Accelerated Article Preview

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

Xingjie Hao, Shanshan Cheng, Degang Wu, Tangchun Wu, Xihong Lin & Chaolong Wang

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Article

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

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

As countries in the world review interventions for containing the COVID-19 pandemic, important lessons can be drawn by studying the full transmission dynamics of SARS-CoV-2 in Wuhan, China, where vigorous non-pharmaceutical interventions have suppressed the local COVID-19 outbreak[1]. Here, we use a modelling approach to reconstruct the full-spectrum dynamics of COVID-19 between January 1, 2020 and March 8, 2020 across five periods marked by events and interventions based on 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 March 8 were unascertained, potentially including asymptomatic and mild-symptomatic cases; and a basic reproduction number R_0 of 3.54 (95% credible interval [CrI]: 3.40-3.67) in the early outbreak, much higher than for SARS and MERS[4,5]. We observe that multi-pronged interventions had considerable positive effects on controlling the outbreak, decreasing the reproduction number to 0.28 (0.23-0.33) and by projection reducing the total infections in Wuhan by 96.0% as of March 8. We furthermore explore the probability of resurgence following lifting of all interventions after 14 days of no ascertained infections, estimating it at 0.32 and 0.06 based on models with 87% and 53% unascertained infections, respectively, highlighting the risk posed by unascertained cases in changing intervention strategies. These results provide important implications for continuing surveillance and interventions to eventually contain COVID-19 outbreaks.

The coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 was detected in Wuhan, China, in December 2019[6]. The high population density together with the increased social activities before the Chinese New Year catalyzed the outbreak, of which the spread was expedited by the massive human movement during the holiday travel season *Chunyun* from January 10, 2020[3]. Shortly after the confirmation of human-to-human transmission, the Chinese authorities implemented the unprecedented *cordons sanitaire* of Wuhan on January 23 to contain the geographic spread, followed by a series of non-pharmaceutical interventions to reduce virus transmission, including suspension of all intra- and inter-city transportation, compulsory mask wearing in public places, cancelation of social gatherings, and home quarantine of mild-symptomatic patients[1]. From February 2, strict stay-at-home policy for all residents, centralized isolation of all patients, and centralized quarantine of suspected cases and close contacts were implemented to stop household and community transmission. Furthermore, a city-wide door-to-door universal symptom survey was carried out during February 17-19 by designated community workers to identify previously undetected symptomatic cases. These drastic interventions, together with the improved medical resources and healthcare manpower from all over the country, have effectively crushed the epidemic curve and reduced the attack rate in Wuhan, shedding light on the global efforts to control the COVID-19 outbreak[1].

Recent studies have revealed important transmission features of COVID-19, including infectiousness of asymptomatic cases[7–10] and presymptomatic cases[2,11,12]. Furthermore, the number of ascertained cases was much smaller than that estimated using international cases exported from Wuhan prior to the travel suspension[3,13,14], implying a substantial number of unascertained cases. Using reported cases from 375 cities in China, a 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

[1]Ministry of Education Key Laboratory of Environment and Health, and 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, and Department of Statistics, Harvard University, Boston, MA, USA. [6]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

Article

important in facilitating interpretation of transmission dynamics and epidemic trajectories.

Based on comprehensive epidemiological data from Wuhan[1], we delineated the full dynamics of COVID-19 in the epicenter by extending the susceptible-exposed-infectious-recovered (SEIR) model to a new model named SAPHIRE (Fig. 1, Methods, Extended Data Tables 1-2). We modelled the outbreak from January 1, 2020 across five time periods based on key events and interventions: January 1 to 9 (before *Chunyun*), January 10 to 22 (*Chunyun*), January 23 to February 1 (*cordons sanitaire*), February 2 to 16 (centralized isolation and quarantine), and February 17 to March 8 (community screening). We assumed a constant population size of 10 million with equal numbers of daily inbound and outbound travelers (500,000 before *Chunyun*, 800,000 during *Chunyun*, and 0 after *cordons sanitaire*)[3]. Furthermore, we assumed that the transmission rate and the ascertainment rate did not change in the first two periods, because few interventions were implemented before January 23, while these rates were allowed to vary in later periods to reflect different intervention strengths. 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 symptom onset to isolation and of the relative transmissibility of unascertained cases to ascertained cases. As expected, estimates of R_e were positively correlated with the specified latent and infectious periods, while 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 [CI]: 0.14-0.42) (Methods). We then fit the daily incidences in Wuhan from January 1 to February 29, assuming the initial ascertainment rate was 0.23, and predicted the trend from March 1 to 8 (Methods). Our model fit the observed data well, except for the outlier on February 1, which might be due to approximate-date records of many patients admitted to the field hospitals set up after February 1 (Fig. 2a). After a series of multi-faceted public health interventions, R_e decreased from 3.54 (95% CrI: 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, respectively (Fig. 2b, Extended Data Tables 4-5). We estimated the cumulative number of infections, including unascertained cases, till March 8 to be 258,728 (204,783-320,145) if the trend of the fourth period was assumed (Fig. 2c), or 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) by fitting data from all five periods (Fig. 2a). Correspondingly, these numbers translate into 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.

