RESEARCH ARTICLE

REVISED Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]

Akira Endo [1,2],

Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group,

Sam Abbott [1,3], Adam J. Kucharski[1,3], Sebastian Funk [1,3]

[1]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK
[2]The Alan Turing Institute, London, NW1 2DB, UK
[3]Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

v3 **First published:** 09 Apr 2020, **5**:67
https://doi.org/10.12688/wellcomeopenres.15842.1
Second version: 03 Jul 2020, **5**:67
https://doi.org/10.12688/wellcomeopenres.15842.2
Latest published: 10 Jul 2020, **5**:67
https://doi.org/10.12688/wellcomeopenres.15842.3

Abstract

Background: A novel coronavirus disease (COVID-19) outbreak has now spread to a number of countries worldwide. While sustained transmission chains of human-to-human transmission suggest high basic reproduction number R_0, variation in the number of secondary transmissions (often characterised by so-called superspreading events) may be large as some countries have observed fewer local transmissions than others.
Methods: We quantified individual-level variation in COVID-19 transmission by applying a mathematical model to observed outbreak sizes in affected countries. We extracted the number of imported and local cases in the affected countries from the World Health Organization situation report and applied a branching process model where the number of secondary transmissions was assumed to follow a negative-binomial distribution.
Results: Our model suggested a high degree of individual-level variation in the transmission of COVID-19. Within the current consensus range of R_0 (2-3), the overdispersion parameter k of a negative-binomial distribution was estimated to be around 0.1 (median estimate 0.1; 95% CrI: 0.05-0.2 for R0 = 2.5), suggesting that 80% of secondary transmissions may have been caused by a small fraction of infectious individuals (~10%). A joint estimation yielded likely ranges for R_0 and k (95% CrIs: R_0 1.4-12; k 0.04-0.2); however, the upper bound of R_0 was not well informed by the model and data, which did not notably differ from that of the prior distribution.
Conclusions: Our finding of a highly-overdispersed offspring distribution highlights a potential benefit to focusing intervention efforts on

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 3 (revision) 10 Jul 2020	✓ report	
version 2 (revision) 03 Jul 2020	✓ report	✓ report
version 1 09 Apr 2020	✓ report	? report

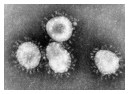
1 **Lin Wang**, Institut Pasteur, Paris, France

2 **Kaiyuan Sun**, National Institutes of Health, Bethesda, USA

Any reports and responses or comments on the article can be found at the end of the article.

superspreading. As most infected individuals do not contribute to the expansion of an epidemic, the effective reproduction number could be drastically reduced by preventing relatively rare superspreading events.

Keywords
COVID-19, SARS-CoV-2, novel coronavirus, overdispersion, superspreading, branching process



This article is included in the Coronavirus (COVID-19) collection.

Corresponding author: Akira Endo (akira.endo@lshtm.ac.uk)

Author roles: Endo A: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Abbott S**: Validation, Writing – Review & Editing; **Kucharski AJ**: Supervision, Writing – Review & Editing; **Funk S**: Methodology, Supervision, Writing – Review & Editing

Competing interests: AE received a research grant from Taisho Pharmaceutical Co., Ltd.

Grant information: SA [210758], AJK [206250] and SF [210758] are supported by the Wellcome Trust. AE is financially supported by The Nakajima Foundation and The Alan Turing Institute.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott S *et al.* **Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]** Wellcome Open Research 2020, **5**:67 https://doi.org/10.12688/wellcomeopenres.15842.3

First published: 09 Apr 2020, **5**:67 https://doi.org/10.12688/wellcomeopenres.15842.1

REVISED **Amendments from Version 2**

A typo in the equation on $p_{80\%}$ was corrected; originally it was

$$1-p_{80\%}=1-\int \mathrm{NB}\left(\lfloor x \rfloor; k, \frac{k}{(R_0+k)}\right) dx,$$

which should have been instead

$$1-p_{80\%}=\int \mathrm{NB}\left(\lfloor x \rfloor; k, \frac{k}{(R_0+k)}\right) dx,$$

The typo was only present in the manuscript and did not affect the analysis or other parts of the manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

A novel coronavirus disease (COVID-19) outbreak, which is considered to be associated with a market in Wuhan, China, is now affecting a number of countries worldwide[1,2]. A substantial number of human-to-human transmission has occurred; the basic reproduction number R_0 (the average number of secondary transmissions caused by a single primary case in a fully susceptible population) has been estimated around 2–3[3–5]. More than 100 countries have observed confirmed cases of COVID-19. A few countries have already been shifting from the containment phase to the mitigation phase[6,7], with a substantial number of locally acquired cases (including those whose epidemiological link is untraceable). On the other hand, there are countries where a number of imported cases were ascertained but fewer secondary cases have been reported than might be expected with an estimated value of R_0 of 2–3.

This suggests that not all symptomatic cases cause a secondary transmission, which was also estimated to be the case for past coronavirus outbreaks (SARS/MERS)[8,9]. High individual-level variation (i.e. overdispersion) in the distribution of the number of secondary transmissions, which can lead to so-called superspreading events, is crucial information for epidemic control[9]. High variation in the distribution of secondary cases suggests that most cases do not contribute to the expansion of the epidemic, which means that containment efforts that can prevent superspreading events have a disproportionate effect on the reduction of transmission.

We estimated the level of overdispersion in COVID-19 transmission by using a mathematical model that is characterised by R_0 and the overdispersion parameter k of a negative binomial branching process. We fit this model to worldwide data on COVID-19 cases to estimate k given the reported range of R_0 and interpret this in the context of superspreading.

