Modeling the COVID-19 Outbreak in China through Multi-source Information Fusion

Lin Wu,[1,*] Lizhe Wang,[2] Nan Li,[3] Tao Sun,[1,4] Tangwen Qian,[1,4] Yu Jiang,[1,4] Fei Wang,[1,*] and Yongjun Xu[1,*]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[2]China University of Geosciences (Wuhan), Wuhan, China

[3]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China

[4]University of Chinese Academy of Sciences, Beijing, China

*Correspondence: wulin@ict.ac.cn (L.W.); wangfei@ict.ac.cn (F.W.); xyj@ict.ac.cn (Y.X.)

https://doi.org/10.1016/j.xinn.2020.100033

Modeling the outbreak of a novel epidemic, such as coronavirus disease 2019 (COVID-19), is crucial for estimating its dynamics, predicting future spread and evaluating the effects of different interventions. However, there are three issues that make this modeling a challenging task: uncertainty in data, roughness in models, and complexity in programming. We addressed these issues by presenting an interactive individual-based simulator, which is capable of modeling an epidemic through multi-source information fusion.

The first challenge when modeling a novel epidemic is that all the reported and inferred data about its dynamics are inevitably affected by some level of uncertainty, resulting in wide ranges and systematic bias. The number of infected cases on any given date is a hidden state in the stochastic process and cannot be observed directly because it is never clear how much time elapses between when infection occurs and when that infection is identified and reported. This gap includes the incubation period and any delay in medical visit, diagnosis, or reporting. To make matters worse, modeling may be influenced by authors' prejudice, interest relationships, or preconceived ideas. Therefore, we argue that scientific research should combine multiple sources of data worldwide rather than be based on a single source, and should treat estimates from global researchers as elastic constraints imposed on models.

The second challenge is roughness in popular models, which is caused by oversimplification. It introduces significant errors and makes it impossible to reduce uncertainty by combining different types of information. The most common epidemic dynamics models are compartment models, such as SEIR (susceptible, exposed, infectious, and removed) and SIR (susceptible, infectious, and removed), which have been adopted widely in the simulation of COVID-19. The state vector of each person in a compartment is simplified as homogeneous and Markovian (memoryless), and transitions among compartments are modeled by differential equations with fixed parameters, such as incubation rate, transmission rate, and recovery rate. However, oversimplified models are not capable of incorporating multi-type uncertain information, such as clinical courses, viral shedding, subclinical transmission, infections, confirmations, deaths, or interventions, so they cannot reduce uncertainty by multi-source information fusion.

The last but far from the smallest challenge is the complexity of programming. This is a problem for researchers and reviewers who do not have a background in computer science. When it comes to individual-based models, implementation is impossible without rich experience in object-oriented programming, which compartmentalizes data into objects and describes object contents and behavior through the declaration of classes. Therefore, scientists have begun to call for sharing model codes so that the results of papers can be replicated and evaluated. However, sharing model codes is not enough given that there are many programming languages and it is no small task to install and configure corresponding development environments and run publicly shared codes.

To tackle the three challenges of modeling epidemic dynamics, we have developed an interactive simulator for individual-based models in this paper. This is described in detail in the Supplemental Information. In contrast to compartment models, individual-based models represent each individual via an independent set of specific characteristics that may change over time. This feature allows a more realistic and informative analysis of an epidemic. It is capable of interactively modeling parameter ranges as probability distributions, heterogeneity as independent objects, and randomness of transmission as stochastic processes through a terminal or webpage without coding. The output of this model consists of daily values of infected cases, confirmed cases, recovered cases, deaths, and effective reproduction numbers. We can fit input parameters and output results with reported and inferred data from multiple sources.

Based on this simulator, we modeled the COVID-19 outbreak in China by fusion of multi-source multi-type uncertain data, incorporating daily confirmed and suspected cases as well as deaths from Centers for Disease Control in China, infection probability per contact from the World Health Organization, date of first diagnosed infection,[1] estimated range of symptom onset case count by January 18, 2020,[2] basic reproduction number before Wuhan lockdown,[3] estimated range of infected cases in Wuhan by January 25, 2020,[4] and February 1, 2020,[5] date of obligatory quarantine in dedicated facilities, incubation period[3], clinical courses,[1] time from illness onset to discharge or death,[6] range of fatality ratio,[7] daily contact patterns,[8] proportion of asymptomatic infections,[9] delay from onset to first medical visit and confirmation,[3] temporal dynamics in viral shedding and transmissibility,[10] and other information from dozens of studies, which are described in detail in the Supplemental Information.

The results of modeling through multi-source information fusion are refined estimates of the COVID-19 dynamics in China as depicted in Figure 1 and the Supplemental Information. According to the results of our simulation, the probability of infection per contact is about 3.65% when the transmissibility peaks around onset of symptoms. When an infected person makes 10 to 20 contacts per day during this period, he or she will infect 0.365 to 0.73 individuals on average. Imposing all constrains described in the Supplemental Information, the basic reproduction number in China before interventions was 2.3 and total infections were approximately 35,000 on January 23, 2020, when the travel ban was issued for Wuhan.

