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

Billy J Quilty[1,2¶], Lloyd AC Chapman[1,3¶*], James D Munday[1,5], Kerry LM Wong[1], Amy Gimma[1], Suzanne Pickering[4], Stuart JD Neil[4], Rui Pedro Galão[4], W John Edmunds[1], Christopher I Jarvis[1], Adam J Kucharski[1] on behalf of the CMMID COVID-19 Working Group

[1] Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine

[2] Center for Global Health, Charité Universitätsmedizin Berlin

[3] School of Mathematical Sciences, Lancaster University

[4] Department of Infectious Diseases, School of Immunology & Microbial Sciences, King's College London

[5] Department of Biosystems Science and Engineering, ETH Zürich, Basel

* Corresponding author

Email: l.chapman4@lancaster.ac.uk

¶ These authors share first authorship

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 viral load 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 basic reproduction number (R_0 = 2.2, 95% CI 2.2-2.3) from first principles and 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 regular testing every 3 days, or pre-event testing with a minimum event size of 10, could reduce the mean reproduction number below 1 with moderate to high levels of uptake (60-80%) for pre-pandemic contact levels. This work demonstrates the potential for using viral load and contact data to predict 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 viral load dynamics. Here, we instead integrate data from real contact surveys conducted in the UK before and during the COVID-19 pandemic with data on 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 closely match real-time estimates from other epidemiological studies, with reductions in secondary infections consistent with observed transmission dynamics during periods of contact restrictions in the UK. We show that variation in numbers of daily contacts is predictive of superspreading and therefore provides a tool with which to monitor superspreading potential 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 ("wrong person" hypothesis) or whether most individuals pass through a highly infectious period which happens to coincide with a period of high contact ("wrong time" 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 viral load trajectories and infectivity combined with data on reported numbers of daily contacts from two social contact surveys in the UK. 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 lateral flow tests (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 characterised as 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),

$$R_i \sim NegBin(k, R),$$

such that the variance in R_i is $R+R^2/k$. 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 lateral-flow antigen tests (LFTs) on reducing R and the potential for superspreading events (by decreasing variation in numbers of secondary infections, i.e. increasing k).

Results

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

Figure 1A 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 9 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 <3% of the total CoMix survey sample.

