

RESEARCH ARTICLE

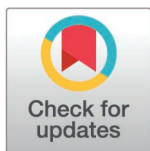
# Disentangling the drivers of heterogeneity in SARS-CoV-2 transmission from data on viral load and daily contact rates

Billy J. Quilty<sup>1,2</sup>, Lloyd A. C. Chapman<sup>1,3</sup>\*, James D. Munday<sup>1,4</sup>, Kerry L. M. Wong<sup>1</sup>, Amy Gimma<sup>1</sup>, Suzanne Pickering<sup>5</sup>, Stuart Neil<sup>5</sup>, Rui Pedro Galao<sup>5</sup>, W. John Edmunds<sup>1</sup>, Christopher I. Jarvis<sup>1</sup>, Adam J. Kucharski<sup>1</sup>

**1** Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom, **2** Center for Global Health, Charité Universitätsmedizin Berlin, Berlin, Germany, **3** School of Mathematical Sciences, Lancaster University, Lancaster, United Kingdom, **4** Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, **5** Department of Infectious Diseases, School of Immunology and Microbial Sciences, King's College London, London, United Kingdom

\* These authors contributed equally to this work.

\* [l.chapman4@lancaster.ac.uk](mailto:l.chapman4@lancaster.ac.uk)



## OPEN ACCESS

**Citation:** Quilty BJ, Chapman LAC, Munday JD, Wong KLM, Gimma A, Pickering S, et al. (2026) Disentangling the drivers of heterogeneity in SARS-CoV-2 transmission from data on viral load and daily contact rates. PLoS Comput Biol 22(8): e1014715. <https://doi.org/10.1371/journal.pcbi.1014715>

**Editor:** Nicholas Geard, University of Melbourne, AUSTRALIA

**Received:** September 26, 2025

**Accepted:** August 13, 2026

**Published:** August 28, 2026

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1014715>

**Copyright:** © 2026 Quilty et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

## Abstract

SARS-CoV-2 transmission is highly overdispersed, with a minority of individuals responsible for the majority of transmission, though the drivers of this heterogeneity are unclear. Here, we assess the contribution of variation in viral load and daily contact rates to this heterogeneity by combining published viral load estimates and contact survey data in a mathematical model to estimate the secondary infection distribution. Using data from the BBC Pandemic and CoMix contact surveys, we estimate the secondary infection distribution throughout the pandemic in the UK in 2020, and the effectiveness of frequent and pre-event rapid testing for reducing superspreading events. We find that individual heterogeneity in contacts rather than individual heterogeneity in shedding is the main driver of observed heterogeneity in the secondary infection distribution. Our results suggest that everyone testing every 3 days would reduce the reproduction number below 1 and be equivalent in terms of impact on secondary infections to everyone testing only before events with a minimum event size of 10 for pre-pandemic contact levels. This work demonstrates the potential for using viral load and contact data to estimate heterogeneity in transmission and the effectiveness of rapid testing strategies for curbing transmission in future pandemics.

## Author summary

SARS-CoV-2 spreads mainly through superspreading, with around 20% of infected individuals responsible for around 80% of secondary infections. Previous studies have inferred this using plausible assumptions about contact rates and

and reproduction in any medium, provided the original author and source are credited.

**Data availability statement:** All data and code used in the study are available online at [https://github.com/bquilty25/superspreading\\_testing](https://github.com/bquilty25/superspreading_testing).

**Funding:** The authors acknowledge funding from the Medical Research Council (MC\_PC\_19065 to WJE); European Commission (EpiPose 101003688 to WJE, CIJ, JDM, KLMW, AG); NIHR (CV220-088 - COMIX to WJE, CIJ, JDM, KLMW, AG; 16/137/109 to BJQ, 16/136/46 to BJQ); Bill & Melinda Gates Foundation (OPP1139859 to BJQ); Wellcome Trust (206250/Z/17/Z to AJK, LACC; WT098049AIA to SJDN); NIHR HPRU in Modelling & Health Economics (NIHR200908 to AJK, LACC); Huo Family Foundation (Huo Family Foundation Award to SP, SJDN); King's College London (King's Together Rapid COVID-19 Call to SJDN, RPG). The funders played no role in the study design, data analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

viral load dynamics. Here, we instead integrate data from real contact surveys conducted in the UK before and during the COVID-19 pandemic with published estimates of viral load trajectories and infectiousness to estimate the average and variation in numbers of secondary infections per case over the first year of the pandemic. Our results are consistent with observed reductions in secondary infections during periods of contact restrictions in the UK. We show that variation in numbers of daily contacts can be used to monitor superspreading risk in real-time during an epidemic. When combined with regular rapid testing, which identifies individuals with high viral loads when they are most infectious, this offers a means of effectively reducing transmission and avoiding costly blanket interventions by targeting restrictions to when and where they are most needed.

## Introduction

Transmission of SARS-CoV-2 occurs primarily through superspreading, with 20% of infections generating around 80% of secondary infections [1]. A review and meta-regression by Chen et al. [2] indicates that substantial variation in the respiratory viral load of individuals infected with SARS-CoV-2 is a driver of overdispersion in secondary infection generation. However, as most studies cited measured viral load at one point over the course of infection, this review could not answer whether this is due to some individuals being more infectious than others generally ("between-person infectiousness variability" hypothesis) or whether most individuals pass through a highly infectious period which happens to coincide with a period of high contact ("between-person contact variability and within-person infectiousness variability" hypothesis). High contact rates are a prerequisite for infecting a large number of people, and hence the potential for superspreading should have varied over the course of the COVID-19 pandemic as contact distributions changed with the enactment and relaxation of restrictions.

In this paper, we reconstruct the secondary infection distribution of SARS-CoV-2 using a model of intra- and inter-host heterogeneity in infectiousness derived from published data on viral load trajectories and infectivity combined with data on reported numbers of daily contacts from two social contact surveys in the UK. As overdispersion in the social contact data has not previously been analysed or reported, we first perform a descriptive analysis of the social contact data. While previous models of superspreading of SARS-CoV-2 have either fitted to summary data on key epidemiological metrics, such as the mean reproduction number and distributions of individual-level numbers of secondary cases from contact tracing studies [3], to estimate social contact rates, or considered a wide range of plausible contact rates [4], here we estimate the secondary infection distribution from first principles using data on social contacts gathered prior to and during the pandemic in the UK. This allows us to characterise variation in the secondary infection distribution over time under different levels of restrictions on contacts. We also consider the impact of rapid lateral flow antigen tests (LFTs), which are able to detect individuals with high viral

loads when they are most likely to be infectious [5,6], taken regularly or before events on the mean reproduction number and the potential for superspreading with differing levels of adherence and background contact rates.

The distribution of the number of secondary infections generated by each infectious individual  $i$ ,  $R_i$ , can be assumed to be negative binomial with mean equal to the mean number of secondary infections  $R$  and overdispersion parameter  $k$  representing the variation in the number of secondary infections, with smaller values of  $k$  representing greater variation,

$$R_i \sim \text{NegBin}(k, R)$$

such that the variance in  $R_i$  is

$$\text{Var}(R_i) = R + \frac{R^2}{k}.$$

Even if the mean number of secondary infections  $R$  is below 1, there may still be a considerable probability of one or more secondary infections if  $k$  is small. We estimate the utility of regular rapid LFTs for reducing  $R$  and the potential for super-spreading events by decreasing variation in numbers of secondary infections, that is, increasing  $k$ .

## Methods

### Ethics statement

This study only involved secondary analysis of publicly available contact survey data and SARS-CoV-2 viral load data. Secondary analysis of these data was approved by the London School of Hygiene and Tropical Medicine Observational Research Ethics Committee (ref 14400 and ref 21795).

### Contact data

**BBC pandemic survey.** The BBC Pandemic contact survey was conducted between September 2017 and December 2018 as part of a BBC Four documentary and involved over 40,000 participants (full details are published elsewhere [7–9]). Participants used an app to record their personal basic demographic information and the number of social contacts they made during the previous 24-hour period, as well as information such as the contact's age, type of interaction, and setting (home, work, school, other). Children under 13 years of age were excluded from participating for ethical reasons.

**CoMix survey.** The CoMix survey was a behavioural survey launched on 24th of March 2020 to gather social and behavioural data to aid the response to the COVID-19 pandemic. The contact survey was based on the POLYMOD contact survey [10]. The sample was broadly representative of the UK adult population. Participants were invited to respond to the survey once every two weeks. Weekly data was collected by running two alternating panels. Parents completed the survey on behalf of children (17 years old or younger). Participants recorded direct, face-to-face contacts made on the previous day, specifying certain characteristics for each contact including the age and sex of the contact, whether contact was physical (skin-to-skin contact), and where contact occurred (e.g., at home, work, while undertaking leisure activities, etc.). Full details have been published elsewhere [11].

From the 24th May 2020 onwards, the CoMix survey included the ability to record an estimated mass contacts count in situations where it would be infeasible to record the detailed information of all contacts made, such as large gatherings. Since the BBC Pandemic survey did not include this option, and this potentially truncated the true contact distribution, we conducted a sensitivity analysis of the effect on  $R$  and  $k$  of imputing a heavier tail for the BBC Pandemic contact distribution. We did this by fitting a generalised Pareto distribution to the number of individuals reporting over 250 contacts in relaxed restrictions time periods (“Relaxed restrictions” and “School reopening” in Table 1) from the CoMix data and sampling proportionally from this distribution for the BBC Pandemic contacts.

**Table 1. Percentage of participants in the CoMix and BBC Pandemic surveys with more than a certain number of contacts in different time periods in the UK from March 2020 to May 2021, and pre-pandemic in 2017-2018.**

| Time period*           | From       | To         | N <sup>†</sup> | Over 5 (%) | Over 10 (%) | Over 20 (%) | Over 50 (%) | Over 100 (%) | Over 200 (%) |
|------------------------|------------|------------|----------------|------------|-------------|-------------|-------------|--------------|--------------|
| Pre-pandemic           | 01/09/2017 | 01/12/2018 | 40162          | 65.8       | 38.9        | 13.7        | 1.6         | 0.2          | 0.0          |
| 1st lockdown           | 23/03/2020 | 03/06/2020 | 15971          | 6.5        | 1.5         | 0.4         | 0.1         | 0.1          | 0.0          |
| 1st lockdown easing    | 04/06/2020 | 29/07/2020 | 10764          | 8.5        | 3.3         | 1.6         | 0.6         | 0.3          | 0.1          |
| Relaxed restrictions   | 30/07/2020 | 03/09/2020 | 15547          | 17.0       | 6.2         | 3.1         | 1.2         | 0.5          | 0.2          |
| School reopening       | 04/09/2020 | 24/10/2020 | 21031          | 15.9       | 8.7         | 6.1         | 2.0         | 0.8          | 0.3          |
| 2nd lockdown           | 05/11/2020 | 02/12/2020 | 10106          | 11.3       | 6.7         | 4.8         | 1.3         | 0.5          | 0.2          |
| 2nd lockdown easing    | 03/12/2020 | 19/12/2020 | 5957           | 15.0       | 8.0         | 5.3         | 1.8         | 0.8          | 0.2          |
| 3rd lockdown           | 05/01/2021 | 07/03/2021 | 21755          | 5.3        | 2.1         | 1.1         | 0.4         | 0.2          | 0.1          |
| 3rd lockdown + schools | 08/03/2021 | 31/03/2021 | 8300           | 10.3       | 5.8         | 3.7         | 1.1         | 0.5          | 0.2          |
| Step 2 + schools       | 16/04/2021 | 16/05/2021 | 11607          | 14.2       | 7.3         | 5.0         | 1.6         | 0.4          | 0.2          |

\* Different pandemic time periods are based on lockdowns and different levels of restrictions.

