PLOS ONE

RESEARCH ARTICLE

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

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]

1 Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, United States of America, **2** Department of Veterans Affairs Salt Lake City Healthcare System, Salt Lake City, Utah, United States of America, **3** Department of Mathematics, University of Utah, Salt Lake City, Utah, United States of America, **4** Department of Finance, University of Utah David Eccles School of Business, Salt Lake City, Utah, United States of America, **5** Department of Family and Preventive Medicine, University of Utah School of Medicine, Salt Lake City, Utah, United States of America

* Damon.Toth@hsc.utah.edu

Citation: Toth DJA, Beams AB, Keegan LT, Zhang Y, Greene T, Orleans B, et al. (2021) High variability in transmission of SARS-CoV-2 within households and implications for control. PLoS ONE 16(11): e0259097. https://doi.org/10.1371/journal.pone.0259097

Editor: Muhammed Elhadi, University of Tripoli, LIBYA

Received: February 19, 2021

Accepted: October 12, 2021

Published: November 10, 2021

Data Availability Statement: The data underlying the results presented in the study are publicly available at https://github.com/damontoth/householdTransmission.

Funding: SCA received funding from the State of Utah (https://www.utah.gov/) under contract number AR3473 to collect the data analyzed in this manuscript. DJAT, ABB, LTK, YZ, and MHS were supported by the Centers for Disease Control and Prevention (https://www.cdc.gov/) under award number U01CK000585 and by the Department of

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.

Methods

We compiled serological SARS-CoV-2 antibody test data and prior SARS-CoV-2 test reporting from members of 9,224 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 0.41% (95% CI 0.32%– 0.51%) chance of acquiring SARS-CoV-2 infection outside their household. Our household secondary attack rate estimate was 36% (27%– 48%), substantially higher than the crude estimate of 16% 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 0.41 (0.33–0.52) to avoid continued growth of the pandemic in Utah.

Veterans Affairs (https://www.va.gov/) under award number I50HX001240. The funders had no role in the study design, data collection and analysis, decision to published, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: DJAT and MHS report grants from Pfizer, Inc., outside the submitted work. LTK reports grants from Pfizer, Inc. and Becton, Dickinson and Company, outside the submitted work. TG reports grants from AstraZeneca, Boehringer Ingelheim, CSL, and Vertex Pharmaceuticals and personal fees from Janssen Pharmaceuticals, Pfizer, Inc., Novartis and AstraZeneca, outside the submitted work. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

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 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 testing specimens from their household contacts using PCR or other methods that detect presence of the virus. 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 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 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 could participate in a survey that included questions about prior SARS-CoV-2 test results (see Supplementary Methods in S1 File for wording of relevant survey questions). Adult household members could fill out surveys on behalf of children of any age in the household. Survey participants age 12 or older could additionally opt to provide serological samples for COVID-19 antibody testing. Serum specimens were analyzed using the Abbott SARS-CoV-2 IgG assay performed on an Abbott Architect i2000 instrument (Abbott Laboratories), with methodology and criteria for a positive antibody result defined according to the manufacturer's instructions. Data included in this analysis were collected between May 4 and August 15, 2020.

The University of Utah Institutional Review Board reviewed the surveillance project that produced the data analyzed in this manuscript and determined it as non-research public health surveillance, waived the requirement for documented consent, and determined that use of these data for analysis to understand the dynamics of SARS-CoV-2 transmission was exempt

from further review (IRB_00132598). Individuals were informed of the project procedures and that participation was voluntary. Participants provided their agreement to participate and were given the chance to opt out of having their data used for future research. The data were analyzed anonymously for this manuscript.

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 test result and received a positive antibody test
- a_{PN}: number who reported a prior positive test result and received a negative antibody test
- a_{NP}: number who reported no prior positive test result and received a positive antibody test
- s_P: number who were surveyed, reported a prior positive test result, and did not receive an antibody test

Those surveyed participants who reported no prior positive test result includes both those who had never been tested and those who had been tested but received no positive results. We did not have sufficient information to properly distinguish those two groups, nor to determine the circumstances of any prior negative tests that might affect the inferred probability of true prior infection.

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 $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_C,f_1,\ldots f_C)$.