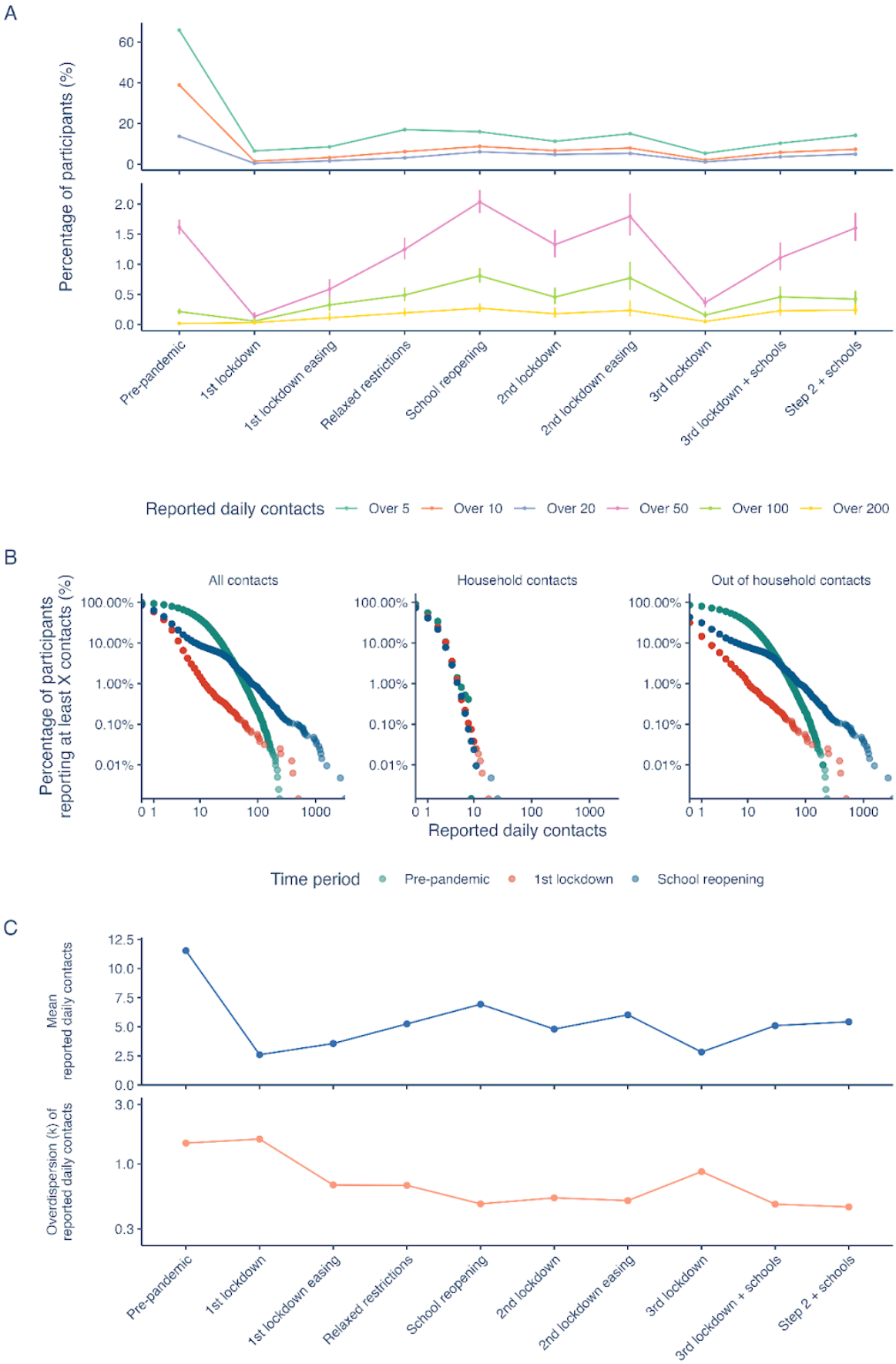

A
Percentage of participants (%)
Pre-pandemic
1st lockdown
1st lockdown easing
Relaxed restrictions
School reopening
2nd lockdown
2nd lockdown easing
3rd lockdown
3rd lockdown + schools
Step 2 + schools
Reported daily contacts
Over 5
Over 10
Over 20
Over 50
Over 100
Over 200
B
All contacts
Household contacts
Out of household contacts
Percentage of participants reporting at least X contacts (%)
Reported daily contacts
Time period
Pre-pandemic
1st lockdown
School reopening
C
Mean reported daily contacts
Overdispersion (k) of reported daily contacts

Figure 1: 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 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 negative binomial distribution fitted to contacts during the different time periods. N.B. The heavier tail in the contact distribution for the "School reopening" period in B compared to 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. We conduct a sensitivity analysis to determine the impact of this on our results (Fig. S3).

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	Over 5 (%)	Over 10 (%)	Over 20 (%)	Over 50 (%)	Over 100 (%)	Over 200 (%)
Pre-pandemic (BBC Pandemic)	01/09/2017	01/12/2018	40162	65.8	38.9	13.7	1.6	0.2	0
Lockdown 1	23/03/2020	03/06/2020	15906	6.6	1.5	0.5	0.1	0.1	0
Lockdown 1 easing	04/06/2020	29/07/2020	10651	8.6	3.3	1.7	0.6	0.3	0.1
Relaxed restrictions	30/07/2020	03/09/2020	15415	17.6	6.4	3.2	1.3	0.5	0.2
School reopening	04/09/2020	24/10/2020	20759	16.7	9	6.2	2.1	0.8	0.3
Lockdown 2	05/11/2020	02/12/2020	10008	11.9	6.9	4.9	1.3	0.5	0.2
Lockdown 2 easing	03/12/2020	19/12/2020	5898	16.1	8.1	5.4	1.8	0.8	0.2
Lockdown 3	05/01/2021	07/03/2021	21542	5.6	2.2	1.2	0.4	0.2	0.1
Lockdown 3 + schools	08/03/2021	31/03/2021	8452	10.8	5.9	3.7	1.2	0.5	0.2
Step 2 + schools	16/04/2021	16/05/2021	1771	16.6	8	5	1.9	0.7	0.3

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

The estimated mean and overdispersion parameter *k* (lower values indicating more overdispersion (variation) in the distribution) of numbers of daily contacts in the UK were lower than that observed

pre-pandemic, with mean daily contacts averaging ~6 during the pandemic compared to ~12 pre-pandemic, and *k* averaging ~0.6 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 ~3 during the first lockdown to ~7 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 (Figure 1C). The overdispersion in contacts, *k,* was similar pre-pandemic and during the first lockdown (1.5 and 1.6 respectively) then became lower (more varied) following the easing of the first lockdown, where *k* averaged ~0.6 (Figure 1C). Stratifying contacts by household/non-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 (Figure S1). 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 45 minutes (IQR: 5, 180 minutes) for out-of-household contacts compared to 720 minutes (12 hours) (95% CI: 270, 1440 minutes) for household contacts (Figure S2).

Variation in viral load and infectivity over the course of infection

Analysis by Kissler et al.[7] 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) (Figure 2A). By mapping viral load to infectivity via a logistic function based on viral load and culture (representing live, infectious virus) from Pickering et al. [8] (Figure 2B), calculating P(infectivity) by day (Figure 2C), then integrating under the infectivity curve, we can reproduce substantial heterogeneity in individual infectiousness as reported by Ke et al.[6], with a >70-fold difference in individual-level infectivity between the 2.5% and 97.5% percentiles of the individual-level distribution (0.08 and 5.74, a.u, respectively) and a shape parameter of a fitted Gamma distribution of 1.42. When simulating whether or not individuals would infect a single contact on each day of their infection, a substantial proportion (18.7%) were estimated to infect contacts for zero days, but the remainder infected contacts for at least one day (Figure 2). Across all individuals, the median duration of infectiousness,

i.e. having high enough viral load to cause infections (culturable virus), was 2 days (95% prediction interval: 0, 6 days). Over 50% of individuals had a peak viral load for which their probability of infectiousness was greater than 80%, 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% as the time window for this level of infectivity is so narrow that it must coincide perfectly with a high number of contacts for a high number of infections to occur.

