The Lancet Public Health

Estimation of the Time-Varying Reproduction Number of COVID-19 Outbreak in China

--Manuscript Draft--

Manuscript Number:	thelancetpublichealth-D-20-00223
Article Type:	Article (Original Research)
Keywords:	serial interval, incubation period, infectious period, reproduction rate
Corresponding Author:	Xiao-Hua Zhou, Ph.D. Peking University CHINA
First Author:	Xiao-Hua Zhou, Ph.D.
Order of Authors:	Xiao-Hua Zhou, Ph.D.
	Chong You
	Yuhao Deng
	Wenjie Hu
	Jiarui Sun
	Qiushi Lin
	Feng Zhou
	Cheng Heng Pang
	Yuan Zhang
	Zhengchao Chen
Manuscript Region of Origin:	CHINA
Abstract:	Background The 2019 novel coronavirus (2019-nCoV) outbreak in Wuhan, China has attracted world-wide attention. As of February 11, 2020, a total of 44,730 cases of pneumonia associated with the 2019-nCoV were confirmed by the National Health Commission (NHC) of China. Methods Three approaches, namely Poisson likelihood-based method (ML), exponential growth rate-based method (EGR) and stochastic Susceptible-Infected-Removed dynamic model-based method (SIR), were implemented to estimate the basic and controlled reproduction numbers. Results A total of 71 chains of transmission together with dates of symptoms onset and 67 dates of infections were identified among 5,405 confirmed cases outside Hubei as reported by February 2, 2020. Based on this information, we find the serial interval having an average of 4.27 days with a standard deviation of 3.44 days and the infectious period having an average of 10.91 days with a standard deviation of 3.95 days. The estimated controlled reproduction numbers, Common.EditSubmissionSteps.Transform.EquationText, produced by all three methods in all analyzed regions of China are significantly smaller compared with the basic reproduction numbers Common.EditSubmissionSteps.Transform.EquationText. Conclusions The controlled reproduction number is declining. It is lower than one in most regions of China, but still larger than one in Hubei Province. Sustained efforts are needed to further keep/reduce the Common.EditSubmissionSteps.Transform.EquationText to

	below one in order to end the current epidemic.

Estimation of the Time-Varying Reproduction Number of COVID-19 Outbreak in China

Chong You[1]*, Yuhao Deng[2]*, Wenjie Hu[2], Jiarui Sun[2], Qiushi Lin[2], Feng Zhou[3], Cheng Heng Pang[4], Yuan Zhang[5], Zhengchao Chen[6], Xiao-Hua Zhou[1,3]**

1 Beijing International Center for Mathematical Research, Peking University, China
2 School of Mathematical Sciences, Peking University, China
3 Department of Biostatistics, School of Public Health, Peking University, China
4 Faculty of Science and Engineering, University of Nottingham Ningbo China, China
5 National Research Institute for Health and Family Planning, China
6 Beijing Obstetrics and Gynecology Hospital, Capital Medical University, China

** Corresponding author: azhou@math.pku.edu.cn
***** Joint first authors

Abstract

Background: The 2019 novel coronavirus (2019-nCoV) outbreak in Wuhan, China has attracted world-wide attention. As of February 11, 2020, a total of 44,730 cases of pneumonia associated with the 2019-nCoV were confirmed by the National Health Commission (NHC) of China.

Methods: Three approaches, namely Poisson likelihood-based method (ML), exponential growth rate-based method (EGR) and stochastic Susceptible-Infected-Removed dynamic model-based method (SIR), were implemented to estimate the basic and controlled reproduction numbers.

Results: A total of 71 chains of transmission together with dates of symptoms onset and 67 dates of infections were identified among 5,405 confirmed cases outside Hubei as reported by February 2, 2020. Based on this information, we find the serial interval having an average of 4.27 days with a standard deviation of 3.44 days, the incubation period having an average of 5.33 days with a standard deviation of 3.36 days and the infectious period having an average of 10.91 days with a standard deviation of 3.95 days. The estimated controlled reproduction numbers, R_c, produced by all three methods in all analyzed regions of China are significantly smaller compared with the basic reproduction numbers R_0.