Methods

Data source

We extracted the number of imported/local cases in the affected countries (Table 1) from the WHO situation report 38[10] published on 27 February 2020, which was the latest report of the number of imported/local cases in each country (as of the situation report 39, WHO no longer reports the number of cases stratified by the site of infection). As in the WHO situation reports, we defined imported cases as those whose likely site of infection is outside the reporting country and local cases as those whose likely site of infection is inside the reporting country. Those whose site of infection was under investigation were excluded from the analysis (Estonia had no case with a known site of infection and was excluded). In Egypt and Iran, no imported cases have been confirmed, which cause the likelihood value to be zero; data in these two countries were excluded. To distinguish between countries with and without an ongoing outbreak, we extracted daily case counts from an online resource[11] and determined the dates of the latest case confirmation for each country (as of 27 February).

Model

Assuming that the offspring distributions (distribution of the number of secondary transmissions) for COVID-19 cases are identically- and independently-distributed negative-binomial distributions, we constructed the likelihood of observing the reported number of imported/local cases (outbreak size) of COVID-19 for each country. The probability mass function for the final cluster size resulting from s initial cases is, according to Blumberg *et al.*[12], given by

$$c(x;s)=P(X=x;s)=\frac{ks}{kx+x-s}\binom{kx+x-s}{x-s}\frac{\left(\frac{R_0}{k}\right)^{x-s}}{\left(1+\frac{R_0}{k}\right)^{kx+x-s}}.$$

If the observed case counts are part of an ongoing outbreak in a country, cluster sizes may grow in the future. To address this issue, we adjusted the likelihood for those countries with ongoing outbreak by only using the condition that the final cluster size of such a country has to be larger than the currently observed number of cases. The corresponding likelihood function is

$$c_o(x;s)=P(X\geq x;s)=1-\sum_{m=0}^{x-1}c(m;s),$$

with a convention $\sum_{m=0}^{-1}c(m;s)=0$. We assumed that the growth of a cluster in a country had ceased if 7 days have passed since the latest reported case (denoted by set A). We applied the final size likelihood $c(x; s)$ to those countries and $c_o(x; s)$ to the rest of the countries (countries with an ongoing outbreak: B). The total likelihood is

$$L(R_0,k)=\prod_{i\in A}P(X=x_i;s_i)\prod_{i\in B}P(X\geq x_i;s_i).$$

Statistical analysis

Varying the assumed R_0 between 0–5 (fixed at an evenly-spaced grid of values), we estimated the overdispersion parameter k using the likelihood function described above. We used the Markov-chain Monte Carlo (MCMC) method to provide 95% credible intervals (CrIs). The reciprocal of k was sampled where the prior distribution for the reciprocal was weakly-informed half-normal (HalfNormal(σ = 10)). We employed the adaptive hit-and-run Metropolis algorithm[13]

Table 1. The number of confirmed COVID-19 cases reported (as of 27 February 2020).

Country	Total cases	Imported cases	Local cases	Site of infection unknown	Deaths	Latest date of case confirmation
South Korea	1766	17	605	1144	13	27/02/2020
Japan	186	39	129	18	3	27/02/2020
Singapore	93	24	69	0	0	27/02/2020
Australia	23	20	3	0	0	26/02/2020
Malaysia	22	20	2	0	0	27/02/2020
Vietnam*	16	8	8	0	0	13/02/2020
Philippines*	3	3	0	0	1	05/02/2020
Cambodia*	1	1	0	0	0	30/01/2020
Thailand	40	23	7	10	0	26/02/2020
India*	3	3	0	0	0	03/02/2020
Nepal*	1	1	0	0	0	24/01/2020
Sri Lanka	1	1	0	0	0	27/01/2020
USA	59	56	2	1	0	26/02/2020
Canada	11	9	1	1	0	27/02/2020
Brazil	1	1	0	0	0	26/02/2020
Italy	400	3	121	276	12	27/02/2020
Germany	21	3	14	4	0	27/02/2020
France	18	8	7	3	2	27/02/2020
UK	13	12	1	0	0	27/02/2020
Spain	12	10	1	1	0	27/02/2020
Croatia	3	2	1	0	0	26/02/2020
Austria	2	2	0	0	0	27/02/2020
Finland	2	2	0	0	0	26/02/2020
Israel	2	2	0	0	0	27/02/2020
Russia*	2	2	0	0	0	31/01/2020
Sweden	2	2	0	0	0	27/02/2020
Belgium*	1	1	0	0	0	04/02/2020
Denmark	1	1	0	0	0	27/02/2020
Estonia†	1	0	0	1	0	27/02/2020
Georgia	1	1	0	0	0	26/02/2020
Greece	1	1	0	0	0	27/02/2020
North Macedonia	1	1	0	0	0	26/02/2020
Norway	1	1	0	0	0	27/02/2020
Romania	1	1	0	0	0	26/02/2020
Switzerland	1	1	0	0	0	27/02/2020
Iran†	141	0	28	113	22	27/02/2020
Kuwait	43	43	0	0	0	27/02/2020
Bahrain	33	33	0	0	0	26/02/2020
UAE	13	8	5	0	0	27/02/2020
Iraq	6	6	0	0	0	27/02/2020
Oman	4	4	0	0	0	27/02/2020
Lebanon	1	1	0	0	0	27/02/2020
Pakistan	2	1	0	1	0	26/02/2020
Afghanistan	1	1	0	0	0	24/02/2020
Egypt*†	1	0	1	0	0	14/02/2020
Algeria	1	1	0	0	0	25/02/2020

* Countries considered to be without an ongoing outbreak

† Countries excluded from analysis

and obtained 500 thinned samples from 10,000 MCMC steps (where the first half of the chain was discarded as burn-in). We confirmed that the final 500 samples have an effective sample size of at least 300, indicating sufficiently low auto-correlation.