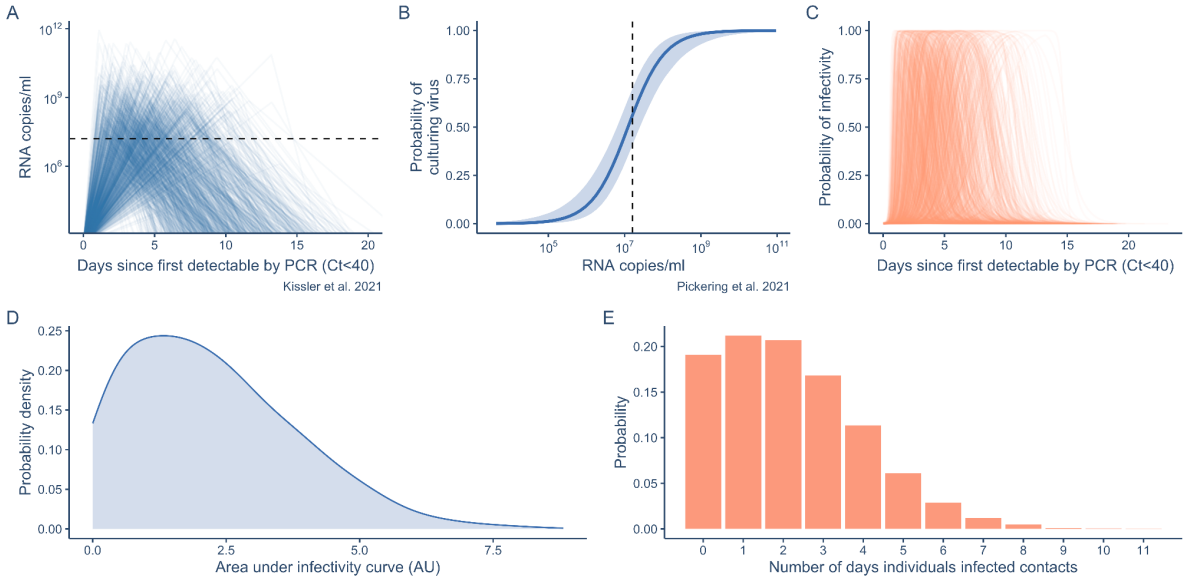


Figure 2: Variation in viral load progression and infectivity. A. Individual-level viral load trajectories. **B.** Logistic model for probability of culturing virus given a certain viral load. **C.** Probability of infectivity over time. **D.** Area under the infectivity curve. **E.** Distribution of the number of days for which individuals infected contacts. Dashed line in A and B represents the RNA copies/ml required for a 50% probability of culturing virus/infectivity.

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

Simulating secondary infections using an individual-based model based on daily contact rates, contact duration, and daily infectivity derived from viral load (see *Methods* section), we estimated an R_0 (mean initial reproduction number based on UK pre-pandemic contact rates) of 2.2 (95% confidence interval (CI) 2.2-2.3), closely matching contemporaneous estimates[9]. The overdispersion in the initial reproduction number, *k,* was estimated at 1.1 (95% CI 1.1-1.2), somewhat higher than other estimates (less heterogeneous)[1]. Limiting our analysis to 2020 prior to the widespread emergence of more transmissible variants of SARS-CoV-2 and mass vaccination, during the first

national lockdown *R* was estimated at 0.4 (95% CI 0.4-0.4) and *k* at 0.5 (95% CI 0.4-0.6). *R* rose as lockdown eased (*R* = 0.6, 95% CI 0.5-0.6) and restrictions were relaxed in the summer of 2020 (*R* = 0.9, 95% CI 0.9-1.0), before rising above 1 as schools reopened in the autumn (*R* = 1.3, 95% CI 1.2-1.5). As with contacts, *k* reduced as the pandemic began, indicating more variability in the secondary infection distribution, and varied between 0.3 and 0.4 after the first lockdown up to the end of 2020 (Figure 3). A sensitivity analysis imputing a heavier tail for the BBC Pandemic contact distribution based on the tails of contact distributions from relaxed restrictions periods during the CoMix survey led to a small increase in our estimate of R_0 from 2.2 to 2.4 and a reduction in *k* from 1.1 to 0.9 (Figure S3).

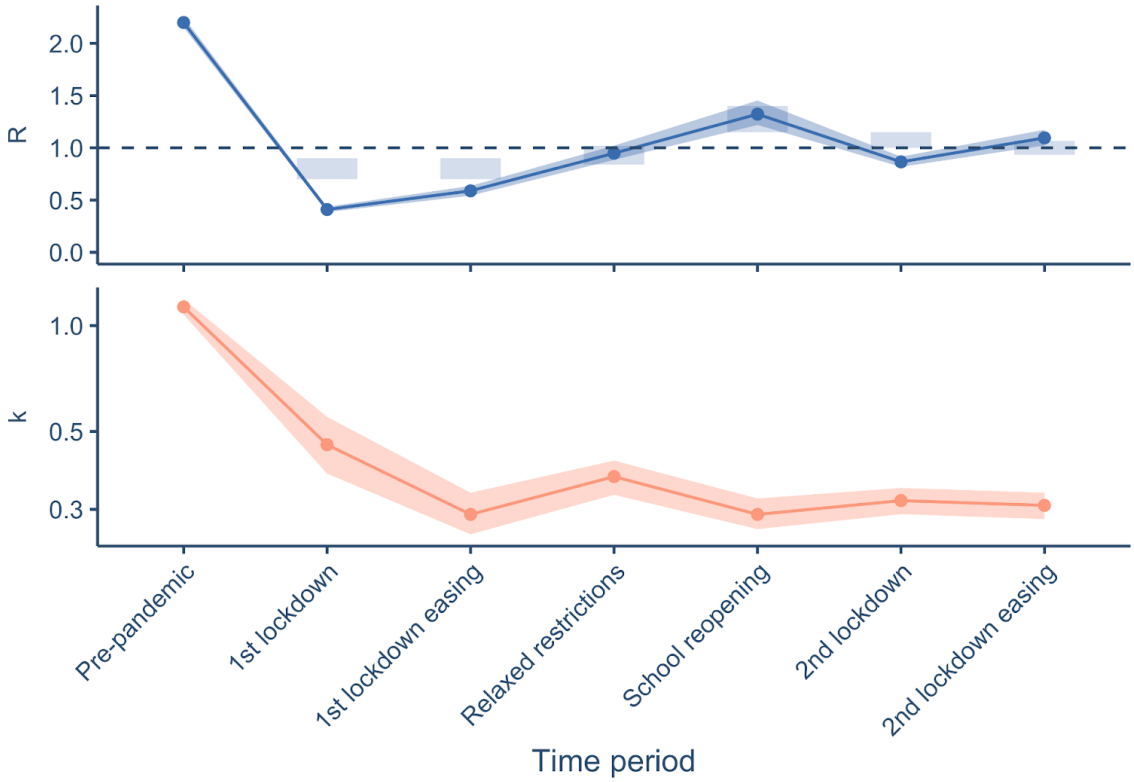


Figure 3. Estimates of the mean and dispersion of negative binomial distributions fitted to simulated secondary infection distributions by time period in 2020. The dispersion parameter *k* gives an indication of the variation in the secondary infection distribution, with smaller values corresponding to greater variation and values less than 1 corresponding to very large variation. Grey boxes in *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. Shaded bands show 95% bootstrap CIs for the *R* and *k* estimates.

