medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Title

High variability in transmission of SARS-CoV-2 within households and implications for control

Authors

Damon J.A. Toth[1,2,3], Alexander B. Beams[1,3], Lindsay T. Keegan[1,2], Yue Zhang[1], Tom Greene[1], Brian Orleans[1], Nathan Seegert[4], Adam Looney[4], Stephen C. Alder[5], Matthew H. Samore[1,2]

Affiliations

[1]Department of Internal Medicine, University of Utah School of Medicine

[2]Department of Veterans Affairs Salt Lake City Healthcare System

[3]Department of Mathematics, University of Utah

[4]Department of Finance, University of Utah David Eccles School of Business

[5]Department of Family and Preventive Medicine, University of Utah School of Medicine

Correspondence

Damon J.A. Toth

295 Chipeta Way

Salt Lake City, Utah 84132

Damon.Toth@hsc.utah.edu

Abstract

Background. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses a high risk of transmission in close-contact indoor settings, which may include households. Prior studies have found a wide range of household secondary attack rates and may contain biases due to simplifying assumptions about transmission variability and test accuracy.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Methods. We compiled serological SARS-CoV-2 antibody test data and prior PCR test reporting from members of more than 9000 Utah households. We paired these data with a probabilistic model of household importation and transmission. We calculated a maximum likelihood estimate of the importation probability, mean and variability of household transmission probability, and sensitivity and specificity of test data. Given our household transmission estimates, we estimated the threshold of non-household transmission required for epidemic growth in the population.

Results. We estimated that individuals in our study households had a 0.38% (95% CI 0.30% – 0.48%) chance of acquiring SARS-CoV-2 infection outside their household. Our household secondary attack rate estimate was 35% (26% – 47%), substantially higher than the crude estimate of 15% unadjusted for imperfect serological test specificity and other factors. We found evidence for high variability in individual transmissibility, with higher probability of no transmissions or many transmissions compared to standard models. With household transmission at our estimates, the average number of non-household transmissions per case must be kept below 0.40 (0.32 – 0.51) to avoid continued growth of the Utah epidemic.

Conclusions. Our findings suggest that crude estimates of household secondary attack rate based on serology data without accounting for false positive tests may underestimate the true average transmissibility, even when test specificity is high. Our finding of potential high variability (overdispersion) in transmissibility of infected individuals is consistent with characterizing SARS-CoV-2 transmission being largely driven by superspreading from a minority of infected individuals. Mitigation efforts targeting large households and other locations where many people congregate indoors might curb continued spread of the virus.

1 Introduction

Since its emergence in 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for COVID-19, has spread rapidly, causing severe morbidity, mortality, and disruption to daily life. As public health officials continue grappling with reducing community spread, it is of increased importance to understand transmission risk in different locations where people mix. Transmission within households may be especially important, given the mounting evidence that indoor environments with close, sustained contact are especially high risk for SARS-CoV-2 transmission [1-3]. Furthermore, with substantial observed decreases in mobility during the

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

pandemic [4], individuals likely are spending a greater proportion of time at home, thus increasing the importance of understanding within-household transmission. Likewise, isolation and quarantine measures recommended to help control COVID-19 frequently occur within homes, increasing risk to susceptible household members [5].

Data collected from members of households with at least one person infected with SARS-CoV-2 have revealed a wide range of within-household transmission estimates. One systematic review and meta-analysis [6] found 24 studies with household data conducted from January-March 2020, mostly in China, with secondary attack rate estimates ranging from 5% to 90% in the individual studies; pooling these data led to an average secondary attack rate estimate of 27% (95% CI: 21% – 32%). Another published review and meta-analysis of more recent data found 22 studies on the secondary attack rate in households, including estimates ranging from 4% to 32% [7]. Pooling these studies, the review found an average secondary attack rate of 17.1% (95% CI: 13.7% – 21.2%). Another review and meta-analysis found 40 household studies with individual study estimates ranging from 4% to 45% [8]. Their pooled analysis found that the household-based secondary attack rate for all household contacts was 19.0% (95% CI: 14.9 – 23.1%). Data from households in the U.S. [9-12] produced secondary attack rate estimates from 11% to 53%.

Most household studies generated data by first identifying index household cases via active or passive surveillance followed by monitoring and PCR testing of their household contacts. These studies may exhibit bias if mild or asymptomatic cases were less likely to be identified as an index household case. By contrast, data for the presence of antibodies among household members provide information on the distribution of final sizes of household outbreaks no longer in progress and in which some or none of the cases were identified at the time. We are aware of only 3 studies that used serological antibody data to estimate household transmission, using data from Spain [13], Brazil [14], and Switzerland [15].

In addition to average transmission rates, heterogeneity and variability in SARS-CoV-2 transmission have also been quantified. The amount of individual-level variation in the number of secondary infections can affect final outbreak size [16]. Large variation (i.e., overdispersion) indicates the presence of superspreading by a minority of individuals who transmit to a disproportionately large number of others [17]. Better understanding of superspreading individuals and locations can greatly enhance efficient targeting of transmission control strategies [18]. Backward contact tracing can efficiently trace sources of acquisition to high-transmission individuals and circumstances when

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

superspreading is present [19], and efforts that target similar circumstances for transmission prevention can have disproportionate benefits [20, 21].

Studies have quantified the variability in the number of SARS-CoV-2 transmissions from infected individuals using the dispersion parameter k, governing the variance of a negative binomially distributed offspring distribution [22-26]. Those studies estimated high overdispersion (low values of k) similar to what was observed during the first SARS-CoV outbreak in 2003 [17]. These estimates were derived from data on transmissions, including superspreading events, occurring in a variety of locations both inside and outside of households. Regarding household transmission specifically, Madewell et al. [8] showed preliminary evidence of overdispersion in household data, with more households than expected experiencing extremes of transmission (i.e., either no transmission or many transmissions) from an introduced case.

In this study, we combine SARS-CoV-2 data from serological antibody tests and self-reported prior PCR tests to estimate within-household transmission of COVID-19 in Utah. Previously published secondary attack rate estimates are largely based on crude formulae which ignore the probabilities of multiple members of a household acquiring infection from the community, multiple generations of transmission within the household (i.e. secondary, tertiary, etc. transmissions), and imperfect test sensitivity and specificity. We addressed these limitations by extending previous models of final household outbreak size distributions [27] to develop a novel probabilistic model of household importation and household transmission combined with test sensitivity and specificity. Our model also quantifies variability in household transmission and the potential extent of overdispersion, to shed light on superspreading phenomena and the implications of household transmission for population-level controllability of COVID-19.

2 Methods

2.1 Data collection from Utah households

Details of our data collection process are described elsewhere [28]. Briefly, the Utah Health & Economic Recovery Outreach project involved selecting households in several counties in Utah by population sampling designed to form a set of households by which average community seroprevalence could be assessed. Any member of selected households age 12 or older were offered the opportunity to participate in a survey and serological COVID-19

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

antibody testing. Household members declining to provide a serological sample could opt to fill out the survey, which included the opportunity to report prior PCR test results for COVID-19. Data included in this analysis were collected in the months of May, June, and July, 2020.

The data are represented as follows. For each household in the dataset, we captured the following 7 values from the data:

- n: total number of people in household
- a: number who were antibody tested
- s: number who responded to the survey but were not antibody tested
- a_{PP}: number who reported a prior positive PCR result and received a positive antibody test
- a_{PN}: number who reported a prior positive PCR result and received a negative antibody test
- a_{NP}: number who reported no prior positive PCR result and received a positive antibody test
- s_P: number who were surveyed, reported a prior positive PCR test, and did not receive an antibody test

Each of the C unique combinations of the above 7 values found at least once in the dataset was indexed as a vector $\mathbf{y}_i$:

$$\mathbf{y}_i = (n_i, a_i, s_i, a_{PPi}, a_{PNi}, a_{NPi}, s_{Pi})$$

We tallied the number of households for which each $\mathbf{y}_i$ occurred in the frequency elements f_i, and represented the entire dataset by the vector $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_C, f_1, \dots f_C)$.

The dataset $\mathbf{y}$ and all codes, written in R version 4.0.3, used for analyses described in the following sections are posted and publicly available at https://github.com/damontoth/householdTransmission.

2.2 Total household infection size model

Here we derive the probabilities M_{kn} for the probability that k out of n total household members ended up infected. If k members of a size-n household were infected, that means that $n - k$ members escaped being infected by a non-household member (called "community" acquisitions) and escaped being infected by any of the n infected within the household. Thus, our model for M_{kn} combines both probabilities and does not depend on the order of occurrence of household transmissions and subsequent community acquisitions after the initial one, as in similar prior formulations

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

[27]. Also following prior formulations, we assume that active infections were not present in the households at the time of antibody data collection (i.e., that household outbreaks had reached final size), consistent with the low prevalence of active infections at the time of data collection [28].

The M_{kn} values depend on 3 parameters. The parameter p_c is the average per-capita probability of community acquisition, p_h is the mean transmission probability from an infected person to a fellow household member, and d_h is the dispersion parameter characterizing variability in transmissibility across infected individuals, with no assumed correlation among members of the same household.

For a given household size $n \geq 2$, the formula for M_{kn} is:

$$M_{kn}(p_c, p_h, d_h) = \begin{cases} (1-p_c)^n, & k = 0 \\ \sum_{i=1}^{k} \binom{n}{i} (p_c)^i (1-p_c)^{n-i} T_{i,k-i,n-i}(p_h, d_h), & k = 1, \dots, n-1 \\ 1 - \sum_{k=0}^{n-1} M_{kn}(p_c, p_h, d_h), & k = n \end{cases}$$

For households of size $n = 1$, note that the expression involving the household transmission parameters does not apply and we have $M_{01} = 1 - p_c$ and $M_{11} = p_c$.

The probability that a household of size n had 0 infections: $M_{0n} = (1-p_c)^n$, is the probability that none of the household members acquired infection from the community and does not depend on the household transmission variables because no household transmissions were possible without a community introduction. For the final number of household infections to be nonzero, there must be at least one community acquisition, which may be followed by within-household transmissions. The $\binom{n}{i}(p_c)^i(1-p_c)^{n-i}$ expression is the binomial probability that i out of the n household members had a community acquisition, and the function T_{xyz} is the probability that x already infected household members lead to a total of y transmissions to z susceptible household members. In other words, T_{xyz} is the probability that the final outbreak size is $x + y$, given that x household members are already infected in a house with z susceptible members. For efficiency of computation, the T_{xyz} values are calculated in order of increasing values of y, i.e. T_{x0z} for each relevant x and z value are calculated first, then the T_{x1z} values, then T_{x2z}. This allows the use of T_{xyz} values for lower values of y to be used in the formula (see Supplementary Material for details):

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

$$T_{xyz}(p_h, d_h) = \begin{cases} H_{x0z}(p_h, d_h), & y = 0 \\ \sum_{i=0}^{y-1} H_{x,y-i,z}(p_h, d_h) T_{y-i,i,z-y+i}(p_h, d_h), & y = 1, \dots, n \end{cases}$$

Within the T_{xyz} formula, the function H_{xyz} is the probability that x infected household members transmit infection *directly* to y out of z fellow household members who are susceptible. The H_{xyz} values are calculated in order of increasing values of x for efficient computation (see Supplementary Material):

$$H_{xyz}(p_h, d_h) = \begin{cases} F_{yz}(p_h, d_h), & x = 1 \\ \sum_{i=0}^{y} H_{x-1,i,z}(p_h, d_h) H_{1,y-i,z-i}(p_h, d_h), & x = 2, \dots, n-1 \end{cases}$$

Finally, the function $F_{yz}(p, d)$ is the probability mass function of the beta-binomial distribution for y successes out of z trials, parameterized by a mean success probability p and a dispersion parameter d. When d is finite and nonzero, F_{yz} is derived from the binomial distribution with success probability that is a beta-distributed random variable with parameters $\alpha = dp, \beta = d(1-p)$, with decreasing variance as d increases. We also make use of the boundary cases $d = 0$ and $d \to \infty$. In the limit $d \to \infty$, holding p constant, F_{yz} becomes the binomial distribution with constant success probability p (Supplementary Material). In the maximal variance limit, $d \to 0$, with p held constant, F_{yz} becomes an "all-or-nothing" distribution where $y = z$ successes occur with probability p and to $y = 0$ success occur with probability $1 - p$ (Supplementary Material):

$$F_{yz}(p, d) = \begin{cases} \begin{cases} 1-p, & y = 0 \\ 0, & 0 < y < z, \\ p, & y = z \end{cases} & d = 0 \\ \binom{z}{y} \frac{\mathrm{B}\big(y + dp, z - y + d(1-p)\big)}{\mathrm{B}\big(dp, d(1-p)\big)}, & 0 < d < \infty \\ \binom{z}{y} p^y (1-p)^{z-y}, & d \to \infty \end{cases}$$

The function B is the beta function. We use F_{yz} within the formula for H_{1yz} to quantify the distribution of household transmissions directly from a single infected household member, where y is the number of transmissions, z is the number of susceptible household members, $p = p_h$, and $d = d_h$.

