Predicting the epidemic trend of COVID-19 in China and across the world using the machine learning approach

Mengyuan Li [1, 2], Zhilan Zhang [1, 2], Shanmei Jiang [1, 2], Qian Liu [1, 2], Canping Chen [1, 2], Yue Zhang [3, 4, 5], Xiaosheng Wang [1, 2, *]

[1] Biomedical Informatics Research Lab, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 211198, China

[2] Big Data Research Institute, China Pharmaceutical University, Nanjing 211198, China

[3] Pinghu hospital of Shenzhen university, Shenzhen, China

[4] Futian Hospital for Rheumatic Diseases, Shenzhen, China

[5] Department of Rheumatology and Immunology, The First Clinical college of Harbin Medical University, Harbin, China

[*] Correspondence to: Xiaosheng Wang, E-mail: xiaosheng.wang@cpu.edu.cn

Abstract

Background: Although the COVID-19 has been well controlled in China, it is rapidly spreading outside China that may lead to catastrophic results globally without implementation of necessary mitigation measures. Because the COVID-19 outbreak has made comprehensive and profound impacts on the world, an accurate prediction of its epidemic trend is significant. Although many studies have predicted the COVID-19 epidemic trend, most of these studies have used the early stage data and focused on the China cases.

Methods: We predicted the COVID-19 epidemic trend in China and across the world using the machine learning approach. We first built the models for predicting the daily numbers of cumulative confirmed cases (CCCs), new cases (NCs), and death cases (DCs) of COVID-19 in China based on the data from Jan 20, 2020, to Mar 1, 2020. Furthermore, we built the models, derived from the models for the China cases, for predicting the epidemic trend across the world (outside China).

Findings: The COVID-19 outbreak will peak on Feb 22, 2020 in China and April 10, 2020 across the world. It will be basically under control early April, 2020 in China and mid-June, 2020 across the world. The total number of COVID-19 cases will reach around 89, 000 in China and 403,000 across the world during the epidemic. Around 4,000 and 18,300 people will died of COVID-19 in China and across the world, respectively. The COVID-19 mortality rate is estimated to be around 4% all over the world.

Interpretation: The COVID-19 outbreak is controllable in the foreseeable future if comprehensive and stringent control measures are taken.

Keywords
COVID-19 outbreak; COVID-19 epidemic trend; predictive model; machine learning;

Background

Although the spread of the corona virus disease 2019 (COVID-19) caused by the 2019 novel coronavirus (SARS-CoV-2) is slowing down in China, it is rapidly growing outside China and how it will evolve remains unclear. So far, SARS-CoV-2 has infected nearly 200,000 people and led to nearly 8,000 deaths worldwide. The COVID-19 outbreak has been declared as a pandemic and is expected to cause one of the most serious global public health problems in recent years [1]. Of note, the impacts of the COVID-19 epidemic are not only limited to global public health, but also to global economy, geopolitics, culture, and society. Thus, an accurate prediction of the epidemic trend of COVID-19 may provide valuable advice on how to effectively prevent and control the spread of COVID-19 to relieve the major social and economic impacts of this disease. Although a number of studies have estimated the epidemic trend of the COVID-19 outbreak, most of these studies have used the early stage data and focused on the cases in China [2-8]. For example, using the SEIR (Susceptible, Exposed, Infectious, and Removed) model, Wang et al. estimated the numbers of COVID-19 cases in Wuhan, China under the conditions of insufficient and sufficient control measures being taken, respectively [3]. Wu et al. predicted the domestic and global spread of SARS-CoV-2 based on travel volume data during Dec, 2019, and Jan, 2020 [4]. Chen et al. evaluated the transmissibility of SARS-CoV-2 using the reservoir-people transmission network model [7]. Yang et al. predicted the epidemic peaks and scales of COVID-19 using the SEIR model and machine learning approach to show the indispensability of the nationwide interventions starting from Jan 23, 2020 [8]. With

COVID-19 remaining active in China and continuing to spread around the world, predicting the COVID-19 epidemic trend in China and across the world based on the updated data is definitely needed.

