How Infections Propagate After Point-Source Outbreaks

An Analysis of Secondary Norovirus Transmission

Jonathan L. Zelner,[a,b,c] *Aaron A. King,*[c,d,e,f] *Christine L. Moe,*[g] *and Joseph N. S. Eisenberg*[c,h]

Background: Secondary transmission after point-source outbreaks is an integral feature of the epidemiology of gastrointestinal pathogens such as norovirus. The household is an important site of these secondary cases. It can become the source of further community transmission as well as new point-source outbreaks. Consequently, time-series data from exposed households provide information for risk assessment and intervention.

Methods: Analysis of these data requires models that can address (1) dependencies in infection transmission, (2) random variability resulting from households with few members, and (3) unobserved state variables important to transmission. We use Monte Carlo maximum likelihood via data augmentation for obtaining estimates of the transmission rate and infectious period from household outbreaks with the 3 above features.

Results: We apply this parameter estimation technique to 153 infection sequences within households from a norovirus outbreak in Sweden and obtain maximum likelihood estimates of the daily rate of transmission ($\hat{\beta} = 0.14$, 95% confidence interval [CI] = 0.08–0.24) and average infectious period ($1/\hat{\gamma} = 1.17$ days, 95% CI = 1.00–1.88). We also demonstrate the robustness of the estimates to missing household sizes and asymptomatic infections.

Conclusions: Maximum likelihood techniques such as these can be used to estimate transmission parameters under conditions of unobserved states and missing household size data, and to aid in the understanding of secondary risks associated with point-source outbreaks.

(*Epidemiology* 2010;21: 711–718)

Norovirus is a highly-infectious gastrointestinal pathogen that affects all age groups.[1] Investigations of primary point-source outbreaks, therefore, often focus on secondary cases.[2,3] Households constitute a particularly important site of these secondary cases, as living in close proximity facilitates a higher effective rate of contact, particularly for diseases where the fecal-oral route is important to transmission. This household transmission contributes to overall disease burden, and individuals infected at the household level may generate infections in the community that result in new point-source outbreaks that infect many people at one time.

From 1997 through 2002, norovirus was responsible for 93% of nonbacterial gastroenteritis outbreaks in the United States.[4] The high incidence of norovirus is attributable both to its low infectious dose[1] and its ability to survive in the environment.[5] As a leading cause of gastroenteritis worldwide,[6] norovirus is an important concern for local public health departments as well the US Environmental Protection Agency (EPA). It is important, therefore, to develop effective intervention and control strategies for norovirus and similar pathogens. These require both reliable estimates of household transmission parameters and effective analytic tools for obtaining these estimates.

Although there have been studies of community norovirus outbreaks,[7] there are no studies that quantify transmission dynamics in the community using a dynamic model. One of the difficulties of these studies is that we often observe only the time of symptom onset for infectious cases. Unobserved events typically include infection and recovery and the times at which these occur. Properly describing the transmission dynamics in household systems necessitates the use of mechanistic models that account for unobserved state variables (eg, the number of infectious and susceptible individuals at any given time), and the more pronounced random variability in outbreaks in small populations.

In this paper, we develop tools to address these challenges and analyze household data collected subsequent to a norovirus outbreak. Götz et al[8] followed a series of 153

Submitted 30 July 2009; accepted 9 March 2010; posted 27 May 2010.

From the [a]Department of Sociology, and [b]Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI; [c]Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI; [d]Departments of Ecology and Evolutionary Biology, and [e]Mathematics, University of Michigan, Ann Arbor, MI; [f]Fogarty International Center, National Institutes of Health, Bethesda, MD; [g]Department of Global Health, School of Public Health, Emory University, Atlanta, GA; and [h]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI.

Supported by US Environmental Protection Agency (EPA) STAR Grant # RD83172701(to J.L.Z., C.L.M. and J.N.S.) and by CAMRA program of the Department of Homeland Security (DHS) and EPA Grant # RD83236201(to J.L.Z. and J.N.S.). RAPIDD program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center of the National Institutes of Health (to A.A.K.).

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Jonathan L. Zelner, Center for the Study of Complex Systems, University of Michigan, 321A West Hall, 1085 S. University Ave, Ann Arbor, MI 48109. E-mail: jzelner@umich.edu.

ISSN: 1044-3983/10/2105-0711
DOI: 10.1097/EDE.0b013e3181e5463a

households exposed to norovirus after a 1999 point-source, food-borne outbreak within a network of daycare centers in Stockholm, Sweden. For each of these households, one person (the household index case) was infectious and symptomatic due to the point-source outbreak, and the time of symptom onset for all subsequent cases was recorded. We denote each of these case sequences as a time series.

We analyze these outbreak data using a dynamic model, and obtain maximum likelihood estimates of the household transmission parameter, β, and the average duration of infectiousness, $1/\gamma$, where γ is the mean daily rate of recovery from infectiousness. We find that the observation of multiple household time-series may provide enough information to mitigate the absence of observed infection times, infectious periods and household sizes.

METHODS

Data

Illness data were obtained from a published study of a food-borne norovirus outbreak in 30 daycare centers in Stockholm, Sweden in 1999.[8] The origin of this outbreak was a single food-service worker who shedded norovirus while preparing lunches that were distributed from a central location to 30 daycare centers throughout Stockholm. At the time of the outbreak this worker was infectious but had no overt symptoms.

Among 775 subjects surveyed after the outbreak, 195 cases of gastroenteritis were identified, 176 as norovirus. Among those subjects with norovirus infections, 23 lived alone, 49 lived in households where transmission occurred, and 104 lived with one or more persons but with no observed transmission. Nineteen subjects were excluded because they lived in households with pre-existing cases of gastroenteritis at the time of the outbreak. The primary dataset used in this analysis consists of time series from the 153 exposed households with 2 or more members.

Data were collected retrospectively for the 9 days following the onset of symptoms in index cases. The data consist of the times that cases became symptomatic, reported to the nearest 12 hours and normalized (with the onset of symptoms in the index case set to time zero). Stool samples were collected from 5 symptomatic individuals, and the presence of norovirus was confirmed via electron microscopy. Remaining cases were diagnosed based on a norovirus screening interview and a confirmed exposure to a household member infected at the point-source event. Figure 1 provides a visual depiction of the household time-series data for exposed households with secondary cases (modified from the paper by Götz et al,[8] Fig. 5).