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 [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). Accounting for the timing of recent household importations, transmissions, and development of detectable antibodies during an ongoing household outbreak would significantly complicate the model equations and would likely have little effect on our overall results, given that the prevalence of active infections at the time of data collection was very low [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:

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

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 S1 File for details):

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

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 S1 File):

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

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 (S1 File). 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 $y = 0$ successes occur with probability $1-p$ (S1 File):

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

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 S1 File. 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 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 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 test results for SARS-CoV-2 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 test for the virus, and π_V, the probability that a surveyed person with no prior infection did not report receiving a positive 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 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 test result and received a negative antibody test
- $S_{Ni} = s_i - s_{Pi}$: number who were surveyed, reported no prior positive test 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:

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

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 prior test 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 prior positive 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 prior positive test). The formula for A is

$$A(u, v, w, x; \phi_V, \phi_A, \pi_V, \pi_A) = f_m(u, v, w, x; \mathbf{p}_I(\phi_V, \phi_A)) f_m(a_{PPi} - u, a_{PNi} - v, a_{NPi} - w, a_{NNi} - x; \mathbf{p}_U(\pi_V, \pi_A))$$

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 $\mathbf{p} = \mathbf{p}_I$) or uninfected (for $\mathbf{p} = \mathbf{p}_U$):

$$\mathbf{p}_I(\phi_V, \phi_A) = (\phi_V\phi_A, \phi_V(1-\phi_A), (1-\phi_V)\phi_A, (1-\phi_V)(1-\phi_A))$$

$$\mathbf{p}_U(\pi_V, \pi_A) = ((1-\pi_V)(1-\pi_A), (1-\pi_V)\pi_A, \pi_V(1-\pi_A), \pi_V\pi_A)$$

The first element of $\mathbf{p}_I$, $\phi_V\phi_A$, is the probability that an antibody-tested person with a prior infection reported a prior positive 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 prior test and the probability that an infected person actually sought and received a SARS-CoV-2 test during the period of infection in which detectable virus was present and reported that positive test on our survey. Elements 2–4 of $\mathbf{p}_I$ are the probabilities that an antibody-tested, prior infected person reported a prior positive test but tested negative for antibodies, did not report a prior positive test and tested positive for antibodies, and did not report a prior positive test and tested negative for antibodies, respectively. The elements of $\mathbf{p}_U$ 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:

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

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 maximized the log likelihood over the 7 unknown parameters (p_c, p_h, d_h, ϕ_V, ϕ_A, π_V, π_A) using the observations (n, a, s, a_{PP}, a_{PN}, a_{NP}, s_P) for each household, to produce the MLE: $\hat{\theta} = (\hat{p}_c, \hat{p}_h, \hat{d}_h, \hat{\phi}_V, \hat{\phi}_A, \hat{\pi}_V, \hat{\pi}_A)$. The log likelihood maximization was performed using the "optim" function in R. 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 p_c, p_h, and d_h, 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 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.

We also compared the performance of our 7-parameter model against simpler models with fewer free parameters, including models assuming $d_h = 0$ and $d_h \to \infty$, models fixing the sensitivity and specificity of the antibody test according to independent data from the test manufacturer (ϕ_A = 89.3% and/or π_A =99.6%) [32], and models assuming perfect specificity of prior test result reporting or antibody testing (π_V = 100% or π_A = 100%). We tested each of these assumptions individually and in various combinations, resulting in models ranging from 3 to 6 optimized parameters. We compared the models using the Akaike information criterion, which aims to balance goodness of fit with model simplicity.

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 in S1 File).

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 in S1 File). 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:

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

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 S1 File.

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 S1 File, is

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

where h_i is the fraction of all households that are size i.

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 $R_c(R_h^* + 1) > 1$, where R_c is the average number of community (non-household) transmissions from an infected person. We derive this formula in the S1 File, 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 (Fig 1). Of these, we retained 9,224 (98.3%) for use in the MLE. The 159 excluded households were removed because the household size was unknown (51) or the reported household size was less than the number of people tested or surveyed in the house (108). In the 9,224 retained households, there were 28,321 (3.07 per household) reported household members, 13,998 (1.52 per household) people who were both surveyed and antibody tested, and another 5,249 (0.57 per household) who were surveyed but not antibody tested. The households in the data were located in 7 of the 29 counties in Utah; the 22 excluded counties account for <14% of Utah's total population S1 Table in S1 File.

