Article

Reconstruction of the full transmission dynamics of COVID-19 in Wuhan

https://doi.org/10.1038/s41586-020-2554-8

Received: 14 April 2020

Accepted: 10 July 2020

Published online: 16 July 2020

Check for updates

Xingjie Hao[1,2,8], Shanshan Cheng[1,2,8], Degang Wu[1,2,8], Tangchun Wu[1,3,4]✉, Xihong Lin[5,6,7]✉ & Chaolong Wang[1,2,4]✉

As countries in the world review interventions for containing the pandemic of coronavirus disease 2019 (COVID-19), important lessons can be drawn from the study of the full transmission dynamics of its causative agent—severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)— in Wuhan (China), where vigorous non-pharmaceutical interventions have suppressed the local outbreak of this disease[1]. Here we use a modelling approach to reconstruct the full-spectrum dynamics of COVID-19 in Wuhan between 1 January and 8 March 2020 across 5 periods defined by events and interventions, on the basis of 32,583 laboratory-confirmed cases[1]. Accounting for presymptomatic infectiousness[2], time-varying ascertainment rates, transmission rates and population movements[3], we identify two key features of the outbreak: high covertness and high transmissibility. We estimate 87% (lower bound, 53%) of the infections before 8 March 2020 were unascertained (potentially including asymptomatic and mildly symptomatic individuals); and a basic reproduction number (R_0) of 3.54 (95% credible interval 3.40–3.67) in the early outbreak, much higher than that of severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS)[4,5]. We observe that multipronged interventions had considerable positive effects on controlling the outbreak, decreasing the reproduction number to 0.28 (95% credible interval 0.23–0.33) and—by projection—reducing the total infections in Wuhan by 96.0% as of 8 March 2020. We also explore the probability of resurgence following the lifting of all interventions after 14 consecutive days of no ascertained infections; we estimate this probability at 0.32 and 0.06 on the basis of models with 87% and 53% unascertained cases, respectively—highlighting the risk posed by substantial covert infections when changing control measures. These results have important implications when considering strategies of continuing surveillance and interventions to eventually contain outbreaks of COVID-19.

COVID-19, caused by SARS-CoV-2, was detected in Wuhan in December 2019[6]. The high population density, together with increased social activities before the Chinese New Year, catalysed the outbreak; the spread of the outbreak was expedited by massive human movement during the *Chunyun* holiday travel season from 10 January 2020[3]. Shortly after the confirmation of human-to-human transmission, the Chinese authorities implemented an unprecedented cordon sanitaire of Wuhan on 23 January to contain the geographical spread of the disease, followed by a series of non-pharmaceutical interventions—including suspension of all intra- and inter-city transportation, compulsory mask wearing in public places, cancellation of social gatherings and the home quarantine of individuals with presumed infections, those with COVID-19 related symptoms and their close contacts[1]—to reduce virus transmission. From 2 February, a strict stay-at-home policy for all residents, and the centralized isolation and quarantine of all patients, individuals suspected to have contracted the virus and their close contacts were implemented to stop household and community transmission. In addition, a city-wide door-to-door universal survey of symptoms was carried out during 17–19 February by designated community workers, to identify previously undetected symptomatic cases. These interventions—together with improved medical resources and the redeployment of healthcare personnel from all over the country—have crushed the epidemic curve and reduced the attack rate in Wuhan, with the potential to shed light on global efforts to control outbreaks of COVID-19[1].

Recent studies have revealed important transmission features of COVID-19, including the infectiousness of asymptomatic[7–10] and

[1]Ministry of Education Key Laboratory of Environment and Health, State Key Laboratory of Environmental Health (Incubating), School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [2]Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [3]Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [4]National Medical Center for Major Public Health Events, Huazhong University of Science and Technology, Wuhan, China. [5]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [6]Department of Statistics, Harvard University, Cambridge, MA, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]These authors contributed equally: Xingjie Hao, Shanshan Cheng, Degang Wu. ✉e-mail: wut@mails.tjmu.edu.cn; xlin@hsph.harvard.edu; chaolong@hust.edu.cn

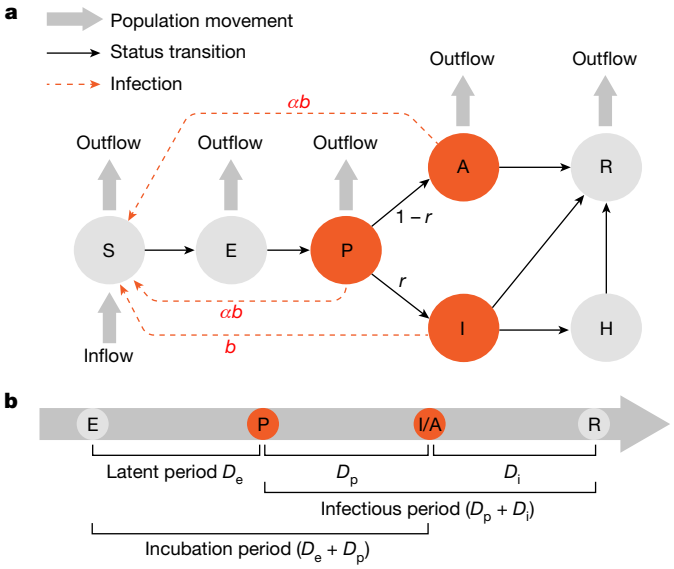


Fig. 1 | Illustration of the SAPHIRE model. We extended the classic SEIR model to include seven compartments: susceptible (S), exposed (E), presymptomatic infectious (P), ascertained infectious (I), unascertained infectious (A), isolation in hospital (H) and removed (R). **a**, Relationship between different compartments. Two parameters of interest are *r* (ascertainment rate) and *b* (transmission rate), which are assumed to vary across time periods. **b**, Schematic disease course of symptomatic individuals. In this model, the unascertained compartment A includes asymptomatic and some mildly symptomatic individuals who were not detected. Although there is no presymptomatic phase for asymptomatic individuals, we treated asymptomatic as a special case of mildly symptomatic and modelled both with a 'presymptomatic' phase for simplicity.

presymptomatic[2,11,12] individuals. Furthermore, the number of ascertained cases was much smaller than that estimated using international cases exported from Wuhan before the travel suspension[3,13,14], which implies a substantial number of unascertained cases. Using reported cases from 375 cities in China, a previous modelling study concluded that a sizeable number of unascertained cases—despite having lower transmissibility—had facilitated the rapid spreading of COVID-19[15]. In addition, accounting for unascertained cases has refined the estimation of case fatality risk of COVID-19[16]. Modelling both ascertained and unascertained cases is important for interpreting transmission dynamics and epidemic trajectories.

On the basis of comprehensive epidemiological data from Wuhan[1], we delineated the full dynamics of COVID-19 in the epicentre by extending the susceptible–exposed–infectious–recovered (SEIR) model to include presymptomatic infectiousness (P), unascertained cases (A) and case isolation in the hospital (H), generating a model that we name SAPHIRE (Fig. 1, Methods, Extended Data Tables 1, 2). We modelled the outbreak from 1 January 2020 across 5 time periods that were defined on the basis of key events and interventions: 1–9 January (before *Chunyun*), 10–22 January (*Chunyun*), 23 January to 1 February (cordon sanitaire), 2–16 February (centralized isolation and quarantine) and 17 February to 8 March (community screening). We assumed a constant population size of 10 million with equal numbers of daily inbound and outbound travellers (500,000 before *Chunyun*, 800,000 during *Chunyun* and 0 after cordon sanitaire)[3]. Furthermore, we assumed that the transmission rate and ascertainment rate did not change in the first two periods (because few interventions were implemented before 23 January), whereas these rates were allowed to vary in later periods to reflect the strengths of different interventions. We estimated these rates across periods by Markov Chain Monte Carlo (MCMC) and further converted the transmission rate into the effective reproduction number (R_e) (Methods).

We first simulated epidemic curves with two periods to validate our parameter estimation procedure (Methods, Extended Data Fig. 1). Our method could accurately estimate R_e and the ascertainment rates when the model was correctly specified, and was robust to misspecification of the duration from the onset of symptoms to isolation and of the relative transmissibility of unascertained versus ascertained cases. As expected, estimates of R_e were positively correlated with the specified latent and infectious periods, and the estimated ascertainment rates were positively correlated with the specified ascertainment rate in the initial state.

