Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population

Zhongqi Li1,2,*
Zhizhong Wang3,*
Huan Song1
Qiao Liu1
Biyu He1
Peiyi Shi1
Ye Ji1
Dian Xu1
Jianming Wang1,2

1Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, People’s Republic of China; 2Key Laboratory of Infectious Diseases, School of Public Health, Nanjing Medical University, Nanjing, People’s Republic of China; 3Department of Epidemiology and Health Statistic, School of Public Health, NingXia Medical University, Yinchuan, People’s Republic of China

*These authors contributed equally to this work

Objective: To investigate suitable forecasting models for tuberculosis (TB) in a Chinese population by comparing the predictive value of the autoregressive integrated moving average (ARIMA) model and the ARIMA-generalized regression neural network (GRNN) hybrid model.

Methods: We used the monthly incidence rate of TB in Lianyungang city from January 2007 through June 2016 to construct a fitting model, and we used the incidence rate from July 2016 to December 2016 to evaluate the forecasting accuracy. The root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and mean error rate (MER) were used to assess the performance of these models in fitting and forecasting the incidence of TB.

Results: The ARIMA (10, 1, 0) (0, 1, 1)12 model was selected from plausible ARIMA models, and the optimal spread value of the ARIMA-GRNN hybrid model was 0.23. For the fitting dataset, the RMSE, MAPE, MAE and MER were 0.5594, 11.5000, 0.4202 and 0.1132, respectively, for the ARIMA (10, 1, 0) (0, 1, 1)12 model, and 0.5259, 11.2181, 0.3992 and 0.1075, respectively, for the ARIMA-GRNN hybrid model. For the forecasting dataset, the RMSE, MAPE, MAE and MER were 0.2805, 8.8797, 0.2261 and 0.0851, respectively, for the ARIMA (10, 1, 0) (0, 1, 1)12 model, and 0.2553, 5.7222, 0.1519 and 0.0571, respectively, for the ARIMA-GRNN hybrid model.

Conclusions: The ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model in predicting the short-term TB incidence in the Chinese population, especially in fitting and forecasting the peak and trough incidence.

Keywords: model, ARIMA, GRNN, tuberculosis, incidence, forecasting

Introduction

Tuberculosis (TB) is an ancient infectious disease that is caused by Mycobacterium tuberculosis (M. tb), and pulmonary tuberculosis is the most common type. According to the Global tuberculosis report 2018, there were approximately 10 million new cases and 1.57 million deaths due to TB in 2017. Globally, TB is the tenth leading cause of death and is the leading cause of death from a single infectious agent.1 Although the global TB incidence is declining at a rate of 2% per year, there is still a long way to go to reach the first milestone of the “End TB” strategy, due to the challenges of multidrug-resistant TB (MDR-TB), TB-HIV dual infection and high incidence of TB in the floating population.2–4 By collecting long-term TB morbidity data and selecting appropriate models, it is possible to predict the trend of TB epidemics, thereby anticipating possible outbreaks and guiding emergency preparedness at an early stage.

The autoregressive integrated moving average (ARIMA) model, which is also known as the Box-Jenkins model, was proposed by George Box and Gwilym Jenkins in the early
1970s. The ARIMA model assigns the combined effects of multiple risk factors that affect the disease occurrence and prevalence to time. This model has become one of the most popular and convenient models in time series analysis, and it has been widely used in the prediction of infectious diseases, such as malaria,\(^5\)–\(^9\) influenza,\(^10\)–\(^13\) hand-foot-mouth disease (HFMD),\(^14\)–\(^16\) hepatitis,\(^17\) dengue fever\(^18\) and hemorrhagic fever.\(^19\) Nevertheless, the ARIMA model has a limitation of preassumed linearity, which often does not conform to real-world problems, as the incidence of infectious diseases is affected by a variety of uncertain factors and usually exhibits nonlinear characteristics.\(^20\),\(^21\) Another commonly used model is the generalized regression neural network (GRNN). This model belongs to the artificial neural network (ANN) family, and its advantages include a small sample size, few parameters for artificial determination and strong nonlinear fitting ability. In recent years, the GRNN model has performed well in predicting epidemics either solely\(^22\) or in combination with the ARIMA model.\(^23\),\(^24\)

The ARIMA model has been used previously to predict the incidence of TB;\(^25\)–\(^27\) but the GRNN model and the ARIMA-GRNN hybrid model have rarely been studied. In the current study, we collected surveillance data of TB in a Chinese population and compared the predictive value of the ARIMA model and the ARIMA-GRNN hybrid model with the aim of providing a tool for decision making in the early warning system for managing TB.

