Clinical Infectious Diseases

MAJOR ARTICLE

Reconstruction of Transmission Pairs for Novel Coronavirus Disease 2019 (COVID-19) in Mainland China: Estimation of Superspreading Events, Serial Interval, and Hazard of Infection

Xiao-Ke Xu,[1,a] Xiao Fan Liu,[2,a] Ye Wu,[3,4,a] Sheikh Taslim Ali,[5,a] Zhanwei Du,[6,a] Paolo Bosetti,[7] Eric H. Y. Lau,[5] Benjamin J. Cowling,[5] and Lin Wang[7,8]

[1]College of Information and Communication Engineering, Dalian Minzu University, Dalian, China, [2]Web Mining Laboratory, Department of Media and Communication, City University of Hong Kong, Hong Kong Special Administrative Region, China, [3]Computational Communication Research Center, Beijing Normal University, Zhuhai, China, [4]School of Journalism and Communication, Beijing Normal University, Beijing, China, [5]World Health Organization Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong Special Administrative Region, China, [6]Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA, [7]Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, Centre National de la Recherche Scientifique (CNRS), Paris, France, and [8]Department of Genetics, University of Cambridge, Cambridge, United Kingdom

Background. Knowledge on the epidemiological features and transmission patterns of novel coronavirus disease (COVID-19) is accumulating. Detailed line-list data with household settings can advance the understanding of COVID-19 transmission dynamics.

Methods. A unique database with detailed demographic characteristics, travel history, social relationships, and epidemiological timelines for 1407 transmission pairs that formed 643 transmission clusters in mainland China was reconstructed from 9120 COVID-19 confirmed cases reported during 15 January–29 February 2020. Statistical model fittings were used to identify the superspreading events and estimate serial interval distributions. Age- and sex-stratified hazards of infection were estimated for household vs nonhousehold transmissions.

Results. There were 34 primary cases identified as superspreaders, with 5 superspreading events occurred within households. Mean and standard deviation of serial intervals were estimated as 5.0 (95% credible interval [CrI], 4.4–5.5) days and 5.2 (95% CrI, 4.9–5.7) days for household transmissions and 5.2 (95% CrI, 4.6–5.8) and 5.3 (95% CrI, 4.9–5.7) days for nonhousehold transmissions, respectively. The hazard of being infected outside of households is higher for people aged 18–64 years, whereas hazard of being infected within households is higher for young and old people.

Conclusions. Nonnegligible frequency of superspreading events, short serial intervals, and a higher risk of being infected outside of households for male people of working age indicate a significant barrier to the identification and management of COVID-19 cases, which requires enhanced nonpharmaceutical interventions to mitigate this pandemic.

Keywords. COVID-19; transmission; superspreading event; serial interval; hazard of infection.

In December 2019, a novel coronavirus disease (COVID-19) emerged in Wuhan of Hubei Province, China. The World Health Organization announced a public health emergency of international significance on 30 January 2020 [1] and classified the threat as a global pandemic on 11 March 2020 [2]. More than 8 million confirmed cases and 440 290 deaths have been reported from >200 countries and territories as of 17 June 2020 [3].

On 23 January 2020, China raised the national emergency response to the highest level, which triggered an unprecedented travel ban starting from the lockdown of Wuhan on 23 January, 14 cities in Hubei province on 24 January, and >30 provinces thereafter. Despite this unprecedented intervention, we estimated that COVID-19 cases had been introduced into 130 (95% credible interval [CrI], 190–369) cities in mainland China prior to the lockdown of Wuhan on 23 January 2020 [4]. Similar findings on the rapid geographic expansion of COVID-19 have also been reported in several recent studies [5–8]. Starting from the last week of January 2020, >260 Chinese cities have implemented intensive social distancing and confinement policies, which brought the epidemic under control [7–10]. However, the epidemic has still caused >10 000 confirmed cases in China outside Hubei Province within a month.

To enhance public health preparedness and awareness, Chinese health authorities have publicly reported detailed records of confirmed COVID-19 cases since mid-January. This provides a unique resource for studying the transmission patterns, routes, and risk factors of COVID-19.

Received 15 April 2020; editorial decision 3 June 2020; accepted 11 June 2020; published online June 18, 2020.

