PLOS COMPUTATIONAL BIOLOGY

METHODS

Detecting and quantifying heterogeneity in susceptibility using contact tracing data

Beth M. Tuschhoff, David A. Kennedy*

Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America

* dak30@psu.edu

OPEN ACCESS

Citation: Tuschhoff BM, Kennedy DA (2024) Detecting and quantifying heterogeneity in susceptibility using contact tracing data. PLoS Comput Biol 20(7): e1012310. https://doi.org/10.1371/journal.pcbi.1012310

Editor: Eric HY Lau, The University of Hong Kong, CHINA

Received: October 6, 2023

Accepted: July 10, 2024

Published: July 29, 2024

Data Availability Statement: All code used to produce and analyze these results is publicly available on a GitHub repository at https://github.com/bmtuschhoff/heterogeneity_in_susceptibility.git.

Funding: DAK and BMT were supported by Institute of General Medical Sciences, National Institutes of Health (R01GM140459) and the UK Biotechnology and Biological Sciences Research Council as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program. DAK was also supported by National Science

Abstract

The presence of heterogeneity in susceptibility, differences between hosts in their likelihood of becoming infected, can fundamentally alter disease dynamics and public health responses, for example, by changing the final epidemic size, the duration of an epidemic, and even the vaccination threshold required to achieve herd immunity. Yet, heterogeneity in susceptibility is notoriously difficult to detect and measure, especially early in an epidemic. Here we develop a method that can be used to detect and estimate heterogeneity in susceptibility given contact by using contact tracing data, which are typically collected early in the course of an outbreak. This approach provides the capability, given sufficient data, to estimate and account for the effects of this heterogeneity before they become apparent during an epidemic. It additionally provides the capability to analyze the wealth of contact tracing data available for previous epidemics and estimate heterogeneity in susceptibility for disease systems in which it has never been estimated previously. The premise of our approach is that highly susceptible individuals become infected more often than less susceptible individuals, and so individuals not infected after appearing in contact networks should be less susceptible than average. This change in susceptibility can be detected and quantified when individuals show up in a second contact network after not being infected in the first. To develop our method, we simulated contact tracing data from artificial populations with known levels of heterogeneity in susceptibility according to underlying discrete or continuous distributions of susceptibilities. We analyzed these data to determine the parameter space under which we are able to detect heterogeneity and the accuracy with which we are able to estimate it. We found that our power to detect heterogeneity increases with larger sample sizes, greater heterogeneity, and intermediate fractions of contacts becoming infected in the discrete case or greater fractions of contacts becoming infected in the continuous case. We also found that we are able to reliably estimate heterogeneity and disease dynamics. Ultimately, this means that contact tracing data alone are sufficient to detect and quantify heterogeneity in susceptibility.

Author summary

Hosts often vary in their likelihood of contracting an infectious disease. This variation is referred to as heterogeneity in susceptibility, and it can have major public health

Foundation grants DEB-1754692 and DEB-2211322. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this article.

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

consequences. However, heterogeneity in susceptibility is notoriously difficult to detect and quantify, and so, it has often been left out of mathematical models and ignored by decision makers. Here, we present a novel method that can be used to detect and quantify heterogeneity in susceptibility using only contact tracing data. The premise is that if heterogeneity is present, the average individual that did not become infected after appearing in a contact network would have lower susceptibility to infection than the average individual that has never appeared in a contact network. By measuring the difference in susceptibility between these two groups of individuals, which we assess with contact tracing data, it is possible to detect and quantify the level of heterogeneity. We demonstrate the application of this method and explore the method's power and accuracy using simulated contact tracing data.

Introduction

At the outset of an epidemic, public health responses depend on estimates of the final epidemic size, the peak number of cases, the timing of the peak, and the herd immunity threshold. Compartmental models such as the susceptible-infected-recovered (SIR) model are commonly used to model infectious disease dynamics and predict outcomes, but there are limitations to this approach [1–4]. Namely, SIR models tend to oversimplify the complexity of disease dynamics, resulting in discrepancies between the model predictions and epidemic data [1]. One of the simplifying assumptions of the standard SIR model is that all host individuals are the same. However, this is often false: individuals can be heterogeneous in many ways [5, 6] including with regard to their likelihood of becoming infected, hereafter referred to as heterogeneity in susceptibility [7].

Heterogeneity in susceptibility can have a large impact on infectious disease dynamics [7–10]. Increased heterogeneity in susceptibility results in a lower peak number of cases, different timing of the peak, smaller final epidemic size, and lower herd immunity threshold [10–12]. As a result, disease control programs [13] and epidemiological models [7, 9, 14, 15] may need to account for heterogeneity in susceptibility if they are to be optimally useful. Accurate early predictions of disease dynamics could give policy makers critical information to make decisions, but heterogeneity in susceptibility is notoriously difficult to measure [16]. Moreover, the effects of heterogeneity in susceptibility are typically small during the earliest phases of epidemics and only become apparent later, making it even more challenging to estimate heterogeneity in susceptibility in real time and account for its effects. It would therefore be useful to develop new methods for quantifying heterogeneity in host susceptibility early in epidemics.

Existing methods to quantify heterogeneity in susceptibility are not adequate for estimation in real time because they rely on using data that are either collected later in epidemics or that typically cannot be collected due to ethical or logistical constraints. Dwyer et al. [7], Ben-Ami et al. [17], and Langwig et al. [9] used laboratory dose-response and field transmission experiments to estimate heterogeneity in susceptibility, but these experimental methods are not feasible for application in real time or for human epidemics in general due to time constraints and ethical concerns. Gomes et al. [15] compared disease incidence across municipalities in several countries to construct Lorenz curves and fit susceptibility risk distributions, but this method requires a substantial amount of data that would not be available early in an epidemic. Smith et al. [18] and Corder et al. [19] used morbidity data to fit models and estimate heterogeneity, but this method cannot be used until later in an epidemic when there is sufficient data to fit curves. Gomes et al. [10] also used curve fitting with mortality data that could be implemented

once at least four months of data were available, but their method is heavily dependent on the underlying model and assumptions. With the recent increased interest in real-time estimation, Anderson et al. [20] developed a method to estimate within-household heterogeneity in susceptibility, but this is not the same as the population-level heterogeneity that drives population-level disease dynamics. Here we develop a novel method to identify and quantify host heterogeneity in susceptibility using contact tracing data, which can be collected early in an epidemic. Contact tracing is often performed to mitigate the spread of pathogens that are otherwise difficult to control [21, 22], and therefore, our method should not require the collection of any data beyond that which would already be collected for other purposes.

Contact tracing typically takes one of two forms: forward and backward. Our method is suitable for use with data from both types. Forward contact tracing attempts to find all the contacts of an infected person to whom the disease could transmit. This is done by identifying infected individuals and all their known contacts. The contacts are then quarantined and monitored for disease. For any contact that becomes infected, the process is repeated with their contacts. Backward contact tracing attempts to identify the contact of an infected person from whom the disease transmitted. In practice, both methods can be employed simultaneously in an effort to maximize the effectiveness of contact tracing efforts [23], and the data on infected individuals and their contacts are typically recorded. When done thoroughly, contact tracing data provide information about the infection status of individuals that have been in contact with an infected individual. As we will explain, when contact tracing data tracks individuals through multiple exposure events, it can be used to quantify heterogeneity in susceptibility given contact through the method that we develop here.

Our method uses the fact that average susceptibility decreases over time in a population with heterogeneity in susceptibility (Fig 1). This is because individuals with high susceptibility are more likely to be infected than individuals with low susceptibility for a given exposure level. Individuals that show up in a second contact tracing network, after not being infected in the first, should therefore have a lower risk of infection than individuals that show up in a network for the first time. In the rest of the paper, we establish our method and analyze its effectiveness for two cases: a population with two discrete susceptibility levels and a population with continuous variation in susceptibility. Notably, the selection of these two cases is arbitrary, and our method is flexible enough that it can be employed for any distribution of heterogeneity in susceptibility.

Methods and results

Our method to detect and quantify heterogeneity in susceptibility exploits the change in average susceptibility over multiple exposure events that would be expected to occur if a population had heterogeneity in susceptibility (Fig 1). Given contact with an infectious individual, individuals with high susceptibility are more likely to be infected than those with low susceptibility. This creates a selection process in which susceptibility should on average decline in a heterogeneous host population following each exposure event. This change in average susceptibility provides a way to identify and estimate the level of heterogeneity early in an epidemic despite the seemingly small effects of heterogeneity at the beginning of epidemics. Notably, no change in average susceptibility should occur in a population that lacks heterogeneity in susceptibility.

This method employs contact tracing data. With contact tracing data, there are multiple contact networks that are each composed of an infected individual and the known contacts the infected individual had during their infectious period. This means each contact network is a set of exposure events where contacts are exposed to a pathogen and have a chance of being infected. In order for our method to work, there must be individuals that show up in at least

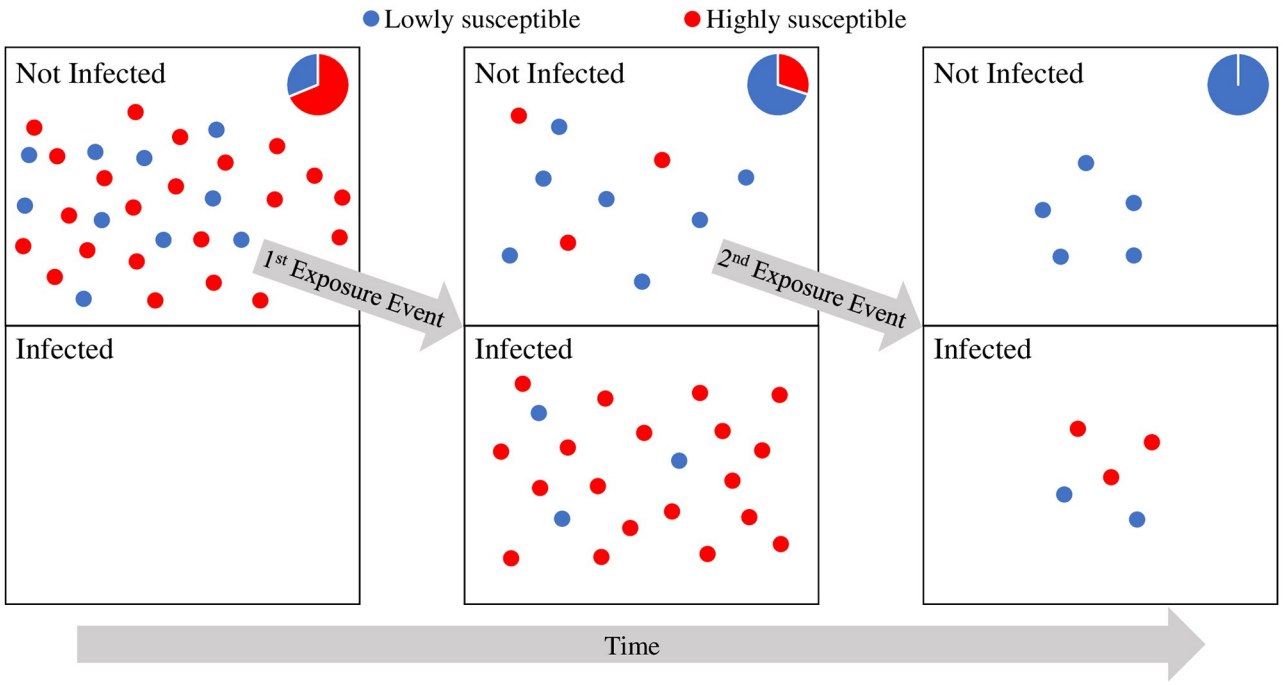

Fig 1. Average susceptibility decreases over exposure events in a heterogeneous population. The figure depicts individuals infected and not infected over two exposure events in a heterogeneous population with more susceptible (red) and less susceptible (blue) individuals. The pie charts show the composition of the not infected population. Average susceptibility in the not infected population decreases after each exposure event as the highly susceptible individuals are infected more frequently than the lowly susceptible individuals. Note that if the population lacked heterogeneity in susceptibility, all individuals would be either red or blue, and thus, susceptibility would not change.

https://doi.org/10.1371/journal.pcbi.1012310.g001

two separate contact networks such that they are exposed but not infected in the first of these networks. At the start of the second exposure event, these individuals would have been previously exposed but not infected (henceforth called focal individuals). This contact network must also contain naive contacts: individuals that have not been previously exposed to the pathogen. The basis of our method is to compare the fraction of naive individuals and focal individuals that become infected in the second contact network; if there is no heterogeneity in susceptibility, focal individuals should have the same susceptibility as naive individuals, whereas if there is heterogeneity in susceptibility, focal individuals should on average be less susceptible than naive individuals. This difference in susceptibility arises due to the selection process for infection of more susceptible individuals (Fig 1). While in practice it is always possible that unknown to contact tracers, an individual classified as naive was exposed in the past, our method is fairly insensitive to such misclassification (S12 Text).

