Estimation of the Household Secondary Attack Rate:  
Binomial Considered Harmful

Yushuf Sharker and Eben Kenah
Biostatistics Department and Emerging Pathogens Institute
University of Florida
May 4, 2017

Abstract

One of the primary goals of household studies of infectious disease transmission is to estimate the household secondary attack rate (SAR), the probability of direct transmission from an index case $A$ to a susceptible household member $B$ during $A$’s infectious period. In a household with $m$ susceptibles and a single index case, the number of secondary infections is often treated as a binomial($m, p$) random variable where $p$ is the SAR. This assumes that all subsequent infections in the household are transmitted directly from the index case. Because a given transmission chain of length $k$ from $A$ to $B$ has probability $p^k$, it is thought that chains of length $k > 1$ can be ignored when $p$ is small. However, the number of transmission chains of length $k$ from $A$ to $B$ can be large, so the total risk of infection through any chain of length $k$ can be much greater than $p^k$. In simulations, we show that estimation of the SAR using a binomial model is biased upward and produces confidence intervals with poor coverage probabilities. Accurate point and interval estimates of the SAR can be obtained using chain binomial models or pairwise survival analysis.
In infectious disease epidemiology, the household secondary attack rate (SAR) is the probability of disease transmission from an infected household member \( A \) to a susceptible member \( B \) during \( A \)’s infectious period. The SAR is used to assess the transmissibility of disease and to evaluate control measures \([1,2]\). Household surveillance data from emerging infections is often used to estimate the SAR \([2,8]\). Almost all of these studies assume that the number of secondary infections in a household is a binomial\((m, p)\) random variable, where \( m \) is the number of susceptible household members and \( p \) is the household SAR. This assumption seems reasonable because a given transmission chain of length \( k \) occurs with probability \( p^k \), which is negligible when \( p \) is small and \( k > 1 \).

This reasoning neglects the fact that the number of possible transmission chains of length \( k \) can be large, and any one of these paths can transmit infection from an index case \( A \) to a susceptible household member \( B \). A transmission chain of length \( k > 1 \) from the \( A \) to \( B \) can be specified by choosing \( k - 1 \) individuals from the \( m - 1 \) other susceptible household members. Each ordering of these \( k - 1 \) individuals determines a unique transmission chain, so the total number of paths of length \( k \) from \( A \) to \( B \) within the household equals the number of permutations of \( k - 1 \) objects chosen from \( m - 1 \) objects, which is

\[
P(m - 1, k - 1) = \frac{(m - 1)!}{(m - k)!}
\]

for \( 1 \leq k \leq m - 1 \). Table 1 shows that the number of paths of length \( k \) from a given susceptible to the index case can grow quickly with household size. Because of the risk of infection along each of these paths, the total risk of infection in susceptible household members can be much greater than the SAR. In a binomial model, this additional risk of infection is attributed erroneously to direct transmission from the index case, so the point estimate of the SAR is biased upward.

There are two ways around this difficulty. In a disease where the latent period (between infection and the onset of infectiousness) is much longer than the infectious period, multiple
generations of infection within the household would be clearly separated. A binomial model could be used to estimate the risk of infection within a time interval designed to capture only the first generation of infection. However, most infectious diseases important to public health can have overlapping generations of infection within households.

A second way around this difficulty is to change the target parameter. Instead of estimating the household SAR, we could estimate the probability that a susceptible is infected by transmission within his or her household, whether or not the infection is transmitted directly from the index case. We’ll call this the household final attack rate (FAR). When the household SAR is constant, the household FAR increases with household size as shown in Figure 1. A constant FAR can occur only when the SAR decreases with household size, so the change in target parameter leads to a different model of transmission unless all households have the same size. While the SAR remains clearly defined in households with multiple index cases or transmission from outside the household, the FAR does not extend so easily to these settings. Here, we assume a single index case and no transmission from outside the household so both the SAR and FAR are clearly defined.

To estimate the FAR, each household must be followed up for the entire duration of the within-household outbreak. An unbiased estimate of the household FAR could then be obtained by calculating the proportion of susceptibles who are infected, as in a binomial model. Each time a susceptible household member is infected, the risk of infection in the remaining susceptibles increases. The binomial variance assumes that infections in different household members are independent, but infections within a household are positively correlated. Because of this correlation, the binomial model underestimates the variance of the number of infections in a household. Thus, binomial confidence intervals will have poor coverage probabilities. To address this issue, cluster adjusted variances have been used to account for correlation among household members [2].