Conclusions: The controlled reproduction number is declining. It is lower than one in most regions of China, but still larger than one in Hubei Province. Sustained efforts are needed to further keep/reduce the R_c to below one in order to end the current epidemic.

1. Introduction

On December 29, 2019, four cases of pneumonia with unknown etiology were reported in Wuhan, the capital city of Hubei Province in Central China. Since then, the outbreak has dramatically worsened over a short span of time and has received considerable global attention. On January 7, 2020, the pathogen of the current outbreak was identified as a novel coronavirus (2019-nCoV), and its gene sequence was quickly submitted to the WHO (The coronavirus was renamed COVID-19 by the WHO on February 12).[1,2] On January 30, the WHO announced the listing of this novel coronavirus-infected pneumonia (NCP) as a "public health emergency of international concern". As of February 11, 2020, the National Health Commission (NHC) of China had confirmed a total of 44,730 cases of COVID-19 in Mainland China, including 1,114 fatalities and 4,771 recoveries.

Since January 19, 2020, strict containment measures, including travel restrictions, contact tracing, entry or exit screening, non-hospital isolation, quarantine and awareness campaigns have been implemented by the Wuhan municipal government and quickly adopted by other cities within China with the aim to minimize virus transmission via human-to-human contact.

In 2009, similar measures were employed in China in response to the outbreak of H1N1 virus breakout.

This article investigates the change in the basic reproduction number R_0 and controlled reproduction number R_c since the outbreak of COVID-19. We have found that the estimated controlled reproduction numbers R_c in all different regions are significantly smaller compared with the basic reproduction numbers R_0, but R_c is still greater than one in Hubei Province.

2. Data

Publicly available data were collected from provincial/municipal health commissions in China and ministry of health in other countries and regions. The following details were collected on each confirmed case: case ID, region, age, gender, date of symptom onset, date of diagnosis, history of travel or previous residency in Wuhan, and, if available, related information regarding contact history with other confirmed cases. In addition, the secondary dataset also contains date of infection and probable transmission chains which were inferred based on history of travel to or previous residency in Wuhan and other related information, when available. These inferences were made as follows:

(1) If the individual has not been to Hubei province recently, but was exposed within a four-day period (i.e., the individual had contact with a confirmed case on one of four consecutive days), then the corresponding date of infection is inferred as the middle of the exposure period;

(2) If the individual previously traveled to Hubei province but returned within four days, then the date of infection is inferred as the middle of the travel period;

(3) If the individual has not recently been in Hubei province, but has been in close contact with an imported case from Hubei, then the individual was determined to be infected by this imported case;

(4) If the individual has not recently been in Hubei province, but has been in close contact with a local case who was clearly infected before the contact, then this individual was determined to be infected by the corresponding local case.

Note: if the individual has been to Hubei Province, the transmission history would not be recorded despite the existence of contact tracing information.

3. Inference about the Generation Time, Incubation period and Infectious Period

Generation time is the time difference between dates of infection of successive cases in a chain of transmission. Infectious period is the duration of which an infected individual can transmit pathogens to a susceptible host. In this study, infectious period is defined as the time difference between date of infection and date of diagnosis as there is strong evidence showing that a diseased individual remains contagious even during the incubation period, and would be immediately isolated upon positive diagnosis hence losing the transmissibility. The

incubation period is defined as time difference between contraction of the disease and symptoms onsets. All are key quantities that depict an epidemic and are essential to estimate the basic/controlled reproductive number, R_0/R_c. Among the 139 chains of transmission identified from 5,405 confirmed cases recorded outside Hubei by February 2, 2020, none of them have their dates of infection acquired, but 71 of them have their dates of symptoms onset available. Hence, the corresponding generation time was approximated by the differences in dates of symptom onset rather than the actual dates of infection, that is the serial interval, see Figure 1.

