High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2

Steven Sanche[1], Yen Ting Lin[1], Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, Ruian Ke

Severe acute respiratory syndrome coronavirus 2 is the causative agent of the ongoing coronavirus disease pandemic. Initial estimates of the early dynamics of the outbreak in Wuhan, China, suggested a doubling time of the number of infected persons of 6–7 days and a basic reproductive number (R_0) of 2.2–2.7. We collected extensive individual case reports across China and estimated key epidemiologic parameters, including the incubation period (4.2 days). We then designed 2 mathematical modeling approaches to infer the outbreak dynamics in Wuhan by using high-resolution domestic travel and infection data. Results show that the doubling time early in the epidemic in Wuhan was 2.3–3.3 days. Assuming a serial interval of 6–9 days, we calculated a median R_0 value of 5.7 (95% CI 3.8–8.9). We further show that active surveillance, contact tracing, quarantine, and early strong social distancing efforts are needed to stop transmission of the virus.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the etiologic agent of the current rapidly growing outbreak of coronavirus disease (COVID-19), originating from the city of Wuhan, Hubei Province, China (*1*). Initially, 41 cases of "pneumonia of unknown etiology" were reported to the World Health Organization by the Wuhan Municipal Health Committee at the end of December 2019 (*2*). On January 8, 2020, the pathogen was identified (*1*), and human-to-human transmission was reported soon after. By January 21, most provinces of China had reported COVID-19 cases. By March 16, the outbreak had led to >170,000 total confirmed cases and >6,500 deaths globally. In a period of 3 months, an outbreak of apparent idiopathic pneumonia had become the COVID-19 pandemic.

Studying dynamics of a newly emerged and rapidly growing infectious disease outbreak, such as COVID-19, is important but challenging because of the limited amount of data available. In addition, unavailability of diagnostic reagents early in the outbreak, changes in surveillance intensity and case definitions, and overwhelmed healthcare systems confound estimates of the growth of the outbreak based on data. Initial estimates of the exponential growth rate of the outbreak were 0.1–0.14/day (a doubling time of 6–7 days), and a basic reproductive number (R_0; defined as the average number of secondary cases attributable to infection by an index case after that case is introduced into a susceptible population) ranged from 2.2 to 2.7 (*1*,*3*–*5*). These estimates were based on 2 broad strategies. First, Li et al. used very early case count data in Wuhan before January 4 (*1*). However, case count data can be confounded by reservoir spillover events, stochasticities in the initial phase of the outbreak, and low surveillance intensity. The epidemic curve based on symptom onset after January 4 showed a very different growth rate (*6*). Second, inference was performed by using international flight data and infected persons reported outside of China (*3*–*5*). Because of the low numbers of persons traveling abroad compared with the total population size in Wuhan, this approach leads to substantial uncertainties (*7*,*8*). Inferences based on a low number of observations are prone to measurement error when data are incomplete or model assumptions are not fully justified; both conditions are common challenges associated with rapid and early outbreak analyses of a new pathogen.

We collected an expanded set of case reports across China on the basis of publicly available information, estimated key epidemiologic parameters, and provided a new estimate of the early epidemic growth rate and R_0 in Wuhan. Our approaches are based on integration of high-resolution domestic travel data and early infection data reported in provinces other than Hubei to infer outbreak dynamics in

Author affiliation: Los Alamos National Laboratory, Los Alamos, New Mexico, USA

DOI: https://doi.org/10.3201/eid2607.200282

[1]These first authors contributed equally to this article.

Wuhan. They are designed to be less sensitive to biases and confounding factors in the data and model assumptions. Without directly using case confirmation data in Wuhan, we avoid the potential biases in reporting and case confirmation in Wuhan, whereas because of the high level of domestic travel before the Lunar New Year in China, inference based on these data minimizes uncertainties and risk for potential misspecifications and biases in data and model assumptions.

Methods

Methodologic Overview

We developed 2 modeling approaches to infer the growth rate of the outbreak in Wuhan from data from provinces other than Hubei. In the first model, the first arrival model, we computed the likelihood of the arrival times of the first known cases in provinces outside of Hubei as a function of the exponential growing population of infected persons in Wuhan before late January. This calculation involved using domestic travel data to compute the probability that an infected person traveled from Wuhan to a given province as a function of the unknown actual number of infected persons in Wuhan and the probability that they traveled. The timings of the arrivals of the first infected persons in different provinces would reflect the rate of the epidemic growth in Wuhan.

In the second model, the case count model, we accounted for the detection of additional persons who were infected in Wuhan and received their diagnoses in other provinces and explicitly modeled those persons by using a hybrid deterministic–stochastic SEIR (susceptible-exposed-infectious-recovered) model. We then fitted this model to new daily case count data reported outside Hubei Province during the period before substantial transmission occurred outside of the province.

By using data collected outside Hubei Province, we minimized the effect of changes in surveillance intensity. By the time cases were confirmed in provinces outside Hubei, all of the provinces of China had access to diagnostic kits and were engaging in active surveillance of travelers out of Wuhan (e.g., using temperatures detectors and digital data to identify infected persons [9]) as the outbreak unfolded. Furthermore, the healthcare systems outside Hubei were not yet overwhelmed with cases and were actively searching for the first positive case, leading to much lower bias in the reporting in each province compared with the time series of confirmed cases in Wuhan.

Data

Individual Case Reports

We collected publicly available reports of 140 confirmed COVID-19 cases (mostly outside Hubei Province). These reports were published by the Chinese Centers for Disease Control and Prevention (China CDC) and provincial health commissions; accession dates were January 15–30, 2020 (Appendix 1 Table 1, https://wwwnc.cdc.gov/EID/article/26/7/20-0282-App1.xlsx). Many of the individual reports were also published on the China CDC official website (http://www.chinacdc.cn/jkzt/crb/zl/szkb_11803) and the English version of the China CDC weekly bulletin (http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm). These reports include demographic information as well as epidemiologic information, including potential periods of infection, and dates of symptom onset, hospitalization, and case confirmation. Most of the health commissions in provinces and special municipalities documented and published detailed information of the first or the first few patients with confirmed COVID-19. As a result, a unique feature of this dataset includes case reports of many of the first or the first few persons who were confirmed to have SARS-CoV-2 virus infection in each province, where dates of departure from Wuhan were available.

Travel Data

We used the Baidu Migration server (https://qianxi.baidu.com) to estimate the number of daily travelers in and out of Wuhan (Appendix 1 Table 2).
The server is an online platform summarizing mobile phone travel data hosted by Baidu Huiyan (https://huiyan.baidu.com).

Calculations of R_0 and Effect of Intervention Strategies

We considered realistic distributions for the latent and infectious periods to calculate R_0. We described the methods we used to calculate R_0 and the effect of intervention strategies on the outbreak (Appendix 2, https://wwwnc.cdc.gov/EID/article/26/7/20-0282-App2.pdf).

Results

Estimating Distributions of Epidemiologic Parameters

We first translated reports from documents or news reports published daily from the China CDC website and official websites of health commissions across provinces and special municipalities in China during January 15–30, 2020. Altogether, we collected

137 individual case reports from China and 3 additional case reports from outside of China (Appendix 1 Table 1).

By using this dataset, we estimated the basic parameter distributions of durations from initial exposure to symptom onset to hospitalization to discharge or death. Our estimate of the time from initial exposure to symptom onset (i.e., the incubation period) is 4.2 days (95% CI 3.5–5.1 days) (Figure 1, panel A), based on 24 case reports. This estimated duration is generally consistent with a recent report by Guan et al. (*10*) showing that the median incubation period is 4 days. Our estimate is ≈1 day shorter than 2 previous estimates (*1*,*11*). One potential caveat of our estimation is that because most of the case reports we collected were from the first few persons detected in each province, this estimation might be biased toward patients with more severe symptoms if they are more likely to be detected.

The time from symptom onset to hospitalization showed evidence of time dependence (Figure 1, panel B; Appendix 2 Figure 1). Before January 18, the time from symptom onset to hospitalization was 5.5 days (95% CI 4.6–6.6 days), whereas after January 18, the duration shortened significantly to 1.5 days (95% CI 1.2–1.9 days) ($p<0.001$ by Mann–Whitney U test). The change in the distribution coincides with news reports of potential human-to-human transmission and upgrading of emergency response level to Level 1 by the China CDC. The emerging consensus about the risk for COVID-19 probably led to substantial behavior changes among symptomatic persons, in terms of seeking more timely medical care during this period. However, because most of the individual reports were collected in provinces other than Hubei, the change in durations might only reflect changes in the rest of China (rather than in Hubei). We also found that the time from initial hospital admittance to discharge was 11.5 days (95% CI 8.0–17.3 days) (Figure 1, panel C) and from initial hospital admittance to death was 11.2 days (95% CI 8.7–14.9 days) (Figure 1, panel D). The time from symptom onset to death was estimated to be 16.1 days (95% CI 13.1–20.2 days).

Estimating the Growth Rate of the Outbreak in Wuhan in January 2020

Moving from empirical estimates of basic epidemiologic parameters to an understanding of the early growth rates of COVID-19 requires model-based inference and data. We first collected real-time travel data during the epidemic by using the Baidu Migration server, which provides real-time travel patterns in China based on mobile-phone positioning services (Figure 2, panel A; Appendix 1 Methods, Table 2). We estimated that, before the January 23 lockdown of the city, ≈40,000–140,000 people in Wuhan traveled to destinations outside Hubei Province each day (Figure 2, panel B). The extensive travel before the Lunar New Year was probably an important driver of the spread of COVID-19 in China.