Strikingly, we estimated low ascertainment rates throughout, which were 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, respectively (Extended Data Table 6). Even with the universal community symptom screening implemented from February 17 to 19, the ascertainment rate was only raised to 0.16. Based on the fitted model using data from January 1 to February 29, we projected the cumulative number of ascertained cases to be 32,577 (30,216-34,986) by March 8, 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 both ascertained and unascertained, peaked at 55,879 (43,582-69,571) on February 2 and dropped afterwards to 701 (436-1,043) on March 8 (Fig. 2f). If the trend remained unchanged, the number of ascertained infections would first become zero on March 27 (95% CrI: March 20 to April 5), while the clearance of all infections would occur on April 21 (April 8 to May 12) (Extended Data Table 7). The first day of zero ascertained case in Wuhan was reported on March 18, indicating enhanced interventions in March.

By stochastic simulations, we investigated the implications of unascertained cases for continuing surveillance and interventions (Methods)[17]. Because of the 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 could be as high as 0.97, and the surge was predicted to occur on day 34 (95% CrI: 27-47) after lifting controls (Fig. 3). If we were to impose a more stringent criterion of lifting controls after 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 (95% CrI: 33-55) after lifting controls (Fig. 3). These results highlighted the risk of ignoring unascertained cases in switching intervention strategies, despite using an over-simplified model.

We performed a series of sensitivity analyses to test the robustness of our results by smoothing the outlier data point on February 1, varying lengths of latent and infectious periods, duration from symptom onset to isolation, ratio of transmissibility of unascertained cases to ascertained cases, and the initial ascertainment rate (Extended Data Tables 4-7, Supplementary Information). Our major findings of remarkable 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 would be 0.08 (0.07-0.10) and 0.23 (0.16-0.28), respectively. If we assumed an extreme scenario with no unascertained cases in the early outbreak (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. In this model, because of the higher ascertainment rate compared to the main analysis, we estimated a lower probability of resurgence of 0.06 when lifting controls after 14 days of no ascertained cases, and the resurgence was expected to occur on day 38 (95% CrI: 29-52) after lifting controls (Fig. 3). A simplified model assuming complete ascertainment anytime performed significantly worse than the full model (Extended Data Table 4, Supplementary Information).

Understanding the proportion of unascertained cases and their transmissibility is critical for prioritization of the surveillance and control measures[17]. Our finding of a large fraction of unascertained cases, despite the strong surveillance in Wuhan, indicated the existence of many asymptomatic or mild-symptomatic cases. It was estimated that asymptomatic cases accounted for 18% of the infections onboard the Diamond Princess Cruise ship[8] and 31% of the infected Japanese evacuated from Wuhan[9]. In addition, 29 of the 33 (88%) infected pregnant women were asymptomatic by universal screening of 210 women admitted for delivery between March 22 and April 4 in New York City[10]. Several reports also highlighted the difficulty in detecting COVID-19 cases: the detection capacity varied from 11% in low surveillance countries to 40% in high surveillance countries[18,19]; modelling of the epidemics outside of Wuhan suggested that the ascertainment rate was 24.4% in China (excluding Hubei province)[14] and 14% in Wuhan prior to travel ban[15]. Consistent with these studies and the emerging serologic studies showing much higher seroprevalence than the reported case prevalence[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 was at the higher end of the range of the estimated basic reproduction number R_0 from other studies using early epidemic data from Wuhan[6,23]. The discrepancy might be due to modelling of unascertained cases, more complete case records in our analysis, and different time periods analyzed. If we modelled from the first COVID-19 case reported in Wuhan, we would estimate a lower R_e of 3.38 (3.28-3.48) before interventions (Extended Data Fig. 2), which is still much higher than those for SARS and MERS[4,5].

Our modelling study 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 synergistically propelled the COVID-19 pandemic, imposing grand challenges to control the outbreak. Still, lessons from Wuhan have demonstrated the effectiveness of vigorous and multifaceted containment efforts. In particular, despite relatively low ascertainment rates due to undetected symptoms of many cases, the outbreak could be controlled by extensive interventions to block the transmission from unascertained cases, such as wearing face masks, social distancing, and quarantining close contacts[1].

Further investigations, such as survey of the seroprevalence of SARS-CoV-2 specific antibodies, are needed to confirm our model estimates given the limitations discussed below. First, due 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, which, however, would lead to underestimation of ascertainment rates, especially for the third and fourth periods when many clinically diagnosed cases were reported[1]. The variation in the estimated ascertainment rates across periods reflected a combined effect of the evolving surveillance, interventions, medical resources, and case definitions across time periods[1,24]. Third, our model assumed homogeneous transmission within the population while ignoring heterogeneity between groups by sex, age, geographic regions 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[27]. Therefore, we might have overestimated the probability of resurgence. Finally, we could not evaluate the impact of individual interventions based on the epidemic curve from a single city, because many interventions were applied simultaneously. Future work by modelling heterogeneous transmission between different groups and joint analysis with data from other cities will lead to 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. *JAMA* **323**, 1915-1923, https://doi.org/10.1001/jama.2020.6130 (2020).
2. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**, 672-675, https://doi.org/10.1038/s41591-020-0869-5 (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, https://doi.org/10.1016/S0140-6736(20)30260-9 (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, https://doi.org/10.1002/jmv.25748 (2020).
5. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966-1970, https://doi.org/10.1126/science.1086616 (2003).
6. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **382**, 1199-1207, https://doi.org/10.1056/NEJMoa2001316 (2020).
7. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* **323**, 1406-1407, https://doi.org/10.1001/jama.2020.2565 (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**, https://doi.org/10.2807/1560-7917. ES.2020.25.10.2000180 (2020).
9. Nishiura, H. *et al.* Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *Int J Infect Dis* **94**, 154-155, https://doi.org/10.1016/j.ijid.2020.03.020 (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, https://doi.org/10.1056/NEJMc2009316 (2020).
11. Tong, Z. D. *et al.* Potential Presymptomatic Transmission of SARS-CoV-2, Zhejiang Province, China, 2020. *Emerg Infect Dis* **26**, 1052-1054, https://doi.org/10.3201/eid2605.200198 (2020).
12. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, https://doi.org/10.1126/science.abb6936 (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, https://doi.org/10.1016/S1473-3099(20)30144-4 (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, https://doi.org/10.1126/science.aba9757 (2020).
15. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489-493, https://doi.org/10.1126/science.abb3221 (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, https://doi.org/10.1038/s41591-020-0822-7 (2020).
17. Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the Epidemiology of Covid-19 - Studies Needed. *N Engl J Med* **382**, 1194-1196, https://doi.org/10.1056/NEJMp2002125 (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, https://doi.org/10.3201/eid2607.200250 (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*, https://doi.org/10.1016/S1473-3099(20)30229-2 (2020).
20. Levesque, J. & Maybury, D. W. A note on COVID-19 seroprevalence studies: a meta-analysis using hierarchical modelling. *medRxiv*, 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*, 10.1016/ S2666-5247 (20)30053-7 (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. & Rocklov, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med*, https://doi.org/10.1093/jtm/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, https://doi.org/10.1016/S2468-2667(20)30089-X (2020).
25. Zhang, J. *et al.* Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481-1486, https://doi.org/10.1126/science.abb8001 (2020).
26. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* **395**, e47, https://doi.org/10.1016/S0140-6736(20)30462-1 (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, https://doi.org/10.1038/nature04153 (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, https://doi.org/10.1126/science.abb6105 (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, https://doi.org/10.1016/S2468-2667(20)30073-6 (2020).

