

1                   **Transmission potential and forecasting the number of**  
2                   **Coronavirus disease 2019 in Hubei Province, China**

3           Ke-wei Wang<sup>1,2,#</sup>, Jie Gao<sup>3#</sup>, Hua Wang<sup>2</sup>, Xiao-long Wu<sup>1</sup>, Qin-fang Yuan<sup>1</sup>, Yang Cheng<sup>2</sup>

4           <sup>1</sup> Department of hospital infection, Affiliated Hospital of Jiangnan University, Wuxi, China

5           <sup>2</sup> Wuxi School of Medicine, Jiangnan University, Wuxi, China

6           <sup>3</sup>Medicine, Shandong Province Maternal and Child Health Care Hospital Affiliated to  
7           Shandong University, Jinan, China

8           #Ke-Wei Wang and Jie Gao contributed equally to this work

9           \*Correspondence: Yang Cheng, MD, Department of public health and Preventive Medicine,

10          Wuxi School of Medicine, Jiangnan University, no. 1800, lihu avenue, wuxi, jiangsu, 214122,  
11          China. woerseng@126.com

12

13

14 **Abstract**

15 **Background:** Coronavirus disease 2019 (COVID-19) was first reported in Wuhan, Hubei  
16 province, China. We aimed to describe the temporal and spatial distribution and the  
17 transmission dynamics of COVID-19 and to assess whether a hybrid model can forecast the  
18 trend of COVID-19 in Hubei Province, China.

19 **Method:** The data of COVID-19 cases were obtained from the websites of Chinese Center for  
20 Disease Control and Prevention, whereas the data on the resident population were obtained  
21 from the websites of Hubei Provincial Bureau of Statistics. The temporal and spatial  
22 distribution and the transmission dynamics of COVID-19 were described. A combination of  
23 autoregressive integrated moving average (ARIMA) and support vector machine was  
24 constructed to forecast the trend of COVID-19.

25 **Results:** A total of 56,062 confirmed COVID-19 cases, which were mainly concentrated in  
26 Wuhan, were reported from January 16 to March 16, 2020 in Hubei Province, China. The daily  
27 number of confirmed cases exponentially increased to 3,156 before February 4, 2020,  
28 fluctuated to 4,823 before February 13, 2020, and then markedly decreased to 1 after March  
29 16, 2020. The highest mean reproduction number  $R(t)$  of 9.48 was recorded on January 16,  
30 2020, after which it decreased to 2.15 on February 2, 2020 and further decreased to less than 1  
31 on February 13, 2020. In the modeling stage, the mean square error, mean absolute error, and  
32 mean absolute percentage error of the hybrid ARIMA–SVM model decreased by 98.59%,  
33 89.19% and 89.68%, and those of SVM decreased by 98.58%, 87.71%, and 88.94%,  
34 respectively, compared with the ARIMA model. Similar results were obtained in the  
35 forecasting stage.

36 **Conclusion:** Public health interventions resulted in the terminal phase of COVID-19 in Hubei  
37 province. The hybrid ARIMA–SVM model may be a reliable tool for forecasting the trend of

38 the COVID-19 epidemic.

39 **Keywords** COVID-19, transmission, SARS-CoV-2, ARIMA–SVM, R(t)

## 40 **Background**

41 In December 2019, coronavirus disease 2019 (COVID-19) was detected in Wuhan, China and  
42 then quickly spread to other cities outside of Wuhan and other countries[1, 2]. As of May 28,  
43 2020, a total of 5,690,182 confirmed COVID-19 cases had been reported in 188 countries or  
44 regions globally [3]. Although recent studies have illustrated the epidemiology of COVID-19  
45 in Wuhan, the capital of Hubei Province, China, the COVID-19 transmission intensity from  
46 Wuhan to other cities of Hubei Province has not been estimated. Therefore, a retrospective  
47 study of the temporal and spatial distribution and the transmission dynamics of COVID-19 in  
48 Hubei Province, China was needed.