We then integrated spatiotemporal domestic travel data to infer the outbreak dynamics in Wuhan by using two mathematical approaches (Appendix 2; conceptual framework depicted in Figure 3, panel A). The first-arrival model uses a unique feature of our case report dataset whereby the dates of departure from Wuhan for many of the first persons who were confirmed with SARS-CoV-2 infection in each province were known (Appendix 1 Table 1). We assumed an exponential growth for the total infected

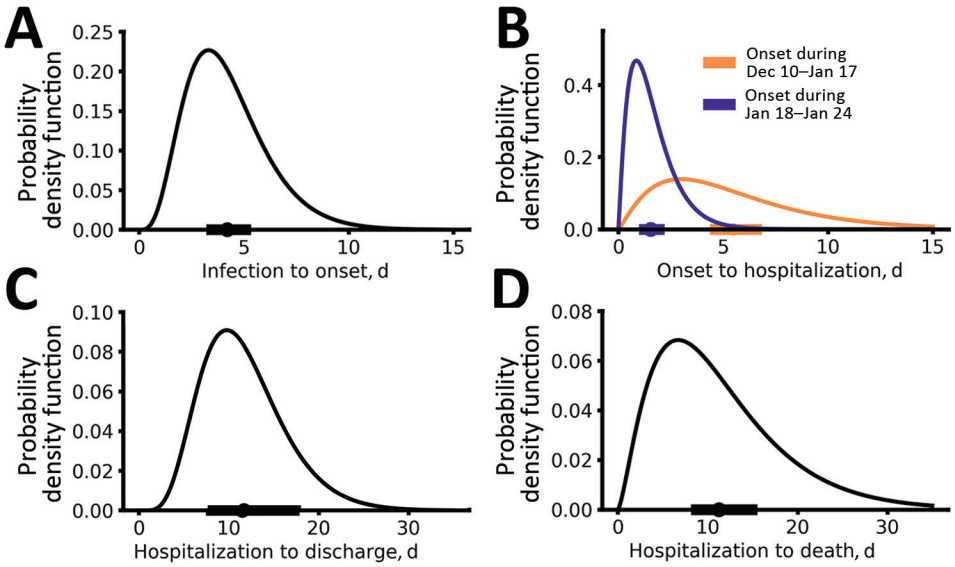


Figure 1. Epidemiologic characteristics of early dynamics of coronavirus disease outbreak in China. Distributions of key epidemiologic parameters: durations from infection to symptom onset (A), from symptom onset to hospitalization (B), from hospitalization to discharge (C), and from hospitalization to death (D). Filled circles and bars on *x*-axes denote the estimated means and 95% CIs.

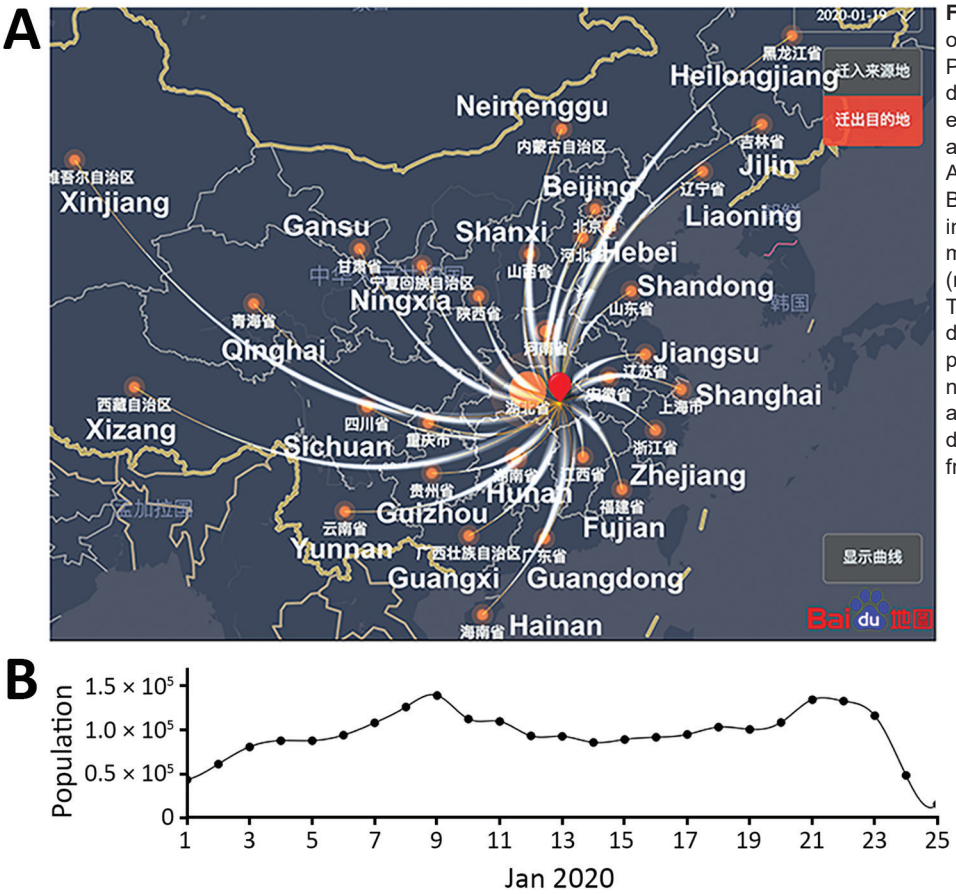


Figure 2. Extremely high level of travel from Wuhan, Hubei Province, to other provinces during January 2020, as estimated by using high-resolution and real-time travel data, China. A) A modified snapshot of the Baidu Migration online server interface showing the human migration pattern out of Wuhan (red dot) on January 19, 2020. Thickness of curved white lines denotes the size of the traveler population to each province. The names of most of the provinces are shown in white. B) Estimated daily population sizes of travelers from Wuhan to other provinces.

population I^* in Wuhan, $I^*(t) = e^{r(t-t_0)}$, where I^* includes infected persons who are asymptomatic or symptomatic, r is the exponential growth rate, and t_0 is the theoretical time of the exponential growth initiation, so that $I^*(t_0) = 1$ in the deterministic model. We call t_0 a "theoretical" time in the sense that it should not be interpreted as the time of first infection in a population. We should expect that t_0 is later than the date of the first infection because multiple spillover events from the animal reservoir might be needed to establish sustained transmission and stochasticity might play a large role in initial dynamics before the onset of exponential growth (*12*–*14*).

We used travel data for each of the provinces (Appendix 1 Table 3) and the earliest times that an infected person arrived in a province, across a total of 26 provinces (Figure 3, panel B), to infer r and t_0 (Appendix 2). Model predictions of arrival times in the 26 provinces fitted the actual data well (Appendix 2 Figure 2). The growth rate r is estimated to be 0.29/day (95% CI 0.21–0.37/day), corresponding to a doubling time of 2.4 days (95% CI 1.9–3.3 days). t_0 is estimated to be December 20, 2019 (95% CI December 11–26). As we show later, there exist larger uncertainties in the estimation of t_0.

We further estimated that the total infected population size in Wuhan was ≈4,100 (95% CI 2,423–6,178) on January 18 (Appendix 2 Figure 3), which is consistent with a recently posted estimate (*7*). The estimated number of infected persons was ≈18,700 (95% CI 7,147–38,663) on January 23 (i.e., the date when Wuhan started its lockdown). We projected that without any control measures, the infected population would be ≈233,400 (95% CI 38,757–778,278) by the end of January.

An alternative model, the case count approach, used daily new case counts of persons who had COVID-19 diagnosed in other provinces but who had been in Hubei Province within 14 days of becoming symptomatic. This model uses data beyond the first appearance of an infected person from Wuhan but also accounts for the stochastic nature of the process by using a hybrid model. In this model, the infected population in Wuhan was described with a deterministic model, whereas the infected persons who traveled from Wuhan to other provinces were tracked with a stochastic SEIR (susceptible-exposed-infectious-recovered) model (*12*). We restricted the data to the period of January

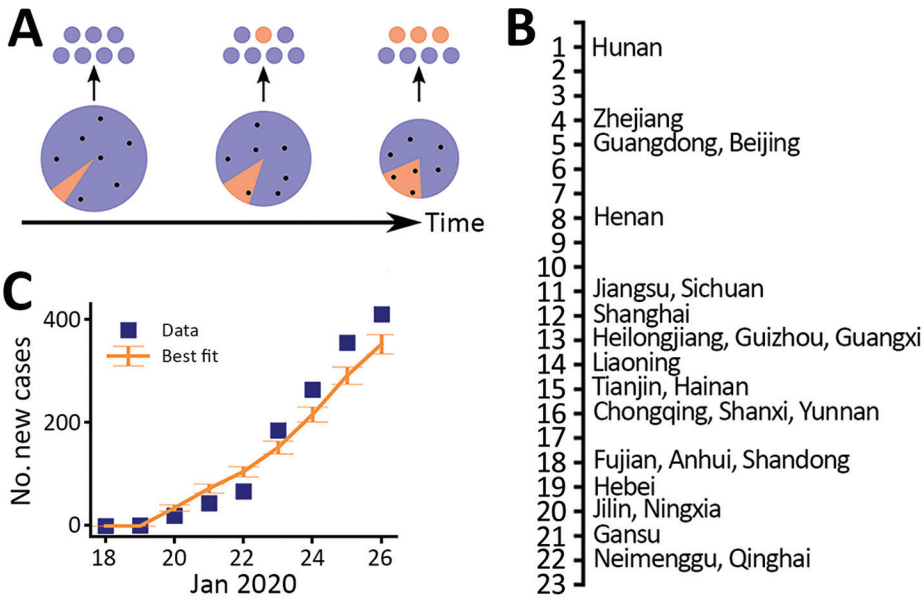


Figure 3. Estimates of the exponential growth rate and the date of exponential growth initiation of the coronavirus disease outbreak in China based on 2 different approaches. A) Schematic illustrating the export of infected persons from Wuhan. Travelers (dots) are assumed to be random samples from the total population (whole pie). Because of the growth of the infected population (orange pie) and the shrinking size of the total population in Wuhan over time, probability of infected persons traveling to other provinces increases (orange dots). B) The dates of documented first arrivals of infected persons in 26 provinces. C) Best fit of the case count model to daily counts of new cases (including only imported cases) in provinces other than Hubei. Error bars indicate SDs.

19–26, when new cases reported were mostly infections imported from Wuhan (i.e., indicative of the dynamics in Wuhan). The transitions of the infected persons from symptom onset to hospitalization and then to case confirmation were assumed to follow the distributions inferred from the case report data (Appendix 2). Simulation of the model using best-fit parameters showed that the model described the observed case counts over time well (Figure 3, panel C). The estimated theoretical time (t_0) is December 16, 2019 (95% CI December 12–21), and the exponential growth rate is 0.30/day (95% CI 0.26–0.34/day). These estimates are consistent with estimates in the first arrival approach (Figure 4; Appendix 2 Figure 4).