a

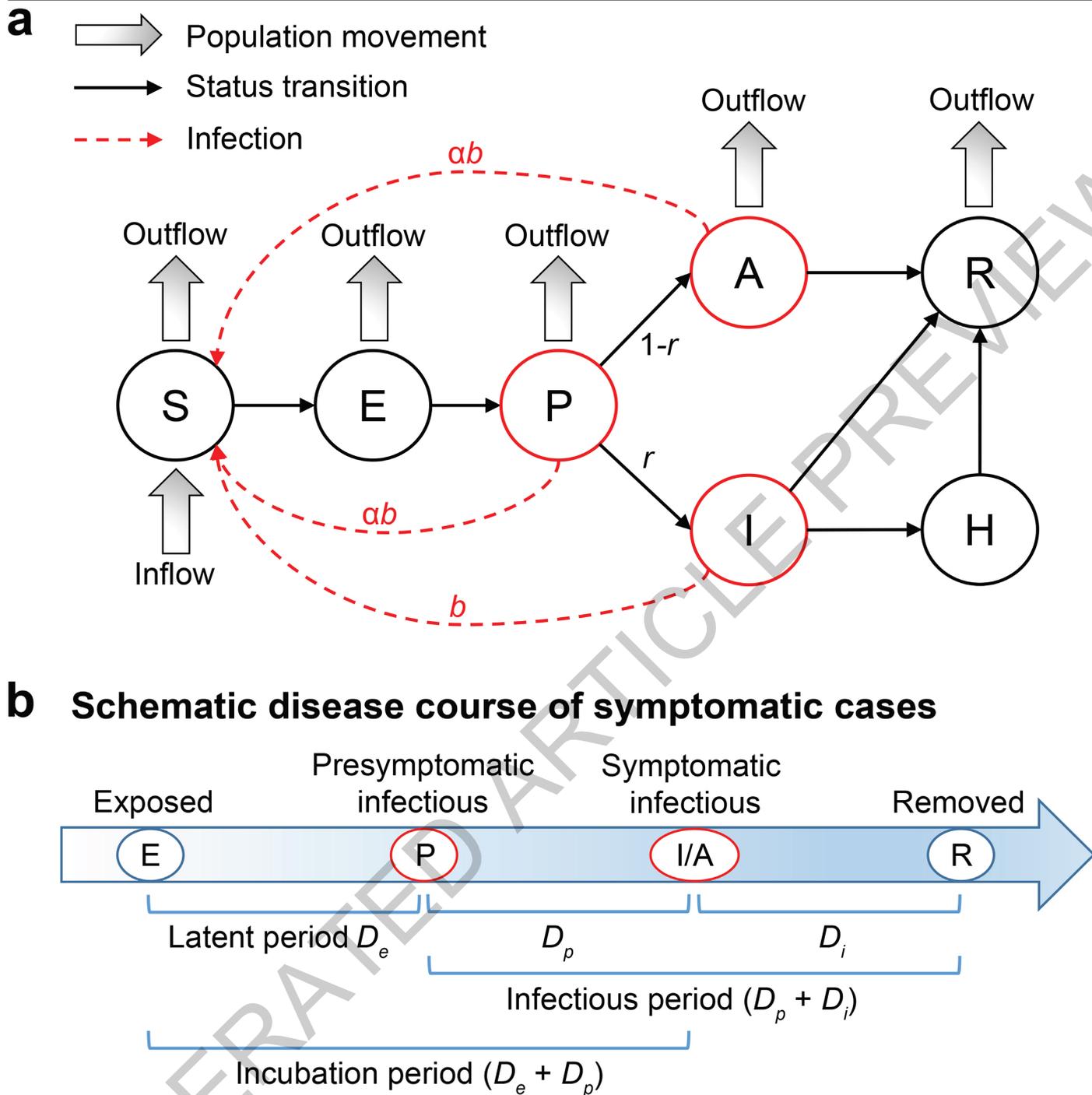**b** Schematic disease course of symptomatic cases

Fig. 1 | Illustration of the SAPHIRE model. We extended the classic SEIR model to include seven compartments, namely *S* (susceptible), *E* (exposed), *P* (presymptomatic infectious), *I* (ascertained infectious), *A* (unascertained infectious), *H* (isolated), and *R* (removed). (a) Relationship between different compartments. Two parameters of interests are *r* (ascertainment rate) and *b* (transmission rate), which are assumed to be varying across time periods.