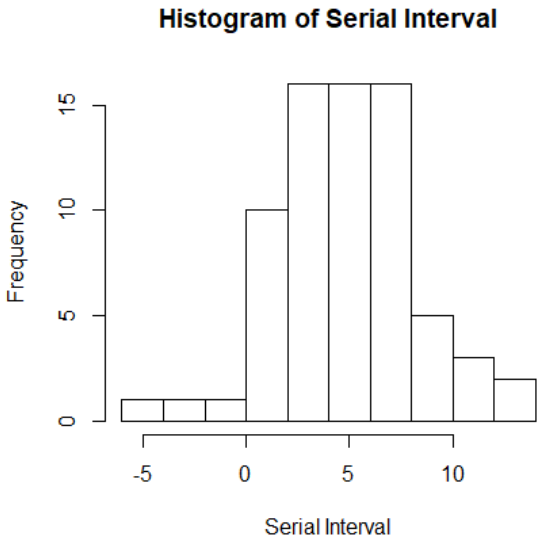

Figure 1: Histogram of serial interval with the average of 4.27 days before correction.

We can see that some serial intervals are negative, which suggests that COVID-19 is contagious during incubation and negative values were caused by different lengths of incubation period between individuals. We do acknowledge that the distribution of serial interval may be biased for estimating generation time, especially when the disease is contagious during incubation, the variance could be overestimated[3]. Further studies should be conducted in regards a better estimation of generation time based on serial interval, but this is out of our scope in this article. In this article, the distribution of generation time would be directly replaced by the empirical distribution of serial interval. The average of the serial intervals is 4.27 days and the standard deviation is 3.44 days (see table 1). Note that the serial interval of SARS-nCoV in Hongkong was 8.4 days on average.[4] In addition, a total of 67 cases in the collected data were able to identify the corresponding dates of infection. Figure 2 plots the histogram of infectious period while Table 2 shows the numerical summary.

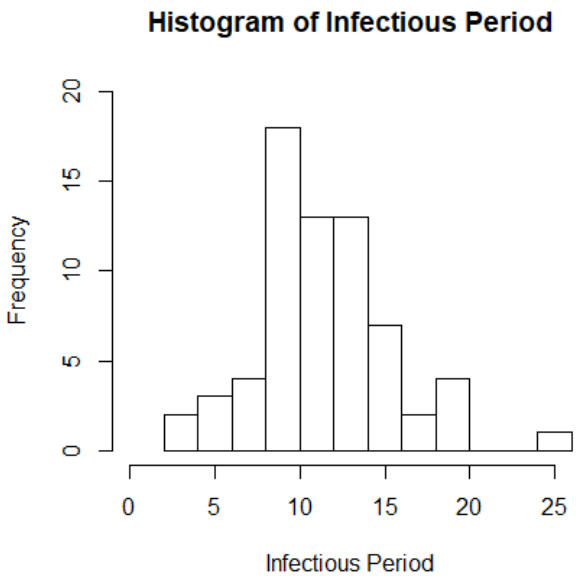

Figure 2: Histogram of infectious period with the average of 10.91 days.

Furthermore, a total of 56 cases in the collected data were able to identify both the dates of infection and date of symptoms onset. Figure 3 plots the histogram of infectious period while Table 3 shows the numerical summary.

