EVALUATION OF TIME-SERIES MODELS FOR PREDICTING STATUTORY NOTIFIABLE INFECTIOUS DISEASES IN SHANDONG PROVINCE, PR CHINA

Gan Wang[1,2] , Longfei Lv[3] and Jin Wang[4]

[1]Shanghai Institute of Infectious Disease and Biosecurity, [2]School of Public Health, Fudan University, Shanghai, PR China; [3]Department of Thoracic and Oncological Surgery, Children's Hospital Affiliated to Shandong University, Jinan, Shandong Province, PR China; [4]Department of Healthcare-associated Infection Management, Qingdao Municipal Hospital, University of Health and Rehabilitation Sciences, Qingdao, PR China

Abstract. In Shandong Province, PR China, three time-series models were utilized to forecast and analyze statutory notifiable infectious diseases. The models' predictive performances also were assessed. Holt-Winters, Prophet and Seasonal AutoRegressive Integrated Moving Average (SARIMA) models were trained using a dataset of infectious diseases' incidents that was gathered between January 2017 and December 2021. The test set consisted of time-series data on incidents from January to December 2022, and the fitting effects were assessed. The monthly incidences of infectious diseases were predicted using the three time-series models, and the best-performing model was chosen based on four different error indicators: mean squared, mean absolute, mean absolute percentage, and symmetric mean absolute percentage errors. The Prophet model, followed by the Holt-Winters and SARIMA models, produced for all indicators the best goodness of fit for acute hemorrhagic conjunctivitis, gonorrhea, mumps, syphilis, tuberculosis, and viral hepatitis. For acquired immunodeficiency syndrome, acute hemorrhagic conjunctivitis, epidemic hemorrhagic fever, gonorrhea, mumps, syphilis, tuberculosis, and viral hepatitis, the Prophet model fared better than the SARIMA model; while for acquired immunodeficiency syndrome, acute hemorrhagic conjunctivitis, brucellosis, epidemic hemorrhagic fever, gonorrhea, mumps, other infectious diarrheal diseases, syphilis, tuberculosis, and viral hepatitis, the Prophet model performed better than the Holt-Winters model. The Prophet model fits curves more accurately by incorporating seasonal and cyclical variations. Taken altogether, the

Prophet model outperforms both the Holt-Winters and SARIMA models in terms of yielding more accurate trend predictions regarding disease outbreaks. This should help public health agencies formulate their prevention and control programs with more precise results.

Keywords: infectious disease, Holt-Winters model, Prophet model, SARIMA model, time-series model

Correspondence: Gan Wang, Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Dongan Road 130, Shanghai 200030, PR China
Tel: +86 18221069833 E-mail: ganwang21@m.fudan.edu.cn

INTRODUCTION

Infectious diseases are illnesses caused by infective organisms that can be transmitted between individuals. These include viral infections such as measles, influenza, and HIV, as well as bacterial infections such as meningitis and cholera. Additionally, they encompass fungal pathogens, protozoa (such as those responsible for malaria), larger parasitic infections, and other worms (Rock *et al*, 2014). It is estimated that one quarter of human deaths worldwide are caused by infectious diseases (Anonymous, 1997). Hence, if the trend and peak of an epidemic can be predicted, timely interventions in the early stages of transmission can be of great benefit to the public at risk. According to the Law of the People's Republic of China on the Prevention and Control of Infectious Diseases (The National People's Congress of the People's Republic of China, 2020), China has 40 infectious diseases divided into three classes A, B and C according to their decreasing degree of severity, comprising 2, 27 and 11 diseases, respectively.

After a severe acute respiratory syndrome (SARS) epidemic in 2003, China established a public health information monitoring system, based on an information network system covering national, provincial, municipal (local), and county (district) disease prevention and control agencies, medical and health institutions, and health administrative departments located in towns, villages and urban communities (Ministry of Health,

2006). As a result, massive time-series data have been accumulated. With the maturity of several time-series analysis models (Yan *et al*, 2019; Ubaid *et al*, 2021; Middya and Roy, 2022), infectious disease prediction has become an important research area (Asfahan *et al*, 2020; El-Din Saad *et al*, 2022; Priyadarshini *et al*, 2023), revealing information on both seasonal fluctuations and long-term trends.

Time-series models have been widely used to predict a number of disease occurrences. Three important models used towards achieving this goal are Holt-Winters, Prophet and Seasonal AutoRegressive Integrated Moving Average (SARIMA) models. The Prophet model, developed by Facebook in 2017, performs time-series prediction and is suitable for generating data showing seasonal and holiday effects, and can rapidly generate high-quality predictions, with visualization tools to assist analysis (Sah *et al*, 2022). Similarly, the SARIMA model can be applied to time-series data to make predictions on seasonality (Sah *et al*, 2022), and the Holt-Winters model, also known as triple exponential smoothing, is used to analyze time-series data to show trends and seasonality (Ye *et al*, 2021). The latter model uses past data, incorporating seasonal and trend factors, and error terms, to predict future trends. This model is commonly applied to such fields as medicine, where it clearly demonstrates seasonal and trend patterns.

Most studies have explored the application of prediction models to a particular infectious disease, neglecting cross-sectional comparisons of multiple models across multiple infectious diseases. This study aimed to address such a gap in the literature. Specifically, the Holt-Winters, Prophet and SARIMA models were applied to predict regularly reported infectious diseases in Shandong Province, PR China. The predictive performances of the three models were evaluated to provide a scientific basis for the surveillance and prevention of infectious diseases' programs.

MATERIALS AND METHODS

Data sources

The data for this study were retrieved from the official website of the Shandong Provincial Health and Wellness Commission and related to the statutory notification of infectious diseases (URL: http://www.shandong.

gov.cn/col/col305260/index.html?vc_xxgkarea=113700000045022274&number=SD341103). All the infectious diseases selected for this study were based on the Law of the People's Republic of China on the Prevention and Control of Infectious Diseases (The National People's Congress of the People's Republic of China, 2020). Monthly reported incidence time-series data of diseases were collated and summarized from January 2017 to December 2022, and an incidence prediction analysis was conducted.