We also performed a joint-estimation of R_0 and k by the MCMC method (with a weakly-informed normal prior $N(\mu = 3, \sigma = 5)$ for R_0 and the weakly-informed half-normal prior (HalfNormal($\sigma = 10$)) for the reciprocal of k.

Statistical analysis was implemented in R-3.6.1 with a package {LaplacesDemon}-16.1.1. The reproducible code for this study is available on GitHub[14].

Proportion responsible for 80% of secondary transmissions

Using the estimated R_0 and k, we computed the estimated proportion of infected individuals responsible for 80% of the total secondary transmissions. Such proportion $p_{80\%}$ is given as

$$1 - p_{80\%} = \int_0^X \mathrm{NB}\left(\lfloor x \rfloor; k, \frac{k}{R_0 + k}\right) dx,$$

where X satisfies

$$1 - 0.8 = \frac{1}{R_0}\int_0^X \lfloor x \rfloor \mathrm{NB}\left(\lfloor x \rfloor; k, \frac{k}{R_0 + k}\right) dx.$$

Here, $\mathrm{NB}\left(x; k, \frac{k}{R_0+k}\right)$ represents the probability mass of a negative-binomial distribution with a mean R_0 and an overdispersion parameter k. This calculation is eased by the following rearrangement:

$$\frac{1}{R_0}\int_0^X \lfloor x \rfloor \mathrm{NB}\left(\lfloor x \rfloor; k, \frac{k}{R_0 + k}\right) dx = \int_0^{X-1} \mathrm{NB}\left(\lfloor x \rfloor; k+1, \frac{k}{R_0 + k}\right) dx.$$

We computed $p_{80\%}$ for each MCMC (Markov-chain Monte Carlo) sample to yield median and 95% CrIs.

Model comparison with a Poisson branching process model

To test if our assumption of overdispersed offspring distribution better describes the data, we compared our negative-binomial branching process model with a Poisson branching process model, which assumes that the offspring distribution follows a Poisson distribution instead of negative-binomial. Since a negative-binomial distribution converges to a Poisson distribution as $k \to \infty$, we approximately implemented a Poisson branching process model by fixing k of the negative-binomial model at 10^{10}. We compared the two models by the widely-applicable Bayesian information criterion (WBIC)[15].

Simulation of the effect of underreporting

We used simulations to investigate potential bias caused by underreporting, one of the major limitations of the present study. Underreporting in some countries may be more frequent than others because of limited surveillance and/or testing capacity, causing heterogeneity in the number of cases that could have affected the estimated overdispersion. See *Extended data* (Supplementary materials)[16] for detailed methods.

The effect of a differential reproduction number for imported cases

Due to interventions targeting travellers (e.g. screening and quarantine), the risk of transmission from imported cases may be lower than that from local cases. As part of the sensitivity analysis in *Extended data*, we estimated k assuming that the reproduction number of imported cases is smaller than that of local cases.

Results

Our estimation suggested substantial overdispersion ($k << 1$) in the offspring distribution of COVID-19 (Figure 1A and Figure 2). Within the current consensus range of R_0 (2–3), k was estimated to be around 0.1 (median estimate 0.1; 95% CrI: 0.05–0.2 for R_0 = 2.5). For the R_0 values of 2–3, the estimates suggested that 80% of secondary transmissions may

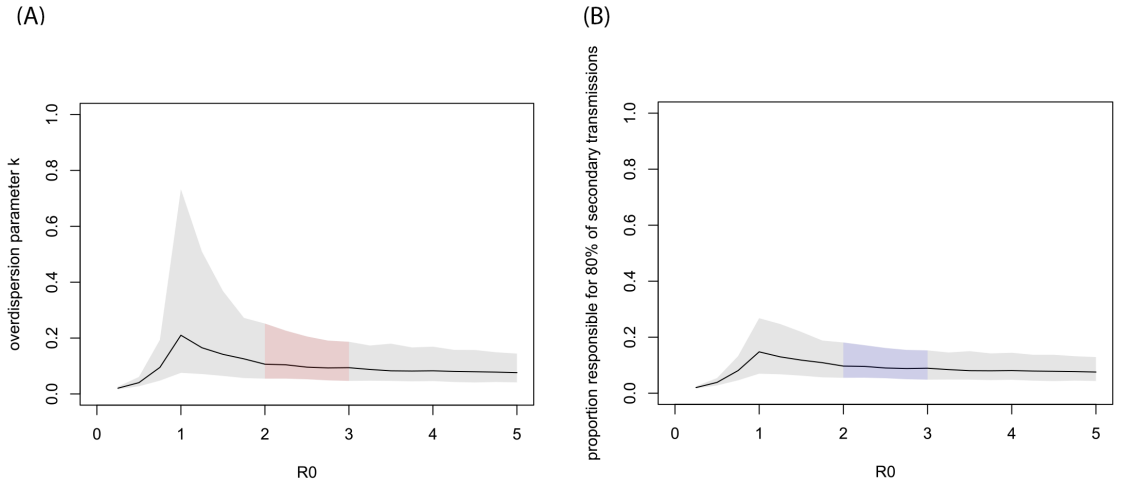


Figure 1. MCMC estimates given assumed R_0 values. (**A**) Estimated overdispersion parameter for various basic reproduction number R_0. (**B**) The proportion of infected individuals responsible for 80% of the total secondary transmissions ($p_{80\%}$). The black lines show the median estimates given fixed R_0 values and the grey shaded areas indicate 95% CrIs. The regions corresponding to the likely range of R_0 (2–3) are indicated by colour.