To compute the number of naive and focal individuals infected, there must be data on which specific individuals are infected and which individuals are showing up in a contact network for a second time, which would be available for example if individuals were identifiable between contact networks. This requirement should be easily met as unique identifiers are often collected or assigned during contact tracing [24, 25]. There must also be a sufficient sample size to detect heterogeneity in susceptibility. Here we explore the effects of sample size, level of heterogeneity, and infection probability on our ability to detect and quantify heterogeneity in susceptibility.

We apply this method to two underlying models describing the distribution of individuals' susceptibilities. In one underlying model (discrete case), it is assumed that the population is composed of two host types where each host type has a different susceptibility or probability of

being infected given contact. Discrete susceptibility types might be expected when heterogeneity in susceptibility is predominantly accounted for by a small number of factors that create groups in the population with distinct susceptibilities. For example, genetic polymorphisms could be selected for that increase resistance to a pathogen, resulting in populations containing a mixture of individuals with and without the mutation such as was seen for HIV [26]. Likewise, prior exposure, whether natural or vaccine-induced, to a pathogen or related pathogen could create more resistant subpopulations such as with milkmaids not developing smallpox after contracting cowpox [27]. Behaviors like handwashing and mask wearing [28, 29] or host nutritional status [30] could also produce approximately binary outcomes for susceptibility to infection.

In the other underlying model (continuous case), it is assumed that the population is composed of hosts with a continuous range of susceptibilities such that each host's probability of being infected given contact is unique. This situation might be expected when there is a complex combination of factors dictating heterogeneity in susceptibility or when the cause of heterogeneity is a trait that continuously varies across individuals. For instance, variability in gene expression, which could be affected by epigenetics, copy number variations, and sequence polymorphisms, is associated with disease susceptibility [31]. In addition, some of the factors that lead to discrete variation in susceptibility could also have a continuous effect such as the degree of cleanliness achieved by handwashing [28] or continuous variation in nutrients. Beyond a complex combination of factors, there could also be situations where a continuously varying trait like body mass [32], the level of antibodies induced in an immune response [33], or age [34] explains the heterogeneity in susceptibility in the population.

Methods

Our method is comprised of two parts: detecting heterogeneity in susceptibility and quantifying it if present. The former is a hypothesis testing problem, and the latter is a parameter estimation problem. For the detection of heterogeneity, we test the hypothesis that there is heterogeneity in susceptibility against the null hypothesis that there is homogeneity in susceptibility. Fig 2 summarizes the steps of our method, and Table 1 provides descriptions of the parameters used.

Detection of heterogeneity in susceptibility. We consider F contact networks that each contain $N_i - 1$ naive individuals and one focal individual where i is the set of contact networks. For simplicity, we assume N_i are equal for all i and thus drop the subscript, but this assumption can be easily relaxed and does not influence any of our conclusions (S11 Text). We therefore have a total of $F(N-1)$ naive individuals and F focal individuals. Also, although we consider contact networks containing one focal individual, there could be more focal individuals in a network in reality as any individuals that have been previously exposed but not infected would be considered focal. For our method, we first compute the fractions of naive, focal, and total individuals infected. The fractions of naive and focal individuals infected are estimates for the probability of a naive or focal individual being infected (p_n and p_f respectively). The fraction of total individuals infected is an estimate for the average probability of being infected ($\bar{p}$). We then calculate the log-likelihood of the data (numbers of individuals infected) under each hypothesis as a sum of the log-likelihoods for the number of each type of individual infected where

$$L_{\text{hom}} = \ln[P(x_n|F(N-1),\bar{p})] + \ln[P(x_f|F,\bar{p})] \quad (1)$$

$$L_{\text{het}} = \ln[P(x_n|F(N-1),p_n)] + \ln[P(x_f|F,p_f)]. \quad (2)$$

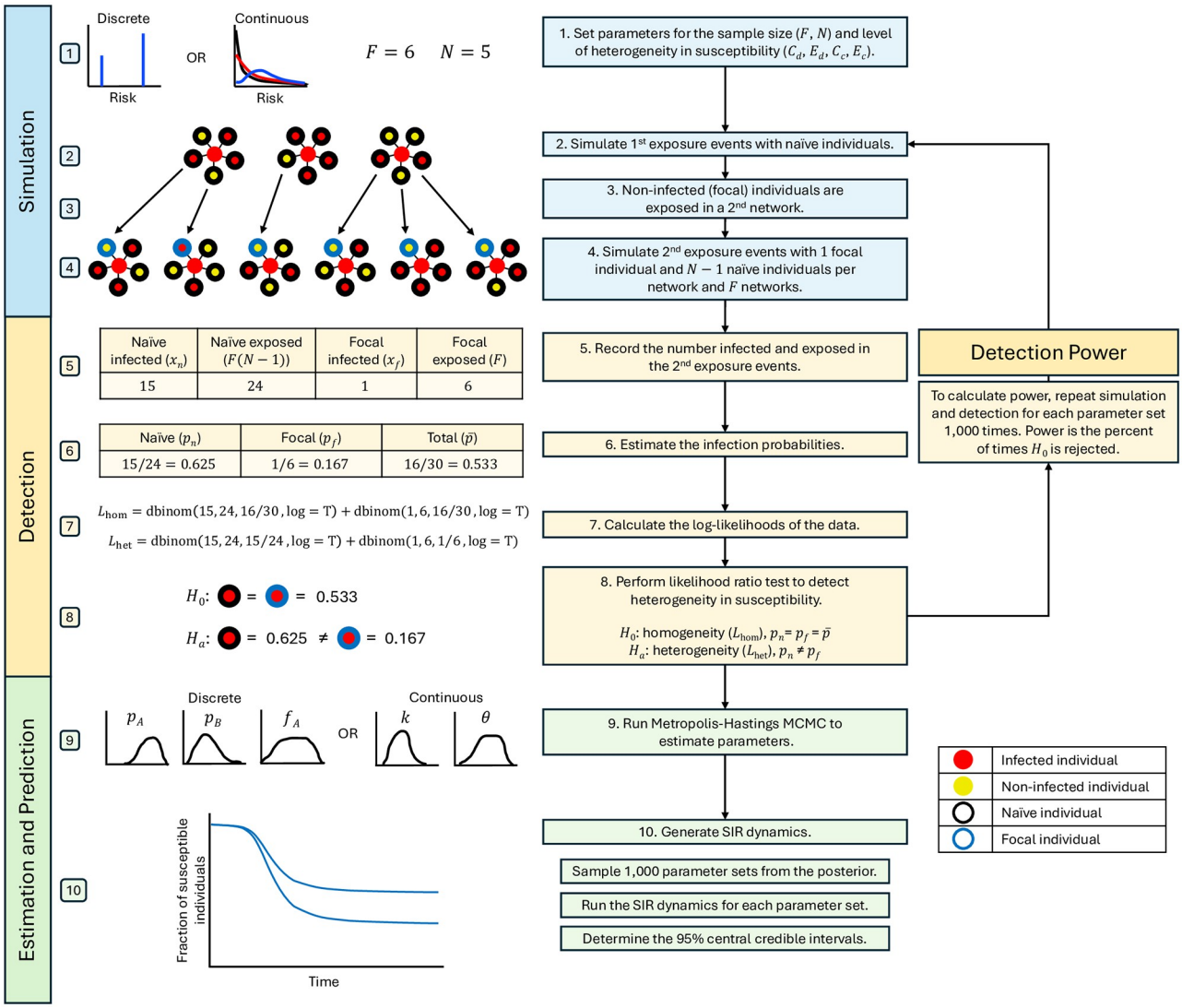

Fig 2. Method flowchart. The figure summarizes the steps of our method from simulating contact tracing data to detecting heterogeneity in susceptibility to estimating the level of heterogeneity and predicting disease dynamics. The diagram depicts a toy example going through the steps of our method. At step 7, we show the log-likelihoods in terms of the R code that would be used to calculate them. With real contact tracing data, the simulation section would be skipped.

https://doi.org/10.1371/journal.pcbi.1012310.g002

L_{hom} is the log-likelihood of the data under the null hypothesis that there is homogeneity in susceptibility, so we assume all individuals have the same probability of being infected, regardless of whether they are naive or focal ($p_n = p_f = \bar{p}$). L_{het} is the log-likelihood under the alternative hypothesis that there is heterogeneity in susceptibility, so we assume naive and focal individuals have different probabilities of being infected due to the infection selection process that occurs when heterogeneity is present ($p_n \neq p_f$). These log-likelihoods are calculated identically regardless of whether the heterogeneity is discrete or continuous. $P(x|y, p)$ is the probability of observing x individuals infected out of y individuals exposed with probability p of being infected and is distributed according to a binomial distribution. The number of naive individuals infected has distribution Binom($y = F(N-1)$, p_n), and the number of focal individuals infected has distribution Binom($y = F$, p_f). x_n and x_f are the numbers of naive and focal

Table 1. Descriptions of all parameters used in our method. The parameters' values are either set by input data, assumed, calculated from other parameters, or estimated via MCMC, which is specified in the Source column. Parameters in the first section are used universally across the cases, those in the second section are used for the discrete case, and those in the third section are used for the continuous case. Note that in the table, all probabilities of infection, risks of infection, and expected fractions infected are conditional on an individual showing up in a contact network and therefore do not depend on the overall force of infection in the population.