As a very small sample size may lead to a large prediction error and loss of meaning, only infectious diseases with monthly surveillance of ≥50 cases were considered for inclusion in the study. Class A infectious diseases (plague and cholera) were excluded, while class B infectious diseases (n=12) (acquired immunodeficiency syndrome (AIDS), bacterial and amoebic dysentery, brucellosis, epidemic hemorrhagic fever, gonorrhea, influenza pulmonary tuberculosis, mumps, scarlet fever, syphilis, viral hepatitis, and whooping cough,) and class C infectious diseases (n=3) (acute hemorrhagic conjunctivitis, hand, foot and mouth disease and other infectious diarrheal diseases) were included.

Prophet model

The Prophet model was developed by Facebook for time-series analysis and allows nonlinear fitting and forecasting of multiperiodic, trend-shifted and outlier data. The Prophet model is calculated using the following equations (Asfahan *et al*, 2020; Lu and Meyer, 2020; Yadav *et al*, 2020; Dash *et al*, 2021; Ai *et al*, 2022):

$$y(t) = g(t) + s(t) + h(t) + \varepsilon \quad (1)$$

$$g(t) = \left(k + a(t)^T \delta\right) \times \left(t + \left(m + a(t)^T \gamma\right)\right) \quad (2)$$

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right)\right) \quad (3)$$

$$h(t) = Z(t)\kappa = \sum_{i=1}^{L} \kappa_i \bullet I_{\{\in D_i\}} \quad (4)$$

where g(t) is the trend term, which represents the acyclic trend in the time series, s(t) the period or seasonal term, which is generally measured in weeks or years, h(t) the holiday term, which indicates holidays, and ε the error term, which represents all factors not taken into account by the other terms.

SARIMA model

The SARIMA model can process time-series data with seasonal components. It adds three hyperparameters (P,D,Q) to the conventional autoregressive

integrated moving average (ARIMA) hyperparameters (p,d,q) and an additional seasonal cycle hyperparameter, s. Hence, SARIMA (p,d,q), namely, (P,D,Q)s, has seven parameters that can be divided into three nonseasonal parameters (p,d,q) and four seasonal parameters $(P,D,Q)s$. P, Q, p, and q denote the maximum lag order of the seasonal, nonseasonal, autoregressive, and moving average operator, respectively, while D and d denote the number of seasonal and nonseasonal differentials respectively. The $[(p,d,q) \times (P,D,Q)s]$ order seasonal time-series model is calculated as follows (Zahidur Rahman and Mohammed, 2022; Priyadarshini *et al*, 2023; Sandie *et al*, 2023):

$$\Phi_p(L)\tilde{\phi}_p(L^s)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\varepsilon_t \quad (5)$$

where is the non-seasonal autoregressive lag polynomial, the seasonal autoregressive lag polynomial, the time series × differenced d times × and seasonally differenced D times, the trend polynomial (including the intercept), the non-seasonal moving average lag polynomial, and the seasonal moving average lag polynomial. SARIMA model contains a seasonal component, unlike the autoregressive moving average in the ARIMA model, which removes the seasonal component by adding a single step to seasonal differencing and introduces a seasonal component into the model. Although the D-seasonal difference is applied, the series may still contain a seasonal autocorrelation component. Therefore, the SARIMA model better expresses the inter-sample autocorrelations in periodic time series.

Holt-Winters model

The Holt-Winters model introduces a winter cycle term (also called a seasonal term) based on the Holt model and can handle fluctuations in monthly (cycle 12), quarterly (cycle 4), and weekly (cycle 7) data as well as other fixed periods (Lynch and Gore, 2021; Garcia *et al*, 2022; El-Din Saad *et al*, 2022). The introduction of multiple winter terms allows the coexistence of multiple periods. The Holt-Winters model is applied to nonstationary series with linear trends and fixed periods, and can be an additive or multiplicative model. An additive model is selected when the magnitude of the seasonal pattern does not depend on the data magnitude. The following additive model was used in this study:

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)} \quad (6)$$

$$\ell_t = \alpha\,(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \quad (7)$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \quad (8)$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \quad (9)$$

where k represents the integer component of (h-1)/m, ensuring that the estimates of seasonal indices utilized for forecasting are derived from the final year of the data sample. The level equation (7) presents a weighted amalgamation between the seasonally adjusted observation (yt - st-m) and the non-seasonal forecast (ℓt-1+bt-1) for the given time t, the trend equation (8) remains consistent with Holt's linear method and the seasonal equation (9) illustrates a weighted blend between the present seasonal index (yt - ℓt-1 - bt-1) and the seasonal index from the corresponding season in the previous year.

Fit evaluation

The Holt-Winters, Prophet, and SARIMA models were assessed for regression using the following evaluation metrics: mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric MAPE (SMAPE). These indicators were calculated utilizing the equations specified below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (10)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (11)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (12)$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2} \quad (13)$$

Data processing and analysis

The PyCharm integrated development environment in Phyton was used for data modeling and analysis. The dataset containing the numbers of monthly cases of infectious diseases from January 2017 to December 2021 was used as the training set to build the Holt-Winters, Prophet and SARIMA models, while the numbers of monthly cases of infectious diseases from January to December 2022 were used to construct the test set to evaluate model fitness (Fig 1).

