BMJ Open

Prediction model comparison for hemorrhagic fever with renal syndrome

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-025773
Article Type:	Research
Date Submitted by the Author:	01-Aug-2018
Complete List of Authors:	Wang, Ya-wen; Chinese Academy of Medical Sciences / Peking Union Medical College, School of Public Health Shen, Zhong-zhou JIANG, Yu; Chinese Academy of Medical Sciences and Peking Union Medical College
Keywords:	autoregressive integrated moving average, generalized regression neural network, hemorrhagic fever with renal syndrome, prediction

Prediction model comparison for hemorrhagic fever with renal syndrome

First author:
Ya-wen Wang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-17810259300
E-mail: ywwang2099@163.com

Second author:
Zhong-zhou Shen
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-18310017094
E-mail: szz90123@163.com

Corresponding author:
Professor Yu Jiang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-13693271887
E-mail: jiangyu@pumc.edu.cn

Key words: autoregressive integrated moving average; generalized regression neural network; hemorrhagic fever with renal syndrome; prediction

Word count: 2818 words.

ABSTRACT

Objectives Hemorrhagic fever with renal syndrome (HFRS) is a serious public health threat in China, accounting for almost 90% cases reported globally. Infectious disease prediction may help in disease prevention despite some uncontrollable influence factors. This study conducted a comparison between a hybrid model with two single models in forecasting the monthly incidence of HFRS in China.

Design Time-series study.

Setting The People's Republic of China

Methods Autoregressive integrated moving average (ARIMA) model, generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model were constructed. The incidence data from January 2011 to December 2017 were adopted to test models' fitting performance. Data from January 2018 to May were used to demonstrate the models' forecasting performance. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were adopted to evaluate these models' effectiveness.

Results The incidence of HFRS in the past seven years was characterized by a slight declining trend and obverse seasonal variation. The ARIMA(2,1,1)(2,1,1)12 model was selected as the optimal model in HFRS fitting. The smooth factor of the basic GRNN model and the hybrid model was 0.03 and 0.043 respectively. The hybrid model was the best in disease forecasting but was underperform in fitting part.

Conclusion The hybrid ARIMA-GRNN model was better than single ARIMA and basic GRNN model in forecasting monthly incidence of HFRS in China. It could be considered as a decision-making tool in HFRS prevention and control.

Strengths and limitations of this study

- This study examined the forecasting performances of autoregressive integrated moving average (ARIMA) model and generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model in forecasting incidence of hemorrhagic fever with renal syndrome (HFRS) in China, as a reference to choose suitable model in infectious disease prediction.
- The reported data we collected may slightly differ from the actual incidence number since reported data came from monitor, it may not include the person who was infected but not tested.
- Many factors could influence the incidence of an infectious disease. But only time factor in study period was considered in our models which may increase forecast error. Thus data should be updated to maintain the model's accuracy.

BACKGROUND

Hantaviruses, a member of family *Bunyaviridae*, contains the most important zoonotic pathogens of humans.[1] Two categories of hantaviruses are Old World (Asia and Europe) virus that cause hemorrhagic fever with renal syndrome (HFRS), and New World (Americas) virus that causes hantavirus pulmonary syndrome (HPS).[2,3] Hantaviruses are spread through the infected mammals' urine, faces, and saliva. People can be infected mainly through respiratory tract, alimentary tract and skin/mucus membrane abrasion. The onset symptoms of HFRS are fever, circulatory collapse with hypotension, hemorrhage and acute kidney injury (AKI).[4,5] The hallmark of HFRS is

capillary leak syndrome, which causes edema and hemorrhage, even threaten people's life.[6,7] Cases of HFRS are widely distributed in eastern Asia, particularly in China, Russia and Korea.[8] It is reported that the number of HFRS human cases in China accounts for almost 90% of the total cases worldwide.[9,10] Some comprehensive control activities such as effective vaccine and rodent elimination have achieved remarkable results, while the incidence of HFRS is still high owing to some uncontrollable factors.[11,12] Thus it is important to forecast the diseases trends and take some control measure.

Statistic models such as linear regression, artificial neural network and grey model have been widely used in infectious diseases forecasting. Reliable forecasting plays an important role in diseases control before pandemic or outbreak. The autoregressive integrated moving average(ARIMA) model is one of the most popular methods in diseases prediction. The principle of the model contains filtering out the high-frequency noise in the data, detecting local trends based on liner dependence and forecasting the develop trends. The limitation of this model is that ARIMA can only analyse the liner part of infectious disease series.[13] However, the non-linear part of epidemic data may not be white noise, which means some information may be lost by ARIMA model. To overcome the inherent defect of ARIMA model, an artificial neural network (ANN) model was adopted. ANN is a conceptualized mathematical non-linear classification model inspired by the behavior of biological networks of neurons.[14,15] The generalized regression neural network (GRNN) is a member of ANN family with unique characteristics of accelerated learning and greater capability for non-linear fitting. The hybrid ARIMA-GRNN model has both advantages of ARIMA model and GRNN model, which means that both the linear part and non-linear part of time series are fitted.

There are some researches showing that the hybrid model provides better incidence forecasting performance than single ARIMA model and basic GRNN model in some infectious diseases, while which model is the best in predicting the incidence of HFRS in China is still unclear. Besides, some studies had compared the forecasting performance of hybrid ARIMA-GRNN model with other models.[16] But the comparison between hybrid model with two single models in HFRS prediction had not been found. This study aims to develop a single ARIMA model, a basic GRNN model and a hybrid ARIMA-GRNN model to predict the monthly incidence of HFRS in China. The fitting and forecasting performance of these three models were compared to determine the best one, which is suggested to be employed in the provision of reference information for HFRS control.

METHODS

Materials Source

The monthly incidence data of HFRS in China from January 2011 to May 2018 were collected from the official website of National Health Commission of the People's Republic of China (Ministry of Health). All HFRS cases must be reported to the National Health Commission through the Infectious Disease Surveillance System within 24 hours. The data was separated into two parts: model building and model forecasting.

Single ARIMA Model

The ARIMA model is usually shown as ARIMA (p, d, q)$(P,D,Q)_S$ while the parameters mean non-seasonal and seasonal order of auto-regression, the degree of difference and moving average

respectively. Besides, the subscript means the length of cyclical pattern. An ARIMA model is developed by time series stationary, parameter estimation and model check.[17]

Time stationary means no fluctuation or periodicity with time goes by. The Augmented Dickey-Fuller (ADF) unit-root test could help estimating whether the time series is stationary or not. Log transformation and differences are frequently adopted to stabilize the time series.

The parameters of p, q, P and Q are determined through the autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph. D is the length of seasonal difference and d is the length of trend difference, these two parameters are determined when original series is stationaried. Generally, more than one plausible models could be combined.

Since the best model must has the highest accuracy in disease prediction, some substandard models are removed. A suitable model must show statistical significance in parameter test and a white noise sequence in residual test. Besides, the best model should have the lowest Akaike information criterion (AIC) value than other combined models.

Basic GRNN Model

The GRNN model is built on the basis of non-liner regression theory. The input layer, pattern layer, summation layer and output layer are involved in the construction of GRNN model.[18] Its inherent function is to identify the relationship between each input value and output value. Smoothing factor is the only parameter of GRNN which means the network could not be affected by human factors. Generally, there are more than one possible values of smoothing factor and the best one should be determined to build an optimal GRNN model.

Initially, the original data are divided into two parts, the training set and the test set. The test set is the last two data or two random data of original series, the rest as the training set. Then the training network was tested for a series of smoothing factors. And following that, the basic GRNN model is established with the best smoothing factor, which must have the lowest root mean square error (RMSE). Finally, all the original data were adopted as input part to predict the future data by the best GRNN model.

Hybrid ARIMA-GRNN Model

The ARIMA model has advantage in extracting and fitting the linear part of the original time series, while the non-liner information in residuals is abandoned. GRNN model is combined for it can analyze the non-liner information and mine the information adequately. The hybrid ARIMA-GRNN model is developed to demonstrate if it has the highest accuracy in HFRS incidence prediction.

In the development of the hybrid model, the input variable is the fitting data of ARIMA model while the output variable is the actual data. Same with the basic GRNN model, the last two samples or two randomly selected samples of original series are set as testing set and the rest are set as training set to find optimal smoothing factor. The smoothing factor must have the minimum RMSE. Finally, the forecasted values of ARIMA model is used as the input data of hybrid model to get the output predictive values.

Model Comparison

The forecasting effects of ARIMA model, GRNN model and hybrid ARIMA-GRNN model are estimated with root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).[19] Excel 2016 was used to build the database, R 3.4.3 software was used to create the ARIMA model, the Matlab R2016a was used to create the basic GRNN model and hybrid ARIMA-GRNN model.

Ethics

Since no primary data collection was undertaken, no patient or public was involved, no formal ethical assessment or informed consent was required.

RESULTS

Single ARIMA model

The monthly incidence data of HFRS in China from January 2011 to December 2017 was used to develop the ARIMA model (Fig 1). As shown in the original time series graph, the HFRS incidence shows seasonal variation (s=12) and a slightly declining trend, which means the time series was not stationary. Trend difference (d=1) and seasonal difference (D=1) were dene to eliminate the instability. The ADF test showed that the differenced time sequence was stationary (t statistics was -4.7201, P=0.01).

Figure 1 Monthly incidence of HFRS in China from January 2011 to December 2017.

The ACF graph and PACF graph (Fig 2) were applied to explore the parameters of the ARIMA model. Four appropriate models were chosen by residual test and were filtered by AIC values. The AIC values of ARIMA$(1,1,1)(1,1,1)_{12}$ ARIMA$(1,1,1)(2,1,1)_{12}$ ARIMA$(2,1,1)(1,1,1)_{12}$ ARIMA$(2,1,1)(2,1,1)_{12}$ were 950.48, 944.68, 940.55, 936.61 respectively. The ARIMA$(2,1,1)(2,1,1)_{12}$ model had the lowest AIC value and was chosen as the most suitable model. The residual test showed white noise (Fig 3).

Figure 2 The ACF and PACF graphs of differenced HFRS incidence series.
Figure 3 Residual white noise test

Basic GRNN model

The samples from January 2011 to December 2017 were adopted to develop the network. The last two samples were used as testing samples while the others were training samples. To determine the optimal smoothing factors, a series of smoothing factors were tested. The smoothing factor with the minimum RMSE of the network was selected as the optimal one. Fig 4 shows the RMSE of these smoothing factors. As shown in Fig 4, the optimal smoothing factor of the one-dimensional input and one-dimensional output GRNN model was 0.03.

Figure 4 The selection of basic GRNN model and hybrid ARIMA-GRNN model

Hybrid ARIMA-GRNN model

The fitted data of ARIMA model from January 2011 to December 2017 were used as the input samples for the GRNN model. The actual HFRS values were used as the output samples to training the hybrid ARIMA-GRNN model. The RMSE of hybrid model was the lowest when the smoothing factor was 0.043 (Fig 4). Thus 0.043 was selected to develop the GRNN model. Subsequently, the forecasting outcomes of ARIMA model from January 2018 to May 2018 were selected as the entry value of the GRNN model, and the output values were the predictive values of the combined ARIMA-GRNN model.

Finally, these three models were adopted to forecast the HFRS incidence in China from January to May 2018. The forecasting performance parameters of the three models for the fitting and forecasting parts are shown in Table 1. The fitting and forecasting curves of the three models and the actual HFRS incidence series are depicted in Fig 5.

Table 1. The fitting and forecasting performance of the three models.

	Fitting part			Forecasting part		
Predicting error	MAPE	MAE	RMSE	MAPE	MAE	RMSE
ARIMA	9.1154	89.0302	138.8356	21.0212	175.7042	220.6269
GRNN	10.7332	134.596	265.7046	19.2029	177.0356	202.1684
ARIMA-GRNN	9.6083	85.0429	140.6426	17.8026	152.3013	196.4682

Figure 5 The fitting and forecasting curves of the three models and the actual HFRS incidence series

DISCUSSION

In this study, a hybrid model was constructed based on traditional ARIMA model and basic GRNN model. Three different models were compared in forecasting performance and the results showed that the hybrid ARIMA-GRNN model was the optimal model in predicting the monthly incidence of HFRS in China. Thus we consider the hybrid model as a decision-making tool to give some suggestion in public health policy decision.

The characteristic of monthly incidence of HFRS in China is suitable for ARIMA model and GRNN model. As shown in the result part, the incidence of HFRS in China has a slight decreasing trend and a bimodal seasonal cases distribution, which are same with other studies.[20,21] The incidence reaches a peak in winter rapidly and has a longer lasting peak in Spring. Autumn to winter peak is the other peak, which is lower than the first one. People are more likely to be exposed to the disease for increased activities in these two seasons, rodent behavior changes with climate change.[22,23] Besides, these two peaks, including the peak values, might vary with different hantaviruses types.

The hybrid ARIMA-GRNN model was superior among three models even with imperfect fitting performance. ARIMA model is one of the most commonly used methods in infectious diseases prediction and has been proved with high accuracy. In this study, the traditional ARIMA model was used as the baseline model for evaluating the performance of other models. The results showed that single ARIMA model and basic GRNN model are better than hybrid model in data fitting because of lower MAE and MAPE. It is recognized to be used in the prediction incidence of infectious disease. Even some unmeasurable factors may impact data fitting, the forecasting performance should be at the first consideration.[18] The MAPE, MAE, RMSE of hybrid model in validation part were lower than single ARIMA model or basic GRNN model. Some studies showed that hybrid ARIMA-GRNN model had less error than single model both in modeling and forecasting stage, which is different with our study. These studies built the hybrid model with tuberculosis incidence or hand-foot-mouth disease incidence.[24,25] so we hypothesis that diseases characteristics may affect the model performance.

