The mixed trunsored model with applications to SARS in detail

Hideo Hirose

Department of Systems Innovation and Informatics
Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology
Iizuka, Fukuoka, 820-8502 Japan
email: hirose@ces.kyutech.ac.jp

Abstract

The trunsored model, which is a new incomplete data model regarded as a unified model of the censored and truncated models in lifetime analysis, can not only estimate the ratio of the fragile population to the mixed fragile and durable populations or the cured and fatal mixed populations, but also test a hypothesis that the ratio is equal to a prescribed value with ease.

Since SARS showed a severe case fatality ratio, our concern is to know such a case fatality ratio as soon as possible after a similar outbreak begins. The epidemiological determinants of spread of SARS can be dealt with as the probabilistic growth curve models, and the parameter estimation procedure for the probabilistic growth curve models may similarly be treated as the lifetime analysis. Thus, we try to do the parameter estimation to the SARS cases for the infected cases, fatal cases, and cured cases here, as we usually do it in the lifetime analysis. Using the truncated data models to the infected and fatal cases with some censoring time, we may estimate the total (or final) numbers of the patients and deaths, and the case fatality ratio may be estimated by these two numbers. We may also estimate the case fatality ratio using the numbers of the patients and recoveries, but this estimate differs from that using the numbers of the patients and deaths, especially when the censoring time is located at early stages.

To circumvent this inconsistency, we propose a mixed trunsored model, an extension

of the trunsored model, which can use the data of the patients, deaths, and recoveries simultaneously. The estimate of the case fatality ratio and its confidence interval are easily obtained in a numerical sense.

This paper mainly treats the case in Hong Kong. The estimated epidemiological determinants of spread of SARS, fitted to the infected, fatal, and cured cases in Hong Kong, could be the logistic distribution function among the logistic, lognormal, gamma, and Weibull models. Using the proposed method, it would be appropriate that the SARS case fatality ratio is roughly estimated to be 17% in Hong Kong. Worldwide, it is roughly estimated to be about 12-18%, if we consider the safety side without the Chinese case. Unlike the questionably small confidence intervals for the case fatality ratio using the truncated models, the case fatality ratio in the proposed model provides a reasonable confidence interval.

Keywords: truncated data; grouped data; generalized logistic distribution; case fatality rate; case fatality ratio; mortality rate; case survival ratio; bootstrap.

1. Introduction

A. Motivation and Objectives

WHO reports Severe Acute Respiratory Syndrome (SARS) outbreak as shown in Appendix (see also [42, 43]). During almost a month from 21 February the SARS virus spread without isolation of probable patients. Taking into account of the short incubation period which is estimated as five to eight days (see [22]), it appears that the virus raged for more than a month without prevention. The number of probable patients appeared to grow exponentially in this period, and then the control of the human-to-human chain of transmission of the disease suppressed the growth rate of spreading. It may be considered that only one seed made a typical epidemic growth curve of the disease spread. Our concern is first what the appropriate probability distribution for the curve is; the logistic, the lognormal, the gamma, or the Weibull distribution may be fitted to the data provided by WHO ([41]).

As SARS showed a severe case fatality ratio (abbreviated CFR here like in [20], but other terms such as case fatality rate in reference [7, 35, 36] or mortality rate in reference

[40] are also used), our second concern is to know the ratio as soon as possible after the outbreak began. Since WHO opens the numbers of probable cases and the fatal cases to the public day by day, we can estimate the CFR by some censoring time T using the conditional likelihoods for both the probable and fatal cases; this approach is considered to be the truncated model approach. However, WHO, in addition to these two data cases, gave us the recovery (or cured) cases, which would be the fruitful information for the parameter estimation of the underlying probability distributions; we can also estimate the CFR using the probable and cured cases. We propose here a new estimation method for the parameters of the underlying distributions and the CFR using the three data sets of probable, fatal, and cured cases together. The trunsored model approach (Hirose [17, 18]) can do this, but the traditional truncated model approach cannot.

The objective of the introduction of the trunsored model was to do hypothesis tests easily (Hirose [17, 18]). This purpose may also be realized in our situation that we use the three data sets together. However, we do not go deeply into such a direction in this paper; we introduce the estimation methods of the underlying probability distribution parameters and of the CFR.

