Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2013 March ; 75(2): 277–303. doi:10.1111/j.1467-9868.2012.01042.x.

Nonparametric survival analysis of infectious disease data

Eben Kenah

Department of Biostatistics and Emerging Pathogens Institute, University of Florida, Gainesville, Florida, USA

Summary

This paper develops nonparametric methods based on contact intervals for the analysis of infectious disease data. The contact interval from person i to person j is the time between the onset of infectiousness in i and infectious contact from i to j , where we define infectious contact as a contact sufficient to infect a susceptible individual. The hazard function of the contact interval distribution equals the hazard of infectious contact from i to j , so it provides a summary of the evolution of infectiousness over time. When who-infects-whom is observed, the Nelson-Aalen estimator produces an unbiased estimate of the cumulative hazard function of the contact interval distribution. When who-infects-whom is not observed, we use an EM algorithm to average the Nelson-Aalen estimates from all possible combinations of who-infected-whom consistent with the observed data. This converges to a nonparametric maximum likelihood estimate of the cumulative hazard function that we call the marginal Nelson-Aalen estimate. We study the behavior of these methods in simulations and use them to analyze household surveillance data from the 2009 influenza A(H1N1) pandemic.

Keywords

Chain-binomial models; Contact intervals; Generation intervals; Infectious disease; Nonparametric methods; Survival analysis

1. Introduction

Infectious diseases remain one of the greatest threats to human health and commerce, and the analysis of infectious disease data is among the most important but neglected applications of statistics. Current statistical methods for infectious disease data fall into two categories: chain-binomial models and generation interval models. In a chain binomial model, the transmission of disease is modeled as a discrete-time process. Methods based on generation intervals use continuous time, but they model the transmission of disease as a process that creates a population of infected individuals rather than a process of spread through an existing susceptible population. In this paper, we propose nonparametric statistical methods that model the spread of disease through an existing population in continuous time. These methods are a nonparametric counterpart to those of Kenah (2011).

If person i infects person j , the generation interval is the time between the infection of i and the infection of j . The serial interval, which is often used as a proxy for the generation interval, is the time between symptom onset in i and symptom onset in j . Data from several recent and historical epidemics, including the 1918 influenza (Mills et al., 2004), severe acute respiratory syndrome (Lipsitch et al., 2003; Wallinga and Teunis, 2004), pandemic influenza (Fraser et al., 2009; McBryde et al., 2009; Yang et al., 2009), and avian influenza (Ferguson et al., 2005, 2006), have been analyzed using methods based on generation or serial interval distributions. These methods are based on the Lotka-Euler equation (Wallinga and Lipsitch, 2007; Roberts and Heesterbeek, 2007) or on a branching-process approximation to the early spread of disease (Wallinga and Teunis, 2004; White and Pagano, 2008). In both of these approaches, infected

people are the offspring of their infectors and the generation interval is the age at which the infector gives birth. Uninfected people and person-time are accounted for indirectly (in the Lotka-Euler equation) or not at all (in the branching process approximation).

Although generation and serial interval distributions are often considered characteristic features of an infectious disease (Fine, 2003), these distributions change systematically over the course of an epidemic as susceptibles are exposed to an increasing number of infectives (Svensson, 2007; Kenah et al., 2008). Given that i infected j , i must have infected j before anyone else. Generation and serial intervals contract because of this competition among infectives to infect the remaining susceptibles. When transmission is rapid within households or schools, generation and serial interval contraction can occur even when the global prevalence of infection is low. Statistical methods based on generation or serial interval distributions fail to account for competing risks of infection from multiple sources.

If there is no rapid transmission within small groups, the generation interval distribution is stable very early in an epidemic, when most new infections have been exposed to only one infective. However, the assumption that generation intervals are independent and identically-distributed (iid) severely constrains the underlying epidemic model. Consider a model with only two people, i and j . Upon infection, there is a constant hazard λ of infectious contact. The infectious period is t_1 with probability p or $t_2 \neq t_1$ with probability $q = 1 - p$. If person i is infected and person j is susceptible, the generation interval X has the probability density function (pdf)

$$g(x) = p \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda t_1}} \mathbf{1}_{x \in (0, t_1]} + q \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda t_2}} \mathbf{1}_{x \in (0, t_2]},$$

where $\mathbf{1}_S$ equals one if S is true and zero otherwise. Now suppose i can transmit infection to a second susceptible k , and assume that transmission between j and k cannot occur (to avoid generation interval contraction). The generation interval Y from i to k has the pdf $g(y)$. However, the joint pdf of X and Y is

$$g(x, y) = p \frac{\lambda^2 e^{-\lambda(x+y)}}{(1 - e^{-\lambda t_1})^2} \mathbf{1}_{x, y \in (0, t_1]} + q \frac{\lambda^2 e^{-\lambda(x+y)}}{(1 - e^{-\lambda t_2})^2} \mathbf{1}_{x, y \in (0, t_2]}$$

because the infectious period of i is either t_1 or t_2 , not both. X and Y are independent if and only if $p = 0$ or $q = 0$. A similar result holds for the latent period from infection to the onset of infectiousness. Suppose i has infectious period τ and becomes infectious immediately with probability p or after a latent period $\varepsilon > 0$ with probability $q = 1 - p$. Then X and Y each have the pdf

$$g(x) = p \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda \tau}} \mathbf{1}_{x \in (0, \tau]} + q \frac{\lambda e^{-\lambda(x-\varepsilon)}}{1 - e^{-\lambda \tau}} \mathbf{1}_{x \in (\varepsilon, \varepsilon + \tau]},$$

but their joint pdf is

$$g(x, y) = p \frac{\lambda^2 e^{-\lambda(x+y)}}{(1 - e^{-\lambda \tau})^2} \mathbf{1}_{x, y \in (0, \tau]} + q \frac{\lambda^2 e^{-\lambda(x+y-2\varepsilon)}}{(1 - e^{-\lambda \tau})^2} \mathbf{1}_{x, y \in (\varepsilon, \varepsilon + \tau]}.$$

Again, X and Y are independent if and only if $p = 0$ or $q = 0$. In general, statistical methods based on iid generation intervals implicitly require constant latent and infectious periods. When serial intervals are used as a proxy for generation intervals, a constant incubation period (from infection to symptom onset) is also required.

Much of the trouble with generation and serial intervals is caused by conditioning on the transmission of infection. Kenah (2011) outlined an alternative analysis of epidemic data based on contact intervals. The contact interval τ_{ij} from an infectious person i to a susceptible person j is the time from the onset of infectiousness in i to the first infectious contact from i to j , where we define infectious contact as a contact sufficient to infect a susceptible individual. It will be right-censored if j is infected by someone other than i or if i stops being infectious before making infectious contact with j . This censoring is handled using methods adapted from survival analysis. The distribution of τ_{ij} provides a concise summary of the evolution of infectiousness over time in person i because its hazard function equals the hazard of infectious contact from i to j while i is infectious. Likelihoods written in terms of contact intervals improve on likelihoods based on generation or serial intervals by using information contributed by uninfected person-time and accounting for exposure to multiple infectives. They improve on chain-binomial models by allowing the use of continuous time, which more accurately reflects the data-generating mechanism underlying infectious disease data.

This paper outlines the nonparametric survival analysis of epidemic data based on contact intervals. The rest of Section 1 defines the general stochastic Susceptible-Exposed-Infectious-Removed (SEIR) model, outlines assumptions about the censoring and observation processes, and discusses the application of these methods to Susceptible-Infectious-Removed (SIR) models. Section 2 outlines the Nelson-Aalen and marginal Nelson-Aalen estimators for infectious disease data. Section 2.1 shows that the Nelson-Aalen estimator produces an unbiased estimate of the cumulative hazard function of the contact interval distribution when who-infects-whom is observed. Section 2.2 describes an expectation-maximization (EM) algorithm that can produce a nonparametric maximum likelihood estimate (MLE) of this cumulative hazard function when who-infects-whom is not observed. Section 3 explores the performance of these estimators in simulations, and Section 4 uses them to analyze household surveillance data from the 2009 in uenza A(H1N1) pandemic. Section 5 discusses the limitations, advantages, and future development of these methods.

1.1. Stochastic SEIR model

Consider a closed population of n individuals assigned indices $1, \dots, n$. Each individual is in one of four possible 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 i 's infectiousness onset time $t_i + \varepsilon_i$, i moves from E to I, beginning an infectious period of length ι_i . At his or her removal time $t_i + \varepsilon_i + \iota_i$, i moves from I to R. Once in R, i can no longer infect others or be infected. The latent period is a nonnegative random variable, the infectious period is a strictly positive random variable, and both have finite mean and variance.

After becoming infectious at time $t_i + \varepsilon_i$, person i makes infectious contact with j 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. We define infectious contact to be sufficient to cause infection in a susceptible person, so $t_j \leq t_{ij}$. Since infectious contact must occur while i is infectious or never, $\tau_{ij}^* = \infty$.

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 that the infectious contact interval τ_{ij}^* is generated as follows: A contact interval τ_{ij} is drawn from a distribution with hazard function $\lambda_{ij}(\tau)$. If $\tau_{ij} \leq \iota_i$ and $C_{ij} = 1$, then $\tau_{ij}^* = \tau_{ij}$. Otherwise, $\tau_{ij}^* = \infty$. In this paper, we assume that the contact intervals in all ordered pairs ij are independent and identically distributed (iid) with finite mean and variance.

An epidemic begins with one or more persons infected from outside the population, which we call imported infections. Imported infections can also occur while an epidemic is ongoing. To avoid modeling the process of infection from outside the population, we assume the set of imported infections is known.

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 $v_j = \infty$ if j is not infected before the end of observation. The transmission network is the directed network with an edge from v_j to j for each infected j . It can be represented by a vector $\mathbf{v} = (v_1, \dots, v_n)$. When who-infects-whom is not observed, let $\mathcal{V}_j$ denote the set of all i who might have infected j , which we call the infectious set of j . Let $\mathcal{V}$ denote the set of all possible $\mathbf{v}$ consistent with the observed data. A $\mathbf{v} \in \mathcal{V}$ can be generated by choosing a $v_j \in \mathcal{V}_j$ for each infected j .