In this study, based on the public data of COVID-19 cases, we predicted the epidemic trend of COVID-19 in China and across the world. We first built the prediction model using the publicly available data for COVID cases. These data included the daily numbers of cumulative confirmed cases (CCCs), new cases (NCs), and death cases (DCs) of COVID-19 in China since January 20, 2020. We trained our model using the data from Jan 20, 2020, to Mar 1, 2020, and predicted the daily numbers of CCCs, NCs, and DCs after Mar 1, 2020. Furthermore, we predicted the epidemic trend of COVID-19 across the world (outside China) using the models derived from the models built based on the China cases.

Methods

Data preparation

We downloaded the statistics of confirmed COVID-19 cases in China from the National Health Committee of China (http://www.nhc.gov.cn/), including the daily numbers of CCCs, NCs, and DCs since January 20, 2020. The data for the statistics of confirmed COVID-19 cases outside China were downloaded from Worldometer (https://www.worldometers.info/coronavirus/#countries).

Predictive models

We built the models for predicting the daily numbers of CCCs, NCs, and DCs of

COVID-19 in China based on the data from Jan 20, 2020, to Mar 1, 2020 using Eureqa (Trial Version 1.24.0, https://www.nutonian.com/products/eureqa-desktop). Eureqa is a machine learning algorithm that can automatically build predictive models from data [9]. We input the daily numbers of CCCs or DCs and their corresponding days to Eureqa to obtain a formula which perfectly shapes the relationship between both variables. Due to insufficient sample size of the datasets for COVID-19 cases outside China because of their recent outbreak, we did not directly use Eureqa to build the predictive models for the COVID-19 epidemic trend outside China. Instead, we derived the predictive models from the models for the China cases on the assumption that the COVID-19 epidemic trend outside China is similar to that in China with a time lag.

Results

Prediction of the numbers of CCCs, NCs, and DCs of COVID-19 in China

We trained the model for predicting the daily number of CCCs of COVID-19 in China using the data from Jan 20, 2020, to Mar 1, 2020 with Eureqa. We obtained the predictive model: $Y = 42641.14 + 30962.75 \times atan(0.16 \times X - 3.43)$, where Y is the number of CCCs and X the time (days). The goodness-of-fit (R^2) of the model equals to 0.997 in the training set. The model predicted that the number of CCCs will peak on Feb 22, 2020 with 76, 983 CCCs and that the epidemic curve will flatten starting from Mar 25, 2020 in China (Fig. 1A). This is consistent with several recent reports [8,10-12]. We predicted that the COVID-19 epidemic will end around mid-May in China. The number of COVID-19 cases is expected to total around 89, 000 when the epidemic ends in China.

On the basis of the predicted daily number of CCCs, we estimated the daily number of NCs on a given day as the number of CCCs that day minus that of CCCs the previous day (Fig. 1B). The number of NCs is predicted to peak on Feb 10, 2020, with 4,944 NCs. At the peak of the COVID-19 outbreak on Feb 22, 2020, the number of NCs is estimated to be 1,050. After Feb 22, 2020, the daily numbers of NCs are expected to be less than 1,000. Starting from Mar 25, 2020, the daily numbers of NCs will drop to less than 100, indicating that the COVID-19 outbreak has been basically under controlled in China.

In addition, using Eureqa, we built a model for predicting the daily DCs in China as follows:

$$Y = 1221.43 + 1033.39 \times atan2\left(X - 29.63, 11.11\right) + 1033.39 \times atan2\left(X - 6.76 - atan2\left(X - 6.17, cos\left(X\right)\right), 28.25\right)$$

, where Y is the number of NCs and X the time (days). The model R^2 equals to 0.998 in the training set. The model predicted around 4,000 deaths caused by COVID-19 during the epidemic in China (Fig. 1C). Furthermore, we estimated that the mortality rate for COVID-19 will be 4.2%, which is slightly higher than the mortality rate of 4% calculated based on the latest data on Mar 17, 2020. However, because the number of NCs grows small while the number of critical cases remains more than 3,000 on Mar 17, 2020, the final actual mortality rate for COVID-19 in China will approach to our estimate. Furthermore, we estimated the daily COVID-19 mortality rate in China based on the daily CCCs and DCs predicted. Because the median time from onset to critical condition is around 10 days[13], we calculated the mortality rate on a given day by

dividing the number of DCs that day by the number of CCCs 10 days before. The daily mortality rate increases over time and grows steady from early April, 2020 (Fig. 1D).