B. Statistical Background

In some lifetime estimation problems, short-term survivors and long-term survivors are mixed: for example, Boag [3], Farewell [10], and Goldman [12] discussed the proportion of patients cured by a particular treatment; Anscombe [1] treated market penetration; Maltz and McCleary [26], and Steinhurst [31] discussed recidivism; Meeker [27] and Hirose [17, 18] applied the model to integrated circuit reliability. Maller and Zhou [25], Zhou and Maller [37], Sun and Zhou [32], Vu, Maller, and Zhou [34], Peng, Dear, and Carriere [29] discussed the model as long-term survivors. Tsodikov, Ibrahim, and Yakovlev recently review the cure rates [33]. In such cases, r events within T are observed from n samples, but the ratio, p_m, of the long-term survivors to the mixed populations is unknown. If n is unknown, the truncated model (e.g., Johnson, Kotz and Balakrishnan [21]; Meeker and Escobar [28]; Wallace, Blischke and Murthy [39]; and Klein and Moeschberger [23]) could be applied. However, the information n may be useful in our situation; one of the advantages to adopt this kind of model is described as the application of the likelihood ratio test in Hirose [17, 18].

The epidemiological determinants of spread of SARS can be dealt with as the probabilistic growth curve models [24], and the parameter estimation procedure for the prob-

abilistic growth curve models may similarly be treated as the lifetime analysis. Thus, we try to do the parameter estimation to the SARS cases for the infected cases, fatal cases, and cured cases here, as we usually do it in the lifetime analysis. To estimate the CFR caused by SARS, the truncated model approach using the infected and fatal growth curves may be fine. However, the recovery rate by the same approach using the infected and cured growth curves may not be consistent with the CFR obtained by using the infected and fatal cases. Thus, the truncated approach cannot have such consistency. A new approach proposed here, *the mixed trunsored model*, can have, however. Donnelly et al. [7] computed the CFR with the admission-to-death and admission-to-discharge distributions, but the proposed method shown here used the infected case distribution in addition.

2. Trunsored model

2.1 Single Trunsored Model

We define a cumulative probability distribution function, $H(t;\psi)$, which is a linear combination of $F(t;\theta)$ and $G(t;\phi)$ given by

$$\begin{aligned} H(t;\psi) = sF(t;\theta) + (1-s)G(t;\phi), \\ (t \geq 0,\ -\infty < s < \infty), \end{aligned} \tag{1}$$

with a combination parameter s, and the corresponding pdf, $h(t;\psi)$, for H is also defined

$$h(t;\psi) = sf(t;\theta) + (1-s)g(t;\phi). \tag{2}$$

Then, the likelihood function for the combined model can be expressed in the form

$$L(\psi) = \{1 - H(T;\psi)\}^{n-r} \cdot \prod_{i=1}^{r} h(t_i;\psi), \tag{3}$$

where t_i denotes the observed times that events occurred. If we assume that the censoring time, T, is smaller than the left endpoint, T_0, of $G(t)$ such that

$$G(T) = 0, \quad g(t_i) = 0, \quad (t_i < T < T_0, \quad i = 1, \cdots, n), \tag{4}$$

i.e., G implies the long-term survivors, then $L(\psi) \to L_{ts}(\theta, s)$, where

$$L_{ts}(\theta, s) = \{1 - sF(T;\theta)\}^{n-r} \cdot \prod_{i=1}^{r} \{sf(t_i;\theta)\}. \tag{5}$$

This is the likelihood for the trunsored model in Hirose [17, 18].

For the sake of comparison, we define two additional likelihood functions for the censored model and the truncated model as

$$L_c(\theta) = \{1 - F(T;\theta)\}^{n-r} \cdot \prod_{i=1}^{r} f(t_i;\theta), \tag{6}$$

$$L_t(\theta) = \prod_{i=1}^{r} \{f(t_i;\theta)/F(T;\theta)\}. \tag{7}$$

2.2 Mixed Trunsored Model

We consider cumulative probability distribution functions, F_j $(j = 1, \cdots, J)$, with trunsored likelihoods such that

$$L_{ts}^{j}(\theta_j, s_j) = \{1 - s_j F_j(T;\theta_j)\}^{n_j - r_j} \cdot \prod_{i=1}^{r_j} \{s_j f_j(t_i;\theta_j)\}, \tag{8}$$

under the restriction that

$$\zeta(s_1, \cdots, s_J) = 0, \tag{9}$$

where n_j $(j = 1, \cdots, J)$ are the number of samples, and r_j $(j = 1, \cdots, J)$ are the number of observed events. If restriction (9) is not imposed, the likelihood equations in (8) can be solved independently; with the restriction, however, we need to solve the likelihood equations simultaneously. In SARS applications, F_1, F_2, and F_3 may correspond to the infected case, fatal case, and cured case growth curves, respectively; restriction (9) implies that the probable cases are divided into exactly two categories: the fatal and the recovered cases as in (10)

$$s_1 = s_2 + s_3. \tag{10}$$

Then, we can estimate the parameters, s_j and θ_j, by maximizing the likelihood function for the mixed trunsored model,

$$L_{mts}(\theta, s) = \prod_{j=1}^{J} L_{ts}^{j}(\theta_j, s_j). \tag{11}$$

If the time of event is not observed and the number of events in some period, e.g., from T_i to T_{i+1}, are observed instead, we consider the grouped data model such that

$$L_{ts}(\theta, s) = \{1 - sF(T;\theta)\}^{n-r} \cdot \prod_{i=1}^{k} [s\{F(T_{i+1}) - F(T_i)\}]. \tag{12}$$

In SARS case, T_i to T_{i+1} may be one day, two days, or three days.