In both models, we assumed perfect detection (i.e., of infected cases outside of Hubei Province). However, a certain fraction of cases probably was not reported. To investigate the robustness of our estimates, we performed extensive sensitivity analyses to test 23 different scenarios of surveillance intensity (Appendix 2). First, we tested the assumption that a constant fraction of infected persons (e.g., persons with mild or no symptoms) (*15*) were not detected. We found that under this assumption, $t_0$would be earlier than our estimate but the estimation of the growth rate remained the same (Appendix 1 Table 4). Second, we tested the assumption that the intensity of surveillance increases over the period of data collection, although this scenario is less likely because of the intensive surveillance implemented outside Hubei Province. We found that our data in general do not support this hypothesis on the basis of corrected Akaike Information criterion scores (Appendix 1 Table 4). However, if the intensity of surveillance outside Hubei Province increased over the period of January, we would predict a lower growth rate than the estimate we just described. For the worst-case scenario considered, we estimated the growth rate of the outbreak to be 0.21/day (Appendix 2).

Other Evidence of a High Growth Rate of the Outbreak in Wuhan

In addition to using 2 modeling approaches, we looked for other evidence of a high outbreak growth rate to cross-validate our estimations. We found that the time series of reported deaths in Hubei, which is less subject to the biases of the confirmed case counts, is simply not consistent with a growth rate of 0.1/day (Appendix 2 Figure 5). As the infected population grows, the number of death cases will grow at the same rate but with a delayed onset corresponding to the time from infection to death. Fitting a simple exponential growth model to the number of reported deaths in Hubei during late January 2020 yields an estimate of 0.22–0.27/day, which is within the 95% CI of the estimation we previously described.

Overall, these analyses suggest that although there exist uncertainties depending on the level of surveillance, the exponential growth rate of the outbreak is probably 0.21–0.3/day. This estimation is much higher than previous reports, in which the growth rate was estimated to be 0.1–0.14/day (*1*,*3*–*5*).

Estimating R_0

The basic reproductive number, R_0, is dependent on the exponential growth rate of an outbreak, as well as additional factors such as the latent period (the time from infection to infectiousness) and the infectious period (*16*,*17*), both of which cannot be estimated directly from the data. Following the approach by Wearing and Rohani (*16*), we found that with a high growth rate of the outbreak, R_0 is in general high and the longer the latent and the infectious periods, the higher the estimated R_0 (Appendix 2 Figure 6).

To derive realistic values of R_0, we used previous estimates of serial intervals for COVID-19. The serial interval is estimated to be ≈7–8 days based on data collected early in the outbreak in Wuhan (*1*). More recent data collected in Shenzhen Province, China, suggests that the serial interval is dependent on the time to hospital isolation (Q. Bi et al., unpub. data, https://doi.org/10.1101/2020.03.03.20028423). When infected persons are isolated after 5 days of symptoms (a probable scenario for the early outbreak in Wuhan, where the public was not aware of the virus and few interventions were implemented), the serial interval is estimated to be 8 days (Q. Bi et al., unpub. data). Thus, these results suggest a serial interval of 7–8 days. With this serial interval, we sampled latent and infectious periods within wide biologically plausible ranges (Appendix 2) and estimated the median R_0 to be 5.8 (95% CI 4.4–7.7) (Figure 5, panel A). To include a wider range of serial interval (i.e., 6–9 days) (Figure 5, panel A; Appendix 2 Figure 6), given the uncertainties in these estimations, we estimated that the median of estimated R_0 is 5.7 (95% CI of 3.8–8.9) (Figure 5, panel B). The estimated R_0 can be lower if the serial interval is shorter. However, recent studies reported that persons can be infectious for a long period, such as 1–3 weeks after symptom onset (*18*; R. Woelfel et al., unpub data. https://doi.org/10.1101/2020.03.05.20030502); thus, we believe that a mean serial interval shorter than 6 days is unlikely during the early outbreak in Wuhan, where infected persons were not rapidly hospitalized.

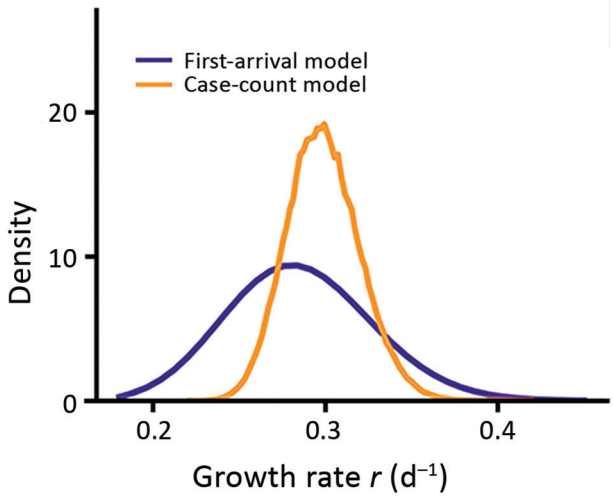


Figure 4. Marginalized likelihoods of growth rate (*r*) for 2 inference approaches to estimates the exponential growth rate of the coronavirus disease outbreak in China.

Implications for Intervention Strategies

The R_0 values we estimated have important implications for predicting the effects of pharmaceutical and nonpharmaceutical interventions. For example, the threshold for combined vaccine efficacy and herd immunity needed for disease extinction is calculated as $1 - 1/R_0$. At $R_0 = 2.2$, this threshold is only 55%. But at $R_0 = 5.7$, this threshold rises to 82% (i.e., >82% of the population has to be immune, through either vaccination or prior infection, to achieve herd immunity to stop transmission).

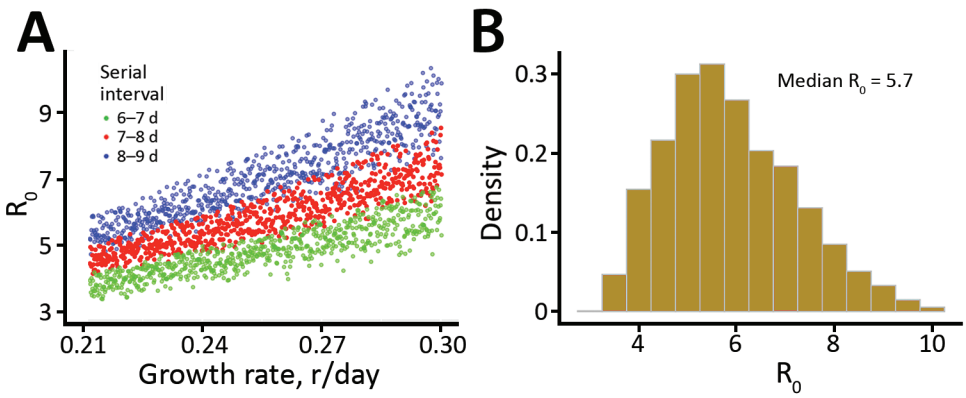


Figure 5. Estimation of the basic reproductive number (R_0), derived by integrating uncertainties in parameter values, during the coronavirus disease outbreak in China. A) Changes in R_0 based on different growth rates and serial intervals. Each dot represents a calculation with mean latent period (range 2.2–6 days) and mean infectious periods (range 4–14 days). Only those estimates falling within the range of serial intervals of interests were plotted. B) Histogram summarizing the estimated R_0 of all dots in panel A (i.e., serial interval ranges of 6–9 days). The median R_0 is 5.7 (95% CI 3.8–8.9).

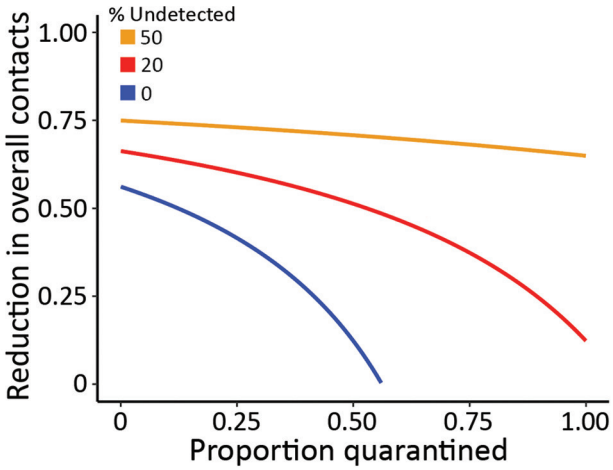


Figure 6. Levels of minimum efforts of intervention strategies needed to control the spread of severe acute respiratory syndrome coronavirus 2, (i.e. reducing the reproductive number to <1), during the coronavirus disease outbreak in China. Strategies considered were quarantine of infected persons and persons who had contact with them (*x*-axis) and population-level efforts to reduce overall contact rates (*y*-axis). Percentages denote the percentages of transmissions driven by infected persons that were not detected by surveillance as a result of asymptomatic infection, mild-to-moderate illness or low surveillance intensity.

We then evaluated the effectiveness for nonpharmaceutical interventions, such as contact tracing, quarantine, and social distancing, by using the framework by Lipsitch et al. (*19*) (Appendix 2). We extended the framework to consider a fraction of transmission occurring from infected persons who would not be identified by surveillance and can transmit effectively (*15*). This fraction is determined by the fraction of actual asymptomatic persons and the extent of surveillance efforts to identify these persons and persons with mild-to-moderate symptoms. Results show that quarantine and contact tracing of symptomatic persons can be effective when the fraction of unidentified persons is low. However, when 20% of transmission is driven by unidentified infected persons, high levels of social distancing efforts will be needed to contain the virus (Figure 6), highlighting the importance of early and effective surveillance, contact tracing, and quarantine. Future field, laboratory, and modeling studies aimed to address the unknowns, such as the fraction of asymptomatic persons, the extent of their transmissibility depending on symptom severity, the time when persons become infectious, and the existence of superspreaders are needed to accurately predict the impact of various control strategies (*20*).

Discussion

In this study, we estimated several basic epidemiologic parameters, including the incubation period (4.2 days), a time dependent duration from symptom onset to hospitalization (changing from 5.5 days in early January to 1.5 days in late January outside Hubei Province), and the time from symptom onset to death (16.1 days). By using 2 distinct approaches, we estimated the growth rate of the early outbreak in Wuhan to be 0.21–0.30 per day (a doubling time of 2.3–3.3 days), suggesting a much faster rate of spread than initially measured. This finding would have important implications for forecasting epidemic trajectories and the effect on healthcare systems as well as for evaluating the effectiveness of intervention strategies.

We found R_0 is likely to be 5.7 given our current state of knowledge, with a broad 95% CI (3.8–8.9). Among many factors, the lack of awareness of this new pathogen and the Lunar New Year travel and gathering in early and mid-January 2020 might or might not play a role in the high R_0. A recent study based on structural analysis of the virus particles suggests SARS-CoV-2 has a much higher affinity to the receptor needed for cell entry than the 2003 SARS virus (*21*), providing a molecular basis for the high infectiousness of SARS-CoV-2.