	Symbol	Description	Source
Universal	F	Number of focal individuals; Number of contact networks	Data input
	N	Number of contacts in each network	Data input
	x_n	Number of naive individuals infected	Data input
	x_f	Number of focal individuals infected	Data input
	p_n	Probability of infection for a naive individual; $\frac{x_n}{F(N-1)}$	Calculated from x_n, F, N
	p_f	Probability of infection for a focal individual; $\frac{x_f}{F}$	Calculated from x_f, F
	$\bar{p}$	Probability of infection for an average individual; $\frac{x_f+x_n}{FN}$	Calculated from x_f, x_n, F, N
	r_i	Risk of infection for the ith individual; $-\ln(1-p_i)$	Calculated from p_A, p_B, f_A or k, θ
	p_i	Probability of infection for the ith individual; $1-e^{-r_i}$	Calculated from r_i
	c	Contact rate for SIR model	Calculated from $R_{0,d}$ or $R_{0,c}$
	γ	Recovery rate for SIR model	Assumed $\gamma = 0.1$
Discrete case	p_A	Probability of infection for a type A, more susceptible individual	Estimated
	p_B	Probability of infection for a type B, less susceptible individual	Estimated
	f_A	Fraction of the initial population that is type A	Estimated
	C_d	Coefficient of variation of the risk of being infected; $\frac{(r_A-r_B)\sqrt{f_A(1-f_A)}}{r_Af_A+r_B(1-f_A)}$	Calculated from r_A, r_B, f_A
	E_d	Expected fraction of naive individuals infected; $p_Af_A + p_B(1-f_A)$	Calculated from p_A, p_B, f_A
	$R_{0,d}$	Basic reproduction number for SIR model; $\frac{(p_Af_A+p_B(1-f_A))c(S_0+I_0)}{\gamma}$	Assumed $R_{0,d} = 3$
	β_A, β_B	Transmission rates for SIR model; $\beta_A = p_Ac$, $\beta_B = p_Bc$	Calculated from p_A, p_B, c
Continuous case	k	Shape parameter for the distribution of individuals' risks	Estimated
	θ	Scale parameter for the distribution of individuals' risks	Estimated
	C_c	Coefficient of variation of the risk of being infected; $\frac{1}{\sqrt{k}}$	Calculated from k
	E_c	Expected fraction of naive individuals infected; $1-(1+\theta)^{-k}$	Calculated from k, θ
	$R_{0,c}$	Basic reproduction number for SIR model; $\frac{\rho c(S_0+I_0)}{\gamma}$	Assumed $R_{0,c} = 3$
	β	Transmission rate for SIR model; ρc	Calculated from ρ, c
	ρ	Probability of infection given contact for SIR model; $1-(1+\theta)^{-k}$	Calculated from k, θ

https://doi.org/10.1371/journal.pcbi.1012310.t001

individuals infected respectively where $x_n \in [0, F(N-1)]$ and $x_f \in [0, F]$. p_n, p_f, and $\bar{p}$ are estimated from the data as $p_n = \frac{x_n}{F(N-1)}$, $p_f = \frac{x_f}{F}$, and $\bar{p} = \frac{x_f+x_n}{FN}$. The log-likelihoods of the data under each hypothesis were compared using a likelihood ratio test with one degree of freedom and significance level $\alpha = 0.05$.

Here, we simulated data to test our method. To do so, we first set parameters dictating the sample size and heterogeneity present in the population. Then, we simulated initial exposure events with N individuals in each network and kept uninfected individuals as our focal individuals. For each focal individual, we then simulated a second exposure event with that focal individual and $N-1$ naive individuals. The susceptibilities of the naive individuals were drawn randomly from the same heterogeneity distribution set for the starting population. We recorded the fraction of each type of individual (i.e. focal or naive) infected in the second exposure event and calculated the log-likelihood of the simulated data under our two hypotheses. Then, we compared the hypotheses using a likelihood ratio test. We ran 1, 000 simulations for each set of parameters to determine our statistical power (the percent of simulations in which we reject the null hypothesis) to detect heterogeneity in susceptibility with that parameter combination. All simulations and data analyses were performed in R version 4.0.3 [35].

For the discrete case, we simulated data using two types of individuals (denoted A and B), but we note that the aforementioned factors could potentially be combined to result in more than two distinct groupings, and similar methods could be applied for these situations. At the beginning of each simulation, we set the probability of being infected for each type of individual, p_A and p_B, where $p_A \in [0, 1]$ and $p_B \in [0, p_A]$. p_A and p_B are the probabilities of infection given contact between that type of individual and a particular infected individual and therefore are independent of group size. We also set the fraction of the starting population that is type A (f_A) where $f_A \in [0, 1]$. All three parameters p_A, p_B, and f_A affect the level of heterogeneity in susceptibility in the population.

We later calculated the coefficient of variation of the risk of being infected for this discrete case (C_d) and the expected fraction of naive individuals infected (E_d) from p_A, p_B, and f_A to better summarize the results. The risks of being infected for type A and B individuals, r_A and r_B respectively, are shown below. These equations are derived from the formula for the probability of being infected $p_i = 1 - e^{-r_i}$, $i = A, B$.

$$r_A = -\ln(1 - p_A) \tag{3}$$

$$r_B = -\ln(1 - p_B). \tag{4}$$

The coefficient of variation is defined as the standard deviation divided by the mean. Hence, C_d is the standard deviation of risk divided by the mean risk (S1 Text) and is given by

$$C_d = \frac{(r_A - r_B)\sqrt{f_A(1 - f_A)}}{r_A f_A + r_B(1 - f_A)}. \tag{5}$$

E_d is the same as the mean probability of being infected $\bar{p}$, which is given by

$$E_d = \bar{p} = p_A f_A + p_B(1 - f_A). \tag{6}$$

We additionally defined the sample size for the simulation by setting the number of individuals in each exposure group N and the number of focal individuals F. For our simulations, we used $N = 5$ and $F = 50$ or 200.

For the continuous case, in contrast to the discrete case just discussed, each individual in the population has a different risk of being infected. Here, we assume that individuals' risks for being infected follow a gamma distribution, but as in the discrete case, other distributions could be used. We chose to use a gamma distribution for illustration purposes because it is flexible and has been used to model heterogeneous populations previously [7, 9].

At the beginning of each simulation, we set the parameters k and θ, respectively the shape and scale of the gamma distribution, that dictate the risk distribution where $k, \theta > 0$. For ease of interpretation, we present our results with respect to the coefficient of variation of risk for continuous variation C_c and expected fraction of naive individuals infected E_c. As in the discrete case, the risk r_i for the ith individual being infected given contact between that individual and a particular infected individual is related to the probability of being infected such that $p_i = 1 - e^{-r_i}$ and thus

$$r_i = -\ln(1 - p_i). \tag{7}$$

As it is gamma distributed, the risk distribution has standard deviation $\sigma = \theta\sqrt{k}$ and mean $\mu = k\theta$. So, C_c can be simplified to

$$C_c = \frac{1}{\sqrt{k}}. \tag{8}$$

E_c is the same as the mean probability of being infected $\bar{p}$ and is derived in [7] as

$$E_c = \bar{p} = 1 - \frac{S_t}{S_0} = 1 - (1+\theta)^{-k}, \tag{9}$$

where S_0 and S_t are the number of susceptible individuals at the beginning and end of an exposure round respectively.

We additionally defined the sample size for the simulation by setting the number of individuals in each exposure group N and the number of focal individuals F. As in the discrete case, we use $N = 5$ and $F = 50$, 200, or 1000.

We tested the ability of our method to detect heterogeneity in susceptibility for each potential combination of f_A, F, $C_d \in [0, 3]$ with step size 0.02, and $E_d \in [0.02, 0.98]$ with step size 0.02 in the discrete case and F, $C_c \in [0, 3]$ with step size 0.02, and $E_c \in [0.02, 0.98]$ with step size 0.02 in the continuous case. This was done for 1, 000 simulations to compute the statistical power of the method. We did not simulate $E_d = 0$, 1 or $E_c = 0$, 1 because such values preclude heterogeneity in susceptibility. We examined C_d, $C_c \in [0, 3]$ because this captures most of the range of published values for the coefficient of variation of risk we could find: 0.0007 to 3.33 [7, 9, 10, 14–19, 36–38].

We also generated sets of contact tracing data by using a Gillespie algorithm to simulate a stochastic dynamic disease model. While these data were substantially more time consuming to generate than our other simulated data and thus not suitable for use in the the full suite of power and estimation analyses conducted, the datasets that were analyzed yielded consistent results to that of our other simulation method (S11 Text).

Quantification of heterogeneity in susceptibility. Given the detection of heterogeneity in susceptibility, the next question is whether that heterogeneity will substantially impact disease dynamics. To determine whether it will, we need to ask whether contact tracing data is sufficient to estimate the parameters of SIR models that include heterogeneity in susceptibility and whether those parameter estimates accurately capture disease dynamics. To do so, we fit the parameters of our underlying risk distributions using simulated contact tracing data as above. Parameter values used to simulate the contact tracing data for the discrete and continuous heterogeneity cases are provided in Table 2.

Table 2. The 95% CIs, medians, and true values for parameters estimated from MCMC in the discrete and continuous cases with $F = 1000$ and $N = 5$.

	Parameter	95% CI	Median	True
Discrete case	p_A	[0.437,0.958]	0.599	0.748
	p_B	[0.005,0.172]	0.085	0.125
	f_A	[0.102,0.543]	0.321	0.2
	C_d	[0.842,1.845]	1.093	1.3
	E_d	[0.236,0.263]	0.249	0.25
Continuous case	k	[0.364,1.024]	0.584	0.592
	θ	[0.321,1.257]	0.647	0.626
	C_c	[0.988,1.657]	1.309	1.3
	E_c	[0.237,0.269]	0.252	0.25

https://doi.org/10.1371/journal.pcbi.1012310.t002

We generated posterior distributions for both models using Metropolis-Hastings MCMC. In the discrete case, our MCMC chain had length 30, 000, 000 with a burn-in of 15, 000, 000 and thinning interval 1, 500. For all three parameters, we used flat priors and uniform proposal distributions. Our proposal distributions were $p_A \sim \text{Unif}(0, 1)$, $p_B \sim \text{Unif}(0, p_A)$, and $f_A \sim \text{Unif}(0, 1)$. There is not a simple, analytic likelihood function for the likelihood of the data given a proposed parameter set, so the likelihood was estimated by simulation with Approximate Bayesian Computation (ABC), where the likelihood estimate was determined by comparing the fraction of simulations that provided results that were within a pre-specified error tolerance of the actual data [39]. To do so, we ran 100 simulations of the number of focal and naive individuals infected across F contact networks for a proposed parameter set. We then calculated the fraction of simulations where the number of individuals infected was within a 1% error tolerance of the number infected in the true data. Note that our results are fairly insensitive to this error tolerance (S4 Text). This simulation was done separately for focal and naive individuals. We then computed the overall log-likelihood as a sum of the logs of those fractions. We assessed convergence of the chains by visually inspecting the resulting trace plots and marginal posterior distributions for each parameter. In the continuous case, our MCMC chain had length 600, 000 with a burn-in of 200, 000 and thinning interval 100. We used an exponential prior Exp(2) for k because known values of C_c suggest that k is likely to be small [7, 9, 10, 14–19, 36–38]. We used a flat prior for θ for all values $[0, \infty)$ and a multivariate lognormal proposal distribution $(k, \theta) \sim \text{MLogNorm}\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.01 & -0.008 \\ -0.008 & 0.05 \end{pmatrix}\right)$. We assessed convergence of the chains by visually inspecting the resulting trace plots and marginal posterior distributions for each parameter [40].

We then used these parameter estimates to generate SIR dynamics. Notably, the system of differential equations describing the discrete and continuous cases differ. For the discrete case, we implemented the following system of ordinary differential equations:

$$\frac{dS_A}{dt} = -\beta_A S_A I \tag{10}$$

$$\frac{dS_B}{dt} = -\beta_B S_B I \tag{11}$$

$$\frac{dI}{dt} = (\beta_A S_A + \beta_B S_B) I - \gamma I. \tag{12}$$

S_A and S_B are the susceptible individuals of types A and B, and I is the infected individuals where I includes infected A and infected B individuals such that $I = I_A + I_B$. At the start of each SIR simulation, we determine the fraction of the population to allocate as A and B from f_A. We also set the basic reproduction number $R_{0,d} = \frac{\bar{\beta}(S_0+I_0)}{\gamma} = \frac{(p_A f_A + p_B(1-f_A))c(S_0+I_0)}{\gamma}$ at an assumed "true" value where $\bar{\beta}$ is the average transmission rate and S_0+ I_0 is the population size. $R_{0,d}$ is often a reasonably well approximated value, and it does not change with heterogeneity in susceptibility as initial average susceptibility remains the same regardless of heterogeneity [41, 42]. β_A and β_B are the transmission rates for types A and B respectively and were calculated as $\beta_A = p_A c$ and $\beta_B = p_B c$ where c is the contact rate. Note that c was calculated from $R_{0,d}$. γ is the recovery rate and was kept constant between the types of individuals at an assumed "true" value.

