu W (2021) Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. BMC Infect Dis 21: 839. doi: 10.1186/s12879-021-06503-y.
21. Yu G, Feng H, Feng S, Zhao J, Xu J (2021) Forecasting hand-foot-and-mouth disease cases using wavelet-based SARIMA-NNAR hybrid model. PLoS One 16: e0246673. doi: 10.1371/journal.pone.0246673.
22. Fang L, Wang D, Pan G (2020) Analysis and estimation of COVID-19 spreading in Russia based on ARIMA model. SN Compr Clin Med 2: 2521-2527. doi: 10.1007/s42399-020-00555-y.
23. He Z, Tao H (2018) Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: a nine-year retrospective study. Int J Infect Dis 74: 61-70. doi: 10.1016/j.ijid.2018.07.003.
24. Lu S (2021) Research on GDP forecast analysis combining BP neural network and ARIMA model. Comput Intell Neurosci 2021: 1026978. doi: 10.1155/2021/1026978.
25. Wang Y, Xu C, Yao S, Zhao Y, Li Y, Wang L, Zhao X (2020) Estimating the prevalence and mortality of coronavirus disease 2019 (COVID-19) in the USA, the UK, Russia, and India. Infect Drug Resist 13: 3335-3350. doi: 10.2147/IDR.S265292.
26. Hyndman R J, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 27: 1-22. doi: 10.18637/jss.v027.i03.
27. Ceylan Z, Bulkan S, Elevli S (2020) Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: a case study for megacity Istanbul. J Environ Health Sci Eng 18: 687-697. doi: 10.1007/s40201-020-00495-8.
28. Liu B, Jin Y, Li C (2021) Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR-SVR-ARMA combined model. Sci Rep 11: 348. doi: 10.1038/s41598-020-79462-0.
29. Yang S, Chen HC, Chen WC, Yang CH (2020) Forecasting outbound student mobility: a machine learning approach. PLoS One 15: e0238129. doi: 10.1371/journal.pone.0238129.
30. Zheng Y, Zhang L, Wang C, Wang K, Guo G, Zhang X, Wang J (2021) Predictive analysis of the number of human brucellosis cases in Xinjiang, China. Sci Rep 11: 11513. doi: 10.1038/s41598-021-91176-5.
31. Zhai M, Li W, Tie P, Wang X, Xie T, Ren H, Zhang Z, Song W, Quan D, Li M, Chen L, Qiu L (2021) Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: a time series predictive analysis. BMC Infect Dis 21: 280. doi: 10.1186/s12879-021-05973-4.
32. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S (2020) An advanced data-driven hybrid model of SARIMA-NNNAR for tuberculosis incidence time series forecasting in Qinghai province, China. Infect Drug Resist 13: 867-880. doi: 10.2147/IDR.S232854.
33. Qi C, Zhang D, Zhu Y, Liu L, Li C, Wang Z, Li X (2020) SARFIMA model prediction for infectious diseases: application to hemorrhagic fever with renal syndrome and comparing with SARIMA. BMC Med Res Methodol 20: 243. doi: 10.1186/s12874-020-01130-8.
34. Shahid F, Zameer A, Muneeb M (2020) Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos Solitons Fractals 140: 110212. doi: 10.1016/j.chaos.2020.110212.

35. Wang YW, Shen ZZ, Jiang Y (2018) Comparison of ARIMA and GM (1,1) models for prediction of hepatitis B in China. PLoS One 13: e0201987. doi: 10.1371/journal.pone.0201987.
36. Zou JJ, Jiang GF, Xie XX, Huang J, Yang XB (2019) Application of a combined model with seasonal autoregressive integrated moving average and support vector regression in forecasting hand-foot-mouth disease incidence in Wuhan, China. Medicine (Baltimore) 98: e14195. doi: 10.1097/MD.0000000000014195.
37. Suntronwong N, Vichaiwattana P, Klinfueng S, Korkong S, Thongmee T, Vongpunsawad S, Poovorawan Y (2020) Climate factors influence seasonal influenza activity in Bangkok, Thailand. PLoS One 15: e0239729. doi: 10.1371/journal.pone.0239729.
38. Chen Y, Leng K, Lu Y, Wen L, Qi Y, Gao W, Chen H, Bai L, An X, Sun B, Wang P, Dong J (2020) Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010-2018. Epidemiol Infect 148: e29. doi: 10.1017/S0950268820000151.

Corresponding author

Daren Zhao, MD.
Department of Medical Administration, Sichuan Provincial Orthopedics Hospital
Chengdu 610041, P. R. China
Tel: 86-028-87026130
Fax: 86-028-8726130
Email: cdzhaodaren@163.com

Conflict of interests: No conflict of interests is declared.

Annex – Supplementary Items

Supplementary Table 1. Monthly influenza incidence data from 2004 to 2018.

