Check for updates

Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China

Joseph T. Wu [1,3 ✉], Kathy Leung[1,3], Mary Bushman[2], Nishant Kishore[2], Rene Niehus[2], Pablo M. de Salazar[2], Benjamin J. Cowling [1], Marc Lipsitch[2] and Gabriel M. Leung[1]

As of 29 February 2020 there were 79,394 confirmed cases and 2,838 deaths from COVID-19 in mainland China. Of these, 48,557 cases and 2,169 deaths occurred in the epicenter, Wuhan. A key public health priority during the emergence of a novel pathogen is estimating clinical severity, which requires properly adjusting for the case ascertainment rate and the delay between symptoms onset and death. Using public and published information, we estimate that the overall symptomatic case fatality risk (the probability of dying after developing symptoms) of COVID-19 in Wuhan was 1.4% (0.9–2.1%), which is substantially lower than both the corresponding crude or naïve confirmed case fatality risk (2,169/48,557 = 4.5%) and the approximator[1] of deaths/ deaths + recoveries (2,169/2,169 + 17,572 = 11%) as of 29 February 2020. Compared to those aged 30–59 years, those aged below 30 and above 59 years were 0.6 (0.3–1.1) and 5.1 (4.2–6.1) times more likely to die after developing symptoms. The risk of symptomatic infection increased with age (for example, at ~4% per year among adults aged 30–60 years).

On 9 January 2020, the novel coronavirus SARS-CoV-2 was officially identified as the cause of the COVID-19 outbreak in Wuhan, China. One of the most critical clinical and public health questions during the emergence of a completely novel pathogen, especially one that could cause a global pandemic, pertains to the spectrum of illness presentation or severity profile. For the patient and clinician, this affects triage and diagnostic decision-making, especially in settings without ready access to laboratory testing or when surge capacity has been exceeded. It also influences therapeutic choice and prognostic expectations. For managers of health services, it is important for rapid forward planning in terms of procurement of supplies, readiness of human resources to staff beds at different intensities of care and generally ensuring the sustainability of the health system through the peak and duration of the epidemic.

At the population level, determining the shape and size of the 'clinical iceberg'[2,3], both above and below the observed threshold (in turn determined by symptomatology, care-seeking behavior and clinical access), is key to understanding the transmission dynamics and interpreting epidemic trajectories. Specifically, delineating the proportion of infections that are clinically unobserved under different circumstances is critical to refining model parameterization. In turn, estimates of both the observed and unobserved infections are essential for informing the development and evaluation of public health strategies, which need to be traded off against economic, social and personal freedom costs. For example, drastic social distancing and mobility restrictions, such as school closures and travel advisories/bans, should only be considered if an accurate estimation of case fatality risk warrants these interventions, which seriously disrupt social and economic stability.

For a completely novel pathogen, especially one with a high (say, >2) basic reproductive number (the expected number of secondary cases generated by a primary case in a completely susceptible population) relative to other recently emergent and seasonal directly transmissible respiratory pathogens[4], assuming homogeneous mixing and mass action dynamics, the majority of the population will be infected eventually unless drastic public health interventions are applied over prolonged periods and/or vaccines become available sufficiently quickly. Even under more realistic assumptions about mixing informed by observed clustering of infections within households and the increasingly apparent role of superspreading events (for example, the *Diamond Princess* cruise ship, Chinese prisons and the church in Daegu, South Korea)[5,6], at least one-quarter to one-half of the population will very likely become infected, absent drastic control measures or a vaccine. Therefore, the number of severe outcomes or deaths in the population is most strongly dependent on how ill an infected person is likely to become, and this question should be the focus of attention.

We therefore extended our previously published transmission dynamics model[4], updated with real-time input data and enriched with additional new data sources, to infer a preliminary set of clinical severity estimates that could guide clinical and public health decision-making as the epidemic continues to spread globally. Estimation of true case numbers—necessary to determine the severity per case—is challenging in the setting of an overwhelmed healthcare system that cannot ascertain cases effectively. Therefore, as in our prior work[4], our approach has been to use a range of publicly available and recently published data sources (numbered 1 to 8 below) to build a picture of the full number of cases and deaths by age group. Briefly, because the healthcare structure has been overwhelmed in Wuhan and milder cases were unlikely to have been tested, we used the prevalence of infection in travelers (both on commercial flights before 19 January and on charter flights from 29 January to 4 February) to estimate the true prevalence of infection in Wuhan; we also used the Wuhan case numbers from only the first 425 cases to estimate the growth rate of the epidemic (assuming that the ascertainment proportion was constant between 10 December 2019 and 3 January 2020) (Fig. 1).

Specifically, we inferred the epidemiologic parameters listed in Extended Data Fig. 1 by fitting an age-structured transmission model to the following data:

1. The epidemic curve of confirmed cases of COVID-19 in Wuhan with no epidemiologic links to Huanan Seafood Wholesale

[1]WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China. [2]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]These authors contributed equally: Joseph T. Wu, Kathy Leung. ✉e-mail: joewu@hku.hk
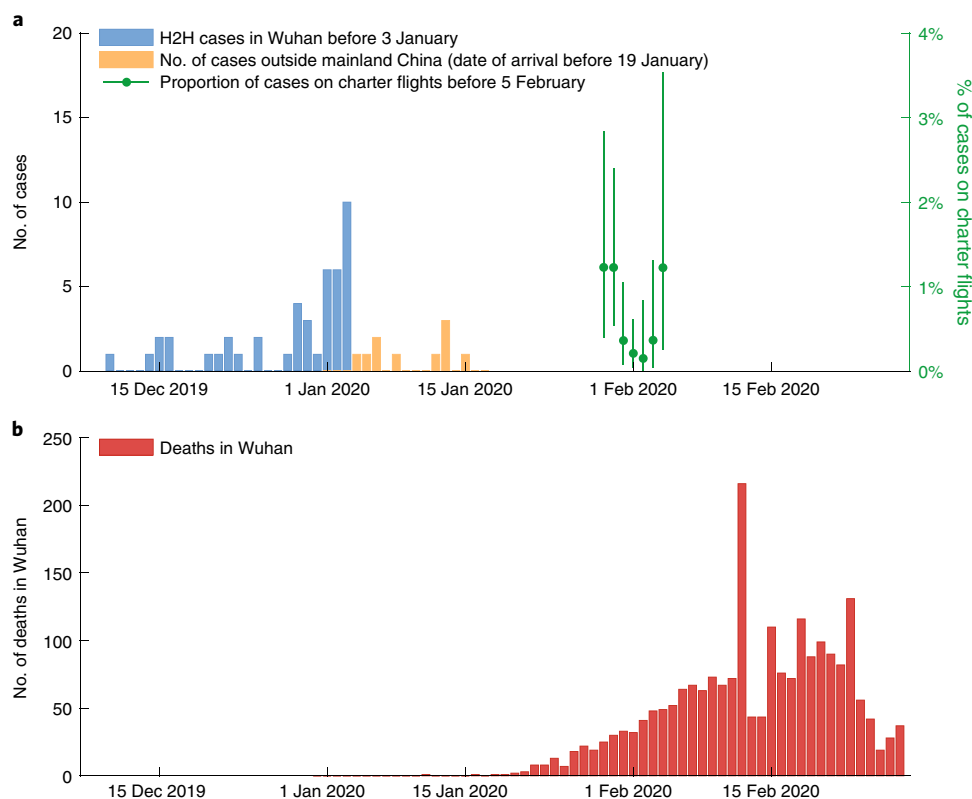
Fig. 1 | Data used in the inference. a, The daily number of confirmed cases in Wuhan (with no epidemiologic links to Huanan Seafood Wholesale Market, i.e., cases due to human-to-human (H2H) transmission) between 1 December 2019 and 3 January 2020 (blue), the daily number of cases exported from Wuhan to cities outside mainland China via air travel between 25 December 2019 and 19 January 2020 (orange) and the proportion of expatriates on charter flights between 29 January and 4 February 2020 who were laboratory-confirmed to be infected (green). The numbers of passengers and confirmed cases who returned to their countries from Wuhan on chartered flights are provided in Supplementary Table 3. Bars indicate the 95% confidence intervals (CIs) of the proportion. **b**, The daily number of deaths in Wuhan reported between 1 December 2019 and 28 February 2020.

Market (which was postulated to be the index zoonotic source of the COVID-19 epidemic) between 10 December 2019 and 3 January 2020 (Fig. 1 and Supplementary Table 1)[7].

2. The number of confirmed cases who departed from the Wuhan international airport to cities outside mainland China via air travel on each day between 25 December 2019 and 19 January 2020 (Fig. 1 and Supplementary Table 2)[4].

3. The number of expatriates and visitors who returned to their countries from Wuhan on charter flights between 29 January and 4 February 2020 and the proportion of passengers on each flight who had laboratory-confirmed infection with COVID-19 (by polymerase chain reaction with reverse transcription, RT-PCR) on arrival (Fig. 1 and Supplementary Table 3).

4. The age distribution of all confirmed cases of COVID-19 in Wuhan as of 11 February 2020[8] (Supplementary Table 4).

5. The age distribution of all death cases of COVID-19 in mainland China as of 11 February 2020[8] (Supplementary Table 5).

6. The cumulative number of deaths among confirmed cases of COVID-19 infection in Wuhan as of 25 February 2020[9] (Supplementary Table 6).

7. The time between onset and death or the time between admission and death for 41 death cases of COVID-19 in Wuhan[10–12] (Supplementary Table 7).

8. The time between the onset dates (that is, serial intervals) of 43 infector–infectee pairs (Supplementary Table 8).

The clinical severity of infectious diseases is typically measured in terms of infection fatality risk (IFR), symptomatic case fatality risk (sCFR) and hospitalization fatality risk (HFR). The case definitions underlying these severity measures are as follows:

1. IFR defines a case as a person who would, if tested, be counted as infected and rendered (at least temporarily) immune, as usually demonstrated by seroconversion or other immune response[13]. Such cases may or may not be symptomatic.

2. sCFR defines a case as someone who is infected and shows certain symptoms.

3. HFR defines a case as someone who is infected and hospitalized. It is typically assumed in such estimates that the hospitalization is for treatment rather than isolation purposes.

Figure 2 summarizes our estimates of age-specific sCFRs and susceptibility to symptomatic infection. Both parameters increase substantially with age. If the probability of developing symptoms after infection, P_{sym}, is 0.5, the sCFR values are 0.3% (0.1–0.7%), 0.5% (0.3–0.8%) and 2.6% (1.7–3.9%) for those aged <30 years, 30–59 years and >59 years, respectively. The overall sCFR is 1.4% (0.9–2.1%). Compared to those aged 30–59 years, those aged <30 years and >59 years are 0.16 (0.15–0.17) and 2.0 (1.95–2.08) times more susceptible to symptomatic infection. Our estimates of sCFRs would be lower if P_{sym} were higher than the baseline value of 0.5; for example, the overall sCFR is 1.3% (0.8–2.3%) and 1.2% (0.7–1.9%) if P_{sym} is 0.75 and 0.95, respectively. Our estimates of age-specific susceptibility are not sensitive to P_{sym}.

Figure 3 summarizes our estimates of the key epidemiologic parameters of COVID-19 in Wuhan. In the baseline scenario
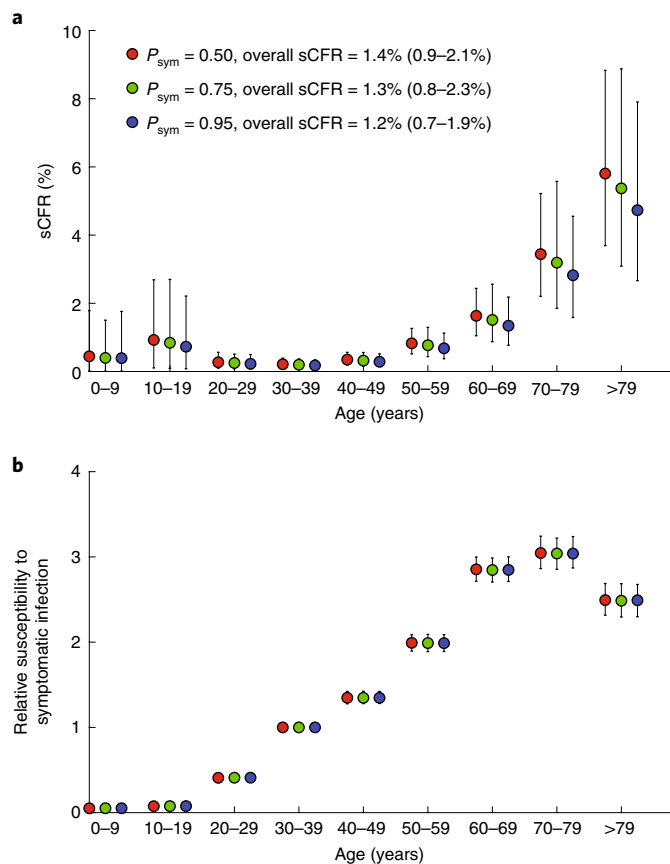
Fig. 2 | Estimates of age-specific sCFR and susceptibility to symptomatic infection for COVID-19 in Wuhan. a, Estimates of age-specific sCFRs assuming P_{sym} is 0.50 (red), 0.75 (green) and 0.95 (blue). **b**, Estimates of relative susceptibility to symptomatic infection by age assuming P_{sym} is 0.50 (red), 0.75 (green) and 0.95 (blue). The markers in both panels show the posterior means and the bars show 95% credible intervals (CrIs).