RESULTS

Descriptive statistical analysis of infectious diseases

Among the 15 statutory notifiable infectious diseases in Shandong

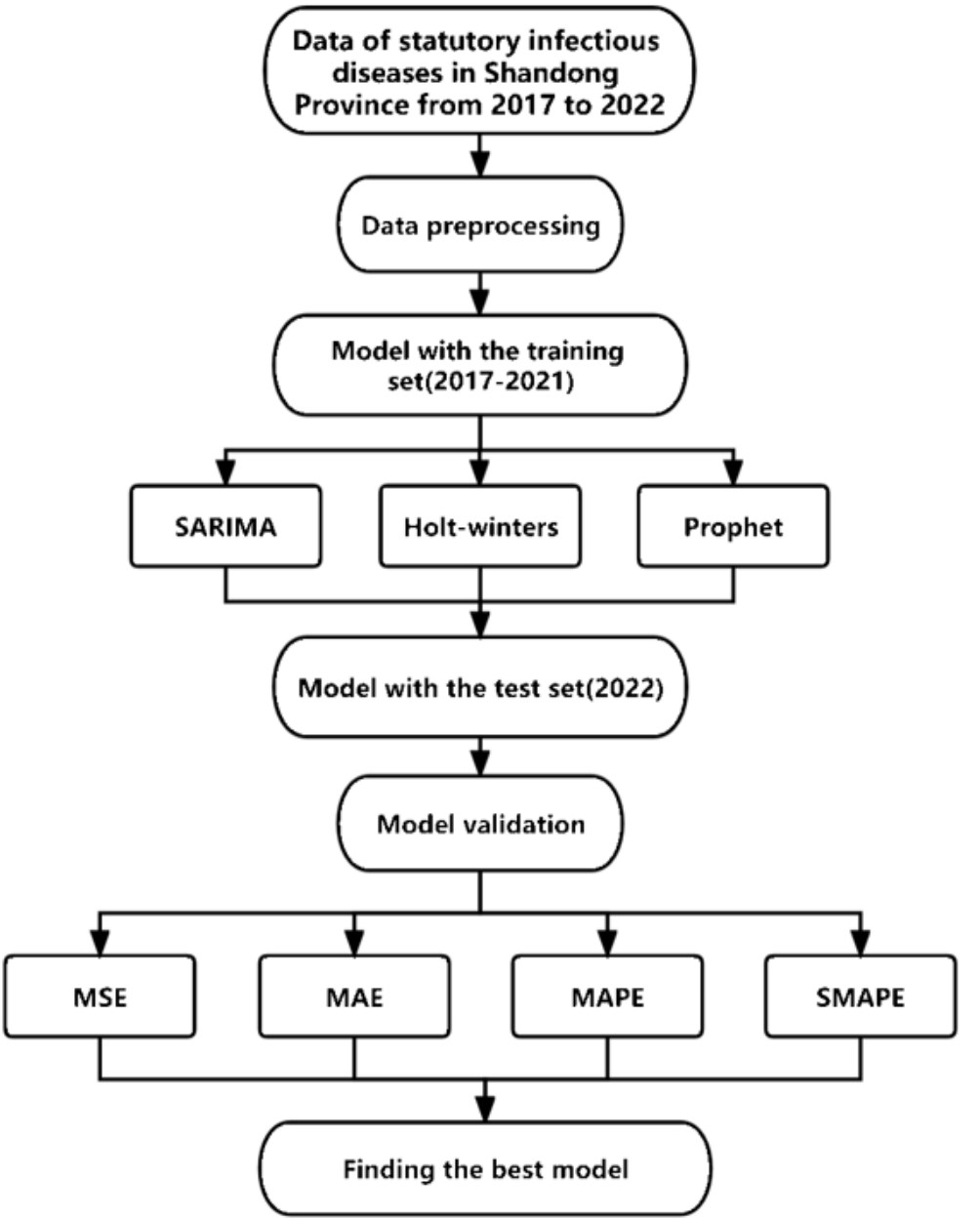

Fig 1 - Flow chart for predicting incidence of infectious diseases

MAE: mean absolute error; MAPE: mean absolute percentage error; MSE: mean squared error; SARIMA: seasonal autoregressive integrated moving average; SMAPE: symmetric mean absolute percentage error

Province, those with a monthly average of >1000 notified cases were syphilis, influenza, tuberculosis, hand, foot and mouth disease, viral hepatitis, and other infectious diarrheal diseases, with monthly cases (mean ± SD) of 1,614 ± 233, 2,246 ± 3,321, 2,672 ± 453, 4,752 ± 6,255, 7,434 ± 996, and 10,244 ± 4170, respectively. Diseases with a monthly average of <100 notified cases were epidemic hemorrhagic fever, AIDS and acute hemorrhagic conjunctivitis, with monthly cases (mean ± SD) of 74 ± 71, 81 ± 30 and 91 ± 19 cases, respectively. The minimum number of notified cases per month was whooping cough and the three diseases with the highest numbers of monthly cases were influenza (20,652), other infectious diarrheal diseases (22,555) and hand, foot, and mouth disease (25,247) (Table 1).

The parameters in the SARIMA model for the 15 statutory notifiable infectious diseases in Shandong Province, *ie* (p,d,q) and $(P,D,Q)s$, were obtained automatically using the training set and function "Auto ARIMA" in Python (Table 2). The basis for building an optimal SARIMA model was finding the minimum value of the Akaike information criterion. The results for the Holt-Winters, Prophet and SARIMA models in predicting statutory notifiable infectious diseases are shown in Fig 2.

Comparisons among the three models

For the prediction of six infectious diseases (acute hemorrhagic conjunctivitis, gonorrhea, mumps, syphilis, tuberculosis, and viral hepatitis), the best of all the goodness-of-fit indicators was obtained with the Prophet model, followed by Holt-Winters and then SARIMA models (Table 3). For viral hepatitis, the MSE value of the Prophet, Holt-Winters and SARIMA models was 1,241,937.789, 2,172,805.942 and 2,605,713.984, the MAE value 837.587, 996.782, and 1170.518, MAPE value 0.145, 0.179, and 0.204, and SMAPE value 12.578, 14.449 and 16.739, respectively. For AIDS, the best goodness-of-fit indicators were obtained with the Prophet model followed by the SARIMA and then the Holt-Winters models, with MSE value of 615.338, 1,007.338 and 1,399.074, respectively (Table 3). For the prediction of whooping cough, the best goodness-of-fit indicators were obtained with the Prophet model followed by the Holt-Winters and then the SARIMA models (Table 3).

