Received: 11 July 2021 | Revised: 16 February 2022 | Accepted: 17 February 2022

DOI: 10.1002/sim.9365

RESEARCH ARTICLE

Statistics in Medicine WILEY

The required size of cluster randomized trials of nonpharmaceutical interventions in epidemic settings

Justin K. Sheen[1] | **Johannes Haushofer**[2,3,4,5] | **C. Jessica E. Metcalf**[1,6] | **Lee Kennedy-Shaffer**[7]

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey

[2]Department of Economics, Stockholm University, Stockholm, Sweden

[3]Research Institute of Industrial Economics, Stockholm, Sweden

[4]Max Planck Institute for Research on Collective Goods, Bonn, Germany

[5]Jain Family Institute, New York

[6]School of Public and International Affairs, Princeton University, Princeton, New Jersey

[7]Department of Mathematics and Statistics, Vassar College, Poughkeepsie, New York

Correspondence
Lee Kennedy-Shaffer, Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY, USA.
Email: lkennedyshaffer@vassar.edu

To control the SARS-CoV-2 pandemic and future pathogen outbreaks requires an understanding of which nonpharmaceutical interventions are effective at reducing transmission. Observational studies, however, are subject to biases that could erroneously suggest an impact on transmission, even when there is no true effect. Cluster randomized trials permit valid hypothesis tests of the effect of interventions on community transmission. While such trials could be completed in a relatively short period of time, they might require large sample sizes to achieve adequate power. However, the sample sizes required for such tests in outbreak settings are largely undeveloped, leaving unanswered the question of whether these designs are practical. We develop approximate sample size formulae and simulation-based sample size methods for cluster randomized trials in infectious disease outbreaks. We highlight key relationships between characteristics of transmission and the enrolled communities and the required sample sizes, describe settings where trials powered to detect a meaningful true effect size may be feasible, and provide recommendations for investigators in planning such trials. The approximate formulae and simulation banks may be used by investigators to quickly assess the feasibility of a trial, followed by more detailed methods to more precisely size the trial. For example, we show that community-scale trials requiring 220 clusters with 100 tested individuals per cluster are powered to identify interventions that reduce transmission by 40% in one generation interval, using parameters identified for SARS-CoV-2 transmission. For more modest treatment effects, or when transmission is extremely overdispersed, however, much larger sample sizes are required.

KEYWORDS

nonpharmaceutical interventions, power, reproduction number, sample size, SARS-CoV-2

1 | INTRODUCTION

Since its emergence in late 2019, the pandemic SARS-CoV-2 virus has spread globally, resulting in millions of deaths.[1] Before vaccines were developed and authorized, policy responses to control the spread of the virus relied on nonpharmaceutical interventions (NPIs). NPIs—including, among other measures, mask mandates, school and business closures, and restrictions on travel—have the potential to reduce transmission of the virus. However, which specific NPIs have an

effect on transmission remains largely uncertain, while their economic or psychological costs may be substantial. This presents a challenge for policy design. Retrospective statistical analyses of reported cases or deaths,[2-4] in some cases supplemented with mobility data,[5,6] have provided one means to estimate these impacts, but the conclusions are far from clear. In the first waves of the pandemic, school closures, for example, were found to have relatively minor effects in various types of studies.[2,3,7] However, more recent analyses suggested that school closure might reduce transmission, perhaps by as much as 15%.[4,8-11]

Ultimately, while this body of work has helped inform the policy landscape, important gaps remain. Different interventions often overlap in timing, and communities that adopt particular interventions may be similar in other ways, resulting in confounding that biases estimates, even under the null hypothesis of no intervention effect, and challenges the validity of hypothesis tests.[12] In addition, since these observational studies generally rely on the number of observations available in existing data sets, their power to detect meaningful effect sizes may be low or uncertain.

Randomized controlled trials (RCTs) are widely used to evaluate the impact of interventions on infectious diseases, as seen in recent vaccine trials with tens of thousands of participants.[13,14] While the first trials for SARS-CoV-2 vaccines focused on estimating direct effects on individual-level protection, cluster randomized trials (cRCTs) can also provide valuable insight into indirect and total effects of a vaccination regimen in a community.[15,16] For NPIs, clusters are the natural scale of analysis, as many interventions are implemented at this scale—for example, by school districts or municipalities—and as both direct and indirect protection are of interest in policy design.

Random allocation of interventions to different communities in parallel has been proposed as an approach to assess whether interventions affect transmission.[17,18] Indeed, because transmission is rapid, effects can be evaluated in a matter of weeks. Because the impact of interventions remains uncertain, trials that allow control clusters to adopt the intervention after only a short lag may achieve equipoise and have relatively high community acceptability during an epidemic. Despite these advantages, these RCTs might still incur substantial costs and require significant coordination, implementation, and testing. These logistical challenges grow as the number of intervention units increases. Thus, the degree to which deploying RCTs to evaluate NPIs is a useful policy tool will depend on the sample size required to detect, with adequate power, a meaningful reduction in transmission.[19]

Simulation approaches have been previously used to evaluate the statistical power of cRCTs evaluating vaccination at both individual and cluster scales[20-23] This allows investigators to size cRCTs to have a desired power to detect a specified true effect size of interest (ie, reduction in transmission). Although estimators have different properties and interpretations depending on the phase of the epidemic—for example, by capturing more indirect effects or having less impact when there is more preexisting immunity—this allows hypothesis tests to be appropriately powered.[22,23] Here, we build on these results to provide estimates for the number of clusters and number of individuals measured within each cluster needed to test the effectiveness of an NPI in a short period of time, with the aim of bounding the feasibility of such analyses. Additionally, our simulation approach considers the use of baseline testing, the effects of variable cluster sizes and overdispersion, and the use of different delays in sampling after intervention and matching techniques in order to improve power.

We provide investigators with several tools to estimate the required sample size of a cRCT powered to test the effect of an NPI on epidemic transmission. These include approximate formulae that can be used to size the trial, and simulation results that inform how power depends on key parameters. For more precise sample size and power calculations, simulations adapted to the context of the trial under consideration will be necessary. The results presented here, however, can provide a baseline for assessing whether a trial may be feasible with a reasonable sample size and can provide a starting point for more specific investigations.

2 | METHODS

2.1 | Trial design

We consider cRCTs where there are $N = N_1 + N_0$ total clusters enrolled, with N_1 in the intervention arm and N_0 in the control arm. At a specified day t after infections have begun in the clusters, one round of sampling is performed, sampling m_0 individuals from each cluster and testing them for the infection. In the simulations, we set the desired average proportion of infectious individuals, $E[I_t]$, as a parameter, and identify the time t so that the average of this value across clusters matches this value. We then use this value t in all clusters, so there will be variability in the proportion infectious across clusters. At that point, the intervention begins in the intervention arm clusters, a randomly-selected half of the

study clusters. A set amount of time later (in our main results, this amount of time is equal to one generation interval, the time from the infection of a primary case to infection of a secondary case infected by that primary case),[24] another round of sampling occurs, this time sampling m_1 individuals from each cluster and testing them. We denote this time point by $t+1$. A range of sampling times beyond one generation interval after intervention are explored in Section 3.2.2. In the approximate formulae derived here, m_0 and m_1 need not be equal; our simulations, however, assume the same number of individuals are tested at both time points for simplicity. We assume the number of individuals tested in each cluster is the same across clusters. We assume that the test accurately identifies infectious individuals. More complex models may be used to adjust for known imperfect sensitivity and specificity, as has been done in other settings.[12,25] In the approximate formulae, each generation of infections is discrete (discrete-generation), such that no two generations of infections will overlap in time with one another, while in the simulations the generations may overlap.

In some cases where clusters are relatively small, it may be reasonable to assume that everyone or nearly everyone in the cluster will have their outcome measured. For example, some schools, universities, long-term care facilities, and workplaces have proposed or implemented universal testing strategies.[26-29] We consider this setting first, followed by settings where a simple random sample (without replacement) is chosen for testing, independently at each time point.