($P_{sym} = 0.5$), the basic reproductive number is 1.94 (1.83–2.06). The mean serial interval is 7.0 (5.8–8.1) days, with a standard deviation of 4.5 (3.5–5.5) days. The mean time from onset to death is 20 (17–24) days, with a standard deviation of 10 (7–14) days. The epidemic doubling time (the time it takes for daily incidence to double) was 5.2 (4.6–6.1) days before Wuhan was quarantined and public health interventions implemented within Wuhan reduced transmissibility by 48% (24–71%). We estimate that only 1.8% (0.9–3.3%) of symptomatic cases that occurred between 10 December 2019 and 3 January 2020 were ascertained. Figure 3 suggests that our estimates of the basic reproductive number, mean generation time and intervention effectiveness would be slightly lower if P_{sym} were higher than the baseline value of 0.5, whereas our estimates of the other parameters are largely insensitive to P_{sym}.

There is a clear and considerable age dependency in symptomatic infection (susceptibility) and outcome (fatality) risks, by multiple folds in each case. Given that we have parameterized the model using death rates inferred from projected case numbers (from traveler data) and observed death numbers in Wuhan, the precise fatality risk estimates may not be generalizable to those outside the original epicenter, especially during subsequent phases of the epidemic. The experience gained from managing those initial patients and the increasing availability of newer, and potentially better, treatment modalities to more patients would presumably lead to fewer deaths, all else being equal. Public health control measures widely imposed in China since the Wuhan alert have also kept case

numbers down elsewhere, so that their health systems are not nearly as overwhelmed beyond surge capacity, thus again perhaps leading to better outcomes[6,8]. Indeed, so far, the death-to-case ratio in Wuhan has been consistently much higher than that among all the other mainland Chinese cities (Extended Data Fig. 2). Given the intensive efforts of case finding and the sharp drop in community transmission of COVID-19 in Chinese cities outside Hubei over the past few weeks, the ascertainment rates in these cities were probably very high. As such, we postulate that confirmed case fatality risk in these cities should be in some ways comparable to our sCFR estimates for Wuhan, which attempt to account for under-ascertainment of cases in Wuhan. Nonetheless, crude case fatality risks estimated from cities outside Wuhan should be, and are, lower than our sCFR estimates for Wuhan, because the former do not account for the delay between onset and death (thus being artefactually lower) and because healthcare outside Hubei is less overwhelmed (thus allowing a truly lower CFR). Indeed, as of 29 February 2020, the crude case fatality risk in areas outside Hubei was 0.85%, which is ~23–41% lower than our sCFR estimates of 1.2–1.4% for Wuhan[9].

Considering the risk estimates in context, Extended Data Fig. 3 compares infection, case and hospitalization fatality risks for pandemic influenza in 1918 and 2009, SARS and MERS. SARS causes moderate to severe disease requiring hospitalization, so the infection fatality risk and case fatality risk are essentially the same as the hospitalization fatality risk. The hospitalization fatality risk for MERS is well documented, although the shape and depth of the clinical iceberg remains less well defined. In contrast, because (1) the majority of COVID-19 infections do not cause severe disease[8] and (2) hospitals in Wuhan have been overwhelmed, presumably having led to prioritized admission of more serious cases, the sCFR will be substantially lower than the HFR. However, despite a lower sCFR, COVID-19 is likely to infect many more (given emerging evidence of presymptomatic transmission[14,15] and growing evidence of extensive community spread in numerous countries[16]), thus ultimately causing many more deaths than SARS and MERS. Compared with the 1918 and 2009 influenza pandemics, our estimates are intermediate but substantially higher than 2009, which was generally regarded as a low-severity pandemic. We find that sCFR is highest in the oldest age group. Unlike any previously reported pandemic or seasonal influenza, we find that risk of symptomatic infection also increases with age, although this may be in part due to preferential ascertainment of older and thus more severe cases. One largely unknown factor at present is the number of asymptomatic, undiagnosed infections. These do not enter our estimates of sCFR, but if such asymptomatic or clinically very mild cases existed and were not detected, the infection fatality risk would be lower than sCFR. Further clarifying this requires new data sources that are not yet available, specifically including age-stratified serologic studies.

