Published in final edited form as:
Comput Stat Data Anal. 2007 August 15; 51(12): 6582–6595.

A Data-Augmentation Method for Infectious Disease Incidence Data from Close Contact Groups

Yang Yang[1,*], **Ira M. Longini Jr.**[1,2], and **M. Elizabeth Halloran**[1,2]

1 *Program of Biostatistics and Biomathematics, Division of Public Health Sciences Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA*

2 *Department of Biostatistics, University of Washington, Seattle, WA 98195, USA yang@scharp.org*

Abstract

A broad range of studies of preventive measures in infectious diseases gives rise to incidence data from close contact groups. Parameters of common interest in such studies include transmission probabilities and efficacies of preventive or therapeutic interventions. We estimate these parameters using discrete-time likelihood models. We augment the data with unobserved pairwise transmission outcomes and fit the model using the EM algorithm. A linear model derived from the likelihood based on the augmented data and fitted with the iteratively re-weighted least squares method is also discussed. Using simulations, we demonstrate the comparable accuracy and lower sensitivity to initial estimates of the proposed methods with data augmentation relative to the likelihood model based solely on the observed data. Two randomized household-based trials of zanamivir, an influenza antiviral agent, are analyzed using the proposed methods.

Keywords

Antiviral agent; Data augmentation; EM algorithm; Infectious disease; Intervention efficacy; Linear model; MLE

1 Introduction

Close contact groups, such as households, are the important places of transmission for many infectious diseases. Data collected from these contact groups provide a basis for evaluating person-to-person transmission risks and effectiveness of intervention methods such as antiviral treatments or vaccine (Halloran, Struchiner and Longini, 1997; Becker, Britton and O'Neill, 2003). Using different levels of information available in the data, various statistical methods have been developed for data analysis. If only the final infection status of participants are known, methods utilizing recursive final-size probabilities can be applied, including likelihood maximization (Longini and Koopman, 1982; Addy, Longini and Haber, 1991), Bayesian approaches (O'Neill and Roberts, 1999), generalized linear models (Magder and Brookmeyer, 1993), and estimating equations with martingale techniques (Becker and Hasofer, 1997). In many modern clinical trials, sequential laboratory tests and symptom diary of participants provide time-to-event data with individual-specific longitudinal exposure information. To take into account exposure and transmission dynamics at the individual level, Rampey et al.

*To whom correspondence should be addressed.

(1992) constructed discrete-time likelihoods based on assumptions about the natural history of the disease such as the distributions of the latent and infectious periods. Yang, Longini and Halloran (2006) extended this method to the more realistic case-ascertained design. Cauchemez et al. (2004) proposed a Bayesian model with the flexibility of estimating the natural history of the disease, but time-dependent covariates have not been accommodated.

The discrete-time likelihoods in Rampey et al. (1992) and Yang et al. (2006) are built solely upon the observed data, including symptom onset dates, laboratory test results and household structure (which individuals live in which households), and involve summing probability components over the latent period. Summations or integrals are commonly seen in likelihoods based solely on the observed data, and such complicated structure may present difficulties for standard analyses or prevent extension by other methods (O'Neill et al., 2000). More importantly, when data are sparse because of rare incidences and/or a multicovariate structure, iterative estimation procedures (e.g., the Newton-Raphson algorithm) using only the observed data may be sensitive to the initial estimates in locating the maximum likelihood estimates (MLEs). This fact can be seen in section 3 and 4 of this paper, and is also mentioned in Yang et al. (2006). Data augmentation is a popular technique to circumvent computational difficulties in classical likelihood methods because likelihood functions conditional on unobserved variables are often simpler (van Dyk and Meng, 2001; Paap, 2002). In a transmission model for infectious diseases, a basic element is the transmission probability given a contact between an infective person and a susceptible person. The contact may be defined in various ways, for example, one day of living in the same household. The outcome of each contact, infection or escape, is generally not observable since a person may make multiple contacts before infection. In this paper, we revise the discrete-time likelihood in Yang et al. (2006) by augmenting the observed symptom onset data with the unobserved transmission outcome for each contact. This likelihood based on the augmented data has a simpler form than the one based on only the observed data and can be maximized with the EM algorithm. To illustrate the potential use of the simple likelihood by a different method, we derive a linear model that can be fitted using the iteratively re-weighted least squares (IRLS) procedure. We show via simulation studies that both the maximum likelihood (ML) and the IRLS methods using the augmented data are less sensitive to initial estimates as compared to the ML method using only the observed data in Yang et al. (2006). We use the proposed approaches to estimate the prophylactic and treatment effectiveness of an influenza antiviral agent in two household trials.

2 Methods

Suppose that the disease under investigation is influenza and the data arise from a clinical trial in which household members are randomized to either an antiviral agent or control when an index case is identified by clinical symptoms. Let us assume the antiviral agent provides temporary protection for susceptible contacts and therapy for cases. In the discrete-time likelihood model setting, risks are evaluated for each susceptible participant in each time interval. Suppose that the time intervals are consecutive days, and define a contact as the exposure of a susceptible person to an infective person in the same household throughout a day. The pairwise transmission probability per contact between a susceptible person i with covariates x_i and an infective person j with covariates x_j in the same household is expressed as $p(x_i, x_j)$. If x_i and x_j are scalars denoting treatment status of antiviral agent (1=yes, 0=no), then one can define efficacy measures $\mathrm{AVE}_S = 1 - \frac{p(1,0)}{p(0,0)}$, $\mathrm{AVE}_I = 1 - \frac{p(0,1)}{p(0,0)}$ and $\mathrm{AVE}_T = 1 - \frac{p(1,1)}{p(0,0)}$, where in the epidemiological literature AVE_S measures the antiviral efficacy in reducing susceptibility, AVE_I measures the efficacy in reducing infectiousness, and AVE_T is called the total effectiveness (Halloran et al., 1997). Let $p = p(0, 0)$ be the baseline daily pairwise transmission probability without any treatment. For notational convenience, a

reparameterization leads to $p\left(x_i, x_j\right) = \theta^{x_i(1-x_j)}\varphi^{(1-x_i)x_j}\eta^{x_i x_j}p$ where $\theta = 1 - \text{AVE}_S$, $\phi = 1 - \text{AVE}_I$ and $\eta = 1 - \text{AVE}_T$. For simplicity, we assume multiplicativity of θ and ϕ such that $\eta = \theta\phi$, and thus $p\left(x_i, x_j\right) = \theta^{x_i}\varphi^{x_j}p$. In Yang et al. (2006), we explored the assumption of multiplicativity for the ML method using only the observed data.

As our interest centers around estimation of transmission probabilities and treatment efficacies, we assume that: 1. the latent period (time from infection to being infectious) coincides with the incubation period (time from infection to the onset of symptoms); and 2. durations of the latent and the infectious periods have known probability distributions. If the latent and the incubation periods do not coincide but are both known, the model can be adjusted for such situation.