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

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, while in the “heterogeneous” cases, contacts and viral load trajectories were allowed to vary according to their full distributions. If both variables are homogeneous, then the secondary infection distribution is approximately Poisson with equal mean and variance (very large values of *k*). If contacts are Poisson distributed but viral load trajectories are allowed to vary between individuals, *k* is 0.8-1.7. If viral load is fixed at the median trajectory, but contacts are overdispersed, values of *k* are lower during pandemic periods and approximately equal to that of overdispersed contacts and variable viral load trajectories, 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 (Figure 4).

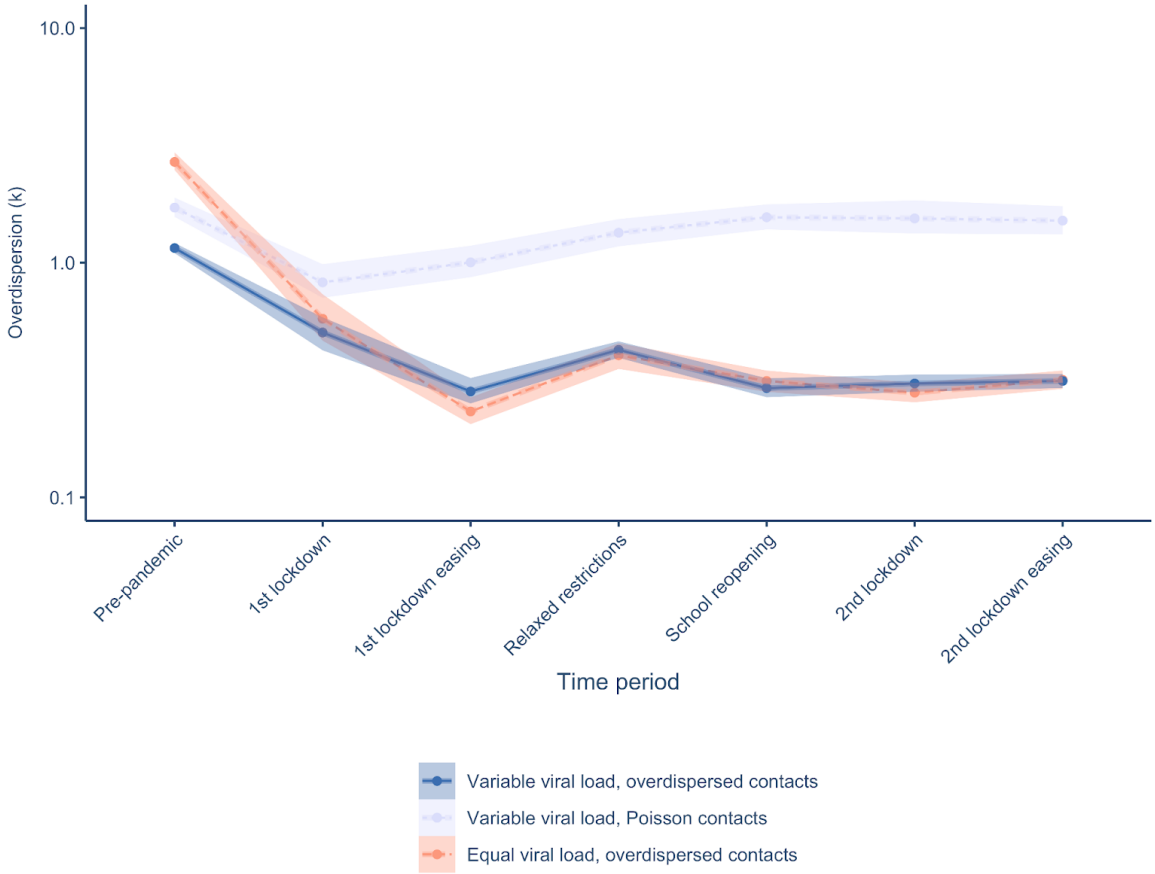


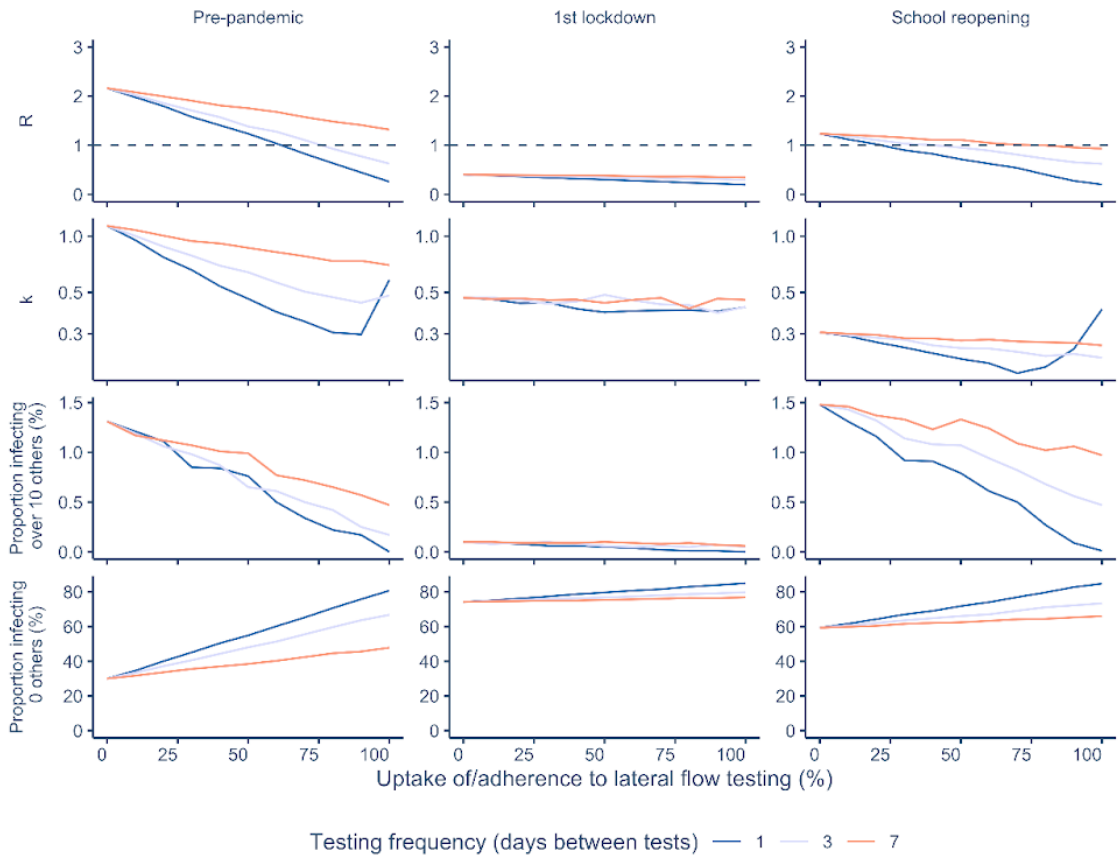
Figure 4. Estimates of the dispersion of negative binomial distributions fitted to secondary infection distributions by period, with different combinations of homogeneous and heterogeneous contact rates and viral load trajectories. The dispersion parameter *k* gives an indication of the variation in 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 not shown due to being very large (i.e., numbers of secondary infections are approximately Poisson distributed). Shaded bands show 95% bootstrap CIs.

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