Table 1

Descriptive analysis of the number of confirmed cases of statutory notifiable infectious diseases in Shandong Province, PR China 2022

Notifiable infectious disease	Confirmed cases	
	Mean ± SD	Minimum - Maximum
Class B		
AIDS	81 ± 30	27 - 168
Viral hepatitis	7,434 ± 996	3,944 - 9,373
Epidemic hemorrhagic fever	74 ± 71	12 - 390
Bacterial and amoebic dysentery	197 ± 140	19 - 580
Pulmonary tuberculosis	2,672 ± 453	1,392 - 3,564
Whooping cough	299 ± 251	1 - 1,317
Scarlet fever	508 ± 456	39 - 1,825
Brucellosis	254 ± 105	57 - 456
Gonorrhea	318 ± 66	112 - 457
Syphilis	1,614 ± 233	787 - 2,032
Influenza	2,246 ± 3,321	194 - 20,652
Mumps	459 ± 187	93 - 967
Class C		
Acute hemorrhagic conjunctivitis	91 ± 19	58 - 129
Other infectious diarrheal diseases	10,244 ± 4,170	3,073 - 22,555
Hand, foot and mouth disease	4,752 ± 6,255	43 - 25,247

AIDS: acquired immunodeficiency syndrome; SD: standard deviation

Table 2

SARIMA (p,d,q) (P,D,Q)s model parameters

Statutory notifiable infectious disease	Best model
Class B	
AIDS	(1,0,0) (0,1,1) [12]
Viral hepatitis	(0,1,0) (0,1,1) [12]
Epidemic hemorrhagic fever	(1,0,0) (2,1,1) [12]
Bacterial and amoebic dysentery	(2,0,0) (0,1,0) [12]
Pulmonary tuberculosis	(0,0,1) (1,1,0) [12]
Whooping cough	(1,0,1) (2,1,0) [12]
Scarlet fever	(3,0,0) (0,1,1) [12]
Brucellosis	(1,0,0) (0,1,1) [12]
Gonorrhea	(1,0,1) (1,1,0) [12]
Syphilis	(0,1,2) (0,1,1) [12]
Class C	
Influenza	(1,0,0) (1,1,0) [12]
Mumps	(2,0,0) (0,1,1) [12]
Acute hemorrhagic conjunctivitis	(1,0,0) (2,1,0) [12]
Other infectious diarrheal diseases	(0,0,1) (1,1,2) [12]
Hand, foot and mouth disease	(2,0,0) (0,1,1) [12]

Note: In SARIMA (Seasonal Autoregressive Integrated Moving Average) models, the first set of digits (1,0,0) represents the order of the autoregressive (AR) part, the second set (0,1,1) the order of the differencing (d) part, and [12] the seasonal differencing period. In this scenario, the model has an autoregressive order of 1, a differencing order of 1, and a seasonal differencing period of 12.

AIDS: acquired immunodeficiency syndrome

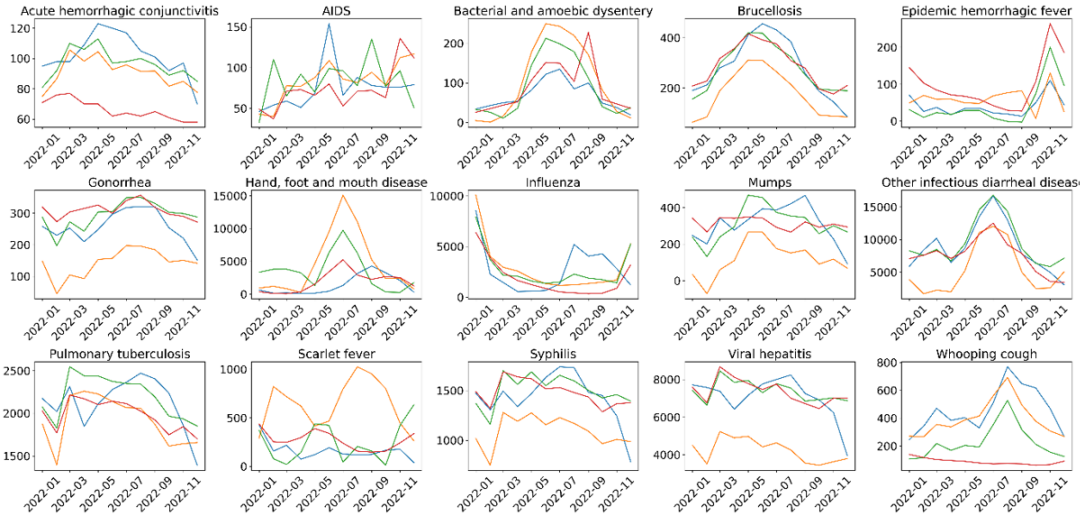

Fig 2 - Results forecasted by the Holt-Winters, SARIMA and Prophet models

The blue curve represents the true value of the monthly incidence of each legally infectious disease for 2022. The orange, red and green lines in the respective subgraphs represent the forecasted values of monthly reported cases for the year 2022 using the Holt-Winters, SARIMA and Prophet models for infectious diseases. The abscissa of each subgraph represents the year and month, while the ordinate represents the number of cases of illness per month.

X-axis represents years and months while the Y-axis represents the number cases

AIDS: acquired immunodeficiency syndrome; SARIMA: seasonal autoregressive integrated moving average.