Our inferences were based on a variety of sources, and have a number of caveats that are highlighted below, but considering the totality of the findings they nevertheless indicate that COVID-19 transmission is difficult to control. With a basic reproductive number of around two, we might expect at least half of the population to be infected, even with aggressive use of community mitigation measures. Perhaps the most important target of mitigation measures would be to 'flatten out' the epidemic curve, reducing the peak demand on healthcare services and buying time for better treatment pathways to be developed. In due course, but almost certainly after the first global wave of infections, vaccines may also be available to protect against infection or severe disease. Although our estimates of sCFR are concerning, these could be reduced if effective antivirals were identified and widely adopted for the treatment of severe cases. Timely data from clinical trials of remdesivir, lopinavir/ritonavir and other potential chemotherapies, as well as supportive care modalities, would be extremely informative.
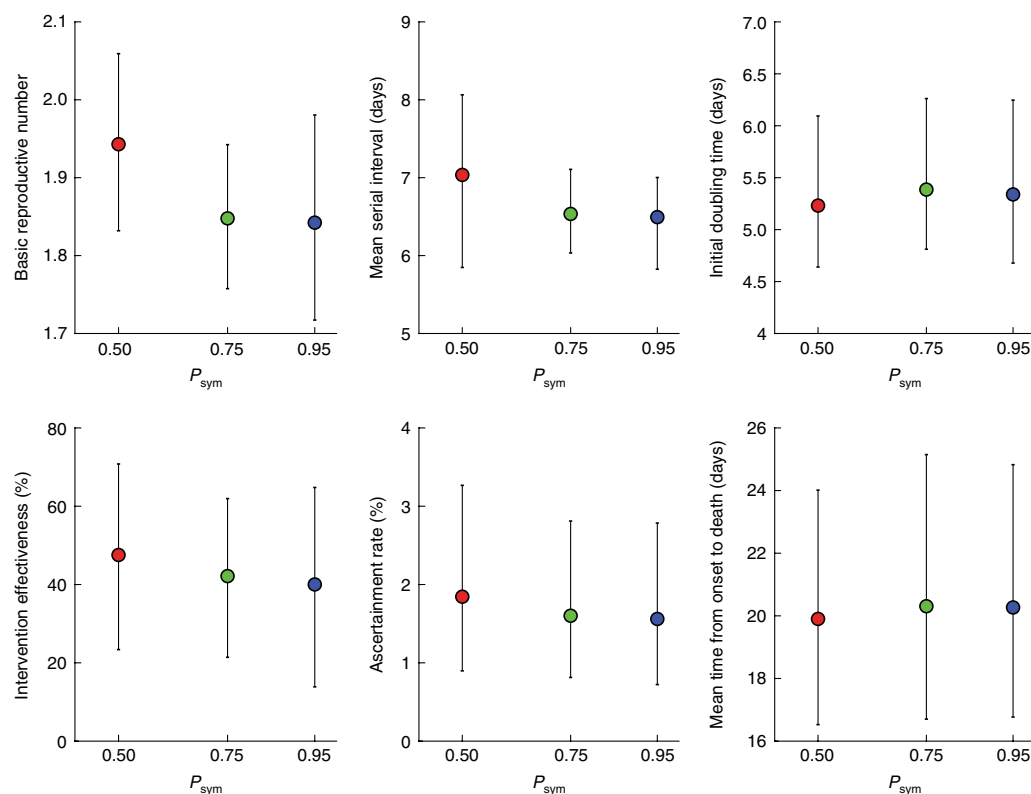
Fig. 3 | Estimates of key epidemiologic parameters of the COVID-19 epidemic in Wuhan. Estimates of basic reproductive number, mean serial interval, initial doubling time, intervention effectiveness, ascertainment rate and the mean time from onset to death, assuming P_{sym} is 0.50 (red), 0.75 (green) and 0.95 (blue). The markers show the posterior means and the bars show 95% CrIs.

Several important caveats are worth mentioning, as follows. First, and most importantly, our modeled estimates have necessarily relied on numerous strong assumptions, given the paucity of definitive data elements such as serosurveys, serial viral shedding studies, robust ascertainment of sufficient transmission chains and incomplete testing of travelers and returnees from Wuhan, all of which need to be underpinned by systematic unbiased sampling of the underlying population and by important age and other sub-groups.

Our estimates of sCFR are inevitably affected by under-ascertainment of cases and deaths of COVID-19. On the one hand, overstretched and overwhelmed healthcare surge capacity in Wuhan could result in sCFRs that are higher than they would be in a less stressed healthcare setting, as presumably the sicker patients would have been prioritized for admission while leaving the milder cases untested and thus unconfirmed. Our prevalence estimates relying on travelers are based on those well enough to travel, so may slightly underestimate prevalence in Wuhan by not including those who are already in a serious condition and perhaps hospitalized. We have accounted for the possibility that travelers may underestimate the prevalence of infection in Wuhan[17] by using our best estimate, from a separate analysis, of the probability of detection for international travelers (38% (22–64%))[17]. On the other hand, the numerator of the number of deaths could also have been undercounted, although much less likely compared to enumerating the denominator, for the same surge capacity reason or due to imperfect test sensitivity, especially during the first month of the outbreak[18]. If deaths in Wuhan were under-ascertained, this would bias our severity estimates downward.

Another caveat concerns one of our key inputs—the infection prevalence among returnees airlifted out of Wuhan on charter flights. Their point prevalence might well be lower than that among local residents, because of a generally more advantaged socioeconomic background, and the sensitivity for detecting infected

individuals among them might not be 100%, as assumed. As such, this would be a lower bound of the cross-sectional disease prevalence. If this were the case, then we would have overestimated the reduction in transmissibility conferred by public health interventions in Wuhan and overestimated the severity. Based on only publicly available data, there is necessarily substantial uncertainty in our estimates of the effectiveness of intra-Wuhan public health interventions in reducing transmissibility. Calculating the instantaneous reproductive number from a set of line lists that are updated daily would be the most reliable method for detecting changes in transmissibility associated with interventions.

There has been refinement of case definitions at both national and provincial levels, such as excluding RT-PCR-test-positive asymptomatics (perhaps, in fact, very mildly symptomatics) from being labeled an officially 'confirmed' case[19] or including test-naïve clinically diagnosed cases with clear epidemiologic links as 'confirmed'[20]. Although these should not affect our estimation given our data sources from the earlier phase of the epidemic, such changes in the reporting criteria may influence the interpretation of future data. Finally, given that Wuhan is no longer the only (albeit the first) location with sustained local spread, it would be important to assess and take into account the experience from elsewhere, both domestically in mainland China and overseas. These secondary epicenters, having learned from the early phase of the Wuhan epidemic, might have had a systematically different epidemiology and response that could impact the parameters estimated here[21–31].

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/s41591-020-0822-7.