Using confirmed cases exported from Wuhan to Singapore (Extended Data Table 3), we conservatively estimated the ascertainment rate during the early outbreak in Wuhan to be 0.23 (95% confidence interval 0.14–0.42; unless specified otherwise, all parenthetical ranges refer to the 95% credible interval) (Methods). We then fit the daily incidences in Wuhan from 1 January to 29 February, assuming the initial ascertainment rate was 0.23, and predicted the trend from 1 March to 8 March (Methods). Our model fit the observed data well, except for the outlier on 1 February; this outlier might be due to the approximate-date records of many patients admitted to the field hospitals set up after 1 February (Fig. 2a). After a series of multifaceted public health interventions, R_e decreased from 3.54 (3.40–3.67) and 3.32 (3.19–3.44) in the first two periods to 1.18 (1.11–1.25), 0.51 (0.47–0.54) and 0.28 (0.23–0.33) in the later three periods (Fig. 2b, Extended Data Tables 4, 5). We estimated the cumulative number of infections, including unascertained cases, up until 8 March to be 258,728 (204,783–320,145) if the trend of the fourth period was assumed (Fig. 2c), 818,724 (599,111–1,096,850) if the trend of the third period was assumed (Fig. 2d) or 6,302,694 (6,275,508–6,327,520) if the trend of the second period was assumed (Fig. 2e), in comparison to the estimated total infections of 249,187 (198,412–307,062) obtained by fitting data from all 5 periods (Fig. 2a). Correspondingly, these numbers translate into a 3.7%, 69.6% and 96.0% reduction of infections by the measures taken in the fifth period, the fourth and the fifth periods combined, and the last three periods combined, respectively.

We estimated low ascertainment rates throughout: 0.15 (0.13–0.17) for the first two periods, and 0.14 (0.11–0.17), 0.10 (0.08–0.12), and 0.16 (0.13–0.21) for the remaining three periods (Extended Data Table 6). Even with the universal screening of the community for symptoms that was implemented from 17 February to 19 February, the ascertainment rate was raised only to 0.16. On the basis of the fitted model using data from 1 January to 29 February, we projected the cumulative number of ascertained cases to be 32,577 (30,216–34,986) by 8 March, close to the reported number of 32,583. This was equivalent to an overall ascertainment rate of 0.13 (0.11–0.16) given the estimated total infections of 249,187 (198,412–307,062). The model also projected that the number of daily active infections (including presymptomatic, ascertained and unascertained infections) peaked at 55,879 (43,582–69,571) on 2 February and dropped afterwards to 701 (436–1,043) on 8 March (Fig. 2f). If the trend remained unchanged, the number of ascertained infections would have first become zero on 27 March (95% credible interval 20 March to 5 April), and the clearance of all infections would have occurred on 21 April (8 April to 12 May) (Extended Data Table 7). The first day of zero ascertained cases in Wuhan was reported on 18 March, indicating enhanced interventions in March.

We used stochastic simulations to investigate the implications of unascertained cases for continuing surveillance and interventions[17] (Methods). Because of latent, presymptomatic and unascertained cases, the source of infection would not be completely cleared shortly after the first day of zero ascertained cases. We found that if control measures were lifted 14 days after the first day of zero ascertained cases, the probability of resurgence, defined as the number of active ascertained cases greater than 100, could be as high as 0.97, and the surge was predicted to occur on day 34 (27–47) after lifting controls (Fig. 3). If we were to impose a more-stringent criterion of lifting controls after

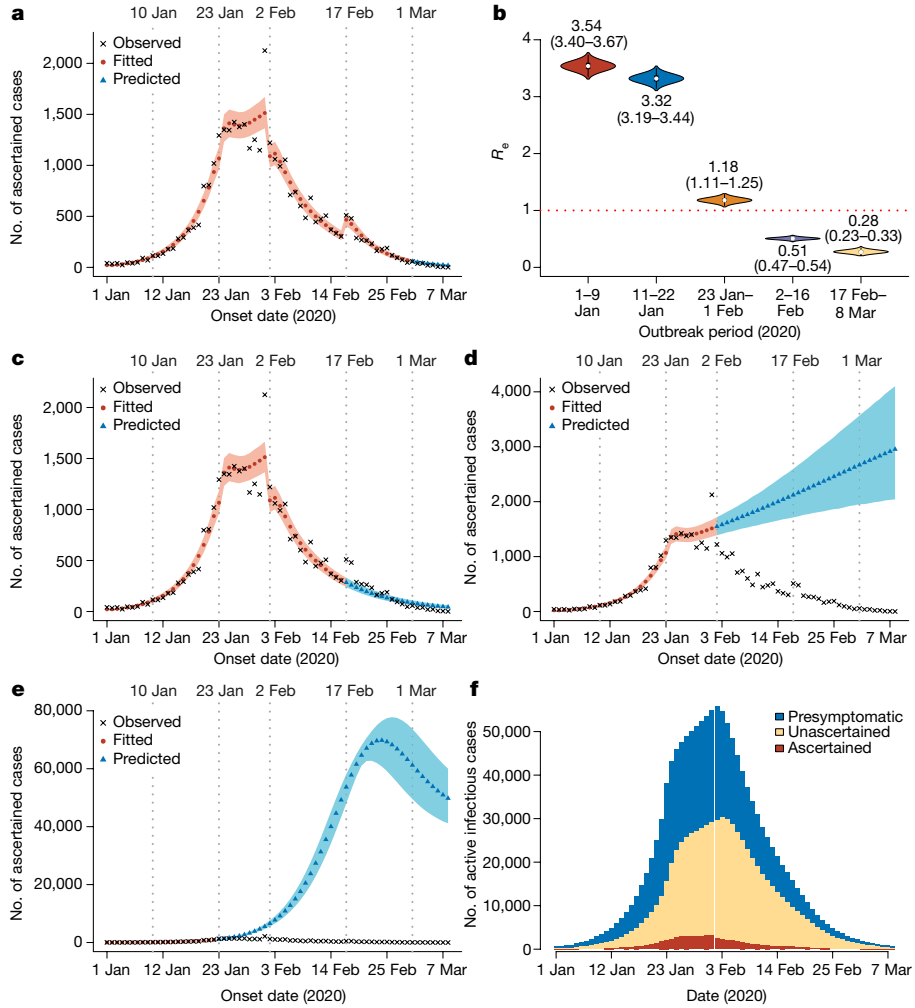


Fig. 2 | Modelling the COVID-19 epidemic in Wuhan. Parameters were estimated by fitting data from 1 January to 29 February. **a**, Prediction using parameters from period 5 (17 February–29 February). **b**, Distribution of R_e estimates from 10,000 MCMC samples. In each violin plot, the white dot represents the median, the thick bar represents the interquartile range and the thin bar represents the minimum and the maximum. The mean and the 95% credible interval (in parentheses) are labelled below or above. **c**, Prediction using parameters from period 4 (2 February–16 February). **d**, Prediction using parameters from period 3 (23 January to 1 February). **e**, Prediction using parameters from period 2 (10 January to 22 January). The shaded areas in **a**, **c**–**e** are 95% credible intervals, and the coloured points are the mean values based on 10,000 MCMC samples. **f**, Estimated number of active infectious cases in Wuhan from 1 January to 8 March.

observing no ascertained cases in a consecutive period of 14 days, the probability of resurgence would drop to 0.32, with possible resurgence delayed to day 42 (33–55) after lifting controls (Fig. 3). These results highlight the risk of ignoring unascertained cases in switching intervention strategies, despite our use of a simplified model.

We performed a series of sensitivity analyses to test the robustness of our results by smoothing the outlier data point on 1 February, as well as varying the lengths of latent and infectious periods, the duration from the onset of symptoms to isolation, the ratio of transmissibility in unascertained versus ascertained cases, and the initial ascertainment rate (Extended Data Tables 4–7, Supplementary Information). Our major findings, of a marked decrease in R_e after interventions and the existence of a substantial number of unascertained cases, were robust. Consistent with simulations, the estimated ascertainment rates were positively correlated with the specified initial ascertainment rate. When we specified the initial ascertainment rate as 0.14 or 0.42, the estimated overall ascertainment rate was 0.08 (0.07–0.10) and 0.23 (0.16–0.28), respectively. If we assume an extreme scenario with no unascertained cases in the early outbreak (which we term model 'S8' (Supplementary Information)), the estimated ascertainment rate would be 0.47 (0.39–0.58) overall, which would represent an upper bound of the ascertainment rate. Because of the higher ascertainment rate (compared to the main analysis) in this model, we estimated a lower probability of resurgence (0.06) when lifting controls after 14 days of no ascertained cases, and the resurgence was expected to occur on day 38 (29–52) after lifting controls (Fig. 3). A simplified model that assumes complete ascertainment at any time performed substantially worse than the full model (Extended Data Table 4, Supplementary Information).