When describing household transmission dynamics, we assume that the onset of symptoms corresponds to the beginning of the infectious period. This is supported by a controlled norovirus dosing trial in which early shedding in the absence of symptoms occurred primarily in persons who never became symptomatic.[9] Our model also allows the infectious period to be longer than the symptomatic period, which is typical of norovirus infections.[9,10]

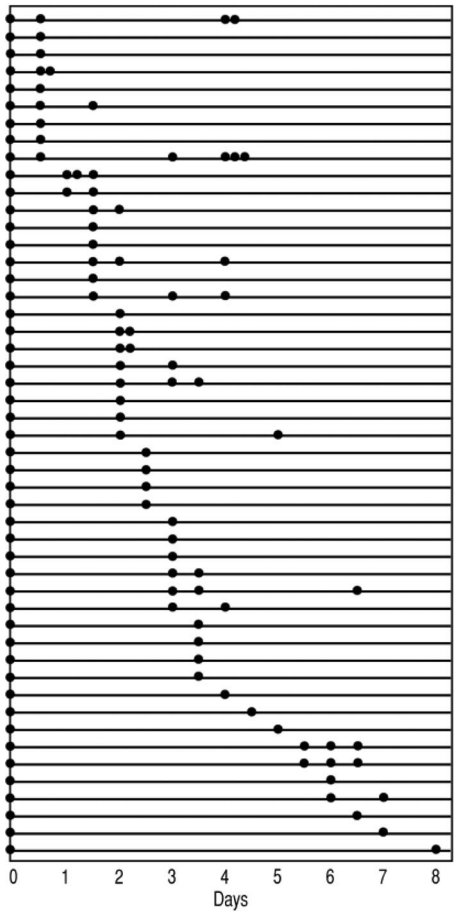


FIGURE 1. Time series for 49 households with secondary infections from Götz et al[8] data. Time of symptom onset, to nearest 12 hours, is denoted by ●.

In addition, we estimate the distribution of the incubation period, using data reported for the Stockholm outbreak[8] on the time lag between the point-source event and the onset of symptoms in the 153 household index cases. A gamma distribution with mean $1/\varepsilon$, and shape parameter ε_s was fit to these incubation time data by maximum likelihood ($1/\hat{\varepsilon} = 1.7$ days; $\hat{\varepsilon}_s = 3.73$ [SE = 0.048]) (Fig. 2). To fit the assumptions of the compartmental transmission model described in the following section, we round the estimated shape parameter to the nearest integer. However, our estimation approach is robust to models with arbitrarily-distributed infectious periods.

When estimating parameters of the infection-process model, we characterize the infectious period as gamma distributed with an unknown mean and shape parameter. Household sizes were not reported in the original outbreak dataset. To address this missing data issue, census data on the distri-

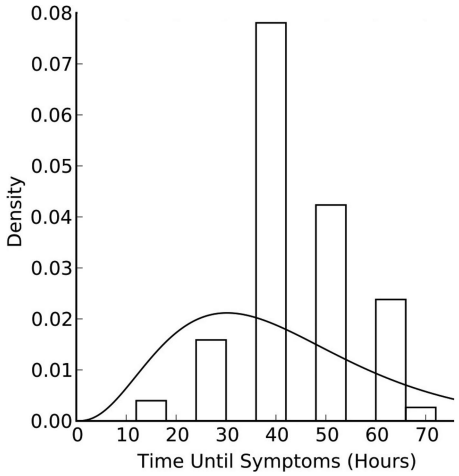


FIGURE 2. Histogram and ML gamma distribution of incubation times from Götz et al[8] data.

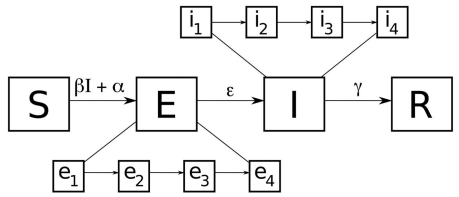


FIGURE 3. Flow diagram showing first and second order compartments in SEIR transmission model. The density-dependent infection rate is β and α is the rate of community transmission. The rate of transition from incubation to symptoms (and infectiousness) and from infectiousness to recovery are ε and γ, respectively.

bution of Swedish household sizes during the study period were incorporated into our analysis.

Because the Stockholm outbreak data include only the time of symptom onset, we are unable to directly estimate the rate at which asymptomatic infections were created. Accounting for asymptomatic infections is important, as they have been estimated to comprise from 12% to 50% of norovirus cases.[11–14] Additional analysis was conducted to assess the impact of increasing levels of asymptomatic infection on our results.

Model

We treat the household infection process as a continuous-time Markov chain, where persons can be in one of 4 states: susceptible (S), exposed/incubating (E), symptomatic/infectious (I) and recovered (R) (Fig. 3). The daily transmission rate, β, is defined as rate of contact at time t multiplied by the probability that contact between a susceptible and an infected person results in transmission. We account for the baseline risk of community and environmental infection through the parameter α, which is measured in terms of the daily risk of infection per susceptible. The incubation and infectious periods are assumed to follow gamma distributions, where each is defined by a mean duration ($1/\varepsilon$, $1/\gamma$) and shape parameter (ε_s, γ_s). The shape parameters for the distributions of the incubation and infectious periods can be thought of as the number of stages that persons pass through before they are either infectious or recovered, respectively. These stages are represented by the first-order compartments in Figure 3.

At any given time, t, the hazard, ω_t, to each susceptible in a household is defined by the force of infection,

$$\omega_t = \beta I_t + \alpha \quad (1)$$

where I_t denotes the total number of infectious persons in a household at time t. Consistent with a Poisson process, we assume that these waiting times are exponentially distributed with mean $1/\omega_t$. Under these assumptions, the probability of observing one or more infections over this interval Δt is the exponential cumulative distribution function.

$$P_{Infection}(t, t + \Delta t) = 1 - \exp(-\omega_t S_t \Delta_t) \quad (2)$$

The classic model for infectious disease dynamics is the flow of hosts among various compartments defined as susceptible, exposed but not infectious, infectious, and recovered (SEIR). To generate sample data for evaluating the statistical method described in the next section, we use the force of infection (Eq. 1), gamma-distributed incubation and infectious periods, and household sizes drawn from the census distribution in a stochastic SEIR simulation model. Implementation details are available in the supplementary materials.

Data Model

First, we define a likelihood function for an infection time series when all 4 individual states (susceptible, exposed/incubating, infectious, and recovered) are observable, and only the transmission parameters β and α are unknown. We then outline a data augmentation method[10] that allows us to extend this likelihood function to the case in which some states are unobserved (Fig. 4).

Likelihood

The household time series is described as a series of system states, $q_{ij} = \{S_{ij},E_{ij},I_{ij},R_{ij}\}$, for each household, i, and state, j, where N_Q is the number of distinct system states in a household time series and $Q_i = \{q_{i,0} \ldots q_{i,N_Q}\}$ is the entire set of states in a household in chronological order (Fig. 4). Beginning times for each system state are denoted t_{ij}. Three state transitions are possible: infection, onset of symptoms (and infectiousness), and recovery. The states of the system immediately before the occurrence of infection events, where infection is defined as a transition into E, are indexed by k and denoted as $v_{ik} \in V_i$, where $V_i \subset Q_i$. The number of infections in a household observation is N_K.