In this paper, we use analytical calculations and simulations to illustrate the dangers of using a binomial model to analyze the transmission of infection within households. Our
results indicate that the household FAR can be estimated accurately as long as the variance is
cluster-adjusted to account for the correlation of infection within households. To estimate the
household SAR, chain binomial models (9,10) or pairwise survival analysis (11,12) should be
used. We use household surveillance data collected by the Los Angeles County Department
of Public Health during the 2009 influenza A (H1N1) pandemic to illustrate the practical
advantages of these methods over a binomial analysis.

Methods

Our analytical calculations and simulations assume a uniform secondary attack rate within
households and no infection from outside the household, which are assumptions shared by
statistical analyses of household transmission based on the binomial model. We used proba-
bility generating functions (PGFs) to calculate the true outbreak size distributions at differ-
ent combinations of number of susceptibles ($m$) and SAR ($p$). We then performed simulations
of household outbreaks at these same combinations of household size and SAR.

Household outbreak size distributions

Assume that each infectious member of a household makes infectious contact with each other
member of the household with probability $p$ during his or her infectious period. Let $p_{mi}$ be
the probability that $i$ out of $m$ susceptibles are infected by within-household transmission in
a household with a single index case. Then

$$g_m(x) = \sum_{i=0}^{m} p_{mi}x^i$$

(2)

is the probability generating function (PGF) for the outbreak size distribution in a house-
hold with $m$ susceptibles and one index case. A household with zero susceptibles has zero
secondary infections with probability one, so $g_0(x) = 1$. 

4
The PGF for the outbreak size distribution in a household with \( m + 1 \) susceptibles can be derived from the PGFs for smaller households. Imagine a household with \( m \) susceptibles of whom \( i \) were infected. Now imagine that the household had one more susceptible. The additional susceptible person would remain uninfected only if he or she escaped infection from all \( i + 1 \) infected household members, which occurs with probability \((1 - p)^{i+1}\). If the additional susceptible is not infected, the total number of infections in the household is \( i \).

With probability \( 1 - (1 - p)^{i+1} \), the additional susceptible is infected and acts like an index case in a household containing the \( m - i \) susceptibles who were not infected. In this case, there are \( i + 1 \) infections plus the number infected among the \( m - i \) remaining susceptibles, which has the PGF \( g_{m-i}(x) \). Therefore,

\[
g_{m+1}(x) = \sum_{i=0}^{m} p_{mi} [(1 - p)^{i+1} x^i + (1 - (1 - p)^{i+1}) x^{i+1} g_{m-i}(x)] \tag{3}
\]

The first few iterations yield

\[
g_0(x) = 1, \tag{4}
\]
\[
g_1(x) = (1 - p) + px, \tag{5}
\]
\[
g_2(x) = (1 - p)^2 + 2p(1 - p)^2 x + (3p^2 - 2p^3)x^2 \tag{6}
\]

which can be checked by hand. We calculated these polynomials using Python code available in the Web Appendix (SARcode.py). As shown in equation \(2\), the coefficient on \( x^i \) in the PGF \( g_m(x) \) is the probability that \( i \) of \( m \) susceptibles are infected in a household outbreak started by a single index case. Using these probabilities, we can calculate the mean and variance of the number of infections among the \( m \) susceptibles.
Household outbreak simulations

We simulated household outbreaks using Erdős-Rényi random graphs \cite{13,14}, where each pair of nodes is connected independently with probability $p$. In our graphs, each node represents a household member and $p$ is the SAR. One node is fixed as the index case, and all household members connected to the index case by a series of edges are infected.

We performed 10,000 simulations for each combination of household size and SAR. In each simulation, there were 200 independent households of the same size. We used logistic regression to calculate the proportion of susceptible household members who were infected with a 95% confidence interval. We then calculated a cluster-adjusted confidence interval using the multi-way clustering method of Cameron, Gelbach, and Miller \cite{15}, which is implemented in the R package \texttt{multiwayvcov} \cite{16}. All confidence intervals were calculated on the logit scale as $\hat{\beta} \pm 1.96 \sigma$ where $\hat{\beta} = \logit(\hat{p})$ is the estimated log odds of infection and $\sigma$ is the unadjusted or cluster-adjusted standard error estimate. Finally, we transformed the confidence intervals to the probability scale and estimated the coverage probabilities for the true household SAR and the true household FAR.