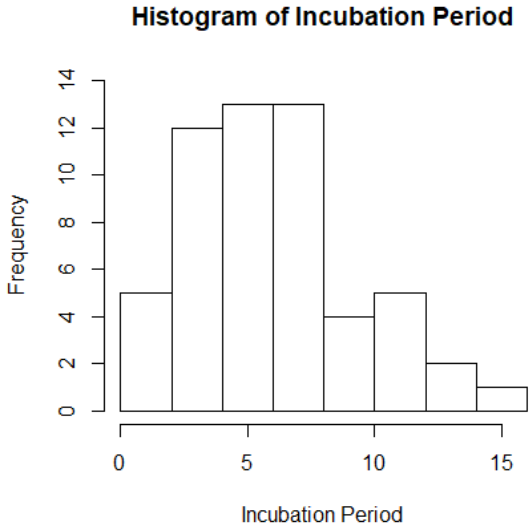

Figure 3: Histogram of infectious period with the average of 5.33 days.

We found that there were no significant demographical differences between the subset of cases used to estimate serial interval and infectious period and the cases in the full dataset. Therefore, the inference made on serial interval, incubation period and infectious period based on the corresponding subsets should be able to represent the full dataset.

4. Estimation of Basic/Controlled Reproduction Number

Definition

The reproduction number R_0 is defined as the (average) number of new infections generated by one infected individual during the entire infectious period in a fully susceptible population.[5] It can be also understood as the average number of infections caused by a typical individual during the early stage of an outbreak when nearly all individuals in the population are susceptible. The basic reproduction number reflects the ability of an infection spreading under no control. When the size of susceptible population is limited, the quantity, effective reproduction number R_e, is used instead of R_0. Similarly, the quantity, controlled reproduction number R_c, should be used to describe the ability of disease spreading when interventions (such as quarantine, isolation, or traffic control) are taking place. Hence a good measure of any intervention is to reduce R_c. Note that the disease will decline and eventually die out if $R_c \leq 1$.

Methods

The basic reproduction number can be estimated through a variety of models.[6] In this section, we have compared three most popular estimates of R_0/R_c as shown below.

(1) Poisson Likelihood-based (ML) method

Let N_t be the number of reported new confirmed cases on day t. Suppose that the serial interval has a maximum of k days and the number of new cases generated by an infected individual is assumed to follow a Poisson distribution with parameter R_0.[7] The probability that the serial interval of an individual in j days is w_j, which can be estimated from the empirical distribution of serial interval or by setting up a discretized Gamma prior on it. Thus, the likelihood function can be reduced into a thinned Poisson

$$L(R, w) = \prod_{t=1}^{T} \frac{e^{-\mu_t} \mu_t^{N_t}}{N_t!}$$

where

$$\mu_t = R \sum_{j=1}^{\min\{k,t\}} N_{t-j} w_j.$$

The reproduction number R can be estimated by maximizing the likelihood function. Note that if the empirical distribution of serial interval is used or $w_j s$ are given, then

$$\hat{R} = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{\min\{k,t\}} N_{t-j} w_j}.$$

(2) Exponential growth rate-based (EGR) method

At the early period of an epidemic, the number of infected cases rises exponentially. Suppose the exponential epidemic growth rate (Malthusian coefficient) is r, which can be estimated by fitting a least square line to the daily number of reported new confirmed cases in a log-scale, namely, $\log(N_t)$. Let $f_G(t)$ denote the probability density function of serial interval. Hence the reproduction number can be calculated according to the Euler-Lotka equation in a moment

generating form[8]

$$\hat{R} = \frac{1}{\int_0^\infty e^{-rt} f_G(t) dt}.$$

(3) Stochastic dynamic model-based method

Here we consider a stochastic Susceptible-Infected-Removed (SIR) model rather than a standard deterministic one. The major advantage of using a stochastic dynamic model is that it affords improved accounting for real variabilities and increases opportunity for quantifying uncertainties.[9] Let $S(t)$, $I(t)$ and $R(t)$ denote the number of susceptible, infectious and recovered population at time t respectively, and note that $N = S(t) + I(t) + R(t)$. Suppose that the infectious period of an individual is a random variable $T \sim Exp(\gamma)$, then the reproduction number $R = \beta E(T) = \beta/\gamma$, where γ and β are the recovery rate and transmission rate respectively in the system of ordinary differential equation (ODE) below,

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I.$$

The maximum likelihood method is used to estimate model parameters where the likelihood is obtained by sequential Monte Carlo method, and parameters are estimated using the Iterated Filtering algorithm (IF2)[10] implemented as `mif` in the `R` package `pomp`[11] where $S(0)$ equals the population of the region, $R(0) = 0$, $I(0)$ is 10 times the average number of confirmed cases from Day 0 to Day 7 and $\gamma = 10.91$ obtained from the collected data described before.