2.1 The Maximum Likelihood Method Based on the Augmented Data

Suppose that the trial is conducted on a population of size N and is observed on a daily basis from day 1 to day T. Let us assume day 1 is the first calendar day of exposure for the whole study population. The observed data for each subject include household membership, the date of symptom onset, laboratory test result, randomized treatment and treatment period as well as other characteristics such as age and gender. On day t, the probability that an infective person j with treatment status $r_j(t)$ (0: untreated, 1: treated) infects a susceptible person i with treatment status $r_i(t)$ in the same household is expressed as

$$p_{ji}(t) = \theta^{r_i(t)}\varphi^{r_j(t)}pf\left(t \mid \tilde{t}_j\right), \tag{1}$$

where $f(t|\tilde{t}_j)$ is the probability that person j stays infectious on day t given the day of symptom onset t_j and is derived from the known distribution of the infectious period. For simplicity in notation, we use t_i to denote the observed symptom onset time for each person, although t_i is right-censored for those who are free of symptoms up to day T. We allow a constant common infective source from outside of the household, by setting $p_{ci}(t) = \theta^{r_i(t)}b$, where c refers to the common source, and b is the baseline probability of being infected by the common source per day. Let $\psi_j = 1$ if the infective source j is a person and 0 if $j = c$. A modification of (1) takes into account the common source as the following

$$p_{ji}(t) = \theta^{r_i(t)}\varphi^{r_j(t)}p^{\psi_j}b^{1-\psi_j}f\left(t \mid \tilde{t}_j\right), \tag{2}$$

where $f_c(t|\tilde{t}_c) = 1$ and $r_c(t) = 0$ for all t. A likelihood involving only the observed data, $\{\tilde{t}_i : 1 \le t \le T, 1 \le i \le N\}$, can be constructed from (2) and the known distribution of the latent period as in Yang et al. (2006).

Let $Y_{ji}(t)$ be the transmission result (1:infection, 0:escape) between an infective source j and a susceptible person i on day t. Let l_{max} and l_{min} be the maximum and minimum duration of the latent period, so that $\underline{t}_i = \tilde{t}_i - l_{max}$ and $\bar{t}_i = \tilde{t}_i - l_{min}$ are the earliest and latest potential infection days for person i. Given the observed symptom onset day t_i, the sequence of $Y_{ji}(t)$'s for $t \ge t_i$ remains unknown. It should be noted that $Y_{ji}(t)$ is a random variable only if $Y_{ji}(\tau) = 0$ for all $\tau < t$, and $Y_{ji}(t)$ is independent of $Y_{ki}(t)$ for the same day t. Define

$$Z_{ji}(t) = Y_{ji}(t)\prod_{k\in D_i,\tau<t}\left(1 - Y_{ki}(\tau)\right)$$

and

$$\bar{Z}_{ji}(t) = \left(1 - Y_{ji}(t)\right)\prod_{k\in D_i,\tau<t}\left(1 - Y_{ki}(\tau)\right),$$

where D_i is the collection of potential infective sources for person i, i.e., people living in the same household with person i plus the external common source. $Z_{ji}(t) = 1$ is the event that person i escapes infection from any source before day t but is infected by source j on day t, while $\bar{Z}_{ji}(t) = 1$ is the event that person i escapes infection from any source before day t and from source j on day t. Let $max_{j\in D_i} Z_{ji}(t)$ indicate if $Z_{ji}(t) = 1$ for any j on day t. The likelihood of the augmented data is

$$\begin{aligned} & L_i\left(b, p, \theta, \varphi \mid \tilde{t}_j, Z_{ji}(t), j \in D_i, t \le T\right) \\ = & \prod_{t=1}^{T}\left\{g\left(\tilde{t}_i \mid t\right)^{max_{j\in D_i} Z_{ji}(t)} \prod_{j\in D_i}\left(p_{ji}(t)\right)^{Z_{ji}(t)}\left(1-p_{ji}(t)\right)^{\bar{Z}_{ji}(t)}\right\} \\ \propto & \prod_{t=1}^{T}\prod_{j\in D_i}\left(p_{ji}(t)\right)^{Z_{ji}(t)}\left(1-p_{ji}(t)\right)^{\bar{Z}_{ji}(t)}\Bigg\}, \end{aligned} \tag{3}$$

where $g(\tilde{t}_i|t)$ denotes the probability of illness onset on day $\tilde{t}_i$ given infection on day t and is derived from the distribution of the latent period. According to our assumption, both $f(t|t_j)$ and $g(t_i|t)$ are known. This likelihood is a product of binomial probability components, much simpler than the one in Yang et al. (2006). To apply the EM algorithm, we need to determine the distributions of $Z_{ji}(t)$ and $\bar{Z}_{ji}(t)$ conditioning on current estimates of b, p, θ and ϕ as well as t_j, $j \in D_i$ (Dempster, Laird and Rubin, 1977). Define $S_i(t)$ as the event that person i has symptom onset on day t, $I_i(t)$ the event that person i is infected on day t and $I_{ji}(t)$ the event that person i is infected by j on day t. Then, the conditional distributions are given by (Appendix A)

$$\Pr\left(Z_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) = \begin{cases} \frac{\Pr(I_{ji}(t))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \times \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right), & t_i \le t < \bar{t}_i \\ 0, \text{otherwise} \end{cases} \tag{4}$$

and

$$\Pr\left(\bar{Z}_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) = \begin{cases} \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right) \times \{\Pr(I_i(t)) - \Pr(I_{ji}(t))\}}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} + \sum_{\tau=t+1}^{\bar{t}_i} \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right) \times \Pr(I_i(\tau))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)}, & t_i \le t < \bar{t}_i \\ 1, t < t_i \\ 0, \text{otherwise} \end{cases} . \tag{5}$$

Given estimates $\left(\hat{b}_{l-1}, \hat{p}_{l-1}, \hat{\theta}_{l-1}, \hat{\varphi}_{l-1}\right)$ from the $(l-1)^{th}$ iteration, in the l^{th} iteration we have

$$\begin{aligned} \Pr\left(I_{ji}(t)\right) &= \begin{cases} \hat{Q}_i(t-1)\hat{\theta}_{l-1}^{r_i(t)}\hat{\varphi}_{l-1}^{r_j(t)}\hat{p}_{l-1} f\left(t \mid \tilde{t}_j\right), & j \in D_i \\ \hat{Q}_i(t-1)\hat{\theta}_{l-1}^{r_i(t)}\hat{b}_{l-1}, & j = c \end{cases} \\ \Pr(I_i(t)) &= \hat{Q}_i(t-1)\left\{1 - \left(1-\hat{\theta}_{l-1}^{r_i(t)}\hat{b}_{l-1}\right)\prod_{j\in D_i}\left(1-\hat{\theta}_{l-1}^{r_i(t)}\hat{\varphi}_{l-1}^{r_j(t)}\hat{p}_{l-1} f\left(t \mid \tilde{t}_j\right)\right)\right\}, \\ \Pr\left(S_i\left(\tilde{t}_i\right)\right) &= \sum_{\tau=t_i}^{\bar{t}_i} \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right) \times \Pr(I_i(\tau)), \\ \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right) &= g\left(\tilde{t}_i \mid \tau\right), \end{aligned}$$