Prediction of the epidemic trend of COVID-19 across the world

The COVID-19 outbreak across the world (outside China) has lagged behind that in China. We tried to predict the epidemic trend of COVID-19 across the world (China excluded). We found that the daily numbers of CCCs within the 14 consecutive days starting from Feb 27, 2020 across the world followed the same distribution to those within the 14 consecutive days starting from Jan 26, 2020 in China (Kolmogorov-Smirnov test (K-S test), p = 1) (Fig. 2A). Thus, we supposed that the COVID-19 epidemic across the world had a similar trend to that in China with a 32-day lag. Accordingly, based on the model for predicting the daily number of CCCs of COVID-19 in China, we generated the model for predicting the daily number of CCCs of COVID-19 across the world:

$$Y = (42641.14 + 30962.75 \times a\tan(0.16 \times (X - 32) - 3.43)) \times \beta$$

, where Y is the number of CCCs, X is the time (days), and β (= 4.5) is the ratio of the population outside China to that in China. Based on this model, we predicted that the number of NCs will peak on Mar 13, 2020, with 22,341 NCs across the world (Fig. 2B). On April 10, 2020, the COVID-19 outbreak will peak across the world with the number of CCCs and NCs being 382,403 and 1,058, respectively (Fig. 2B). After April 10, 2020, the daily numbers of NCs will be less than 1,000. Starting from June 15, 2020, the daily number of NCs will drop to 100, indicating that the COVID-19 outbreak is

basically under control across the world. The total number of COVID-19 cases is estimated to reach 403,216 during the epidemic across the world.

In addition, we found that the daily numbers of DCs within the 7 consecutive days starting from Feb 29, 2020 outside China followed the same distribution to those within the 7 consecutive days starting from Jan 27, 2020 in China (K-S test, $p = 1$) (Fig. 2C). Accordingly, based on the model for predicting the daily number of DCs in China, we built the model for predicting the daily number of DCs across the world:

$$Y = \left(1033.39 \times atan2\left(\left(X - 33\right) - 6.76 - atan2\left(\left(X - 33\right) - 6.17,\ cos\left(\left(X - 33\right)\right)\right),\ 28.25\right) + 1033.39 \times atan2\left(\left(X - 33\right) - 29.63,\ 11.11\right) + 1221.43\right) \times \beta$$

, where Y is the number of DCs, X is the time (days), and β (= 4.5) is the ratio of the population outside China to that in China. Using this model, we predicted that a total of 18,381 people will died of the COVID-19 disease when the COVID-19 outbreak is basically under control across the world (Fig. 2D). The mortality rate for COVID-19, estimated by our predictive models, will be 4.56% across the world during the COVID-19 outbreak. This number is close to that (4.2%) for China.

Discussion

The COVID-19 outbreak has made comprehensive and profound impacts on the world. Although this disease appears to be well controlled in China, the recent dramatic increase in new cases and deaths outside China indicates that the COVID-19 outbreak may have catastrophic results globally without implementation of necessary mitigation measures. However, the experience from China suggests that the COVID-19 outbreak

is controllable if effective strategies are employed [8]. Because the COVID-19 outbreak outside China is in the initial or exponential expansion phase depending on different regions or countries, currently it is difficult to predict the major turning points in the COVID-19 epidemic outside China based on their data. Therefore, we assumed that the COVID-19 outbreak outside China follows a similar pattern to that in China, and accordingly derived predictive models for the COVID-19 epidemic outside China from the models built based on the China cases. Our models predicted that the COVID-19 outbreak will peak across the world around April 10, 2020, and is basically under control in mid-June (Table 1). It should be noted that these predictions are made under the premise that the countries outside China implement comprehensive and stringent control measures like China, such as city lockdown, traffic control, and concentrated medical support for seriously-infected areas. Otherwise, the epidemic outside China could follow a different trend from China, e.g., prolonged outbreak that results in more unnecessary deaths and cases.

Our model predicted that the number of NCs will peak on Mar 13, 2020 outside China. This is consistent with a recent prediction of the epidemic trend of COVID-19 in Italy [14]. We estimated the mortality rate for COVID-19 to be 4.56% outside China. This number may varies depending on different regions or countries. For example, the current mortality rate is 7.9% in Italian versus 0.28% in German. It should be noted that the estimated mortality rate could exceed the actual mortality rate considering that a number of asymptomatic or mildly-symptomatic cases might not be identified. Actually,

a study of 1, 099 China cases indicated that the COVID-19 mortality rate in China is 1.4% [15]. If it is true, the total number of COVID-19 cases will reach around 260,000, a number far exceeding that reported presently.