The basic GRNN model was developed as a new potential tool for infectious diseases incidence prediction in recent years. In this study, one-dimensional input and one-dimension output GRNN model was built and same as the hybrid model. The fitting error of GRNN model is

higher than the other two model, but the forecasting error just higher than the hybrid model. Here also some studies showed that basic GRNN model performs better than ARIMA model in disease prediction.[16] Same as other disease prediction model, the disease control department could assess the disease developing trend with the help of the hybrid ARIAM-GRNN model. In a short term, the prediction values have same trend with the actual values. It means if the predictive values continue to rise, an outbreak should be alerted. Besides, disease prediction model is developed to evaluate the effectiveness of diseases intervention strategies like vaccine. An effective control measure will make the actual values lower than the predicted results. Something noteworthy is that these two functions are based on short terms. The incidence of infectious disease is influenced by some uncontrollable factors, like HFRS is infected by weather, climate, human activities and so on.[26-28] These factors may keep stable in a short period and might change in a long run.

Several limitations of this study should be noted. As is shown above, the prediction model was merely developed for short-term forecasting. Maintaining the prediction performance for months or years requires constantly update of data and model. Besides, this study only considered the incidence of HFRS in China, weather the hybrid model is suitable for other countries is still unclear and larger samples are needed to test.

Conclusions

The hybrid ARIMA-GRNN model is superior than the single ARIMA model and basic GRNN model both in fitting part and forecasting part in monthly incidence of HFRS in China. The data should keep update to maintain the forecasting performance. This hybrid model should be considered as a decision-making tool in HFRS prevention and control.

Supporting information

S1 Table. The data of HFRS incidence in China from January 2011 to May 2018.

Contributors YJ, YW and ZS designed the study. ZS extracted the data and constructed the database. YW and ZS analyzed the data. YW drafted the manuscript. YJ and ZS made critical revision to the manuscript. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data available.

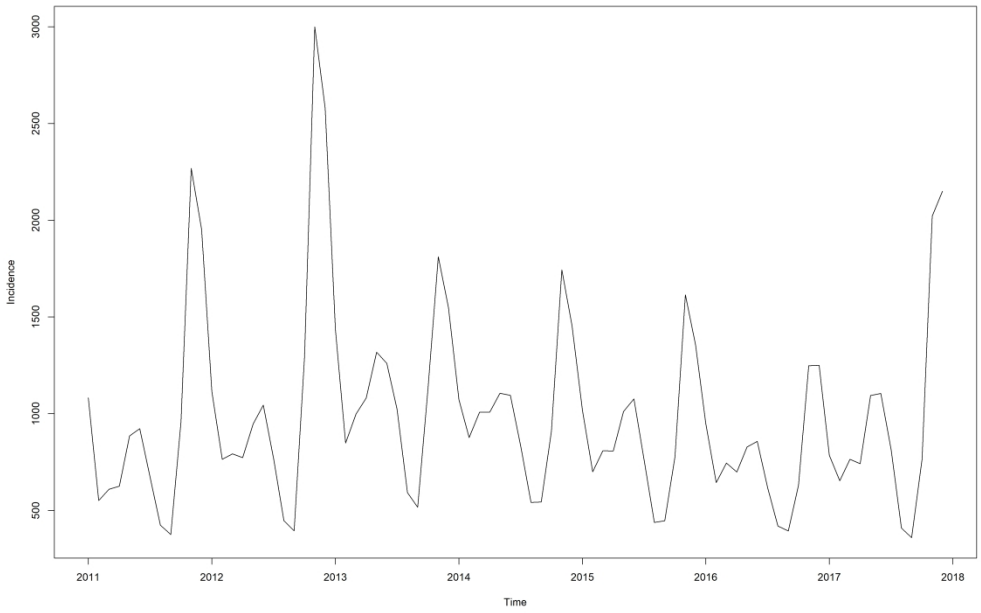
REFERENCES

1. Schmaljohn CS, Hasty SE, Dalrymple JM, LeDuc JW, Lee HW, von Bonsdorff CH, et al. Antigenic and genetic properties of viruses linked to hemorrhagic fever with renal syndrome. Science. 1985;227:1041-1044.
2. Ehelepola NDB, Basnayake BMLS, Sathkumara SMBY, Kaluphana KLR. Two Atypical Cases of Hantavirus Infections from Sri Lanka. Case Reports in Infectious Diseases. 2018;2018:1-6. doi: 10.1155/2018/4069862.
3. Avsic-Zupanc T, Saksida A, Korva M. Hantavirus infections. Clin Microbiol Infect. 2015;1-11. doi: 10.1111/1469-0691.12291.

4. Latus J, Schwab M, Tacconelli E, Pieper FM, Wegener D, Dippon J, et al. Clinical course and long-term outcome of hantavirus-associated nephropathia epidemica, Germany. Emerg Infect Dis. 2015;21:76-83. doi: 10.3201/eid2101.140861.
5. Vaheri A, Strandin T, Hepojoki J, Sironen T, Henttonen H, Makela S, et al. Uncovering the mysteries of hantavirus infections. Nat Rev Microbiol. 2013;11:539-550. doi: 10.1038/nrmicro3066.
6. Hepojoki J, Vaheri A, Strandin T. The fundamental role of endothelial cells in hantavirus pathogenesis. Front Microbiol. 2014;5:727. doi: 10.3389/fmicb.2014.00727.
7. Pal E, Korva M, Resman RK, Kejzar N, Bogovic P, Kurent A, et al. Sequential assessment of clinical and laboratory parameters in patients with hemorrhagic fever with renal syndrome. Plos One. 2018;13:e197661. doi: 10.1371/journal.pone.0197661.
8. Bi P, Parton KA. El Niño and incidence of hemorrhagic fever with renal syndrome in China. JAMA. 2003;289:176-177.
9. Zhang S, Wang S, Yin W, Liang M, Li J, Zhang Q, et al. Epidemic characteristics of hemorrhagic fever with renal syndrome in China, 2006-2012. Bmc Infect Dis. 2014;14:384. doi: 10.1186/1471-2334-14-384.
10. Du H, Wang PZ, Li J, Bai L, Li H, Yu HT, et al. Clinical characteristics and outcomes in critical patients with hemorrhagic fever with renal syndrome. Bmc Infect Dis. 2014;14:191. doi: 10.1186/1471-2334-14-191.
11. Zhang WY, Wang LY, Liu YX, Yin WW, Hu WB, Magalhaes RJ, et al. Spatiotemporal Transmission Dynamics of Hemorrhagic Fever with Renal Syndrome in China, 2005−2012. Plos Negl Trop Dis. 2014;8:e3344. doi: 10.1371/journal.pntd.0003344.
12. He X, Wang S, Huang X, Wang X. Changes in age distribution of hemorrhagic fever with renal syndrome: An implication of China's expanded program of immunization. Bmc Public Health 2013, 13:394. doi: 10.1186/1471-2458-13-394.
13. Petukhova T, Ojkic D, McEwen B, Deardon R, Poljak Z. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza a virus frequency in swine in Ontario, Canada. Plos One 2018, 13(6):e198313. doi: 10.1371/journal.pone.0198313.
14. Yosipof A, Guedes RC, Garcia-Sosa AT. Data mining and machine learning models for predicting drug likeness and their disease or organ category. Front Chem. 2018;6:162. doi: 10.3389/fchem.2018.00162.
15. Nair TM. Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia. Comput Biol Chem. 2018;75:222-230. doi: 10.1016/j.compbiolchem.2018.05.011.
16. Wu W, Guo J, An S, Guan P, Ren Y, Xia L, et al. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in jiangsu province, china. Plos One. 2015;10:e135492. doi: 10.1371/journal.pone.0135492.
17. Rubaihayo J, Tumwesigye NM, Konde-Lule J, Makumbi F. Forecast analysis of any opportunistic infection among HIV positive individuals on antiretroviral therapy in Uganda. Bmc Public Health. 2016;16:766. doi: 10.1186/s12889-016-3455-5.
18. Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, et al. Application of a combined model with autoregressive integrated moving average (ARIMA) and generalized regression neural network (GRNN) in forecasting hepatitis incidence in heng county, china. Plos One. 2016;11:e156768. doi:

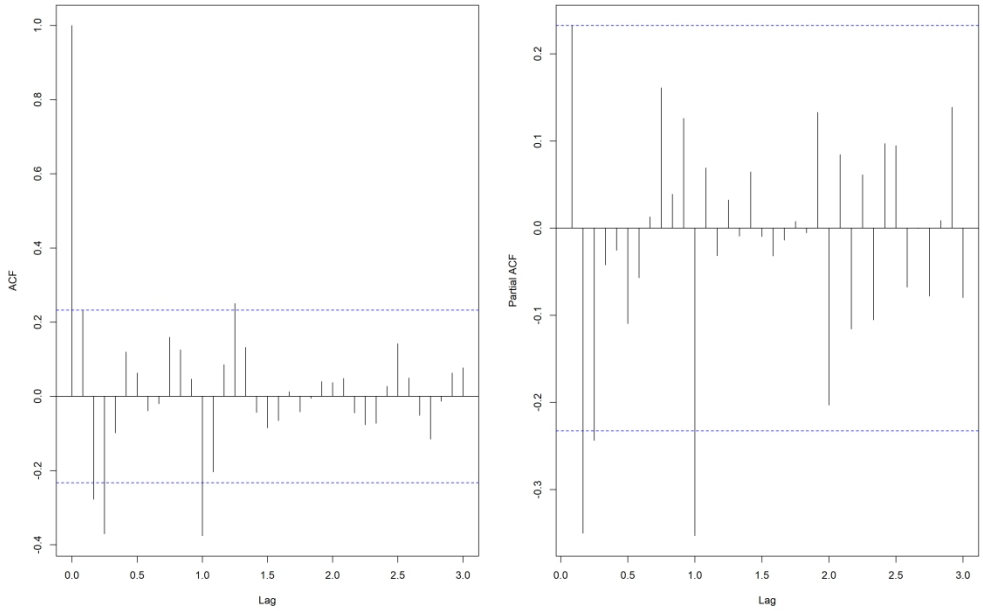
10.1371/journal.pone.0156768.

19. Gan R, Chen N, Huang D. Comparisons of forecasting for hepatitis in Guangxi Province, China by using three neural networks models. Peerj. 2016;4:e2684. doi: 10.7717/peerj.2684.
20. Hansen A, Cameron S, Liu Q, Sun Y, Weinstein P, Williams C, et al. Transmission of haemorrhagic fever with renal syndrome in china and the role of climate factors: A review. Int J Infect Dis. 2015;33:212-218. doi: 10.1016/j.ijid.2015.02.010.
21. Liu YX, Feng D, Zhang Q, Jia N, Zhao ZT, De Vlas SJ, et al. Key differentiating features between scrub typhus and hemorrhagic fever with renal syndrome in northern China. Am J Trop Med Hyg. 2007;76:801-805.
22. Park YH. Absence of a seasonal variation of hemorrhagic fever with renal syndrome in yeoncheon compared to nationwide korea. Infect Chemother. 2018;50:120-127. doi: 10.3947/ic.2018.50.2.120.
23. Mills JN, Gage KL, Khan AS. Potential influence of climate change on vector-borne and zoonotic diseases: A review and proposed research plan. Environ Health Perspect. 2010;118:1507-1514. doi: 10.1289/ehp.0901389.
24. Wang H, Tian CW, Wang WM, Luo XM. Time-series analysis of tuberculosis from 2005 to 2017 in China. Epidemiol Infect. 2018;146:935-939. doi: 10.1017/S0950268818001115.
25. Peng Y, Yu B, Wang P, Kong DG, Chen BH, Yang XB. Application of seasonal auto-regressive integrated moving average model in forecasting the incidence of hand-foot-mouth disease in Wuhan, China. J Huazhong Univ Sci Technolog Med Sci. 2017;37:842-848. doi: 10.1007/s11596-017-1815-8.
26. Joshi YP, Kim E, Cheong H. The influence of climatic factors on the development of hemorrhagic fever with renal syndrome and leptospirosis during the peak season in Korea: An ecologic study. Bmc Infect Dis. 2017;17:406. doi: 10.1186/s12879-017-2506-6.
27. Han SS, Kim S, Choi Y, Kim S, Kim YS. Air pollution and hemorrhagic fever with renal syndrome in South Korea: An ecological correlation study. Bmc Public Health. 2013;13:347. doi: 10.1186/1471-2458-13-347.
28. Xiang J, Hansen A, Liu Q, Tong MX, Liu X, Sun Y, et al. Impact of meteorological factors on hemorrhagic fever with renal syndrome in 19 cities in China, 2005-2014. Sci Total Environ. 2018;636:1249-1256. doi: 10.1016/j.scitotenv.2018.04.407.



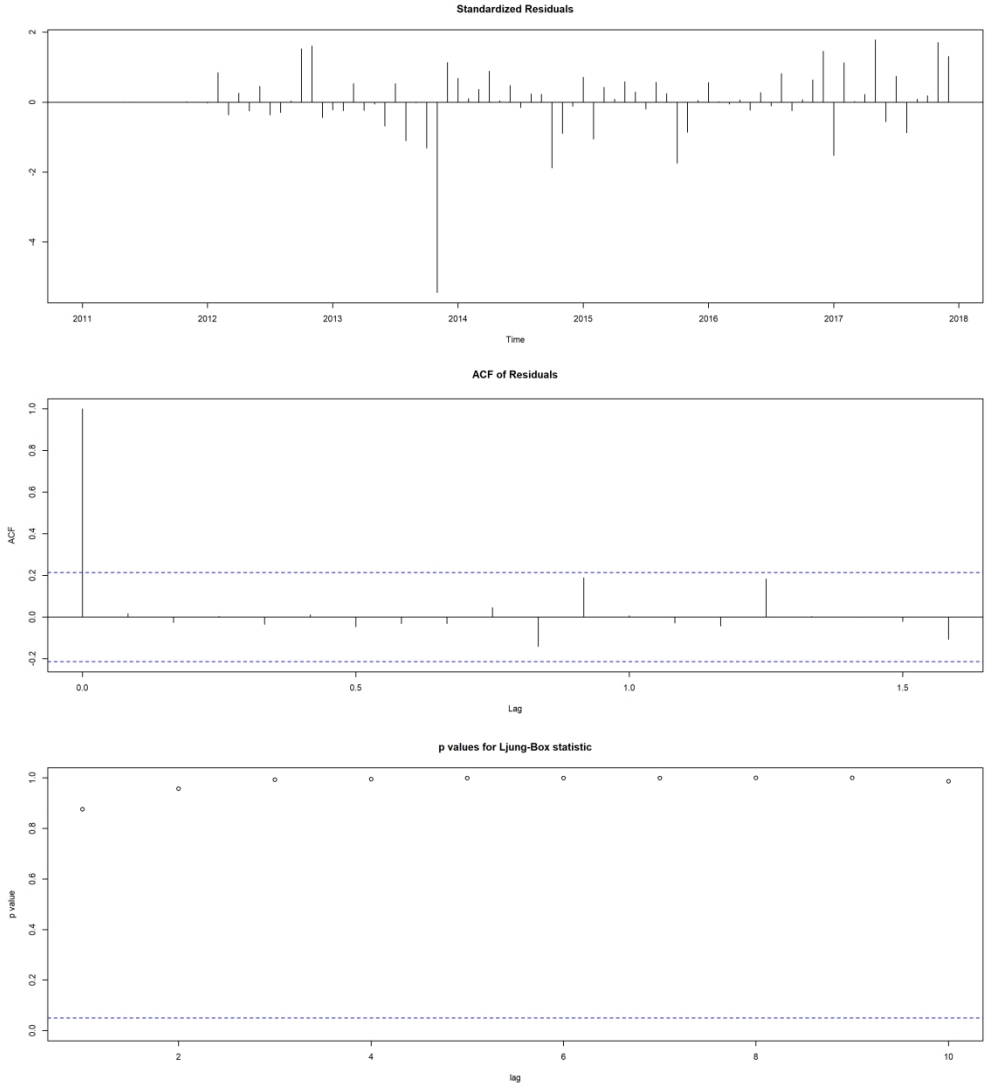
Monthly incidence of HFRS in China from January 2011 to December 2017.