Of the 13,998 antibody tests in the retained households, 178 (1.27%) were positive. Of those 178 people with a positive antibody test, 58 (32.6%) reported receiving a prior positive test. Of the 19,247 people who were antibody tested or surveyed only, 119 (0.62%) reported receiving a prior positive test. This broke down to 0.53% (75 / 13,998) for those who were antibody tested and 0.84% (44 / 5,249) for those who were surveyed but not antibody tested. The rate of testing positive for antibodies among those reporting a prior positive test was 77.3% (58 / 75). The interval between the reported prior positive test date and the antibody test date did not exhibit

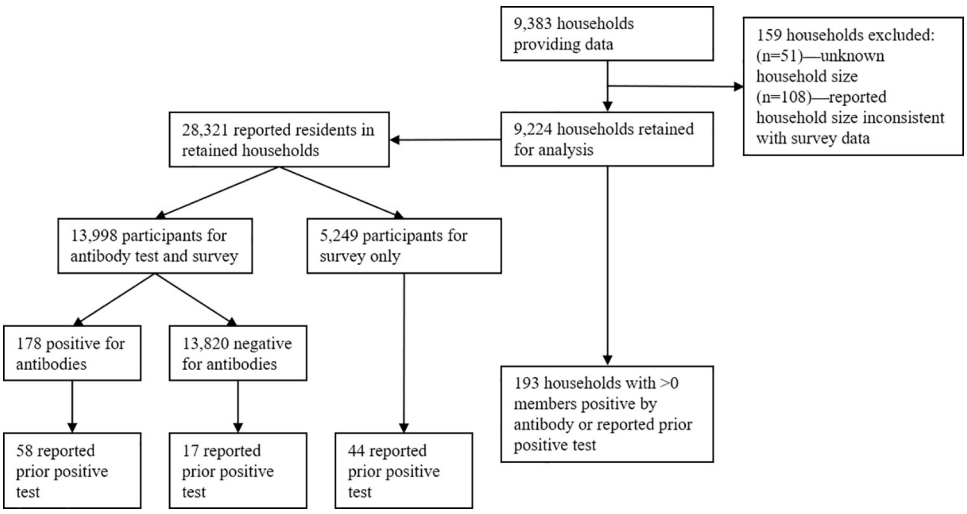

Fig 1. Data summary flowchart. Flow diagram for data from participating households and household members.

https://doi.org/10.1371/journal.pone.0259097.g001

a strong correlation to the fraction of testing antibody positive, other than perhaps the 3 individuals reporting a very recent (less than 1 week) positive test all testing negative for antibodies (S2 Table in S1 File). The rate of survey participants agreeing to antibody testing was lower for those who reported a prior positivetest compared to those who did not: 63.0% (75 of 119) vs. 72.8% (13,923 of 19,128), a small but statistically significant ($P < 0.01$) difference in proportion.

Of the retained households, 193 (2.1%) had at least one household member who either tested positive for antibodies or reported a prior positive test. There were 159 households with exactly 1 positive member (by either antibody test or reported prior test or both), 26 households with 2 positives, 6 with 3 positives, 1 with 4 positives, and 1 with 6 positives. In all, there were $C = 273$ unique $\mathbf{y}_i$ vectors representing household data described in section 2.1.

The crude secondary attack rate measure derived from antibody testing only (fraction of antibody-tested housemates of antibody-positive household members who were also antibody positive) was 14.9% (29 / 194). The crude secondary attack rate estimate from reported prior test data only (fraction of surveyed housemates of people reporting a prior positive test who also reported a prior positive test) was 23.0% (31 / 135). When combining both types of data, the crude secondary attack rate estimate (fraction of surveyed / tested housemates of any antibody-positive or reported-prior-positive person who were positive by either or both measures) was 15.6% (46 / 295).

We tallied demographic statistics of the set of surveyed individuals (S3 Table in S1 File). The distribution of reported ages skewed older than Utah's overall population age distribution, and females were slightly overrepresented (52.0%). The distribution of surveyed individuals' race, Hispanic origin, and education level also differed from the overall Utah and U.S. distributions.

3.2 Maximum likelihood estimates

Our MLE procedure produced simultaneous estimates for all 7 parameters (Table 1). The MLE for p_c, the per-person community acquisition probability from outside the household, was 0.41% (0.32%– 0.51%). For within household transmission probability, the MLE produced an average secondary attack rate estimate $p_h = 36\%$ (27%– 48%). The MLE for the dispersion

Table 1. Maximum likelihood estimates.