1.2. Censoring and observation

Throughout this paper, we assume that we can observe τ_{ij} only if j is infected by i at time $t_i + \epsilon_i + \tau_{ij}$. C_{ij} indicates whether infectious contact from i to j is possible, so τ_{ij} can be observed only if $C_{ij} = 1$. Three processes can right-censor τ_{ij} :

- Let $\mathcal{I}_i(\tau) = 1_{\tau \in (0, \iota_i]}$ indicate whether i remains infectious at infectiousness age τ . Then i makes infectious contact with j at infectiousness age τ_{ij} only if $\mathcal{I}_i(\tau_{ij}) = 1$. Since ι_i is independent of τ_{ij} , $\mathcal{I}_i(\tau)$ independently censors τ_{ij} .
- Let $\mathcal{S}_{ij}(\tau) = 1_{t_i + \epsilon_i + \tau \leq t_j}$ indicate whether j remains susceptible when i reaches infectiousness age τ . Then j can be infected by i at infectiousness age τ_{ij} only if $\mathcal{S}_{ij}(\tau_{ij}) = 1$. Since τ_{ij} is unaffected by infectious contact with j by infectives other than i , $\mathcal{S}_{ij}(\tau)$ independently censors τ_{ij} .
- Assume that infection in j can be observed until a time T_j chosen independently of τ_{ij} , and let $\mathcal{Y}_{ij}(\tau) = 1_{t_i + \epsilon_i + \tau \leq T_j}$. Then infectious contact from i to j at infectiousness age τ_{ij} can be observed only if $\mathcal{Y}_{ij}(\tau) = 1$, and $\mathcal{Y}_{ij}(\tau)$ independently censors τ_{ij} .

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

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

is a left-continuous process that indicates the risk of an observed infectious contact from i to j . Since $\mathcal{I}_i(\tau)$, $\mathcal{S}_{ij}(\tau)$, and $\mathcal{Y}_{ij}(\tau)$ independently censor τ_{ij} , so does $Y_{ij}(\tau)$. Some independence assumptions made for $\mathcal{I}_i(\tau)$, $\mathcal{S}_{ij}(\tau)$, and $\mathcal{Y}_{ij}(\tau)$ can be relaxed without destroying the independent censoring of τ_{ij} ; see Appendix A.

1.3. SIR models

The purpose of the latent period in SEIR model is to allow a random waiting time before the beginning of infectiousness. Biologically, this corresponds to a period during which the infection is establishing itself in its host but unable to jump to another host. The methods of

this paper also apply to Susceptible-Infectious-Removed (SIR) models, which have no latent period. The equations in this paper can be adapted to an SIR model by setting $\varepsilon_i = 0$ for all i . In particular, the infectiousness age of i at time t becomes $t - t_i$.

An SEIR model with a fixed latent period ε can be rewritten as an SIR model with $\lambda_0(\tau) = 0$ for $\tau \in (0, \varepsilon]$. If the latent period is assumed to be fixed but unknown, using an SIR model instead of an SEIR model removes the need to observe the latent period in infected individuals. When the latent period varies between individuals, the SEIR model cannot be rewritten as an SIR model with a single function $\lambda_0(\tau)$ describing the evolution of infectiousness over time. In this sense, the SEIR model is more general than the SIR model. In practice, SIR models may be biologically plausible for many infectious diseases.

2. Methods

In this section, we extend the Nelson-Aalen estimator (Altshuler, 1970; Nelson, 1972; Aalen, 1978) to get two nonparametric estimators of the cumulative hazard function of the contact interval distribution. We assume that the contact interval distribution is continuous with the hazard function $\lambda_0(\tau)$ and the cumulative hazard function

$$\Lambda_0(\tau) = \int_0^\tau \lambda_0(u) du. \quad (2)$$

We call τ the infectiousness age. By the definition of the contact interval, $\lambda_0(\tau)$ is the hazard of infectious contact from an infective i at infectiousness age τ to any given j such that $C_{ij} = 1$. The hazard of infectious contact is a familiar idea in infectious disease epidemiology. The contact interval distribution is simply the distribution defined by this hazard function. This change of perspective makes survival analysis a natural framework for the analysis of infectious disease data.

To derive parametric estimators of the contact interval distribution, Kenah (2011) used counting processes and martingales defined in absolute time. To derive nonparametric estimators, we use counting processes and martingales defined in infectiousness age. Whereas absolute-time processes are defined for all ordered pairs ij , infectiousness-age processes are defined only for those ij in which $t_i < \infty$. This change of time scale is allowed because the distribution of the contact interval τ_{ij} does not depend on the infection time t_i of i .

In Section 2.1, we show that the Nelson-Aalen estimator generates an unbiased estimate of $\Lambda_0(\tau)$ when who-infects-whom is observed. When who-infects-whom is not observed, we average the Nelson-Aalen estimates from all $\mathbf{v} \in \mathcal{V}$ to get a marginal Nelson-Aalen estimator, which we describe in Section 2.2. Since the probability of each $\mathbf{v} \in \mathcal{V}$ is unknown, we use an EM algorithm (Dempster et al., 1977) to iteratively reweight each possible $\mathbf{v}$, obtaining a sequence of marginal Nelson-Aalen estimates that converges to a nonparametric MLE of $\Lambda_0(\tau)$. This method adapts easily to a situation where who-infects-whom is partially observed. In Section 2.3, we derive approximate estimators for mass-action SEIR models.

2.1. Nonparametric estimation when who-infects-whom is observed

First, suppose there is no censoring, so we observe the contact interval in each ordered pair ij after the onset of infectiousness in i . Then

$$\mathcal{N}_{ij}(\tau) = 1_{\tau \geq \tau_{ij}} \quad (3)$$

is a counting process with a jump of size one at τ_{ij} , and

$$\mathcal{M}_{ij}(\tau) = \mathcal{N}_{ij}(\tau) - \int_0^\tau \lambda_0(u) du \quad (4)$$

is a zero-mean martingale. In the presence of censoring, we have the observed counting process

$$N_{ij}(\tau) = \int_0^\tau Y_{ij}(u) d\mathcal{N}_{ij}(u) \quad (5)$$

and the counting-process martingale

$$M_{ij}(\tau) = \int_0^\tau Y_{ij}(u) d\mathcal{M}_{ij}(u). \quad (6)$$

The combined observations of all ij for a fixed j give us the counting process

$$N_{\cdot j}(\tau) = \sum_{i \neq j} N_{ij}(\tau), \quad (7)$$

which has a jump of size one at $\tau = t_j - t_{vj} - \varepsilon_{vj}$ if j is infected. Since it is a sum of zero-mean martingales,

$$M_{\cdot j}(\tau) = \sum_{i \neq j} M_{ij}(\tau) = N_{\cdot j}(\tau) - \int_0^\tau \lambda_0(u) Y_{\cdot j}(u) du \quad (8)$$

is a zero-mean martingale, where

$$Y_{\cdot j}(\tau) = \sum_{i \neq j} Y_{ij}(\tau) \quad (9)$$

is the number of infectives who could made an observed infectious contact with j at infectiousness age τ . The combined observations of all j give us the counting process

$$N(\tau) = \sum_{j=1}^n N_{\cdot j}(\tau), \quad (10)$$

which counts the number of observed infectious contacts occurring at infectiousness age τ . Since it is a sum of zero-mean martingales,

$$M(\tau) = \sum_{j=1}^n M_{\cdot j}(\tau) = N(\tau) - \int_0^\tau \lambda_0(u) Y(u) du \quad (11)$$

is a zero-mean martingale, where

$$Y(\tau) = \sum_{j=1}^n Y_{\cdot j}(\tau) \quad (12)$$

is the total number of contact intervals of length τ that were observed. Note that $Y(\tau)$ is left-continuous and monotonically decreasing, as are $Y_{ij}(\tau)$ and $Y_j(\tau)$.

We are now ready to show that the standard Nelson-Aalen estimator produces an unbiased estimate of $\Lambda_0(\tau)$. The Nelson-Aalen estimate of $\Lambda_0(\tau)$ is

$$\widehat{\Lambda}(\tau) = \int_0^\tau \frac{1_{Y(u) > 0}}{Y(u)} dN(u) \quad (13)$$

Let $\mathcal{T} = \max\{\tau: Y(\tau) > 0\}$, and let $\tau \wedge \mathcal{T} = \min(\tau, \mathcal{T})$. Then

$$\widehat{\Lambda}(\tau) - \Lambda_0(\tau \wedge \mathcal{T}) = \int_0^{\tau \wedge \mathcal{T}} \frac{1_{Y(u) > 0}}{Y(u)} dM(u), \quad (14)$$

is a zero-mean martingale. Therefore, $\widehat{\Lambda}(\tau)$ is an unbiased estimate of $\Lambda_0(\tau)$ for $\tau \in [0, \mathcal{T}]$. In addition to being unbiased, $\widehat{\Lambda}(\tau)$ maximizes the log likelihood

$$\ell(\Lambda) = \sum_{j: t_j \leq T_j} \ln \left(\frac{d\Lambda(t_j - t_{v_j} - \varepsilon_{v_j})}{Y(t_j)} \right) - \int_0^\infty Y(u) d\Lambda(u) \quad (15)$$

over all step functions $\Lambda(\tau)$ and is asymptotically efficient (Van der Vaart, 1998). For $\tau \in [0, \mathcal{T}]$, an unbiased estimate of $\text{Var}(\widehat{\Lambda}(\tau))$ is

$$\widehat{\sigma}^2(\tau) = \int_0^{\tau \wedge \mathcal{T}} \frac{1_{Y(u) > 0}}{Y(u)^2} dN(u), \quad (16)$$

which is the optional variation process of (14). Using the martingale central limit theorem and a log transformation, we get the approximate pointwise $1 - \alpha$ confidence limits

$$\widehat{\Lambda}(\tau) \exp \left(\pm \frac{\widehat{\sigma}(\tau)}{\widehat{\Lambda}(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right), \quad (17)$$

where Φ is the cumulative distribution function of the standard normal distribution. This confidence interval is valid for τ at which $Y(\tau)$ is large; see Appendix B.

A Kaplan-Meier estimate (Kaplan and Meier, 1958) of the survival function of the contact interval distribution can be obtained from the increments of the Nelson-Aalen estimate using the product integral (Kalbfleisch and Prentice, 2002; Aalen et al., 2009):

$$\widehat{S}(\tau) = \pi_0^\tau \left(1 - \frac{1_{Y(u) > 0}}{Y(u)} dN(u) \right) \quad (18)$$

with the approximate pointwise $1 - \alpha$ log confidence limits

$$\ln \widehat{S}(\tau) \exp \left(\pm \frac{\widehat{\sigma}(\tau)}{\widehat{\Lambda}(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right). \quad (19)$$

2.2. Nonparametric estimation when who-infects-whom is not observed

Now suppose that we can observe infection but not who-infects-whom. We cannot calculate the Nelson-Aalen estimate because we do not know which contact intervals end in infectious contact and which are censored. However,

$$\Lambda_0(\tau) = \mathbb{E}[\widehat{\Lambda}(\tau)] = \mathbb{E}[\mathbb{E}[\widehat{\Lambda}(\tau) | \text{observed data}]] \quad (20)$$

by the law of iterated expectation, so we can still obtain an unbiased estimate of $\Lambda_0(\tau)$. Let $\widehat{\Lambda}_{\mathbf{v}}(\tau)$ denote the value of $\widehat{\Lambda}(\tau)$ that we would have calculated had we observed the transmission network $\mathbf{v}$. Then

$$\widetilde{\Lambda}(\tau) = \sum_{\mathbf{v} \in \mathcal{V}} \widehat{\Lambda}_{\mathbf{v}}(\tau) \Pr(\mathbf{v} | \text{observed data}) \quad (21)$$

is an unbiased estimate of $\Lambda_0(\tau)$ for $\tau \in [0, \mathcal{T}]$. This is the marginal Nelson-Aalen estimator.

