STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2525–2537
Published online 12 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2123

The estimation of SARS incubation distribution from serial interval data using a convolution likelihood

Anthony Y. C. Kuk[1,*,†] and Stefan Ma[2]

[1]*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore*
[2]*Ministry of Health, Singapore*

SUMMARY

The incubation period of SARS is the time between infection of disease and onset of symptoms. Knowledge about the distribution of incubation times is crucial in determining the length of quarantine period and is an important parameter in modelling the spread and control of SARS. As the exact time of infection is unknown for most patients, the incubation time cannot be determined. What is observable is the serial interval which is the time from the onset of symptoms in an index case to the onset of symptoms in a subsequent case infected by the index case. By constructing a convolution likelihood based on the serial interval data, we are able to estimate the incubation distribution which is assumed to be Weibull, and justifications are given to support this choice over other distributions. The method is applied to data provided by the Ministry of Health of Singapore and the results justify the choice of a ten-day quarantine period. The indirect estimate obtained using the method of convolution likelihood is validated by means of comparison with a direct estimate obtained directly from a subset of patients for whom the incubation time can be ascertained. Despite its name, the proposed indirect estimate is actually more precise than the direct estimate because serial interval data are recorded for almost all patients, whereas exact incubation times can be determined for only a small subset. It is possible to obtain an even more efficient estimate by using the combined data but the improvement is not substantial. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: convolution; incubation distribution; quarantine period; serial interval; severe acute respiratory syndrome; Weibull distribution

1. INTRODUCTION

Severe acute respiratory syndrome (SARS) is an illness caused by a coronavirus, called SARS-associated coronavirus (SARS-CoV), see References [1–3]. As a disease, SARS has a high case-fatality rate. According to the WHO, a total of 8422 people worldwide became sick with SARS during the 2003 outbreak. Of these, 916 died. Up to now, there is no definitive cure or

*Correspondence to: Anthony Y. C. Kuk, Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546, Singapore.
†E-mail: stakuka@nus.edu.sg

Received June 2004
Accepted October 2004

vaccine for SARS. As a result, health officials have to rely on isolation and quarantine as the two main preventive measures. The people to be isolated include the suspected, probable and confirmed SARS cases so as to prevent further transmission of the disease. As people infected with SARS have to go through an incubation period before they become symptomatic, it is also important to quarantine people who may have been exposed to SARS-CoV, for example, the contacts of cases, so that they can be isolated as soon as they show possible signs of the disease. Obviously, the suitable length of the quarantine period is directly related to the length of the incubation period. The distribution of SARS incubation period is also an important input parameter in a mathematical model to describe the dynamics and spread of the disease and for assessing the potential impact of interventions.

Incubation period is the time between infection of a disease and onset of symptoms. Estimation of the SARS incubation distribution has been hampered by the fact that the dates of infection cannot be determined exactly for the majority of patients. As a result, Donnelly *et al.* [4] have to restrict their attention to a small subset of 57 SARS patients in Hong Kong with short and defined periods of exposure to known SARS cases. In their own words, the resulting estimate is 'based on a limited number of observations to date, and has high variance and may reflect biases in reporting, different routes of transmission, or varying infectious doses of the virus'.

Meltzer [5] attempts to deal with the problem that many patients have multiple possible incubation periods by using a simulation approach whereby the incubation period of each patient is simulated uniformly from his/her set of possible values. It is claimed that the simulation approach can be used to give frequency distribution and confidence intervals for the various quantities of interest. The argument, is, however, flawed from a statistical point of view. This is because the use of simulations only accounts for the Monte Carlo variation of simulations treating the observed sample as fixed but not the sampling variability that takes into account the randomness of samples. To be precise, suppose there are n patients in the sample and the incubation period of patient i could only be determined up to m_i possible values $t_{i1},\ldots,t_{im_i}$. The estimate obtained from Meltzer's simulations, say, $\tilde{F}$ is just an approximation of the distribution $\hat{F}$ that assigns probability mass $1/nm_i$ to $t_{ij}, i=1,\ldots,n$, $j=1,\ldots,m_i$. What has been accounted for is the variability of $\tilde{F}$ as an approximation of $\hat{F}$ but not the variability of $\hat{F}$. In fact, given that $\hat{F}$ can be computed, there is no need to do any simulations at all.

While the exact date of infection and hence the incubation period cannot be determined for most patients, what is observable is the serial interval which is defined in Reference [6] as the time from onset of symptoms in an index case to the onset of symptoms in a subsequent case infected by the index patient. By constructing a convolution likelihood based on the serial interval data in Singapore, we are able to estimate the incubation distribution parametrically. To be specific, the Weibull distribution is selected as a suitable distribution for SARS incubation from a four-parameter family of generalized F distributions (see Sections 2.2.7 and 3.9.1 of Reference [7]) that contains the log-normal, log-logistic, gamma, Weibull and reciprocal Weibull distributions as special cases. Based on the resulting estimate, the choice of a 10-day incubation period in Singapore is well justified. The validity of the convolution likelihood estimate is demonstrated by means of comparison with a direct estimate obtained directly from a subset of patients for whom the incubation time can be ascertained. Whereas the direct estimate is only based on 50 incubation times, the convolution likelihood estimate

is based on 198 serial intervals and hence is more precise. It is also possible to construct a maximum likelihood estimate based on the ascertained incubation times for the 50 patients plus the serial times for the remaining 148 patients.

2. CONVOLUTION LIKELIHOOD BASED ON SERIAL INTERVAL DATA

A serial interval is the time from onset of symptoms in an index case to the onset of symptoms in a subsequent case infected by the index case. During the 2003 outbreak in Singapore, a total of 206 probable SARS cases were diagnosed using the initial case definition issued by WHO in March 2003, and 32 additional cases were confirmed later by laboratory test. Serial intervals can be determined for 198 patients. Index these 198 patients by $i = 1, \ldots, 198$. Let s_i be the observed serial time between onset of symptoms in patient i and onset of symptoms in his/her infector. The observed serial time s_i for patient i can be decomposed as

$$s_i = u_i + t_i$$

where u_i is the time between the date of onset of symptoms in the index case and the unobserved date of infection of patient i, and t_i the incubation time between infection and onset of symptoms for patient i. Note that we can only observe the sum s_i, but not the components u_i and t_i. Assuming that the two components are statistically independent, the density of the sum s_i is given by the convolution of the densities of u_i and t_i, and the incubation times t_i are assumed to be independent and identically distributed according to a density $f_T(t)$. Furthermore, let d_i be the duration between onset of symptoms and isolation for the spreader who transmitted the disease to patient i. We assume that a person infected with SARS will not transmit the disease to others before onset of symptoms and after isolation. Under this assumption, the time d_i defined earlier can be interpreted as the duration of infectiousness of the person who transmitted SARS to patient i. This is a reasonable assumption as there is no compelling evidence so far of transmission from asymptomatic persons [6] and isolation measure, if strictly adhered to, should be effective in preventing transmission. To obtain the density of u_i that is to be convoluted with that of t_i, we condition on the duration of infectiousness d_i of the spreaders. Assuming constant infectiousness throughout the duration d_i, the conditional density $f_{U_i}(u_i \mid d_i)$ of the infection time u_i is uniformly distributed between 0 and d_i. We have also assumed independence between the incubation times t_i of patients and the duration times of their infectors so that the conditional distribution of t_i given d_i is still given by $f_T(t)$. Combining, the density of $s_i = u_i + t_i$ given the duration of infectiousness d_i, is given by the convolution

$$\begin{aligned} f_{S_i}(s_i \mid d_i) &= \int_{t=0}^{\infty} f_T(t) f_{U_i}(s_i - t \mid d_i)\, \mathrm{d}t \quad \begin{cases} \displaystyle\int_{t=0}^{s_i} f_T(t) \frac{1}{d_i}\, \mathrm{d}t & \text{if } s_i < d_i \\ \displaystyle\int_{t=s_i-d_i}^{s_i} f_T(t) \frac{1}{d_i}\, \mathrm{d}t & \text{if } s_i \geqslant d_i \end{cases} \\ &= \begin{cases} \dfrac{1}{d_i} F_T(s_i) & \text{if } s_i < d_i \\ \dfrac{1}{d_i}\{F_T(s_i) - F_T(s_i - d_i)\} & \text{if } s_i \geqslant d_i \end{cases} \end{aligned} \tag{1}$$

A simpler way to derive (1) proposed by a referee is to think in terms of interval censoring of the incubation times t_i. Since $s_i = u_i + t_i$, we have $t_i = s_i - u_i$, that is to say, the incubation time t_i is the difference between the serial interval s_i and the infection time u_i. Now the infection time u_i must be within the period of infectiousness, that is, between 0 and d_i, it follows that the incubation time $t_i = s_i - u_i$ must be in between $s_i - d_i$ and s_i, i.e. $t_i \in (s_i - d_i, s_i)$. But of course the incubation time cannot be negative and so we can conclude further that $t_i \in (0, s_i)$ if $s_i < d_i$. In the absence of further knowledge about the time of infection, which is equivalent to the uniform distribution assumption for u_i, the likelihood contribution from case i is either $F_T(s_i)$, or $F_T(s_i) - F_T(s_i - d_i)$, depending on whether $s_i < d_i$ or $s_i \geqslant d_i$. This argument also leads to (1), apart from the factor $1/d_i$, which does not depend on parameters and hence makes no difference in maximum likelihood estimation. We follow the convolution approach since it is arguably more rigorous than the censored data approach and the first line of (1) is general enough to include the case of non-uniform $f_U(u)$.

It follows from (1) that the density of the serial interval s_i has closed form if the cumulative distribution function $F_T(t)$ of the incubation times t_i has closed form. A popular choice for $F_T(t)$ is the Weibull distribution function

$$F_T(t) = 1 - \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\} \tag{2}$$

The popularity of Weibull distribution in survival analysis stems from the fact that it is flexible in shape, allows both decreasing and increasing hazard over time, and it is the only distribution that is compatible with both the proportional hazards and accelerated failure time regression models. Substituting (2) into (1), we get

$$f_{S_i}(s_i \mid d_i; \alpha, \beta) = \begin{cases} \dfrac{1}{d_i}\left[1 - \exp\left\{-\left(\dfrac{s_i}{\alpha}\right)^{\beta}\right\}\right] & \text{if } s_i < d_i \\ \dfrac{1}{d_i}\left[\exp\left\{-\left(\dfrac{s_i - d_i}{\alpha}\right)^{\beta}\right\} - \exp\left\{-\left(\dfrac{s_i}{\alpha}\right)^{\beta}\right\}\right] & \text{if } s_i \geqslant d_i \end{cases}$$

By maximizing the likelihood

$$\prod_i f_{S_i}(s_i \mid d_i; \alpha, \beta) \tag{3}$$

based on the serial interval data s, we obtain the estimates $\hat{\alpha}(s) = 5.44$, $\hat{\beta}(s) = 1.91$. The associated standard errors are 0.27 and 0.16 which are obtained by the usual method of inverting the information matrix. The estimated incubation distribution is shown in Figure 1. Under the Weibull assumption, the mean incubation time is $E(T) = \mu = \alpha\Gamma(1 + \beta^{-1})$ which is estimated by $\hat{\mu} = \hat{\alpha}\Gamma(1 + \hat{\beta}^{-1}) = 4.83$ days. The standard error 0.235 of $\hat{\mu}$ is obtained from the variance–covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ using the delta method. It follows that an approximate 95 per cent confidence interval for μ is (4.37, 5.29) days.

Turning to percentiles, the 95th percentile of the incubation time is given by $\xi = \alpha(-\log(0.05))^{1/\beta}$ and estimated by $\hat{\xi} = \hat{\alpha}(-\log(0.05))^{1/\beta} = 9.66$ days (SE = 0.50) as shown in Figure 1. A one-sided 95 per cent confidence interval is (0, 10.49) days. Thus we are 95 per cent confident that the 95th percentile of the incubation distribution is at most 10.49

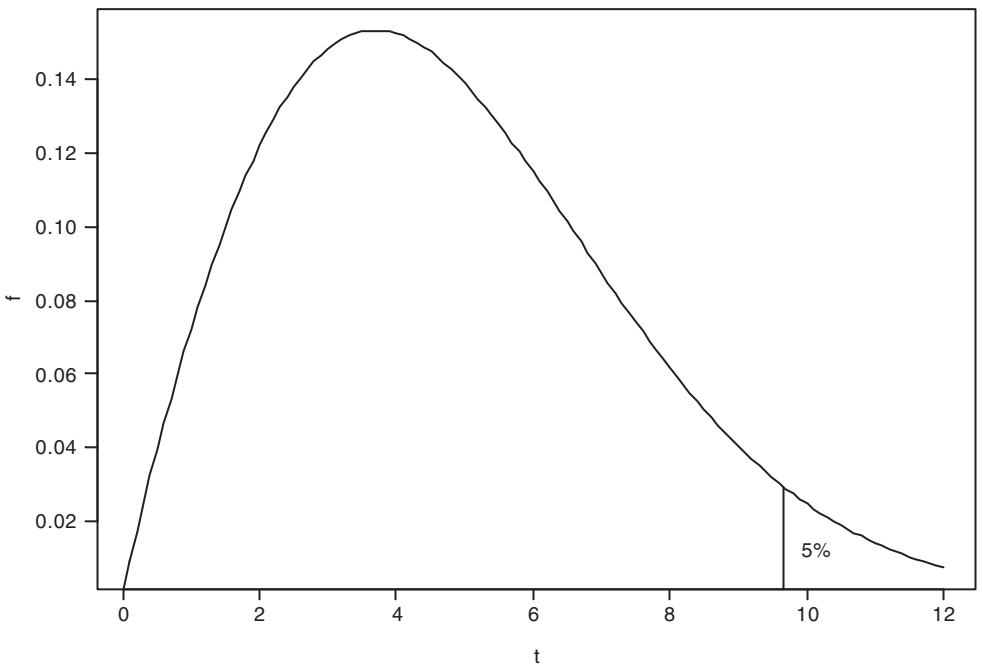

Figure 1. Maximum convolution likelihood estimate of the incubation distribution of SARS based on serial interval data and the assumption of Weibull distribution.

days. As commented earlier, a 10-day quarantine period was used in Singapore in 2003, the probability that the incubation time of a SARS patient is less than or equal to 10 days is

$$P(T \leqslant 10) = F_T(10) = 1 - \exp\left\{-\left(\frac{10}{\alpha}\right)^{\beta}\right\}$$

under the Weibull assumption and can be estimated by

$$\hat{P}(T \leqslant 10) = 1 - \exp\left\{-\left(\frac{10}{\hat{\alpha}}\right)^{\hat{\beta}}\right\} = 0.959 \ (\mathrm{SE} = 0.0133)$$

A one-sided 95 per cent confidence interval for this probability is (0.937, 1.0). In other words, we are 95 per cent confident that at least 93.7 per cent of SARS patients have incubation time of at most 10 days. This suggests that a 10-day quarantine period should be sufficient, bearing in mind that some patients have started incubation before they get quarantined.

Other common choices of $F_T(t)$ include the log-normal and log-logistic distribution. The convolution density of s_i is again given by (1) and parameter estimates can be obtained by maximizing the convolution likelihood. Figure 2 displays the estimated incubation distribution for various choices of the functional form of $F_T(t)$. As expected, there is not much difference between the log-normal and log-logistic estimates, but they are both quite different from the

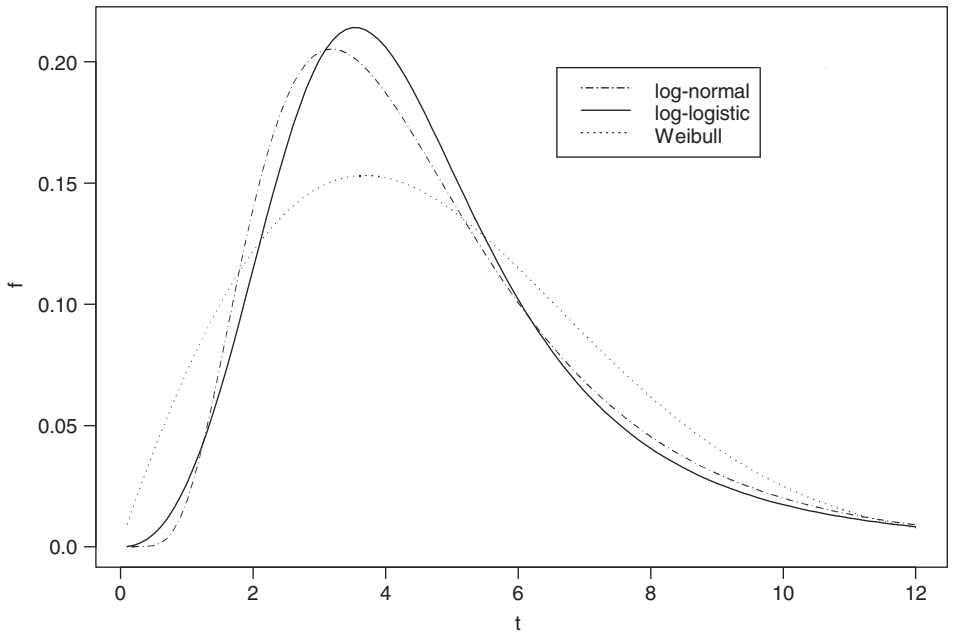

Figure 2. Estimates of the incubation distribution of SARS under different distributional assumptions.

Weibull estimate. We will consider model validation in the next section to provide support for the Weibull model.

3. MODEL VALIDATION USING ASCERTAINED INCUBATION TIMES

With strenuous effort and intensive contact tracing, the Ministry of Health of Singapore was able to more or less determine the dates of infection and hence ascertain the incubation times for a subset D of 50 patients. Since the incubation times t_i are directly observable for the patients in D, we can estimate the incubation distribution directly by maximizing the likelihood

$$\prod_{i \in D} f_T(t_i; \alpha, \beta) \tag{4}$$

based on t_i, for $i \in D$, to obtain the estimates $\hat{\alpha}(t_D) = 5.80$ and $\hat{\beta}(t_D) = 2.59$. This maximization can be implemented by, for example, the SAS procedure LIFEREG. The histogram of the 50 incubation times and the fitted Weibull distribution are shown in Figure 3. The Pearson goodness of fit statistic for fitting a Weibull distribution to the histogram is 5.03 on 7 degrees of freedom (p value $= 0.66$) and so the Weibull model cannot be rejected.

A good way to check the validity of the convolution likelihood method proposed in the last section is to compare the direct estimates with the convolution likelihood estimates based on the same subset of patients. To be precise, we are comparing the direct esti-

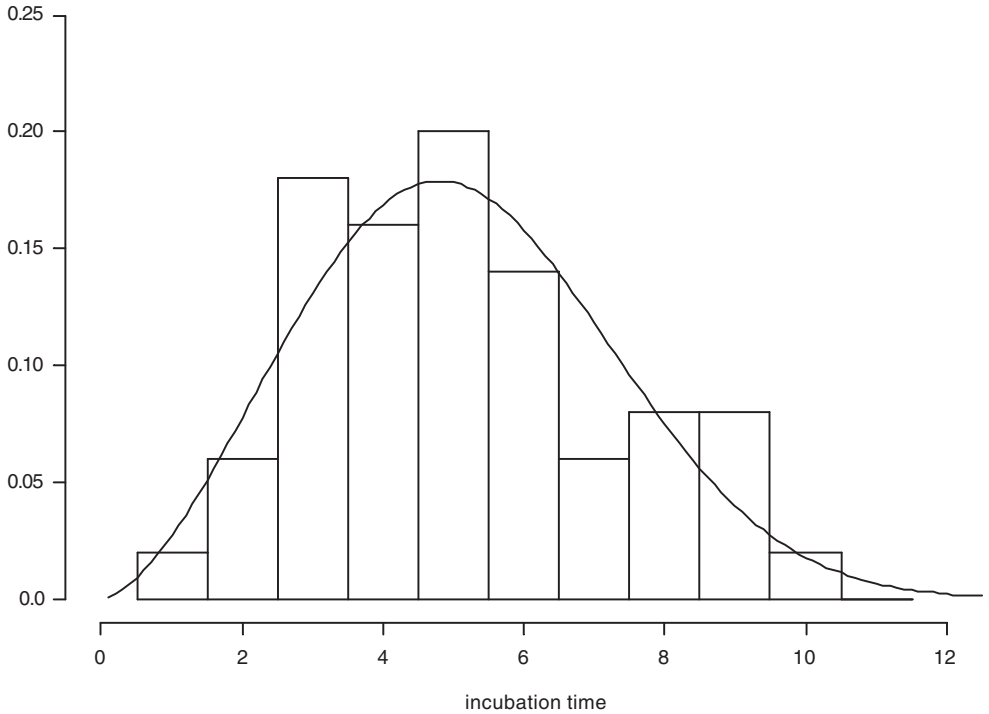

Figure 3. Histogram of ascertained incubation times for a subset of patients and the corresponding Weibull density estimate.

mates $\hat{\alpha}(t_D) = 5.80$ (SE $= 0.38$), $\hat{\beta}(t_D) = 2.59$ (SE $= 0.32$) that maximize (4) with the estimates $\hat{\alpha}(s_D) = 6.08$ (SE $= 0.47$), $\hat{\beta}(s_D) = 2.83$ (SE $= 0.52$) that maximize convolution likelihood

$$\prod_{i \in D} f_{S_i}(s_i \mid d_i; \alpha, \beta) \tag{5}$$

based on serial interval data. Note that the product is taken over the same subset D as in (4) instead of over all patients as done in (3). In terms of standard error, it comes as no surprise that estimating the incubation distribution directly using $t_D = \{t_i, i \in D\}$ is better than estimating it indirectly from the serial interval data $s_D = \{s_i, i \in D\}$ through convolution likelihood. In terms of the actual values of the estimates, there is not much difference. As can be seen in Figure 4, there is good visual agreement between the direct and indirect Weibull estimates. The standardized differences between the indirect and direct estimates of α and β are 0.639 and 0.454, respectively which are not statistically significant. This lends support to the appropriateness of the Weibull assumption and the validity of the convolution likelihood method, as both are necessary for the indirect estimate to be close to the direct estimate. In contrast, there is a much more discernable visual difference in Figure 5 between the direct and indirect log-normal estimate. The standardized differences of the parameter estimates are now 0.731 and -1.353. Figure 5 suggests that the log-normal distribution is a less appropriate

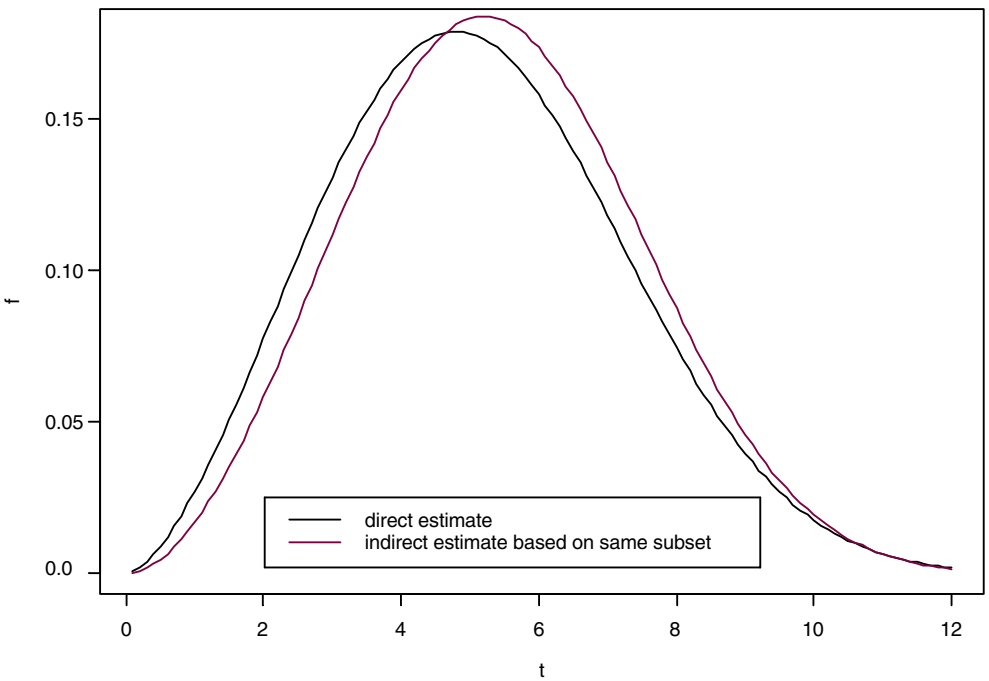

Figure 4. Direct and indirect Weibull estimates of the incubation distribution of SARS based on the same subset of patients.

choice for describing SARS incubation in Singapore. The results for log-logistic distribution are similar and will not be reported here to save space.

It is also possible to estimate the parameters α and β based on the ascertained incubation times for the 50 patients in D plus the serial times for the remaining 148 patients by maximizing the combined likelihood

$$\prod_{i \in D} f_T(t_i; \alpha, \beta) \prod_{i \notin D} f_{S_i}(s_i \mid d_i; \alpha, \beta) \tag{6}$$

to get the combined data estimates $\hat{\alpha}(t_D, s_{D^c}) = 5.50$, $\hat{\beta}(t_D, s_{D^c}) = 2.02$. By making use of the additional data $s_{D^c} = \{s_i, i \notin D\}$, the combined data estimates $\hat{\alpha}(t_D, s_{D^c})$, $\hat{\beta}(t_D, s_{D^c})$ are expected to be more efficient than the estimates $\hat{\alpha}(t_D)$, $\hat{\beta}(t_D)$ based on $t_D = \{t_1, i \in D\}$ only, and this is reflected by the smaller standard errors reported in Table I. Compared with the estimates $\hat{\alpha}(s) = \hat{\alpha}$
$(s_D, s_{D^c}) = 5.44$, $\hat{\beta}(s) = \hat{\beta}(s_D, s_{D^c}) = 1.91$ that we obtained in Section 2 by maximizing the likelihood (3) based on the full set of serial interval data, the values of the new estimates $\hat{\alpha}(t_D, s_{D^c}) = 5.50$, $\hat{\beta}(t_D, s_{D^c}) = 2.02$ are not that much different. The standard errors are also similar. Thus using the ascertained incubation times $t_D = \{t_i, i \in D\}$ instead of the serial intervals $s_D = \{s_i, i \in D\}$ does not lead to substantial improvement in efficiency when we have additional data $s_{D^c} = \{s_i, i \notin D\}$, and the two sets of estimates are equally good for practical purposes. A possible explanation for this is that there are only 50 cases in D compared with 148 cases in D^c.

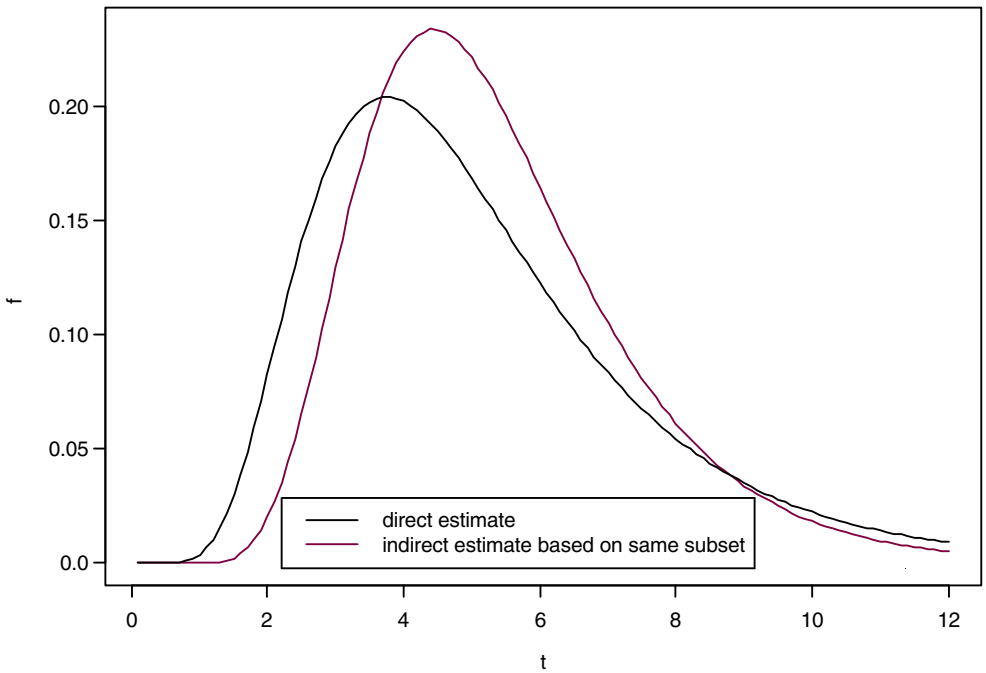

Figure 5. Direct and indirect log-normal estimates of the incubation distribution of SARS based on the same subset of patients.

Table I. A comparison of the maximum likelihood estimates based on the ascertained incubation times for 50 cases, the serial intervals for the remaining 148 cases, the combined data and the serial intervals for all 198 cases.

Data	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	Log-like
t_D	5.80 (0.38)	2.59 (0.32)	5.15 (0.35)	−30.786
s_{D^c}	5.25 (0.32)	1.75 (0.17)	4.68 (0.28)	−397.504
$t_D + s_{D^c}$	5.50 (0.25)	2.02 (0.15)	4.87 (0.22)	−431.805
$s = s_D + s_{D^c}$	5.44 (0.27)	1.91 (0.16)	4.83 (0.24)	−509.533

From Table I, we can also compare the estimates $\hat{\alpha}(t_D)$, $\hat{\beta}(t_D)$ based on the 50 cases with ascertained incubation times with the estimates $\hat{\alpha}(s_{D^c})$, $\hat{\beta}(s_{D^c})$ based on the remaining 148 cases. Assume $t_i \sim$Weibull (α_1, β_1) for $i \in D$ with corresponding mean $\mu_1 = \alpha_1\Gamma(1+\beta_1^{-1})$ and $t_i \sim$Weibull (α_2, β_2) with mean $\mu_2 = \alpha_2\Gamma(1+\beta_2^{-1})$ for $i \notin D$. A 95 per cent confidence interval for $\mu_1 - \mu_2$ is $5.15 - 4.68 \pm 1.96\sqrt{0.35^2 + 0.28^2} = (-0.41, 1.35)$ which contains zero and so the two mean estimates are not significantly different from one another. A more omnibus test of H_0: $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ versus H_1: $\alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2$ is the likelihood ratio test $W = 2\{431.805 - (30.786 + 397.504)\} = 7.03$ on two degrees of freedom. Thus there is some evidence that there could be some difference between the two underlying distributions but

their means appear to be compatible. As commented by Donnelly *et al.* [4], the ascertained incubation times are subject to 'biases in reporting, different routes of transmission, or varying infectious doses of the virus', and so the distribution of the ascertained incubation times may be somewhat different from the actual incubation distribution. This is another reason why the estimates $\hat{\alpha}(s_D, s_{D^c}), \hat{\beta}(s_D, s_{D^c})$ based on serial interval data alone reported in Section 2 might be more reliable, but as explained in the last paragraph, the results do not change much if we use $\hat{\alpha}(t_D, s_{D^c}), \hat{\beta}(t_D, s_{D^c})$ for this data set because of the small proportion of ascertained cases.

4. THE GENERALIZED F DISTRIBUTION

Another way to compare the Weibull, log-normal and log-logistic models is to embed them all within a wider family. A convenient choice is the four-parameter family of generalized F distributions (see Sections 2.2.7 of Reference [7]). More specifically, we assume that $\log T = \eta + \sigma W$, where W is distributed like the logarithm of an F variate with $2m_1$ and $2m_2$ degrees of freedom. It is known that $(m_1, m_2) = (1, 1)$ corresponds to a log–logistic distribution for T, $m_2 = \infty$ corresponds to the generalized gamma distribution which further reduces to the gamma distribution if $\sigma = 1$. A Weibull distribution for T corresponds to an extreme value distribution for W and this corresponds to $(m_1, m_2) = (1, \infty)$. The log-normal distribution is obtained by letting $(m_1, m_2) \to (\infty, \infty)$. To get around the difficulty of infinite degrees of freedom, Prentice [8] proposed the re-parameterization (see also Section 3.9.1 of Reference [7])

$$q = (m_1^{-1} - m_2^{-1})(m_1^{-1} + m_2^{-1})^{-1/2}$$

$$p = 2(m_1 + m_2)^{-1}$$

so that the log-logistic model corresponds to $p = 1$, $q = 0$. The log-normal, Weibull, reciprocal Weibull and generalized gamma distributions all belong to the three-parameter subfamily with $p = 0$ and $q = 0, 1, -1$ and > 0, respectively. When the three-parameter family is fitted to the subset of 50 incubation times, the log-likelihood function of q, with η and σ profiled out, is given in Figure 6. It can be seen that the maximum likelihood estimate of q is around 0.8 which is quite close to the q value of 1 for the Weibull model. By inverting the likelihood ratio test, an approximate 95 per cent confidence interval for q is roughly from 0.06 to 1.83. Thus the Weibull distribution is compatible with the observed data but not so for the log-normal distribution. If the three-parameter family is fitted to the combined data consisting of the observed incubation times for the 50 cases as well as the serial intervals for the remaining 148 cases, the maximum likelihood estimate of q is 0.57. Due to the use of additional data, the 95 per cent confidence interval of q that results from test inversion is tighter and ranges from 0.07 to 1.15, which is again compatible with the Weibull but not the log-normal model.

5. DISCUSSION

We have made the simplifying assumption that individuals are equally infectious throughout the period from the onset of symptoms to isolation. A consequence of this assumption is

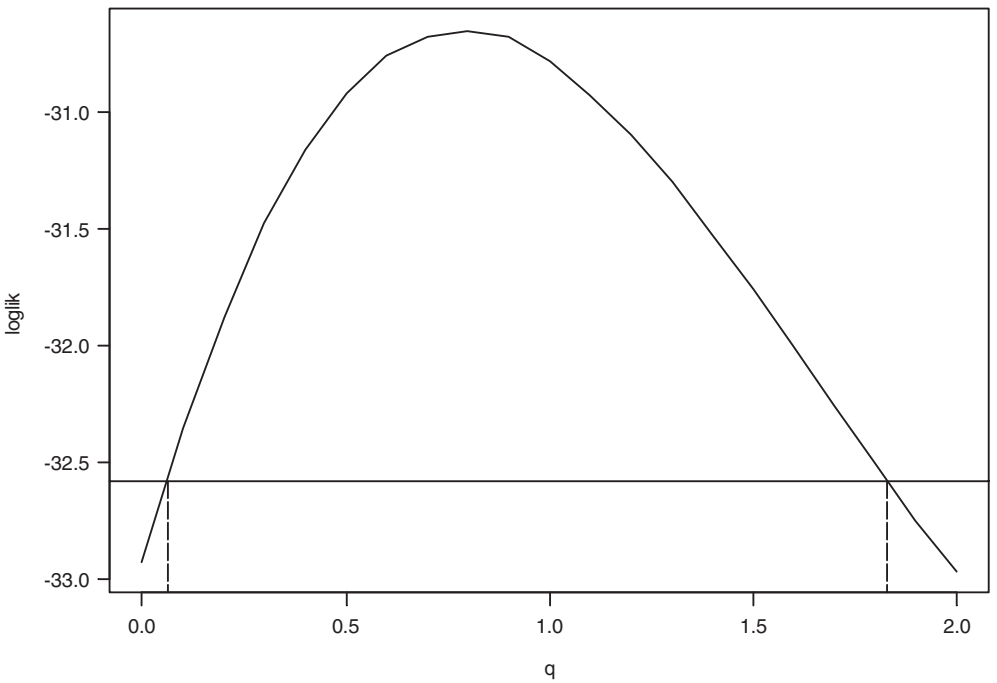

Figure 6. Profile log-likelihood of the generalized F model as a function of q with p fixed at 0 and η, σ profiled out.

that given the duration of infectiousness of an index case, the times of infection of those infected by him/her are uniformly distributed. In other words, we are assuming that u_i/d_i are uniformly distributed in the unit interval. A Pearson chi-square test of the goodness of fit of the uniform distribution is conducted using the subset of patients for whom the infection times u_i, and hence also u_i/d_i, can be determined. Partitioning the unit interval into 5 equal sub-intervals, the value of the chi-square statistic is 4.22 on 4 degrees of freedom which is not statistically significant (p value $= 0.38$). We do not expect the uniform assumption to be strictly true, as the clinicians we have spoken to believe that the viral load levels, and hence the infectiousness, of SARS patients will typically reach its peak in a number of days after onset of symptoms and then decrease. Nevertheless, we have obtained fairly reasonable estimate of the incubation distribution by using the uniform density $f_{U_i}(u_i \mid d_i) = 1/d_i$ to construct the convolution density of $s_i = u_i + t_i$. The validity of the resulting convolution likelihood estimate has also been established through comparison with a direct estimate, see Figure 4. In principle, we could assume that infections are transmitted by a symptomatic person according to a Poisson process with intensity function $\lambda(u)$. It follows that given that there is an infection during the duration d, the infection time is distributed according to the normalized density

$$f_U(u \mid d) = \frac{\lambda(u)}{\int_0^d \lambda(t)\,\mathrm{d}t}$$

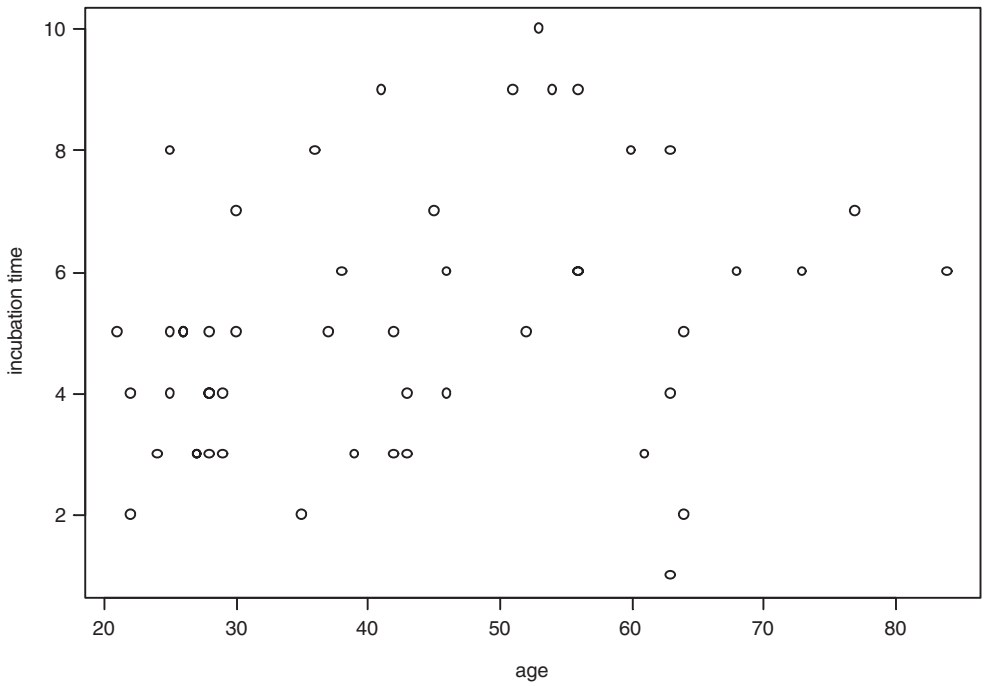

Figure 7. Scatterplot of ascertained incubation time versus age for a subset of patients.

rather than the uniform distribution. Unlike the case of uniformly distributed u_i, the convolution density $f_{S_i}(s_i \mid d_i) = \int_{t=0}^{\infty} f_T(t) f_{U_i}(s_i - t \mid d_i)\,\mathrm{d}t$ in this general case has no closed form even though numerical maximization of the convolution likelihood might be possible. Given that the uniform distribution assumption seems to be producing sensible estimates, it is not clear what are the benefits of using a more general distribution for u_i and whether these benefits will outweigh the computational difficulties involved. Other researchers [9] have also found constant infectiousness a useful and convenient working assumption.

A plausible conjecture is that the distribution of incubation times is age dependent and the proposed approach is flexible enough to incorporate this. However, a scatter plot of incubation time versus age for the subset data suggests that there is not much relationship between the two, see Figure 7. The Pearson correlation of 0.29 is also quite weak. For these reasons, we will not explore this possibility any further.

6. CONCLUSION

In conclusion, we have proposed in this paper a convolution method for estimating the SARS incubation distribution from serial interval data. For the SARS data in Singapore, the Weibull model seems to be more appropriate than other distributions like the log-normal and log-logistic. The validity of the indirect estimate obtained from serial interval data is established by way of comparison with a direct estimate based on the ascertained incubation times

for a subset of patients. An advantage of the maximum convolution likelihood estimates of α and β is that they are based on 198 serial intervals (SE = 0.27 and 0.16) and hence more precise than the direct estimates which are based on only 50 incubation times (SE = 0.38 and 0.32). It is possible to obtain a combined estimate that makes use of the incubation times for the 50 patients in the ascertained subset and the serial intervals for the remaining 148 cases. However, the resulting estimate is not that much different from that based on the full set of 198 serial intervals. The proposed method is general enough to allow non-uniform distribution for infection times and age dependence for incubation times but it appears that there is no strong reason for doing these.

Our calculation shows that the use of a 10-day quarantine period in Singapore is safe. Inevitably, some of the individuals quarantined as a precautionary measure actually do not have SARS. During the 2003 outbreak in Singapore, 7863 contacts were served home quarantine orders while a further 4331 were put on daily telephone surveillance for 10 days [10]. Out of all these, only 58 persons turned out to have SARS. The next challenge is how to be more discriminatory in identifying contacts who are at risk, so as to reduce the number of people to be placed under quarantine, without loss of effectiveness of the measure.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments and particularly to first referee for pointing out the connection with interval censoring. Preliminary results of this paper have been presented at a meeting of the Consortium on the Mathematical Modelling of Infectious Diseases, Ministry of Health, Singapore, and the authors would like to thank committee members for their helpful suggestions and comments. The first author would also like to thank the Ministry for allowing him access to the SARS data.

REFERENCES

1. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, Nicholls J, Yee WKS, Yan WW, Cheung MT, Cheng VCC, Chan KH, Tsang DNC, Yung RWH, Ng TK, Yuen KY, SARS study group. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 2003; **361**:1319–1325.
2. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S *et al*. A novel coronavirus associated with severe acute respiratory syndrome. *New England Journal of Medicine* 2003; **348**:1953–1966.
3. Drosten C, Günther S, Preiser W, van der Werf S, Brode HR, Becker S *et al*. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England Journal of Medicine* 2003; **348**:1967–1976.
4. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S *et al*. Epidemiological determinants of spread of casual agent of severe acute respiratory syndrome in Hong Kong. *Lancet* 2003; **361**:1761–1766.
5. Meltzer MI. Multiple contact dates and SARS incubation periods. *SARS Epidemiology* 2004; **10**:207–209.
6. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. Transmission dynamics and control of severe acute respiratory syndrome. *Science* 2003; **300**:1966–1970.
7. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
8. Prentice RL. Discrimination among some parametric models. *Biometrika* 1975; **62**:607–614.
9. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ *et al*. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* 2003; **300**:1961–1966.
10. Tan CC. Public health response: a view from Singapore. In *SARS*: *The First New Plague of the 21st Century*. Blackwell: London, to appear.