References

1. Ghani, A. et al. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *Am. J. Epidemiol.* **162**, 479–486 (2005).
2. Wong, J. Y. et al. Case fatality risk of influenza A (H1N1pdm09): a systematic review. *Epidemiology* **24**, 830–841 (2013).
3. Yu, H. et al. Human infection with avian influenza A H7N9 virus: an assessment of clinical severity. *Lancet* **382**, 138–145 (2013).
4. 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* (2020); https://doi.org/10.1016/S0140-6736(20)30260-9
5. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* (2020); https://doi.org/10.1016/S0140-6736(20)30462-1
6. World Health Organization. *Report of the WHO–China Joint Mission on Coronavirus Disease 2019* (*COVID-19*), *16–24 February 2020* (2020); https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf
7. Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Eng. J. Med.* (2020); https://doi.org/10.1056/NEJMoa2001316
8. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China. *China CDC Weekly* **2**, 113–122 (2020).
9. Chinese Center for Disease Control and Prevention. *Dashboard of Reported 2019-nCoV Cases* (2020); http://2019ncov.chinacdc.cn/2019-nCoV/
10. Data Platform of Shanghai Observer. *Line List of 2019-nCoV Confirmed Fatal Cases* (*from publicly available information*) (2020); http://data.shobserver.com/www/index.html#/home
11. Wuhan Municipal Health Commission. *Wuhan Municipal Health Commission's Briefing on the Current Pneumonia Epidemic in the City* (2020); http://wjw.wuhan.gov.cn/
12. Hubei Municipal Health Commission. *Hubei Municipal Health Commission's Briefing on the Current Pneumonia Epidemic in the Province* (2020); http://wjw.hubei.gov.cn/fbjd/dtyw/
13. Wong, J. Y. et al. Infection fatality risk of the pandemic A(H1N1)2009 virus in Hong Kong. *Am. J. Epidemiol.* **177**, 834–840 (2013).
14. Zou, L. et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Eng. J. Med.* (2020); https://doi.org/10.1056/NEJMc2001737
15. Pan, Y., Zhang, D., Yang, P., Poon, L. L. M. & Wang, Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Inf. Dis.* (2020); https://doi.org/10.1016/S1473-3099(20)30113-4
16. Callaway, E. Time to use the p-word? Coronavirus enters dangerous new phase. *Nature* (2020); https://www.nature.com/articles/d41586-020-00551-1
17. Niehus, R., De Salazar, P. M., Taylor, A. & Lipsitch, M. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. Preprint at *medRxiv* https://doi.org/10.1101/2020.02.13.20022707 (2020).
18. The State Council of The People's Republic of China. *Press Conference of the Joint Prevention and Control Mechanism of the State Coucil* (2020); http://www.gov.cn/xinwen/2020-02/09/content_5476513.htm
19. National Health Commission of People's Republic of China. *Notice of the General Office of the National Health Commission on the Distribution of the Plan of Prevention and Control of the Pneumonia Caused by the Novel Coronavirus* (*Version 4*) (2020); http://www.gov.cn/zhengce/zhengceku/2020-02/07/content_5475813.htm
20. Hubei Municipal Health Commission. *Situation of the Epidemic of Pneumonia caused by the Novel Coronavirus in Hubei, as of Feb 12* (2020); http://wjw.hubei.gov.cn/fbjd/dtyw/202002/t20200213_2025581.shtml
21. Jia, N. et al. Case fatality of SARS in mainland China and associated risk factors. *Tropical Med. Int. Health* **14**, 21–27 (2009).
22. Donnelly, C. A. et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet* **361**, 1761–1766 (2003).
23. Leung, G. M. et al. The epidemiology of severe acute respiratory syndrome in the 2003 Hong Kong epidemic: an analysis of all 1755 patients. *Ann. Intern. Med.* **141**, 662–673 (2004).
24. Lau, E. H. et al. A comparative epidemiologic analysis of SARS in Hong Kong, Beijing and Taiwan. *BMC Infect. Dis.* **10**, 50 (2010).
25. Taubenberger, J. K. & Morens, D. M. 1918 influenza: the mother of all pandemics. *Rev. Biomed.* **17**, 69–79 (2006).
26. Collins, S. D. Age and sex incidence of influenza in the epidemic of 1943–44, with comparative data for preceding outbreaks: based on surveys in Baltimore and other communities in the Eastern States. *Public Health Rep. (1896–1970)* **59**, 1483–1503 (1944).
27. Andreasen, V., Viboud, C. & Simonsen, L. Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *J. Inf. Dis.* **197**, 270–278 (2008).
28. Donaldson, L. J. et al. Mortality from pandemic A/H1N1 2009 influenza in England: public health surveillance study. *BMJ* **339**, b5213 (2009).
29. Oh, M. D. et al. Middle East respiratory syndrome: what we learned from the 2015 outbreak in the Republic of Korea. *Korean J. Intern. Med.* **33**, 233–246 (2018).
30. Wong, J. Y. et al. Hospitalization fatality risk of influenza A (H1N1)pdm09: a systematic review and meta-analysis. *Am. J. Epidemiol.* **182**, 294–301 (2015).
31. Abbad, A. et al. Middle East respiratory syndrome coronavirus (MERS-CoV) neutralising antibodies in a high-risk human population, Morocco, November 2017 to January 2018. *Euro. Surveill.* **24**, 1900244 (2019).

Methods

We made the following assumptions in the model:

1. The population of Wuhan is stratified into $m = 9$ age groups: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79 and >79. The relative susceptibility to infection of age group i is α_i with respect to those aged 30–39 years (that is, $\alpha_4 = 1$). The sCFR of age group i is sCFR$_i$.

2. The probability density function (pdf) of the incubation period, $f_{\text{incubation}}$, is gamma, with a mean of 6.5 days and standard deviation of 2.6 days[32].

3. The pdf of the time between onset and death, $f_{\text{onset-to-death}}$, is gamma. We inferred the values of the mean and standard deviation of $f_{\text{onset-to-death}}$ (Extended Data Fig. 1).

4. The pdf of the generation time, f_{GT}, is gamma and the same as that of the serial interval. We inferred the values of the mean and standard deviation of f_{GT} (Extended Data Fig. 1).

5. The infection-symptomatic probability (P_{sym}; the proportion of infections that progress to develop symptoms) is the same for all age groups. We assume $P_{\text{sym}} = 0.50$ in the baseline scenario and 0.75 and 0.95 in alternate scenarios.

6. The sensitivity of detecting symptomatic cases exported from mainland China is $P_{\text{det}} = 38\%$ (22%–64%) for cities that reported case importation between 25 December 2019 and 19 January 2020 (Supplementary Table 2)[17].

7. Inbound and outbound mobility in Wuhan had been reduced by ~90% for mainland Chinese cities (https://qianxi.baidu.com/) and 99% for international cities since Wuhan was quarantined on 23 January 2020.

8. The diagnostic test for the charter flight passengers is 100% sensitive and 100% specific for detecting COVID-19 infections.

9. Recent phylogenetic analyses suggest that the most recent common ancestor of the sequenced COVID-19 genomes emerged between 23 October and 16 December 2019 (http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347; accessed 12 Feb 2020). As such, we assume that the epidemic in Wuhan was seeded by a single zoonotic event that generated z_0 infections on 15 November 2019. We inferred the value of z_0 (Extended Data Fig. 1).

10. Public health interventions in Wuhan reduced local transmissibility by φ_0. We inferred the value of φ_0 (Extended Data Fig. 1).

11. Given that the epidemic curve in Wuhan was weeks ahead of that in other mainland Chinese cities, we ignored the effect of case importation at Wuhan.