Understanding the proportion of unascertained cases and their transmissibility is critical for the prioritization of the surveillance and control measures[17]. Our finding of a large fraction of unascertained cases—despite the high level of surveillance in Wuhan—indicates the existence of many asymptomatic or mildly symptomatic individuals. It was previously estimated that asymptomatic individuals accounted for 18% of the infections on board the Diamond Princess Cruise ship[8] and 31% of the infected Japanese individuals who were evacuated from Wuhan[9]. In addition, in a cohort of 210 women admitted for delivery between 22 March and 4 April in New York City (USA), 29 of 33 (88%) pregnant women infected with SARS-CoV-2 were asymptomatic[10]. Several reports have also highlighted the difficulty of detecting cases of COVID-19: the detection capacity varied from 11% in low-surveillance

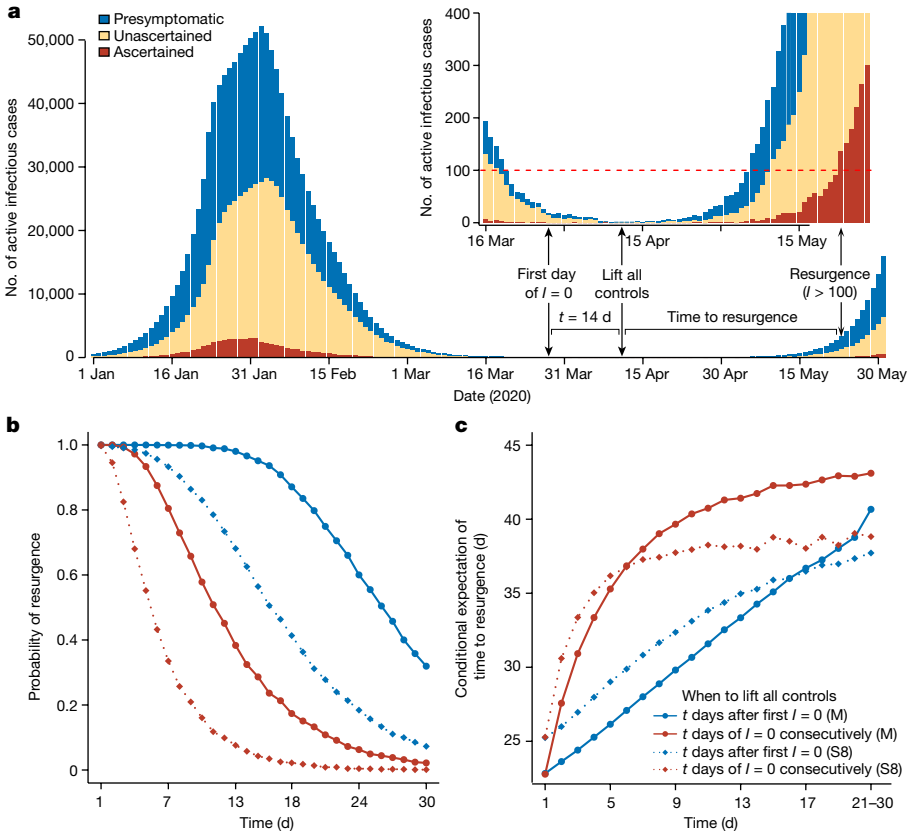


Fig. 3 | Risk of resurgence after lifting controls. We consider the main model (M) and the sensitivity analysis (S8) (Methods). In model M, we assume the initial ascertainment rate $r_0 = 0.23$, and thus an overall ascertainment rate of 0.13. In S8, we assume no unascertained cases initially and thus an overall ascertainment rate of 0.47. For each model, we simulated epidemic curves on the basis of 10,000 sets of parameters from MCMC, and set the transmission rate (b), ascertainment rate (r) and population movement (n) to their values in the first period after lifting controls. Resurgence was defined as reaching over 100 active ascertained infections. **a**, Illustration of a simulated curve under the main model, with control measures lifted 14 days after the first day of no ascertained cases. The inset is an enlarged plot from 16 March to 28 May. **b**, Probability of resurgence if control measures were lifted t days after the first day of no ascertained cases, or after observing zero ascertained cases for t days consecutively. **c**, Expectation of time to resurgence, conditional on the occurrence of resurgence. We grouped the final 10 days (t = 21 to 30) to calculate the expected time to resurgence because of their low probability of resurgence. Key applies to both **b** and **c**.

countries to 40% in high-surveillance countries[18,19], and the modelling of epidemics outside of Wuhan has suggested that the ascertainment rate was 24.4% in China (excluding Hubei province)[14] and 14% in Wuhan before the travel ban[15]. Consistent with these studies and emerging serological studies that show that seroprevalence is much higher than the reported case prevalence in cities and countries worldwide[20–22], our analyses of data from Wuhan indicated an overall ascertainment rate between 8% and 23% (Extended Data Table 6, excluding the extreme scenario of model S8).

Our R_e estimate of 3.54 (3.40–3.67) before any interventions is at the higher end of the range of the estimated R_0 values of other studies that used early epidemic data from Wuhan[6,23]. This discrepancy might be due to the modelling of unascertained cases, more-complete case records in our analysis and/or to the different time periods analysed. If we modelled from the first case of COVID-19 reported in Wuhan, we would estimate a lower R_e of 3.38 (3.28–3.48) before interventions (Extended Data Fig. 2), which remains much higher than those of SARS and MERS[4,5].

Our modelling study has delineated the full-spectrum dynamics of the COVID-19 outbreak in Wuhan, and highlighted two key features of the outbreak: high covertness and high transmissibility. These two features have synergistically propelled the COVID-19 pandemic, and imposed considerable challenges to attempts to control the outbreak. However, the Wuhan case study demonstrates the effectiveness of vigorous and multifaceted containment efforts. In particular, despite the relatively low ascertainment rates (owing to mild or absent symptoms of many infected individuals), the outbreak was controlled by interventions such as wearing face masks, social distancing and quarantining close contacts[1], which block transmission that stems from unascertained cases.

Given the limitations of our model as discussed below, further investigations—such as a survey of the seroprevalence of SARS-CoV-2-specific antibodies—are needed to confirm our estimates. First, owing to the delay in laboratory tests, we might have missed some cases and therefore underestimated the ascertainment rate (especially for the last period). Second, we excluded clinically diagnosed cases without laboratory confirmation to reduce false-positive diagnoses; however, this leads to an underestimation of ascertainment rates—especially for the third and fourth periods, during which many clinically diagnosed cases were reported[1]. The variation in the estimated ascertainment rates across periods reflects a combined effect of the evolving surveillance, interventions, medical resources and case definitions across time periods[1,24]. Third, our model assumes homogeneous transmission within the population and ignores heterogeneity between groups by sex, age, geographical region and socioeconomic status[25]. Furthermore, individual variation in infectiousness—such as superspreading events[26]—is known to result in a higher probability of stochastic extinction given a fixed population R_e (ref. [27]). We might therefore have overestimated the probability of resurgence. Finally, we could not evaluate the effect of individual interventions on the basis of an epidemic curve from a single city, because many interventions were applied

simultaneously. Future work that models heterogeneous transmission between different groups, and joint analysis with data from other cities, will provide deeper insights into the effectiveness of different control strategies[28,29].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2554-8.