A limitation of this study is that we did not take into account other factors which are associated with the spread and outbreak of COVID-19 in addition to demographics in deriving the predictive models for the global cases. These factors include politics, economy, culture, education, health facilities, geographical position, race etc. Overall, the African countries have a low number of COVID-19 cases; the number of COVID-19 cases is also small in the south Asian countries, including India, Pakistan, Bangladesh, and Sri Lanka, despite of their high population density. In contrast, the east Asian countries, including China, South Korea, and Japan, have reported a large number of cases; Europe has replaced China as the center of COVID-19 outbreak. The reason why there are notably different COVID-19 epidemic size between different regions or countries is worth further investigation.

Conclusions

The COVID-19 outbreak is controllable in the foreseeable future if comprehensive and stringent control measures are taken. Our prediction for the world cases is based on the assumption that other countries take effective control measures similar to China, and therefore should be cautiously optimistic.

List of abbreviations

CCCs: cumulative confirmed cases, **NCs**: new cases; **DCs**: death cases; **COVID-19**: the corona virus disease 2019; **SARS-CoV-2**: the 2019 novel coronavirus; **K-S test**: Kolmogorov-Smirnov test.

Conflicts of Interest

The authors declare that they have no competing interests.

Funding Statement

This work was supported by the China Pharmaceutical University (grant numbers 3150120001 to XW).

Acknowledgments

Not applicable.

References

ntial

1. Gates B. Responding to Covid-19 - A Once-in-a-Century Pandemic? *The New England journal of medicine* 2020.
2. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *The New England journal of medicine* 2020.
3. Wang H, Wang Z, Dong Y, et al. Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China. *Cell discovery* 2020; **6**: 10.
4. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet (London, England)* 2020; **395**(10225): 689-97.
5. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* 2020.
6. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 2020.
7. Chen TM, Rui J, Wang QP, Zhao ZY, Cui JA, Yin L. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious diseases of poverty* 2020; **9**(1): 24.

8. Yang Z, Zeng Z, Wang K. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J Thorac Dis* 2020.
9. Schmidt M, Lipson H. Distilling Free-Form Natural Laws from Experimental Data. **324**(5923): 81-5.
10. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling. *arXiv preprint arXiv:200206563* 2020.
11. Gao J, Tian Z, Yang X. Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *BioScience Trends* 2020.
12. Chang Y-C, Tung Y-A, Lee K-H, et al. Potential therapeutic agents for COVID-19 based on the analysis of protease and RNA polymerase docking. 2020.
13. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. *The Lancet Infectious Diseases* 2020.
14. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? *The Lancet* 2020.
15. Guan W-j, Ni Z-y, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine* 2020.

Figures

Fig. 1. Prediction of the COVID-19 epidemic trend in China. The predicted daily numbers of cumulative confirmed cases (A), new cases (B), and death cases (C) of COVID-19 in China. D. The estimated daily COVID-19 mortality rate in China. The mortality rate on a given day = the number of death cases that day / the number of cumulative confirmed cases 10 days before. NCs: new cases.

Fig. 2. Prediction of the COVID-19 epidemic trend in world. A. The daily numbers of cumulative confirmed cases of COVID-19 within the 14 consecutive days starting from Feb 27, 2020 outside China follow the same distribution to those within the 14 consecutive days starting from Jan 26, 2020 in China (Kolmogorov-Smirnov test (K-S test), $p = 1$). B. The predicted daily numbers of cumulative confirmed cases and new cases of COVID-19 in world. C. The daily numbers of death cases of COVID-19 within the 7 consecutive days starting from Feb 29, 2020 outside China followed the same

distribution to those within the 7 consecutive days starting from Jan 27, 2020 in China (K-S test, $p = 1$). D. The predicted daily numbers of death cases of COVID-19 outside China. NCs: new cases.