(b) Schematic disease course of a symptomatic case. In this model, the unascertained compartment *A* includes asymptomatic and some mild-symptomatic cases who were not detected. While there is no presymptomatic phase for asymptomatic cases, we treated asymptomatic as a special case of mild-symptomatic and modeled both with a "presymptomatic" phase for simplicity.
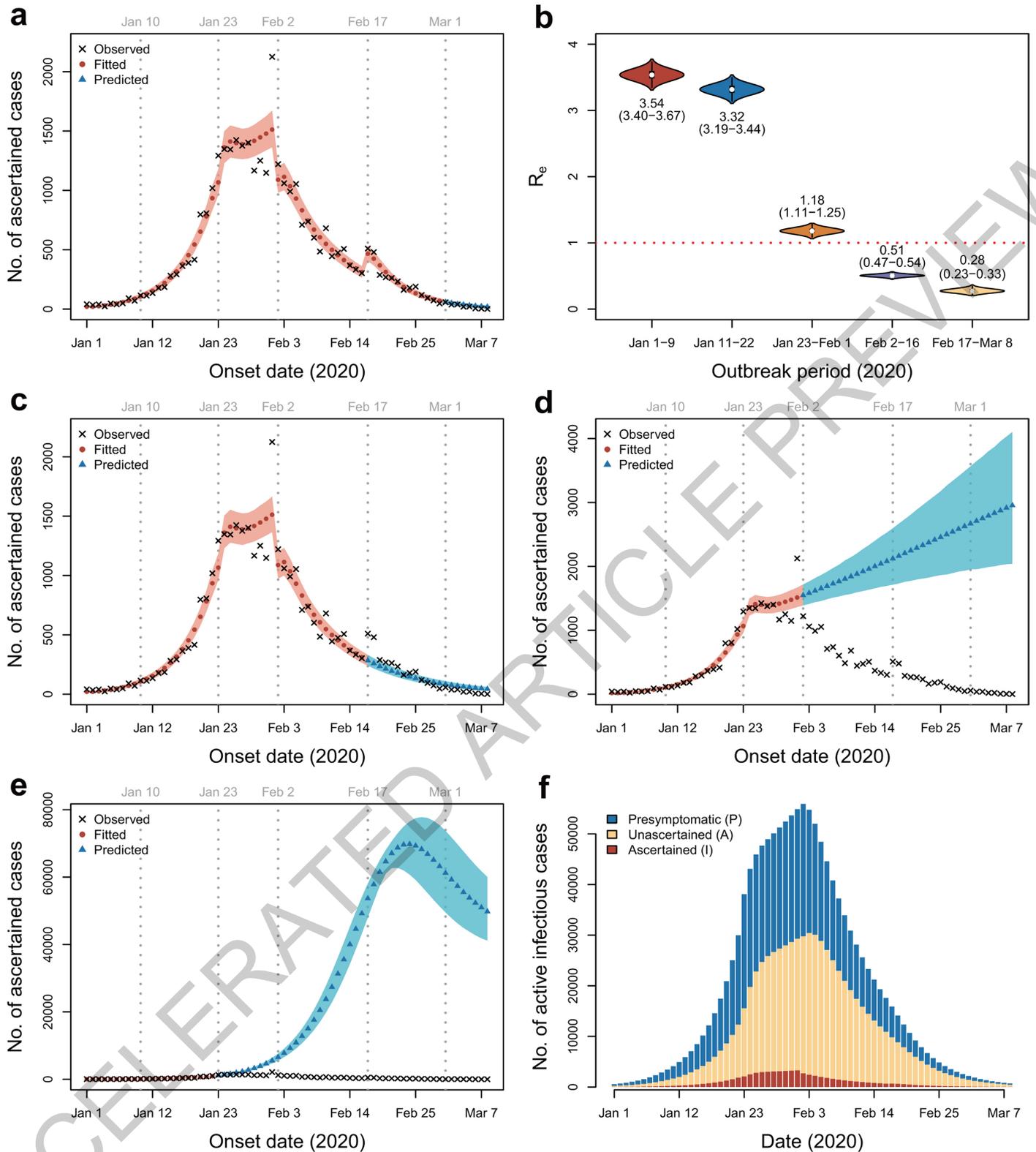
Fig. 2 | Modelling the COVID-19 epidemic in Wuhan. Parameters were estimated by fitting data from January 1 to February 29. (a) Prediction using parameters from period 5 (February 17 to 29). (b) Distribution of R_e estimates from 10,000 MCMC samples. In each violin plot, the white dot represents the median, the thick bar represents the interquartile range (IQR), and the thin bar represents the minimum and the maximum. The mean and the 95% CrI (in parentheses) are labelled below or above. (c) Prediction using parameters from period 4 (February 2 to 16). (d) Prediction using parameters from period 3 (January 23 to February 1). (e) Prediction using parameters from period 2 (January 10 to 22). The shaded areas in (a, c, d