Simulations were implemented in Python version 3.5.1 (\url{www.python.org}) using NumPy version 1.11 (\url{www.numpy.org}), NetworkX version 1.11 (\url{networkx.github.io}), and pandas version 0.18.0 (\url{www.pandas.pydata.org}). Statistical analysis was performed in R version 3.3 (\url{www.r-project.org}) using RPy2 version 2.7.0 (\url{rpy2.bitbucket.io}). These software packages are free and open-source, and the code is available in the Web Appendix (SARcode.py).

Household data analysis

We use influenza A (H1N1) household surveillance data collected by the Los Angeles County Department of Public Health (LACDPH) between April 22 and May 19, 2009 to give a practical example of biased estimation of the household SAR using a binomial model. The data was collected using the following protocol \cite{11}:  

1. Nasopharyngeal swabs and aspirates were taken from individuals who reported to the LACDPh or other health care providers with acute febrile respiratory illness (AFRI), defined as a fever ≥ 100°F plus cough, core throat, or runny nose. These specimens were tested for influenza, and the age, gender, and symptom onset date of the AFRI patient were recorded.

2. Patients whose specimens tested positive for pandemic influenza A(H1N1) or for influenza A of undetermined subtype were enrolled as index cases. Each of them was given a structured phone interview to collect information about his or her household contacts. They were asked to report the symptom onset date of any AFRI episodes among their household contacts.

3. When necessary, a follow-up interview was given 14 days after the symptom onset date of the index case to assess whether any additional AFRI episodes had occurred in the household, including their illness onset date.

For simplicity, we assume all AFRI episodes among household members were caused by influenza A(H1N1) and that all household members except the index case were susceptible to infection. All analyses use natural history assumptions adapted from Yang et al. [17] and identical to those in Kenah [18,19]. In the primary analysis, we assumed an incubation period of 2 days, a latent period of 0 days, and an infectious period of 6 days. In a sensitivity analysis, we also considered 8-day and 10-day infectious periods.

We estimated the household SAR for 2009 pandemic influenza A(H1N1) using unadjusted and cluster-adjusted binomial models, parametric pairwise survival analysis, and a chain binomial model. All analysis was conducted in R version 3.3 (www.r-project.org). Pairwise survival models are implemented in the R package transtat, which is available on GitHub (github.com/ekenah/transtat). The code is available in the Web Appendix (LAsar.R).

To see how well the SAR estimates fit the data, we simulated outbreaks in the Los Angeles households using SAR point estimates from the binomial model, the chain binomial
model, and pairwise survival models. We performed 2,000 simulations for each estimate. In each simulation, we calculated the number of infections among susceptible household members. These simulations were implemented in Python version 3.5.1 (www.python.org) using NetworkX version 1.11 (networkx.github.io). Statistical analysis was performed in R version 3.3 (www.r-project.org) using RPy2 version 2.7.0 (rpy2.bitbucket.io). The code is available in the Web Appendix (SAR.rgraph.py).

Binomial models  The binomial analysis was conducted in two different ways. First, we used an intercept-only logistic regression model to get unadjusted and multi-way cluster adjusted (15) confidence intervals exactly as in the simulation studies. Second, we used a generalized estimating equations (GEE) model to get a second set of cluster-adjusted confidence intervals (20).

Pairwise survival analysis  Pairwise survival analysis estimates failure times in ordered pairs consisting of an infectious individual and a susceptible household member. The pair AB is at risk of transmission starting with the onset of infectiousness in A, and failure occurs if A infects B. This failure time, called a contact interval is right-censored if B is infected by someone other than A or if observation of the pair stops. To account for uncertainty about who-infected-whom, the overall likelihood is the sum of the likelihoods for all possible combinations of who-infected-whom consistent with the data (11). The survival function \( S(\tau, \theta) \) is the probability that the contact interval is greater than \( \tau \), where \( \theta \) is a parameter vector. If \( \theta_0 \) is the true value of the parameter and the infectious period is \( \iota \), then the household SAR is \( 1 - S(\iota, \theta_0) \). To get a point estimate of the SAR, the unknown true value \( \theta_0 \) is replaced by a maximum likelihood estimate \( \hat{\theta} \).