For each $\mathbf{v} \in \mathcal{V}$, we can calculate $\Pr(\mathbf{v} | \text{observed data})$ if we know $\lambda_0(\tau)$ (Kenah et al., 2008). Given that j is infected at time t_j , the probability that j was infected by person i is $p_{ij}(\lambda_0)$, where

$$p_{ij}(\lambda) = \frac{\lambda(t_j - t_i - \varepsilon_i) 1_{i \in \mathcal{V}_j}}{\sum_{k \in \mathcal{V}_j} \lambda(t_j - t_k - \varepsilon_k)}. \quad (22)$$

For any hazard function $\lambda(\tau)$, the probability of a transmission network $\mathbf{v} = (v_1, \dots, v_n)$ given the observed data and λ is

$$\Pr(\mathbf{v} | \text{observed data}, \lambda) = \prod_{j: t_j < T_j} p_{v_j j}(\lambda), \quad (23)$$

so the infector $v_j \in \mathcal{V}_j$ of each infected j can be chosen independently of all v_i , $i \neq j$. Equations (22) and (23) assume a continuous contact interval distribution, which ensures that simultaneous infectious contacts have probability zero.

A marginal Nelson-Aalen estimate can be calculated using a known or estimated $\lambda(\tau)$. We outline this calculation and then use it as the basis of an EM algorithm that converges to a nonparametric MLE of $\Lambda_0(\tau)$. Let

$$\widetilde{N}_{\cdot j}(\tau | \lambda) = \mathbb{E} [N_{\cdot j}(\tau) | \lambda, \text{observed data}], \quad (24)$$

be the cadlag (right-continuous with left-hand limits) step function with a jump of size $p_{ij}(\lambda)$ at $\tau = t_j - t_i - \varepsilon_i$ for each $i \in \mathcal{V}_j$. Now let

$$\widetilde{N}(\tau | \lambda) = \sum_{j=1}^n \widetilde{N}_{\cdot j}(\tau | \lambda). \quad (25)$$

By equations (21) through (23), the marginal Nelson-Aalen estimate given $\lambda(\tau)$ is

$$\widetilde{\Lambda}(\tau | \lambda) = \int_0^\tau \frac{1_{Y(u) > 0}}{Y(u)} d\widetilde{N}(u | \lambda). \quad (26)$$

If $\lambda(\tau) = \lambda_0(\tau)$, this is an unbiased estimate of $\Lambda_0(\tau)$ for each $\tau \in [0, \mathcal{T}]$. Though it would be pointless to calculate $\widetilde{\Lambda}(\tau)$ if we knew $\lambda_0(\tau)$, equation (26) can be used in the following EM algorithm:

- a. Begin with an initial $\lambda^{(0)}(\tau)$, which could be a constant (equivalent to assuming exponential contact intervals) or a parametric estimate. Calculate an initial marginal Nelson-Aalen estimate, $\widetilde{\Lambda}(\tau | \lambda^{(0)})$. Set $k = 0$.

- b. Use $\tilde{\Lambda}(\tau|\lambda^{(k)})$ to calculate a new hazard function estimate $\lambda^{(k+1)}(\tau)$
- c. Update the probabilities in equation (22) using $\lambda^{(k+1)}(\tau)$.
- d. Use the updated probabilities to calculate $\tilde{\Lambda}(\tau|\lambda^{(k+1)})$ via equation (26). Set $k = k+1$.
- e. Repeat the smoothing step (b), the E-step (c) and the M-step (d) until the sequence $\tilde{\Lambda}(\tau|\lambda^{(k)})$ converges.

The limit of this sequence is the marginal Nelson-Aalen estimate $\tilde{\Lambda}(\tau)$.

To show that this is an EM algorithm, we must prove that $\tilde{\Lambda}(\tau|\lambda^{(k)})$ increases the expected log likelihood given $\lambda^{(k)}(\tau)$. Averaging the log likelihood from equation (15) over all $\mathbf{v} \in \mathcal{V}$ using the infector probabilities $p_{ij}(\lambda^{(k)})$, we get

$$\tilde{\ell}(\Lambda|\lambda^{(k)}) = \sum_{j:t_j \leq T} \sum_{i \in \mathcal{V}_j} p_{ij}(\lambda^{(k)}) \ln \left(\frac{d\Lambda(t_j - t_i - \varepsilon_i)}{Y(t_j - t_i - \varepsilon_i)} \right) - \int_0^\infty Y(u) d\Lambda(u). \quad (27)$$

Differentiating with respect to $d\Lambda(t_j - t_i - \varepsilon_i)$, we find that $\tilde{\ell}(\tau|\lambda^{(k)})$ is maximized by

$$d\Lambda(t_j - t_i - \varepsilon_i) = \frac{p_{ij}(\lambda^{(k)})}{Y(t_j - t_i - \varepsilon_i)}, \quad (28)$$

exactly as in $\tilde{\Lambda}(\tau|\lambda^{(k)})$. Therefore, $\tilde{\Lambda}(\tau|\lambda^{(k)})$ maximizes the expected log likelihood given $\lambda^{(k)}(\tau)$. Before convergence of the EM algorithm, $d\tilde{\Lambda}(\tau|\lambda^{(k)}) \neq d\lambda^{(k)}(\tau)$. At convergence, we have equality. In this sense, the EM algorithm converges to a self-consistent nonparametric estimate of $\Lambda_0(\tau)$.

The variance $\tilde{\sigma}^2(\tau)$ of $\tilde{\Lambda}(\tau)$ can be estimated using the conditional variance formula.

Conditioning on the transmission network $\mathbf{v}$, we get

$$\tilde{\sigma}^2(\tau) = \mathbb{E}[\tilde{\sigma}_{\mathbf{v}}^2(\tau)] + \text{Var}(\widehat{\Lambda}_{\mathbf{v}}(\tau)), \quad (29)$$

where $\widehat{\Lambda}_{\mathbf{v}}(\tau)$ is the Nelson-Aalen estimate from equation (13) and $\tilde{\sigma}_{\mathbf{v}}^2(\tau)$ is the variance estimate from equation (16) that we would have calculated had we observed the transmission network $\mathbf{v}$. Let $\tilde{\lambda}(\tau)$ denote the hazard function estimate corresponding to $\tilde{\Lambda}(\tau)$, and let $\tilde{N}(\tau) = \tilde{N}(\tau|\tilde{\lambda})$. The first term of equation (29) reduces to

$$\mathbb{E}[\tilde{\sigma}_{\mathbf{v}}^2(\tau)] = \int_0^\tau \frac{1_{Y(u)>0}}{Y(u)^2} d\tilde{N}(u). \quad (30)$$

Appendix C shows that the second term reduces to

$$\text{Var}(\widehat{\Lambda}_{\mathbf{v}}(\tau)) = \int_0^\tau \frac{1_{Y(u)>0}}{Y(u)^2} d\tilde{N}(u) - \sum_{j:t_j < T_j} \left(\int_0^\tau \frac{1_{Y(u)>0}}{Y(u)} d\tilde{N}_{\cdot j}(u) \right)^2, \quad (31)$$

where $\tilde{N}_{\cdot j}(\tau) = \tilde{N}_{\cdot j}(\tau|\tilde{\lambda})$. Therefore,

$$\tilde{\sigma}^2(\tau) = 2 \int_0^\tau \frac{1_{Y(u)>0}}{Y(u)^2} d\tilde{N}(u) - \sum_{j: t_j < \tau} \left(\int_0^\tau \frac{1_{Y(u)>0}}{Y(u)} d\tilde{N}_{\cdot j}(u) \right)^2 \quad (32)$$

Using the martingale central limit theorem and a log transformation, we get the approximate pointwise $1 - \alpha$ confidence limits

$$\tilde{\Lambda}(\tau) \exp \left(\pm \frac{\tilde{\sigma}(\tau)}{\tilde{\Lambda}(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right). \quad (33)$$

This confidence interval is valid for τ where $Y(\tau)$ is large. Since $Y(\tau)$ does not depend on who-infected-whom, the martingale central limit theorem should hold for $\tilde{\Lambda}(\tau)$ under conditions very similar to those required for the Nelson-Aalen estimate $\hat{\Lambda}(\tau)$; see Appendix B.

A marginal Kaplan-Meier estimator of the contact interval survival function can be obtained from the increments of the marginal Nelson-Aalen estimator using the product integral (Kalbfleisch and Prentice, 2002; Aalen et al., 2009):

$$\tilde{S}(\tau) = \pi_0^\tau \left(1 - \frac{1_{Y(u)>0}}{Y(u)} d\tilde{N}(\tau) \right) \quad (34)$$

with the approximate pointwise $1 - \alpha$ log confidence limits

$$\ln \tilde{S}(\tau) \exp \left(\pm \frac{\tilde{\sigma}(\tau)}{\tilde{\Lambda}(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right). \quad (35)$$

2.2.1. Partially-observed transmission networks—Suppose we observe the infectors of a subset of all infected persons. For each j with an observed v_j , let $\tilde{N}_{\cdot j}(\tau) = N_{v_{jj}}(\tau)$. Taking $\mathcal{V}_j = \{v_j\}$ for these j , equations (21) through (26) define a marginal Nelson-Aalen estimator that takes this additional information into account. With these changes, equations (32) and (33) can be used to get confidence intervals.