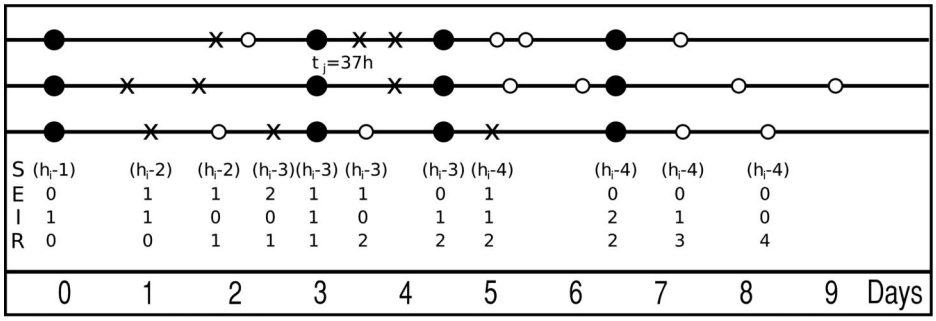


FIGURE 4. Three hypothetical infection histories where the only observed state transition is the onset of symptoms (denoted by ●). Each of the 3 example histories illustrate different possibilities for the 2 unobserved state transitions, infection (denoted by **x**) and recovery (denoted by ○). Values, q_{ij}, under the bottom series are the complete state of the system in household h at state i, where S, E, I, R, are the number of individuals in the susceptible, incubation, infectious and recovered states, respectively; h_j is the number of individuals in household j.

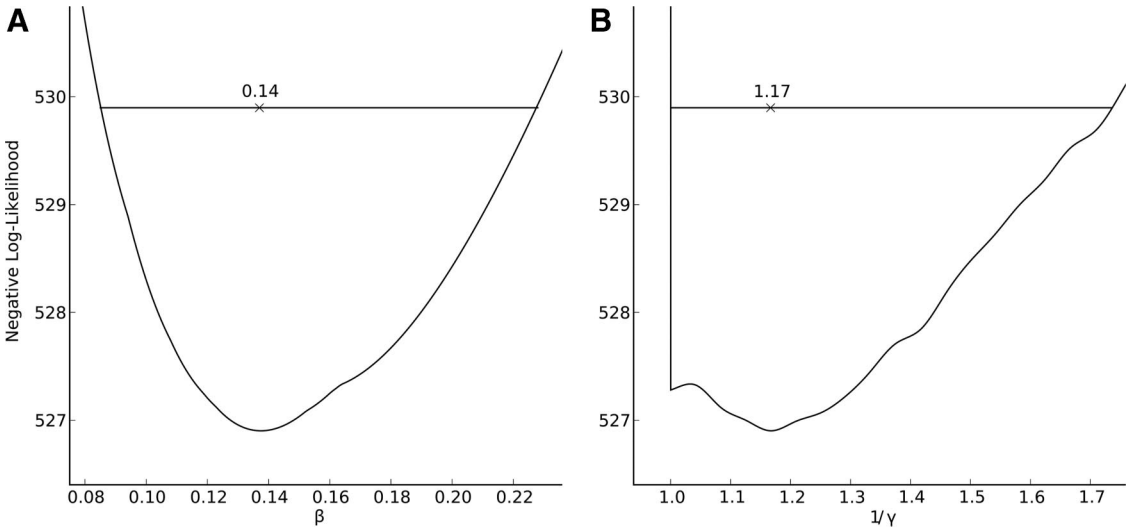


FIGURE 5. Profile likelihood plot of Stockholm outbreak data. Transmission rate (β) and mean infectious period ($1/\gamma$) are on the x-axis in panels A and B, respectively. On the y-axis is negative log-likelihood values for a given β or γ when it is held fixed and the other parameters of interest are optimized. "x" denotes the location of the maximum likelihood estimates and the horizontal bar shows the width of the 95% CI.

With this notation, $q_{i,0}$ corresponds to the state of household i immediately after the onset of symptoms in the index case, and $v_{i,0}$ corresponds to the state of the household immediately before the first household infection.

Assuming that the times of infection, symptom onset, and recovery are known, we can formulate the household likelihood function as the product of 2 terms: (1) the likelihood of observing no new cases during the Δt between all state transitions (ℓ_a) and (2) the likelihood of infection at the time when new infection events are observed (ℓ_b).

The expected number of new infections for a given household, i, at state j, is given by:

$$\lambda(S_{ij},I_{ij},\beta,\alpha) = S_{ij}(\beta I_{ij} + \alpha) \tag{3}$$

The first term, ℓ_a, is the probability of observing no infections over all of the time intervals between states:

$$\ell_{i,a} = \prod_{j=0}^{NQ-1} \exp(-\lambda(S_{ij},I_{ij},\beta,\alpha)(t_{j+1} - t_j)) \tag{4}$$

The second, ℓ_b, describes the joint likelihood of all observed infection events, ie, the product of all instantaneous infection probabilities at times when infection events are observed:

$$\ell_{i,b} = \prod_{k=1}^{N_K} \lambda(S_{ik},I_{ik},\beta,\alpha) \tag{5}$$

Based on these definitions, the likelihood of the data for household i, given β and α, is:

$$\ell_i = \ell_{i,a} \times \ell_{i,b} \quad (6)$$

The product of the likelihoods for all observed households is taken to be the likelihood of the entire observed outbreak, O:

$$\ell_O = \prod_{i \in \mathrm{H}} \ell_{\mathrm{i}} \quad (7)$$

Data Augmentation

The observed data consist of the times of symptom onset in new cases, represented by increments to the household infectious-state variable $I_{i,}$ and, by consequence, decrements to the state variable $E_{i.}$ We do not observe infection events for household cases; this is represented by an increment to the household incubating state E_i and a decrement in the number of susceptibles S_i. We also do not observe recovery from infectiousness, represented by an increment to the household immune state R_i (and decrement in I_i). Because all states are necessary to characterize the transmission dynamics of the system, but only transitions into state I are observed, a method is needed to evaluate the likelihood. To address this missing-data problem, we generate an augmented household time series by sampling from our incubation and infectious period distributions (mean, shape = $1/\varepsilon,\varepsilon_s$ and $1/\gamma,\gamma_s$, respectively) for each case, as described by Cooper et al.[15] We account for right-censoring by following the convention that all recovery times greater than the observation period, t_f, are truncated to be equal to t_f. This returns the correct likelihood of the data when sampled recovery times are outside the observation window. In this way, we create an outbreak realization with all states accounted for. Using this augmented dataset, we can calculate the likelihood. We repeat this process many times, resampling new times from the distributions and calculating a new likelihood each time. The mean of this set of sampled likelihoods approximates the true likelihood of the household time series. This procedure is equivalent to Monte Carlo numerical integration with importance sampling[16] and is depicted visually in Figure 4. (See papers by Rampey et al[17] and Rhodes[18] for alternative approaches to estimating transmission parameters with this type of data.)