Finally, we assessed the hypothesis that if those with the highest viral loads are most likely to cause superspreading events, and lateral flow tests are most sensitive for those with the highest viral loads, then lateral flow tests should be able to prevent superspreading events, here defined as infecting over 10 contacts. We modelled regular (1-, 3-, and 7-day frequency) and pre-event lateral flow testing (before meeting >10, >20, >50 others) to determine the comparative effectiveness of rapid testing to control transmission. We considered three indicative timepoints before (BBC Pandemic) and during (1st lockdown (March-June) and the reopening of schools (September 2020)) the pandemic, and varied the rate of uptake or adherence to the specified policy. For pre-pandemic levels of contacts,

uptake must exceed 60% 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 (Figure 5A). 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 50% (Figure 5A). Pre-event testing acts similarly, with high uptake necessary to reduce *R* below 1 for pre-pandemic levels of contact, but is effective in reducing *R* below 1 even if testing only when attending events with more than 20 others for moderate levels of adherence during relaxed restrictions with schools open (Figure 5B). 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 a substantial increase in the proportion of individuals that infect no one. This results in a decrease in *k* (greater overdispersion in the secondary infection distribution) (Figure 5).

A


Pre-pandemic
1st lockdown
School reopening
R
k
Proportion infecting over 10 others (%)
Proportion infecting 0 others (%)
Uptake of/adherence to lateral flow testing (%)
Testing frequency (days between tests) 1 3 7