3. Probability distributions

We consider four typical probability distribution models for the growth curves: the generalized logistic distribution (GL) [44], the extended lognormal distribution (ELN) [15], the extended gamma distribution (EGM) [14], and the generalized extreme-value distribution (GEV) [13], to allow the negative and positive skewness in the distribution functions [16]; the number of parameters are three including the location parameter.

The logistic distribution with two parameter is often used as the growth model because this distribution is derived from the differential equation for the biological models; the generalized logistic curve [44], also known as Richards' curve [30], is a widely-used and flexible function for growth modeling by including the shape parameter in the model. The probability density function and the cumulative distribution function for GL are expressed by,

$$f_{GL}(x;\sigma,\mu,\beta) = \frac{\beta\exp(-z)}{\sigma\{1+\exp(-z)\}^{\beta+1}}, \tag{13}$$

$$F_{GL}(x;\sigma,\mu,\beta) = \frac{1}{\{1+\exp(-z)\}^{\beta}}, \tag{14}$$

$$(z=(x-\mu)/\sigma, \quad -\infty < x <\infty, \quad -\infty<\mu<\infty, \quad \sigma>0, \quad \beta>0).$$

This distribution is negatively skewed when $\beta<1$, and is positively skewed when $\beta>1$. It is symmetric when $\beta=1$, as is known to two parameter logistic distribution.

As mentioned in section 1, probabilistic growth curves of the spread of SARS fitted to the infected cases, fatal cases, and cured cases can similarly be treated to the lifetime distributions, we deal with three typical probability distribution models used in the lifetime analysis. The density functions for ELN, EGM, and GEV are expressed by,

$$f_{ELN}(x;\sigma,\mu,\lambda) = \frac{1}{\sqrt{2\pi}\sigma\{1+\lambda z\}}\exp\Big(-\frac{[\log\{1+\lambda z\}]^2}{2\lambda^2}\Big), \tag{15}$$

$$f_{EGM}(x;\sigma,\mu,\lambda) = \frac{1}{\sigma|\lambda|\Gamma(\lambda^{-2})}\left(\frac{1+\lambda z}{\lambda^2}\right)^{\lambda^{-2}-1}\exp\left\{-\left(\frac{1+\lambda z}{\lambda^2}\right)\right\}, \tag{16}$$

$$f_{GEV}(x;\sigma,\mu,\lambda) = \frac{1}{\sigma}\big(1+\lambda z\big)^{1/\lambda-1}\exp\Big\{-\big(1+\lambda z\big)^{1/\lambda}\Big\}, \tag{17}$$

with

$$\sigma>0, \quad \lambda\neq 0, \quad 1+\lambda z>0, \quad z=\frac{x-\mu}{\sigma}. \tag{18}$$

These three distribution models are the extension models from the log-normal (LN), gamma (GM), and Weibull (WB) distributions, respectively, with densities,

$$f_{LN}(x;\alpha,\tau,\gamma) = \frac{1}{\sqrt{2\pi}(x-\alpha)\tau}\exp\Big[-\frac{\{\log(\frac{x-\alpha}{\gamma})\}^2}{2\tau^2}\Big], \quad (x>\alpha, \quad \tau>0, \quad \gamma>0) \tag{19}$$

$$f_{GM}(x;\alpha,\beta,\gamma) = \frac{1}{\gamma\Gamma(\beta)}\left(\frac{x-\alpha}{\gamma}\right)^{\beta-1}\exp\left\{-\left(\frac{x-\alpha}{\gamma}\right)\right\}, \quad (x\geq\alpha, \quad \beta>0, \quad \gamma>0), \tag{20}$$

$$f_{WB}(x;\eta,\beta,\gamma) = \frac{\beta}{\eta}\left(\frac{x-\gamma}{\eta}\right)^{\beta-1}\exp\left\{-\left(\frac{x-\gamma}{\eta}\right)^{\beta}\right\}, \quad (x\geq\gamma, \quad \eta>0, \quad \beta>0). \tag{21}$$

4. Applications to SARS

4.1 WHO Data

WHO opened the daily number of probable cases from March 17, 2003, to July 11, 2003, to the public [41]; On September 26, 2003, summary of probable SARS cases with onset of illness from November 1, 2002, to July 31, 2003, is additionally opened. As mentioned earlier, the outbreak began by only one seed in Hong Kong; the growth curves for infected cases, fatal cases, and cured cases in Hong Kong are smooth and natural comparing to those in other districts such as China, Taiwan and Canada; for example in Canada, two successive asynchronous outbreaks occurred. Here, we deal with a rather simple case such as the case in Hong Kong as a primary analysis. The cumulative numbers of infected patients, deaths, and recovered persons from March 17, 2003, to July 11, 2003, are shown in Table 1.