We obtain a likelihood estimate for an entire outbreak by augmenting all households 10^4 times and estimating their joint likelihood (Eq. 7). Because we are sampling incubation and infectious periods proportionally from their joint distribution, the expectation of this set of likelihoods approximates the likelihood of the data, given the parameters vector $\theta = \{\alpha,\beta,1/\varepsilon,\varepsilon_s,1/\gamma,\gamma_s\}$.

In the Stockholm outbreak dataset, the number of people in each household is unobserved. We account for these missing data with household size data obtained from a national census[19] and combine this with information from the household observations; the number of household members must be equal to or greater than the number of observed cases. We combine the census distribution with this lower bound on size for each household to construct a conditional distribution of sizes for each household. When an augmented household time series is generated, a size is sampled from this distribution, allowing us to incorporate and bound our uncertainty regarding household sizes when estimating the likelihood. In the following section we will demonstrate that this does not have a significant negative impact on our results. For details on the implementation of the data augmentation procedure, see the eAppendix (http://links.lww.com/EDE/A400).

The Table lists the 2 parameterizations used in the analysis. Parameter set 1 uses case and incubation-period data from the Stockholm outbreak. We estimate the transmission parameter, β, as well as the mean, $1/\gamma$, and shape parameter γ_s of the distribution of the infectious period. We constrain our parameter search to values of $1/\gamma > 1$ day, as durations of symptomatic shedding less than 1 day are biologically implausible.[10,11] Parameter set 2 consists of the population parameter values of a single 153-household outbreak realization from the stochastic model, with household sizes drawn from the census distribution. With these simulated data, we estimate β and $1/\gamma$ under 2 conditions: known household sizes and unknown household sizes.

RESULTS

Figure 5 contains the maximum likelihood estimates and confidence intervals of both the main transmission parameter ($\hat{\beta} = 0.14$ [95% confidence interval {CI} = 0.08–0.24]; Fig. 5A) and average duration of infectiousness ($1/\hat{\gamma} = 1.17$ days [1.00–1.88]; Fig. 5B) for the Stockholm outbreak. We also estimated the shape parameter for the duration of infectiousness ($\gamma_s = 1.0$ [1.0–2.0]; not pictured). Figure 6 is a contour plot showing a 2-dimensional likelihood profile with respect to β and $1/\gamma$. Each cell contains the likelihood corresponding to the optimized value of γ_s for each $(\beta,1/\gamma)$ pair. We also estimate the parameters when $\alpha = 0.01$ and obtain similar results ([$\hat{\beta} = 0.13$ [0.07–0.22]; $1/\hat{\gamma} = 1.0$ days [1.0–1.33]; $\gamma_s = 1.0$ [1.0–2.0]; not pictured). Thus there is likely some bias in our estimated beta due to environmental infection, but this bias is small.

To examine the impact of unknown household sizes, we created a simulated dataset with parameters $\beta = 0.14$ (transmission rate), $\alpha = 0.001$ (background transmission rate), $1/\varepsilon = 1.5$ days, $\varepsilon_s = 4.0$, (incubation period), $1/\gamma = 1.17$ days, $\gamma_s = 1.0$ (duration of infectiousness) (Table, Parameter Set 2). We then estimated 2 of these parameters, the transmission rate and average duration of infectiousness, under 2 conditions: (1) where actual household sizes are explicitly included in the estimation (dashed line: $\hat{\beta}_{unknownHH} =$

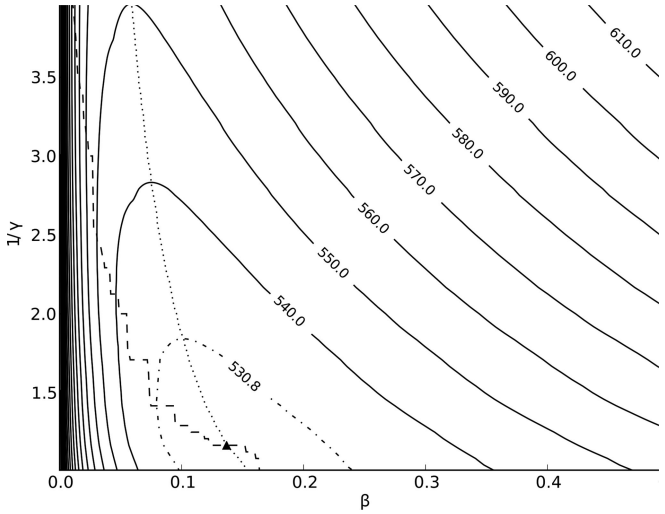


FIGURE 6. Two-dimensional likelihood profile for Stockholm outbreak data. A filled triangle denotes the location of the maximum likelihood estimates. Solid contours bound regions of lesser or equal negative-log-likelihood (NLL) than the contour label. The dash-dotted ellipsoid bounds the 95% confidence region. The dashed line represents the relationship between each value of the transmission rate (β) and the corresponding maximum likelihood estimate of the value of infectiousness period ($1/\gamma$) when β is held fixed. The dotted line represents this relationship in reverse, with points along the x-axis corresponding to maximum likelihood values of β for each γ.

TABLE. Household, Pathogen, and Transmission Parameter Sets

			Parameter Set	
Parameter	**Description**	**Units**	**1**	**2**
β	Within-household infectivity	Infections/day	EST	0.14
α	Community infectivity	Infections/day	0.001	0.001
$1/\epsilon$	Average incubation period	Days	1.7	1.7
ε_s	Incubation period shape		4	4
$1/\gamma$	Average infectious period	Days	EST	1.17
γ_s	Infectious period shape		EST	1
h	Household size	Persons	Unknown	Known

EST indicates parameters to be estimated.

0.139 [95% CI = 0.087–0.273], $1/\hat{\gamma}_{unknownHH}$ = 1.21 days [0.625–1.88], Fig. 7A); and (2) where household sizes are drawn from the census distribution (solid line: $\hat{\beta}_{unkownHH}$ = 0.133 [0.079–0.259] $1/\hat{\gamma}_{unknownHH}$ = 1.21 days [0.63–1.88], Fig. 7B).

Asymptomatic Infection

To understand the impact of unobserved asymptomatic infections, we performed a simulation-based sensitivity analysis that allows us to predict the value of the transmission parameter, β, for varying proportions of asymptomatic infections, τ.