[a]X.-K. X., X.-F. L., Y. W., S. T. A., and Z. D. contributed equally to this work.

Correspondence: L. Wang, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK (lw660@cam.ac.uk).

Clinical Infectious Diseases® 2020;71(12):3163–7

DOI: 10.1093/cid/ciaa790

METHODS

Data Collection

In mainland China, 27 provincial and 264 urban health commissions have publicly posted 9120 confirmed case reports online during 15 January–29 February 2020, which accounts for 72% of all cases confirmed in mainland China outside Hubei province. We compiled a unique line-list database using these reports, which contains detailed information about demographic features, social relationships, travel history, and key epidemiological timelines (eg, dates of symptom onset, confirmation, and hospitalization). In contrast to several published COVID-19 data repositories [11–16], which focus on describing information about individual cases, our database allows the reconstruction of transmission pairs and clusters by inferring potential causal associations among different cases. The entire dataset of transmission pairs is available at our GitHub (https://github.com/linwangidd/covid19_transmissionPairs_China). See the Supplementary Materials for more details.

Statistical Analysis

We reconstructed 1407 transmission pairs using the epidemiological evidence among reported cases. The section "Reconstruction of Transmission Pairs" in the Supplementary Materials specifies how we identified a pair or a group of confirmed cases using information about their close contacts, stratified transmission pairs into household and nonhousehold settings using information about familial relationships, and determined the direction of transmission between infector and infectee using information about travel histories. For each transmission pair, we term the infector the "primary case" and the infectee the "secondary case." We also consider connected chains of confirmed cases, in which we term the original case the "index" and the entire chain of cases, including the index, the "transmission cluster" (Figure 1A).

We categorized each transmission pair by the social relationship between primary and secondary cases (eg, familial members of the same household, nonhousehold relatives, colleagues, classmates, friends, and other face-to-face contacts). Considering that during the Spring Festival travel season (10 January–18 February 2020), several billion human movements can occur because of the tradition of Chinese New Year (to visit and live with their original families), we considered any transmission pair with immediate familial relationships (eg, a person's spouse, parents, and children) as a household transmission pair, and with other familial relationships (eg, a person's siblings with age >17 years) or close contacts with no familial information (eg, classmates, colleagues) as a nonhousehold transmission pair. The numbers of household (662) and nonhousehold (745) transmission pairs are almost even.

Following Lloyd-Smith et al [17], we defined the threshold of observing superspreading events (SSEs) as the 99th percentile of the offspring distribution for the number of secondary cases caused by a primary case. Household and nonhousehold transmissions were combined together for computing the offspring distribution. To estimate the threshold of observing SSE, we used a Poisson, exponential, and power-law distribution to fit the empirical offspring distribution via the Distribution Fitter App in Matlab R2020a [18]. Since the power-law distribution gives the smallest Akaike information criterion compared to the Poisson and exponential distributions (Supplementary Table 2), the threshold of observing SSEs is set as 3.78, which indicates the occurrence of an SSE if 4 or more secondary cases were infected by a single primary case.

For each transmission pair with known symptom onset times for both primary and secondary cases, we computed the empirical serial interval as the number of days between the symptom onset dates of the primary case and of the secondary case [19]. Due to the presence of negative-valued serial intervals and the skewness of the empirical distribution (Figure 1B), we estimated the serial interval distributions by fitting a normal distribution via the Markov chain Monte Carlo (MCMC) method with Gibbs sampling and noninformative flat prior. We confirmed the convergence of MCMC chains via trace plot and diagnosis, and obtained the posterior estimates of parameters by running 100 000 iterations with a burn-in of 40 000 iterations and a thinning interval of 10. Fitting serial interval data with a Gumbel or logistic distribution gives similar estimates (Supplementary Tables 3–5).

We estimated the age-stratified hazard of infection $\gamma_{H/N}(a,b)$ for household vs nonhousehold transmissions by the ratio between the probability $P_H(a,b)$ that a secondary case of age group b was infected by a primary case of age group a within the same household and the probability $P_N(a,b)$ that a secondary case of age group b was infected by a primary case of age group a outside of households, that is, $\gamma_{H/N}(a,b) = P_H(a,b)/P_N(a,b)$. If $\gamma_{H/N}(a,b) > 1$, then the infection within households has a higher risk than the infection outside of households for secondary cases of age group b being infected by primary cases of age group a.

