Biometrics 000, 000–000 DOI: 000
000 0000

# Pairwise Accelerated Failure Time Models with External Sources of Infection and Epidemiologic Studies of Infectious Disease Transmission

**Yushuf Sharker**

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.

*email:* yushuf.sharker@yale.edu

**and**

**Eben Kenah**

Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, Ohio, U.S.A.

*email:* kenah.1@osu.edu

Summary: Pairwise survival analysis handles dependent happenings in infectious disease transmission data by analyzing failure times in ordered pairs of individuals. The contact interval in the pair $ij$ is the time from the onset of infectiousness in $i$ to infectious contact from $i$ to $j$, where an infectious contact is sufficient to infect $j$ if he or she is susceptible. The contact interval distribution determines transmission probabilities and the infectiousness profile of infected individuals. Many important questions in infectious disease epidemiology involve the effects of covariates (e.g., age or vaccination status) on transmission. These effects on infectiousness and susceptibility can be parametrized via the rate parameter or hazard function of the contact interval distribution. When infectious disease transmission is studied in households or other groups of close contacts, there is usually a risk of infection from outside the group. This is a competing risk that can contain information about the covariate effects on susceptibility. Here, we generalize extend parametric pairwise survival analysis in two ways to allow estimation of these effects: First, we introduce an accelerated failure time model that allows the contact interval rate parameter to depend on infectiousness covariates for $i$, susceptibility covariates for $j$, and pairwise covariates. Second, we show how internal infections (caused by individuals under observation) and external infections (caused environmental or community sources) can be handled simultaneously. In simulations, we show that these methods produce valid point and interval estimates. We also find that accounting for external infection and valid epidemiologic study design are critical to consistent estimation. Finally, we use these methods to analyze household surveillance data from Los Angeles County during the 2009 influenza A(H1N1) pandemic.

Key words: Accelerated failure time models; Epidemiology; Infectious diseases; Regression; Survival analysis

This paper has been submitted for consideration for publication in *Biometrics*

## 1. Introduction

Many of the most important questions in infectious disease epidemiology involve the effects of covariates on the risk of transmission. For transmission from an infectious individual $i$ to a susceptible individual $j$, there are three possible types of covariates: Covariates for $i$ could affect his or her infectiousness, covariates for $j$ could affect his or her susceptibility, and pairwise covariates (e.g., membership in the same household) could affect the risk of transmission independently of the infectiousness of $i$ or susceptibility of $j$. Estimation of these effects can inform the design of public health responses to emerging infections and be used to evaluate vaccine efficacy (Halloran et al., 1997, 2010).

The analysis of infectious disease transmission data is complicated by the dependencies among infection outcomes in different individuals (Halloran and Struchiner, 1991). There are two critical elements to this exposure When exposure to infection differs systematically in individuals with different covariate values, estimates of covariate effects on susceptibility that do not account for this dependence can be biased even in randomized trials (Halloran and Struchiner, 1991; Halloran et al., 2010; Eck et al., 2018; Morozova et al., 2018).

In pairwise survival analysis, these dependencies are handled by analyzing failure times in ordered pairs of individuals rather than individuals (Kenah et al., 2008; Kenah, 2011). In the ordered pair $ij$, the *contact interval* is the time from the onset of infectiousness in $i$ to infectious contact from $i$ to $j$, where an infectious contact is defined to be a contact sufficient to infect $j$ if he or she is susceptible. The survival function of the contact interval distribution can be used to calculate the probability of transmission from $i$ to $j$, and its hazard function shows how the infectiousness of $i$ changes during his or her infectious period.

The contact interval from $i$ to $j$ is right-censored if any of the following occur prior to infectious contact from $i$ to $j$: $i$ recovers from infectiousness, $j$ is infected from a source other than $i$, or observation of the pair $ij$ ends. When who-infects-whom is observed, standard para-

metric and nonparametric methods from survival analysis can be used to estimate the contact interval distribution. When who-infects-whom is not observed, parametric likelihoods can be integrated over all possible transmission trees (Kenah, 2011) or nonparametric estimates from all possible transmission trees can be averaged using an expectation-maximization (EM) algorithm (Kenah, 2013). These methods assume that the contact interval distribution is the same in all infectious-susceptible pairs at risk of transmission.

To allow estimation of covariate effects on the hazard of transmission, Kenah (2015) developed a semiparametric regression model in which the hazard of infectious contact from $i$ to $j$ was

$$h_{ij}(\tau) = e^{\beta^\top X_{ij}} h_0(\tau) \tag{1}$$

where $\beta$ is a coefficient vector, $h_0(\tau)$ is an unspecified baseline hazard for the contact interval, and $X_{ij}$ is a vector that can include susceptibility covariates for $i$, infectiousness covariates for $j$, and pairwise covariates. This allows point and interval estimation of hazard ratios associated with covariates, but it still assumes that all transmission occurred between individuals under observation. The inability to account for external sources of infection is a limitation that must be addressed before pairwise survival analysis can become a practical tool for infectious disease epidemiology.

In this paper, we develop a pairwise accelerated failure time (AFT) model that allows the rate parameter of the contact interval distribution to depend on covariates while accounting for the risk of infection from external sources. Coefficient estimates from this model are consistent and asymptotically normal, which we show both analytically and in a large simulation study of various epidemiologic study designs. We apply the model to household surveillance data collected by the Los Angeles County Department of Public Health during the 2009 influenza A(H1N1) pandemic, and we find that accounting for external infection greatly improves statistical power. The pairwise AFT model has the potential to become an

important new tool in the design and analysis of vaccine trials, outbreak investigations, and other studies of infectious disease transmission.

### 1.1 *Stochastic S(E)IR models*

At any time, each individual $i \in \{1, \ldots, n\}$ is in one of four states: susceptible (S), exposed (E), infectious (I), or removed (R). Person $i$ moves from S to E at his or her *infection time* $t_i$, with $t_i = \infty$ if $i$ is never infected. After infection, $i$ has a *latent period* of length $\varepsilon_i$ during which he or she is infected but not infectious. At time $t_i + \varepsilon_i$, $i$ moves from E to I, beginning an *infectious period* of length $\iota_i$. At time $t_i + \varepsilon_i + \iota_i$, $i$ moves from I to R, where he or she can no longer infect others or be infected. The latent period $\varepsilon_i$ is a nonnegative random variable, the infectious period $\iota_i$ is a strictly positive random variable, and both have finite mean and variance. The time elapsed since the onset of infectiousness in $i$ at time $t_i + \varepsilon_i$ is the *infectious age* of $i$. An SIR model is an SEIR model with no latent period.

After becoming infectious at time $t_i + \varepsilon_i$, person $i$ makes infectious contact with $j \neq i$ at time $t_{ij} = t_i + \varepsilon_i + \tau^*_{ij}$. The *infectious contact interval* $\tau^*_{ij}$ is a strictly positive random variable with $\tau^*_{ij} = \infty$ if infectious contact never occurs. Because infectious contact can only occur while $i$ is infectious, either $\tau^*_{ij} \in (0, \iota_i]$ or $\tau^*_{ij} = \infty$. Because we define infectious contact to be sufficient to infect a susceptible person, $t_j \leq t_{ij}$ for all $i$ and $j$.

An *internal infection* occurs when an individual is infected by another individual in the observed population. For each ordered pair $ij$, let $C_{ij} = 1$ if infectious contact from $i$ to $j$ is possible and $C_{ij} = 0$ otherwise. We assume the infectious contact interval $\tau^*_{ij}$ is generated as follows: A *contact interval* $\tau_{ij}$ is drawn from a distribution with hazard function $h_{ij}(\tau)$. If $\tau_{ij} \leq \iota_i$ and $C_{ij} = 1$, then $\tau^*_{ij} = \tau_{ij}$. Otherwise, $\tau^*_{ij} = \infty$. The contact interval distribution can be used to calculate transmission probabilities and to calculate infectiousness as a function of infectious age.