2.2.2. Estimating the hazard function—Obtaining $\lambda^{(k+1)}(\tau)$ from $\tilde{\Lambda}(\tau|\lambda^{(k)})$ in step (b) of the EM algorithm is nontrivial. In general, the increments of $\tilde{\Lambda}(\tau|\lambda^{(k)})$ cannot be used directly as $\lambda^{(k+1)}(\tau)$. To see this, consider an infected person j . For each $i \in \mathcal{V}_j$, let $p_{ij}^{(k)}$ denote the estimated probability that i infected j in the k^{th} iteration of the EM algorithm, with $p_{ij}^{(0)}$ denoting the initial estimate. After the first iteration, we get $p_{ij}^{(1)} \propto p_{ij}^{(0)} \left(Y(\tau_{ij}) \right)^{-1}$. After k iterations,

$$p_{ij}^{(k)} \propto \frac{p_{ij}^{(0)}}{Y(\tau_{ij})^k}. \quad (36)$$

Since $Y(\tau)$ is decreasing in τ , the EM algorithm would converge to a marginal Nelson-Aalen estimate where the $i \in \mathcal{V}_j$ with the longest τ_{ij} is assigned probability one of being v_j . This problem arises because equation (22) assumes a continuous contact interval distribution. Since the contact interval distribution is continuous, the intervals between the increments are used

of $\tilde{\Lambda}(\tau|\lambda^{(k)})$ contain information about $\lambda(\tau)$ that is ignored if these increments are used directly as an estimate of $\lambda(\tau)$. Therefore, obtaining $\lambda^{(k+1)}(\tau)$ from $\tilde{\Lambda}(\tau|\lambda^{(k)})$ requires some form of smoothing. The algorithm depends on the smoothed estimate of $\lambda(\tau)$ only to calculate the probabilities p_{ij} in equation (22), so it is not highly sensitive to the smoothing method used. In our experience, cubic smoothing splines and kernel smoothers produce nearly identical results.

2.3. Approximate Nelson-Aalen estimates for mass-action SEIR models

In a mass-action model, $C_{ij} = 1$ for all ordered pairs ij but the hazard of infectious contact is inversely proportional to the population size n . If $\Lambda_n(\tau)$ is the cumulative hazard function of the contact interval distribution in a population of size n , then

$$\Lambda_n(\tau) = \frac{\Lambda_*(\tau)}{n-1}, \quad (37)$$

where we call $\Lambda_*(\tau)$ the normalized cumulative hazard function of the contact interval distribution. With knowledge of n , the methods of the previous two sections could be used to estimate $\Lambda_n(\tau)$ and $\Lambda_*(\tau)$. In Kenah (2011), it was shown that the full parametric likelihoods for mass-action SEIR models can be approximated by likelihoods that depend only on data about infected individuals. Here, we derive approximate nonparametric estimates of $\Lambda_*(\tau)$ that depend only on data about infected individuals. For a fixed number m of infections, these approximations become exact in the limit as $n \rightarrow \infty$. Thus, they can be used to analyze data from the early stages of an epidemic, when $m \ll n$.

Until now, we have assumed that each ordered pair ij is under observation as long as infection in j can be observed. Since everyone can infect everyone else in a mass-action model, it is simplest to assume that the entire population remains under observation until a time T . Then $\mathcal{Y}_{ij}(\tau)$ depends only on i , so let

$$\mathcal{Y}_i(\tau) = 1_{t_i + e_i + \tau \leq T}. \quad (38)$$

Since $C_{ij} = 1$ for all ij ,

$$Y(\tau) = \sum_i \mathcal{I}_i(\tau) \mathcal{Y}_i(\tau) \sum_{j \neq i} \mathcal{I}_{ij}(\tau) \quad (39)$$

by equations (1), (9), and (12). Let $Y_*(\tau) = \sum_i \mathcal{I}_i(\tau) \mathcal{Y}_i(\tau)$ denote the number of infectives who remain infectious and under observation at infectiousness age τ . Then

$$(n-m)Y_*(\tau) \leq Y(\tau) \leq (n-1)Y_*(\tau). \quad (40)$$

for all $\tau \in [0, \mathcal{T}]$ because $n-m \leq \sum_{j \neq i} \mathcal{I}_{ij}(\tau) \leq n-1$ for all i .

Let $\widehat{\Lambda}_n(\tau)$ be the Nelson-Aalen estimate of $\Lambda(\tau)$ obtained in a population of size n when who-infects-whom is observed, and let

$$\widehat{\Lambda}_*(\tau) = \int_0^\tau \frac{1_{Y_*(u) > 0}}{Y_*(u)} dN(u). \quad (41)$$

Since we assume $n \gg m$, $Y(u) > 0$ if and only if $Y_*(u) > 0$. Therefore,

$$\widehat{\Lambda}_*(\tau) \leq (n-1)\widehat{\Lambda}_n(\tau) \leq \frac{n-1}{n-m}\widehat{\Lambda}_*(\tau), \quad (42)$$

so $\widehat{\Lambda}_*(\tau) \rightarrow (n-1)\widehat{\Lambda}_n(\tau)$ for a fixed m as $n \rightarrow \infty$. Since $\widehat{\Lambda}_n(\tau)$ is a consistent estimator of $\Lambda_n(\tau)$, this implies that $\widehat{\Lambda}_*(\tau)$ is a consistent estimator of $\Lambda_*(\tau)$, where we take limits first as $n \rightarrow \infty$ and then as $m \rightarrow \infty$. Similarly,

$$\widehat{\sigma}_*^2(\tau) = \int_0^\tau \frac{1_{Y_*(u)>0}}{Y_*(u)^2} dN(u). \quad (43)$$

is a consistent estimate of the variance of $\widehat{\Lambda}_*(\tau)$. Using the martingale central limit theorem and a log transformation, we get the asymptotic $1 - \alpha$ confidence limits

$$\widehat{\Lambda}_*(\tau) \exp \left(\pm \frac{\widehat{\sigma}_*(\tau)}{\widehat{\Lambda}_*(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right). \quad (44)$$

This confidence interval is valid for τ such that $Y_*(\tau)$ is large. A Kaplan-Meier estimate can be obtained from the increments of $\widehat{\Lambda}_*(\tau)$ using the product integral, and log confidence limits can be obtained using the multiplier in equation (44).

When who-infects-whom is not observed, we have the approximate marginal Nelson-Aalen estimate

$$\widetilde{\Lambda}_*(\tau|\lambda_*) = \int_0^\tau \frac{1_{Y_*(u)>0}}{Y_*(u)} d\widetilde{N}(u|\lambda_*). \quad (45)$$

Since the probabilities p_{ij} in equation (22) depend on $\lambda(\tau)$ only up to a multiplicative constant, $\widetilde{\Lambda}_*(\tau|\lambda_*)$ can replace $\widetilde{\Lambda}(\tau|\lambda)$ in the EM algorithm from Section 2.2. Let $\widetilde{\Lambda}_*(\tau)$ denote the marginal Nelson-Aalen estimate to which this algorithm converges, let $\widetilde{\lambda}_*(\tau)$ denote the corresponding hazard function, and let $\widetilde{N}(\tau) = \widetilde{N}(\tau|\widetilde{\lambda}_*)$. Then the variance of $\widetilde{\Lambda}_*(\tau)$ is

$$\widetilde{\sigma}_*^2(\tau) = 2 \int_0^\tau \frac{1_{Y_*(u)>0}}{Y_*(u)^2} d\widetilde{N}^*(u) - \sum_{j=1}^n \left(\int_0^\tau \frac{1_{Y_*(u)>0}}{Y_*(u)} d\widetilde{N}_{\cdot j}(u) \right)^2, \quad (46)$$

where $\widetilde{N}_{\cdot j}(\tau) = \widetilde{N}_{\cdot j}(\tau|\widetilde{\lambda}_*)$. The asymptotic $1 - \alpha$ confidence limits are

$$\widetilde{\Lambda}_*(\tau) \exp \left(\pm \frac{\widetilde{\sigma}_*(\tau)}{\widetilde{\Lambda}_*(\tau)} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right). \quad (47)$$

This confidence interval is valid for τ such that $Y_*(\tau)$ is large. A marginal Kaplan-Meier estimate can be obtained from the increments of $\widetilde{\Lambda}_*(\tau)$ using a product integral, and log confidence limits can be obtained using the multiplier in equation (47).

3. Simulations

To evaluate the performance of the methods from Section 2, we conducted simulations of network-based and mass-action epidemics. In each simulation, we used data on the first 1,000 infections from an epidemic in a population of 100,000 to calculate the Nelson-Aalen and marginal Nelson-Aalen estimates of $\Lambda_0(\tau)$ and the Kaplan-Meier and marginal Kaplan-Meier

estimates of the corresponding survival function. For each of these estimators, we looked at whether the confidence interval contained the true value of the estimated function at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of all possible (i.e., censored and uncensored) contact intervals. After 1,000 simulations, we calculated the coverage probability for each estimator at each quantile. For each coverage probability, we calculated an exact 95% confidence interval.

All models had a latent period of zero and an exponential infectious period with mean one. The smoothing step of the EM algorithm was done via an inverse-variance weighted cubic smoothing spline with a smoothing parameter chosen by generalized cross-validation (Wegman and Wright, 1983). Convergence of the EM algorithm was monitored by calculating the mean of absolute value of the differences between the current and previous marginal Nelson-Aalen estimates at the 5th, 10th, 15th, . . . , 85th, 90th, and 95th percentiles of all possible contact intervals. We call this the L1 difference. All EM algorithms began by assigning equal probability to all possible transmission networks, which is equivalent to assuming an exponential contact interval distribution. We ran the algorithm for a minimum of 5 iterations, and it was continued until an L1 difference less than a specified tolerance was achieved or the 50th iteration was completed.

All epidemic models were written in Python (www.python.org) using the NumPy and SciPy packages (www.scipy.org). For network-based models, our contact networks were generated using the NetworkX package (networkx.lanl.gov). Statistical analysis was conducted in R (www.r-project.org) via the RPy2 package (rpy.sourceforge.net). The code for the models and estimators is available as Online Supplementary Information.

3.1. Network-based simulations

In all network-based simulations, the contact network was 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 100,000 nodes, each node was connected to its 10 nearest neighbors. Each edge was then rewired to a randomly chosen node with probability .1, so 34.9% of nodes are connected only to their ten nearest neighbors, 38.7% have one long-distance edge, 19.4% have two, 5.7% have three, and 1.3% have four or more. Since this network structure gave us high clustering, infected nodes typically had several possible infectors. A new contact network was built for each simulation, so the results do not reflect the idiosyncrasies of any particular network realization.