We find that, starting from our maximum likelihood estimate of $\beta = 0.14$ when $\tau = 0$, the predicted value of β increases linearly by approximately 0.035 units for each 10% increase in τ (Fig. 8). For further details on the design and implementation of this analysis, see the eAppendix (http://links.lww.com/EDE/A400).

DISCUSSION

Using a collection of household-exposure and illness-onset time series, we have obtained estimates (and their confidence intervals) for the household person-to-person infection rate and average infectious period for norovirus. We also predict the value of the transmission parameter β as a function of the proportion of asymptomatic infections. We obtained these estimates despite the absence of potentially important data, including infection times, recovery times, and household sizes. The inclusion of census data with household-specific lower bounds (due to the number of observed cases) allowed us to obtain an accurate estimate of household force of infection in the absence of directly observed household sizes.

Although the pattern of contact in households tends to fit the standard mass-action assumption in susceptible-infected-removed models,[20] their typically small sizes require careful consideration of the influence of random variability on results, obviating the use of deterministic models.[21,22] This is a topic that has received considerable attention, and there is an extensive literature on techniques for fitting stochastic models to outbreak data[18,23,24] in a variety of settings (eg, communities,[25] schools[26] and households[27]). Using household-level infection data at the end of an outbreak, Longini et al[24] generated estimates of household and community parameters for the distribution of final household outbreak sizes. However, because their method was developed to explain final-size data from public health reports and does not use temporal information, it provides only limited insights regarding the interaction between infectivity and the durations of the incubation and recovery periods in outbreak time-series.

Hohle et al[28] present a technique that could be useful with household time-series data. They use Bayesian inference to estimate transmission parameters in spatially heterogeneous SEIR models, and innovate on previous Markov-chain-Monte-Carlo-based techniques by allowing variability in the incubation period. Two significant drawbacks of Bayesian approaches are that: (1) even when care is taken to use noninformative prior distributions, these priors can condition estimates,[29] and (2) the results can be difficult to interpret, particularly with respect to reproducibility.[30] We have presented an alternative, frequentist approach that produces maximum likelihood parameter estimates and allows a straightforward exploration of the likelihood surface.

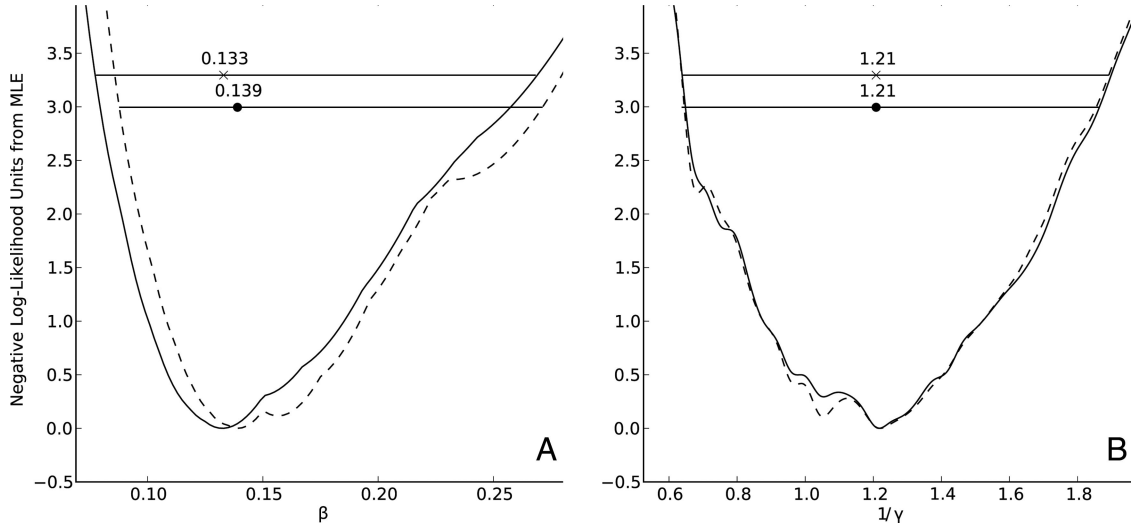


FIGURE 7. Likelihood profiles for simulated data, with respect to transmission rate, β (A) and mean infectious period ($1/\gamma$) (B). The dashed line is a profile where household sizes are known (location of the maximum likelihood estimates is denoted by ●) and the dash-dotted line estimates the parameters in the case where household sizes are uncertain (MLE: "x"), and the horizontal bars span the 95% CI for both cases.

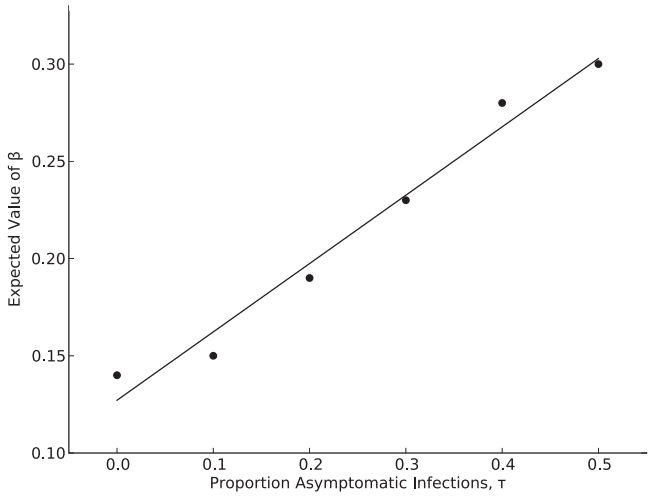


FIGURE 8. Expected household transmission rate, β, by increasing proportion of asymptomatic infections, τ, note that the expected value of β when $\tau = 0$ is 0.14.

Community transmission is undoubtedly more complicated than our representation. Fixing the community transmission parameter, α, to a value 2 orders of magnitude smaller than the household transmission parameter, β, makes the strong assumption that the within-household transmission process is dominant. We show that our results are not very sensitive to this assumption, and we argue that the assumption is reasonable with respect to our data because all households in the Stockholm dataset had a known source of exposure—an index case infected by the point-source outbreak—and all secondary cases identified in households occurred in a plausible temporal sequence. A better estimate of the rate of community transmission requires focused attention on the mechanisms behind this process, which is outside of the scope of both our dataset and this paper. This is an important focus for future research. In addition, the data used in this analysis come from only 9 days of observation, resulting in right-censoring. While our inferences for the transmission rate and effective duration of infectiousness in the course of a household outbreak are valid, they are not generalizable to community or regional scales.

Reliable transmission parameter estimates are critical to risk assessments and exploratory modeling for public health policy. The impact of interventions on norovirus prevalence and persistence can be better assessed in a model such as ours that includes realistic feedback in the transmission process and empirically-derived transmission parameters.