**Materials and methods**

**Study area and data collection**

We selected Lianyungang as the study site. This city is located in the northeastern part of Jiangsu province in China with an area of approximately 7.6 thousand square kilometers and a permanent population of 4.52 million in 2017. Surveillance data of TB from January 2007 to December 2016 were extracted from the Lianyungang Center for Disease Control and Prevention (CDC). Population data were obtained from the Lianyungang Statistical Yearbook. We used the incidence data from January 2007 to June 2016 as the model-constructing dataset and the incidence data from July 2016 to December 2016 as the model-validating dataset.

**Construction of ARIMA model**

The ARIMA model is written in shorthand as ARIMA \((p, d, q) (P, D, Q)\)\(_s\), where \(p, d\) and \(q\) represent the autoregressive order, the number of nonseasonal differences and the moving average order, respectively, and \(P, D\) and \(Q\) represent the seasonal autoregressive order, the number of seasonal differences and the seasonal moving average order, respectively, and \(s\) indicates the length of the cyclical pattern. The construction of the ARIMA model included the following four steps: First, we determined that the series was nonstationary from the incidence series plot, which exhibited a long-term trend and seasonal fluctuations. The existence of a long-term trend was also demonstrated by the Mann-Kendall trend test. We applied one nonseasonal difference and one seasonal difference to stabilize the series, and the series was confirmed to be stationary through the difference analysis according to the Augmented Dickey–Fuller (ADF) test. Second, we identified optional parameters \((p, q, P, Q)\) to establish one or more alternative models by referring to the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the stationary series. We first determined the parameters of the seasonal part of the ARIMA model \((P, Q)\) and then defined the parameters of the nonseasonal part \((p, q)\). The model with the lowest corrected Akaike’s information criterion (AICc) and Bayesian information criterion (BIC) was regarded as the optimal model. Third, we used the maximum likelihood method to estimate the parameters and the Ljung-Box test to examine the residual series of the optimal model. The residual series should be white noise, demonstrating that the model completely extracted the information from the original data. Moreover, the ACF and PACF plots of the residual series should not show any significant correlation. Finally, the optimal model was applied to predict the TB incidence and compare this prediction with the validating dataset.\(^12\),\(^16\),\(^19\)

**ARIMA-GRNN hybrid model**

The GRNN, which is a branch of the artificial neural network (ANN), is a feedforward neural network based on the nonlinear regression theory. This network examines the relationship between each pair of the input vector \(X\) and the observed output \(Y\) and finally deduces the inherent function, which can be summarized with the following equation:

\[
E[Y/X] = \left( \int_{-\infty}^{\infty} Y f(X, Y)dY \right)/(\int_{-\infty}^{\infty} f(X, Y)dY),
\]

where \(X\) is the input vector \([X_1, X_2, \ldots, X_n]^T\) and \(Y\) is the predicted vector \([Y_1, Y_2, \ldots, Y_n]^T\) of GRNN. \(E[Y/X]\) is the expected value of the output \(Y\) with a given input vector \(X\), and \(f(X, Y)\) is the joint probability density of \(X\) and \(Y\).
The structure of GRNN consists of the following four layers: the input layer, pattern layer, summation layer and output layer. The number of neurons in the input layer is equal to the dimension of the input vector in the learning samples. Each neuron in the input layer is a simple distribution unit that can directly submit the input variables to the pattern layer. The number of neurons in the pattern layer is equal to the number of learning samples, and each neuron corresponds to a different sample. The transfer function of neurons of the pattern layer is calculated with the following equation:

$$P_i = e^{-rac{(x-x_i)^T(x-x_i)}{2\sigma^2}}, \quad i = 1, 2, \ldots, n,$$

where $P_i$ is the output of the neurons in the pattern layer, $X = [X_1, X_2, \ldots, X_n]^T$ is the input vector; $x_i$ is the learning samples of the $i$-th neurons; $n$ is the number of input; $i$ is the number of neurons; and $\sigma$ is the spread factor. There are two types of summation for the neurons of the summation layer. One type of summation is to arithmetically sum the output of neurons of the pattern layer. The weighted connection between the pattern layer and each neuron is 1. The formula of the transfer function is as follows: $S_D = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} e^{-rac{(x-x_i)^T(x-x_i)}{2\sigma^2}}, \quad i = 1, 2, \ldots, n.$