We used Weibull(.5, 1), exponential(1), and Weibull(2, 1) contact interval distributions, where Weibull(s ; r) denotes a Weibull distribution with shape parameter s and rate parameter r . The corresponding cumulative hazard functions are:

$$\Lambda_0(\tau) = \sqrt{\tau} \quad \text{for the Weibull (.5, 1) distribution} \quad (48)$$

$$\Lambda_0(\tau) = \tau \quad \text{for the exponential(1) distribution, and} \quad (49)$$

$$\Lambda_0(\tau) = \tau^2 \quad \text{for the Weibull(2, 1) distribution.} \quad (50)$$

Note that the exponential(1) distribution is the same as the Weibull(1, 1) distribution. The hazard of infectious contact decreases throughout the infectious period in the Weibull(.5, 1) case, remains constant in the exponential(1) case, and increases in the Weibull(2, 1) case.

The L1 tolerance for the EM algorithm was set to .0005. This was approximately the minimum tolerance consistently achieved by the EM algorithm before the L1 differences became a seemingly chaotic sequence of small numbers.

Results The EM algorithm for the marginal Nelson-Aalen estimate converged easily in all simulations. For models with exponential(1) contact intervals, the desired tolerance was achieved within 5 iterations for all models. For Weibull(.5, 1) contact intervals, the maximum number of iterations was 6. For the Weibull(2, 1) contact intervals, the maximum was 8. Table 1 shows the coverage probabilities and exact 95% confidence intervals for the Nelson-Aalen, marginal Nelson-Aalen, Kaplan-Meier, and marginal Kaplan-Meier estimators in models with Weibull(.5, 1) and Weibull(2, 1) contact interval distributions. Coverage probabilities for the Nelson-Aalen and Kaplan-Meier estimators were close to .95 across the range of available data for both models. Coverage probabilities for the marginal Nelson-Aalen and Kaplan-Meier estimators were slightly lower but above .90 in all cases. Coverage probabilities for the model with exponential contact intervals were above .93 for all estimators and data quantiles (not shown). Figures 1 and 2 show good agreement between the estimated and true cumulative hazard and survival functions of the contact interval distribution across the range of available data.

3.2. Mass-action simulations

For mass-action models, we used Weibull(.5, 5), exponential(2), and Weibull(2, 1) normalized contact interval distributions. The corresponding cumulative hazard functions are:

$$\Lambda_*(\tau) = \sqrt{5\tau} \quad \text{for the Weibull } (.5, 5) \quad \text{distribution} \quad (51)$$

$$\Lambda_*(\tau) = 2\tau \quad \text{for the exponential } (2) \quad \text{distribution, and} \quad (52)$$

$$\Lambda_*(\tau) = \tau^2 \quad \text{for the Weibull } (2, 1) \quad \text{distribution.} \quad (53)$$

In a mass-action model, a Weibull(.5, 1) distribution for the normalized contact interval distribution produces $R_0 = .89$, so no epidemics occur. Changing the rate parameter to 5 produces $R_0 = 1.98$, which is very close to the $R_0 = 2$ of the other two models. The hazard of infectious contact decreases throughout the infectious period in the Weibull(.5, 5) case, remains constant in the exponential(1) case, and increases in the Weibull(2, 1) case.

The L1 tolerance for the EM algorithm was set to .005. This was approximately the minimum tolerance consistently achieved by the EM algorithm before the L1 differences became a chaotic sequence of small numbers.

Results The EM algorithm for the marginal Nelson-Aalen estimate took longer to converge for mass-action models than for network-based models, despite the greater tolerance. For models with exponential contact intervals, the maximum number of iterations was 19. For Weibull(2, 1) contact intervals, the maximum was 29. For Weibull(.5, 5) contact intervals, the maximum was 32. Table 2 shows coverage probabilities and exact 95% confidence intervals for the Nelson-Aalen, marginal Nelson-Aalen, Kaplan-Meier, and marginal Kaplan-Meier estimators in models with Weibull(.5, 5) and Weibull(2, 1) normalized contact interval distributions. Coverage probabilities for the Nelson-Aalen and Kaplan-Meier estimators were close to .95 across the range of available data for both models. In contrast, the marginal Nelson-Aalen and Kaplan-Meier estimators were spectacular failures. Figures 3 and 4 show good agreement between the estimated and true survival and cumulative hazard functions when who-infects-whom is observed (left) but poor agreement when who-infects-whom is not observed (right).

The number of possible infectors for each infection is much larger in a mass-action model than in a network-based model. This could cause problems for the EM algorithm in several ways. It is possible that the EM algorithm works but converges very slowly. Slow convergence may be overwhelmed by numerical instability, or the iterations may have been stopped prematurely. It is also possible that the asymptotic approximation to the EM algorithm is not itself an EM algorithm. Whatever the reason, the marginal Nelson-Aalen estimator does not work for mass-action models. Fortunately, the approximate parametric likelihoods for mass-action models in Kenah (2011) work well even when who-infects-whom is not observed.

4. Data analysis: Pandemic influenza A(H1N1) in Los Angeles County, 2009

In this section, we use the marginal Nelson-Aalen estimator to analyze pandemic influenza A (H1N1) (pH1N1) surveillance data collected by the Los Angeles County Department of Public Health between April 22 and May 19, 2009. The data was collected according to the following protocol (Sugimoto et al., 2011):

- a. Individuals who presented to healthcare providers or the county health department with acute febrile respiratory illness (AFRI, defined as fever $\geq 100^{\circ}\text{F}$ plus cough, sore throat, or runny nose) had nasopharyngeal swabs and aspirates tested for pandemic influenza A(H1N1). Those who tested positive are the index cases.
- b. The index cases were interviewed by telephone and asked to report AFRI episodes among household contacts, including dates of illness onset. Additional AFRI episodes among household contacts were ascertained during follow-up interviews 14 days after the illness onset of the index case. All cases of AFRI in the household after 10 days prior to symptom onset in the index case were assumed to be pH1N1 cases. The earliest pH1N1 case in each household is the primary case; this is usually (but not always) the index case.

There were 58 households with a total of 299 members. In these households, there were 62 primary cases (four households had co-primary cases) and 35 secondary cases. In 51 of the 58 households, the index case was a primary case. This is a good example of household surveillance data from the early stages of an epidemic.

As in the simulations, our goal is to estimate the cumulative hazard function of the contact interval distribution. We compare the marginal Nelson-Aalen estimate with parametric estimates obtained using the methods from Kenah (2011). We use the corresponding marginal Kaplan-Meier estimate to estimate the probability that an infected person makes infectious contact with a given household member during his or her infectious period, which we call the household infectious contact probability.

Our natural history assumptions are adapted from Yang et al. (2009). 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. This means that a person i with onset of symptoms on day t_i^{sym} is infected on day $t_i = t_i^{\text{sym}} - 2$, has onset of infectiousness on day t_i , and recovers from infectiousness on day $t_i + 6$. In discrete time, the first day on which i is able to infect other persons is the day after his or her onset of infectiousness, which is $t_i + 1 = t_i^{\text{sym}} - 1$ under our assumptions. If day 0 is the infection time, then we have the first secondary transmissions on day 1, the onset of symptoms on day 2, and removal on day 6. In a sensitivity analysis, we vary the incubation period from 1 day to 3 days, the infectious period from 5 days to 7 days, and the latent period from 0 days to 1 day.

4.1. Results

Figure 5 shows the marginal Nelson-Aalen estimate of the cumulative hazard function of the contact interval distribution, along with approximate 95% confidence limits and parametric estimates assuming exponential and Weibull contact interval distributions. A cumulative hazard estimate assuming a gamma contact interval distribution (not shown) was almost exactly the same as that assuming a Weibull distribution.

The exponential, Weibull, and marginal Nelson-Aalen point estimates of the cumulative hazard at 6 days post-infection are .0687, .0692, and .0729, respectively. The corresponding survival probabilities are .9334, .9331, and .9296. Thus, all three estimates indicate a household infectious contact probability of .07 over the course of the infectious period. The approximate 95% confidence interval for the marginal Nelson-Aalen estimate is (.0503, .1056). The corresponding 95% confidence interval for the marginal Kaplan-Meier estimate of the survival probability is (.8997, .9509). Therefore, our nonparametric estimate of the household infectious contact probability is .07 (.05, .10).

While the nonparametric and parametric estimates agree on the household infectious contact probability, they differ in their estimates of distribution of infectiousness over time. The exponential estimate inherently predicts a constant hazard of infectious contact over the entire infectious period. The Weibull estimate also predicts little variation in the hazard of infectious contact over the infectious period, with slightly lower infectiousness near the beginning and slightly higher infectiousness near the end. The marginal Nelson-Aalen curve has much larger jumps on days 1, 2, and 3 following infection than on days 5 and 6, which places the highest infectiousness on the day prior to, the day of, and the day after the onset of symptoms. According to this estimate, only 12% of infectious contacts occur > 2 days after the onset of symptoms and only 6% occur > 3 days after symptom onset.

4.2. Sensitivity analysis

Figure 6 shows the results of a sensitivity analysis in which we vary the assumed (a) incubation period, (b) latent period, and (c) infectious period. In panel (d), we vary all three of these intervals simultaneously to generate the greatest possible variation. The results are insensitive to the incubation period and infectious period but sensitive to the latent period. However, there is good evidence that little or no transmission occurs two or more days prior to the onset of symptoms (Donnelly et al., 2011).

4.3. Previous pH1N1 household data analyses

Our nonparametric estimate of .07 (.05, .10) for the household infectious contact probability is similar to the .06 (.03, .11) estimated using data collected by Public Health{Seattle and King County during April-May, 2009 (Sugimoto et al., 2011). This estimate was obtained using a chain-binomial model to estimate the probability of infectious contact with a given household member on each day of the infectious period and then calculating the probability that infectious contact occurs at least once, making it directly comparable with our estimate.

It is more difficult to compare our estimated household infectious contact probability with estimates of the household secondary attack rate (SAR), which are usually obtained by calculating the proportion of all susceptible members of study households who have the onset of a given set of symptoms within a given time period relative to symptom onset in the index case. Among these estimates for pH1N1 are: .145 (.129, .164) for in uenza-like illness (ILI) within 7 days after the index case symptom onset (Carcione et al., 2011), .13 for acute respiratory infection (ARI) and .9 for ILI within 1-9 days after the index case symptom onset (Morgan et al., 2010), .113 (.088, .137) for ILI up to 23 days after index case symptom onset (France et al., 2010), and .13 for ARI and .10 for ILI 7 days before or after the index case

symptom onset (Cauchemez et al., 2009). We did the following simulations to see if our estimated household infectious contact probability of .07 is consistent with these estimates:

- a. For each of the 58 households in the LA data, we simulated an epidemic using its observed number of index cases and a household infectious contact probability of .07.
- b. After the epidemics in all households completed, we calculated the proportion of all household contacts who were infected.

Repeating this 10,000 times, we got a mean household SAR of .12 with a bootstrap percentile 95% confidence interval of (.06, .19), which places us solidly within the range of previous household SAR estimates.