Although the analysis presented here focuses on the transmission of an infectious pathogen in a specific epidemiologic and social context, the methods employed are relevant to other problems in epidemiology and medicine, in which unobserved variables strongly affect outcomes. We have focused on unobserved within-host disease states and household sizes, but other important variables, including contact structures and environmental reservoirs, are often difficult to observe or missing from otherwise-useful public-health surveillance data.

For example, social and economic factors are likely to increase within-household transmission of pathogens such as tuberculosis and shigellosis,[31] by increasing host susceptibility to physical and social stress via mechanisms such as allostatic load and household overcrowding.[32] Administrative records often include important information on the timing,

geographic distribution, and infectious contacts of cases[33] but because of their focus on immediate control, often lack direct observations of contacts that do not result in infections. Consequently, we lack information on how those who become ill and those who escape infection differ in contact patterns and other factors important in transmission. Our work suggests that case-data missing such information can be combined with reasonable, empirically grounded models of contact structures to yield important and useful insights even in the absence of a full dataset. The next step is to apply this approach to different pathogens in more complicated social settings.

ACKNOWLEDGMENTS

We thank Meghan Milbrath for helpful input over many drafts as well as Rick Riolo, Michael Bommarito and the University of Michigan Center for the Study of Complex Systems for technical assistance and the use of computational resources.

REFERENCES

1. Teunis P, Moe CL, Liu P, et al. Norwalk virus: How infectious is it? *J Med Virol*. 2008;80:1468–1476.
2. Widdowson MA, Glass R, Monroe S, et al. Probable transmission of norovirus on an airplane. *JAMA*. 2005;293:1859–1860.
3. Tsang O, Wong A, Chow C, et al. Clinical characteristics of nosocomial norovirus outbreaks in Hong Kong. *J Hosp Infec*. 2008;69:135–140.
4. Atmar RL, Estes MK. The epidemiologic and clinical importance of norovirus infection. *Gastroenterol Clin North Am*. 2006;35:275–290.
5. Patel M, Hall A, Vinje J, et al. Noroviruses: A comprehensive review. *J Clin Virol*. 2009;44:1–8.
6. Caul EO. Viral gastroenteritis: small round structured viruses, calciviruses and astroviruses Part I. The clinical and diagnostic perspective. *J Clin Pathol*. 1996;49:874–880.
7. Lopman BA, Adak GK, Reacher MH, et al. Two epidemiologic patterns of norovirus outbreaks: surveillance in England and Wales, 1992–2000. *Emerg Infect Dis*. 2003;9:71–77.
8. Götz H, Ekdahl K, Lindbäck J, et al. Clinical spectrum and transmission characteristics of infection with Norwalk–like virus: findings from a large community outbreak in Sweden. *Clin Infect Dis*. 2001;33:622–628.
9. Atmar RL, Opekun AR, Gilger MA, et al. Norwalk virus shedding after experimental human infection. *Emerg Infect Dis*. 2008;14:1553–1557.
10. Kirkwood C, Streitberg R. Calicivirus shedding in children after recovery from diarrhoeal disease. *J Clin Virol*. 2008;43:346–348.
11. Gallimore CI, Cubitt D, du Pleiss N, et al. Asymptomatic and symptomatic excretion of noroviruses during a hospital outbreak of gastroenteritis. *J Clin Microbiol*. 2004;42:2271–2274.
12. Ozawa K, Oka T, Takeda N, et al. Norovirus infections in symptomatic and asymptomatic food handlers in Japan. *J Clin Microbiol*. 2007;45:3996–4005.
13. Goller J, Dimitriadis A, Tan A, et al. Long-term features of norovirus gastroenteritis in the elderly. *J Hosp Infect*. 2004;58:286–291.
14. Parashar UD, Dow L, Fankhauser RL, et al. An outbreak of viral gastroenteritis associated with consumption of sandwiches: implications for the control of transmission by food handlers. *Epidemiol Infect*. 1998;121:615–621.
15. Cooper B, Medley G, Bradley S, et al. An augmented data method for the analysis of nosocomial infection data. *Am J Epidemiol*. 2008;168:548–557.
16. Robert CP, Casella G. *Monte Carlo Statistical Methods*. New York: Springer; 2004.
17. Rampey AH, Longini IM, Haber M, et al. A discrete-time model for the statistical analysis of infectious disease incidence data. *Biometrics*. 1992;48:117–128.
18. Rhodes PH. Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *J R Stat Soc Series B Stat Methodol*. 1996;58:751–761.
19. United Nations Population Survey, United Nations (New York), 2008. Available at: http://www.un.org/esa/population/unpop.htm.
20. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford Science Publications; 1992.
21. Koopman JS. Modeling infection transmission. *Annu Rev Public Health*. 2004;25:303–326.
22. Matthews L, Woolhouse M. New approaches to quantifying the spread of infection. *Nat Rev Microbiol*. 2005;3:529–536.
23. Ionides EL, Breto C, King A. Inference for nonlinear dynamical systems. *Proc Natl Acad Sci USA*. 2006;103:18438–18443.
24. Longini IM, Koopman JS, Monto A, et al. Estimating household and community transmission parameters for influenza. *Am J Epidemiol*. 1982;115:736–751.
25. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. *Nature*. 2008;454:877–880.
26. O'Neill PD, Marks PJ. Bayesian model choice and infection route modelling in an outbreak of Norovirus. *Stat Med*. 2005;24:2011–2024.
27. Longini IM, Koopman JS. Household and community transmission parameters from final distributions of infections in households. *Biometrics*. 1982;38:115–126.
28. Hohle M, Jorgenson E, O'Neill PD. Inference in disease transmission experiments by using stochastic epidemic models. *Appl Stat*. 2005;54:349–66.
29. Press SJ. *Subjective and Objective Bayesian Statistics*. New York: Wiley; 2003.
30. Lele S, Dennis B, Lutscher F. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol Lett*. 2007;10:551–563.
31. Wallace R. A synergism of plagues: "planned shrinkage," contagious housing destruction, and aids in the Bronx. *Environ Res*. 1988;47:1–33.
32. House JS, Landis KR, Umberson D. Social relationships and health. *Science*. 1988;241:540–545.
33. Jones RC, Liberatore M, Fernandez JR, et al. Use of a prospective space-time scan statistic to priorities shigellosis case investigations in an urban jurisdiction. *Public Health Rep*. 2006;121:131–139.

eAppendix

1. Stochastic SEIR Transmission Model Implementation

A sample outbreak is initialized by creating 153 households, with sizes h_i, drawn from the census distribution of household sizes. The initial household state is set to $q_{i,0} = \{(h_i - 1),0,1,0\}$, indicating that only the index case is symptomatic, all other household members being susceptible. The transmission model is summarized in the algorithm below, where S, E, I and R are the number of individuals in each state and the model is initialized at t=0:

```
If E + I > 0:
    For s in S:
        Draw x from Uniform(0,1]
        If x <= 1 − exp(−(βI + α)dt):
            S = S − 1
            E = E + 1
            Draw symptom onset time from Gamma(1/ε, ε_S)
            Draw recovery time from Gamma(1/γ, γ_S)
    t = t + dt

At end of step, transition from E → I and I → R those who have
symptom onset or recovery time <= t
```

eAlgorithm 1

The model is stepped forward in hourly increments ($dt = 1/24$), which gives a reasonable approximation of a continuous time infection process. Rates are expressed in terms of days but scaled to the appropriate time step.