The other type of summation is to sum the weighted output of neurons of the pattern layer. The weighted connection between the $i$-th neuron of the pattern layer and the $j$-th neuron of the summation layer is equal to the $j$-th element in the $i$-th output samples of $Y_i$. The formula of transfer function is as follows:

$$S_{Nj} = \sum_{i=1}^{n} Y_{ij}P_i = \sum_{i=1}^{n} Y_{ij}e^{-rac{(x-x_i)^T(x-x_i)}{2\sigma^2}}, \quad j = 1, 2, \ldots, k,$$

where $k$ is the dimension of the output vector. The number of neurons of the output layer and the dimension of the input vector of the learning samples are the same. The formula of the output of the $j$-th neurons is as follows:

$$y_j = \frac{S_{Nj}}{S_D}, \quad j = 1, 2, \ldots, k.$$  

In contrast to the traditional error back propagation algorithm, the learning algorithm of GRNN does not need to adjust for the connection weight between neurons during the training process; rather, the GRNN changes the spread factor to adjust for the transfer function of each unit to obtain the training process; rather, the GRNN changes the spread factor to adjust for the connection weight between neurons during the training process. Two months were randomly selected as the testing samples to determine the optimal spread value. We let the spread value increase incrementally within a certain range, and then we calculated the root mean square error (RMSE) between the output incidence and the actual incidence of the two testing samples. The minimum RMSE value corresponded to the optimal spread value. Once the optimal spread value was identified, we used the forecasting incidence of the ARIMA model as the input values and applied the trained GRNN model to predict the future TB incidence.

**Comparison of the two models**

The indices including RMSE, mean absolute percentage error (MAPE), mean absolute error (MAE) and mean error rate (MER) were used to evaluate the performance of the models in fitting and forecasting the TB incidence in the study site.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2}.$$  

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_i - \hat{X}_i| + 100}{X_i}.$$  

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{X}_i|.$$  

$$\text{MER} = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_i - \hat{X}_i|}{X_i} \cdot X_i$$

is the actual incidence at time $i$, $\hat{X}_i$ is the fitting or forecasting incidence at time $i$, $\bar{X}_i$ is the mean of the actual incidence, and $n$ is the number of samples.
Statistical software
We used the packages of “forecast,” “ggplot2,” “trend” and “tsseries” of R3.5.1 (https://www.r-project.org/) to construct the ARIMA model, and we used the neural network toolbox of MATLAB R2017a (MathWorks, Massachusetts, USA) to construct the ARIMA-GRNN hybrid model.

Ethics statement
This study was approved by the Ethics Committee of Nanjing Medical University. After informed consent was obtained from all participants, questionnaires were used to collect demographic data.

Results
Construction of the ARIMA model
The monthly TB incidence during January 2007 and June 2016 in Lianyungang is shown in Figure 1. Based on the incidence series plot, we observed a slight declining trend ($P < 0.001$) and seasonal fluctuations ($s =12$). The peak incidence primarily occurred in March, April, May, November and December. The trough was more common in January and February. We applied both nonseasonal (d =1) difference and seasonal difference (D =1) to eliminate numerical instabilities. The plot of the original incidence series after one nonseasonal and one seasonal difference is shown in File S1, from which we preliminarily judged that the declining trend and seasonal fluctuations were eliminated. The ADF test remained significant ($P < 0.001$) after the nonseasonal and seasonal difference, indicating a stationary incidence series. The declining trend proved to be eliminated by the nonseasonal difference according to the Mann-Kendall trend test ($P =0.85$). The seasonal fluctuations were demonstrated to be eliminated by the seasonal difference according to the ACF and PACF plots; only lag 12 exhibited a significant spike in the ACF plot, whereas lag 12 and 24 exhibited no significant spikes in the PACF plot (Figure 2A). For the seasonal part of the ARIMA model, there was a significant spike at lag 12 in the ACF plot, but there was no significant spike at lag 24 ($Q =1$) in the ACF plot or at lag 12 or 24 in the PACF plot ($P =0$). For the nonseasonal part of the ARIMA model, in the first cycle, there were three significant spikes (lag 1, lag 10 and lag 11) in the ACF plot and four significant spikes (lag 1, lag 2, lag 10 and lag 11) in the PACF plot. We initially considered the following nine possibilities: p =0 and q =1; p =1 and q =0; p =1 and q =1; p =2 and q =0; p =2 and q =1; p =10 and q =0; p =11 and q =0; p =0 and q =10; p =0 and q =11, since the ACF and PACF plots did not show an obvious pattern. The AICc and BIC values and the Ljung-Box test results of residual series of these nine plausible ARIMA models are listed in Table 1. We selected ARIMA (10,1,0) (0,1,1)$_{12}$ as the optimal model, as this model had resulted in the minimum AICc and BIC values, indicating that the residual series represented white noise. The parameter estimation of this model is shown in Table 2. The ACF and PACF plots of residual series were also demonstrated to be white noise, since their correlation coefficients were not beyond