406x270mm (300 x 300 DPI)



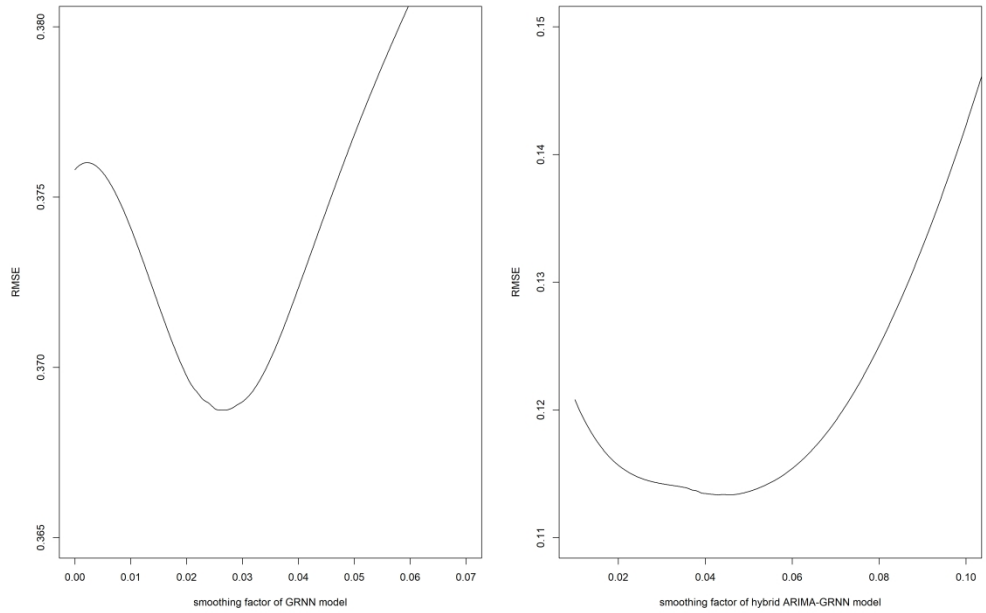
The ACF and PACF graphs of differenced HFRS incidence series.

406x256mm (300 x 300 DPI)



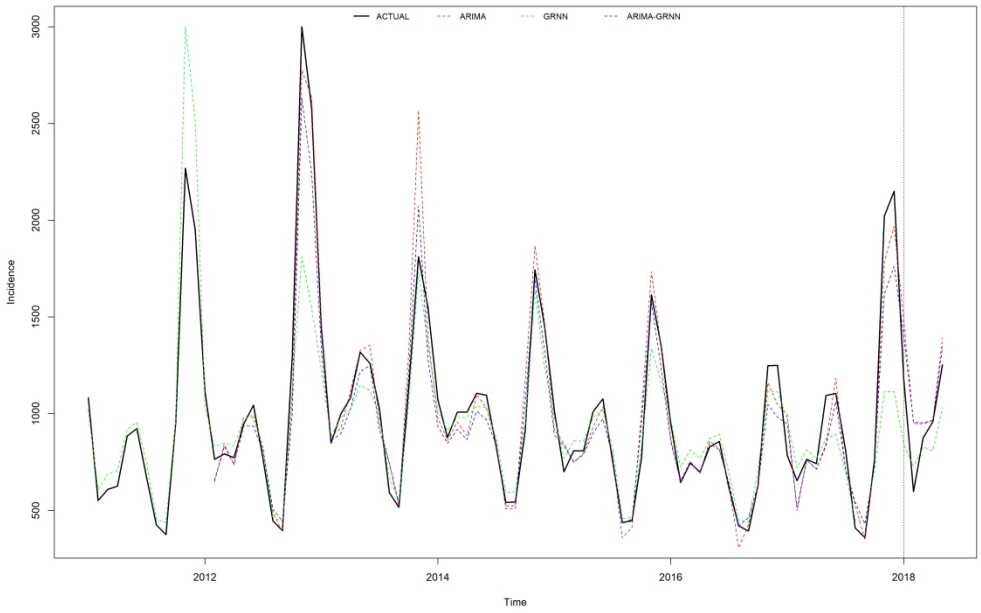
Residual white noise test

321x355mm (300 x 300 DPI)



The selection of basic GRNN model and hybrid ARIMA-GRNN model

406x270mm (300 x 300 DPI)



The fitting and forecasting curves of the three models and the actual HFRS incidence series

406x270mm (300 x 300 DPI)

S1 Table The data of HFRS incidence in China from January 2011 to May 2018

	2011	2012	2013	2014	2015	2016	2017	2018
Jan.	1082	1115	1437	1072	1016	949	785	1180
Feb.	551	765	848	876	700	644	654	598
Mar.	609	793	998	1008	809	745	764	874
Apr.	625	773	1081	1008	808	699	742	959
May	885	947	1318	1106	1011	828	1094	1253
Jun.	923	1044	1260	1095	1077	857	1105	
Jul.	670	765	1023	836	760	617	812	
Aug.	424	447	592	541	438	420	409	
Sept.	376	395	517	544	447	394	359	
Oct.	959	1296	1136	910	777	631	764	
Nov.	2268	3000	1811	1744	1614	1248	2021	
Dec.	1951	2578	1547	1454	1355	1250	2150	

BMJ Open

Comparison of autoregressive integrated moving average model and generalized regression neural network model for prediction of hemorrhagic fever with renal syndrome in China: a time-series study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-025773.R1
Article Type:	Research
Date Submitted by the Author:	15-Jan-2019
Complete List of Authors:	Wang, Ya-wen; Chinese Academy of Medical Sciences / Peking Union Medical College, School of Public Health Shen, Zhong-zhou JIANG, Yu; Chinese Academy of Medical Sciences and Peking Union Medical College
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Infectious diseases, Public health
Keywords:	autoregressive integrated moving average, generalized regression neural network, hemorrhagic fever with renal syndrome, prediction

Comparison of autoregressive integrated moving average model and generalized regression neural network model for prediction of hemorrhagic fever with renal syndrome in China: a time-series study

First author:
Yawen Wang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-17810259300
E-mail: ywwang2099@163.com

Second author:
Zhongzhou Shen
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-18310017094
E-mail: szz90123@163.com

Corresponding author:
Yu Jiang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-13693271887
E-mail: jiangyu@pumc.edu.cn

Key words: autoregressive integrated moving average; generalized regression neural network; hemorrhagic fever with renal syndrome; prediction

Word count: 3290 words.

ABSTRACT

Objectives Hemorrhagic fever with renal syndrome (HFRS) is a serious threat to public health in China, accounting for almost 90% cases reported globally. Infectious disease prediction may help in disease prevention despite some uncontrollable influence factors. This study conducted a comparison between a hybrid model and two single models in forecasting the monthly incidence of HFRS in China.

Design Time-series study.

Setting The People's Republic of China

Methods Autoregressive integrated moving average (ARIMA) model, generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model were constructed by R 3.4.3 software. The monthly reported incidence of HFRS from January 2011 to December 2017 were adopted to evaluate models' fitting performance. Data from January to May 2018 were used to demonstrate the models' forecasting performance. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were adopted to evaluate these models' effectiveness.

Results The monthly incidence of HFRS in the past several years showed a slight downtrend and obvious seasonal variation. A total of four plausible ARIMA models were built and ARIMA(2,1,1)(2,1,1)$_{12}$ model was selected as the optimal model in HFRS fitting. The smooth factors of the basic GRNN model and the hybrid model were 0.027 and 0.043 respectively. The RMSE of ARIMA(2,1,1)(2,1,1)$_{12}$ model, basic GRNN model and hybrid model were 138.8356, 265.7046 and 140.6426 respectively in fitting part and 220.6269, 202.1648 and 196.4682 respectively in forecasting part. The single ARIMA model was better in fitting while hybrid model was the best in prediction.

Conclusions The hybrid ARIMA-GRNN model was better than single ARIMA and basic GRNN model in forecasting monthly incidence of HFRS in China. It could be considered as a decision-making tool in HFRS prevention and control.

Strengths and limitations of this study

- The monthly incidence of hemorrhagic fever with renal syndrome (HFRS) in China showed an uptrend since January 2018, so it is crucial to predict the development of HFRS and prevent it outbreak.
- This study evaluated the performance of autoregressive integrated moving average (ARIMA) model and generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model in forecasting incidence of HFRS in China, the results could give a reference to choose suitable model in HFRS prediction.
- The reported data we collected may slightly differ from the actual incidence number since reported data came from monitor, it may not include the person who was infected but not went to test.
- Many factors could influence the incidence of HFRS but only time factor in study period was considered in our models, thus data should be updated to maintain the model's accuracy.
- There are lots of prediction model and this study only compared three of them, further comparison is needed to choose the best model for HFRS forecasting.

BACKGROUND

Hantavirus is a member of family *Bunyaviridae* which contains the most important zoonotic pathogens of humans.[1] Two categories of hantaviruses are Old World (Asia and Europe) virus that causes hemorrhagic fever with renal syndrome (HFRS), and New World (Americas) virus that causes hantavirus pulmonary syndrome (HPS).[2, 3] Hantaviruses are spread through the infected mammals' urine, faces, and saliva. People can be infected mainly through respiratory tract, alimentary tract and skin/mucus membrane abrasion. The onset symptoms of HFRS are fever, circulatory collapse with hypotension, hemorrhage and acute kidney injury (AKI).[4, 5] The hallmark of HFRS is capillary leak syndrome, which causes edema and hemorrhage and threatens people's life.[6, 7] Cases of HFRS are widely distributed in eastern Asia, particularly in China, Russia and Korea.[8] It is reported that the number of HFRS cases in China accounts for almost 90% of the total cases worldwide.[9, 10] Some comprehensive control activities such as effective vaccine and rodent elimination have achieved remarkable effects, while the incidence of HFRS is still high owing to some uncontrollable factors.[11, 12] Thus it is important to forecast the diseases trends and get early warning before disease outbreak.

Statistic models such as linear regression, artificial neural network and grey model have been widely used in time series forecasting.[13, 14] Reliable forecasting plays an important role in infectious diseases control before pandemic or outbreak. Autoregressive integrated moving average (ARIMA) model is one of the most popular methods in diseases prediction. The principle of ARIMA model contains filtering out the high-frequency noise in the data, detecting local trends based on linear dependence and forecasting the development trends. The limitation of this model is that ARIMA can only analyze the linear part of infectious disease series.[15] However, the non-linear part of epidemic data may not be white noise, which means some information may be lost by ARIMA model. To overcome the inherent defect of ARIMA model, an artificial neural network (ANN) model was adopted. ANN is a conceptualized mathematical non-linear classification model inspired by the behavior of biological networks of neurons.[16, 17] The generalized regression neural network (GRNN) is a member of ANN family and has unique ability of accelerated learning and greater capability for non-linear fitting. The hybrid ARIMA-GRNN model has both advantages of ARIMA model and GRNN model, it means that both the linear part and non-linear part of time series could be fitted by this hybrid model.

Some researches indicated that the hybrid model had better incidence forecasting performance than single ARIMA model and basic GRNN model in infectious diseases,[18] while the best model in predicting the incidence of HFRS in China is still unclear. Besides, some studies had compared the performance of hybrid ARIMA-GRNN model with other models[19] despite the comparison between the hybrid model with two single models in HFRS prediction is rare. This study aims to develop a single ARIMA model, a basic GRNN model and a hybrid ARIMA-GRNN model to fit and predict the monthly incidence of HFRS in China. The fitting and forecasting performance of these three models were compared to determine the best one, which is suggested to be employed in the provision of reference information for HFRS control.

METHODS

Data sources

The monthly reported incidence data of HFRS in China from January 2011 to May 2018 were collected from the official website of National Health Commission of the People's Republic of

China (Ministry of Health). All HFRS cases in mainland China must be reported to the National Health Commission through the Infectious Disease Surveillance System within 24 hours. The data was separated into model building part and model forecasting part. According to some researches, the data from January 2011 to December 2017 were adopted to build model while data from January to May 2018 were used for model verification.

Single ARIMA model

The ARIMA model is usually shown as ARIMA (p, d, q)$(P,D,Q)_S$ while the parameters mean non-seasonal and seasonal order of auto-regression, the degree of difference and moving average respectively, the subscript means the length of cyclical pattern. An ARIMA model is developed by time series stationary, parameter estimation and model check.[20]

Time series stationary is the first requirement for ARIMA model establishment, it means no fluctuation or periodicity over time. The Augmented Dickey-Fuller (ADF) unit-root test could help estimating whether the time series is stationary or not. Log transformation and differences are frequently adopted to stabilize the time series.

The parameter D is the length of seasonal difference and d is the length of trend difference, these two parameters are determined when original series is stable. The parameters of p, q, P and Q are determined by researcher's personal experience through the autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph of stationary series. Generally, more than one values may be given to each parameter so that several plausible models could be combined.

