

# Supplementary material

## Infection model

### Viral load

The viral load of each index case (in Ct) was constructed from a piecewise linear model of their Ct values by assuming that their Ct was 40 at exposure and cessation of shedding and then sampling their days from exposure to peak viral load,  $d_p$ , peak viral load  $Ct_{peak}$ , and days from peak to cessation,  $d_r$ , from the following distributions from Kissler et al.<sup>22</sup>:

$$\begin{aligned}d_p &\sim N_{[1,14]}(2.0, 13.4) \\ Ct_{peak} &\sim N_{[0,40]}(20.2, 19.5) \\ d_r &\sim N_{[1,30]}(6.3, 6.8)\end{aligned}$$

where  $N_{[a,b]}(\mu, \sigma^2)$  denotes a truncated normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and lower and upper limits  $a$  and  $b$ .

### Infectiousness

The probability of infectiousness for a given viral load was estimated by fitting a logistic regression model to the probability of culturing virus at that viral load,  $p_{culture}$ <sup>13</sup>:

$$\begin{aligned}culture &\sim \text{Bernoulli}(p_{culture}) \\ \text{logit}(p_{culture}) &= \beta_0 + \beta_1 vl\end{aligned}$$

where *culture* is the binary outcome of whether or not virus was cultured, and *vl* is the viral load (in RNA copies/ml) using the estimated parameter values ( $\beta_0 = -14.8$ ,  $\beta_1 = 2.1$ ) to calculate the probability of infectiousness over the course of each individual's viral load trajectory (Figure 1). The Ct value was converted to RNA copies/ml via the following formula<sup>12</sup>:

$$vl = \frac{Ct - b_{conv}}{m_{conv}} + \log_{10}(250)$$

where  $m_{conv} = -3.60971$  and  $b_{conv} = 40.93733$  are the slope and intercept of a linear regression of Ct value on  $\log_{10}$ -transformed RNA concentration from a standard curve.

### Contacts and contact durations

We sampled numbers of household and non-household conversational contacts from empirical contact distributions (the BBC Pandemic contact survey<sup>7,8,14</sup> and CoMix contact surveys<sup>9</sup>) with index case  $i$  having  $N_{HH,i}$  household (HH) contacts and  $N_{NHH,i,t}$  daily non-household (NHH) contacts (work, school and "other" contacts), where  $t$  is the time since  $i$ 's exposure in days. Household contacts were sampled once per index case, whereas non-household contacts were sampled daily. Each household and non-household contact had a duration,  $dur_{HH,i,j,t}$  and  $dur_{NHH,i,j,t}$ , defined as the proportion of a 24-hour period spent at home or outside of the home respectively, independently sampled from the CoMix survey (Figure S2).

## Infections

The infection process for contact between an index case  $i$  and their contact  $j$  was modelled as Bernoulli with the probability of infection equal to the infectiousness of the index case on the day of contact multiplied by the duration of contact for both household and non-household infections:

$$\begin{aligned}infection_{HH,i,j,t} &\sim \text{Bernoulli}(p_{culture,i,t} dur_{HH,i,j,t}) \\infection_{NHH,i,j,t} &\sim \text{Bernoulli}(p_{culture,i,t} dur_{NHH,i,j,t})\end{aligned}$$