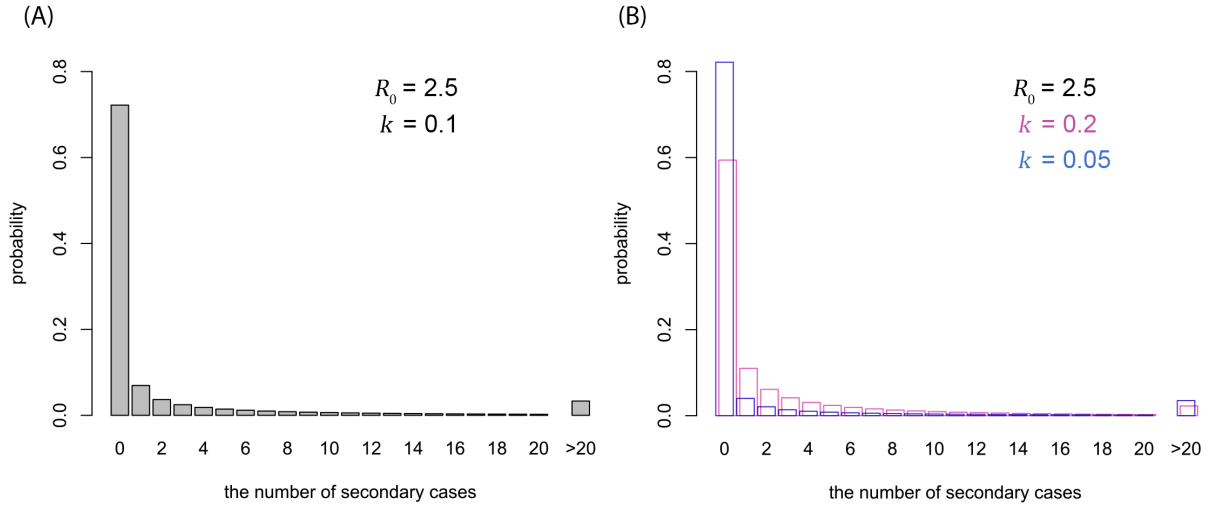


Figure 2. Possible offspring distributions of COVID-19. (**A**) Offspring distribution corresponding to R_0 = 2.5 and k = 0.1 (median estimate). (**B**) Offspring distribution corresponding to R_0 = 2.5 and k = 0.05 (95% CrI lower bound), 0.2 (upper bound). The probability mass functions of negative-binomial distributions are shown.

have been caused by a small fraction of infectious individuals (~10%; Figure 1B).

The result of the joint estimation suggested the likely bounds for R_0 and k (95% CrIs: R_0 1.4–12; k 0.04–0.2). The upper bound of R_0 did not notably differ from that of the prior distribution (=13.5), suggesting that our model and the data only informed the lower bound of R_0. This was presumably because the contribution of R_0 to the shape of a negative-binomial distribution is marginal when k is small (*Extended data*, Figure S1)[16]. A scatterplot (*Extended data*, Figure S2)[16] exhibited a moderate correlation between R_0 and k (correlation coefficient -0.4).

Model comparison between negative-binomial and Poisson branching process models suggested that a negative-binomial model better describes the observed data; WBIC strongly supported the negative-binomial model with a difference of 11.0 (Table 2). The simulation of the effect of underreporting suggested that possible underreporting is unlikely to cause underestimation of overdispersion parameter k (*Extended data*, Figure S3)[16]. A slight increase in the estimate of k was observed when the reproduction number for imported cases was assumed to be lower due to interventions (*Extended data*, Table S1).

Table 2. Model comparison between negative-binomial and Poisson branching process models.

Model	Parameter 95% CrIs		WBIC	ΔWBIC
	R_0	k		
Negative-binomial	1.4–12	0.04-0.2	45.6	0
Poisson	0.95–1.2	10^{10} (fixed)	56.6	11.0

Discussion

Our results suggested that the offspring distribution of COVID-19 is highly overdispersed. For the likely range of R_0 of 2–3, the overdispersion parameter k was estimated to be around 0.1, suggesting that the majority of secondary transmission may be caused by a very small fraction of individuals (80% of transmissions caused by ~10% of the total cases). These results are consistent with a number of observed superspreading events observed in the current COVID-19 outbreak[17], and also in line with the estimates from the previous SARS/MERS outbreaks[8].

The overdispersion parameter for the current COVID-19 outbreak has also been estimated by stochastic simulation[18] and from contact tracing data in Shenzhen, China[19]. The former study did not yield an interpretable estimate of k due to the limited data input. In the latter study, the estimates of R_c (the effective reproduction number) and k were 0.4 (95% confidence interval: 0.3–0.5) and 0.58 (0.35–1.18), respectively, which did not agree with our findings. However, these estimates were obtained from pairs of cases with a clear epidemiological link and therefore may have been biased (downward for R_0 and upward for k) if superspreading events had been more likely to be missed during the contact tracing.

Although cluster size distributions based on a branching process model are useful in inference of the offspring distribution from limited data[12,20], they are not directly applicable to an ongoing outbreak because the final cluster size may not yet have been observed. In our analysis, we adopted an alternative approach which accounts for possible future

growth of clusters to minimise the risk of underestimation. As of 27 February 2020, the majority of the countries in the dataset had ongoing outbreaks (36 out of 43 countries analysed, accounting for 2,788 cases of the total 2,816). Even though we used the case counts in those countries only as the lower bounds of future final cluster sizes, which might have only partially informed of the underlying branching process, our model yielded estimates with moderate uncertainty levels (at least sufficient to suggest that k may be below 1). Together with the previous finding suggesting that the overdispersion parameter is unlikely to be biased downwards[21], we believe our analysis supports the possibility of highly-overdispersed transmission of COVID-19.