Since the best model must have the highest accuracy in disease prediction, some substandard models are excluded. A suitable model must show statistical significance in parameter test and get white noise sequence in residual test. Besides, the best model should have the lowest Akaike information criterion (AIC) value than other combined models.

Basic GRNN model

The GRNN model is built based on non-linear regression theory. The input layer, pattern layer, summation layer and output layer are involved in the construction of GRNN model.[21] Its inherent function is to identify the relationship between each input value and output value. Initially, the original data are divided into training set and test set. The test set can be the last two data or two random data of original series, the rest are adopted as the training set. Smoothing factor is the only parameter of GRNN which means the network could not be affected by human. A series of smoothing factors were tested by a circular program through Matlab software. Generally, there are more than one possible value of smoothing factor and the best one must have the lowest root mean square error (RMSE). Finally, all the original data were adopted as input part to predict the future data by the GRNN model which was built with the best smoothing factor.

Hybrid ARIMA-GRNN Model

The ARIMA model has advantage in extracting and fitting the linear part of the original time series, while the non-linear information in residual is abandoned. GRNN model is combined thanks to its capacity in data mining, so that the limitation of ARIMA model could be overcome. The hybrid ARIMA-GRNN model is developed to demonstrate if it has the highest accuracy in HFRS incidence prediction.

To develop the hybrid model, the input values are the fitting data of ARIMA model while the output values are actual data. Same with the basic GRNN model, the last two samples or two randomly selected samples of original series are used as testing set and the rest are used as training set to find the best smoothing factor and rebuilt the GRNN model. Finally, the forecasted values of

ARIMA model is used as the input data of hybrid model to get the output predictive values.

Model comparison

The forecasting effects of ARIMA model, GRNN model and hybrid ARIMA-GRNN model are estimated with RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE).[22] Excel 2016 was used to build the database, R 3.4.3 software was used to create the ARIMA model, the Matlab R2016a software was used to create the basic GRNN model and hybrid ARIMA-GRNN model.

Patient and public involvement

In this study, no patients or public was involved.

Ethics

Since no primary data collection was undertaken, no patient or public was involved, no formal ethical assessment or informed consent was required.

RESULTS

Single ARIMA model

The monthly incidence data of HFRS in China from January 2011 to December 2017 was used to develop the ARIMA model (figure 1). As shown in the original time series graph, the HFRS incidence showed seasonal variation and the period was 12 months (s=12). A slightly declining trend can be seen and it means the time series was not stationary. Trend difference (d=1) and seasonal difference (D=1) were done to eliminate the instability. The ADF test showed that the differenced time sequence was stationary (t statistics was -4.7201, P=0.0100).

Figure 1 Monthly incidence of HFRS in China from January 2011 to December 2017.

The ACF graph and PACF graph (figure 2) were applied to explore the parameters of the ARIMA model. Four appropriate models were chosen by residual test and filtered by AIC value. The AIC values of ARIMA$(1,1,1)(1,1,1)_{12}$ ARIMA$(1,1,1)(2,1,1)_{12}$ ARIMA$(2,1,1)(1,1,1)_{12}$ ARIMA$(2,1,1)(2,1,1)_{12}$ were 950.48, 944.68, 940.55 and 936.61 respectively. The ARIMA$(2,1,1)(2,1,1)_{12}$ model had the lowest AIC value and was chosen as the most suitable model in HFRS prediction. The residual test showed white noise (figure 3).

Figure 2 The ACF and PACF graphs of differenced HFRS incidence series.

Figure 3 Residual white noise test

Basic GRNN model

The samples from January 2011 to December 2017 were adopted to develop the network. The last two samples were used as testing samples while the others were training samples. To determine the optimal smoothing factors, a series of smoothing factors were tested. The smoothing factor with the minimum RMSE was selected as the optimal one. Figure 4 shows the RMSE of these smoothing factors and it can be found that the optimal smoothing factor of the one-dimensional input and one-dimensional output GRNN model was 0.027.

Figure 4 The selection of basic GRNN model and hybrid ARIMA-GRNN model

Hybrid ARIMA-GRNN model

The fitted data of ARIMA model from January 2011 to December 2017 were used as the input samples for the GRNN model and the actual HFRS values were used as the output samples to training the hybrid ARIMA-GRNN model. The RMSE of hybrid model was the lowest when the smoothing factor was 0.043 (figure 4), so 0.043 was selected to develop the GRNN model. Subsequently, the forecasting outcomes of ARIMA model from January 2018 to May 2018 were selected as the entry value of the ARIMA-GRNN model, and the output values were the predictive values of the hybrid model.

Finally, all three models had forecasted the HFRS incidence in China from January to May 2018. The forecasting performance parameters of the three models for the fitting and forecasting parts are shown in Table 1. The curves of the three models and the actual HFRS incidence series are depicted in figure 5. In this figure, the curves were divided into fitting part and forecasting part by a vertical dashed line, the left is fitting part while the forecasting part is on right.

Table 1. The fitting and forecasting performance of three models.

	Fitting part			Forecasting part		
Predicting error	MAPE	MAE	RMSE	MAPE	MAE	RMSE
ARIMA	9.1154	89.0302	138.8356	21.0212	175.7042	220.6269
GRNN	10.7332	134.596	265.7046	19.2029	177.0356	202.1684
ARIMA-GRNN	9.6083	85.0429	140.6426	17.8026	152.3013	196.4682

Figure 5 The fitting and forecasting curves of three models and the actual HFRS incidence series

DISCUSSION

In this study, a hybrid model was constructed based on traditional ARIMA model and basic GRNN model. Three different models were compared in fitting and forecasting performance and the results showed that the hybrid ARIMA-GRNN model was the best model in predicting the monthly reported incidence of HFRS in China. Thus we consider the hybrid model as a decision-making tool to give some suggestion in public health policy decision.

The characteristic of monthly incidence of HFRS in China is suitable for ARIMA model and GRNN model. As shown in the results, the incidence of HFRS in China has a slight decreasing trend and a bimodal seasonal cases distribution, which are same with other studies.[23, 24] The incidence reaches peak in winter rapidly and has a longer lasting peak in Spring. Autumn to winter peak is the other peak, which is lower than the winter to spring one. Two reasons may could explain this seasonal distribution. People are more likely to be exposed to the disease due to increased activities in these two seasons and rodent behavior changes with climate change.[25, 26] Besides, the distribution and peak value might change with different hantaviruses types.

The hybrid ARIMA-GRNN model was superior among three models even with imperfect fitting performance. ARIMA model is one of the most commonly used methods in infectious diseases prediction and has been proved with high accuracy. In this study, the traditional ARIMA model was used as the basic model for evaluating the performance of other models. The results showed that single ARIMA model and basic GRNN model were better than hybrid model in data fitting according to lower MAE and MAPE. Even some unmeasurable factors may impact data

fitting, the forecasting performance should be at the first consideration.[21] The MAPE, MAE, RMSE of hybrid model in validation part were lower than single ARIMA model or basic GRNN model. Some studies built the hybrid model with tuberculosis incidence or hand-foot-mouth disease incidence[19,27, 28] and the results showed that hybrid ARIMA-GRNN model had less error than single model both in modeling and forecasting stage, which is different with our study. Thus we hypothesis that diseases characteristics may affect the model performance and the best predictive model of each infectious disease is different. Model in this study could only fit the incidence of HFRS in China, its performance in other diseases or other nation needs further research.

The time series prediction model was developed as a new potential tool for infectious diseases incidence prediction in recent years. In this study, hybrid ARIMA-GRNN model was chose as a potential outbreak warning tool. Same with other disease prediction models, the disease control department could assess the disease developing trend with the help of the hybrid ARIAM-GRNN model. In a short term, the prediction values have same trend with the actual values. It means if the predictive values continue to rise, an outbreak should be alerted. Besides, disease prediction model is developed to evaluate the effectiveness of diseases intervention strategies like vaccine. An effective control measure will make the actual values lower than the predicted results. Something noteworthy is that these two functions are based on short terms. The incidence of infectious disease is influenced by some uncontrollable factors and HFRS is infected by weather, climate, human activities and so on.[29-31] These factors may keep stable in a short period and might change in a long run.

Several limitations of this study should be noted. As is shown above, the prediction model was merely developed for short-term forecasting. Maintaining the prediction performance for months or years requires constantly update of data and model. Here we build three new models whose fitting data were HFRS incidence from January 2011 to December 2015 and data from January 2016 to May 2018 were used to verification (Table S1). It showed that model with new data has higher accuracy. Besides, this study only considered the incidence of HFRS in China, weather the hybrid model is suitable for other countries is still unclear and larger samples are needed to test. At last, HFRS incidence data in this manuscript was total incidence in China, we can not explore the performance of these models in provincial incidence prediction. Spatial factor is an important factor that can affect HFRS development, so the applicability of results in this research need further study.

CONCLUSIONS

The hybrid ARIMA-GRNN model is superior than the single ARIMA model and basic GRNN model both in fitting and forecasting of monthly incidence of HFRS in China. The data should keep update to maintain the forecasting performance. This hybrid model should be considered as a decision-making tool in HFRS prevention and control.

Supporting information

S1 Table. The fitting and forecasting performance of three new models.

Acknowledgement We would like to express our gratitude to professor Yin Yang for carefully revise of overall readability. We also thank peer reviewers for carefully revising and useful

comments.

Contributors YJ, YW and ZS designed the study. ZS extracted the data and constructed the database. YW and ZS analyzed the data. YW drafted the manuscript. YJ and ZS made critical revision to the manuscript. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All original data was available at the official website of National Health Commission of the People's Republic of China (http://www.nhfpc.gov.cn/jkj/new_index.shtml).

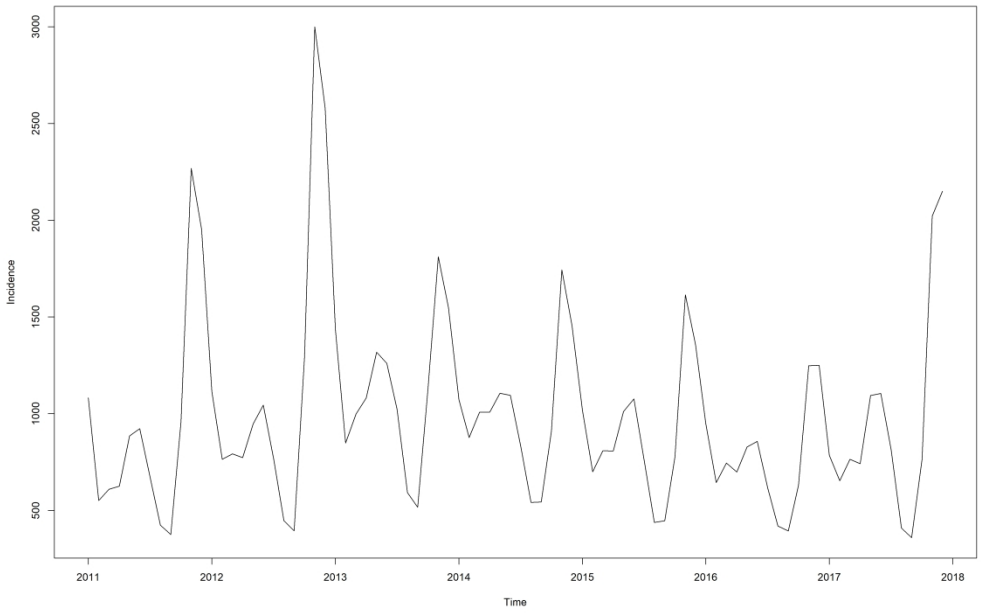
REFERENCES

1. Schmaljohn CS, Hasty SE, Dalrymple JM, et al. Antigenic and genetic properties of viruses linked to hemorrhagic fever with renal syndrome. *Science.* 1985;227(4690):1041-1044.
2. Ehelepola NDB, Basnayake BMLS, Sathkumara SMBY, Kaluphana KLR. Two Atypical Cases of Hantavirus Infections from Sri Lanka. *Case Reports in Infectious Diseases.* 2018;2018:1-6. DOI: 10.1155/2018/4069862.
3. Avsic-Zupanc T, Saksida A, Korva M. Hantavirus infections. *Clin Microbiol Infect.* 2015. DOI: 10.1111/1469-0691.12291.
4. Latus J, Schwab M, Tacconelli E, et al. Clinical course and long-term outcome of hantavirus-associated nephropathia epidemica, Germany. *Emerg Infect Dis.* 2015;21(1):76-83. DOI: 10.3201/eid2101.140861.
5. Vaheri A, Strandin T, Hepojoki J, et al. Uncovering the mysteries of hantavirus infections. *Nat Rev Microbiol.* 2013;11(8):539-550. DOI: 10.1038/nrmicro3066.
6. Hepojoki J, Vaheri A, Strandin T. The fundamental role of endothelial cells in hantavirus pathogenesis. *Front Microbiol.* 2014;5:727. DOI: 10.3389/fmicb.2014.00727.
7. Pal E, Korva M, Resman RK, et al. Sequential assessment of clinical and laboratory parameters in patients with hemorrhagic fever with renal syndrome. *Plos One.* 2018;13(5):e197661. DOI: 10.1371/journal.pone.0197661.
8. Bi P, Parton KA. El Nino and incidence of hemorrhagic fever with renal syndrome in China. *JAMA.* 2003;289(2):176-177.
9. Zhang S, Wang S, Yin W, et al. Epidemic characteristics of hemorrhagic fever with renal syndrome in China, 2006-2012. *BMC Infect Dis.* 2014;14:384. DOI: 10.1186/1471-2334-14-384.
10. Du H, Wang PZ, Li J, et al. Clinical characteristics and outcomes in critical patients with hemorrhagic fever with renal syndrome. *BMC Infect Dis.* 2014;14:191. DOI: 10.1186/1471-2334-14-191.
11. Zhang WY, Wang LY, Liu YX, Yin WW, Hu WB, Magalhaes RJ, et al. Spatiotemporal Transmission Dynamics of Hemorrhagic Fever with Renal Syndrome in China, 2005−2012. *Plos Negl Trop Dis*. 2014;8:e3344. DOI: 10.1371/journal.pntd.0003344.
12. He X, Wang S, Huang X, Wang X. Changes in age distribution of hemorrhagic fever with renal syndrome: an implication of China's expanded program of immunization. *BMC Public Health.* 2013;13:394. DOI: 10.1186/1471-2458-13-394.
13. Wang YW, Shen ZZ, Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. *Plos One.* 2018;13(9):e201987. DOI: 10.1371/journal.pone.0201987.