For the continuous case, we implemented the following system of ordinary differential equations derived in [16]:

$$\frac{dS}{dt} = -\beta SI\left(\frac{S}{S_0}\right)^{C_c^2} \tag{13}$$

$$\frac{dI}{dt} = \beta SI\left(\frac{S}{S_0}\right)^{C_c^2} - \gamma I. \tag{14}$$

S_0 is the number of susceptible individuals at the beginning of the simulation, S is the number of susceptible individuals at time t, and I is the number of infected individuals. At the start of each simulation, we set the basic reproduction number $R_{0,c} = \frac{\rho c(S_0+I_0)}{\gamma}$ at an assumed "true" value where ρ is the average probability of being infected given contact for a naive individual, c is the contact rate, $S_0 + I_0$ is the population size, and γ is the recovery rate. ρ was computed from the sampled parameters as $\rho = 1 - (1 + \theta)^{-k}$, c was calculated from $R_{0,c}$, and γ was fixed at an assumed "true" value. β is the transmission rate and was calculated as $\beta = \rho c$.

For each case, we randomly sampled 1, 000 parameter sets from the posterior distribution to run SIR model simulations, and we compared this to the dynamics generated by the "true" parameter set used to generate our contact tracing data. Using these simulations, we determined 95% central credible intervals for the SIR dynamics for each model by finding the 2.5% and 97.5% percentiles of the 1, 000 simulated dynamics at each time point over the epidemic. For our SIR simulations, we set $R_{0,d} = R_{0,c} = 3$, $S_0 = 20, 000$, $I_0 = 10$, and $\gamma = 0.1$.

Results

Detection of heterogeneity in susceptibility. Figs 3 and 4 illustrate that the sample size, level of heterogeneity, and fraction of individuals infected affect our power to detect heterogeneity in susceptibility. This is because these factors ultimately affect the likelihoods used to test for heterogeneity in terms of the difference between the probabilities of infection for naive and focal individuals (p_n and p_f) and the variability in the likelihood ratio test statistic (S3 Text). More precisely, these figures show that as the number of focal individuals F increases from 50 to 200, there is greater power to detect lower levels of heterogeneity (lower values of C_d, C_c). This additionally allows for greater power across a wider range of E_d and E_c. This was to be expected because higher sample sizes, particularly of the previously exposed, focal individuals, decreases variability in our estimates of p_n, p_f, and $\bar{p}$. Notably, changing the total number of hosts in each contact network N had very little effect on our results (S2 Text).

The level of heterogeneity in susceptibility present is described by the coefficient of variation of the risk distribution C_d or C_c. As C_d and C_c increase, there is more power to detect heterogeneity in susceptibility as there is more heterogeneity in the population. In the discrete case, for a given C_d, there is also more power to detect heterogeneity as f_A approaches 0.5. This is because as f_A approaches 0.5, the population is more evenly split between the two types of individuals, allowing for a greater difference between p_A and p_B and, therefore, p_n and p_f.

Lastly, the impact of the expected fraction of naive individuals infected (E_d, E_c) on power differs between the two underlying models. There is greater power to detect heterogeneity when an intermediate fraction of individuals is infected in the discrete case and when a greater fraction of individuals is infected in the continuous case. In the discrete case, E_d is determined by p_A, p_B, and f_A as per Eq 6. The only way to have a large fraction of individuals infected is if both p_A and p_B are large. Hence, when E_d is high, p_A and p_B must both be close to 1. For similar reasons, when E_d is low, p_A and p_B must both be close to 0. Even though the risks r_A and r_B

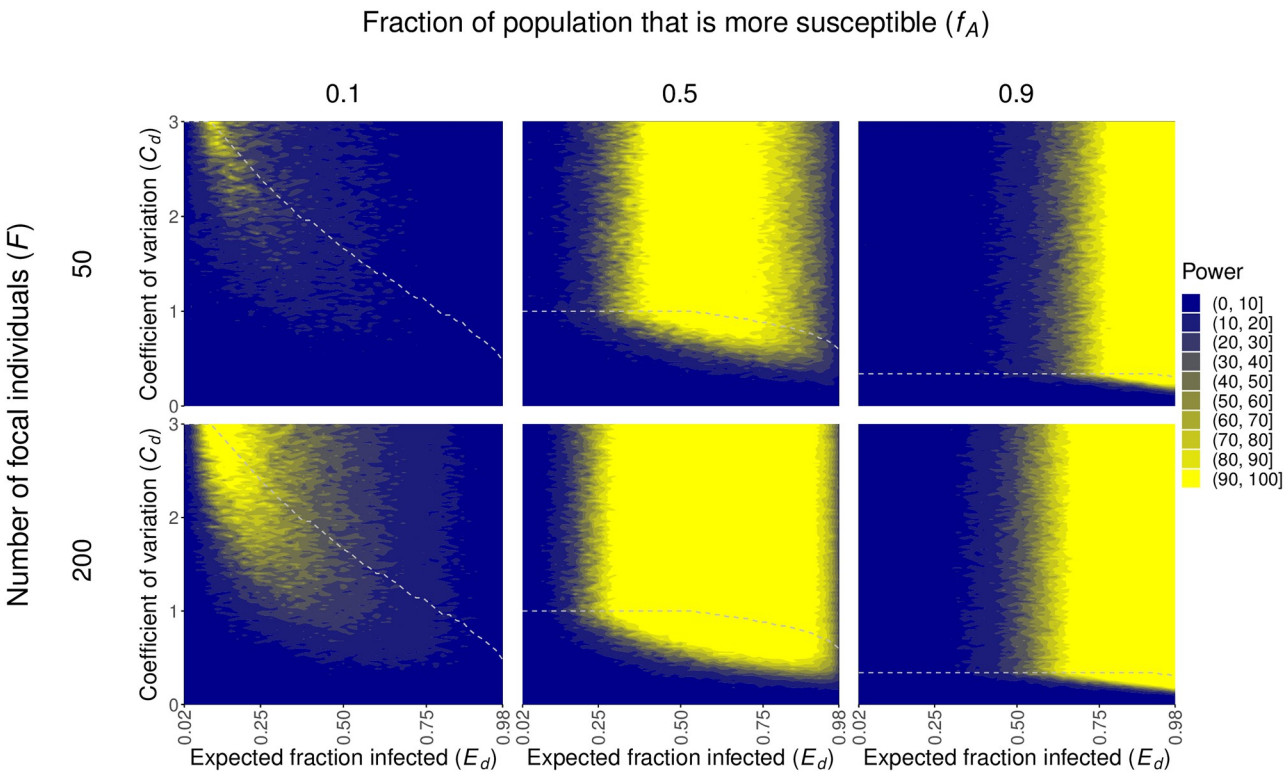

Fig 3. Increased heterogeneity in susceptibility (larger C_d and $f_A \to 0.5$), intermediate fractions of individuals infected (intermediate E_d), and increased sample sizes (larger F) enhance our power to detect heterogeneity in susceptibility in the discrete case. The plots show the power to detect heterogeneity in susceptibility in the discrete case, calculated as described in the text, across different numbers of focal individuals F and fraction of the population that is type A and more susceptible f_A. The areas above the gray dashed lines represent parameter space that gives computationally indistinguishable probabilities of infection p_A and p_B, and therefore power, to the parameter combination with the same E_d and highest C_d below the line. This occurs because risks of infection can be changed to increase C_d without bound, whereas probabilities are bounded. $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g003

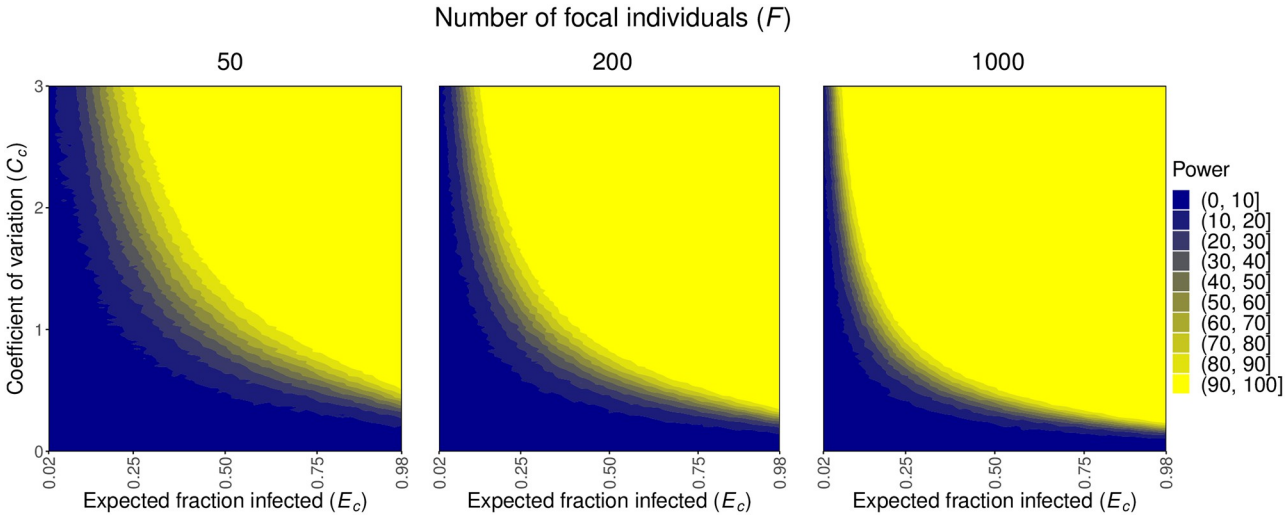

Fig 4. Increased heterogeneity in susceptibility (larger C_c), greater fractions of individuals infected (larger E_c), and increased sample sizes (larger F) enhance our power to detect heterogeneity in susceptibility in the continuous case. The plots show the power to detect heterogeneity in susceptibility in the continuous case, calculated as described in the text, across different numbers of focal individuals F. $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g004