A number of limitations need to be noted in this study. We used the confirmed case counts reported to WHO and did not account for possible underreporting of cases. Heterogeneities between countries in surveillance and intervention capacities, which might also be contributing to the estimated overdispersion, were not considered (although we investigated such effects by simulations; see *Extended data*, Figure S3)[16]. Reported cases whose site of infection classified as unknown, which should in principle be counted as either imported or local cases, were excluded from analysis. Some cases with a known site of infection could also have been misclassified (e.g., cases with travel history may have been infected locally). The distinction between countries with and without ongoing outbreak (7 days without any new confirmation of cases) was arbitrary. However, we believe that our conclusion is robust because the distinction does not change with different thresholds (4–14 days), within which the serial interval of SARS-CoV-2 is likely to fall[22,23].

Our finding of a highly-overdispersed offspring distribution suggests that there is benefit to focusing intervention efforts on superspreading. As most infected individuals do not contribute to the expansion of transmission, the effective reproduction number could be drastically reduced by preventing relatively rare superspreading events. Identifying characteristics of settings that could lead to superspreading events will play a key role in designing effective control strategies.

Data availability

Source data

Zenodo: Extended data: Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. https://doi.org/10.5281/zenodo.3740348[16].

This project contains the following source data taken from references 10 and 11:

- bycountries_27Feb2020.csv. (Imported/local case counts by country from WHO situation report 38[10].)
- dailycases_international_27Feb2020.csv. (Daily case counts by country from COVID2019.app[11].)

Extended data

Zenodo: Extended data: Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. https://doi.org/10.5281/zenodo.3911576[16].

This project contains the following extended data

- supplementarymaterials.pdf. (Supplementary material: Estimating the amount of superspreading using outbreak sizes of COVID-19 outside China.)
- figS1.tif. (Figure S1. Offspring distributions for different R_0 values. The probability mass functions of negative-binomial distributions are shown. The overdispersion parameter k is fixed at 0.1.)
- figS2.tif. (Supplementary Figure 2. Scatter plot of MCMC samples from a joint estimation of R_0 and k. The dotted line represents the threshold $R_0 = 1$)
- figS3.tif. (Supplementary Figure 3. Estimates of over-dispersion from simulations with underreporting. (A) Maximum-likelihood estimates (MLEs) of over-dispersion parameter k with different distributions for country-specific reporting probability q_i (including constant $q_i = 1$). Both imported and local cases are assumed to be reported at probability q_i in country i. The blue dotted line indicates the true value $k = 0.1$. (B) MLEs where imported cases were assumed to be fully reported and local cases were reported at probability q_i. (C) Probability density functions for beta distributions used in the simulation.)

Code availability

The reproducible code is available at: https://github.com/akira-endo/COVID19_clustersize.

Archived code at time of publication: https://doi.org/10.5281/zenodo.3741743[14].

Acknowledgements

This study was greatly motivated and inspired by the analysis published online by Kyra Grantz, C. Jessica E. Metcalf and Justin Lessler (https://hopkinsidd.github.io/nCoV-Sandbox/DispersionExploration.html). We thank the authors for insightful inputs and contribution. We also thank Seth Blumberg for valuable feedback.

Members of the Centre for Mathematical Modelling of Infectious Diseases (CMMID) COVID-19 Working Group (random order):

Rosalind M Eggo, Billy J Quilty, Nikos I Bosse, Kevin van Zandvoort, James D Munday, Stefan Flasche, Alicia Rosello, Mark Jit, W John Edmunds, Amy Gimma, Yang Liu, Kiesha Prem, Hamish Gibbs, Charlie Diamond, Christopher I Jarvis, Nicholas Davies, Fiona Sun, Joel Hellewell, Timothy W Russell, Thibaut Jombart, Samuel Clifford, Petra Klepac, Graham Medley, Carl A B Pearson

CMMID COVID-19 working group funding statements:

Rosalind M Eggo (HDR UK (MR/S003975/1)), Billy J Quilty (National Institute for Health Research (NIHR) (16/137/109)), Kevin van Zandvoort (Elrha's Research for Health in Humanitarian Crises (R2HC) Programme), James D

Munday (Wellcome Trust (210758/Z/18/Z)), Stefan Flasche (Wellcome Trust (208812/Z/17/Z)), Alicia Rosello (NIHR (PR-OD-1017-20002)), Mark Jit (Gates (INV-003174), NIHR (16/137/109)), Amy Gimma (RCUK/ ESRC (ES/P010873/1)), Yang Liu (Gates (INV-003174), NIHR (16/137/109)), Kiesha Prem (Gates (INV-003174)), Hamish Gibbs (NIHR (ITCRZ 03010)), Charlie Diamond (NIHR (16/137/109)), Christopher I Jarvis (RCUK/ESRC (ES/P010873/1)), Nicholas Davies (NIHR (HPRU-2012-10096)), Fiona Sun (NIHR EPIC grant (16/137/109)), Joel Hellewell (Wellcome Trust (210758/Z/18/Z)), Timothy W Russell (Wellcome Trust (206250/Z/17/Z)), Thibaut Jombart (RCUK/ESRC (ES/P010873/1), UK PH RST, NIHR HPRU Modelling Methodology), Samuel Clifford (Wellcome Trust (208812/Z/17/Z)), Petra Klepac (Gates (INV-003174))