<sup>†</sup> N is the number of survey responses in each time period.

<https://doi.org/10.1371/journal.pcbi.1014715.t001>

## Analysis of social contact data

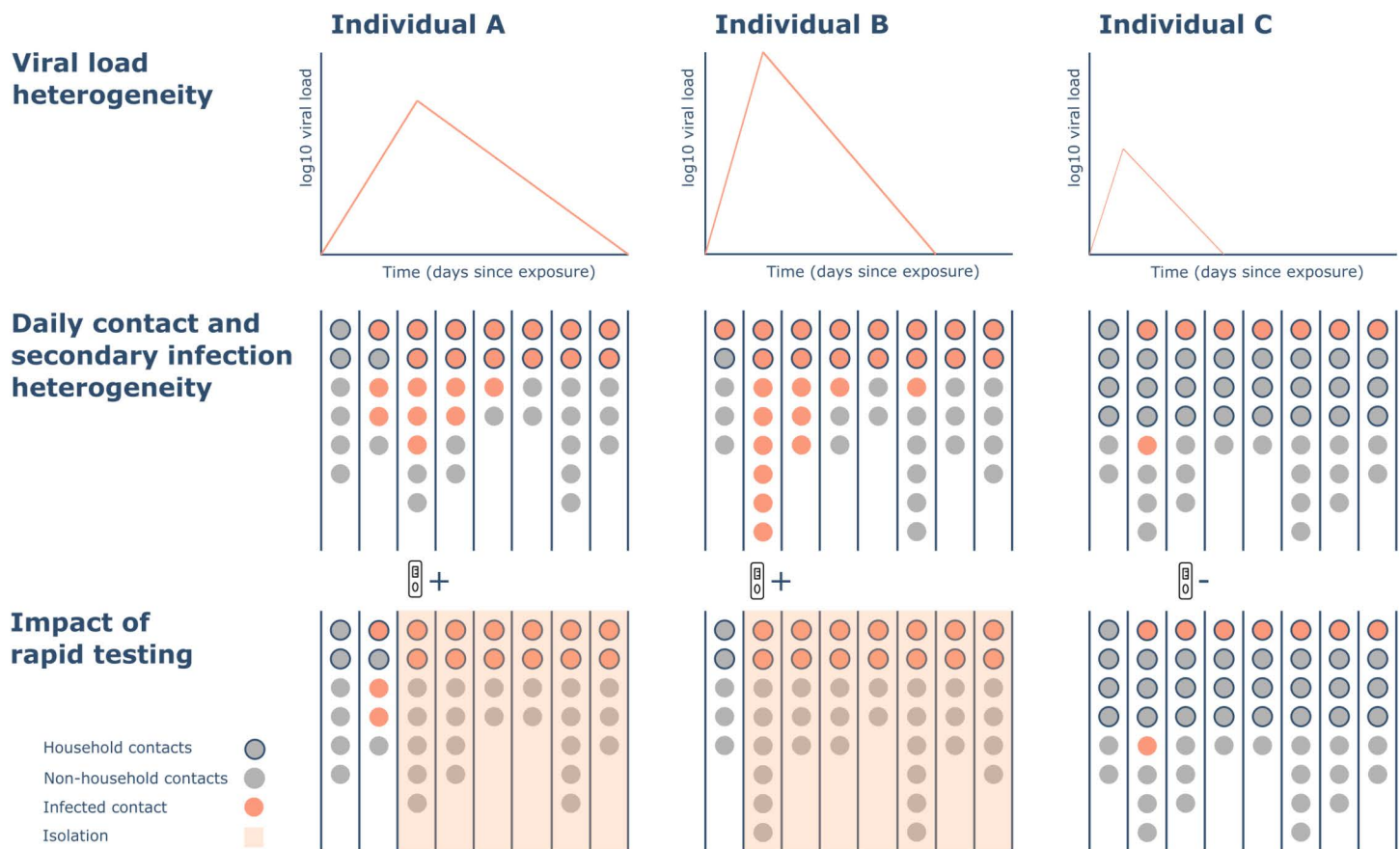
We compared contact distributions (overall and stratified by household and out-of-household) for the BBC Pandemic survey [7–9] and CoMix survey [11,12]. To account for differences in numbers of participants per time period, we calculated and plotted the percentage of participants reporting more than 5, 10, 20, 50, 100, and 200 contacts in the 24-hour period prior to filling in the survey for the BBC Pandemic survey (here referred to as Pre-pandemic) and nine indicative time periods between March 2020 and May 2021 from the CoMix survey representing different levels of restrictions, as defined in previous work [11,12] (Table 1). We fitted negative binomial distributions to numbers of daily contacts in the different time periods to estimate the mean and dispersion of the contact distributions. To assess whether the negative binomial adequately captured the contact distribution, or whether an excess of zero-contact individuals required a zero-inflated extension, we additionally fitted a zero-inflated negative binomial (ZINB) distribution to the per-individual daily contact counts for three indicative time periods (Pre-pandemic, 1st lockdown, School reopening) and compared models using the Akaike information criterion (AIC) and a likelihood ratio test (LRT). The negative binomial was found to be sufficient in all periods (zero-inflation parameter  $\pi < 6 \times 10^{-8}$ ,  $\Delta\text{AIC} = +2.0$ , LRT  $p > 0.97$ ; S2 Table).

## Reconstruction of secondary infection distribution

We reconstructed the secondary infection distribution of SARS-CoV-2 in the UK over the course of 2020 as described below. A schematic of the model is shown in Fig 1. We restricted our analysis to the first six time periods of the CoMix survey in 2020 (23rd March to 19th December) prior to the widespread emergence of more transmissible variants and mass vaccination.

## Viral load trajectories

To predict the distribution of individual-level transmission at different points in time, we simulated 10,000 individual respiratory viral load trajectories of index cases over the course of infection (Fig 3A). The viral load of each index case was constructed from a piecewise linear function of their Ct values defined by a proliferation phase (days from exposure to peak,  $d_p$ ), a clearance phase (days from peak to cessation,  $d_r$ ) and a peak Ct value,  $\text{Ct}_p$  (Fig 1). The three parameters were drawn from distributions given in Kissler et al. [13]:



**Fig 1. Schematic of the model.** The viral load and number of daily contacts (circles) varies from person to person and over time, influencing the number of secondary infections (stratified by household (outlined) and non-household (no outline)) they generate. The viral load at time of testing also determines the likelihood they will test positive and subsequently isolate.

<https://doi.org/10.1371/journal.pcbi.1014715.g001>

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

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  $b$ . The Ct value at exposure and cessation of shedding was assumed to be 40 (negative).

The Ct value was converted to viral load vl (in log10 RNA copies/ml) via the following formula [13]:

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

where  $m_{conv} = -3.60971$  and  $b_{conv} = 40.93733$  are the slope and intercept of a linear regression of Ct value on log10-transformed RNA concentration from a standard curve.

## Infectiousness

The probability of infectiousness for a given viral load (in Ct) was estimated by fitting a logistic regression model to the probability of culturing virus at that viral load [14] (Fig 3B),  $p_{\text{culture}}$ :

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

where culture is the binary outcome of whether or not virus was cultured. The probability of infectiousness over the course of each individual's viral load trajectory was then calculated using the estimated regression coefficients  $\beta_0$  and  $\beta_1$  to produce infectiousness trajectories (Fig 3C). Uncertainty in  $\beta_0$  and  $\beta_1$  was accounted for by drawing values for each individual from a bivariate normal distribution centred at the maximum likelihood estimate  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = (-14.8, 2.1)$ :

$$\beta \sim \mathcal{N}(\hat{\beta}, \Sigma)$$

where  $\Sigma$  is the estimated covariance matrix from the logistic regression. Drawing values for the regression coefficients in this way may be taken to represent a degree of variation in infectiousness between individuals for the same viral load, as Ke et al. [6] concluded was necessary to explain observed heterogeneity in infectiousness. We use viral culture positivity as our proxy for infectiousness because it indicates the presence of replication-competent virus and is therefore believed to provide a better indicator of transmission potential than Ct values from PCR measurements [15]. It also allows us to make use of empirical relationships linking viral load, viral culture positivity, and test detection: (i) viral load determines the probability of culturing viable virus, with the logistic relationship fitted directly from data in Pickering et al. [14] (Fig 3B); and (ii) LFT positivity closely tracks viable virus in controlled human challenge studies [5].

## Contacts and contact durations

For the pre-pandemic time period and each of the pandemic time periods, we randomly sampled the number of household (HH) contacts  $N_{\text{HH},i}$  and daily non-household (NHH) contacts  $N_{\text{NHH},i,t}$  for index case  $i$ , where  $t$  is the time since exposure, from the empirical contact distribution for that time period from the BBC Pandemic contact survey [7–9] and CoMix contact survey [11,12]. The number of household and non-household contacts were sampled jointly by sampling a survey participant at random, to preserve correlation in individual-level numbers of household and non-household contacts. The numbers of non-household contacts on different days were sampled from the repeat survey responses of the chosen participant for the given time period, to maintain correlation in individual-level numbers of non-household contacts over time. Each household and non-household contact had a duration,  $\text{dur}_{\text{HH},i,j,t}$  and  $\text{dur}_{\text{NHH},i,j,t}$ , defined as the proportion of a 24-hour period that the contact lasted (S1 Fig). For the pandemic time periods the duration was sampled from the contact durations for the same survey participant as the contact numbers, where these were available. This ensured that correlation between number of contacts and contact duration was maintained (S2 Fig). For the pre-pandemic time period, the duration was sampled from the contact durations in the POLYMOD survey, as contact duration data is not publicly available for the BBC Pandemic survey. Where the contact duration was missing, it was randomly sampled from the pre-pandemic or pandemic contact duration distribution according to whether it was a pre-pandemic or pandemic contact.