We used the exponential, Weibull, and log-logistic distributions. For the exponential distribution, \( S(\tau, \lambda) = \exp(-\lambda \tau) \) where \( \lambda \) is the called the rate parameter. For the Weibull distribution, \( S(\tau, \lambda, \kappa) = \exp[-(\lambda \tau)^\kappa] \) where \( \lambda \) is the rate and \( \kappa \) is called the shape parameter. For the log-logistic distribution, \( S(\tau, \lambda, \kappa) = [1 + (\lambda \tau)^\kappa]^{-1} \) for rate \( \lambda \) and shape \( \kappa \). For all
three distributions, \( \lambda > 0 \) and \( \kappa > 0 \) so we defined our likelihoods in terms of their natural logarithms \( \ln \lambda \) and \( \ln \kappa \). Standard maximum likelihood estimation was used to get point estimates and a covariance matrix for these parameters. To get a 95% confidence interval for the SAR, we sampled \( \ln \lambda \) and \( \ln \kappa \) from their approximate multivariate normal distribution, calculated the household SAR for each one, and took the 2.5% and 97.5% quantiles of this distribution as confidence limits.

**Chain binomial model** The chain binomial model assumes that each susceptible household member escapes infection from each infectious household member with probability \( \alpha_0 \) each day. An individual \( B \) who is exposed to \( k \) infectious household members on day \( t \) will escape infection on day \( t \) with probability \( \alpha_0^k \) and will be infected on day \( t \) with probability \( 1 - \alpha_0^k \). The likelihood contribution of \( B \) is the product of these likelihood contributions over all days where \( B \) was at risk of infection, and the overall likelihood is the product of the likelihood contributions of all susceptibles who were at risk of infection for at least one day. If the infectious period is \( \iota \) days, then the household SAR is \( 1 - \alpha_0^\iota \). To get a point estimate of the SAR, the unknown true \( \alpha \) is replaced by a point estimate \( \hat{\alpha} \). Because \( \alpha \in [0, 1] \), our likelihood was defined in terms of \( \text{logit}(\alpha) = \ln \frac{\alpha}{1-\alpha} \). Standard maximum likelihood estimation was used to get a point estimate \( \hat{\alpha} \) and a variance estimate. Confidence intervals were calculated on the logit scale and transformed back to the probability scale. For simplicity, we have assumed that the probability of escaping infection from an infectious household member does not depend on how long he or she has been infectious or on any pairwise or individual-level covariates. More sophisticated models can allow the escape probability to vary with the time since infection or with covariates \([9][10]\).
Results

Household outbreak simulations

Figure 1 shows the household FAR calculated using PGFs (lines) and from simulations (symbols) as a function of the true SAR. There is excellent agreement between the analytical calculations and the simulations. Both show that the household FAR is always larger than the household SAR when there is more than one susceptible. Thus, a binomial model will produce a point estimate of the SAR that is biased upward whenever there is more than one susceptible household member. At a fixed SAR, the difference between the SAR and the FAR increases with household size.

Figure 2 shows the household SAR coverage probabilities for unadjusted and cluster-adjusted binomial 95% confidence intervals. The coverage probabilities are always lower than 95%. Even for small households, the coverage probabilities decrease rapidly as the true SAR increases. Cluster adjustment increased the coverage probabilities only slightly. With or without adjustment for clustering by household, a binomial model cannot produce reliable point or interval estimates of the household SAR.

The simulation results in Figure 1 show that a binomial model can produce accurate point estimates of the household FAR. Figure 3 shows coverage probabilities of unadjusted and cluster-adjusted 95% confidence intervals for the household FAR. The coverage probabilities for unadjusted confidence intervals are always below 95%, and they decrease with increasing household size and SAR. Adjustment for clustering by household corrects this problem, producing coverage probabilities near 95% for all household sizes. Thus, a binomial model can produce reliable point and interval estimates of the household FAR as long as clustering within households is taken into account.
Household data analysis

In the LACDPH pandemic influenza A(H1N1) data, there were 58 households with a total of 299 members. There were 99 infections, of which 62 were index cases (4 of the 58 households had co-primary cases) and 27 were household contacts with an AFRI. The median household size was 5 (range: 2-20), so multiple infection pathways within households is a practical—not just theoretical—problem for estimation of the household SAR.