The above formulas are derived in the Supplementary Material. Elements of this model appear in other publications. Longini and Koopman [27] derived a formula for M_{kn} for the model with no variability among households or

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

individuals, equivalent to our model with $d_h \to \infty$. While they provided a more efficient formula that takes advantage of the properties of that special case, we confirmed that our calculation scheme above reproduces the results of their formula. Becker [29] published explicit formulas for the final size of household outbreaks after a single introduction to households up to size 5 using the beta-binomial chain model, equivalent to our T_{xyz} for $x = 1$ and z up to 4. We confirmed that our scheme for calculating T_{xyz} produces the same results as their example formulas for arbitrary values of p_h and d_h.

2.3 Likelihood model

We sought to use our data to simultaneously estimate the 3 parameters (p_c, p_h, d_h) using maximum likelihood estimation (MLE). However, applying the M_{kn} formula directly to our data would be problematic because the true number of infections k in each household are not known with certainty. The data include two sources of COVID-19 test information by which prior infection status of a portion of individual household members can be probabilistically inferred: antibody test results and surveys in which participants could report results of a prior PCR test.

Antibody test results are subject to imperfect sensitivity and specificity due to false negative tests and false positive tests, respectively. To account for these, we added two additional parameters to be estimated by the MLE: ϕ_A, the probability that an antibody-tested person with a prior infection tested positive for antibodies, and π_A, the probability that an antibody-tested person with no prior infection tested negative for antibodies.

Prior PCR test results for COVID-19 virus reported on the survey also do not perfectly identify those with prior infections. To quantify this imperfection, we introduced two more parameters to be estimated by the MLE: ϕ_V, the probability that a surveyed person with a prior infection reported receiving a positive viral PCR test, and π_V, the probability that a surveyed person with no prior infection did not report receiving a positive viral PCR test.

Some household members received a survey but no antibody test and other members received neither. The M_{kn} formula depends on the total household size n, which for many households includes individuals with missing data. For households with at least one but not all members infected $(1 \leq k \leq n - 1)$ and in which less than n member were full participants, the likelihood formula required the probability that different portions of the k infected members were among those who were antibody tested or surveyed only. To arrive at our formula, we assumed that

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the antibody-tested and surveyed-only portion of a household were a random sample of household members with respect to their prior infection status. I.e., we assumed that those individuals in a participating household with and without prior infections were equally likely to participate in the study and equally likely to agree to antibody testing.

In all we have 7 variables to be estimated by MLE, encapsulated in the following vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (p_c, p_h, d_h, \phi_V, \phi_A, \pi_V, \pi_A)$$

The log likelihood of the dataset $\mathbf{y}$ described in Section 2.1 with variable set $\boldsymbol{\theta}$ is then

$$\ln \mathcal{L}(\mathbf{y}|\boldsymbol{\theta}) = f_1 \ln \mathcal{L}(\mathbf{y}_1|\boldsymbol{\theta}) + \cdots + f_C \ln \mathcal{L}(\mathbf{y}_C|\boldsymbol{\theta})$$

To present the formula for $\mathcal{L}(\mathbf{y}_i|\boldsymbol{\theta})$, the likelihood of a particular $\mathbf{y}_i$, we first define the following quantities calculated from the core elements of $\mathbf{y}_i$ listed in Section 2.1:

- $a_{NNi} = a_i - a_{PPi} - a_{PNi} - a_{NPi}$: number who reported no prior positive PCR result and received a negative antibody test
- $s_{Ni} = s_i - s_{Pi}$: number who were surveyed, reported no prior positive PCR result, and did not receive an antibody test
- $q_i = n_i - a_i - s_i$: number untested for antibodies and not surveyed

Then we have:

$$\mathcal{L}(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{u=0}^{a_{PPi}} \sum_{v=0}^{a_{PNi}} \sum_{w=0}^{a_{NPi}} \sum_{x=0}^{a_{NNi}} A(u, v, w, x; \phi_V, \phi_A, \pi_V, \pi_A) \sum_{y=0}^{s_{Pi}} \sum_{z=0}^{s_{Ni}} S(y, z; \phi_V, \pi_V) \sum_{k=u+v+w+x+y+z}^{u+v+w+x+y+z+q_i} H(k, u+v+w+x, y+z) M_{kn}(p_c, d_c, p_h, d_h)$$

In the formula, the function A quantifies the probability of observing the given set of test result combinations among antibody-tested people $(a_{PPi}, a_{PNi}, a_{NPi}, a_{NNi})$, given that (u, v, w, x) of them had a prior infection, respectively. E.g., u is the number of the a_{PPi} household member who had an infection (true positives), v is the number of the a_{PNi} household members who had an infection (true positive by PCR and false negative by antibody test), w is the number of the a_{NPi} household members who had an infection (true positive by antibody test and did not report a positive PCR test), and x is the number of the a_{NNi} household members who had an infection (false negative by antibody test and did not report a positive PCR test). The formula for A is

$$A(u, v, w, x; \phi_V, \phi_A, \pi_V, \pi_A) = f_m\big(u, v, w, x; \mathbf{p}_{\mathrm{I}}(\phi_V, \phi_A)\big) f_m\big(a_{PPi} - u, a_{PNi} - v, a_{NPi} - w, a_{NNi} - x; \mathbf{p}_{\mathrm{U}}(\pi_V, \pi_A)\big)$$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

The function $f_m(\mathbf{r};\mathbf{p})$ is the probability mass function for the multinomial distribution, where the number of trials is the sum of the elements of $\mathbf{r}$, which are the number of infected or uninfected antibody-tested people who received each of the four possible test result combinations. The vector $\mathbf{p}$ contains the probability of each of the four test result combinations given that the person was infected (for $\mathbf{p} = \mathbf{p}_\mathrm{I}$) or uninfected (for $\mathbf{p} = \mathbf{p}_\mathrm{U}$):

$$\mathbf{p}_\mathrm{I}(\phi_V, \phi_A) = \left(\phi_V\phi_A, \phi_V(1-\phi_A), (1-\phi_V)\phi_A, (1-\phi_V)(1-\phi_A)\right)$$

$$\mathbf{p}_\mathrm{U}(\pi_V, \pi_A) = \left((1-\pi_V)(1-\pi_A), (1-\pi_V)\pi_A, \pi_V(1-\pi_A), \pi_V\pi_A\right)$$

The first element of $\mathbf{p}_\mathrm{I}$, $\phi_V\phi_A$, is the probability that an antibody-tested person with a prior infection reported a positive PCR test (with probability ϕ_V) and also had a positive antibody test result (with probability ϕ_A). Note that ϕ_A represents the sensitivity of the antibody test, but ϕ_V includes both the sensitivity of the PCR test and the probability that an infected person actually sought and received a PCR test during the period of infection in which detectable virus was present and reported that positive test on our survey. Elements 2–4 of $\mathbf{p}_\mathrm{I}$ are the probabilities that an antibody-tested, prior infected person reported a positive PCR test but tested negative for antibodies, did not report a positive PCR test and tested positive for antibodies, and did not report a positive PCR test and tested negative for antibodies, respectively. The elements of $\mathbf{p}_\mathrm{U}$ are the corresponding probabilities for individuals with no prior infection.

The function S quantifies the probability of the survey-only data (s_{Pi}, s_{Ni}) given that y of the s_{Pi} individuals had a prior infection and z of the s_{Ni} individuals had a prior infection:

$$S(y, z; \phi_V, \pi_V) = f_b(y; y+z, \phi_V) f_b(s_{Ni} - z; s_{Ni} - z + s_{Pi} - y, \pi_V)$$

The function $f_b(q; r, p)$ is the probability mass function for the binomial distribution, for q successes given that there were r independent trials with probability p for success of each trial.

The function $H(k, k_a, k_s)$ in the likelihood equation is the probability that, when k of n_i individuals in the household were infected, k_a infected individuals were among the a_i individuals antibody tested and k_s infected individuals were among the s_i individuals surveyed but not antibody tested:

$$H(k, k_a, k_s) = f_h(k_a; k, n_i - k, a_i) f_h(k_s; k - k_a, n_i - k - (a_i - k_a), s_i)$$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

The function $f_h(b; c, d, e)$ is the probability mass function of the hypergeometric distribution for the number b of infected people selecting to be antibody-tested or surveyed-only, given that there were c infected people and d uninfected people available for selection in the household, and e people were tested or surveyed-only. These terms account for individuals in households who received neither an antibody test nor a survey, who may have included infected individuals. Our use of the hypergeometric distribution led from our assumption that, if some members of the household had a prior infection and others didn't, the antibody-tested / surveyed individuals were a random sample from the household with respect to their prior infection status.

2.4 Likelihood optimization and uncertainty

We minimized the log likelihood over the 7 unknown parameters $(p_c, p_h, d_h, \phi_V, \phi_A, \pi_V, \pi_A)$ using the observations $(n, a, s, a_{PP}, a_{PN}, a_{NP}, s_P)$ for each household, to produce the MLE: $\hat{\theta} = (\hat{p}_c, \hat{p}_h, \hat{d}_h, \hat{\phi}_V, \hat{\phi}_A, \hat{\pi}_V, \hat{\pi}_A)$. We derived approximate confidence interval boundaries for an individual parameter θ_i using the likelihood ratio test, using the statistic $2\log(\mathcal{L}(\hat{\theta})/\mathcal{L}(\theta))$, where θ consists of θ_i freely varying and the other 6 elements of θ held at their optimal value. We defined a 95% confidence interval boundary where θ_i produces a value for this statistic equal to the 95th percentile of the chi-squared distribution with 1 degree of freedom. We also plotted 2-dimensional confidence region boundaries for each of the 21 possible (θ_i, θ_j) parameter pairs by allowing each pair to vary freely together while holding the other 5 at their optimal values. We calculated the boundary in the (θ_i, θ_j) parameter plane where the likelihood ratio statistic equals the 95th percentile of the chi-squared distribution with 2 degrees of freedom. To calculate P-values at which certain fixed parameter values could be rejected in favor of the MLE, we used the chi-squared distribution with degrees of freedom equal to the number of fixed parameters.

Additionally, we developed a simulation model to produce synthetic data sets on which to test our likelihood model. We ran the simulation for the same number of households with the same sizes and participation rates for survey and antibody testing as in the actual data (fixed values of n, a, and s for each household). We randomized importations to households and simulated transmissions using the MLE values of the three epidemiological parameters, randomized survey and antibody test results using the MLE sensitivity and specificity values, and maximized the likelihood against the simulated data. We repeated this process for 500 simulated data sets and recorded the median estimated value of each variable, for comparison against the MLE value that generated the data. We also used the

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

500 sets of simulation-based estimates as a parametric bootstrap to generate 95% confidence estimates for each variable, for comparison against the intervals generated from the likelihood ratio test.

Finally, we tested an alternate model that allows the community acquisition probability to vary by household, such that some households may have a higher per-capita acquisition rate than others applied to each household member. To quantify this probability in the alternate model, we employed the beta-binomial distribution for the number of community acquisitions in a household of a given size (see Supplementary Methods).

2.5 Household transmission variability

We quantified the implications of our household transmission variability estimates by calculating the probability of transmission extremes, compared to those produced by the classic binomial transmission model ($d_h = \infty$). Specifically, we calculated the probability that an initially infected individual transmits to no one or everyone in households of sizes from 2 to 10. For households of size n, the probability of no transmissions from the index infection is $F_{0,n-1}(p_h, d_h)$ and the probability the index person transmits directly to the entire household is $F_{n-1,n-1}(p_h, d_h)$. We used our overall MLE values for $\hat{p}_h$ and $\hat{d}_h$ to calculate these values for each n, with confidence intervals using our parametric bootstrap results. For comparison to the binomial model we applied $d_h = \infty$, paired with the alternate MLE of p_h under that constraint.