How contagious SARS-CoV-2 is in other countries remains to be seen. Given the rapid rate of spread as seen in current outbreaks in Europe, we need to be aware of the difficulty of controlling SARS-CoV-2 once it establishes sustained human-to-human transmission in a new population (*20*). Our results suggest that a combination of control measures, including early and active surveillance, quarantine, and especially strong social distancing efforts, are needed to slow down or stop the spread of the virus. If these measures are not implemented early and strongly, the virus has the potential to spread rapidly and infect a large fraction of the population, overwhelming healthcare systems. Fortunately, the decline in newly confirmed cases in China and South Korea in March 2020 and the stably low incidences in Taiwan, Hong Kong, and Singapore strongly suggest that the spread of the virus can be contained with early and appropriate measures.

Acknowledgments

We thank Alan Perelson, Christiaan van Dorp, and Ruy Ribeiro for suggestions and critical reading of the manuscript and Weili Yin for help with collecting and translating documents from provincial health commission websites.

S.S. and R.K. received funding from the Defense Advanced Research Projects Agency (grant no. HR0011938513) and the Laboratory Directed Research and Development Rapid

Response Program through the Center for Nonlinear Studies at Los Alamos National Laboratory. C.X. received funding from Laboratory Directed Research and Development Program. E.R.S. received funding from the National Institutes of Health (grant no. R01AI135946). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: R.K. and N.H. conceived the project; R.K. collected data; S.S., Y.T.L., C.X., and R.K. performed analyses; S.S., Y.T.L., E.R.S., N.H., and R.K. wrote and edited the manuscript.

Authors declare no competing interests. All data are available in the main text and in Appendices 1 and 2.

About the Author

Dr. Sanche is a postdoctoral research associate at Los Alamos National Laboratory, Los Alamos, New Mexico, USA. His primary research interest lies in complex disease dynamics inferred from data science and mathematical modeling. Dr. Lin is also a postdoctoral research associate at Los Alamos National Laboratory. His primary research interest lies in applied stochastic processes, biological physics, statistical inference, and computational system biology.

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382:1199–207. https://doi.org/10.1056/NEJMoa2001316
2. WHO. Pneumonia of unknown cause—China [cited 2020 Jan 30]. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china
3. 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. 2020;395:689–97. https://doi.org/10.1016/S0140-6736(20)30260-9
4. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surveill. 2020;25:25. https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000058
5. Du Z, Wang L, Cauchemez S, Xu X, Wang X, Cowling BJ, et al. Risk for transportation of coronavirus disease from Wuhan to other cities in China. Emerg Infect Dis. 2020;26:1049–52. https://doi.org/10.3201/eid2605.200146
6. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China [in Chinese]. Zhonghua Liu Xing Bing Xue Za Zhi. 2020;41:145–51.
7. Imai N, Dorigatti I, Cori A, Donnelly C, Riley S. Ferguson. Report 2: estimating the potential total number of novel coronavirus cases in Wuhan City, China [cited 2020 Feb 2]. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-update-epidemic-size-22-01-2020.pdf
8. Imai N, Dorigatti I, Cori A, Riley S, Ferguson NM. Estimating the potential total number of novel coronavirus cases in Wuhan City, China [cited 2020 Feb 2]. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/2019-nCoV-outbreak-report-17-01-2020.pdf
9. Hongyang L. Railway corporation using big data to trace potential virus carrier. ChinaDailyNews [cited 2020 Feb 1]. https://www.chinadaily.com.cn/a/202001/30/WS5e329ca2a310128217273b89.html
10. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al.; China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020 Feb 28 [Epub ahead of print]. https://doi.org/10.1056/NEJMoa2002032
11. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med. 2020 Mar 10 [Epub ahead of print]. https://doi.org/10.7326/M20-0504
12. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Oxford Science Publications. Oxford: Oxford University Press; 1991. p. 768.
13. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438:355–9. https://doi.org/10.1038/nature04153
14. Romero-Severson EO, Ribeiro RM, Castro M. Noise is not error: detecting parametric heterogeneity between epidemiologic time series. Front Microbiol. 2018;9:1529. https://doi.org/10.3389/fmicb.2018.01529
15. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N Engl J Med. 2020;382:970–1. https://doi.org/10.1056/NEJMc2001468
16. Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005;2:e174. https://doi.org/10.1371/journal.pmed.0020174
17. Lloyd AL. The dependence of viral parameter estimates on the assumed viral life cycle: limitations of studies of viral load data. Proc Biol Sci. 2001;268:847–54. https://doi.org/10.1098/rspb.2000.1572
18. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395:1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3
19. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003;300:1966–70. https://doi.org/10.1126/science.1086616
20. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci U S A. 2004;101:6146–51. https://doi.org/10.1073/pnas.0307506101
21. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367:1260–3. https://doi.org/10.1126/science.abb2507

Address for correspondence: Ruian Ke, T-6 Theoretical Biology and Biophysics, Mailstop K710, Los Alamos National Laboratory, NM 87544, USA; email: rke@lanl.gov

Article DOI: https://doi.org/10.3201/eid2607.200282

High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2

Appendix 2

Travel Data

We used the Baidu Migration server (https://qianxi.baidu.com/) to estimate the number of daily travelers in and out Wuhan (Appendix 1 Table 2, https://wwwnc.cdc.gov/EID/article/26/7/20-0282-App1.xlsx). The server an online platform summarizing mobile phone travel data hosted by Baidu Huiyan (https://huiyan.baidu.com). Baidu Huiyan is a widely used positioning system in China. It processes >120 billion positioning requests daily through GPS, WIFI and other means (https://huiyan.baidu.com). Specifically, we extracted from the server the Immigration Index and Emigration Index for Wuhan based on cell phone positioning data. The indexes are linearly related to the number of travelers going in and out of Wuhan, respectively. We also extracted the fraction of individuals who went to or came from a particular province. It has been reported that there were 5 million people going out of Wuhan between January 10, i.e., the start of the Chinese New Year travel rush, and January 25 (https://www.washingtonpost.com/world/asia_pacific/china-coronavirus-live-updates/2020/01/30/1da6ea52-4302-11ea-b5fc-eefa848cde99_story.html; accessed Feb. 2, 2020). This allowed us to calibrate the Emigration Index and estimated the number of daily travelers to or from a particular province, and thus the fraction of people traveling to or from a particular province (Appendix 1 Table 3).

Estimating Distributions of Epidemiologic Parameters from Individual Case Reports

We used the first confirmed cases in provinces other than Hubei to inform the time between patient infection and the onset of symptoms ($n = 24$). These individuals had all traveled

to Wuhan a short time preceding symptoms onset. Since these individuals were the first cases detected in the province, it is likely that the infection occurred during their recent stay in Wuhan. We approximated the time of infection as the middle time point of their stay. Because the delays between infection and symptoms onset vary between patients, we modeled the delay using a gamma distribution, as its support is nonnegative and it permits relatively large delays as compared to the median. Figure 1 in the main text presents results from fitting the distribution to the data (https://wwwnc.cdc.gov/EID/article/26/7/20-0282-F1.htm).

The fitting procedure was performed by maximizing the likelihood of observed delays between infection and symptoms onset. For a single observation, the individual likelihood is the gamma density function evaluated at the infection-to-onset delay. Some of the delays were censored, i.e., bounded by a certain value. For example, in some cases, only the times of infection and hospitalization were reported, and the time of symptom onset was missing in the case report. In such cases, we assumed that the missing onset time is bounded between times of infection and hospitalization. Then, the likelihood for this observation is equal to the cumulative gamma distribution evaluated at this censored value, i.e., the time when the patient was hospitalized. The maximum likelihood estimates (MLEs) are the shape and scale parameters that maximize the sum over all observations of the individual log-likelihoods. We used differential evolution in scipy.optimize library (Python) to perform maximization. A stochastic algorithm was implemented in the optimization procedure to avoid being trapped in local minima (*1*). The likelihood-based confidence intervals was computed by methods reported in Raue et al (*2*).

A similar approach was adopted to fit distributions to the time between symptom onset and hospitalization ($n = 96$), between hospitalization and discharge ($n = 6$), and between hospitalization and death ($n = 23$). The reported dates for these events was obtained directly from official sources. Data from cases originating from all over China and neighboring countries were used for distribution fitting. Detailed patient-level data are provided in Appendix 1 Table 1.

The "First-Arrival" Model: Inferring Disease Dynamics in Wuhan Using the First-Arrival Times at Other Provinces

In this model, we used the first-arrival time of a patient who traveled from Wuhan to a specific other province and was later confirmed to have been infected by SARS-CoV-2. The

rationale behind our approach is that an increasing fraction of people infected in Wuhan increases the likelihood that one such case is exported to the other provinces. Hence, how soon new cases are observed in other provinces can inform the disease progression in Wuhan. We hypothesize that this information is more reliable because the infected population in Wuhan needs to sufficiently large to allow probable export of one infected individual. The flow of expected cases depends on the flow of travelers to each province and on the proportion of the Wuhan population that is infected by the virus.

We first estimated the daily number of travelers from Wuhan to each of the China provinces. For this purpose, we used Wuhan's daily migration index to other provinces and the daily distribution of traveler destinations from Wuhan (see Data Collection). When assuming linearity between the migration index and the total number of exported individuals, it can be estimated that a migration index of 1 is approximately equal to 5 million individuals over the sum of migration indexes from January 10 to January 25, 2020 (it was reported that 5 million individuals left Wuhan during that period; see Data Collection section). The total number of daily Wuhan travelers to a province at a certain date was then set equal to the number of travelers estimated from the migration index times the fraction of the population having traveled to this province. Results from estimation are reported in Appendix 1 Table 2.

An infected traveler may be pre-symptomatic, i.e., this individual may have been exposed to the virus (E) and not have developed symptoms or be already symptomatic (I). In fact, for many individuals, infection onset was recorded days after the time of their departure from Wuhan (see Appendix 1 Table 1). Assuming travelers represent a random sample of the whole population, it follows that the probability that a traveler is infected is equal to the number of exposed or infected individuals in Wuhan ($I^* = E + I$) over the total Wuhan population ($N(t)$). The total population size varied during the infection period. We estimated the population size by using the daily inflow and outflow of individuals from Wuhan (see Appendix 1 Table 2). To represent the beginning of an outbreak, we modeled an exponential increase in the size of exposed and infected population over time t:

$$I^*(t) = e^{r(t-t_0)} \qquad \text{(Equation 1)}$$

where r is the infection growth rate and t_0 is the time of onset of exponential outbreak.