49  
50 Considering the miscalculation of the early epidemic trend, the COVID-19 pandemic has  
51 become a large economic burden globally. Many statistical methods have been used to predict  
52 the trend of the disease[4, 5]. Autoregressive integrated moving average (ARIMA) is the most  
53 popular time series used to forecast infectious diseases. The ARIMA facilitates a linear  
54 relationship between the present and past data points and errors in a time series[6]. However,  
55 infectious diseases are characterized by several challenges, such as uncertainty, complexity,  
56 and non-linear pattern[7, 8]. The artificial neural network (ANN) is good for forecasting via  
57 complex nonlinear mapping without prior knowledge of problem solving[9, 11].

58  
59 Therefore, this study was designed to describe the temporal and spatial distribution and the  
60 transmission dynamics of COVID-19 between January 16 and March 13, 2020 in Hubei  
61 Province, China. We also created a hybrid ARIMA-support vector machine (SVM) model to

62 forecast the trend of COVID-19 in Hubei Province, China. Our findings may serve as important  
63 references for the prevention and control of the outbreak.

## 64 **Methods**

### 65 **Data collection**

66 The time series of the observations of confirmed COVID-19 cases from January 16 to March  
67 16, 2020 was obtained from the websites of Chinese Center for Disease Control and Prevention  
68 (<http://2019ncov.chinacdc.cn/2019-nCoV/>). The information collected included the number of  
69 confirmed cases in Hubei Province, China. Data of resident population were obtained from the  
70 websites of Hubei Provincial Bureau of Statistics (<http://tjj.hubei.gov.cn/>).

71

72 A confirmed case was defined as a suspected case with a positive result for severe acute  
73 respiratory syndrome coronavirus 2 (SARS-CoV-2) according to real-time reverse  
74 transcriptase–polymerase chain reaction (RT-PCR) assay or high-throughput sequencing of  
75 nasal and pharyngeal swab specimens[12]. The real-time reproduction number,  $R(t)$ , is the  
76 expected number of secondary cases that each infected individual would infect if the conditions  
77 remained as they were at time  $t$ . A Bayesian statistical framework was used to calculate  $R(t)$  on  
78 the basis of the number of COVID-19 cases, the number of secondary cases, the serial interval,  
79 and the five-day moving average[12, 13].  $R(t)$  and its 95% credible interval (CI) for the entire  
80 period were calculated.

81

82 Three models, namely, ARIMA, SVM, and ARIMA–SVM, were constructed to forecast the  
83 trend of COVID-19. We compared the performance of these models in the train dataset and the  
84 test dataset. Initially, 58 days of data between January 16 and March 13, 2020 were analyzed  
85 to construct the single ARIMA model. Considering the small sample size, the predictive

86 validity of the models was calculated using three days of data from March 14, 2020 to March  
87 16, 2020.

### 88 **ARIMA model**

89 The ARIMA model was first developed in 1970[14]. ARIMA was calculated using the  
90 following formula: ARIMA ( $p, d, q$ ), where  $p$  denotes the number of parameters in the  
91 autoregressive (AR) model,  $d$  represents the degree of difference, and  $q$  stands for the number  
92 of parameters in the moving average (MA) model. A logarithmic transform ARIMA model ( $p,$   
93  $d, q$ ) was constructed by the autonomous identification in R version 3.6.2. Several ARIMA  
94 models were identified. The optimum model was selected using Akaike's information criterion  
95 (AIC). White noise tests were used to determine the independence and normal distribution of  
96 the residual data.