and e) are 95% CrI and the colored points are the mean values based on 10,000 MCMC samples. (f) Estimated number of active infectious cases in Wuhan from January 1 to March 8.
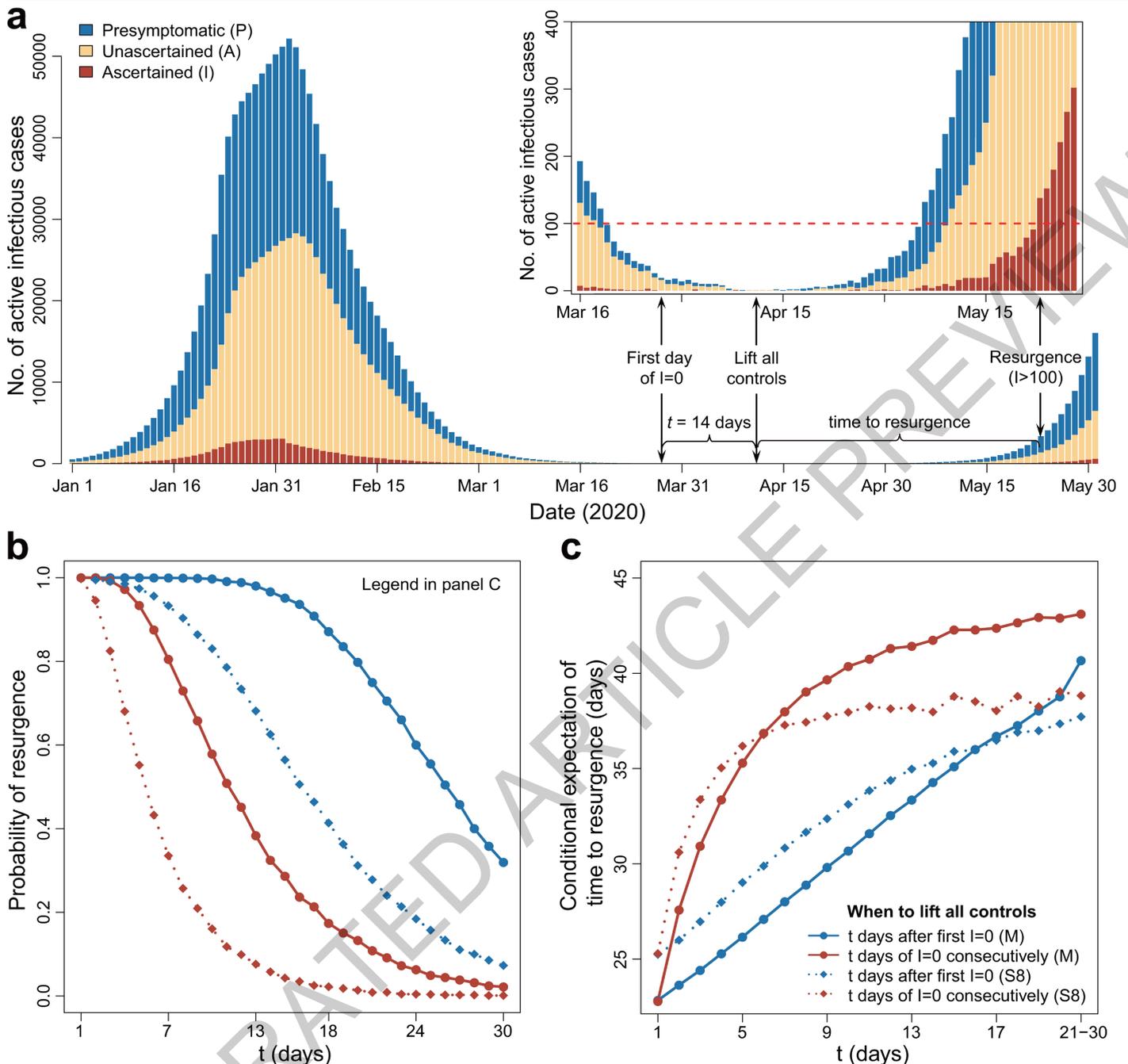
Fig. 3 | Risk of resurgence after lifting controls. We considered the main model (M) and the sensitivity analysis S8 (**Methods**). In model M, we assumed the initial ascertainment rate r_0=0.23, and thus had an overall ascertainment rate of 0.13. In model S8, we assumed no unascertained cases initially and thus had an overall ascertainment rate of 0.47. For each model, we simulated epidemic curves based on 10,000 sets of parameters from MCMC, assuming transmission rate b, ascertainment rate r, and population movement n were resumed to values before *Chunyun* after lifting controls. A resurgence was defined by the time of 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 inserted panel is a zoom-in plot from March 16 to May 28. (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 last 10 days (t=21 to 30) to calculate the expected time to resurgence because of their low probability of resurgence.