Table 2 shows point estimates and 95% confidence intervals for the household SAR. As expected, a binomial model produces a much higher estimate than pairwise survival or the chain binomial models. Adjustment for clustering produced a wider confidence interval, and multiway cluster adjustment and GEE produced very similar results. The pairwise survival and chain binomial models produced nearly identical point estimates of the household SAR. Among the pairwise survival models, the exponential had the lowest Akaike Information Criterion (AIC) at 280.5 versus 281.9 for the Weibull model and 281.8 for the log-logistic model. The exponential model produced a much narrower confidence interval than the other two pairwise survival models. Among chain binomial models, the closest equivalent to the exponential pairwise survival model is the model with a constant $\alpha$, so we did not attempt to fit more complex models. The confidence interval from the chain binomial model is wider than the exponential pairwise survival model but narrower than those of the Weibull and log-logistic models. Similar results were obtained for 8-day and 10-day infectious periods, with slightly higher estimates for the household SAR (not shown).

Figure 4 shows histograms of the simulated numbers of infections for the binomial estimate, the chain binomial model, and the exponential and log-logistic pairwise survival models. In each panel, a vertical black line indicates the 27 observed infections. Simulations using the binomial estimate always produced many more infections than were observed. Simulations using estimates from the chain binomial and pairwise survival models predict a number of infections with a distribution centered near the observed number. The true household SAR, to the extent one exists, is almost certainly far below the binomial estimates and
close to the estimates from the other models.

**Discussion**

Even when the SAR is small, it is important to account for the risk of infection through multiple generations of infection within households. Unless generations of infection are clearly separated in time, a binomial estimate of the SAR will be biased upward and have a confidence interval with low coverage probability. A binomial model can estimate the household FAR accurately if cluster-adjusted confidence intervals are used. However, most epidemiologic questions involving person-to-person transmission are framed more naturally in terms of the SAR.

Using logistic regression, it is possible to generalize an analysis based on a constant FAR to associate the risk of infection in a household outbreak with individual-level or household-level covariates. Similarly, an analysis based on the SAR can be generalized to allow the probability or hazard of transmission to depend on individual-level, pairwise, and household-level covariates. In the ordered pair \((ij)\), the individual-level covariates of \(i\) can be associated with infectiousness and the individual-level covariates of \(j\) can be associated with susceptibility, allowing simultaneous estimation of covariate associations with both (19). Analysis based on the SAR allows more detailed scientific insight into disease transmission than an analysis based on the FAR.

The household SAR can be clearly defined and estimated even when there are multiple index cases, a risk of infection from outside the household, or loss to follow-up (10). The chain binomial model can include a probability of infection from the environment or community outside the household in each time unit, and pairwise survival analysis can include a hazard of infection from the environment or community. Both models can explicitly account for right-censored observations or delayed entry (11). Models based on the household FAR are difficult or impossible to extend to these situations.
Infectious disease data from households or other groups of close contacts should always be analyzed using chain binomial models or pairwise survival analysis to estimate hazards or probabilities of transmission. The primary obstacle to the adoption of these methods has been the lack of available software. Chain binomial models are available in the free software package TranStat (www.cidid.org/transtat), which incorporates several advanced methods (21, 22) and has been used in analyses of influenza (17), Zika (23), and Ebola virus (24) transmission. The free and open source transtat package for R, which we used to analyze the LA household data above, includes parametric pairwise survival models and will be expanded to include nonparametric (18) and semiparametric (19) pairwise survival models as well as chain binomial models. We hope the use of statistical methods designed specifically for person-to-person transmission will help infectious disease epidemiologists contribute even more effectively to science and public health.