B

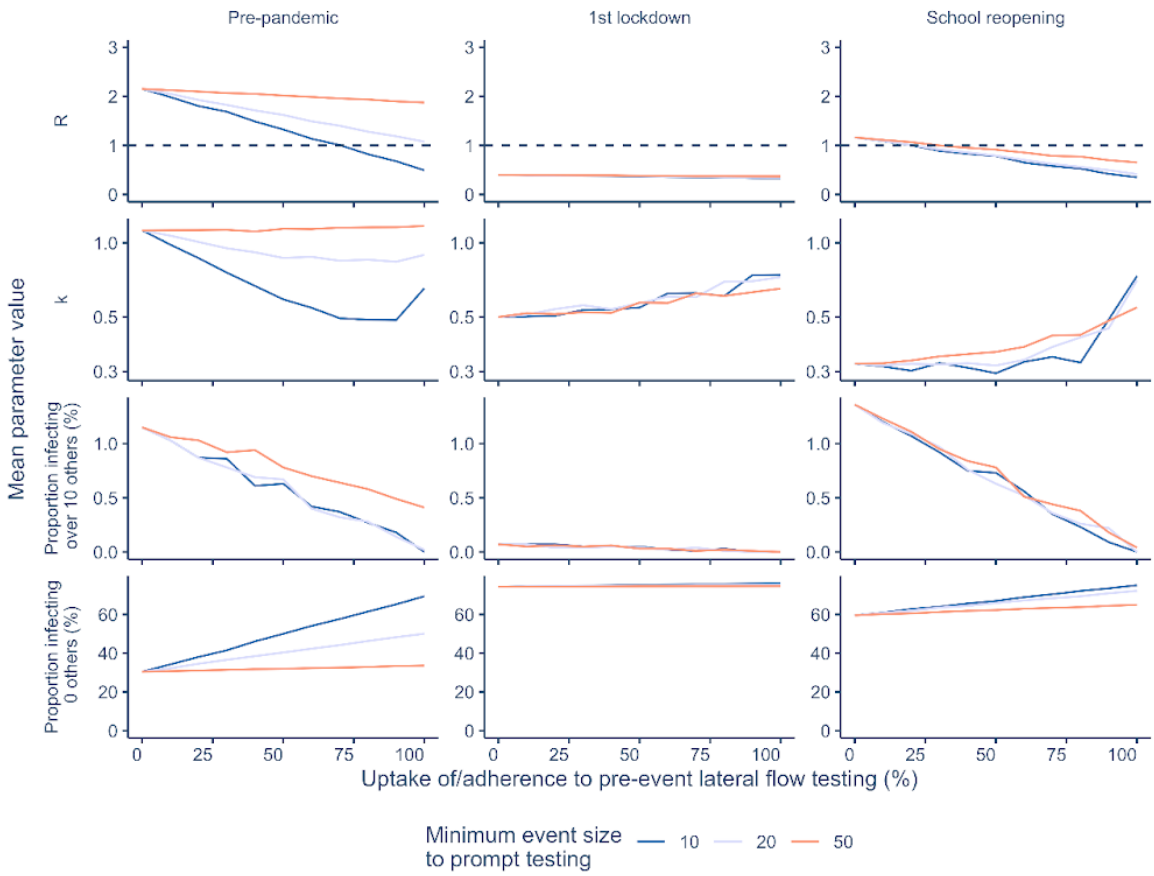

Pre-pandemic
1st lockdown
School reopening
Mean parameter value
R
k
Proportion infecting over 10 others (%)
Proportion infecting 0 others (%)
Uptake of/adherence to pre-event lateral flow testing (%)
Minimum event size to prompt testing 10 20 50

Figure 5: Effect of testing on *R*, *k*, and the proportion infecting over 10 or 0 others for varying levels of uptake/adherence and background contact rates. A. Regular testing. B. Pre-event testing.

Discussion

Using an individual-based model incorporating viral load trajectories and reported daily contacts, we find that we can reconstruct secondary infection distributions for SARS-CoV-2 from first principles over the course of the pandemic in the UK in 2020, closely matching contemporaneous estimates. Daily contact rates became both lower on average (from ~12 per day to ~6 per day) and more overdispersed during the pandemic in the UK in 2020 as some individuals maintained high contact rates (e.g., essential workers) while others had their number of daily contacts reduced to zero (those working from home). This drove a decrease in the mean number of secondary infections per infected individual (reproduction number), but an increase in the dispersion in the reproduction number. Our estimates of the degree of overdispersion in the reproduction number are similar to those reported by other authors, which are mostly between 0.3 and 0.6 [10,11].

Our analysis also shows that dynamic shifts in the mean and overdispersion of the reproduction number (Figure 3) closely track changes in the mean and overdispersion of daily numbers of contacts (Figure 1C). This suggests that contact overdispersion can be reliably measured in contact surveys and is predictive of superspreading, and therefore worth following in real-time during a pandemic, 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.

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 >50% estimated to have viral loads capable of infecting many others on at least one day of their infection. 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 (Figure 4). This indicates that superspreading is more a case of “wrong place, wrong time” than “wrong person”. Hence if reductions

in contact rates can be targeted specifically at infected individuals during this window of high infectivity (e.g., by daily testing of their contacts[12]) 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[4,6,13–16]. 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[8]. Our use of reported heterogeneity in viral load trajectories[7] 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].

Like others [6,16,17], 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 as the relative increase in those infecting zero others exceeded the decrease in those infecting many others (here, over 10 others). Hence, assessment of superspreading 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 (e.g., 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. This indicates that rapid testing could be an effective, minimally disruptive intervention to reduce transmission if uptake/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.[16] similarly found that daily lateral flow testing could

significantly reduce numbers of secondary infections (preventing ~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.

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[18–20] and POLYMOD[21] surveys. However, previous surveys required participants to list contacts individually to include other information such as contact age, sex, occupation, meaning that it was difficult to include mass contacts such as those made at a large gathering, thus truncating the true contact distribution (Figure 1B). 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- 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 our estimate of R_0 and a small reduction in k. We assume that the probability of shedding infectious virus is equal to the probability of culturing virus, which in turn is dependent on intra-host viral load kinetics over the course of infection[8,22]. We do not account for variability in viral load or infectiousness due to factors that are not well accounted for in the National Basketball Association cohort in Kissler et al.[7], such as age, which there is some evidence has an effect on infectiousness[6]. We also do not account for correlation of the timing of peak viral load and infectiousness with symptom onset and intensity[23], 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]. 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 (VOCs) such as Alpha, Delta, and Omicron. We assumed within-household contacts for each individual were the same each day, whereas out-of-household contacts were

sampled randomly for each individual per day, which may underestimate the overall 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[25]; 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 VOCs 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 (less overdispersed transmission) during the Omicron wave[26,27]. 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[28,29], 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 (higher *k*)[30]. Together, these studies suggest that while overdispersion may be weakening with widespread immunity and new VOCs, 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 respiratory viruses that exhibit heterogeneous transmission, including influenza, RSV, and adenoviruses. 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.

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 over 50% of infected individuals being sufficiently infectious on 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.

Methods

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[7,8]). 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).

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[21]. 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 [31].

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 negative exponential distribution to the number of individuals reporting over 250 contacts in relaxed restrictions time periods ("Relaxed restrictions", "School reopening" and "Step 2 + schools") from the CoMix data and sampling proportionally from this distribution for the BBC Pandemic contacts.

Analysis of social contact data

We calculated 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 nine indicative time periods between March 2020 and May 2021 representing different levels of restrictions (as defined in previous work [31,32]).

We compared contact distributions (overall and stratified by household/out of household) for the BBC Pandemic [18] survey (here referred to as Pre-pandemic) and CoMix surveys [31,32]. For plotting contact distributions, we calculated the percentage of participants in each time period who reported a certain number of contacts to account for differences in numbers of participants per time period. 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.