Value	MLE (95% CI)	Parametric bootstrap: median (95% range)
Mean community acquisition probability (p_c)	0.41% (0.32%–0.51%)	0.41% (0.30%– 0.55%)
Mean per-capita household transmission probability (p_h)	36% (27%– 48%)	36% (25%– 51%)
Per-capita household transmission dispersion (d_h)	0.43 (0.02–2.0)	0.38 (0–2.2)
Probability infected person reported a prior positive test (ϕ_V)	72% (62%– 82%)	72% (63%– 82%)
Probability infected person tested positive for antibodies (ϕ_A)	86% (75%– 93%)	86% (77%– 95%)
Probability uninfected person did not report a prior positive test (π_V)	99.94% (99.88%–99.98%)	99.94% (99.87%– 99.99%)
Probability uninfected person tested negative for antibodies (π_A)	99.3% (99.2%–99.5%)	99.3% (99.2%– 99.5%)

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.

https://doi.org/10.1371/journal.pone.0259097.t001

parameter d_h, quantifying variability in transmissibility by person, was 0.43 (0.02–2.0). 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 = 0.001 (Table 2).

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	Free Parameters	ΔAIC
None	0.41%	36%	0.43	72%	86%	99.94%	99.3%	−1173.84	-	7	-
$d_h = 0$	0.44%	41%	(0)	68%	80%	99.95%	99.3%	−1174.85	0.16	6	0.02
$d_h = \infty$	0.34%	32%	(∞)	76%	89%	99.91%	99.3%	−1179.30	0.00096	6	8.91
$\phi_A = 89.3\%$	0.39%	36%	0.55	73%	(89%)	99.93%	99.3%	−1174.11	0.46	6	−1.46
$\pi_V = 100\%$	0.51%	34%	0.16	70%	77%	(100%)	99.3%	−1176.01	0.037	6	2.34
$\pi_A = 99.6\%$	0.57%	31%	0.20	60%	83%	99.96%	(99.6%)	−1181.08	0.00014	6	12.47
$\pi_A = 100\%$	1.2%	18%	0.02	37%	78%	100%	(100%)	−1201.26	<0.0001	6	52.84
(d_h, ϕ_A)	0.41%	43%	(0)	70%	(89%)	99.93%	99.3%	−1176.85	0.050	5	2.01
(d_h, π_V)	0.52%	36%	(0)	68%	75%	(100%)	99.3%	−1176.17	0.097	5	0.66
(d_h, π_A)	0.61%	33%	(0)	58%	79%	99.97%	(99.6%)	−1181.35	0.00055	5	11.02
(ϕ_A, π_V)	0.49%	30%	0.36	73%	(89%)	(100%)	99.3%	−1180.62	0.0011	5	9.55
(ϕ_A, π_A)	0.53%	30%	0.39	62%	(89%)	99.94%	(99.6%)	−1181.76	0.00036	5	11.84
(π_V, π_A)	0.65%	30%	0.05	58%	76%	(100%)	(99.6%)	−1181.72	0.00038	5	11.75
(d_h, ϕ_A, π_V)	0.51%	34%	(0)	70%	(89%)	(100%)	99.3%	−1183.29	0.00029	4	12.89
(d_h, ϕ_A, π_A)	0.55%	34%	(0)	59%	(89%)	99.94%	(99.6%)	−1183.87	0.00017	4	14.05
(d_h, π_V, π_A)	0.66%	30%	(0)	58%	76%	(100%)	(99.6%)	−1181.73	0.0013	4	9.78
(ϕ_A, π_V, π_A)	0.62%	26%	0.31	62%	(89%)	(100%)	(99.6%)	−1186.34	<0.0001	4	18.99
(d_h, ϕ_A, π_V, π_A)	0.64%	29%	(0)	60%	(89%)	(100%)	(99.6%)	−1188.39	<0.0001	3	21.11

Values in parentheses 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). ΔAIC value is the difference in Akaike information criterion compared to that of the model in the top row: positive ΔAIC means the model in the top row has lower AIC and is favored by this criterion.

https://doi.org/10.1371/journal.pone.0259097.t002

Our MLE result for ϕ_V, the probability that a surveyed person with a prior infection reported a prior positive test, was 72% (62%– 82%). The ϕ_V value can be interpreted as the case ascertainment fraction, i.e. fraction of individuals with SARS-CoV infections who were identified with a positive test during their infection. Our result may be high 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 [31]. Our finding 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.