4.2 Appropriate Distribution Model using the Truncated Model

To find the most appropriate probability distribution model in the four models introduced previously, we first fit the four models to SARS data for the infected, fatal, and cured cases. Using the truncated model of (7) with censoring time on July 11, 2003, the maximum values of the log-likelihood functions are obtained as shown in Table 2, resulting that the generalized logistic model has the largest likelihood values for the infected,

fatal, and cured cases. The difference of the likelihood values between the log-normal and the gamma is not so large; however, the difference between the generalized logistic and the log-normal and that between the generalized logistic and the Weibull are significantly large. We use the generalized logistic model from now on.

The estimated cumulative probability distribution functions of the generalized logistic distribution and the empirical distribution functions for the patients, fatal, and cured cases are shown in Figure 1; circles, triangles, and squares in the figure express the empirical functions for patients, fatal, and cured cases, respectively, and the dashed lines are estimated distribution functions. It appears that the shapes of the three probability distribution functions are almost the same; only the location parameter seems to be different. We therefore may assume that the shape and scale parameters for these three distributions are the same; under such an assumption, the maximum likelihood estimates for the parameters in (13) and (14) are $\hat{\sigma} = 12.559$, $\hat{\lambda} = 3.2697$, $\hat{\mu}_1 = 3.9973$ (infected case), $\hat{\mu}_2 = 25.316$ (fatal case), $\hat{\mu}_3 = 25.343$ (cured case), and the corresponding log-likelihood value is -13532.9, which is smaller than the value of sum of the three independently obtained maximum log-likelihood values, -13515.0, for the patients, fatal, and cured cases, where time $t = 0$ is set to the date on March 16, 2003; see Table 2. Here, we use the notation of $\theta_j = (\sigma, \lambda, \mu_j)^T$.

(INSERT TABLE 1, 2 AND FIGURE 1 ABOUT HERE.)

4.3 Case Fatality Ratio by the Truncated Model Approach

The observed numbers of the patients and deaths are considered to be grouped (day by day) and right truncated. By computing both the total expected numbers of patients and deaths, it seems that we can estimate the CFR as shown below, but the estimate seems to be questionable.

(a) Inconsistency of the estimate

Using the truncated model likelihood to the infected patients, we can estimate the total number of patients, m_1, in the future. If the estimated parameter is $\hat{\theta}_1$, then $\hat{m}_1$ can be estimated by

$$\hat{m}_1 = r_1 / F_1(T_1; \hat{\theta}_1), \tag{22}$$

where, T_1 is the censoring time. Similarly, the total number of fatal cases, $\hat{m}_2$, and the total number of cured cases, $\hat{m}_3$, are also calculated easily, if parameters, $\hat{\theta}_2$ and $\hat{\theta}_3$, are obtained.

The CFR, p_f, and the case survival ratio (abbreviated CSR here, p_s, are estimated by

$$\hat{p}_f = \hat{m}_2/\hat{m}_1, \quad \hat{p}_s = \hat{m}_3/\hat{m}_1, \tag{23}$$

where the CSR is defined by the number of survivors divided by the number of patients in this paper.

As mentioned above, the best fit probability distribution model is the generalized logistic distribution, thus we may obtain the CFR by applying the truncated models with the generalized logistic distribution to the infected and fatal cases. When we set the censoring time, $T = T_1 = T_2$, on July 11, 2003, and we suppose that the scale and shape parameters are the same for patients, deaths, and recoveries, then we can obtain the estimates, $\hat{m}_1 = 1,755.71$ and $\hat{m}_2 = 298.66$; thus, the CFR, $\hat{p}_f$, becomes 17.01%. If we use the estimate of the total number of cured cases, $\hat{m}_3 = 1,436.17$, then the CSR $\hat{p}_s = 81.80\%$ (i.e., $\hat{p}_f = 18.20\%$) is obtained. Here, these two ratios under the truncated model approach are obtained by solving the simultaneous likelihood equations,

$$\frac{\partial \log L_t(\theta_j)}{\partial \theta_j} = 0, \quad (j = 1, 2, 3), \tag{24}$$

where $\theta_j = (\sigma, \lambda, \mu_j)^T$ because we supposed that $\sigma_j = \sigma, \lambda_j = \lambda,\ (j = 1, 2, 3)$; the number of unknown parameters are 5 (σ, λ, μ_1, μ_2, μ_3). However, the sum of the CFR, obtained by using the fatal and infected cases, and the CSR, obtained by using the cured and infected cases, is not equal to 1. If we set the censoring time on May 25, 2003, this discrepancy becomes markedly large; we obtain $\hat{m}_1 = 1,740.23$, $\hat{m}_2 = 278.90$, and $\hat{m}_3 = 1,346.46$, then the estimated CFR and the CSR are, $\hat{p}_f = 16.03\%$ and $\hat{p}_s = 77.37\%$. It would be crucial to get rid of this inconsistency even in earlier stages, i.e., the censoring time is earlier.