1. Pan, A. et al. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *J. Am. Med. Assoc.* **323**, 1915–1923 (2020).
2. He, X. et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
3. Wu, J. T., Leung, K. & Leung, G. M. 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 (2020).
4. Wang, Y., Wang, Y., Chen, Y. & Qin, Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *J. Med. Virol.* **92**, 568–576 (2020).
5. Lipsitch, M. et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).
6. Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
7. Bai, Y. et al. Presumed asymptomatic carrier transmission of COVID-19. *J. Am. Med. Assoc.* **323**, 1406–1407 (2020).
8. Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* **25**, 2000180 (2020).
9. Nishiura, H. et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int. J. Infect. Dis.* **94**, 154–155 (2020).
10. Sutton, D., Fuchs, K., D'Alton, M. & Goffman, D. Universal screening for SARS-CoV-2 in women admitted for delivery. *N. Engl. J. Med.* **382**, 2163–2164 (2020).
11. Tong, Z. D. et al. Potential presymptomatic transmission of SARS-CoV-2, Zhejiang Province, China, 2020. *Emerg. Infect. Dis.* **26**, 1052–1054 (2020).
12. Ferretti, L. et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020).
13. Kucharski, A. J. et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558 (2020).
14. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
15. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
16. Wu, J. T. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506–510 (2020).
17. Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the epidemiology of COVID-19 – studies needed. *N. Engl. J. Med.* **382**, 1194–1196 (2020).
18. De Salazar, P. M., Niehus, R., Taylor, A., Buckee, C. O. & Lipsitch, M. Identifying locations with possible undetected imported severe acute respiratory syndrome coronavirus 2 cases by using importation predictions. *Emerg. Infect. Dis.* **26**, 1465–1469 (2020).
19. Niehus, R., De Salazar, P. M., Taylor, A. R. & Lipsitch, M. Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *Lancet Infect. Dis.* **20**, 803–808 (2020).
20. Levesque, J. & Maybury, D. W. A note on COVID-19 seroprevalence studies: a meta-analysis using hierarchical modelling. Preprint at https://doi.org/10.1101/2020.05.03.20089201 (2020).
21. To, K. K.-W. et al. Seroprevalence of SARS-CoV-2 in Hong Kong and in residents evacuated from Hubei province, China: a multicohort study. *Lancet Microbe* **1**, E111–E118 (2020).
22. Xu, X. et al. Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. *Nat. Med.* https://doi.org/10.1038/s41591-020-0949-6 (2020).
23. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, taaa021 (2020).
24. Tsang, T. K. et al. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. *Lancet Public Health* **5**, e289–e296 (2020).
25. Zhang, J. et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481–1486 (2020).
26. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* **395**, e47 (2020).
27. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
28. Tian, H. et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020).
29. Prem, K. et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* **5**, e261–e270 (2020).

Methods

Data of cases of COVID-19 in Wuhan

We analysed the daily incidence data of COVID-19, presented in figure 1 of ref. [1]. In brief, information on cases of COVID-19 from 8 December 2019 to 8 March 2020 were extracted from the municipal Notifiable Disease Report System on 9 March 2020. The date of the onset of symptoms (the self-reported date of developing symptoms, such as a fever, cough or other respiratory symptoms) and the date of confirmed diagnosis were collected. For the consistency of case definition throughout the periods, we included only 32,583 individuals who had a laboratory-confirmed positive test for SARS-CoV-2 by the real-time reverse-transcription polymerase-chain-reaction (RT–PCR) assay or high-throughput sequencing of nasal and pharyngeal swab specimens. SAS software (version 9.4) was used in data collection.

Estimation of initial ascertainment rate using cases exported to Singapore

As of 10 May 2020, a total of 24 confirmed cases of COVID-19 in Singapore were reported to be imported from China, among which 16 were imported from Wuhan before the cordon sanitaire on 23 January; the first case arrived in Singapore on 18 January (Extended Data Table 3). Based on VariFlight Data (https://data.variflight.com/en/), the total number of passengers who travelled from Wuhan to Singapore between 18 January and 23 January 2020 was 2,722. Therefore, the infection rate among these passengers was 0.59% (95% confidence interval 0.30–0.88%). These individuals had an onset of symptoms between 21 January and 30 January 2020. In Wuhan, a total of 12,433 confirmed cases involved individuals who were reported to have experienced an onset of symptoms in the same period—equivalent to a cumulative infection rate of 0.124% (95% confidence interval 0.122–0.126%), assuming a population size of 10 million for Wuhan. By further assuming complete ascertainment of early cases in Singapore (which is well-known for its high level of surveillance[18,19]), the ascertainment rate during the early outbreak in Wuhan was estimated to be 0.23 (95% confidence interval 0.14–0.42), corresponding to 0.77 (95% confidence interval 0.58–0.86) of the infections being unascertained. This represents a conservative estimate for two reasons: (1) the assumption of perfect ascertainment in Singapore ignored potential asymptomatic individuals;[8,9] and (2) the number of imported cases in which individuals experienced symptom onset between 21 January and 30 January was underestimated owing to the suspension of flights after lockdown in Wuhan. Without direct information to estimate the initial ascertainment rate before 1 January 2020, we used these results based on Singapore data to set the initial value and the prior distribution of ascertainment rates in our model, and performed sensitivity analyses under various assumptions.

The SAPHIRE model

We extended the classic SEIR model to a SAPHIRE model (Fig. 1, Extended Data Table 1), which incorporates three additional compartments to account for presymptomatic infectious individuals (P), unascertained cases (A) and cases isolated in the hospital (H). We chose to analyse data from 1 January 2020, when the Huanan Seafood Market was disinfected, and thus did not model the zoonotic force of infection[3]. We assumed a constant population size (N) = 10,000,000, with equal numbers of daily inbound and outbound travellers (n), in which n = 500,000 for 1–9 January, 800,000 for 10–22 January (owing to *Chunyun*) and 0 after the cordon sanitaire from 23 January[3]. We divided the population into susceptible (S), exposed (E), P, A, ascertained infectious (I), H and removed (R) individuals. We introduced compartment H because ascertained cases would have a shorter effective infectious period owing to isolation, especially when medical resources were improved[1]. We use italicized letters to denote the number of individuals in each corresponding compartment. The dynamics of these compartments across time (t) are described by the following set of ordinary differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = n - \frac{bS(\alpha P + \alpha A + I)}{N} - \frac{nS}{N} \tag{1}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{bS(\alpha P + \alpha A + I)}{N} - \frac{E}{D_\mathrm{e}} - \frac{nE}{N} \tag{2}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{E}{D_\mathrm{e}} - \frac{P}{D_\mathrm{p}} - \frac{nP}{N} \tag{3}$$

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{(1-r)P}{D_\mathrm{p}} - \frac{A}{D_\mathrm{i}} - \frac{nA}{N} \tag{4}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{rP}{D_\mathrm{p}} - \frac{I}{D_\mathrm{i}} - \frac{I}{D_\mathrm{q}} \tag{5}$$

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{I}{D_\mathrm{q}} - \frac{H}{D_\mathrm{h}} \tag{6}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{A+I}{D_\mathrm{i}} + \frac{H}{D_\mathrm{h}} - \frac{nR}{N} \tag{7}$$

in which b is the transmission rate for ascertained cases (defined as the number of individuals that an ascertained case can infect per day); α is the ratio of the transmission rate of unascertained cases to that of ascertained cases; r is ascertainment rate; D_e is the latent period; D_p is the presymptomatic infectious period; D_i is the symptomatic infectious period; D_q is the duration from illness onset to isolation; and D_h is the isolation period in hospital. R_e could be computed as

$$R_\mathrm{e} = \alpha b\left(D_\mathrm{p}^{-1} + \frac{n}{N}\right)^{-1} + (1-r)\alpha b\left(D_\mathrm{i}^{-1} + \frac{n}{N}\right)^{-1} + rb\left(D_\mathrm{i}^{-1} + D_\mathrm{q}^{-1}\right)^{-1} \tag{8}$$

in which the three terms represent infections contributed by presymptomatic individuals, unascertained cases and ascertained cases, respectively. We adjusted the infectious periods of each type of case by taking population movement $\left(\frac{n}{N}\right)$ and isolation (D_q^{-1}) into account.