## Infections

The infection process for contact between index case  $i$  and their contact  $j$  was modelled as Bernoulli for both household and non-household infections

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

with the probability of  $i$  infecting  $j$  on day  $t$  in location  $X$ ,  $p_{X,ij,t}$ , where  $X \in \{HH, NHH\}$ , given by

$$p_{X,ij,t} = 1 - \exp(-\beta p_{\text{culture},i,t} \text{dur}_{X,ij,t})$$

where  $\beta$  is a scaling parameter which converts the product of the relative infectiousness of  $i$  on the day of contact,  $p_{\text{culture},i,t}$ , and the duration of contact,  $\text{dur}_{X,ij,t}$ , into a transmission rate. This process was iterated over each day of the index case's infection up to a maximum infection duration of 44 days in the baseline analysis. While we modelled infection events for household members on each day of the index case's infection, we only counted the first infection that occurred for each household member. In other words, household members could only be infected once. 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:

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

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. 95% percentile bootstrap confidence intervals (CIs) for the  $R$  and  $k$  estimates were calculated by bootstrapping the simulated numbers of secondary infections with 1000 bootstrap samples, calculating  $R$  and  $k$  for each bootstrap sample and then finding the 2.5 and 97.5 percentiles of the distributions of the bootstrap estimates. We calibrated  $\beta$  such that the basic reproduction number  $R_0$ , the reproduction number for the pre-pandemic time period, was 2.5, based on estimates from the literature [12,16,17], for the baseline analysis and all sensitivity analyses varying heterogeneity in contact rates and viral load. We obtained values of  $\beta$  between 1.1 and 4.7.

### Contribution of contact and viral load heterogeneity to heterogeneity in secondary infections

To investigate the contribution of heterogeneity in contact rates and heterogeneity in viral load between individuals to variation in the secondary infection distribution, we calculated the overdispersion in secondary infections for different combinations of “homogeneous” and “heterogeneous” contact rates and viral load trajectories. In the “homogeneous” cases, contacts were Poisson distributed with mean equal to the mean daily contacts for each time period and all individuals had the median viral load trajectory (i.e., the trajectory with the median proliferation and clearance durations and median peak Ct value), while in the “heterogeneous” cases, contacts and viral load trajectories were allowed to vary according to their full distributions.

We performed a sensitivity analysis with greater variation in viral load trajectories across individuals, by doubling the standard deviations of the distributions for the proliferation time, clearance time and peak Ct value, to test the sensitivity of the results to the degree of viral load heterogeneity. For this sensitivity analysis, we considered onward transmissions up to 88 days post exposure.

### Simulation of interventions

We assessed the hypothesis that if those with the highest viral loads are most likely to cause superspreading events, and LFTs are most sensitive for those with the highest viral loads, then lateral flow testing should be able to prevent superspreading events, which we defined as infecting over 10 contacts. We determined the probability of testing positive on an LFT at a given viral load,  $p_{\text{test}}$ , by fitting a logistic regression model to data on LFT outcomes for isolates with different viral loads [14]. Detection was then modelled via a Bernoulli draw with this detection probability, that is,

$$\text{test} \sim \text{Bernoulli}(p_{\text{test}})$$

$$\text{logit}(p_{\text{test}}) = \alpha_0 + \alpha_1 v_i$$

where  $\alpha_0 = -9.1$  (95% CI: -13.3, 5.0) and  $\alpha_1 = 1.8$  (95% CI: 1.1, 2.6) were the estimated coefficients from the logistic regression. We then modelled index cases testing regularly with LFTs at different frequencies (every 1, 3, and 7 days) or only before events of a certain size (before meeting >10, >20, >50 others) to determine the comparative effectiveness of rapid testing to control transmission. We assumed that a certain proportion of individuals self-isolated at home upon first testing positive, such that their non-household contacts were decreased to zero after the date of the positive test, but their number of household contacts was unchanged. The remaining individuals were assumed to continue making contacts as normal. We varied the proportion of individuals who self-isolated upon testing positive from 0-100% to model varying uptake/adherence. We assumed timing of testing was not affected by symptom onset and intensity, as we wished to model regular and event-triggered testing regardless of symptoms. We considered three indicative timepoints before (BBC Pandemic, 2018) and during the pandemic (1st lockdown (March-June 2020) and the reopening of schools (September 2020)). We reran the analysis for regular lateral flow testing with Poisson-distributed contacts instead of overdispersed contacts to look at the effect of contact heterogeneity on the impact of regular testing.

The code and data for this study can be found at [https://github.com/bquilty25/superspreading\\_testing](https://github.com/bquilty25/superspreading_testing).

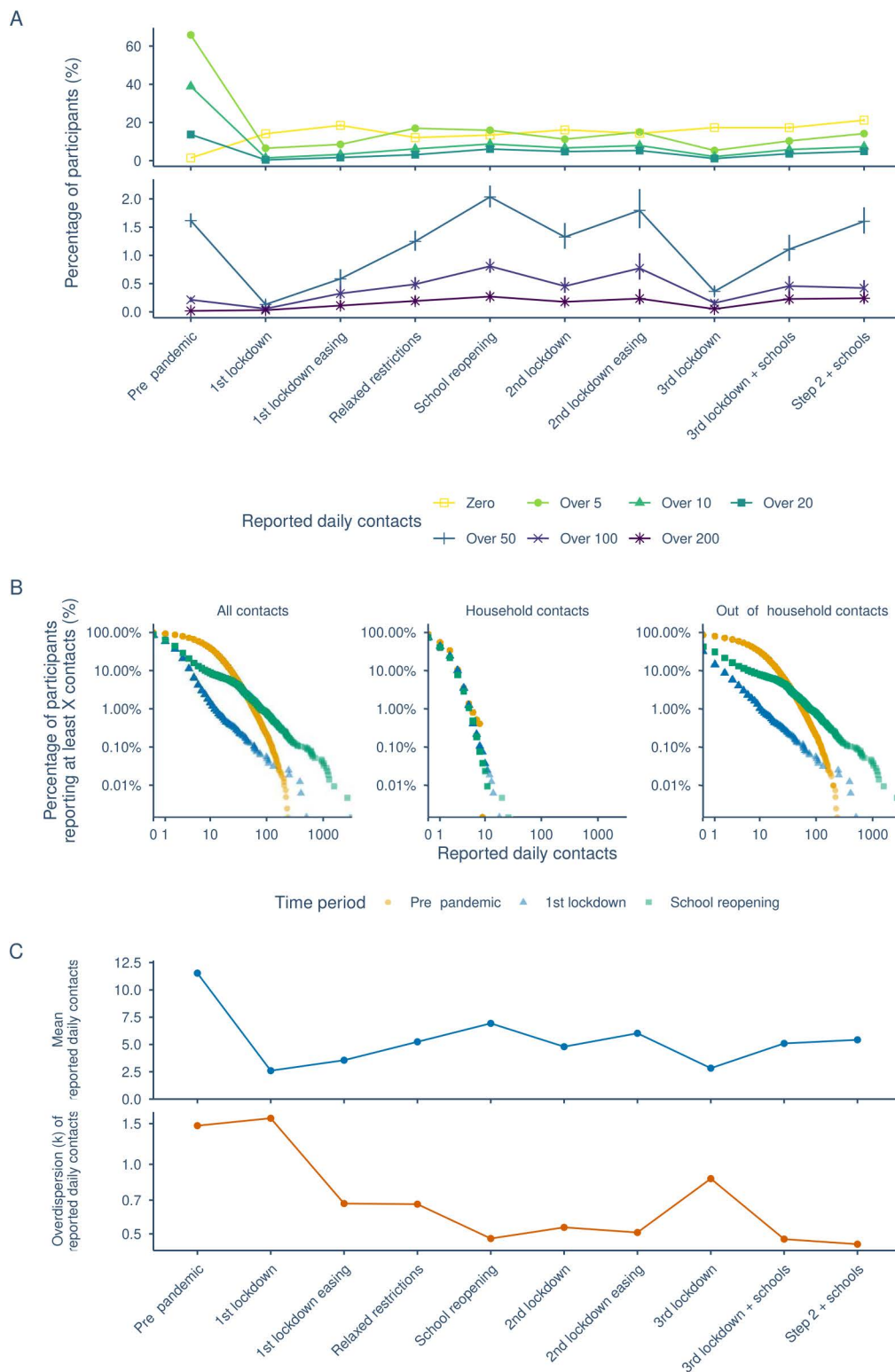
## Results

### Changes in the distribution of social contacts with changes in the intensity of restrictions

The age distribution of all participants in the BBC Pandemic survey and of the participants in each of the nine periods of the CoMix survey is shown in [S3 Fig](#). Relative to the mid-2020 age distribution of the UK population, children and older adults (70+ years) are under-represented in the BBC Pandemic data. The age distribution of CoMix participants is representative of the general population (with slight under-representation of children  $\leq 4$  years and over-representation of adults aged 50–69 years) and remains fairly stable across the different time periods.

[Fig 2A](#) and [Table 1](#) show the proportions of survey participants in the UK reporting more than a certain number of contacts (5, 10, 20, 50, 100 and 200) in the previous day over time, both pre-pandemic (BBC Pandemic survey) and from the CoMix survey for the nine time periods during the survey with different levels of restrictions. The percentage of individuals reporting more than 20 contacts in a day was substantially lower during the pandemic compared to the pre-pandemic period (when it was 13.7%), varying from a low of 0.4% during the first lockdown from March to June 2020 to a peak of 6.1% in September 2020, when restrictions were most relaxed and schools reopened. The percentage of individuals reporting over 100 and 200 contacts was lower in the pre-pandemic BBC Pandemic contact survey compared to periods of relaxed restrictions due to differences in the reporting of high contact events. Overall, people reporting more than 50, 100, or 200 contacts made up less than 3% of the total CoMix survey sample.