where $\hat{Q}_i(t-1)$ is the estimated cumulative escape probability based on $\left(\hat{b}_{l-1}, \hat{p}_{l-1}, \hat{\theta}_{l-1}, \hat{\varphi}_{l-1}\right)$ The likelihood history before day $\underline{t}_i$ can be dropped from $\Pr(I_{i,j}(t))$ and $\Pr(I_i(t))$, since $\hat{Q}_i\left(\underline{t}_i - 1\right)$ is the common factor and will eventually be cancelled out in the calculations of (4)

and (5). The implementation of the EM algorithm is straightforward. In the E-step, (4) and (5) are calculated and plugged into the logarithm of (3) to obtain

$$\log\left(L_i\left(b,p,\theta,\varphi \mid \tilde{t}_j, Z_{ji}(t), j\in D_i, t\le T\right)\right) \propto \sum_{t=1}^{T}\sum_{j\in D_i}\left\{\Pr\left(Z_{ji}(t)=1 \mid b,p,\theta,\varphi,\tilde{t}_i\right)\log\left(p_{ji}(t)\right)+\Pr\left(\bar{Z}_{ji}(t)=1 \mid b,p,\theta,\varphi,\tilde{t}_i\right)\log\left(1-p_{ji}(t)\right)\right\} \quad (6)$$

which is maximized in the M-step.

Variances of the parameter estimates can be evaluated using Louis' method (Louis, 1982). Let $\boldsymbol{Z}$ be the collection of $Z_{ji}(t)$, and $\boldsymbol{t}$ the collection of t_i, for all i, j and t, so that $\boldsymbol{t}$ is the observed data and $\boldsymbol{Z}$ is the partially latent data. Let $\lambda = \{b, p, \theta, \boldsymbol{\phi}\}$. Louis' method states that

$$\frac{\partial^2 \log\left(L\left(\lambda \mid \tilde{t}\right)\right)}{\partial\lambda^2} = \mathrm{E}_{\tilde{Z}\mid t,\lambda}\left\{-\frac{\partial^2 \log\left(L\left(\lambda \mid \tilde{t}, Z\right)\right)}{\partial\lambda^2}\right\} + \mathrm{VAR}_{\tilde{Z}\mid t,\lambda}\left\{-\frac{\partial \log\left(L\left(\lambda \mid \tilde{t}, Z\right)\right)}{\partial\lambda}\right\}.$$

The first component on the right side can be evaluated analytically based on (6), while the second component can be estimated via sampling from the distribution of $\boldsymbol{Z}$ conditioning on $\tilde{\boldsymbol{t}}$ and $\widehat{\lambda}$.

2.2 The Linear Model Based on the Augmented Data

A linear model is a natural consequence of modeling the daily pairwise transmissions. Taking the logarithm on both sides of (2),

$$\begin{aligned}\log\left(p_{ji}(t)\right) &= \log(b)+\psi_j\log\tfrac{p}{b}+r_i(t)\log(\theta)+r_j(t)\log(\varphi)+\log\left(f\left(t \mid \tilde{t}_j\right)\right)\\ &= \beta_0+\beta_1\psi_j+\beta_2 r_i(t)+\beta_3 r_j(t)+\log\left(f\left(t \mid \tilde{t}_j\right)\right).\end{aligned} \quad (7)$$

The response of this model is $Y_{ji}(t)$ since $p_{ji}(t)=\Pr\left(Y_{ji}(t)=1 \mid Y_{ki}(\tau)=0, k\in D_i, \tau<t\right)$. From (6), it is clear that one should assign weights $\Pr\left(Z_{ji}(t)=1 \mid b,p,\theta,\varphi,\tilde{t}_i\right)$ to the outcome $Y_{ji}(t) = 1$ and $\Pr\left(\bar{Z}_{ji}(t)=1 \mid b,p,\theta,\varphi,\tilde{t}_i\right)$ to the outcome $Y_{ji}(t) = 0$. As the weights need to be calculated from pre-estimated parameters, we use the iteratively re-weighted least squares (IRLS) method to fit the model.

To apply the IRLS method, suppose the conditional expected frequencies of $Y_{ji}(t)$'s have been summarized into H binomial proportions P_h, $h = 1, \ldots, H$, with the H covariate patterns defined by $r_i(t)$, $r_j(t)$, ψ_j and $f\left(t \mid \tilde{t}_j\right)$. We fit model (7) by minimizing the objective function $\Sigma_{h=1}^{H} w_h\left\{\log\left(\tilde{P}_h\right)-\log(P_h)\right\}^2$, the squared difference between the observed proportion $\tilde{P}_h$ and the mean proportion P_h. Let n_h be the number of observations in the h^{th} pattern. The weight for the h^{th} pattern $w_h = \mathrm{VAR}^{-1}\left(\log\left(\widehat{P}_h\right)\right)$ could be estimated from either $\tilde{P}_h$ (data-based) or the fitted response $\hat{P}_h$ (model-based). Our simulations suggest that combinations such as the arithmetical mean $\frac{1}{2}\left\{\frac{n_h\times\tilde{P}_h}{1-\tilde{P}_h}+\frac{n_h\times\widehat{P}_h}{1-\widehat{P}_h}\right\}$ or the geometric mean $n_h\sqrt{\frac{\tilde{P}_h\widehat{P}_h}{\left(1-\tilde{P}_h\right)\left(1-\widehat{P}_h\right)}}$ provide estimates close to the MLEs. If $\tilde{P}_h = 0$, we replace P_h by $\hat{P}_h$ from the previous iteration.

Let $\widehat{\beta}_0, \ldots, \widehat{\beta}_3$ be the WLS estimates of the coefficients in model (7), then the WLS estimates of the parameters at the l^{th} iteration are
$\widehat{b}_l = exp\left(\widehat{\beta}_0\right), \widehat{p}_l = exp\left(\widehat{\beta}_0 + \widehat{\beta}_1\right), \widehat{\theta}_l = exp\left(\widehat{\beta}_2\right)$, and $\widehat{\varphi}_l = exp\left(\widehat{\beta}_3\right)$.

We then update the parameters and re-fit the model until the estimates converge. We have generalized the linear model method to populations with heterogeneity in the transmission probabilities (Appendix B).

At each iteration, the variances of $\hat{b}_l$, $\hat{p}_l$, $\widehat{\theta}_l$ and $\widehat{\varphi}_l$ estimated from the linear model have been averaged over the conditional distribution of *Z*. With the loss of randomness in *Z*, the final estimates will under-estimate the true variances. Since
$\mathrm{VAR}\left(\widehat{\lambda}\right) = \mathrm{E}\left(\mathrm{VAR}\left(\widehat{\lambda} \mid Z\right)\right) + \mathrm{VAR}\left(\mathrm{E}\left(\widehat{\lambda} \mid Z\right)\right)$, similar to the Louis' method for the ML method, one can employ the following adjustment procedure to approximate $\mathrm{VAR}\left(\widehat{\lambda}\right)$:

- Sample ***Z*** from $\Pr\left(Z \mid t, \widehat{\lambda}\right)$, where $\widehat{\lambda}$ is the final parameter estimates.
- Use the sampled ***Z*** as the weights to fit model (7) and obtain new point estimates of the parameters and their variances.
- Repeat the previous steps for a sufficient number of times. The sample average of the newly-estimated variances approximates $\mathrm{E}\left(\mathrm{VAR}\left(\widehat{\lambda} \mid Z\right)\right)$, and the sample variance of the newly-estimated parameters approximates $\mathrm{VAR}\left(\mathrm{E}\left(\widehat{\lambda} \mid Z\right)\right)$.