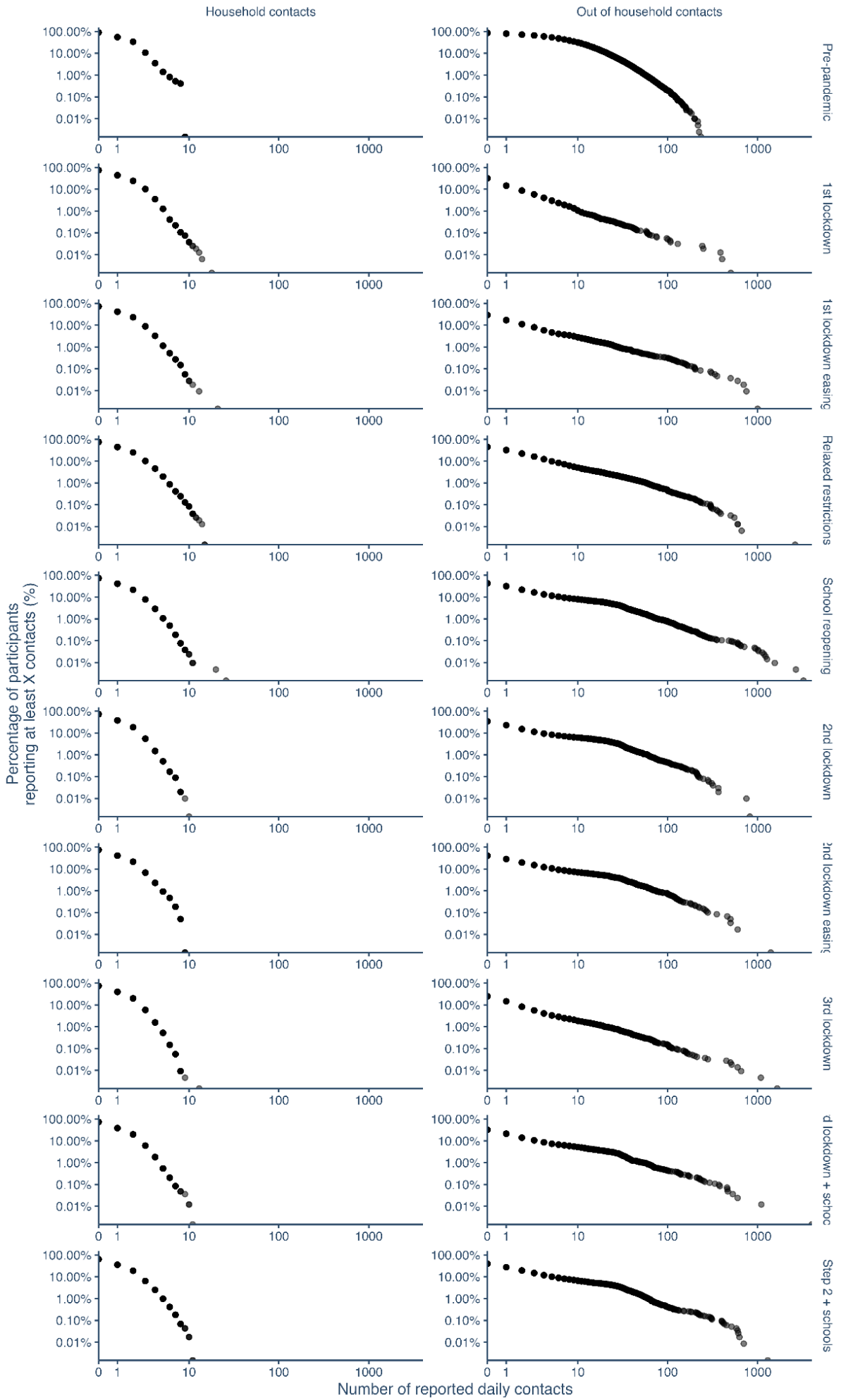
We assumed uniform susceptibility of individuals in the model which does not vary by, for example, age.

The number of secondary infections for index case  $i$  (i.e. their individual level reproduction number),  $R_i$ , was therefore given by:

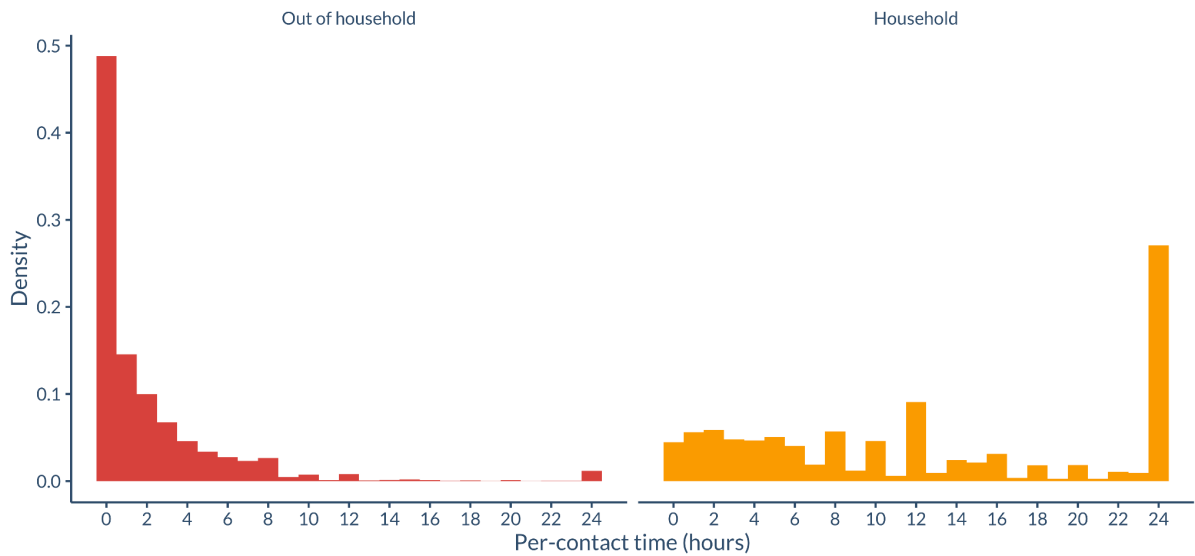
$$R_i = \sum_{t=1}^{T_i} \left( \sum_{j=1}^{N_{HH,i,t}} infection_{HH,i,j,t} + \sum_{j=1}^{N_{NHH,i,t}} infection_{NHH,i,j,t} \right)$$