The incubation and infectious periods are conceptualized as a sequence of e_s and i_s second-order compartments, with the probability of transition between these compartments for each individual equal to $(\varepsilon \cdot \varepsilon_s)dt$ and $(\gamma \cdot \gamma_s)dt$. This process yields $E \rightarrow I$ and $I \rightarrow R$ transition rates that are gamma distributed with means e, g and shape

parameters e_s, g_s, respectively. Transmission rates are also scaled in terms of dt (see Equation 1).

2. Asymptomatic Infections

To assess the effect of unobserved asymptomatic infections, we implemented the stochastic SEIR model outlined above, with an additional parameter, τ, that controls the proportion of new infections that are asymptomatic:

```
If E + I > 0:
    For s in S:
        Draw x from Uniform(0,1]
        If x <= 1 − exp(−(βI + α)dt):
            Draw y from Uniform(0,1]
            If y <= τ:
                S = S − 1
                R = R + 1
            Else:
                S = S − 1
                E = E + 1
                Draw symptom onset time from Gamma(1/ε, ε_S)
                Draw recovery time from Gamma(1/γ, γ_S)

    t = t + dt

At end of step, transition from E → I and I → R those who have symptom
onset or recovery time <= t
```

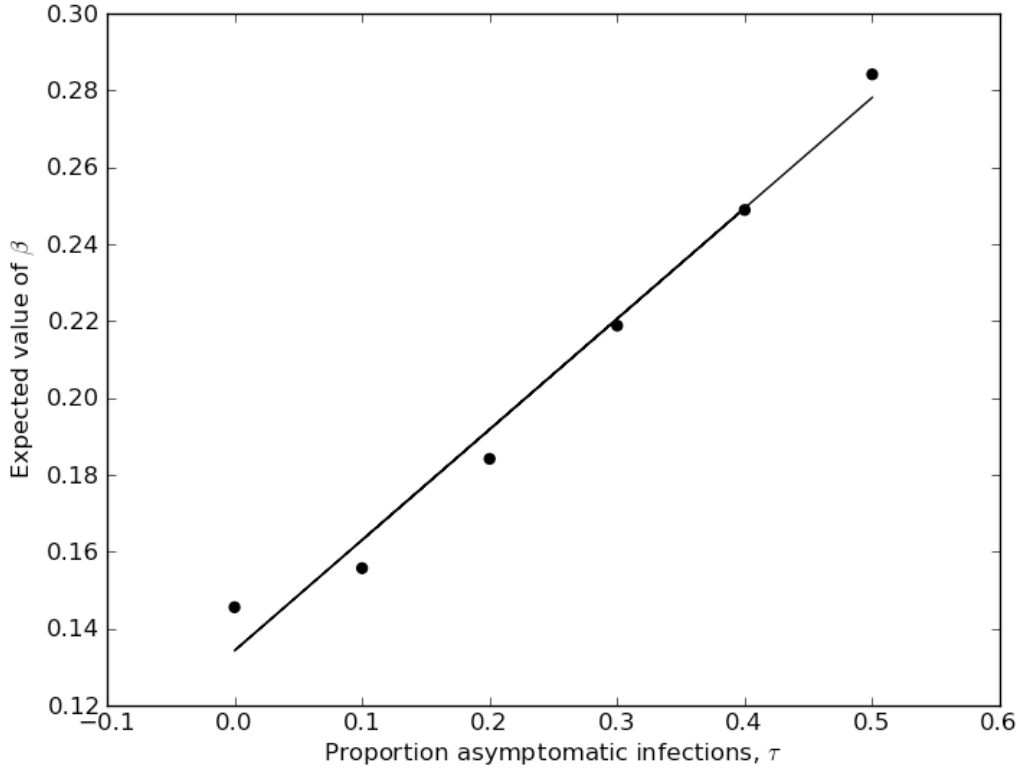
eAlgorithm 2

Asymptomatic infections are, in this simplified model, immediately moved to the immune class. This is because they are significantly less infectious than symptomatic infections, e.g., (10), and can be expected to generate cases on a longer timescale than our window of observation (9 days). Although they are unlikely to contribute significantly to observed within-household transmission dynamics, we expect that they are important to

the community-level persistence of norovirus and, as such, need to be accounted for in the estimate of rate of transmission. In this context, then, asymptomatic cases can be thought of as censored data that bias our estimate of the force of infection.

When simulating outbreaks, we fix the background infection rate and the distribution of the incubation and infectious periods,($\alpha = 0.001$, $1/e = 1.7$ days, $e_S = 4.0$, $1/g = 1.14$ days, $g_S = 1.0$) and allow the transmission parameter, β, and proportion of asymptomatic infections, τ, to vary. We then sample all 126 parameter combinations from $\beta = \{.10, .11, \ldots, .30\}$ and $\tau = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. We draw 20 stochastic realizations of each parameter set and estimate the mean ML value of β (i.e., average over the 20 runs) for each (τ, β) combination, as though $\tau = 0$. This gives a predicted value of β for each level of τ. Starting from our ML estimate of 0.14 for β when $\tau = 0$, the predicted value of β increases linearly by 0.035 units for each 10% for increase in τ (Figure 8).

We test the sensitivity of these results to the assumption that asymptomatic individuals do not contribute to household transmission by allowing asymptomatic infections to be 10% as infectious as symptomatic ones. We find broadly similar results, with the predicted value of β increasing linearly by 0.028 units for each 10% increase in τ (eFigure 1).



eFigure 1. Relationship of proportion asymptomatic to expected value of β when asymptomatic infections are 10% as infectious as symptomatic infections.

3. Missing Household Sizes

Since all households in our dataset consist of two or more people, the minimum household size, *h*, is 2. We start with the empirical distribution of household sizes from a 1990 census of household sizes in Sweden (see eTable), denoted as *C,* where *C(h)* is the probability of observing a household of size *h* in the total population .