The estimated mean and overdispersion parameter  $k$  (with lower values indicating more overdispersion in the distribution) of numbers of daily contacts in the UK were lower than that observed pre-pandemic, with mean daily contacts averaging 4.7 during the pandemic compared to 11.5 pre-pandemic, and  $k$  averaging 0.7 during the pandemic compared to 1.5 pre-pandemic, indicating individuals having on average lower, but more varied, numbers of daily contacts. These values also changed considerably across the different periods of restrictions in a similar pattern to the proportion of participants with high numbers of contacts, with the mean number of daily contacts ranging from 2.6 during the first lockdown to 6.9 when schools reopened in September 2020, with a similar drop during the third lockdown, but less of a drop during the second lockdown when schools remained open ([Fig 2C](#)). The overdispersion in contacts,  $k$ , was similar pre-pandemic and during the first lockdown (1.5 and 1.6 respectively) then became lower, indicating more variability, following the easing of the first lockdown, with  $k$  between 0.5 and 0.7 in the remaining 2020 periods ([Fig 2C](#)). Stratifying contacts by



**Fig 2. Changes in the distribution of social contacts in the UK from pre-pandemic to May 2021.** A. Percentage of participants in the BBC Pandemic and CoMix contact surveys reporting zero, over 5, 10, 20, 50, 100, and 200 daily contacts in 2018 (pre-pandemic) and nine time periods during

the pandemic in the UK between March 2020 and May 2021. Median and 95% binomial confidence intervals shown. **B.** Distribution of the number of reported daily contacts for three indicative timepoints before and during the pandemic in the UK in 2020. **C.** Mean and overdispersion parameter,  $k$ , of a negative binomial distribution fitted to contacts during the different time periods. The heavier tail in the contact distribution for the School reopening period in panel B compared with the Pre-pandemic period is likely due to mass contacts not being recorded in the BBC Pandemic contact survey, which may have truncated the contact distribution. See [S4 Fig](#) for the BBC Pandemic contact distribution with an imputed heavier tail, which is used to perform a sensitivity analysis.

<https://doi.org/10.1371/journal.pcbi.1014715.g002>

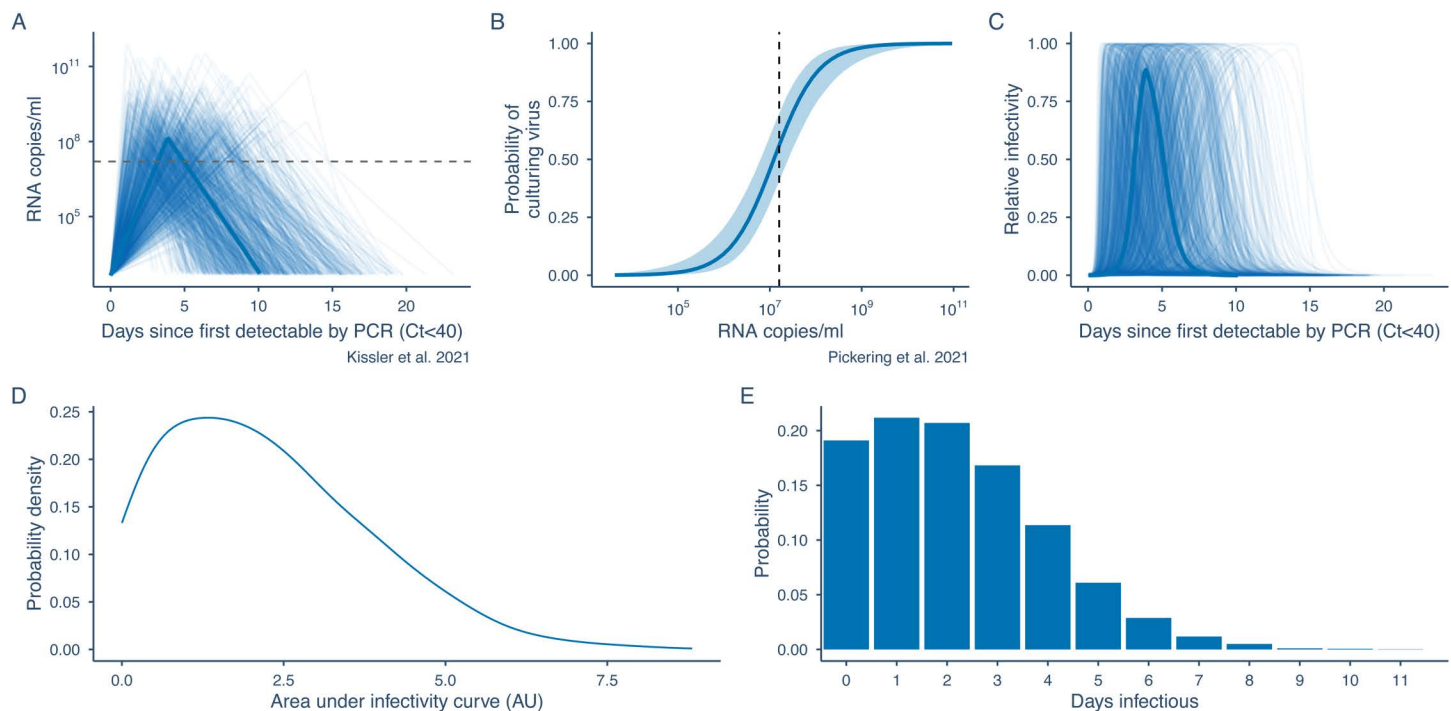
household and out-of-household revealed contact rates within the household remained stable throughout, with changing non-household contact rates driving much of the variation in the mean and  $k$  of daily contacts over the course of the pandemic ([S5 Fig](#)). Both the mean and  $k$  of out-of-household daily contacts remained lower than pre-pandemic contact rates. Contact durations in CoMix differed significantly between household and non-household contacts, with a median duration of 30 minutes (IQR: 5 – 180 minutes) for out-of-household contacts compared to 480 minutes (8 hours) (IQR: 180 – 1080 minutes) for household contacts ([S1 Fig](#)). [S6 Fig](#), in which contacts are stratified by age as well household/out-of-household, shows that both household and out-of-household contacts in children (<18 years) remained higher than those in adults throughout 2020 and early 2021, with significant increases in out-of-household contacts of children during periods when schools were open. Contact levels were similar for males and females across all time periods.

### Variation in viral load and infectivity over the course of infection

Analysis by Kissler et al. [[13](#)] of densely sampled viral load trajectories in individuals infected with SARS-CoV-2 suggested substantial variation between individuals in the duration of proliferation and clearance phases of infection, and lowest Ct value (inversely correlated with peak viral load) ([Fig 3A](#)). Mapping viral load to infectivity via a logistic function based on viral load and culture from Pickering et al. [[14](#)] ([Fig 3B](#)), calculating the probability of infectivity by day ([Fig 3C](#)), then integrating under the infectivity curve to calculate total infectious virus shed ([Fig 3D](#)), reproduces substantial heterogeneity in individual infectiousness as reported by Ke et al. [[6](#)]. We found a 73-fold difference in individual-level infectivity between the 2.5% and 97.5% percentiles of the distribution of the area under the infectivity curve (0.08 and 5.74 arbitrary units, respectively). Fitting a Gamma distribution to this distribution gave a shape parameter of 1.42 (1.6 for Ke et al.). A substantial proportion (19.4%) of individuals were estimated to be infectious for zero days, but the remainder were infectious for at least one day ([Fig 3E](#)). Across all individuals, the median duration of infectiousness around peak viral load, that is, having high enough viral load to cause infections through culturable virus, was 2 days (95% prediction interval: 0, 6 days). The majority of individuals passed through at least one day of high infectiousness (peak culture probability greater than 0.6 for 68% of individuals, greater than 0.8 for 53%), and were therefore theoretically capable of causing superspreading events given sufficient contacts of sufficient duration. However, the percentage of infected individuals who can actually initiate a superspreading event is much lower than 50% because the time window for this level of infectivity is so narrow that it must coincide perfectly with a high number of contacts of sufficient duration for a high number of infections to occur.

### Predicting transmission dynamics from heterogeneity in contact rates and viral load progression

Simulating secondary infections using the stochastic model described above, we estimated the overdispersion,  $k$ , in the initial secondary infection distribution as 0.48 (95% CI 0.47 – 0.51) ([Fig 4](#)), in line with other estimates [[18,19](#)]. During the first national lockdown,  $R$  was estimated at 0.37 (95% CI 0.35 – 0.41) and  $k$  at 0.25 (95% CI 0.21 – 0.29).  $R$  rose as lockdown eased ( $R=0.66$ , 95% CI 0.56 – 0.8) and restrictions were relaxed in the summer of 2020 ( $R=1.0$ , 95% CI 0.92 – 1.1) and again as schools reopened in the autumn ( $R=1.5$ , 95% CI 1.3 – 1.6).  $R$  then decreased again during the second lockdown to 1.0 (95% CI 0.92 – 1.2). As with contacts,  $k$  reduced as the pandemic began, indicating more variability in the secondary infection distribution, and varied between 0.12 and 0.17 after the first lockdown up to the end of 2020. A sensitivity analysis imputing a heavier tail for the BBC Pandemic contact distribution based on the tails of contact distributions



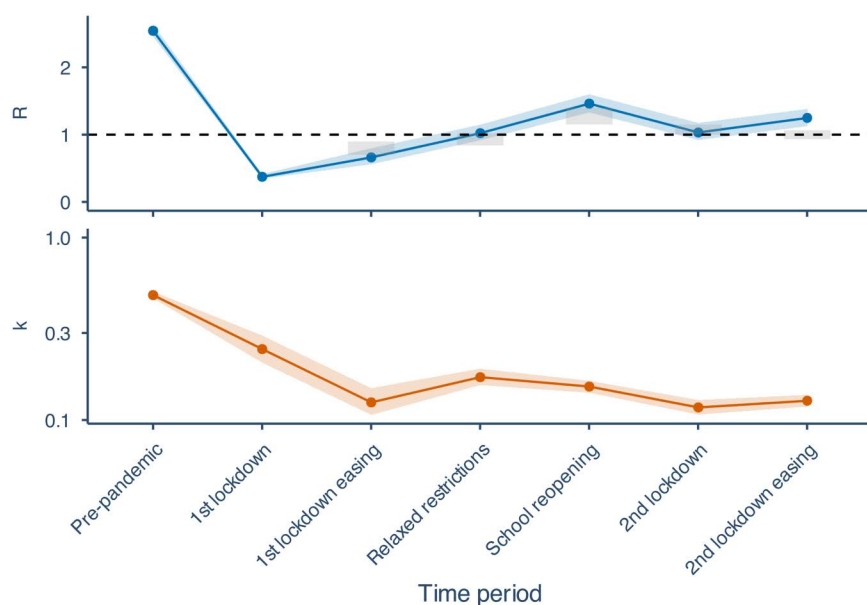
**Fig 3. Variation in viral load progression and infectivity.** **A.** Individual-level viral load trajectories. **B.** Logistic model for probability of culturing virus given a certain viral load. **C.** Relative infectivity over time. **D.** Area under the infectivity curve (in arbitrary units, AU). **E.** Distribution of the number of days for which individuals were infectious. Solid lines in A and C represent median viral load and infectivity trajectories. The dashed line in A and B represents the RNA copies per ml required for a 50% probability of culturing virus or infectiousness. The shaded band in B represents the uncertainty in the probability of culturing virus given a certain viral load from the fitted logistic regression model from Pickering et al. [14]. S7 Fig shows the difference in viral load trajectories and infectivity for the sensitivity analysis with double the standard deviation in the viral load trajectory parameters.