We also calculated an example of a dynamic transmission model that produces a distribution of household transmission probabilities close to that produced by our MLE beta distribution, using the method of moments. Specifically, if an infected person's duration of infectiousness is assumed to be fixed and transmissibility to a housemate is modeled as a gamma distribution with shape parameter k, then we solve for the value k that produces the same mean and variance for the transmission probability as that of the beta distribution with mean p_h and dispersion d_h (Supplementary Methods). We solved for k using our MLE $\hat{p}_h$ and $\hat{d}_h$ values, and we derived a confidence interval for k using the pairs of (p_h, d_h) estimates from our parametric bootstrap analysis.

2.6 Within-household reproduction number

We calculated the within-household reproduction number R_h, defined as the expected number of household transmissions directly from a community acquirer with all fellow household members susceptible:

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

$$R_h = p_h(\mu + \sigma^2/\mu - 1),$$

where μ and σ^2 are the mean and variance of the household size distribution, and p_h is the secondary attack rate as determined by our MLE. This equation for R_h is derived as in Ball et al. [30] and detailed in the Supplementary Material.

Additionally, we derived an alternate household reproduction number R_h^* defined as the expected total number of transmissions in the household of an infected person who acquired infection in the community and has no initially non-susceptible housemates. This differs from R_h in that it counts all potential downstream transmissions in the household stemming from the index community acquirer. The formula for R_h^*, derived in the Supplementary Material, is

$$R_h^* = \sum_{i=1}^{N-1} \frac{(i+1)h_{i+1}}{\mu} \sum_{j=1}^{i} jT_{1ji}(p_h, d_h)$$

To investigate the implications of household transmission for population-wide transmission control, we use a threshold condition delineating subcritical and supercritical transmission in the population. Supercritical transmission occurs when $R_c > 1/(R_h^* + 1)$, where R_c is the average number of community (non-household) transmissions from an infected person. We derive this formula in the Supplementary Material, following Ball et al. [30]. We estimated R_h, R_h^*, and the threshold value for R_c by applying our MLE estimates of p_h and d_h to the above formulas and their confidence intervals by applying the (p_h, d_h) pairs from each parametric bootstrap estimate.

3. Results

3.1 Data summary

We compiled data from 9,383 households. Of these, we retained 9,091 (96.9%) for use in the MLE. The 292 excluded households were removed because the household size was unknown (192) or the reported household size was less than the number of people tested or surveyed in the house (100). In the 9,091 retained households, there were 27,998 (3.08 per household) reported household members, 13,712 (1.51 per household) people who were both surveyed and antibody tested, and another 5,262 (0.58 per household) who were surveyed but not antibody tested.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Of the 13,712 antibody tests in the retained households, 170 (1.24%) were positive. Of those 170 people with a positive antibody test, 55 (32.4%) reported receiving a positive PCR test. Of the 18,974 people who were antibody tested or surveyed only, 117 (0.62%) reported receiving a positive PCR test. This broke down to 0.53% (72 / 13,712) for those who were antibody tested and 0.86% (45 / 5,622) for those who were surveyed but not antibody tested. The rate of testing positive for antibodies among those reporting a prior positive PCR test was 76.4% (55 / 72). The rate of survey participants agreeing to antibody testing was lower for those who reported a prior positive PCR test compared to those who did not: 61.5% (72 of 117) vs. 72.7% (5217 of 18,857), a small but statistically significant ($P < 0.01$) difference in proportion.

Of the retained households, 189 (2.1%) had at least one household member who either tested positive for antibodies or reported a prior positive PCR test. There were 157 households with exactly 1 positive member (by either antibody test or reported PCR test or both), 25 households with 2 positives, 5 with 3 positives, 1 with 4 positives, and 1 with 6 positives. In all, there were $C = 270$ unique $\mathbf{y}_i$ vectors representing household data described in section 2.1.

The crude SAR measure derived from antibody testing only (fraction of antibody-tested housemates of antibody-positive household members who were also antibody positive) was 13.7% (26 / 190). The crude SAR estimate from reported PCR data only (fraction of surveyed housemates of people reporting a positive PCR test who also reported a positive PCR test) was 22.6% (30 / 133). When combining both types of data, the crude SAR estimate (fraction of surveyed / tested housemates of any antibody-positive or reported PCR-positive person who were positive by either or both measures) was 14.8% (43 / 290).

3.2 Maximum likelihood estimates

The MLE for p_c, the per-person community acquisition probability from outside the household, was 0.38% (0.30% – 0.48%). For within household transmission probability, the MLE produced an average secondary attack rate estimate $p_h = 35\%$ (26% – 47%). The MLE for the dispersion parameter d_h, quantifying variability in transmissibility by person, was 0.54 (0.06 – 2.6). The boundary case $d_h = \infty$, representing the classic binomial household transmission model with no variability in individual infectiousness [27], could be rejected with $P = 0.002$.

Our MLE result for ϕ_V, the probability that a surveyed person with a prior infection reported testing PCR-positive, was 76% (66% – 86%). For ϕ_A, the probability that a prior-infected person's antibody test was positive, the MLE

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

was 85% (75% – 93%). For π_V, the probability that a surveyed person with no prior infection reported no PCR-positive test, the MLE was 99.93% (99.87% – 99.98%). For π_A, the probability that an antibody-tested person with no prior infection tested negative for antibodies was 99.3% (99.1% – 99.4%).

When optimizing the likelihood equation against 500 synthetic data sets simulated using the MLE variable assumptions, the median estimates of each parameter were very close to the MLE values (Table 1). The confidence intervals derived from these bootstrap estimates were similar to those derived from the likelihood ratio test, though the bootstrap intervals were somewhat wider for the three parameters governing importation and transmission. Likewise, the likelihood ratio-based intervals reported in Table 1 expanded modestly when we calculated 2-dimensional confidence regions based on each pair of estimated parameters, with most regions exhibiting close to symmetric shapes around the MLE (Supplemental Figures). Notably, the 95% confidence regions involving the transmission dispersion parameter d_h can extend to the high-variability boundary $d_h = 0$, a result that is also reflected by the fact that the MLE for the model with fixed $d_h = 0$ cannot be rejected with 95% confidence (P = 0.09) (Table 2).

Our alternate model that employed a beta-binomial distribution for the number of household acquisitions, using a new dispersion parameter d_c estimated as an additional variable in the MLE, found $d_c = 2.4$ (0.95 – 9.6), with somewhat altered estimates of the other parameters (Table S1) compared to those in Table 1. However, the log likelihood of the model in Table 1, which is equivalent to the alternate model with $d_c = \infty$, is sufficiently close to that of the alternate model that $d_c = \infty$ cannot be rejected by the likelihood ratio test and is favored by the Bayesian information criterion. However, if overdispersion in household community acquisitions does occur, the uncertainty ranges of the transmission variables p_h and d_h become large (see Supplementary Results).

3.3 Household transmission variability

We quantified the implications of our key finding of high transmission variability within households of persons infected with COVID-19 by calculating the probability of transmission extremes. Compared to our overall MLE, the classic binomial transmission model ($d_h = \infty$) produced a similar average secondary attack rate estimate of $p_h =$ 31% (24% – 41%). However, the binomial model produces substantially lower probabilities that an infected individual transmits to no one or everyone in larger households (Table 3).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

For example, our MLE model estimates that an infected member of a 8-member household would have a 44% (16% – 68%) chance of transmitting to no one, but a 17% (1% – 44%) chance of transmitting infection directly to all 7 housemates. By contrast, the no-variability binomial model estimate would be substantially lower for each extreme: 7% (3% – 15%) chance of transmitting to no one and 0.03% (0.004% – 0.2%) chance of transmitting to everyone (Table 3).

We calculated an example of a dynamic transmission model that would produce the same mean and variance of a person's transmission probability to a household member that is produced by our MLE beta distribution. If an infected person's duration of infectiousness is assumed to be fixed and transmissibility to a housemate is modeled as a gamma distribution with shape k, then $k = 0.21$ (95% CI 0 – 1.3) when the mean and variance are matched, regardless of the infectious duration (Supplementary Methods). This estimate of k is comparable to the dispersion parameter k of the negative binomial distribution commonly used to characterize overall variability in the number of transmissions from individuals, which can be derived from the Poisson distribution with a mean that is gamma-distributed with shape parameter k [17].

3.4 Within-household reproduction numbers

Our estimate of the household reproduction number R_h, the expected number of household transmissions from a community acquirer with no other infected fellow household members, depends on our estimate of p_h and the mean μ and variance σ^2 of the household size distribution. From our data we found $\mu = 3.08$ and $\sigma^2 = 3.13$, so our estimate is $R_h = 1.09$ (0.76 – 1.49). Our estimate of the alternate household reproduction number R_h^*, the expected total number of transmissions in the household of a community acquirer, is $R_h^* = 1.48$ (0.95 – 2.13).

The supercritical threshold for R_c, the average number of non-household transmissions by an infected individual, is approximated by $1/(R_h^* + 1)$ (see Methods section 2.5 and Supplementary Material). Using our estimate for R_h^* in Utah, this formula suggests that R_c must be kept below approximately 0.40 (0.32 – 0.51) to avoid increasing growth of COVID-19 infections in the population.

4. Discussion

The key findings of our analyses stem from our simultaneous estimation of the average and variability of SARS-CoV-2 household transmission, household importation, and test data accuracy. Our novel combination of those

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

interacting features within our model revealed two important epidemiological insights. First, we found that accounting for test error, especially the specificity of the serological antibody test, produced a substantially higher estimate for the household secondary attack rate. Second, we found evidence of substantial variability of transmissibility within households, which has important implications for understanding broad transmission patterns and mitigation strategies.

An important implication of the first finding is that assuming perfect test accuracy may be a source of underestimation for the household secondary attack rate in other studies. Our maximum likelihood estimate was 35% (26% – 47%), which is higher than recent pooled estimates of 17–19% from the most recent meta-analyses of worldwide household studies [7, 8]. These and other published studies have generally estimated the secondary attack rate by a simple calculation of the fraction of tests that were positive among household contacts of known cases. When we applied that calculation to our combined data, we found a crude secondary attack rate estimate of 14.8%. We traced the major source of this substantial underestimate to the assumption of perfect test specificity inherent in the crude formula. When we ran our MLE under the assumption of perfect specificity (no false positives) for the antibody test, the result for secondary attack rate was 16.6%, much closer to the crude estimate. Notably, the estimate of community acquisition probability when assuming perfect specificity was substantially *higher* than our overall MLE: 1.2% vs. 0.4%. Thus, our model suggests that allowing for false positives can shift the attribution of infections toward household transmissions and away from acquisitions outside the household.

Of note, we did not make independent assumptions for the serological test specificity or other test accuracy measures, but instead optimized them simultaneously with the epidemiological parameters in our likelihood model. For the antibody test specificity, our MLE of 99.3% (99.1% – 99.4%) was within the uncertainty range of the test manufacturer's estimate of 99.6% (99.0% – 99.9%) based on 4 positive tests from 997 samples collected prior to September 2019 [31]. Thus, our model independently produced a plausible estimate of the test specificity; if instead we had assumed the manufacturer's point specificity estimate directly, our secondary attack rate estimate would be 30%, still substantially higher than the crude estimate.

Our second major finding of overdispersion of household transmission stemmed from our use of the beta-binomial distribution to quantify the number of household transmissions from infected individuals. We quantified individual-level variability in transmissibility using a dispersion parameter d_h, and the optimal value occurred at low dispersion

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

(high variability; $d_h = 0.54$). The more commonly used binomial model, a special case of our model at minimal variability ($d_h \to \infty$), was rejected, suggesting that transmission patterns are not well captured by that simplifying assumption.

Our dispersion parameter estimate is not directly comparable to another commonly used dispersion parameter, often named k, that characterizes variability in the total number of transmissions (whether household or not) from each infected person as a parameter of the negative binomial distribution [17]. We attempted to relate d_h to k in the context of simple model in which the only source of variability is a person's transmissibility per unit time in contact with others, finding $k = 0.21$. This estimate is similar with point estimates for SARS-CoV-2 of $k = 0.1$ [22], $k = 0.25$ [23], and $k = 0.33$ [24]. This similarity perhaps suggests that variability in infectivity per time is a major driver of overall transmission variability for SARS-CoV-2. This could be consistent with findings that viral shedding is highly variable by individuals with SARS-CoV-2 infections, both during asymptomatic and symptomatic phases of disease, suggesting that heterogeneous transmissibility may be largely explained by overdispersion in levels of viral shedding by individuals [32]. However, other studies suggest that SARS-CoV-2 transmission overdispersion in the wider population beyond households may be less driven by biological heterogeneity and more by heterogeneous social contact behavior [33].