3 Simulation Study

To compare the ML and IRLS methods using the augmented data with the ML method using only the observed data, we conducted simulations under two scenarios: with a large number of cases and with sporadic cases. A pseudo-community composed of households of size two or larger with 1000 people was generated according to the distributions of age and household sizes from the US Census 2000. The distribution of the simulated household sizes is {2 : 67%, 3 : 13%, 4 : 10%, 5 : 7%, 6 : 2%, 7 : 1%}. Simulated epidemics were stopped on day 100, the typical length of the influenza season for a community. The empirical latent and infectious period distributions, from which $f(t|t_i)$ and $g(t_i|t)$ were derived, were obtained from Elveback, Fox and Ackerman (1976) and given in Table 1. Our simulations were implemented with individual-level randomization of treatments, where individuals including index cases in the same household may receive different treatments. In the Newton-Raphson procedure for likelihood maximization, we apply the complementary log-log transformation for *b* and *p* and the log transformation for θ and ϕ to help improve convergence. One thousand stochastic replications were carried out for each scenario investigated.

We first set the values of the parameters to $b = 0.005$, $p = 0.1$, $\theta = 0.4$, $\phi = 0.7$. Under this setting, on average 69% of the households and 51% of the contacts were attacked in simulated epidemics, and 20% of the contacts were infected when receiving treatment. The three iterative procedures were initiated from the true values of the parameters and, with adequate numbers of events, converged most of the time. By convergence we mean that the estimates of all four parameters converge to reasonable values. Specifically, estimates of *b* and *p* in $(10^{-10}, 1)$ and estimates of θ and ϕ in $(10^{-10}, 10)$ are considered reasonable. Given convergence, the MLEs obtained from only the observed data are exactly the same as those obtained from the augmented data, and the estimates of the SDs are also similar. Therefore, we present only the MLEs obtained from the augmented data. Table 2 shows mean parameter estimates, Monte Carlo standard deviations (SD of point estimates), mean model-estimated SDs and coverage rates of 95% confidence intervals (CI) based on model-estimated SDs for the two approaches

using the augmented data. The IRLS method yielded about the same estimates of the parameters and SDs as the MLEs. The small differences between the IRLS estimates and the MLEs for b, $\mathrm{SD}(\hat{b})$ and $\mathrm{SD}(\widehat{\theta})$ decrease as the sample size increases (not shown).

To compare the sensitivity of the three methods to starting parameter values when data are sparse, we reduced the true values of b from 0.005 to 0.002 and p from 0.1 to 0.01 so as to reduce transmissions within households. Under this setting, the average attack rates decreased to 39% for households and to 12% for contacts, and only 10% of the contacts were infected when receiving treatment. We ran simulations under different starting values of b and p, as $\log(p_{ji}(t))$ is generally more sensitive to the transmission probabilities than to the efficacies. Simulation results including convergence rates and parameter estimates are compared in Table 3. Clearly, the ML method using only the observed data is highly sensitive to initial values of b and p. The convergence rate of the ML method using only the observed data was comparable to the methods using the augmented data when the iteration started from the true parameters, but dropped dramatically when starting from larger values ($b = 0.02$, $p = 0.1$) or smaller values ($b = 0.0002$, $p = 0.001$) of the probability parameters. In contrast, the convergence rate was relatively stable for the approaches using the augmented data, regardless of the starting values. Parameter estimates and associated Monte Carlo standard deviations were similar across methods, except that the IRLS method appeared to overestimate θ to a larger extent compared to the ML methods. All methods overestimated ϕ as a consequence of sparse data. In addition, the ML methods overestimated, while the IRLS method underestimated, the standard deviation of ϕ. For example, when starting from true values of b and p, the mean standard errors are 1.10, 1.16 and 0.78 (not shown in Table 3) for the MLE based on the observed data, the MLE based on the augmented data and the IRLS estimate of ϕ respectively, in contrast to Monte Carlo standard deviations 0.95, 0.96 and 0.93.

As seen in Table 3, sparse data generally lead to biased and unstable efficacy estimates for the parametric methods, particularly for the IRLS method. At the same time, sparse data also increase the chance of non-convergence for the standard likelihood maximization algorithms. Household-level randomization, in which individuals in the same household receive the same treatments, provides much less information for estimating θ and ϕ separately compared to individual-level randomization with the same population size. More discussion on trial design issues can be found in Donner (1998),Datta, Halloran and Longini (1999),Halloran et al. (2006) and Yang et al. (2006).

4 Data Analysis

Two randomized multi-center efficacy trials of zanamivir, an inhaled influenza antiviral agent, were conducted during October 1998 - April 1999 (Hayden et al., 2000) and June 2000 - April 2001 (Monto et al., 2002). In both trials, households were randomized to zanamivir or placebo but only eligible household members (aged 5+ years) were treated. In the later trial, index cases were not treated. Characteristics of the two trials are given in Table 4.

The earlier trial adopted a typical household-level randomization, providing information about $\mathrm{AVE}_T = 1 - \theta\phi$, if we assume multiplicativity between θ and ϕ, and the later trial contains information mainly about AVE_S. Neither trial alone provides any information about AVE_I, and thus we combine the two trials to estimate AVE_S and AVE_I simultaneously. While transmission probabilities and antiviral efficacies might differ from center to center, the limited sample size prohibits estimation of centerspecific parameters. As a result, we assume all the centers in both trials share the same parameters. The two reference papers used slightly different definition for clinical symptoms. We used the one in Monto et al.(2002) for both trials, i.e., presence of at least two of temperature≥ 37.8° C or feverishness (counted as one), cough, headache, sore throat and myalgia. As it is well known that influenza is more transmissible

among children, we assume age-specific transmission probabilities in two age groups, children (< 18) and adults (≥ 18). Our primary endpoint is laboratory-confirmed influenza with clinical symptoms (clinical infection). Households in both trials were followed from the ascertainment time of index cases, for which selection bias was adjusted for based on Yang et al. (2006) and Appendix C. In such adjustment, index cases were excluded from analyses regardless of laboratory results, but their effects on the exposure level of the contacts were considered.