Tables

Table 1. Comparison of the COVID-19 epidemic between China and world

Comparisons	China	World
Number of cumulative confirmed cases	89,276	403,216
Number of deaths	4,092	18,381
Epidemic duration (months）	5	9
Time taken for the epidemic to peak (months)	2,5	3
Time taken for the peak to end (months)	2.5	6

Figure 1

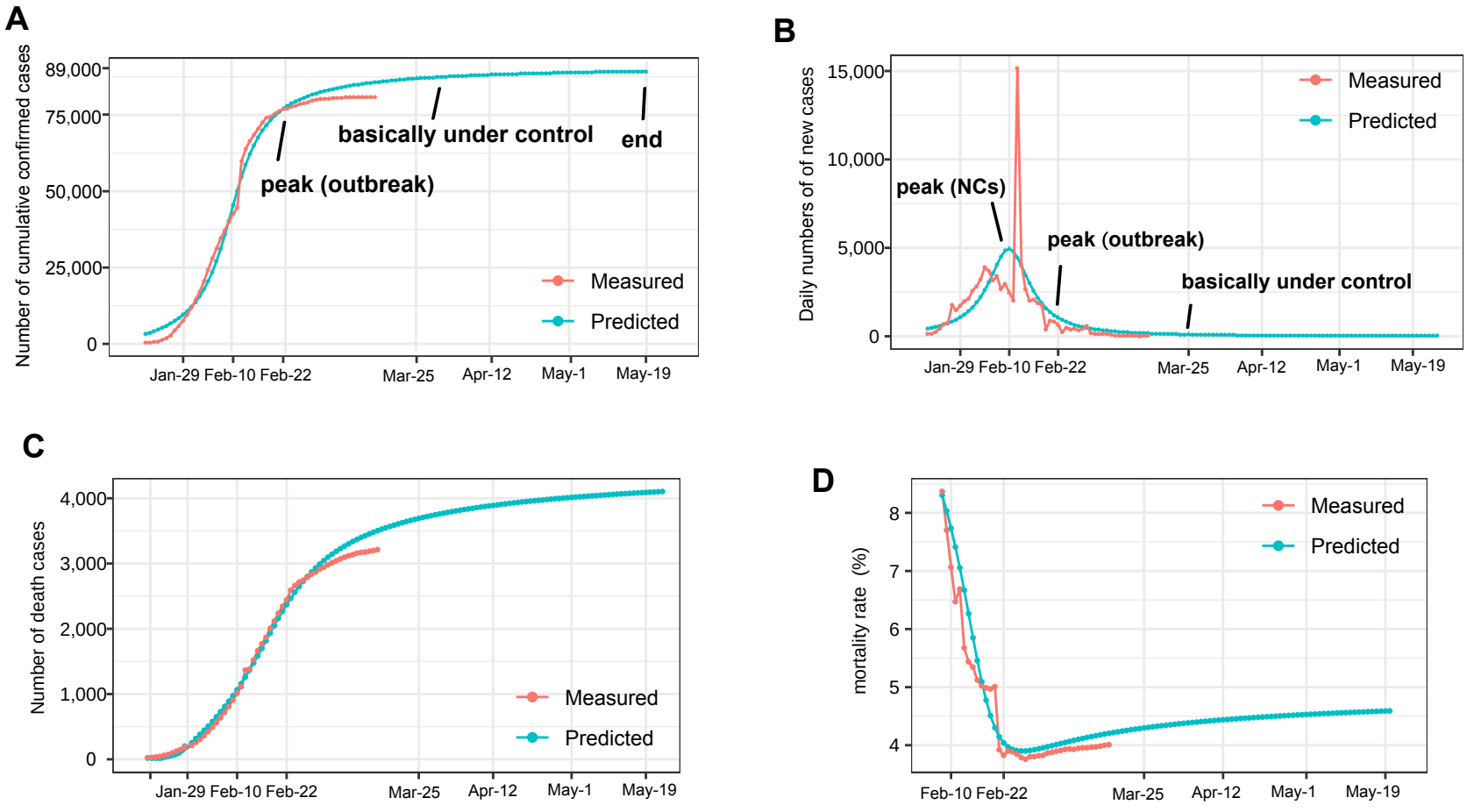
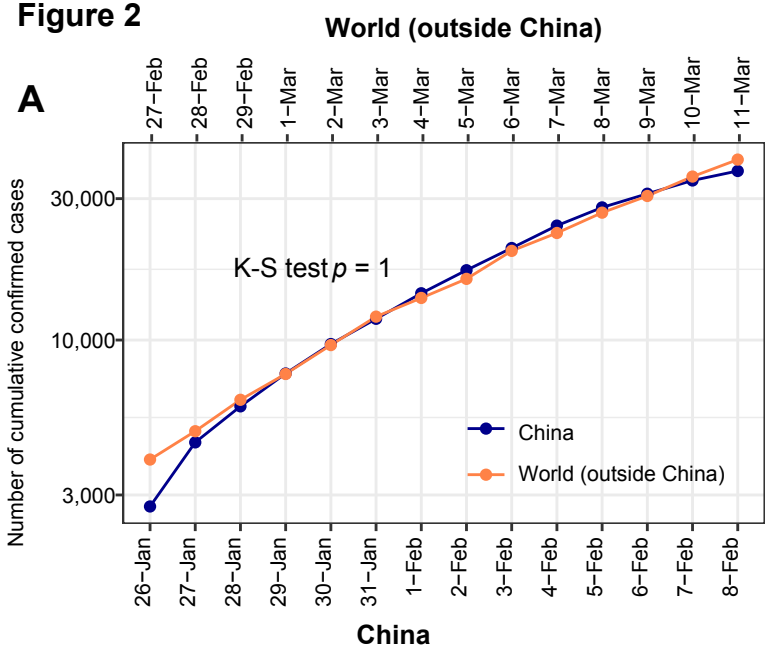
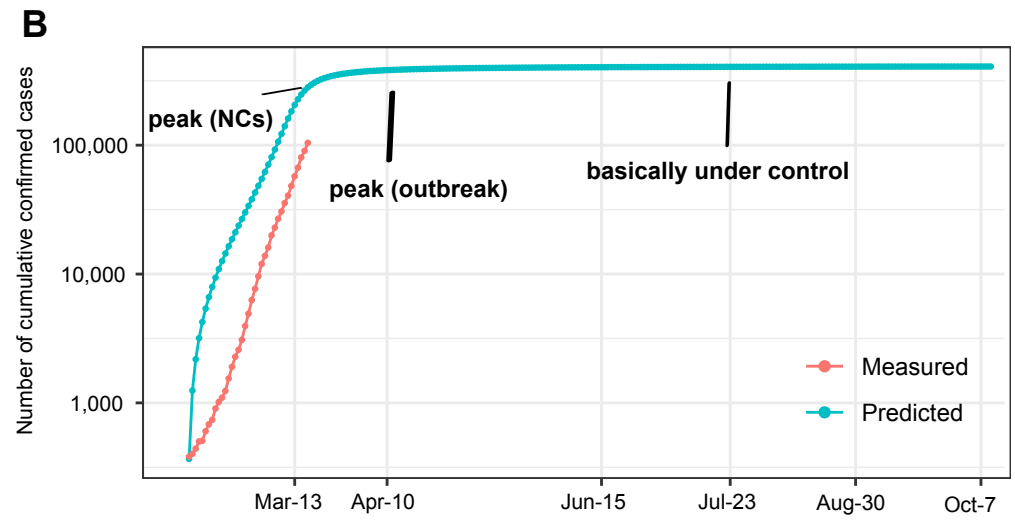
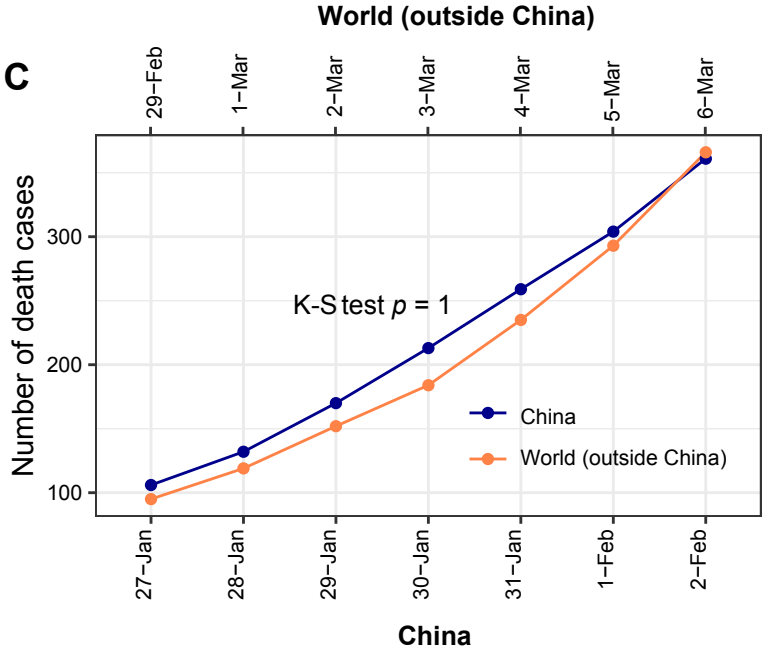
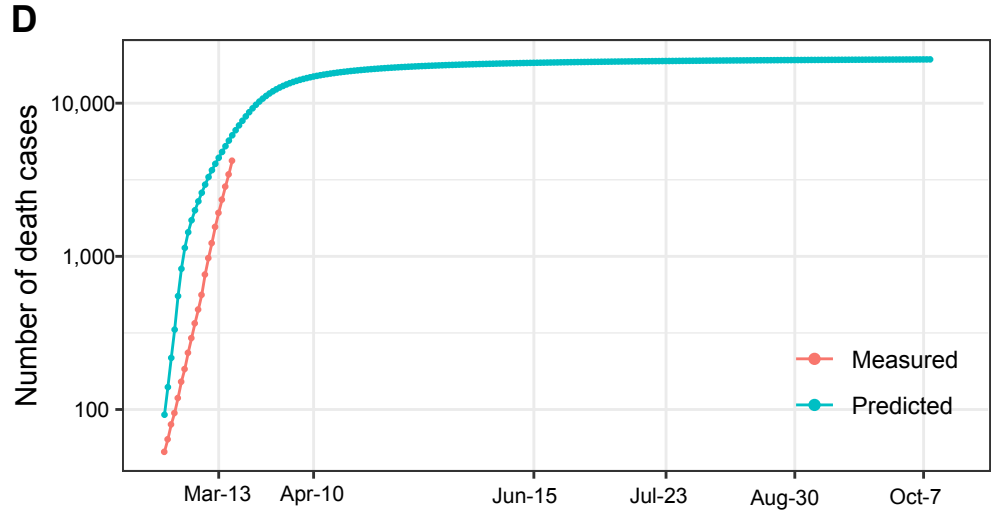