Results

In this section we have estimated the basic reproduction number R_0 and the controlled reproduction number R_c. Since January 19, 2020, various containment measures have been strictly implemented, especially after the State Council agreed to include NCP into the Management of the Infectious Diseases Law and the Health and Quarantine Law on January 20. Based on an average10.91-day infectious period estimate from our collected data, we expect a flatter rate of increment starting on January 29. Figure 4 plots the number of daily new cases on a log-scale against date, and, as anticipated, the trend supports this estimate. Therefore, the quantities R_0 and R_c are estimated based on collected data in two separate periods, i.e., from January 21 (the starting date of daily updates of confirmed cases nationwide) to January 28, and from January 29 to February 5 (the end date of this study) respectively.

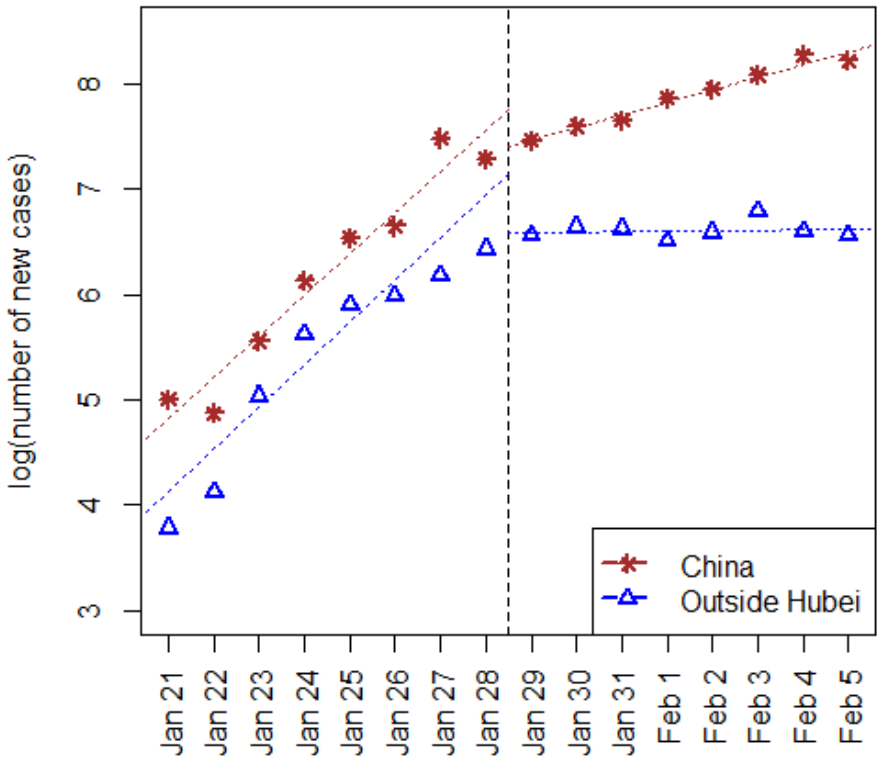

Figure 4: Visualization of daily numbers of new confirmed case along with date, as obvious change of rate occurred on January 29 as expected.