References

1. Zhu N, Zhang D, Wang W, *et al.*: **A Novel Coronavirus from Patients with Pneumonia in China, 2019.** *N Engl J Med.* 2020; **382**(8): 727–733.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Lai CC, Shih TP, Ko WC, *et al.*: **Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges.** *Int J Antimicrob Agents.* 2020; **55**(3): 105924.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Zhao S, Lin Q, Ran J, *et al.*: **Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak.** *Int J Infect Dis.* 2020; **92**: 214–217.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Zhang S, Diao M, Yu W, *et al.*: **Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis.** *Int J Infect Dis.* 2020; **93**: 201–204.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Abbott S, Hellewell J, Munday J, *et al.*: **The transmissibility of novel Coronavirus in the early stages of the 2019-20 outbreak in Wuhan: Exploring initial point-source exposure sizes and durations using scenario analysis [version 1; peer review: 1 approved].** *Wellcome Open Res.* 2020; **5**: 17.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Headquarters for Novel Coronavirus Disease Control; Ministry of Health Labour and Welfare: **Basic Policies for Novel Coronavirus Disease Control.** 2020.
Reference Source
7. Department of Health and Social Care, Hancock M: **Press release: Government outlines further plans to support health and social care system in fight against COVID-19.** 2020; [cited 9 Mar 2020].
Reference Source
8. Kucharski AJ, Althaus CL: **The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission.** *Euro Surveill.* 2015; **20**(25): 14–8.
PubMed Abstract | Publisher Full Text
9. Lloyd-Smith JO, Schreiber SJ, Kopp PE, *et al.*: **Superspreading and the effect of individual variation on disease emergence.** *Nature.* 2005; **438**(7066): 355–359.
PubMed Abstract | Publisher Full Text | Free Full Text
10. World Health Organization: **Coronavirus disease 2019 (COVID-19) Situation Report – 38.** 2020.
Reference Source
11. **COVID2019.app - LIVE stats and graphs.** 2020; [cited 4 Mar 2020].
Reference Source
12. Blumberg S, Funk S, Pulliam JR: **Detecting differential transmissibilities that affect the size of self-limited outbreaks.** Wilke CO, editor. *PLoS Pathog.* 2014; **10**(10): e1004452.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Chen MH, Schmeiser B: **Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers.** *J Comput Graph Stat.* 1993; **2**(3): 251–272.
Publisher Full Text
14. Endo A, Abbott S, Kucharski AJ, *et al.*: **Estimating the amount of superspreading using outbreak sizes of COVID-19 outside China (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3741743
15. Watanabe S: **A Widely Applicable Bayesian Information Criterion.** 2013; **14**: 867–897.
Publisher Full Text
16. Endo A, Abbott S, Kucharski AJ, *et al.*: **Extended data: Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China.** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3911576
17. Liu Y, Eggo RM, Kucharski AJ: **Secondary attack rate and superspreading events for SARS-CoV-2.** *Lancet.* 2020; **395**(10227): e47.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Riou J, Althaus CL: **Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020.** *Euro Surveill.* 2020; **25**(4): 2000058.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Bi Q, Wu Y, Mei S, *et al.*: **Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts.** *medRxiv.* 2020; 2020.03.03.20028423.
Publisher Full Text
20. Blumberg S, Lloyd-Smith JO: **Inference of R(0) and transmission heterogeneity from the size distribution of stuttering chains.** Ferguson N, editor. *PLoS Comput Biol.* 2013; **9**(5): e1002993.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Lloyd-Smith JO: **Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases.** Rees M, editor. *PLoS One.* 2007; **2**(2): e180.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Li Q, Guan X, Wu P, *et al.*: **Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia.** *N Engl J Med.* 2020; **382**(13): 1199–1207.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Nishiura H, Linton NM, Akhmetzhanov AR: **Serial interval of novel coronavirus (COVID-19) infections.** *Int J Infect Dis.* 2020; **93**: 284–286.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 13 July 2020

https://doi.org/10.21956/wellcomeopenres.17714.r39515

✔ **Lin Wang**
Mathematical Modelling of Infectious Diseases Unit, UMR2000, CNRS, Institut Pasteur, Paris, France

Nice revision.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modeling, epidemiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 07 July 2020

https://doi.org/10.21956/wellcomeopenres.17694.r39407

✔ **Lin Wang**
Mathematical Modelling of Infectious Diseases Unit, UMR2000, CNRS, Institut Pasteur, Paris, France

The authors have addressed all comments raised in previous review.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modeling, epidemiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 July 2020

https://doi.org/10.21956/wellcomeopenres.17694.r39406

Kaiyuan Sun
Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

The authors have fully addressed the suggestions and I do not have further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modeling.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 May 2020

https://doi.org/10.21956/wellcomeopenres.17377.r38601

Kaiyuan Sun
Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

In this study, the authors estimate the over-dispersion of SARS-CoV-2 transmission using imported and local COVID-19 cases reported during the early phase of the epidemic through fitting the outbreak size distribution of a over-dispersed branching process. The analysis is well executed. The manuscript is well

written and the results are clearly presented. However, the data used in the study may have a few potential bias representing several alternative scenarios that I recommend the authors to explore:

1. The number of imported cases is likely an underestimate of the true number of cases as screening of travelers is unlikely to reach high detection rate for SARS-CoV-2[1]. Certain studies estimated the reporting rate is only around 30-40%, even for countries with high surveillance intensity[2]. This is likely different from the reporting rate of local cases (see point 3). I recommend the authors explore reporting rate of imported cases and local cases separately.

2. The over-dispersion estimated for SARS in the previous study is under un-controlled epidemic scenario[3]. However, for imported cases detected through travel screening, certain control measures is likely in-place such as isolation/quarantine, which will reduce the effective reproduction number, thus in Figure 1, the effective R0 range for imported cases could extend to <1.