Acknowledgments

Brit Oiulfstad, Dee Ann Bagwell, Brandon Dean, Laurene Mascola, and Elizabeth Bancroft of the Los Angeles County Department of Public Health generously provided the household influenza surveillance data. The subtitle of the paper was inspired by Edsger Dijkstra’s letter “Go To Statement Considered Harmful” (Communications of the ACM 11:147–148, 1968). This research was supported by National Institute of Allergy and Infectious Diseases (NIAID) grants R01 AI116770 (YS and EK) and R03 AI124017 (YS and EK) and National Institute of General Medical Sciences (NIGMS) grant U54 GM111274 (EK). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIAID, NIGMS, or the National Institutes of Health.
References

[1] JB Terry, CJ Flatley, DJ van den Berg, GG Morgan, M Trent, JA Turahui, MC Greenwood, PW Corben, and GJ Bell. A field study of household attack rates and the effectiveness of macrolide antibiotics in reducing household transmission of pertussis. *Communicable Diseases Intelligence Quarterly Report*, 39(1):E27–33, 2014.

[2] M Elizabeth Halloran, Marie Pierre Préziosi, and Haitao Chu. Estimating vaccine efficacy from secondary attack rates. *Journal of the American Statistical Association*, 98(461):38–46, 2003.

[3] Rachel Savage, Michael Whelan, Ian Johnson, Elizabeth Rea, Marie LaFreniere, Laura C Rosella, Freda Lam, Tina Badiani, Anne-Luise Winter, Deborah J Carr, et al. Assessing secondary attack rates among household contacts at the beginning of the influenza A (H1N1) pandemic in Ontario, Canada, April-June 2009: A prospective, observational study. *BMC Public Health*, 11(1):234, 2011.

[4] Dale Carcione, CM Giele, LS Goggin, K SH Kwan, DW Smith, GK Dowse, DB Mak, and P Effler. Secondary attack rate of pandemic influenza A (H1N1) 2009 in Western Australian households, 29 may-7 august 2009. *Euro Surveill*, 16(3):19765, 2011.

[5] Oliver W Morgan, Sharyn Parks, Trudi Shim, Patricia A Blevins, Pauline M Lucas, Roger Sanchez, Nancy Walea, Fleetwood Loustalot, Mark R Duffy, Matthew J Shim, et al. Household transmission of pandemic (H1N1) 2009, San Antonio, Texas, USA, April-May 2009. *Emerg Infect Dis*, 16(4):631–637, 2010.

[6] Anne Marie France, Michael Jackson, Stephanie Schrag, Michael Lynch, Christopher Zimmerman, Matthew Biggerstaff, and James Hadler. Household transmission of 2009 influenza A (H1N1) virus after a school-based outbreak in New York City, April-May 2009. *Journal of Infectious Diseases*, 201(7):984–992, 2010.
[7] Simon Cauchemez, Christl A Donnelly, Carrie Reed, Azra C Ghani, Christophe Fraser, Charlotte K Kent, Lyn Finelli, and Neil M Ferguson. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *New England Journal of Medicine*, 361(27):2619–2627, 2009.

[8] Philippe De Wals, L Hertoghe, I Borlee-Grimee, S De Maeyer-Cleempoel, G Reginster-Haneuse, A Dachy, André Bouckaert, and MF Lechat. Meningococcal disease in Belgium. Secondary attack rate among household, day-care nursery and pre-elementary school contacts. *Journal of Infection*, 3:53–61, 1981.

[9] Niels G Becker. *Analysis of infectious disease data*, volume 33. CRC Press, 1989.

[10] Alvin H. Rampey, Jr, Ira M. Longini, Jr, Michael Haber, and Arnold S. Monto. A discrete-time model for the statistical analysis of infectious disease incidence data. *Biometrics*, 48:117–128, 1992.

[11] Eben Kenah. Contact intervals, survival analysis of epidemic data, and estimation of $R_0$. *Biostatistics*, 12:548–566, 2011.

[12] Eben Kenah. Nonparametric survival analysis of epidemic data. *Journal of the Royal Statistical Society, Series B*, 75:277–303, 2013.

[13] Béla Bollobás. *Random graphs*. Springer, 1998.

[14] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[15] A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.

[16] Nathaniel Graham, Mahmood Arai, and Bjørn Hagstrømer. *multiwayvcov: Multi-Way Standard Error Clustering*, 2016. R package version 1.2.3.
[17] Yang Yang, Jonathan Sugimoto, M. Elizabeth Halloran, Nicole E. Basta, Dennis L. Chao, Laura Matrajt, Gail Potter, Eben Kenah, and Ira M. Longini, Jr. The transmissibility and control of pandemic influenza A(H1N1) virus. *Science*, 326:729–733, 2009.