Each cluster j then has two values associated with it: $Y_{j,t}$, the number of sampled individuals who test positive in the first round of sampling (preintervention); and $Y_{j,t+1}$, the number of sampled individuals who test positive in the second round of sampling (postintervention). We conduct analysis using test statistics based on the quantity $\frac{Y_{j,t+1}/m_1}{Y_{j,t}/m_0}$ from each cluster, for example, by comparing the mean of this statistic in the intervention arm to that in the control arm. This statistic estimates the growth rate of infections, which is related to the reproduction number—R_t, the number of secondary infections that arise from a typical primary case at time t—in the cluster,[30] so the difference in means estimates the reduction in the growth rate of new infections.[17]

This statistic may not always be unbiased, due to adjustments made for zero-case clusters, asymmetric effects of the progress of the epidemic prior to time t, and the continuous-time nature of transmissions. However, in a randomized trial, under the null hypothesis of no effect of intervention, the statistic has zero expectation. Thus, hypothesis tests based on this statistic are valid. In the sample size calculations to follow, we size the trial based on the hypothesized true effect of intervention on the infection growth rate. For the discrete-generation approximations considered, this is equal to the effect of intervention on the reproduction number. While the power and sample size calculations presented here are based on hypothesis tests using this statistic, they may be reasonable approximations for other test statistics using the same information. While Type I error is preserved through internal validity, other statistics may have greater power, especially if they avoid any bias in the estimator. In addition, other sampling schemes are possible, including sampling only at time $t+1$ or additionally using serologic sampling to estimate the number of susceptible individuals at either or both time points. We focus on the setting using only virologic testing at the two time points for power calculations as it may be broadly feasible to implement.

2.2 | Epidemic spread assumptions

Both the development of approximate sample size formulae and the simulations that follow depend on certain assumptions about the epidemic process. First of all, we assume that clusters are independent; that is, there is no transmission between clusters. This may be reasonable if clusters are sufficiently geographically distinct.

Secondly, we assume that once a cluster has its initial infections, the pathogen spreads according to a standard susceptible-exposed-infectious-recovered (SEIR) model. In the approximations, we consider only a single discrete time step (equal to one generation interval). We assume that at time t, the proportion of individuals who are infectious is I_t. The reproduction number at time t, R_t, is the mean number of individuals who will be infected in the next time step (move from S to I) for each infectious individual. R_t can be thought of as the result in each community of a common basic reproduction number, R_0, changed as the number of susceptible individuals changed in that community up to time t, with overdispersion parameterized by k. This parameter allows for the epidemic spread to encompass settings ranging from very little variation in transmission across individuals ($k \geq 1$) to a small number of infected individuals being responsible for the vast majority of onward transmission ($k \leq 0.1$), sometimes referred to as "superspreading".[31,32] The number of individuals in the community is assumed to be sufficiently large compared to the number of infectious individuals at time t such that no individual is infected by two infectious individuals in the same generation. More details can be found in Appendix 1.A.

TABLE 1 Parameters used in simulations

Parameter	Values		
R_0	1.5, 2.0		
Overdispersion parameter (k)	0.1, 0.4, 0.7		
Effect size (Reduction in transmission) (Δ)	20%, 40%		
Cluster population (n)	100	1000	10,000
Number sampled per cluster (m)	10, 50, 100	100, 1000	100, 1000
Mean infection prevalence at t ($E[I_t]$)	2%	0.5%	0.5%[a]
Initial infection prevalence (I_0)	1%	0.4%	0.4%

[a]0.45% for $n = 10{,}000$ when $k = 0.1$. Results when $k = 0.1$, $n = 10{,}000$, $R_0 = 1.5$ not included as $E[I_t]$ did not rise above 0.45% for this parameter combination.

In the simulations, we use a continuous-time stochastic SEIR model. We assume that each exposed individual's incubation period is drawn from an exponential distribution with a mean of 5.51 days.[33] We match the simulations done elsewhere, assuming the mean infectious period across individuals is 5 days.[22,23] We use an exponential rather than gamma distribution because its memoryless property allows for interrupted modeling without disturbing in-progress infectious periods; the difference is minor and the exponential distribution has been used elsewhere as an approximation.[34] We assume an approximate negative binomial degree distribution for the network structure of each cluster that is overdispersed by parameter k.[31] We assume a mean degree of 15 within each cluster as has been found in cRCTs for other infectious diseases[22,35] and suggested for respiratory diseases based on large-scale contact surveys.[36] We use a configuration model (CM) algorithm to generate a graph according to this distribution.[37] We then remove unnecessary edges, such as self-loops and multiple edges connecting nodes, as these are irrelevant for this setting, leaving a contact structure that nearly but does not exactly follow a negative binomial degree distribution. When $n = 100$, it is possible for a drawn degree from the negative binomial distribution to be greater than n. For these nodes, their maximum degree by the end of the algorithm is capped at 99. We also assume a fixed initial number of infections to seed the epidemic according to cluster size (see Table 1). Note that because the epidemics progress stochastically within each cluster, the R_t and I_t values vary between clusters.

In the approximations, the intervention can affect the reproduction number and the overdispersion parameter in the clusters where it is implemented. In the simulations, the intervention affects the transmission rate, but does not affect the contact structure in the cluster and thus not the overdispersion of contacts. Different assumptions should be used if the intervention primarily affects the distribution of the number of contacts each person has.

2.3 | Approximate sample size formulae

We consider analysis based on the test statistic comparing the means of $\frac{Y_{j,t+1}/m_1}{Y_{j,t}/m_0}$ (the ratio of the proportion infected post-intervention to pre-intervention, which approximates the reproduction number when measured one generation interval apart) in the intervention vs. control arms. We present two results based on approximate distributions of the test statistic: (i) an approximate sample size required (in terms of the number of clusters per arm) when there is full testing within each cluster at times t and $t + 1$; and (ii) an approximate sample size required when there is sampled testing within each cluster at those times. The former is useful for small cluster sizes, where full or near-full testing is feasible. The latter assumes that the number tested is small compared to the total cluster population, although it focuses on the variability due to post-intervention sampling and ignores the variability due to sampling at time t to simplify the approximations and formulae. The latter is thus likely to underestimate the required sample size in many settings, except where the proportion sampled is non-negligible (ie, greater than 10%). The approximations do not directly account for changes in the number susceptible, so both methods are most accurate for short lags between t and $t + 1$. Details of these assumptions can be found in Appendix 1.

Both of these results are based on a Welch's two-sample t-test for the comparison of two means, with unequal variances. For this test, the required sample size (number of clusters) in each arm, N_i, to detect a difference in means of Δ with power $1 - \beta$ at two-sided significance level α solves:[38]

$$N_i = \frac{(\sigma^{2,I} + \sigma^{2,C})(t_{2N_i-2,1-\alpha/2} + t_{2N_i-2,1-\beta})^2}{\Delta^2}, \tag{1}$$

where $\sigma^{2,i}$ is the variance of the observations in intervention arm i, where $i = I$ for the intervention arm and $i = C$ for the control arm, and $t_{DF,\phi}$ is the ϕ-quantile of the t distribution with DF degrees of freedom. We present variance estimates that can be used in this calculation, with derivations presented in Appendix 1.

If the full cluster populations are tested at each time point, then an approximate sample size can be calculated for the difference in the means of $\frac{Y_{j,t+1}}{Y_{j,t}}$ between the intervention arms, where $Y_{j,t}$ is the number of individuals who test positive at time t in cluster j. The t-test can then be used, with effect size $\Delta = R_t^I - R_t^C$ estimated by the difference in means, and variances approximated for each arm i by:

$$\sigma^{2,i} \approx R_t^i \left(1 + \frac{R_t^i}{k^i}\right) \frac{1}{nE[I_{j,t}]}, \tag{2}$$

where R_t^i is the time-varying reproduction number at time t in intervention arm i (again, I for the NPI and C for control), k^i is the overdispersion parameter of transmission in intervention arm i, n is the population (number of individuals) in each cluster, and $E[I_{j,t}]$ is the mean (across clusters) proportion of individuals who are infectious at time t.

If the variance of the proportion of individuals who are infectious at time t across clusters, $Var[I_{j,t}]$, can be estimated as well, then the calculation should use variances approximated by:

$$\sigma^{2,i} \approx R_t^i \left(1 + \frac{R_t^i}{k^i}\right) \left(\frac{1}{nE[I_{j,t}]} + \frac{Var[I_{j,t}]}{nE[I_{j,t}]^3}\right). \tag{3}$$

When a small proportion of the population of each cluster is tested at each time point instead (specifically, m_0 individuals per cluster at time t and m_1 individuals per cluster at time $t+1$), the effect size $\Delta = R_t^I - R_t^C$ can be approximated by the difference in the means of $\frac{Y_{j,t+1}/m_1}{Y_{j,t}/m_0}$. The variances can be approximated by:

$$\sigma^{2,i} \approx \frac{R_t^i}{m_1} \left\{ \left[1 + \frac{m_1 - 1}{n}\left(1 + \frac{R_t^i}{k^i}\right)\right] \left[\frac{1}{E[I_{j,t}]}\right] - R_t^i \right\}. \tag{4}$$

Again, if we can estimate the variance of the proportion of individuals who are infectious at time t by $Var[I_{j,t}]$, then the variances can be approximated by:

$$\sigma^{2,i} \approx \frac{R_t^i}{m_1} \left\{ \left[1 + \frac{m_1 - 1}{n}\left(1 + \frac{R_t^i}{k^i}\right)\right] \left[\frac{1}{E[I_{j,t}]} + \frac{Var[I_{j,t}]}{E[I_{j,t}]^3}\right] - R_t^i \right\}. \tag{5}$$

R functions to calculate these values are available at http://www.github.com/jsheen/NPI.