where  $T_i$  is index case  $i$ 's duration of infection (defined as having infection detectable by PCR).

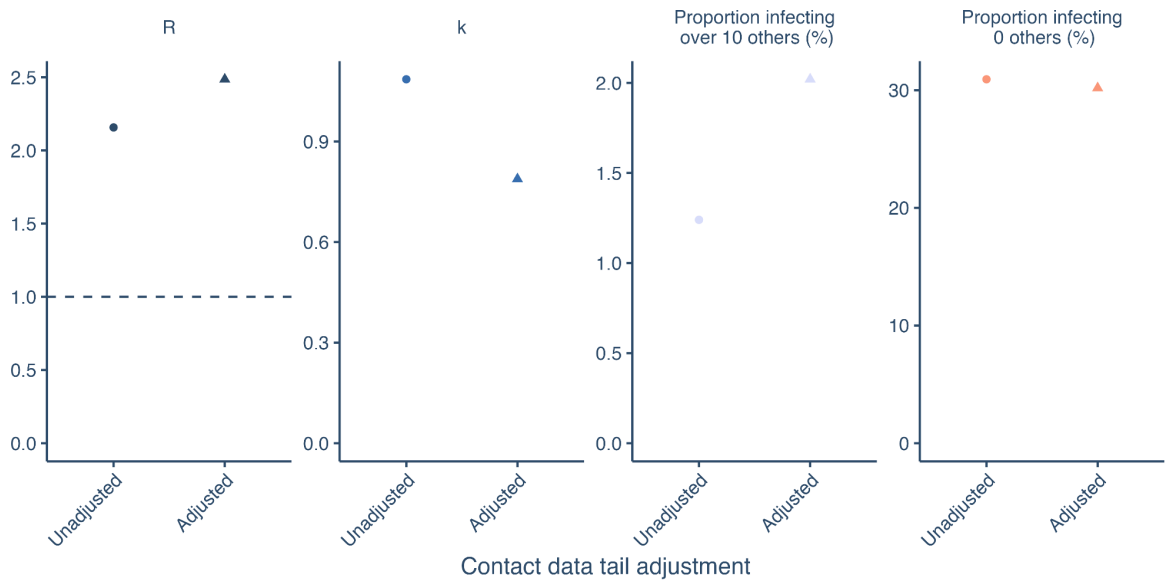
We then estimated the mean,  $R$ , and overdispersion,  $k$ , of the number of secondary infections by fitting a negative binomial distribution to the  $R_i$  values across all index cases.



**Figure S1. Distribution of the number of reported daily contacts for all timepoints before (BBC Pandemic contact survey) and during (CoMix contact survey) the pandemic in the UK between March 2020 and May 2021.**



**Figure S2: Distribution of the duration of contact for household and out-of-household contacts from the CoMix contact survey in the UK.**



**Figure S3: Sensitivity analysis of the effect on  $R$ ,  $k$ , and the proportion of infected individuals infecting over 10 and 0 others imputing a heavier tail of the contact distribution for the pre-pandemic BBC Pandemic contact survey, which lacked the ability to record raw counts of high numbers of contacts.**