Results are given in Table 5. For this data set, both ML methods converge and thus give the same MLEs. Prophylaxis with zanamivir led to significantly preventive efficacy against clinical infection by $\widehat{\mathrm{AVE}}_S = 0.75$ (95% C.I.=(0.56, 0.86)). Hence, a susceptible person taking zanamivir has his chance of developing influenza illness reduced by 75% per daily exposure to an untreated symptomatic infected person. Zanamivir did not show significant efficacy in reducing the infectiousness of infected people with $\widehat{\mathrm{AVE}}_I = 0.23$ (95% C.I.=(−1.33, 0.75)). Assuming multiplicativity of θ and ϕ, the total efficacy AVE_T reached 0.81 (95% C.I.=(0.50, 0.93)). Based on final data of clinical influenza illness provided in Hayden et al. (2000) and Monto et al. (2002), similar AVE_T (0.80; 95% C.I.=(0.53, 0.91)) and AVE_S (0.84; 95%C.I.= (0.61, 0.90)) were reported by Halloran et al. (2006). They also reported AVE_S (0.75; 95% C.I.=(0.54, 0.86)), AVE_I (0.19; 95% C.I.=(−1.60, 0.75)) and AVE_T (0.87; 95% C.I.=(0.63, 0.95)) based on secondary attack rates (SAR) during 2-7 days since the ascertainment of index cases. These results differ in their interpretation.

The estimated probability of infection from the common source per daily exposure is 0.0028 for children and 0.0010 for adults. Within households, the daily pairwise transmission probability is also higher in children $(\widehat{p}_{cc} = 0.040)$ than in adults $(\widehat{p}_{aa} = 0.032)$. These estimates of transmission probabilities are comparable to those found in two trials of oseltamivir, another influenza antiviral agent, conducted about the same time in North America and Europe (Yang et al., 2006).

The IRLS estimates are fairly close to the MLEs except for p_{aa} and θ. In addition, the IRLS method might have under-estimated the SD for ϕ. The two trials combined together still do not provide sufficient information for estimating ϕ as suggested by the large SD for the MLE of ϕ. Starting estimates for all three methods were provided by a non-iteratively evaluated linear model (Appendix D). With a complementary log-log transformation for probability parameters and a log transformation for efficacy parameters, all three methods converge very well. Without such transformation, the Newton-Raphson procedure applied to the observed data converges if started from the IRLS estimates or the MLEs obtained via data augmentation but not from the noniteratively obtained estimates, which confirms the relative robustness of the methods using data augmentation to starting estimates.

5 Discussion

By augmenting the observed sequential symptom onsets in close contact groups with unobserved daily pairwise transmission outcomes, we identified a likelihood that has a simpler form than the one based solely on observed data and that can be maximized via the EM algorithm. Reilly and Lawlor (1999) used a similar approach to study hepatitis C infection in women with know exposure to anti-D immunoglobulin in sequential years before testing. However, the presence of multiple infective sources in the same time interval and the involvement of latent and infectious periods of influenza make our situation more complex. This simple form of the likelihood offers the flexibility of using other potential methods, for instance, the Fisher-scoring method instead of the Newton-Raphson algorithm for iterative maximization. As another example, we derived from this likelihood a linear model fitted with the IRLS method in combination with the EM-analogous algorithm. In a simulation study, the

two approaches using the augmented data performed better than the ML method using the observed data in terms of robustness to initial estimates, especially for sparse data. The IRLS method is the most robust to initial estimates, and asymptotically provides estimates of the same quality as the MLEs. The IRLS estimates are likely biased and have larger variances when data are sparse, but can serve as good initial estimates for the ML methods.

We have assumed known distributions for the latent and infectious periods and the coincidence between the latent and the incubation periods, which may not be realistic for some infectious diseases. If these assumptions do not hold, estimates could be biased and misleading. Cauchemez et al. (2004) used a Bayesian hierarchical model to allow estimation of the latent and infectious periods, assuming that the latent and the incubation periods were equal, but such estimation requires a sufficient number of cases. In addition, our models are limited to symptomatic infections. However, asymptomatic influenza infections can provide further information about the efficacies and transmission probabilities from a virological point of view, although such ”silent” cases complicate the likelihood to a large extent. A future research topic of potential public health interest would be to extend our data augmentation scheme to a Bayesian framework that can estimate the natural history of the disease and take into account asymptomatic cases.

In the data analysis, index cases were excluded regardless of their laboratory test results. According to the rationale of adjustment for selection bias, i.e., conditioning on the symptom status (caused by true infection) of the index case on the ascertainment day, a test-negative index case should be viewed as a susceptible and followed the same way as for contacts. However, not all clinical trials required symptom diary for index cases after enrollment, e.g., in the 2000-2001 trial of zanamivir. Households with test-negative index cases are generally excluded from calculations of SARs; but in our case the inclusion of the contacts in these households can improve estimation of b and θ and of p to a lesser extent. This issue could be resolved by improving the follow-up of index cases.

In this paper we have assumed fixed antiviral effects and non-random susceptibility. If sufficient data are available, random effects on the transmission probabilities as well as the antiviral efficacies could be considered to address potential heterogeneity among centers, households, or individuals (Longini and Halloran, 1996; Halloran, Préziosi and Chu, 2003).

With the potential for pandemic influenza, a rising global concern, zanamivir is one of the major available influenza antivirals agents (Hayden, 2001). Our estimates can be used in modeling research to evaluate the effects of intervention options at different levels of contact groups (Longini et al., 2004; Longini et al., 2005; Germann et al., 2006). This research also emphasizes the need for proper study design for the parameters to be adequately estimated.

Acknowledgements

This work was partially supported by National Institute of Allergy and Infectious Diseases grant R01-AI32042. The data on the clinical trials of zanamivir were provided by GlaxoSmithKline Laboratories Inc.

Appendix A: Conditional Expected Frequency of Transmission Status

Define $\bar{I}_{ji}(t)$ as the event that a susceptible person i escapes infection from infective source j on day t. Note that the following basic facts hold:

- $I_i(t) \cap I_{ji}(t) = I_{ji}(t)$.
- $\Pr\left(I_{ji}(t) \mid I_i(t) \cap S_i\left(\tilde{t}_i\right)\right) = \Pr\left(I_{ji}(t) \mid I_i(t)\right)$.

- $\bar{I}_{ji}(t) \cap I_i(\tau) = I_i(\tau)$ for $\tau > t$.
- $\Pr\left(\bar{I}_{ji}(t) \cap I_i(\tau)\right) = 0$ for $\tau < t$.