Parameter settings and initial states

Parameter settings for the main analysis are summarized in Extended Data Table 2. We set α = 0.55 according to ref. [15], assuming lower transmissibility for unascertained cases. Compartment P contains both ascertained and unascertained cases in the presymptomatic phase. We set the transmissibility of P to be the same as unascertained cases, because it has previously been reported that the majority of cases are unascertained[15]. We assumed an incubation period of 5.2 days and a presymptomatic infectious period of D_p = 2.3 days[2,6]. Thus, the latent period was D_e = 5.2 − 2.3 = 2.9 days. Because presymptomatic infectiousness was estimated to account for 44% of the total infections from ascertained cases[2], we set the mean of total infectious period as $(D_\mathrm{p} + D_\mathrm{i}) = \frac{D_\mathrm{p}}{0.44}$ = 5.2 days, assuming constant infectiousness across the presymptomatic and symptomatic phases of ascertained cases[12]—thus, the mean symptomatic infectious period was D_i = 2.9 days. We set a long isolation period of D_h = 30 days, but this parameter has no effect on our fitting procedure and the final parameter estimates. The duration from the onset of symptoms to isolation was estimated to be D_q = 21, 15, 10, 6 and 3 days as the median time length from onset to confirmed diagnosis in period 1–5, respectively[1].

On the basis of the settings above, we specified the initial state of the model on 31 December 2019 (Extended Data Table 1). The initial

number of ascertained symptomatic cases $I(0)$ was specified as the number of ascertained cases in which individuals experienced symptom onset during 29–31 December 2019. We assumed the initial ascertainment rate was r_0, and thus the initial number of unascertained cases was $A(0) = r_0^{-1}(1-r_0)I(0)$. We denoted $P_I(0)$ and $E_I(0)$ as the numbers of ascertained cases in which individuals experienced symptom onset during 1–2 January 2020 and 3–5 January 2020, respectively. Then, the initial numbers of exposed and presymptomatic individuals were set as $E(0) = r_0^{-1}E_I(0)$ and $P(0) = r_0^{-1}P_I(0)$, respectively. We assumed $r_0 = 0.23$ in our main analysis, on the basis of the point estimate using the Singapore data (described in 'Estimation of initial ascertainment rate using cases exported to Singapore').

Estimation of parameters in the SAPHIRE model

Considering the time-varying strength of control measures, we assumed $b = b_{12}$ and $r = r_{12}$ for the first two periods, $b = b_3$ and $r = r_3$ for period 3, $b = b_4$ and $r = r_4$ for period 4, and $b = b_5$ and $r = r_5$ for period 5. We assumed that the observed number of ascertained cases in which individuals experienced symptom onset on day d—denoted as x_d—follows a Poisson distribution with rate $\lambda_d = rP_{d-1}D_p^{-1}$, in which P_{d-1} is the expected number of presymptomatic individuals on day $(d-1)$. We fit the observed data from 1 January to 29 February ($d = 1, 2, \ldots, D$, and $D = 60$) and used the fitted model to predict the trend from 1 March to 8 March. Thus, the likelihood function is

$$L(b_{12}, b_3, b_4, b_5, r_{12}, r_3, r_4, r_5) = \prod_{d=1}^{D} \frac{e^{-\lambda_d}\lambda_d^{x_d}}{x_d!} \quad (9)$$

We estimated b_{12}, b_3, b_4, b_5, r_{12}, r_3, r_4 and r_5 by MCMC with the delayed rejection adaptive metropolis algorithm implemented in the R package BayesianTools (version 0.1.7)[30]. We used a non-informative flat prior of Unif(0,2) for b_{12}, b_3, b_4 and b_5. For r_{12}, we used an informative prior of Beta(7.3,24.6) by matching the first two moments of the estimate using Singapore data (described in 'Estimation of initial ascertainment rate using cases exported to Singapore'). We reparameterized r_3, r_4 and r_5 by

$$\text{logit}(r_3) = \text{logit}(r_{12}) + \delta_3$$

$$\text{logit}(r_4) = \text{logit}(r_3) + \delta_4$$

$$\text{logit}(r_5) = \text{logit}(r_4) + \delta_5$$

in which $\text{logit}(r) = \log\left(\frac{r}{1-r}\right)$. In the MCMC, we sampled δ_3, δ_4 and δ_5 from the prior of $N(0,1)$. We set a burn-in period of 40,000 iterations and continued to run 100,000 iterations with a sampling step size of 10 iterations. We repeated MCMC with three different sets of initial values and assessed the convergence by the trace plot and the multivariate Gelman–Rubin diagnostic[31] (Supplementary Information). Estimates of parameters were presented as posterior means and 95% credible intervals from 10,000 MCMC samples. All of the analyses were performed in R (version 3.6.2) and the Gelman–Rubin diagnostic was calculated using the gelman.diag function in the R package coda (version 0.19.3).

Stochastic simulations

We used stochastic simulations to obtain the 95% credible interval of a fitted or predicted epidemic curve. Given a set of parameter values from MCMC, we performed the following multinomial random sampling:

$$(U_{S\to E}, U_{S\to O}, U_{S\to S}) \sim \text{Multinomial}(S_{t-1}; p_{S\to E}, p_O, 1-p_{S\to E}-p_O)$$

$$(U_{E\to P}, U_{E\to O}, U_{E\to E}) \sim \text{Multinomial}(E_{t-1}; p_{E\to P}, p_O, 1-p_{E\to P}-p_O)$$

$$(U_{P\to I}, U_{P\to A}, U_{P\to O}, U_{P\to P}) \sim \text{Multinomial}(P_{t-1}; p_{P\to I}, p_{P\to A}, p_O, 1-p_{P\to I}-p_{P\to A}-p_O)$$

$$(U_{I\to H}, U_{I\to R}, U_{I\to I}) \sim \text{Multinomial}(I_{t-1}; p_{I\to H}, p_{I\to R}, 1-p_{I\to H}-p_{I\to R})$$

$$(U_{A\to R}, U_{A\to O}, U_{A\to A}) \sim \text{Multinomial}(A_{t-1}; p_{A\to R}, p_O, 1-p_{A\to R}-p_O)$$

$$(U_{H\to R}, U_{H\to H}) \sim \text{Multinomial}(H_{t-1}; p_{H\to R}, 1-p_{H\to R})$$

$$(U_{R\to O}, U_{R\to R}) \sim \text{Multinomial}(R_{t-1}; p_O, 1-p_O)$$

in which O denotes the status of outflow population, $p_O = nN^{-1}$ denotes the outflow probability and other quantities are status transition probabilities, including $p_{S\to E} = b(\alpha P_{t-1} + \alpha A_{t-1} + I_{t-1})N^{-1}$, $p_{E\to P} = D_e^{-1}$, $p_{P\to I} = rD_p^{-1}$, $p_{P\to A} = (1-r)D_p^{-1}$, $p_{I\to H} = D_q^{-1}$, $p_{I\to R} = p_{A\to R} = D_i^{-1}$ and $p_{H\to R} = D_h^{-1}$. The SAPHIRE model described by equations (1)–(7) is equivalent to the following stochastic dynamics:

$$S_t - S_{t-1} = n - U_{S\to E} - U_{S\to O} \quad (10)$$

$$E_t - E_{t-1} = U_{S\to E} - U_{E\to P} - U_{E\to O} \quad (11)$$

$$P_t - P_{t-1} = U_{E\to P} - U_{P\to A} - U_{P\to I} - U_{P\to O} \quad (12)$$

$$A_t - A_{t-1} = U_{P\to A} - U_{A\to R} - U_{A\to O} \quad (13)$$

$$I_t - I_{t-1} = U_{P\to I} - U_{I\to H} - U_{I\to R} \quad (14)$$

$$H_t - H_{t-1} = U_{I\to H} - U_{H\to R} \quad (15)$$

$$R_t - R_{t-1} = U_{A\to R} + U_{I\to R} + U_{H\to R} - U_{R\to O} \quad (16)$$

We repeated the stochastic simulations for all 10,000 sets of parameter values sampled by MCMC to construct the 95% credible interval of the epidemic curve by the 2.5 and 97.5 percentiles at each time point.

Prediction of epidemic ending date and the risk of resurgence

Using the stochastic simulations described in 'Stochastic simulations', we predicted the first day of no new ascertained cases and the date of clearance of all active infections in Wuhan, assuming continuation of the same control measures as the last period (that is, same parameter values).

We also evaluated the risk of outbreak resurgence after lifting control measures. We considered lifting all controls (1) at t days after the first day of zero ascertained cases, or (2) after a consecutive period of t days with no ascertained cases. After lifting controls, we set the transmission rate b, ascertainment rate r and population movement n to be the same as the first period, and continued the stochastic simulation to the stationary state. Time to resurgence was defined as the number of days from lifting controls to when the number of active ascertained cases (I) reached 100. We performed 10,000 simulations with 10,000 sets of parameter values sampled from MCMC (as described in 'Estimation of parameters in the SAPHIRE model'). We calculated the probability of resurgence as the proportion of simulations in which resurgence occurred, as well as the time to resurgence conditional on the occurrence of resurgence.