<https://doi.org/10.1371/journal.pcbi.1014715.g003>

from relaxed restriction periods during the CoMix survey led to a small increase in  $R_0$  from 2.5 (95% CI 2.4 – 2.6) to 2.8 (95% CI 2.6 – 3.0) and a small reduction in  $k$  from 0.48 (95% CI 0.46 – 0.51) to 0.44 (95% CI 0.40 – 0.49) (S1 Table).

### Contribution to heterogeneity in secondary infections due to variation in contact rates and viral load progression

If both contact rates and viral load trajectories are treated as homogeneous, with Poisson-distributed contacts and no variation in viral load trajectory between individuals, then the secondary infection distribution is approximately Poisson with equal mean and variance, corresponding to very large values of  $k$ . If contacts are Poisson distributed but viral load trajectories are allowed to vary between individuals,  $k$  is 1.5 – 1.9. If viral load is fixed at the median trajectory, but contacts are overdispersed, values of  $k$  are lower (0.11 – 0.74) and approximately equal to that of overdispersed contacts and variable viral load trajectories (0.11 – 0.51), indicating that overdispersion in numbers of daily contacts, as the denominator in the infection process, contributes more to superspreading (overdispersion in the secondary infection distribution) than variable viral load trajectories (Fig 5). The two overdispersed-contact scenarios (equal and variable viral load) differ only pre-pandemic ( $k=0.74$  versus 0.51). This is due to individuals in the upper tail of infectiousness generating disproportionately many secondary infections when contacts are plentiful, as a result of the right-skewed area-under-the-curve distribution (>73-fold range) (Fig 3D). During restriction periods this difference disappears, as suppressed contact rates create a ceiling on secondary cases regardless of viral load, and the proportion infecting zero others and more than ten others converges across the two scenarios from the first lockdown onwards (Fig 5). In the sensitivity analysis with greater variation in individuals' viral load trajectories, with double the standard deviation in the viral load trajectory parameters, the



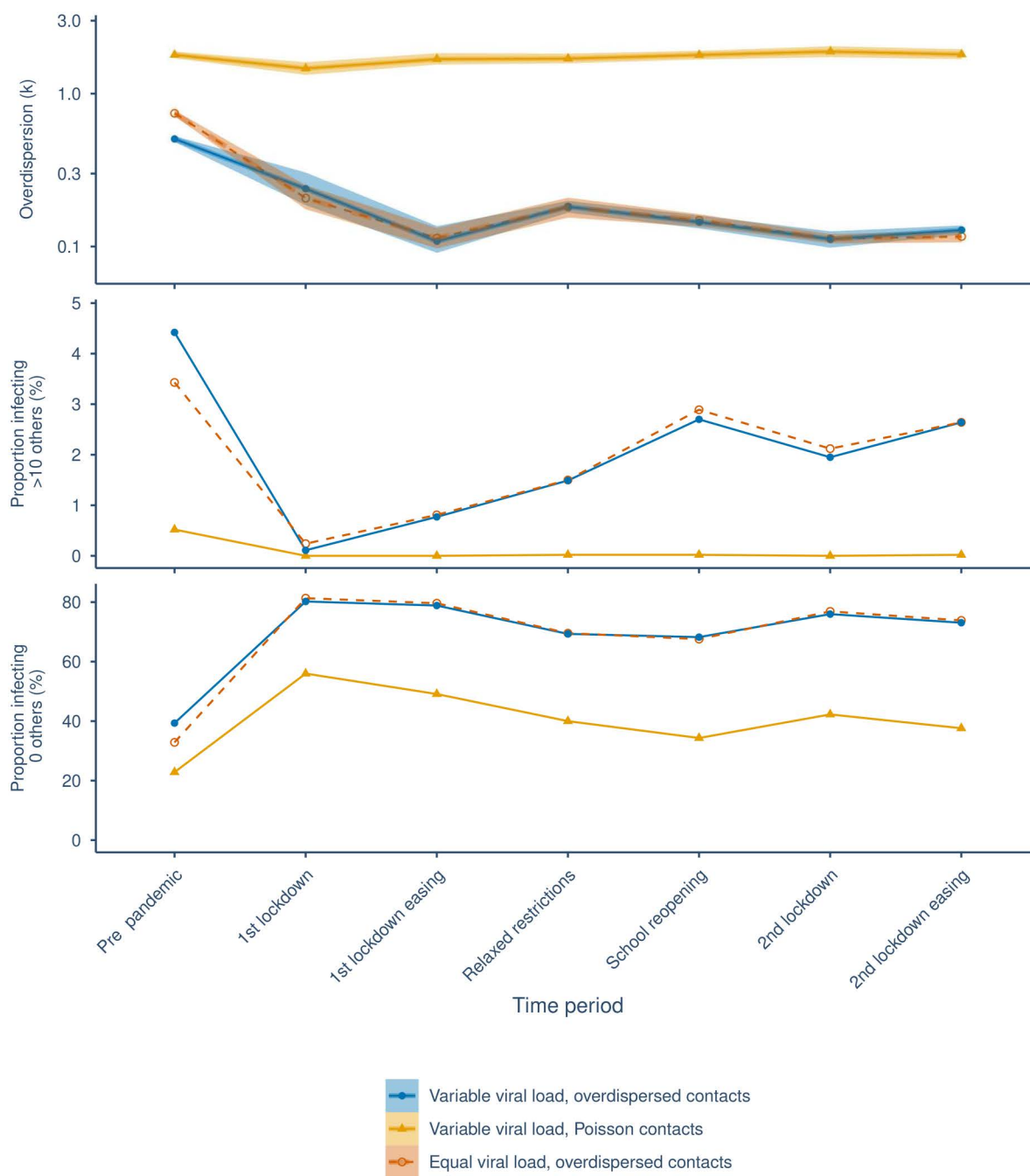
**Fig 4. Estimates of the mean and dispersion of negative binomial distributions fitted to simulated secondary infection distributions by time period in 2020.** The dispersion parameter  $k$  gives an indication of the variation in the secondary infection distribution, with smaller values corresponding to greater variation and values less than 1 corresponding to very large variation. Grey boxes in the  $R$  plot show the average upper and lower bounds (90% confidence interval) of the consensus estimates published by the Scientific Pandemic Influenza group on Modelling in the UK for the specified time periods [17]. Shaded bands show 95% bootstrap confidence intervals for the  $R$  and  $k$  estimates.

<https://doi.org/10.1371/journal.pcbi.1014715.g004>

$k$  estimates for the scenario with Poisson contacts (0.76 – 0.92) and variable viral loads were closer to those for overdispersed contacts and variable viral loads (0.08 – 0.33), but still higher (S8 Fig). The finding that contact heterogeneity plays a greater role in overdispersion in secondary infections than viral load heterogeneity therefore appears to be robust to much greater variation in peak viral load and duration of viral shedding than found in the cohort in Kissler et al. [13].

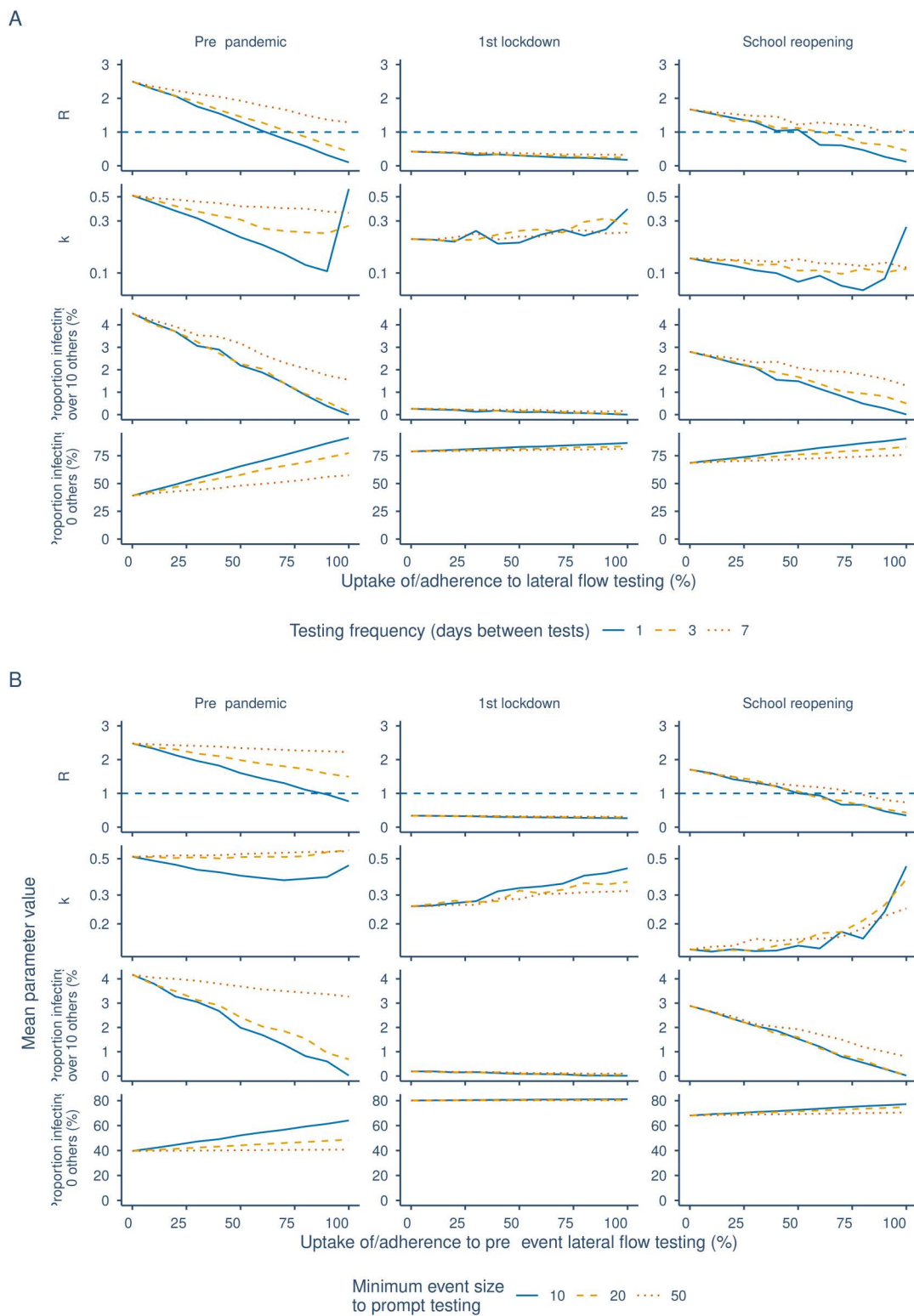
### Impact of rapid testing (regular testing vs. pre-event testing)