Methods

Data of COVID-19 cases in Wuhan

We analyzed the daily incidence data of COVID-19 presented in Figure 1 of Pan *et al*[1]. Briefly, information of COVID-19 cases from December 8, 2019 till March 8, 2020 were extracted from the municipal Notifiable Disease Report System on March 9, 2020. Date of symptom onset (the self-reported date of developing symptoms such as fever, cough, or other respiratory symptoms) and date of confirmed diagnosis were collected. For the consistency of case definition throughout the periods, we only included 32,583 laboratory-confirmed cases who tested positive 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. Software SAS (version 9.4) was used in data collection.

Estimation of initial ascertainment rate using cases exported to Singapore

As of May 10, 2020, a total of 24 confirmed COVID-19 cases in Singapore were reported to be imported from China, among which 16 were imported from Wuhan before the *cordons sanitaire* on January 23 and the first case arrived in Singapore on January 18 (Extended Data Table 3). Based on VariFlight Data (https://data.variflight.com/en/), the total number of passengers from Wuhan to Singapore between January 18 and 23, 2020 was 2,722. Therefore, the cumulative infection rate among the passengers was 0.59% (=16/2722, 95% CI: 0.30-0.88%). These cases had symptom onset between January 21 and 30, 2020. In Wuhan, a total of 12,433 confirmed cases were reported to have symptom onset in the same period, equivalent to a cumulative infection rate of 0.124% (95% CI: 0.122-0.126%) by 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 excellent surveillance strength[18,19], the ascertainment rate during the early outbreak in Wuhan was estimated to be 0.23 (95% CI: 0.14–0.42), corresponding to 0.77 (95% CI: 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 cases[8,9]; and (2) the number of imported cases with onset between January 21 and 30 was censored due to suspension of flights after Wuhan lockdown. Without direct information to estimate the initial ascertainment rate before January 1, 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 susceptible-exposed-infectious-recovered (SEIR) model to a SAPHIRE model (Fig. 1, Extended Data Table 1), which incorporates three additional compartments to account for presymptomatic infectiousness (P), unascertained cases (A), and case isolation in the hospital (H). We chose to analyze data from January 1, 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 number of daily inbound and outbound travelers n, where n=500,000 for January 1 to 9, 800,000 for January 10 to 22 due to *Chunyun*, and 0 after *cordons sanitaire* from January 23[3]. We divided the population into S susceptible, E exposed, P presymptomatic infectious, A unascertained infectious, I ascertained infectious, H isolated, and R removed individuals. We introduced an isolated compartment H because ascertained cases would have shorter effective infectious period due to isolation, especially when medical resources were improved[1]. Dynamics of these compartments across time t were described by the following set of ordinary differential equations:

$$\frac{dS}{dt} = n - \frac{bS(\alpha P + \alpha A + I)}{N} - \frac{nS}{N} \tag{1}$$

$$\frac{dE}{dt} = \frac{bS(\alpha P + \alpha A + I)}{N} - \frac{E}{D_e} - \frac{nE}{N} \tag{2}$$

$$\frac{dP}{dt} = \frac{E}{D_e} - \frac{P}{D_p} - \frac{nP}{N} \tag{3}$$

$$\frac{dA}{dt} = \frac{(1-r)P}{D_p} - \frac{A}{D_i} - \frac{nA}{N} \tag{4}$$

$$\frac{dI}{dt} = \frac{rP}{D_p} - \frac{I}{D_i} - \frac{I}{D_q} \tag{5}$$

$$\frac{dH}{dt} = \frac{I}{D_q} - \frac{H}{D_h} \tag{6}$$

$$\frac{dR}{dt} = \frac{A+I}{D_i} + \frac{H}{D_h} - \frac{nR}{N} \tag{7}$$

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

$$R_e = \alpha b\left(D_p^{-1} + \frac{n}{N}\right)^{-1} + (1-r)\alpha b\left(D_i^{-1} + \frac{n}{N}\right)^{-1} + rb(D_i^{-1} + D_q^{-1})^{-1} \tag{8}$$

where the three terms represent infections contributed by presymptomatic, unascertained, and ascertained cases, respectively. We adjusted the infectious periods of each type of cases 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 were summarized in Extended Data Table 2. We set α=0.55 according to Li *et al*.[15], assuming lower transmissibility for unascertained cases. Compartment P contains both ascertained and unascertained cases at the presymptomatic phase. We set the transmissibility of P to be the same as unascertained cases because it has been reported that the majority of the cases were 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 of ascertained cases[2], we set the mean of total infectious period as $(D_p + D_i) = \frac{D_p}{0.44} = 5.2$ days assuming constant infectiousness across presymptomatic and symptomatic phases for the 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 impact on our fitting procedure and the final parameter estimates. The duration from symptom onset 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 each period, respectively[1].

Based on the settings above, we specified the initial state of the model on December 31, 2019 (Extended Data Table 1). The initial number of ascertained symptomatic cases $I(0)$ was specified as the number of ascertained cases with onset during December 29 to 31, 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 with onset during

Article

January 1 to 2, 2020 and during January 3 to 5, 2020, respectively. Then, the initial numbers of exposed cases and presymptomatic cases 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 based on the point estimate using the Singapore data (described above).

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 the observed number of ascertained cases with symptom onset on day d, denoted as x_d, follows a Poisson distribution with rate $\lambda_d = rP_{d-1}D_p^{-1}$, where P_{d-1} is the expected number of presymptomatic cases on day (d-1). We fit the observed data from January 1 to February 29 (d=1, 2, ..., D, and D=60) and used the fitted model to predict the trend from March 1 to 8. Thus, the likelihood function is

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

We estimated $b_{12}, b_3, b_4, b_5, r_{12}, r_3, r_4$ and r_5 by Markov Chain Monte Carlo (MCMC) with the Delayed Rejection Adaptive Metropolis (DRAM) 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 above). We reparameterized r_3, r_4, and r_5 by

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

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

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

where $\text{logit}(r) = \log\left(\frac{r}{1-r}\right)$. 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 (Supplementary Information)[31]. Estimates of parameters were presented as posterior means and 95% credible intervals (CrIs) from 10,000 MCMC samples. All 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% CrI of a fitted/predicted epidemic curve. Given a set of parameter values from MCMC, we performed the following multinomial random sampling:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We repeated the stochastic simulations for all 10,000 sets of parameter values sampled by MCMC to construct the 95% CrI 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 above, 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 conol measures as the last period (*i.e.*, 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 above). We calculated the probability of resurgence as the proportion of simulations in which a 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 (Eqs. 10–16) 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 Eq. 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 Eq. 8). Lengths of both periods were set to 15 days, and the initial ascertainment rate was set to r_0=0.3, while 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, while setting all other parameters and initial values the same as the true values. We translated b_1 and b_2 into $(R_e)_1$ and $(R_e)_2$ according to Eq. 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, while keeping 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 displayed results for 100 repeated simulations.