Our nonparametric estimate of the distribution of infectiousness over time is consistent with estimates of the serial interval distribution for pH1N1 by Donnelly et al. (2011), who found high infectiousness on days 0, 1, and 2 following symptom onset with only 18% of transmission occurring > 2 days after the onset of symptoms and only 5% occurring > 3 days after symptom onset. A similar pattern of high infectiousness early in the infectious period was found by Cauchemez et al. (2009), who also used a serial interval distribution. Though theoretical and practical problems with the serial interval distribution were noted in Section 1, the similarity between our results and those of Donnelly et al. (2011) and Cauchemez et al. (2009) suggests that the marginal Nelson-Aalen estimate captured an important feature of pH1N1 transmission that was missed by the parametric contact interval distribution estimates.

5. Discussion

The nonparametric estimators proposed in Section 2 have many of the same limitations as the parametric estimators in Kenah (2011) and some of their own. These include:

- a. Several technical issues need further exploration. A more rigorous theoretical analysis of convergence and efficiency of the marginal Nelson-Aalen estimator, including the smoothing step, is needed. Convergence criteria need to be more carefully examined, and confidence intervals for small samples need to be developed.
- b. We assumed that the contact interval distribution is identically distributed for all ordered pairs ij such that $C_{ij} = 1$, which is unrealistic. For instance, age was found to have an effect on pH1N1 transmission in most of the papers cited in Section 4 (Cauchemez et al., 2009; France et al., 2010; Morgan et al., 2010; Carcione et al., 2011; Sugimoto et al., 2011), and hazards of infectious contact between household members seem to depend on the size of the household (Cauchemez et al., 2009). The effects of covariates on the transmission of disease is a central concern of infectious disease epidemiology that we have not addressed here.
- c. We have assumed that contact intervals are independent of infectious periods, which is unrealistic. Both are affected by the interaction between the host and the pathogen.
- d. We have assumed that infection times, infectiousness onset times, and removal times are observed, which is unrealistic. Usually, only symptom onset times are observed. In the analysis of the Los Angeles County household data, we were forced to assume constant latent, incubation, and infectious periods because our data included only symptom onset times. The implicit assumption of constant latent, incubation, and infectious periods in methods based on generation and serial intervals was one of our main criticisms in Section 1. The data analysis in Section 4 merely made these same assumptions explicit.

- e. The SEIR framework is limited to acute, immunizing infectious diseases that spread person-to-person, so our methods do not apply directly to tuberculosis, HIV/AIDS, foodborne and waterborne diseases, pneumococcal and meningococcal diseases, and other infectious diseases of major public health importance.

Nonetheless, survival analysis based on contact intervals has important theoretical and practical advantages over methods based on generation or serial intervals, including greater flexibility in the underlying transmission model, greater clarity about data requirements and analytical assumptions, the use of information contributed by uninfected person-time, and validity throughout the course of an epidemic. The theory and methods of survival analysis offer many possibilities for overcoming the limitations listed above.

The asymptotic analysis required to address limitation (a) is complicated by the smoothing step. In simulations, the performance of the marginal Nelson-Aalen estimator did not appear sensitive to the smoothing method. However, a smoother designed specifically for hazard functions, such as that of Müller and Wang (1994), may improve performance. Another technical limitation is that the variance estimates based on equation (29) treat the estimated transmission network probabilities as known, ignoring a component of the variance that may be important in small samples. These and other technical issues need to be resolved with more rigorous theoretical analysis and further simulation studies.

Limitation (b) can be addressed by extending the marginal Nelson-Aalen estimate into a semiparametric regression model for the effects of covariates on the hazard of infectious contact. Suppose the hazard of infectious contact from person i to person j at infectiousness age τ of i is

$$\lambda_{ij}(\tau) = \lambda_0(\tau) \exp(\beta^T X_{ij}), \quad (54)$$

where β is a parameter vector and X_{ij} is a vector of infectiousness covariates for i , susceptibility covariates for j , and pairwise covariates for ij . The methods of Section 2 could be adapted to nonparametrically estimate $\lambda_0(\tau)$ by adding a step to the EM algorithm that maximized the expected log likelihood over β . This would yield a semiparametric estimator of the effects of covariates on infectiousness and susceptibility that is similar in spirit to the Cox regression model (Cox, 1972). In infectious disease epidemiology, these covariate effects are usually a higher priority than the distribution of infectiousness over time, so the marginal Nelson-Aalen estimator may find its most important applications behind the scenes of this regression model.

Limitation (c) could be addressed in two ways. The simplest would be to identify a covariate process $X(\tau)$ such that $\mathcal{J}_i(\tau)$ and $\mathcal{N}_{ij}(\tau)$ are independent given $X(\tau^-)$. This covariate could be incorporated into the semiparametric regression model mentioned above. If such covariates do not exist or are not observable, multivariate survival methods could be used to estimate the joint distribution of the contact interval and the infectious period.

The missing data problems mentioned in limitation (d) are acknowledged, not created, by using contact intervals instead of generation or serial intervals. One of the most important advantages of the statistical framework proposed in this paper is that it allows a rigorous discussion of the missing-data problems that plague infectious disease epidemiology. Unobserved infection, infectiousness onset, and removal times could be handled using a semiparametric Bayesian framework (Sinha and Dey, 1997) or a particle sampler (Lee et al., 2005). A semiparametric Bayesian approach would use a prior process instead of a marginal Nelson-Aalen estimator. A particle sampler would take advantage of the fact that the marginal Nelson-Aalen estimate is an efficient numerical method of obtaining a likelihood maximized over all contact interval distributions for a given set of infection, infectiousness onset, and removal times.

Limitation (e) could be addressed by allowing individuals to experience multiple infections or infections of different types (e.g., new infection, carriage, relapse). As noted in Appendix A, complex infectiousness and susceptibility processes are allowed as long as independent censoring of contact intervals by the infectiousness, susceptibility, and observation processes remains valid.

By reformulating the analysis of infectious disease data in terms of survival analysis based on contact intervals, the methods proposed here and in Kenah (2011) point the way to more practical and rigorous statistical methods for infectious disease epidemiology. Each of the ingredients of this approach have been applied to infectious diseases before: counting processes and martingales by Becker (1989) and Rhodes et al. (1996), the EM algorithm by Becker (1997), and summation over possible transmission networks by Wallinga and Teunis (2004). The marginal Nelson-Aalen estimator combines these in a novel and useful way. In a separate paper, we will show that these methods reproduce classical chain-binomial models (see Rampey et al., 1992) for epidemics in discrete time.

Acknowledgments

This work benefited greatly from the comments of M. Elizabeth Halloran, Ira M. Longini, Jr., Yang Yang, Jonathan D. Sugimoto, and participants in the Dhaka University Statistics Department Alumni Association International Conference on the Theory and Application of Statistics (December 26/28, 2010 in Dhaka, Bangladesh), especially Md. Abu Yushuf Sharker of ICDDR,B. Brit Oulfsstad, Dee Ann Bagwell, Brandon Dean, Laurene Mascola, and Elizabeth Bancroft of the Los Angeles County Department of Public Health generously allowed the use of their data in Section 4. Office space, computing facilities, and administrative support were provided by the Fred Hutchinson Cancer Research Center. This research was supported by National Institute of General Medical Sciences (NIGMS) grant F32GM085945 and National Institute of Allergy and Infectious Diseases (NIAID) grant K99AI095302. The content is solely the responsibility of the author and does not necessarily represent the official views of NIGMS, NIAID, or the National Institutes of Health.

A. Independent censoring by infectiousness, susceptibility, and observation

In Section 2, we avoided technical discussion of the history (increasing family of α -fields) underlying the counting processes and martingales, which is essential to defining independent censoring of τ_{ij} by $\mathcal{I}_i(\tau)$, $\mathcal{S}_{ij}(\tau)$, and $\mathcal{O}_{ij}(\tau)$. A good introduction to these concepts is given in Chapter of Aalen et al. (2009).

First, we define the uncensored history generated by all contact intervals. Let $(\mathcal{G}_\tau)$ denote the history generated by $\mathcal{N}_{ij}(\tau)$ for all ij such that $C_{ij} = 1$ and i became infectious. Informally, this means that $\mathcal{G}(\tau)$ contains the value of $\mathcal{N}_{ij}(u)$ for all $u \in [0, \tau]$ and all ij in which an infectious contact interval would have been observed in the absence of censoring. The intensity process of $\mathcal{N}_{ij}(\tau)$ with respect to $(\mathcal{G}_\tau)$ is

$$\lambda_{ij}^{\mathcal{G}}(\tau) = \frac{\mathbb{E}[\mathrm{d}\mathcal{N}_{ij}(\tau) | \mathcal{G}_{\tau-}]}{\mathrm{d}\tau} = \begin{cases} \lambda_0(\tau) & \text{if } \tau_{ij} \geq \tau \\ 0 & \text{if } \tau_{ij} < \tau \end{cases}, \quad (55)$$

where $\mathcal{G}_{\tau-}$ is the uncensored history just before infectiousness age τ . The first equality is valid for models in which the contact interval distribution can depend on ij . The second equality assumes that the contact interval hazard function is $\lambda_0(\tau)$ for all ij , as in Section 2.

Next, we define the history generated by both the contact interval and censoring processes. Let $(\mathcal{H}_\tau)$ be the history generated by $\mathcal{N}_{ij}(\tau)$, $\mathcal{I}_i(\tau)$, $\mathcal{S}_{ij}(\tau)$, and $\mathcal{O}_{ij}(\tau)$. Then $\mathcal{H}_\tau$ contains the values of $\mathcal{N}_{ij}(u)$, $\mathcal{I}_i(u)$, $\mathcal{S}_{ij}(u)$ and $\mathcal{O}_{ij}(u)$ for all $u \in [0, \tau]$ and all ij in which an infectious contact

interval would have been observed in the absence of censoring. The intensity process of $\mathcal{N}_{ij}(\tau)$ with respect to $(\mathcal{H}_\tau)$ is

$$\lambda_{ij}^{\mathcal{H}}(\tau) = \frac{\mathbb{E} \left[d\mathcal{N}_{ij}(\tau) | \mathcal{H}_{\tau^-} \right]}{d\tau}. \quad (56)$$

Censoring of τ_{ij} by $\mathcal{I}_i(\tau)$, $\mathcal{I}_{ij}(\tau)$, and $\mathcal{Y}_{ij}(\tau)$ is independent if

$$\lambda_{ij}^{\mathcal{H}}(\tau) = \lambda_{ij}^{\mathcal{G}}(\tau). \quad (57)$$

for all τ and all ij in which an infectious contact interval could have been observed in the absence of censoring. Informally, this means that the censoring processes up to infectiousness age τ contain no further information about τ_{ij} given that $\tau_{ij} \geq \tau$.