If the minimum possible number of individuals, i.e., the number of infections observed in a household, $h_{\min}$, is less than or equal to 2, the entire empirical distribution is used to sample a household size. If $h_{\min}$ = , the number of cases observed is set as the minimum household size, with values smaller than $h_{\min}$ assigned a density of zero. We

assume that the case data provide no additional information on the distribution of the remaining household sizes, so the remaining sizes on the interval $h_{\min} \le h \le 10$ are assigned a uniform density.

This information is combined with the census data in the top row of eTable for each size to generate a distribution from which we can sample household sizes for h ≥ $h_{\min}$:

$$P(h \mid C, h_{\min}) = \frac{C(h)}{\sum_{h \ge h_{\min}}^{10} C(h)}$$

eEquation

In order to sample random variates from this distribution, we compute the conditional CDF of the household size distribution and draw a random number on the interval (0,1], and select the smallest value of *h* where the CDF is less than equal to the random number.

The second row of eTable shows the probability distribution resulting from this sampling procedure. We find that the expected household size increases slightly from 3.73 to 3.87 individuals, with most of this change accounted for by a decrease in the density of households of size 2 to slightly larger ones.

	# Household Members								
	2	3	4	5	6	7	8	9	10
Census Density	0.325	0.193	0.248	0.108	0.027	0.041	0.024	0.017	0.017
Sampled Density	0.283	0.192	0.265	0.115	0.031	0.047	0.027	0.018	0.019

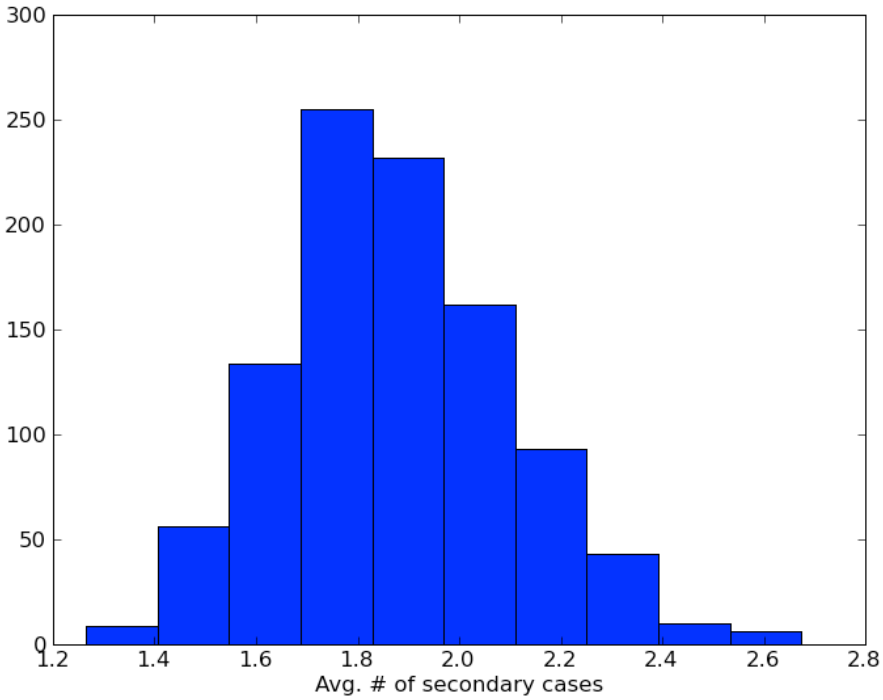
eTable. Empirical Probability Distribution of Household Sizes

4. Model Validation

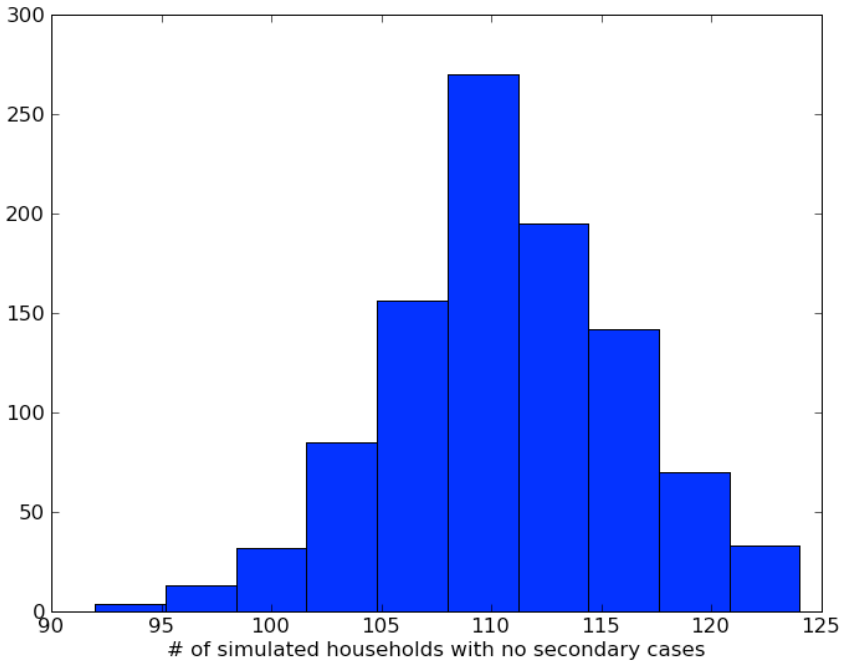
In order to validate the SEIR model used for simulation and parameter estimation, we performed additional simulation analysis using a Gillespie[1] algorithm-based implementation of the model described in eAlgorithm 1, which is an exact, continuous-time simulation of the transmission model.

In each simulation, there are 153 households, the sizes of which are drawn from *C*, the empirical distribution of household sizes. At *t=0*, each household has a single index case. Model parameters are the same as those obtained from our statistical analysis (β = 0.14, $1/\gamma$=1.17 days, γ_s = 1.0). For each of 1000 simulations, we record the number of households with no secondary cases, i.e., where there is stochastic die-out, and the average number of cases in households with secondary cases.

We find that our simulation results are in good agreement with the Stockholm data for both outbreak size (Simulated mean = 1.9 cases, SD = .2, vs. 1.6 for Stockholm data; eFigure 2) and the number of simulated households in which there are no secondary cases (Simulated mean = 110.5 households, SD = 5.5 vs. 104 households for Stockholm data; eFigure 3).



eFigure 2. Histogram of average number of secondary cases in simulated household outbreaks.



eFigure 3. Histogram of number of households with no secondary cases.

5. Computational Details

Data augmentation software was implemented in C++ and Python 2.6 using *Boost.Python* and the *Numpy* and *Scipy* numerical and scientific computing libraries. Plots were generated with *Matplotlib* 0.98 graphing and plotting tools for Python. All diagrams were created in *Inkscape* 0.47.

All results presented here come from 10^4 independent samples for each parameter combination.

References

1.Gillespie, D.T. (1976). "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". *Journal of Computational Physics* **22** (4): 403–434