Simulation study for method validation

To validate the method, we performed two-period stochastic simulations (equations (10) to (16)) with transmission rate $b = b_1 = 1.27$, ascertainment rate $r = r_1 = 0.2$, daily population movement $n = 500{,}000$, and duration from illness onset to isolation $D_q = 20$ days for the first period (so that $R_e = 3.5$ according to equation (8)), and $b = b_2 = 0.41$, $r = r_2 = 0.4$, $n = 0$ and $D_q = 5$ for the second period (so that $R_e = 1.2$ according to equation (8)). Lengths of both periods were set to 15 days, and the

initial ascertainment rate was set to $r_0 = 0.3$, and the other parameters and initial states were set as those in our main analysis (Extended Data Tables 1, 2). We repeated stochastic simulations 100 times to generate 100 datasets. For each dataset, we applied our MCMC method to estimate b_1, b_2, r_1 and r_2, and set all other parameters and initial values to be the same as the true values. We translated b_1 and b_2 into $(R_e)_1$ and $(R_e)_2$ according to equation (8), and focused on evaluating the estimates of $(R_e)_1$, $(R_e)_2$, r_1 and r_2. We also tested the robustness to misspecification of the latent period D_e, presymptomatic infectious period D_p, symptomatic infectious period D_i, duration from illness onset to isolation D_q, ratio of transmissibility between unascertained and ascertained cases α, and initial ascertainment rate r_0. In each test, we changed the specified value of a parameter (or initial state) to be 20% lower or higher than its true value, and kept all other parameters unchanged. When we changed the value of r_0, we adjusted the initial states $A(0)$, $P(0)$ and $E(0)$ according to Extended Data Table 1.

For each simulated dataset, we ran the MCMC method with 20,000 burn-in iterations and an additional 30,000 iterations. We sampled parameter values from every 10 iterations, resulting in 3,000 MCMC samples. We took the mean across 3,000 MCMC samples as the final estimates and display results for 100 repeated simulations.

Sensitivity analyses for the real data

We designed nine sensitivity analyses to test the robustness of our results from real data. For each of the sensitivity analyses, we fixed parameters and initial states to be the same as the main analysis except for those mentioned below. For analysis (S1), we adjust the reported incidences from 29 January to 1 February to their average. We suspect the spike of incidences on 1 February might be caused by approximate-date records among some patients admitted to the field hospitals after 2 February. The actual dates for illness onset for these patients were likely to be spread between 29 January and 1 February. For analysis (S2), we assume an incubation period of 4.1 days (lower 95% confidence interval from ref. [6]) and a presymptomatic infectious period of 1.1 days (the lower 95% confidence interval from ref. [2] is 0.8 days, but our discrete stochastic model requires $D_p > 1$), equivalent to set $D_e = 3$ and $D_p = 1.1$, and adjust $P(0)$ and $E(0)$ accordingly. For analysis (S3), we assume an incubation period of 7 days (upper 95% confidence interval from ref. [6]) and a presymptomatic infectious period of 3 days (upper 95% confidence interval from ref. [2]), equivalent to set $D_e = 4$ and $D_p = 3$, and adjust $P(0)$ and $E(0)$ accordingly. For analysis (S4), we assume the transmissibility of the unascertained cases is $\alpha = 0.46$ (lower 95% confidence interval from ref. [15]) of the ascertained cases. For analysis (S5), we assume the transmissibility of the unascertained cases is $\alpha = 0.62$ (upper 95% confidence interval from ref. [15]) of the ascertained cases. For analysis (S6), we assume the initial ascertainment rate is $r_0 = 0.14$ (lower 95% confidence interval of the estimate using Singapore data) and adjust $A(0)$, $P(0)$ and $E(0)$ accordingly. For analysis (S7), we assume the initial ascertainment rate is $r_0 = 0.42$ (upper 95% confidence interval of the estimate using Singapore data) and adjust $A(0)$, $P(0)$ and $E(0)$ accordingly. For analysis (S8), we assume the initial ascertainment rate is $r_0 = 1$ (theoretical upper limit) and adjust $A(0)$, $P(0)$ and $E(0)$ accordingly. For analysis (S9), we assume no unascertained cases by fixing $r_0 = r_{12} = r_3 = r_4 = r_5 = 1$. We compared this simplified model to the full model using the Bayes factor.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data analysed in this study are available on GitHub at https://github.com/chaolongwang/SAPHIRE.

Code availability

Codes are available on GitHub at https://github.com/chaolongwang/SAPHIRE.

30. Haario, H., Laine, M., Mira, A. & Saksman, E. DRAM: efficient adaptive MCMC. *Stat. Comput.* **16**, 339–354 (2006).
31. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434–455 (1997).

Acknowledgements We thank H. Tian from Beijing Normal University for comments. This study was supported by the National Natural Scientific Foundation of China (91843302), the Fundamental Research Funds for the Central Universities (2019kfyXMBZ015), and the 111 Project (X.H., S.C., D.W., C.W., T.W.). X.L. is supported by Harvard University.

Author contributions T.W., X.L. and C.W. designed the study. X.H., S.C., X.L. and C.W. developed statistical methods. X.H., S.C. and D.W. performed data analysis. C.W. wrote the first draft of the manuscript. All authors reviewed and edited the manuscript.

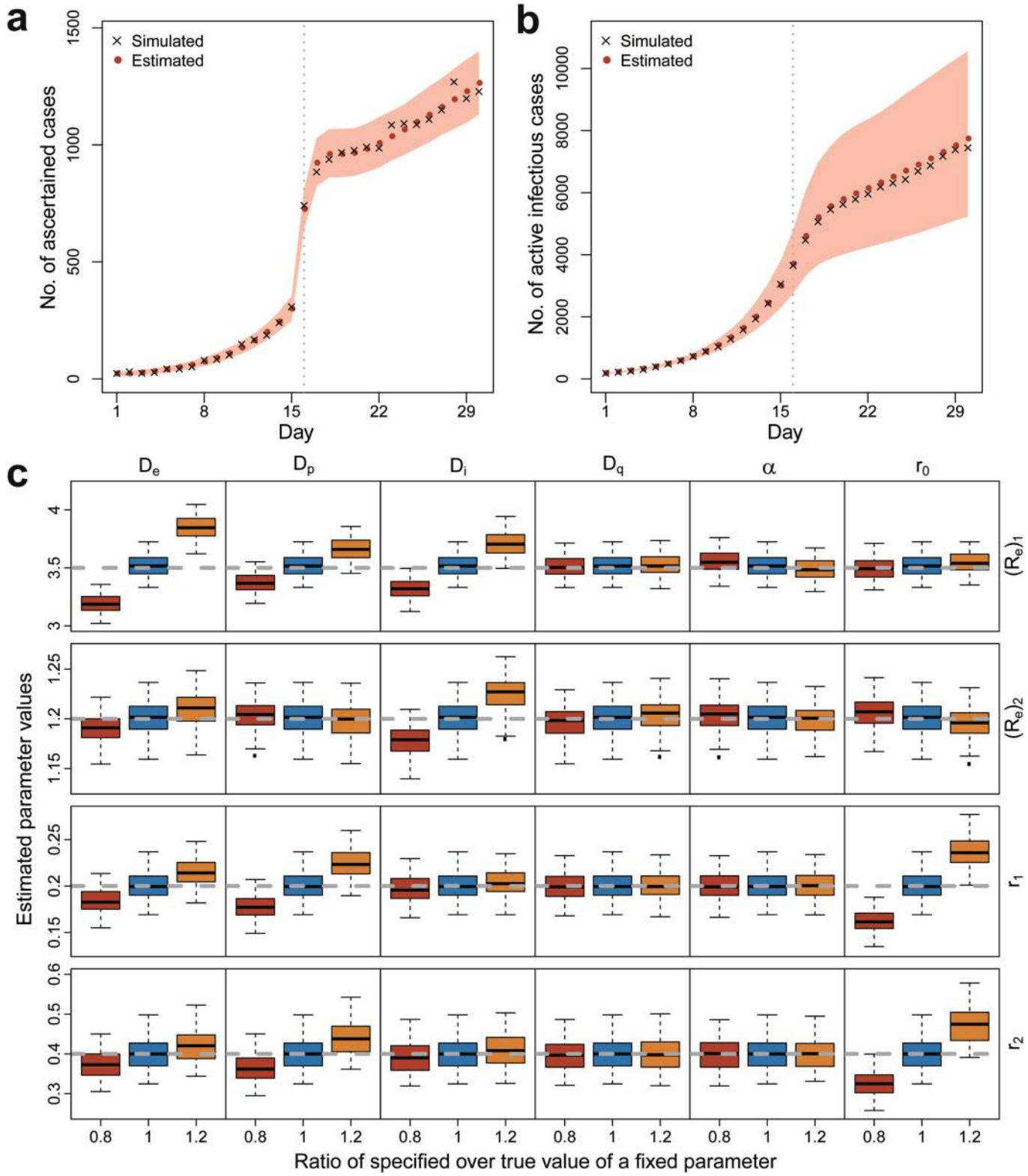
Competing interests The authors declare no competing interests.