The independence assumptions made in Section 1.2 were a simple way of guaranteeing independent censoring of τ_{ij} . These assumptions can be relaxed as long as equation (57) is not violated. For example, the observation process $\mathcal{Y}_{ij}(\tau)$ can be defined more flexibly than it was in Section 2.1. Let $\mathcal{T}_{ij}$ the infectiousness age of i at which we stop observing the pair ij , so $\mathcal{Y}_{ij}(\tau) = 1_{\tau \leq \mathcal{T}_{ij}}$. When who-infects-whom is observed, $\mathcal{T}_{ij}$ can be any stopping time with respect to a history generated by $\mathcal{I}_i(\tau)$, $\mathcal{I}_{ij}(\tau)$, and other processes that independently censor τ_{ij} . This flexibility extends to the marginal Nelson-Aalen estimator as long as $\mathcal{Y}_{ij}(t_j - t_i - \varepsilon_i) = 1$ for all i that might have infected j . Without this, we cannot assign probabilities to the possible $i \in \mathcal{V}_j$ using an estimate of $\Lambda_0(\tau)$. This condition is satisfied by the $\mathcal{Y}_{ij}(\tau)$ in Section 2.1, but other observation processes are possible.

B. Martingale central limit theorem

Here, we use the version of the martingale central limit theorem given in Aalen et al. (2009). Let $Y^{(m)}(\tau)$ denote the $Y(\tau)$ calculated from observations of m ordered pairs ij in which transmission might have taken place, and let $\widehat{\Lambda}^{(m)}(\tau)$ be the corresponding Nelson-Aalen estimate. Then

$$E^{(m)}(\tau) = \sqrt{m} \left(\widehat{\Lambda}^{(m)}(\tau) - \Lambda_0(\tau) \right) = \int_0^\tau \frac{\sqrt{m}}{Y^{(m)}(u)} 1_{Y^{(m)}(u) > 0} dM(u). \quad (58)$$

is a zero-mean martingale for each m . The predictable variation process for $E^{(m)}(\tau)$ is

$$\langle E^{(m)} \rangle(\tau) = \int_0^\tau \frac{m}{Y^{(m)}(u)} 1_{Y^{(m)}(u) > 0} \lambda_0(u) du. \quad (59)$$

For $\epsilon > 0$, let $E_\epsilon^{(m)}(\tau)$ be a martingale that contains all jumps of $E^{(m)}(\tau)$ of size $> \epsilon$. Since all jumps in $M(\tau)$ have size one,

$$E_\epsilon^{(m)}(\tau) = \int_0^\tau \frac{\sqrt{m}}{Y^{(m)}(u)} 1_{Y^{(m)}(u) > 0} 1_{\frac{\sqrt{m}}{Y^{(m)}(u)} > \epsilon} du, \quad (60)$$

and its predictable variation process is

$$\langle E_\epsilon^{(m)} \rangle (\tau) = \int_0^\tau \frac{m}{Y^{(m)}(u)} 1_{Y^{(m)}(u) > 0} 1_{\frac{\sqrt{m}}{Y^{(m)}(u)} > \epsilon} \lambda_0(u) \, du. \quad (61)$$

By the martingale central limit theorem, $E^{(m)}(\tau)$ will converge in distribution to a mean-zero Gaussian martingale if the following two conditions are satisfied:

- a. $\langle E^{(m)} \rangle (\tau) \xrightarrow{P} V(\tau)$ for all $\tau \in [0, \mathcal{T}]$ as $m \rightarrow \infty$, where $V(\tau)$ is a strictly increasing continuous function with $V(0) = 0$
- b. $\langle E_\epsilon^{(m)} \rangle (\tau) \xrightarrow{P} 0$ for all $\tau \in [0, \mathcal{T}]$ and all $\epsilon > 0$ as $m \rightarrow \infty$.

Both conditions will be satisfied if

$$\frac{Y^{(m)}(\tau)}{m} \rightarrow \bar{Y}(\tau) \quad \text{for all } \tau \in [0, \mathcal{T}], \quad (62)$$

where $\bar{Y}(\tau)$ is a strictly positive function. Then $E^{(m)}(\tau)$ converges in distribution to the mean-zero martingale

$$E^{(\infty)}(\tau) = W \left(\int_0^\tau \frac{\lambda_0(u)}{\bar{Y}(u)} \, du \right) \quad \text{for } \tau \in [0, \mathcal{T}], \quad (63)$$

where W is a Wiener process. Rates of convergence other than $\sqrt{m}$ are possible. If $m^{-2\alpha} Y^{(m)}(\tau)$ converges to a positive deterministic function, then $m^\alpha (\widehat{\Lambda}^{(m)}(\tau) - \Lambda_0(\tau))$ converges to a mean-zero Gaussian martingale.

For asymptotic normality, it does not matter whether we observe many pairs ij from the same epidemic or from multiple epidemics. The approximate confidence intervals from Section 2 can be used for data from a single large epidemic, as in Section 3, or from a large household surveillance study, as in Section 4.

Since $Y(\tau)$ is not affected by who-infected-whom, the martingale central limit theorem should hold for the marginal Nelson-Aalen estimate $\widehat{\Lambda}(\tau)$ under conditions very similar to those given above. However, a rigorous asymptotic analysis of $\widehat{\Lambda}(\tau)$ is complicated by fact that we need the smoothed hazard to approach the true hazard as $m \rightarrow \infty$.

C. Variance of $\Lambda^{\mathbf{v}}(\tau)$

By definition,

$$\text{Var}(\widehat{\Lambda}_{\mathbf{v}}(\tau)) = \mathbb{E}[\widehat{\Lambda}_{\mathbf{v}}(\tau)^2] - \mathbb{E}[\widehat{\Lambda}_{\mathbf{v}}(\tau)]^2. \quad (64)$$

Let $\mathcal{V}_j(\tau) = \{i \in \mathcal{V}_j : t_j - t_i - \epsilon_i \leq \tau\}$. Since each person has only one infector $v_j \in \mathcal{V}_j$ for any realization of $\mathbf{v}$ and the infectors of different individuals can be chosen independently,

$$\mathbb{E}[\widehat{\Lambda}_{\mathbf{v}}(\tau)^2] = \int_0^\tau \frac{1_{Y(u) > 0}}{Y(u)^2} \, d\widetilde{N}(\tau) + \left(\sum_{j=1}^n \sum_{i \in \mathcal{V}_j(\tau)} \frac{1_{Y(t_j - t_i - \epsilon_i) > 0}}{Y(t_j - t_i - \epsilon_i)} p_{ij}(\bar{\lambda}) \right)^2 \quad (65)$$

$$-\sum_{j=1}^n \left(\sum_{i \in \mathcal{Y}_j(\tau)} \frac{1_{Y(t_j - t_i - \varepsilon_i) > 0}}{Y(t_j - t_i - \varepsilon_i)} p_{ij}(\bar{\lambda}) \right)^2 \quad (66)$$

For the second term of equation (64), we get

$$\mathbb{E}[\widehat{\Lambda}_{\mathbf{v}}(\tau)]^2 = \left(\sum_{j=1}^n \sum_{i \in \mathcal{Y}_j(\tau)} \frac{1_{Y(t_j - t_i - \varepsilon_i) > 0}}{Y(t_j - t_i - \varepsilon_i)} p_{ij}(\bar{\lambda}) \right)^2 \quad (67)$$

Therefore,

$$\text{Var}(\widehat{\Lambda}_{\mathbf{v}}(\tau)) = \int_0^\tau \frac{1_{Y(u) > 0}}{Y(u)^2} d\widetilde{N}(\tau) - \sum_{j=1}^n \left(\sum_{i \in \mathcal{Y}_j(\tau)} \frac{1_{Y(t_j - t_i - \varepsilon_i) > 0}}{Y(t_j - t_i - \varepsilon_i)} p_{ij}(\bar{\lambda}) \right)^2, \quad (68)$$

which is equivalent to equation (31).

References

- Aalen O. Nonparametric inference for a family of counting processes. *Annals of Statistics*. 1978; 6:701–726.
- Aalen, OO.; Borgan, O.; Gjessing, H. *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag; New York: 2009. *Statistics for Biology and Health*.
- Altshuler B. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*. 1970; 6:1–11.
- Becker, NG. *Analysis of Infectious Disease Data*. Chapman & Hall/CRC; Boca Raton, FL: 1989. *Monographs on Statistics and Applied Probability*.
- Becker NG. Uses of the EM algorithm in the analysis of data on HIV/AIDS and other diseases. *Statistical Methods in Medical Research*. 1997; 6:24–37. [PubMed: 9185288]
- Carcione D, Giele GM, Goggin LS, Kwan KS, Smith DW, Dowse GK, Mak DB, Effler P. Secondary attack rate of pandemic influenza A(H1N1) 2009 in Western Australia, 29 May–7 August 2009. *Eurosurveillance*. 2011; 16:19765. [PubMed: 21262182]
- Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, Kent CK, Finelli L, Ferguson NM. Household transmission of 2009 pandemic influenza A(H1N1) virus in the United States. *New England Journal of Medicine*. 2009; 361:2619–2627. [PubMed: 20042753]
- Cox DR. 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 incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–38.
- Donnelly CA, Finelli L, Cauchemez S, Olsen SJ, Doshi S, Jackson ML, Kennedy ED, Kamimoto L, Marchbanks TL, Morgan OW, Patel M, Swerdlow DL, Ferguson NM, the PH1N1 Household Investigations Working Group. Serial intervals and the temporal distribution of secondary infections within households of 2009 pandemic influenza A(H1N1): implications for influenza control recommendations. *Clinical Infectious Diseases*. 2011; 52:S123–S130. [PubMed: 21342883]
- Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*. 2005; 437:209–214. [PubMed: 16079797]
- Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. *Nature*. 2006; 442:448–452. [PubMed: 16642006]
- Fine PEM. The interval between successive cases of an infectious disease. *American Journal of Epidemiology*. 2003; 158:1039–1047. [PubMed: 14630599]