For π_V, the probability that a surveyed person with no prior infection reported no prior positive test, the MLE was 99.94% (99.88%– 99.98%). This result is consistent with the low probability of false positives among viral tests, which to our knowledge were exclusively PCR-based in Utah prior to our data collection. It is possible that some false positives in our survey data occurred by erroneous reporting, i.e. survey respondents reporting a prior positive 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 (Table 2), which suggests that studies producing epidemiological estimates relying on a 100% viral test specificity assumption should test robustness of conclusions to small deviations from that assumption.

For ϕ_A, the probability that a prior-infected person's antibody test was positive, the MLE was 86% (75%– 93%), similar to the test manufacturer's finding that 109 of 122 (89.3%) PCR-positive subjects were positive for antibodies [32]. Assuming $\phi_A = 89.3\%$ directly and optimizing the other 6 parameters produces a slightly better AIC than the full 7-parameter model (Table 2). The manufacturer's results included only symptomatic subjects and were highly dependent on the number of days post-symptom onset at which the serological sample was taken. Because the symptom histories of the antibody-tested people in our data are largely uncertain, it is difficult to determine how consistent our result is with the manufacturer's data.

The MLE for π_A, the probability that an antibody-tested person with no prior infection tested negative for antibodies, was 99.3% (99.2%– 99.5%), which is 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 [32]. Models assuming the manufacturer's point specificity estimate of 99.6%, produced inferior AIC to the full model (Table 2). When we ran our MLE under the assumption of perfect specificity (no false positives) for the antibody test ($\pi_A = 100\%$), the result for secondary attack rate reduced from 36% to 18%, which is closer to the crude estimate described in Section 3.1, and the results for community acquisition probability *increased* from 0.4% to 1.2% (Table 2). 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. We also found that assuming perfect specificity of the antibody test dramatically reduced the estimate of ϕ_V from 72% to 37% (Table 2), which suggests that ignoring false positives in serology data could cause an underestimate of the case ascertainment rate if the serology data are used for that purpose.

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 in S1 File). 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 high confidence (P = 0.16) (Table 2). However, none of the alternate models assuming $d_h = 0$ produced a superior AIC to the full model (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 $d_c = 2.1$ (0.89–7.5), with somewhat altered estimates of the other parameters (S4 Table in S1 File) 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 neither model is favored by the Akaike information criterion. 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 in S1 File).

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 $p_h = 32\%$ (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).

For example, our MLE model estimates that an infected member of an 8-member household would have a 46% (22%– 70%) chance of transmitting to no one, but a 20% (3%– 50%) 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%– 14%) chance of transmitting to no one and 0.03% (0.005%– 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

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	64% (49%– 75%)	68% (59%– 76%)	36% (25%– 51%)	32% (24%– 41%)
3	57% (42%– 71%)	46% (35%– 57%)	29% (14%– 49%)	10% (6%– 17%)
4	53% (34%– 69%)	32% (21%– 43%)	26% (9%– 49%)	3% (1%– 7%)
5	51% (30%– 69%)	22% (12%– 33%)	24% (6%– 50%)	1% (0.3%– 3%)
6	49% (26%– 69%)	15% (7%– 25%)	22% (5%– 50%)	0.3% (0.08%– 1%)
7	47% (24%– 69%)	10% (4%– 19%)	21% (3%– 50%)	0.1% (0.02%– 0.5%)
8	46% (22%– 70%)	7% (3%– 14%)	20% (3%– 50%)	0.03% (0.005%– 0.2%)
9	45% (20%– 70%)	5% (2%– 11%)	20% (2%– 50%)	0.01% (0.001%– 0.08%)
10	44% (19%– 70%)	3% (0.9%– 8%)	19% (2%– 51%)	0.003% (0.0003%– 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.

https://doi.org/10.1371/journal.pone.0259097.t003

with shape k, then $k = 0.18$ (95% CI 0–0.7) when the mean and variance are matched, regardless of the infectious duration (Supplementary Methods in S1 File). 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]. Our estimate of k is similar to point estimates for SARS-CoV-2 of $k = 0.1$ [22], $k = 0.25$ [23], and $k = 0.33$ [24].