Pairwise comparison of the models

Holt-Winters versus SARIMA models

The Holt-Winters model outperformed the SARIMA model in predicting the monthly incidence of acute hemorrhagic conjunctivitis, gonorrhea, hand, foot and mouth disease, mumps, pertussis, scarlet fever, syphilis, tuberculosis, and viral hepatitis, whereas the converse was obtained for AIDS, brucellosis and other infectious diarrheal diseases. For example, the MSE value of the SARIMA and Holt-Winters model

Table 3

Validity evaluation indicators MSE, MAE, MAPE, and SMAPE for fitting models of statutory notifiable infectious diseases

Notifiable infectious disease	SARIMA				Holt-Winters				Prophet			
	MSE	MAE	MAPE	SMAPE	MSE	MAE	MAPE	SMAPE	MSE	MAE	MAPE	SMAPE
Class B												
AIDS	1,007.338	24.823	0.331	28.960	1,399.074	31.842	0.444	35.376	615.338	21.255	0.286	25.990
Viral hepatitis	2,605,713.984	1,170.518	0.204	16.739	2,172,805.942	996.782	0.179	14.449	1,241,937.789	837.587	0.145	12.578
Epidemic hemorrhagic fever	6,331.830	69.087	1.938	95.153	6,342.173	57.053	1.438	78.298	1,192.690	24.781	0.612	76.636
Bacterial and amoebic dysentery	3,690.376	38.209	0.517	36.051	1,095.522	28.201	0.521	37.618	3,793.428	55.376	1.090	107.538
Pulmonary tuberculosis	207,390.244	361.831	0.190	16.356	146,547.504	277.356	0.151	13.026	76,802.104	225.224	0.117	10.908
Whooping cough	84,631.838	255.729	0.525	74.008	37,423.572	167.199	0.364	34.144	59,371.126	229.759	0.518	71.565
Scarlet fever	128,161.412	277.926	3.206	108.765	72,306.791	200.569	2.420	69.975	75,376.454	235.680	2.557	132.668
Brucellosis	2,381.754	37.022	0.204	15.926	6,140.607	60.809	0.316	23.094	2,248.564	38.256	0.211	17.135
Gonorrhea	7,234.989	71.776	0.326	26.312	3,716.843	44.294	0.214	17.446	2,963.858	42.727	0.201	16.826
Syphilis	122,609.331	244.303	0.217	16.932	108,852.323	226.602	0.203	15.990	51,618.744	171.805	0.149	12.815

Table 3 (cont)

Notifiable infectious disease	SARIMA				Holt-Winters				Prophet			
	MSE	MAE	MAPE	SMAPE	MSE	MAE	MAPE	SMAPE	MSE	MAE	MAPE	SMAPE
Class C												
Influenza	10,579,095.945	2148.285	0.680	73.613	7,248,188.744	1,812.943	0.968	61.684	3,729,679.636	1,605.930	1.032	67.814
Mumps	20,761.429	106.824	0.566	32.447	8,905.637	79.023	0.359	28.037	8,120.787	76.069	0.346	27.755
Acute hemorrhagic conjunctivitis	702.420	23.185	0.221	25.840	227.177	10.685	0.114	11.119	143.588	10.143	0.100	10.302
Other infectious diarrheal diseases	3,354,436.618	1,431.329	0.216	18.257	10,858,549.624	2,855.016	0.350	37.449	2,558,095.196	1,187.025	0.217	16.630
Hand, foot and mouth disease	24,662,471.890	3170.966	6.711	101.830	7,127,416.593	1,788.698	3.754	78.500	17,082,346.916	3,714.165	9.748	169.930

AIDS: acquired immunodeficiency syndrome; MAE: mean absolute error; MAPE: mean absolute percentage error; MSE: mean squared error; SARIMA: seasonal autoregressive integrated moving average; SMAPE: symmetric mean absolute percentage error

for predicting the incidence of scarlet fever was 128,161.412 and 723,06.791 respectively (Table 3).

SARIMA versus Prophet models

The Prophet model outperformed the SARIMA model in predicting the monthly incidence of acute hemorrhagic conjunctivitis, AIDS, epidemic hemorrhagic fever, gonorrhea, mumps, pertussis, syphilis, tuberculosis, and viral hepatitis, whereas the converse was obtained for bacterial and amoebic dysentery. For example, the MAE value of the SARIMA and Prophet models for predicting the incidence of epidemic hemorrhagic fever was 69.087 and 24.781 respectively (Table 3).

Holt-Winters versus Prophet models

The Prophet model outperformed the Holt-Winters model in predicting the monthly incidence of acute hemorrhagic conjunctivitis, AIDS, brucellosis, epidemic hemorrhagic fever, gonorrhea, mumps, other infectious diarrheal diseases, syphilis, tuberculosis, and viral hepatitis, whereas the converse was obtained for bacterial and amoebic dysentery, hand, foot and mouth disease, pertussis, and scarlet fever. For example, the MAPE value of the Holt-Winters and Prophet models for mumps was 0.359 and 0.346 respectively (Table 3).

DISCUSSION

Predicting epidemic outbreaks is a key concern in public health. It allows national health systems to plan for healthcare supply and demand and to avoid overloading or crowding out healthcare resources (Martin-Moreno *et al*, 2022). The mathematical models used to predict epidemiological trends are typically divided into conventional and data-driven models. Conventional epidemiological models include SIR, SQUIDER, SEIR, SEIRS, SIRV, SIRD, SIRS, and SUCQ (D: detected; E: exposed; I: infectious; Q: quarantine; R: recovered; S: susceptible; U: undetected; V: vaccinated) (Ahmetolan *et al*, 2020; Calafiore *et al*, 2020; Khan *et al*, 2020; Kissler *et al*, 2020; Rocchi *et al*, 2020; Struben, 2020; Zhao and Chen, 2020). Data-driven models include ARIMA, machine learning, deep learning, genetic evolutionary programming, long short-term memory (LSTM), and global epidemics and mobility models (Salgotra *et al*, 2020; Shakeel *et al*, 2021). Our study focused on evaluating the ability of different data-driven models to predict infectious diseases.

