A New Hybrid Model Using an Autoregressive Integrated Moving Average and a Generalized Regression Neural Network for the Incidence of Tuberculosis in Heng County, China

Wudi Wei,† Junjun Jiang,† Lian Gao, Bingyu Liang, Jiegang Huang, Ning Zang, Chuanyi Ning, Yanyan Liao, Jingzhen Lai, Jun Yu, Fengxiang Qin, Hui Chen, Jiming Su, Li Ye, and Hao Liang

Guangxi Key Laboratory of AIDS Prevention and Treatment and Guangxi Universities Key Laboratory of Prevention, Guangxi Medical University, 22 Shuangyong Road, Nanning, Guangxi, China; Department of Infectious Diseases, Heng County Centers for Disease Control and Prevention, 16 Gongyuan Road, Heng County, China; Life Sciences Institute, Guangxi Medical University, 22 Shuangyong Road, Nanning, China; Geriatrics Digestion Department of Internal Medicine, The First Affiliated Hospital of Guangxi Medical University, 6 Shuangyong Road, Nanning, China

INTRODUCTION

Guangxi Zhuang Autonomous Region has the second largest burden of tuberculosis cases in China. Tuberculosis has significant economic impact in Guangxi and may be an obstacle to Guangxi’s development. The tuberculosis epidemic of Heng County is particularly problematic in Guangxi. The morbidity of tuberculosis was significant, and ranked second in Nationally Notifiable Infectious Diseases of the country.

Heng County is a key area in the Guangxi Beibu Gulf Economic Zone, with a large transient population that promotes the spread of tuberculosis. Currently, the prevalence of human immunodeficiency virus/tuberculosis coinfections represent a serious new challenge to the public health system. In 2011, about 1.1 million (13%) of the 8.7 million people infected with tuberculosis worldwide were HIV-positive. In addition, it is estimated that about one-third of people who are infected with tuberculosis are latent carriers. Furthermore, multidrug-resistant tuberculosis is prevalent, especially in rural areas. To cope with the challenges to the public health system posed by the spread of tuberculosis, a series of preventative measures should be used in locations such as Heng County. Among these measures, a highly accurate epidemiologic model of the disease that can predict the trends in incidence that can help the public health authorities control the epidemic would be useful.

At this time, the existing mathematical models, specifically the autoregressive integrated moving average (ARIMA) model, are used to predict incidence of infectious diseases. The generalized regression neural network (GRNN) model, which is generally used to analyze the nonlinear data sets, could also be used. Combination models using the basic assumptions of both ARIMA and GRNN have been used for predicting disease incidence rates. The previous studies have also shown that a combined model may provide the advantages from both earlier models and increase forecasting accuracy. However, Wu and others found that the predictions of the ARIMA–GRNN model were not as good as those of the ARIMA model for the incidence of hemorrhagic fever with renal syndrome. Therefore, traditional ARIMA–GRNN models may not be useful for predicting the incidence of all infectious diseases, and the traditional hybrid models may be improved under some circumstances.

To our knowledge there is no research with respect to forecasting the incidence of tuberculosis using these statistical models. In this study, we developed ARIMA models, traditional ARIMA–GRNN models, and new ARIMA–GRNN models to develop an optimal model for tuberculosis incidence forecasting in Heng County. We expect that the model will play an important role in prevention and control of tuberculosis.

MATERIALS AND METHODS

Materials source. We obtained all the necessary data including the incidence of tuberculosis from January 2005 to December 2013 and population information from the Heng County Center for Disease Control and Prevention (CDC) and the Heng County Statistics Bureau. All tuberculosis cases must be reported to the Heng County CDC through the network system within 12 hours of diagnosis. Thus, the tuberculosis incidence data for Heng County are known to be reliable and complete because there is compulsory reporting. The original data can be gained from the corresponding author.

We divided the data into two parts so that we could compare the modeling and predictive performances of the two combined models and the ARIMA model. The incidence rates from...
January 2005 to June 2013 were used to fit the models, whereas the rest were used to validate the four models.

Any ethical considerations in this study were limited. No informed consent was required or was feasible because all the data used for the modeling came from publicly accessible secondary data sources, and no personal identifiers were entailed.