3. It's also quite likely that local transmission is heavily under-reported during February as well. A way to gauge this under-detection is to see when each country reported the first few deaths due to COVID-19. Assuming an infection fatality of 1% will suggest a few hundred cumulative infections about 2 weeks before the detection of death. The authors already listed a number of deaths at the same date of case reporting, I recommend the authors also reports the number of deaths 2-weeks later (or the delay from case detection to death that the authors finds appropriate) and comment on the possible rate of under reporting for local cases, and together with point 1, how it may affects the estimates of k.

References

1. Gostic K, Gomez A, Mummah R, Kucharski A, et al.: Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *eLife*. 2020; **9**. Publisher Full Text
2. Niehus R, De Salazar P, Taylor A, Lipsitch M: Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. *medRxiv*. 2020. Publisher Full Text
3. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM: Superspreading and the effect of individual variation on disease emergence.*Nature*. 2005; **438** (7066): 355-9 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modeling.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 30 Jun 2020

Akira Endo, London School of Hygiene & Tropical Medicine, London, UK

Thank you for your constructive suggestions. We have discussed potential biases as suggested and performed additional analysis where applicable. Please find our responses below.

1. The number of imported cases is likely an underestimate of the true number of cases as screening of travelers is unlikely to reach high detection rate for SARS-CoV-2. Certain studies estimated the reporting rate is only around 30-40%, even for countries with high surveillance intensity. This is likely different from the reporting rate of local cases (see point 3). I recommend the authors explore reporting rate of imported cases and local cases separately.

> We agree that the number of imported cases is likely to be under ascertained. In our sensitivity analysis, we explored how the estimated value of k may be biased with different probabilities of reporting. We found that the same level of underreporting for both imported and local cases tend to result in overestimation of k, while the estimate was hardly affected if the imported cases are fully reported but the local cases are underreported. As the scenario suggested by the reviewer may likely to lie in between these two extreme scenarios we explored, we expect that such scenario would lead to the overestimation of k (but not as much as our “equally underreported” scenario; Figure S3A).

2. The over-dispersion estimated for SARS in the previous study is under un-controlled epidemic scenario. However, for imported cases detected through travel screening, certain control measures is likely in-place such as isolation/quarantine, which will reduce the effective reproduction number, thus in Figure 1, the effective R0 range for imported cases could extend to <1.

> We have included the suggested scenario as part of our sensitivity analysis. Assuming that R0 for local cases is 2.5, we varied the effective reproduction number for the imported cases (0.5, 0.8 and 1.2) and estimated the value of k. We found that the estimates of k were larger than those in our baseline analysis for R0 = 2.5 if the assumed reproduction number for the imported cases (R_I) was below 1 (k = 0.3 for R_I = 0.5; k = 0.2 for R_I = 0.8). For R_I = 1.2, the estimate was similar to our baseline analysis (k = 0.1). We have added the description of this additional analysis in the supplementary document.

3. It's also quite likely that local transmission is heavily under-reported during February as

well. A way to gauge this under-detection is to see when each country reported the first few deaths due to COVID-19. Assuming an infection fatality of 1% will suggest a few hundred cumulative infections about 2 weeks before the detection of death. The authors already listed a number of deaths at the same date of case reporting, I recommend the authors also reports the number of deaths 2-weeks later (or the delay from case detection to death that the authors finds appropriate) and comment on the possible rate of under reporting for local cases, and together with point 1, how it may affects the estimates of k.

> As the reviewer suggests, the number of deaths is a useful measure to assess underreporting. Assuming the average infection fatality 1% for the initial deaths may be subject to bias because the earliest cases may have specific age profiles that result in a different fatality. However, when averaged over the dataset, the overall case fatality may suggest the possible degree of underreporting in the dataset. As the mean lags of 8-13 days from case confirmation to death have been used for early outbreaks in existing studies [1,2], we referred to the WHO situation report 45 (5 March) and 52 (12 March), published on the 7^{th} and 14^{th} day from the situation report 38 we used in the analysis [3,4]. The total number of deaths in the countries included in our analysis was 168 and 1,065, respecctively. Given 2,815 total confirmed cases as of February 27^{th}, these suggest ascertainment ratios of 16.8% and 2.6%, respectively (assuming the true infection fatality risk is 1%). However, these ratios may be underestimates because of the rapid growth in both cases and deaths. It was suggested that the lag distribution from confirmation to death has a large variation (coefficient of variation 50%-100% [1,2]), and early-reported deaths from the cases confirmed later than February 27^{th} may have inflated the number of deaths in situation reports 45 and 52. In either case, we believe that the assumed reporting probability in our sensitivity analysis (Figure S3C) was consistent overall with these observations. Although we did not include the above calculation in the manuscript because it is only a rough estimation in which we are not completely confident, we cited Niehus et al. to show that our assumed range of the reporting probability in the sensitivity analysis was plausible:
(Supplementary document, Section 3) *An existing study suggested 38% as an optimistic global estimate of the detection probability for imported cases from Wuhan, China, with a substantial variation between countries [1].*

1. Mizumoto K, Chowell G. Estimating Risk for Death from Coronavirus Disease, China, January–February 2020. Emerg Infect Dis. 2020;26(6):1251-1256. https://dx.doi.org/10.3201/eid2606.200233
2. Russell TW, Hellewell J, Jarvis CI, et al. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill*. 2020;25(12):2000256. doi:10.2807/1560-7917.ES.2020.25.12.2000256
3. World Health Organization: Coronavirus disease 2019 (COVID-19) Situation Report – 45. 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200305-sitrep-45-cov
4. World Health Organization: Coronavirus disease 2019 (COVID-19) Situation Report – 52. 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200312-sitrep-52-cov

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 April 2020

https://doi.org/10.21956/wellcomeopenres.17377.r38418

Lin Wang
Mathematical Modelling of Infectious Diseases Unit, UMR2000, CNRS, Institut Pasteur, Paris, France

In this manuscript, Endo *et al*. estimated the overdispersion of COVID-19 transmission outside of China. The authors collected the number of imported and local cases in each affected country from the World Health Organization situation report. Using likelihood-based inference, they fitted a negative-binomial or Poisson offspring distribution to the empirical data. In summary, this study is scientifically sound and well presented. I only have a few suggestions.