Hybrid models have also been widely used in recent years for the prediction and analysis of the novel coronavirus disease (COVID-19) pandemic, providing a synergy between data-driven and machine-learning models. Common machine learning algorithms include support vector machines, random forests, decision trees, gradient boosting trees, adaptive boosting, and extreme gradient boosting (Hassan *et al*, 2022). The cutting-edge time-series forecasting models are deep learning models, such as recurrent and convolutional neural networks, transformers, and neural basis expansion analysis for time series. Conventional ARIMA models are less efficient when dealing with large-scale time series and fail to extract similarities across different series. As various time series are fitted separately and independently, deep learning models can learn from multiple series and greatly improve the efficiency of the models in processing large-scale data. Therefore, artificial intelligence may be a better choice than conventional models for predicting the incidence of infectious diseases.

Based on the reported cases of 15 infectious diseases, the incidence of these diseases generally follows seasonal fluctuations, which provides a good time-series data basis for their prediction. In our study, the monthly numbers of notified cases of 15 infectious diseases in Shandong Province, PR China, were predicted using the Holt-Winters, Prophet and SARIMA models. The prediction performance of each model for different diseases was comprehensively evaluated. The Prophet model provided the best overall performance. From the long-term trend of infectious disease notifications from 2017 to 2022, the monthly AIDS incidence showed an upward trend. On the other hand, the monthly incidence of bacterial and amoebic dysentery, tuberculosis, scarlet fever, mumps, and hand, foot, and mouth disease showed a downward trend, indicating better control of these infectious diseases.

Every infectious disease prediction model has its advantages and disadvantages. Unlike the ARIMA model, Prophet does not require data to have evenly spaced intervals, so there is no need for special operations such as interpolation. This model can automatically handle missing values and outliers in the studied sequence, and predict the future trends of time

series (Shen *et al*, 2020). The Prophet model uses certain parameters that require manual configuration, such as the flexibility of seasonality and trends, and the selection of seasonal patterns. The accurate selection and adjustment of these parameters may necessitate professional knowledge and experience. In addition, the Prophet model has some limitations when it comes to long-term predictions. The assumption of unsaturated susceptible population in the Prophet model undermines the long-term prediction of infectious diseases. In reality, there must be an upper limit to the susceptible population. Therefore, a more intricate model is needed for accurate long-term forecasting purposes (Asfahan *et al*, 2020).

In the SARIMA model, the parameters possess discernible interpretations, such as the autoregressive (AR) component, the differencing (I) component and the moving average (MA) component. Thus, the comprehension and elucidation of these factors can influence the model (Malki *et al*, 2022). In addition, the SARIMA model necessitates high data quality, entailing the handling of missing values and outliers and ensuring data stationarity. In the case of poor data quality, the model's efficacy may be compromised. Selecting appropriate SARIMA model parameters presents a challenging task, typically requiring the aid of graphs and statistical tests for parameters' selection (Priyadarshini *et al*, 2023; Sandie *et al*, 2023). Our study showed that SARIMA model exhibited significant deviations in predicting the incidence of hemorrhagic conjunctivitis and whooping cough. The SARIMA model assumes that the time series data are stationary and satisfies certain conditions such as autocorrelation. However, in the period between 2020 and 2021, the incidence of whooping cough did not exhibit seasonal fluctuations and the number of cases affected by the epidemic remained relatively flat. Additionally, the incidence of acute hemorrhagic conjunctivitis in 2017 was at a historically low level compared to the following four years. These unstable data make it difficult for the SARIMA model to accurately fit and predict the data. Thus, improvements will be required to address these problems.

The Holt-Winters model showed remarkable resilience in handling anomalies and short-

term fluctuations, along with other forms of noise. Nevertheless, this model necessitates a stationary data sequence and encounters challenges when handling data with limited historical information (Lynch and Gore, 2021; Garcia *et al*, 2022). The predictions of the Holt-Winters model exhibit noteworthy deviations when estimating the incidence rates of gonorrhea and viral hepatitis. Additionally, the Holt-Winters model often disregards external factors that might significantly influence the transmission of infectious diseases. For instance, the imposition of China's epidemic prevention and control policies led to exceptionally low values for both gonorrhea and viral hepatitis in January 2020, resulting in unstable seasonal fluctuations. Consequently, the Holt-Winters model encounters difficulty in accurately fitting the new data. To enhance the precision, it is imperative to consider these external factors.

The Prophet model has been shown to outperform the SARIMA model in predicting the monthly incidence of AIDS, with the MSE, MAE and MAPE for the SARIMA (1,0,1) (0,1,1) model being 0.0073, 0.0657 and 47.8470, respectively (Lu and Meyer, 2020), and those for the Prophet model being 0.0060, 0.0602 and 44.8336, respectively (Luo *et al*, 2022). These results are consistent with the findings of our study. The Prophet model performed well in predicting the incidence of dengue fever, being the best time-series fit model for predicting morbidity over 10 years in nine cities in Maharashtra, India (Patil and Pandya, 2021). It outperformed the Holt-Winters, autoregressive, moving average, ARIMA, and SARIMA models. The Prophet model also performed well in another study on the incidence prediction of dengue fever (Valencia *et al*, 2021). Time-series data of confirmed dengue cases in the metropolitan area of Panama from 1999 to 2014 were used for training, and those from 2015 to 2017 were used for testing. The Prophet model with exogenous variables demonstrates root MSE and MAPE value of 25.76 and 108.44 respectively compared to 26.16 and 59.68 respectively using a recurrent LSTM neural network model (Valencia *et al*, 2021).

Joint models can outperform single models. In a study on brucellosis in Algeria, a hybrid neural network autoregression/ SARIMA model showed superior predictive performance (Akermi *et*

al, 2022). This model has greatly assisted veterinarians and health policymakers in implementing effective and targeted interventions. In the prediction of COVID-19, researchers in India found that, although stacked models require larger datasets for training, a coupled stacked LSTM/gated recurrent unit model outperforms the Prophet and ARIMA models in terms of the coefficient of determination and root MSE value (Sah *et al*, 2022).