(b) Paradox of the error

Using the bootstrap method [8, 9] with 1,000 resampling, we can obtain the confidence interval for the CFR. When we set the censoring time on May 25, 2003, the 95% confidence interval for the CFR is computed as $13.60\% \leq p_f \leq 17.40\%$. This value seems to be acceptable. If the censoring time is set to the right far enough, e.g., on July 11, 2003, however, the estimated number of patients, $\hat{m}_1 = 1,755.71$, and the estimated number of deaths, $\hat{m}_2 = 298.66$, become very close to the observed numbers of patients, 1755, and deaths, 298, by that time; in other resampling cases, the results are much the same. Then,the 95% confidence interval for the CFR is computed as $16.90\% \leq p_f \leq 17.09\%$

(heavily skewed as shown in Figure 2). Such very small confidence intervals are also reported elsewhere ([6]). After the outbreaks are completely ceased, e.g., based on data as of the December 31, 2003, the CFR might be computed with extremely small variance, if we use the conditional likelihood. For example, in Hong Kong, the CFR would become to be just $299/1,755(= 17.0370370...\%)$ if no new patients, deaths, and recoveries were observed at all after December 31, 2003; similarly in Taiwan, just $37/346(= 10.69...\%)$ is expected; in Singapore, just $33/238(= 13.86...\%)$; in Canada, just $43/251(= 17.13...\%)$. However, the number of deaths in Hong Kong, for example, may differ from that in other situations; for example, the number of deaths 299 could be 301 by chance; then, the CFR would be changed to some other value $(301/1,755(= 17.15099715...\% > 17.09\%))$. Assuming that the CFR of SARS is supposed to be some constant value, then the number of deaths would be varied by chance. The CFRs in various districts could be fluctuated, but they would be covered by some interval, say $[0.1, 0.2]$. This is the reason why I think that the very small confidence intervals obtained by using the truncated model are paradoxical.

(INSERT FIGURE 2 ABOUT HERE.)

4.4 Mixed Trunsored Model Approach and the Case Fatality Ratio

Based on the truncated model, inconsistent estimates for the CFR and paradoxical confidence intervals are computed. To circumvent these flaws, we next use the proposed method, the mixed trunsored model.

All the patients are divided exactly into two categories: fatal cases and cured cases. This means that $p_f + p_s = 1$. This restriction cannot be imposed to the truncated model approach straightforwardly. The trunsored model approach using (8-12), however, can do this; we only need to impose the restriction that $s_3 = s_1 - s_2$. The CFR and the CSR are calculated by

$$p_f = s_2/s_1, \quad p_s = s_3/s_1 = 1 - p_f. \tag{25}$$

Setting n_j $(j = 1, 2, 3)$ to some numbers, e.g., the actual population in Hong Kong (this is about 6,810,000 persons in 2003 [4]), the estimated parameters, under the assumption that $\sigma_j = \sigma$ $(j = 1, 2, 3)$ and $\lambda_j = \lambda$ $(j = 1, 2, 3)$, are $\hat{\sigma} = 12.560$, $\hat{\lambda} = 3.2708$, $\hat{\mu}_1 = 3.9911$, $\hat{\mu}_2 = 25.310$, $\hat{\mu}_3 = 25.337$, $\hat{s}_1 = 0.21147 \times 10^{-3}$, $\hat{s}_2 = 0.44223 \times 10^{-4}$, and the corresponding log-likelihood value is $-46,577$ when we set the censoring time on July 11, 2003; thus, $\hat{p}_f = 1 - \hat{p}_s = 17.30\%$ is obtained. If we set the censoring time on May 25, 2003, the CFR

is computed as $\hat{p}_f = 17.16\%$, which is almost the same value as that when the censoring time is July 11, 2003. The values of the estimates, $\hat{s}_j$ $(j = 1, 2, 3)$, are not important by themselves; they change their values by setting n_j $(j = 1, 2, 3)$ to other values, but $\hat{p}_f$ and $\hat{p}_s$ are hardly affected by these values.