Results for the time before February 2020 show significant deviations between reported and simulated numbers of cases. According to the summary published by the Wuhan New Pneumonia Prevention and Control Headquarters, there were three causes of errors in the number of reported cases. The first cause was delayed, missed, and false reports in the early stage of the outbreak resulting from the overwhelmed medical system. The second cause was a lack of testing and hospitalization capacity before February 20, 2020. The third cause was the imperfection of statistics systems during the beginning of outbreak. By June 23, 2020, there had been several corrections on reported cases by Centers for Disease Control in China. On February 12, 2020, 13,332 PCR-negative cases with symptoms of COVID-19 were added to the confirmed cases. On February

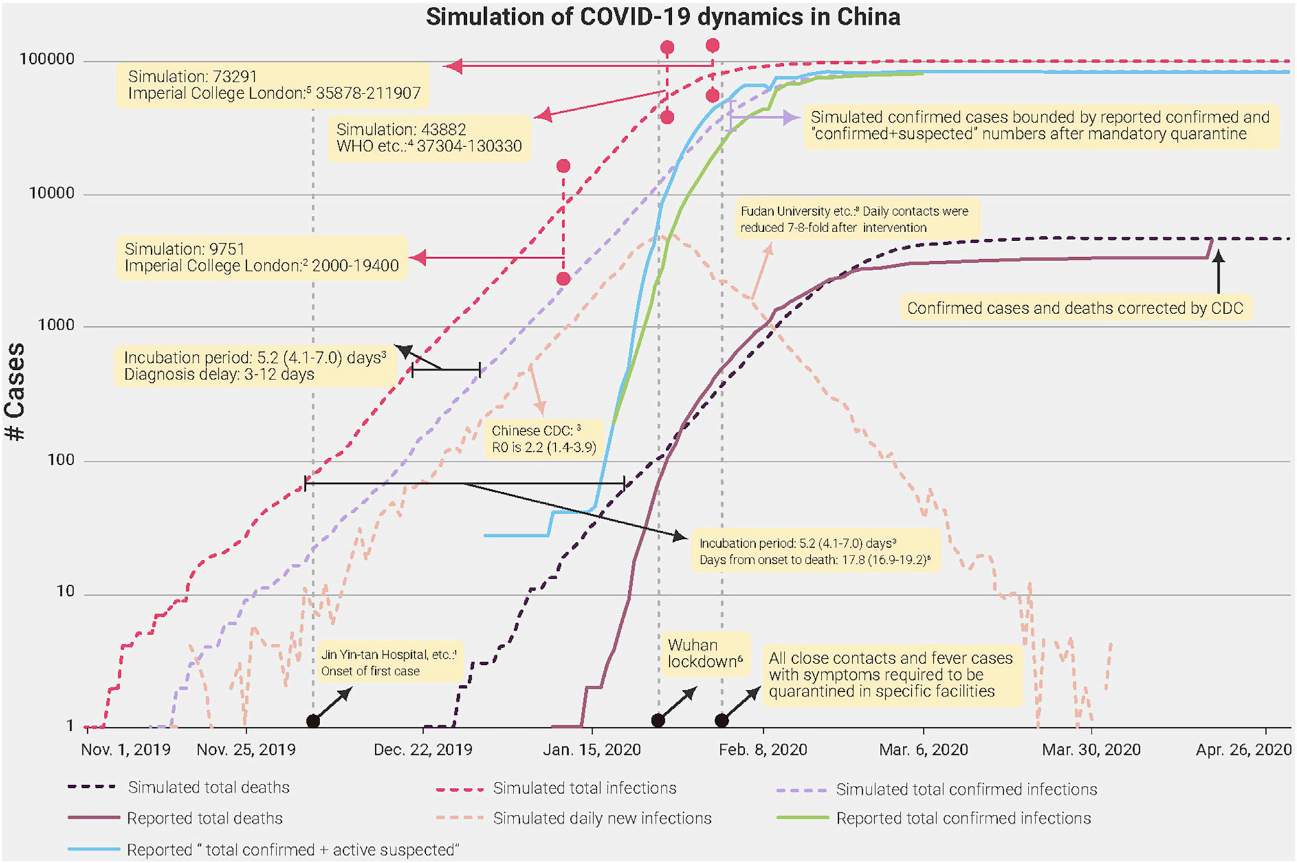

Figure 1. Simulation and Reported Cases of COVID-19 in China from November 1, 2019, to April 28, 2020, under Logarithmic Coordinates. The exponential growth of the number of daily new infections ended shortly after the lockdown of Wuhan.

13, 2020, repeated case entries were removed. On April 17, 2020, confirmed cases and deaths were corrected thoroughly, adding 325 confirmed cases and 1,290 deaths. By multi-source information fusion, we are able to make a refined estimate of the overall process of COVID-19's spread in China.

REFERENCES

1. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497–506.
2. Imai, N., Dorigatti, I., Cori, A., Donnelly, C., Riley, S., and Ferguson, N.M. (2020). Estimating the Potential Total Number of Novel Coronavirus Cases in Wuhan City, China (Imperial College London).
3. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S.M., Lau, E.H.Y., Wong, J.Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N. Engl. J. Med. *382*, 1199–1207.
4. Wu, J.T., Leung, K., and Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet *395*, 689–697.
5. Thompson, H., Imai, N., Dighe, A., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dorigatti, I., et al. (2020). Estimating Infection Prevalence in Wuhan City from Repatriation Flights (Imperial College London).
6. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect. Dis. *20*, 669–677.
7. Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., et al. (2020). Estimating the burden of SARS-CoV-2 in France. Science *369*, 208–211.
8. Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., Wu, Q., Merler, S., Viboud, C., Vespignani, A., et al. (2020). Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science *368*, 1481–1486.
9. Mizumoto, K., Kagaya, K., Zarebski, A., and Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance *25*, 2000180.
10. He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. *26*, 672–675.

ACKNOWLEDGMENTS

This work is partially supported by NSFC no. 61902376, NSFC no. 61602487, and NSFC no. 81701067.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.xinn.2020.100033.