Additional information

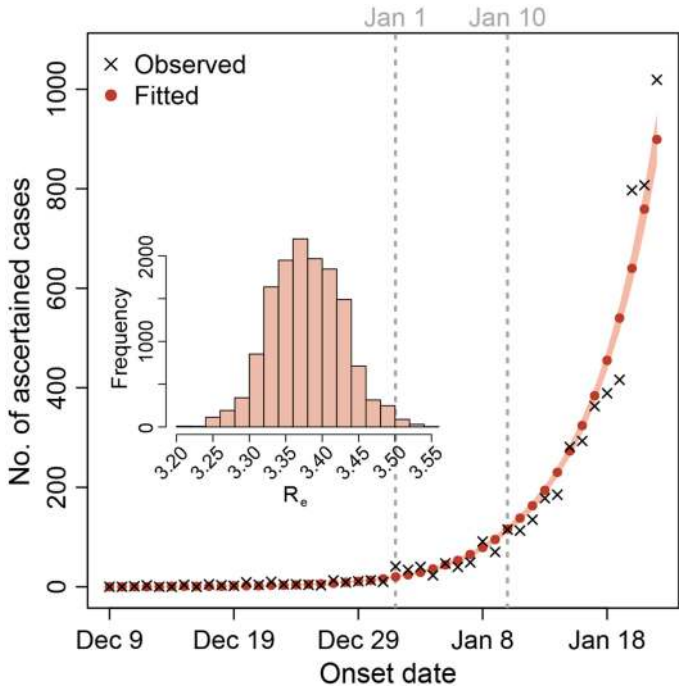
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2554-8.

Correspondence and requests for materials should be addressed to T.W., X.L. or C.W.

Peer review information *Nature* thanks David Fisman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Evaluation of the method on simulated data with two periods. a, **b**, Illustration of one simulated dataset. We estimated b_1, b_2, r_1 and r_2 when the other parameters were specified to their true values. The red points represent the mean estimates and the shaded areas indicate 95% credible intervals from 3,000 MCMC samples. **c**, Summary of results from 100 simulations. Each row represents an estimated parameter as indicated on the right, including $(R_e)_1$, $(R_e)_2$, r_1 and r_2. The grey dashed line in each row represents the true value of the parameter to be estimated. Each column represents a specified parameter as indicated on the top, including D_e, D_p, D_i, D_q, α and r_0, which we specified as the true values or 20% lower or higher than the true values. Each box summarizes estimates from 100 replicates, of which the median is indicated by the horizontal line, the interquartile range is indicated by the lower and upper bounds, and the minimum and the maximum are indicated by the whiskers.



Extended Data Fig. 2 | Estimation of R_0 using daily incidence data, starting from 9 December. Following the main analysis, we assumed $r_0 = 0.23$ and set $I(0) = 1$, $A(0) = 3$, $E(0) = 17$ and $P(0) = H(0) = R(0) = 0$ accordingly. We assumed transmission rate b, ascertainment rate r and duration from illness onset to hospitalization D_q (set to 21 days) were the same until 22 January 2020. All the other settings were the same as in the main analysis. The shaded area in the plot indicates 95% credible intervals estimated by the deterministic model with 10,000 sets of parameter values sampled from MCMC. Unlike other analyses, we did not construct 95% credible intervals by stochastic simulations, because stochastic variation at the early days had very large effects, owing to low counts. The inserted histogram shows the distribution of the estimated R_0 from 9 December 2019 to 22 January 2020, for which the mean estimate was 3.38 (95% credible interval 3.28–3.48).

Extended Data Table 1 | Notations of compartments and the initial values for the main analysis

Notation	Meaning	Initial value	Note
S	Number of susceptible individuals	9,999,021	$S = N - E - P - A - I - H - R$
E	Number of exposed cases	478	$E(0) = r_0^{-1}E_1(0)$, where $E_1(0)$ was the number of ascertained cases with onset during day $(D_p + 1)$ and day $(D_p + D_e)$ (Jan 3-5, 2020)*
P	Number of presymptomatic cases	326	$P(0) = r_0^{-1}P_1(0)$, where $P_1(0)$ was the number of ascertained cases with onset during day 1 and day D_p (Jan 1-2, 2020)*
I	Number of ascertained cases	34	Number of ascertained cases with onset within D_i days before day 1 (Dec 29-31, 2019)
A	Number of unascertained cases	114	$A(0) = r_0^{-1}(1 - r_0)I(0)$*
H	Number of isolated cases	27	Number of cases reported before day 1 (Jan 1, 2020)
R	Number of removed individuals	0	Number of cases recovered before day 1 (Jan 1, 2020)

*The initial ascertainment rate r_0 was assumed to be 0.23 in the main analysis. Day 1 is 1 January 2020.

Extended Data Table 2 | Parameter settings for five periods in the main analysis

Parameter	Meaning	Jan 1-9	Jan 10-22	Jan 23-Feb 1	Feb 2-16	Feb 17-Mar 8
b	Transmission rate of ascertained cases	b_{12}	b_{12}	b_3	b_4	b_5
r	Ascertainment rate	r_{12}	r_{12}	r_3	r_4	r_5
α	Ratio of transmission rate for unascertained over ascertained cases	0.55	0.55	0.55	0.55	0.55
D_e	Latent period	2.9	2.9	2.9	2.9	2.9
D_p	Presymptomatic infectious period	2.3	2.3	2.3	2.3	2.3
D_i	Symptomatic infectious period	2.9	2.9	2.9	2.9	2.9
D_q	Duration from illness onset to isolation	21	15	10	6	3
D_h	Isolation period	30	30	30	30	30
N	Population size	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000
n	Daily inbound and outbound size	500,000	800,000	0	0	0

Extended Data Table 3 | COVID-19 cases exported from Wuhan to Singapore before 23 January 2020

Case ID	Arrival date	Symptom onset	Confirmed date
1	2020/1/20	2020/1/21	2020/1/23
2	2020/1/21	2020/1/21	2020/1/24
3	2020/1/20	2020/1/23	2020/1/24
4	2020/1/22	2020/1/23	2020/1/25
5	2020/1/18	2020/1/24	2020/1/27
6	2020/1/19	2020/1/25	2020/1/27
7	2020/1/23	2020/1/24	2020/1/27
8	2020/1/19	2020/1/24	2020/1/28
9	2020/1/19	2020/1/24	2020/1/29
10	2020/1/20	2020/1/21	2020/1/29
11	2020/1/22	2020/1/27	2020/1/29
12	2020/1/22	2020/1/26	2020/1/29
13	2020/1/21	2020/1/28	2020/1/30
16	2020/1/22	2020/1/23	2020/1/31
18	2020/1/22	2020/1/30	2020/2/1
26	2020/1/21	2020/1/28	2020/2/4

This information is from https://co.vid19.sg/singapore/dashboard.