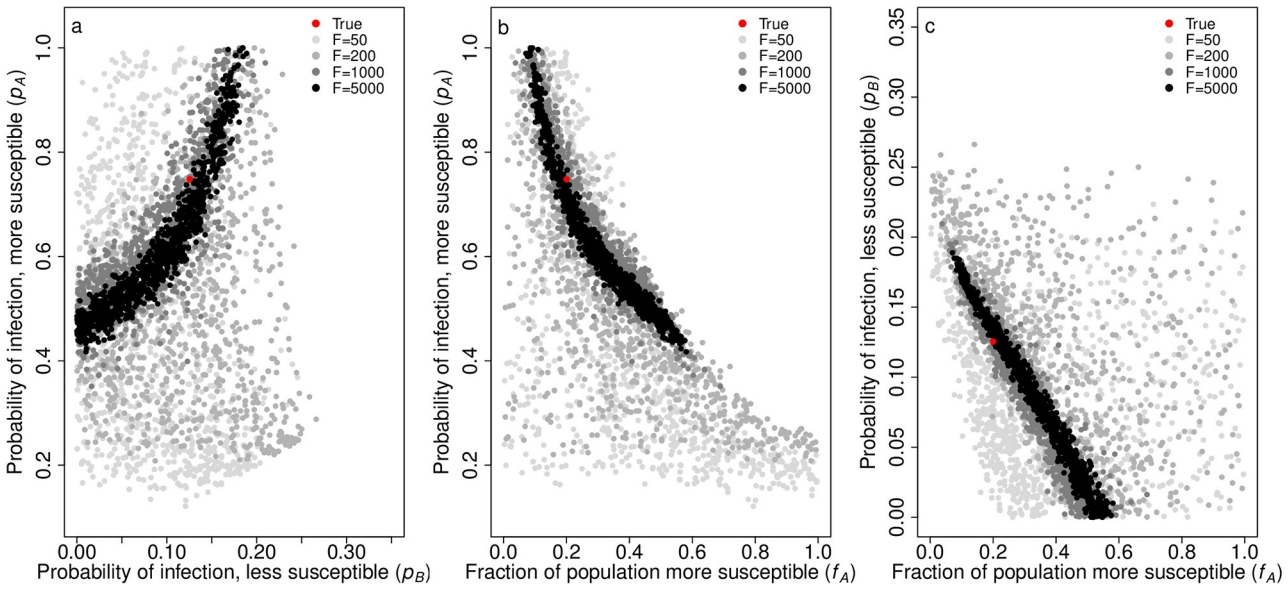

Fig 5. Parameter estimates for p_A, p_B, and f_A in the discrete case capture the true values and are highly correlated. The plots show the correlation in the parameter estimates for a) p_A vs. p_B, b) p_A vs. f_A, and c) p_B vs. f_A with different numbers of focal individuals F. These are the parameters that determine the distribution of individuals' susceptibilities in the discrete case. The red dots represent the true parameters used to generate our simulated data, and the gray dots depict 1, 000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $p_A = 0.748$, $p_B = 0.125$, $f_A = 0.2$, and $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g005

associated with these values may have varying levels of heterogeneity, the individuals themselves will have very similar infection outcomes, making it difficult to detect heterogeneity in susceptibility. Therefore, heterogeneity in susceptibility is better detected when an intermediate fraction of individuals is infected in the discrete case. In contrast, power increases in the continuous case with greater fractions of individuals infected (larger values of E_c). This is because there is more selection for who is infected as more individuals are infected, so the average population susceptibility will decrease more drastically, making it easier to detect heterogeneity in susceptibility.

Quantification of heterogeneity in susceptibility. We then explored the method's ability to estimate model parameters as well as predict the associated SIR dynamics. We performed this analysis for a particular parameter combination that led to $C_d = C_c = 1.3$ and $E_d = E_c = 0.25$. These values were chosen because they represent a biologically realistic scenario based on previous literature [7, 9, 10, 14–19, 36–38, 43–48]. In the discrete case, we used C_d and E_d and set $f_A = 0.2$ to calculate the true values $p_A = 0.748$ and $p_B = 0.125$. In the continuous case, we used C_c and E_c to calculate the true values $k = 0.592$ and $\theta = 0.626$.

We determined our 95% CIs for parameter estimation of the underlying parameters with $F = 1000$ and $N = 5$ to be those shown in Table 2. Note that the true values for p_A, p_B, and f_A as well as for k and θ are captured by these intervals. Admittedly, these parameter estimates are somewhat broad. Upon further investigation, we found the broad intervals to be due to high correlation in our parameter estimates, indicating low identifiability (Figs 5 and 6). However, acceptable estimates do not span the entire ranges of the parameters and encapsulate the true parameters, so there is some information about their values in the data. As we will discuss, this partial identifiability does not hinder us from making precise predictions about the impact of the heterogeneity in susceptibility on the disease dynamics.

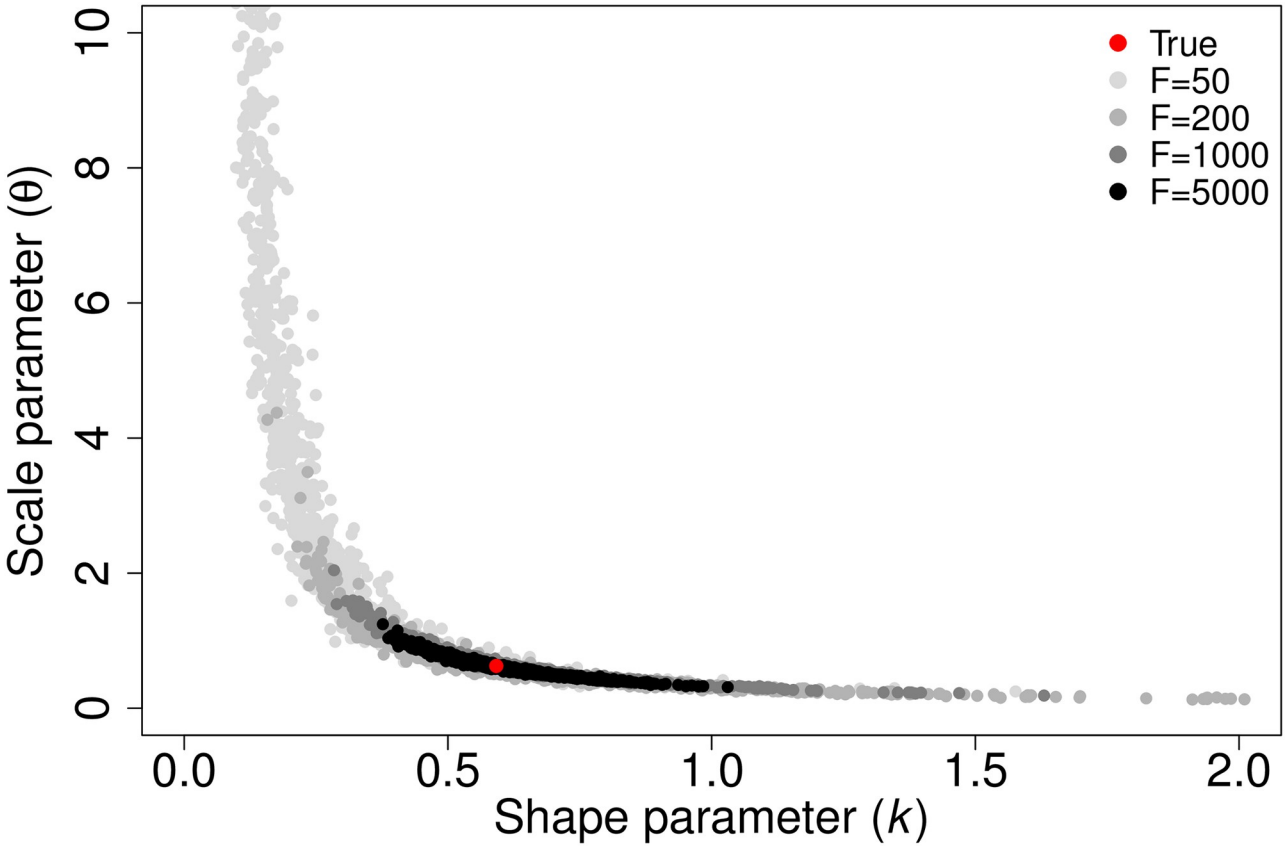

Fig 6. Parameter estimates for k and θ in the continuous case capture the true values and are highly correlated. This plot shows the correlation in the parameter estimates for k and θ that determine the gamma distribution of individuals' susceptibilities in the continuous case with different numbers of focal individuals F. The red dot represents the true parameters used to generate our simulated data, and the gray dots depict 1, 000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $k = 0.592$, $\theta = 0.626$, and $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g006

Using Eqs 5, 6, 8 and 9, we calculated and plotted the posterior distributions for C_d and E_d and C_c and E_c (Fig 7). With $F = 1000$ and $N = 5$, we determined the 95% CIs to be those shown in Table 2, which capture the true values. In the discrete case, the range of potential estimates for C_d is somewhat broad, but there is a strong ability to accurately and precisely estimate E_d. However, in the continuous case, there is a strong ability to accurately and precisely estimate both C_c and E_c. With increasing values of F from 50 to 5000, estimates for C_c and E_c become more precise.

We then investigated the SIR dynamics for these parameter sets with different sample sizes (F and N). We also investigated the dynamics with different error tolerances allowed for ABC in the discrete case. For both underlying models, with $N = 5$ and $F = 50$, 200, 1000, or 5000, the true dynamics are captured by the 95% CIs (Fig 8). Additionally, for $F > 200$ in the discrete case and for all F in the continuous case, the estimated disease dynamics do not overlap those where there is assumed to be no heterogeneity in susceptibility. Hence, despite low identifiability in the parameter estimates, we are able to use this method to make accurate and precise predictions about the effect of heterogeneity in susceptibility on disease dynamics. This is because there is interdependence among the parameters (Figs 5 and 6), and so, while individual parameters may be only partially identifiable, combinations of them can be precisely estimated,

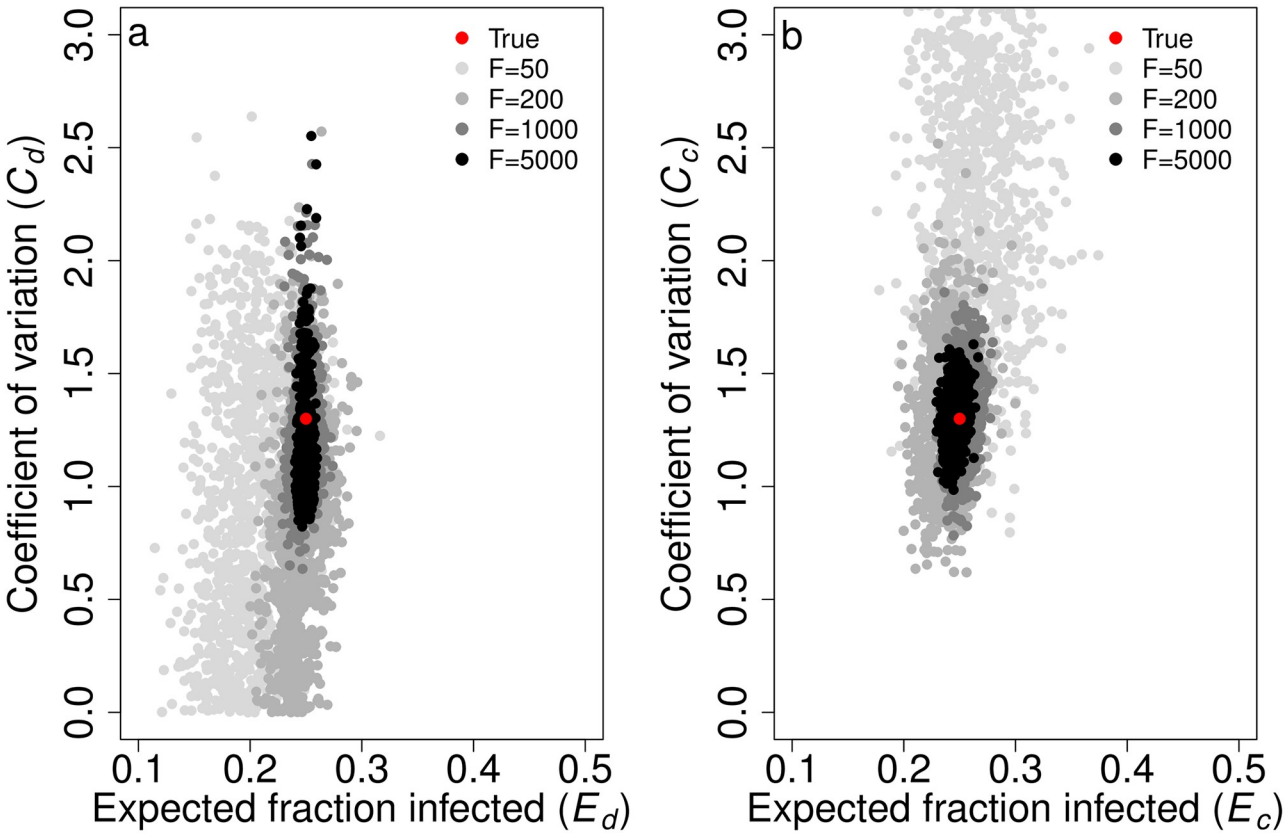

Fig 7. Parameter estimates for the coefficient of variation of risk (C_d, C_c) and expected fraction of naive individuals infected (E_d, E_c) capture the true values and become more precise with increasing numbers of focal individuals *F*. The plots show the parameter estimates for *C* and *E* with different numbers of focal individuals *F* in a) the discrete case and b) the continuous case. The red dots represent the true parameters used to generate our simulated data, and the gray dots depict 1, 000 parameter sets from our posterior distribution for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). $C_d = C_c = 1.3$, $E_d = E_c = 0.25$, $f_A = 0.2$, and $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g007

leading to relatively precise estimates of the level of heterogeneity in susceptibility *C* and the fraction of naive individuals infected *E*.