Sensitivity analyses for the real data

We designed nine sensitivity analyses to test the robustness of our real data results. 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. (S1) Adjust the reported incidences from January 29 to February 1 to their average. We suspect the spike of incidences on February 1 might be caused by approximate-date records among some patients admitted to the field hospitals after February 2. The actual illness onset dates for these patients were likely to be spread between January 29 and February 1. (S2) Assume an incubation period of 4.1 days (lower 95% CI from ref. [6]) and presymptomatic infectious period of 1.1 days (lower 95% CI 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. (S3) Assume an incubation period of 7 days (upper 95% CI from ref. [6]) and presymptomatic infectious period of 3 days (upper 95% CI from ref. [2]), equivalent to set $D_e = 4$ and $D_p = 3$, and adjust $P(0)$ and $E(0)$ accordingly. (S4) Assume the transmissibility of the unascertained cases is $\alpha = 0.46$ (lower 95% CI from ref. [15]) of the ascertained cases. (S5) Assume the transmissibility of the unascertained cases is $\alpha = 0.62$ (upper 95% CI from ref. [15]) of the ascertained cases. (S6) Assume the initial ascertainment rate is $r_0 = 0.14$ (lower 95% CI of the estimate using Singapore data) and adjust $A(0)$, $P(0)$ and $E(0)$ accordingly. (S7) Assume the initial ascertainment rate is $r_0 = 0.42$ (upper 95% CI of the estimate using Singapore data) and adjust $A(0)$, $P(0)$ and $E(0)$ accordingly. (S8) Assume the initial ascertainment rate is $r_0 = 1$ (theoretical upper limit) and adjust $A(0)$, $P(0)$, and $E(0)$ accordingly. (S9) 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. *Statistics and Computing* **16**, 339-354, https://doi.org/10.1007/s11222-006-9438-0 (2006).
31. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat* **7**, 434-455 (1997).