These assumptions were reflected in the following susceptible–infected–recovered (SIR) model for simulating the COVID-19 epidemic in Wuhan, where $S_i(t)$, and $R_i(t)$ are the number of susceptible and recovered individuals in age group i at time t, and $I_i(t, \tau)$ is the number of infected individuals in age group i at time t who were infected at time $t - \tau$:

$$\frac{dS_i(t)}{dt} = -\alpha_i S_i(t)\pi(t) + \frac{N_i(t)}{N(t)} L_{\text{inbound}}(t) - \frac{S_i(t)}{N(t)} L_{\text{outbound}}(t)$$

$$\frac{\partial I_i(t, \tau)}{\partial t} + \frac{\partial I_i(t, \tau)}{\partial \tau} = -f_{\text{GT}}(\tau) I_i(t, \tau) - \frac{I_i(t, \tau)}{N(t)} L_{\text{outbound}}(t)$$

$$I_i(t, 0) = \alpha_i S_i(t)\pi(t)$$

$$\frac{dR_i(t)}{dt} = \int_0^t f_{\text{GT}}(\tau) I_i(t, \tau) d\tau - \frac{R_i(t)}{N(t)} L_{\text{outbound}}(t)$$

$$N_i(t) = S_i(t) + \int_0^t I_i(t, \tau) d\tau + R_i(t), \quad N(t) = \sum_{i=1}^m N_i(t)$$

$$\pi(t) = \frac{\beta(1 - \varphi(t))}{N(t)} \sum_{i=1}^m \int_0^t I_i(t, \tau) d\tau$$

$$\varphi(t) = \begin{cases} 0 & \text{before 23 January 2020} \\ \varphi_0 & \text{otherwise} \end{cases}$$

$$L_{\text{outbound}}(t) = L_{\text{W,I}}(t) + L_{\text{W,C}}(t)$$

The next-generation matrix for this SIR model is

$$\frac{\beta T_G}{N} \begin{bmatrix} \alpha_1 N_1 & \cdots & \alpha_1 N_1 \\ \vdots & \ddots & \vdots \\ \alpha_m N_m & \cdots & \alpha_m N_m \end{bmatrix}$$

where T_G is the mean generation time. The basic reproductive number R_0 is the largest eigenvalue of this matrix, which is $\frac{\beta T_G}{N} \sum_{i=1}^m \alpha_i N_i$. The incidence rates of infection, onset and death for age group i at time t are calculated as follows:

$$A_{i,\text{infection}}(t) = \alpha_i S_i(t)\pi(t)$$

$$A_{i,\text{onset}}(t) = P_{\text{sym}} \int_0^t A_{i,\text{infection}}(u) f_{\text{incubation}}(t - u) du$$

$$A_{i,\text{death}}(t) = \text{sCFR}_i \int_0^t A_{i,\text{onset}}(u) f_{\text{onset-to-death}}(t - u) du$$

The number of new cases (onset) and the cumulative number of cases in age group i on day d are $\omega_{d,i} = \int_{d-1}^d A_{i,\text{onset}}(t)dt$ and $\Omega_{d,i} = \int_0^d A_{i,\text{onset}}(t)dt$, respectively. The cumulative number of death cases in age group i up to time t is $D_i(t) = \int_0^t A_{i,\text{death}}(u)du$. Let $\omega_d = \sum_{i=1}^m \omega_{d,i}$, $\Omega_d = \sum_{i=1}^m \Omega_{d,i}$ and $D(t) = \sum_{i=1}^m D_i(t)$ be the summation of the number of new cases, the cumulative number of cases and the cumulative number of deaths across all age groups up to time t, respectively. Similarly, $I(t) = \sum_{i=1}^m \int_0^t I_i(t, \tau)d\tau$ is the total number of infected individuals at time t.

We inferred the parameters listed in Extended Data Fig. 1 assuming that the remaining parameters are fixed at the values shown in Extended Data Fig. 4. We use θ to denote the set of parameters that are subject to inference (Extended Data Fig. 1). The likelihood function is a product of several components associated with the data in Supplementary Tables 1–8:

$$L(\theta) = \prod_{k=1}^8 L_k(\theta)$$

The formulation of each component was as follows:

1. The number of observed international case exportations on each day is assumed to be an imperfect Poisson observation of the number of infected travelers leaving Wuhan on that day who had or would develop symptoms. Let x_d be the observed number of such international case exportations on day d between 25 December 2019 ($D_{s,1}$) and 19 January 2020 ($D_{e,1}$) based on the data in Supplementary Table 2. We assume that travel behavior is not affected by disease and hence such case exportation occurs according to a non-homogeneous process with rate $\lambda(t) = P_{\text{sym}} \frac{L_{\text{W,I}}(t)}{N(t)} I(t)$. Let P_{det} be the probability that an infected traveler who has or will develop symptoms is detected in the destination country. The expected number of detected case exportations on day d is $\lambda_d = P_{\text{det}} \int_{d-1}^d \lambda(u) du$ and hence $x_d \approx \text{Poisson}(\lambda_d)$. As such, the likelihood function associated with the data in Supplementary Table 2 is

$$L_1(\theta) = \int_0^1 \prod_{d=D_{s,1}}^{D_{e,1}} \frac{e^{-\lambda_d} \lambda_d^{x_d}}{x_d!} g(P_{\text{det}}) dP_{\text{det}}$$

where g is the posterior distribution of P_{det} from a separate study that had a mean of 38% and a 95% credible interval of 22–64%[17].

2. Let y_d be the observed number of confirmed cases of COVID-19 in Wuhan with no epidemiologic links to Huanan Seafood Wholesale Market (which is presumed to be the index zoonotic source of the COVID-19 epidemic) on day d between 10 December 2019 ($D_{s,2}$) and 3 January 2020 ($D_{e,2}$) based on the data in Supplementary Table 1[7]. These cases are assumed to be a Poisson observation of the true number of newly symptomatic cases on that day, with ascertainment rate ε, which remained fixed over this time period. As such, assuming $y_d \approx \text{Poisson}(\varepsilon \omega_d)$, the likelihood function for the data in Supplementary Table 1 is

$$L_2(\theta) = \prod_{d=D_{s,2}}^{D_{e,2}} \frac{e^{-\varepsilon \omega_d}(\varepsilon \omega_d)^{y_d}}{y_d!}$$

3. We consider the test results of entry screening among expatriates and visitors on returning to their countries from Wuhan on charter flights between 29 January 2020 ($D_{s,3}$) and 4 February 2020 ($D_{e,3}$). Let m_d^{all} be the number of such passengers on day d who were tested regardless of symptoms (for example, Japan, Germany, South Korea and so on; Supplementary Table 3) and m_d^{sym} be the number of such passengers on day d who were probably tested only if they showed symptoms (for example, United States, United Kingdom, Thailand, Australia and so on; Supplementary Table 3). Let u_d^{all} and u_d^{sym} be the respective observed number of passengers who were confirmed to be infected based on the data in Supplementary Table 3. The prevalence of infection and symptoms among travelers are assumed to reflect a representative binomial sample of the same quantities in the Wuhan population on their day of departure. The likelihood function associated with the data in Supplementary Table 3 is

$$L_3(\theta) = \prod_{d=D_{s,3}}^{D_{e,3}} \binom{m_d^{\text{all}}}{u_d^{\text{all}}} q_d^{u_d^{\text{all}}} (1 - q_d)^{m_d^{\text{all}} - u_d^{\text{all}}} \binom{m_d^{\text{sym}}}{u_d^{\text{sym}}} (P_{\text{sym}} q_d)^{u_d^{\text{sym}}} (1 - P_{\text{sym}} q_d)^{m_d^{\text{sym}} - u_d^{\text{sym}}}$$

where $q_d = I(d)/N(d)$ is the proportion of individuals who were infected on day d.