2.4 | Simulation setup

For each parameter combination described in Table 1, we first create a simulation bank of cluster simulations with and without an enacted nonpharmaceutical intervention. The simulated epidemic is described above. Note that while we report R_0 in the table, it is used primarily to progress the epidemic until the point of intervention. For practical purposes, the relevant parameters for the trial are R_t, the time-varying reproduction number, and I_t, the proportion of infectious individuals, at the time of intervention. To get realistic values of these parameters, we use relatively low R_0 values in the context of SARS-CoV-2, which may actually reflect the impact of other interventions and behavior changes. We examine the sensitivity of results to this parameter throughout. The per-contact daily transmission rate, β, is set empirically to give the desired R_0 at the beginning of the simulation, or can be set to give the desired R_t and I_t at the time of intervention. For ease of comparison, we report the reproduction numbers used for each result rather than the transmission rate itself.

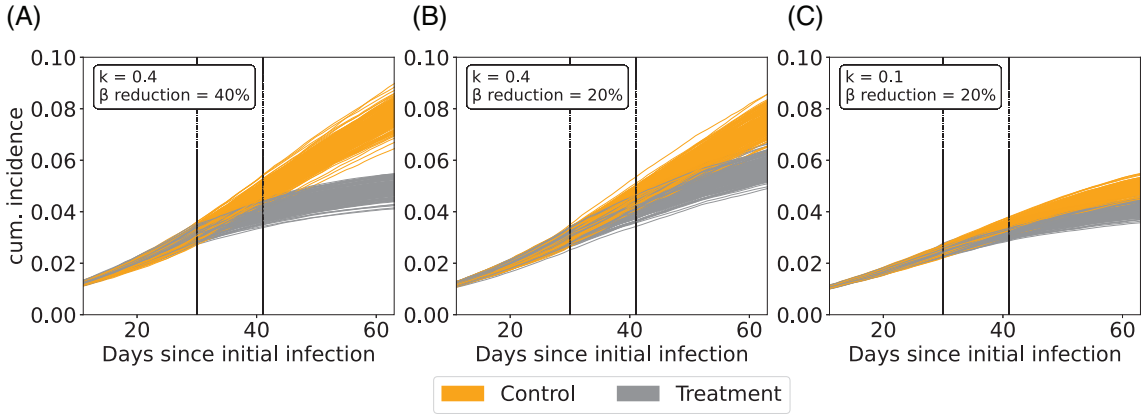

FIGURE 1 Simulated trajectories of a randomized control trial of a nonpharmaceutical intervention. Each arm has 100 clusters, with average basic reproduction number under control $R_0 = 1.5$, overdispersion parameter $k = 0.4$ (A,B) or $k = 0.1$ (C, indicating more overdispersed transmission), and cluster size $n = 1000$. On day 30, the intervention begins, which reduces the transmission rate, β, by 40% (A) or 20% (B,C). The dashed lines represent the day of intervention (t) and the day of sampling, one generation interval after intervention ($t + 1$)

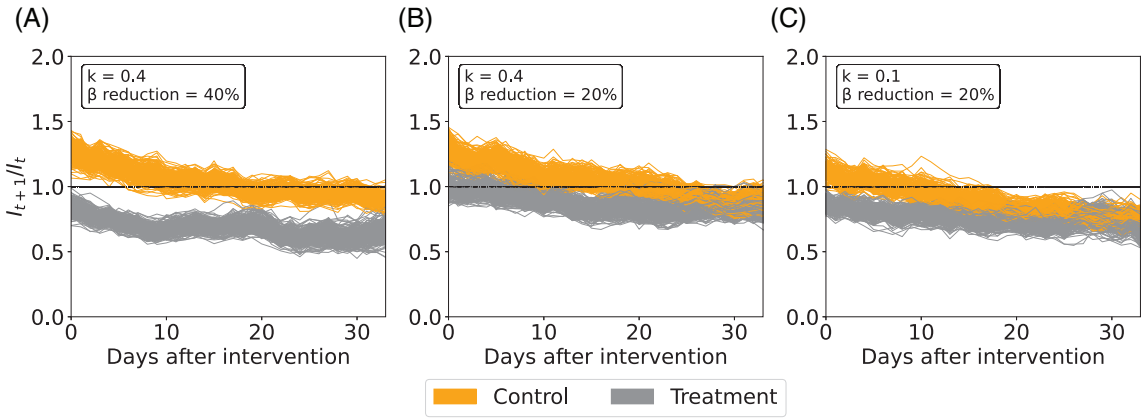

FIGURE 2 The growth rate of infections decreases during the observation period of 33 days after nonpharmaceutical intervention. Simulations shown have average basic reproduction number under control $R_0 = 1.5$, overdispersion parameter $k = 0.4$ (A,B) or $k = 0.1$ (C, indicating more overdispersed transmission), and cluster size $n = 1000$. Intervention (treatment) reduces the transmission rate, β, by 40% (A) or 20% (B,C). The y-axis shows the growth rate at day t, calculated as the ratio of I_{t+1}—the number of infectious individuals one generation interval (11 days) later—to I_t, the number of infectious individuals on day t. The dashed line represents the threshold of a growth rate of 1; that is, each infectious individual at I_t is able to replace themselves at I_{t+1}

To approximately align clusters on epidemic time, time t is defined as the first day when the mean proportion of infectious people across clusters is approximately equal to the target $E[I_t]$. In an epidemic, it would be reasonable to aim to test interventions at similar epidemic time points, but the remaining variability in the number infected allows us to explore the impact of this variance on power. After t is identified, the simulation bank is created by interrupting the simulation for each cluster at time t. At time t, we continue the cluster simulation both with and without an enacted NPI intervention for one generation interval (11 days) and record the number of infectious individuals at this point (denoted time $t + 1$). The generation interval is equal to the ceiling of the sum of the average incubation period and average infectious period.[39,40] At times t and $t + 1$, m individuals are sampled and tested within each cluster. We create 3000 simulations for the simulation bank. Only simulated clusters with at least one infectious individual at time t are kept, so we implicitly assume there is at least one infectious individual in each cluster. Example simulation trajectories of cumulative incidence of cases and of growth rates of infectious individuals are shown in Figures 1 and 2, respectively.

To find the number of clusters in each arm of the trial needed to achieve approximately 80% power when $\alpha = 0.05$, we use a binary search algorithm with a minimum and maximum number of clusters of 1 and 1000, respectively. At each

iteration of the algorithm, 10,000 trial simulations are performed by choosing a given number of clusters from each of the treatment arm and the control arm in the simulation bank. The empirical power is then calculated after sampling m individuals from each cluster and a two-sample Welch's t-test is performed on the quantity $\log\left(\frac{Y_{j,t+1}+1}{Y_{j,t}+1}\right)$. Note that to avoid undefined test statistic values, we add one infected individual at each time point to each cluster. This binary search algorithm implicitly assumes that the relationship between power and number of recruited clusters for the trial monotonically increases. Algorithm pseudocode for the creation of the simulation bank and binary search algorithm for N can be found in Appendix 2, along with comments on the stability of the algorithm (see Figure S1).

We further extend these simulation results in two ways: (i) we provide N when the postintervention testing time point is either two or three generation intervals after t instead of one generation interval in Section 3.2.2; and (ii) we provide simulation results when matching clusters into pairs depending on the number of susceptible individuals at time t, as well as the number of noninfectious individuals at time t, within each cluster, using a matched-pairs t-test in Section 3.2.3. We use a greedy matching algorithm to randomly assign one cluster of each pair to either treatment or control.

We use the *EoN* python package to simulate the epidemic.[41] Code used for simulations is provided at http://www.github.com/jsheen/NPI.

3 | RESULTS

We present results from both the approximate sample size calculations and simulations that target an empirical power of 80% to detect a specified effect size with two-sided significance level $\alpha = 0.05$ for the parameter combinations described in Table 1.

3.1 | Approximate sample size requirements

The required sample size depends on features of the transmission of infection, the sizes of the cluster populations and samples, and the effect size studied. These relationships are illustrated using the approximate variance formulae.