### 97 **SVM**

98 SVM is a machine learning tool that is used to carry nonlinear classification and regression of  
99 data[15]. For the classification of subjects, SVM constructs an optimal classification  
100 hyperplane by mapping data to a high dimensional space. In this study, steps of establishing  
101 the SVM model for non-linear regression were as follows:

- 102 1) The number of daily new confirmed COVID-19 cases in Hubei Province was  $y_t$ , and the  
103 sliding window is used to train SVM for  $y_t$ . The value of  $y_{(t+1)}$  was predicted  
104 using SVM constructed by  $y_t, y_{(t-1)}, \dots, y_{(t-k+1)}$ . Considering the small sample size,  
105 four loops of data were selected, and the test dataset only included the data of cases in the  
106 last 3 days.
- 107 2) These optimal parameters of SVM were determined by gridding search to minimize the  
108 mean square error (MSE) between the predicted and actual values for the training model.
- 109 3) SVM was used to predict the daily number of confirmed cases in Hubei province.

### 110 **ARIMA-SVM model**

111 The hybrid model developed in this study combined ARIMA (linear approach) and SVM (non-  
112 linear approach). The three-step methodology was as follows. In the first step, an ARIMA  
113 model was used to analyze the linear dimension. In the second step, the SVM model was used  
114 to model the residuals derived from the ARIMA model.  $e_t$  represented the residual parameter  
115 at time  $t$  in the ARIMA model, in which  $e_t = y_t - \hat{y}_t$ , where  $\hat{y}_t$  denotes the forecast value, and  
116  $y_t$  represents the daily number at time  $t$ . In the third step, the predicted error ( $\hat{e}_t$ ), denoting the  
117 estimation of  $e_t$ , was obtained using the SVM model (nonlinear approach). The ARIMA–SVM  
118 model yields the predicted value  $\hat{\mu}_t = \hat{y}_t + \hat{e}_t$ .

### 119 **Forecast evaluation**

120 The performance of the model was determined on the basis of the similarity between the  
121 forecast values and test results and the observed data. Three different measures were used to  
122 compare the performance of the ARIMA, SVM, and ARIMA–SVM models, including, MSE,  
123 mean absolute error (MAE), and mean absolute percentage error (MAPE), respectively. Lower  
124 values suggest higher prediction accuracy. In other words, an optimal prediction model was  
125 obtained by minimizing the MSE, MAE, and MAPE.

126

$$127 \quad MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2$$

$$128 \quad MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|$$

$$129 \quad MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{y_t}$$

130

### 131 **Data processing and analysis**

132 ArcGIS software version 10.6 (Environmental Systems Research Institute Inc) was used to plot  
133 the geographical locations of daily new confirmed COVID-19 cases across Hubei province

134 from 16 January 2020 to 16 March 2020, and we carried out descriptive analyses of the dates  
135 of illness onset. Other analyses were carried in R version 3.6.2. In all analyses, a two-tailed  
136 significance level of 0.05 was used.

### 137 **Ethical review**

138 The study protocol and utilization of COVID-19 case data were obtained from the websites of  
139 Chinese Center for Disease Control and Prevention (<http://2019ncov.chinacdc.cn/2019-nCoV/>)  
140 and Hubei Provincial Bureau of Statistics (<http://tjj.hubei.gov.cn/>). No ethical issues were  
141 identified.

## 142 **Results**

### 143 **Temporal patterns and reproduction number ( $R(t)$ ) of the COVID-19 cases**

144 A total of 56, 062 confirmed COVID-19 cases were reported from January 16 to March 16,  
145 2020 in Hubei province, China. The daily confirmed case exponentially increased to 3,156  
146 before February 4, 2020, fluctuatingly increased to 4,823 before February 13, 2020, and then  
147 markedly decreased to 1 after March 16, 2020 (Figure 1). The occurrence of COVID-19 was  
148 initially concentrated in Wuhan before January 23, 2020 and spread rapidly to all areas in Hubei  
149 province by January 29 (Figure 2). Although Wuhan had the highest rate, other areas except  
150 for Enshi in Hubei Province had a rate of above 0.5 case per 10,000 people between January  
151 29 and February 18. After February 19, the daily new COVID-19 cases gradually decreased in  
152 Hubei province and were concentrated in Wuhan. The highest mean reproduction number  $R(t)$   
153 of 9.48 was recorded on January 16, 2020, then dropped to 2.15 on February 2, 2020. This  
154 further decreased to less than 1 on February 13, 2020 (Figure 3).