![Figure 1](https://example.com/figure1.png) Monthly reported TB incidence from January 2007 to June 2016 in Lianyungang.
the confidence borders (Figure 2B). Then, we applied the ARIMA \((10,1,0) (0,1,1)_12\) model to predict the TB incidence during July 2016 and December 2016. The predictive data are listed in Table 3.

![ACF and PACF plots](image)

**Figure 2** ACF and PACF plots. (A): The ACF and PACF plots of TB incidence series after the application of one nonseasonal difference and one seasonal difference; (B): The ACF and PACF plots of residual series of the ARIMA \((10,1,0) (0,1,1)_12\) model. **Abbreviations:** ACF, autocorrelation function; PACF, partial autocorrelation function.

### Table 1 AICc and BIC values and the Ljung-Box test results of residual series of plausible ARIMA models

<table>
<thead>
<tr>
<th>Model</th>
<th>AICc</th>
<th>BIC</th>
<th>(p^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA ((0,1,1) (0,1,1)_12)</td>
<td>226.99</td>
<td>234.59</td>
<td>0.02</td>
</tr>
<tr>
<td>ARIMA ((1,1,0) (0,1,1)_12)</td>
<td>239.70</td>
<td>247.30</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ARIMA ((1,1,1) (0,1,1)_12)</td>
<td>228.74</td>
<td>238.78</td>
<td>0.01</td>
</tr>
<tr>
<td>ARIMA ((2,1,0) (0,1,1)_12)</td>
<td>232.00</td>
<td>242.05</td>
<td>0.03</td>
</tr>
<tr>
<td>ARIMA ((2,1,1) (0,1,1)_12)</td>
<td>229.33</td>
<td>241.78</td>
<td>0.02</td>
</tr>
<tr>
<td>ARIMA ((10,1,0) (0,1,1)_12)</td>
<td>226.11</td>
<td>233.94</td>
<td>0.51</td>
</tr>
<tr>
<td>ARIMA ((11,1,0) (0,1,1)_12)</td>
<td>228.43</td>
<td>258.24</td>
<td>0.49</td>
</tr>
<tr>
<td>ARIMA ((0,1,10) (0,1,1)_12)</td>
<td>233.97</td>
<td>261.80</td>
<td>0.06</td>
</tr>
<tr>
<td>ARIMA ((0,1,11) (0,1,1)_12)</td>
<td>230.31</td>
<td>260.13</td>
<td>0.45</td>
</tr>
</tbody>
</table>

**Note:** *Ljung-Box test.*

**Abbreviations:** AICc, corrected Akaike's information criterion; BIC, Bayesian information criterion.

### Construction of the ARIMA-GRNN hybrid model

Due to the one nonseasonal difference and one seasonal difference, the information of thirteen months was lost in the construction of the ARIMA-GRNN hybrid model. The fitting incidence of the ARIMA model and the actual incidence between February 2008 and June 2016 were used as the input and output values of the GRNN model, respectively. We randomly selected data in March 2008 and January 2014 as the testing samples to determine the optimal spread value. We observed that the spread value gradually increased from 0.01 to 0.5 with an interval of 0.01; the corresponding RMSE values between the output and actual incidence of the two testing samples are shown in Figure 3. Since the RMSE reached the minimum value when the spread value was 0.23, we set the optimal spread value as 0.23. Eventually, we used the predicted incidence of ARIMA model as the input values and applied the...
Table 2 Estimation of parameters of the ARIMA (10,1,0) (0,1,1)_{12} model