3.1.1 | Approximations under full measurement

In cases where the outcome (infection) is measured in everyone in each cluster, we can use Equation (2) to estimate the variance. The variances (and thus required sample sizes) increase as the reproduction numbers, R_t^i, decrease or as the overdispersion parameter, k^i, decreases. Note that decreasing k^i corresponds to more overdispersion and thus more variability in the change in the number of infections over one generation. In addition, the variance increases as the average proportion of infected individuals per cluster at time t decreases. These relationships are plotted in Figure 3, which shows the number of clusters per arm required for 80% power to detect a reduction in transmission of 40% at significance level $\alpha = 0.05$. Figure 3A shows that, for a given cluster size and expected proportion of infections at enrollment, there is a slight decrease in the required sample size as R_t^C increases and a substantial decrease in the required sample size as k increases (less overdispersion). Figure 3B shows that, for fixed R_t^C and k ($R_t^C = 1.5$ and $k = 0.4$ are shown), the required number of clusters decreases as the expected proportion of the population infected at enrollment, $E[I_t]$, increases and as the cluster size (and thus number sampled) increases. When the expected number of infections per cluster at time t falls below approximately two, the required sample size increases dramatically.

3.1.2 | Approximations accounting for sampling

When sampling within each cluster is accounted for, similar relationships are observed between R_t^C, k, and the required sample size calculated using Equations (1) and (4). Figure 4A shows these relationships for a cluster size of $n = 10{,}000$ and sampling $m = 100$ individuals, with $E[I_t] = 0.005$, and with an effect size (transmission reduction) of 40%. Because of the larger cluster size, the spread of infections is more deterministic, leading to a smaller effect of overdispersion. Figure 4B shows how the effect size affects the required sample size for fixed $k = 0.4$.

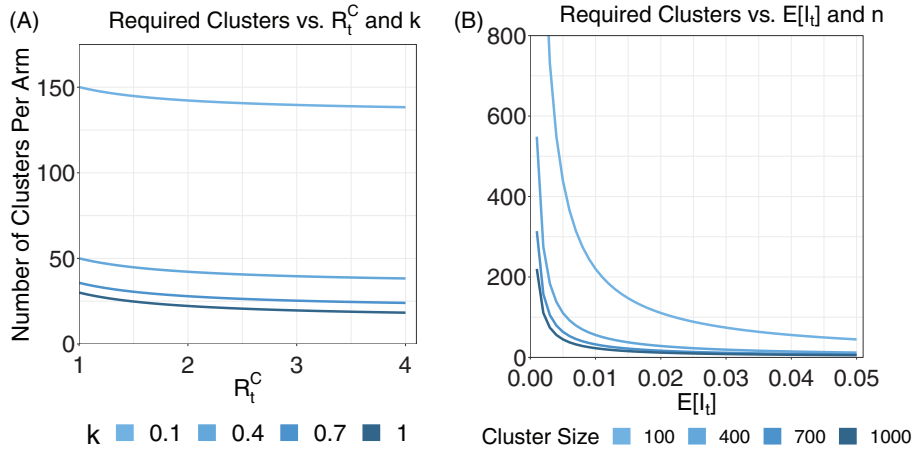

FIGURE 3 When all individuals are tested in each cluster, the required number of clusters per arm decreases as the reproduction number, overdispersion parameter, proportion infected at time of intervention, or cluster size increase. Approximate number of clusters per arm required for 80% power to detect an effect size of 40% at $\alpha = 0.05$ vs (A) reproduction number under control (R_t^C) and overdispersion parameter (k) for fixed cluster size $n = 1000$ and expected proportion of the population infectious at enrollment $E[I_t] = 0.005$ and vs (B) $E[I_t]$ and n for fixed $R_t^C = 1.5$ and $k = 0.4$. Note that lower values of k correspond to more overdispersed transmission

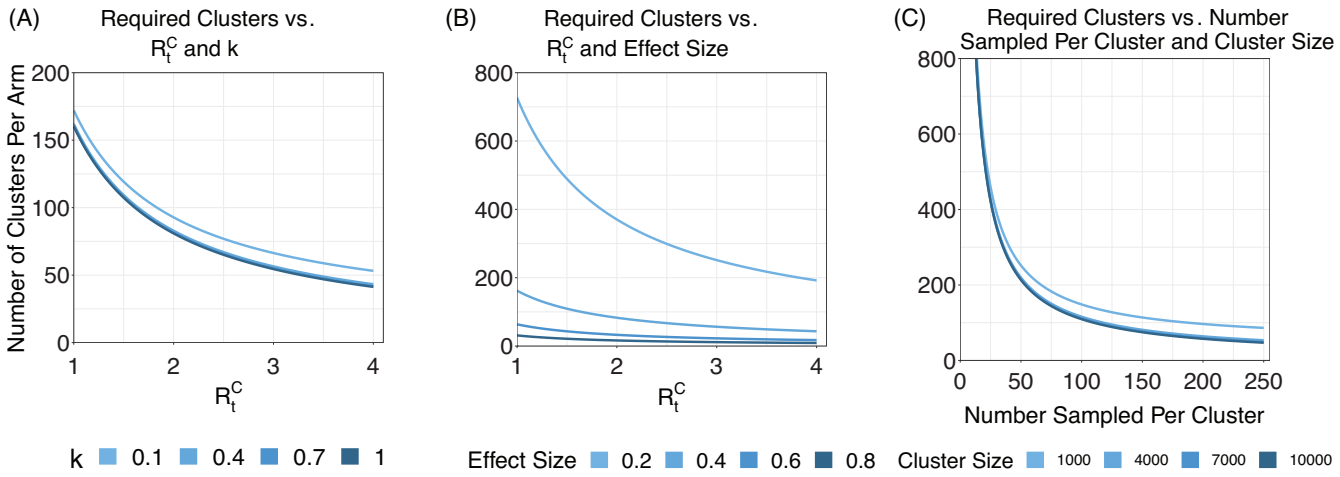

FIGURE 4 When a sample of individuals in each cluster are tested, the approximate required number of clusters per arm decreases as the reproduction number, overdispersion parameter, effect size, cluster size, or number of individuals sampled per cluster increase. Approximate number of clusters per arm required (as calculated by Equations (1) and (4)) for 80% power to detect an effect size of 40% (A, C) or as specified (B) at $\alpha = 0.05$ vs (A) reproduction number under control (R_t^C) and overdispersion parameter (k) for fixed cluster size $n = 10{,}000$ and number sampled per cluster $m = 100$; vs (B) R_t^C and effect size for fixed $n = 10{,}000$, $m = 100$, and $k = 0.4$; and vs (C) n and m for fixed $R_t^C = 1.5$, and $k = 0.4$. In all panels, the expected proportion of the population infectious at enrollment is $E[I_t] = 0.005$

With this approximation, we can also examine the relationship between the number of individuals sampled per cluster, the cluster size, and the required sample size. Figure 4C illustrates these relationships when $R_t^C = 1.5$, $k = 0.4$, the reduction in transmission due to intervention is 40%, and $E[I_t] = 0.005$. Note that these approximations ignore any finite sample corrections. When the number sampled per cluster is a large proportion of the cluster size (ie, $1 - \frac{m_1}{n}$ is meaningfully less than 1), this difference is likely to be meaningful.[42] For reference, the approximate sample size if the full cluster is tested with these parameters ranges from six clusters per arm if $n = 10{,}000$ to 45 clusters per arm if $n = 1000$, which represent (approximately) the minimum number of clusters for less-than-complete sampling.

There are two key effects of an increase in cluster size, holding all other parameters fixed: (i) stochastic effects in epidemic spread are less pronounced, leading to more similar epidemic trajectories across clusters; and (ii) the number

of individuals sampled represents a smaller proportion of the cluster population. The former effect tends to decrease the variance of the test statistic, while the latter effect tends to increase the variance of the test statistic. Thus, it is difficult to describe a general rule governing the relationship between cluster size and required sample size. In the approximations shown in 3C, the latter effect is ignored, so the estimated required number of clusters decreases as cluster size increases.

As the number of individuals sampled per cluster increases, there is a reduction in the required number of clusters per arm. However, this exhibits diminishing returns as it increases, indicating an eventual tradeoff between the number of clusters required per arm and the total number of samples required per arm, as is common in cRCTs.[21] This figure likely underestimates the value of increased testing per cluster, especially for relatively large fractions sampled, as it ignores finite population corrections.

3.2 | Sample size requirements from simulations

Estimated required sample sizes to get the desired empirical power are calculated in simulations as well. A full set of results from these simulations are shown in Supplementary Tables S1 to S17. Here, we focus on the relationships between the key parameters and the required sample size.

3.2.1 | Sample sizes sampling one generation after intervention

Similar to the approximation results, simulation results demonstrate that, in general, the required sample size decreases as the overdispersion parameter increases, the initial reproduction number under control R_0^C decreases, effect size increases, and the number of individuals sampled per cluster increases (Figure 5).