The level of within-household transmission variability captured by the parameter d_h affects the contribution of household transmission toward threshold levels of overall transmission. Threshold conditions are often expressed using a reproduction number (R), the average number of transmissions from each infected person. The average number of household transmissions directly from an initially infected household member (R_h) is independent of d_h, but d_h does affect the average number of household transmissions in the next generation, i.e. by someone who acquired infection from a housemate. When transmission variability is higher, the household transmission potential of a household acquirer is lower, reducing to zero in the "all-or-nothing" limit $d_h = 0$. To capture this effect, we introduced an alternate reproduction number R_h^*, which is the average number of total household transmissions after the initial introduction, when final household outbreak size has been reached.

Neither $R_h > 1$ nor $R_h^* > 1$ are sufficient threshold conditions for epidemic growth, which requires some level of between-household transmission to be sustained. Given our estimate of $R_h^* = 1.48$ (0.95 – 2.13), we can estimate the critical value of R_c, the average number of non-household community transmission that would push transmission for

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the population above the supercritical threshold for a growing epidemic, with the threshold condition $R_c > 1/(R_h^* + 1)$. Thus, we estimate that R_c must be kept below approximately 0.40 (0.32 – 0.51) to avoid a growing epidemic in Utah if household transmission continues to be well characterized by our model. As this result depended on the average household size in our data, it is notable that Utah has the highest state-average household size in the United States. The average household size in Utah is 3.1, about 20% higher than the national average household size. Thus, our R_h estimate may be high compared to other locations. A lower value of R_h would lead to a higher threshold value for R_c.

Our findings also provide further insight for interpretation of COVID-19 test results. In addition to the serological test specificity discussed above, our model also produced estimates of the sensitivity of this test. The MLE of $\phi_A =$ 85% (75% – 93%) is difficult to compare to the manufacturer's data, which were based on testing symptomatic, PCR-positive people for antibodies. Overall, 109 of 122 (89%) of those tests were positive for antibodies, but the results were highly dependent on the number of days post-symptom onset at which the serological sample was taken [31]. Because the symptom histories of the antibody-tested people in our data is largely uncertain, we cannot evaluate whether our result is consistent with the manufacturer's data, or whether our result can be reliably applied to other data.

We also produced estimates of the sensitivity and specificity of our survey data with respect to the probability that a reported prior-positive PCR test (or lack thereof) reflects the existence (or lack) of a prior infection. Our result of $\phi_V =$ 76% (66% – 86%) is our estimate of the percentage of individuals with SARS-CoV infections who tested PCR-positive test during their infection and were willing to report it on our survey. This may be a high identification fraction compared to other areas of the U.S.: one study estimated that less than 60% of symptomatic cases in the U.S. were identified during February-June 2020 [34]. Our result may reflect unusually successful case ascertainment efforts in Utah during the Spring and early Summer of 2020, perhaps partly owing to slower emergence compared to other regions. Notably, we found that a model assuming perfect specificity of the antibody test ($\pi_A =$ 100%) produced a dramatically lower estimate of $\phi_V =$ 37% (Table 2), which suggests that crude estimates based on serology data may underestimate case ascertainment without considering false positives.

Finally, our estimate for very high specificity $\pi_V =$ 99.93% (99.87% – 99.98%) of survey-reported PCR testing reflects the low probability of PCR tests producing false positives. It is possible that some false positives in our

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

survey data occurred by erroneous reporting, i.e. survey respondents reporting a prior positive PCR test that did not occur, rather than via errors in testing procedure. Even though our MLE for this parameter was in excess of 99.9%, we found that an alternate model assuming $\pi_V = 100\%$ produced notably different estimates of some of the other parameters, which suggests that studies producing epidemiological estimates relying on a 100% PCR specificity assumption should test robustness of conclusions to small deviations from that assumption.

This study has several limitations. Our estimate of high household transmission variability may not be robust to alternate assumptions for the way community acquisition risk varies by household. For example, some households could have been comprised of families with both parents working essential jobs during Spring/Summer 2020, with children attending in-person day care or camps, thus placing the entire household at much higher risk of community acquisition compared to households working / caring for children at home. Also, households could have high collective community acquisition probability via attending multi-household gatherings of extended family or other social groups. In these ways, households conceivably could vary considerably in their infection numbers for reasons that don't involve within-household transmission.

We tested the implications of this alternate possibility for household variability in our model by allowing variability in community acquisition by household using an additional dispersion parameter to the MLE model (Supplementary Results). Interestingly, the MLE for the transmission dispersion parameter d_h still occurred at high variability in household transmission ($\hat{d}_h = 0.30$) under this alternate model. Furthermore, the improvement in likelihood was not substantial, such that the more complicated model would not be favored by the likelihood ratio test nor the Bayesian information criterion. However, larger uncertainty ranges under the alternate model suggest that we may not be able to definitively rule out the possibility that variability in community acquisition risk by household plays a substantial role in explaining overall variability in household infection numbers.

It is also possible that household transmission variability could be driven by properties of households such as contact behavior, underlying health composition of household members, physical properties of the domicile such as size and ventilation, or other properties that could increase transmission risk of all household members together. Possible variability in person-to-person transmission probability by household, rather than by individual, is not accounted for in our model. Using a beta distribution for this probability across different households to arrive at an alternate final size distribution would require integrating the beta distribution over the full final size distribution equations

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

produced by the binomial-chain model, which would be complicated for larger households. Alternatively, one could model a functional relationship between observed properties of a household in the dataset and its average transmission probability, while retaining dispersion occurring at the individual level. We have not attempted this with our data; we suspect that the sample size of outbreaks in households with a given feature would not be large enough to draw meaningful conclusions.

Another limitation lies in our potentially inaccurate assumptions used to quantify the probability of prior infections among those with missing data within participating households. Most non-participating individuals within participating households were children under 12, who were not offered antibody tests nor surveyed. Thus, our model carries the assumption that children under 12 have equal community acquisition rates, susceptibility to acquisition from another household member, and transmissibility to other household members compared to study participants. Our assumption is consistent with studies finding similar transmission rates to and from children compared to adults. In a study of COVID-19 clusters linked to day care centers within our study area in Utah [35], 42% of the cases occurred in children, who represented 60% of the people with epidemiological contacts to the facilities. The infected children (median age 7) transmitted infection to at least 26% of their non-facility contacts, close to our household estimate. Another study found that children under 10 in China were as likely to be infected as adults [36]. However, other studies suggest that children may be less likely to acquire infection than adults [37], and one study found very low household secondary attack from infected children in South Korea [38]. A study similar to ours found lower rates of importation and household acquisition among children aged 5-9 compared to older groups, although confidence intervals overlapped [15]. If substantial differences existed between children under 12 and our study participants, one or more of our estimates could be biased.

In addition, many eligible participants older than 12 chose not to participate, either declining the serological antibody test only (but still filling out a survey) or declining to participate at all. Comparing full participants to survey-only participants, we found that participants reporting a prior positive PCR were less likely to agree to antibody testing, though the difference was not large (61.5% vs. 72.7%). It is unknown whether a prior confirmed or suspected infection affected eligible household members' decision to agree or decline to fill out a survey.

We also have not adjusted for potential biases related to non-participation rates of entire households that were selected and approached for inclusion in the study. Our data collection included a complicated sampling design

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

across several different strata, and weights were introduced partly to account for different rates of nonresponse across the different strata. For simplicity we ignored these details and sampling weights for the analysis presented here. Thus, households with higher COVID-19 risk may be overrepresented or underrepresented in our data relative to their frequency in the broader population of households with Utah.

Although these potential limitations, which also exist for other analyses of household transmission from serological data [13-15], remain in our analysis, we believe our model has addressed other limitations of existing models that may be more substantial. Our improvements to household secondary attack rate estimates, including factoring out non-household community acquisitions and tertiary transmissions, inclusion of overdispersion estimates, and careful consideration of the impact of imperfect test sensitivity and specificity, have produced improved insights into this important measure.

In conclusion, we found evidence of a relatively high secondary attack rate and high overdispersion in transmission of SARS-COV-2 in Utah households during a time when overall community prevalence was low. Other published household secondary attack rates may be underestimated without accounting for imperfect test sensitivity and specificity. Controllability of the virus may depend on mitigating transmission from a minority of highly infectious individuals in large households and other household-like locations where several people congregate indoors for extended periods.

References

1. Leclerc QJ, Fuller NM, Knight LE, Group CC-W, Funk S, Knight GM. What settings have been linked to SARS-CoV-2 transmission clusters? Wellcome Open Res **2020**; 5: 83.
2. Nishiura H, Oshitani H, Kobayashi T, et al. Closed environments facilitate secondary transmission of coranvirus disease 2019 (COVID-19). MedRxiv **2020**; 2020.02.28.20029272.
3. Qian H, Miao T, Liu L, Zheng X, Luo D, Li Y. Indoor transmission of SARS-CoV-2. Indoor Air **2020**; ina.12766.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

4. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. Lancet Infect Dis **2020**; 20(11): 1247-54.

5. Zhu Y, Wang C, Dong L, Xiao M. Home quarantine or centralized quarantine, which is more conducive to fighting COVID-19 pandemic? Brain Behav Immun **2020**; 87: 142-3.

6. Lei H, Xu X, Xiao S, Wu X, Shu Y. Household transmission of COVID-19-a systematic review and meta-analysis. J Infect **2020**; 81(6): 979-97.

7. Fung HF, Martinez L, Alarid-Escudero F, et al. The household secondary attack rate of SARS-CoV-2: A rapid review. Clin Infect Dis **2020**; ciaa1558.

8. Madewell ZJ, Yang Y, Longini IM, Halloran ME, Dean NE. Household transmission of SARS-CoV-2: a systematic review and meta-analysis. JAMA Netw Open **2020**; 3(12): e2031756.

9. Yousaf AR, Duca LM, Chu V, et al. A prospective cohort study in non-hospitalized household contacts with SARS-CoV-2 infection: symptom profiles and symptom change over time. Clin Infect Dis **2020**; ciaa1072.

10. Dawson P, Rabold EM, Laws RL, et al. Loss of Taste and Smell as Distinguishing Symptoms of COVID-19. Clin Infect Dis **2020**; ciaa799.

11. Rosenberg ES, Dufort EM, Blog DS, et al. COVID-19 Testing, Epidemic Features, Hospital Outcomes, and Household Prevalence, New York State-March 2020. Clin Infect Dis **2020**; 71(8): 1953-9.

12. Grijalva CG, Rolfes MA, Zhu Y, et al. Transmission of SARS-COV-2 Infections in Households - Tennessee and Wisconsin, April-September 2020. MMWR Morb Mortal Wkly Rep **2020**; 69(44): 1631-4.

13. Pollan M, Perez-Gomez B, Pastor-Barriuso R, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet **2020**; 396(10250): 535-44.

14. Silveira MF, Barros AJD, Horta BL, et al. Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. Nat Med **2020**; 26(8): 1196-9.

15. Bi Q, Lessler J, Eckerle I, et al. Household Transmission of SARS-COV-2: Insights from a Population-based Serological Survey. MedRxiv **2020**; 2020.11.04.20225573.

16. Gomes MGM, Corder RM, King JG, et al. Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. medRxiv **2020**; 2020.04.27.20081893.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

17. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature **2005**; 438(7066): 355-9.

18. Frieden TR, Lee CT. Identifying and Interrupting Superspreading Events-Implications for Control of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis **2020**; 26(6): 1059-66.

19. Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Leclerc QJ, et al. Implication of backward contact tracing in the presence of overdispersed transmission in COVID-19 outbreaks [version 1; peer review: 2 approved]. Wellcome Open Res **2020**; 5(239).

20. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. Lancet **2020**; 395(10227): e47.

21. Kain MP, Childs ML, Becker AD, Mordecai EA. Chopping the tail: how preventing superspreading can help to maintain COVID-19 control. Epidemics **2021**; 34: 100430.

22. Endo A, Centre for the Mathematical Modelling of Infectious Diseases C-WG, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res **2020**; 5: 67.

23. Zhang Y, Li Y, Wang L, Li M, Zhou X. Evaluating Transmission Heterogeneity and Super-Spreading Event of COVID-19 in a Metropolis of China. Int J Environ Res Public Health **2020**; 17(10).

24. Adam DC, Wu P, Wong JY, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nat Med **2020**; 26(11): 1714-9.

25. Wong F, Collins JJ. Evidence that coronavirus superspreading is fat-tailed. Proc Natl Acad Sci U S A **2020**; 117(47): 29416-8.