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 $\mu = 3.07$ and $\sigma^2 = 3.12$, so our estimate is $R_h = 1.12$ (0.78–1.56). Our estimate of the alternate household reproduction number R_h^*, the expected total number of transmissions in the household of a community acquirer, is $R_h^* = 1.45$ (0.94–2.05).

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 2.6 and S1 File). Using our estimate for R_h^* in Utah, this formula suggests that R_c must be kept below approximately 0.41 (0.33–0.52) 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 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% (27%– 48%), 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 15.6%. We traced the major source of this substantial underestimate to the assumption of perfect test specificity inherent in the crude formula.

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 (high variability; $d_h = 0.43$). 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 converted 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 $k = 0.18$, similar to other published results for SARS-CoV-2. 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 [33]. 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 [34].

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 sustained transmission in a community, which requires some level of between-household transmission to be maintained. Given our estimate of $R_h^* = 1.45$ (0.94–2.05), we can estimate the critical value of R_c, the average number of non-household community transmission that would push transmission for 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 0.41 (0.33–0.52) to avoid continued case growth 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.The potential contribution of interventions to reduce household transmission may also be important. Using the terms defined above, if $R_c < 1$ but $R_c(R_h^* + 1) > 1$, then overall transmission is above-threshold but could be pushed below-threshold by reducing household transmission alone, such that $R_h^* < 1/R_c - 1$. Methods to reduce household transmission might include increased used of at-home testing to earlier detect potential asymptomatic or pre-symptomatic transmitters, paired with increased use of masks, disinfectants, and/or distancing within homes of an infectious person [35].

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 in S1 File). Interestingly, the MLE for the transmission dispersion parameter d_h still occurred at high variability in household transmission ($\hat{d}_h = 0.21$) 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 Akaike 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 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, but this could be an important direction of future work enhanced by a larger dataset.

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. Older participants could fill out surveys on behalf of children of any age, including reporting of prior positive tests, but participation in that option was low. Thus, our assumption that non-participants had equal community acquisition rates, susceptibility to acquisition from another household member, and transmissibility to other household members compared to study participants would be violated if children were substantially different from adults in one or more of those quantities. 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 [36], 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 [37]. However, other studies suggest that children may be less likely to acquire infection than adults [38], and one study found very low household secondary attack from infected children in South Korea [39]. 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 SARS-CoV-2 test were less likely to agree to antibody testing, though the difference was not large (63.0% vs. 72.8%). It is unknown whether a prior confirmed or suspected infection affected eligible household members' decision to agree or decline to fill out a survey. The full set of surveyed participants had different distributions of reported age, sex, race, Hispanic origin, and education level compared to the wider population, and future work could assess the implications of those differences for extrapolating COVID-19 risk to other households.

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 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. Also, while the 7 included Utah counties represent >86% of the state population, there may be important differences in households from the 22 excluded counties. 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 in Utah.

Other potential limitations due to simplifying assumptions could be addressed in future work by relaxing those assumptions, such as assuming different prior infection probabilities for those who reported prior negative tests vs. those who reported never being tested and including the probability of active infections at the time of serological testing. 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. While the likelihood equations resulting from our model are somewhat complicated, we have demonstrated that the complications introduced by including 7 unknown model parameters are justified by systematically comparing its performance against simpler models, using criteria that seek to balance goodness of fit with simplicity. Furthermore, we have provided full mathematical specification and computational code for reproducibility. The ability to explicitly calculate the likelihood for our model is an advantage for optimization speed and further mathematical analysis, and extensions to the epidemiological household model can readily be simulated to explore potential improvements.

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.

Supporting information

S1 File. Supplementary material.
(DOCX)

Author Contributions

Conceptualization: Damon J. A. Toth, Matthew H. Samore.

Data curation: Damon J. A. Toth, Brian Orleans, Nathan Seegert, Adam Looney, Matthew H. Samore.

Formal analysis: Damon J. A. Toth.

Funding acquisition: Stephen C. Alder.

Investigation: Damon J. A. Toth, Alexander B. Beams, Lindsay T. Keegan, Matthew H. Samore.

Methodology: Damon J. A. Toth.

Software: Damon J. A. Toth, Alexander B. Beams.

Supervision: Stephen C. Alder, Matthew H. Samore.