The CFR under the mixed trunsored model approach with 7 (σ, λ, μ_1, μ_2, μ_3, s_1, s_2) unknown parameters are shown in Figure 3 when we vary the censoring time T. The estimated value of the CFR at time t in the figure means that the estimate is obtained under the assumption that the censoring time T is equal to t. In the truncated model, the CFRs are obtained by two estimates: one is by using the numbers of the patients and deaths, and the other is by using the the numbers of the patients and recoveries. In Figure 3, these two CFRs under the truncated model approach are also shown. We can see that the estimated CFRs in the mixed trunsored model keep almost a constant value in a wide range of censoring time, while the CFRs in the truncated model do not, as mentioned above.

(INSERT FIGURE 3 ABOUT HERE.)

The 95% confidence intervals for the estimates of the CFR using the bootstrap method are computed as $15.51\% \leq p_f \leq 19.13\%$ and $13.73\% \leq p_f \leq 19.04\%$ when the censoring time is set to on July 11, 2003, and on May 25, 2003, respectively. The corresponding standard deviations, $SD(\hat{p}_f)$, are 0.92% and 1.35%, respectively. These values are considered to be reasonable and acceptable; see the next section. The histogram of the bootstrapped estimates for the CFR, when the censoring time is on July 11, 2003, is shown in Figure 4. The frequency distributions of the bootstrapped estimates for the CFRs at various censoring times are shown in Figure 5. We can see that the confidence interval of the CFR at earlier estimating stage, e.g., 70th day from March 17, 2003, i.e., May 25, 2003, is wider than that at the final stage, but they are not so different from each other.

(INSERT FIGURES 4 AND 5 ABOUT HERE.)

5. Discussion

5.1 Robustness against the Amount of n_j

The confidence intervals for the CFR are obtained under the assumption that $n_j = 6,810,000$ $(j = 1, 2, 3)$; other values of n_j $(j = 1, 2, 3)$ will provide different confidence

intervals, but the confidence intervals are not affected much as long as the values of n_j $(j = 1, 2, 3)$ are not so small. For example, using $n_j = 681,000$ $(j = 1, 2, 3)$, the 95% confidence intervals for the CFR are computed as $15.52\% \leq p_f \leq 19.11\%$ and $13.64\% \leq p_f \leq 19.20\%$ when the censoring time is set to on July 11, 2003, and on May 25, 2003, respectively.

5.2 Approximate Standard Deviation of the Case Fatality Ratio

The variance of a ratio X/Y is approximately obtained by

$$Var\Big(\frac{X}{Y}\Big) \approx \Big(\frac{E(X)}{E(Y)}\Big)^2 \times \Big(\frac{Var(X)}{E(X)^2} - 2\frac{Cov(X,Y)}{E(X)E(Y)} + \frac{Var(Y)}{E(Y)^2}\Big), \tag{26}$$

where X and Y are random variables [2]. We assume that $X = s_2$ and $Y = s_1$. When the censoring time is late enough, then $E(X)$ and $E(Y)$ become s_2 and s_1, and $Var(X)$ and $Var(Y)$ become approximately $s_2(1 - s_2)/n_2$ and $s_1(1 - s_1)/n_1$. Using $Cov(X, Y) = \rho\sqrt{Var(X)Var(Y)}$, (26) is approximately reduced to

$$Var(\hat{p}_f) \approx \hat{p}_f^2 \times \Big(\frac{1}{\hat{n}_p} - \frac{2\rho}{\sqrt{\hat{n}_p\hat{n}_d}} + \frac{1}{\hat{n}_d}\Big), \tag{27}$$

where $\hat{n}_p$ and $\hat{n}_d$ are the estimates for the numbers of patients and deaths; ρ denotes the correlation coefficient, $Corr(X, Y)$, between X and Y. Since $\hat{n}_p$ and $\hat{n}_d$ are estimated as $1,741.3$ and 301.16, the approximate standard deviation of the CFR, $SD(\hat{p}_f)$, varies $0.00582 \leq SD(\hat{p}_f) \leq 0.01079$ according to the value of the correlation coefficient, $0 \leq \rho \leq 1$, which is consistent to the standard deviation obtained by the bootstrap in the mixed trunsored model.

Using the number of patients, deaths, and recoveries by the date of the December 31, 2003 in various infected districts, approximate CFRs and their 95% confidence intervals are computed by (27); they are shown in Table 3 and Figure 6. In the figure, the solid and dashed lines express the 95% confidence intervals when $\rho = 0$ and when $\rho = 1$, respectively. A very rough interval for the CFR, $[12, 18]\%$, includes points in the 95% confidence intervals of Canada, Hong Kong, Taiwan, Singapore, and Viet Nam, but does not include points in the 95% confidence interval of China. According to [41], 325 cases have been discarded in Taiwan since 11 July, 2003 because Laboratory information was insufficient or incomplete for 135 discarded cases, of which 101 died. World-wide, the CFR of about 9.6% (including Chinese cases) has been announced by media. However, this

estimate should be treated cautiously; this is caused mainly by the Chinese CFR, and this value, about 6.6%, is very different from those in other countries. There would be reasons for such a very different value of the CFR. One reason would be that Chinese infected cases were counted circumspectly. However, a noticeable reference is also seen (see [5]), in which Chinese medicine is found to improve the case survival rate in the treatment of SARS. In any case, it would be appropriate that the SARS CFR is estimated without the Chinese case if we consider the safety side. In such a case, it is roughly estimated to be about 12-18%, worldwide.