For pre-pandemic levels of contacts, uptake of regular LFT testing must exceed 70% to reduce  $R$  below 1 if testing daily and must be above 80% if testing every 3 days, but even 100% uptake will not reduce  $R$  below 1 if testing weekly (Fig 6A). For lockdown levels of contacts, testing does not appreciably reduce  $R$  below its already low level. For contact rates equivalent to those under relaxed restrictions with schools open, testing every 3 days will reduce  $R$  below 1 if uptake exceeds 70% (Fig 6A). Pre-event testing acts similarly, with high uptake (> 90%) and testing for events with more than 10 others necessary to reduce  $R$  below 1 for pre-pandemic levels of contact, but only testing when attending events with more than 20 others and 60% adherence required during relaxed restrictions with schools open (Fig 6B). Both regular testing and pre-event testing reduce the rate of superspreading as defined as the proportion of individuals infecting over 10 others. However, there is also an increase in the proportion of individuals that infect no one. In some cases, this results in a decrease in  $k$  (greater overdispersion in the secondary infection distribution) (Fig 6). From the sensitivity analysis comparing the impact of regular testing with and without overdispersion in contacts (S9 Fig), we can see that ignoring overdispersion in contacts reduces the estimated impact of testing in terms of reducing  $R$  and decreasing the proportion of superspreading events. This demonstrates the importance of accounting for contact overdispersion when estimating the impact of interventions on secondary infections.



**Fig 5. Estimates of the dispersion of negative binomial distributions fitted to secondary infection distributions by period, with different combinations of homogeneous and heterogeneous contact rates and viral load trajectories, along with the proportion of individuals infecting zero others and more than ten others.** The dispersion parameter  $k$  gives an indication of the variation in numbers of secondary infections, with smaller values corresponding to greater variation and values less than 1 corresponding to very large variation. Values of  $k$  for equal viral load and equal contacts are not shown because they are very large, meaning the numbers of secondary infections are approximately Poisson distributed. Shaded bands show 95% bootstrap confidence intervals. The sensitivity analysis with greater variation in viral loads is shown in [S8 Fig](#).

<https://doi.org/10.1371/journal.pcbi.1014715.g005>



**Fig 6. Effect of regular and pre-event lateral flow testing on  $R$ ,  $k$ , and the proportion infecting over 10 or 0 others for varying levels of uptake or adherence (horizontal axes) and background contact rates (columns). A. Regular testing. B. Pre-event testing.**

<https://doi.org/10.1371/journal.pcbi.1014715.g006>

## Discussion

We have estimated the secondary infection distribution for SARS-CoV-2 over the course of the pandemic in the UK in 2020 using a stochastic model incorporating viral load trajectories and reported daily contacts. Our estimates of the mean number of secondary infections are consistent with contemporaneous estimates (Fig 4). Daily contact rates became both lower on average, from 11.5 per day to 4.7 per day, and more overdispersed during the pandemic in the UK in 2020. While certain groups, such as essential workers, maintained high contact levels, the "stay at home" guidance and the shift towards remote working reduced daily contacts for many others, often to zero. This drove a decrease in the mean number of secondary infections per infected individual, the reproduction number, but an increase in the dispersion in the number of secondary infections. Our estimates of the degree of overdispersion in the secondary infection distribution are similar to those reported by other authors, which are mostly between 0.3 and 0.6 [18,19], though lower estimates of around 0.1 have been reported from outbreak cluster size data [1].

Our analysis shows that changes in the mean and overdispersion of daily numbers of contacts over time (Fig 2C) can be used to estimate dynamic shifts in the mean and overdispersion of the reproduction number (Fig 4). This motivates collecting contact data during epidemics, as contact overdispersion could provide a potential indicator of superspreading risk in real-time, along with the effective reproduction number. Changes in contact overdispersion would indicate when and where the population is at risk of superspreading events, and enable implementation of targeted, rather than blanket, restrictions on contact levels to limit outbreaks. However, ground-truth data on secondary infection distributions, e.g., from contact tracing studies, would first be required to validate estimates of the mean and overdispersion of the secondary infection distribution.

Contact heterogeneity was found to contribute more to superspreading than heterogeneity in viral load, as the number of contacts an individual makes places an upper limit on the number of people they can infect, and, despite some individuals having much higher infectious potential than others, most infected individuals pass through a high viral load period, with 68% having a peak culture probability greater than 0.6 and 53% greater than 0.8. Viral load heterogeneity does still contribute significantly to heterogeneity in transmission since viral load varies by orders of magnitude over time within individuals and between individuals (Fig 5). This indicates that superspreading is more a result of between-person contact variability and within-person infectiousness variability than between-person infectiousness variability. Hence if reductions in contact rates can be targeted specifically at individuals ahead of or during their window of high infectivity, for example by encouraging contacts of cases to undergo regular self-testing and self-isolation upon a positive rapid test result [20], then this may result in reductions in transmission while minimising the burden of quarantine. Other studies investigating the superspreading nature of SARS-CoV-2 have come to similar conclusions [3,4,6,21–23]. However, as far as we are aware our study is the first to estimate the relative roles of contact and viral load heterogeneity and the impact of changes in contact heterogeneity using real-world contact distributions from before and during the pandemic and data on viral load trajectories and infectiousness [14]. Our use of reported heterogeneity in viral load trajectories [13] also contributes to a wider estimated distribution for the number of days individuals are likely infectious, which closely matches empirical daily sampling [6] and human challenge studies [5]. Our approach of using culture positivity as a direct virological proxy for infectiousness is complementary to that of Marc et al. [24], who inferred the viral-load-infectiousness relationship from observed transmission outcomes and similarly found that viral load alone does not fully explain transmission heterogeneity, with a residual random effect standard deviation of 85%. Consistent with our findings, Kendall et al. [25] found that contact rates explained more daily variability in  $R$  than per-contact transmissibility (46% versus 30%) using NHS COVID-19 app data that tracked CoMix closely; Ferretti et al. [26] similarly found that contact duration drove transmission risk more than proximity, supporting the use of duration-weighted contacts in our model.

Like others [6,23,27], we hypothesised that lateral flow testing, by detecting individuals with high viral loads when they were most infectious, would reduce transmission through reducing the potential for superspreading. This manifested as a decrease in the proportion infecting over 10 others and a substantial increase in the proportion infecting zero others. This,

perhaps counter-intuitively, resulted in a decrease in  $k$  in some cases, as the relative increase in those infecting zero others exceeded the decrease in those infecting many others, here defined as over 10 others. Hence, assessment of super-spreading solely via the metric of the overdispersion parameter  $k$  may conceal changes in both the upper and lower tail of the secondary infection distribution. Both regular testing and pre-event testing were effective in reducing  $R$  given high enough frequency or a low enough event size threshold, respectively, as long as uptake or adherence was high. Testing had the highest relative impact on transmission when contact rates were high, for example at pre-pandemic levels, as there were more potentially preventable exposures, meaning rapid testing could reduce  $R$  below the growth threshold of 1 while otherwise maintaining relatively normal contact rates. In contrast, testing during lockdown would have less impact as  $R$  was already below 1. Having everyone in the population test every 3 days would bring  $R$  below 1 and be approximately equivalent in terms of impact on secondary infections to having people test only before events of minimum size 10 for pre-pandemic contact levels. This indicates that rapid testing could be an effective, minimally disruptive intervention to reduce transmission if uptake or adherence could be maximised through incentivising use. Our findings align with those of Ke et al. [6], who estimated using a within-host model of infection that daily antigen testing could reduce infectiousness by 85% but that impact reduced significantly with decreasing testing frequency. However, our analysis goes further by estimating the impact of different rapid testing frequencies on the reproduction number accounting for the full distribution of daily contacts. Hart et al. [23] similarly found that daily lateral flow testing could significantly reduce numbers of secondary infections, preventing around 70% of infections assuming complete isolation of infected individuals following detection, although they used a full multiscale model of within-host dynamics and transmission and focused on the Omicron variant, for which the effectiveness of testing may be different due to differences in within-host dynamics and pre-existing immunity.

In their analysis of the CoMix data, Gimma et al. [11] found that school-aged children consistently had higher numbers of contacts across both lockdowns and periods of relaxed restrictions than working-age adults, who in turn had higher numbers of contacts than older adults (60+ years). Contact levels also varied with employment status, with part-time-employed participants having more contacts than full-time or self-employed participants. Another analysis of the CoMix data that focused on individuals with high numbers of contacts [28] identified that most reports of high numbers of contacts came from work and school settings and tended to be reported by participants in certain professions (teaching, retail, administration, construction, medicine, social work, hospitality). Some of this work is considered essential, and keeping schools open where possible is a high priority, so opportunities for reducing these contacts may be limited. This highlights the potential value of regular rapid testing as a way of preventing superspreading even when some individuals have many contacts.

Our analysis has some limitations. The CoMix contact survey was designed to be comparable to previous contact surveys in the UK, namely the BBC Pandemic [7–9] and POLYMOD [10] surveys. However, previous surveys required participants to list contacts individually to include other information such as contact age, sex, and occupation, meaning that it was difficult to include mass contacts such as those made at a large gathering, thus truncating the true contact distribution (Fig 2B). From 18 May 2020 the CoMix survey introduced the option to record mass contacts as a count rather than listing each contactee in a separate entry. This means pre-pandemic and early-pandemic contact distributions are not directly comparable to those conducted later. However, we conducted a sensitivity analysis to check how this could impact our estimates of numbers of secondary infections, by imputing a heavier tail for the BBC Pandemic contact distribution based on the tails of the contact distributions during periods of relaxed restrictions in the CoMix survey, and observed only a small increase in  $R_0$  and a small reduction in  $k$ . Additionally, variation between participants in how direct contacts are interpreted and reported could contribute to artificial overdispersion in the contact distribution and hence in our estimates of secondary infection overdispersion. Another limitation of the BBC Pandemic contact data is that working-age adults were overrepresented among the participants, while children under 13 years were excluded (S3 Fig). It is possible that this may have led to some bias in the reported contact distributions. However, the similarity of contact patterns observed

in the BBC Pandemic and UK POLYMOD datasets [8] suggests that the BBC Pandemic data is broadly representative of pre-pandemic contact patterns in the whole UK population.