### 155 **Optimized ARIMA model**

156 After the logarithmic transform for the number of new confirmed COVID-19 cases,  
157 autonomous identification in R version 3.6.2 was used to construct ARIMA. Based on the

158 results of the goodness-of-fit test statistics, we confirmed ARIMA (0, 2, 1) as the optimal model,  
 159 with the lowest AIC (83.10) of the six candidate models. The parameter estimates of the  
 160 ARIMA model are shown in Table 1. The new confirmed COVID-19 cases from March14,  
 161 2020 to March 16, 2020 were predicted using the constructed ARIMA model (0, 2, 1). The QQ  
 162 graph and box-ljung test ( $\chi^2= 1.0268, p = 0.3109$ ) of the residuals suggested that the residuals  
 163 were normal and were not correlated (Figure 4). Hence, the model fitting data was sufficient  
 164 and can be used for prediction.

165 **Table 1** Model estimation of the ARIMA (0, 2, 1) model

Variable	Coefficient	Standard error	t statistics	P values
MA (1)	-0.8506	0.0603	-14.106	<0.0001

166

167 **Single SVM model**

168 The SVM model facilitated the prediction of the number of confirmed cases in Hubei Province,  
 169 China by representing nonlinear and complex data. According to the steps of establishing SVM  
 170 model, the previous four-time data were used as input parameters into the model. Considering  
 171 that the input data had a small sample size, we did not use cross-validation to determine the  
 172 model. We adopted the eps-regression model and used different kernel functions. The radial  
 173 was used as the kernel function, and we obtained the best delay predicting result, in which the  
 174 number of support vectors was eight, the kernel parameter was  $\gamma= 1e-04$ , and insensitive loss  
 175 coefficient was  $\varepsilon=0.1$ . The prediction results were compared with the initial results, as shown  
 176 in Figures 1 and 5.

177 **The hybrid ARIMA-SVM**

178 The residual series from ARIMA was used as the target series for the SVM. Similar to the  
 179 single SVM model, the SVM model for residual series was constructed with five support  
 180 vectors, kernel parameter of  $\gamma = 0.001$ , and insensitive loss coefficient of  $\varepsilon=0.1$ . The

181 prediction results were compared with the initial results, as shown in Figures 1 and 5.

182

### 183 **Prediction accuracy**

184 The parameters of MAE, MSE, and MAPE in the ARIMA model and the SVM and The hybrid  
185 ARIMA-SVM model in the train dataset and the test dataset were shown in Table 2. In the  
186 modelling stage using the train dataset, the MSE (7,144.944), MAE (37.333), and MAPE  
187 (0.042) of the ARIMA-SVM were lower than those of the ARIMA and SVM models.  
188 Furthermore, in the modeling stage, the MSE, MAE, and MAPE of the hybrid ARIMA-SVM  
189 model decreased by 98.59%, 89.19% and 89.68%, whereas those of SVM decreased by 98.58%,  
190 87.71%, and 88.94%, respectively, compared with the ARIMA model. In the forecasting stage  
191 using the test dataset, the hybrid ARIMA-SVM model had lower MSE, MAE, and MAPE than  
192 the ARIMA model, whereas the SVM had higher MSE, MAE, and MAPE than the ARIMA  
193 and hybrid ARIMA–SVM models. Therefore, the ARIMA–SVM model was superior to the  
194 ARIMA and SVM models in accurately forecasting the number of COVID-19 cases in Hubei  
195 Province, China.