Then,

$$\begin{aligned}\Pr\left(Z_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) &= \Pr\left(I_{ji}(t) \mid S_i\left(\tilde{t}_i\right)\right) = \Pr\left(I_i(t) \cap I_{ji}(t) \mid S_i\left(\tilde{t}_i\right)\right) \\ &= \Pr\left(I_{ji}(t) \mid I_i(t) \cap S_i\left(\tilde{t}_i\right)\right) \times \Pr\left(I_i(t) \mid S_i\left(\tilde{t}_i\right)\right) \\ &= \Pr\left(I_{ji}(t) \mid I_i(t)\right) \times \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right) \times \Pr(I_i(t))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &= \frac{\Pr(I_{ji}(t) \cap I_i(t))}{\Pr(I_i(t))} \times \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right) \times \Pr(I_i(t))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &= \frac{\Pr(I_{ji}(t))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \times \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right)\end{aligned} \tag{8}$$

and

$$\begin{aligned}\Pr\left(\bar{Z}_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) = \Pr\left(\bar{I}_{ji}(t) \mid S_i\left(\tilde{t}_i\right)\right) &= \sum_{\tau=t}^{\bar{t}_i} \Pr\left(\bar{I}_{ji}(t) \cap I_i(\tau) \mid S_i\left(\tilde{t}_i\right)\right) \\ &= \sum_{\tau=t}^{\bar{t}_i} \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid \bar{I}_{ji}(t) \cap I_i(\tau)\right) \times \Pr\left(\bar{I}_{ji}(t) \cap I_i(\tau)\right)}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &= \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid \bar{I}_{ji}(t) \cap I_i(t)\right) \times \Pr\left(\bar{I}_{ji}(t) \cap I_i(t)\right)}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &\quad + \sum_{\tau=t+1}^{\bar{t}_i} \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right) \times \Pr(I_i(\tau))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &= \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right) \times \{\Pr(I_i(t)) - \Pr(I_{ji}(t))\}}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)} \\ &\quad + \sum_{\tau=t+1}^{\bar{t}_i} \frac{\Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right) \times \Pr(I_i(\tau))}{\Pr\left(S_i\left(\tilde{t}_i\right)\right)}.\end{aligned} \tag{9}$$

Appendix B: Generalization of the Linear Model to Heterogeneous Populations

For a heterogeneous population composed of k risk categories of people (e.g., age groups), let p_{vu} be the pairwise transmission probability per unprotected contact between a susceptible individual in category u and an infective person in category v. Further, let b_u be the probability of infection from the common source for category u. Assume that the AVE_S and the AVE_I are the same for all categories for notational simplicity. The models can be easily generalized to situations with heterogeneous efficacies as well. There are k parameters for common source transmission probabilities and k^2 parameters for household transmission probabilities.

Let group k be the reference stratum. The model in matrix form derived from (7) would be

$$\log\left(p_{ji}(t)\right) = \beta^{(b)\tau}\mathbf{I}_i + \mathbf{J}_i{}^{\tau}\beta^{(p)}\mathbf{I}_i + \beta^{(\theta)}r_i(t) + \beta^{(\varphi)}r_j(t) + \log\left(f_j\left(t \mid \tilde{t}_j\right)\right), \tag{10}$$

where $\beta^{(\theta)} = \log(\theta)$, $\beta^{(\phi)} = \log(\phi)$, and

$$\begin{aligned}
\mathbf{I}_i &= (I_{\{i\in 1\}},\ldots,I_{\{i\in k-1\}},1)^\tau,\\
\mathbf{J}_i &= \left(\psi_j I_{\{j\in 1\}},\ldots,\psi_j I_{\{j\in k-1\}},1\right)^\tau,\\
\beta^{(b)} &= \left(\beta_1^{(b)},\ldots,\beta_k^{(b)}\right)^\tau = \left(\log\left(\frac{b_1}{b_k}\right),\ldots,\log\left(\frac{b_{k-1}}{b_k}\right),\log(b_k)\right)^\tau,\\
\beta^{(p)} &= \left\{\beta_{vu}^{(p)}\right\}_{k\times k} = \begin{Bmatrix} \log\left(\frac{p_{11}p_{kk}}{p_{1k}p_{k1}}\right) & \cdots & \log\left(\frac{p_{1(k-1)}p_{kk}}{p_{1k}p_{k(k-1)}}\right) & \log\left(\frac{p_{1k}}{p_{kk}}\right)\\ \vdots & \ddots & \vdots & \vdots\\ \log\left(\frac{p_{(k-1)1}p_{kk}}{p_{(k-1)k}p_{k1}}\right) & \cdots & \log\left(\frac{p_{(k-1)(k-1)}p_{kk}}{p_{(k-1)k}p_{k(k-1)}}\right) & \log\left(\frac{p_{(k-1)k}}{p_{kk}}\right)\\ \log\left(\frac{p_{k1}b_k}{p_{kk}b_1}\right) & \cdots & \log\left(\frac{p_{k(k-1)}b_k}{p_{kk}b_{k-1}}\right) & \log\left(\frac{p_{kk}}{b_k}\right) \end{Bmatrix}.
\end{aligned}$$

Appendix C: Adjustment for Selection Bias in Case-ascertained Follow-up Design

In a prospective follow-up design, exposure to risks of infection starts on day 1. However, in real clinical trials, households are generally enrolled when one or more index cases are identified by symptom onsets, to which we refer as a case-ascertained design. To reduce bias caused by such selective enrollment, Yang et al. (2006) suggest that the individual likelihood contributions be conditioned on observed symptom status up to the symptom onset day of the index case. The consequences of such adjustment are the following:

- Index cases do not contribute to the likelihood.
- The likelihood calculation for person i starts from the day $\underline{t}_{d_i}+1$, where d_i denotes the index case in the household of person i.
- The individual log-likelihood is subtracted by $\log(A_i)$ where

$$\mathcal{A}_i = \sum_{t=\underline{t}_{d_i}+1}^{\bar{t}_{d_i}} \left\{ \left(\prod_{\tau=\underline{t}_{d_i}+1}^{t-1} e_i(\tau) \right) (1-e_i(t))\ \Pr\left(\tilde{t} > \tilde{t}_{d_i} \mid t\right) \right\} + \prod_{t=\underline{t}_{d_i}+1}^{\bar{t}_{d_i}} e_i(t). \tag{11}$$

For the ML method using the augmented data, the same adjustment can be applied. For the linear model method, such a conditional adjustment is difficult. However, since minimizing the weighted least squares is analogous to maximizing the log-likelihood, it is natural to use the same adjusting term to penalize the objective function

$$\sum_{h=1}^{H} \omega_h \left\{ \log\left(\tilde{\mathcal{P}}_h\right) + \log(\mathcal{P}_h) \right\}^2 + \sum_i \log(\mathcal{A}_i(\beta)),$$

where A_i is re-expressed as functions of $\beta = (\beta_0, \ldots, \beta_3)$. Denote the covariate matrix by X, the diagonal weight matrix by W and the observed response vector by $\log\left(\tilde{P}\right)$, then at the l^{th} iteration,

$$\widehat{\beta}_l = \left(X' W_{l-1} X\right)^{-1} \left\{ X' W_{l-1} \log\left(\tilde{\mathcal{P}}\right) - \frac{1}{2} \sum_i \frac{d \log\left(\mathcal{A}_i\left(\widehat{\beta}_{l-1}\right)\right)}{d \widehat{\beta}_{l-1}} \right\}.$$

Appendix D: Non-iteratively Fitted Linear Model for Initial Estimates

The ML and IRLS methods require initial estimates to start the iteration. Iteration could be avoided if we model $I_i(t)$ instead of $I_{ji}(t)$, i.e., infection status of person i on day t instead of pairwise transmission, and assume equal $\Pr(I_i(t))$ for all $\underline{t}_i \le t \le \bar{t}_i$.

Let $N_i(t)$ be the number of treated infective individuals and $M_i(t)$ be the number of untreated infective individuals that a susceptible person i is exposed to within the household on day t. Given $N_i(t)$ and $M_i(t)$, the probability that person i is infected on day t is given by

$$p_i(t) = 1 - (1-b)^{1-r_i(t)}(1-\theta b)^{r_i(t)} \times (1-p)^{M_i(t)(1-r_i(t))}(1-\theta p)^{M_i(t)r_i(t)}(1-\varphi p)^{N_i(t)(1-r_i(t))}(1-\theta\varphi p)^{N_i(t)r_i(t)}.$$

A reparameterization leads to

$$\log(1-p_i(t)) = \beta_0 + \beta_1 r_i(t) + \beta_2 M_i(t) + \beta_3 r_i(t) M_i(t) + \beta_4 N_i(t) + \beta_5 r_i(t) N_i(t). \quad (12)$$

where

$$\beta_0 = \log(1-b), \quad \beta_1 = \log\left(\tfrac{1-\theta b}{1-b}\right), \quad \beta_2 = \log(1-p), \quad \beta_3 = \log\left(\tfrac{1-\theta p}{1-p}\right),$$
$$\beta_4 = \log(1-\varphi p), \quad \text{and} \quad \beta_5 = \log\left(\tfrac{1-\theta\varphi p}{1-\varphi p}\right).$$

Let $Y_i(t)$ indicate the infection status (1:infection, 0:escape) for person i on day t. Similar to Section 2.1, define

$$Z_i(t) = Y_i(t)\prod_{\tau<t}(1-Y_i(\tau))$$

and

$$\bar{Z}_i(t) = \prod_{\tau\le t}(1-Y_i(\tau)).$$

$Z_i(t) = 1$ is the event that person i escapes infection from any source until day t, while $\bar{Z}_i(t) = 1$ is the event that person i escapes infection from any source up to day t. Assume that $\Pr(I_i(t))$ is equal for all $t \in \left[\tilde{t}_i, \bar{t}_i\right]$. Then the conditional probabilities

$$\Pr\left(Z_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) = \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(t)\right),$$
$$\Pr\left(\bar{Z}_{ji}(t) = 1 \mid b, p, \theta, \varphi, \tilde{t}_i\right) = \sum_{\tau=t+1}^{\bar{t}_i} \Pr\left(S_i\left(\tilde{t}_i\right) \mid I_i(\tau)\right),$$

do not involve unknown parameters, and can be used as the weights for fitting (12). While $N_i(t)$ and $M_i(t)$ are generally unknown, they can be obtained by randomly sampling the duration of infectious period for each infective individual according to the known empirical distribution f. Alternatively, all possible combinations of $N_i(t)$ and $M_i(t)$ can contribute to model (12) with the weights multiplied by the joint probability $\Pr(N_i(t), M_i(t))$ derived from f.

Model (12) gives rise to multiple estimators for the efficacy parameters because of the increase in parameter dimension:

$$\widehat{\theta}_1 = \frac{1-\exp\left(\widehat{\beta}_0+\widehat{\beta}_1\right)}{1-\exp\left(\widehat{\beta}_0\right)}, \quad \widehat{\theta}_2 = \frac{1-\exp\left(\widehat{\beta}_2+\widehat{\beta}_3\right)}{1-\exp\left(\widehat{\beta}_2\right)} \quad \text{and} \quad \widehat{\theta}_3 = \frac{1-\exp\left(\widehat{\beta}_4+\widehat{\beta}_5\right)}{1-\exp\left(\widehat{\beta}_4\right)}$$

for θ and

$$\widehat{\varphi}_1 = \frac{1-\exp\left(\widehat{\beta}_4\right)}{1-\exp\left(\widehat{\beta}_2\right)} \quad \text{and} \quad \widehat{\varphi}_2 = \frac{1-\exp\left(\widehat{\beta}_4+\widehat{\beta}_5\right)}{1-\exp\left(\widehat{\beta}_2+\widehat{\beta}_3\right)}$$

for ϕ. The average of the multiple estimates weighted by reciprocal standard errors can serve as the initial estimate, e.g., $\widehat{\theta} = \Sigma_{i=1}^{3}\omega_i\widehat{\theta}_i$, where $\omega_i = \frac{\frac{1}{s.e.(\widehat{\theta}_i)}}{\Sigma_{j=1}^{3}\frac{1}{s.e.(\widehat{\theta}_j)}}$.

References

Addy CL, Longini IM, Haber MJ. A generalized stochastic model for the analysis of infectious disease final size data. Biometrics 1991;47:961–974. [PubMed: 1742449]

Becker, NG. Analysis of Infectious Disease Data. Chapman and Hall; New York, NY: 1989.

Becker NG, Hasofer AM. Estimation in Epidemics with Incomplete Observations. Journal of the Royal Statistical Society, Series B 1997;59:415–429.

Becker NG, Britton T, O'Neill PD. Estimating vaccine effects on transmission of infection from household outbreak data. Biometrics 2003;59:467–475. [PubMed: 14601747]

Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boëlle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statist. Med 2004;23:3469–3487.

Datta S, Halloran ME, Longini IM. Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. Biometrics 1999;55:792–798. [PubMed: 11315008]

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.

Donner A. Some aspects of the design and analysis of cluster randomized trials. Statistics in Medicine 1998;47:95–113.

Elveback LR, Fox JP, Ackerman E. An influenza simulation model for immunization studies. American Journal of Epidemiology 1976;103:152–165. [PubMed: 814808]

Germann TC, Kadau K, Longini IM, Macken CA. Mitigation strategies for pandemic influenza in the United States. Proceedings of the National Academy of Science of the U. S. A 2006;103:5935–5940.

Halloran ME, Struchiner CJ, Longini IM. Study designs for different efficacy and effectiveness aspects of vaccination. American Journal of Epidemiology 1997;146:789–803. [PubMed: 9384199]

Halloran ME, Préziosi M-P, Chu H. Estimating vaccine efficacy from secondary attack rates. Journal of the American Statistical Association 2003;98:38–46.

Halloran ME, Hayden FG, Yang Y, Longini IM, Monto AS. Antiviral effects on influenza viral transmission and pathogenicity: observations from household-based trials. American Journal of Epidemiology 2006;165:212–222. [PubMed: 17088311]

Hayden FG, Gubareva LV, Monto AS, Klein TC, Elliott MJ, Hammond JM, Sharp SJ, Ossi MJ, Zanamivir Family Study Group. Inhaled zanamivir for the prevention of influenza in families. New England Journal of Medicine 2000;343:1282–1289. [PubMed: 11058672]

Hayden FG. Perspectives on antiviral use during pandemic influenza. Philosophical transactions of the Royal Society of London, Series B, Biological sciences 2001;356:1877–1884.

Longini IM, Koopman JS. Household and Community Transmission Parameters from Final Distributions of Infections in Households. Biometrics 1982;38:115–126. [PubMed: 7082755]

Longini IM, Halloran ME. A frailty mixture model for estimating vaccine efficacy. Journal of the Royal Statistical Society, Series C 1996;45:165–173.

Longini IM, Halloran ME, Nizam A, Yang Y. Containing pandemic influenza with antiviral agents. American Journal of Epidemiology 2004;159:623–633. [PubMed: 15033640]

Longini IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DAT, Halloran ME. Containing pandemic influenza at the source. Science 2005;309:1083–1087. [PubMed: 16079251]

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

Magder L, Brookmeyer R. Analysis of infectious disease data from partner studies with unknown source of infection. Biometrics 1993;49:1110–1116. [PubMed: 8117904]

Monto AS, Pichichero ME, Blanckenberg SJ, Ruuskanen O, Cooper C, Fleming DM, Kerr C. Zanamivir prophylaxis: an effective strategy for the prevention of influenza types A and B within households. Journal Infectious Diseases 2002;186:1582–1588.

O'Neill P, Roberts GO. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society, Series A 1999;162:121–129.

O'Neill P, Balding DJ, Becker NG, Eerola M, Mollison D. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series C 2000;49:517–542.

Paap R. What are the advantages of MCMC based inference in latent variable models? Statistica Neerlandica 2002;56:2–22.

Rampey AH, Longini IM, Haber MJ, Monto AS. A discrete-time model for the statistical analysis of infectious disease incidence data. Biometrics 1992;48:117–128. [PubMed: 1316178]

Reilly M, Lawlor E. A likelihood-based method for identifying contaminated lots of blood product. International Journal of Epidemiology 1999;28:787–792. [PubMed: 10480712]

van Dyk DA, Meng X. The art of data augmentation. Journal of Computational and Graphical Statistics 2001;10:1–50.

Welliver R, Monto AS, Carewicz O, Schattemanet E, Hassman M, Hedrick J, Jackson HC, Huson L, Ward P, Oxford JS. Effectiveness of oseltamivir in preventing influenza in household contacts: a randomized controlled trial. Journal of the American Medical Associtation 2001;285:748–754.

Yang Y, Longini IM, Halloran ME. Design and evaluation of prophylactic interventions using infectious disease incidence data from close contact groups. Journal of the Royal Statistical Society, Series C 2006;55:317–330.

Table 1

Empirical cumulative distributions of the latent period and the infectious period for influenza (Elveback et al., 1976).

Latent Period		Infectious Period	
Duration (days)	**Cumulative Probability**	**Duration (days)**	**Cumulative Probability**
0	0	≤ 2	0
1	0.2	3	0.3
2	0.8	4	0.7
3	1.0	5	0.9
		6	1.0

Table 2

Comparison between MLEs and IRLS estimates based on the augmented data. Results are based on 1000 simulations.[†‡]

Parameter	Mean of Point Estimates		Monte Carlo SD		Mean of SD Estimates		Coverage of 95% CI	
	MLE	IRLS	MLE	IRLS	MLE	IRLS	MLE	IRLS
b	0.0051	0.0051	0.00028	0.00028	0.00028	0.00027	95.5	95.2
p	0.10	0.10	0.011	0.011	0.011	0.011	94.9	96.1
θ	0.40	0.40	0.067	0.067	0.067	0.069	95.4	95.7
ϕ	0.71	0.71	0.13	0.13	0.13	0.13	95.4	94.7

[†] True parameters are set to b=0.005, p=0.1, θ=0.40, ϕ=0.70.

[‡] MLEs are the same for observed and augmented data.

Table 3

Comparing sensitivity to initial estimates between the ML method using observed data and the approaches using the augmented data when data are sparse. Results are based on 1000 simulations.[†]

Initial Values (b_0, p_0)[‡]	Method[§]	Conv. Rate (/1000)	Parameters[§§]			
			b	p	θ	ϕ
(0.002, 0.01)						
	ML(Obs)	903	0.0020 (0.00016)	0.010 (0.0049)	0.42 (0.25)	0.98 (0.95)
	ML(Aug)	889	0.0020 (0.00016)	0.010 (0.0048)	0.42 (0.24)	1.01 (0.96)
	IRLS(Aug)	937	0.0020 (0.00016)	0.011 (0.0047)	0.48 (0.24)	1.07 (0.93)
(0.02, 0.1)						
	ML(Obs)	524	0.0020 (0.00016)	0.010 (0.0048)	0.41 (0.24)	1.19 (1.11)
	ML(Aug)	878	0.0020 (0.00016)	0.010 (0.0049)	0.42 (0.24)	0.99 (1.00)
	IRLS(Aug)	920	0.0020 (0.00016)	0.011 (0.0048)	0.48 (0.24)	1.07 (1.00)
(0.0002, 0.001)						
	ML(Obs)	92	0.0020 (0.00016)	0.010 (0.0054)	0.38 (0.23)	1.04 (0.79)
	ML(Aug)	864	0.0020 (0.00015)	0.010 (0.0047)	0.44 (0.26)	1.03 (1.08)
	IRLS(Aug)	928	0.0020 (0.00015)	0.011 (0.0047)	0.49 (0.24)	1.08 (0.90)

[†] True parameters are set to b=0.002, p=0.01, θ=0.40, ϕ=0.70.

[‡] Initial values for θ and ϕ are set to the true values.

[§] Obs: observed data, Aug: augmented data.

[§§] Values in the parentheses are Monte Carlo standard deviations.

Table 4

Two randomized multi-center trials of zanamivir, an influenza antiviral agent

	Hayden et al., 2000	Monto et al., 2002
Time of trial	Oct. 1998 - Apr. 1999	Jun. 2000 - Apr. 2001
Households	336	484
Population	1186	1770
Index case randomization	Yes	No
Duration of medication		
Index case	5 days	N/A
Contact	10 days	10 days
Follow up (symptom diary)	14 days	14 days
Infected/Symptomatic(index)†	164/336	281/484
Infected/Exposed(contacts)†		
Control	52/435	76/626
Zanamivir	17/415	27/660

Numbers may slightly differ from references due to different criteria of data inclusion for analysis.

† Laboratory-confirmed infections with clinical symptoms

Table 5

Estimates of efficacies and transmission probabilities by age (1-17 vs. 18+) for pooled zanamivir trials conducted in 1998-1999 and 2000-2001. Results are obtained by approaches using the augmented data.

	IRLS		MLE		
Parameter	**Point Estimate**	**SD**	**Point Estimate**	**SD**	**95% CI**
b_c †	0.0024	0.00052	0.0028	0.00063	(0.0017, 0.0042)
b_a	0.00086	0.00030	0.0010	0.00039	(0.00045, 0.0021)
p_{cc} †	0.040	0.0074	0.040	0.0077	(0.027, 0.057)
p_{ca}	0.028	0.0045	0.029	0.0048	(0.021, 0.040)
p_{ac}	0.023	0.0071	0.020	0.0071	(0.009, 0.037)
p_{aa}	0.040	0.011	0.032	0.011	(0.016, 0.058)
AVE_S	0.68	0.086	0.75	0.072	(0.56, 0.86)
AVE_I	0.24	0.38	0.23	0.44	(−1.33, 0.75)
AVE_T			0.81	0.094	(0.50, 0.93)

† Subscript *c* denotes child (1-17), *a* denotes adult (18+), and *ca* denotes child-to-adult transmission.