26. Wang L, Didelot X, Yang J, et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nat Commun **2020**; 11(1): 5006.

27. Longini I, Koopman JS. Household and community transmission parameters from final distributions of infections in households. Biometrics **1982**; 38(1): 115-26.

28. Samore M, Looney A, Orleans B, et al. SARS-CoV-2 seroprevalence and detection fraction in Utah urban populations from a probability-based sample. MedRxiv **2020**; 2020.10.26.20219907.

29. Becker N. An epidemic chain model. Biometrics **1980**; 36(2): 249-54.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

30. Ball F, Mollison D, Scalia-Tomba G. Epidemics with two levels of mixing. The Annals of Applied Probability **1997**; 7(1): 46-89.

31. SARS-CoV-2 IgG Architect - Instructions for Use. Available at: https://www.fda.gov/media/137383/download. Accessed 1/25/2021.

32. Chen PZ, Bobrovitiz N, Premji Z, Koopmans M, Fisman DN, Gu FX. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. medRxiv **2020**; 2020.10.13.20212233.

33. Susswein Z, Bansal S. Characterizing superspreading of SARS-CoV-2: from mechanism to measurement. medRxiv **2020**; 2020.12.08.20246082.

34. Russell TW, Golding N, Hellewell J, et al. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC Med **2020**; 18(1): 332.

35. Lopez AS, Hill M, Antezano J, et al. Transmission Dynamics of COVID-19 Outbreaks Associated with Child Care Facilities - Salt Lake City, Utah, April-July 2020. MMWR Morb Mortal Wkly Rep **2020**; 69(37): 1319-23.

36. Bi Q, Wu Y, Mei S, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect Dis **2020**; 20(8): 911-9.

37. Li X, Xu W, Dozier M, et al. The role of children in transmission of SARS-CoV-2: A rapid review. J Glob Health **2020**; 10(1): 011101.

38. Kim J, Choe YJ, Lee J, et al. Role of children in household transmission of COVID-19. Arch Dis Child **2020**; archdischild-2020-319910.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Table 1. Maximum likelihood estimates

Value	MLE estimate (95% CI)	Parametric bootstrap: median (95% range)
Mean community acquisition probability (p_c)	0.38% (0.30% – 0.48%)	0.38% (0.26% – 0.51%)
Mean per-capita household transmission probability (p_h)	35% (26% – 47%)	35% (24% – 48%)
Per-capita household transmission dispersion (d_h)	0.54 (0.05 – 2.5)	0.55 (0 – 4.6)
Probability infected person reported a positive PCR test (ϕ_V)	76% (66% – 86%)	76% (66% – 85%)
Probability infected person tested positive for antibodies (ϕ_A)	85% (75% – 93%)	85% (76% – 94%)
Probability uninfected person did not report a positive PCR test (π_V)	99.93% (99.87% – 99.98%)	99.93% (99.87% – 99.99%)
Probability uninfected person tested negative for antibodies (π_A)	99.3% (99.1% – 99.4%)	99.3% (99.2% – 99.4%)

Confidence intervals for MLE derived from the likelihood ratio test, varying each individual parameter while fixing other parameters at their MLE values. Parametric bootstrap was based on MLE fits to 500 different synthetic data sets generated from stochastic simulations using the MLE parameter values.

Table 2. Comparison of MLE for alternate models

Fixed values	$\hat{p}_c$	$\hat{p}_h$	$\hat{d}_h$	$\hat{\phi}_V$	$\hat{\phi}_A$	$\hat{\pi}_V$	$\hat{\pi}_A$	Log likelihood	Rejection P value
None	0.385%	35.4%	0.536	76.3%	85.2%	99.93%	99.3%	−1143.92	-
$d_h = 0$	0.427%	40.7%	**0***	71.2%	79.1%	99.95%	99.3%	−1145.35	0.09
$d_h = \infty$	0.326%	31.3%	∞*	79.2%	88.1%	99.91%	99.3%	−1148.67	0.002
$\pi_v = 100\%$	0.487%	31.6%	0.220	73.8%	76.0%	**100%***	99.3%	−1146.15	0.035
$\pi_a = 99.6\%$	0.550%	29.8%	0.254	62.3%	81.9%	99.96%	**99.6%***	−1152.32	<0.0001
$\pi_a = 100\%$	1.24%	16.6%	0.0322	36.8%	77.5%	100%	**100%***	−1173.39	<0.0001

*Values that were fixed for the model in that row; other values were optimized by MLE. P values were derived from the likelihood ratio test, compared to the likelihood of the overall MLE in the top row (twice the difference in log likelihood compared to the chi-squared distribution with one degree of freedom).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Table 3. Effect of transmission overdispersion on probability that first infected person transmits to no one or everyone in the household

Household size	Transmit to none: overall model MLE (95% CI)	Transmit to none: binomial model MLE (95% CI)	Transmit to all: overall model MLE (95% CI)	Transmit to all: binomial model MLE (95% CI)
2	65% (52% – 75%)	69% (60% – 76%)	35% (24% – 48%)	31% (24% – 41%)
3	57% (42% – 71%)	47% (35% – 58%)	27% (12% – 44%)	10% (6% – 16%)
4	52% (32% – 69%)	32% (21% – 45%)	24% (7% – 43%)	3% (1% – 7%)
5	50% (25% – 68%)	22% (13% – 34%)	21% (3% – 43%)	1% (0.3% – 3%)
6	47% (21% – 68%)	15% (7% – 26%)	20% (2% – 43%)	0.3% (0.07% – 1%)
7	46% (18% – 68%)	11% (4% – 20%)	19% (2% – 43%)	0.09% (0.02% – 0.4%)
8	44% (16% – 68%)	7% (3% – 15%)	18% (0.9% – 43%)	0.03% (0.004% – 0.2%)
9	43% (14% – 68%)	5% (2% – 12%)	17% (0.7% – 43%)	0.009% (0.0009% – 0.07%)
10	42% (13% – 68%)	3% (0.9% – 9%)	16% (0.5% – 43%)	0.003% (0.0002% – 0.03%)

Probabilities in this table are for a single infected household member transmitting directly to no one or everyone else in the household. The "transmit to all" values do not include the probability of multiple-generation transmission chains that eventually infect all household members. Confidence intervals for the overall MLE-based estimates were derived from applying (p_h, d_h) pairs from our parametric bootstrap analysis to the beta-binomial transmission equations.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Supplemental Methods

Alternate model with variability in household importation

For the alternate model, the formula for M_{kn} for a given household size $n \geq 2$ becomes

$$M_{kn}(p_c, d_c, p_h, d_h) = \begin{cases} F_{0n}(p_c, d_c), & k = 0 \\ \sum_{i=1}^{k} F_{in}(p_c, d_c) T_{i,k-i,n-i}(p_h, d_h), & k = 1, \dots, n-1 \\ 1 - \sum_{k=0}^{n-1} M_{kn}(p_c, d_c, p_h, d_h), & k = n \end{cases}$$

For households of size $n = 1$, $M_{01}(p_c, d_c) = F_{01}(p_c, d_c)$ and $M_{11}(p_c, d_c,) = 1 - F_{01}(p_c, d_c)$. The function $F_{yz}(p, d)$ is defined in the main text (probability mass function of the beta-binomial distribution with boundary case definitions at $d = 0$ and $d \to \infty$), where in this case y is the number of community acquisitions and z is the total number of household members. The main-text model is a special case of this alternate model, with $d_c \to \infty$.

The likelihood equation is the same as in the main text, but with the additional element d_c in the vector $\boldsymbol{\theta}$ of variables to be optimized:

$$\boldsymbol{\theta} = (p_c, d_c, p_h, d_h, \phi_V, \phi_A, \pi_V, \pi_A)$$

and $M_{kn}(p_c, p_h, d_h)$ in the likelihood equation is replaced with $M_{kn}(p_c, d_c, p_h, d_h)$ as defined above.

We found the MLE and single-parameter confidence intervals using the same procedure described in the main text, and further assessed the uncertainty the estimate of d_h by solving for the MLE of the other 7 variables when fixing it at its boundary values 0 and ∞. We also compared the likelihood at the MLE of the alternate model to that of the main text model using the likelihood ratio test, to determine whether the main text model result could be rejected in favor of the alternate model by this criterion. As an additional comparison, we used the Bayesian information criterion to score the alternate model against the main text model, using $n = 9091$ as the number of data points (number of households) and $k = 7$ parameters for the main-text model and $k = 8$ parameters for the alternate model.

Beta-binomial distribution at limits of dispersion parameter: $d \to \infty$ and $d \to 0$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Our likelihood equations make use of the beta-binomial probability distribution, parameterized with an average probability p and a dispersion parameter d. The probability mass function F for positive, finite values of d is

$$F_{yz}(p,d) = \binom{z}{y}\frac{\mathrm{B}\big(y+dp, z-y+d(1-p)\big)}{\mathrm{B}\big(dp, d(1-p)\big)}, \qquad y = 0,1,\ldots,z$$

We use $F_{yz}(p,d)$ to quantify the distribution of household transmissions directly from a single infected household member, where y is the number of transmissions, z is the number of susceptible household members, $p = p_h$, and $d = d_h$. In our alternate model we also use $F_{yz}(p,d)$ to quantify the distribution of community acquisitions among members of a household from non-household members, where y is the number of community acquisitions, z is the total number of household members, $p = p_c$, and $d = d_c$.

Here, we derive the formula for $F_{yz}(p,d)$ at the boundaries of the range of possible values for d: $d \to \infty$ and $d \to 0$. To do this, we rewrite $F_{yz}(p,d)$ in an alternate form. First, using the property $\mathrm{B}(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$:

$$F_{yz}(p,d) = \binom{z}{y}\frac{\Gamma(d)}{\Gamma(z+d)}\frac{\Gamma(y+dp)}{\Gamma(dp)}\frac{\Gamma\big(z-y+d(1-p)\big)}{\Gamma\big(d(1-p)\big)}$$

Then using the property, for positive integer n, $\Gamma(z+n) = z(z+1)\cdots(z+n-1)\Gamma(z)$:

$$F_{yz}(p,d) = \begin{cases} \dfrac{d(1-p)(d(1-p)+1)\cdots(d(1-p)+z-1)}{d(d+1)\cdots(d+z-1)} & y = 0 \\ \binom{z}{y}\dfrac{[dp(dp+1)\cdots(dp+y-1)][d(1-p)(d(1-p)+1)\cdots(d(1-p)+z-y-1)]}{d(d+1)\cdots(d+z-1)} & 0 < y < z \\ \dfrac{dp(dp+1)\cdots(dp+z-1)}{d(d+1)\cdots(d+z-1)} & y = z \end{cases}$$

In the following we rewrite the numerators and denominators in powers of d. Only the lowest and highest powers of d will matter for taking the limits that follow, so the other terms within "⋯" are not shown.

$$F_{yz}(p,d) = \begin{cases} \dfrac{(1-p)(z-1)!\,d + \cdots + (1-p)^z d^z}{(z-1)!\,d + \cdots + d^z} & y = 0 \\ \binom{z}{y}\dfrac{p(1-p)(y-1)!\,(z-y-1)!\,d^2 + \cdots + p^y(1-p)^{z-y}d^z}{(z-1)!\,d + \cdots + d^z} & 0 < y < z \\ \dfrac{p(z-1)!\,d + \cdots + p^z d^z}{(z-1)!\,d + \cdots + d^z} & y = z \end{cases}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

When taking the limit $d \to \infty$, we note that each fraction has highest order d^z in both numerator and denominator, so the limit will be the ratio of the coefficients of that term:

$$\lim_{d\to\infty} \binom{z}{y} \frac{\mathrm{B}\big(y + dp, z - y + d(1-p)\big)}{\mathrm{B}\big(dp, d(1-p)\big)} = \binom{z}{y} p^y (1-p)^{z-y}$$

When taking the limit $d \to 0$, we note that the fraction for cases $y = 0$ and $y = z$ have lowest order term d in both the numerator and denominator, so the limit will be the ratio of the coefficients on those terms. The fraction for the $0 < y < z$ case has only powers of d^2 and higher in the numerator, and a nonzero d term in the denominator, so the limit is 0:

$$\lim_{d\to 0} \binom{z}{y} \frac{\mathrm{B}\big(y + dp, z - y + d(1-p)\big)}{\mathrm{B}\big(dp, d(1-p)\big)} = \begin{cases} 1-p, & y = 0 \\ 0, & 0 < y < z \\ p, & y = z \end{cases}$$

Direct transmission probabilities H_{xyz}

Here we derive the formulae for H_{xyz}: the probability of y transmissions to z susceptible household members directly from x infected members