196

197

Table 2 Comparison of the modeling and forecasting performance of ARIMA, SVM and ARIMA-SVM model

Model	train dataset			test dataset		
	MSE	MAE	MAPE	MSE	MAE	MAPE
ARIMA	505840.134	345.324	0.407	2.282	1.418	0.892
SVM	7165.682	42.431	0.045	30.355	5.045	2.194
ARIMA-SVM	7144.944	37.333	0.042	2.071	1.417	0.776

198

199

200

201 **Discussion**

202 In this study, 56,062 confirmed COVID-19 cases, which were mainly concentrated in Wuhan,  
203 were reported from January 16 to March 16, 2020 in Hubei Province, China. The temporal and  
204 spatial distribution and the transmission dynamics of COVID-19 were also described. Three  
205 models, namely, ARIMA, SVM, and ARIMA-SVM, were constructed using time series to  
206 predict the number of COVID-19 cases. The values of MSE, MAE, and MAPE of the ARIMA–  
207 SVM model were the least among the three models, suggesting the superiority of the model in  
208 predicting the daily number of COVID-19 cases in Hubei Province, China.

209

210 Considering the outbreak of COVID-19 in Wuhan, China, Hubei Province's response was  
211 immediate and emphatic. To stop the spread of SARS-CoV-2, the government imposed public  
212 health interventions, including city lockdown, traffic suspension, home quarantine, compulsory  
213 mask wearing in public places, and cancellation of social gatherings after January 23, 2020. A  
214 stay-at-home policy was implemented, and Fang Cang Shelter hospitals were used after  
215 February 4, 2020. Door-to-door and individual-to-individual symptom screening was  
216 performed for all residents after February 17, 2020 in Hubei Province[12, 16, 17]. In Hubei  
217 Province, the COVID-19 epidemic was of approximately 60 days' duration from four cases  
218 reported on January 16, 2020 to the peak value of mean 2201.76 cases daily between January  
219 29 and February 18, 2020 and then to the mean of 233.78 between February 19 and March 16,  
220 2020. The end of epidemics COVID-19 was about to come true in Hubei Province[12, 16].  
221 These changes of time trend on new COVID-19 cases indicated the effectiveness of some  
222 public health interventions implemented in Hubei province in preventing the spread of COVID-  
223 19[16-19]. Additionally, with the implementation of some public health interventions, the  $R(t)$   
224 gradually decreased and eventually reached less than 1 after February 13, 2020. This result  
225 further emphasized that some public health interventions effectively stopped the spread of

226 COVID-19.

227

228 The early and accurate prediction of COVID-19 can provide an important reference for  
229 outbreak control. ARIMA, which explores the dynamic development and change of infectious  
230 disease over time, has been widely used to early predict kinds of infectious diseases, such as  
231 tuberculosis, schistosoma, malaria, and bacterial dysentery[20-23]. However, ARIMA cannot  
232 adjust some confounding factors that influence infectious disease and estimate some  
233 parameters when the data are complex; hence, it results in the inaccurate prediction of  
234 outbreaks in small samples[14]. Since COVID-19 was first detected in Wuhan, various public  
235 health interventions have been used to improve the control of the outbreak[12]. These  
236 interventions inevitably disrupted the original onset state of the disease and have led to the  
237 instability and discontinuity of the time series data, thus making ARIMA unsuitable.  
238 Furthermore, the ARIMA based on linear model cannot capture the information of nonlinear  
239 data of time series. With the development of machine learning, the prediction accuracy of the  
240 disease can be improved substantially through machine learning[9]. Among the machine  
241 learning methods, the SVM has excellent generalization ability and can help solve many  
242 unpredictable problems well. The SVM can perform classification and regression by  
243 training[15]. Nowadays, the SVM is widely used in time series by fitting small sample data,  
244 high-dimensional data, non-linear data, and other complex data with high prediction accuracy.