14. Cao H, Wang J, Li Y, et al. Trend analysis of mortality rates and causes of death in children under 5 years old in Beijing, China from 1992 to 2015 and forecast of mortality into the future: an entire population-based epidemiological study. *BMJ Open.* 2017;7(9):e15941. DOI: 10.1136/bmjopen-2017-015941.
15. Petukhova T, Ojkic D, McEwen B, Deardon R, Poljak Z. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. *Plos One.* 2018;13(6):e198313. DOI: 10.1371/journal.pone.0198313.
16. Yosipof A, Guedes RC, Garcia-Sosa AT. Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. *Front Chem.* 2018;6:162. DOI: 10.3389/fchem.2018.00162.
17. Nair TM. Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia. *Comput Biol Chem.* 2018;75:222-230. DOI: 10.1016/j.compbiolchem.2018.05.011.
18. Wei W, Jiang J, Gao L, et al. 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. *Am J Trop Med Hyg.* 2017;97(3):799-805. DOI: 10.4269/ajtmh.16-0648.
19. Wu W, Guo J, An S, et al. Comparison of Two Hybrid Models for Forecasting the Incidence of Hemorrhagic Fever with Renal Syndrome in Jiangsu Province, China. *Plos One.* 2015;10(8):e135492. DOI: 10.1371/journal.pone.0135492.
20. Rubaihayo J, Tumwesigye NM, Konde-Lule J, Makumbi F. Forecast analysis of any opportunistic infection among HIV positive individuals on antiretroviral therapy in Uganda. *BMC Public Health.* 2016;16(1):766. DOI: 10.1186/s12889-016-3455-5.
21. Wei W, Jiang J, Liang H, et al. Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. *Plos One.* 2016;11(6):e156768. DOI: 10.1371/journal.pone.0156768.
22. Gan R, Chen N, Huang D. Comparisons of forecasting for hepatitis in Guangxi Province, China by using three neural networks models. *Peerj.* 2016;4:e2684. DOI: 10.7717/peerj.2684.
23. Hansen A, Cameron S, Liu Q, et al. Transmission of Haemorrhagic Fever with Renal Syndrome in China and the Role of Climate Factors: A Review. *Int J Infect Dis.* 2015;33:212-218. DOI: 10.1016/j.ijid.2015.02.010.
24. Liu YX, Feng D, Zhang Q, et al. Key differentiating features between scrub typhus and hemorrhagic fever with renal syndrome in northern China. *Am J Trop Med Hyg.* 2007;76(5):801-805.
25. Park YH. Absence of a Seasonal Variation of Hemorrhagic Fever with Renal Syndrome in Yeoncheon Compared to Nationwide Korea. *Infect Chemother.* 2018;50(2):120-127. DOI: 10.3947/ic.2018.50.2. 120.
26. Mills JN, Gage KL, Khan AS. Potential influence of climate change on vector-borne and zoonotic diseases: a review and proposed research plan. *Environ Health Perspect.* 2010;118(11):1507-1514. DOI: 10.1289/ehp.0901389.
27. Wang H, Tian CW, Wang WM, Luo XM. Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiol Infect.* 2018;146(8):935-939. DOI: 10.1017/S0950268818001115.
28. Peng Y, Yu B, Wang P, Kong DG, Chen BH, Yang XB. Application of seasonal auto-regressive

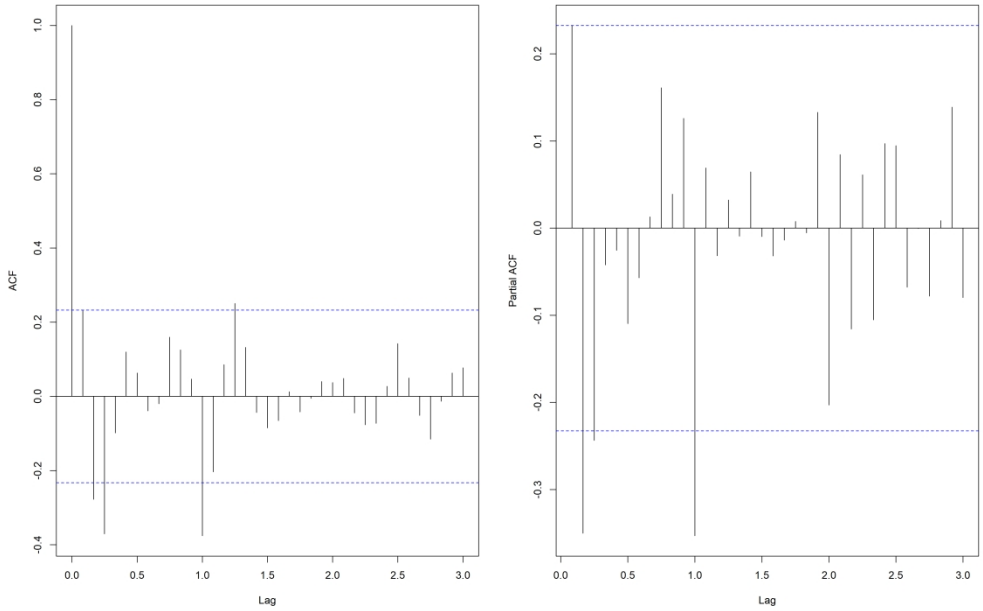
integrated moving average model in forecasting the incidence of hand-foot-mouth disease in Wuhan, China. *J Huazhong Univ Sci Technolog Med Sci.* 2017;37(6):842-848. DOI: 10.1007/s11596-017-1815-8.

29. Joshi YP, Kim E, Cheong H. The influence of climatic factors on the development of hemorrhagic fever with renal syndrome and leptospirosis during the peak season in Korea: an ecologic study. *BMC Infect Dis.* 2017;17(1). DOI: 10.1186/s12879-017-2506-6.
30. Han SS, Kim S, Choi Y, Kim S, Kim YS. Air pollution and hemorrhagic fever with renal syndrome in South Korea: an ecological correlation study. *BMC Public Health.* 2013;13:347. DOI: 10.1186/1471-2458-13-347.
31. Xiang J, Hansen A, Liu Q, et al. Impact of meteorological factors on hemorrhagic fever with renal syndrome in 19 cities in China, 2005-2014. *Sci Total Environ.* 2018;636:1249-1256. DOI: 10.1016/j.scitotenv.2018.04.407.



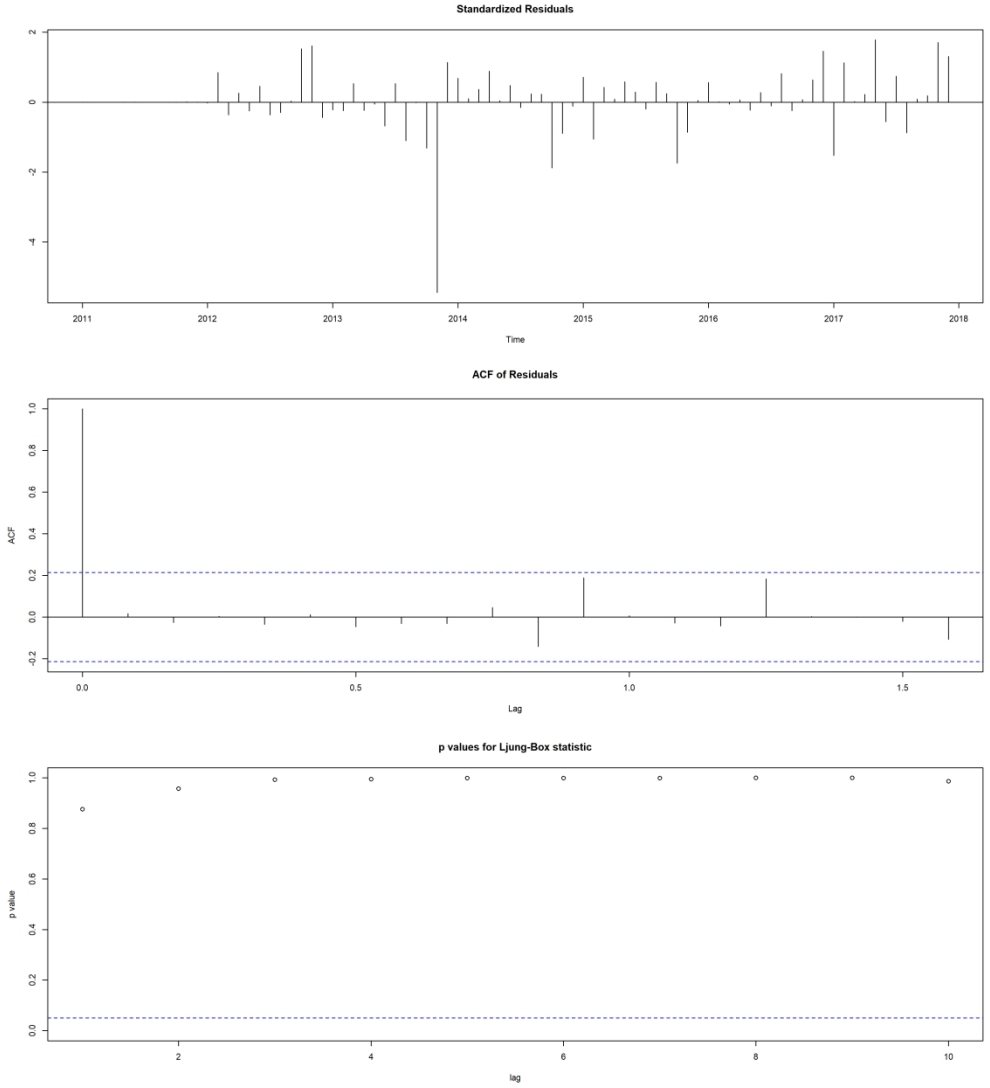
Monthly incidence of HFRS in China from January 2011 to December 2017.

406x270mm (300 x 300 DPI)



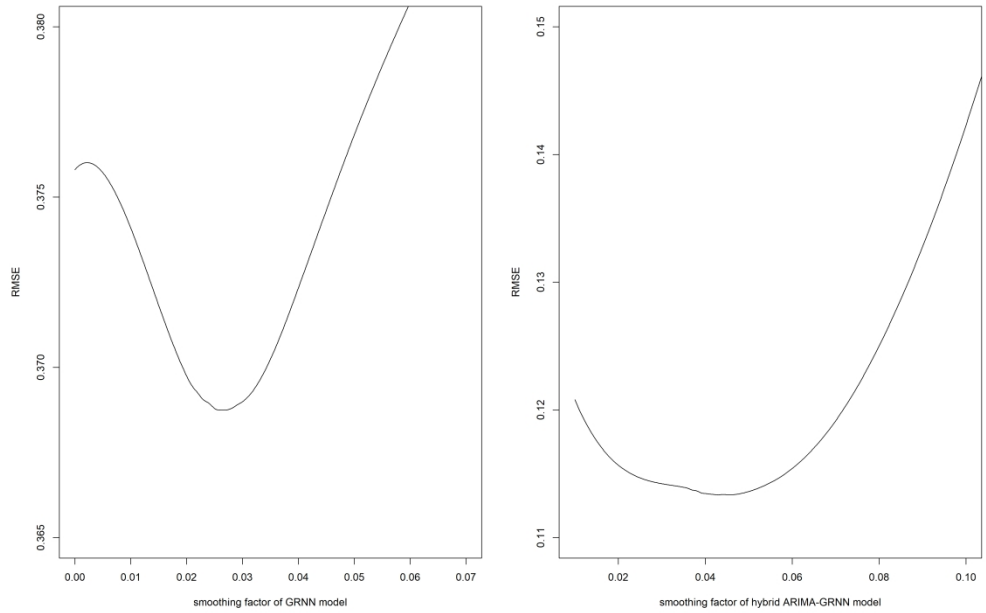
The ACF and PACF graphs of differenced HFRS incidence series.