<table>
<thead>
<tr>
<th>Model parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoregressive, lag 1</td>
<td>-0.7873</td>
<td>0.0895</td>
<td>8.7966</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Autoregressive, lag 2</td>
<td>-0.4982</td>
<td>0.1092</td>
<td>4.5623</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Autoregressive, lag 3</td>
<td>-0.3388</td>
<td>0.1133</td>
<td>2.9903</td>
<td>0.0035</td>
</tr>
<tr>
<td>Autoregressive, lag 4</td>
<td>-0.3657</td>
<td>0.1193</td>
<td>3.0654</td>
<td>0.0028</td>
</tr>
<tr>
<td>Autoregressive, lag 5</td>
<td>-0.4631</td>
<td>0.1280</td>
<td>3.6180</td>
<td>0.0005</td>
</tr>
<tr>
<td>Autoregressive, lag 6</td>
<td>-0.3039</td>
<td>0.1348</td>
<td>2.2545</td>
<td>0.0263</td>
</tr>
<tr>
<td>Autoregressive, lag 7</td>
<td>-0.3378</td>
<td>0.1393</td>
<td>2.4250</td>
<td>0.0171</td>
</tr>
<tr>
<td>Autoregressive, lag 8</td>
<td>-0.4064</td>
<td>0.1326</td>
<td>3.0649</td>
<td>0.0028</td>
</tr>
<tr>
<td>Autoregressive, lag 9</td>
<td>-0.4120</td>
<td>0.1213</td>
<td>3.3965</td>
<td>0.0010</td>
</tr>
<tr>
<td>Autoregressive, lag 10</td>
<td>-0.4196</td>
<td>0.0938</td>
<td>4.4733</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Seasonal moving average, lag 12</td>
<td>-1.0000</td>
<td>0.1995</td>
<td>5.0125</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 3 Predicted TB incidence by the ARIMA and ARIMA-GRNN hybrid models from July to December 2016

<table>
<thead>
<tr>
<th>Month</th>
<th>Observed incidence (1/100,000)</th>
<th>ARIMA model</th>
<th>ARIMA-GRNN hybrid model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted incidence (1/100,000)</td>
<td>Relative error (%)</td>
<td>Predicted incidence (1/1,000,000)</td>
</tr>
<tr>
<td>July</td>
<td>2.7805</td>
<td>2.6083</td>
<td>6.1931</td>
</tr>
<tr>
<td>August</td>
<td>2.6302</td>
<td>2.5434</td>
<td>3.3001</td>
</tr>
<tr>
<td>September</td>
<td>2.7429</td>
<td>2.6608</td>
<td>2.9932</td>
</tr>
<tr>
<td>October</td>
<td>2.3484</td>
<td>1.9059</td>
<td>18.8426</td>
</tr>
<tr>
<td>November</td>
<td>2.5551</td>
<td>3.0261</td>
<td>18.4337</td>
</tr>
<tr>
<td>December</td>
<td>2.8932</td>
<td>2.9949</td>
<td>3.5151</td>
</tr>
</tbody>
</table>

Figure 3 Selection of the optimal spread value for the ARIMA-GRNN hybrid model.
trained GRNN model to predict the incidence of TB from July 2016 to December 2016 (Table 3).

**Comparison of the two models**

The performance of the single ARIMA model and the ARIMA-GRNN hybrid model in fitting and forecasting TB incidence was compared based on the RMSE, MAPE, MAE and MER (Table 4). Although the hybrid model was slightly inferior to the single ARIMA model in forecasting the TB incidence of the last two months in 2016, generally speaking, the hybrid model performed better. As shown in Figure 4, the hybrid model was more accurate than the single ARIMA model, especially in terms of fitting and forecasting the peak and trough incidence.

**Discussion**

Although the incidence of TB has declined in recent years, China is still one of the thirty countries with the highest TB burden. Accurately predicting the incidence of TB is essential for policy-makers to make effective interventions in a timely manner and to allocate health resources reasonably. The early detection of the peak incidence is conducive to raising awareness. In this study, we used the incidence data of TB in a Chinese population to construct a predictive model, and we observed that the ARIMA-GRNN hybrid model was superior to the single ARIMA model. These findings indicate the potential value of the hybrid model in forecasting the short-term TB incidence in the study area.