An *external infection* occurs when an individual is infected from a source outside the

observed population. Let $C_{0j}$ indicate whether individual $j$ is at risk of external infectious contact. Let the *external infectious contact time* $t^*_{0j}$ denote the first time that an individual $j$ receives infectious contact from outside the observed population, with $t^*_{0j} = \infty$ if this never occurs. We assume that the external infectious contact time is generated as follows: A *external contact time* $t_{0j}$ is drawn from a distribution with hazard function $h_{0j}(t)$. If $C_{0j} = 1$, then $t^*_{0j} = t_{0j}$. Otherwise, $t^*_{0j} = \infty$. The external contact time distribution can be used to calculate the probability of infection from external sources and to calculate the risk of external infection as a function of time.

### 1.2 *Exposure and infectious sets*

For each internal infection $j$, let $v_j$ denote the index of his or her infector. Let $v_j = 0$ if $j$ is an external infection and $v_j = \infty$ if $j$ is not infected. When $v_j$ is observed for all infected $j$, we say that *who-infected-whom* (WIW) is observed. Otherwise, we say that WIW is not observed—even if $v_j$ is observed for a subset of infected $j$.

For each individual $j$, his or her *exposure set* is

$$\mathcal{W}_j = \{i < \infty : (t_i + \varepsilon_i < t_j \text{ or } i = 0) \text{ and } C_{ij} = 1\}, \tag{2}$$

which is the set of all sources of infection to whom $j$ was exposed while susceptible. If $j$ is infected by an unknown infector, his or her *infectious set* is

$$\mathcal{V}_j = \{i < \infty : (t_i + \varepsilon_i < t_j \leq t_i + \varepsilon_i + \iota_i \text{ or } i = 0) \text{ and } C_{ij} = 1\}, \tag{3}$$

which is the set of possible sources that caused his or her infection. If the infector of $j$ is known, then $\mathcal{V}_j = \{v_j\}$. If $j$ is not infected, $\mathcal{V}_j = \varnothing$ (the empty set).

### 1.3 *Infectious disease data*

Our epidemiologic data contain the times of all S $\rightarrow$ E (infection), E $\rightarrow$ I (infectiousness onset), and I $\rightarrow$ R (removal) transitions in the observed population between time 0 and a

time $T$ that is a stopping time with respect to the observed data. For all ordered pairs $ij$ in which $i$ is infected or $i = 0$, we observe $C_{ij}$.

The contact interval $\tau_{ij}$ can be observed only if $j$ is infected by $i$ at time $t_{ij} = t_i + \varepsilon_i + \tau_{ij}$. This can happen only if $C_{ij} = 1$ and the pair $ij$ is at risk of transmission at time $t_{ij}$. Contact intervals can be right-censored by the end of infectiousness in $i$, by the infection of $j$ from a source other than $i$, and by the end of observation. For $i \neq 0$, let $I_i(t) = \mathbb{I}_{t-t_i-\varepsilon_i \in (0,\iota_i]}$ indicate whether $i$ remains infectious at time $t$, and let $I_0(t)$ indicate whether external infectious contact is possible at time $t$. Let $S_j(t) = \mathbb{I}_{t \leq t_j}$ indicate whether $j$ remains susceptible at time $t$, and let $\mathcal{O}(t) = \mathbb{I}_{t \leq T}$ indicate whether observation is ongoing at time $t$. Since $I_i(t)$, $S_j(t)$, and $\mathcal{O}(t)$ are left-continuous,

$$Y_{ij}(t) = C_{ij} I_i(t) S_j(t) \mathcal{O}(t) \tag{4}$$

is a left-continuous process that indicates the risk of an observed infectious contact from $i$ to $j$ at time $t$. The assumptions made in the stochastic S(E)IR model above ensure that censoring of $\tau_{ij}$ and $t_{0j}$ is independent.

## 2. Methods

In our models, any parametric failure time distribution could be used. For simplicity, we focus on the following three, which have simple closed-form survival and hazard functions: the exponential distribution with rate $\lambda$, the Weibull distribution with rate $\lambda$ and shape $\gamma$, and the log-logistic distribution with rate $\lambda$ and shape $\gamma$. The internal and external transmission models can use the same failure time distribution or different distributions. Let the parameters of the internal failure time distribution be $(\lambda_{\text{int}}, \gamma_{\text{int}})$ and the parameters of the external distribution be $(\lambda_{\text{ext}}, \gamma_{\text{ext}})$.

The internal and external transmission models generally work on different time scales. In a pair $ij$ with $i \neq 0$, the time origin is the onset of infectiousness in $i$, which can differ from

pair to pair. A pair $0j$ is at risk of transmission when $j$ is susceptible and external infectious contact is possible. Typically, a common time origin will be specified for all external pairs in a single population under observation.

### 2.1 *Internal and external rate parameters*

When $i \neq 0$, the rate parameter of the contact interval distribution in the pair $ij$ is

$$\lambda_{ij} = e^{\beta_{\text{int}}^{\top} X_{ij}} \lambda_0 \tag{5}$$

where $\beta_{\text{int}}$ is an unknown coefficient vector, $\lambda_0$ is a baseline rate, and $X_{ij}$ is a covariate vector that can include infectiousness covariates for $i$, susceptibility covariates for $j$, and pairwise covariates. This is equivalent to an AFT model where $\exp(-\beta_{\text{int}}^{\text{T}} X_{ij})$ is the acceleration factor (Kalbfleisch and Prentice, 2002).

The rate parameter for the external contact time for individual $j$ is

$$\lambda_{0j} = e^{\beta_{\text{ext}}^{\top} X_{0j}} \mu_0 \tag{6}$$

where $\beta_{\text{ext}}$ is an unknown coefficient vector, and $\mu_0$ is the baseline external rate parameter, and $X_{0j}$ is a covariate vector that can include susceptibility covariates for $j$ and environmental or community covariates.

There may be overlap between the internal coefficient vector $\beta_{\text{int}}$ and the external coefficient vector $\beta_{\text{ext}}$. For example, vaccine status could affect the rate parameters for both models. To handle this, we parameterize the combined model as

$$\lambda_{ij} = e^{\beta^{\top} X_{ij}} \lambda_0^{1-\mathbb{I}_{i=0}} \mu_0^{\mathbb{I}_{i=0}} \tag{7}$$

where the coefficient vector $\beta$ includes coefficients unique to the internal model, coefficients unique to the external model, and shared coefficients. The components of $X_{ij}$ used only in the internal model are set to zero when $i = 0$, and the components of $X_{ij}$ used only in the external model are set to zero when $i \neq 0$. The distinction between internal and external rows in the data set is maintained using an *external pair indicator* $\zeta$, which equals one when $i = 0$

and zero otherwise. If a covariate in $X_{ij}$ is shared by the internal and external transmission models, it can be allowed to have different coefficients in the two models by including an interaction term with $\zeta$. We call these *external interaction terms*.

## 2.2 *Maximum likelihood estimation*

The likelihood and its score process can be derived in a manner similar to that of Kenah (2011). Let $\theta$ be a coefficient vector containing the log rate ratios $\beta$, the log baseline rate parameters $\ln \lambda_0$ and $\ln \mu_0$, and the log shape parameters $\ln \gamma_{\text{int}}$ and $\ln \gamma_{\text{ext}}$ as needed. Let $h_{ij}(t, \theta)$ and $S_{ij}(t, \theta)$ be the hazard and survival functions with rate $\lambda_{ij}$ from equation (7). Let $\theta_0$ denote the true value of $\theta$.

*Who-infects-whom observed.* Let $\mathcal{N}_{ij}(t) = \mathbb{I}_{t \geq t_{ij}}$ count the first infectious contact from $i$ to $j$. Assume $j$ is susceptible at time $t = 0$, so $\mathcal{N}_{ij}(0) = 0$. Then $\mathcal{M}_{ij}(t, \theta_0)$ is a mean-zero martingale, where

$$\mathcal{M}_{ij}(t, \theta) = \mathcal{N}_{ij}(t) - \int_0^t h_{ij}(u - t_i - \varepsilon_i, \theta) C_{ij} I_i(u) \, \mathrm{d}u \tag{8}$$

for $i \neq 0$ and

$$\mathcal{M}_{0j}(t, \theta) = \mathcal{N}_{0j}(t) - \int_0^t h(u, \theta) C_{ij} I_i(u) \, \mathrm{d}u. \tag{9}$$

We observe infectious contacts from $i$ to $j$ only while $j$ is still susceptible and $ij$ is under observation, which gives us the observed counting process

$$N_{ij}(t) = \int_0^t Y_{ij}(u) \, \mathrm{d}\mathcal{N}_{ij}(u). \tag{10}$$

Similarly, let

$$M_{ij}(t, \theta) = \int_0^t Y_{ij}(u) \, \mathrm{d}\mathcal{M}_{ij}(u, \theta). \tag{11}$$

Then $M_{ij}(t, \theta_0)$ is a mean-zero martingale because it is the integral of a predictable process with respect to $\mathcal{M}_{ij}(u, \theta_0)$.