406x256mm (300 x 300 DPI)



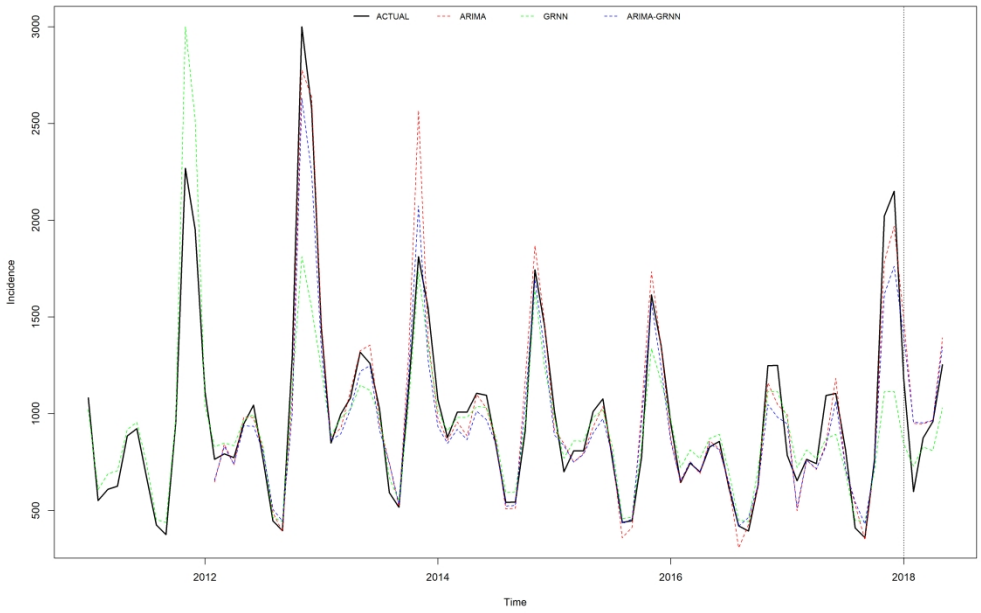
Residual white noise test

321x355mm (300 x 300 DPI)



The selection of basic GRNN model and hybrid ARIMA-GRNN model

406x270mm (300 x 300 DPI)



The fitting and forecasting curves of the three models and the actual HFRS incidence series

406x270mm (300 x 300 DPI)

S1 Table The fitting and forecasting performance of three new models

	Fitting part			Forecasting part		
Predicting error	MAPE	MAE	RMSE	MAPE	MAE	RMSE
ARIMA	10.2735	105.6382	155.7399	27.9010	259.9656	359.9456
GRNN	33.4315	325.6638	512.7855	41.1768	299.8275	402.3660
ARIMA-GRNN	22.5002	213.6670	248.8867	21.3148	221.3138	336.5332

Table 1. Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis*

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	2-3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both	3
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	3-4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	3-4
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres	NA
	5b	D;V	Describe eligibility criteria for participants	NA
	5c	D;V	Give details of treatments received, if relevant	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	NA
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted	NA
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	5
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors	NA
Sample size	8	D;V	Explain how the study size was arrived at	4
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	NA
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses	4
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	4
	10c	V	For validation, describe how the predictions were calculated	4-5
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models	5

	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	3
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	NA
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome)	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis	5-6
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point)	5-6
	15b	D	Explain how to use the prediction model	6-7
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model	6
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance)	6
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data)	7
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data	6-7
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	6-7
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research	7
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	7
Funding	22	D;V	Give the source of funding and the role of the funders for the present study	NA

Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

BMJ Open

Comparison of autoregressive integrated moving average model and generalized regression neural network model for prediction of hemorrhagic fever with renal syndrome in China: a time-series study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-025773.R2
Article Type:	Research
Date Submitted by the Author:	13-Mar-2019
Complete List of Authors:	Wang, Ya-wen; Chinese Academy of Medical Sciences and Peking Union Medical College, School of Public Health Shen, Zhong-zhou; Chinese Academy of Medical Sciences and Peking Union Medical College, School of Public Health JIANG, Yu; Chinese Academy of Medical Sciences and Peking Union Medical College, School of Public Health
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Infectious diseases, Public health
Keywords:	autoregressive integrated moving average, generalized regression neural network, hemorrhagic fever with renal syndrome, prediction

Comparison of autoregressive integrated moving average model and generalized regression neural network model for prediction of hemorrhagic fever with renal syndrome in China: a time-series study

First author:
Yawen Wang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-17810259300
E-mail: ywwang2099@163.com

Second author:
Zhongzhou Shen
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-18310017094
E-mail: szz90123@163.com

Corresponding author:
Yu Jiang
Address: School of Public Health, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China.
Tel: 86-13693271887
E-mail: jiangyu@pumc.edu.cn

Key words: autoregressive integrated moving average; generalized regression neural network; hemorrhagic fever with renal syndrome; prediction

Word count: 3631 words.

ABSTRACT

Objectives Hemorrhagic fever with renal syndrome (HFRS) is a serious threat to public health in China, accounting for almost 90% cases reported globally. Infectious disease prediction may help in disease prevention despite some uncontrollable influence factors. This study conducted a comparison between a hybrid model and two single models in forecasting the monthly incidence of HFRS in China.

Design Time-series study.

Setting The People's Republic of China

Methods Autoregressive integrated moving average (ARIMA) model, generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model were constructed by R 3.4.3 software. The monthly reported incidence of HFRS from January 2011 to May 2018 were adopted to evaluate models' performance. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were adopted to evaluate these models' effectiveness. Spatial stratified heterogeneity (SSH) of the time series was tested by month and another GRNN model was built with a new series.

Results The monthly incidence of HFRS in the past several years showed a slight downtrend and obvious seasonal variation. A total of four plausible ARIMA models were built and ARIMA(2,1,1)$(2,1,1)_{12}$ model was selected as the optimal model in HFRS fitting. The smooth factors of the basic GRNN model and the hybrid model were 0.027 and 0.043 respectively. The single ARIMA model was the best in fitting part (MAPE=9.1154, MAE=89.0302, RMSE=138.8356) while the hybrid model was the best in prediction (MAPE=17.8335, MAE=152.3013, RMSE=196.4682). GRNN model was revised by building model with new series and the forecasting performance of revised model (MAPE=17.6095, MAE=163.8000, RMSE=169.4751) was better than original GRNN model (MAPE=19.2029, MAE=177.0356, RMSE=202.1684).

Conclusions The hybrid ARIMA-GRNN model was better than single ARIMA and basic GRNN model in forecasting monthly incidence of HFRS in China. It could be considered as a decision-making tool in HFRS prevention and control.

Strengths and limitations of this study

- The monthly incidence of hemorrhagic fever with renal syndrome (HFRS) in China showed an uptrend since January 2018, so it is crucial to predict the development of HFRS and prevent it outbreak.
- This study evaluated the performance of autoregressive integrated moving average (ARIMA) model and generalized regression neural network (GRNN) model and hybrid ARIMA-GRNN model in forecasting incidence of HFRS in China, the results could give a reference to choose suitable model in HFRS prediction.
- The reported data we collected may slightly differ from the actual incidence number since reported data came from monitor, it may not include the person who was infected but not went to test.
- Many factors could influence the incidence of HFRS but only time factor in study period was considered in our models, thus data should be updated to maintain the model's accuracy. Besides, there are lots of prediction models and this study only compared three of them, further comparison is needed to choose the best model for HFRS forecasting.
- Spatial stratified heterogeneity (SSH) should be tested in time series prediction research,

applying the prediction model in each spatial is an important way to improve the model's performance.

BACKGROUND

Hantavirus is a member of family *Bunyaviridae* which contains the most important zoonotic pathogens of humans.[1] Two categories of hantaviruses are Old World (Asia and Europe) virus that causes hemorrhagic fever with renal syndrome (HFRS), and New World (Americas) virus that causes hantavirus pulmonary syndrome (HPS).[2, 3] Hantaviruses are spread through the infected mammals' urine, faces, and saliva. People can be infected mainly through respiratory tract, alimentary tract and skin/mucus membrane abrasion. The onset symptoms of HFRS are fever, circulatory collapse with hypotension, hemorrhage and acute kidney injury (AKI).[4, 5] The hallmark of HFRS is capillary leak syndrome, which causes edema and hemorrhage and threatens people's life.[6, 7] Cases of HFRS are widely distributed in eastern Asia, particularly in China, Russia and Korea.[8] It is reported that the number of HFRS cases in China accounts for almost 90% of the total cases worldwide.[9, 10] Some comprehensive control activities such as effective vaccine and rodent elimination have achieved remarkable effects, while the incidence of HFRS is still high owing to some uncontrollable factors.[11, 12] Thus it is important to forecast the diseases trends and get early warning before disease outbreak.

Statistic models such as linear regression, artificial neural network and grey model have been widely used in time series forecasting.[13, 14] Reliable forecasting plays an important role in infectious diseases control before pandemic or outbreak. Autoregressive integrated moving average (ARIMA) model is one of the most popular methods in diseases prediction. The principle of ARIMA model contains filtering out the high-frequency noise in the data, detecting local trends based on linear dependence and forecasting the development trends. The limitation of this model is that ARIMA can only analyze the linear part of infectious disease series.[15] However, the non-linear part of epidemic data may not be white noise, which means some information may be lost by ARIMA model. To overcome the inherent defect of ARIMA model, an artificial neural network (ANN) model was adopted. ANN is a conceptualized mathematical non-linear classification model inspired by the behavior of biological networks of neurons.[16, 17] The generalized regression neural network (GRNN) is a member of ANN family and has unique ability of accelerated learning and greater capability for non-linear fitting. The hybrid ARIMA-GRNN model has both advantages of ARIMA model and GRNN model, it means that both the linear part and non-linear part of time series could be fitted by this hybrid model.

Some researches indicated that the hybrid model had better incidence forecasting performance than single ARIMA model and basic GRNN model in infectious diseases,[18] while the best model in predicting the incidence of HFRS in China is still unclear. Besides, some studies had compared the performance of hybrid ARIMA-GRNN model with other models[19] despite the comparison between the hybrid model with two single models in HFRS prediction is rare. This study aims to develop a single ARIMA model, a basic GRNN model and a hybrid ARIMA-GRNN model to fit and predict the monthly incidence of HFRS in China. The fitting and forecasting performance of these three models were compared to determine the best one, which is suggested to be employed in the provision of reference information for HFRS control.

METHODS

Data sources

The monthly reported incidence data of HFRS in China from January 2011 to May 2018 were collected from the official website of National Health Commission of the People's Republic of China (Ministry of Health). All HFRS cases in mainland China must be reported to the National Health Commission through the Infectious Disease Surveillance System within 24 hours. The data was separated into model building part and model forecasting part. According to some researches, the data from January 2011 to December 2017 were adopted to build model while data from January to May 2018 were used for model verification.

Single ARIMA model

The ARIMA model is usually shown as ARIMA (p, d, q)$(P,D,Q)_S$ while the parameters mean non-seasonal and seasonal order of auto-regression, the degree of difference and moving average respectively, the subscript means the length of cyclical pattern. An ARIMA model is developed by time series stationary, parameter estimation and model check.[20]

Time series stationary is the first requirement for ARIMA model establishment, it means no fluctuation or periodicity over time. The Augmented Dickey-Fuller (ADF) unit-root test could help estimating whether the time series is stationary or not. Log transformation and differences are frequently adopted to stabilize the time series.

The parameter D is the length of seasonal difference and d is the length of trend difference, these two parameters are determined when original series is stable. The parameters of p, q, P and Q are determined by researcher's personal experience through the autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph of stationary series. Generally, more than one values may be given to each parameter so that several plausible models could be combined.

Since the best model must have the highest accuracy in disease prediction, some substandard models are excluded. A suitable model must show statistical significance in parameter test and get white noise sequence in residual test. Besides, the best model should have the lowest Akaike information criterion (AIC) value than other combined models.

Basic GRNN model

The GRNN model is built based on non-linear regression theory. The input layer, pattern layer, summation layer and output layer are involved in the construction of GRNN model.[21] Its inherent function is to identify the relationship between each input value and output value. Initially, the original data are divided into training set and test set. The test set can be the last two data or two random data of original series, the rest are adopted as the training set. Smoothing factor is the only parameter of GRNN which means the network could not be affected by human. A series of smoothing factors were tested by a circular program through Matlab software. Generally, there are more than one possible value of smoothing factor and the best one must have the lowest root mean square error (RMSE). Finally, all the original data were adopted as input part to predict the future data by the GRNN model which was built with the best smoothing factor.

Hybrid ARIMA-GRNN Model

The ARIMA model has advantage in extracting and fitting the linear part of the original time series, while the non-linear information in residual is abandoned. GRNN model is combined thanks to its capacity in data mining, so that the limitation of ARIMA model could be overcome. The hybrid ARIMA-GRNN model is developed to demonstrate if it has the highest accuracy in HFRS incidence prediction.

To develop the hybrid model, the input values are the fitting data of ARIMA model while the output values are actual data. Same with the basic GRNN model, the last two samples or two randomly selected samples of original series are used as testing set and the rest are used as training set to find the best smoothing factor and rebuilt the GRNN model. Finally, the forecasted values of ARIMA model is used as the input data of hybrid model to get the output predictive values.

Model revision

Spatial stratified heterogeneity (SSH) refers to the phenomenon that within strata are more similar than between strata.[22] SSH is an unavoidable confounder in global model application, especially in areas with huge region.[23] The "spatial" not only refer to geospatial meaning, but also mathematical meaning, such as gender, region and education level. In this study, SSH was tested by month to demonstrate if there were different strata in HFRS incidence series. The prediction model will be built in different strata if SSH test is significant.

Model comparison

The forecasting effects of ARIMA model, GRNN model and hybrid ARIMA-GRNN model are estimated with RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE).[24] Excel 2016 was used to build the database, R 3.4.3 software was used to create the ARIMA model, the Matlab R2016a software was used to create the basic GRNN model and hybrid ARIMA-GRNN model. GeoDetector software was used for SSH test.

Patient and public involvement

In this study, no patients or public was involved.

Ethics

Since no primary data collection was undertaken, no patient or public was involved, no formal ethical assessment or informed consent was required.

RESULTS

Single ARIMA model

The monthly incidence data of HFRS in China from January 2011 to December 2017 was used to develop the ARIMA model (figure 1). As shown in the original time series graph, the HFRS incidence showed seasonal variation and the period was 12 months (s=12). A slightly declining trend can be seen and it means the time series was not stationary. Trend difference (d=1) and seasonal difference (D=1) were done to eliminate the instability. The ADF test showed that the differenced time sequence was stationary (t statistics was -4.7201, P=0.0100).

Figure 1 Monthly incidence of HFRS in China from January 2011 to December 2017.

The ACF graph and PACF graph (figure 2) were applied to explore the parameters of the ARIMA model. Four appropriate models were chosen by residual test and filtered by AIC value. The AIC values of ARIMA$(1,1,1)(1,1,1)_{12}$ ARIMA$(1,1,1)(2,1,1)_{12}$ ARIMA$(2,1,1)(1,1,1)_{12}$ ARIMA$(2,1,1)(2,1,1)_{12}$ were 950.48, 944.68, 940.55 and 936.61 respectively. The ARIMA$(2,1,1)(2,1,1)_{12}$ model had the lowest AIC value and was chosen as the most suitable model in HFRS prediction. The residual test showed white noise (figure 3).

Figure 2 The ACF and PACF graphs of differenced HFRS incidence series.

Figure 3 Residual white noise test

Basic GRNN model

The samples from January 2011 to December 2017 were adopted to develop the network. The last two samples were used as testing samples while the others were training samples. To determine the optimal smoothing factors, a series of smoothing factors were tested. The smoothing factor with the minimum RMSE was selected as the optimal one. Figure 4 shows the RMSE of these smoothing factors and it can be found that the optimal smoothing factor of the one-dimensional input and one-dimensional output GRNN model was 0.027.