We found the continuous case provided more accurate and precise predictions of disease dynamics than the discrete case, but the 95% CIs narrowed with higher sample sizes in both cases (Fig 8). In the discrete case, as *F* increased, there was a limit to how narrow the 95% CIs became. $F > 1000$ did not substantially improve the predicted dynamics relative to those for $F = 1000$. Likewise, the number of non-focal individuals had relatively little impact on our predicted dynamics, yielding nearly identical results for $N = 5$ and $N = 100$ (S2 Text). In the continuous case, as *F* increased, the 95% CIs narrowed and converged around the true dynamics. With $N = 5$ versus $N = 100$, there was not a substantial difference in the 95% CIs (S2 Text).

To assess the accuracy of our ABC method for parameter estimation in the discrete case, we examined the SIR dynamics with different error tolerances of 10%, 1%, or 0%. We did so with $N = 5$ and $F = 200$ and 1000. Changing the error tolerance did not substantially impact the precision of the 95% CIs in any of the cases explored (S4 Text).

We also attempted to predict disease dynamics with the wrong underlying model of individuals' risks as it may be unknown which model is correct in a real system. To do so, we generated data under the discrete case then predicted SIR dynamics assuming the continuous case and vice versa. Notably, the 95% CIs from the incorrectly assumed underlying models did not

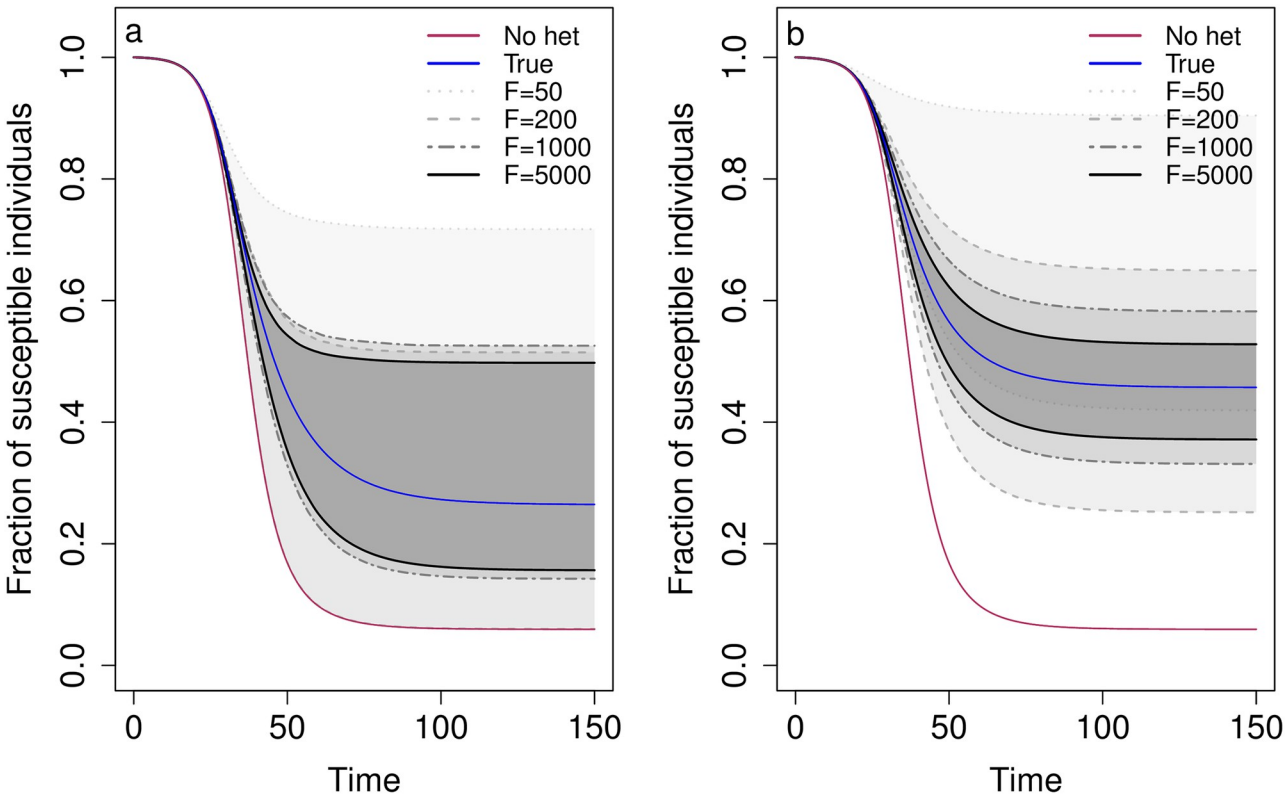

Fig 8. Predicted SIR dynamics capture the true dynamics and the 95% CIs narrow as the number of focal individuals *F* increases. The plots show the predicted SIR dynamics in a) the discrete case and b) the continuous case with different numbers of focal individuals *F*. Specifically, the fraction of susceptible individuals $\frac{S}{S_0}$ is shown over the course of an epidemic. Shaded regions represent 95% CIs determined from 1, 000 posterior samples for $F = 50$ (light gray), 200 (medium gray), 1000 (dark gray), and 5000 (black). The blue lines show the true dynamics for the parameters used to generate the contact tracing data, and the red lines show the corresponding dynamics if there is homogeneity in susceptibility. $C_d = C_c = 1.3$, $E_d = E_c = 0.25$, $f_A = 0.2$, and $N = 5$.

https://doi.org/10.1371/journal.pcbi.1012310.g008

capture the true dynamics, meaning that caution should be taken in ensuring that an accurate model of heterogeneity is assumed before trusting the precise disease dynamics that would be expected to arise from a given set of parameter estimates (S5 Text). Nevertheless, we stress that the ability to detect the presence of heterogeneity is independent of the underlying model and will not be affected by an incorrect model.

Discussion

As we saw play out during the COVID-19 pandemic, early epidemiological model predictions of disease dynamics can be crucial in informing public health policy. There are numerous imperfect assumptions made by standard SIR models, and a great deal of work has been aimed at trying to improve such models. Heterogeneity in susceptibility, differences between hosts in their likelihood of becoming infected given contact, can be critically important to disease dynamics [7–10]. However, current methods to estimate this heterogeneity rely on data that are collected late in an epidemic or that are unable to be collected due to ethical or logistical constraints. Here we have developed a method to detect and estimate heterogeneity using contact tracing data which, in theory, could allow epidemiologists to incorporate the effects of heterogeneity in susceptibility into their models even before the effects of such heterogeneity are observable at the population scale. Using a simulation-based approach, we found that contact

tracing data alone have enough information to be used to detect and quantify heterogeneity in susceptibility. For our method, power to detect heterogeneity increases with larger sample sizes and greater heterogeneity present as well as intermediate fractions infected in the discrete case (E_d) and high fractions infected in the continuous case (E_c).

Few studies have estimated heterogeneity in susceptibility in any infectious disease systems. Performing a standard literature search, we were able to find 45 estimates of heterogeneity in susceptibility explicitly provided from only 8 unique systems [7, 9, 10, 14–19, 36–38] with only 5 of those estimates pertaining to 3 human disease systems. There are additionally some studies that find heterogeneity in susceptibility but do not provide an estimate [19, 49–51] though one could be calculated if all pertinent information were available, increasing the number of systems for which heterogeneity in susceptibility is known to at least 10. While our list of estimates may not be entirely exhaustive, our method may be useful for expanding the set of systems for which heterogeneity in susceptibility can be detected and estimated. To determine whether our method is sufficiently powered, we need to know whether the values of the expected fraction infected E and the coefficient of variation of risk C are in a parameter space where our method would likely be suitable. Of the estimates for C that we found in the literature, 41 (91%) of them were greater than 0.5 and 20 (44%) were greater than or equal to 1.5. With 200 focal individuals ($F = 200$), $f_A = 0.5$, and $C = 1.5$, we have at least 80% power to detect heterogeneity in susceptibility when E_d is between 0.28 and 0.92 or when E_c is between 0.26 and 0.98. With $F = 1000$ and $C = 1.5$, we have at least 80% power when E_d is between 0.18 and 0.98 or when E_c is between 0.14 and 0.98 (Figs 3 and 4). In studies examining contact tracing data, we found secondary attack rates, which provide conservative estimates of E, to often be around 0.2 and sometimes as high as 0.733 [43–48]. Our method should therefore be sufficiently powered for many systems.

The precision in our prediction of SIR dynamics is also affected by the nature of the heterogeneity in susceptibility. Our estimates of how heterogeneity affects disease dynamics are less precise when there are discrete differences in risk between hosts, as opposed to continuous variation in risk (Fig 8). This is because, in addition to C_d and E_d, the fraction of the initial population that is the more susceptible type of individual, f_A, is critical for determining the trajectory of the epidemic. With the same C_d and E_d, the final epidemic size can differ depending on f_A (S6 Text). Hence, the need to estimate the additional parameter f_A in the discrete case with the same data results in wider 95% CIs. However, we can generate narrow 95% CIs and more precise parameter estimates in the discrete case if there is prior knowledge of at least one of the parameters p_A, p_B, or f_A (S7 Text).

We found that using the correct underlying model is important for accurately predicting disease dynamics, but not for the detection of heterogeneity in the first place. The underlying model used for parameter estimation should therefore be carefully chosen to reflect prior understanding of the potential drivers of heterogeneity in susceptibility in the system. The process for initial detection of heterogeneity in susceptibility is the same regardless of the underlying model (Eqs 1 and 2). Therefore, we can reliably detect heterogeneity in susceptibility without knowledge of the distribution of individuals' risks.

One strength of our method is that it allows for estimation of heterogeneity in susceptibility in real time, early in an epidemic with no data other than contact tracing data. Admittedly, the use of these data in real time will depend on the speed with which the necessary data can be collected and communicated, but existing methods to quantify heterogeneity are not adequate for real time usage even with immediate access to the data. Ben-Ami et al. [17] and Langwig et al. [9] used experimental dose-response curves to estimate heterogeneity in susceptibility, and Dwyer et al. [7] used a combination of laboratory dose-response experiments, field transmission experiments, and models fit to mortality data to investigate heterogeneity. Although

these experimental methods can provide good estimates of heterogeneity in susceptibility, they are not feasible for application in real time or for human epidemics in general due to time constraints and ethical concerns. Gomes et al. [15] compared disease incidence across municipalities in several countries to quantify heterogeneity for tuberculosis. This was done by ordering the municipalities by incidence rate and plotting the percentage of cumulative tuberculosis cases versus cumulative population sizes to construct Lorenz curves and thereby fit susceptibility risk distributions. This method, however, requires a considerable amount of data with ten or more years of data used in this study. Smith et al. [18] and Corder et al. [19] used malaria morbidity data to fit models of malaria and estimate heterogeneity. This method cannot be used until later in an epidemic when sufficient data are collected to fit curves. Gomes et al. [10] also used curve fitting with mortality data to estimate heterogeneity in susceptibility for COVID-19. They were able to estimate heterogeneity in real time once at least four months of data were available. While our method is in principle able to estimate heterogeneity in a similar time frame provided robust contact tracing, we also note that their method is heavily dependent on the underlying model and assumptions, and the authors advise not to trust the precision of their estimates. In addition, Gomes et al. [10] were unable to disentangle heterogeneity in contact rate from heterogeneity in underlying susceptibility. Our method estimates heterogeneity in underlying susceptibility, and the remaining heterogeneity in contact rate can be determined from the contact network data. Anderson et al. [20] used household study data to estimate heterogeneity in susceptibility. While this method is suitable for use in real time, and can be applied to human infectious diseases, the method notably is designed to estimate heterogeneity within households, which is not the same as the population-level heterogeneity that drives population-level disease dynamics.