**ARIMA model construction.** The ARIMA (p,d,q) (P,D,Q)s model is a frequently used time series model that contains seven different parameters: 1) the order of autoregression (p), 2) the degree of difference (d), 3) the order of the moving average (q), 4) the seasonal autoregression lags (P), 5) the degree of seasonal difference (D), 6) the seasonal moving average lags (Q), and 7) the length of any cyclical pattern(s). We used the augmented Dickey-Fuller (ADF) unit root test to first determine the stationary value of the data. Second, log transformation and differencing were used to maintain the stationary time series if the sequence is not stationary.23 Finally, we used the Box–Jenkins strategy to construct the ARIMA models including three steps: identification, estimation, and diagnostic checking.25 The $R^2$, the Schwarz–Bayesian information criterion, and the Akaike information criterion were used to select the most appropriate model.24

**Construction of the traditional ARIMA–GRNN model.** The GRNN model is one of the various artificial neural network models that were developed by Specht18; it contains four layers, input layer, pattern layer, summation layer, and output layer. Compared with other networks, the GRNN model has strong ability for nonlinear mapping and good learning ability skills. The forecasting performance of the network is good and stable even if the sample size is small. Moreover, the construction of the GRNN model is straightforward and only has a regulation parameter, the smoothing factor. The equivalent nonlinear regression formula shows how the GRNN model works:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \exp\left(\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(\frac{D_i^2}{2\sigma^2}\right)}$$

where $X$ means the input vector $(X_1, X_2, \ldots, X_i)$, which consists of $n$ predictor variables; $Y$ denotes the output values predicted by the GRNN; $\hat{Y}(X)$ is the expected value of the output $Y$ given an input vector $X$; and $\sigma$ is the smoothing factor.26 Thus, the smoothing factor is very significant contributor in the GRNN model.

In this study, the forecasted incidence of the ARIMA model and the actual incidence were used as the input and output variables, respectively, to construct the traditional ARIMA–GRNN model. The smoothing factor is the only parameter of the model that directly affects the precision of the predictions. Specht18 proposed a method to find the optimal smoothing factor. We randomly selected two samples (each sample is likely to be selected, and the probability is the same) as the testing data to develop the GRNN model. Subsequently, the model was examined by a series of variables to find the optimal smoothing factor for which the minimum mean square error (MSE) of the model was the lowest. Finally, the

**Table 1**

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>SBC</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA(1,1)(2,1,1)$_2$</td>
<td>0.0811</td>
<td>0.2161</td>
<td>0.4734</td>
</tr>
<tr>
<td>ARIMA(0,1,0)(2,1,1)$_2$</td>
<td>0.1353</td>
<td>0.2357</td>
<td>0.4274</td>
</tr>
<tr>
<td>ARIMA(0,1,2)(2,1,2)$_2$</td>
<td>-0.0080</td>
<td>0.1927</td>
<td>0.5476</td>
</tr>
</tbody>
</table>

AIC = Akaike information criterion; ARIMA = autoregressive integrated moving average; SBC = Schwarz Bayesian information criterion.
predictions made by ARIMA model were used as the input values of the combined model, and then those values were optimized. These values were used as the forecasting values of the ARIMA–GRNN model. The flow chart of the traditional hybrid model is presented in the Supplemental Appendix 1.

Construction of the new ARIMA–GRNN model. Unlike traditional ARIMA–GRNN model construction, in the new hybrid models, the absolute error sequence from the ARIMA model is first generated with the formula $e_t = Y_t - A_t$, where $e_t$, $Y_t$, and $A_t$ denote the absolute error, the actual incidence, and the forecasting incidence, respectively. Subsequently, this absolute error sequence was used to construct the GRNN model. We used GRNN model that was developed to predict the absolute error of the next 6 months' tuberculosis incidence rate. Finally, we obtained the final forecasting data through the formula $H_t = Y_{t1} - e_{t1}$, where $H_t$, $Y_{t1}$, and $e_{t1}$ denote the final forecasting data, the actual incidence, and the predicted absolute error from GRNN model, respectively. The flow chart of the new hybrid model was presented in the Supplemental Appendix 2.

Construction of the GRNN model. The GRNN model with N-dimensional input and one-dimensional output was used to predict the absolute errors. As mentioned earlier, the smoothing factor directly affects the prediction precision of model. Thus, the two parameters (the N and the smoothing factor) play an important role in constructing the GRNN model. The values of N can be obtained graphically or via a series of tests. To determine the best value of the two parameters, the initial data in the set were used to develop the model, and the last two samples were used for data to test the model. We then developed several GRNN models, with a series of smoothing factors and values of N, to find the most suitable model in which the MSE was the lowest. The flow chart of the basic GRNN model is presented in the Supplemental Appendix 3.

Comparison with the four models for performance in simulations. We selected mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as model evaluation indexes to estimate the fitting and prediction effects of the four models. EViews 8.0 (IHS Global, Inc., Southeastern, PA) was used to create the ARIMA model, and the traditional and new ARIMA–GRNN models were performed with Matlab2012b (MathWorks, Natick, MA).