(INSERT TABLE 3 AND FIGURE 6 ABOUT HERE.)

6. Concluding remarks

The epidemiological determinants of spread of SARS can be dealt with as the probabilistic growth curve models, and the parameter estimation procedure for the probabilistic growth curve models may similarly be treated as the lifetime analysis. Thus, we try to do the parameter estimation to the SARS cases for the infected cases, fatal cases, and cured cases, here, as we usually do it in the lifetime analysis. The truncated data model approach using the infected and fatal cases can estimate the case fatality ratio of the disease, but it also estimates the case fatality ratio using the numbers of the patients and recoveries; these estimates differ from each other in early censoring time stage. To circumvent this inconsistency, and to obtain reasonable estimates, the mixed trunsored model, which is an extension of the censored and truncated unified model, is found to be useful in estimating the case fatality ratio of SARS, when we use the data of the patients, deaths, and recoveries together. Using the proposed method, it would be appropriate that the SARS case fatality ratio is roughly estimated to be about 12-18% worldwide, if we consider the safety side without the Chinese case. Unlike the questionably small confidence intervals for the case fatality ratio using the truncated models, the case fatality ratio in the proposed model provides a reasonable confidence interval.

References

[1] F.J. Anscombe, "Estimating a mixed-exponential response law", *Journal of the American Statistical Association*, **56**, (1961) 493-502.

[2] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis, Theory and Practice* MIT Press (1975).

[3] J.W. Boag, "Maximum likelihood estimates of the proportion of patients cured by cancer therapy", *Journal of the Royal Statistical Society - Series B*, **11**, (1948) 11-53.

[4] Bureau of East Asian and Pacific Affairs, http://www.state.gov/r/pa/ei/bgn/2747.htm (2004)

[5] Z. Chen and T. Nakamura, "Statistical evidence for the usefulness of Chinese medicine in the treatment of SARS". *Phytotherapy Research*, **18**, (2004) 592-594.

[6] Z. Chen and T. Nakamura, "Statistical estimation method and its reliability of SARS". *Japanese Federation of Statistical Science Association Convention Record*, (2005) 33-339. (in Japanese)

[7] C.A. Donnelly, A.C. Ghani, G.M. Leung, et al., "Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong". *Lancet*, **361**, (2003) 1761-1766.

[8] B. Efron, "Bootstrap methods, another look at the jackknife", *Annals of Statistics*, **7**, (1979) 1-26.

[9] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans* Society of Industrial and Applied Mathematics, Philadelphia (1982).

[10] V.T. Farewell, "A model for a binary variable with time-censored data", *Biometrika*, **64**, (1977) 43-46.

[11] V.T. Farewell, and R.L. Prentice, "A study of distribution shape in life testing", *Technometrics*, **19**, (1977) 69-76.

[12] A.I. Goldman, "Survivorship analysis when cure is a possibility", a Monte Carlo study, *Statistics in medicine*, **3**, (1984) 153-167.

[13] H. Hirose, "Parameter estimation in the extreme-value distributions using the continuation method", *Transactions of Information Processing Society of Japan*, **35**, (1994) 1674-1681.

[14] H. Hirose, "Maximum likelihood parameter estimation in the three-parameter gamma distribution", *Computational Statistics and Data Analysis*, **20**, (1995) 343-354.

[15] H. Hirose, "Maximum likelihood estimation in the three-parameter log-normal distri-

bution using the continuation method", *Computational Statistics and Data Analysis*, **24**, (1997) 139-152.

[16] H. Hirose, "Maximum likelihood parameter estimation by model augmentation with applications to the extended four-parameter generalized gamma distribution", *Mathematics and Computers in Simulation*, **54**, (2000) 81-97.

[17] H. Hirose, "Trunsored data analysis with applications to field data", *Hawaii International Conference on Statistics and Related Fields*, (2002) June 5-9, Honolulu.

[18] H. Hirose, "The Trunsored model and its applications to lifetime analysis, unified censored and truncated model", *IEEE Transactions on Reliability*, **54** (2005) 11-21.

[19] H. Hirose, "The mixed trunsored model with applications to SARS", submited.