We estimated the sex-specific hazard of infection for household vs nonhousehold transmissions by the ratio between the probability that a secondary case of gender b was infected by a primary case of gender a within the same household and the probability that a secondary case of gender b was infected by a primary case of gender a via nonhousehold transmission.

RESULTS

We in total reconstructed 643 transmission clusters formed by 1407 transmission pairs (Figure 1A). The size is <5 for 587 transmission clusters, whereas the size exceeds 20 for the largest cluster. We identified 34 primary cases as the superspreaders, with 5 SSEs occurring within households. Stratification by

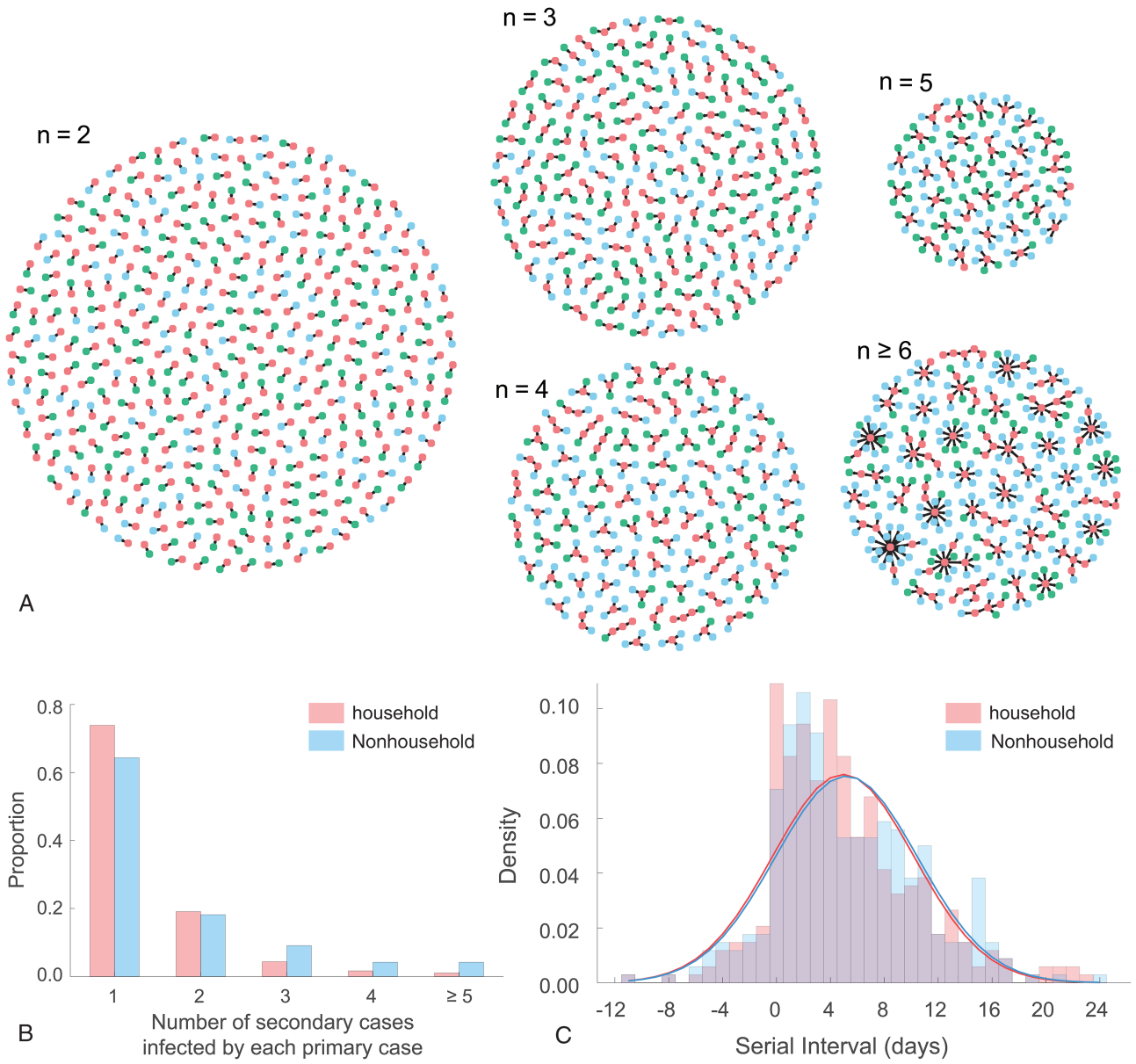

Figure 1. COVID-19 transmission clusters. *A*, Six hundred forty-three transmission clusters, stratified by the size of cluster n. Red, green, and blue nodes denote primary cases, household secondary cases, and nonhousehold secondary cases, respectively. *B*, Empirical offspring distribution for the number of secondary cases infected by each of the 809 primary cases. *C*, Empirical and estimated serial interval distributions for household or nonhousehold transmissions. Pink and blue bars indicate the empirical distributions of serial interval data for household and nonhousehold transmissions, respectively. Red curve indicates the estimated serial interval distributions for household transmissions, based on the posterior median estimates of the mean and standard deviation (SD) of normal distribution. Blue curve indicates the estimated serial interval distributions for nonhousehold transmissions, based on the posterior median estimates of the mean and SD of normal distribution. Fitting results with alternative distributions (eg, Gumbel or Logistic distribution) are available in Supplementary Tables 3–5.

household setting demonstrates that 356, 92, and 34 primary cases infected 1, 2, and ≥3 secondary cases within households, respectively, and 276, 78, and 75 primary cases infected 1, 2, and ≥3 secondary cases outside of households, respectively (Figure 1B).

Fitting a normal distribution to serial interval data for household transmissions estimates the mean and standard deviation (SD) of serial intervals as 5.0 (95% CrI, 4.4–5.5) days and 5.2 (95% CrI, 4.9–5. 7) days, respectively. Given the posterior median estimates of the mean and SD, the median serial interval distribution is estimated to be 5.0 (interquartile range [IQR], 1.5–8.5) days for household transmissions. Fitting a normal distribution to serial interval data for nonhousehold transmissions estimates the mean and SD of serial intervals as 5.2 (95% CrI, 4.6–5.8) days and 5.3 (95% CrI, 4.9–5.7) days, respectively. Given the posterior median estimates of the mean and SD, the median serial interval distribution is estimated to be 5.2 (IQR, 1.6–8.8) days for nonhousehold transmissions. See Supplementary Table