Equation 1 allows a simple analytic expression of the likelihood of arrival times for the first cases in each of the provinces other than Hubei. For a specific province, indexed by i, we modeled the arrival of new cases in each province during short time intervals as a Poisson random process $X_t^{(i)}$. Note that the rate parameter of this Poisson process, $\lambda(t) = I^*(t)\,\kappa_i(t)/N(t)$ depends on the time-varying sum of exposed and symptomatic populations $I^*(t)$, the time varying flow of population $\kappa_i(t)$ transported from Wuhan to the province i and the time varying population size. It can be shown mathematically (*3*) that the probability that no exposed or symptomatic traveler arrived to province i during a short time interval $(t, t+\Delta t)$, $\Delta t \ll 1$ is:

$$\mathbb{P}\left\{X_{t+\Delta t}^{(i)} - X_t^{(i)} = 0\right\} \approx \exp\left(-\frac{I^*(t)\kappa_i(t)}{N(t)}\Delta t\right) \qquad \text{(Equation 2)}$$

We assume no delay was incurred due transportation in our model. Equation 2 is valid for any $t > 0$, and because the overall process is Markovian, we can formulate the probability that the time of arrival of the first case in province i, $T^{(i)}$, is later than t by:

$$\mathbb{P}\left\{T^{(i)} > t\right\} = \lim_{\Delta t \to 0} \prod_{j=1}^{M} \mathbb{P}\left\{X_{j\Delta t}^{(i)} - X_{(j-1)\Delta t}^{(i)} = 0\right\} = \exp\left(-\int_{t_0}^{t} \frac{I^*(s)\kappa_i(s)}{N(s)} ds\right) \qquad \text{(Equation 2)}$$

where $[t_0, t)$ was partitioned into M equal intervals of $\Delta t = j(t - t_0)/M$, and we convert the Riemannian sum into an integral in the limit of $M \to \infty$ $(\Delta t \to 0)$. Finally, we apply $\mathrm{d}/\mathrm{d}t$ to $1 - \mathbb{P}\left\{T^{(i)} > t\right\}$ to obtain the probability density function (PDF) of the first-arrival time of province i:

$$\mathrm{PDF}_i(t) = \frac{I^*(t)\kappa_i(t)}{N(t)} \exp\left(-\int_{t_0}^{t} \frac{I^*(s)\kappa_i(s)}{N(s)} ds\right) \qquad \text{(Equation 3)}$$

The form of the probability density function Equation 4 was used to estimate the likelihood of observed arrival times in each province as a function of the growth rate r and outbreak initiation time t_0. This likelihood was maximized, again using differential_evolution in scipy.optimize (*1*), and the confidence intervals for r and t_0 were obtained through profile likelihood (*2*). Numerical integration was performed by discretizing time in daily time intervals, since both the flow of travelers and the population size in Wuhan were estimated daily.

Sensitivity Analyses for the "First-Arrival" Model

Under the 'first-arrival' model, it is assumed that all infected individuals since their arrival from Wuhan were eventually recorded/detected, i.e., 100% detection probability. However, it is possible that some first cases were missed by surveillance. Additionally, the model did not account for the possibility that detection efforts increased across provinces as the number of cases in Wuhan soared. Here, we perform sensitivity analyses to test the robustness of our estimation against these possibilities.

The model formulation above needed a small modification to perform here sensitivity analyses. The event Y: "no new arrival before time t is later diagnosed with the infection" is now equivalent to "no arrival of an infected individual before time t," "one infected arrival before time t remained undiagnosed," "two infected arrivals before time t remained undiagnosed," etc. For a Poisson process with fixed parameter λ, the probability of Y can be expressed as:

$$\mathbb{P}(Y) = \mathrm{e}^{-\lambda} + \sum_{k=1}^{\infty} \frac{(1-p)^k \lambda^k e^{-\lambda}}{k!} = e^{-\lambda p} \qquad \text{(Equation 4)}$$

where p is the probability of detection. It follows that the modified PDF formulation for sensitivity analyses is:

$$\mathrm{PDF}_i(t) = \frac{I^*(t)\kappa_i(t)\,p}{N(t)} \exp\left(-\int_{t_0}^{t} \frac{I^*(s)\kappa_i(s)\,p}{N(s)}\,ds\right) \qquad \text{(Equation 5)}$$

This PDF was used instead of equation 4 to obtain maximum likelihood estimates of the growth rate and outbreak initiation date for sensitivity analyses.

Results from Sensitivity Analyses

We evaluated the sensitivity of the growth rate estimate to these detection scenario uncertainties. A total of 23 detection scenarios were considered. As an illustration, Appendix 2 Figure 7 below describes two of these scenarios (purple and orange lines).

As shown in Appendix 2 Figure 7, we considered a start date of detection. We also considered the possibility that this date was December 25 or 31, which corresponds to the date of arrival of the first detected case in other provinces. After the start date, either the detection changed to a constant over time (the blue line), or the detection rate increases over time (the orange line). The detection probabilities were either constant over time (purple line) or increased

from early to late January (orange line). In scenarios with increasing detection over time, we considered that probabilities linearly increased either from 5% to 35%, or from 10% to 70%. The 7-fold increase was based on a recent paper from China CDC (*4*) showing that the case fatality ratio changes from 15% to 2% from early January to late January. This suggests a roughly 7-fold change in identifying infected individuals, assuming the true case fatality ratio shall be constant over time. Note that because this change reflects changes in Wuhan, rather than changes in non-Hubei provinces (from which data was used for inference), we think this 7-fold change is a maximal change given the high surveillance intensity outside of Hubei.

All considered scenarios along with their corresponding maximum likelihood estimate of the growth rate are reported in Appendix 1 Table 4. When the probability of detection was set to a constant level after the start date of detection (scenarios 1–12), the estimated growth rates are robust in the range of 0.28 to 0.29/day. t_0 changed in a wide range between Dec 3 and 21, 2019. When the probability of detection of a case was set to 10%, the estimated growth rate remained 0.29/day, but the estimated outbreak initiation date was Dec 12, 2019. When the probability of detection changes over time (scenarios 13–20), the estimated growth rates are in the range between 0.21–0.25/day.

An additional analysis we did was to fixed t_0 to December 1st and estimate detection levels (scenarios 21 to 23). Growth rate estimates in these cases are between 0.21 and 0.23/day, the detection level changed 2–3 folds.

Overall, growth rate estimates varied from 0.21 to 0.3 across scenarios. An evaluation of the AIC suggests that models with constant levels of detection better fit the data. This cannot be attributed to model parsimony as the number of estimated parameters was the same across all scenarios (two parameters). This could indicate relatively constant awareness in non-Hubei provinces where no or very few cases had been detected.

The “Case Count” Model: the SEIR-Type Hybrid Stochastic Model

Model 1 fitted the time of arrival of the first confirmed case of each province. We used a different approach and a different dataset to infer disease dynamics. In particular, we constructed a hybrid stochastic model for inferring the disease dynamics in Wuhan using daily counts of individuals who contracted the infection in Wuhan and were diagnosed outside Hubei province.

The model is hybrid in the sense that we will couple a deterministic and exponential growth to describe the outbreak in Wuhan and an agent-based model which describes the discrete population dynamics of the patients after they left Hubei to other provinces. In Appendix 2 Figure 8, we present a schematic diagram of the hybrid meta-population model.

Deterministic and Exponential Dynamics in Wuhan

We assume an exponential growth of the number of exposed (E_W, W for *W*uhan) and symptomatic (I_W) populations in Wuhan over time, $E_W(t) = E_W(0)e^{rt}$ and $I_W(t) = I_W(0)e^{rt}$ from the onset. The overall growth rate r is dominated by the largest eigenvalue of a sequential compound process, and given an r value, the ratio $\phi \coloneqq E(0)/I(0)$ is asymptotic constant (*4*). Thus, given a growth rate parameter r and an initial condition $E(t_0) + I(t_0) = 1$, we numerically compute the exposed population $E(t) = \phi(r)\left(1 + \phi(r)\right)^{-1} \exp\left(r(t - t_0)\right)$ and the symptomatic population $I(t) = \left(1 + \phi(r)\right)^{-1} \exp\left(r(t - t_0)\right)$.

Agent-Based Model for Patients Who Have Left Wuhan to Other Provinces

We assume that between 1/1 and 1/26, the populations in Wuhan are large and the dynamics can be reasonably approximate by the above deterministic and exponentially growing curves. However, the initial propagation of the disease to other provinces in China involves only a small population of exposed (E_O, O for *O*thers) or symptomatic individuals who left Hubei province. In addition, the transitions between different phases of these patients, from exposed (E_O) to symptomatic (I_O), over to hospitalized (H_O), and finally to be confirmed by laboratory examinations (C_O) in other provinces are also variable (as we quantified in Figure 1, panels C–F). Consequently, the resulting population dynamics in other provinces is highly stochastic. We thus adopt an agent-based modeling approach and rely on kinetic Monte Carlo Sampling techniques detailed below to simulate the population dynamics in other provinces. With this approach, we aim to generate samples of 1) each individual patient who left Wuhan at a specific date, and 2) the individual's health status as the time progresses (susceptible, exposed, or symptomatic). The goal is to accumulate a large amount of Monte Carlo samples, by which we can compute the key summary statistics, i.e., the average case reported on each day between 1/18 and 1/26, to be compared against to the data. We achieve this by the following algorithmic procedures.

1. Generate random number of infected populations leaving Wuhan. We collected migration index which quantifies the fraction of total populations (14 million) in Wuhan that

traveled to other provinces on each date $t_i = 1, \ldots, 26$ (see Appendix 1 Table 3). Assuming independence of an individual's health state (susceptible, exposed, or symptomatic) and the individual's migration decision (leaving to other provinces or not), on each date t_i, the exposed and symptomatic populations leaving Hubei can be modeled by two Binomial distributions, $B_E = \text{Binomial}\big(E_W(t_i), \mu(t_i)\big)$ and $B_I = \text{Binomial}\big(I_W(t_i), \mu(t_i)\big)$. Here, $E_W(t)$ and $I_W(t)$ are the exposed and symptomatic population in Wuhan, and are assigned to the nearest integers to the previously prescribed exponential growth, given model parameters (r, t_0). Thus, to generate one stochastic sample path (realization), we generate Binomially-distributed random populations leaving Hubei on each day between 1/1 and 1/26 (both included), and model each of these *in silico* patients' health states by the following procedures.