[18] Eben Kenah. Non-parametric survival analysis of infectious disease data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):277–303, 2013.

[19] Eben Kenah. Semiparametric relative-risk regression for infectious disease transmission data. *Journal of the American Statistical Association*, 110:313–325, 2015.

[20] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22, 1986.

[21] Yang Yang, Jr. Longini, Ira M., and M. Elizabeth Halloran. A resampling-based test to detect person-to-person transmission of infectious diseases. *Annals of Applied Statistics*, 1:211–228, 2007.

[22] Yang Yang, Jr. Longini, Ira M., M. Elizabeth Halloran, and V Obenchain. A hybrid EM and Monte Carlo EM algorithm and its application to analysis of transmission of infectious diseases. *Biometrics*, 68:1238–1249, 2012.

[23] Rojas D. P., Dean N. E., Yang Y., Kenah E., Quintero J., Tomasi S., Ramirez E. L., Kelly Y., Castro C., Carrasquilla G., Halloran M. E., and Longini I. M. The epidemiology and transmissibility of Zika virus in Girardot and San Andres Island, Colombia, September 2015 to January 2016. *Eurosurveillance*, 21:30283, 2016.

[24] Fang L-Q, Yang Y, Jiang J-F, Yao H-W, Kargbo D, Li X-L, Jiang B-G, Kargbo B, Tong Y-G, Wang Y-W, Liu K, Kamara A, Dafae F, Kamu A, Jiang R-R, Sun Y, Sun R-X, Chen W-J, Ma, M-J, Dean, NE, Thomas H, Longini IM, Halloran ME, and Cao W-C. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. *PNAS*, 113:4488–4493, 2016.
### Table 1: Number of paths from the index case to a susceptible.

| $m$ | 2 | 3 | 4 | 5 |
|-----|---|---|---|---|
| 4   | 3 | 6 | 6 | 0 |
| 6   | 5 | 20| 60| 120|
| 10  | 9 | 72| 504| 3024|

### Table 2: Estimates of the household SAR with 95% confidence limits. The pairwise survival and chain binomial analyses assume a 2-day incubation period and 6-day infectious period.

| Model                          | $\hat{p}$ | Lower limit | Upper limit |
|--------------------------------|------------|-------------|-------------|
| Binomial (unadjusted)          | 0.148      | 0.106       | 0.197       |
| Binomial (multiway cluster adjusted) | 0.148      | 0.087       | 0.239       |
| Binomial (GEE)                 | 0.148      | 0.088       | 0.238       |
| Pairwise survival (exponential) | 0.066      | 0.062       | 0.071       |
| Pairwise survival (Weibull)    | 0.067      | 0.048       | 0.114       |
| Pairwise survival (log-logistic) | 0.069      | 0.049       | 0.119       |
| Chain binomial                 | 0.067      | 0.046       | 0.094       |
Figure 1: The household FAR as a function of the SAR for households with different numbers of susceptibles ($m$) from analytical calculations (lines) and simulations (symbols). We assume a single index case, so the total household size is $m + 1$. When there is more than one susceptible, the FAR is always greater than the SAR.
Figure 2: Coverage probabilities of binomial 95% confidence intervals for the household SAR with different numbers of susceptibles \((m)\). Gray lines are coverage probabilities for unadjusted confidence intervals, and black lines are coverage probabilities for cluster-adjusted confidence intervals. Both unadjusted and adjusted confidence intervals have low coverage.
Figure 3: Coverage probabilities of binomial 95\% confidence intervals for the household FAR with different numbers of susceptibles \( (m) \). Gray lines are coverage probabilities for unadjusted confidence intervals, and black lines are coverage probabilities for cluster-adjusted confidence intervals. Coverage of the HAR is much higher than coverage of the SAR, but unadjusted confidence intervals still have low coverage. Confidence intervals adjusted for within-household clustering have coverage probabilities close to 95\%.
Figure 4: Simulated numbers of infections in the LA households based on SAR estimates from different models. The vertical black lines indicate the 27 observed infections. The binomial estimate produces far more infections than observed, but the other household SAR estimates fit the observed data well.