Date	Seasonal Influenza incidence 1/100,000
2004/01	0.1906
2004/02	0.4797
2004/03	0.8994
2004/04	0.7932
2004/05	0.2356
2004/06	0.1751
2004/07	0.1274
2004/08	0.1502
2004/09	0.3227
2004/10	0.1486
2004/11	0.1488
2004/12	0.1365
2005/01	0.1651
2005/02	0.0907
2005/03	0.3580
2005/04	0.6711
2005/05	0.3047
2005/06	0.2472
2005/07	0.0968
2005/08	0.1011
2005/09	0.1913
2005/10	0.3465
2005/11	0.5676
2005/12	0.3736
2006/01	0.2566
2006/02	0.3653
2006/03	1.2105
2006/04	0.9385
2006/05	0.3968
2006/06	0.3019
2006/07	0.2326
2006/08	0.1385
2006/09	0.1061
2006/10	0.1128
2006/11	0.1263
2006/12	0.2160
2007/01	0.4823
2007/02	0.1514
2007/03	0.2724
2007/04	0.2778
2007/05	0.1985
2007/06	0.2741
2007/07	0.1398
2007/08	0.1128
2007/09	0.1580
2007/10	0.1631
2007/11	0.2119
2007/12	0.3296
2008/01	0.3304
2008/02	0.1994
2008/03	0.5944
2008/04	0.2893
2008/05	0.2259
2008/06	0.1530
2008/07	0.1765
2008/08	0.1991
2008/09	0.2308
2008/10	0.2065
2008/11	0.2449

Date	Seasonal Influenza incidence 1/100,000
2008/12	0.3051
2009/01	0.2655
2009/02	0.3392
2009/03	0.6474
2009/04	0.5131
2009/05	0.5618
2009/06	0.6657
2009/07	0.5559
2009/08	1.2184
2009/09	3.1481
2009/10	1.9185
2009/11	3.2861
2009/12	1.8184
2010/01	0.7867
2010/02	0.4750
2010/03	0.6356
2010/04	0.4417
2010/05	0.2770
2010/06	0.1961
2010/07	0.1981
2010/08	0.2938
2010/09	0.3652
2010/10	0.3076
2010/11	0.3946
2010/12	0.4610
2011/01	0.4366
2011/02	0.4468
2011/03	0.5291
2011/04	0.4033
2011/05	0.3070
2011/06	0.2201
2011/07	0.1931
2011/08	0.2364
2011/09	0.3134
2011/10	0.4137
2011/11	0.5401
2011/12	0.8923
2012/01	0.8042
2012/02	1.4195
2012/03	1.4760
2012/04	0.6966
2012/05	0.6169
2012/06	0.4574
2012/07	0.5273
2012/08	0.4768
2012/09	0.4681
2012/10	0.5371
2012/11	0.6614
2012/12	0.9240
2013/01	1.1644
2013/02	0.6952
2013/03	0.7987
2013/04	0.7936
2013/05	0.6314
2013/06	0.4591
2013/07	0.3892
2013/08	0.4546
2013/09	0.6489
2013/10	0.7142
2013/11	0.9512
2013/12	1.8910
2014/01	2.7549

Date	Seasonal Influenza incidence 1/100,000
2014/02	1.8674
2014/03	1.9214
2014/04	0.9155
2014/05	0.9706
2014/06	1.4475
2014/07	1.0105
2014/08	0.8060
2014/09	0.6648
2014/10	0.6399
2014/11	0.9684
2014/12	1.9377
2015/01	1.6940
2015/02	1.0248
2015/03	1.4425
2015/04	1.0239
2015/05	1.0722
2015/06	2.5383
2015/07	1.4422
2015/08	0.8646
2015/09	0.6612
2015/10	0.6915
2015/11	0.8058
2015/12	1.1044
2016/01	1.9691
2016/02	2.4015
2016/03	5.7429
2016/04	3.0531
2016/05	1.1633
2016/06	0.6811
2016/07	0.5242
2016/08	0.5758
2016/09	0.8250
2016/10	1.0398
2016/11	1.6015
2016/12	2.7954
2017/01	2.1184
2017/02	1.6812
2017/03	2.2032
2017/04	1.6320
2017/05	1.3589
2017/06	1.7014
2017/07	4.2799
2017/08	2.9382
2017/09	1.8321
2017/10	1.2227
2017/11	2.3188
2017/12	9.8126
2018/01	19.4372
2018/02	9.5330
2018/03	4.8001
2018/04	1.9192
2018/05	1.5839
2018/06	1.1623
2018/07	1.0131
2018/08	0.8517
2018/09	0.9550
2018/10	1.1024
2018/11	2.0909
2018/12	10.6363

The data used or analyzed during the current study are available from the China Public Health Science Data Center website (https://www.phsciencedata.cn/Share/index.jsp). Anyone meeting the requirements can gain access to them. The data were relatively uninvolved in detailed patient personal information. The authors confirm they did not have any special access privileges that other would not have.