3 for more results. Notably, 25 of 339 household and 25 of 340 nonhousehold transmission pairs reported negative-valued serial intervals, implying presymptomatic transmission.

We performed several sensitivity analyses on estimating serial interval distributions (Supplementary Tables 4 and 5), such as the stratification of transmission pairs by the location of primary cases (imported vs local), and estimating transmission pairs with more clear epidemiological evidence (eg, primary case linked only to a single secondary case). Results of these sensitivity analyses are consistent with those estimated with all transmission pairs.

Hazard of being infected within households was higher for young (<18 years) and elderly (>65 years) people, whereas the hazard of being infected outside of households was higher for the age group 18–64 years (Table 1). Primary cases of elderly (>65 years) people were more prone to cause household infections. Hazard of infection between different sexes was higher for household than for nonhousehold transmission (Table 2).

Table 2. Sex-specific Hazard of Infection for Household Versus Nonhousehold Transmissions

		Secondary Cases		
		Male	Female	Total
Primary Cases	Male	0.6	1.6	1.0
	Female	1.2	0.7	0.9
	Total	0.8	1.2	1.0

DISCUSSION

We have built a line-list database with detailed demographic information, travel history, epidemiological timelines, and social relationships for 1407 transmission pairs that formed 643 transmission clusters in mainland China outside Hubei province. We identified 34 primary cases as superspreaders. The majority of SSEs were observed for nonhousehold transmissions, which is consistent with a recent study [21] on transmission settings of COVID-19 (eg, hospitals, residential care, prisons, boarding schools, cruise ships). This indicates the importance of nonpharmaceutical interventions (eg, isolation, quarantine, social distancing, and confinement [7, 22–24]) in mitigating the COVID-19 epidemic.

Household studies are helpful to identify risk factors for certain demographic groups [25, 26]. The analysis of the age-stratified and sex-specific hazard of infection suggests a higher risk of infection within households for young (<18 years of age), elderly (>65 years of age), and female people. The higher risk of being infected outside of households for male people aged between 18 and 64 years may indicate their role in driving household secondary infections, perhaps because these were travelers of working age from Wuhan.