4. We assume that all deaths from COVID-19 infection in Wuhan were confirmed. Let G be the cumulative number of death cases in Wuhan as of 25 February 2020 (time T). We assume $G \approx \text{Poisson}(D(T))$ and hence the likelihood function associated with this data is

$$L_4(\theta) = \frac{\mathrm{e}^{-D(T)} D(T)^G}{G!}$$

5. We assume that the age distribution of confirmed cases is a multinomial sampling process from the age distribution of true cases. Let c_i be the observed number of confirmed cases in age group i in Wuhan based on the data in Supplementary Table 4. The likelihood function for the data in Supplementary Table 4 is

$$L_5(\theta) = \frac{(c_1 + c_2 + c_3)!}{c_1! c_2! c_3!} \prod_{i=1}^{m} \left(\frac{\Omega_{T,i}}{\Omega_T} \right)^{c_i}$$

6. We assume that the age distribution of confirmed deaths is a multinomial sampling process from the age distribution of true deaths. Given that most COVID-19 deaths were Wuhan-related, we assume that the age distribution of confirmed deaths for Wuhan is the same as that for mainland China[8]. Let b_i be the observed number of death cases in age group i in Wuhan based on the data in Supplementary Table 5. The likelihood function for the data in Supplementary Table 5 is

$$L_6(\theta) = \frac{(b_1 + b_2 + b_3)!}{b_1! b_2! b_3!} \prod_{i=1}^{m} \left(\frac{D_i(T)}{D(T)} \right)^{b_j}$$

7. With regard to the data in Supplementary Table 7, let A be the set of death cases whose onset dates are known, and B the set comprising the remaining cases. Let v_j be the observed time delay between onset and death for the jth case in A and let v_j^L be the observed time between hospital admission and death (which serves as a lower bound for the delay between onset and death) for the jth case in B. The likelihood function for the data in Supplementary Table 7 is

$$L_7(\theta) = \prod_{j \in A} f_{\text{onset-death}}\left(v_j | \theta \right) \prod_{j \in B} \left(1 - F_{\text{onset-death}}\left(v_j^L | \theta \right) \right)$$

where $f_{\text{onset-death}}$ and $F_{\text{onset-death}}$ are the pdf and cumulative density function (cdf) of the time between onset and death (assumed to be gamma-distributed with mean μ_D and standard deviation σ_D).

8. With regard to the data in Supplementary Table 8, let A be the set of infector–infectee pairs for whom the serial interval (time elapsed between their onset dates) is known and B the set comprising the remaining pairs for whom only the ranges of their serial intervals are known. Let s_j be the observed value of the serial interval for the jth pair in A, and $\left(s_j^L, s_j^U \right)$ be the observed range of the serial interval for the jth pair in B. For some infector–infectee pairs, the travel history and onset dates of the infector impose a lower bound on the serial interval (Supplementary Table 8). Let s_j^* be such a lower bound for the jth pair. The likelihood function for the data in Supplementary Table 8 is

$$L_8(\theta) = \prod_{j \in A} \frac{f_{\text{SI}}\left(s_j | \theta \right)}{1 - F_{\text{SI}}\left(s_j^* | \theta \right)} \prod_{j \in B} \frac{F_{\text{SI}}\left(s_j^U | \theta \right) - F_{\text{SI}}\left(s_j^L | \theta \right)}{1 - F_{\text{SI}}\left(s_j^* | \theta \right)}$$

where f_{SI} and F_{SI} are the pdf and cdf of the serial interval. We assume that the serial interval and the generation time have the same pdf.

We estimated the model parameters θ using Markov chain Monte Carlo methods with Gibbs sampling and non-informative flat priors. Point estimates and statistical uncertainty are presented using posterior means and 95% CrIs, respectively.

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

Data availability
We collated epidemiological data from publicly available data sources (news articles, press releases and published reports from public health agencies). All the epidemiological information that we used is documented in the main text, the extended data and supplementary tables.

Code availability
The codes are available upon request to the corresponding author.

References

32. Backer, J. A., Klinkenberg, D. & Wallinga, J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Euro. Surveill.* **25**, 2000062 (2020).

Author contributions
J.T.W., M.L. and G.M.L. contributed to conceptualization, data analysis, results interpretation and manuscript writing. K.L. contributed to conceptualization, data collection, data analysis, results interpretation and manuscript writing. M.B., N.K., R.N. and P.M.d.S. contributed to data analysis and results interpretation. B.J.C. contributed to results interpretation.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0822-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0822-7.

Correspondence and requests for materials should be addressed to J.T.W.

Peer review information Joao Monteiro was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Parameters	Description
R_0	Basic reproductive number, which was linearly related to the transmission rate β
φ_0	Relative reduction in transmissibility after vs. before lockdown of Wuhan on 23 January 2020
z_0	The number of infections generated by the seeding zoonotic event
$sCFR_i$	The symptomatic case-fatality risk for age group i, $i = 1, \ldots, m$
α_i	The relative susceptibility to infection of age group i compared to those aged 30-39 years, $i = 1, \ldots, m$
(μ_D, σ_D)	Mean and standard deviation for the time between onset and death
(μ_{SI}, σ_{SI})	Mean and standard deviation of the serial interval
ε	The proportion of human-to-human infections that were confirmed between 1 December 2019 and 4 January 2020