The number of disease cases can also be indirectly predicted by changes in their associated effects. In Sabah, Malaysia, researchers used climate predictors and Internet search queries to predict cases of hand, foot and mouth disease (Jayaraj and Hoe, 2022). Correlations are first estimated by measuring model fitting followed by indicator validation based on the root MSE and MAPE values. The SARIMA model has a mean temperature regression lag term of 0 and a Google search a regression lag term of 1. The model produces the most accurate prediction 2 weeks in advance (root MSE = 18.77 and MAPE = 0.242), outperforming the model's highest accuracy 2 weeks earlier and seems to be promising model for disease preparedness in Sabah.

COVID-19 was included in the national statutory management of infectious diseases. The National Health Commission of the People's Republic of China included pneumonia deriving from COVID-19 as a class B infectious disease under the Law of the People's Republic of China on the Prevention and Control of Infectious Diseases on 20 January 2020 and adopted preventive and control measures for class A infectious diseases (National Health Commission of the People's Republic of China, 2020). However, COVID-19 was not included in this study because it did not exist in the study period of 2017-2019.

However, the adoption of time-series prediction models has several limitations (Malki *et al*, 2022). Selecting appropriate parameters for a model can be a formidable challenge as different parameter settings may yield divergent outcomes. In the context of infectious disease prediction models, the focus tends to be solely on the impact of the disease itself, disregarding other external factors that may influence transmission, such as government interventions or population mobility. The prognostication of infectious disease models in our study relies upon historical data, thus potentially

constraining their predictive capabilities when confronted with new emerging outbreaks lacking such historical information.

There exist numerous approaches to enhance the accuracy of predictive models (Toharudin *et al*, 2023; Lynch and Gore, 2021). The Prophet model encompasses several adjustable parameters, including seasonal adjustment, trend flexibility parameter and seasonal flexibility parameters. It is advisable to make necessary parameter adjustments based on the characteristics of the data and the requirements of the problem at hand. Employing techniques such as cross-validation can help in selecting the optimal parameters. To enhance the Holt-Winters model, future research should employ statistical methods to analyze the characteristics of seasonal effects, such as calculating the average, variance or standard deviation of seasonal components. These indicators can be used to observe whether these statistical properties remain stable under different data amplitudes; if the statistical properties of the seasonal component remain stable, it can be assumed that the seasonal pattern is not dependent on data amplitude, and an additive model may be used in the analysis. Conversely, if the statistical properties of the seasonal component do vary with data amplitude, it would be appropriate to utilize a multiplicative model in the analysis. If the SARIMA model fails to adequately fit the data or has low predictive accuracy, users should consider incorporating exogenous variables, exploring alternative time series models (such as SARIMAX or VAR), or integrating machine learning techniques to enhance model performance.

Overall, the Prophet model has advantages in both flexibility and ease of use, making it suitable for multiple types and trends of infectious diseases. On the other hand, the Holt-Winters and SARIMA models are more appropriate for handling data with distinctive seasonality and trends. The performance of a specific model in infectious disease forecasting depends on the characteristics of the data and the accurate selection of model parameters. Therefore, when using any model for infectious disease prediction, an evaluation should be conducted based on the actual situation, in order to select the most appropriate model.

CONFLICT OF INTEREST DISCLOSURE

The authors report no financial relationships with commercial interests.

REFERENCES

Ahmetolan S, Bilge AH, Demirci A, Peker-Dobie A, Ergonul O. What can we estimate from fatality and infectious case data using the susceptible-infected-removed (SIR) model? A case study of COVID-19 pandemic. *Front Med (Lausanne)* 2020; 7: 556366.

Ai YA, He F, Lancaster E, Lee J. Application of machine learning for multi-community COVID-19 outbreak predictions with wastewater surveillance. *PLoS One* 2022; 17(11): e0277154.

Akermi SE, L'Hadj M, Selmane S. Epidemiology and time series analysis of human brucellosis in Tebessa Province, Algeria, from 2000 to 2020. *J Res Health Sci* 2022; 22(1): e00544.

Anonymous. The World Health Report 1997--conquering suffering, enriching humanity. *World Health Forum* 1997; 18(3-4): 248-60.

Asfahan S, Gopalakrishnan M, Dutt N, *et al*. Using a simple open-source automated machine learning algorithm to forecast COVID-19 spread: a modelling study. *Adv Respir Med* 2020; 88(5): 400-5.

Calafiore GC, Novara C, Possieri C. A modified SIR model for the COVID-19 contagion in Italy, 2020 [cited 2023 Jun 21]. Available from: URL: https://arxiv.org/pdf/2003.14391.pdf

Dash S, Chakraborty C, Giri SK, Pani SK. Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics. *Pattern Recognit Lett* 2021; 151: 69-75.

El-Din Saad NG, Ghoniemy S, Faheem H, Seada NA. An evaluation of time series-based modeling and forecasting of infectious diseases progression using statistical versus compartmental methods, 2022 [cited 2023 Apr 23]. Available from: URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9756060

Garcia KKS, Abrahao AA, Oliveira AFD, *et al*. Malaria time series in the extra-Amazon region of Brazil: epidemiological scenario and a two-year prediction model. *Malar J* 2022; 21(1): 157.

Hassan A, Prasad D, Rani S, Alhassan M. Gauging the impact of artificial intelligence and mathematical modeling in response to the COVID-19 pandemic: a systematic review. *Biomed Res Int* 2022; 2022: 7731618.

Jayaraj VJ, Hoe VCW. Forecasting HFMD cases using weather variables and Google search queries in Sabah,

Malaysia. *Int J Environ Res Public Health* 2022; 19(24): 16880.

Khan ZS, Van Bussel F, Hussain F. A predictive model for COVID-19 spread - with application to eight US states and how to end the pandemic. *Epidemiol Infect* 2020; 148: e249.

Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 2020; 368(6493): 860-8.

