STATISTICS IN MEDICINE
*Statist. Med.* (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3659

# Estimating incubation period distributions with coarse data

Nicholas G. Reich[1,*,†], Justin Lessler[2], Derek A. T. Cummings[2] and Ron Brookmeyer[1]

[1]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, U.S.A.*
[2]*Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, U.S.A.*

SUMMARY

The incubation period, the time between infection and disease onset, is important in the surveillance and control of infectious diseases but is often coarsely observed. Coarse data arises because the time of infection, the time of disease onset or both are not known precisely. Accurate estimates of an incubation period distribution are useful in real-time outbreak investigations and in modeling public health interventions. We compare two methods of estimating such distributions. The first method represents the data as doubly interval-censored. The second introduces a data reduction technique that makes the computation more tractable. In a simulation study, the methods perform similarly when estimating the median, but the first method yields more reliable estimates of the distributional tails. We conduct a sensitivity analysis of the two methods to violations of model assumption and we apply these methods to historical incubation period data on influenza A and respiratory syncytial virus. The analysis of reduced data is less computationally intensive and performs well for estimating the median under a wide range of conditions. However for estimation of the tails of the distribution, the doubly interval-censored analysis is the recommended procedure. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: doubly interval-censored data; incubation period; coarse data; influenza

## 1. INTRODUCTION

Defined as the length of time between infection with a pathogen and the onset of symptoms [1], the incubation period is an important aspect of disease natural history, knowledge of which aids in the surveillance and control of infectious disease. In the recent severe acute respiratory syndrome

*Correspondence to: Nicholas G. Reich, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, U.S.A.
†E-mail: nreich@jhsph.edu

Contract/grant sponsor: U.S. Department of Homeland Security; contract/grant number: N00014-06-1-0991
Contract/grant sponsor: NIH; contract/grant number: U01-GM070708

*Received 30 December 2008*
*Accepted 8 June 2009*

(SARS) outbreak, early estimates of the SARS incubation period informed quarantine policies that helped to stop the spread of the disease without pharmaceutical interventions. The incubation period also can play a crucial role in identifying health-care-associated infections [2, 3]. Models designed to measure the impact of public health interventions in outbreak investigations rely heavily on accurate measures of the incubation period [4, 5]. With the latent period (the time separating infection and infectiousness) the incubation period helps to determine the theoretical effectiveness of outbreak interventions targeted at symptomatic individuals [6]. In most of these applications, a full characterization of the incubation period distribution is vital, as the tails of the distributions (e.g. the 5th and 95th percentiles) play important roles.

Coarse data arise when we 'observe only a subset of the complete-data sample space in which the true, unobservable data lie' [7]. Often incubation period data are coarse because one or both of the infection time and the time of symptom onset are not observed exactly but are observed to lie within an interval of time. For diseases where the incubation period can be short (e.g. less than 2 days), the common unit of observation for symptom onset, one calendar day, is a considerable interval of time relative to the incubation period and should be treated as coarse data.

This paper compares two parametric approaches for estimating the incubation period distribution using coarse data. One method uses all available information about both exposure and disease onset contained in an observation, representing the data as doubly interval-censored. The second introduces a data reduction technique that makes the computation considerably more tractable.

We conducted a simulation study to contrast the performance of the two methods and we performed a sensitivity analysis to examine the methods' robustness to violations of model assumption, in particular the distribution of infection times. Sets of coarsened observations on the incubation period of influenza A and respiratory syncytial virus (RSV), culled from a literature review [8], are analyzed using both methods. In providing point estimates and confidence intervals for the 5th and 95th percentiles, these methods quantify important aspects of the incubation period distribution and our uncertainty in estimating them, enhancing the utility of the estimates in clinical and epidemiological settings.

## 2. APPROACHES TO ANALYZING COARSE INCUBATION PERIOD DATA

### *2.1. Doubly interval-censored data*

Let $E$ and $S$ be the true times of the infecting exposure and symptom onset for a given individual. Hence, $T = S - E$ is the true incubation period. We focus here on data where $E$ and $S$ are observed to fall within a finite interval. A typical observation consists of four points in time, $X = (E_L, E_R, S_L, S_R)$, where subscripts L and R denote the left and right boundaries on the possible infection and symptom onset times (see Figure 1(A)). When the observed data consist of an interval around $E$ and another around $S$, this is called doubly interval-censored data. The first method to analyze doubly interval-censored data was a non-parametric self-consistency algorithm developed by Turnbull in 1974 [9]. Since then, methods to analyze this type of data have been used to model the time between infection with HIV and the onset of AIDS both parametrically [10] and non-parametrically [11–13]. In addition, the doubly interval-censored likelihood has been used in a Bayesian analysis of SARS incubation period data [14] and time to developing caries in children [15].



DOI: 10.1002/sim

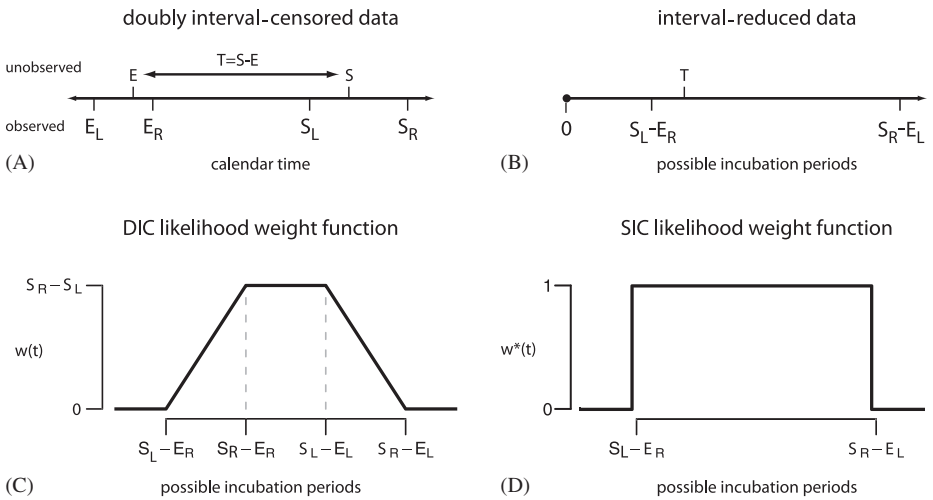


Figure 1. This figure displays typical incubation period data and the likelihood weight functions associated with each type of observation. Graph A shows a typical doubly interval-censored observation along with the unobserved exact times of the infecting exposure and symptom onset. Graph B shows an interval-reduced observation generated from a doubly interval-censored observation along with the unobserved exact incubation period. Graph C plots the doubly interval-censored likelihood weight function $w(t)$, which assigns less weight to fringe values in the interval of possible incubation periods. Graph D displays the weight function $w^*(t)$ that appears in the likelihood for single interval-censored and interval-reduced data.

Let the incubation period $T$ be a non-negative continuous random variable with probability density function (p.d.f.) $f_\theta(t)$, dependent on a parameter $\theta$. We further define $h_\lambda(e)$ to be the p.d.f. of the infecting exposure time $E$, dependent on a parameter $\lambda$, and $g(s)$ to be the p.d.f. of $S$. We assume $E$ to be independent of the incubation period $T$.

From the above definitions and assumptions, it follows that $g(s|e) = f_\theta(s-e|e) = f_\theta(s-e)$ because given an observed value of $E$ and the density of $T$ we can specify the distribution of $S$. Using $p(\cdot)$ to represent an undefined p.d.f., we can express the joint p.d.f. of $E$ and $S$ as

$$p(e,s) = p(e)\,p(s|e) = h_\lambda(e)\,g(s|e) = h_\lambda(e)\,f_\theta(s-e)$$

Therefore, the likelihood for a doubly interval-censored observation is

$$L(\theta,\lambda;X) = \int_{E_L}^{E_R}\int_{S_L}^{S_R} h_\lambda(e)\,f_\theta(s-e)\,\mathrm{d}s\,\mathrm{d}e \tag{1}$$

and distributional assumptions on both $E$ and $T$ allow us to calculate the likelihood of our observed data.

In cases where one of $E$ or $S$ is observed exactly, the data is single interval-censored with $(T_L, T_R) = (S-E_R, S-E_L)$ or $(S_L-E, S_R-E)$. The likelihood for single interval-censored data can be expressed as

$$L(\theta;T_L,T_R) = \int_{T_L}^{T_R} f_\theta(t)\,\mathrm{d}t = F_\theta(T_R) - F_\theta(T_L) \tag{2}$$



DOI: 10.1002/sim

If each of the infection and symptom onset time is observed precisely, then the data are exact. The likelihood for an exact incubation period observation, $T$, is

$$L(\theta; T) = f_\theta(T)$$

Every observation in a data set is categorized as doubly interval-censored, single interval-censored or exact. Let $\delta_i$ be an indicator of whether an observation $i$ is a doubly interval-censored observation. Similarly, let $\omega_i$ be an indicator of whether observation $i$ is represented as a single interval-censored observation. Therefore the likelihood for a given observation $X_i$, assuming $\lambda$ is known, is

$$\begin{aligned} L(\theta; X_i, \lambda) = &\left\{ \int_{E_{\mathrm{Li}}}^{E_{\mathrm{Ri}}} \int_{S_{\mathrm{Li}}}^{S_{\mathrm{Ri}}} h_\lambda(e) f_\theta(s-e) \, \mathrm{d}s \, \mathrm{d}e \right\}^{\delta_i} \\ &\times \{F_\theta(T_{\mathrm{Ri}}) - F_\theta(T_{\mathrm{Li}})\}^{\omega_i} \\ &\times \{f_\theta(T_i)\}^{(1-\delta_i)(1-\omega_i)} \end{aligned}$$

For a complete data set, the full likelihood of the unknown parameters given $n$ observations is

$$L(\theta; \mathbf{X}, \lambda) = \prod_{i=1}^{n} L(\theta; X_i, \lambda) \tag{3}$$

To find the maximum likelihood estimates of $\theta$, we can use numerical or exact techniques to calculate the maxima of this likelihood function.

One possible parametric model for incubation period data is the accelerated failure time model. A simple form of this model is

$$\ln(T) = \mu + W \tag{4}$$

where $\mu$ is an intercept and $W$ is an error variable following some distribution. Based on the work of Cowling *et al.* [16] and Sartwell [17], we assume that the natural logarithms of $T$ are distributed normally. Letting $W$ to be a Normal random variable with mean 0 and variance $\sigma^2$, we then have that $T \sim \mathrm{LogNormal}(\mu, \sigma)$ and $f_\theta(t)$ in the likelihood equations above is the p.d.f. of a log normal distribution with $\theta = (\mu, \sigma)^T$. In a log normal distribution, the parameters $\mu$ and $\sigma$ are the mean and standard deviation of the data on the log scale. More interpretable parameters for the incubation period are the median and the dispersion, defined as $m = e^\mu$ and $d = e^\sigma$ respectively. When the dispersion is 1, $\sigma$ is zero and there is no variation. Using the rules of a normal distribution, about two thirds of the data should be contained within a range of $m/d$ to $m \cdot d$.

### *2.2. A data reduction technique*

A doubly interval-censored observation such as $X = (E_{\mathrm{L}}, E_{\mathrm{R}}, S_{\mathrm{L}}, S_{\mathrm{R}})$ can be reduced to a single interval of possible incubation period values. The smallest possible incubation period, $T_{\mathrm{L}}$, is the time between the latest possible infection time and the earliest possible symptom onset, i.e. $T_{\mathrm{L}} = S_{\mathrm{L}} - E_{\mathrm{R}}$. Likewise, the largest possible incubation period, $T_{\mathrm{R}}$, is the length of time separating the earliest exposure and latest onset of disease, $T_{\mathrm{R}} = S_{\mathrm{R}} - E_{\mathrm{L}}$. We will refer to data simplified in this way as interval-reduced data (see Figure 1(B)). An interval-reduced observation can be represented in an analysis with the single interval-censored likelihood, as in equation (2), although this is not equivalent to representing the same data as doubly interval-censored (see Section 3 below). Therefore, a data set with doubly interval-censored data can be analyzed using not only the



DOI: 10.1002/sim

correct full likelihood for each observation but can also be analyzed after changing every doubly interval-censored observation to interval-reduced data.

In many studies, the process by which the data are observed and recorded is not made explicit. For example, an incubation period is reported to be between 3 and 5 days but no description is given about the intervals of possible exposure and symptom onset. In this situation, it is unclear if the data are single interval-censored or if they are interval-reduced from a doubly interval-censored observation. As we show in the Appendix, the single interval-censored likelihood for interval-reduced data is not equivalent to the doubly interval-censored likelihood for the same observation, under standard assumptions about the exposure distribution. Several studies have explored single interval-censored modeling techniques, but none have looked at the impact of changing doubly interval-censored data to interval-reduced data [16, 18–20].

## 3. COMPARISON OF PROCEDURES: LIKELIHOOD CONSIDERATIONS

A standard assumption is that the exposure is distributed uniformly on the observed interval. If we assume $h_{\lambda}(e)$ to be uniform for the likelihood in equation (1), then the likelihoods for a doubly interval-censored observation and an interval-reduced observation represented as single interval-censored data are not equal (see Appendix for details). The doubly interval-censored likelihood in equation (1) has an extra non-constant term in it when compared with the interval-reduced data likelihood in equation (2). When using the doubly interval-censored likelihood, more weight is given to central values in the interval of possible incubation periods. This is clearly seen by displaying the two likelihoods in a common format:

$$\text{doubly interval-censored likelihood (1)} \propto \int_{-\infty}^{\infty} f_{\theta}(t)w(t)\,\mathrm{d}t$$

$$\text{interval-reduced likelihood (2)} \propto \int_{-\infty}^{\infty} f_{\theta}(t)w^{*}(t)\,\mathrm{d}t$$

Figures 1(C) and (D) depict the weight functions, $w(t)$ and $w^{*}(t)$, which are derived in the Appendix. The weight function for the doubly interval-censored likelihood has a trapezoidal shape, down-weighting the fringes of the possible incubation period interval. To intuitively understand this trapezoidal shape, consider that the only way a value of $T = S_{\mathrm{L}} - E_{\mathrm{R}}$ could occur is if $E = E_{\mathrm{R}}$ and $S = S_{\mathrm{L}}$. However, if $T$ is in the middle of possible values (where $w(t)$ is flat) then there is a range of possible values for each $E$ and $S$ that would together produce such a $T$.

Since doubly interval-censored data are represented differently in the likelihood if they are represented as interval-reduced data, simulation studies and an analysis of original data were used to compare the two different representations of the data in practice.

## 4. COMPARISON OF PROCEDURES: A SIMULATION STUDY

### *4.1. Simulation methodology*

Data sets were generated to have doubly interval-censored observations on the incubation period. For each combination of simulation parameters, 1000 data sets were generated with log normal



DOI: 10.1002/sim

incubation periods. In all simulations, the dispersion parameter ($d$) of the log normal distribution was fixed at 1.6 and the distribution of exposure times, $h_{\lambda}(e)$, was assumed to be uniform. The following simulation parameters varied across simulations: median incubation period ($m$) of 2 and 4 days; sample size ($n$) of 20, 50 and 100; and the length of exposure window ($w$) of 1 and 2 days. These incubation period distribution parameters were chosen to be similar to previous estimates of several common respiratory viral infections such as influenza, RSV or SARS. The degrees of censoring were chosen based on the commonly observed patterns in the data we had collected.

To generate the data for a single data set, the following algorithm was repeated $n$ times for a given set of parameters. Fix $E_L = 0$ and $E_R = w$; draw $E$ from a Uniform$(0, w)$ and $T$ from a LogNormal$(m, d)$. We define $S = T + E$ and $S$ is assumed to be observed to within 1 day; that is, we let $S_L = \lfloor S \rfloor$ and $S_R = \lceil S \rceil$. The window lengths for exposure and symptom onset were chosen based on frequently observed data patterns in the literature review.

Log normal models as in equation (4) were fit to the data and for each data set, the 5th, 50th, 95th percentiles and the dispersion were estimated.

Each data set was fit to the accelerated failure time model twice. One analysis converted all doubly interval-censored data to the simpler interval-reduced form and maximized the single interval-censored likelihood (i.e. fixing $\omega_i = 1$ for all $i$) using the `survreg()` function in the Survival package of R [21]. The second analysis maximized the doubly interval-censored likelihood using the `optim()` function in R. The doubly interval-censored method assumed that the exposure was uniform on the observed window and was modified so that at any time the algorithm did not converge, it created interval-reduced data and tried to fit the data using the single interval-censored likelihood. If convergence was still not achieved, it fit the data using an imputation where each observation was taken to be the midpoint of the single interval. This ensured that a complete set of estimates was produced for each data set. In practice, the doubly interval-censored analysis converged 97 per cent of the time. The default convergence criteria of the `optim()` function was used. That is, the log-likelihood function in (3) was assumed to be maximized at the $(n+1)$st iteration if $\ln[L(\hat{\theta}_{n+1})] - \ln[L(\hat{\theta}_n)] < \varepsilon$, where the default value for $\varepsilon$ is $1 \times 10^{-8}$.

For the single interval-censored analysis, standard errors for $\mu$ and $\ln\sigma$ were returned by `survreg()`. In the doubly interval-censored analysis, standard errors for $\mu$ and $\sigma$ were obtained by inverting the numerical approximation to the Hessian matrix at the maximum value on the two-dimensional likelihood surface. In both cases, the Delta Method was used to convert these into standard errors for the desired estimates [22]. Using these standard errors, 95 per cent confidence intervals were constructed for the incubation period parameters. For data sets with the same parameters, the 95 per cent confidence interval coverage rates were calculated as the number of times the confidence interval contained the true parameter (e.g. the median) divided by the number of data sets.

### *4.2. Simulation results*

Figure 2 shows the empirical 95 per cent confidence interval coverage rates for medians of 2 and 4 days and dispersion of 1.6 across both the doubly interval-censored and single interval-censored estimation algorithms. The coverage of the 5th percentile in these scenarios ranged from 81 to 96 per cent; coverage for the median ranged from 93 to 96 per cent; and coverage for the 95th percentile ranged between 74 and 95 per cent. As the sample size increased, the doubly interval-censored confidence interval coverage approached 95 per cent, and was within 1 per cent point of 95 per cent when $n$ was 100.



DOI: 10.1002/sim

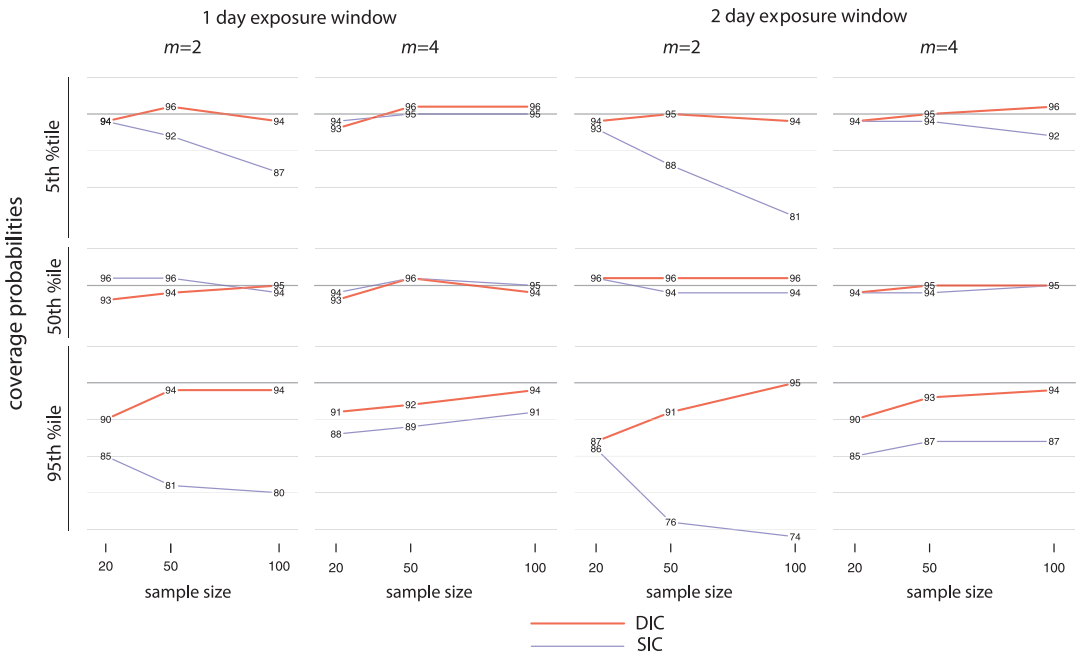


Figure 2. This figure displays the 95 per cent confidence interval coverage for the doubly interval-censored and single interval-censored simulations. Shown here are the results for the cases when the true incubation periods were generated with medians of 2 and 4 days ($m=2$ vs $m=4$) and the dispersion was 1.6. The doubly interval-censored (DIC) coverages are traced by the heavy solid lines and the single interval-censored (SIC) by the lighter dotted lines.

Estimates of average bias revealed that the doubly interval-censored estimates became less biased in estimating the median and dispersion as the sample size increased. In data sets generated with median of 4 days, dispersion of 1.6 and an exposure window size of 2 days, the doubly interval-censored method underestimated the dispersion by on average 2 per cent with $n=20$ and by less than 1 per cent with $n=100$. The single interval-censored estimates remain slightly biased away from the true values, even with the larger sample sizes. In the same data sets, the single interval-censored analysis underestimated the dispersion by on average 4 per cent with $n=20$ and 3 per cent with $n=100$. Similar patterns held in estimation of the median. Across all sample sizes, the doubly interval-censored method overestimated the median by on average between 1 and 2 per cent while the single interval-censored method overestimated the median by on average between 5 and 6 per cent. (Detailed results on bias are not shown.)

Median lengths of confidence intervals over the sets of 1000 data sets were recorded and are shown in Table I. Confidence intervals for the 5th and 50th percentiles from the single interval-censored analysis tend to be slightly larger or of the same length as those from the doubly interval-censored analysis and smaller for the 95th percentile. As the sample size increases, the median confidence interval lengths shrink. The median confidence interval lengths increase as the percentile being estimated increases. The confidence intervals are wider when the data are more coarse.



DOI: 10.1002/sim

Table I. Median length of CIs across simulations (in units of days). The rows are indexed by the median incubation period ($m$), which is measured in days.

| | | One day coarseness | | | | | | Two day coarseness | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $n=20$ | | $n=50$ | | $n=100$ | | $n=20$ | | $n=50$ | | $n=100$ | |
| Percentile | $m$ | SIC* | DIC† | SIC | DIC | SIC | DIC | SIC | DIC | SIC | DIC | SIC | DIC |
| 5th | 2 | 0.99 | 0.83 | 0.57 | 0.50 | 0.40 | 0.35 | 1.28 | 1.04 | 0.74 | 0.62 | 0.51 | 0.43 |
| | 4 | 1.47 | 1.36 | 0.89 | 0.83 | 0.62 | 0.58 | 1.69 | 1.51 | 1.02 | 0.92 | 0.71 | 0.64 |
| 50th | 2 | 1.01 | 0.96 | 0.60 | 0.59 | 0.43 | 0.41 | 1.22 | 1.08 | 0.72 | 0.67 | 0.51 | 0.47 |
| | 4 | 1.77 | 1.76 | 1.10 | 1.09 | 0.77 | 0.77 | 1.83 | 1.81 | 1.16 | 1.14 | 0.82 | 0.80 |
| 95th | 2 | 2.84 | 3.01 | 1.68 | 1.87 | 1.20 | 1.32 | 3.10 | 3.15 | 1.78 | 2.06 | 1.26 | 1.44 |
| | 4 | 5.52 | 5.69 | 3.47 | 3.58 | 2.46 | 2.54 | 5.39 | 5.69 | 3.49 | 3.65 | 2.49 | 2.62 |

*SIC columns display results from the single interval-censored analyses.
†DIC columns display results from the doubly interval-censored analyses.

In the single interval-censored estimation of the tails, the combination of a non-negligible amount of bias and shrinking average confidence interval length leads to some of the poorer coverage rates at larger sample sizes.

### *4.3. Sensitivity analysis*

We conducted a sensitivity analysis of the two methods by violating key assumptions in the generation of new data sets. To test the assumption that $h_i(e)$ is uniform, data sets were generated where the exposure was not uniform on $(E_L, E_R)$. In addition, to test the assumption of a specific parametric model, data sets were generated with uniform exposure times but Weibull incubation periods.

Two non-uniform exposure distributions were used. In the 'diurnal exposure' scenario, infection times were generated according to a piecewise uniform diurnal distribution in which 90 per cent of the probability of infection was assigned to between the 'waking' hours of a day, between 6 a.m. and 10 p.m. (see Figure 3(A)). The diurnal distribution was chosen because it may characterize the patterns of exposure in everyday life. In the 'spike exposure' scenario, data with uniform infection times were mixed with observations that had a piecewise uniform exposure distribution. The piecewise uniform exposure distribution is characterized by a spike in the probability of infection early in the exposure window: 90 per cent of the probability of infection for a given exposure window was assigned to the first 6 hours of the interval (see Figure 3(B)). These observations represent data points (like ones often seen in the literature review) where there was a clear period of high exposure early in the exposure window followed by less certain exposures for the rest of the window. It could also roughly represent a period of sustained exposure to one infected individual whose infectiousness wears off over time. Half of each mixed data set was generated with the spiked infection times; the other half had uniform infection times.

Results from the exposure distribution sensitivity analyses are presented in Figure 4. The diurnal exposure distribution did not have a large impact on the efficiency in estimating the key percentiles when using the doubly interval-censored likelihood. The doubly interval-censored coverage probabilities ranged from 87 to 96 per cent. However, when the data were converted to the interval-reduced format and analyzed using the single interval-censored likelihood, the coverage was not



DOI: 10.1002/sim

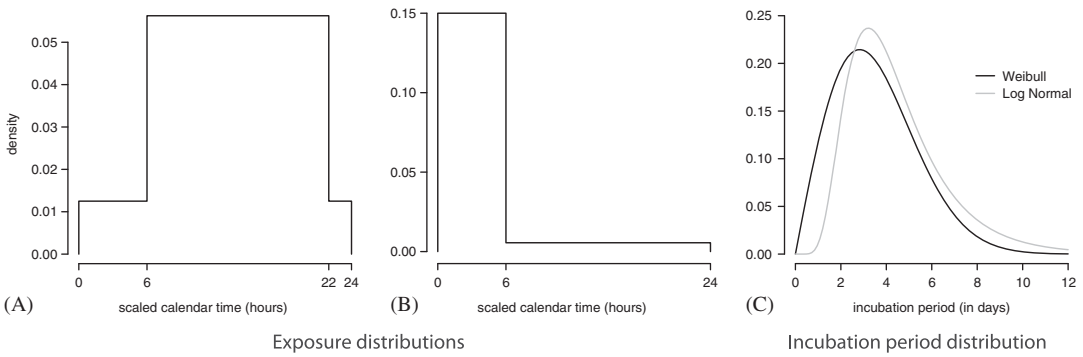


Figure 3. Distributions used in the sensitivity analyses. Graph A shows the diurnal exposure distribution that assigns 90 per cent probability of infection between the hours of 6 a.m. and 10 p.m. on the observed day of infection. Graph B shows the spike exposure distribution that assigns 90 per cent probability of infection between the hours of midnight and 6 a.m on the observed day of infection. Graph C shows the Weibull distribution with scale = 4 and shape = 2 that was used to test the sensitivity to parametric assumptions. A log normal distribution with median 4 and dispersion 1.6 is shown in gray for comparison.

as reliable. The single interval-censored coverage probabilities ranged from 71 to 96 per cent. These coverages were significantly lower than the doubly interval-censored coverage probabilities when estimating the 5th and 95th percentiles.

Confidence interval coverage in data sets with the spiked exposure distribution was mixed. The doubly interval-censored coverage probabilities ranged from 69 to 94 per cent. The single interval-censored coverage probabilities ranged from 68 to 98 per cent. Coverage of the median, which was consistent for all other methods considered in this paper, was notably low for both methods and in both cases dropped farther from 95 per cent as the sample size increased.

A 'parametric' sensitivity test generated data with uniform exposure, but with incubation periods distributed according to a Weibull distribution with scale = 4 and shape = 2 (see Figure 3(C)). In general, the methods appear more sensitive to the parametric assumption than to the uniform exposure distribution assumption. When Weibull incubation periods were analyzed as if they were log normal, the coverage probabilities were similar for the doubly interval-censored and single interval-censored analyses (detailed results are not shown). However, coverage of the true 5th percentile was poor: as low as 17 (single interval-censored) and 34 per cent (doubly interval-censored) with sample size of 100. Coverages of the median ranged from 70 to 89 per cent and were lower on average for larger sample sizes. The coverage probability for the 95th percentile was high, due to overly wide confidence intervals.

## 5. ANALYSIS OF THE INCUBATION PERIOD OF INFLUENZA A AND RSV

Influenza A and RSV are common respiratory viral infections that cause significant morbidity and mortality worldwide. Data on the incubation period of influenza A and RSV were extracted from published research. Twenty-four cases with sufficient data on the incubation period of RSV were found in the literature. Seventeen of these observations were doubly interval-censored and the remaining seven were reported as single interval-censored. One hundred fifty-one observations



DOI: 10.1002/sim

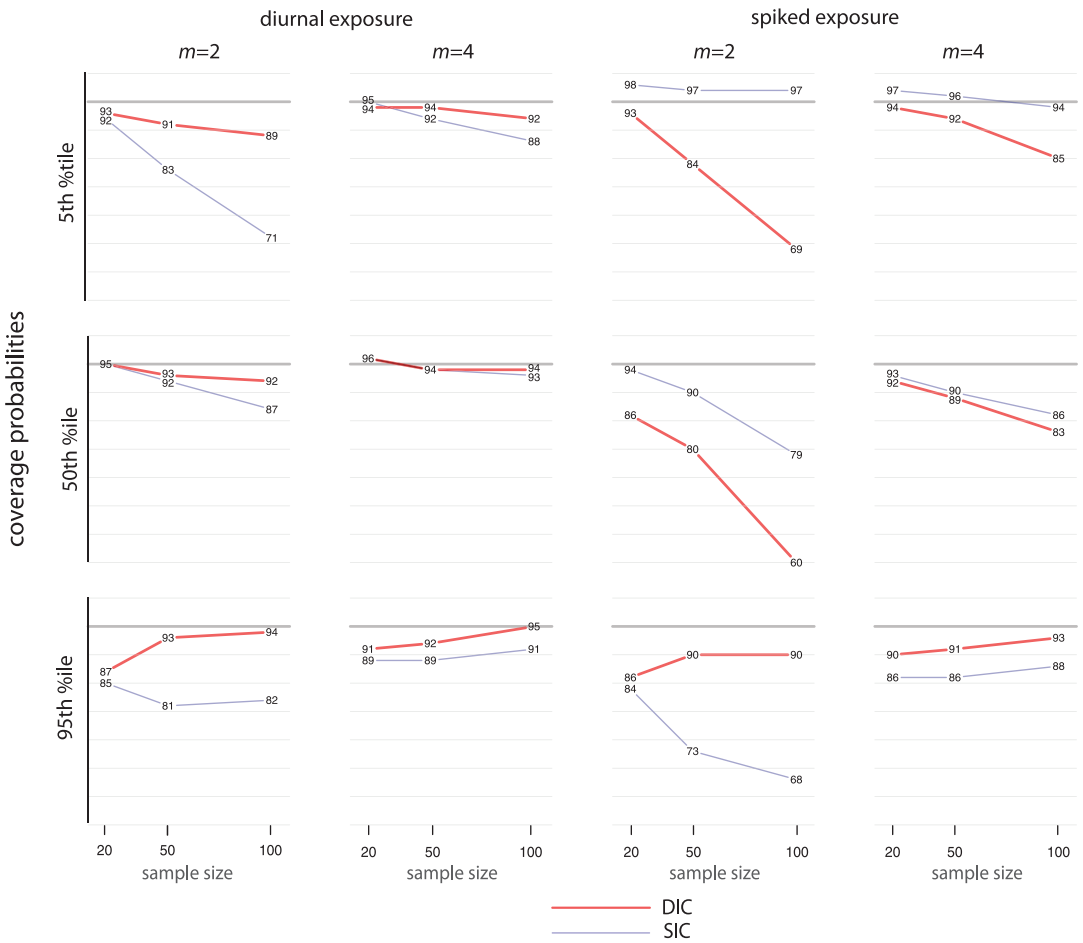


Figure 4. The 95 per cent confidence interval coverage for the exposure distribution sensitivity analysis. Shown here are the results for the cases when the true incubation periods were generated with medians of 2 and 4 days ($m=2$ vs $m=4$) and the dispersion was 1.6. The doubly interval-censored (DIC) coverages are traced by the heavy solid lines and the single interval-censored (SIC) by the lighter dotted lines. The left columns display results for the diurnal exposure distribution and the right columns display results for the spike exposure distribution.

of influenza A were found. Seventy-six of these observations were doubly interval-censored, 73 were single interval-censored and 2 were exact (reported to the nearest hour). The infection and symptom onset times were either not known or not reported exactly for many of the observations. For RSV, the exposure window size ranged from 1.0 to 7.4 days (only 6 exposure windows were greater than 1.0) and the symptom onset window size was 1 day for 18 of the cases and 0 for the remaining 6. For influenza A, the exposure window size ranged from 0 to 4 days (only 11 were greater than 1.0) and the symptom onset window size ranged from 0 to 2 days (7 equalled 2 days exactly, all the rest were 1 day or less). Details on the literature review methodology and a complete description of the data have been given elsewhere [8].



DOI: 10.1002/sim

Table II. Results from analyzing real data on influenza A and RSV.

| | Percentiles | | | |
|---|---|---|---|---|
| | 5th | 50th | 95th | Dispersion |
| Influenza A ($n=151$) | | | | |
| Doubly interval-censored | 0.73 (0.64, 0.82) | 1.46 (1.35, 1.57) | 2.94 (2.60, 3.27) | 1.53 (1.44, 1.61) |
| Single interval-censored | 0.73 (0.64, 0.81) | 1.43 (1.33, 1.54) | 2.83 (2.50, 3.16) | 1.51 (1.43, 1.60) |
| RSV ($n=24$) | | | | |
| Doubly interval-censored | 3.05 (2.43, 3.67) | 4.41 (3.90, 4.92) | 6.39 (5.32, 7.47) | 1.25 (1.14, 1.36) |
| Single interval-censored | 3.11 (2.44, 3.78) | 4.41 (3.89, 4.94) | 6.26 (5.18, 7.34) | 1.24 (1.12, 1.35) |

The doubly interval-censored rows show the estimates when using the complete doubly interval-censored likelihood. The single interval-censored rows show the estimates when doubly interval-censored observations are changed to interval-reduced data. Ninety-five per cent confidence intervals are given in parentheses after the point estimate. The units are in days.

Each of these data sets was fit to a log normal distribution twice. One analysis fit the model using the complete likelihood for doubly interval-censored observations assuming that the exposure was distributed uniformly on the observed exposure interval. A second analysis fit the model after converting all doubly interval-censored observations to the interval-reduced format. Estimates of the parameters of the distributions are reported in Table II.

To check the log normal assumption, we fit Weibull, log-logistic and exponential models to the interval-reduced version of the influenza A and RSV data. For each disease, we compared the log-likelihood of the four different fitted parametric models. In each case, the log normal models had the greatest log-likelihood of the four, giving a marginally better fit to the data than the Weibull and log-logistic models and a noticeably better fit than the exponential model. The log normal models have also been shown to match well with non-parametric estimates of the data distribution [8].

The doubly interval-censored analysis estimated the median incubation period for influenza A to be 1.46 days (95 per cent CI: 1.35, 1.57), while the single interval-censored analysis estimated the median as 1.43 days (95 per cent CI: 1.33, 1.54). The dispersion was estimated to be 1.53 (95 per cent CI: 1.41, 1.61) by the doubly interval-censored analysis and 1.51 (95 per cent CI: 1.43, 1.60) by the single interval-censored analysis. The difference in dispersions led to the doubly interval-censored analysis having slightly larger estimates of the 95th percentile (2.94 vs 2.83 days), although the point estimates of the 5th percentile were equivalent (0.73 days). Confidence interval lengths were comparable for the tail estimates.

The doubly interval-censored analysis estimated the median incubation period for RSV to be 4.41 days (95 per cent CI: 3.90, 4.92), while the single interval-censored analysis estimated the median as 4.41 days (95 per cent CI: 3.89, 4.94). The dispersion was estimated to be 1.25 (95 per cent CI: 1.14, 1.36) by the doubly interval-censored analysis and 1.24 (95 per cent CI: 1.12, 1.35) by the single interval-censored analysis. The difference in dispersions led to the doubly interval-censored analysis having slightly larger estimates of the 95th percentile (6.39 vs 6.26 days) and slightly smaller estimates of the 5th percentile (3.05 vs 3.11 days).

In these case-studies, the two methods never differ in their estimates of any percentile by more than 4 hours, a difference that is not significant in a clinical setting. These results confirm the prevailing conventional wisdom about the incubation period of influenza A and RSV and they provide a level of precision and quantification of uncertainty that is not found in the existing literature.



DOI: 10.1002/sim

## 6. DISCUSSION

The incubation period plays a vital role in the surveillance and control of infectious diseases. Methods that can efficiently estimate parameters of an incubation period distribution provide results that are useful in real-time outbreak investigations, in identifying possible sources of exposure and in modeling the efficacy of public health interventions to stop or slow the spread of infectious disease outbreaks. Such methods must be able to handle coarse data because both ends of incubation periods are often imprecisely observed. The methods should also be robust to model assumptions because in an emerging outbreak, especially with a new pathogen, the validity of the underlying assumptions may be uncertain.

We present in this paper two methods of obtaining estimates for an incubation period distribution and we applied these methods to obtain estimates of the distributions for influenza A and RSV. Reducing doubly interval-censored data to use standard single interval-censored data analysis techniques will yield fairly accurate estimates of the central tendency, even with small sample sizes. However, for accurate estimation of the tails of the distribution, we show the doubly interval-censored analysis to be a more reliable method than the interval-reduced data method, especially for large sample sizes. As is shown in Figure 2, increasing the sample size tends to decrease the confidence interval coverage of tail percentiles in an analysis of interval-reduced data.

In 2003, the World Health Organization recommended that 200 observations would be needed to estimate the incubation period distribution [23]; however, that recommendation was not based on a rigorous statistical analysis that accounted for coarse data arriving in a real-time epidemic. Our methods can be used to investigate the adequacy of the WHO sample size recommendations. A focused statistical investigation could provide evidence-based sample size guidelines for estimating both the center and the tails of the distribution under different levels of coarse data.

One limitation of this work is that our examples explore only a small range of the possible incubation period parameter space and of degrees of coarse data. A theoretical or simulation-based study focused on determining a specific relationship between the degree of coarseness present in the data and the efficiency in estimating the parameters of interest would be a valuable contribution to this area of research.

The doubly interval-censored method is the preferred method for all applications that involve the tails of the distribution, e.g. estimating the optimal length of quarantine. The interval-reduced method may be more appropriate when quick estimates of central tendency are needed, as it performs well for estimating the median incubation period under a wide range of conditions. It may also be appropriate when good doubly interval-censored data are unavailable or are too computationally burdensome. However, results from these analyses should be viewed with care in the light of the results presented here, which indicate some bias and poor confidence interval coverage when using the interval-reduced method.

Performance of these methods depend somewhat on the parametric model assumptions. As the sensitivity analysis shows, an incorrect choice of a parametric distribution could lead to inefficient estimates of the incubation period distribution. More work like that of Cowling *et al.* [16] and Nishiura [24] is needed to study the evidence on which distributions best characterize the available data on incubation periods.

The diurnal exposure does not appear to have a large impact on the observed confidence interval coverage, which is not surprising given that it is not that large of a departure from a uniform distribution. However, when the uniform exposure distribution is contaminated with a spike distribution, the performance of both methods suffers, especially with larger sample sizes.



DOI: 10.1002/sim

These results suggest that if prior knowledge indicates that the exposure may not be uniform, then that information should be incorporated into the analytic techniques. An adjustment to the exposure distribution, $h_\lambda(e)$, in the likelihood may increase the accuracy of the analysis. Further work in this area could suggest methods of specifying other exposure distributions. In addition, developing a theoretical and practical understanding of whether the exposure time distribution could be estimated separately from or concurrently with the incubation period distribution would be a valuable contribution.

Both methods rely on the assumption that the time of infection is independent of the incubation period. This assumption could be violated if immune-compromised individuals in an outbreak were more likely to be infected earlier in calendar time and were also more likely to have shorter incubation periods. Investigating the degree to which a violation of this independence assumption could impact the accuracy of estimation could be a fertile area for further investigation. Applying the methods to real-time surveillance data in emerging outbreaks with new pathogens in order to adaptively estimate the incubation period distribution is also an area of inquiry that will have important implications for controlling outbreaks of infectious disease.

## APPENDIX A: PROOF OF NON-EQUIVALENCE IN INTERVAL-CENSORED LIKELIHOODS

*Claim*

When $h_\lambda(e)$ is assumed to be uniform, the likelihoods for doubly interval-censored data and interval-reduced data representing the same observation are not equivalent.

*Proof*

Let a doubly interval-censored observation be defined as $X=(E_L, E_R, S_L, S_R)$. Recall, from equation (1) that the likelihood of this observation is

$$L(\theta, \lambda; X)=\int_{E_L}^{E_R}\int_{S_L}^{S_R} h_\lambda(e)\, f_\theta(s-e)\,\mathrm{d}s\,\mathrm{d}e$$

The interval-reduced version of this observation is $X^*=(T_L, T_R)=(S_L-E_R, S_R-E_L)$ and we recall equation (2) to express the likelihood of $X^*$ as

$$L(\theta; X)=F_\theta(T_R)-F_\theta(T_L)$$

Since $h_\lambda(e)$ is defined without reference to observed data we assume that the probability of infection is uniform on some wide interval $(A, B)$ where $A<<E_L$ and $B>>E_R$, i.e. we define the range of $h_\lambda(e)$ to be outside the range of observable data. Therefore, $h_\lambda(e)=1/(B-A)\cdot I\{A<e<B\}$.

We calculate directly the doubly interval-censored likelihood:

$$L(\theta, \lambda; X) = \int_{E_L}^{E_R}\int_{S_L}^{S_R} h_\lambda(e)\, f_\theta(s-e)\,\mathrm{d}s\,\mathrm{d}e$$

$$(h_\lambda(e)\text{is uniform and substituting } t=s-e)$$



DOI: 10.1002/sim

$$= \frac{1}{B-A} \int_{E_L}^{E_R} \int_{S_L-e}^{S_R-e} f_\theta(t)\,\mathrm{d}t\,\mathrm{d}e$$

(replace inner integral limits with an indicator)

$$\propto \int_{E_L}^{E_R} \int_{-\infty}^{\infty} f_\theta(t) I(S_L - e < t < S_R - e)\,\mathrm{d}t\,\mathrm{d}e$$

(using Fubini's theorem since $f_\theta(\cdot)$ and $I(\cdot) \leqslant 1$)

$$\propto \int_{-\infty}^{\infty} f_\theta(t) \int_{E_L}^{E_R} I(S_L - t < e < S_R - t)\,\mathrm{d}e\,\mathrm{d}t$$

$$\left(\text{let } w(t) = \int_{E_L}^{E_R} I(S_L - t < e < S_R - t)\,\mathrm{d}e\right)$$

$$\propto \int_{-\infty}^{\infty} f_\theta(t) w(t)\,\mathrm{d}t$$

To determine the behavior of this function, we need to look at its possible values for different values of $t$. The function $w(t)$ depends on which of the exposure interval and symptom onset interval is larger. Without loss of generality, we assume that symptom window is smaller than exposure window (i.e. $S_R - S_L \leqslant E_R - E_L$).

Based on the definition of $w(t)$, we observe that

$$w(t) = \begin{cases} 0 & \text{if } t < S_L - E_R \\ E_R - S_L + t & \text{if } S_L - E_R < t < S_R - E_R \\ S_R - S_L & \text{if } S_R - E_R < t < S_L - E_L \\ S_R - E_L - t & \text{if } S_L - E_L < t < S_R - E_L \\ 0 & \text{if } t > S_R - E_L \end{cases}$$

Figure 1(C) is a graph of $w(t)$, and it shows that the weight function follows a trapezoidal shape. We observe that $w(t)$ can be pulled outside of the integral above (yielding equivalent likelihoods) $\Longleftrightarrow$ $w(t)$ were constant over the interval $(S_L - E_R, S_R - E_L)$ and zero outside this interval, i.e. not dependent on $t$. For example:

$$L(\theta, \lambda; X) \propto \int_{-\infty}^{\infty} f_\theta(t) w(t)\,\mathrm{d}t$$

(iff $w(t)$ is constant w.r.t. $t$)

$$\propto \int_{S_L - E_R}^{S_R - E_L} f_\theta(t)\,\mathrm{d}t$$

$$\propto F_\theta(S_R - E_L) - F_\theta(S_L - E_R)$$

$$\propto L(\theta; X^*)$$



DOI: 10.1002/sim

Therefore: $h_\lambda(e)$ is uniform $\Longrightarrow$ $w(t)$ is not constant $\Longrightarrow$ the likelihoods are not equivalent.
□

ACKNOWLEDGEMENTS

Nicholas G. Reich, Justin Lessler and Ron Brookmeyer were funded by a grant from the U.S. Department of Homeland Security (N00014-06-1-0991). Derek A. T. Cummings work on this project was funded by a grant from the NIH (U01-GM070708) and he holds a Career Award at the Scientific Interface from the Burroughs Welcome Fund.

REFERENCES

1. Brookmeyer R. Incubation period of infectious diseases. *Encyclopedia of Biostatistics*, vol. 3. Wiley: New York, 1998; 2011–2016.
2. Lessler J, Brookmeyer R, Perl TM. An evaluation of classification rules based on date of symptom onset to identify health-care associated infections. *American Journal of Epidemiology* 2007; **166**(10):1220–1229. DOI: 10.1093/aje/kwm188.
3. Karanfil L, Conlon M, Lykens K, Masters C, Forman M, Griffith M, Townsend T, Perl T. Reducing the rate of nosocomially transmitted respiratory syncytial virus. *AJIC*: *American Journal of Infection Control* 1999; **27**(2):91. DOI: 10.1016/S0196-6553(99)70087-8.
4. 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**(7056):209–214. DOI: 10.1038/nature04017.
5. Longini IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DAT, Halloran ME. Containing pandemic influenza at the source. *Science* 2005; **309**(5737):1083–1087. DOI: 10.1126/science.1115717.
6. Fraser C, Riley S, Anderson R, Ferguson NM. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the U.S.A.* 2004; **101**:6146–6151. DOI: 10.1073/pnas.0307506101.
7. Heitjan D, Rubin D. Ignorability and coarse data. *The Annals of Statistics* 1991; **19**(4):2244–2253. DOI: 10.1214/aos/1176348396.
8. Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE, Cummings DAT. A systematic review of the incubation periods of acute respiratory viral infections. *The Lancet Infectious Diseases* 2009; **9**(5):291–300. DOI: 10.1016/S1473-3099(09)70069-6.
9. Turnbull B. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* 1974; **69**(345):169–173. DOI: 10.2307/2285518.
10. Brookmeyer R, Goedert J. Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* 1989; **45**(1):325–335. DOI: 10.2307/2532057.
11. De Gruttola V, Lagakos S. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* 1989; **45**(1):1–11. DOI: 10.2307/2532030.
12. Kim M, De Gruttola V, Lagakos S. Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* 1993; **49**:13. DOI: 10.2307/2532598.
13. Sun J. Empirical estimation of a distribution function with truncated and doubly-interval censored data and its application to AIDS studies. *Biometrics* 1995; **51**(3):1096–1104. DOI: 10.2307/2533008.
14. McBryde ES, Gibson G, Pettitt AN, Zhang Y, Zhao B, McElwain DLS. Bayesian modelling of an epidemic of severe acute respiratory syndrome. *Bulletin of Mathematical Biology* 2006; **68**(4):889–917. DOI: 10.1007/s11538-005-9005-4.
15. Komárek A, Lesaffre E. Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association* 2008; **103**(482):523–533. DOI: 10.1198/016214507000000563.
16. Cowling BJ, Muller MP, Wong IOL, Ho L, Louie M, McGeer A, Leung G. Alternative methods of estimating an incubation distribution: examples from severe acute respiratory syndrome. *Epidemiology* 2007; **18**(2):253–259. DOI: 10.1097/01.ede.0000254660.07942.fb.
17. Sartwell PE. The distribution of incubation periods of infectious disease. *American Journal of Hygiene* 1950; **51**:310–318.

18. Lindsey J. A study of interval censoring in parametric regression models. *Lifetime Data Analysis* 1998; **4**(4): 329–354. DOI: 10.1023/A:1009681919084.
19. Lindsey J, Ryan L. Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine* 1998; **17**(2):219–238. DOI: 10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O.
20. Odell P, Anderson K, D'Agostino R. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992; **48**(3):951–959. DOI: 10.2307/2532360.
21. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5**(3):299–314. DOI: 10.2307/1390807.
22. Casella G, Berger R. *Statical Inference* (2nd edn). Duxbury: Pacific Grove, CA, 2002.
23. World Health Organization. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS), May 2003. Available from: http://www.who.int/csr/sars/WHOconsensus.pdf [accessed 17 March 2008].
24. Nishiura H. Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerging Themes in Epidemiology* 2007; **4**:2. DOI: 10.1186/1742-7622-4-2.