RESULTS

Basic ARIMA model. We used the monthly incidence from January 2005 to June 2013 to develop the model (Figure 1). The ADF test showed that the sequence was not stationary (ADF test: $t = -1.7075$, $P = 0.0830$). A log transformation, nonseasonal ($d = 1$), and seasonal difference ($D = 1$) were used to eliminate numerical instabilities. Finally, the time series was stationary (ADF test: $t = -12.7095$, $P < 0.0001$).

We developed a series of possible ARIMA models to determine the best version. After removing the models that did not pass the parameter or residual test, three models were
selected: ARIMA (1,1,0)(2,1,1)$_12$, ARIMA (0,1,0)(2,1,1)$_12$, and ARIMA (0,1,2)(2,1,2)$_12$. The parameters of the three models are shown in Table 1. We selected the ARIMA (0,1,2)(2,1,2)$_12$ model as the best one according to the results of the testing of parameters. The residual test of this model showed a white noise sequence ($P > 0.05$). The parameter test results are shown on Table 2.

Traditional ARIMA–GRNN model. The forecasting incidence values from the ARIMA model were selected as inputs of the GRNN model, and the actual incidence values were used for subsequent testing of the final model. To determine the best smoothing factor, the incidence of January 2005 and May 2013 were randomly used as the testing samples for the GRNN model. When the value of the smoothing factor was 0.004, the MSE of the model was lowest (Figure 2). Thus, we developed the GRNN in which the smoothing factor was 0.004. Then we used the constructed hybrid model to predict the disease incidence from July to December 2013.

New ARIMA–GRNN model. The absolute error sequence from February 2008 to June 2013 is shown in Figure 3. The absolute error sequence contained two periodic sequences: the sequence from February 2008 to December 2009 with a cycle of 6 months, and sequence from January 2010 to June 2013 with a cycle of 37 months. We selected the second sequence to construct the GRNN model to prevent interference from the first sequence. The absolute error of November 2012 and December 2012 were used as validation samples and the rest of these data were used to model the model. We chose the sequence number of the cycles as the $N$ value ($N = 37$). After many repeated tests, when the smoothing factor was 0.3, the model had the lowest MSE (Figure 4).

Basic GRNN model. We also used the monthly incidence rates from January 2005 to June 2013 to develop the GRNN model. The incidence rates from May to June 2013 were used to test the model, and the balance of the incidence data were used to refine the model. A smoothing factor of 0.1–3.0 with an interval of 0.1, and an $N$ value of 1–98 with an interval of 1, were selected to find the minimum MSE for the GRNN model. When the combinations of smoothing factor and $N$ value approached 1.6 and 26, respectively, the minimum MSE of the GRNN model was the lowest. Then the developed GRNN model was used to forecast the incidence of next 6 months.

Comparison with the four models in simulation performance. Four models were selected to forecast the tuberculosis incidence in Heng County from July 2013 to December 2013 (Table 3). The modeling and forecasting curves of the four models are shown in Supplemental Figure 1 and Figure 5. The performance parameters of the four models in fitting and validation are shown in Table 4.
DISCUSSION

Until now, there were no validated prediction models for tuberculosis incidence in Heng County. In this study, we developed a basic ARIMA model, a basic GRNN model, a traditional ARIMA–GRNN model, and a new ARIMA–GRNN model to fit the data and used these results to then predict the tuberculosis incidence rate. The MSE, MAE, and MAPE of new ARIMA–GRNN model were the lowest in forecasting. Interestingly, although the values of the three parameters of the traditional ARIMA–GRNN model were lower than the basic ARIMA model in the fitting stage, the forecasting parameters were the highest, results that are similar to those of previous research.24 Although the basic GRNN model had apparent perfect fitting ability, the forecast effect from that model was not very good. Thus, the basic GRNN model was overfitting. This is a congenital defect and up to now we do not have a good solution to fix it.

The effect of new ARIMA–GRNN model was better than basic ARIMA model. In this new hybrid model, the liner information was first extracted from the data by the ARIMA model. Unlike the traditional hybrid model, we directly focused on a residual sequence that contained the nonlinear information. An improved method was used to construct the GRNN model. It gave full consideration to the two parameters (the value of N and the smoothing factor) and developed the network with a 37-dimensional input and one-dimensional output to extract the information from the residual sequence. Previous studies only focused on the smoothing factor, which is hard to find the most suitable model.20–23,25 In this study, the traditional hybrid model with a one-dimensional input and a one-dimensional output could not make full use of the original information and might generate bias. With the new ARIMA–GRNN combined model, the linear and nonlinear information were extracted adequately. The performance of the fitting part and forecasting part of the model obviously improved.