2. Generate the progression of the health state for each patient: We assume that each hypothetical patient generated by the above procedure would stochastically, identically and independently progress toward to be confirmed (C_O) and reported in one of the other provinces. If an individual was exposed (E_O) when s/he left Hubei at t_i, we generate a Gamma distributed random time $\Delta t_{E\to I} \sim \Gamma(\alpha_1, \beta_1)$ and update the individual's health state to symptomatic (I_O) at time $t_i + \Delta t_{E\to I}$. We chose a time-dependent waiting-time distribution for the progression from symptomatic sate I_O to reflect the two regimes we observed from the data (see main text): If $t_i + \Delta t_{E\to I}$ is before 1/18 (included), we generate a Gamma distributed random time $\Delta t_{I\to H} \sim \Gamma\big(\alpha_{2,1}, \beta_{2,1}\big)$ to model the waiting time for an infected patient to be hospitalized (otherwise, if it is later than 1/18, $\Delta t_{I\to H} \sim \Gamma\big(\alpha_{2,2}, \beta_{2,2}\big)$). Consequently, the patient's state is changed to H_O at time $t_i + \Delta t_{E\to I} + \Delta t_{I\to H}$. If $t_i + \Delta t_{E\to I} + \Delta t_{I\to H}$ is before 1/19, the patient would wait in the "H" state until 1/19 when the policy of case confirmation was announced and institutionalized. Then, the confirmation process is modeled by another Gamma distributed random time $\Delta t_{H\to C} \sim \Gamma(\alpha_3, \beta_3)$. The patient is then confirmed and reported at time $t_i + \Delta t_{E\to I} + \Delta t_{I\to H} + \Delta \mathrm{t}_{H\to C}$, and we add one more case report at the next integer (date of January). Similar procedure applied to a patient who had already progressed to the I_W state before s/he left Hubei on date t_i, with the exception that the first random waiting time is neglected—the patient's confirmation time would be $t_i + \Delta t_{I\to H} + \Delta_{H\to C}$. We repeat the procedure for each *in-silico* patient who left Wuhan between 1/1 and 1/26 (both included), and register the time when these patients were reported between 1/18 and 1/26 (both included).

Parameter Estimation and Uncertainty Quantification of (r, t_0)

It is our task to infer the unknown parameters, exponential growth rate r and exponential growth onset time t_0 by the number of confirmed cases reported between 1/18 and 1/26. This is possible because the information of the unknown parameters (r, t_0) have an impact of the deterministic growths of the exposed $E_W(t)$ and symptomatic population $I_W(t)$, which in turn have an impact on the random populations which have left Hubei on each date. These populations follow statistically quantified processes until the final confirmation outside of Hubei, and can be compared against the reported data.

An error measure is devised to assess the quality of fit of the model given a set of parameters (r, t_0) by the following procedures. For each parameter set, we generate $2^{13} = 8192$ Monte Carlo samples. On each date t_i, the jth sample reports a random number $n_C^{MC}(t_i|r, t_0, j)$ of confirmed new cases. We thus average over all the samples and obtain an averaged number of newly confirmed cases on a date t_i, $n_C^{MC}(t_i|r, t_0) \coloneqq \sum_{j=1}^{8192} n_C^{MC}(t_i|r, t_0, j)$, and compare it to the actual data $n_C^{Data}(t_i)$. We quantify the quality of the fit by computing the sum of the squared residuals:

$$\varepsilon^2(r, t_0) \coloneqq \sum_{t_i=18}^{26} \left[n_C^{MC}(t_i|r, t_0) - n_C^{Data}(t_i)\right]^2 \qquad \text{(Equation 6)}$$

A 100×100 grid-based parameter scan is performed to identify the parameters in the region $0.22 < r < 0.42$ and $-20 \leq t_0 \leq -5$ for identifying the best-fit parameters:

$$r^*, t_0^* \coloneqq \operatorname{argmin}_{\{r, t_0\}} \varepsilon^2(r, t_0) \qquad \text{(Equation 7)}$$

As for uncertainty quantification, we formulate the logarithm of the likelihood $\mathcal{L}$ of a parameter set (r, t_0) as

$$\log \mathcal{L}(\alpha, \tau) \coloneqq -n \frac{\varepsilon^2(r, t_0)}{\varepsilon^2(r^*, t_0^*)} \qquad \text{(Equation 8)}$$

Here, $n = 9$ is the number of data points we use to fit the model. The assumption we make to formulate the above likelihood is that 1) the data (number reported new cases on date t_i) is normally distributed with a mean which equals to the Monte Carlo mean reported new cases in

our model, and 2) the variance of the noise is identically and t_i-independently distributed, and the variance is equal to the mean squared residuals of the best-fit model.

We can then formulate a likelihood ratio test, which quantifies how likely a set of parameters (r, t_0) is in comparison to the best-fit parameters (r^*, t_0^*):

$$\mathbb{P}\{r, t_0 \mid Data\} \sim \exp\left[-n\left(1 - \frac{\varepsilon^2(r, t_0)}{\varepsilon^2(r^*, t_0^*)}\right)\right] \qquad \text{(Equation 9)}$$

In Bayesian inference, what we computed is essentially the joint posterior distribution of the model parameters (r, t_0), provided a uniform prior distribution on the region of our interests. We present this joint distribution in Appendix 2 Figure 5. Finally, because the joint posterior is narrowly distributed, we can numerically compute the marginalized posterior,

$$\begin{aligned} \mathbb{P}\{r \mid Data\} &\sim \int \mathbb{P}\{r, t_0 \mid Data\}\, \mathrm{d}t_0 \\ \mathbb{P}\{t_0 \mid Data\} &\sim \int \mathbb{P}\{r, t_0 \mid Data\}\, \mathrm{d}r \end{aligned} \qquad \text{(Equation 10)}$$

which is reported in Figure 4 and used to calculate the bounds of centered 95% probability mass to estimate the confidence interval of the growth rate r.

Calculation of R_0 from Estimated Exponential Growth Rates

Assuming gamma distributions for the latent and infectious periods, Wearing et al. (*5*) have shown that the value of R_0 can be calculated from estimated exponential growth rate, r, of an outbreak as:

$$R_0 = \frac{r\left(\frac{r}{\sigma m} + 1\right)^m}{\gamma\left[1 - \left(\frac{r}{\gamma n} + 1\right)^{-n}\right]} \qquad \text{(Equation 12)}$$

where $1/\sigma$ and $1/\gamma$ are the mean latent and infectious periods, respectively, and m and n are the shape parameters for the gamma distributions for the mean latent and infectious periods, respectively.

To quantify the uncertainty of R_0, we assumed that m = 4.5 (same as the shape parameter we estimated for the incubation period), n = 3. The parameters (r, σ, γ) are assumed to be mutually independent and we generate random samples according to ranges of variations defined in Appendix 1 Table 5 to compute the resulting R_0. We generated 10^4 parameters, accepted those that result in a serial interval within the range of interests, and then computed their respective R_0 using Equation 12. We used the 97.5% and 2.5% percentile of the generate data to quantify the 95% confidence interval.

Calculation of the Impact of Intervention Strategies

Using a susceptible–exposed (noninfectious)– infectious–recovered (SEIR) type compartmental model, Lipsitch et al. (*6*) evaluated the impact of quarantine of symptomatic cases and their contacts to prevent further transmission. Assuming that only symptomatic individuals transmit the pathogen, they showed that the reproductive number after the intervention, R_{int}, can be expressed as:

$$R_{int} = \frac{R(1-q)D_{int}}{D} \qquad \text{(Equation 11)}$$

where R is the reproductive number before intervention, q is the percentage of infected individuals being quarantined, D_{int} and D are the mean durations of infectious period after intervention and without intervention, respectively.

Here in our model, we adopted this formulation; however, we assumed that a fraction, f, of infected individuals are asymptomatic and can transmit. In this case, quarantine of symptomatic individuals only reduces the contribution of these individuals toward the reproductive number. Thus, we can calculate the reproductive number under quarantine, R_q, as:

$$R_q = fR + (1-f)R_{int} = R\left(f + (1-f)(1-q)\frac{D_{int}}{D}\right) \qquad \text{(Equation 12)}$$

We also considered another form of control measure, i.e., the population-level control measure that reduces overall number of daily contacts in the population by ε. These measures include closing down of transportation systems, work and/or school closure, etc. Since R depends

linearly on the number of daily contacts, we calculate the combined impact of the individual-based quarantine and the population level control measure as:

$$R_{combine} = (1-\varepsilon)[fR + (1-f)R_{int}] = (1-\varepsilon)R\left(f + (1-f)(1-q)\frac{D_{int}}{D}\right) \quad \text{(Equation 13)}$$

In our calculations, we assume that the mean duration of infectious period of COVID-19 to be 10 days, i.e. D = 10 days. We further assume that intervention can reduce infectious period to 4 days, i.e., $D_{int} = 4$ days, based on data on the time from symptom onset to hospitalization from Singapore (*7*) and that individuals may transmit the virus before symptom onset. Since Singapore has one of the best surveillance systems for emerging infectious diseases like COVID-19, the value of D_{int} used here shall represent the best scenario for case isolation intervention. We set the value of R to be the median estimate of R_0, i.e., $R_0 = 5.7$.