Our method is unable to precisely estimate the individual parameters that define the risk distributions (i.e., p_A, p_B, f_A in the discrete case and k, θ in the continuous case), but our method is able to reliably predict disease dynamics. This seeming paradox arises because the disease dynamics depend on combinations of parameters rather than individual parameters. Notably, our method is substantially better at estimating the composite parameters describing the coefficient of variation of risk C and the expected fraction of naive individuals infected E. Nevertheless, our method does require a substantial amount of data (200 individuals showing up in contact networks for a second time). This requirement could be mitigated by pooling contact network data from multiple locations in order to more quickly collect sufficient data. It may also be possible to combine our method with another, like that of [10], to reduce the data required by either method. By combining our method with another, it may also be possible to more precisely estimate the individual parameters (i.e., p_A, p_B, f_A in the discrete case and k, θ in the continuous case) that our method has limited ability to precisely estimate.

There are additionally several considerations to address with regard to working with contact tracing data. Perhaps most prominently, contact tracing data tend to be messy and imperfect. Our method as described above assumes perfect data. However, our method can be readily modified to account for imperfect data. We can imagine multiple ways in which contact tracing data may be imperfect. Some important considerations are that: a) individuals may be mislabeled as uninfected when they are infected (false negatives), b) individuals may be mislabeled as infected when they are uninfected (false positives), and c) individuals may be missing from the contact networks despite being contacts (missing contacts). If there are false negatives, our method may overestimate the level of heterogeneity because our estimate of p_f may be biased lower. This is because, assuming infection confers at least partial immunity, focal individuals that were actually infected previously (i.e. false negatives) will be less likely to be infected than focal individuals that were true negatives. To counteract this issue, we developed a version of the method that corrects for false negatives by adjusting the likelihood

calculations for both detecting and estimating heterogeneity. For estimating parameters and predicting disease dynamics, adjusting the method to correct for false negatives fixes the issue (S8 Text). For detecting heterogeneity in susceptibility, adjusting the likelihood calculation corrects for the impact of false negatives except when the expected fraction infected E_d is very close to 1. We do not think this will be a major issue as E_d is typically less than 0.5 [43–48]. If there are false positives, our method may underestimate the level of heterogeneity because our estimate of p_f may be biased higher. This is because a high false positive rate will have a larger impact on individuals with a low susceptibility than those with a high susceptibility. Hence, focal individuals, which are on average less susceptible, and naive individuals will appear to have more similar infection probabilities. However, false positive rates are often small, close to 1–2% [52, 53], so this issue is not a huge concern for our method unless false positive rates are known to be unusually large. If there are many missing contacts, our method may have reduced power to detect heterogeneity because our estimates of p_n and p_f may both be biased lower, resulting in a smaller difference between them. This is because individuals that we observe as being exposed one (naive) or two (focal) times total but that were actually previously exposed more times may be less likely to be infected than true naive and focal individuals. These missed individuals may have gained immunity through infection or may be on average less susceptible through the infection selection process. Nevertheless, with only about 50% of exposures captured, our method is able to accurately estimate the level of heterogeneity and predict disease dynamics. Additionally, the reduction in detection power caused by missing contacts is lessened with increasing sample size (S12 Text). Early in an epidemic there is also a low chance of substantial missed individuals showing up in the second contact networks that our method evaluates. So, missing individuals should have only a negligible effect on the method's power in these early stages. While we have considered these three ways in which contact tracing data may be imperfect, it is highly likely that each set of contact tracing data will have its own set of peculiarities. Note that these peculiarities, if known, can readily be accounted for using our ABC method since any process may be used for simulation. Known imperfections in the data should therefore not bias estimates although they may still reduce power or increase required sample sizes.

We have presented our method using contact tracing data that were simulated as static contact networks, all of the same size N. In reality, contact networks would come from a dynamic epidemic and have varying N. To verify our method's performance on this more realistic data, we generated contact tracing data from a stochastic, individual-based SIR model. We recovered the same results for both detection and estimation with the dynamically generated contact networks in the discrete and continuous cases with the exception that the method may need to be modified late in an epidemic after individuals have been exposed many times (S11 Text). Therefore, the method can be reliably applied to contact tracing data from a dynamic epidemic and is not impacted by variable contact network sizes.

Another important point is that our method, as presented, assumes no forms of heterogeneity other than heterogeneity in susceptibility. One other source of heterogeneity is heterogeneity in transmission [54]. Heterogeneity in transmission is differences between hosts in their likelihood of transmitting a pathogen once infected. If this heterogeneity arises due to variation in the number of contacts that individuals have, then heterogeneity in transmission poses no problems for our method. It would simply mean that each contact network would have a unique value for N (S11 Text). We note that this variation in contact rate is the typical mechanism through which heterogeneity in transmission is assumed to act [54]. However, if heterogeneity in transmission arises due to differences between hosts in their likelihood of transmission given contact, our method may have less power to detect heterogeneity in susceptibility and may yield less precise or faulty conclusions about the disease dynamics (S9 Text).

Our method, in its base form, is thus not suitable in these cases. A first step to identify whether this is a concern is to perform a goodness of fit test before implementing the method to determine whether there is evidence of heterogeneity in transmission given contact (S9 Text). If this source of heterogeneity is detected, then our base method should not be used, and the generalized version of our method that simultaneously accounts for both sources of heterogeneity should instead be employed (S10 Text).

There may additionally be heterogeneity in exposure strength among contacts within a network such that individuals experience different forces of infection. This could be due to factors like differences in exposure time or type of contact (e.g., contacts that shared a taxi, were at the same party, etc.). This added heterogeneity may reduce the power of our method to detect heterogeneity in susceptibility as different contact types may have different transmission probabilities, providing varying levels of information that our current method disregards. To alleviate the potential impact of this heterogeneity, it may be necessary to break apart contact networks into specific exposure events or by some relevant factor, such as the duration of contact, and either weigh the type of contact differently or only use equivalent contact types.

Finally, we note that exposure could change individuals' susceptibilities. Individuals exposed in a first contact network could receive a small dose of the pathogen such that their immune system is stimulated without them becoming infected. This could decrease their susceptibility, meaning that some focal individuals have lower susceptibilities because they developed immunity, not because they were innately less susceptible [55]. However, this will have the same effect as heterogeneity in susceptibility of slowing down the epidemic and could even be considered a form of heterogeneity in susceptibility. In the same way, vaccination could affect susceptibility and be a cause of heterogeneity in susceptibility [9].

The earliest practice of tracing diseases dates back to the 1500s when doctors would track the spread of syphilis [56], and the earliest known example of contact tracing dates to 1576 during a bubonic plague pandemic [57]. Since then, the practice of contact tracing has spread, and it is now used widely, ranging from diseases such as influenza to HIV [43–48]. Recently, contact tracing data has transitioned from paper copies to electronic databases. Regardless, all of these sources of data could be used with our method provided they include focal individuals that are identifiable between contact networks, specify which individuals are infected, and have a sufficient sample size. Using our method, it should therefore, without collecting any new data, be possible to estimate heterogeneity in susceptibility in various locations and time periods for dozens of disease systems in which it has never been estimated previously.

Supporting information

S1 Text. Derivation of C_d.
(PDF)

S2 Text. Changing N.
(PDF)

S3 Text. Difference between probabilities of infection for naive and focal hosts (p_n and p_f) and variability in the likelihood ratio test statistic.
(PDF)

S4 Text. Changing error tolerance for Approximate Bayesian Computation (ABC) in the discrete case.
(PDF)

S5 Text. Effect of assuming the wrong underlying model.
(PDF)

S6 Text. Final epidemic size in the discrete case with different fractions of the population that are more susceptible (f_A).
(PDF)

S7 Text. Effect of having informative priors for p_A, p_B, or f_A on predicted disease dynamics.
(PDF)

S8 Text. False negatives.
(PDF)

S9 Text. Heterogeneity in transmission.
(PDF)

S10 Text. Estimating heterogeneity in susceptibility in the presence of heterogeneity in transmission given contact.
(PDF)

S11 Text. Contact tracing data from a dynamic epidemic.
(PDF)

S12 Text. Missing contacts.
(PDF)

Acknowledgments

We thank the Read, McGraw, and Kennedy labs for stimulating discussions.

Author Contributions

Conceptualization: Beth M. Tuschhoff, David A. Kennedy.

Data curation: Beth M. Tuschhoff.

Formal analysis: Beth M. Tuschhoff.

Funding acquisition: David A. Kennedy.

Investigation: Beth M. Tuschhoff.

Methodology: Beth M. Tuschhoff, David A. Kennedy.

Project administration: David A. Kennedy.

Resources: David A. Kennedy.

Software: Beth M. Tuschhoff.

Supervision: David A. Kennedy.

Validation: Beth M. Tuschhoff.

Visualization: Beth M. Tuschhoff.

Writing – original draft: Beth M. Tuschhoff.

Writing – review & editing: Beth M. Tuschhoff, David A. Kennedy.

