

# Supplementary material: Estimating the amount of superspreading using outbreak sizes of COVID-19 outside China

Akira Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Sam Abbott, Adam J Kucharski, Sebastian Funk

## 1. Negative-binomial offspring distributions for different $R_0$ values

We compared negative-binomial offspring distributions for different  $R_0$  values where the overdispersion parameter  $k$  is fixed at 0.1 (Figure S1). When  $k$  is small, different  $R_0$  values barely change the offspring distribution except for the mass for 0 and for large ( $> 20$ ) secondary cases.

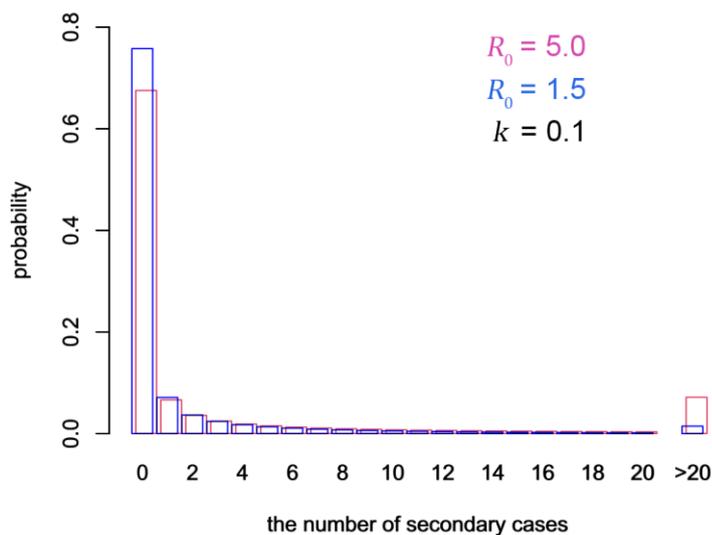


Figure S1. Offspring distributions for different  $R_0$  values. The probability mass functions of negative-binomial distributions are shown. The overdispersion parameter  $k$  is fixed at 0.1.

## 2. Joint estimation of $R_0$ and $k$

We 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  $R_0$  to prevent a divergence; the prior for the reciprocal of  $k$  was a weakly-informed half-normal (HalfNormal( $\sigma = 10$ ))). The estimated range of  $R_0$  was wide (median 4.4; 95% CrI 1.4-12)

and the upper bound did not notably differ from that of the prior distribution ( $=13.5$ ). The estimated range of  $k$  was low (median 0.08; 95% CrI 0.04-0.2), suggesting a highly heterogeneous offspring distribution. A scatterplot (Figure S2) exhibited a moderate correlation between  $R_0$  and  $k$  (correlation coefficient 0.4).

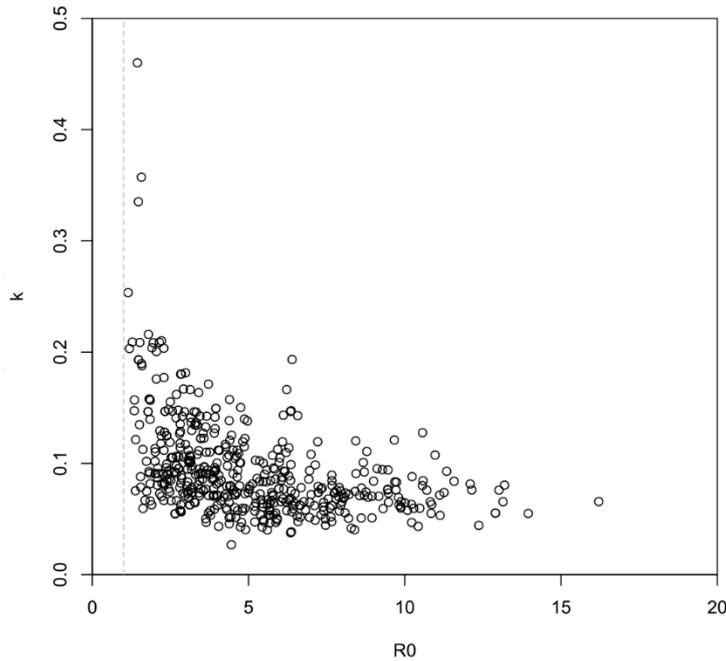


Figure S2. Scatter plot of MCMC samples from a joint estimation of  $R_0$  and  $k$ . The dotted line represents the threshold  $R_0 = 1$ .

### 3. Simulation of the effect of underreporting

One of the major limitations of the present study is potential underreporting in the dataset.

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. We used simulations to investigate potential bias caused by underreporting. First, we assumed that the same probability of reporting applies to both imported and local cases in a country. We represented the data-generating process in the presence of underreporting as a binomial sampling. Let  $s_i$  and  $x_i^0$  be the observed and true number of cases in country  $i$ , respectively.

$$s_i \sim \text{Binom}(x_i^0, q_i),$$

where  $q_i$  is the reporting probability for country  $i$ . When  $s_i$  is observed, by assuming that the prior probability for  $x_i^0$  is (improper-) uniformly distributed, we get

$$x_i^0 - s_i \sim \text{NegBinom}(s_i + 1, q_i). \quad (*)$$

We generated simulation datasets in the following steps.