Reconstruction of secondary infection distribution

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 as a piecewise function defined by a proliferation phase (days from exposure to peak), a clearance phase (days from peak to cessation) and a peak viral load, with these three parameters drawn from distributions given in Kissler et al. 2021[7] (Figure 6). Viral load (in cycle threshold (Ct) units) was assumed to be Ct 40 (negative) at exposure and cessation of shedding, with a peak viral load of Ct 20.2 (95% CI: 18.7, 21.7). All parameters were assumed to be normally distributed. We then

estimated the probability of infectiousness for a given viral load (in Ct) by fitting a logistic regression model to the probability of culturing virus at that viral load[8] (Figure 2B), producing an infectiousness trajectory (Figure 2C). Parameter uncertainty was accounted for by drawing values of the logistic regression coefficients (intercept and viral load coefficient) for each individual from a bivariate normal distribution centred at the maximum likelihood estimates with covariance given by 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.

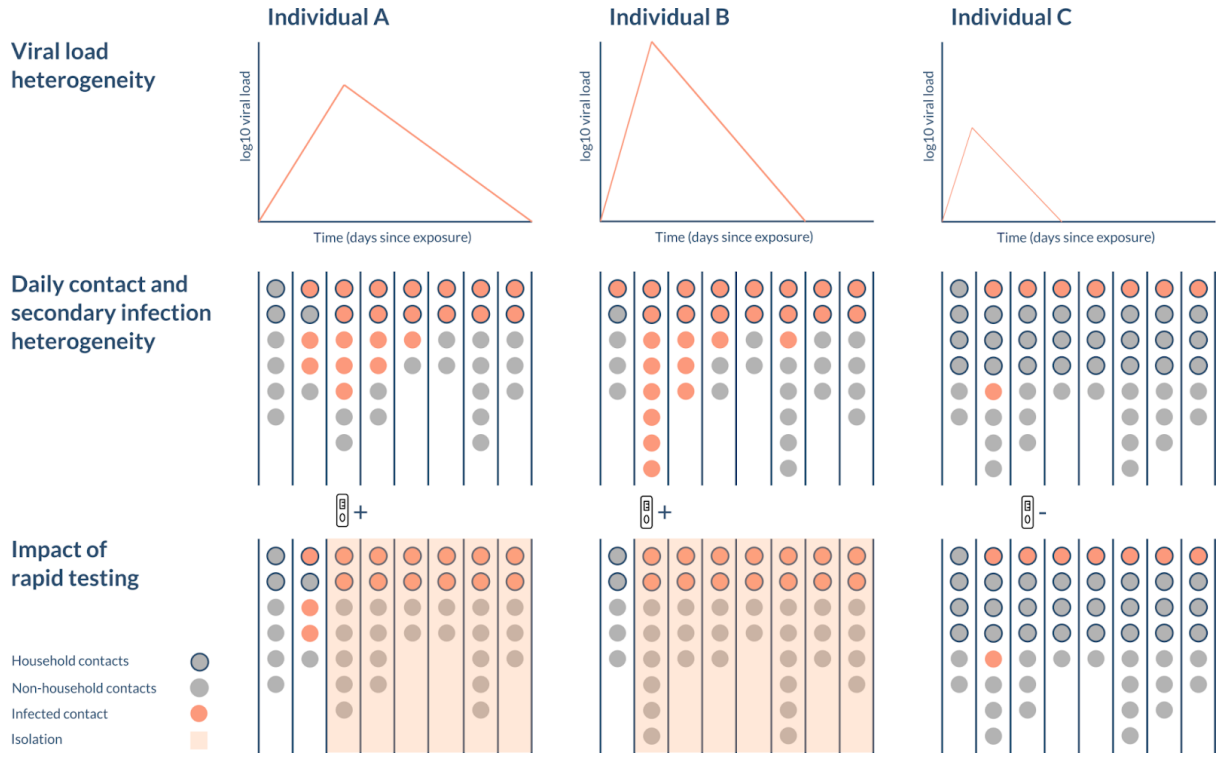


Figure 6: Schematic of the model. The viral load and number of daily contacts (circles) varies from person to person and over time, influencing the number of secondary infections (stratified by household (outlined) and non-household (no outline)) they generate. The viral load at time of testing also determines the likelihood they will test positive and subsequently isolate.

To simulate secondary infections, we first randomly sampled a number (and duration) of contacts from empirical contact distributions (the BBC Pandemic contact survey[18–20] and CoMix contact survey[31,32]) such that each index case had a certain number of household contacts and certain number of daily non-household contacts (work, school and “other” contacts). Each household and

non-household contact had a duration, defined as the proportion of a 24-hour period spent at home or outside of the home respectively that the contact lasted, also independently sampled from the CoMix survey (Figure S2). Household contacts were sampled once per index case, whereas non-household contacts were sampled daily. The infection process for each contact 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 (see Supplementary Material for further details). We assumed uniform susceptibility of individuals in the model which does not vary by, for example, age.

We then estimated the corresponding *R* (mean number of secondary infections) and *k* (overdispersion in the number of secondary infections) by fitting a negative binomial distribution to the numbers of secondary infections across all individuals. 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.

Simulation of interventions

We estimated the impact of regular testing every 3 days with LFTs, with detection calculated by fitting a logistic regression model to the probability of detection with LFTs given viral load[8], and individuals self-isolating at home upon their first positive test (i.e, we reduced the number of work, school and casual contacts to zero after the date of the positive test while leaving home contacts unchanged).

The code and data for this study can be found at https://github.com/bquilty25/superspreading_testing.

References

1. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
2. Chen, P. Z. *et al.* Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. *eLife* **10**, e65774 (2021).
3. Goyal, A., Reeves, D. B., Cardozo-Ojeda, E. F., Schiffer, J. T. & Mayer, B. T. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. *eLife* **10**,

467 e63537 (2021).

468 4. Susswein, Z. & Bansal, S. Characterizing superspreading of SARS-CoV-2 : from mechanism to
469 measurement. *medRxiv* 2020.12.08.20246082 (2020) doi:10.1101/2020.12.08.20246082.