As described in the main text, we first define the probabilities for transmissions directly from $x = 1$ infected household member:

$$H_{1yz} = F_{yz}(p_h, d_h) = \begin{cases} \begin{cases} 1-p_h, & y = 0 \\ 0, & 0 < y < z \\ p_h, & y = z \end{cases}, & d_h = 0 \\ \binom{z}{y} \dfrac{\mathrm{B}\big(y + d_h p_h, z - y + d_h(1-p_h)\big)}{\mathrm{B}\big(d_h p_h, d_h(1-p_h)\big)}, & 0 < d_h < \infty \\ \binom{z}{y} {p_h}^y (1-p_h)^{z-y}, & d_h = \infty \end{cases}$$

Next consider $z = 2$ infected household members. The probability that $y = 0$ transmissions occur is the probability that both infected members transmit to 0 others: $H_{20z} = H_{10z}H_{10z}$. To calculate the probability that $y > 0$ transmissions occur from the two infected members, it is convenient to consider the two infected individuals having transmission opportunities in sequence, say A followed by B. If A transmits to any household members, this reduces the number of susceptible members remaining for B to infect. For example, the probability that $y = 1$ is the probability that A transmits to 0 of z and B transmits to 1 of z, plus the probability that A transmits to 1 of z and B

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

transmits to 0 of $z-1$ remaining susceptible members: $H_{21z} = H_{10z}H_{11z} + H_{11z}H_{1,0,z-1}$. The calculation follows a similar pattern for $y = 2$: $H_{22z} = H_{10z}H_{12z} + H_{11z}H_{1,1,z-1} + H_{12z}H_{1,0,z-2}$. It follows that:

$$H_{2yz} = \sum_{i=0}^{y} H_{1iz}H_{1,y-i,z-i}$$

Now for $x = 3$ infected household members, we can use the fact that we have already calculated H_{2yz}, which covers the transmission probabilities from two of the three infected members, and then we include the probability that third member transmits to any remaining susceptible members that the first two did not infect:

$$H_{3yz} = \sum_{i=0}^{y} H_{2iz}H_{1,y-i,z-i}$$

Following this pattern, we continue calculating H_{xyz} for each x in increasing sequence:

$$H_{xyz} = \sum_{i=0}^{y} H_{x-1,i,z}H_{1,y-i,z-i}$$

Total transmission probabilities T_{xyz}

Here we derive the formulae for T_{xyz}: the probability of y total transmissions to z initially susceptible members from x initially infected members. In other words, T_{xyz} is the probability that the final household outbreak size is $x + y$, given that x household members were initially infected and z household members were initially susceptible.

First, we note that $T_{x0z} = H_{x0z}$ for all possible (x, z) pairs, because if the initial x infected members do not transmit to anyone ($y = 0$), the household outbreak is over and the final size has been reached. Next we consider the probability of $y = 1$ total transmissions, which occurs when the initial x infected members transmit directly to 1 other (with probability H_{x1z}), who then does not subsequently transmit to any of the remaining $z-1$ susceptible members (with probability $H_{1,0,z-1} = T_{1,0,z-1}$). Hence,

$$T_{x1z} = H_{x1z}T_{1,0,z-1}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

For $y = 2$ total transmissions, we must include the probability that the initial x infected members transmit directly to 2 others who then transmit to none and the probability that the initial x infected members transmit directly to 1 other who then produces an outbreak among the remaining susceptible members with 1 total transmission:

$$T_{x2z} = H_{x2z}T_{2,0,z-2} + H_{x1z}T_{1,1,z-1}$$

Following similar logic for $y = 3$, and making use of the T_{x0z}, T_{x1z}, and T_{x2z} values already calculated, we arrive at:

$$T_{x3z} = H_{x3z}T_{3,0,z-3} + H_{x2z}T_{2,1,z-2} + H_{x1z}T_{1,2,z-1}$$

The general formula calculated for increasing values of y is:

$$T_{xyz} = \sum_{i=0}^{y-1} H_{x,y-i,z}T_{y-i,i,z-y+i}$$

Within-household reproduction numbers and threshold condition

We define the within-household reproduction number R_h as the expected number of household transmissions directly from an infected person who acquired infection in the community and has no non-susceptible housemates. Let h_i be the fraction of households with size i, up to a maximum size N. Then the mean μ and variance σ^2 of the household size distribution are

$$\mu = \sum_{i=1}^{N}(ih_i), \qquad \sigma^2 = \sum_{i=1}^{N}(i^2h_i) - \mu^2$$

Let c_i be the probability that a randomly chosen person has i housemates. The probability that a randomly chosen person lives in a house of *total* size i (including themselves) is ih_i/μ, so

$$c_i = (i+1)h_{i+1}/\mu$$

Then R_h is p_h times the mean number of housemates of a randomly chosen person:

$$R_h = p_h\sum_{i=1}^{N-1}(ic_i) = p_h\left(\sum_{i=1}^{N}(i^2h_i) - \sum_{i=1}^{N}(ih_i)\right)\Big/\mu = p_h(\mu + \sigma^2/\mu - 1)$$

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

We also define an alternate within-household reproduction number R_h^* as the expected total number of transmissions in the household of an infected person who acquired infection in the community and has no initially non-susceptible housemates. Given that a person acquiring infection in the community has i susceptible housemates, the probability that j of their housemates become infected before the household outbreak terminates is $T_{1ji}(p_h, d_h)$, as defined in 2.2 of the main text. The expected total number of transmissions in their household will be then be $\sum_{j=1}^{i} jT_{1ji}(p_h, d_h)$. So, the household reproduction number formula is:

$$R_h^* = \sum_{i=1}^{N-1} c_i \sum_{j=1}^{i} jT_{1ji}(p_h, d_h) = \sum_{i=1}^{N-1} \frac{(i+1)h_{i+1}}{\mu} \sum_{j=1}^{i} jT_{1ji}(p_h, d_h)$$

For the high-variability boundary case at $d_h = 0$, we have that

$$\sum_{j=1}^{i} jT_{1ji}(p_h, 0) = ip_h$$

because $T_{1ji}(p_h, 0) = p_h$ when $j = i$ and 0 for other nonzero values of j (reflecting all-or-nothing transmission). It follows that $R_h = R_h^*$ when $d_h = 0$. This makes intuitive sense because in the all-or-nothing scenario, when the index person transmits, all in the household are infected directly, and there is no one left to infect in subsequent generations, so the final household outbreak size is entirely reflected in R_h.

We next investigate the implications of our R_h^* estimate for population-wide transmission control. The threshold condition delineating subcritical and supercritical transmission in the population occurs when the maximal eigenvalue of the matrix

$$\begin{bmatrix} R_c & R_h^* \\ R_c & 0 \end{bmatrix}$$

exceeds one [30]. Here, R_c is defined as the average number of community transmissions per infected individual (i.e., average number of transmissions to people not in the infected individual's household). The occurrence of R_c in both rows of the matrix reflects an assumption that its value applies to the transmissibility of people who acquire their own infection in the community and in their household. The zero element in the lower-right corner of the matrix reflects the fact that the R_h^* people on average who acquire infection in their household do not transmit

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

further in their household, by definition, because R_h^* was derived from the final household outbreak size equations encompassed in T_{xyz}.

The maximal eigenvalue exceeding one produces the following threshold condition:

$$\frac{R_c}{2}\left(1+\sqrt{1+4\,R_h^*/R_c}\right)>1$$

This is equivalent to

$$R_c>\frac{1}{R_h^*+1}$$

Relationship between beta distributed probability and dynamic transmission parameters

If an infected person's duration of infectiousness is τ and the transmission rate to a contact is β, the probability that transmission to the contact occurs is $p=1-e^{-\beta\tau}$. We assume τ is fixed and β is a gamma distributed random variable with shape k and rate r. Then, the first and second moments of the random variable p are:

$$\mathrm{E}[p]=\int_0^\infty \frac{r^k}{\Gamma(k)}x^{k-1}e^{-rx}(1-e^{-\tau x})dx=1-\left(\frac{r}{r+\tau}\right)^k$$

$$\mathrm{E}[p^2]=\int_0^\infty \frac{r^k}{\Gamma(k)}x^{k-1}e^{-rx}(1-e^{-\tau x})^2dx=1-2\left(\frac{r}{r+\tau}\right)^k+\left(\frac{r}{r+2\tau}\right)^k$$

The variance is then

$$\mathrm{Var}[p]=\mathrm{E}[p^2]-(\mathrm{E}[p])^2=\left(\frac{r}{r+2\tau}\right)^k-\left(\frac{r}{r+\tau}\right)^{2k}$$

We then equate the mean and variance to those of the beta distribution with mean p_h and dispersion d_h, which we used in our MLE model in the main text.

$$p_h=1-\left(\frac{r}{r+\tau}\right)^k$$

$$\frac{p_h(1-p_h)}{d_h+1}=\left(\frac{r}{r+2\tau}\right)^k-\left(\frac{r}{r+\tau}\right)^{2k}$$

Combining those two equations yields

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

$$\left((1-p_h)d_h+1\right)^{1/k}\left(2-(1-p_h)^{1/k}\right)-(d_h+1)^{1/k}=0$$

$$r=\frac{\tau(1-p_h)^{1/k}}{1-(1-p_h)^{1/k}}$$

We solved the first equation for k, which is independent of the assumption for τ, using our MLE estimates of p_h and d_h. We applied each of the (p_h, d_h) pairs from our parametric bootstrap analysis to this equation to derive the confidence interval for k.

Supplemental Results

The alternate model produced an estimate for the new dispersion parameter $d_c = 2.4$ (0.95 – 9.6) and altered estimates for the other 7 parameters compared to their values for the main text model (Table S1). The log likelihood at this MLE was about 1 greater than the log likelihood produced by the main text result (Table S2), suggesting that the main text model (equivalent to the alternate model with $d_c = \infty$) cannot be rejected with high confidence in favor of the alternate model by the likelihood ratio test (P = 0.16). The Bayesian information criterion (BIC) for the alternate model is 2358.8 compared to 2351.6 for the main text ($d_c = \infty$) model, which favors the main text model with a BIC difference of 7.1.

The conclusion of high household transmission variability from the main-text model is consistent under this alternate model, with the MLE occurring at low value of the dispersion parameter $d_h = 0.30$, and the low-variability binomial model $d_h = \infty$ can be rejected with P = 0.02. However, uncertainty ranges become wider at higher levels of overdispersion in household risk of community acquisition. This is illustrated by the fact that a model assuming no household transmission ($p_h = 0$), i.e. all household cases explained by acquisitions outside the households with high overdispersion ($d_c = 0.5$), cannot be rejected with very high confidence (P = 0.068). Thus, the alternate explanation for the distribution of household cases may not be definitively ruled out by our data.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Table S1. Alternate model results: allowing variability in importation probability by household

Value	MLE estimate (95% interval)
Mean community acquisition probability (p_c)	0.52% (0.40% – 0.65%)
Household community acquisition dispersion (d_c)	2.4 (0.95 – 9.6)
Mean per-capita household transmission probability (p_h)	28% (16% – 42%)
Per-capita household transmission dispersion (d_h)	0.30 (0 – 4.2)
Probability that surveyed, infected person reported a positive viral PCR test (ϕ_V)	76% (65% – 85%)
Probability that antibody test of person with prior infection was positive (ϕ_A)	86% (76% – 94%)
Probability that surveyed, uninfected person did not report a positive viral PCR test (π_V)	99.92% (99.86% – 99.97%)
Probability that antibody test of person with no prior infection was negative (π_A)	99.3% (99.1% – 99.4%)

Confidence intervals for MLE derived from the likelihood ratio test, varying each individual parameter while fixing other parameters at their MLE values.