To our knowledge, this is the first study to develop a mathematical model for Heng County’s tuberculosis incidence forecasting. Moreover, unlike in other regions, the tuberculosis incidence curve of this county showed some distinctive characteristics. For instance, seasonal patterns were not obvious and there was an obvious peak incidence in 2008. The specialization may make the traditional ARIMA–GRNN models unsuitable for prediction. Therefore, the new ARIMA–GRNN model is a powerful tool that may allow the Heng County public health authorities to reduce the transmission and control outbreaks of tuberculosis. Only when health policy makers can understand incidence trends of tuberculosis can they distribute the necessary resources properly and organize effective medical services to promote the health of the population of the county. If the ARIMA–GRNN results indicate that the incidence of tuberculosis will increase, more health resources can be mobilized in advance. Thus, the local government will be able to take the initiative in the struggle against tuberculosis outbreaks with the help of the new model. Moreover, the novel ARIMA–GRNN model can also evaluate the effects of new intervention measures, such as the use of a tuberculosis vaccine, directly observed antimicrobial treatment, or the effects of short courses of antibiotic combinations, and/or other interventions. The measures will be shown to have been effective when the forecasted values are higher than the actual values that are observed after

![FIGURE 5. The forecasting curves of the four models and the actual tuberculosis incidence series. This figure appears in color at www.ajtmh.org.](image)

<table>
<thead>
<tr>
<th>Prediction error</th>
<th>Fitting part</th>
<th>Validation part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Basic ARIMA</td>
<td>1.5692</td>
<td>1.0396</td>
</tr>
<tr>
<td>Traditional ARIMA–GRNN</td>
<td>0.5993</td>
<td>0.4007</td>
</tr>
<tr>
<td>New ARIMA–GRNN</td>
<td>0.0007</td>
<td>0.0007</td>
</tr>
<tr>
<td>Basic GRNN</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

ARIMA = autoregressive integrated moving average; GRNN = generalized regression neural network; MAE = mean absolute error; MAPE = mean absolute percentage error; MSE = the mean square error.
an intervention has been tested in the population. In general, the new combined model should make a positive contribution to tuberculosis control measures and disease prevention in Heng County. After being documented, these benefits may be extended to other regions.

Several limitations of the new hybrid model are possible. First, it should be updated frequently to prevent the precision of its predictions from eroding and becoming unacceptably low, because it is a short-time forecast model for incidence rates, and can only predict incidence rates for the succeeding 6 months. When predicting 1 year, the result is not satisfactory (Supplemental Table 1). Second, it is a single-factor model that only analyzes the time factor. Other complex elements affecting the emergence and transmission of tuberculosis were not extracted by the model. As shown in Figure 5, the prediction error for the month of September was larger than for the others, and this may have been caused by unknown influences or factors, and the latest data (2014–2016) are unavailable at present due to confidentiality and sensitivity issues. Finally, our study is only focused on tuberculosis cases in Heng County. Further research will be needed to confirm whether the new ARIMA–GRNN hybrid model is suitable for other regions or other diseases.

CONCLUSIONS

In summary, traditional ARIMA–GRNN models could not evaluate the data on tuberculosis epidemiology in Heng County effectively. The new ARIMA–GRNN model is a potentially unique new tool to assist the relevant public health authorities working in Heng County to better control the ongoing epidemic of tuberculosis.

Received August 8, 2016. Accepted for publication June 5, 2017.

Note: Supplemental appendices, figure, and table appear at www.ajtmh.org.

Acknowledgments: We would like to express our gratitude to all the staff from the Heng County Center for Disease Control and Prevention in Guangxi, China, for providing the epidemiological data about the tuberculosis cases they had tabulated.

Financial support: The study was supported by National Natural Science Foundation of China (NSFC, 81560326, 81460305, 31360033, 81460511), Guangxi Scientific and Technological Development Project (Gui Ke Gong 14124003-1), Key Project of Guangxi Universities Scientific Research (20132D012, YB2014062), Guangxi University “100-Talent” Program & Guangxi university innovation team and outstanding scholars program (Gui Jiao Ren 2014[7]).

Authors’ addresses: Wudi Wei, Junjun Jiang, Bingyu Liang, Jiegang Huang, Jingzhen Lai, Jun Yu, Fengxiang Qin, and Jiming Su, Guangxi Medical University, Guangxi Key Laboratory of AIDS Prevention and Treatment and Guangxi Universities Key Laboratory of Prevention and Control of Highly Prevalent Disease, Guangxi Medical University, Nanning, Guangxi, China, E-mails: yeli@gxmu.edu.cn and lianghao@gxmu.edu.cn.

REFERENCES