245

246 To utilize the advantages of various models, many studies have constructed the combination  
247 of kinds of models to carry out a time series for infectious diseases, such as tuberculosis,  
248 schistosoma, malaria, and bacterial dysentery[5, 8, 20, 24]. Compared with the single model,  
249 the hybrid model can fully utilize all kinds of sample information, in more systematic and  
250 comprehensive way than the single model. A study has reported the application of the ARIMA

251 model on the COVID-19 epidemic dataset from the website of Johns Hopkins university[25].  
252 However, this study only visualizes the predicted epidemiological trend of the prevalence and  
253 incidence of COVID-19 and lacks the values of prediction accuracy, including MSE, MAE,  
254 and MAPE, suggesting the need for further comparison and data collection in real time. By  
255 contrast, our study, which included 56,062 confirmed COVID-19 cases (the laboratory  
256 confirmation of SARS-CoV-2infection in the biosamples) from January 16 to March 16, 2020  
257 in Hubei Province, can meet the requirements for further comparison. Another study that  
258 evaluates the spatial dependency and temporal dynamics of COVID-19 demonstrates that  
259 incidence rates are concentrated in Wuhan Metropolitan area, but the prediction of the trend of  
260 COVID-2019 is lacking[26]. In this study, ARIMA was used to fit the linear part of the time  
261 series of daily new confirmed cases of COVID-19 in Hubei Province, whereas the SVM model  
262 was used to fit the nonlinear part of the time series. The hybrid ARIMA–SVM model with the  
263 advantages of ARIMA and SVM was established to predict the number of COVID-19 cases.  
264 The COVID-19 is a new infectious disease, which may be influenced by population density,  
265 environmental changes, public health interventions, and climate[27, 28]. Therefore, a multi-  
266 factor prediction model for COVID-19 is needed. However, considering the emergence of this  
267 infectious disease outbreak, we did not have enough time to understand the transmission of the  
268 disease. The combined model may be a reliable forecasting tool based on the predicted effect  
269 of the hybrid ARIMA-SVM model, which was better than those of ARIMA and SVM alone.

## 270 **Conclusions**

271 With the implementation of some public health interventions, the number of daily new COVID-  
272 19 cases and the  $R(t)$  first increased rapidly and then decreased gradually from January 16,  
273 2020 to March 16, 2020 in Hubei Province, China. This change emphasized that some public  
274 health interventions effectively stopped the spread of COVID-19. Furthermore, the hybrid

275 ARIMA-SVM model may be a reliable forecasting tool for the trend of COVID-19 in Hubei  
276 Province, China. Our findings may have public health implications for the prevention and  
277 control of infectious disease, especially COVID-19.

#### 278 **Declarations**

#### 279 **Abbreviations**

280 COVID-19: coronavirus disease 2019; SARS-CoV-2: Severe acute respiratory syndrome  
281 coronavirus 2; RT-PCR: Real-time fluorescent reverse transcription polymerase chain reaction;  
282 R(t): real-time reproduction number; CI: credible intervals; SARS: Severe Acute Respiratory  
283 Syndrome; MERS: Middle East Respiratory Syndrome.

#### 284 **Ethics approval and consent to participate**

285 Not applicable.

#### 286 **Consent for publication**

287 Not applicable.

#### 288 **Availability of data and materials**

289 Not applicable.

#### 290 **Competing interests**

291 The authors declare that they have no competing interests.

#### 292 **Funding**

293 This work was supported by The National Natural Science Foundation of China (81871681),  
294 The National First-class Discipline Programme of Food Science and Technology  
295 (JUFSTR20180101), The Fundamental Research Funds for the Central Universities funded by  
296 the Ministry of Education of China (JUSRP51710A). The funders had no role in study plan,  
297 design, data collection and analysis, interpretation of the results, decision to publish, or  
298 preparation of the manuscript.

299 **Authors' contributions**

300 YC had full access to all of the data in the study and take responsibility for the integrity of the  
301 data and the accuracy of the data analysis. KW and JG conceived the study. HW, XW and QY  
302 analyzed data and made the figures. KW and YC wrote the first draft of the manuscript. All  
303 authors interpreted the results, and approved the final version for submission.