We used data on viral loads over time from a detailed study of individuals in the National Basketball Association cohort in Kissler et al. [13], which is unlikely to be representative of the full variation in viral load trajectories in the general population, since the cohort consisted mostly of young, healthy, male professional athletes. Our analysis could therefore underestimate the variability in secondary infections due to variability in viral load with factors such as age, for which there is some evidence of an effect on infectiousness [6], co-morbidities, and sex. In particular, the NBA cohort does not include individuals with very long infections. Ghafari et al. [29], for instance, found that some individuals in the Office for National Statistics COVID-19 Infection Survey in the UK had persistent infections lasting more than 60 days. However, they constituted only 0.1 to 0.5% of all infections, with PCR-derived viral loads around 100-fold lower than at symptom onset, substantially reducing their transmission potential. Furthermore, PCR measurements were only taken from ONS infection survey participants weekly for a month and then monthly up to a year, whereas measurements were near daily in the NBA cohort and crucially captured the early stage of infection, and therefore permit much higher resolution modelling of onward transmission. We conducted a sensitivity analysis with much greater variation in viral load trajectories to account for greater variation in the general population and found that it did not qualitatively change our conclusion that contact heterogeneity remains the primary driver of heterogeneity in transmission.

We assume that the probability of shedding infectious virus is equal to the probability of culturing virus, but the exact relationship between viral culture positivity and infectiousness is not known [30], so our results must be interpreted with this limitation in mind. Furthermore, the specific proportions of infectious individuals we report depend on the fitted logistic dose-response curve between viral load and culture probability; alternative functional forms would yield different estimates. A structural asymmetry in our comparison is that contact heterogeneity is measured directly from survey data, whereas infectiousness heterogeneity is captured only via viral load as a proxy. Sources of between-individual variation in infectiousness that are decoupled from within-host viral dynamics, such as variation in respiratory droplet production, speaking volume, or interpersonal distance preferences, are not captured, and if substantial would increase the true contribution of infectiousness heterogeneity relative to our estimates. We partially address this by drawing the logistic regression coefficients linking viral load to infectiousness for each individual from the estimated covariance distribution, and by examining robustness in a sensitivity analysis with twice the standard deviation in viral load trajectory parameters (S8 Fig); however, these measures do not capture variation in transmission efficiency that is entirely independent of viral load. We also do not account for correlation of the timing of peak viral load and infectiousness with symptom onset and intensity [31], which may affect contact rates and numbers of secondary infections [24].

We focus on an index case and their infections in one other generation, which may underestimate second-order effects that considering a full contact network structure would capture [6]. For instance, we would expect individuals that acquire infection to typically have higher contact rates than the average member of the population, since having more contacts increases acquisition risk, and therefore to infect more individuals. However, incorporating this effect would require additional assumptions about how acquisition risk scales with contact number and network structure that we do not have the data to support. We do not consider other interventions that may have an additional impact on  $R$ , such as vaccination, contact tracing, or self-isolation upon symptom onset. We also do not consider the impact of variants with increased transmissibility, and hence limit our analysis to 2020 before the widespread emergence of variants of concern such as Alpha, Delta, and Omicron.

We assumed within-household contacts for each individual were the same each day, and sampled daily non-household contacts from repeat responses for the same survey participant, which may underestimate variation in overall daily contact rates and hence infection potential of specific individuals. We assume that self-isolating individuals are unable to fully self-isolate from their household members as reported by the majority of those surveyed by the ONS in England in April 2021 [32]; further decreases in  $R$  may be possible if self-isolating individuals isolate themselves from household members.

There is some evidence that heterogeneity in secondary infections has changed as new SARS-CoV-2 variants of concern have emerged and population immunity has increased. Studies examining heterogeneity in SARS-CoV-2 transmission over time in Japan and Europe have found that overdispersion remained after the emergence of Alpha, Delta, and Omicron, with  $k$  generally being below 0.5 but varying somewhat across waves and being higher, that is, less overdispersed transmission, during the Omicron wave [33,34]. Meta-analyses of literature estimates of  $k$  have shown a wide range of estimates across settings, clusters and variants, but no clear trend in  $k$  over time [35,36], though some studies suggested slight increases in  $k$  with the emergence of Delta. A modelling study has suggested that non-pharmaceutical interventions such as lockdowns have imposed selective pressure on SARS-CoV-2 towards more homogeneous transmission, corresponding to higher  $k$  [37]. Together, these studies suggest that while overdispersion may be weakening with widespread immunity and new variants of concern, it remains a feature of SARS-CoV-2 transmission.

Thus, while our analysis is restricted to 2020, our findings are still relevant in the current era of endemic COVID-19 and to other pathogens that exhibit heterogeneous transmission, including other SARS coronaviruses, Ebola, measles, influenza, RSV, and adenoviruses [3,38,39]. Moreover, the question of the relative contribution of variation in numbers of contacts and variation in infectiousness to heterogeneity in transmission is a universal one for infectious diseases. By demonstrating that superspreading potential is closely linked to the overdispersion of social contacts and the timing of peak infectiousness, we have shown the importance of monitoring not only the mean effective reproduction number but also its variability. Measurement of contact distributions thus offers a real-time surveillance tool for identifying superspreading risk, supporting pandemic preparedness through enabling targeted interventions against emerging pathogens. The modelling framework is generalisable: provided densely sampled pathogen load trajectories and pathogen-specific culture or test sensitivity data are available to characterise infectiousness, and a contact survey with duration information is available to characterise the contact distribution, the same approach can be applied to estimate heterogeneity in the secondary infection distribution for other pathogens.

Our results suggest superspreading for SARS-CoV-2 can be best explained as a random sample from the tail of the contact and shedding distribution: it occurs when an infected individual makes a high number of contacts during a highly infectious period lasting approximately 2 days on average, with the majority of infected individuals being sufficiently infectious (68% with a peak culture probability over 0.6 and 53% over 0.8) for at least one day to be capable of causing a superspreading event, given they make a high number of contacts. Changes in the number of contacts observed throughout the pandemic in the CoMix contact survey were able to explain changes in the reproduction number, with contact rates becoming more heterogeneous during the pandemic given lockdowns and changes in working practices. For future pandemics, regular or pre-event lateral flow testing may be an efficient way to target individuals when most infectious and hence minimise the burden of non-pharmaceutical interventions while maximising reduction in transmission rates, provided uptake is moderate to high.

## Supporting information

**S1 Fig. Distribution of the duration of contact for household and out-of-household contacts from the CoMix contact survey in the UK.**

(PDF)

**S2 Fig. Correlation between number of contacts and mean contact duration for CoMix survey participants from March 2020 to May 2021.**

(PDF)

**S3 Fig. Age distribution of participants in the BBC Pandemic (Pre-pandemic) contact survey and each of the nine time periods from the CoMix contact survey from March 2020 to May 2021 relative to the mid-2020 UK age distribution [40].**

(PDF)

**S4 Fig. Contact distributions for three indicative timepoints before and during the pandemic in the UK with an imputed heavier tail in the Pre-pandemic (BBC Pandemic) contact distribution (cf. Fig 2B).** The heavier tail was imputed from the tails of contact distributions during periods of relaxed restrictions in the CoMix survey. We conducted a sensitivity analysis with this distribution to determine the effect of the heavier tail on  $R_0$  and  $k$  (see S1 Table).

(PDF)

**S5 Fig. Distribution of the number of reported daily contacts for all time periods before (BBC Pandemic contact survey) and during (CoMix contact survey) the pandemic in the UK between March 2020 and May 2021.**

(PDF)

**S6 Fig. Contact distributions by age and gender across the nine time periods from the CoMix contact survey from March 2020 to May 2021.**

(PDF)

**S7 Fig. Comparison of viral load trajectories and infectivity for the sensitivity analysis with double the standard deviation in viral load trajectory parameters and those used in the main analysis.** A. Individual-level viral load trajectories. B. Logistic model for probability of culturing virus given a certain viral load. C. Relative infectivity over time. D. Area under the infectivity curve, in arbitrary units. E. Distribution of the number of days for which individuals were infectious. Solid lines in A and C represent median viral load and infectivity trajectories. The dashed line in A and B represents the RNA copies per ml required for a 50% probability of culturing virus or infectiousness. The shaded band in B represents the uncertainty in the probability of culturing virus given a certain viral load from the fitted logistic regression model from Pickering et al. [14].

(PDF)

**S8 Fig. Sensitivity of estimates of dispersion of the secondary infection distribution in different time periods to the degree of variation in viral load trajectories.** Left: standard deviation in distributions for viral load trajectory parameters from Kissler et al. [13]. Right: standard deviation of viral load parameters doubled.

(PDF)

**S9 Fig. Comparison of the impact of regular lateral flow testing on  $R$ ,  $k$ , and the proportion infecting over 10 or 0 others with Poisson-distributed contacts and overdispersed contacts, for varying levels of uptake or adherence and background contact rates.**

(PDF)

**S1 Table. Sensitivity analysis of the effect on  $R$  and  $k$  when 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.**

(CSV)

**S2 Table. Sensitivity analysis of the effect on the mean and dispersion in contact rates comparing a negative binomial distribution to a zero-inflated negative binomial distribution.**

(CSV)

## Author contributions

**Conceptualization:** Billy J Quilty, Lloyd A. C. Chapman, Christopher I Jarvis, Adam J Kucharski.

**Data curation:** Billy J Quilty, Lloyd A. C. Chapman, James D Munday, Kerry L M Wong, Amy Gimma, Christopher I Jarvis.

**Formal analysis:** Billy J Quilty, Lloyd A. C. Chapman, Christopher I Jarvis.

**Funding acquisition:** W. John Edmunds, Christopher I Jarvis, Adam J Kucharski.

**Investigation:** Billy J Quilty, Lloyd A. C. Chapman.

**Methodology:** Billy J Quilty, Lloyd A. C. Chapman, Christopher I Jarvis, Adam J Kucharski.

**Project administration:** W. John Edmunds, Christopher I Jarvis, Adam J Kucharski.

**Resources:** Suzanne Pickering, Stuart J D Neil, Rui Pedro Galão.

**Software:** Billy J Quilty, Lloyd A. C. Chapman.

**Supervision:** Lloyd A. C. Chapman, Christopher I Jarvis, Adam J Kucharski.

**Validation:** Billy J Quilty, Lloyd A. C. Chapman.

**Visualization:** Billy J Quilty, Lloyd A. C. Chapman, James D Munday.

**Writing – original draft:** Billy J Quilty, Lloyd A. C. Chapman, Adam J Kucharski.

**Writing – review & editing:** Billy J Quilty, Lloyd A. C. Chapman, James D Munday, Kerry L M Wong, Amy Gimma, Suzanne Pickering, Stuart J D Neil, Rui Pedro Galão, W. John Edmunds, Christopher I Jarvis, Adam J Kucharski.