Figure 5A shows that as overdispersion decreases (k increases), the required sample size decreases. Moreover, as shown in Tables S1 to S3, as k increases, R_0^C at time of intervention generally increases as well, even for fixed R_0^C at the start of the simulated outbreak. Because of this, the relationship between k and R_0^C and the required sample size is even more pronounced in the simulation results. Figure 5B illustrates that the required number of clusters will generally decrease as a greater percentage of the cluster is sampled. This indicates that if testing is easy to conduct, the required number of clusters for a trial can be reduced by increasing the sampling within each cluster. Conversely, if the total number of individuals to be sampled in the trial is fixed (ie, limited number of tests available), and the number sampled from each cluster and number of clusters required are allowed to vary, Figure 5C illustrates that it is more efficient to sample fewer individuals from a greater number of communities than it is to sample more individuals from a smaller number of communities. This relationship is less clear for small clusters ($n = 100$) where nearly full sampling can occur.

For clusters of size $n = 100$ or 1000, the day of NPI intervention for some parameter sets occurred less than four weeks after the start of the epidemic (when $I_t = 2\%$ and 0.5% respectively)—but interventions may not always be able to be implemented this quickly. To account for longer delays between the start of the epidemic and day of intervention, we further extend our results by reporting the sample sizes when the day of intervention is one month after the first day of infection in Tables S4 and S5. We also investigate the effect of a different mean number of contacts on the simulation results. Figure S3 shows a sensitivity analysis, indicating relatively small differences in the required sample size as this parameter changes unless cluster size is small compared to the average number of contacts.

3.2.2 | Sample sizes with greater lags after intervention

The formulae and results described above all tested the effect of intervention after one generation interval. Discretizing on this time scale will provide an estimate that approximates the change in the reproduction number, R_t, although occurrences of secondary infections prior to the full generation interval will bias estimates of transmission upwards, and occurrences of primary infections occurring after the full generation interval will bias estimates downwards.[43] Furthermore, because the direct effects continue and indirect effects may increase on short time scales, the effect size in cRCTs in epidemics can increase over time.[22,23] Eventually, however, the exhaustion of susceptible individuals will lead to a reduction in the effect size as incidence rates become more similar between intervention and control clusters. We explore the effects of the time interval used in our simulations by increasing the lag between intervention and evaluation. The

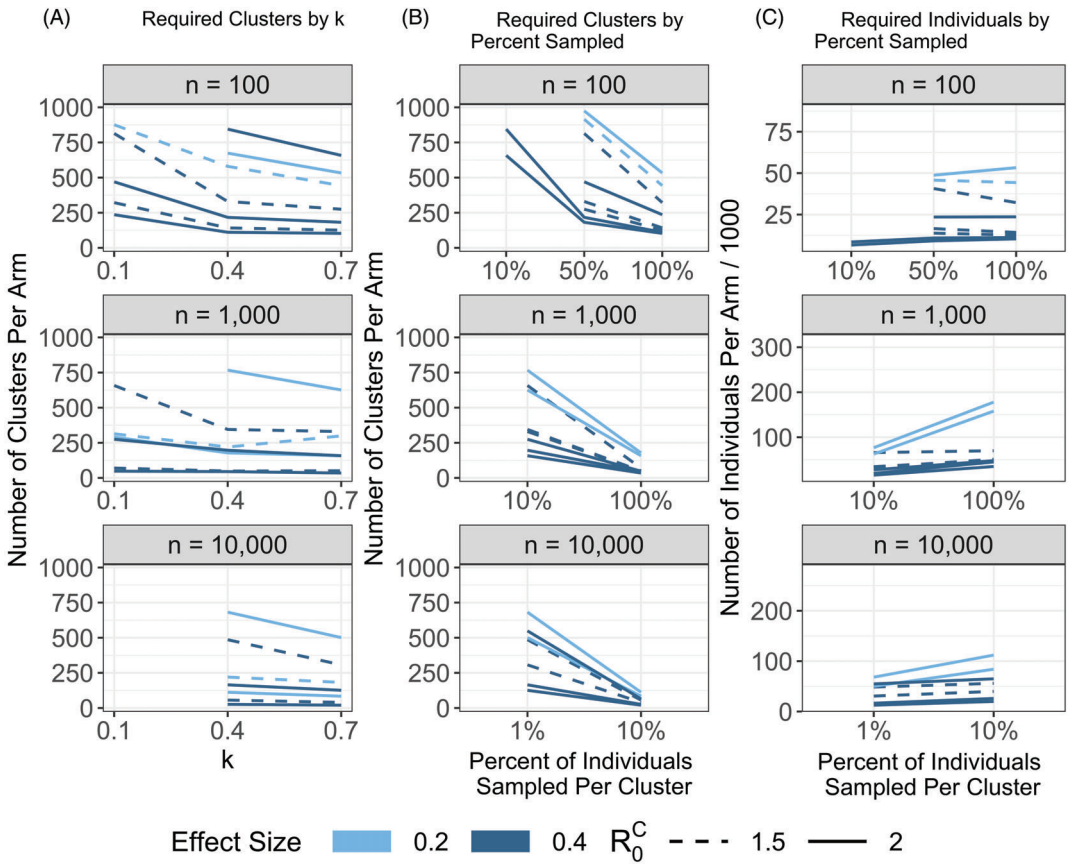

FIGURE 5 The required number of clusters per arm to achieve a desired empirical power in simulations depends on the overdispersion parameter, reproduction number, effect size, cluster size, and percent of individuals in each cluster who are sampled. Number of clusters per arm required (A,B) and number of individuals per arm required (C) to achieve 80% empirical power with a significance level of $\alpha = 0.05$ in 10,000 simulated trials vs overdispersion parameter k (A) or percent of individuals sampled per cluster (B,C). Effect size and basic reproduction number at the start of the outbreak, R_0^C, are varied within each panel. In A, each percent of individuals sampled has a unique line, leading to larger differences even when effect size is held fixed. Points were excluded when the sample size was greater than 1000 clusters per arm. Parameter combinations within each subplot that solely differ in percent of individuals sampled (A) or overdispersion parameter k (B,C) will have the same color and line type

approximations are not well-suited to assess these sample sizes as their assumptions become less reasonable over longer time scales.

Figure 6 shows a drastic increase in power—or decrease in the required sample size—if sampling occurs two generation intervals after intervention compared to sampling one generation interval after intervention.

However, these results change when we increase the lag further, to sampling three generation intervals after intervention. There is generally a more modest decrease in the required sample size for extending from two to three generation intervals than from one to two generation intervals. Second, for small clusters ($n = 100$), increasing the lag between intervention and day of sampling can even increase the required sample size for certain parameter combinations. As the epidemic progresses, there is eventually a point where the depletion of susceptible individuals leads to a decline in power from increasing the lag, although this time point will depend on the precise combination of parameters. We also find this decrease in power occurs more often when there are fewer people sampled from each cluster. Full results are shown in Tables S6 to S11.

3.2.3 | Sample sizes after matching clusters

A common approach to increasing power in cRCTs is to match or stratify clusters on baseline covariates to increase balance and reduce variability.[21,44] In this case, clusters can be matched on the number of susceptible individuals at the