304 **Acknowledgements**

305 Not applicable.

306 **References**

- 307 1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al.* A Novel Coronavirus from Patients  
308 with Pneumonia in China, 2019. *N Engl J Med* 2020;382(8):727-733.
- 309 2. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health  
310 concern. *Lancet* 2020;395(10223):470-473.
- 311 3. Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering  
312 (CSSE) at Johns Hopkins University (JHU), <https://coronavirus.jhu.edu/map.html>.  
313 Accessed: 28 May 2020
- 314 4. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density  
315 ensembles. *PLoS Comput Biol* 2018;14(2):e1005910.
- 316 5. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods  
317 for clinical prediction models. *Stat Methods Med Res* 2018;27(1):185-197.
- 318 6. Zheng A, Fang Q, Zhu Y, Jiang C, Jin F, Wang X. An application of ARIMA model for  
319 predicting total health expenditure in China from 1978-2022. *J Glob Health*  
320 2020;10(1):010803.
- 321 7. Xie J, Kanghuai Z, Cai Y, Wang N, Wang H, Zhao Y, *et al.* Application of ARIMA model  
322 in monitoring the use rate of antibiotics in outpatients in 2010 - 2018. *Int J Clin Pharmacol*  
323 *Ther* 2020;58(5):282 - 288.

- 324 8. Fang X, Liu W, Ai J, He M, Wu Y, Shi Y, *et al.* Forecasting incidence of infectious  
325 diarrhea using random forest in Jiangsu Province, China. *BMC Infect Dis* 2020;20(1):222.
- 326 9. Guo Z, He K, Xiao D. Early warning of some notifiable infectious diseases in China by  
327 the artificial neural network. *R Soc Open Sci* 2020;7(2):191420.
- 328 10. Dakappa PH, Prasad K, Rao SB, Bolumbu G, Bhat GK, Mahabala C. Classification of  
329 Infectious and Noninfectious Diseases Using Artificial Neural Networks from 24-Hour  
330 Continuous Tympanic Temperature Data of Patients with Undifferentiated Fever. *Crit Rev*  
331 *Biomed Eng* 2018;46(2):173-183.
- 332 11. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, *et al.* A Deep  
333 Learning Approach to Antibiotic Discovery. *Cell* 2020;180(4):688-702.e13.
- 334 12. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, *et al.* Association of Public Health  
335 Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. *JAMA*  
336 2020; 323(19):1 - 9.
- 337 13. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate  
338 time-varying reproduction numbers during epidemics. *Am J Epidemiol* 2013;178(9):1505-  
339 12.
- 340 14. Sato RC. Disease management with ARIMA model in time series. *Einstein (Sao Paulo)*  
341 2013;11(1):128-31.
- 342 15. Li Y, Zhang T. Deep neural mapping support vector machines. *Neural Netw* 2017;93:185-  
343 194.
- 344 16. Leung K, Wu JT, Liu D, Leung GM. First-wave COVID-19 transmissibility and severity  
345 in China outside Hubei after control measures, and second-wave scenario planning: a  
346 modelling impact assessment. *The Lancet* 2020;395(10233):1382-1393.
- 347 17. Adhikari SP, Meng S, Wu Y, Mao Y, Ye R, Wang Q, *et al.* Epidemiology, causes, clinical  
348 manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19)

349 during the early outbreak period: a scoping review. *Infectious diseases of poverty*  
350 2020;9(1):29-12.

351 18. Deng S, Peng H. Characteristics of and Public Health Responses to the Coronavirus  
352 Disease 2019 Outbreak in China. *Journal of clinical medicine* 2020;9(2):575.

353 19. Ye Q, Wang B, Mao J, Fu J, Shang S, Shu Q, *et al.* Epidemiological analysis of COVID-  
354 19 and practical experience from China. *J Med Virol* 2020; doi:10.1002/jmv.25813.

355 20. Wang KW, Deng C, Li JP, Zhang YY, Li XY, Wu MC. Hybrid methodology for  
356 tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural  
357 network. *Epidemiol Infect* 2017;145(6):1118-1129.