Extended Data Fig. 1 | Model parameters that were subject to statistical inference. Epidemiologic parameters fitted in the model.
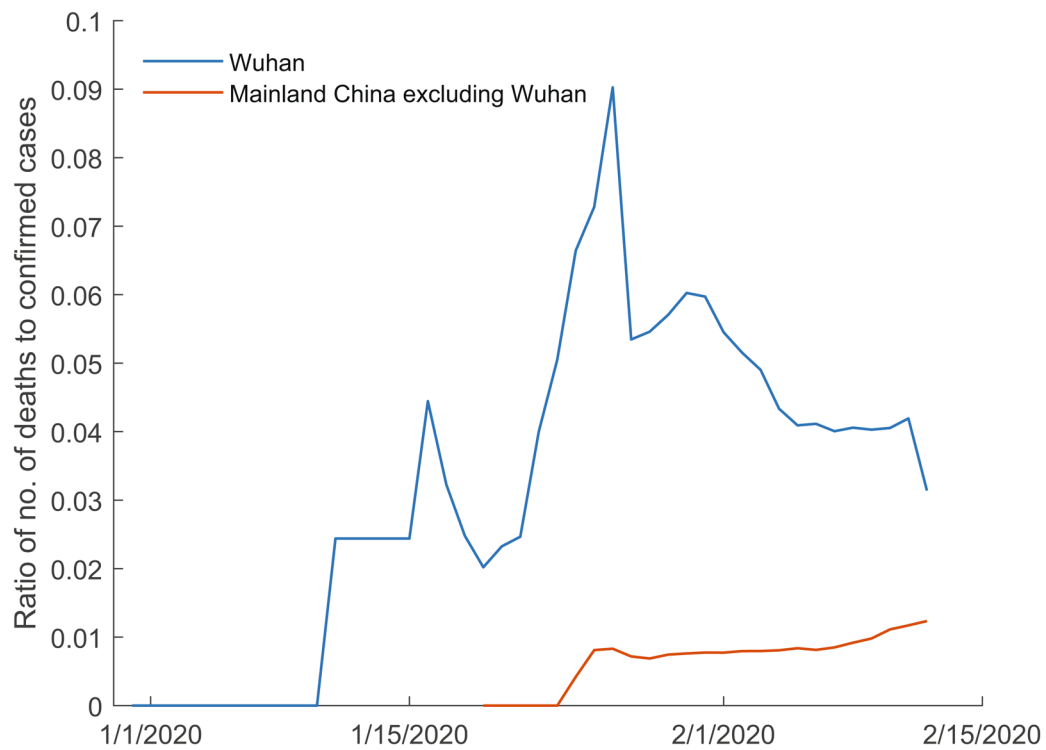
Extended Data Fig. 2 | The ratio of no. of deaths to confirmed cases (crude confirmed case-fatality ratio) in Wuhan and in cities of mainland China other than Wuhan. Blue line shows the ratio of the number of deaths to the number of confirmed cases in Wuhan and the red line shows the ratio locations within mainland China outside Wuhan.

Parameter	SARS*	MERS†	1918 influenza pandemic	2009 influenza pandemic
Infection fatality risk (IFR) (risk of death among all infections)	Worldwide (WHO)** [1] 9.6% (774/8096) Mainland China** [21] 6.4% (343/5327)	--	Worldwide [25,26] 2.5% Copenhagen [27] 1.7%	Hong Kong [13] <60 yrs: <0.1% ≥60 yrs: 1.1% (0.2-4.7)
Symptomatic case fatality risk (sCFR) (risk of death among symptomatic infections)	Hong Kong*** [22,23] Overall: 17.2% (302/1755) <60 yrs: 13.2% (9.8-16.8) ≥60 yrs: 43.3% (35.2-52.4)	--	--	United Kingdom [28] Overall: 0.026% (0.011-0.066) 5-14 yrs: 0.011% (0.003-0.036) ≥65 yrs: 0.98% (0.30-3.2)
Hospitalization fatality risk (HFR) (risk of death among infections that require hospitalization for medical reasons, not only for case isolation)	Taiwan*** [24] Overall: 27.6% (180/664) <60 yrs: 15.3% (72/470) ≥60 yrs: 48.6% (88/181)	Worldwide (WHO)*** 34.4% (858/2494) Saudi Arabia*** [29] 40.7% (726/1783) South Korea*** [29] 20.4% (38/186)		North America [30] Overall: 2.6% (1.6, 3.9) ≤19 yrs: 0.8% (0.5, 1.1) 20-64 yrs: 5.4% (3.5, 7.5) ≥65 yrs: 10.7% (5.3, 17.6)

* IFR=CFR=HFR given virtually every infected person required hospitalization and seroprevalence amongst close contacts and in general community approximated zero
** Among probable cases
*** Among laboratory-confirmed cases
† There are more infections "undetected" in MERS compared with SARS. The seroprevalence amongst individuals exposed to camels was estimated as 6.2% in Arabian Peninsula [31].

Extended Data Fig. 3 | A summary of severity estimates among pandemic influenza strains and coronaviruses with pandemic potential in the past.
Severity estimates of SARS (2002-3), MERS (2014-), 1918 influenza pandemic (1918-20) and 2009 influenza pandemic (2009-10).

Data	Description	Value
$L_{inbound}(t)$	Daily number of persons entering Wuhan (assumed all susceptible)	International: 3,546; decreased by 99% after Wuhan was quarantined on 23 January 2020 Mainland China: 487,310 and 810,500 before and during *chunyun*; decreased by 90% after Wuhan was quarantined on 23 January 2020
$L_{W,I}(t)$	Daily number of persons leaving Wuhan to destinations outside mainland China	3,633; decreased by 99% after Wuhan was quarantined on 23 January 2020
$L_{W,C}(t)$	Daily number of persons leaving Wuhan to destinations in mainland China	502,013 and 717,226 before and during *chunyun*; decreased by 90% after Wuhan was quarantined on 23 January 2020
$N_i(0)$	Number of Wuhan residents in age group i	See Table S9
P_{sym}	Infection-symptomatic probability	0.5 in the baseline scenario; 0.75 and 0.95 in alternative scenarios

Extended Data Fig. 4 | Model parameters that were assumed to be constant. Assumed constants in the model.

nature research

Corresponding author(s): Joseph T. Wu

Last updated by author(s): Mar 5, 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 Authors & Referees 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	R 3.6.2 MATLAB 2019b
Data analysis	R 3.6.2 MATLAB 2019b

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/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

All raw data have been provided in supplementary tables.

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

Life sciences study design

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

Sample size	Not applicable. The study is a mathematical modeling study. We have provided the full information of data in Figures, Extended Data and Supplementary Tables.
Data exclusions	Not applicable. The study is a mathematical modeling study. We have provided the full information of data in Figures, Extended Data and Supplementary Tables.
Replication	All the data have been provided in Figures, Extended Data and Supplementary Tables. The codes will be provided upon the request to the corresponding author.
Randomization	Not applicable. The study is a mathematical modeling study. We have provided the full information of data in Figures, Extended Data and Supplementary Tables.
Blinding	Not applicable.Not applicable. The study is a mathematical modeling study. We have provided the full information of data in Figures, Extended Data and Supplementary Tables.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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