470 5. Killingley, B. *et al.* Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge.
471 Preprint at https://doi.org/10.21203/rs.3.rs-1121993/v1 (2022).

472 6. Ke, R. *et al.* Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial
473 heterogeneity in infectiousness. *Nat. Microbiol.* **7**, 640–652 (2022).

474 7. Kissler, S. M. *et al.* Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic
475 and public health strategies. *PLOS Biol.* **19**, e3001333 (2021).

476 8. Pickering, S. *et al.* Comparative performance of SARS-CoV-2 lateral flow antigen tests and
477 association with detection of infectious virus in clinical specimens: a single-centre laboratory
478 evaluation study. *Lancet Microbe* **2**, e461–e471 (2021).

479 9. Davies, N. G. *et al.* Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and
480 demand for hospital services in the UK: a modelling study. *Lancet Public Health* **5**, e375–e385
481 (2020).

482 10. Lau, M. S. Y. *et al.* Characterizing superspreading events and age-specific infectiousness of
483 SARS-CoV-2 transmission in Georgia, USA. *Proc. Natl. Acad. Sci.* **117**, 22430–22435 (2020).

484 11. Kremer, C. *et al.* Quantifying superspreading for COVID-19 using Poisson mixture distributions.
485 *Sci. Rep.* **11**, 14107 (2021).

486 12. Quilty, B. J. *et al.* Quarantine and testing strategies in contact tracing for SARS-CoV-2: a
487 modelling study. *Lancet Public Health* (2021) doi:10.1016/S2468-2667(20)30308-X.

488 13. Goyal, A., Reeves, D. B., Cardozo-Ojeda, E. F., Schiffer, J. T. & Mayer, B. T. Viral load and
489 contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. *eLife* **10**,
490 e63537 (2021).

491 14. Sun, K. *et al.* Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*
492 **371**, eabe2424 (2021).

493 15. Ke, R., Zitzmann, C., Ho, D. D., Ribeiro, R. M. & Perelson, A. S. In vivo kinetics of SARS-CoV-2
494 infection and its relationship with a person's infectiousness. *Proc. Natl. Acad. Sci.* **118**,
495 e2111477118 (2021).

496 16. Hart, W. S. *et al.* Analysis of the risk and pre-emptive control of viral outbreaks accounting for
497 within-host dynamics: SARS-CoV-2 as a case study. *Proc. Natl. Acad. Sci.* **120**, e2305451120

(2023).

17. Middleton, C. & Larremore, D. B. Modeling the transmission mitigation impact of testing for infectious diseases. *Sci. Adv.* **10**, eadk5108 (2024).

18. Klepac, P., Kissler, S. & Gog, J. Contagion! The BBC Four Pandemic - The model behind the documentary. *Epidemics* **24**, 49–59 (2018).

19. Klepac, P. *et al. Contacts in Context: Large-Scale Setting-Specific Social Mixing Matrices from the BBC Pandemic Project*. http://medrxiv.org/lookup/doi/10.1101/2020.02.16.20023754 (2020) doi:10.1101/2020.02.16.20023754.

20. Kucharski, A. J. *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.* S1473309920304576 (2020) doi:10.1016/S1473-3099(20)30457-6.

21. Mossong, J. *et al.* Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLOS Med.* **5**, e74 (2008).

22. Ke, R. *et al. Daily Sampling of Early SARS-CoV-2 Infection Reveals Substantial Heterogeneity in Infectiousness*. http://medrxiv.org/lookup/doi/10.1101/2021.07.12.21260208 (2021) doi:10.1101/2021.07.12.21260208.

23. Marks, M. *et al.* Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *Lancet Infect. Dis.* **21**, 629–636 (2021).

24. Marc, A. *et al.* Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. *eLife* **10**, e69302 (2021).

25. Office for National Statistics. Coronavirus and self-isolation after testing positive in England. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/bulletins/coronavirusandselfisolationaftertestingpositiveinengland/8to13march2021 (2021).

26. Ko, Y. K. *et al.* Time-varying overdispersion of SARS-CoV-2 transmission during the periods when different variants of concern were circulating in Japan. *Sci. Rep.* **13**, 13230 (2023).

27. Hodcroft, E. B., Wohlfender, M. S., Neher, R. A., Riou, J. & Althaus, C. L. Estimating Re and overdispersion in secondary cases from the size of identical sequence clusters of SARS-CoV-2. 2024.05.26.24307940 Preprint at https://doi.org/10.1101/2024.05.26.24307940 (2024).

28. Du, Z. *et al.* Systematic review and meta-analyses of superspreading of SARS-CoV-2 infections. *Transbound. Emerg. Dis.* **69**, e3007–e3014 (2022).

29. 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* **23**, 1003 (2023).

30. Nielsen, B. F., Eilersen, A., Simonsen, L. & Sneppen, K. Lockdowns exert selection pressure on overdispersion of SARS-CoV-2 variants. *Epidemics* **40**, 100613 (2022).
31. Jarvis, C. I. *et al.* Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med.* **18**, 124 (2020).
32. 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 Med.* **19**, e1003907 (2022).

Acknowledgements

The authors acknowledge funding from the Medical Research Council (MC_PC_19065 - WJE); European Commission (EpiPose 101003688 - KLMW, AG, WJE); NIHR (CV220-088 - COMIX; 16/137/109 - BJQ, 16/136/46 - BJQ); Bill & Melinda Gates Foundation (OPP1139859 - BJQ); Wellcome Trust (206250/Z/17/Z - AJK, LACC; WT098049AIA - SJDN); NIHR HPRU in Modelling & Health Economics (NIHR200908 - AJK, LACC); King's Together Rapid COVID-19 Call - SJDN, RPG; Huo Family Foundation Award - SP, SJDN.

Ethics

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