358 21. Yan L, Wang H, Zhang X, Li MY, He J. Impact of meteorological factors on the incidence  
359 of bacillary dysentery in Beijing, China: A time series analysis (1970-2012). *PLoS One*  
360 2017;12(8):e0182937.

361 22. Wu JY, Zhou YB, Chen Y, Liang S, Li LH, Zheng SB, *et al.* Three Gorges Dam: Impact  
362 of Water Level Changes on the Density of Schistosoma-Transmitting Snail *Oncomelania*  
363 *hupensis* in Dongting Lake Area, China. *PLoS Negl Trop Dis* 2015;9(6):e0003882.

364 23. Ouedraogo B, Inoue Y, Kambire A, Sallah K, Dieng S, Tine R, *et al.* Spatio-temporal  
365 dynamic of malaria in Ouagadougou, Burkina Faso, 2011-2015. *Malar J* 2018;17(1):138.

366 24. Liu W, Bao C, Zhou Y, Ji H, Wu Y, Shi Y, *et al.* Forecasting incidence of hand, foot and  
367 mouth disease using BP neural networks in Jiangsu province, China. *BMC Infect Dis*  
368 2019;19(1):828.

369 25. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the  
370 ARIMA model on the COVID-2019 epidemic dataset. *Data Brief* 2020;29:105340.

371 26. Yang W, Deng M, Li C, Huang J. Spatio-Temporal Patterns of the 2019-nCoV Epidemic  
372 at the County Level in Hubei Province, China. *Int J Environ Res Public Health* 2020;17(7).

373 27. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus

374 Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From  
375 the Chinese Center for Disease Control and Prevention. JAMA 2020;323(13):1239-1242.  
376 28. E GA, C BS, S BR, de Groot RJ, Christian D, A GA, *et al.* The species Severe acute  
377 respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-  
378 CoV-2. Nature microbiology 2020;5(4):536-544.

379

380

### 381 **Figure titles and legends**

382

383 **Figure 1** Daily number of laboratory-confirmed COVID-19 cases in Hubei Province, China  
384 from 16 January 2020 to 13 March 2020.

385

386 **Figure 2** Geographical distribution of 56,062 laboratory-confirmed COVID-2019 cases in  
387 Hubei province, China, with dates of illness onset from 16 January 2020 to 16 March 2020.  
388 Provinces are shaded according to confirmed case rate per 10,000 people during different  
389 periods. According to the mean of COVID-19 cases per day and public health interventions  
390 implemented, the outbreak was divided into five periods as follows: (A) mean of 63.5 in 16  
391 and January 23; (B) city lockdown with traffic suspension and home quarantine and the mean  
392 of 672.83 between January 14 and January 29; (C) mean of 2,154.00 between January 29 and  
393 February 5; (D) Fang Cang Shelter hospitals and the mean of 2,317.46 between February 6 and  
394 February 18; (E) number of COVID-19 cases per day under 1,000 and the mean of 233.78  
395 between February 19 and March 16.

396

397 **Figure 3** Mean reproduction number ( $R[t]$ ) with 95% CI based on confirmed COVID-19 cases  
398 in Hubei Province, China. The effective reproduction number  $R(t)$  is defined as the mean  
399 number of secondary cases generated by a typical primary case at time  $t$  in a population for the

400 whole period over a five-day moving average. The blue horizontal line indicates  $R(t) = 1$ , below  
401 which sustained transmission is unlikely when anti-transmission measures are sustained,  
402 indicating that the outbreak is under control. A black dotted line indicates the 95% credible  
403 intervals of the mean reproduction number.

404

405 **Figure 4** Normal QQ plot of standardized residuals for the ARIMA (0, 2, 1) model.

406

407 **Figure 5** Daily number of COVID-19 cases in Hubei Province, China from March 14 to March  
408 16, 2020.

409