- France AM, Jackson M, Schrag S, Lynch M, Zimmerman C, Biggerstaff M, Hadler J. Household transmission of 2009 influenza A(H1N1) virus after a school-based outbreak in New York City, April–May 2009. *Journal of Infectious Diseases*. 2010; 201:984–992. [PubMed: 20187740]
- Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*. 2009; 324:1557–1561. [PubMed: 19433588]
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. Second ed.. John Wiley & Sons; Hoboken, NJ: 2002. Wiley Series in Probability and Statistics..
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Kenah E. Contact intervals, survival analysis of epidemic data, and estimation of R_0 . *Biostatistics*. 2011; 12:548–566. [PubMed: 21071607]
- Kenah E, Lipsitch M, Robins JM. Generation interval contraction and epidemic data analysis. *Mathematical Biosciences*. 2008; 213:71–79. [PubMed: 18394654]
- Lee BL, Kosorok MR, Fine JP. The profile sampler. *Journal of the American Statistical Association*. 2005; 100:960–969.
- Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. Transmission dynamics and control of severe acute respiratory syndrome. *Science*. 2003; 300:1966–1970. [PubMed: 12766207]
- McBryde ES, Bergeri I, van Gemert C, Rotty J, Headley EJ, Simpson K, Lester RA, Hellard M, Fielding JE. Early transmission characteristics of influenza A(H1N1)v in Australia: Victorian State, 16 May–3 June 2009. *Eurosurveillance*. 2009; 14:19363. [PubMed: 19883544]
- Mills C, Robins JM, Lipsitch M. Transmissibility of 1918 pandemic influenza. *Nature*. 2004; 432:904–906. [PubMed: 15602562]
- Morgan OW, Parks S, Shim T, Blevins PA, Lucas P, Sanchez R, Walea N, Loustalot F, Duffy MR, Shim MJ, Guerra S, Guerra F, Mills G, Verani J, Alsip B, Lindstrom S, Shu B, Emery S, Cohen AL, Menon M, Fry AM, Dawood F, Fonseca VP, Olsen SJ. Household transmission of pandemic (H1N1) 2009, San Antonio, Texas, USA, April–May 2009. *Emerging Infectious Diseases*. 2010; 16:631–637. [PubMed: 20350377]
- Müller H-G, Wang J-L. Hazard rate estimation under random censoring with varying kernels and bandwidth. *Biometrics*. 1994; 50:61–76. [PubMed: 8086616]
- Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics*. 1972; 14:945–966.
- Ramsey AH Jr, Longini IM Jr, Haber M, Monto AS. A discrete-time model for the statistical analysis of infectious disease incidence data. *Biometrics*. 1992; 48:117–128. [PubMed: 1316178]
- Rhodes PH, Halloran ME, Longini IM 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.
- Roberts MG, Heesterbeek JAP. Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. *Journal of Mathematical Biology*. 2007; 55:803–816. [PubMed: 17684743]
- Sinha D, Dey DK. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*. 1997; 92:1195–1211.
- Sugimoto JD, Borse NN, Ta ML, Stockman LJ, Fischer GE, Yang Y, Halloran ME, Longini IM Jr, Duchin JS. The effect of age on transmission of 2009 pandemic influenza A(H1N1) in a camp and associated households. *Epidemiology*. 2011; 22:180–187. [PubMed: 21233714]
- Sugimoto, JD.; Yang, Y.; Halloran, ME.; Dean, B.; Ojulfstad, B.; Bagwell, DA.; Mascola, L.; Bancroft, E.; Longini, IM, Jr. Accounting for unobserved immunity and asymptomatic infection in the early household transmission of the pandemic influenza A (H1N1) 2009. 2011. Submitted
- Svensson Å. A note on generation times in epidemic models. *Mathematical Biosciences*. 2007; 208:300–311. [PubMed: 17174352]
- Van der Vaart, AW. *Asymptotic Statistics*. Cambridge University Press; Cambridge, UK: 1998.

- Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B*. 2007; 274:599–604. [PubMed: 17476782]
- Wallinga J, Teunis P. 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 DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. *Nature*. 1998; 393:440–442. [PubMed: 9623998]
- Wegman EJ, Wright IW. Splines in statistics. *Journal of the American Statistical Association*. 1983; 78:351–365.
- White LF, Pagano M. 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 Y, Sugimoto J, Halloran ME, Basta NE, Chao DL, Matrajt L, Potter G, Kenah E, Longini IM Jr. The transmissibility and control of pandemic influenza A(H1N1) virus. *Science*. 2009; 326:729–733. [PubMed: 19745114]

Fig. 1.

Estimates of the cumulative hazard function (top) and survival function (bottom) for network-based models with a Weibull(.5, 1) contact interval distribution, when who-infects-whom is observed (left) and unobserved (right). The true cumulative hazard and survival functions are indicated with dashed lines. The point clouds are point estimates at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of all possible contact intervals in each of 1,000 simulations. The point clouds for the 5th and 10th percentiles and the 90th and 95th percentiles run together, leaving five separate clouds.

Fig. 2.

Estimates of the cumulative hazard function (top) and survival function (bottom) for network-based models with a Weibull(2, 1) contact interval distribution, when who-infects-whom is observed (left) and unobserved (right). The true cumulative hazard and survival functions are indicated with dashed lines. The point clouds are point estimates at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of all possible contact intervals in each of 1,000 simulations. The point clouds for the 5th and 10th percentiles run together, leaving six separate clouds.

Fig. 3.

Estimates of the cumulative hazard function (top) and survival function (bottom) for mass-action models with a Weibull(.5, 5) normalized contact interval distribution using the methods from Section 2.3, when who-infects-whom is observed (left) and unobserved (right). The true cumulative hazard and survival functions are indicated with dashed lines. The point clouds are point estimates at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of the possible contact intervals in each of 1,000 simulations. The point clouds for the 5th and 10th percentiles and the 90th and 95th percentiles run together, leaving five separate clouds.

Fig. 4.

Estimates of the cumulative hazard function (top) and survival function (bottom) for mass-action models with a Weibull(2, 1) normalized contact interval distribution using the methods from Section 2.3, when who-infects-whom is observed (left) and unobserved (right). The true cumulative hazard function (top) and survival function (bottom) are indicated with dashed lines. The point clouds are point estimates at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of the possible contact intervals in each of 1,000 simulations. The point clouds for the 5th and 10th percentiles run together, leaving six separate clouds.

Influenza A(H1N1) contact intervals in LA county

Fig. 5.

A marginal Nelson-Aalen estimate of the contact interval cumulative hazard function based on the Los Angeles County household data. For comparison, we have parametric estimates of the cumulative hazard function assuming exponential and Weibull contact interval distributions. The assumed latent period for all analyses was 2 days, so symptom onset occurs on day 2 after the onset of infectiousness. Note the large jumps on days 1, 2, and 3 following the onset of infectiousness and the small jumps on days 5 and 6.

Fig. 6.

Sensitivity analyses for the nonparametric analysis of the LA county household data. There is little sensitivity to the assumed incubation and infectious periods. There is greater sensitivity to the assumed latent period, but there is good evidence for transmission on the day prior to symptom onset and little evidence of significant transmission 2 days prior to symptom onset (Cauchemez et al., 2009; Donnelly et al., 2011).

Table 1

95% confidence interval coverage probabilities in network-based simulations.

Data percentile	Contact interval distribution			
	Weibull(.5, 1)		Weibull(2, 1)	
	Nelson-Aalen	marginal NA	Nelson-Aalen	marginal NA
5	.949 (.933, .962)	.946 (.930, .959)	.960 (.946, .971)	.903 (.883, .921)
10	.954 (.939, .966)	.952 (.937, .964)	.963 (.949, .974)	.901 (.881, .919)
25	.957 (.943, .969)	.947 (.931, .960)	.952 (.937, .964)	.929 (.911, .944)
50	.952 (.937, .964)	.940 (.923, .954)	.944 (.928, .957)	.924 (.906, .940)
75	.948 (.932, .961)	.944 (.928, .957)	.958 (.944, .970)	.958 (.944, .970)
90	.935 (.918, .949)	.944 (.928, .957)	.956 (.941, .968)	.939 (.922, .953)
95	.934 (.917, .949)	.938 (.921, .952)	.942 (.926, .956)	.935 (.918, .949)
	Kaplan-Meier		Kaplan-Meier	
	marginal KM		marginal KM	
5	.949 (.933, .962)	.946 (.930, .959)	.960 (.946, .971)	.903 (.883, .921)
10	.954 (.939, .966)	.952 (.937, .964)	.964 (.951, .975)	.901 (.881, .919)
25	.957 (.943, .969)	.947 (.931, .960)	.952 (.937, .964)	.929 (.911, .944)
50	.952 (.937, .964)	.940 (.923, .954)	.944 (.928, .957)	.925 (.907, .941)
75	.948 (.932, .961)	.945 (.929, .958)	.960 (.946, .971)	.957 (.942, .969)
90	.934 (.917, .949)	.944 (.928, .957)	.955 (.940, .967)	.940 (.923, .954)
95	.933 (.916, .948)	.938 (.921, .952)	.943 (.927, .957)	.936 (.919, .950)

Table 2

95% confidence interval coverage probabilities for mass-action simulations.

Data percentile	Scaled contact interval distribution			
	Weibull(.5, 5)		Weibull(2, 1)	
	Nelson-Aalen	marginal NA	Nelson-Aalen	marginal NA
5	.949 (.933, .962)	.003 (.001, .009)	.689 (.659, .718)	0 (0, .004)
10	.957 (.943, .969)	.004 (.001, .010)	.960 (.946, .971)	0 (0, .004)
25	.953 (.938, .965)	.011 (.006, .020)	.963 (.949, .974)	0 (0, .004)
50	.951 (.936, .964)	.056 (.043, .072)	.944 (.928, .957)	0 (0, .004)
75	.934 (.917, .949)	.842 (.818, .864)	.963 (.949, .974)	.001 (.000, .006)
90	.932 (.915, .947)	.091 (.074, .111)	.959 (.945, .970)	.178 (.155, .203)
95	.941 (.925, .955)	.035 (.024, .048)	.945 (.929, .958)	.004 (.001, .010)
	Kaplan-Meier		Kaplan-Meier	
	marginal KM		marginal KM	
5	.950 (.935, .963)	.003 (.001, .009)	.689 (.659, .718)	0 (0, .004)
10	.958 (.944, .970)	.004 (.001, .010)	.960 (.946, .971)	0 (0, .004)
25	.953 (.938, .965)	.011 (.006, .020)	.963 (.949, .974)	0 (0, .004)
50	.951 (.936, .964)	.056 (.043, .072)	.944 (.928, .957)	0 (0, .004)
75	.933 (.916, .948)	.842 (.818, .864)	.963 (.949, .974)	.001 (.000, .006)
90	.930 (.912, .945)	.091 (.074, .111)	.959 (.945, .970)	.178 (.155, .203)
95	.941 (.925, .955)	.035 (.024, .048)	.938 (.921, .952)	.004 (.001, .010)