References

1. Keeling M, Danon L. Mathematical modelling of infectious diseases. Br Med Bull. 2009; 92:33–42. https://doi.org/10.1093/bmb/ldp038 PMID: 19855103
2. Roberts M, Andreasen V, Lloyd A, Pellis L. Nine challenges for deterministic epidemic models. Epidemics. 2015; 10:49–53. https://doi.org/10.1016/j.epidem.2014.09.006 PMID: 25843383
3. Tolles J, Luong T. Modeling epidemics with compartmental models. JAMA. 2020; 323(24):2515–6. https://doi.org/10.1001/jama.2020.8420 PMID: 32459319
4. Dhar A. What one can learn from the SIR model. In: Indian Academy of Sciences Conference Series. 2020; 3(1).
5. Woolhouse ME, Dye C, Etard JF, Smith T, Charlwood J, Garnett G, et al. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proc Natl Acad Sci USA. 1997; 94(1):338–42. https://doi.org/10.1073/pnas.94.1.338 PMID: 8990210
6. VanderWaal KL, Ezenwa VO. Heterogeneity in pathogen transmission: mechanisms and methodology. Funct Ecol. 2016; 30(10):1606–22. https://doi.org/10.1111/1365-2435.12645
7. Dwyer G, Elkinton JS, Buonaccorsi JP. Host heterogeneity in susceptibility and disease dynamics: tests of a mathematical model. Am Nat. 1997; 150(6):685–707. https://doi.org/10.1086/286089 PMID: 18811331
8. Gomes MGM, Lipsitch M, Wargo AR, Kurath G, Rebelo C, Medley GF, et al. A missing dimension in measures of vaccination impacts. PLoS Pathog. 2014; 10(3):e1003849. https://doi.org/10.1371/journal.ppat.1003849 PMID: 24603721
9. Langwig KE, Wargo AR, Jones DR, Viss JR, Rutan BJ, Egan NA, et al. Vaccine effects on heterogeneity in susceptibility and implications for population health management. mBio. 2017; 8:e00796–17. https://doi.org/10.1128/mBio.00796-17 PMID: 29162706
10. Gomes MGM, Ferreira MU, Corder RM, King JG, Souto-Maior C, Penha-Gonçalves C, et al. Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. J Theor Biol. 2022; 540. 111063. https://doi.org/10.1016/j.jtbi.2022.111063 PMID: 35189135
11. Aguas R, Corder RM, King JG, Goncalves G, Ferreira MU, Gomes MGM. Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv [Preprint]. 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.07.23.20160762v5
12. Montalbán A, Corder RM, Gomes MGM. Herd immunity under individual variation and reinfection. J Math Biol. 2022; 85. 2. https://doi.org/10.1007/s00285-022-01771-x PMID: 35773525
13. Anderson RM, May RM. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. Math Med Biol. 1984; 1(3):233–66. https://doi.org/10.1093/imammb/1.3.233 PMID: 6600104
14. King JG, Souto-Maior C, Sartori LM, Maciel-de Freitas R, Gomes MGM. Variation in *Wolbachia* effects on *Aedes* mosquitoes as a determinant of invasiveness and vectorial capacity. Nat Commun. 2018; 9. 1483. https://doi.org/10.1038/s41467-018-03981-8 PMID: 29662096
15. Gomes MGM, Oliveira JF, Bertolde A, Ayabina D, Nguyen TA, Maciel EL, et al. Introducing risk inequality metrics in tuberculosis policy development. Nat Commun. 2019; 10. 2480. https://doi.org/10.1038/s41467-019-10447-y PMID: 31171791
16. Elderd BD, Dushoff J, Dwyer G. Host-pathogen interactions, insect outbreaks, and natural selection for disease resistance. Am Nat. 2008; 172(6):829–42. https://doi.org/10.1086/592403 PMID: 18976065
17. Ben-Ami F, Ebert D, Regoes RR. Pathogen dose infectivity curves as a method to analyze the distribution of host susceptibility: a quantitative assessment of maternal effects after food stress and pathogen exposure. Am Nat. 2010; 175(1):106–15. https://doi.org/10.1086/648672 PMID: 19911987
18. Smith D, Dushoff J, Snow R, Hay S. The entomological inoculation rate and *Plasmodium falciparum* infection in African children. Nature. 2005; 438:492–5. https://doi.org/10.1038/nature04024 PMID: 16306991
19. Corder RM, Ferreira MU, Gomes MGM. Modelling the epidemiology of residual *Plasmodium vivax* malaria in a heterogeneous host population: A case study in the Amazon Basin. PLoS Comput Biol. 2020; 16(3):e1007377. https://doi.org/10.1371/journal.pcbi.1007377 PMID: 32168349
20. Anderson TL, Nande A, Merenstein C, Raynor B, Oommen A, Kelly BJ, et al. Quantifying individual-level heterogeneity in infectiousness and susceptibility through household studies. Epidemics. 2023; 44. 100710. https://doi.org/10.1016/j.epidem.2023.100710 PMID: 37556994
21. Eames KT, Keeling MJ. Contact tracing and disease control. Proc R Soc B. 2003; 270:2565–71. https://doi.org/10.1098/rspb.2003.2554 PMID: 14728778
22. Hossain AD, Jarolimova J, Elnaiem A, Huang CX, Richterman A, Ivers LC. Effectiveness of contact tracing in the control of infectious diseases: a systematic review. Lancet Public Health. 2022; 7(3):E259–73. https://doi.org/10.1016/S2468-2667(22)00001-9 PMID: 35180434

23. Bradshaw WJ, Alley EC, Huggins JH, Lloyd AL, Esvelt KM. Bidirectional contact tracing could dramatically improve COVID-19 control. Nat Commun. 2021; 12. 232. https://doi.org/10.1038/s41467-020-20325-7 PMID: 33431829

24. Simmhan Y, Rambha T, Khochare A, Ramesh S, Baranawal A, George JV, et al. GoCoronaGo: privacy respecting contact tracing for COVID-19 management. J Indian Inst Sci. 2020; 100:623–46. https://doi.org/10.1007/s41745-020-00201-5 PMID: 33199945

25. Sowmiya B, Abhijith V, Sudersan S, Sakthi Jaya Sundar R, Thangavel M, Varalakshmi P. A survey on security and privacy issues in contact tracing application of Covid-19. SN Comput Sci. 2021; 2:1–11. https://doi.org/10.1007/s42979-021-00520-z PMID: 33728414

26. Huang Y, Paxton WA, Wolinsky SM, Neumann AU, Zhang L, He T, et al. The role of a mutant CCR5 allele in HIV–1 transmission and disease progression. Nat Med. 1996; 2:1240–3. https://doi.org/10.1038/nm1196-1240 PMID: 8898752

27. Barquet N, Domingo P. Smallpox: the triumph over the most terrible of the ministers of death. Ann Intern Med. 1997; 127:635–42. https://doi.org/10.7326/0003-4819-127-8_Part_1-199710150-00010 PMID: 9341063

28. Larson E. Skin hygiene and infection prevention: more of the same or different approaches? Clin Infect Dis. 1999; 29(5):1287–94. https://doi.org/10.1086/313468 PMID: 10524977

29. Van der Sande M, Teunis P, Sabel R. Professional and home-made face masks reduce exposure to respiratory infections among the general population. PLoS One. 2008; 3(7):e2618. https://doi.org/10.1371/journal.pone.0002618 PMID: 18612429

30. Chandra R. Nutritional deficiency and susceptibility to infection. Bull World Health Organ. 1979; 57 (2):167–77. PMID: 108017

31. Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. PLoS Comput Biol. 2010; 6(8):e1000910. https://doi.org/10.1371/journal.pcbi.1000910 PMID: 20865155

32. Dobner J, Kaser S. Body mass index and the risk of infection-from underweight to obesity. Clin Microbiol Infect. 2018; 24(1):24–8. https://doi.org/10.1016/j.cmi.2017.02.013 PMID: 28232162

33. Plotkin SA. Correlates of vaccine-induced immunity. Clin Infect Dis. 2008; 47(3):401–9. https://doi.org/10.1086/589862 PMID: 18558875

34. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020; 26(8):1205–11. https://doi.org/10.1038/s41591-020-0962-9 PMID: 32546824

35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

36. Dwyer G, Dushoff J, Elkinton JS, Levin SA. Pathogen-driven outbreaks in forest defoliators revisited: building models from experimental data. Am Nat. 2000; 156(2):105–20. https://doi.org/10.1086/303379 PMID: 10856195

37. Ben-Ami F, Regoes RR, Ebert D. A quantitative test of the relationship between parasite dose and infection probability across different host–parasite combinations. Proc R Soc B. 2008; 275:853–9. https://doi.org/10.1098/rspb.2007.1544 PMID: 18198145

38. Pessoa D, Souto-Maior C, Gjini E, Lopes JS, Ceña B, Codeço CT, et al. Unveiling time in dose-response models to infer host susceptibility to pathogens. PLoS Comput Biol. 2014; 10(8):e1003773. https://doi.org/10.1371/journal.pcbi.1003773 PMID: 25121762

39. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4):2025–35. https://doi.org/10.1093/genetics/162.4.2025 PMID: 12524368

40. Kennedy DA, Dukic V, Dwyer G. Combining principal component analysis with parameter line-searches to improve the efficacy of Metropolis–Hastings MCMC. Environ Ecol Stat. 2015; 22:247–74. https://doi.org/10.1007/s10651-014-0297-0

41. Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A. Beyond R_0: heterogeneity in secondary infections and probabilistic epidemic forecasting. J R Soc Interface. 2020; 17:20200393. https://doi.org/10.1098/rsif.2020.0393 PMID: 33143594

42. Shaw CL, Kennedy DA. What the reproductive number R_0 can and cannot tell us about COVID-19 dynamics. Theor Popul Biol. 2021; 137:2–9. https://doi.org/10.1016/j.tpb.2020.12.003 PMID: 33417839

43. De Serres G, Shadmani R, Duval B, Boulianne N, Déry P, Fradet MD, et al. Morbidity of pertussis in adolescents and adults. J Infect Dis. 2000; 182(1):174–9. https://doi.org/10.1086/315648 PMID: 10882595

44. Rieder H. Contacts of tuberculosis patients in high-incidence countries. Int J Tuberc Lung Dis. 2003; 7: S333–6. PMID: 14677818

45. Taylor MM, Rotblatt H, Brooks JT, Montoya J, Aynalem G, Smith L, et al. Epidemiologic investigation of a cluster of workplace HIV infections in the adult film industry: Los Angeles, California, 2004. Clin Infect Dis. 2007; 44(2):301–5. https://doi.org/10.1086/510487 PMID: 17173235

46. Lessler J, Reich NG, Cummings DA, of Health NYCD, Team MHSII. Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. N Engl J Med. 2009; 361:2628–36. https://doi.org/10.1056/NEJMoa0906089 PMID: 20042754

47. Ajelli M, Parlamento S, Bome D, Kebbi A, Atzori A, Frasson C, et al. The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: epidemiology and impact of interventions. BMC Med. 2015; 13. 281. https://doi.org/10.1186/s12916-015-0524-z PMID: 26607790

48. Koh WC, Naing L, Chaw L, Rosledzana MA, Alikhan MF, Jamaludin SA, et al. What do we know about SARS-CoV-2 transmission? A systematic review and meta-analysis of the secondary attack rate and associated risk factors. PLoS One. 2020; 15(10):e0240205. https://doi.org/10.1371/journal.pone.0240205 PMID: 33031427

49. Flasche S, Hens N, Boëlle PY, Mossong J, van Ballegooijen WM, Nunes B, et al. Different transmission patterns in the early stages of the influenza A (H1N1) v pandemic: A comparative analysis of 12 European countries. Epidemics. 2011; 3(2):125–33. https://doi.org/10.1016/j.epidem.2011.03.005 PMID: 21624784

50. Franco N, Coletti P, Willem L, Angeli L, Lajot A, Abrams S, et al. Inferring age-specific differences in susceptibility to and infectiousness upon SARS-CoV-2 infection based on Belgian social contact data. PLoS Comput Biol. 2022; 18(3):e1009965. https://doi.org/10.1371/journal.pcbi.1009965 PMID: 35353810

51. Zhu W, Wen Z, Chen Y, Gong X, Zheng B, Liang X, et al. Age-specific transmission dynamics under suppression control measures during SARS-CoV-2 Omicron BA. 2 epidemic. BMC Public Health. 2023; 23(1):743. https://doi.org/10.1186/s12889-023-15596-w PMID: 37087436

52. Yang S, Rothman RE. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. Lancet Infect Dis. 2004; 4(6):337–48. https://doi.org/10.1016/S1473-3099(04)01044-8 PMID: 15172342

53. Cohen AN, Kessel B, Milgroom MG. Diagnosing SARS-CoV-2 infection: the danger of over-reliance on positive test results. medRxiv [Preprint]. 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.04.26.20080911v4

54. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438:355–9. https://doi.org/10.1038/nature04153 PMID: 16292310

55. Leon AE, Hawley DM. Host responses to pathogen priming in a natural songbird host. EcoHealth. 2017; 14:793–804. https://doi.org/10.1007/s10393-017-1261-x PMID: 28766063

56. Cohn SK Jr. Syphilis: Naming and Blaming? In: Epidemics: Hate and Compassion from the Plague of Athens to AIDS. Oxford University Press; 2018.

57. Cohn SK Jr. Plague Disputes, Challenges of the 'Universals'. In: Cultures of Plague: Medical thinking at the end of the Renaissance. Oxford University Press; 2009.