References

1. Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim. 1997;11:341–59. https://doi.org/10.1023/A:1008202821328

2. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25:1923–9. PubMed https://doi.org/10.1093/bioinformatics/btp358

3. Cox DR, Oakes D. Analysis of survival data. Boca Raton (Florida): Chapman & Hall/CRC; 1984.

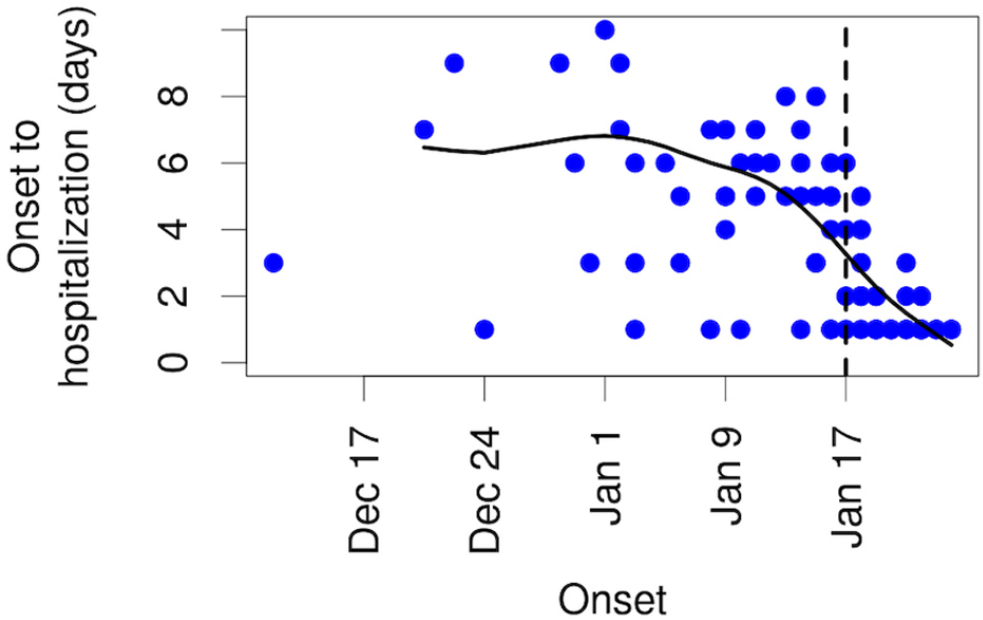
4. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China [in Chinese]. Zhonghua Liu Xing Bing Xue Za Zhi. 2020;41:145–51. PubMed

5. Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. PLoS Med. 2005;2:e174. PubMed https://doi.org/10.1371/journal.pmed.0020174

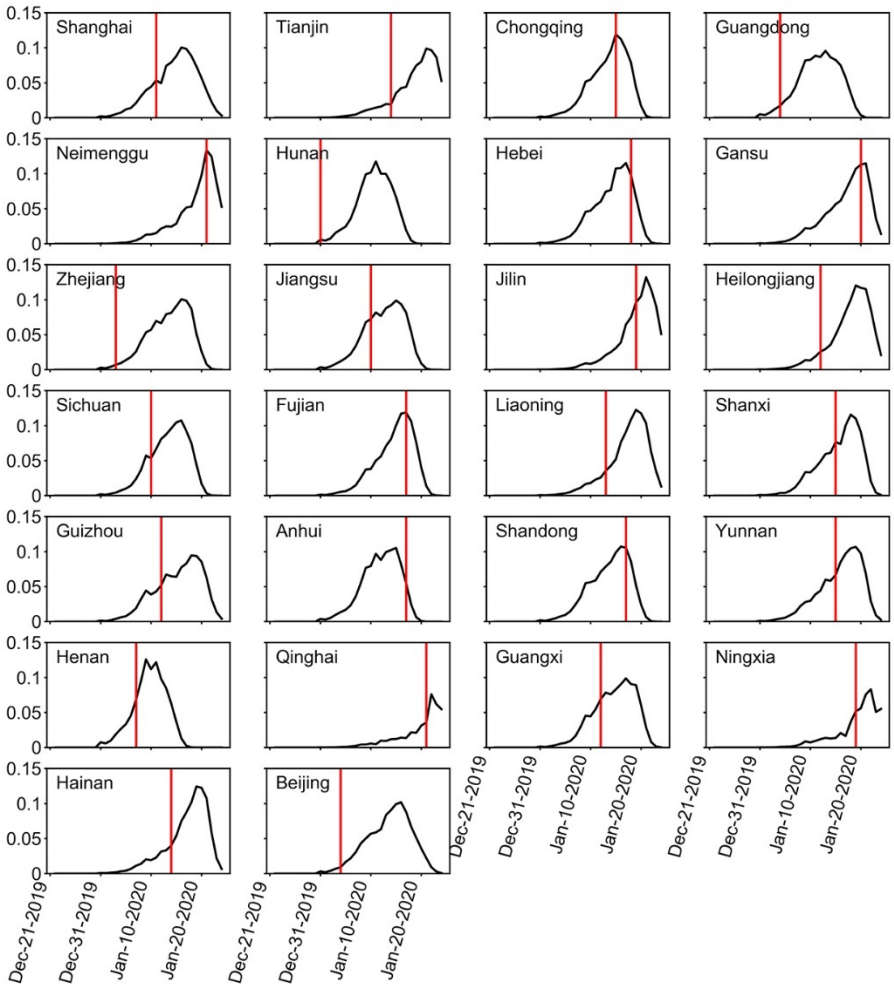
6. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003;300:1966–70. PubMed https://doi.org/10.1126/science.1086616

7. Ng Y, Li Z, Chua YX, Chaw WL, Zhao Z, Er B, et al. Evaluation of the effectiveness of surveillance and containment measures for the first 100 patients with COVID-19 in Singapore—January 2–

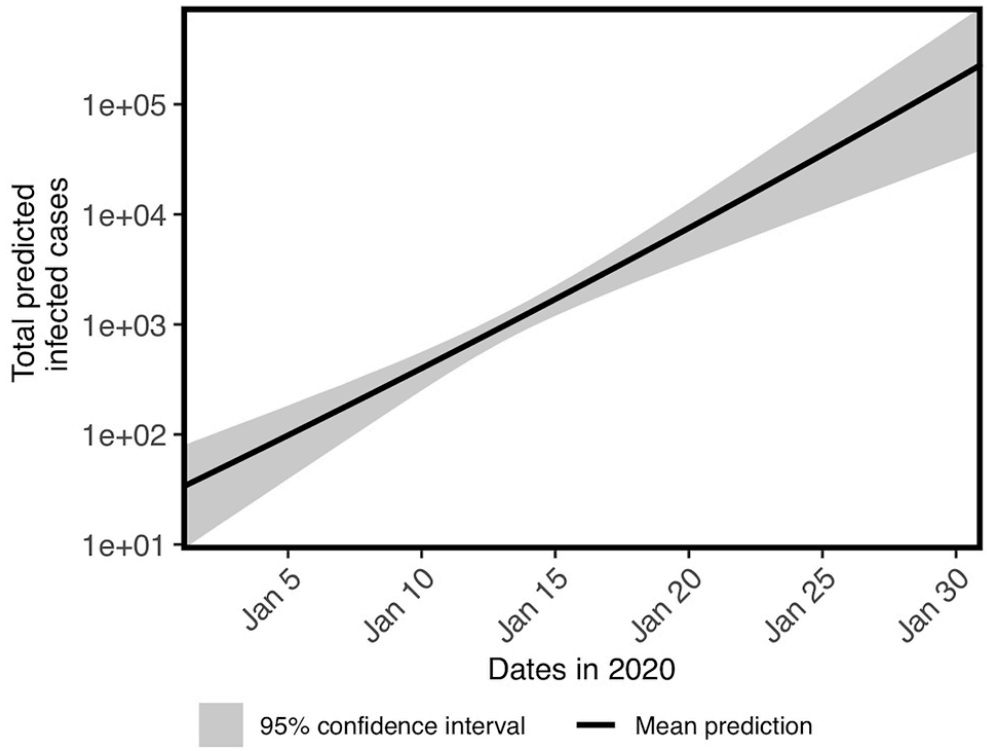
February 29, 2020. MMWR Morb Mortal Wkly Rep. 2020;69:307–11. PubMed https://doi.org/10.15585/mmwr.mm6911e1



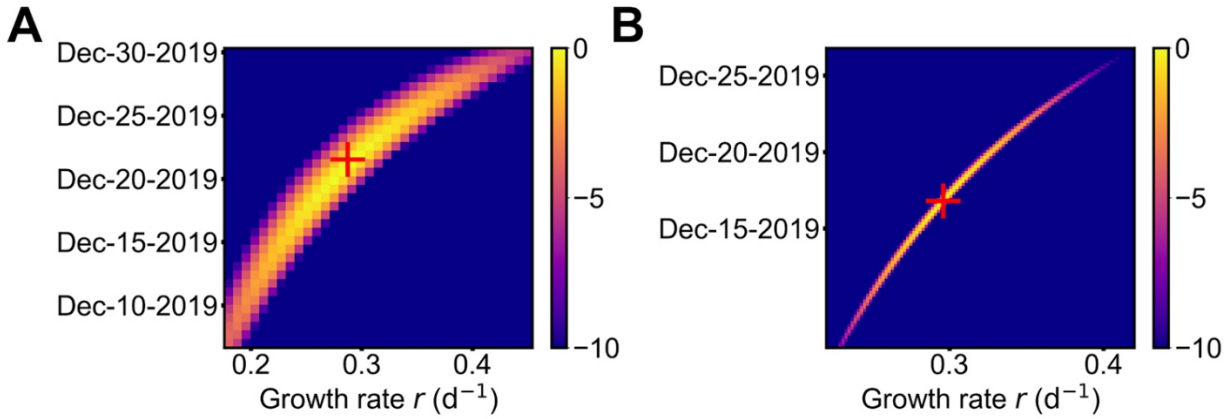
Appendix 2 Figure 1. The duration from symptom onset to hospitalization (y-axis) decreases over time during the outbreak.



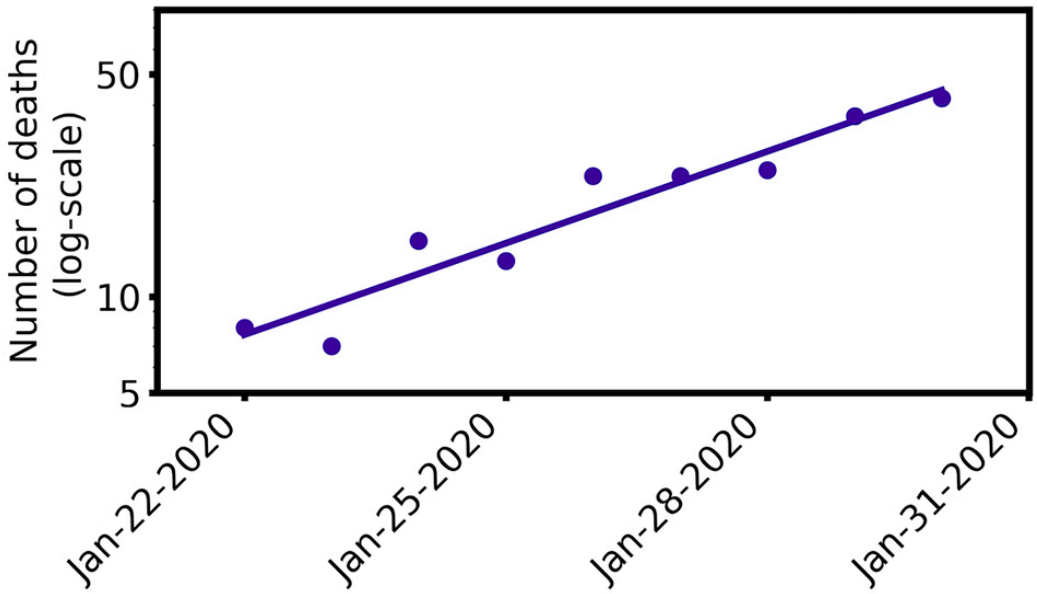
Appendix 2 Figure 2. Predictions of the 'first arrival' model using best-fit parameters agree well with data. Probability densities of times of first arrival of infected cases in each province based on our maximum likelihood estimate (curves) and documented times of first arrival of infected individuals in our case report dataset (lines).