Validation: Damon J. A. Toth, Alexander B. Beams, Yue Zhang, Tom Greene.

Visualization: Damon J. A. Toth.

Writing – original draft: Damon J. A. Toth, Lindsay T. Keegan.

Writing – review & editing: Alexander B. Beams, Lindsay T. Keegan, Yue Zhang, Tom Greene, Brian Orleans, Nathan Seegert, Adam Looney, Stephen C. Alder, Matthew H. Samore.

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. https://doi.org/10.12688/wellcomeopenres.15889.2 PMID: 32656368
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. https://doi.org/10.1111/ina.12766 PMID: 33131151
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. https://doi.org/10.1016/S1473-3099(20)30553-3 PMID: 32621869
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. https://doi.org/10.1016/j.bbi.2020.05.009 PMID: 32387346
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. https://doi.org/10.1016/j.jinf.2020.08.033 PMID: 32858069
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. https://doi.org/10.1001/jamanetworkopen.2020.31756 PMID: 33315116
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. https://doi.org/10.1093/cid/ciaa549 PMID: 32382743

12. Grijalva CG, Rolfes MA, Zhu Y, et al. Transmission of SARS-COV-2 Infections in Households—Tennessee and Wisconsin, Apr-Sep 2020. MMWR Morb Mortal Wkly Rep 2020; 69(44): 1631–4. https://doi.org/10.15585/mmwr.mm6944e1 PMID: 33151916

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. https://doi.org/10.1016/S0140-6736(20)31483-5 PMID: 32645347

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. https://doi.org/10.1038/s41591-020-0992-3 PMID: 32641783

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. https://doi.org/10.1101/2020.04.27.20081893 PMID: 32511451

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. https://doi.org/10.1038/nature04153 PMID: 16292310

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. https://doi.org/10.3201/eid2606.200495 PMID: 32187007

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). https://doi.org/10.12688/wellcomeopenres.16344.3 PMID: 33154980

20. Liu Y, Eggo RM, Kucharski AJ. Secondary attack rate and superspreading events for SARS-CoV-2. Lancet 2020; 395(10227): e47. https://doi.org/10.1016/S0140-6736(20)30462-1 PMID: 32113505

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. https://doi.org/10.1016/j.epidem.2020.100430 PMID: 33360871

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. https://doi.org/10.12688/wellcomeopenres.15842.3 PMID: 32685698

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. https://doi.org/10.1038/s41591-020-1092-0 PMID: 32943787

25. Wong F, Collins JJ. Evidence that coronavirus superspreading is fat-tailed. Proc Natl Acad Sci U S A 2020; 117(47): 29416–8. https://doi.org/10.1073/pnas.2018490117 PMID: 33139561

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. https://doi.org/10.1038/s41467-020-18836-4 PMID: 33024095

27. Longini I, Koopman JS. Household and community transmission parameters from final distributions of infections in households. Biometrics 1982; 38(1): 115–26. PMID: 7082755

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. PMID: 7407313

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. 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. https://doi.org/10.1186/s12916-020-01790-9 PMID: 33087179

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

33. Chen PZ, Bobrovitz N, Premji Z, Koopmans M, Fisman DN, Gu FX. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. Elife 2021; 10. https://doi.org/10.7554/eLife.65774 PMID: 33861198

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

35. Wang Y, Tian H, Zhang L, et al. Reduction of secondary transmission of SARS-CoV-2 in households by face mask use, disinfection and social distancing: a cohort study in Beijing, China. BMJ Glob Health 2020; 5(5). https://doi.org/10.1136/bmjgh-2020-002794 PMID: 32467353

36. Lopez AS, Hill M, Antezano J, et al. Transmission Dynamics of COVID-19 Outbreaks Associated with Child Care Facilities—Salt Lake City, Utah, Apr-Jul 2020. MMWR Morb Mortal Wkly Rep 2020; 69(37): 1319–23. https://doi.org/10.15585/mmwr.mm6937e3 PMID: 32941418

37. 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. https://doi.org/10.1016/S1473-3099(20)30287-5 PMID: 32353347

38. 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. https://doi.org/10.7189/jogh.10.011101 PMID: 32612817

39. Kim J, Choe YJ, Lee J, et al. Role of children in household transmission of COVID-19. Arch Dis Child 2020; archdischild-2020-319910. https://doi.org/10.1136/archdischild-2020-319910 PMID: 32769089