The estimates of R_0 and R_c by Poisson likelihood (ML) and exponential growth rate (EGR) in selected regions of China are listed in Table 4 and Table 5. Despite the disagreement between different estimation methods, all three methods indicate notable reductions from R_0 to R_c which suggests an improvement in the current situation. This is possibly due to the effective interventions and prompt actions by the local and central governments to minimize further spreading. We also notice that EGR yields smaller estimates of R_c compared to other methods. This might be because the number of infected patients does not grow exponentially after such strict containment measures.

Furthermore, the time-varying controlled reproduction number $R_c(t)$ can be estimated through the Poisson likelihood (ML) method where t is from February 1 to February 10, 2020. For each Day t, the number of daily reported new cases from Day $t-7$ to Day t is used to estimate $\hat{R}_c(t)$. Figure 5 plots the estimated controlled reproduction number $\hat{R}_c(t)$ along with its 95% CI for selected regions of China. Note that the estimated $\hat{R}_c(t)$ reflects the average spreading ability of the epidemic in a short period prior to Day t. As a result, the real-time $R_c(t)$ might be overestimated if the general trend of $R_c(t)$ is declining.

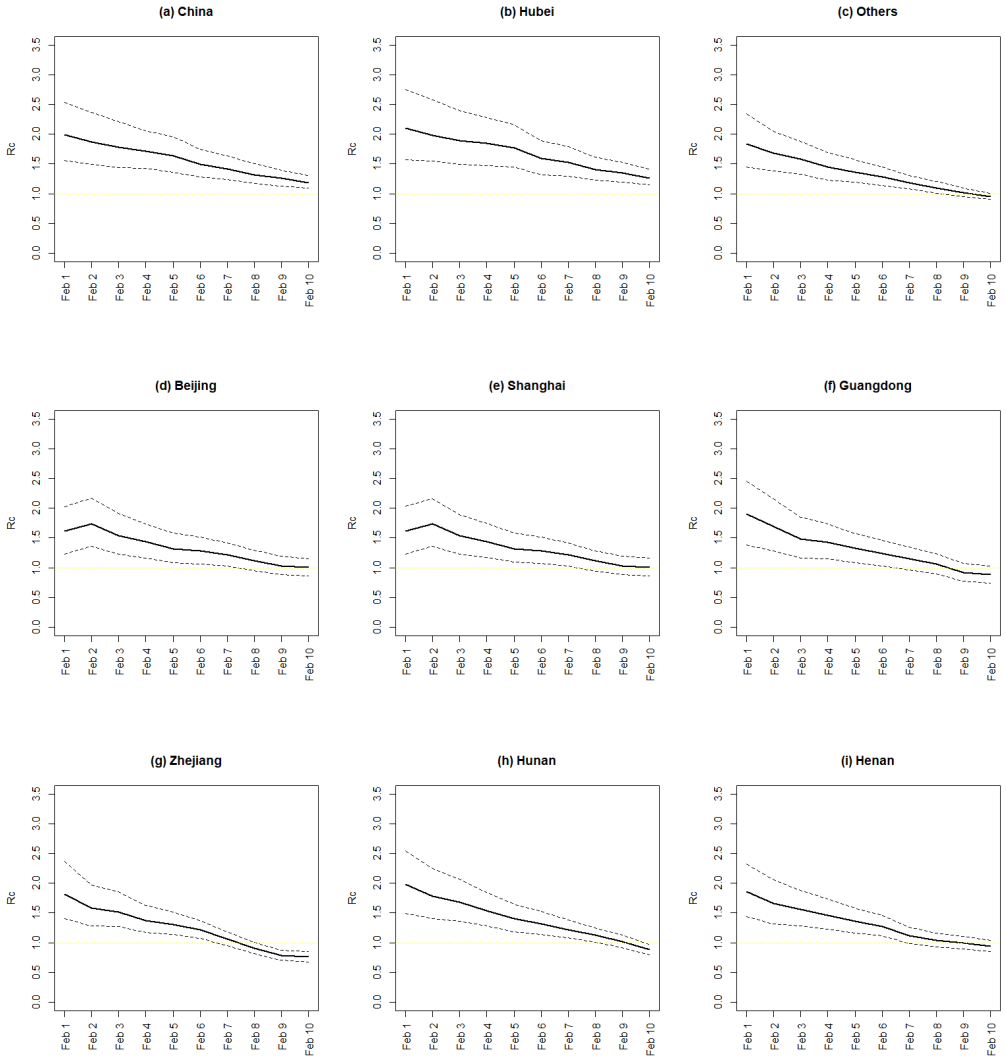

Figure 5: The estimated controlled reproduction number in (a) China, (b) Hubei, (c) Other provinces excluding Hubei, (d) Beijing, (e) Shanghai, (f) Guangdong, (g) Zhejiang, (h) Hunan, and (i) Henan. The dashed line is the 95% confidence interval.

5. Conclusion

Despite the continuous increase in new confirmed cases each day, the estimated controlled reproduction numbers R_c produced by all three methods in all analyzed regions are significantly smaller compared with the basic reproduction numbers R_0. As discussed in Section 4, the real-time controlled reproduction number may be even lower than the estimated values in Figure 4. Nonetheless, further effort is needed to further reduce R_c below one in Hubei Province.

6. Discussion

The dataset used in this study is based on the confirmed cases reported by the NHC of China. However, during our period of data collection, the official guidelines for diagnosis and treatment of COVID-2019 underwent four updates. The criteria of confirmation have evolved from the original "whole genome sequencing of the respiratory excretion" to "positive viral nucleic acid results by the RT-PCR of the respiratory excretion or viral gene sequence", and, as of now, the inclusion of positive nucleic acid results of the blood sample. The confirmation process has been simplified by the removal of the accreditation process by the national expert committee for confirmed cases. The fourth edition of the official guidelines for diagnosis and treatment granted the accrediting authority to municipalities.