Figure 4 The selection of basic GRNN model and hybrid ARIMA-GRNN model

Hybrid ARIMA-GRNN model

The fitted data of ARIMA model from January 2011 to December 2017 were used as the input samples for the GRNN model and the actual HFRS values were used as the output samples to training the hybrid ARIMA-GRNN model. The RMSE of hybrid model was the lowest when the smoothing factor was 0.043 (figure 4), so 0.043 was selected to develop the GRNN model. Subsequently, the forecasting outcomes of ARIMA model from January 2018 to May 2018 were selected as the entry value of the ARIMA-GRNN model, and the output values were the predictive values of the hybrid model.

Finally, all three models had forecasted the HFRS incidence in China from January to May 2018. The forecasting performance parameters of the three models for the fitting and forecasting parts are shown in Table 1. The curves of the three models and the actual HFRS incidence series are depicted in figure 5. In this figure, the curves were divided into fitting part and forecasting part by a vertical dashed line, the left is fitting part while the forecasting part is on right.

Table 1. The fitting and forecasting performance of three models.

	Fitting part			Forecasting part		
Predicting error	MAPE	MAE	RMSE	MAPE	MAE	RMSE
ARIMA	9.1154	89.0302	138.8356	21.0212	175.7042	220.6269
GRNN	10.7332	134.596	265.7046	19.2029	177.0356	202.1684
ARIMA-GRNN	9.6083	85.0429	140.6426	17.8335	152.3013	196.4682

Figure 5 The fitting and forecasting curves of three models and the actual HFRS incidence series

Model revision

HFRS incidence time series from January 2011 to December 2017 was partitioned to 12 strata according to their months and SSH was tested. The results showed a q statistic with 0.776 and a p value with 0.000, the SSH was significant. Given these results, the prediction was applied in each strata.

A total of 12 new time series were established and each one has data with same month of each year. The sample size of each series was 7. Since the ARIMA model requires a series with large sample size, thus we built GRNN model to explore whether the strata help improve the model's performance. The verification data were actual HFRS incidence from January to May 2018, thus we built five revised GRNN models with new series. The relative error of these revised

GRNN models were showed in Table 2. The average relative error of revised GRNN model was 17.61%, which was lower than 17.83% of original GRNN model. The MAPE, MAE and RMSE of revised model were 17.6095, 163.8000, 169.4751, respectively. These results indicated that the revised model was better than original GRNN model and application of prediction model in different strata was important to model's performance improvement.

Table 2. The relative error of GRNN model in HFRS forecasting.

	Actual value	Original GRNN model		Revised GRNN model	
		Forecasted value	Relative error(%)	Forecasted value	Relative error(%)
Jan.	1180	843	28.56	1016	13.90
Feb.	598	729	21.91	765	27.93
Mar.	874	828	5.26	998	14.19
Apr.	959	809	15.64	1081	12.72
May	1253	1030	17.80	1011	19.31

DISCUSSION

In this study, a hybrid model was constructed based on traditional ARIMA model and basic GRNN model. These three different models were compared in fitting and forecasting performance and the results showed that the hybrid ARIMA-GRNN model was the best model in predicting the monthly reported incidence of HFRS in China. The hybrid model might be a potential decision-making tool to give some suggestion in public health policy decision. However, focusing on spatial SSH and developing prediction model in different strata help improve model's performance. A hybrid ARIMA-GRNN model built with data of same month of each year might be better than the existing hybrid model.

The characteristic of monthly incidence of HFRS in China is suitable for ARIMA model and GRNN model. As shown in the results, the incidence of HFRS in China has a slight decreasing trend and a bimodal seasonal cases distribution, which are same with other studies.[25, 26] The incidence reaches peak in winter rapidly and has a longer lasting peak in Spring. Autumn to winter peak is the other peak, which is lower than the winter to spring one. Two reasons may could explain this seasonal distribution. People are more likely to be exposed to the disease due to increased activities in these two seasons and rodent behavior changes with climate change.[27, 28] Besides, the distribution and peak value might change with different hantaviruses types.

The hybrid ARIMA-GRNN model was superior among three models even with imperfect fitting performance. ARIMA model is one of the most commonly used methods in infectious diseases prediction and has been proved with high accuracy. In this study, the traditional ARIMA model was used as the basic model for evaluating the performance of other models. The results showed that single ARIMA model and basic GRNN model were better than hybrid model in data fitting according to lower MAE and MAPE. Even some unmeasurable factors may impact data fitting, the forecasting performance should be at the first consideration.[21] The MAPE, MAE, RMSE of hybrid model in validation part were lower than single ARIMA model or basic GRNN model. Some studies built the hybrid model with tuberculosis incidence or hand-foot-mouth disease incidence[19,29, 30] and the results showed that hybrid ARIMA-GRNN model had less error than single model both in modeling and forecasting stage, which is different with our study. Thus

we hypothesis that diseases characteristics may affect the model performance and the best predictive model of each infectious disease is different. Model in this study could only fit the incidence of HFRS in China, its performance in other diseases or other nation needs further research.

The time series prediction model was developed as a new potential tool for infectious diseases incidence prediction in recent years. In this study, hybrid ARIMA-GRNN model was chose as a potential outbreak warning tool. Same with other disease prediction models, the disease control department could assess the disease developing trend with the help of the hybrid ARIAM-GRNN model. In a short term, the prediction values have same trend with the actual values. It means if the predictive values continue to rise, an outbreak should be alerted. Besides, disease prediction model is developed to evaluate the effectiveness of diseases intervention strategies like vaccine. An effective control measure will make the actual values lower than the predicted results. Something noteworthy is that these two functions are based on short terms. The incidence of infectious disease is influenced by some uncontrollable factors and HFRS is infected by weather, climate, human activities and so on.[31-33] These factors may keep stable in a short period and might change in a long run.

SSH is unavoidable in prediction model application and developing model in different strata is a common way to deal with this confounding. In this study, we partitioned the original series to 12 different series by month in order to relieve confounding. Due to the little sample size of each series, seven data are not enough to build ARIMA. Thus the traditional ARIMA model and hybrid ARIMA-GRNN model could not be revised. The GRNN model requires less about sample size so it was revised. Five revised GRNN model showed a better forecasting performance than original GRNN model. These results alert the SSH confounder in time series prediction model application, especially in huge region or diverse territory and the results also remind us of building model in same strata. According to these results, it could be inferred that revised ARIMA model and hybrid model may have better performance than existing models. More data are needed to revised these two models.

Several limitations of this study should be noted. As is shown above, the prediction model was merely developed for short-term forecasting. Maintaining the prediction performance for months or years requires constantly update of data and model. Here we build three new models whose fitting data were HFRS incidence from January 2011 to December 2015 and data from January 2016 to May 2018 were used to verification (Table S1). It showed that model with new data has higher accuracy. Besides, this study only analyzed the incidence of HFRS in China from January 2011 to December 2017 and the sample size is not enough when building model in different strata. Although the revised GRNN model demonstrated that SSH should be considered, the ARIMA-GRNN model were not revised due to little sample size. A time series with more data than this study is required to revised the hybrid model and improve the model's performance. At last, HFRS incidence data in this manuscript was total incidence in China, we can not explore the performance of these models in provincial incidence prediction. Spatial factor is an important factor that can affect HFRS development, so the applicability of results in this research need further study.

CONCLUSIONS

The hybrid ARIMA-GRNN model is superior than the single ARIMA model and basic GRNN

model both in fitting and forecasting of monthly incidence of HFRS in China. The data should keep update to maintain the forecasting performance. This hybrid model should be considered as a decision-making tool in HFRS prevention and control.

Supporting information

S1 Table. The fitting and forecasting performance of three new models.

Acknowledgement We would like to express our gratitude to professor Yin Yang for carefully revise of overall readability. We also thank peer reviewers for carefully revising and useful comments.

Contributors YJ, YW and ZS designed the study. ZS extracted the data and constructed the database. YW and ZS analyzed the data. YW drafted the manuscript. YJ and ZS made critical revision to the manuscript. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All original data was available at the official website of National Health Commission of the People's Republic of China (http://www.nhfpc.gov.cn/jkj/new_index.shtml).

REFERENCES

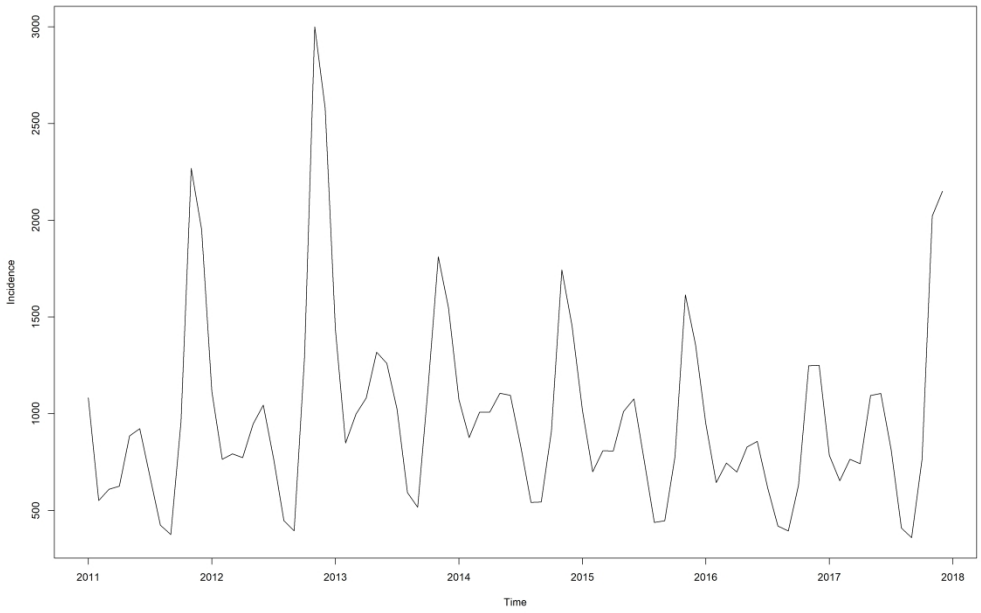
1. Schmaljohn CS, Hasty SE, Dalrymple JM, et al. Antigenic and genetic properties of viruses linked to hemorrhagic fever with renal syndrome. *Science.* 1985;227(4690):1041-1044.
2. Ehelepola NDB, Basnayake BMLS, Sathkumara SMBY, Kaluphana KLR. Two Atypical Cases of Hantavirus Infections from Sri Lanka. *Case Reports in Infectious Diseases.* 2018;2018:1-6. DOI: 10.1155/2018/4069862.
3. Avsic-Zupanc T, Saksida A, Korva M. Hantavirus infections. *Clin Microbiol Infect.* 2015. DOI: 10.1111/1469-0691.12291.
4. Latus J, Schwab M, Tacconelli E, et al. Clinical course and long-term outcome of hantavirus-associated nephropathia epidemica, Germany. *Emerg Infect Dis.* 2015;21(1):76-83. DOI: 10.3201/eid2101.140861.
5. Vaheri A, Strandin T, Hepojoki J, et al. Uncovering the mysteries of hantavirus infections. *Nat Rev Microbiol.* 2013;11(8):539-550. DOI: 10.1038/nrmicro3066.
6. Hepojoki J, Vaheri A, Strandin T. The fundamental role of endothelial cells in hantavirus pathogenesis. *Front Microbiol.* 2014;5:727. DOI: 10.3389/fmicb.2014.00727.
7. Pal E, Korva M, Resman RK, et al. Sequential assessment of clinical and laboratory parameters in patients with hemorrhagic fever with renal syndrome. *Plos One.* 2018;13(5):e197661. DOI: 10.1371/journal.pone.0197661.
8. Bi P, Parton KA. El Nino and incidence of hemorrhagic fever with renal syndrome in China. *JAMA.* 2003;289(2):176-177.
9. Zhang S, Wang S, Yin W, et al. Epidemic characteristics of hemorrhagic fever with renal syndrome in China, 2006-2012. *Bmc Infect Dis.* 2014;14:384. DOI: 10.1186/1471-2334-14-384.
10. Du H, Wang PZ, Li J, et al. Clinical characteristics and outcomes in critical patients with hemorrhagic fever with renal syndrome. *Bmc Infect Dis.* 2014;14:191. DOI:

10.1186/1471-2334-14-191.

11. Spatiotemporal Transmission Dynamics of Hemorrhagic Fever with Renal Syndrome in China, 2005−2012. DOI: 10.1371/journal.pntd.0003344.
12. He X, Wang S, Huang X, Wang X. Changes in age distribution of hemorrhagic fever with renal syndrome: an implication of China's expanded program of immunization. *Bmc Public Health.* 2013;13:394. DOI: 10.1186/1471-2458-13-394.
13. Wang YW, Shen ZZ, Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. *Plos One.* 2018;13(9):e201987. DOI: 10.1371/journal.pone.0201987.
14. Cao H, Wang J, Li Y, et al. Trend analysis of mortality rates and causes of death in children under 5 years old in Beijing, China from 1992 to 2015 and forecast of mortality into the future: an entire population-based epidemiological study. *Bmj Open.* 2017;7(9):e15941. DOI: 10.1136/bmjopen -2017-015941.
15. Petukhova T, Ojkic D, McEwen B, Deardon R, Poljak Z. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. *Plos One.* 2018;13(6):e198313. DOI: 10.1371/journal.pone.0198313.
16. Yosipof A, Guedes RC, Garcia-Sosa AT. Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. *Front Chem.* 2018;6:162. DOI: 10.3389/fchem.2018.00162.
17. Nair TM. Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia. *Comput Biol Chem.* 2018;75:222-230. DOI: 10.1016/j.compbiolchem.2018.05.011.
18. Wei W, Jiang J, Gao L, et al. 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. *Am J Trop Med Hyg.* 2017;97(3):799-805. DOI: 10.4269/ajtmh.16-0648.
19. Wu W, Guo J, An S, et al. Comparison of Two Hybrid Models for Forecasting the Incidence of Hemorrhagic Fever with Renal Syndrome in Jiangsu Province, China. *Plos One.* 2015;10(8):e135492. DOI: 10.1371/journal.pone.0135492.
20. Rubaihayo J, Tumwesigye NM, Konde-Lule J, Makumbi F. Forecast analysis of any opportunistic infection among HIV positive individuals on antiretroviral therapy in Uganda. *Bmc Public Health.* 2016;16(1):766. DOI: 10.1186/s12889-016-3455-5.
21. Wei W, Jiang J, Liang H, et al. Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. *Plos One.* 2016;11(6):e156768. DOI: 10.1371/journal.pone.0156768.
22. Wang J, Zhang T, Fu B. A measure of spatial stratified heterogeneity. *Ecol Indic.* 2016;67: 250-256.DOI: 10.1016/j.ecolind.2016.02.052.
23. Li J, Xu F, Sun Z, Wang J. Regional differences and spatial patterns of health status of the member states in the "Belt and Road" Initiative. *Plos One.* 2019;14(1):e211264.DOI: 10.1371/journal.pone.0211264.
24. Gan R, Chen N, Huang D. Comparisons of forecasting for hepatitis in Guangxi Province, China by using three neural networks models. *Peerj.* 2016;4:e2684. DOI: 10.7717/peerj.2684.
25. Hansen A, Cameron S, Liu Q, et al. Transmission of Haemorrhagic Fever with Renal Syndrome in China and the Role of Climate Factors: A Review. *Int J Infect Dis.* 2015;33:212-218. DOI:

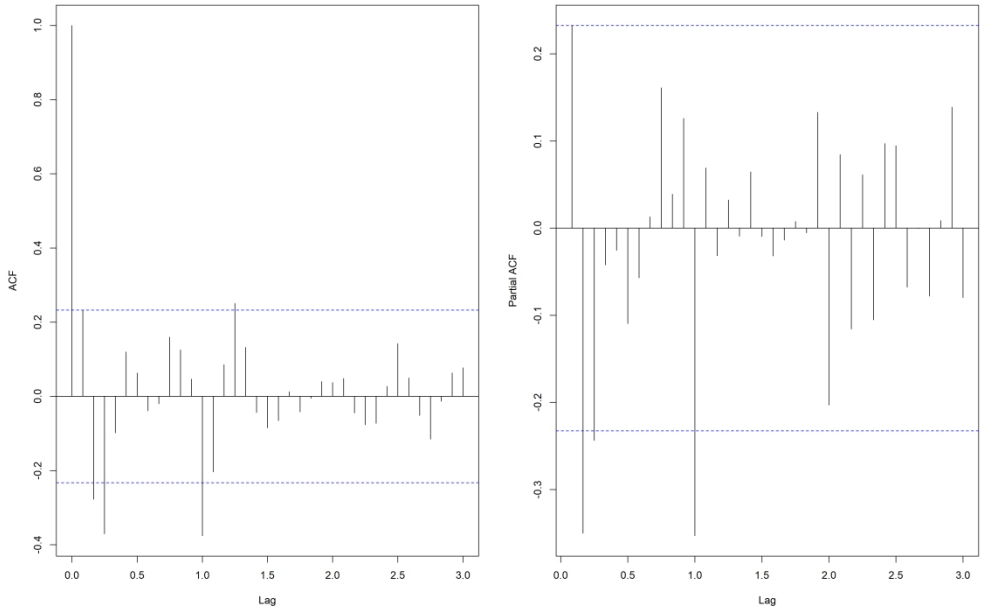
10.1016/j.ijid.2015.02.010.

26. Liu YX, Feng D, Zhang Q, et al. Key differentiating features between scrub typhus and hemorrhagic fever with renal syndrome in northern China. *Am J Trop Med Hyg.* 2007;76(5):801-805.
27. Park YH. Absence of a Seasonal Variation of Hemorrhagic Fever with Renal Syndrome in Yeoncheon Compared to Nationwide Korea. *Infect Chemother.* 2018;50(2):120-127. DOI: 10.3947/ic.2018.50.2. 120.
28. Mills JN, Gage KL, Khan AS. Potential influence of climate change on vector-borne and zoonotic diseases: a review and proposed research plan. *Environ Health Perspect.* 2010;118(11):1507-1514. DOI: 10.1289/ehp.0901389.
29. Wang H, Tian CW, Wang WM, Luo XM. Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiol Infect.* 2018;146(8):935-939. DOI: 10.1017/S0950268818001115.
30. Peng Y, Yu B, Wang P, Kong DG, Chen BH, Yang XB. Application of seasonal auto-regressive integrated moving average model in forecasting the incidence of hand-foot-mouth disease in Wuhan, China. *J Huazhong Univ Sci Technolog Med Sci.* 2017;37(6):842-848. DOI: 10.1007/s11596-017-1815-8.
31. Joshi YP, Kim E, Cheong H. The influence of climatic factors on the development of hemorrhagic fever with renal syndrome and leptospirosis during the peak season in Korea: an ecologic study. *Bmc Infect Dis.* 2017;17(1). DOI: 10.1186/s12879-017-2506-6.
32. Han SS, Kim S, Choi Y, Kim S, Kim YS. Air pollution and hemorrhagic fever with renal syndrome in South Korea: an ecological correlation study. *Bmc Public Health.* 2013;13:347. DOI: 10.1186/1471-2458-13-347.
33. Xiang J, Hansen A, Liu Q, et al. Impact of meteorological factors on hemorrhagic fever with renal syndrome in 19 cities in China, 2005-2014. *Sci Total Environ.* 2018;636:1249-1256. DOI: 10.1016/j.scitotenv.2018.04.407.



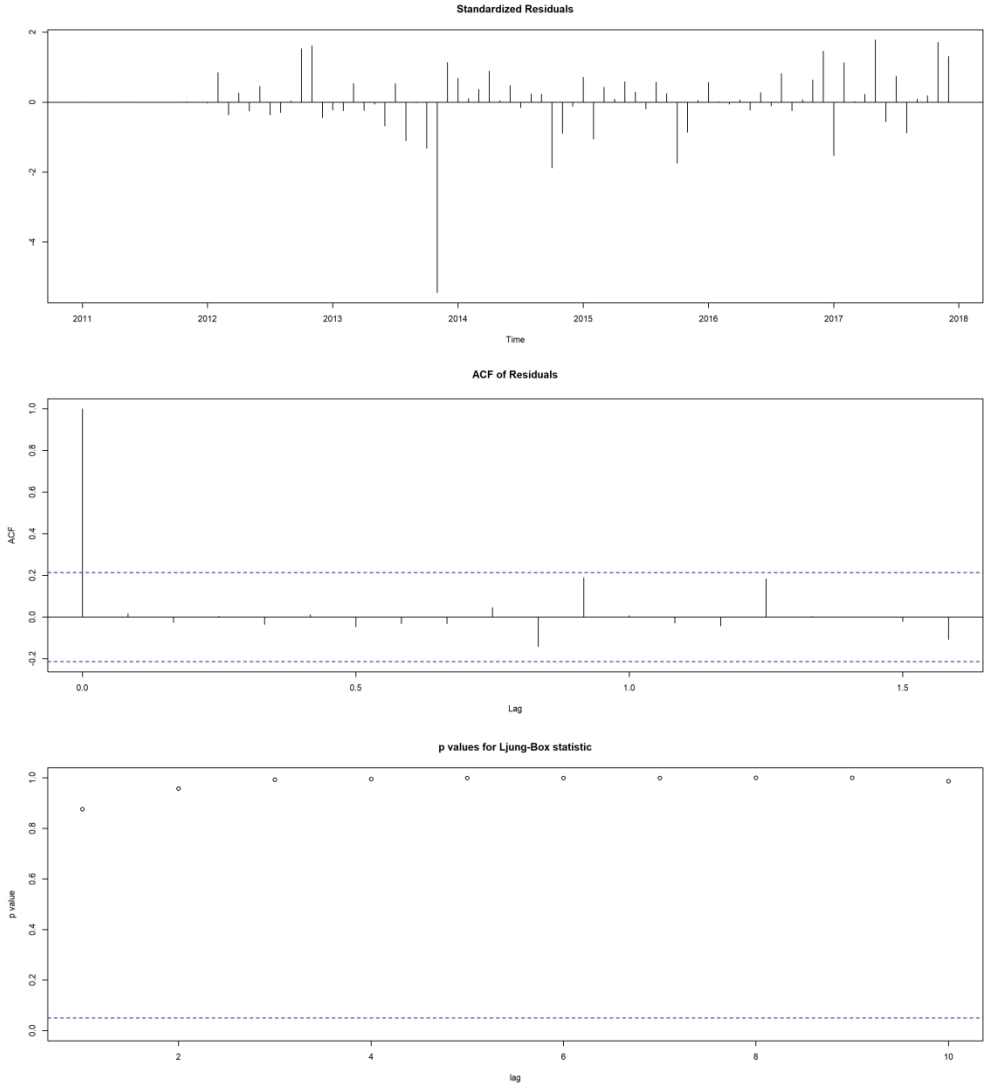
Monthly incidence of HFRS in China from January 2011 to December 2017.

406x270mm (300 x 300 DPI)



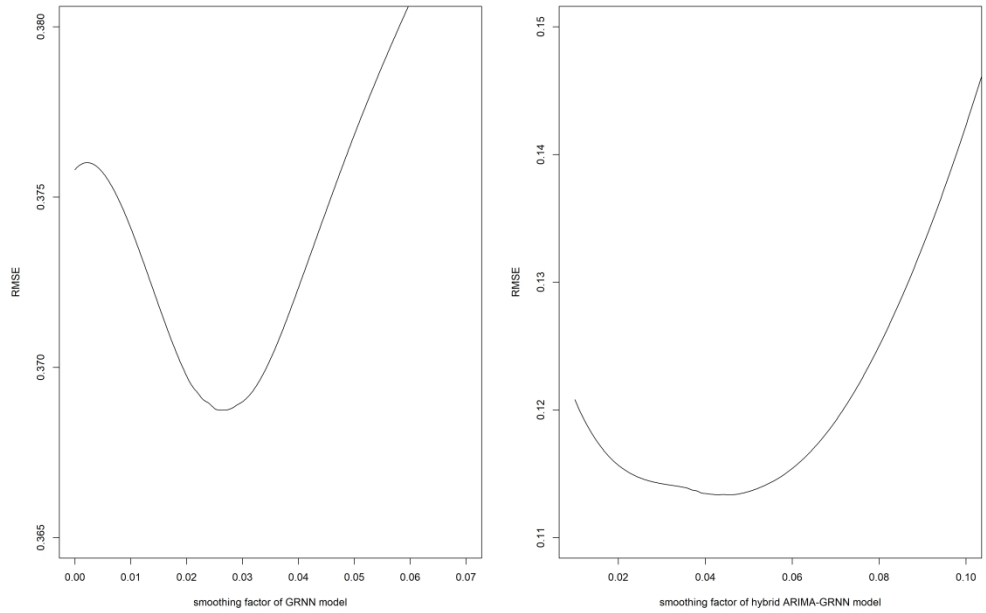
The ACF and PACF graphs of differenced HFRS incidence series.

406x256mm (300 x 300 DPI)



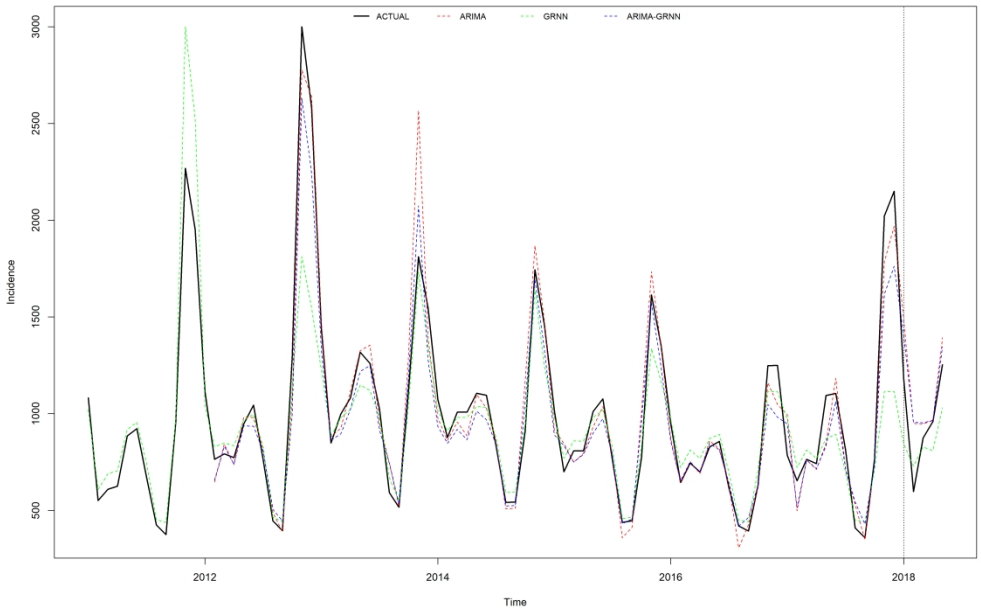
Residual white noise test

321x355mm (300 x 300 DPI)



The selection of basic GRNN model and hybrid ARIMA-GRNN model

406x270mm (300 x 300 DPI)



The fitting and forecasting curves of the three models and the actual HFRS incidence series

406x270mm (300 x 300 DPI)

S1 Table The fitting and forecasting performance of three new models

	Fitting part			Forecasting part		
Predicting error	MAPE	MAE	RMSE	MAPE	MAE	RMSE
ARIMA	10.2735	105.6382	155.7399	27.9010	259.9656	359.9456
GRNN	33.4315	325.6638	512.7855	41.1768	299.8275	402.3660
ARIMA-GRNN	22.5002	213.6670	248.8867	21.3148	221.3138	336.5332

Table 1. Checklist of Items to Include When Reporting a Study Developing or Validating a Multivariable Prediction Model for Diagnosis or Prognosis*

Section/Topic	Item	Development or Validation?	Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model, or both	3
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	4
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres	NA
	5b	D;V	Describe eligibility criteria for participants	NA
	5c	D;V	Give details of treatments received, if relevant	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	NA
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted	NA
Predictors	7a	D;V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	4-5
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors	NA
Sample size	8	D;V	Explain how the study size was arrived at	4
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	NA
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses	4
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	4
	10c	V	For validation, describe how the predictions were calculated	4-5
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models	5

	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors	3
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	NA
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome)	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis	5-6
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point)	5-6
	15b	D	Explain how to use the prediction model	6-7
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model	6
Model updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance)	6-7
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data)	7
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data	7
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	7
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research	7-8
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	8
Funding	22	D;V	Give the source of funding and the role of the funders for the present study	NA

Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.