We identified 50 transmission pairs (~3.5%) with a secondary case reporting symptom onset earlier than the primary case (ie, negative-valued serial intervals), which is consistent with recent clinical reports [27, 28] and epidemiological studies [29, 30]. We estimated that the mean serial interval is around 5 days for both household and nonhousehold infections, which is considerably shorter than the mean serial interval estimated for severe acute respiratory syndrome (eg, 8.4 days [31]) and Middle East respiratory syndrome (eg, 7.6 days [32]).

Our findings have several limitations. First, the household sizes and primary cases with no secondary infections were not provided from the original public case reports. This may give rise to biased estimates if we estimate the household reproduction number and secondary attack rate from the raw data. Field surveys will be helpful to adjust such biases. Second, the information on nosocomial infections and public gathering settings was not available from original case reports, so the observation of SSEs may be less common from our dataset. Third, caution is needed when attempting to generalize the age-stratified hazard of infection to other demographic settings. For example, in our study (Table 1), the fact that children (<18 years of age) never acquired COVID-19 from other children at home may be more a reflection of the usual household composition in Chinese cities (single child living with parents) than the transmission characteristics of the virus. China had a lower proportion of households with multiple children [33], which may reduce the risk of transmission between children, especially during lockdown and school closure.

In sum, the notable threat of SSEs, short serial intervals, and a higher risk of being infected outside of households for adult

Table 1. Age-stratified Hazard of Infection for Household Versus Nonhousehold Transmissions

		Secondary Cases, Age, y				
		0–17	18–49	50–64	≥65	Total
Primary Cases, Age, y	0–17	0.0	0.8	0.8	1.1	0.7
	18–49	6.3	0.7	0.9	2.0	1.1
	50–64	1.7	0.9	0.7	0.6	0.8
	≥65	2.3	1.4	0.6	2.1	1.3
	Total	3.5	0.8	0.8	1.4	1.0

men of working age (18–64 years) indicate a significant barrier to the identification and management of COVID-19 cases; enhanced nonpharmaceutical interventions will be required to mitigate this pandemic.

Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Author contributions. X.-K. X., X. F. L., Y. W., S. T. A., and L. W. conceived the study. S. T. A., Z. D., E. H. Y. L., B. J. C., and L. W. had roles in the study design. X.-K. X., X. F. L., and Y. W. had roles in the data collection, data analysis, and data interpretation. X. F. L., S. T. A., Z. D., P. B., E. H. Y. L., B. J. C., and L. W. had roles in the data interpretation, statistical modeling, and writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgments. The authors thank all health workers and volunteers in China who collected, prepared, and shared data throughout this outbreak. They are in particular grateful to Simon Cauchemez, Henrik Salje, Lauren Ancel Meyers, Juliette Paireau, Qifang Bi, Bingyi Yang, and Lanfang Hu for comments and suggestions. The statistical code and data set are available at: https://github.com/linwangidd/covid19_transmissionPairs_China.

Disclaimer. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the French Agence Nationale de la Recherche, European Commission, European Research Council, National Institutes of Health, National Natural Science Foundation and National Social Science Foundation of China, or the government of Hong Kong Special Administrative Region. The funders had no role in study design, data collection and analysis, preparation of the manuscript, or the decision to submit the manuscript for publication.

Financial support. This study was partially supported by the Investissement d'Avenir program, Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases program (ANR-10-LABX-62-IBEID); European Union V.E.O. project, European Union's Horizon 2020 research and innovation program under (grant number 101003589) European Research Council (804744); National Institutes of Health (U01 GM087719); Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources of China (KF-2019–04–034); National Natural Science Foundation of China (61773091, 11875005, 61976025, 11975025); Major Project of the National Social Science Fund of China (19ZDA324); and Health and Medical Research Fund, Food and Health Bureau, Government of the Hong Kong Special Administrative Region, China (COVID190118).