Table S2. Comparison of alternate model MLE to main-text and other models

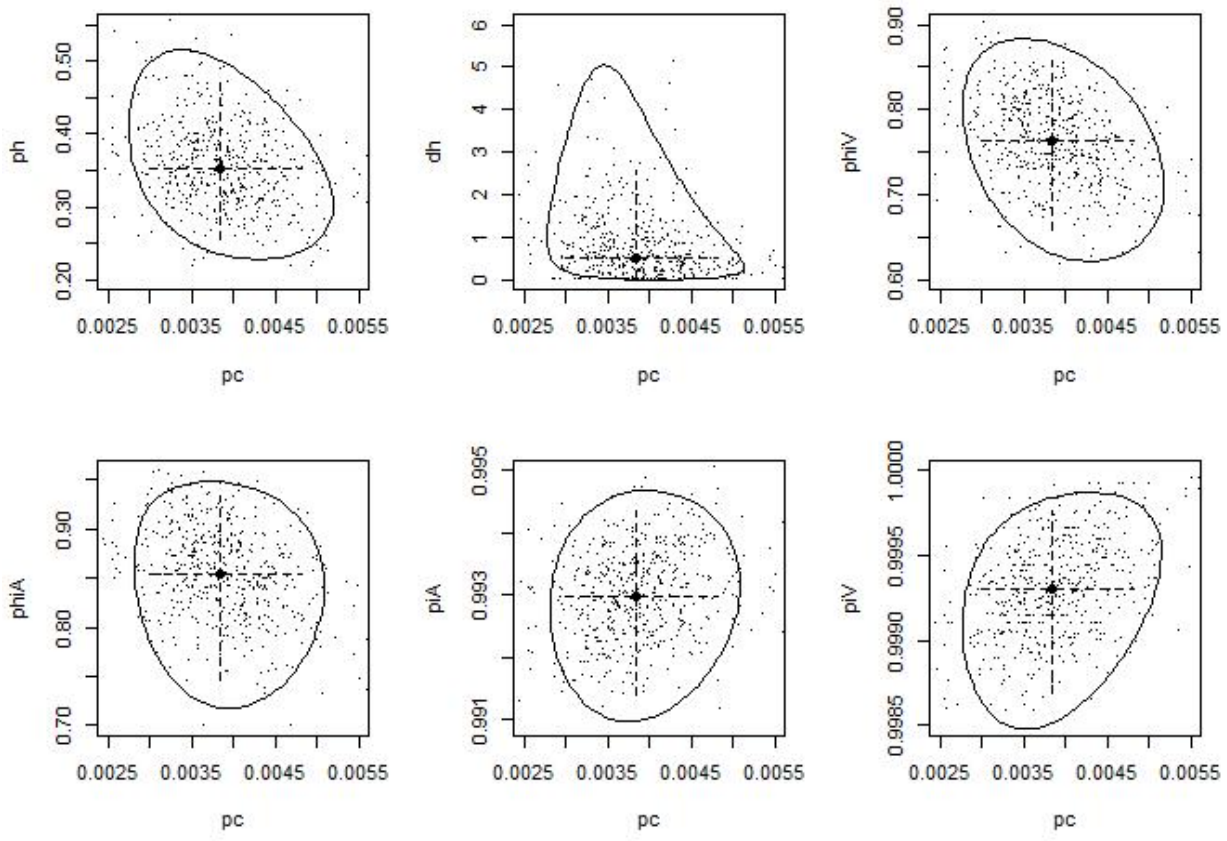
Fixed values	$\hat{p}_c$	$\hat{d}_c$	$\hat{p}_h$	$\hat{d}_h$	$\hat{\phi}_V$	$\hat{\phi}_A$	$\hat{\pi}_V$	$\hat{\pi}_A$	Log likelihood	# of optimized parameters	Rejection P value
None	0.52%	2.4	28%	0.30	76%	86%	99.92%	99.3%	−1142.93	8	-
$d_c = \infty$	0.38%	∞*	35%	0.54	76%	85%	99.93%	99.3%	−1143.92	7	0.16
$d_h = 0$	0.57%	1.7	24%	**0***	75%	86%	99.92%	99.3%	−1143.11	7	0.55
$d_h = \infty$	0.73%	0.5	0%	∞*	76%	87%	99.91%	99.3%	−1145.62	7	0.020
$p_h = 0\%$	0.73%	0.5	**0%***	N/A	76%	87%	99.91%	99.3%	−1145.62	6	0.068

*Values that were fixed for the model in that row; other relevant values were optimized by MLE. P values were derived from the likelihood ratio test, compared to the likelihood of the overall MLE in the top row: twice the difference in log likelihood compared to the chi-squared distribution with degrees of freedom equal to the difference in the number of optimized parameters. The model with household transmission probability p_h fixed at 0% optimized only 6 parameters because the transmission dispersion parameter d_h is irrelevant with no transmission, hence we used 2 degrees of freedom for the reference chi-squared distribution when comparing the likelihood to the full 8-parameter model. For the model with d_h fixed at ∞ (binomial transmission model), the optimum occurred at the boundary p_h = 0, producing the same likelihood as the adjacent model but different P-value as the reference chi-squared distribution has 1 degree of freedom.

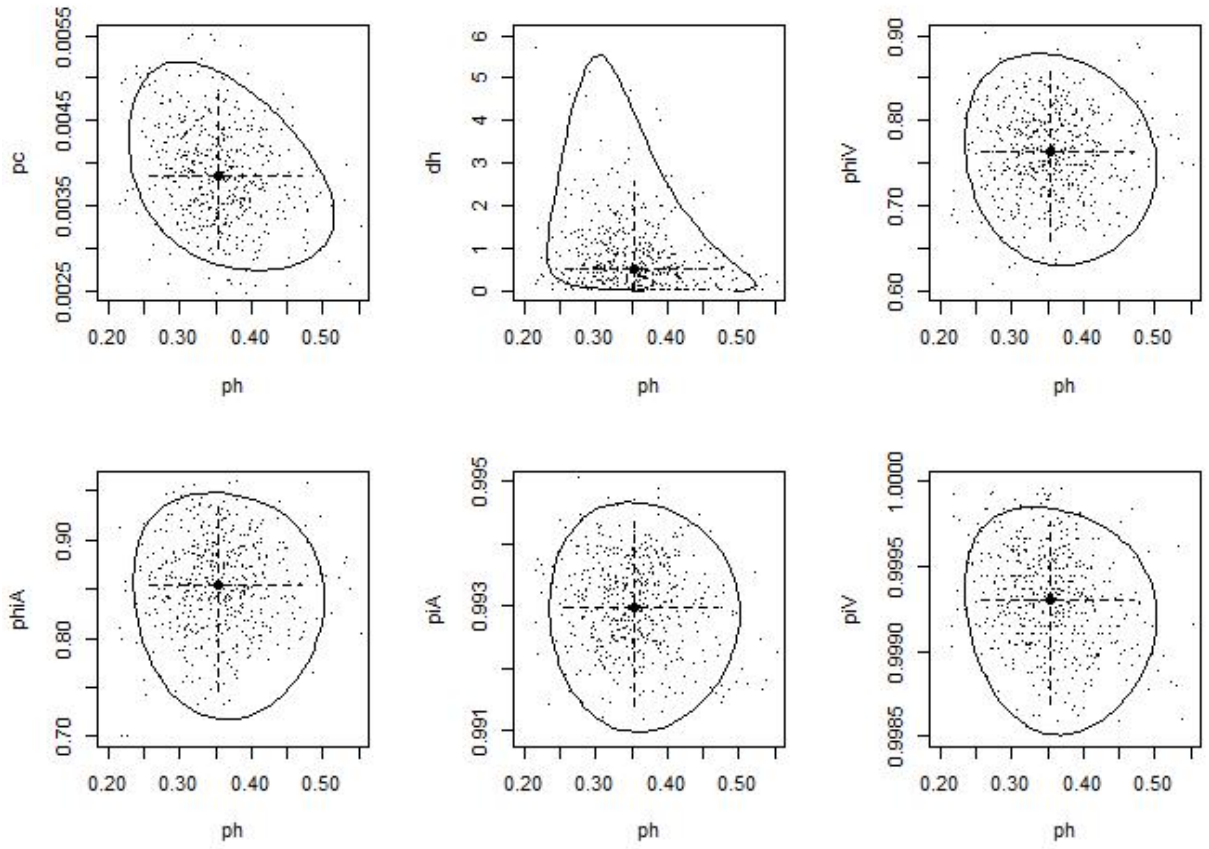
medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figures S1-S7. Two-dimension confidence regions for each pair of estimated parameters.

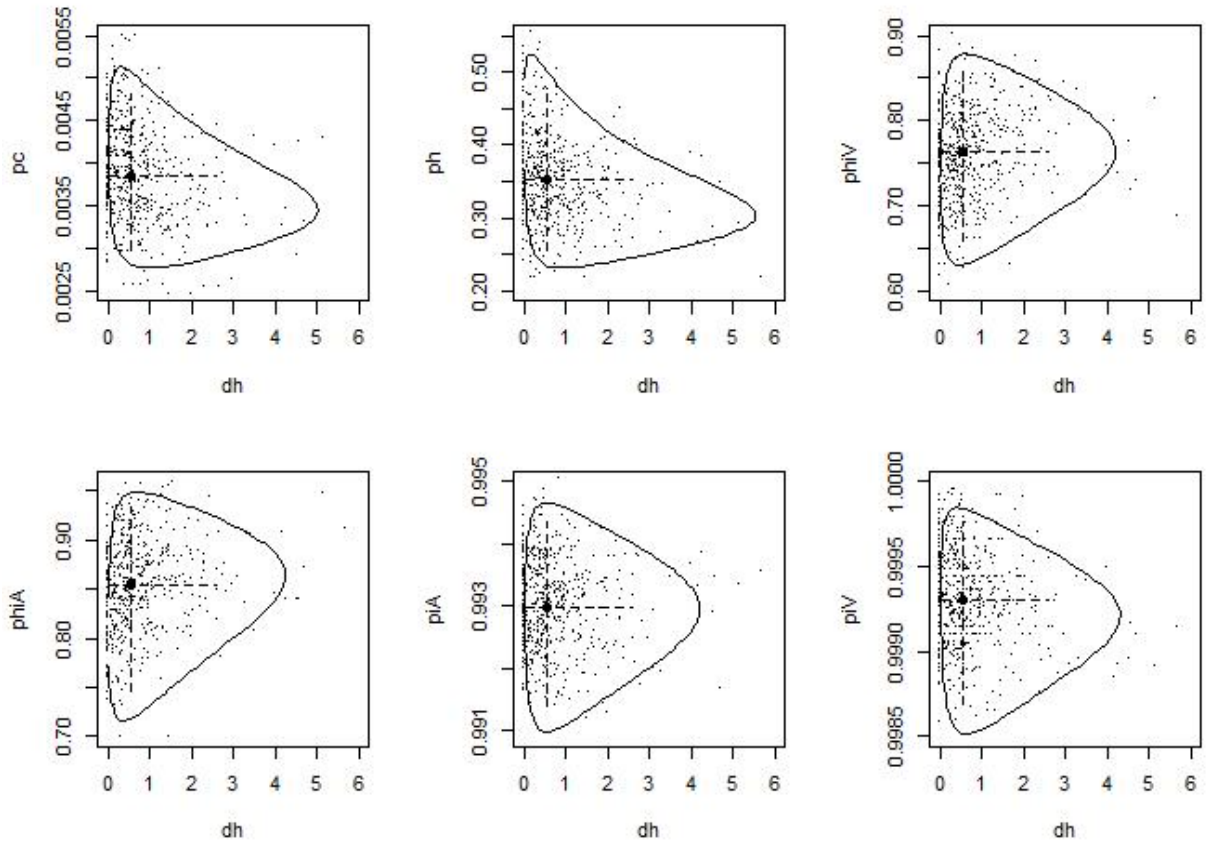
Solid curves are the 2-dimensional confidence regions derived from the likelihood ratio test, comparing the likelihood ratio statistic to the 95th percentile of the chi-squared distribution with 2 degrees of freedom. Large solid circle is the MLE estimate and dashed lines are the confidence intervals for each individual parameter derived from the likelihood ratio test (Table 1 main text). Small dots are the MLE estimates from each of 500 simulated data sets generated using parameter values set at the MLE from the actual data (parametric bootstrap).