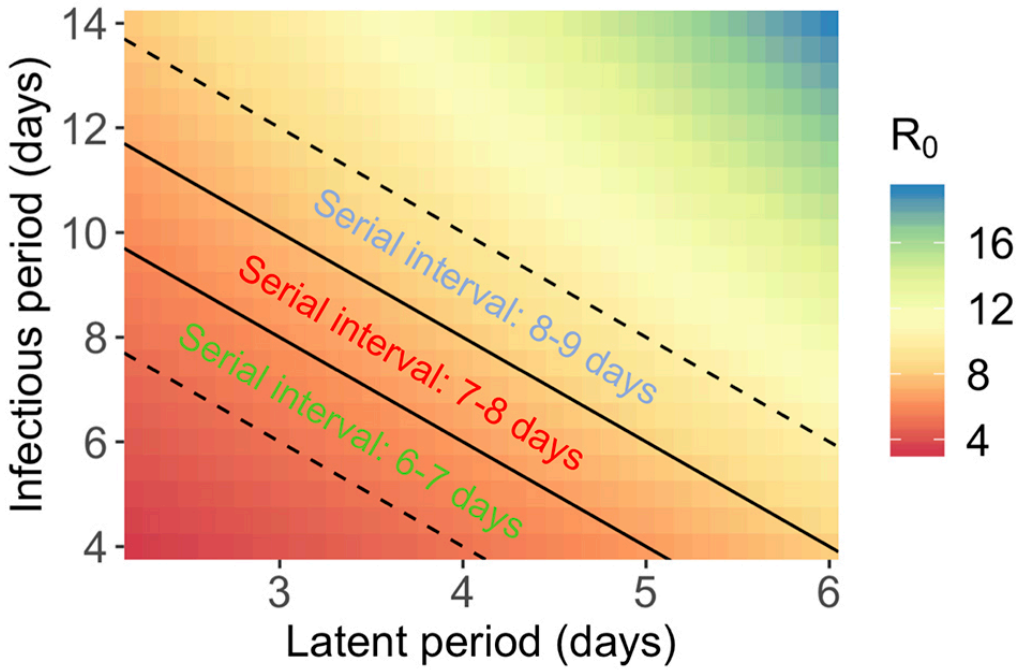
Appendix 2 Figure 3. Projections of numbers of infected individuals in Wuhan between January 1 and 30, 2020 using the likelihood profile of parameter values in the 'first arrival' approach. Projections after the lock-down of Wuhan on January 23 were hypothetical scenarios assuming no control measures are implemented.



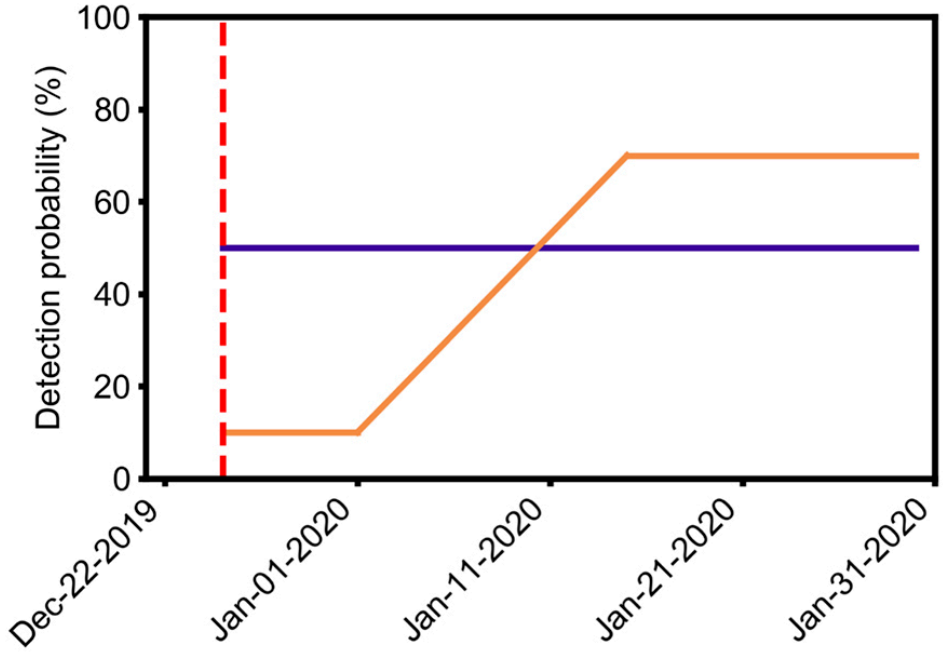
Appendix 2 Figure 4. Log-likelihood profiles of the estimated exponential growth rate of the outbreak, r (x-axis) and the date of exponential growth initiation (y-axis) from the 'first arrival' model (A) and the 'case count' model (B).



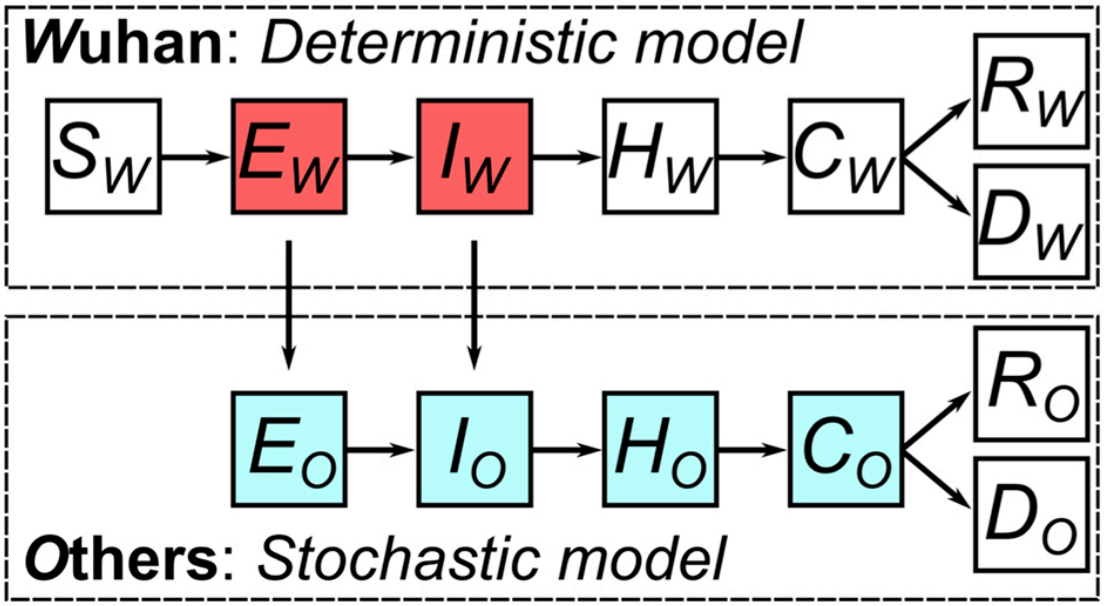
Appendix 2 Figure 5. The growth rate of the number of daily new death cases (on a log scale) in Hubei province in late January 2020 is estimated to be 0.27/day for cases collected between Jan. Twenty-two and 29, 2020. Dots and the blue line denote data and a fitted regression line, respectively. Note, there is a decrease in the growth rate after Jan 29, possibly reflecting intervention efforts or overwhelmed hospital system. When we include the data points on Jan. Thirty and 31, we get a growth rate of 0.22/day. We think the estimation using early data points are a better reflection of the early infection dynamics; however, we report a growth rate of the new death counts to be between 0.22 and 0.27/day.



Appendix 2 Figure 6. Heatmap showing how R_0 changes with the mean durations of latent period and infectious period. The mean latent period is varied between 2.2 days and 6 days. The lower bound include the possibility that infected individual becomes infectious 2–3 days before symptom onset. The mean infectious period is varied between 4–14 days. The outbreak growth rate, r, is set to 0.29/day. Solid and dashed lines denote serial interval of 6, 7, 8 and 9 days, where we assumed the serial interval is the sum of the latent period and the half of the infectious period.



Appendix 2 Figure 7. Illustration of two detection scenarios (blue and orange lines) considered in sensitivity analysis. In both illustrated scenarios, it was considered impossible that a case who had arrived from Wuhan before Dec. 25, 2019 could be later detected with coronavirus (red dashed line), i.e., 0% detection before Dec 25. In the blue scenario, the detection probability changes from 0 to 50% after Dec 25; whereas the detection probability changes from 0 to 10% after Dec 25 and increases from 10% to 70% linearly between Jan. One and 15, 2020.



Appendix 2 Figure 8. Schematic diagram of the proposed meta-population model. Schematic diagram of the hybrid stochastic model. The model is a variant of the SEIR model with two geographic compartment, Wuhan (subscripted W) and other provinces (subscripted O). In Wuhan, a susceptible patient in compartment S_W is first exposed and progresses to an exposed state (E_W), progressed to be infected (I_W), hospitalized (H_W), and then became a confirmed case (C_W), and either recovered (R_W) or deceased (D_W). A portion of ill population (E_W and I_W) moved to other provinces and followed a similar progression. Because these populations are small and thus the dynamics are stochastic, we adopt an agent based approach to simulate the disease dynamics ($E_O(t)$, $I_O(t)$, $H_O(t)$ and $C_O(t)$) in other provinces. The case reports on each day in other provinces were compared against the model's output, $C_O(t)$ to constrain the unknown initial onset and growth rate in Wuhan.