Lu JY, Meyer S. Forecasting flu activity in the United States: benchmarking an endemic-epidemic beta model. *Int J Environ Res Public Health* 2020; 17(4): 1381.

Luo Z, Jia X, Bao J, *et al*. A combined model of SARIMA and Prophet Models in forecasting AIDS incidence in Henan Province, China. *Int J Environ Res Public Health* 2022; 19(10): 5910.

Lynch CJ, Gore R. Short-range forecasting of COVID-19 during early onset at county, health district, and state geographic levels using seven methods: comparative forecasting study. *J Med Internet Res* 2021; 23(3): e24925.

Malki A, Atlam E, Hassanien AE, Ewis A, Dagnew G, Gad I. SARIMA model-based forecasting required number of COVID-19 vaccines globally and empirical analysis of peoples' view towards the vaccines. *Alex Eng J* 2022; 61(12): 12091-110.

Martin-Moreno JM, Alegre-Martinez A, Martin-Gorgojo V, Alfonso-Sanchez JL, Torres F, Pallares-Carratala V. Predictive models for forecasting public health scenarios: practical experiences applied during the first wave of the COVID-19 pandemic. *Int J Environ Res Public Health* 2022; 19(9): 5546.

Middya AI, Roy S. Pollutant specific optimal deep learning and statistical model building for air quality forecasting. *Environ Pollut* 2022; 301: 118972.

Ministry of Health. Notice of the Ministry of Health on amending the "Measures for the Management of Public Health Emergencies and Infectious Disease Epidemic Monitoring Information Reporting" (Ministry of Health Order No. 37), 2006 [cited 2023 Jun 16]. Available from: URL: https://www.gov.cn/zwgk/2006-09/08/content_382018.htm [in Chinese]

National Health Commission of the People's Republic of China. Announcement No.1 of the National Health Commission of the People's Republic of China concerning pneumonia infection, 2020 [cited 2023 Jun 17]. Available from: URL: https://www.gov.cn/zhengce/zhengceku/2020-01/21/content_5471164.htm [in Chinese]

Patil S, Pandya S. Forecasting dengue

hotspots associated with variation in meteorological parameters using regression and time series models. *Front Public Health*. 2021; 9: 798034.

Priyadarshini I, Mohanty P, Kumar R, Taniar D. Monkeypox outbreak analysis: an extensive study using machine learning models and time series analysis. *Computers* 2023; 12(2): 36.

Rocchi E, Peluso S, Sisti D, Carletti M. A possible scenario for the COVID-19 epidemic, based on the SI(R) Model. *SN Compr Clin Med* 2020; 2(5): 501-3.

Rock K, Brand S, Moir J, Keeling MJ. Dynamics of infectious diseases. *Rep Prog Phys* 2014; 77(2): 026602.

Sah S, Surendiran B, Dhanalakshmi R, Mohanty SN, Alenezi F, Polat K. Forecasting COVID-19 pandemic using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU models in India. *Comput Math Methods Med* 2022; 2022: 1556025.

Salgotra R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. *Chaos Solitons Fractals* 2020; 138: 109945.

Sandie AB, Tejiokem MC, Faye CM, *et al*. Observed versus estimated actual trend of COVID-19 case numbers in Cameroon: a data-driven modelling. *Infect Dis Model* 2023; 8(1): 228-39.

Shakeel SM, Kumar NS, Madalli PP, Srinivasaiah R, Swamy DR. COVID-19 prediction models: a systematic literature review. *Osong Public Health Res Perspect* 2021; 12(4): 215-29.

Shen J, Valagolam D, McCalla S. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) in Seoul South Korea. *PeerJ* 2020; 8: 9961.

Struben J. The coronavirus disease (COVID-19) pandemic: simulation-based assessment of outbreak responses and postpeak strategies. *Syst Dyn Rev* 2020; 36(3): 247-93.

The National People's Congress of the People's Republic of China. Law of the People's Republic of China on the prevention and control of infectious diseases, 2020 [cited 2023 Mar 15]. Available from: URL: http://www.npc.gov.cn/npc/c2/c238/202001/t20200122_304251.html [in Chinese]

Toharudin T, Pontoh RS, Caraka RE, Zahroh S, Lee Y, Chen RC. Employing long short-term memory and Facebook prophet model in air temperature forecasting. *Commun Stat Simul Comput* 2023; 52(2): 279-90.

Ubaid A, Hussain F, Saqib M. Container shipment demand forecasting in the Australian shipping industry: a case study of Asia-Oceania trade lane. *J Mar Sci Eng* 2021; 9(9): 968.

Valencia VN, Diaz Y, Pascale JM, Boni MF, Sanchez-Galan JE. Assessing the effect of climate variables on the incidence of dengue cases in the metropolitan region of Panama City. *Int J Environ Res Public Health* 2021; 18(22): 12108.

Yadav D, Maheshwari H, Chandra U. Outbreak prediction of COVID-19 in most susceptible countries. *Glob J Environ Sci Manag* 2020; 6(SI): 11-20.

Yan Y, Li B, Xiao JY, Liang YD, Shang YW, Zhou KD. Comparative study on prediction algorithms for power grid system access failure times, 2019 [cited 2023 Jun 16]. Available from: URL: https://iopscience.iop.org/article/10.1088/1755-1315/252/3/032183/pdf

Ye GH, Alim M, Guan P, Huang DS, Zhou BS, Wu W. Improving the precision of modeling the incidence of hemorrhagic fever with renal syndrome in mainland China with an ensemble machine learning approach. *PLoS One* 2021; 16(3): e0248597.

Zahidur Rahman M, Mohammed N. Strategic assessment considering the future outbreak and hospital scenario, 2022 [cited 2023 Jun 20]. Available from: URL: https://link.springer.com/chapter/10.1007/978-3-031-20429-6_54

Zhao S, Chen H. Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant Biol* 2020; 8(1): 11-9.