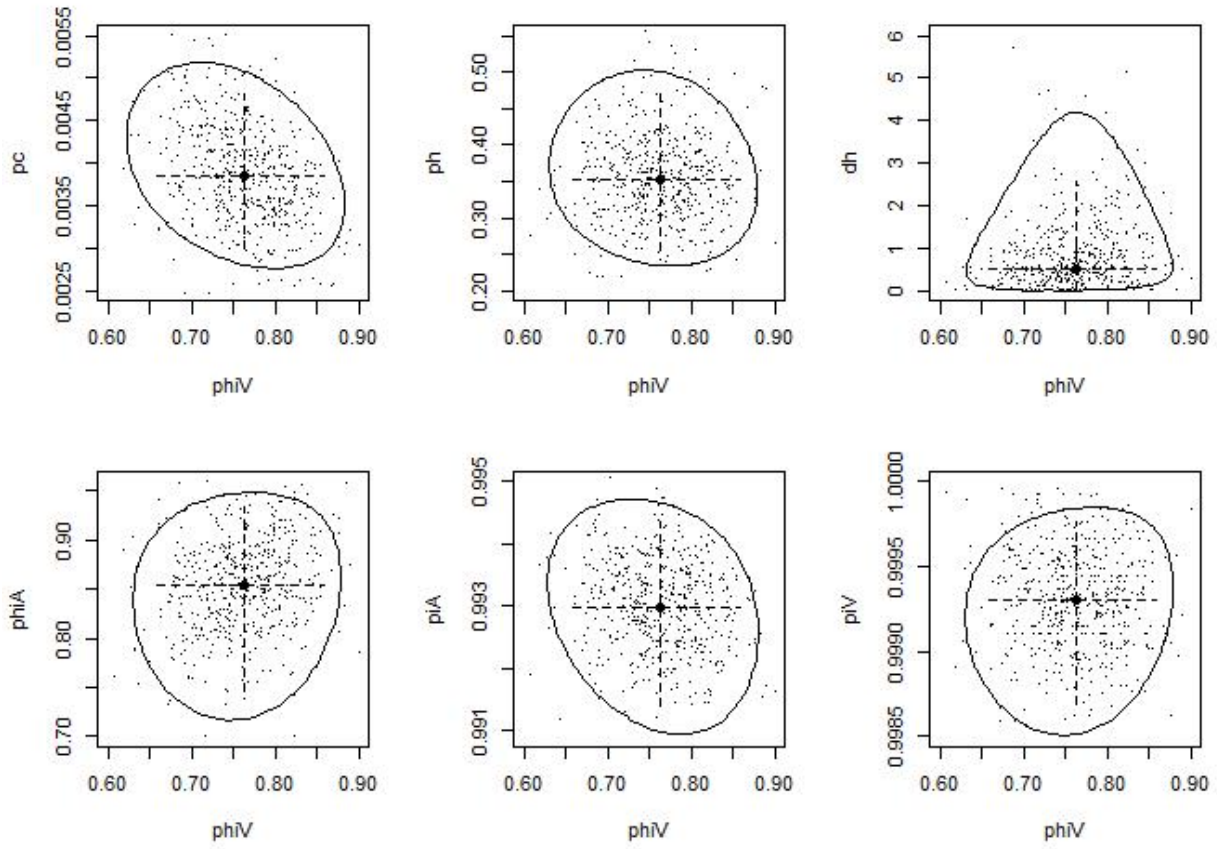
medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



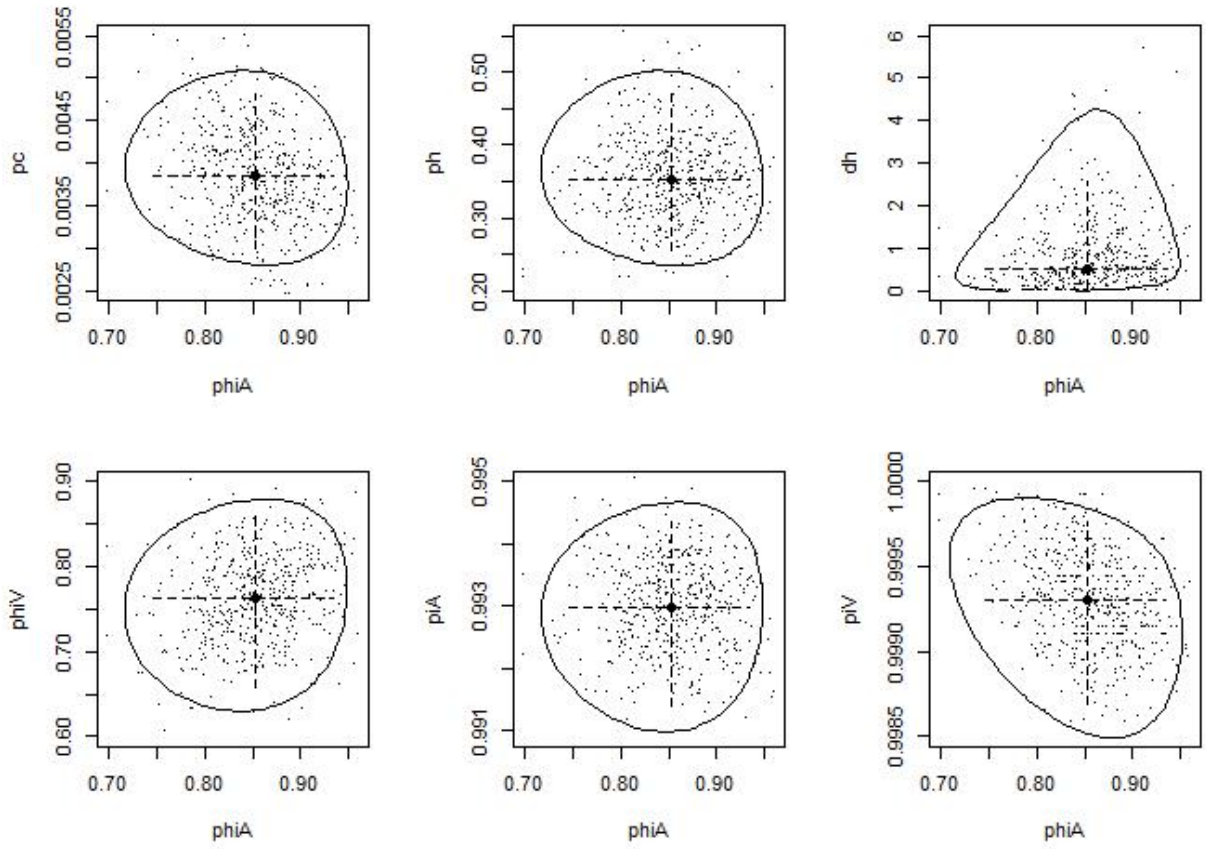
medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



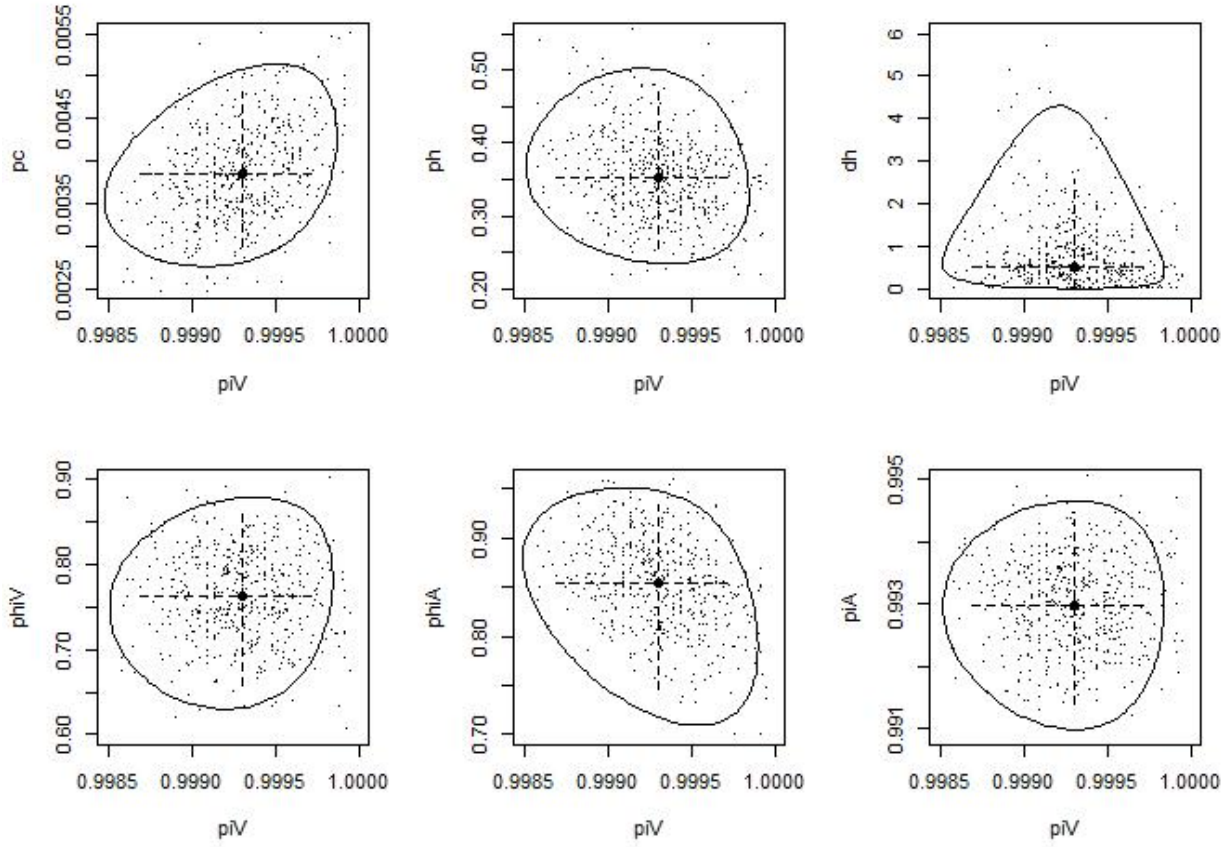
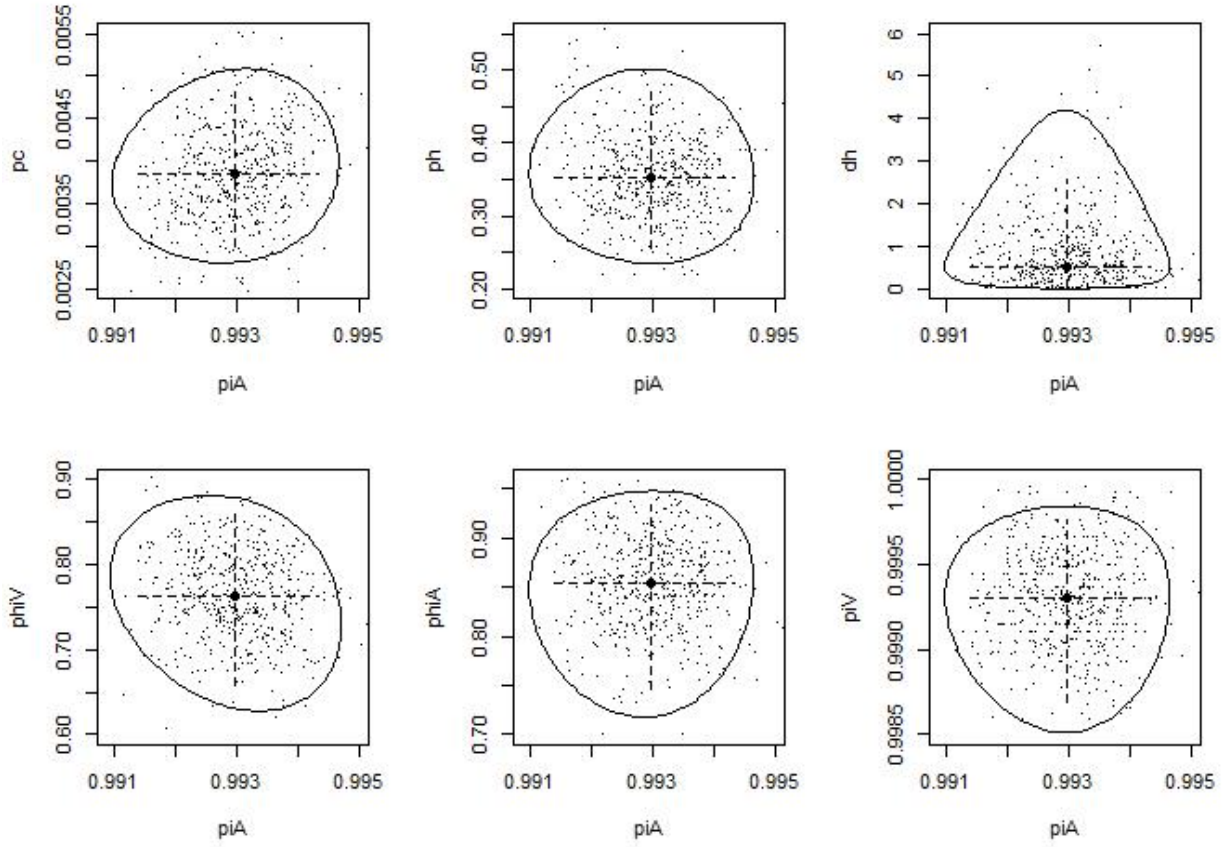
medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.29.20248797; this version posted February 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.