Author contributions TW, XL and CW designed the study. XH, SC, XL and CW developed statistical methods. XH, SC, and DW performed data analysis. CW wrote the first draft of the manuscript. All authors reviewed and edited the manuscript.
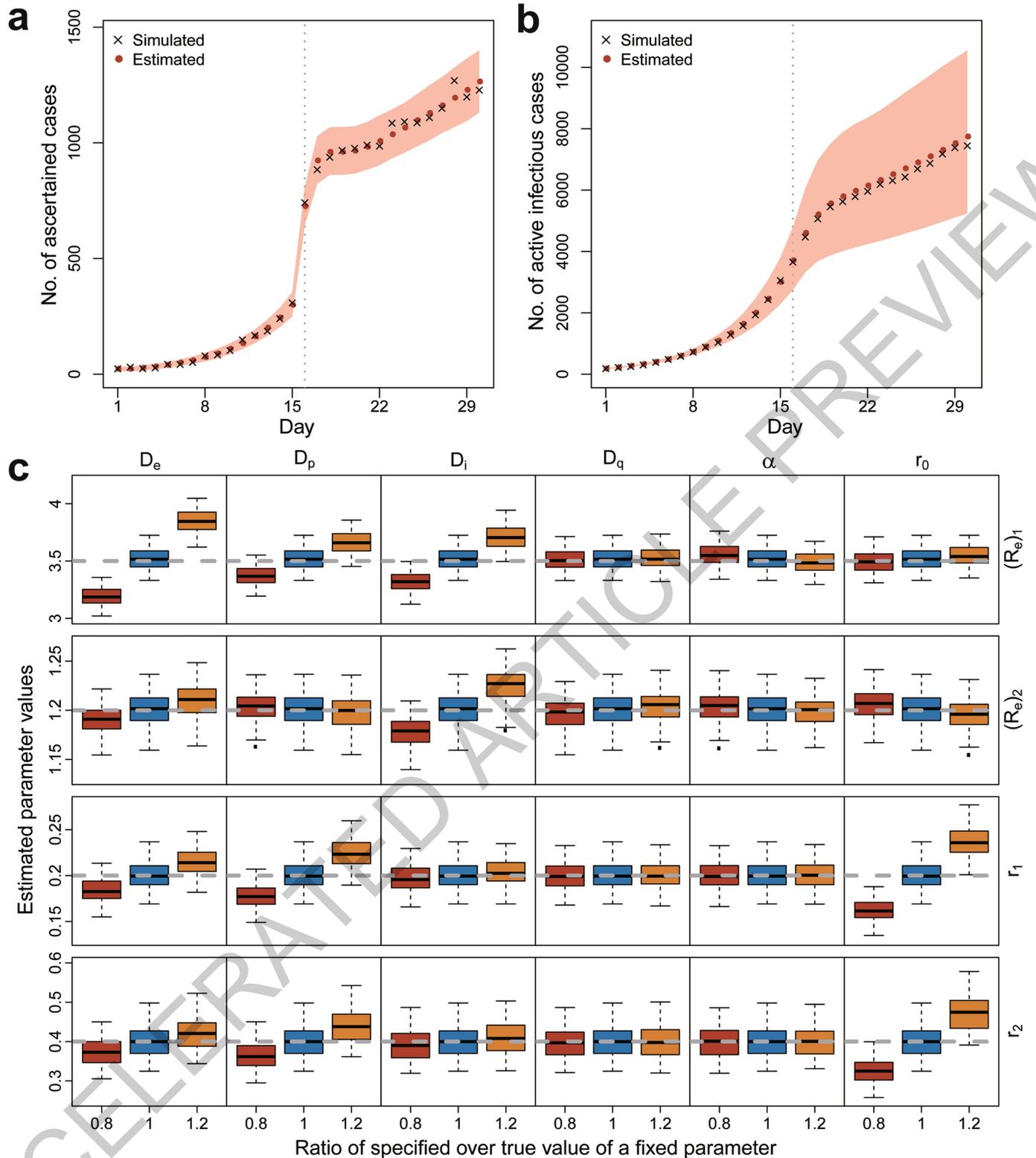
Extended Data Fig. 1 | Evaluation of the method on simulated data with two periods. (a-b) Illustration of one simulated dataset. We estimated b_1, b_2, r_1, and r_2 when the other parameters were specified to their true values. The red points represent the mean estimates while the shaded areas indicate 95% CrIs 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 IQR is indicated by the lower and upper bounds, and the minimum and the maximum are indicated by the whiskers.
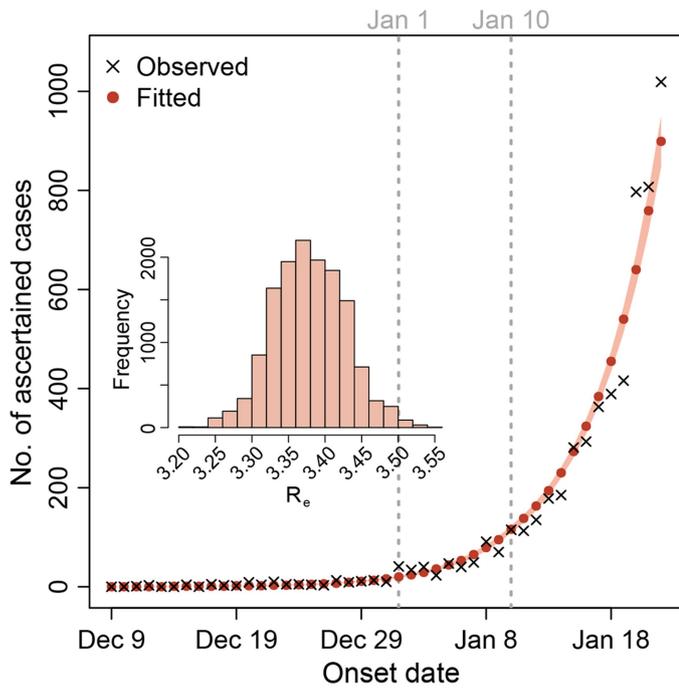
Extended Data Fig. 2 | Estimation of R_0 using daily incidence data starting from December 9. Following the main analysis, we assumed r_0=0.23 and set $I(0)$=1, $A(0)$=3, $E(0)$=17 and $P(0)$=$H(0)$=$R(0)$=0 accordingly. We assumed transmission rate b, ascertainment rate r, and duration from illness onset to hospitalization D_q (set to 21 days) were the same until January 22, 2020. All the other settings were the same as in the main analysis. The shaded area in the plot indicates 95% CrIs estimated by the deterministic model with 10,000 sets of parameter values sampled from MCMC. Unlike other analyses, we did not construct 95% CrIs by stochastic simulations, because stochastic variation at the early days had very large impacts due to low counts. The inserted histogram shows the distribution of the estimated R_0 from December 9, 2019 to January 22, 2020, for which the mean estimate was 3.38 (95% CrI: 3.28-3.48).

Article

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	$E(0) = r_0^{-1}E_I(0)$, where $E_I(0)$ 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	$P(0) = r_0^{-1}P_I(0)$, where $P_I(0)$ 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 was January 1, 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	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

Article

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

Source: 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)	0.4 (0.38-0.43)	0.17 (0.16-0.19)	0.1 (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)	0.1 (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 (0.1-0.13)	0.04 (0.02-0.06)	588.73
S4	1.53 (1.46-1.61)	0.47 (0.44-0.5)	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 (0.07-0.1)	553.49
S6	1.34 (1.28-1.39)	0.41 (0.38-0.44)	0.18 (0.17-0.19)	0.1 (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)	0.1 (0.08-0.11)	555.40
S8	1.2 (1.14-1.27)	0.36 (0.34-0.39)	0.17 (0.16-0.18)	0.1 (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 were displayed as mean (95% CrI) based on 10,000 MCMC samples.
†Deviance Information Criterion (DIC) was presented for model comparison. Nevertheless, DIC of S1 is not comparable to the others because the data of S1 were modified by smoothing the outlier data point on February 1.

Article

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

Analysis	Estimated effective reproduction number R_e*				
	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 were displayed as mean (95% CrI) 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 were displayed as mean (95% CrI) based on 10,000 MCMC samples.

Article

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 were displayed as mean date (95% CrI) based on 10,000 stochastic simulations with parameter values from MCMC sampling.

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

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

nature research

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