[12] In addition, the medical resources in Hubei province especially in Wuhan has received a remarkable boost. All of these changes might result in a temporary surge of confirmed cases and lead to an overestimation of R_c, especially in Hubei Province.

Furthermore, the current containment measures mainly aim to cut the transmission from human to human via respiratory droplets. However, other transmission pathways, including fecal-oral transmission and aerosol transmission, could not yet be excluded based on current evidence. If other transmission mechanisms do exist, the R_c values would remain high in the future unless further measures would intersect these transmission pathways.

Acknowledgments

We thank Taojun Hu, Xueqing Liu and Yuying Li from School of Public Health, Peking University for assistance of data collection.

Tables

Table 1: numerical summary of serial intervals.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
-5.00	2.00	4.00	4.268	6.00	13.00

Table 2: numerical summary of infectious period.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
2.00	8.00	11.00	10.91	13.00	25.00

Table 3: numerical summary of incubation period.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
0.00	2.00	5.00	5.33	7.00	14.00

Table 4: Estimates and 95% confidence intervals of basic reproduction number in some selected provinces (or cities) of China, from Jan 21 to Jan 28, 2020.

	ML	EGR	SIR
China	3.024 (2.020, 4.422)	3.738 (2.502, 6.006)	5.4 (4.5, 6.2)
Hubei	3.334 (2.118, 5.056)	3.758 (2.052, 7.440)	5.5 (4.2, 6.8)
Others	2.681 (1.754, 3.941)	3.687 (2.281, 6.474)	5.1 (3.9, 6.3)
Beijing	2.278 (1.436, 3.281)	2.164 (1.129, 4.116)	2.3 (1.1, 3.8)
Shanghai	2.144 (1.373, 3.002)	1.610 (0.636, 3.441)	2.4 (1.0, 3.8)
Guangdong	2.488 (1.692, 3.493)	2.691 (1.337, 5.335)	3.7 (2.7, 4.9)
Zhejiang	2.506 (1.649, 3.626)	3.132 (1.335, 6.958)	5.0 (3.3, 7.0)
Hunan	2.767 (1.795, 4.086)	4.872 (2.171, 11.136)	5.3 (4.3, 7.0)
Henan	2.888 (1.783, 4.400)	5.452 (2.650, 11.771)	6.4 (3.5, 10.2)

Table 5: Estimates and 95% confidence intervals of controlled reproduction number in some selected provinces (or cities) of China, from Jan 29 to Feb 5, 2020.