Potential conflicts of interest. B. J. C. reports honoraria from Sanofi Pasteur and Roche. All other authors report no potential conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. World Health Organization. Statement on the second meeting of the International Health Regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-nCoV). **2020**. Available at: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov.
2. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19, 11 March 2020. Available at: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.
3. World Health Organization. Coronavirus disease 2019 (COVID-19), situation report–149. **2020**. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200617-covid-19-sitrep-149.pdf?sfvrsn=3b3137b0_4.
4. Du Z, Wang L, Cauchemez S, et al. Risk for transportation of 2019 novel coronavirus disease from Wuhan to other cities in China. Emerg Infect Dis **2002**; 26:1049–52.
5. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet **2020**; 395:689–97.
6. Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science **2020**; 368:395–400.
7. Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science **2020**; 368:637–42.
8. Kraemer MUG, Yang C-H, Guitierrez B, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science **2020**; 368:493–7.
9. Kucharski AJ, Russell TW, Diamond C, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis **2020**; 20:553–8.
10. Lai S, Ruktanonchai NW, Zhou L, et al. Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak in China. medRxiv [Preprint]. Posted 6 March **2020**. doi:10.1101/2020.03.03.20029843.
11. Xu B, Kraemer MUG; Open COVID-19 Data Curation Group. Open access epidemiological data from the COVID-19 outbreak. Lancet Infect Dis **2020**; 20:534.
12. Xu B, Gutierrez B, Mekaru S, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data **2020**; 7:106.
13. Zhang J, Litvinova M, Wang W. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study [manuscript published online ahead of print 2 April 2020]. Lancet Infect Dis **2020**. doi:10.1016/S1473-3099(20)30230–9.
14. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis **2020**; 20:533–4.
15. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit Health **2020**; 2:e201–8.
16. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med **2020**; 172:577–82.
17. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature **2005**; 438:355–9.
18. MathWorks. Working with probability distributions. **2020**. Available at: https://fr.mathworks.com/help/stats/working-with-probability-distributions.html. Accessed 22 June 2020.
19. Vink MA, Bootsma MC, Wallinga J. Serial intervals of respiratory infectious diseases: a systematic review and analysis. Am J Epidemiol **2014**; 180:865–75.
20. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. Lancet **2020**; 395:e47.
21. Leclerc QJ, Fuller NM, Knight LE, et al. What settings have been linked to SARS-CoV-2 transmission clusters? [version 1; peer review: 1 approved with reservations]. Wellcome Open Res **2020**; 5:83. doi:10.12688/wellcomeopenres.15889.1.
22. Cowling BJ, Ali ST, Ng TWY, et al. Impact assessment of non-pharmaceutical interventions against COVID-19 and influenza in Hong Kong: an observational study. Lancet Public Health **2020**; 5:e279–88.
23. Koo JR, Cook AR, Park M, et al. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. Lancet Infect Dis **2020**; 20:678–88.
24. Prem K, Liu Y, Russell TW, et al; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group. 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 **2020**; 5:e261–70.
25. Cowling BJ, Leung GM. Epidemiological research priorities for public health control of the ongoing global novel coronavirus (2019-nCoV) outbreak. Euro Surveill **2020**; 25. doi:10.2807/1560–7917.ES.2020.25.6.2000110.
26. Lipsitch M, Swerdlow DL, Finelli L. Defining the epidemiology of Covid-19 - studies needed. N Engl J Med **2020**; 382:1194–6.
27. Bai Y, Yao L, Wei T, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA **2020**; 323:1406–7.
28. Pan X, Chen D, Xia Y, et al. Asymptomatic cases in a family cluster with SARS-CoV-2 infection. Lancet Infect Dis **2020**; 20:410–1.
29. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial interval of COVID-19 among publicly reported confirmed cases. Emerg Infect Dis **2020**; 26:1341–3.
30. He X, Lau EHY, Wu P, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med **2020**; 26:672–5.
31. Lipsitch M, Cohen T, Cooper B, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science **2003**; 300:1966–70.
32. Assiri A, McGeer A, Perl TM, et al. KSA MERS-CoV Investigation Team. Hospital outbreak of Middle East respiratory syndrome coronavirus. N Engl J Med **2013**; 369:407–16.
33. United Nations. Household size and composition around the world 2017. Available at: https://digitallibrary.un.org/record/3799696/files/household_size_and_composition_around_the_world_2017_data_booklet.pdf. Accessed 22 June 2020.