TB incidence in the study site exhibited a downward trend from 2007 to 2016, which may be attributed to the government’s commitment to controlling TB, an increased budget and an improved public health system. We observed distinct seasonal fluctuations in the incidence of TB in this area. In contrast to certain areas that have one peak of incidence in the spring, this study site showed another peak in late autumn and early winter. Different seasonal characteristics of the TB incidence can be observed in different areas. For example, studies in Japan and Spain showed that spring was the season

<table>
<thead>
<tr>
<th>Diagnostic statistic</th>
<th>Fitting performance</th>
<th>Forecasting performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARIMA</td>
<td>ARIMA-GRNN</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.5594</td>
<td>0.5259</td>
</tr>
<tr>
<td>MAPE</td>
<td>11.5000</td>
<td>11.1281</td>
</tr>
<tr>
<td>MAE</td>
<td>0.4202</td>
<td>0.3992</td>
</tr>
<tr>
<td>MER</td>
<td>0.1132</td>
<td>0.1075</td>
</tr>
</tbody>
</table>

Abbreviations: RMSE, root mean square error; MAPE, mean absolute percentage error; MAE, mean absolute error; MER, mean error rate.

![Figure 4 Fitting and forecasting curves of the ARIMA and ARIMA-GRNN hybrid models and the actual reported TB incidence.](image-url)
with a higher TB incidence, whereas in the UK, the incidence peak was in the summer.

The ARIMA model is a combination of an autoregressive model and a moving average model, which can analyze both nonseasonal and seasonal time series. After logarithmic transformation and/or difference adjustment, even nonstationary time series can also be analyzed by the ARIMA model, which accounts for the capacity of this model to forecast disease. For example, Liu et al used the ARIMA (1,0,1) (0,1,0)_{12} model to predict the HFMD incidence in Sichuan province, China. Wangdi et al used the ARIMA (2,1,1) (0,1,1)_{12} model to predict malaria in endemic districts of Bhutan. Wang et al used the ARIMA (1,1,1) (1,1,0)_{12} model to predict influenza morbidity in Ningbo, China. Although the ARIMA model has a relative high prediction accuracy, it has limitations in processing nonlinear data. Alternatively, the GRNN model can make up for this disadvantage because of its powerful nonlinear fitting ability. The GRNN model is a specific form of radial basis function neural networks (RBFNN). First, the model’s network structure is relatively simple, including only two hidden layers (the pattern layer and the summation layer). Second, the model’s network training is notably easy. When the training samples pass through the hidden layer, the network training has already been completed, which avoids a long training time and high computational cost. Third, because of the simple network structure, it is not necessary to estimate or guess the number of hidden layers and hidden units of the network. In addition, only one free parameter (the spread) needs to be determined, which is the smoothing parameter of the RBF.

The occurrence of tuberculosis is usually affected by various factors, and it is often difficult to identify all the characteristics of the disease sequence. Simply using the linear or nonlinear model cannot extract adequate information. In the current study, we observed that the ARIMA-GRNN hybrid model has a higher prediction accuracy compared with the single ARIMA model, which is consistent with previous studies. Since the construction of the ARIMA-GRNN hybrid model is based on the ARIMA model, its prediction accuracy is affected by the performance of the ARIMA model. Adding more variables that affect TB transmission into the model can improve the accuracy of prediction models. Additionally, it is noteworthy that the ARIMA-based model is a short-term prediction model, and its application to predicting long-term trends needs to be performed with caution. In addition, the application value of this hybrid model needs to be demonstrated in additional studies in other areas.

In conclusion, the ARIMA-GRNN hybrid model was shown to be superior to the single ARIMA model in predicting the short-term TB incidence, especially in fitting and forecasting the peak and trough incidence.

Acknowledgments
This work was supported by the National Key R&D Program of China (2017YFC0907000), National Natural Science Foundation of China (81473027) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
All authors contributed to data analysis, drafting or revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure
The authors report no conflicts of interest in this work.

References


Supplementary material

Figure S1. Incidence series after the application of one nonseasonal difference and one seasonal difference.