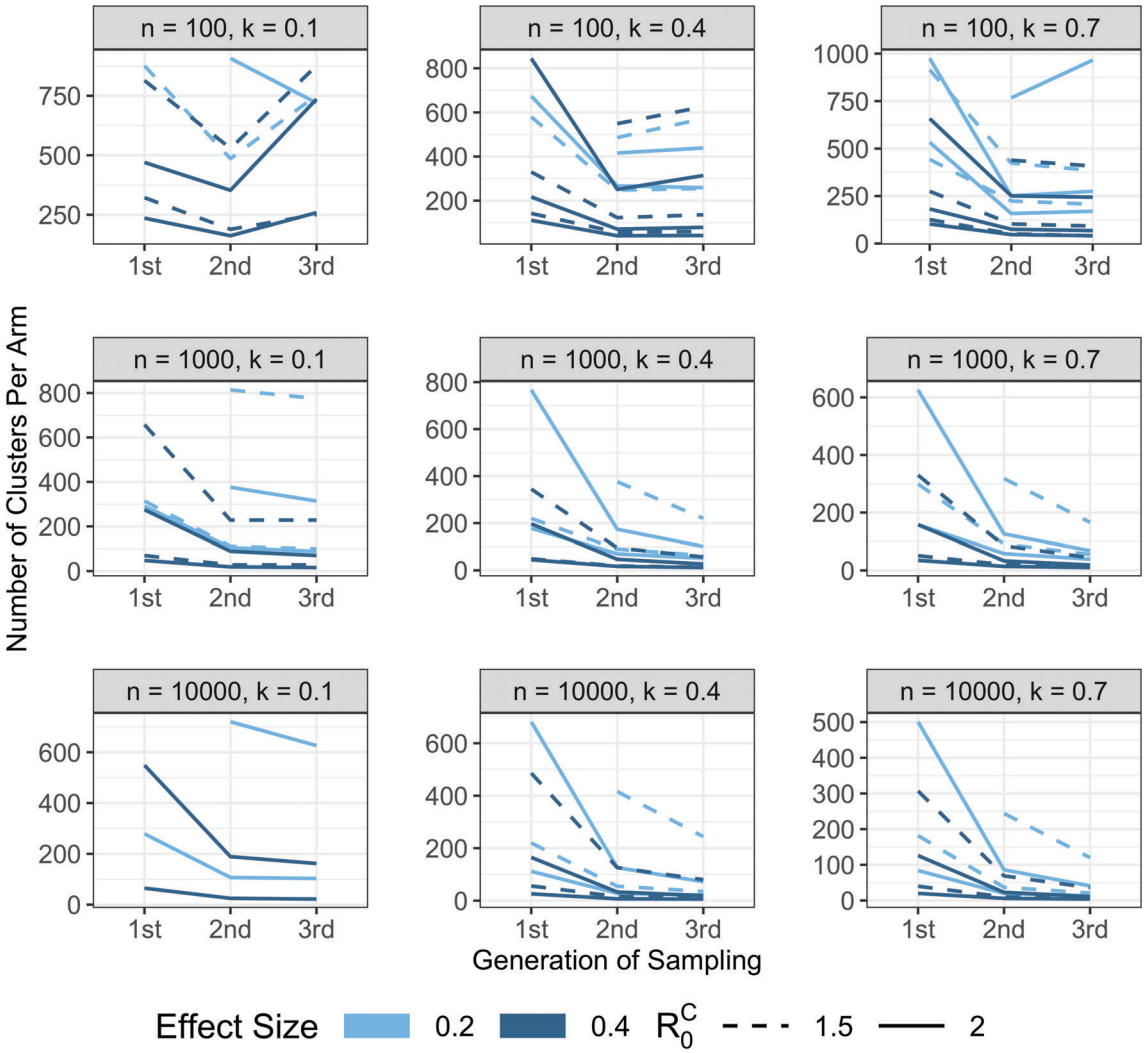

FIGURE 6 Required sample size generally decreases as time of sampling after intervention increases, except when the number of true infections is low three generation intervals after intervention for smaller cluster sizes. Rows correspond to cluster size, n, and columns correspond to overdispersion parameter, k. Effect size, basic reproduction number at the start of the outbreak, R_0^C, and number of individuals sampled per cluster, m, are varied within each panel. On average, sampling two or three generation intervals after intervention required a sample solely 35% and 33% as large, respectively, as after one generation interval. Points were excluded when the sample size was greater than 1000 clusters per arm. Parameter combinations within each subplot that solely differ in the number of individuals sampled per cluster will have the same color and line type

time of intervention (assuming the availability of serological tests) or on the number uninfected at the time of intervention (using the prevalence data collected at time t). This matches clusters on $I_{j,t}$, reducing the effect of the variability in that parameter on the variance of the effect estimate.

We assess the effect of this matching in the simulations, using a matched pairs t-test where clusters are matched on the number of susceptible individuals at the time of intervention (see Tables S12 to S14). We find evidence of modest benefits from matching when cluster sizes are large. The required sample size generally decreases depending on the parameter combination used; however, on average, the change in required number of clusters for $n = 1000$ and $n = 10{,}000$ was modest: 4% and 2% reductions in required sample size, respectively.

For clusters of size 100, the required sample size improved for all parameter combinations where we were able to solve for the number of clusters. The average reduction in required number of clusters was 15%. Because of the smaller cluster size, matching on the number of sampled susceptible individuals may be more reliable and thus more informative than matching with larger cluster sizes. More specifically, the number of sampled susceptible individuals in a smaller cluster

compared to a larger cluster may give more information about the number of cases in the following generations because it has a larger impact on the transmission dynamics.

We also assess the effect of matching when clusters are matched on the number of noninfectious individuals at the time of intervention (see Tables S15 to S17). We find larger benefits from matching on this parameter: the decreases in the required number of clusters for $n = 100$, 1000, and 10,000 were 15%, 12%, and 29% on average, respectively. In general, for each cluster size, the benefits were greater for smaller sample sizes.

When comparing matching clusters on number of susceptible individuals to the number of noninfectious individuals at time of intervention, we find that when $n = 100$ and 1000, when not all individuals were tested, matching on the number of noninfectious individuals reduced the sample size compared to matching on the number of susceptible individuals (by 24% and 28% for $n = 100$ and 1000, respectively). When all individuals in each cluster were tested, matching on number of noninfectious individuals increased the sample size compared to matching on susceptible individuals (by 28% and 7% for $n = 100$ and 1000, respectively). When $n = 10{,}000$, matching on the number of noninfectious individuals reduced the sample size compared to matching on susceptible individuals by an average of 27%. Thus, using serology testing to capture the number of susceptible individuals at time t may be more useful when all participants are tested; when clusters are not fully sampled, in contrast, the number of sampled noninfectious individuals (equivalently, the number of sampled infectious individuals) at time t is more useful for matching.

Importantly, we assess the benefits of matching when time of intervention occurs according to a prespecified $E[I_t]$ in Table 1. Time of intervention will change the number of individuals of each condition (susceptible, exposed, infectious, and recovered), thus affecting the benefits of matching. The effect of time of intervention on the benefits of matching is not explored.

Gains in power may also be achieved using stratification or adjustment by measured covariates related to the transmission of infections within each community. For cluster-level covariates, analysis would then proceed by regression analysis. The reduction in sample size depends on the correlation between the covariate and the outcome, and we refer readers elsewhere to determine the expected reduction that could be applied to calculated sample sizes.[21,38]

3.3 | Comparison of approximation and simulation sample sizes

The approximation and simulation approaches generate different ways of considering the issue of required sample sizes, with the former illustrating the impact of different contextual features (overdispersion, cluster size) and the latter accounting for more of the variability in the epidemic spread process. Direct comparison between the sample size requirements derived from approximation formulae and from simulations are difficult, primarily because of the progression of the epidemic up to the time of intervention in the simulations. As previously susceptible edges of the contact network have already been infected in some clusters, the distribution of infections in the next generation varies from cluster to cluster. Similar effects may occur with the contact networks, where overdispersion falls over time as the most highly-connected individuals are most likely to have already been infected.[45,46]

Despite this, the approximations and simulations shown here generally provide required sample sizes that are comparable in magnitude for many settings. When all individuals are tested in each cluster, the approximation performs very well and closely matches the results of simulations. When the variance of infections at time t across clusters is ignored, this is likely to result in some underestimation of the variance and thus of the required sample size; the same is true because variation in the R_t^C values across clusters is ignored by the approximation.

When only a sample of individuals is tested in each cluster, the approximations diverge further from the simulations. This can occur for a variety of reasons in addition to the parameter mismatch described above: the approximation does not account for sampling prior to the intervention and it does not fully account for the variance in the number of infectious individuals at the time of intervention, $Var[I_t]$, and the variance in the actual reproduction number, R_t, across clusters at the time of intervention. Figure S2 shows the wide spread in the number of infectious individuals per cluster, which reflects the overdispersion of the contact structure of the simulated networks. When the outbreak parameters can be well-estimated a priori, then, the simulations account for certain heterogeneities that the approximations do not. Because of this, approximations can be used to get an estimate of the feasibility of a trial but should not be the only consideration in powering a trial.

Moreover, due to the stochasticity, some clusters may have zero identified cases at time $t + 1$. This is accounted for in the analysis of the simulations by adding one case to each time point, which may introduce some bias. Additionally, the test statistic inherently is based on a fixed, discrete generation interval, which is not the case in the simulation or in

reality. This may lead to the test statistic not estimating the reduction in R_t consistently; however, since the null remains the same in either case, the hypothesis tests remain valid.

4 | PROPOSED APPROACH FOR SAMPLE SIZE CALCULATION

Because of their different required assumptions, parameters, and approximations, both the approximation formulae and the simulations should be considered with a range of plausible estimates for the parameters when designing a trial. We propose that investigators considering a cRCT for an NPI estimate a required sample size using the following procedure:

1. Calculate approximations to the number of clusters required per arm using the two approximation formulae presented here for likely parameters in their setting. If these are well beyond the point of feasibility for the study, the desired power may not be achievable. If any parameters can be manipulated (eg, by only enrolling high-incidence clusters or changing the cluster size for implementation), consider other combinations that may reduce the required sample size.
2. Consult the simulation results in Tables S1 to S17 to find the parameter combination most similar or combinations which bound the likely parameters for the setting of interest. Extrapolate an estimated sample size from these results and again evaluate the feasibility of this sample size.
3. Conduct a simulation study using the best estimates for the transmission dynamics of the setting of interest using the sample size estimated in steps 1 to 2 and the planned analysis method. Determine if the empirical power from this simulation study approximately matches the desired power.