When we observe infectious contacts from $i$ to $j$ between time 0 and time $T$, we get the

log likelihood

$$\ell^*_{ij}(\theta) = \int_0^T \ln h(u - t_i - \varepsilon_i, \theta)\, \mathrm{d}N_{ij}(u) - \int_0^T h(u - t_i - \varepsilon_i, \theta) Y_{ij}(u)\, \mathrm{d}u. \tag{12}$$

when $i \neq 0$ and

$$\ell^*_{0j}(\theta) = \int_0^T \ln h(u, \theta)\, \mathrm{d}N_{0j}(u) - \int_0^T h(u, \theta) Y_{0j}(u)\, \mathrm{d}u. \tag{13}$$

These are standard survival likelihoods: The first term is a log hazard if $i$ infects $j$ and zero otherwise, and the second term is the negative cumulative hazard of infectious contact. The score process is

$$U^*_{ij}(t, \theta) = \int_0^t \left( \frac{\partial}{\partial \theta} \ln h(u - t_i - \varepsilon_i, \theta) \right) \mathrm{d}M_{ij}(u, \theta), \tag{14}$$

when $i \neq 0$ and

$$U^*_{0j}(t, \theta) = \int_0^t \left( \frac{\partial}{\partial \theta} \ln h(u, \theta) \right) \mathrm{d}M_{ij}(u, \theta). \tag{15}$$

Both score processes are mean-zero martingales when $\theta = \theta_0$.

Now fix $j$. If we observe all pairs $ij$ from time 0 until time $T$, the log likelihood is

$$\ell^*_{\cdot j}(\theta) = \sum_{i \neq j} \ell^*_{ij}(\theta) \tag{16}$$

with score process

$$U^*_{\cdot j}(t, \theta) = \sum_{i \neq j} U^*_{ij}(t, \theta). \tag{17}$$

The score process $U^*_{\cdot j}(t, \theta_0)$ is a mean-zero martingale because it is a sum of independent mean-zero martingales.

When we observe who-infected-whom, the log likelihood is $\ell^*(\theta) = \sum_{j=1}^n \ell^*_{\cdot j}(\theta)$ and its score process is $U^*(t, \theta) = \sum_{j=1}^n U^*_{\cdot j}(t, \theta)$. Because it is a sum of independent mean-zero martingales, $U^*(t, \theta_0)$ is a mean-zero martingale. Differentiating $\ell^*(\theta)$, evaluating at $\theta_0$, and taking expectations yields

$$E\Big[ -\frac{\partial^2}{\partial \theta^2} \ell^*(\theta_0) \Big] = E\big[ \langle U^*(\theta_0) \rangle(T) \big], \tag{18}$$

where $\langle U(\theta_0) \rangle(\tau)$ is the predictable variation process of $U(\tau, \theta_0)$.

*Who-infects-whom not observed.* When who-infects-whom is not observed, we cannot see each $N_{ij}(t)$. Instead, we see $N_{\cdot j}(t) = \sum_{i \neq j} N_{ij}(t)$. The total hazard of infectious contact with $j$ at time $t$ is

$$h_{\cdot j}(t, \theta) = h(t, \theta) C_{0j} + \sum_{i:0 \neq i \neq j} h(t - t_i - \varepsilon_i, \theta) C_{ij} I_i(t), \tag{19}$$

so the process

$$M_{\cdot j}(t, \theta) = N_{\cdot j}(t) - \int_0^t h_{\cdot j}(u, \theta) S_j(u) O(t) \, \mathrm{d}u = \sum_{i \neq j} M_{ij}(t, \theta). \tag{20}$$

is a mean-zero martingale when $\theta = \theta_0$. When $j$ is observed from time 0 to time $T$, the log likelihood is

$$\ell_{\cdot j}(\theta) = \int_0^T \ln h_{\cdot j}(u, \theta) \, \mathrm{d}N_{\cdot j}(u) - \int_0^T h_{\cdot j}(u, \theta) S_j(u) \, \mathrm{d}u \tag{21}$$

and its score process is

$$U_{\cdot j}(t, \theta) = \int_0^t \left[ \frac{\partial}{\partial \theta} \ln h_{\cdot j}(u, \theta) \right] \mathrm{d}M_{\cdot j}(u, \theta), \tag{22}$$

which is a mean-zero martingale when $\theta = \theta_0$.

The complete-data log likelihood when we do not observe who-infected-whom is $\ell(\theta) = \sum_{j=1}^n \ell_{\cdot j}(\theta)$ with score process $U(t, \theta) = \sum_{j=1}^n U_{\cdot j}(t, \theta)$. Because it is a sum of independent mean-zero martingales, $U(t, \theta_0)$ is a mean-zero martingale. Differentiating $\ell(\theta)$, evaluating at $\theta_0$, and taking expectations yields

$$E\Big[ - \frac{\partial^2}{\partial \theta^2} \ell(\theta_0) \Big] = E\big[ \langle U(\theta_0) \rangle(T) \big], \tag{23}$$

where $\langle U(\theta) \rangle(\tau)$ is the predictable variation process of $U(\tau, \theta)$.

### 2.3 *Pairwise asymptotics*

The arguments above establish the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\theta}$ as the number of observed infections $m \to \infty$ (Kenah, 2011) as long as the rate of increase in the number of susceptibles at risk of infection is at least as fast as the rate of increase in the number of pairs at risk of transmission (Kenah, 2015). In practice,

maximum likelihood estimation should work as long as we observe a sufficient number of infections in susceptible individuals.

## 3. Simulations

The proposed pairwise AFT regression model was tested through 10,000 network-based simulations for each of five different baseline internal contact interval distributions: exponential(1), Weibull($\gamma = 0.5$, $\lambda = 1$), Weibull(1.5, 1), log-logistic(0.5, 1), and log-logistic(1.5, 1). In all simulations, the external infectious contact time distribution was exponential(1). The infectious period was fixed to one (i.e., $\iota_i = 1$ for each $i$).

In each simulation, we generated an undirected network representing 300 households of size 6. Each household was a complete graph of size 6, and the households were not connected to each other. Once a household member was infected, other members of the household could be infected by transmission within the household or by an external source. Each epidemic was followed until 500 infections occurred, which guaranteed at least 200 infections in individuals who were not index cases (the first case detected in a household).

Each individual $i$ was assigned an independent Bernoulli(0.5) covariate $X_i$. The rate parameter for the contact interval distribution in the pair $ij$ was

$$\lambda_{ij} = \exp\left(\beta_{\text{inf}} X_i + \beta_{\text{sus}} X_j + \mathbb{I}_{i\neq 0} \ln \lambda_0 + \mathbb{I}_{i=0} \ln \mu_0\right), \tag{24}$$

where we set $X_0 = 0$. For each simulation, the true values of $\beta_{\text{inf}}$ and $\beta_{\text{sus}}$ were independent samples from a uniform$(-1, 1)$ distribution.

In each simulation, we analyzed data sets under four different epidemiologic study designs. Analysis of within-household transmission is the same for all four study designs, but they differ in their inclusion of individuals (i.e., pairs $0j$) at risk of external infectious contact. The study designs are:

**Complete cohort:** Follow-up for all 1,800 individuals starts at time zero, which is the time origin for external infectious contact times.

**Contact tracing (CT) with delayed entry:** Follow-up of each individual begins at the infection time of the index case in his or her household. Time at risk of external infectious contact prior the start of follow-up is left-truncated, and individuals in households with no infections are excluded from the study.

**CT without delayed entry:** Follow-up of all members of households where at least one infection occurs starts retroactively at time zero. Individuals in households with no infections are excluded from the study.

**Ignoring external infection:** All pairs $0j$ are excluded from the study. This is equivalent to assuming that, in each household, all infections after the primary case are caused by within-household transmission.

We expect the first two study designs (the "valid" study designs) to yield valid parameter estimates and the last two (the "flawed" study designs) to yield biased parameter estimates. Under each study design, data were analyzed both with and without knowledge of who-infected-whom. In all eight analyses of each simulation, we obtained maximum likelihood point estimates of $\beta_{\text{inf}}$, $\beta_{\text{sus}}$, $\ln \lambda_0$, $\ln \gamma_{\text{int}}$, and $\ln \mu_0$. For all parameters, we calculated 95% Wald and likelihood ratio (LR) confidence intervals. All regression models used an exponential distribution for external rows and the correct parametric family (exponential, Weibull, or log-logistic) for internal rows.

Simulations were implemented in Python version 3.5.1 (`python.org`) using NumPy version 1.11 (`numpy.org`), NetworkX version 1.11 (`networkx.github.io`), and the module `transtat_models` version 0.1.1 (available at `github.com/ekenah/transtat_models`). Statistical analysis was conducted in `R` version 3.3 (`r-project.org`) using `transtat` version 0.2.6 (available at `github.com/ekenah/transtat`). The `transtat` package allows pairwise

AFT models to be specified using standard `R` model syntax and provides Wald and likelihood ratio confidence intervals for coefficient estimates. All of these software packages are free and open-source. Simulation code in Python code and `R` code for the analysis of each simulation is available in the Supplementary Material. Updated versions of these packages are maintained at `github.com/ekenah`.

## 4. Los Angeles County influenza data

We use influenza A (H1N1) household surveillance data collected by the Los Angeles County Department of Public Health (LACDPH) in April and May, 2009 to give a practical example of pairwise AFT modeling of infectious disease transmission data. The data was collected using the following protocol (Sugimoto et al., 2011):

(1) Between April 14 and May 18, nasopharyngeal swabs and aspirates were taken from individuals who reported to the LACDPH or other local health care providers with acute febrile respiratory illness (AFRI), defined as a fever $\geq 37.8$°C plus at least one of cough, sore throat, or rhinorrhea (runny nose). These specimens were tested for influenza using reverse transcriptase polymerase chain reaction (RT-PCR).

(2) Patients whose specimens tested positive for pandemic influenza A(H1N1) or for influenza A of undetermined subtype were invited to participate in a phone interview. These interviews used a standard questionnaire developed by the LACDPH to to collect information about his or her household contacts, including sex, age, and antiviral prophylaxis use. For index cases under 18 years of age, an adult proxy was interviewed.

(3) The initial interview and, when necessary, a follow-up interview were used to obtain the symptom onset dates of AFRI episodes in the household up to 14 days after the symptom onset date of the index case. All interviews were completed between April 30 and June 1.

For simplicity, we assume all AFRI episodes among household members were caused by

influenza A(H1N1). All index cases are assumed to be external infections, and all other household members are assumed to be susceptible to infection from both within-household transmission and external sources. The study design is contact tracing with delayed entry.

The primary analysis assumed an incubation period of 2 days, a latent period of 0 days, and an infectious period of 6 days. These natural history assumptions are adapted from Yang et al. (2009). Households were identified upon clinical presentation of an index case, so household members were considered to be at risk of infection from the infection time of the index case (which depends on the assumed incubation period) until 14 days after the infection time of the index case. In a sensitivity analysis, we varied assumptions about the latent and infectious periods.

The covariates used in our analysis were sex (`male` = 1 for males and `male` = 0 for females), age category (`adult` = 1 for ages $\geq$ 18 years and `adult` = 0 otherwise), and antiviral prophylaxis. Antiviral prophylaxis was assumed to be initiated on the day following the symptom onset of the index case in each household, so it was handled as a time-dependent covariate. Each pair had covariate values for the infectious individual (`male_inf`, `adult_inf`, and `proph_inf`) and for the susceptible individual (`male_sus`, `adult_sus`, and `proph_sus`). In external pairs, all infectiousness covariates were set to zero. We considered exponential, Weibull, and log-logistic distributions for the internal and external contact time distributions. All models were fit using the Broyden, Fletch, Goldfarb, and Shanno method (`BFGS` in the `R` function `optim`) with starting parameter values taken from an initial fit using exponential contact intervals and external contact times.

Statistical analysis was conducted in `R` version 3.3 (`www.r-project.org`) using `transtat` version 0.2.6. The data set and analysis code are available in the Supplementary Material.

## 5. Results

### 5.1 *Household simulations*

For all parameters, Table 1 shows the mean squared error (MSE) for point estimates and Table 2 shows the coverage probabilities for Wald and likelihood ratio (LR) 95% confidence intervals. In all cases, results are shown for all study designs with and without observation of who-infected-whom (WIW).

[Table 1 about here.]

[Table 2 about here.]

*Log rate ratios.* Figure 1 shows scatterplots of the true and estimated $\beta_{\text{inf}}$, and Figure 2 shows scatterplots of the true and estimated $\beta_{\text{sus}}$. When WIW is observed, $\beta_{\text{inf}}$ and $\beta_{\text{sus}}$ are estimated with no apparent bias under all four study designs. All four designs have similarly low MSE, and both Wald and LR 95% confidence interval coverage probabilities are nominal. When WIW is not observed, both invalid study designs produce questionable point and interval estimates. CT without delayed entry shows high MSE for $\beta_{\text{inf}}$ and slightly low coverage probabilities for $\beta_{\text{inf}}$ and $\beta_{\text{sus}}$. Ignoring external infection produces slightly high MSE for $\beta_{\text{sus}}$ and very low coverage probabilities for $\beta_{\text{sus}}$ and $\beta_{\text{inf}}$. Point and interval estimates for the valid study designs behave well whether or not WIW is observed.

[Figure 1 about here.]

[Figure 2 about here.]

*Intercepts.* When WIW is observed, all four study designs show little bias in estimates of the internal model intercept $\ln \lambda_0$. Both flawed study designs have slightly higher MSEs than the valid study designs. Confidence interval coverage probabilities are low for CT without delayed entry but acceptable for all other study designs. When WIW is not observed, there is a clear negative bias under CT without delayed entry and a positive bias under ignoring

external infection. The flawed study designs have substantially higher MSEs than the valid study designs, and their confidence interval coverage probabilities are very low. The valid study designs have Wald coverage probabilities that are slightly too low, but their LR coverage probabilities are nominal.

[Figure 3 about here.]

When WIW is observed, estimates of the external model intercept $\ln \mu_0$ under CT without delayed entry have a large positive bias and high MSE. Estimates under the valid study designs are unbiased and have low MSE. When WIW is not observed, estimates under the complete cohort design are unbiased and have low MSE. Estimates under CT with delayed entry are unbiased but have high MSE due to large variance. Estimates under CT without delayed entry have a large positive bias but a much lower variance, so they end up with a lower MSE than estimates under CT without delayed entry. Whether or not WIW is observed, Wald and LR coverage probabilities are extremely low under CT without delayed entry but nominal under the valid study designs.

[Figure 4 about here.]

*Shape parameter.* The internal log shape parameter, $\ln \gamma_{\text{int}}$, was estimated in all simulations with Weibull or log-logistic contact interval distributions. When WIW is observed, all four study designs show low MSE. Confidence interval coverage probabilities are nominal under the valid study designs and slightly low under the flawed study designs. When WIW is not observed, the valid study designs still show low MSE and acceptable (though slightly lower) coverage probabilities. The flawed study designs show much larger MSE and coverage probabilities that are too low—especially when external infection is ignored.

*Summary.* When WIW was observed, estimation of $\beta_{\text{inf}}$, $\beta_{\text{sus}}$, and $\ln \gamma_{\text{int}}$ was surprisingly robust. Estimation of $\ln \lambda_0$ was acceptable under all study designs except ignoring external

infection, which overestimates $\ln \lambda_0$ because all infections are attributed to transmission within the household. Estimation of $\ln \mu_0$ was acceptable only under the valid study designs.

When WIW was not observed, estimation of all parameters became much more sensitive to epidemiologic study design. The valid designs produced unbiased point estimates and confidence intervals with adequate coverage probabilities, but the flawed study designs did not. LR confidence intervals performed slightly better than Wald confidence intervals.

The Appendix, which is available in the Supplementary Material, shows these tables and figures separately for each baseline contact interval distribution, and it also includes boxplots of $\ln \gamma_{\text{int}}$ estimates for Weibull and log-logistic contact interval distributions. The same general pattern was seen in each subset of simulations.

### 5.2 *Los Angeles County influenza data*

The household data collected by the Los Angeles County Department of Public Health included 299 individuals in 58 households. There were 99 probable influenza infections, of which 62 were index cases—four households had co-primary cases with symptom onsets on the same day. There were three people missing data on sex, four people missing data on age, and 56 people missing data on antiviral prophylaxis. The 62 individuals with missing data came from 17 households with 36 infections, of which 19 were index cases. Because we assume all household members can infect or be infected by other household members, we excluded the entire household if any of its members was missing data. In the complete-cases data set, we have 41 households with 63 infections, of which 43 were index cases.

Using the complete-cases data set, we fit a model with all six covariates using all nine possible combinations of internal and external contact time distributions. The top panel of Table 3 shows the resulting Akaike Information Criterion (AIC) values. The three minimum AIC values occur for exponential internal contact intervals. Among these three, the lowest AIC occurs for log-logistic external contact times. Using exponential interval contact in-

tervals and log-logistic external contact times, we built a model using backwards selection to achieve the minimum AIC. This removed all covariates except for three: age category for infectiousness (`adult_inf`), age category for infectiousness (`adult_sus`), and prophylaxis use by susceptibles (`proph_sus`). The AIC of this model was 203.59.

[Table 3 about here.]

We then checked for external interaction terms, which allow a covariate to have different coefficients in the internal and external transmission models. An external interaction term with `adult_sus` had a p-value of 0.89 and increased the AIC to 205.57. An external interaction term with `proph_sus` had a p-value of 0.87 but reduced the AIC to 202.37. A likelihood ratio test comparing this model to a model with no main effect or interaction for `proph_sus` yielded a p-value of 0.041. Though there is some indication that the coefficient on `proph_sus` might differ in the internal and external transmission models, we did not keep the interaction term in the model. Our final model is summarized at the top of Table 4.

The predicted infectiousness rate ratio for adults compared to children is 3.88 $(0.70, 90.32)$, and the predicted susceptibility rate ratio for adults compared to children is 0.62 $(0.23, 1.20)$. The model suggests that adults are more infectious and less susceptible than children, but the small number of transmission events observed makes these results inconclusive. The predicted rate ratio for susceptibility associated with antiviral prophylaxis is 0.34 $(0.10, 0.76)$, so the model strongly suggests that antiviral prophylaxis in susceptibles reduces their risk of infection. We found no clear evidence of differences in infectiousness or susceptibility by sex, and we found no clear evidence of an effect of antiviral prophylaxis on infectiousness.

[Table 4 about here.]

Table 5 shows the predicted household secondary attack rate (SAR) by the age of the infectious individual, the age of the susceptible individual, and antiviral prophylaxis in the susceptible individual. The higher infectiousness and lower susceptibility of adults is readily

apparent, as is the protective effect of antiviral prophylaxis. Because the predicted household SAR depends on multiple parameters in the regression model, we used Wald confidence intervals for simplicity.

[Table 5 about here.]

To see how accounting for external sources of infection affected our analysis, we re-fit our final model using only data on infectious-susceptible pairs within households. This model is summarized at the bottom of Table 4. The two models give similar results, but accounting for external infection gave us greater statistical power to estimate the effects of age and antiviral prophylaxis. Using Weibull or log-logistic internal contact interval distributions did not restore the statistical power lost by ignoring external sources of infection (not shown). Table 6 shows the results of a sensitivity analysis where we varied assumptions about the latent and infectious periods. The infectiousness rate ratio for age category and its p-value are highly sensitive to the assumed latent and infectious periods. The susceptibility rate ratio for age category and its p-value are somewhat more stable. The susceptibility rate ratio for antiviral prophylaxis and its p-value are remarkably stable. The rate ratio varies from 0.35 to 0.41, and its p-value varies from 0.011 to 0.026. The loss of statistical power when we fail to account for external sources of infection is consistent throughout the sensitivity analysis.

[Table 6 about here.]

## 6. Discussion

The results of the simulation study show that the pairwise AFT model produces valid point and interval estimates when the epidemiologic study design is valid and the statistical model is correctly specified. When external sources of infection are ignored and WIW is not observed, $\ln \lambda_0$ is overestimated because all infections are attributed to within-household transmission. Under CT without delayed entry, $\ln \mu_0$ is overestimated whether or not WIW is

observed because we include all infections of index cases but only a subset of the person-time at risk that gave rise to these cases. The overestimation of $\ln \mu_0$ leads to underestimation of $\ln \lambda_0$ (or overestimation of both $\ln \lambda_0$ and $\ln \gamma_{\text{int}}$).

The high variance of $\ln \mu_0$ estimates under CT with delayed entry is caused by the relatively small number of infections from external sources that occur in households that already have an index case. If observation of the household starts upon identification of an index case, the infection of the index case cannot be included as an infection event. It might be useful to recruit households without index cases to get better estimates of the risk of infection from external sources. One possibility would be to use a test-negative design to recruit households (or other contact groups), similar to that used to recruit individuals in observational studies of vaccine effectiveness (Foppa et al., 2013; Jackson and Nelson, 2013). Households of possible index cases who test negative for the infection of interest could be retained in a study to help estimate the risk of infection from external sources. This design guarantees that these "control" households would have been included if they had a true index case, reducing the chance of selection bias or confounding by health-seeking behavior.

When WIW is observed, estimation of $\beta_{\text{inf}}$ and susceptibility $\beta_{\text{sus}}$ was surprisingly robust under flawed epidemiologic study designs. Though the corresponding estimates of $\ln \lambda_0$ and $\ln \mu_0$ were not accurate, these are nuisance parameters in some contexts. Thus, information about WIW could be very useful for understanding the effects of covariates on transmission. Pathogen phylogenies have been shown to reduce the variance of transmission parameter estimates by providing partial information about WIW (Kenah et al., 2016), but our simulation results suggest that they could also help reduce bias. If phylogenetics can determine which infections originate from outside a household, it might be possible to avoid modeling the risk of infection from outside the household: Infections from external sources would become censoring events for the internal transmission model.

Our analysis of the LACDPH influenza A(H1N1) data strongly suggests that antiviral prophylaxis reduced susceptibility to infection in household contacts. We found inconclusive evidence of higher infectiousness and lower susceptibility to influenza among adults. Our classification of age was very crude, so it is possible that a more sophisticated model would find more conclusive evidence of age effects on susceptibility and infectiousness. These results were obtained with only 20 infection events in 41 households, and the ability to account for external sources of infection was critical to this efficiency.

The simulations and data analysis also pointed to several limitations of pairwise survival analysis and the pairwise AFT model:

- Pairwise survival analysis will require greater sophistication in handling missing data than we demonstrated here. In our analysis, we removed entire households when any member was missing a covariate. We also assumed a fixed incubation, latent and infectious periods to avoid treating these times as missing. Bayesian methods with data imputation (O'Neill and Roberts, 1999) would be a more principled and efficient way to handle missing data.
- With no scientific basis for choosing parametric families, we simply compared models with different combinations of contact interval and external infectious contact time distributions using the AIC. An extension of the semiparametric model of Kenah (2015) would not require the choice of parametric families for these distributions.
- The pairwise AFT model showed some numerical instability. It was important to identify a good starting point for maximization of the likelihood, either with stochastic annealing (as in the simulations) or with a fit using a simple model (as in the data analysis). We dealt successfully with this problem ad hoc, but it needs further investigation.

The pairwise AFT model can be viewed as an extension of the longitudinal chain-binomial model (Rampey et al., 1992; Becker and Britton, 1999) to continuous time. It comes with several advantages: It allows flexibility in the infectiousness profile without a large number

of nuisance parameters, and it can be specified, fit, and interpreted in a manner similar to standard regression models. Unlike almost all individual-level analyses traditionally used in infectious disease epidemiology, the pairwise AFT model conditions on exposure to infection.

The pairwise AFT model allows the simultaneous estimation of rate ratios for susceptibility and infectiousness while accounting for the risk of external infection and conditioning on exposure to infection. It has the potential to improve the design and analysis of vaccine trials, outbreak investigations, and other studies of infectious disease transmission. We hope its availability in an `R` package will make it accessible to practicing epidemiologists.

## Acknowledgements

The authors would like to thank Forrest Crawford (Yale School of Public Health) and Jonathan Sugimoto (Fred Hutchinson Cancer Research Center) for their comments, and we are grateful to the Los Angeles County Department of Public Health (LACDPH) for allowing the use of their data. EK was supported by National Institute of General Medical Sciences (NIGMS) grant U54 GM111274. EK and YS were supported by National Institute of Allergy and Infectious Diseases Grant (NIAID) grants R01 AI116770 and R03 AI124017. YS was supported by National Institutes of Health (NIH) grant DP2HD09179. The content is solely the responsibility of the authors and does not represent the official views of LACDPH, NIGMS, NIAID, or the NIH.

## References

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society, Series B* **61,** 287–307.

Eck, D. J., Morozova, O., and Crawford, F. W. (2018). Randomization for the direct effect of an infectious disease intervention in a clustered study population. *arXiv preprint arXiv:1808.05593* .

Foppa, I. M., Haber, M., Ferdinands, J. M., and Shay, D. K. (2013). The case test-negative design for studies of the effectiveness of influenza vaccine. *Vaccine* **31,** 3104–3109.

Halloran, M. E., Longini, Jr., I. M., and Struchiner, C. (2010). *Design and Analysis of Vaccine Studies*. Statistics for Biology and Health. Springer-Verlag, New York.

Halloran, M. E. and Struchiner, C. (1991). Study designs for dependent happenings. *Epidemiology* **2,** 331–338.

Halloran, M. E., Struchiner, C., and Longini, Jr, I. M. (1997). Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *American Journal of Epidemiology* **146,** 789–803.

Jackson, M. L. and Nelson, J. C. (2013). The test-negative design for estimating influenza vaccine effectiveness. *Vaccine* **31,** 2165–2168.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, second edition.

Kenah, E. (2011). Contact intervals, survival analysis of epidemic data, and estimation of R0. *Biostatistics* **12,** 548–566.

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

Kenah, E. (2015). Semiparametric relative-risk regression for infectious disease transmission data. *Journal of the American Statistical Association* **110,** 313–325.

Kenah, E., Britton, T., Halloran, M. E., and Longini, Jr., I. M. (2016). Molecular infectious disease epidemiology: Survival analysis and algorithms linking phylogenies to transmission trees. *PLoS Computational Biology* **12,** e1004869.

Kenah, E., Lipsitch, M., and Robins, J. M. (2008). Generation interval contraction and epidemic data analysis. *Mathematical biosciences* **213,** 71–79.

Morozova, O., Cohen, T., and Crawford, F. W. (2018). Risk ratios for contagious outcomes. *Journal of The Royal Society Interface* **15,** 20170696.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A* **162,** 121–129.

Rampey, Jr, A. H., Longini, Jr, I. M., Haber, M., and Monto, A. S. (1992). A discrete-time model for the statistical analysis of infectious disease incidence data. *Biometrics* **48,** 117–128.

Sugimoto, J. D., Yang, Y., Halloran, M. E., Dean, B., Oiulfstad, B., Bagwell, D. A., Mascola, L., Bancroft, E., and Longini, Jr., I. M. (2011). Accounting for unobserved immunity and asymptomatic infection in the early household transmission of the pandemic influenza a(h1n1)Unpublished.

Yang, Y., Sugimoto, J., Halloran, M. E., Basta, N. E., Chao, D. L., Matrajt, L., Potter, G., Kenah, E., and Longini, Jr, I. M. (2009). The transmissibility and control of pandemic influenza A(H1N1) virus. *Science* **326,** 729–733.

## Supporting Information

Online supporting information includes the following:

appendix.pdf Simulation results for each baseline contact interval distribution. (pdf)

datsim.py Python script for simulations. (text)

sim_data_analysis.R `R` script for analysis of simulation data. Requires example data files `pdata.csv`, `xdata.csv`, and `coef.csv`. (text)

LAdata_2018-11.csv Los Angeles County Department of Public Health household influenza transmission data. (text)

AFT_LAanalysis.R `R` code for LACDPH household data analysis. (text)

These are available with this paper at the Biometrics website on Wiley Online Library.

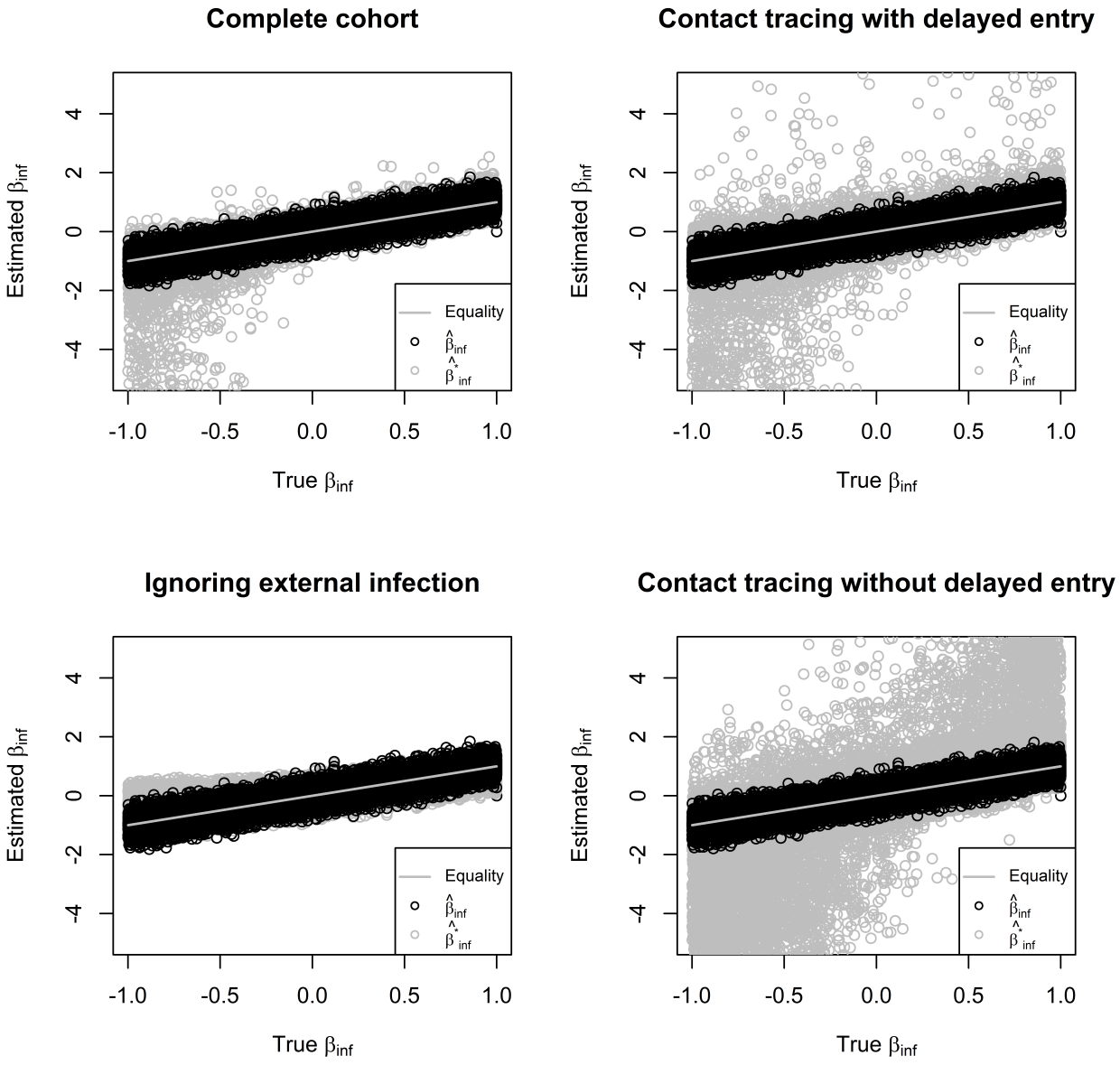


**Figure 1.** Scatterplot of infectiousness coefficient ($\beta_{\text{inf}}$) estimates from all simulations. The true values were independent random samples from a uniform(-1, 1) distribution.

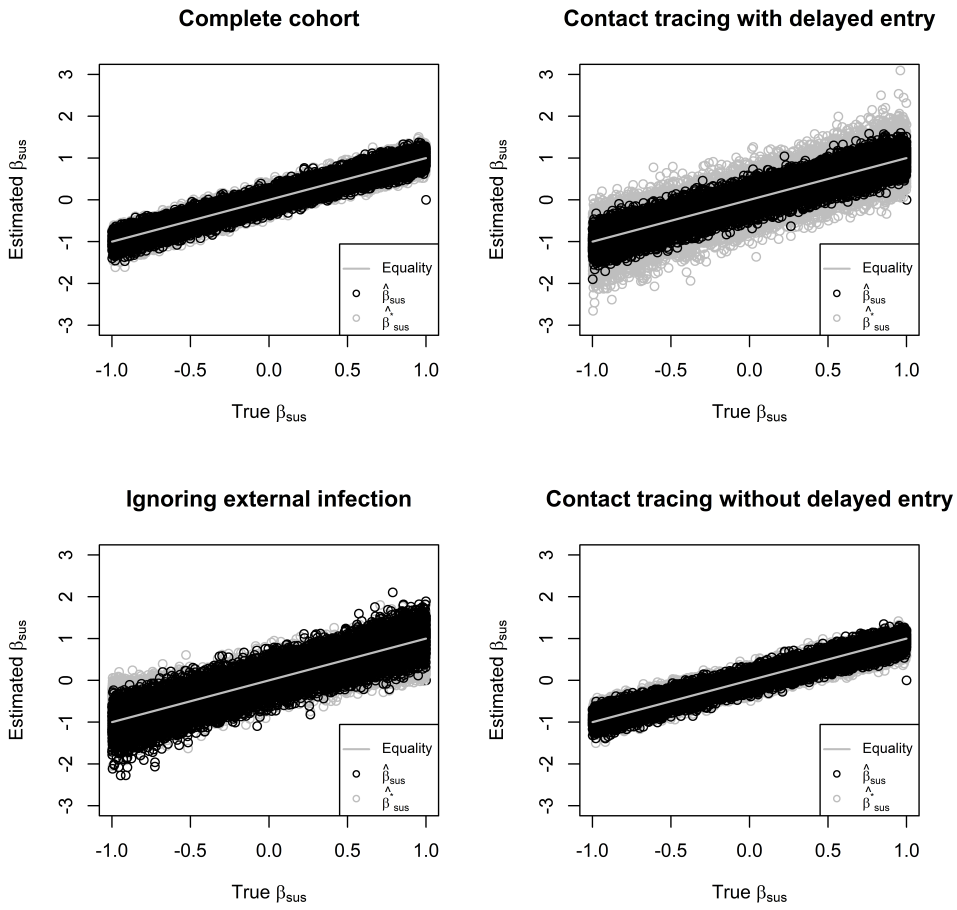


**Figure 2.** Scatterplot of susceptibility coefficient ($\beta_{\text{sus}}$) estimates from all simulations. The true values were independent random samples from a uniform(-1, 1) distribution.

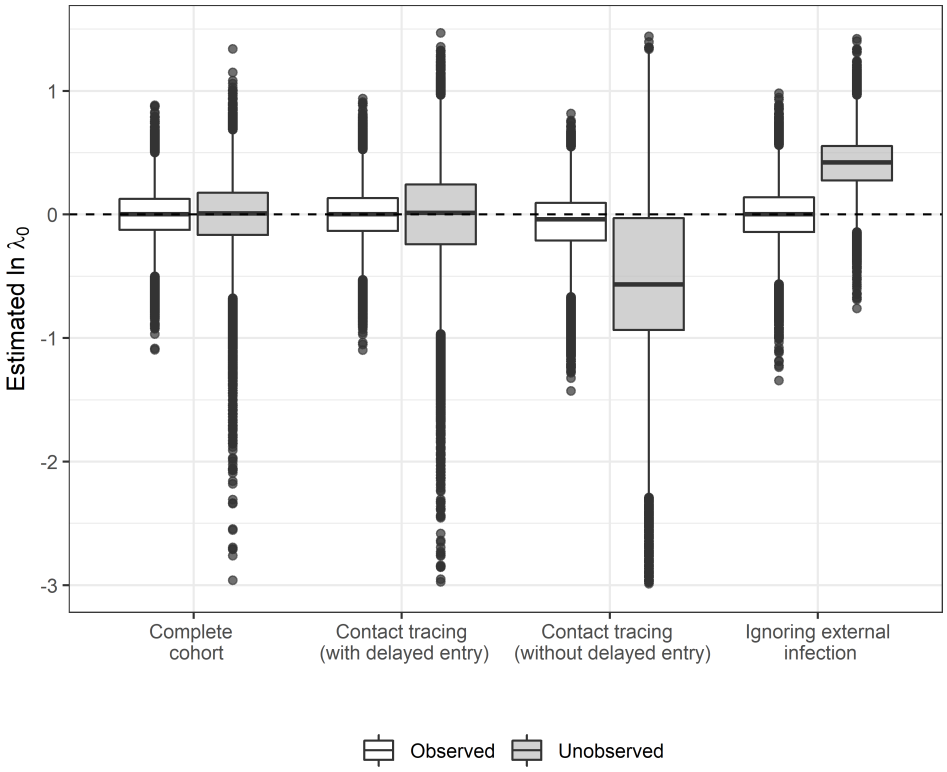


**Figure 3.** Boxplot of log baseline internal rate parameter ($\ln \lambda_0$) estimates from all simulations. The true parameter value was zero. The edges of the boxes correspond to the first and third quartiles. The whiskers extend to the highest and lowest data points within 1.5 IQR (interquartile range) of the nearest quartile.

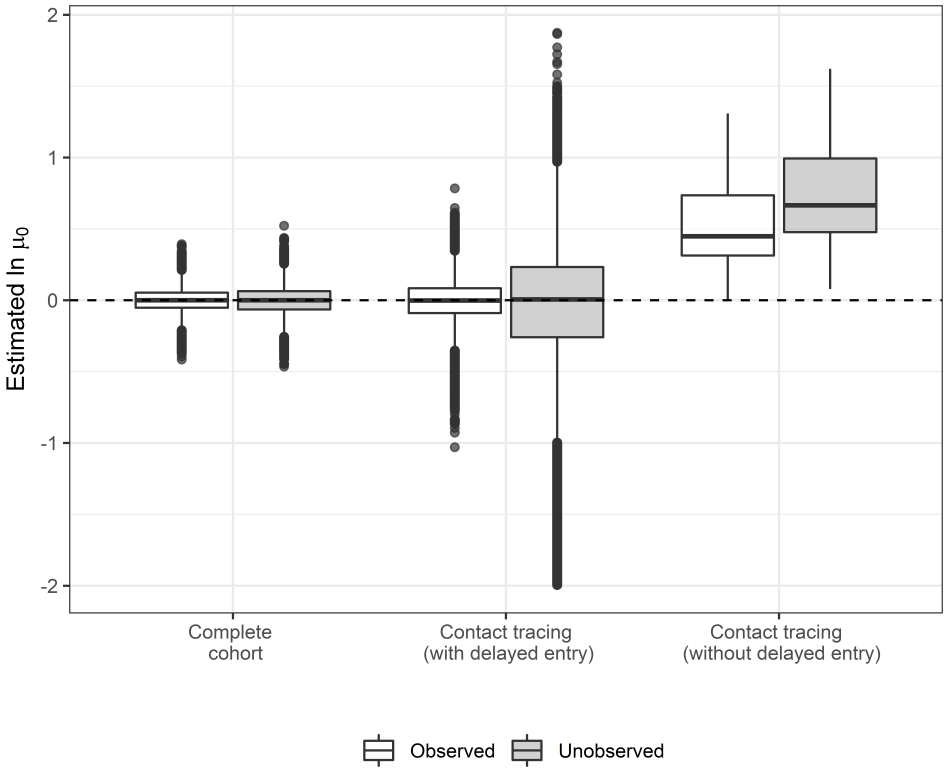


**Figure 4.** Boxplot of log baseline external rate parameter ($\ln \mu_0$) estimates from all simulations. The true parameter value was zero. The edges of the boxes correspond to the first and third quartiles. The whiskers extend to the highest and lowest data points within 1.5 IQR (interquartile range) of the nearest quartile.

**Table 1**
*Mean squared error (MSE) for point estimates in simulations*

| Study design | $\beta_{\text{inf}}$ | $\beta_{\text{sus}}$ | $\ln \lambda_0$ | $\ln \mu_0$ | $\ln \gamma_{\text{int}}$ |
|---|---|---|---|---|---|
| **Who-infected-whom observed** | | | | | |
| Complete cohort | 0.04 | 0.01 | 0.04 | 0.01 | 0.01 |
| Contact tracing with delayed entry | 0.04 | 0.03 | 0.04 | 0.02 | 0.01 |
| Contact tracing without delayed entry | 0.03 | 0.01 | 0.08 | 0.32 | 0.01 |
| Ignoring external infection | 0.04 | 0.05 | 0.05 | - | 0.01 |
| **Who-infected-whom not observed** | | | | | |
| Complete cohort | 0.06 | 0.02 | 0.07 | 0.01 | 0.03 |
| Contact tracing with delayed entry | 0.10 | 0.08 | 0.15 | 1.02 | 0.05 |
| Contact tracing without delayed entry | 0.24 | 0.02 | 0.62 | 0.59 | 0.19 |
| Ignoring external infection | 0.13 | 0.12 | 0.20 | - | 0.14 |

**Table 2**
*Coverage probabilities for 95% Wald and LR confidence intervals in simulations*

| Type | Study design | $\beta_{\text{inf}}$ | $\beta_{\text{sus}}$ | $\ln \lambda_0$ | $\ln \mu_0$ | $\ln \gamma_{\text{int}}$ |
|---|---|---|---|---|---|---|
| | **Who-infected-whom observed** | | | | | |
| Wald | Complete cohort | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 |
| | Contact tracing with delayed entry | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 |
| | Contact tracing without delayed entry | 0.95 | 0.94 | 0.88 | 0.01 | 0.91 |
| | Ignoring external infection | 0.95 | 0.95 | 0.94 | - | 0.94 |
| LR | Complete cohort | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| | Contact tracing with delayed entry | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| | Contact tracing without delayed entry | 0.95 | 0.94 | 0.88 | 0.01 | 0.94 |
| | Ignoring external infection | 0.95 | 0.95 | 0.95 | - | 0.91 |
| | **Who-infected-whom not observed** | | | | | |
| Wald | Complete cohort | 0.94 | 0.95 | 0.92 | 0.95 | 0.94 |
| | Contact tracing with delayed entry | 0.93 | 0.94 | 0.91 | 0.94 | 0.92 |
| | Contact tracing without delayed entry | 0.91 | 0.92 | 0.55 | 0.00 | 0.62 |
| | Ignoring external infection | 0.59 | 0.65 | 0.46 | - | 0.37 |
| LR | Complete cohort | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 |
| | Contact tracing with delayed entry | 0.94 | 0.94 | 0.94 | 0.95 | 0.93 |
| | Contact tracing without delayed entry | 0.90 | 0.92 | 0.52 | 0.00 | 0.61 |
| | Ignoring external infection | 0.59 | 0.66 | 0.50 | - | 0.38 |

**Table 3**
*AIC values for regression models using sex, age category, and antiviral prophylaxis for susceptibility and infectiousness.*

| **Internal contact intervals** | **External contact intervals** | | |
|---|---|---|---|
| | Exponential | Weibull | Log-logistic |
| Exponential | 207.74 | 208.01 | 207.66 |
| Weibull | 209.25 | 209.87 | 209.58 |
| Log-logistic | 209.22 | 209.82 | 209.52 |

**Table 4**
*Summary of final regression model with LR confidence intervals and p-values. Coefficients for covariates are log rate ratios, and $\gamma_{ext}$ is the log-logistic shape parameter.*

| Coefficient | Estimate | 95% CI | p-value |
|---|---|---|---|
| **Accounting for external infection** | | | |
| $\ln \lambda_0$ | $-4.90$ | $(-23.55, -3.38)$ | $< 0.0001$ |
| `adult_inf` | 1.36 | $(-0.36, 4.50)$ | 0.131 |
| `adult_sus` | $-0.48$ | $(-1.46, 0.18)$ | 0.139 |
| `proph_sus` | $-1.06$ | $(-2.26, -0.28)$ | 0.012 |
| $\ln \mu_0$ | $-4.10$ | $(-6.04, -3.59)$ | 0.021 |
| $\ln \gamma_{\text{ext}}$ | 0.80 | $(-0.76, 1.50)$ | 0.191 |
| **Ignoring external infection** | | | |
| $\ln \lambda_0$ | $-4.10$ | $(-5.46, -3.04)$ | $< 0.0001$ |
| `adult_inf` | 0.76 | $(-0.45, 2.10)$ | 0.218 |
| `adult_sus` | $-0.77$ | $(-1.87, 0.37)$ | 0.178 |
| `proph_sus` | $-0.83$ | $(-2.14, 0.30)$ | 0.155 |

| | Transmission from | | | |
|---|---|---|---|---|
| **Transmission to** | Child | | Adult | |
| Child untreated | 4.4% | (0.5%, 34.6%) | 15.9% | (6.4%, 36.4%) |
| Child on prophylaxis | 1.5% | (0.2%, 13.7%) | 5.8% | (2.0%, 15.9%) |
| Adult untreated | 2.7% | (0.3%, 20.1%) | 10.2% | (4.3%, 22.8%) |
| Adult on prophylaxis | 0.9% | (0.1%, 7.9%) | 3.6% | (1.2%, 10.5%) |

**Table 5**

*Predicted household secondary attack rates with Wald 95% confidence intervals.*

| | **Accounting for external infection** | | | **Ignoring external infection** | | |
|---|---|---|---|---|---|---|
| Coefficient | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| **Latent period = 1 day** | | | | | | |
| `adult_inf` | 0.08 | $(-1.24, 1.37)$ | 0.894 | 0.06 | $(-0.99, 1.09)$ | 0.909 |
| `adult_sus` | $-0.63$ | $(-1.52, 0.07)$ | 0.074 | $-0.58$ | $(-1.61, 0.52)$ | 0.286 |
| `proph_sus` | $-1.05$ | $(-2.26, -0.23)$ | 0.013 | $-0.95$ | $(-2.24, 0.13)$ | 0.087 |
| **Infectious period = 5 days** | | | | | | |
| `adult_inf` | 1.73 | $(-0.43, 4.87)$ | 0.140 | 0.58 | $(-0.68, 1.95)$ | 0.366 |
| `adult_sus` | $-0.41$ | $(-1.40, 0.20)$ | 0.167 | $-0.91$ | $(-2.08, 0.25)$ | 0.121 |
| `proph_sus` | $-1.02$ | $(-2.18, -0.29)$ | 0.011 | $-0.69$ | $(-2.02, 0.47)$ | 0.246 |
| **Infectious period = 7 days** | | | | | | |
| `adult_inf` | 0.29 | $(-0.99, 1.73)$ | 0.649 | 0.20 | $(-0.88, 1.28)$ | 0.707 |
| `adult_sus` | $-0.57$ | $(-1.38, 0.15)$ | 0.103 | $-0.44$ | $(-1.44, 0.64)$ | 0.406 |
| `proph_sus` | $-0.90$ | $(-1.96, -0.12)$ | 0.026 | $-0.73$ | $(-1.89, 0.30)$ | 0.167 |

**Table 6** *Log rate ratios in sensitivity analysis with LR confidence intervals and p-values.*