## References

- Endo A, Abbott S, Kucharski AJ, Funk S, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* 2020;5:67. <https://doi.org/10.12688/wellcome-openres.15842.3> PMID: 32685698
- Chen PZ, Bobrovitz N, Premji Z, Koopmans M, Fisman DN, Gu FX. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. *eLife.* 2021;10:e65774. <https://doi.org/10.7554/eLife.65774> PMID: 33861198
- Goyal A, Reeves DB, Cardozo-Ojeda EF, Schiffer JT, Mayer BT. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. *eLife.* 2021;10:e63537. <https://doi.org/10.7554/eLife.63537> PMID: 33620317
- Susswein Z, Bansal S. Characterizing superspreading of SARS-CoV-2: from mechanism to measurement. *medRxiv.* 2020. <https://doi.org/10.1101/2020.12.08.20246082>
- Killingley B, Mann AJ, Kalinova M, Boyers A, Goonawardane N, Zhou J, et al. Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nat Med.* 2022;28(5):1031–41. <https://doi.org/10.1038/s41591-022-01780-9> PMID: 35361992
- Ke R, Martinez PP, Smith RL, Gibson LL, Mirza A, Conte M, et al. Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity in infectiousness. *Nat Microbiol.* 2022;7(5):640–52. <https://doi.org/10.1038/s41564-022-01105-z> PMID: 35484231
- Klepac P, Kissler S, Gog J. Contagion! The BBC Four Pandemic - The model behind the documentary. *Epidemics.* 2018;24:49–59. <https://doi.org/10.1016/j.epidem.2018.03.003> PMID: 29576516
- Klepac P, et al. Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project. *medRxiv.* 2020. <https://doi.org/10.1101/2020.02.16.20023754>
- Kucharski AJ, Klepac P, Conlan AJK, Kissler SM, Tang ML, Fry H, et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *Lancet Infect Dis.* 2020;20(10):1151–60. [https://doi.org/10.1016/S1473-3099\(20\)30457-6](https://doi.org/10.1016/S1473-3099(20)30457-6) PMID: 32559451
- Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* 2008;5(3):e74. <https://doi.org/10.1371/journal.pmed.0050074> PMID: 18366252
- Gimma A, et al. Changes in social contacts in England during the COVID-19 pandemic between March 2020 and March 2021 as measured by the CoMix survey: a repeated cross-sectional study. *PLOS Medicine.* 2022;19:e1003907. <https://doi.org/10.1371/journal.pmed.1003907>
- Jarvis CI, Van Zandvoort K, Gimma A, Prem K, CMMID COVID-19 working group, Klepac P, et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med.* 2020;18(1):124. <https://doi.org/10.1186/s12916-020-01597-8> PMID: 32375776
- Kissler SM, Fauver JR, Mack C, Olesen SW, Tai C, Shiue KY, et al. Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLoS Biol.* 2021;19(7):e3001333. <https://doi.org/10.1371/journal.pbio.3001333> PMID: 34252080
- Pickering S, et al. Comparative performance of SARS-CoV-2 lateral flow antigen tests and association with detection of infectious virus in clinical specimens. *Lancet Microbe.* 2021;2:e461–71. [https://doi.org/10.1016/S2666-5247\(21\)00143-9](https://doi.org/10.1016/S2666-5247(21)00143-9)
- Drain PK, Dalmat RR, Hao L, Berner MJ, Budiawan E, Morton JF, et al. Duration of viral infectiousness and correlation with symptoms and diagnostic testing in non-hospitalized adults during acute SARS-CoV-2 infection: A longitudinal cohort study. *J Clin Virol.* 2023;161:105420. <https://doi.org/10.1016/j.jcv.2023.105420> PMID: 36913789
- Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *Lancet Public Health.* 2020;5(7):e375–85. [https://doi.org/10.1016/S2468-2667\(20\)30133-X](https://doi.org/10.1016/S2468-2667(20)30133-X) PMID: 32502389

17. UK Health Security Agency. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK. 2020. Available from: <https://www.gov.uk/guidance/the-r-value-and-growth-rate>
18. Lau MSY, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *Proc Natl Acad Sci U S A*. 2020;117(36):22430–5. <https://doi.org/10.1073/pnas.2011802117> PMID: [32820074](#)
19. Kremer C, Torneri A, Boesmans S, Meuwissen H, Verdonchot S, Vanden Driessche K, et al. Quantifying superspreading for COVID-19 using Poisson mixture distributions. *Sci Rep*. 2021;11(1):14107. <https://doi.org/10.1038/s41598-021-93578-x> PMID: [34238978](#)
20. Quilty BJ, Clifford S, Hellewell J, Russell TW, Kucharski AJ, Flasche S, et al. Quarantine and testing strategies in contact tracing for SARS-CoV-2: a modelling study. *Lancet Public Health*. 2021;6(3):e175–83. [https://doi.org/10.1016/S2468-2667\(20\)30308-X](https://doi.org/10.1016/S2468-2667(20)30308-X) PMID: [33484644](#)
21. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*. 2021;371(6526):eabe2424. <https://doi.org/10.1126/science.abe2424> PMID: [33234698](#)
22. Ke R, Zitzmann C, Ho DD, Ribeiro RM, Perelson AS. In vivo kinetics of SARS-CoV-2 infection and its relationship with a person's infectiousness. *Proc Natl Acad Sci U S A*. 2021;118(49):e2111477118. <https://doi.org/10.1073/pnas.2111477118> PMID: [34857628](#)
23. Hart WS, Park H, Jeong YD, Kim KS, Yoshimura R, Thompson RN, et al. Analysis of the risk and pre-emptive control of viral outbreaks accounting for within-host dynamics: SARS-CoV-2 as a case study. *Proc Natl Acad Sci U S A*. 2023;120(41):e2305451120. <https://doi.org/10.1073/pnas.2305451120> PMID: [37788317](#)
24. Marc A, Kerioui M, Blanquart F, Bertrand J, Mitjà O, Corbacho-Monné M, et al. Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. *eLife*. 2021;10:e69302. <https://doi.org/10.7554/eLife.69302> PMID: [34569939](#)
25. Kendall M, Ferretti L, Wymant C, Tsallis D, Petrie J, Di Francia A, et al. Drivers of epidemic dynamics in real time from daily digital COVID-19 measurements. *Science*. 2024;385(6710):eadm8103. <https://doi.org/10.1126/science.adm8103> PMID: [38991048](#)
26. Ferretti L, Wymant C, Petrie J, Tsallis D, Kendall M, Ledda A, et al. Digital measurement of SARS-CoV-2 transmission risk from 7 million contacts. *Nature*. 2024;626(7997):145–50. <https://doi.org/10.1038/s41586-023-06952-2> PMID: [38122820](#)
27. Middleton C, Larremore DB. Modeling the transmission mitigation impact of testing for infectious diseases. *Sci Adv*. 2024;10(24):eadk5108. <https://doi.org/10.1126/sciadv.adk5108> PMID: [38875334](#)
28. Chapman L, Kucharski A, Munday J, Gimma A, Wong K, Edmunds J. Analysis of individuals with a high number of contacts in the CoMix study. 2021. Available from: [https://assets.publishing.service.gov.uk/media/610c0fbee90e0706cc852394/S1281\\_CoMix\\_Report\\_-\\_high\\_numbers\\_of\\_contacts\\_SAGE\\_91.pdf](https://assets.publishing.service.gov.uk/media/610c0fbee90e0706cc852394/S1281_CoMix_Report_-_high_numbers_of_contacts_SAGE_91.pdf)
29. Ghafari M, Hall M, Golubchik T, Ayoubkhani D, House T, MacIntyre-Cockett G, et al. Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature*. 2024;626(8001):1094–101. <https://doi.org/10.1038/s41586-024-07029-4> PMID: [38383783](#)
30. Puhach O, Meyer B, Eckerle I. SARS-CoV-2 viral load and shedding kinetics. *Nat Rev Microbiol*. 2023;21(3):147–61. <https://doi.org/10.1038/s41579-022-00822-w> PMID: [36460930](#)
31. Marks M, Millat-Martinez P, Ouchi D, Roberts CH, Alemany A, Corbacho-Monné M, et al. Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *Lancet Infect Dis*. 2021;21(5):629–36. [https://doi.org/10.1016/S1473-3099\(20\)30985-3](https://doi.org/10.1016/S1473-3099(20)30985-3) PMID: [33545090](#)
32. Office for National Statistics. Coronavirus and self-isolation after testing positive in England. 2021. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/bulletins/coronavirusandselfisolationaftertestingpositiveinengland/8to13march2021>
33. Ko YK, Furuse Y, Otani K, Yamauchi M, Ninomiya K, Saito M, et al. Time-varying overdispersion of SARS-CoV-2 transmission during the periods when different variants of concern were circulating in Japan. *Sci Rep*. 2023;13(1):13230. <https://doi.org/10.1038/s41598-023-38007-x> PMID: [37580339](#)
34. Hodcroft EB, Wohlfender MS, Neher RA, Riou J, Althaus CL. Estimating Re and overdispersion in secondary cases from the size of identical sequence clusters of SARS-CoV-2. *PLoS Comput Biol*. 2025;21(4):e1012960. <https://doi.org/10.1371/journal.pcbi.1012960> PMID: [40233303](#)
35. Du Z, Wang C, Liu C, Bai Y, Pei S, Adam DC, et al. Systematic review and meta-analyses of superspreading of SARS-CoV-2 infections. *Transbound Emerg Dis*. 2022;69(5):e3007–14. <https://doi.org/10.1111/tbed.14655> PMID: [35799321](#)
36. Wegehaupt O, Endo A, Vassall A. Superspreading, overdispersion and their implications in the SARS-CoV-2 (COVID-19) pandemic: a systematic review and meta-analysis of the literature. *BMC Public Health*. 2023;23(1):1003. <https://doi.org/10.1186/s12889-023-15915-1> PMID: [37254143](#)
37. Nielsen BF, Eilersen A, Simonsen L, Sneppen K. Lockdowns exert selection pressure on overdispersion of SARS-CoV-2 variants. *Epidemics*. 2022;40:100613. <https://doi.org/10.1016/j.epidem.2022.100613> PMID: [35939969](#)
38. Leung NHL. Transmissibility and transmission of respiratory viruses. *Nat Rev Microbiol*. 2021;19(8):528–45. <https://doi.org/10.1038/s41579-021-00535-6> PMID: [33753932](#)
39. 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. <https://doi.org/10.1038/nature04153> PMID: [16292310](#)
40. Office for National Statistics. Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland: Mid-2020; 2022. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforuk-englandandwalesandscotlandandnorthernireland>