To illustrate this process, we consider a recent cRCT investigating the effects of community-level mask promotion in Bangladesh on masking and symptomatic seroprevalence of COVID-19.[47] In this study, 600 villages were randomized into two arms: an intervention arm that received various mask promotion strategies including free mask provision and role modeling by community leaders, and a control arm that had no specific promotion. An average of 570 adults live in each village, and nearly all reported their follow-up symptom status. Of 27,000 reporting COVID-like symptoms, 11,000 consented to serologic testing. While this study used symptomatic seroprevalence over eight weeks as the primary outcome, we illustrate our proposed method for a primary outcome of virologic test-confirmed infection and a shorter duration, that is, of less than two weeks. This may result in increased effect sizes, as the authors note that symptomatic seroprevalence may identify individuals who were infected prior to the trial onset, and had symptoms for non-COVID reasons during the trial.[47]

We consider a similar trial with $N_0 = N_1 = 300$ clusters per arm and $n = 570$ individuals per cluster. Due to limited testing in these villages, case counts and reproduction number estimates prior to the study were not reliable and not used by the original study authors. However, during this study, reported case counts were high and rising.[47] This indicates that assuming a reproduction number over 1 (eg, 1.2) in the control arm (R_t^C) is reasonable, as is an initial proportion infectious ($E[I_{j,t}]$) of 0.5%, calculated from the final estimated symptomatic seroprevalence (approximately 1.25% accounting for some omitted individuals), doubled to account for asymptomatic infections, and spread across approximately 5 generation intervals covered by the trial duration. Lacking specific knowledge of epidemiologic conditions in the area, we choose a moderate overdispersion parameter of $k^C = k^I = 0.4$, perhaps reflecting existing mitigation measures that prohibit large gatherings. While many assumptions go into these values, they may be reasonable a priori estimates for key parameters.

Using the approximation formulae with these parameters and the variance from Equation (2), if all individuals in each cluster are tested, 80% power to detect a 40% reduction in R_t would require 83 clusters in each arm. If 100 individuals in each cluster are tested, the approximate sample size, calculated using Equation (4), rises to 212 clusters per arm. This results in approximately 40,000 total tests. Testing 50 individuals per cluster instead raises the required clusters per arm to 342 and reduces the total tests required to approximately 34,000. Note that, unlike the serologic testing used in the study, the virologic testing considered in these methods would not require a blood sample and thus might be cheaper, more acceptable, and easier to administer to more individuals per cluster. An investigator considering such a study could weigh the benefits of the outcome and approach used by Abaluck et al against the benefits of a shorter-duration study using virologic testing that would require more tests.

We then consider the simulation results in Tables S1 and S2, using the same parameters. For clusters of sizes 100 and 1000 and testing 100 individuals per cluster, the required numbers of clusters per arm are 111 and 345, respectively. Because the simulations do not guarantee R_t^C at time of enrollment, only R_0, these sample sizes come from the closest approximations to $R_t^C = 1.2$ of 1.33 and 1.42 for clusters of size 100 and 1000, respectively, keeping overdispersion and

effect size the same and testing 100 individuals in each cluster. For our intermediate size of 570, then, 300 clusters per arm is likely reasonable. The higher cluster size number can be considered a conservative estimate of the sample size required if cluster sizes vary up to a maximum of 1000. Both the simulations and approximations can be used to assess the sensitivity of these sample sizes to the assumed parameters: an increase in overdispersion from $k = 0.4$ to $k = 0.1$ would require about 1.5 to 2 times the sample size, while an increase in the reproduction number under control from $R_t^C = 1.2$ to $R_t^C = 1.5$ would reduce the requirement by 10%-20%.

These approximations and estimates suggest that such a trial would be adequately powered to detect a 40% reduction in R_t using 200 to 400 clusters per arm and testing 100 or fewer individuals per cluster. Since the investigators clearly had the resources for a trial of that magnitude, planning could proceed. More information on village sizes and contact patterns could be programmed into the simulation, allowing for a more precise power calculation. While the conducted study used a primary outcome (symptomatic seropositivity) different from the one proposed here, this calculation also suggests that it was adequately powered for a meaningful reduction in transmission, although the estimated effect size on that scale was not reported.[47]

Moreover, this simulation could be used to assess the power of various estimation methods and primary outcomes; for example, it could compare results using a virologic outcome to those using a serologic or symptomatic serologic outcome. And pilot studies and detailed information from the trial setting could be incorporated to further improve the accuracy of these estimates. Future studies should consider the implications of the sample size methods discussed here, as well as the results and experiences of completed studies such as Abaluck et al.[47]

5 | DISCUSSION

To determine whether cRCTs are a practical tool to test the impact of NPIs in epidemic settings, we developed two approximate sample size formulae. We compared these results to simulated outbreaks and developed a simulation bank that can be used to further refine estimates of the required sample size for cRCTs. The simulations can be adapted to specific settings to provide more precise sample size estimation and improve the design of cRCTs.

As an example, we have shown that for settings with communities of 10,000 people, R_t of 1.5 in the absence of intervention, and k of 0.4, 80% power to detect a reduction in R_t of 40% due to intervention can be achieved with approximately 220 total clusters (22,000 sampled individuals) in the trial. While this is certainly a large sample size, cRCTs of that order of magnitude have been conducted for large-scale policy interventions,[47,48] and individual RCTs with thousands or tens of thousands of participants have occurred to evaluate NPIs and vaccines during this pandemic.[13,14,49,50] In particular, if large-scale random testing of individuals is occurring that can be incorporated into the study, sampling large numbers of individuals per cluster may be feasible. As R_t increases, overdispersion decreases, or the effect size increases, this sample size can be reduced while maintaining power.

For communities of 100 people who are all tested (eg, a workplace), with the same transmission parameters, 80% power to detect a reduction in R_t of 40% due to intervention can also be achieved with approximately 220 total clusters (22,000 sampled individuals) in the trial. If the overdispersion were more extreme, the required sample size would increase drastically; for example, for $k = 0.1$, these trials would require approximately 720 clusters, or 72,000 sampled individuals. If, in addition, 80% power to detect a reduction in R_t of 20% is desired, the sample sizes increases dramatically, requiring approximately 3500 clusters or 350,000 sampled individuals. These may not be feasible for many settings, saving researchers from conducting an underpowered study.

These results use a simple estimator based on virologic testing at two time points, one before and one after the intervention.[17] More work is needed to determine the properties of this estimator, especially in cases where the epidemic fades out in certain clusters and as the lag between the two testing times changes, as both may bias estimates away from the true transmission reduction. Other estimators may have more desirable properties in estimating specific estimands of interest or in precision. We focus here only on the power of hypothesis tests, the ability to reject the null of no intervention effect for interventions that reduce the reproduction number by a specific amount.

The approximation formulae are limited by the fact that they ignore the variability in the number of active infections at time t, and the method that accounts for sampling ignores finite population corrections and sampling variability at time t. In addition, these methods ignore the variability in previous infections and the effect those have on future spread on the network, which may serve to overstate the variability in dispersion, especially for small cluster sizes or late time points in the epidemic.[45,46] The simulations require a specific data-generating process and assume that the epidemic unfolds according to the SEIR model up to the point of intervention. It also assumes that overdispersion in transmission is

caused by the contact network structure, which may ignore biological mechanisms of overdispersion.[45] The simultaneous implementation of other NPIs that affect transmission may affect the validity of this model, and more precise modeling should be used to get better sample size estimates for specific settings. In addition, the possibility of imported infections to the trial communities is ignored here, as well as the effect of the intervention on reducing those.

In addition to changing test statistics, other methods may be used to reduce the sample size required to achieve the desired power. Increasing the time between intervention and evaluation can increase the power to some extent, although this may make the trial more logistically challenging and make interpretation of effect estimates more challenging. Matching and stratification on cluster-level variables may reduce the variability of results and improve power, again changing the interpretation of estimates.[21,44,51] If the number of clusters are limited but a large number of tests are available, repeated cross-sectional testing may also improve power; this design also allows investigation of time-varying effects.[21]

Further work is required to improve the sizing of large-scale cRCTs in outbreak settings. In particular, analysis of data on the variability of infections at different time points during outbreaks among relevant clusters would enable validation of the assumptions made in these approaches. This data validation would improve both the closed-form approximations used here and the validity of simulations conducted to assess power and sample size. In addition, understanding the variability in R_t across clusters, and covariates or data that can be used to predict R_t in a given cluster, will enable a better understanding of the mechanism of effects of NPIs on transmission. This improved understanding of the estimand will improve sample size and power calculations and potentially point to more efficient estimators.