	ML	EGR	SIR
China	2.282 (1.726, 2.985)	1.652 (1.457, 1.906)	2.3 (2.1, 2.5)

Hubei	2.503 (1.908, 3.321)	1.935 (1.642, 2.367)	2.8 (2.5, 3.1)
Others	1.840 (1.465, 2.332)	1.071 (0.915, 1.263)	1.2 (1.1, 1.4)
Beijing	2.260 (1.644, 3.013)	1.619 (0.861, 2.986)	2.1 (1.0, 3.3)
Shanghai	1.876 (1.407, 2.446)	1.073 (0.673, 1.616)	1.2 (0.7, 2.0)
Guangdong	1.899 (1.480, 2.482)	1.131 (0.718, 1.712)	1.2 (0.8, 1.8)
Zhejiang	1.480 (1.199, 1.838)	0.654 (0.343, 1.014)	1.0 (0.4, 1.7)
Hunan	1.698 (1.333, 2.151)	0.963 (0.677, 1.302)	1.3 (1.0, 1.8)
Henan	2.091 (1.597, 2.657)	1.403 (1.054, 1.941)	1.5 (1.1, 2.0)

[1] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*.

[2] Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*.

[3] Britton, Tom, and Gianpaolo Scalia Tomba. "Estimation in emerging epidemics: Biases and remedies." *Journal of the Royal Society Interface* 16.150 (2019): 20180670.

[4] Lipsitch M, Cohen T, Cooper B, et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science* 2003; 300**:** 1966–70.

[5] Anderson, R. M., Anderson, B., & May, R. M. (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press.

[6] Nikbakht, R., Baneshi, M. R., Bahrampour, A., & Hosseinnataj, A. (2019). Comparison of methods to Estimate Basic Reproduction Number (R0) of influenza, Using Canada 2009 and 2017-18 A (H1N1) Data. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, *24*.

[7] Forsberg White, L., & Pagano, M. (2008). A likelihood‐ based method for real‐ time estimation of the serial interval and reproductive number of an epidemic. *Statistics in medicine*, *27*(16), 2999-3016.

[8] Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1609), 599-604.

[9] King Aaron A., Domenech de Cellès Matthieu, Magpantay Felicia M. G. and Rohani Pejman. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B.* 282

[10] Ionides, E. L., Nguyen, D., Atchad´e, Y., Stoev, S. & King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps, *Proceedings of the National Academy of Sciences of the U.S.A*. 112, 719–724.

[11] King, A. A., Ionides, E. L., Bret´o, C. M., Ellner, S., Kendall, B., Wearing, H., Ferrari, M. J., Lavine, M. & Reuman, D. C. (2010). *pomp: Statistical inference for partially observed Markov processes (R package)*.

[12] National Health Commission of the People's Republic of China, http://www.nhc.gov.cn/xcs/zhengcwj/list_gzbd.shtml

Authors' contributions:

Chong You: data collection, writing

Yuhao Deng: writing, data analysis

Wenjie Hu: data analysis

Jiarui Sun: data analysis

Qiushi Lin: data collection

Feng Zhou: data collection

Cheng Heng Pang: writing

Yuan Zhang: writing

Zhengchao Chen: wrting

Xiao-Hua Zhou: overall design

Conflict of interest statements:

We have no financial relationships (regardless of amount of compensation) with any entities. There is no conflict of interest.

Role of funding source: NA

Ethics committee approval : not required