A
Number of cumulative confirmed cases
89,000
75,000
50,000
25,000
0
peak (outbreak)
basically under control
end
Measured
Predicted
Jan-29 Feb-10 Feb-22
Mar-25
Apr-12
May-1
May-19
B
Daily numbers of of new cases
15,000
10,000
5,000
0
peak (NCs)
peak (outbreak)
basically under control
Measured
Predicted
Jan-29 Feb-10 Feb-22
Mar-25
Apr-12
May-1
May-19
C
Number of death cases
4,000
3,000
2,000
1,000
0
Measured
Predicted
Jan-29 Feb-10 Feb-22
Mar-25
Apr-12
May-1
May-19
D
mortality rate (%)
8
7
6
5
4
Measured
Predicted
Feb-10 Feb-22
Mar-25
Apr-12
May-1
May-19

Figure 2


A
World (outside China)
27−Feb
28−Feb
29−Feb
1−Mar
2−Mar
3−Mar
4−Mar
5−Mar
6−Mar
7−Mar
8−Mar
9−Mar
10−Mar
11−Mar
Number of cumulative confirmed cases
30,000
10,000
3,000
K-S test p = 1
China
World (outside China)
26−Jan
27−Jan
28−Jan
29−Jan
30−Jan
31−Jan
1−Feb
2−Feb
3−Feb
4−Feb
5−Feb
6−Feb
7−Feb
8−Feb
China


B
Number of cumulative confirmed cases
100,000
10,000
1,000
peak (NCs)
peak (outbreak)
basically under control
Measured
Predicted
Mar-13
Apr-10
Jun-15
Jul-23
Aug-30
Oct-7


C
World (outside China)
29−Feb
1−Mar
2−Mar
3−Mar
4−Mar
5−Mar
6−Mar
Number of death cases
300
200
100
K-S test p = 1
China
World (outside China)
27−Jan
28−Jan
29−Jan
30−Jan
31−Jan
1−Feb
2−Feb
China


D
Number of death cases
10,000
1,000
100
Measured
Predicted
Mar-13
Apr-10
Jun-15
Jul-23
Aug-30
Oct-7