Randomized trials are key to achieving valid hypothesis tests of the effect of interventions in infectious disease outbreaks. Cluster randomized trials can be used to test the total effect of nonpharmaceutical interventions by comparing the infection trajectory in intervention communities to that in control communities. This analysis demonstrates that in some cases, reasonable power to detect meaningful effect sizes can be achieved for such trials, and it provides investigators with tools to estimate the sample size required.

ACKNOWLEDGEMENT

C. Jessica E. Metcalf and Johannes Haushofer were supported by the Princeton Catalysis Initiative.

DATA AVAILABILITY STATEMENT

The simulated data, code, and results that support the findings of this study are openly available in GitHub at http://www.github.com/jsheen/NPI.

ORCID

Justin K. Sheen https://orcid.org/0000-0001-7842-1364
Lee Kennedy-Shaffer https://orcid.org/0000-0001-7604-3638

REFERENCES

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;20(5):533-534.
2. Flaxman S, Mishra S, Gandy A, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*. 2020;584(7820):257-261.
3. Hsiang S, Allen D, Annan-Phan S, et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*. 2020;584(7820):262-267.
4. Li Y, Campbell H, Kulkarni D, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. *Lancet Infect Dis*. 2021;21(2):193-202.
5. Chang S, Pierson E, Koh PW, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*. 2021;589(7840):82-87.
6. Ruktanonchai NW, Floyd J, Lai S, et al. Assessing the impact of coordinated COVID-19 exit strategies across Europe. *Science*. 2020;369(6510):1465-1470.
7. Vlachos J, Hertegård E, Svaleryd HB. The effects of school closures on SARS-CoV-2 among parents and teachers. *Proc Natl Acad Sci*. 2021;118(9):e2020834118.
8. Brauner JM, Mindermann S, Sharma M, et al. Inferring the effectiveness of government interventions against COVID-19. *Science*. 2021;371(6531):eabd9338.
9. Haug N, Geyrhofer L, Londei A, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav*. 2020;4(12):1303-1312.
10. Liu Y, Morgenstern C, Kelly J, Lowe R, Jit M. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Med*. 2021;19(1):1-12.

11. Courtemanche CJ, Le AH, Yelowitz A, Zimmer R. School reopenings, mobility, and COVID-19 spread: evidence from Texas. Working paper 28753, NBER Working Paper Series; 2021.
12. Accorsi EK, Qiu X, Rumpler E, et al. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. *Eur J Epidemiol*. 2021;1-18.
13. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New Engl J Med*. 2020;383(27):2603-2615.
14. Baden LR, El Sahly HM, Essink B, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New Engl J Med*. 2021;384(5):403-416.
15. Halloran ME, Longini IM, Struchiner CJ. *Design and Analysis of Vaccine Studies*. New York, NY: Springer; 2010.
16. Lipsitch M, Dean NE. Understanding COVID-19 vaccine efficacy. *Science*. 2020;370(6518):763-765.
17. Haushofer J, Metcalf CJE. Which interventions work best in a pandemic? *Science*. 2020;368(6495):1063-1065.
18. Weijer C, Hemming K, Hey SP, Lynch HF. Reopening schools safely in the face of COVID-19: can cluster randomized trials help? *Clin Trials*. 2021;18(3):371-376.
19. Haber NA, Wieten SE, Smith ER, Nunan D. Much ado about something: a response to "COVID-19: underpowered randomised trials, or no randomised trials?". *Trials*. 2021;22:780.
20. Bellan SE, Pulliam JR, Pearson CA, et al. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *Lancet Infect Dis*. 2015;15(6):703-710.
21. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. 2nd ed. Boca Raton, FL: CRC Press; 2017.
22. Hitchings MD, Lipsitch M, Wang R, Bellan SE. Competing effects of indirect protection and clustering on the power of cluster-randomized controlled vaccine trials. *Am J Epidemiol*. 2018;187(8):1763-1771.
23. Kennedy-Shaffer L, Lipsitch M. Statistical properties of stepped wedge cluster-randomized trials in infectious disease outbreaks. *Am J Epidemiol*. 2020;189(11):1324-1332.
24. Held L, Hens N, O'Neill P, Wallinga J, eds. *Handbook of Infectious Disease Data Analysis*. Boca Raton, FL: CRC Press; 2019.
25. Larremore DB, Fosdick BK, Bubar KM, et al. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *eLife*. 2021;10:e64206.
26. Paltiel AD, Zheng A, Walensky RP. Assessment of SARS-CoV-2 screening strategies to permit the safe reopening of college campuses in the United States. *JAMA Netw Open*. 2020;3(7):e2016818.
27. Stein-Zamir C, Abramson N, Shoob H, et al. A large COVID-19 outbreak in a high school 10 days after schools' reopening, Israel, May 2020. *Eurosurveill*. 2020;25(29):2001352.
28. Bigelow BF, Tang O, Barshick B, et al. Outcomes of universal COVID-19 testing following detection of incident cases in 11 long-term care facilities. *JAMA Intern Med*. 2021;181(1):127-129.
29. Mack CD, DiFiori J, Tai CG, et al. SARS-CoV-2 transmission risk among National Basketball Association players, staff, and vendors exposed to individuals with positive test results after COVID-19 recovery during the 2020 regular and postseason. *JAMA Intern Med*. 2021;181(7):960-966.
30. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc B*. 2007;274:599-604.
31. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005;438(7066):355-359.
32. Althouse BM, Wenger EA, Miller JC, et al. Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2; 2020. arXiv Preprint 2020. arXiv:2005.13689.
33. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med*. 2020;172(9):577-583.
34. Kiss IZ, Miller JC, Simon PL. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Cham, Switzerland: Springer; 2017.
35. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *Lancet*. 2017;389(10068):505-518.
36. Klepac P, Kucharski AJ, Conlan AJK, et al. Contacts in context: large-scale setting-specific social mixing matrices from the BBC pandemic project; 2020. medRxiv Preprint. doi:10.1101/2020.02.16.20023754
37. Newman ME. The structure and function of complex networks. *SIAM Rev*. 2003;45(2):167-256.
38. Ahn C, Heo M, Zhang S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. Boca Raton, FL: CRC Press; 2015.
39. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press; 1992.
40. Fine PE. The interval between successive cases of an infectious disease. *Am J Epidemiol*. 2003;158(11):1039-1047.
41. Miller JC, Ting T. EoN (Epidemics on Networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *J Open Source Softw*. 2019;4(44):1731.
42. Lohr SL. *Sampling: Design and Analysis*. 2nd ed. Boston, MA: Brooks/Cole; 2010.
43. Ferrari MJ, Bjøornstad ON, Dobson AP. Estimation and inference of R0 of an infectious pathogen by a removal method. *Math Biosci*. 2005;198(1):14-26.
44. Kennedy-Shaffer L, Hughes MD. Sample size estimation for stratified individual and cluster randomized trials with binary outcomes. *Stat Med*. 2020;39(10):1489-1513.

45. Gomes MGM, Corder RM, King JG, et al. Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold; 2020. medRxiv Preprint. doi:10.1101/2020.04.27.20081893.
46. Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science*. 2020;369:846-849.
47. Abaluck J, Kwong LH, Styczynski A, et al. Impact of community masking on COVID-19: a cluster-randomized trial in Bangladesh. *Science*. 2022;375(6577):eabi9069.
48. Egger D, Haushofer J, Miguel E, Niehaus P, Walker MW. General equilibrium effects of cash transfers: experimental evidence from Kenya. Working paper 26600, NBER Working Paper Series; 2019.
49. Bundgaard H, Bundgaard JS, Raaschou-Pedersen DET, et al. Effectiveness of adding a mask recommendation to other public health measures to prevent SARS-CoV-2 infection in Danish mask wearers. *Ann Intern Med*. 2021;174(3):335-343.
50. Revollo B, Blanco I, Soler P, et al. Same-day SARS-CoV-2 antigen test screening in an indoor mass-gathering live music event: a randomised controlled trial. *Lancet Infect Dis*. 2021;21(10):1365-1372.
51. Benkeser D, Díaz I, Luedtke A, Segal J, Scharfstein D, Rosenblum M. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*. 2021;7(4):1467-1481.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sheen JK, Haushofer J, Metcalf CJE, Kennedy-Shaffer L. The required size of cluster randomized trials of nonpharmaceutical interventions in epidemic settings. *Statistics in Medicine*. 2022;41(13):2466-2482. doi: 10.1002/sim.9365