1. Set  $s_i$  as the number of observed imported cases from the WHO situation report (Table 1 in the main text); sample reporting probability  $q_i$  for each country from a beta distribution (see Figure S3C) and then sample  $x_i^0$  based on Equation (\*).
2. Sample two generations of cases where  $x_i^0$  is the number of index cases using a negative-binomial-distrusted offspring distribution. Namely, for  $t = 1, 2$ ,

$$x_i^t \sim \text{NegBinom}\left(kx_i^{t-1}, \frac{k}{R_0 + k}\right).$$

We used  $R_0 = 2.5$  and  $k = 0.1$  for our simulations.

3. Sample the observed number of local cases by binomial sampling:

$$X_i^t \sim \text{Binom}(x_i^t, q_i). \quad (t = 1, 2)$$

4. Apply the likelihood-based model in the main text to the observed imported/local cases:  $(s_i, X_i^1 + X_i^2)$ , where countries with non-zero  $X_i^2$  are treated as “countries with an ongoing outbreak”.

We used the maximum-likelihood approach here (as opposed to MCMC used in the main analysis) for simplicity.  $R_0 = 2.5$  was assumed to be known and overdispersion parameter  $k$  was estimated. We ran 500 simulations for each assumed distribution of  $q_i$  and plotted the estimates (Figure S3A). Lower reporting probability introduced an upward bias in the estimates.

Next, we repeated the simulation with another scenario where the imported cases were assumed to be fully reported (100% reporting probability for imported cases) due to their awareness of the travel history. This can be implemented by skipping step 1 and using  $s_i$  as  $x_i^0$ . The degree of bias introduced in this simulation was relatively small (Figure S3B).

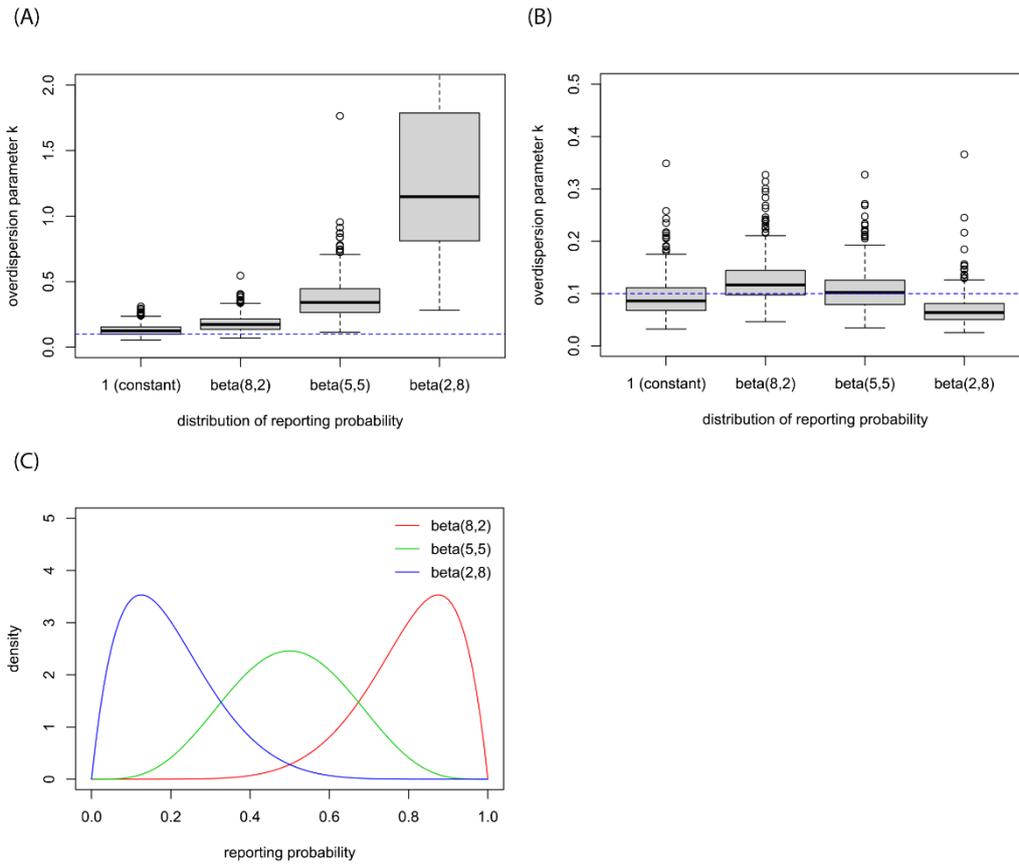


Figure S3. Estimates of overdispersion from simulations with underreporting. (A) Maximum-likelihood estimates (MLEs) of overdispersion parameter  $k$  with different distributions for country-specific reporting probability  $q_i$  (including constant  $q_i = 1$ ). Both imported and local cases are assumed 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.