1. The authors may wish to add one or two sentences about the convergence of MCMC chains, such as the diagnosis used.

2. If I understood correctly, a thinning interval of 10 is used to sample the raw chains. With this thinning interval, is the auto-correlation sufficiently small?

3. As to the statistical model, it seems that the authors assumed that all imported cases arrived and triggered the local epidemic at the same time. If the cases arrived at different time points, will the inferred results be different? This manuscript might be useful to understand the effect of continuous seeding: Characterizing the dynamics underlying global spread of epidemics[1].

References

1. Wang L, Wu J: Characterizing the dynamics underlying global spread of epidemics. *Nature Communications*. 2018; **9** (1). Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modeling, epidemiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 30 Jun 2020

Akira Endo, London School of Hygiene & Tropical Medicine, London, UK

Thank you for your comments. We have added some technical details for better clarity as suggested. Please find our responses to each point below.

1. The authors may wish to add one or two sentences about the convergence of MCMC chains, such as the diagnosis used.
2. If I understood correctly, a thinning interval of 10 is used to sample the raw chains. With this thinning interval, is the auto-correlation sufficiently small?

> We used the effective sample size to assess the convergence and (the weakness of) autocorrelation. For clarity, we have added a line in Statistical analysis section: “We confirmed that the final 500 samples have an effective sample size of at least 300, indicating sufficiently low auto-correlation.”

3. As to the statistical model, it seems that the authors assumed that all imported cases arrived and triggered the local epidemic at the same time. If the cases arrived at different time points, will the inferred results be different? This manuscript might be useful to understand the effect of continuous seeding: Characterizing the dynamics underlying global spread of epidemics.

> We agree with the reviewer that continuous seeding is an important issue in time-series epidemic analysis. However, our approach was not sensitive to the assumption that all the imported cases arrived at the same point in time. Because we only imposed the condition that the final cluster size has to be at least the size of the currently observed number of cases, the temporal distribution of the imported cases does not change the likelihood function we used.

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 1

Reader Comment 18 Jun 2020

Richard Falk, No affiliation, San Rafael, CA, USA

When countries have mitigations in place including wearing masks, this lowers the probability of transmission events from the more casual contacts and shifts the likelihood to more crowded close-contact indoor poor air exchange environments where they were always risky but now represent the main situations that exceed the "probability gate" limited by mask-wearing and physical distancing elsewhere. That is, **there is a bias towards super-spreading events**.

Also, when countries put mitigations in place this shifts the probability of transmission to be more with those infected individuals with higher viral load (not necessarily measured in NP swabs but viral load where it counts in the lungs bringing virus-laden mucus to the vocal folds and epiglottis) or higher droplet/aerosol output because those with moderate viral load have much lower transmission such as from wearing masks or maintaining physical distancing. That is, **there is a bias towards super-spreader individuals**.

This latter point requires a more subtle and complex analysis because it also requires non-linear limits on super-spreading individuals and these exist in the form of 1) the dose-response curve that flattens towards 100% probability of infection (i.e. you cannot infect someone more than once -- that is, you can't count your very high viral load as infecting them three times as much) and 2) the limited number of contacts an individual has during the contagious period. So as mitigations are removed, the super-spreading individuals infect at a relative lower proportional rate than moderate viral-load individuals -- that is, if the moderate viral-load individuals doubled in their R value when taking off their masks, the super-spreading individuals would be less than doubled (i.e. they are saturating in both the dose-response and opportunities).

This means that the estimated "k" value would be higher during the unmitigated exponential growth phase of disease transmission and this appears to be the case in China where most transmission was within families. After mitigation and the more restrictive the mitigation, the lower the "k" estimate would be because only those rarer individuals and events would occur and have one person infect many but have most of those infected not infect many others.

So **while it is reasonable to conclude that super-spreading individuals and/or super-spreader events are driving the majority of transmissions after mitigations are in place, it would be incorrect to conclude that this was (as much of) the situation when there were few mitigations and there was wider community spread**. If mitigations are generally followed, then this can result in outbreaks that quickly die off unless the super-spreading individuals or events were numerous enough to create chains (i.e. if the R value of a super-spreading individual was more than 10 and they represent 10% of people who are infected then the overall $R > 1$ would continue the spread).

It would also be good to be able to distinguish between super-spreader individuals vs. super-spreading events because the latter is more readily controlled via public policy by limiting crowds and requiring wearing masks. So far it appears that outbreaks are largely super-spreading events and while they could be caused from individuals with higher viral load there does not appear to be evidence of people wearing masks indoors maintaining social distance and having proper air exchange causing outbreaks.

Competing Interests: No competing interests.

Reader Comment 06 Jun 2020

Alex Cranberg, Aspect Managemebt Corp, Houston, USA

These results and conclusions may be significantly affected by heterogeneity deriving from susceptibility as opposed to heterogeneity of transmission. Have you modeled the difference between heterogeneity of spreading vs receptivity?

Competing Interests: No competing interests were disclosed.