Extended Data Table 4 | Estimated transmission rates from the main and sensitivity analyses

Analysis	Estimated transmission rates*				DIC†
	b_{12}	b_3	b_4	b_5	
Main	1.31 (1.25-1.37)	0.40 (0.38-0.43)	0.17 (0.16-0.19)	0.10 (0.08-0.12)	554.07
S1	1.31 (1.25-1.37)	0.37 (0.35-0.39)	0.17 (0.16-0.18)	0.10 (0.08-0.11)	387.63
S2	1.51 (1.43-1.57)	0.53 (0.51-0.56)	0.25 (0.24-0.27)	0.15 (0.13-0.17)	539.15
S3	1.46 (1.39-1.53)	0.34 (0.31-0.37)	0.11 (0.10-0.13)	0.04 (0.02-0.06)	588.73
S4	1.53 (1.46-1.61)	0.47 (0.44-0.50)	0.21 (0.19-0.22)	0.11 (0.09-0.13)	554.57
S5	1.18 (1.12-1.24)	0.36 (0.34-0.38)	0.16 (0.15-0.17)	0.09 (0.07-0.10)	553.49
S6	1.34 (1.28-1.39)	0.41 (0.38-0.44)	0.18 (0.17-0.19)	0.10 (0.08-0.12)	555.08
S7	1.27 (1.21-1.33)	0.39 (0.36-0.41)	0.17 (0.16-0.18)	0.10 (0.08-0.11)	555.40
S8	1.20 (1.14-1.27)	0.36 (0.34-0.39)	0.17 (0.16-0.18)	0.10 (0.08-0.12)	595.58
S9	0.93 (0.92-0.94)	0.26 (0.25-0.27)	0.17 (0.16-0.17)	0.16 (0.14-0.18)	808.38

*The estimates are displayed as mean (95% credible interval) based on 10,000 MCMC samples.

†The deviance information criterion (DIC) is presented for model comparison. Nevertheless, the DIC of S1 is not comparable to the others because the data of S1 were modified by smoothing the outlier data point on 1 February.

Extended Data Table 5 | Estimated R_e for different periods from the main and sensitivity analyses

	Estimated effective reproduction number R_e*				
Analysis	Jan 1-9	Jan 10-22	Jan 23-Feb 1	Feb 2-16	Feb 17-Mar 8
Main	3.54 (3.40-3.67)	3.32 (3.19-3.44)	1.18 (1.11-1.25)	0.51 (0.47-0.54)	0.28 (0.23-0.33)
S1	3.54 (3.40-3.67)	3.32 (3.19-3.44)	1.09 (1.02-1.16)	0.50 (0.47-0.54)	0.28 (0.23-0.32)
S2	3.21 (3.09-3.32)	3.03 (2.92-3.13)	1.23 (1.16-1.29)	0.57 (0.54-0.60)	0.33 (0.29-0.37)
S3	4.37 (4.19-4.55)	4.07 (3.91-4.24)	1.13 (1.04-1.22)	0.38 (0.34-0.41)	0.14 (0.08-0.19)
S4	3.56 (3.42-3.68)	3.34 (3.21-3.45)	1.18 (1.11-1.25)	0.51 (0.47-0.54)	0.27 (0.23-0.32)
S5	3.52 (3.39-3.66)	3.30 (3.18-3.43)	1.18 (1.11-1.25)	0.51 (0.47-0.54)	0.27 (0.23-0.32)
S6	3.52 (3.38-3.65)	3.29 (3.17-3.42)	1.19 (1.12-1.27)	0.51 (0.48-0.55)	0.28 (0.23-0.33)
S7	3.59 (3.46-3.72)	3.38 (3.26-3.49)	1.17 (1.10-1.24)	0.50 (0.47-0.53)	0.27 (0.23-0.32)
S8	3.79 (3.68-3.90)	3.58 (3.48-3.68)	1.15 (1.08-1.22)	0.50 (0.47-0.53)	0.27 (0.23-0.32)
S9	3.42 (3.40-3.45)	3.25 (3.23-3.27)	0.92 (0.88-0.95)	0.54 (0.51-0.56)	0.44 (0.38-0.49)

*The estimates are displayed as mean (95% credible interval) based on 10,000 MCMC samples.

Extended Data Table 6 | Estimated ascertainment rates from the main and sensitivity analyses

Analysis	Estimated ascertainment rate*				
	r_{12}	r_3	r_4	r_5	Overall
Main	0.15 (0.13-0.17)	0.14 (0.11-0.17)	0.10 (0.08-0.12)	0.16 (0.13-0.21)	0.13 (0.11-0.16)
S1	0.15 (0.12-0.17)	0.15 (0.12-0.18)	0.11 (0.09-0.14)	0.19 (0.14-0.24)	0.13 (0.11-0.16)
S2	0.14 (0.12-0.17)	0.15 (0.12-0.18)	0.10 (0.08-0.13)	0.17 (0.13-0.22)	0.14 (0.11-0.17)
S3	0.14 (0.12-0.16)	0.13 (0.10-0.16)	0.09 (0.07-0.11)	0.16 (0.12-0.20)	0.12 (0.10-0.15)
S4	0.15 (0.13-0.17)	0.14 (0.12-0.17)	0.10 (0.08-0.12)	0.17 (0.13-0.21)	0.13 (0.11-0.16)
S5	0.15 (0.13-0.17)	0.14 (0.11-0.17)	0.10 (0.08-0.12)	0.16 (0.12-0.21)	0.13 (0.11-0.16)
S6	0.09 (0.08-0.10)	0.09 (0.07-0.11)	0.06 (0.05-0.08)	0.10 (0.08-0.13)	0.08 (0.07-0.10)
S7	0.26 (0.22-0.30)	0.25 (0.20-0.30)	0.18 (0.14-0.22)	0.29 (0.22-0.37)	0.23 (0.16-0.28)
S8	0.55 (0.47-0.62)	0.50 (0.41-0.60)	0.35 (0.28-0.43)	0.59 (0.46-0.74)	0.47 (0.39-0.58)

*The estimates are displayed as mean (95% credible intervals) based on 10,000 MCMC samples.

Extended Data Table 7 | Prediction of the ending date of COVID-19 epidemic in Wuhan from the main and sensitivity analyses

Analysis	First day of no ascertained infections*	Clearance of all infections†
Main	Mar 27 (Mar 20 to Apr 5)‡	Apr 21 (Apr 8 to May 12)
S1	Mar 27 (Mar 20 to Apr 4)	Apr 20 (Apr 7 to May 11)
S2	Mar 28 (Mar 21 to Apr 5)	Apr 22 (Apr 8 to May 13)
S3	Mar 25 (Mar 18 to Apr 2)	Apr 19 (Apr 5 to May 8)
S4	Mar 27 (Mar 20 to Apr 4)	Apr 21 (Apr 8 to May 12)
S5	Mar 27 (Mar 20 to Apr 4)	Apr 21 (Apr 8 to May 13)
S6	Mar 27 (Mar 20 to Apr 4)	Apr 24 (Apr 11 to May 15)
S7	Mar 27 (Mar 20 to Apr 4)	Apr 17 (Apr 4 to May 7)
S8	Mar 26 (Mar 19 to Apr 4)	Apr 10 (Mar 29 to Apr 30)
S9	Apr 5 (Mar 26 to Apr 18)	Apr 20 (Apr 4 to May 16)

*First day of no ascertained infections means the first day of $I = 0$.

†Clearance of all infections means the first day of $E = P = A = I = 0$.

‡The estimates are displayed as mean date (95% credible interval) based on 10,000 stochastic simulations with parameter values from MCMC sampling.

nature research

Corresponding author(s): Chaolong Wang, Xihong Lin, Tangchun Wu

Last updated by author(s): Jun 27, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SAS version 9.4 was used for data collection.
Data analysis	Data analysis was performed in R (version 3.6.2), alongside with third-party R packages BayesianTools (version 0.1.7) and coda (version 0.19.3). R codes are available on Github via link: https://github.com/chaolongwang/SAPHIRE.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data are available on Github via link: https://github.com/chaolongwang/SAPHIRE.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study contained 32,583 laboratory-confirmed COVID-19 cases between 18 Dec 2019 and 8 Mar 2020 in Wuhan. We included all laboratory-confirmed COVID-19 cases that were reported by 8 Mar 2020.
Data exclusions	For the consistency of case definition throughout different time periods, we excluded COVID-19 cases diagnosed by clinical symptoms without SARS-CoV-2 virus confirmation by the real-time reverse-transcription-polymerase-chain-reaction (RT-PCR) assay or high-throughput sequencing of nasal and pharyngeal swab specimens. Exclusion criterion was pre-established.
Replication	Not applicable because this study is retrospective and observational.
Randomization	Not applicable because this study is retrospective and observational.
Blinding	Not applicable because this study is retrospective and observational.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging