Published in final edited form as:

J Am Stat Assoc. 2015 March 1; 110(509): 313–325. doi:10.1080/01621459.2014.896807.

Semiparametric Relative-risk Regression for Infectious Disease Transmission Data

Eben Kenah

Eben Kenah is Assistant Professor, Biostatistics Department, University of Florida, Gainesville, FL, 326110-7450 (ekenah@ufl.edu)

Abstract

This paper introduces semiparametric relative-risk regression models for infectious disease data. The units of analysis in these models are pairs of individuals at risk of transmission. The hazard of infectious contact from *i* to *j* consists of a baseline hazard multiplied by a relative risk function that can be a function of infectiousness covariates for *i*, susceptibliity covariates for *j*, and pairwise covariates. When who-infects-whom is observed, we derive a profile likelihood maximized over all possible baseline hazard functions that is similar to the Cox partial likelihood. When who-infects-whom is not observed, we derive an EM algorithm to maximize the profile likelihood integrated over all possible combinations of who-infected-whom. This extends the most important class of regression models in survival analysis to infectious disease epidemiology. These methods can be implemented in standard statistical software, and they will be able to address important scientific questions about emerging infectious diseases with greater clarity, flexibility, and rigor than current statistical methods allow.

Keywords

Survival analysis; Epidemiology; EM algorithm; Chain-binomial model

1 INTRODUCTION

Infectious diseases are an important threat to human health and commerce, and understanding the transmission of disease is crucial to the design of public health interventions. The statistical analysis of infectious disease data is complicated by the fact that infections are inherently dependent, especially when they are transmitted directly from person to person (Becker, 1989; Andersson and Britton, 2000). Epidemiologists have dealt with this problem in three ways. The most common approach is to model susceptibility to disease using standard statistical methods such as logistic or Cox regression, ignoring disease transmission. A second approach is to use chain binomial models (Rampey et al., 1992), which estimate the probability of escaping infectious contact from infected members

SUPPLEMENTARY MATERIALS

Supplementary material available online at the *Journal of the American Statistical Association* website includes:

Appendices for Section 2 Outline of sufficient conditions for consistency and asymptotic normality (Appendix A) and derivation of the asymptotic variance of baseline hazard estimates (Appendix B).

Simulation code for Section 3 Python module used for the simulations, including statistical analyses in R.

of groups such as households, classrooms, or hospital wards. The third approach is to model the spread of disease as a branching process where infectees are the offspring of their infectors (Wallinga and Teunis, 2004; White and Pagano, 2008). The time between the infections of an infector and an infectee is called a generation interval. In this approach, the generation intervals are assumed to be independent and identically distributed (iid).

To understand transmission, it is crucial to separate the effects of covariates on infectiousness and susceptibility from their association with exposure to infected people (Rhodes et al., 1996). Regression models that ignore transmission cannot do this. When disease transmission is modeled as a branching process, uninfected people do not exist and cannot be exposed to infected people. The failure to account for uninfected person-time and competing risks of infection cause several problems with this approach (Svensson, 2007; Kenah et al., 2008). The assumption that generation intervals are iid is difficult to relax, making estimation of covariate effects difficult (Kenah, 2013). Chain binomial models are a statistically sound response to the problem of dependence and can be used to estimate covariate effects. However, their use is limited in two ways: First, they are not implemented in standard statistical software—a problem solved partially by the publicly-available package TranStat (www.epimodels.org/midas/transtat.do). Second, they use discrete time. Since infectious disease data are usually recorded by the day or week, this is not unnatural. However, continuous-time models corrected for ties may offer a more flexible modeling framework.

Kenah (2011) extended parametric methods from survival analysis to infectious disease data by modeling the contact interval. In the ordered pair *ij*, the contact interval τ_{ij} is the time between the onset of infectiousness in *i* and the first infectious contact from *i* to *j*, where infectious contact is a contact sufficient to infect a susceptible individual. It is right-censored if the infectious period of *i* ends before *i* makes infectious contact with *j* or if *j* is infected by someone other than *i*. These methods solve the problem of dependence by treating ordered pairs of individuals, not the individuals themselves, as the units of analysis. Kenah (2013) showed that the contact interval distribution could be estimated nonparametrically by adapting the Nelson-Aalen estimator from standard survival analysis. These methods assume a homogeneous population where the contact interval distribution is the same for all pairs *ij* in which transmission from *i* to *j* is possible. They are unable to estimate covariate effects on transmission, which is a primary goal of vaccine trials, outbreak investigations, and many other studies of infectious disease.

The goal of this paper is to extend the methods of Kenah (2013) to develop a relative-risk regression model similar to that of Cox (1972). This model will allow semiparametric estimation of the effects of covariates on the hazard of infectious contact in pairs of individuals. For the ordered pair *ij*, the covariate vector can include infectiousness covariates for *i*, susceptibility covariates for *j*, and pairwise covariates. This semiparametric regression model will allow many of the most important scientific questions in infectious disease epidemiology to be addressed with greater clarity, flexibility, and rigor.

1.1 Stochastic S(E)IR epidemic model

Consider a closed population of n individuals assigned indices 1 . . . n. Each individual 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 ε_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 ${}_i$. At time $t_i + \varepsilon_i + l_i$, i moves from I to R. Once in R, i can no longer infect others or be infected. The states and notation are illustrated at the top of Figure 1. The latent period is a nonnegative random variable, the infectious period is a strictly positive random variable, and both have finite mean and variance.

An epidemic begins with one or more persons infected from outside the population, which we call *imported infections*. The methods in this paper require that the set of imported infections is known. For simplicity, we assume that epidemics begin with one or more imported infections at time 0 and there are no other imported infections.

After becoming infectious at time $t_i + \varepsilon_i$, person i makes infectious contact with $j \neq i$ at time $t_{ij}=t_i+\varepsilon_i+\tau^*_{ij}$, where the *infectious contact interval* τ^*_{ij} is a strictly positive random variable with $\tau^*_{ij}=\infty$ if infectious contact never occurs. Since infectious contact must occur while i is infectious or never, $\tau^*_{ij} \in (0, \iota_i]$ or $\tau^*_{ij}=\infty$. We define infectious contact to be a contact sufficient infect a susceptible person, so $t_j \leq t_{ij}$ for all $i \neq j$. The infectious contact interval is illustrated at the bottom of Figure 1.

For each ordered pair ij, let $C_{ij} = 1$ if infectious contact from i to j is possible and $C_{ij} = 0$ otherwise. These C_{ij} could be the entries in an adjacency matrix for a static contact network. We assume that the infectious contact interval τ^*_{ij} is generated in the following way: A *contact interval* τ_{ij} is drawn from a distribution with hazard function $\lambda_{ij}(\tau)$. If $\tau_{ij} \leq l_i$ and $C_{ij} = 1$, then $\tau^*_{ij}=\tau_{ij}$. Otherwise, $\tau^*_{ij}=\infty$. In this paper, we assume the contact intervals in all ordered pairs ij are independent and have finite mean and variance.

1.2 Observation and censoring

Our population has size n, and we observe the times of all S $\rightarrow$ E (infection), E $\rightarrow$ I (onset of infectiousness), and I $\rightarrow$ R (removal) transitions in the population between time 0 and time T . For all ordered pairs ij such that i is infected, we observe C_{ij}. We first consider the case where who-infects-whom is observed and then consider the more realistic case where it is not.

We assume that we can observe τ_{ij} only if j is infected by i at time $t_i+\varepsilon_i+\tau_{ij}$. Clearly, τ_{ij} can be observed only if $C_{ij} = 1$. We also have right-censoring of τ_{ij}:

1. Since infectious contact can occur only while i is infectious, τ_{ij} can be right-censored by the infectious period l_i of i. Let $\mathscr{I}_i\left(\tau\right)=\mathbf{1}_{\tau \in (0,\iota_i]}$ indicate whether i remains infectious at infectious age τ.

2. Since *j* is susceptible to infection by *i* only if he or she has not been infected by anyone else, τ_{ij} can be right-censored by $t_{i'j} - t_i - \varepsilon_i$ for $i' \neq i$. Let $\mathscr{S}_{ij}(\tau) = \mathbf{1}_{t_i + \varepsilon_i + \tau \leq t_j}$ indicate whether *j* remains susceptible at infectious age τ of *i*.

Let T denote the time at which observation ends. Then τ_{ij} can be right-censored by the end of observation at infectious age $T - t_i - \varepsilon_i$ of i. Let $\mathscr{Y}_i(\tau) = \mathbf{1}_{t_i + \varepsilon_i + \tau \leq T}$ indicate whether observation is ongoing when *i* reaches infectious age τ.

Since $\mathscr{I}_i(\tau)$, $\mathscr{S}_{ij}(\tau)$, and $\mathscr{Y}_{ij}(\tau)$ are left-continuous,

$$Y_{ij}(\tau) = C_{ij} \mathscr{I}_i(\tau) \mathscr{S}_{ij}(\tau) \mathscr{Y}_i(\tau) \quad (1)$$

is a left-continuous process that indicates the risk of an observed infectious contact from *i* to *j* at infectious age τ of *i*. The assumptions made in the stochastic S(E)IR model above ensure that $\mathscr{I}_i(\tau)$ and $\mathscr{S}_{ij}(\tau)$ independently censor τ_{ij}. We also require that *T* is a stopping time with respect to the observed data such that, for all *i*, $\mathscr{Y}_i(\tau)$ independently censors τ_{ij} for each *j* exposed to infectious contact from *i*. The possible censoring scenarios are illustrated in Figure 2.

1.3 Transmission trees and infectious sets

Following Wallinga and Teunis (2004), let v_j denote the index of the person who infected person *j*, with $v_j = 0$ for imported infections and $v_j = \infty$ for persons not infected prior to the end of observation. The *transmission tree* is the directed network with an edge from v_j to *j* for each *j* such that $t_j \leq T$. It can be represented by a vector $\mathbf{v} = (v_1, \ldots, v_n)$. Let $\mathscr{V}_j = \{i : C_{ij} \mathscr{I}_i(t_j) = 1\}$ denote the set of possible infectors of person *j*, which we call the *infectious set* of *j*. Let $\mathscr{V}$ denote the set of all **v** consistent with the observed data. A $\mathbf{v} \in \mathscr{V}$ can be generated by choosing a $v_j \in \mathscr{V}_j$ for each non-imported infection *j*.

2 METHODS

Kenah (2013) described the nonparametric estimation of the contact interval distribution for a homogeneous population. Here, we consider a semiparametric relative-risk model like that of Prentice and Self (1983). Let

$$\lambda_{ij}(\tau) = r\left(\beta_0^{\mathrm{T}} X_{ij}(\tau)\right) \lambda_0(\tau), \quad (2)$$

where $\lambda_0(\tau)$ is an unspecified baseline hazard function, $r : \mathbb{R} \to (0, \infty)$ is a relative risk function, β_0 is an unknown $b \times 1$ coefficient vector, and $X_{ij}(\tau)$ is a $b \times 1$ predictable covariate process taking values in a set $\mathscr{X}$. The covariates $X_{ij}(\tau)$ can include individual-level covariates predicting the infectiousness of *i* or the susceptibility of *j* as well as pairwise covariates (e.g., membership in the same household) that predict the hazard of infectious contact from *i* to *j*.

We assume that *r* has continuous first and second derivatives, $r(0) = 1$, and $\ln r(\beta^{\mathrm{T}} X)$ is bounded on $\mathscr{X}$. Letting $r(x) = \exp(x)$ gives us a loglinear relative risk regression model like that of Cox (1972), and letting *r*(*x*) = 1 + *x* gives us a linear relative risk regression model. To fit these semiparametric models, we adapt the nonparametric estimators from Kenah (2013) to account for the relative risk function.

2.1 Who-infects-whom is observed

Let $N_{ij}(\tau) = \mathbf{1}_{\tau_{ij} \leq \tau}$ indicate whether an observed infectious contact from *i* to *j* has occurred by infectious age τ in *i*, and let $N(\tau) = \sum_{j=1}^{n} \sum_{i \neq j} N_{ij}(\tau)$. Note that we must observe who-infected-whom in order to calculate $N(\tau)$.

Let $\Lambda_0(\tau) = \int_0^\tau \lambda_0(u)\,\mathrm{d}u$. Given β, the Breslow estimator (Breslow, 1972) of $\Lambda_0(\tau)$ is

$$\hat{\Lambda}(\beta,\tau) = \int_0^\tau \frac{1}{Y(\beta,u)} \mathrm{d}N(u), \quad (3)$$

where

$$Y(\beta,u) = \sum_{j=1}^{n} \sum_{i \neq j} r\left(\beta^\mathsf{T} X_{ij}(u)\right) Y_{ij}(u). \quad (4)$$

The Breslow estimator has two desirable properties. First, $\hat{\Lambda}(\beta_0,\tau)$ is an unbiased estimator of $\Lambda_0(\tau)$. Let

$$M(\beta,\tau) = N(\tau) - \int_0^\tau Y(\beta,u)\,\lambda_0(u)\,\mathrm{d}u. \quad (5)$$

Then for all τ such that $Y(\tau) > 0$,

$$\hat{\Lambda}(\beta_0,\tau) - \Lambda_0(\tau) = \int_0^\tau \frac{\mathbf{1}_{Y(u)>0}}{Y(\beta_0,u)} \mathrm{d}M(\beta_0,u) \quad (6)$$

is a mean-zero martingale when $\beta = \beta_0$. Second, $\hat{\Lambda}(\beta,\tau)$ maximizes the log likelihood

$$\ell(\beta,\Lambda) = \sum_{j=1}^{n} \sum_{i \neq j} ln\left(r\left(\beta^\mathsf{T} X_{v_j j}\left(\tau_{v_j j}\right)\right) \mathrm{d}\Lambda\left(\tau_{v_j j}\right)\right) - \int_0^\infty Y(\beta,u)\,\mathrm{d}\Lambda(u) \quad (7)$$

over all step functions $\Lambda(\tau)$. Substituting $\hat{\Lambda}(\beta,\tau)$ into *l*(β, Λ) profile likelihood

$$\ell\left(\beta,\hat{\Lambda}\right) = \left(\sum_{j=1}^{n} ln \frac{r\left(\beta^\mathsf{T} X_{v_j j}\left(\tau_{v_j j}\right)\right)}{Y\left(\beta,\tau_{v_j j}\right)}\right) - \tau, \quad (8)$$

where $\mathscr{T} = max\{\tau : Y(\tau) > 0\}$. The first term is similar to the log partial likelihood from Cox (1972) and the second term does not depend on β. Dropping the second term, let

$$pl(\beta) = \sum_{j=1}^{n} ln \frac{r\left(\beta^\mathsf{T} X_{v_j j}\left(\tau_{v_j j}\right)\right)}{Y\left(\beta,\tau_{v_j j}\right)} \quad (9)$$

be the log partial likelihood for *β*. This derivation of the partial likelihood as a profile likelihood follows that of Johansen (1983). Let $\hat{\beta}$ denote the value of *β* that maximizes *pl*(*β*), and let $\hat{\Lambda}_0(\tau) = \hat{\Lambda}\left(\hat{\beta}, \tau\right)$ denote the corresponding Breslow estimate of the baseline cumulative hazard.

2.2 Partial likelihood score process

We can rewrite *pl*(*β*) as a sum of stochastic integrals:

$$pl(\beta) = \sum_{j=1}^{n} \sum_{i \neq j} \int_0^{\infty} ln \frac{r\left(\beta^{\mathsf{T}} X_{ij}(u)\right)}{Y(\beta, u)} \mathrm{d}N_{ij}(u). \quad (10)$$

The corresponding score process is

$$U(\beta, \tau) = \sum_{j=1}^{n} \sum_{i \neq j} \int_0^{\tau} \frac{\partial}{\partial \beta} ln \quad r\left(\beta^{\mathsf{T}} X_{ij}(u)\right) - E(\beta, u)\, \mathrm{d}N_{ij}(u), \quad (11)$$

where

$$E(\beta, u) = \frac{\sum_{j=1}^{n} \sum_{i \neq j} r\left(\beta^{\mathsf{T}} X_{ij}(u)\right) Y_{ij}(u) \frac{\partial}{\partial \beta} ln\, r\left(\beta^{\mathsf{T}} X_{ij}(u)\right)}{\sum_{j=1}^{n} \sum_{i \neq j} r\left(\beta^{\mathsf{T}} X_{ij}(u)\right) Y_{ij}(u)}. \quad (12)$$

is the expected value of $\frac{\partial}{\partial \beta} ln\, r\left(\beta^{\mathsf{T}} X_{ij}(u)\right)$ over the risk set at *u* when each pair is weighted by its hazard of infectious contact at *u*. By the Doob-Meyer decomposition, there is a mean-zero martingale $M_{ij}(u)$ for each *ij* such that

$$\mathrm{d}N_{ij}(u) = r\left(\beta_0^{\mathsf{T}} X_{ij}(u)\right) \lambda_0(u) Y_{ij}(u)\, \mathrm{d}u + \mathrm{d}M_{ij}(u). \quad (13)$$

Expanding equation (11) using this decomposition and simplifying, we get

$$U(\beta_0, \tau) = \sum_{j=1}^{n} \sum_{i \neq j} \int_0^{\tau} \frac{\partial}{\partial \beta} ln \quad \frac{r\left(\beta_0^{\mathsf{T}} X_{ij}(u)\right)}{Y(\beta_0, u)} \mathrm{d}M_{ij}(u). \quad (14)$$

Since it is a sum of integrals of predictable processes with respect to martingales, $U(\beta_0, \tau)$ is a mean-zero martingale.

2.3 Observed and expected information

Let $v^{\otimes 2} = v v^{\mathsf{T}}$ for a column vector *v*. Since the $N_{ij}(\tau)$ do not jump simultaneously in continuous time, the predictable variation process of $U(\beta_0, \tau)$ is

$$\langle U(\beta_0) \rangle (\tau) = \int_0^{\tau} V(\beta_0, u)\, Y(\beta_0, u)\, \lambda_0(u)\, \mathrm{d}u, \quad (15)$$

where

$$V\left(\beta,u\right)=\sum_{j=1}^{n}\sum_{i\neq j}\left(\frac{\partial}{\partial\beta}ln\frac{r\left(\beta^{\top}\ X_{ij}\left(u\right)\right)}{Y\left(\beta,u\right)}\right)^{\otimes 2}\frac{r\left(\beta^{\top}\ X_{ij}\left(u\right)\right)Y_{ij}\left(u\right)}{Y\left(\beta,u\right)}\quad(16)$$

is the variance of $\frac{\partial}{\partial\beta}\,ln\,r\left(\beta^{\top}\ \ X_{ij}\left(u\right)\right)$ over the risk set at *u* when each pair *ij* is weighted by its hazard of infectious contact at *u*.

Let $I\left(\beta\right)=-\frac{\partial^{2}}{\partial\beta^{2}}pl\left(\beta\right)$ be the observed information. Then

$$\begin{aligned}I\left(\beta\right)=&\sum_{j=1}^{n}\sum_{i\neq j}\int_{0}^{\infty}\left(\frac{\partial}{\partial\beta}ln\quad r\left(\beta^{\top}\ X_{ij}\left(u\right)\right)\right)^{\otimes 2}\\&-E(\beta,u)^{\otimes 2}\mathrm{d}N_{ij}\left(u\right)\\&-\sum_{j=1}^{n}\sum_{i\neq j}\int_{0}^{\infty}\frac{\frac{\partial^{2}}{\partial\beta^{2}}r\left(\beta^{\top}\ X_{ij}\left(u\right)\right)}{r\left(\beta^{\top}\ X_{ij}\left(u\right)\right)}\qquad(17)\\&-\frac{\frac{\partial^{2}}{\partial\beta^{2}}Y\left(\beta,u\right)}{Y\left(\beta,u\right)}\mathrm{d}N_{ij}\left(u\right).\end{aligned}$$

Expanding *I*(*β*$_0$) via the Doob-Meyer decomposition (13) and simplifying, we get

$$I\left(\beta_{0}\right)=\int_{0}^{\infty}V\left(\beta_{0},u\right)Y\left(\beta_{0},u\right)\lambda_{0}\left(u\right)\mathrm{d}u+\sum_{j=1}^{n}\sum_{i\neq j}\int_{0}^{\infty}\frac{\partial^{2}}{\partial\beta^{2}}ln\frac{r\left(\beta_{0}^{\top}\ X_{ij}\left(u\right)\right)}{Y\left(\beta_{0},u\right)}\mathrm{d}M_{ij}\left(u\right).\quad(18)$$

The second term has expectation zero, so *I*(*β*$_0$) is an unbiased estimate of Var(*U*(*β*$_0$, ∞)).

Another estimate Var(*U*(*β*$_0$, ∞)) is obtained by substituting the increments of the Breslow estimator (3) for λ$_0$(*u*) d*u* in equation (15). This gives us the estimated expected information

$$\mathscr{I}\left(\beta\right)=\int_{0}^{\infty}V\left(\beta,u\right)\mathrm{d}N\left(u\right).\quad(19)$$

Expanding $\mathscr{I}\left(\beta_{0}\right)$ using the Doob-Meyer decomposition and simplifying, we get

$$\mathscr{I}\left(\beta_{0}\right)=\int_{0}^{\infty}V\left(\beta_{0},u\right)Y\left(\beta_{0},u\right)\lambda_{0}\left(u\right)\mathrm{d}u+\int_{0}^{\infty}V\left(\beta_{0},u\right)\mathrm{d}M\left(u\right),\quad(20)$$

where $M\left(\tau\right)=\sum_{j=1}^{n}\sum_{i\neq j}M_{ij}\left(\tau\right)$. The second term has expectation zero, so $\mathscr{I}\left(\beta_{0}\right)$ is also an unbiased estimate of the variance of *U*(*β*$_0$, ∞). $\mathscr{I}\left(\beta_{0}\right)$ may be a better estimator of Var(*U*(*β*$_0$, ∞)) than *I*(*β*$_0$) because it is guaranteed to be positive semidefinite (Prentice and Self, 1983) and it depends only on aggregates over risk sets (Aalen et al., 2009).

When *r*(*x*) = exp(*x*) as in the Cox model, $I\left(\beta\right)=\mathscr{I}\left(\beta\right)$ for all *β*. For general *r*(*x*), *I*(*β*$_0$) and $\mathscr{I}\left(\beta_{0}\right)$ are asymptotically equivalent under weak regularity conditions (see Appendix A online).

2.4 Large-sample estimation of β_0 and $\Lambda_0(\tau)$

Appendix A, available online, outlines sufficient conditions for the asymptotic normality of $U(\beta_0, \tau)$ and $\hat{\beta}$ as $m \to \infty$, where m is the number of pairs ij at risk of transmission. Under these conditions, hypothesis tests and confidence intervals for β_0 can be obtained using score, Wald, or likelihood ratio statistics. These conditions are very similar to those for the standard Cox model (Prentice and Self, 1983) except for the additional requirement that both the number of susceptibles and the number of pairs be large such that each susceptible is exposed to a number of infectors $< < m$. When a given susceptible j is infected, all pairs ij are censored. If there were many pairs but few susceptibles, each susceptible would be exposed to a very high hazard of infection and most pairs would be censored very quickly. To take an extreme case, imagine a single susceptible exposed to m infecteds. The number of pairs at risk of transmission is m but the susceptible will be infected almost immediately when m is large. After this, there are no more pairs at risk of transmission and we can learn nothing further about β_0 or $\Lambda_0(\tau)$.

Given $\hat{\beta}$ the Breslow estimator of $\Lambda_0(\tau)$ is $\hat{\Lambda}_0(\tau) = \hat{\Lambda}_0\left(\hat{\beta}, \tau\right)$. Its variance is consistently estimated by

$$\hat{\sigma}_0^2(\tau) = \left(\frac{\partial}{\partial\beta}\hat{\Lambda}\left(\hat{\beta}, \tau\right)\right)^{\top} I\left(\hat{\beta}\right)^{-1}\left(\frac{\partial}{\partial\beta}\hat{\Lambda}\left(\hat{\beta}, \tau\right)\right) + \int_0^\tau \frac{1}{Y\left(\hat{\beta}, u\right)^2} \mathrm{d}N(u), \quad (21)$$

which is derived in Appendix B.1. $I\left(\hat{\beta}\right)$ can be replaced by $\mathscr{I}\left(\hat{\beta}\right)$. Using the martingale central limit theorem and a log transformation, we get the approximate pointwise $1 - \alpha$ confidence limits

$$\hat{\Lambda}_0(\tau)\, exp\left(\pm\frac{\hat{\sigma}_0(\tau)}{\hat{\Lambda}_0(\tau)}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right). \quad (22)$$

Point and interval estimates for the baseline survival function can be obtained using the relationship $S_0(\tau) = \exp(-\Lambda_0(\tau)$

2.5 Who-infects-whom is not observed

When we do not observe who-infected-whom, we do not know which contact intervals are observed and which are censored. It impossible to calculate the partial likelihood $pl(\beta)$ or the Breslow estimate $\hat{\Lambda}(\beta, \tau)$. In this section, we show how an EM algorithm similar to that of Kenah (2013) can be used to obtain consistent and asymptotically normal estimates of β_0 and $\Lambda_0(\tau)$.

Given β, $\lambda(\tau)$, and the observed information, the probability that j was infected by i is

$$p_{ij}(\beta, \lambda) = \frac{r\left(\beta^{\top} X_{ij}(t_j - t_i - \varepsilon_i)\right)\lambda(t_j - t_i - \varepsilon_i)\mathbf{1}_{i \in \mathscr{V}_j}}{\sum_{k \in \mathscr{V}_j} r\left(\beta^{\top} X_{kj}(t_j - t_k - \varepsilon_k)\right)\lambda(t_j - t_k - \varepsilon_k)}, \quad (23)$$

and the infectors of different infected persons can be chosen independently (Kenah et al., 2008). The probability of a transmission network $\mathbf{v} = (v_1, \ldots, v_n)$ given β, $\lambda(\tau)$, and the observed data is

$$Pr\left(\mathbf{v}|\beta,\lambda,\text{observed}\ \ \text{data}\right)=\prod_{j:0<v_j<\infty} p_{v_j j}\left(\beta.\lambda\right). \quad (24)$$

Note that the last two equations assume a continuous contact interval distribution, so simultaneous infectious contacts have probability zero.

Let $pl_{\mathbf{v}}(\beta)$ be the log partial likelihood that we would have calculated had we observed the transmission network $\mathbf{v}$. Given a coefficient vector β^* and a baseline hazard function $\lambda^*(\tau)$, the expected log likelihood is

$$\widetilde{pl}_{\beta^*,\lambda^*}\left(\beta\right)=\sum_{\mathbf{v}\in\mathscr{V}} pl_{\mathbf{v}}\left(\beta\right)pr\left(\mathbf{v}|\beta^*,\lambda^*,\text{observed}\ \ \text{data}\right)=\sum_{j=1}^{n}\sum_{i\neq j}\int_0^{\mathscr{T}} ln\frac{r\left(\beta^{\mathsf{T}}\ X_{ij}\left(u\right)\right)}{Y\left(\beta,u\right)}\mathrm{d}\widetilde{N}_{ij}\left(u|\beta^*,\lambda^*\right), \quad (25)$$

where $\tilde{N}_{ij}(\tau|\beta^*, \lambda^*) = p_{ij}(\beta^*, \lambda^*)\mathbf{1}_{\tau\geq t_j-t_i-\varepsilon_i}$. Now let $N(\tau|\mathbf{v})$ be the value of $N(\tau)$ that we would have calculated had we observed the transmission network $\mathbf{v}$. The corresponding Breslow estimate is

$$\hat{\Lambda}_{\mathbf{v}}\left(\beta,\tau\right)=\int_0^{\tau}\frac{1}{Y\left(\beta,u\right)}\mathrm{d}N\left(u|\mathbf{v}\right). \quad (26)$$

The marginal Breslow estimate given β^* and $\lambda^*(\tau)$ is

$$\widetilde{\Lambda}_{\beta^*,\lambda^*}\left(\beta,\tau\right)=\sum_{\mathbf{v}\in\mathscr{V}}\hat{\Lambda}_{\mathbf{v}}\left(\beta,\tau\right)Pr\left(\mathbf{v}|\beta^*,\lambda^*,\text{observed}\ \ \text{data}\right)=\int_0^{\tau}\frac{1}{Y\left(\beta,u\right)}\mathrm{d}\widetilde{N}\left(u|\beta^*,\lambda^*\right), \quad (27)$$

where $\widetilde{N}\left(\tau|\beta^*,\lambda^*\right)=\sum_{j=1}^{n}\sum_{i\neq j}\widetilde{N}_{ij}\left(\tau|\beta^*,\lambda^*\right)$.

For the relative risk function $r(x) = \exp(x)$, the expected log partial likelihood $\widetilde{pl}_{\beta^*,\lambda^*}\left(\beta\right)$ is the log partial likelihood of a weighted Cox regression model (Therneau and Grambsch, 2000) with two copies of each pair ij: an uncensored copy with weight $p_{ij}(\beta^*, \lambda^*)$ and a censored copy with weight $1 - p_{ij}(\beta^*, \lambda^*)$. The baseline hazard estimate from this model is the marginal Breslow estimate $\widetilde{\Lambda}_{\beta^*,\lambda^*}\left(\widetilde{\beta},\tau\right)$, where $\widetilde{\beta}=arg\ max_{\beta}\,\widetilde{pl}_{\beta^*,\lambda^*}\left(\beta\right)$.

2.6 EM algorithm

When who-infects-whom is not observed, the semiparametric regression model can be fit using the ECM algorithm of Meng and Rubin (1993), which is an extension of the EM algorithm of Dempster et al. (1977). In each iteration, we first estimate β_0 using the expected log partial likelihood and then calculate the marginal Breslow estimator of $\Lambda_0(\tau)$. We use these new estimates to re-weight the possible $\mathbf{v}$. The entire process is described in Algorithm 1.

Algorithm 1 ECM algorithm for semiparametric estimation of β_0 and $\Lambda_0(\tau)$.

To show that this is an ECM algorithm, we must show that the CM1 and CM2 steps are conditional maximizations of the expected log likelihood. The CM1 step is a conditional maximization by definition, so it remains to show that the CM2 step is a conditional maximization. Given a coefficient vector β^* and a hazard function λ^*, the expected log likelihood is

$$\tilde{\ell}_{\beta^*,\lambda^*}(\beta,\Lambda)=\sum_{j=1}^{n}\sum_{i\neq j}p_{ij}(\beta^*,\lambda^*)\,ln\left(r\left(\beta^{\mathsf{T}}X_{ij}(t_j-t_i-\varepsilon_i)\right)\mathrm{d}\Lambda(t_j-t_i-\varepsilon_i)\right)-\int_0^{\infty}Y(\beta,u)\,\mathrm{d}\Lambda(u)\,. \quad (28)$$

Differentiating with respect to $\mathrm{d}\Lambda(t_j - t_i - \varepsilon_i)$ for each *i* and *j* shows that, for a fixed β, $l\tilde{\beta}^*,_{\lambda^*}$ (β, Λ) is maximized over all step functions $\Lambda(\tau)$ by setting

$$\mathrm{d}\Lambda(t_j-t_i-\varepsilon_i)=\frac{p_{ij}(\beta^*,\lambda^*)}{Y(\beta,t_j-t_i-\varepsilon_i)}, \quad (29)$$

exactly as in the marginal Breslow estimator $\tilde{\Lambda}_{\beta^*,\lambda^*}(\beta,\tau)$. Therefore, Algorithm 1 is an ECM algorithm. When it is known that $\beta = 0$, it reduces to the EM algorithm in Kenah (2013). Therefore, the convergence of both $\beta^{(k)}$ and $\Lambda^{(k)}(\tau)$ should be monitored.

2.7 Large-sample estimation of β_0

Let $\tilde{\beta}$ denote the estimate of β_0 to which the ECM algorithm converges, and let $\tilde{\lambda}(\tau)$ denote the corresponding estimate of $\lambda_0(\tau)$. Let $U_{\mathbf{v}}(\tau, \beta)$ and $I_{\mathbf{v}}(\beta)$ denote the score and the observed information that we would have calculated had we observed the transmission network **v**. Using the methods of Louis (1982), the observed information is

$$\tilde{I}\left(\tilde{\beta}\right)=\mathbb{E}_{\tilde{\beta},\tilde{\lambda}}\left[I_{\mathbf{v}}\left(\tilde{\beta}\right)\right]-\mathbb{E}_{\tilde{\beta},\tilde{\lambda}}\left[U_{\mathbf{v}}\left(\tilde{\beta},\infty\right)^{\otimes 2}\right], \quad (30)$$

where $\mathbb{E}_{\beta,\lambda}[\cdot]$ denotes an expectation taken under the assumption that the true coefficient vector is β and the true baseline hazard function is $\lambda(\tau)$. The first term in (30) is

$$-\sum_{j=1}^{n}\sum_{i\neq j}\int_0^{\tau}\frac{\partial^2}{\partial\beta^2}ln\frac{r\left(\tilde{\beta}^{\mathsf{T}}X_{ij}(u)\right)}{Y(\beta,u)}\mathrm{d}\tilde{N}_{ij}(u)\,, \quad (31)$$

where $\tilde{N}_{ij}(u)=\tilde{N}_{ij}\left(u|\tilde{\beta},\tilde{\lambda}\right)$. This is the observed information matrix from a weighted regression model where each *ij* has an uncensored copy with weight $p_{ij}\left(\tilde{\beta},\tilde{\lambda}\right)$ and a censored copy with weight 1 − $1-p_{ij}\left(\tilde{\beta},\tilde{\lambda}\right)$. To evaluate the second term in (30), let

$$\tilde{U}_{\cdot j}(\beta,\tau)=\sum_{i\neq j}\int_0^{\tau}\frac{\partial}{\partial\beta}ln\frac{r\left(\beta^{\mathsf{T}}X_{ij}(u)\right)}{Y(\beta,u)}\mathrm{d}\tilde{N}_{ij}(u)\,, \quad (32)$$

be the expected score contribution from all pairs with *j* as a susceptible. Since $\sum_{j=1}^{n}\widetilde{U}_{\cdot j}\left(\widetilde{\beta},\infty\right)=0$, each infected person *j* has only one infector in any **v**, and the infectors of different individuals can be chosen independently, we have

$$\mathbb{E}_{\widetilde{\beta},\widetilde{\lambda}}\left[U\left(\widetilde{\beta},\infty\right)^{\otimes 2}\right]=\sum_{j=1}^{n}\sum_{i\neq j}\int_{0}^{\infty}\left(\frac{\partial}{\partial\beta}ln\frac{r\left(\widetilde{\beta}^{\mathsf{T}}X_{ij}\left(u\right)\right)}{Y\left(\widetilde{\beta},u\right)}\right)^{\otimes 2}\mathrm{d}\widetilde{N}_{ij}\left(u\right)-\sum_{j=1}^{n}\widetilde{U}_{\cdot j}\left(\widetilde{\beta},\infty\right)^{\otimes 2} \quad (33)$$

2.8 Large-sample estimation of $\Lambda_0(\tau)$

Let $\widetilde{\Lambda}_0\left(\tau\right)$ be the marginal Breslow estimate obtained after convergence of the ECM algorithm. Appendix B.2, available online, derives the variance estimate

$$\widetilde{\sigma}_0^2\left(\tau\right)=\left(\frac{\partial}{\partial\beta}\widetilde{\Lambda}_{\widetilde{\beta},\widetilde{\lambda}}\left(\widetilde{\beta},\tau\right)\right)^{\mathsf{T}}\widetilde{I}\left(\widetilde{\beta}\right)^{-1}\left(\frac{\partial}{\partial\beta}\widetilde{\Lambda}_{\widetilde{\beta},\widetilde{\lambda}}\left(\widetilde{\beta},\tau\right)\right) \quad (34)$$

$$+2\int_{0}^{\tau}\frac{1}{Y\left(\widetilde{\beta},u\right)^{2}}\mathrm{d}\widetilde{N}\left(u\right)-\sum_{j=1}^{n}\left(\int_{0}^{\tau}\frac{1}{Y\left(\widetilde{\beta},u\right)}\mathrm{d}\widetilde{N}_{\cdot j}\left(u\right)\right)^{2}, \quad (35)$$

where $\tilde{N}_{\cdot j}(u) = \Sigma_{i\neq j}\,\tilde{N}_{\cdot j}(u)$. Using the martingale central limit theorem and a log transformation, we get the approximate pointwise $1-\alpha$ confidence limits

$$\widetilde{\Lambda}_0\left(\tau\right)exp\left(\pm\frac{\widetilde{\sigma}_0\left(\tau\right)}{\widetilde{\Lambda}_0\left(\tau\right)}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right). \quad (36)$$

Point and interval estimates for the baseline survival function can be obtained using the relationship $S_0(\tau) = \exp\left(-\Lambda_0(\tau)\right)$

3 SIMULATIONS

The performance of the methods from Section 2 was tested with a series of 12000 network-based epidemic simulations. All epidemics took place on a Watts-Strogatz small-world network (Watts and Strogatz, 1998), which mimics the high clustering and low diameter of real human contact networks. Starting with a ring of 50000 nodes, each node was connected to its 10 nearest neighbors and each edge was rewired to a randomly chosen node with probability 0.1. A new contact network was built for each simulation.

All epidemic models were written in Python 2.7 (www.python.org) using the packages NetworkX 1.6 (networkx.lanl.gov), NumPy 1.6, and SciPy 0.9 (www.scipy.org). Statistical analysis was done in in R 2.15 (www.r-project.org) via the Rpy2 2.2 package (rpy.sourceforge.net). The code for the models is available as Online Supplementary Information.

3.1 Transmission model

The transmission model had a latent period of zero and an exponential infectious period with mean one. The baseline contact interval distribution was Weibull(α, γ), where α is the shape parameter and γ is the rate parameter. 6000 simulations had a Weibull(0.5, 0.2) distribution, which has $\Lambda_0(\tau) = (0.2\tau)^{0.5}$. The other 6000 had a Weibull(2, 0.6) distribution, which has $\Lambda_0(\tau) = (0.6\tau)^2$. These distributions gave a basic reproduction number (expected number of infectious contacts made by a typical infectious person) $R_0 \approx 3$ in a null model.

In the transmission model, each person *i* had an infectiousness covariate X_i^{inf} and a susceptibility covariate X_i^{sus}. Each pair *ij* connected by an edge had a pairwise covariate X_{ij}^{pair}. All covariates were independent Bernoulli(.5) random variables. For a connected pair *ij*, the hazard of infectious contact from *i* to *j* at infectious age τ of *i* was

$$\lambda_{ij}(\tau) = exp\left(\beta_{inf}X_i^{inf} + \beta_{sus}X_j^{sus} + \beta_{pair}X_{ij}^{pair}\right)\lambda_0(\tau) \quad (37)$$

For each parameter β, there were 4000 simulations where its true value was chosen from a uniform distribution on (−1, 1). Of these, 2000 simulations used the Weibull(0.5, 0.2) baseline hazard and 2000 used the Weibull(2, 0.6) baseline hazard. Of the 2000 simulations for each baseline hazard, 1000 had the other two β set to 0 and 1000 had the other two β set to 1.

Each simulated epidemic began with a single person infected at time 0. Data from the next 1000 infections was used to fit two regression models, one using information on who-infected-whom as in Section 2.1 and one using an EM algorithm as in Section 2.5. The EM algorithm used a minimum of 2 and a maximum of 25 iterations. At each iteration, a weighted Cox model was run using the last parameter estimates as the initial parameter estimates. Convergence was defined as a change less than 0.002 in the expected log likelihood (tighter convergence criteria yielded nearly identical parameter estimates). After convergence, a Cox model was run using the final weights and initial parameters $\beta_{\text{inf}} = \beta_{\text{sus}} = \beta_{\text{pair}} = 0$.

After each simulation, we recorded the true value, estimate, and 95% confidence interval endpoints for each in the model and the baseline hazard at the 10th, 25th, 50th, 75th, and 90th percentiles of all censored and observed contact intervals. We also recorded the α and γ of the baseline hazard function and the number of EM iterations.

3.2 Results

Figure 3 shows good agreement between the estimated and true β_{inf}, β_{sus}, and β_{pair} for both $\hat{\beta}$ (who-infects-whom observed) and $\tilde{\beta}$ (who-infects-whom unobserved). Table 1 shows excellent 95% confidence interval coverage probabilities for all combinations of baseline hazards and parameters. The $\tilde{\beta}$ estimates had slightly lower coverage probabilities than the $\hat{\beta}$ estimates. The lower right panel of Figure 3 shows that this was achieved with relatively few iterations. The median number of iterations was 6, and 98% of simulations required ≤ 10 iterations. Only 2 out of 12000 simulations failed to converge within 25 iterations.

Figures 4 and 5 show good agreement between the estimated and true base-line hazard for both $\hat{\Lambda}_0(\tau)$ (who-infects-whom observed) and $\tilde{\Lambda}_0(\tau)$ (who-infects-whom unobserved). The smoothed means show almost no bias in $\hat{\Lambda}_0(\tau)$ or $\tilde{\Lambda}_0(\tau)$. Table 2 shows good 95% confidence interval coverage probabilites for both base-line hazards and all percentiles except for $\alpha = 2$ at the 10th and 25th percentiles. When $\alpha = 2$, the baseline hazard of infectious contact is $\lambda_0(\tau) = 1.2\tau$. At low values of τ, the hazard of infectious contact is very small, so almost all of the contact intervals will be censored. Since the percentiles are calculated for all censored and observed contact intervals, there may be too few observed intervals at the 10th and 25th percentiles when $\alpha = 2$ for the large-sample normal approximation to be valid.

Figure 6 shows the widths of the confidence intervals when who-infects-whom is not observed in terms of the width of the confidence interval when who-infects-whom is observed for β_{inf}, β_{sus}, β_{pair}, and $\Lambda_0(\tau)$. For β_{inf} and β_{pair}, the precision gained by observing who-infects-whom is roughly equivalent to a 20-40% increase in sample size. The baseline hazard plays an important role in how much precision is gained, with a larger gain for $\alpha = 0.5$ than for $\beta = 2$. There is no gain in precision for β_{sus} because observing who-infects-whom does not add to our knowledge of who was infected. Seeing who-infects-whom only slightly improves the precision of baseline hazard estimates.

Observing who-infects-whom allows point estimates that are closer to the truth and interval estimates with better coverage probabilities. However, the EM algorithm can recover a great deal of information when who-infects-whom is not observed, making the iterative regression model of Section 2.5 a promising tool for infectious disease epidemiology.

4 DATA ANALYSIS

To show how the methods of Section 2 can be applied, we will look at the effect of antiviral prophylaxis and age on the transmission of pandemic influenza A(H1N1) in Los Angeles County in 2009. The Los Angeles County Department of Public Health (LACDPH) collected household surveillance data between April 22 and May 19 according to the following protocol (Sugimoto et al., 2011):

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 the following information about his or her household contacts: age, gender, type of contact (household, intimate, in-home daycare, non-home daycare), and high risk status (pregnant, child on long-term aspirin therapy, immuno-suppressed, or history of a chronic cardiac, pulmonary, renal, liver, or neurologic condition). The interviewer also recorded whether prophylactic antiviral medication was being taken by

the 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.

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. For simplicity, we assume these were all influenza A(H1N1) cases and that all household members were susceptible to infection.

Our natural history assumptions were adapted from Yang et al. (2009) and are identical to those in Kenah (2013). 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. Under these assumptions, a person *j* with symptom onset at time t_j^{sym} was infected at time $t_j = t_j^{sym} - 2$ and will stop being infectious at time $t_j + 6 = t_j^{sym} + 4$. Under these assumptions, person *j* can transmit infection on days $t_j + 1$ to $t_j + 6$. In a sensitivity analysis, we vary the latent period from 0 to 1 days, and the infectious period from 5 to 7 days.

We modeled influenza transmission within households, not between households or from outside the observed households. In each household, infected household members who had no possible infector within the household according to our natural history assumptions were assumed to be imported infections. We assumed that any infected household member could infect any susceptible household member. We used the regression model of Section 2.5 to estimate influenza transmission hazard ratios for the following covariates:

- $\text{age}_{\text{inf}} = 0$ if the infectious person is < 18 years old and 1 otherwise,
- $\text{age}_{\text{sus}} = 0$ if the susceptible is < 18 years old and 1 otherwise,
- $\text{proph}_{\text{sus}} = 0$ if the susceptible is not on antiviral prophylaxis and 1 otherwise.

Since antiviral prophylaxis was initiated after the initial case in each household, it was considered only as a susceptibility covariate. All statistical analysis was done in R 2.15 (www.r-project.org).

4.1 Results

There were 114 people aged < 18 years and 185 aged ≥ 18 years, with no missing age data. There were 91 people taking antiviral prophylaxis and 152 not taking prophylaxis, with missing prophylaxis data for 56 people. When who-infects-whom is not observed, a complete-case analysis requires the removal of all rows corresponding to infectious-susceptible pairs *ij* where $i \in \mathscr{V}_j$ and any member of $\mathscr{V}_j$ is missing data. Otherwise, the remaining members of $\mathscr{V}_j$ get too much credit for the infection of *j*.

In the main analysis, there were 70 people infected from outside the household (i.e., no possible infector in the household), 16 with 1 possible infector, 7 with 2 possible infectors, 4 with 4 possible infectors, and 2 with 8 possible infectors, giving us $1^{16} \times 2^7 \times 4^4 \times 8^2 =$

2097152 possible transmission trees. The pairwise data contains 443 infectious-susceptible pairs with a total of 2455 pair-days at risk of infection. Of these, 16 × 1 + 7 × 2 + 4 × 4 + 2 × 8 = 62 rows represent possible infection events. All models used the Efron approximation (Efron, 1977) for the partial likelihood with tied failure times.

The top panel of Table 3 shows the results of seven models. All of the models including prophylaxis suggested that antiviral prophylaxis reduced the hazard of infectious contact by about 60%, with low p-values. Hazard ratio point estimates for the main effects of age in all models suggest that adults are more infectious and less susceptible than children. However, evidence for this result is very weak. Only one of the age effects was statistically significant in univariable models, and none were significant in any multivariable model. Multivariable and stratified models with interaction terms for age and antiviral prophylaxis suggest a stronger effect of antiviral prophylaxis on transmission to and from adults than on transmission to and from children. However, the evidence for this result is also weak; these coefficients had high p-values and wide confidence intervals. The bottom panel of Table 3 shows the results of a sensitivity analysis using the multivariable model without interaction. Varying the latent and infectious periods has little effect on the results of the model.

Figure 7 shows estimates of the cumulative transmission probability based on the multivariable and stratified models without interaction. The results of the two models are similar, but the stratified model showed lower probabilities of transmission from children and higher probabilities of transmission from adults. All four panels clearly show the estimated effect of antiviral prophylaxis. All curves show bigger jumps on the first four days after infection than on days 5 and 6, which is consistent with the results of Kenah (2013). Comparing the top and bottom rows shows that children are estimated to be less infectious than adults. Comparing the left and right columns shows that children are estimated to be more susceptible than adults. As noted above, these differences are not statistically significant.

This data analysis has been intended primarily to illustrate the flexibility of the regression modeling framework for analyzing transmission data. There are several important limitations of the analysis itself. With only 29 within-household transmissions, the large-sample normal approximations may not hold and there is limited power to estimate the effects of age and antiviral prophylaxis. The age classification is crude, so it may not accurately capture the effects of age. The prophylaxis variable was missing for many pairs and was binary, allowing no consideration of the timing of prophylaxis relative to exposure. Analyses of the household transmission of influenza A(H3N2) found greater child-to-child than adult-to-adult transmission (Addy et al., 1991). In our analysis of influenza A(H1N1), children appeared less infectious and more susceptible than adults, but these differences were not statistically significant. If not due to random noise, such a result could reflect a difference between the H3N2 and H1N1 subtypes of influenza A or a bias caused the failure to account for infection from outside the household. In any case, this analysis shows that the model needs to be extended to model infection from outside the household and to handle missing data.

5 DISCUSSION

The semiparametric relative-risk regression model proposed here has several important advantages over the chain binomial model. It can be fit using standard statistical software with parameter interpretations that resemble the Cox model. Standard software can be used to convert the results into curves representing the cumulative probability of transmission in pairs of individuals with specific characteristics. It does not make any parametric assumptions about the baseline hazard of infectious contact, and it allows many of the same extensions as the Cox model, including stratification, interaction, and time-dependent covariates. This flexibility and ease of use will make it an important tool for infectious disease epidemiology. To realize this potential, there are several limitations that remain to be addressed.

We assumed that the set of imported infections is known. The chain binomial model handles unknown imported infections by including a per-time-unit probability of escaping infection from outside the household. In the semipara-metric regression model, this could be achieved by fitting two models in each step of the EM algorithm: a pairwise contact interval model within the household and an individual-level model in absolute time for infection from outside the household. At each step, the weights would be recalculated based on covariates, coefficient estimates, the baseline hazard of the contact interval distribution, and the baseline hazard of infection from outside the household.

We assumed that infection times, latent periods, and infectious periods were all observed. We can usually observe only the clinical course of the disease, so these times must be imputed. In Section 4, we had missing data on covariates. Simple missing data (such as antiviral prophylaxis) could be handled by extending the EM algorithm to calculate the expected log likelihood over the possible values of the missing data as well as who-infected-whom. More complex missing data (such as infection and removal times) could be handled using data augmentation in a profile sampler (Lee et al., 2005), getting a posterior distribution for the model coefficients while treating the baseline hazards as a nuisance parameter.

We assumed that all possible infectors of each person were observed. Unobserved infectors could occur because of incomplete contact tracing or asymptomatic infection. The possible bias caused by unobserved sources of infection needs to be studied, and methods for controlling it analytically or assessing its severity in a sensitivity analysis need to be developed.

We assumed a static contact network where the C_{ij} were binary and constant. In reality, people are exposed to close contacts at home, at work, at school, and at other locations in a dynamic process. The extension of these methods to dynamic contact networks is possible but nontrivial. We could allow $C_{ij}(\tau)$ to be a time-dependent process in the infectious age τ of i. The contact interval distribution would then be defined as the distribution of the contact interval that would occur if $C_{ij}(\tau) = 1$ for all τ. For estimation, we would have to observe the process $C_{ij}(\tau)$ for each ij.

Some of our assumptions must be relaxed to capture the natural history of complex diseases. We assumed an SEIR framework best suited to acute, immunizing diseases that spread directly from person to person. Many foodborne and waterborne diseases, pneumococcal and meningococcal diseases, and other infectious diseases of major public health importance do not fit easily into this framework. To extend the proposed regression model to complex diseases, we could allow individuals to experience multiple events (e.g., first infection, second infection) or to experience different types of events (e.g., colonization, infection, relapse). We assumed that contact intervals were independent of infectious periods even though both are affected by the same host-pathogen interaction. In some cases, there may be a covariate process $X(\tau)$ such that $I_i(\tau)$ and $\mathcal{N}_{ij}(\tau)$ are conditionally independent given $X(\tau^-)$. Otherwise, infectious contact and the infectious period must be modeled as a multivariate survival process.

Several technical issues need further attention. The smoothing step is crucial to the fitting the regression model when who-infects-whom is not observed. Here, we used cubic smoothing splines because they were convenient and worked well. However, these do not guarantee that the smoothed hazard function is monotonically increasing and lack a convenient interpretation in terms of the likelihood. A penalized likelihood estimator that guarantees monotonicity, such as that of Anderson and Senthilselvan (1980), would be more consistent with the EM algorithm. Model diagnostics, goodness-of-fit tests, and small-sample methods for point and interval estimation need to be developed, and a more rigorous study of the model asymptotics needs to be done.

Despite these limitations, the semiparametric relative-risk regression model presented here is a powerful new framework for the analysis of infectious disease transmission data. Placing statistical methods for infectious disease epidemiology on the broad and deep theoretical foundation of survival analysis will help clarify study design and causal inference for communicable diseases and allow statistical methods to develop in concert with advances in molecular biology. Ultimately, these improvements may lead to more efficient vaccine trials and a better-informed public health response to future outbreaks and epidemics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by National Institute of Allergy and Infectious Diseases (NIAID) grant K99/R00 AI095302. The content is solely the responsibility of the author and does not necessarily represent the official views of NIAID or the National Institutes of Health. The author thanks Tom Britton, Yang Yang, M. Elizabeth Halloran, and Ira M. Longini, Jr. for their comments and suggestions.

References

Aalen, Odd O.; Borgan, Ørnulf; Gjessing, Hakon. Statistics for Biology and Health. Springer-Verlag; New York: 2009. Survival and Event History Analysis: A Process Point of View..

Addy, Cheryl L.; Longini, Ira M., Jr; Haber, Michael. A generalized stochastic model for the analysis of infectious disease final size data. Biometrics. 1991; 47:961–974. [PubMed: 1742449]

Anderson JA, Senthilselvan A. Smooth estimates for the hazard function. Journal of the Royal Statistical Society, Series B. 1980; 42:322–327.

Andersson, Håkan; Britton, Tom. Lecture Notes in Statistics. Springer; New York: 2000. Stochastic Epidemic Models and Their Statistical Analysis..

Becker, Niels G. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC; Boca Raton, FL: 1989. Analysis of Infectious Disease Data..

Breslow N. Cox DR. Contribution to discussion of paper. Journal of the Royal Statistical Society B. 1972; 34:216–217.

Cox, David R. Regression models and life-tables. Journal of the Royal Statistical Society, Series B. 1972; 34:187–220.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from in complete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39:1–38.

Efron, Bradley. The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association. 1977; 72:557–565.

Soren, Johansen. An extension of Cox's regression model. International Statistical Review. 1983; 51:165–174.

Kenah, Eben. Contact intervals, survival analysis of epidemic data, and estimation of R_0. Biostatistics. 2011; 12:548–566. [PubMed: 21071607]

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

Kenah, Eben; Lipsitch, Marc; Robins, James M. Generation interval contraction and epidemic data analysis. Mathematical Biosciences. 2008; 213:71–79. [PubMed: 18394654]

Leng Lee, Bee; Kosorok, Michael R.; Fine, Jason P. The profile sampler. Journal of the American Statistical Association. 2005; 100:960–969.

Louis, Thomas A. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B. 1982; 44:226–233.

Meng, Xiao-Li; Rubin, Donald. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80:267–278.

Prentice, Ross L.; Self, Steven G. Asymptotic distribution theory for Cox-type regression models with general relative risk form. Annals of Statistics. 1983; 11:804–813.

Rampey, Alvin H., Jr; Longini, Ira M., Jr; Haber, Michael; Monto, Arnold S. A discrete-time model for the statistical analysis of infectious disease incidence data. Biometrics. 1992; 48:117–128. [PubMed: 1316178]

Rhodes, Philip H.; Elizabeth Halloran, M.; Longini, Ira M., Jr. Counting process models for infectious disease data: Distinguishing exposure to infection from susceptibility. Journal of the Royal Statistical Society B. 1996; 58:751–762.

Sugimoto, Jonathan D.; Yang, Yang; Elizabeth Halloran, M.; Dean, Brandon; Oiulfstad, Brit; Ann Bagwell, Dee; Mascola, Laurene; Bancroft, Elizabeth; Longini, Ira M., Jr. Accounting for unobserved immunity and asymptomatic infection in the early household transmission of the pandemic influenza A (H1N1) 2009. Submitted to American Journal of Epidemiology. 2011

Svensson, Åke. A note on generation times in epidemic models. Mathematical Biosciences. 2007; 208:300–311. [PubMed: 17174352]

Therneau, Terry M.; Grambsch, Patricia M. Statistics for Biology and Health. Springer-Verlag; New York: 2000. Modeling Survival Data: Extending the Cox Model..

Wallinga, Jacco; Teunis, Peter. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology. 2004; 160:509–516. [PubMed: 15353409]

Watts, Duncan J.; Strogatz, Steven H. Collective dynamics of ‘small-world’ networks. Nature. 1998; 393:440–442. [PubMed: 9623998]

Forsberg White L, Pagano Marcello. A likelihood-based method for real-time estimate of the serial interval and reproductive number of an epidemic. Statistics in Medicine. 2008; 27:2999–3016. [PubMed: 18058829]

Yang, Yang; Sugimoto, Jonathan; Elizabeth Halloran, M.; Basta, Nicole E.; Chao, Dennis L.; Matrajt, Laura; Potter, Gail; Kenah, Eben; Longini, Ira M., Jr. The transmissibility and control of pandemic influenza A(H1N1) virus. Science. 2009; 326:729–733. [PubMed: 19745114]

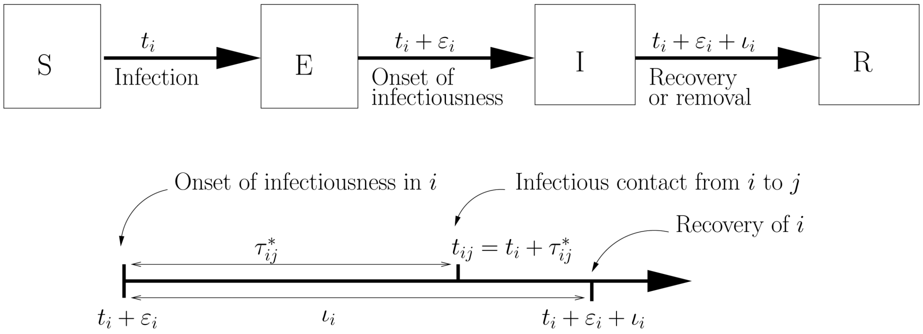
S
t_i
Infection
E
$t_i + \varepsilon_i$
Onset of infectiousness
I
$t_i + \varepsilon_i + \iota_i$
Recovery or removal
R
Onset of infectiousness in i
Infectious contact from i to j
Recovery of i
τ^*_{ij}
$t_{ij} = t_i + \tau^*_{ij}$
$t_i + \varepsilon_i$
ι_i
$t_i + \varepsilon_i + \iota_i$

Figure 1.

Notation for the stochastic SEIR model natural history (top) and infectious contact process (bottom). In the bottom diagram, the infectious contact interval τ^*_{ij} is equal to the contact interval τ_{ij} because $\tau_{ij} \le l_i$. Otherwise, we would have $\tau^*_{ij}=\infty$ and no infectious contact from *i* to *j* would occur.

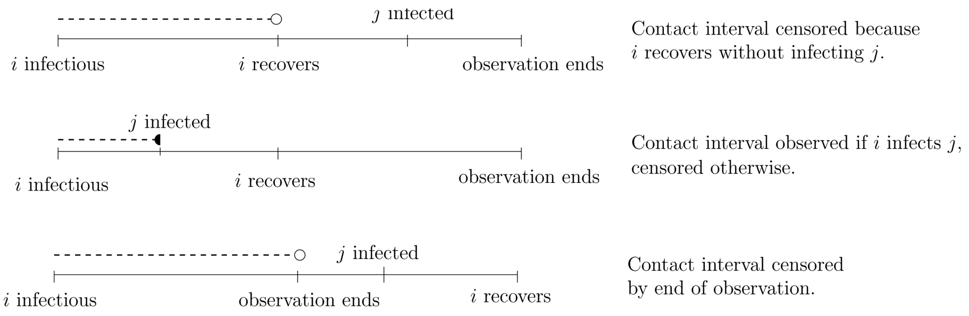

Figure 2.

The three censoring processes for the contact interval τ_{ij}. The onset of infectiousness in *i* occurs at time $t_i + \varepsilon_i$, and the infection of *j* occurs at t_j. At the top, τ_{ij} is censored because $\mathscr{I}_i(t_j - t_i - \varepsilon_i) = 0$. In the middle, τ_{ij} is observed if $\mathscr{S}_{ij}(t_j - t_i - \varepsilon_i) = 1$ and censored otherwise. At the bottom, τ_{ij} is censored because $\mathscr{Y}_i(t_j - t_i - \varepsilon_i) = 0$.

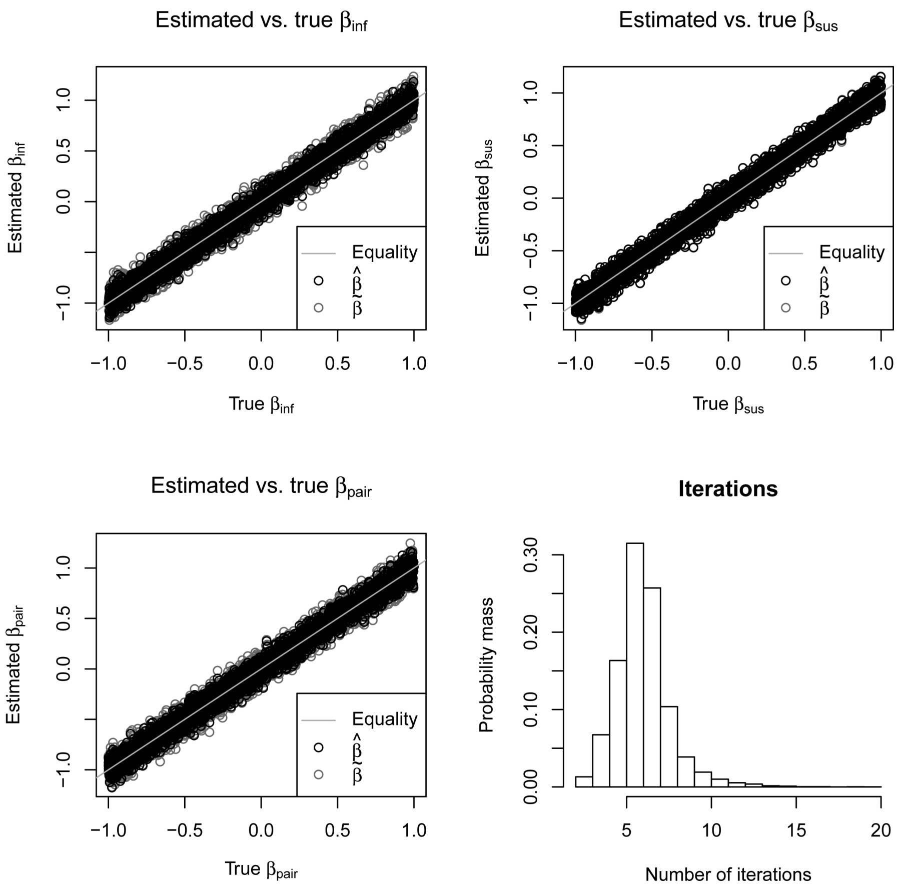

Figure 3.

The top two panels and the bottom left panel show $\hat{\beta}$ (black circles) and $\widetilde{\beta}$ (gray circles) versus true β for β_{inf}, β_{sus}, and β_{pair}. The bottom right panel shows a histogram of the number of EM iterations required for convergence.

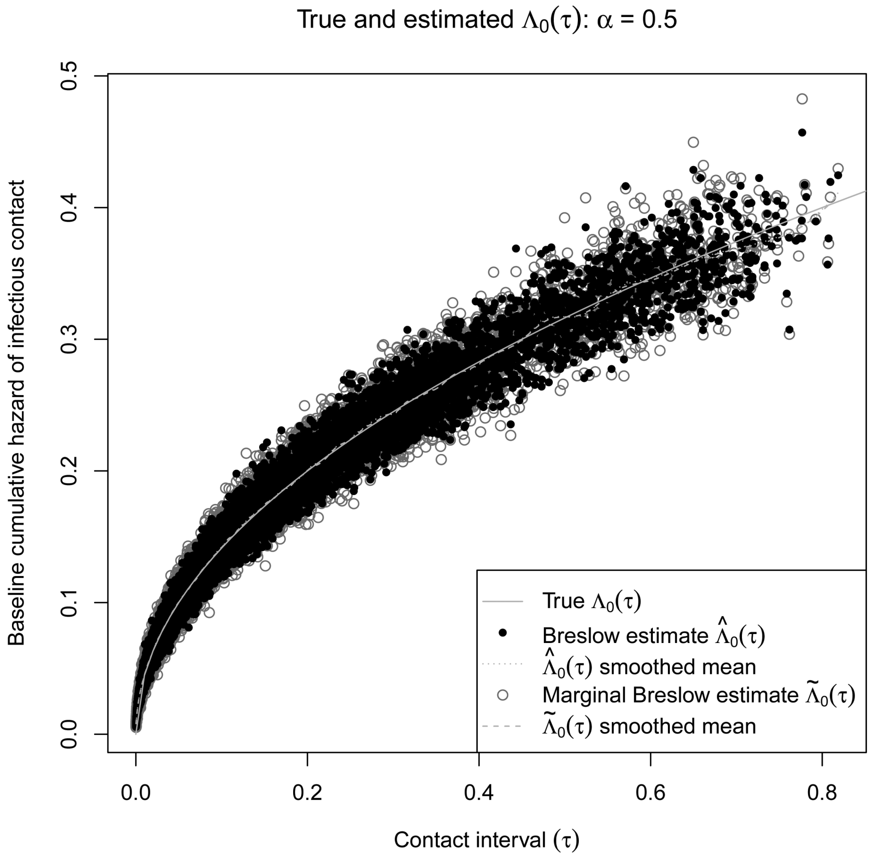

Figure 4.

$\hat{\Lambda}_0(\tau)$ and $\tilde{\Lambda}_0(\tau)$ versus true $\Lambda_0(\tau)$ for the 6000 simulations with a Weibull(0.5, 0.2) baseline contact interval distribution. For each simulation and each estimator, a circle is shown for the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of all possible contact intervals. The smoothed means were calculated using cubic smoothing splines.

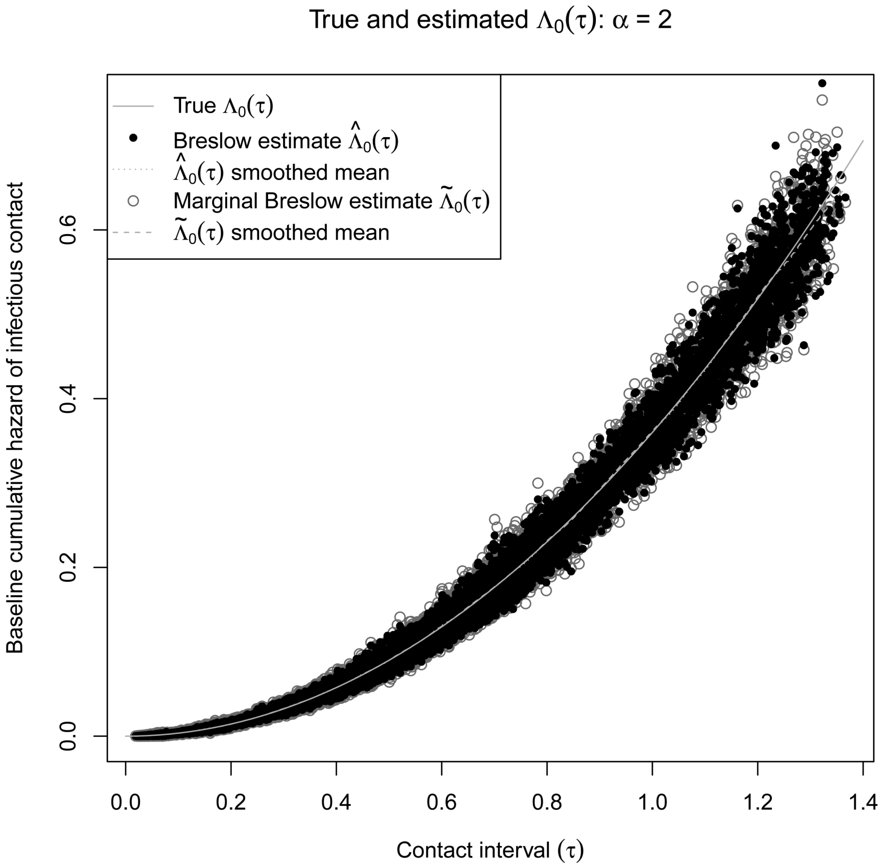

Figure 5.

$\hat{\Lambda}_0(\tau)$ and $\tilde{\Lambda}_0(\tau)$ versus true $\Lambda_0(\tau)$ for the 6000 simulations with a Weibull(2, 0.6) baseline contact interval distribution. For each simulation and each estimator, a circle is shown for the 10th, 25th, 50th, 75th, and 90th percentiles of all possible contact intervals. The smoothed means were calculated using cubic smoothing splines.

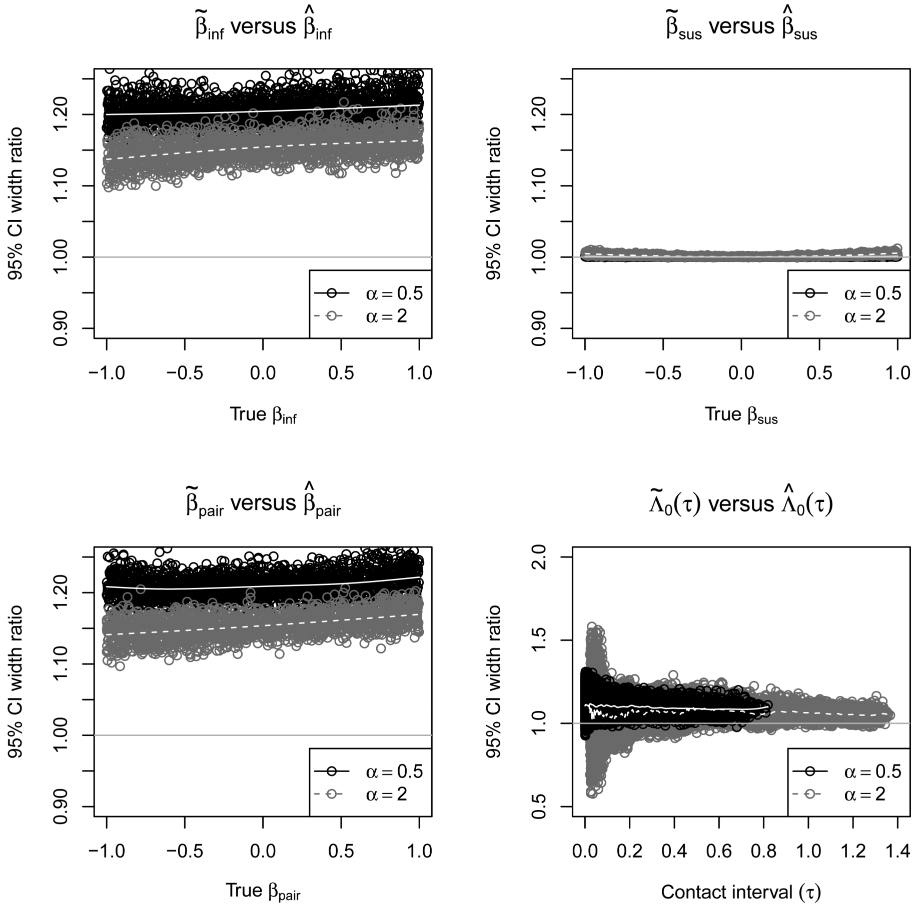

Figure 6.

The width of 95% confidence intervals when who-infects-whom is not observed divided by its width when who-infects-whom is observed for β_{inf}, β_{sus}, β_{pair}, and $\Lambda_0(\tau)$. The solid gray lines show smoothed means for $\alpha = 0.5$ and dashed gray lines show smoothed means for $\alpha = 2$. The smoothed means were calculated using cubic smoothing splines.

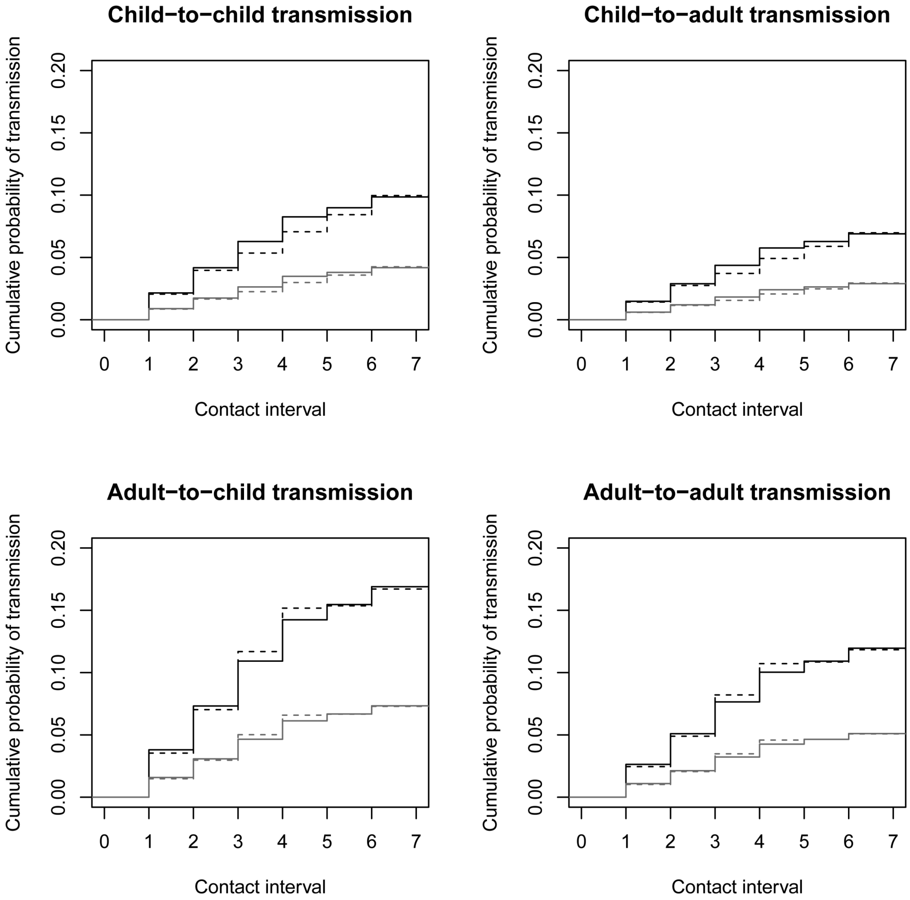

Figure 7.

Household transmission of 2009 pandemic influenza A(H1N1) in Los Angeles County. Each panel shows separate curves for susceptible contacts with (gray lines) and without (black lines) antiviral prophylaxis. The solid lines are based on the multivariable model, and the dotted lines are based on the model stratified by age$_{\text{inf}}$.

Table 1

Hazard ratio 95% confidence interval coverage probabilities in simulations. Each probability is based on the results of 1000 simulations.

Baseline hazard	Parameter: β_{inf}			
	$\beta_{\text{sus}}=\beta_{\text{pair}}=0$		$\beta_{\text{sus}}=\beta_{\text{pair}}=1$	
	β_{inf}	β_{inf}	β_{inf}	β_{inf}
$\alpha = .5$	.952	.937	.937	.945
$\alpha = 2$	.955	.952	.957	.941
Baseline hazard	Parameter: β_{sus}			
	$\beta_{\text{inf}} = \beta_{\text{pair}} = 0$		$\beta_{\text{inf}} = \beta_{\text{pair}} = 1$	
	β_{sus}	β_{sus}	β_{sus}	β_{sus}
$\alpha = .5$	.951	.950	.939	.939
$\alpha=2$	.952	.948	.945	.946
Baseline hazard	Parameter: β_{pair}			
	$\beta_{\text{inf}} = \beta_{\text{sus}} = 0$		$\beta_{\text{inf}} = \beta_{\text{sus}} = 0$	
	β_{pair}	β_{pair}	β_{pair}	β_{pair}
$\alpha = .5$	.942	.927	.955	.946
$\alpha=2$	.942	.929	.951	.951

Table 2

95% confidence interval coverage probabilities in simulations. Each probability is based on the results of 6000 simulations.

Baseline hazard Quantile	$\alpha = .5$		α=2	
	$\hat{\Lambda}_0(\tau)$	$\tilde{\Lambda}0(\tau)$	$\hat{\Lambda}_0(\tau)$	$\tilde{\Lambda}0(\tau)$
10%	.949	.938	.957	.875
25%	.949	.940	.951	.905
50%	.950	.941	.955	.934
75%	.949	.936	.949	.939
90%	.949	.941	.953	.939

Table 3

Hazard ratios with 95% confidence intervals and p-values for different models of the 2009 pandemic influenza A(H1N1) household surveillance data from Los Angeles County. Likelihood ratio p-values comparing models with and without interaction terms are also given. The multivariable and stratified models without interaction were used as the final models.

	Main effects			**Interactions**	
	age_{inf}	**age_{sus}**	**$\text{prophy}_{\text{sus}}$**	**Variables**	**HR (p-value)**
Regression model					
Univariable	1.53 (0.66, 3.54) p = .321	0.41 (0.20, 0.85) p = .016	0.43 (0.18, 1.02) p = . 057		
Multivariable	1.78 (0.69, 4.62) p = .234	0.69 (0.29, 1.64) p= .399	0.41 (0.17, 0.98) p = . 046		
Multivariable + interaction	1.59 (0.32, 7.84) p = .570	0.63 (0.14, 2.73) p= .532	0.04 (0.00, 9.62) p = . 253	age_{inf}:age_{sus} age_{inf} :$\text{proph}_{\text{sus}}$ age_{sus} :$\text{proph}_{\text{sus}}$	0.66 (p = .71) 9.28 (p= .45) 2.72 (p = .36) LR p = .101
Stratified	strata	0.69 (0.29, 1.64) p = .401	0.41 (0.17, 0.99) p = . 047		
Stratified + interaction	strata	0.52 (0.29, 1.64) p = .219	0.23 (0.05, 1.16) p = . 075	age_{inf} :proph*sus*	2.37 (p = .38) LR p = .353
Sensitivity analysis (multivariable model without interaction)					
Latent period 1 day	1.44 (0.64, 3.26) p = .378	0.83 (0.36, 1.93) p = .670	0.35 (0.15, 0.80) p = .013		
Infectious period 5 days	1.59 (0.60, 4.20) p = .348	0.64 (0.27, 1.55) p = .322	0.45 (0.18, 1.07) p = .073		
7 days	1.45 (0.62, 3.40) p = .378	0.89 (0.38, 2.04) p = .670	0.34 (0.17, 0.87) p = .013		