[20] N.P. Jewell, X.D. Lei, et al., "Estimation of the case fatality ratio with competing risks data: an application to severe acute respiratory syndrome (SARS)". *U. C. Berkeley Division of Biostatistics Working Paper Series*, 176. (2005)

[21] N.L. Johnson, and S. Kotz, and Balakrishnan, N. (1994), *Continuous Univariate Distributions, Vol.1, 2nd ed.* Wiley, New York (1994).

[22] B.S. Kamps, and C. Hoffmann, SARSReference, *http://sarsreference.com/*, Flying Publisher (2003)

[23] J.P. Klein, and M.L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data, 2'nd ed.* Springer, New York (2004).

[24] D. Lai , "Monitoring the SARS epidemic in China: A time series analysis", *Journal of Data Science*, **3**, (2005) 279-293.

[25] R.A. Maller, and S. Zhou, *Survival analysis with long-term survivors* Wiley, New York (1996).

[26] M.D. Maltz, and R. McCleary, "The mathematics of behavioral change, recidivism and construct validity", *Evaluation Quarterly*, **1**, (1977) 421-438.

[27] W.Q. Meeker, "Limited failure population life tests, application to integrated circuit reliability", *Technometrics*, **29**, (1987) 51-65.

[28] W.Q. Meeker, and L.A. Escober, *Statistical Methods for Reliability Data* Wiley, New York (1998).

[29] Y. Peng, K.B.G. Dear, and K.C. Carriere, "Testing for the presence of cured patients, a simulation study", *Statistics in Medicine*, **20**, (2001) 1783-1796.

[30] F.J. Richards, "A flexible growth function for empirical use", *Journal of Experimental Botany* , **10**, (1959) 290-300.

[31] W.R. Steinhurst, "Hypothesis tests for limited failure survival distributions", *Evaluation Review*, **5**, (1981) 699-711.

[32] L.Q. Sun, and X. Zhou, "Survival function and density estimation for dependent data", *Statistics & Probability Letters*, **52**, (2001) 47-57.

[33] A.D. Tsodikov, Ibrahim, J.G., and Yakovlev, Y., "Estimating cure rates from survival data, an alternative to two-component mixture models", *Journal of the American Statistical Association*, **98**, (2004) 1063-1078.

[34] H.T.V. Vu, R.A. Maller, and X. Zhou, "Asymptotic properties of a class of mixture models for failure data, the interior and boundary cases", *Annals of Institute of Statistical Mathematics*, **50**, (1998) 627-653.

[35] P. Yip, H. Eric, et al., "A comparison study of real-time case fatality rates: severe acute respiratory syndrome in Hong Kong, Singapore, Toronto and Beijing, China". *Journal of the Royal Statistical Society, A*, **168** (2005a) 233-243.

[36] P. Yip, H. Eric, et al., "A chain multinomial model for estimating the real-time case fatality rate of a disease, with an application to severe acute respiratory syndrome". *American Journal of Epidemiology*, **161** (2005b) 700-706.

[37] S. Zhou, and R.A. Maller, "Likelihood ratio test for the presence of immunes in a censored sample", *Statistics*, **27**, (1995) 181-201.

[38] G, Zhou, G. Yan, "Severe acute respiratory syndrome epidemic in Asia". *Emerging Infectious Diseases*, **9**, (2003) 1608-1610.

[39] R. Wallace, D.N. Blischke, and P. Murthy, *Reliability* Wiley, New York (2000).

[40] http://en.wikipedia.org/wiki/SARS

[41] WHO, http://www.who.int/csr/sars/country/en/ (2003)

[42] WHO, http://www.who.int/csr/sars/postoutbreak/en/ (2003)

[43] WHO, http://www.who.int/features/2003/07/en/ (2003)

[44] W.K. Wong and G. Bian, "Estimating parameters in autoregressive models with asymmetric innovations", *Statistics & Probability Letters*, **71**, (2005) 61-70.

Appendix

WHO (2003) reports SARS outbreak as follows (see [42, 43]):

First recognized as a global threat in mid-March 2003, SARS was successfully contained in less than four months. On 5 July 2003, WHO reported that the last human chain

of transmission of SARS had been broken. While much has been learned about this syndrome since March 2003, including its causation by a new coronavirus (SARS-CoV), our knowledge about the epidemiology and ecology of SARS coronavirus infection and of this disease remains limited. Resurgence of SARS remains a distinct possibility and does not allow for complacency.

The earliest cases are now known to have occurred in mid-November in Guangdong Province, China. SARS was first carried out into the world at large on 21 February, 2003, when an infected medical doctor from Guangdong checked into room 911 on the 9th floor of the Metropole Hotel in Hong Kong. That single hotel floor became the setting for the international spread of SARS. At least 14 guests and visitors carried the virus with them to the hospital systems of Toronto, Hong Kong, Viet Nam, and Singapore. The earliest and most severe outbreaks in Toronto, Hong Kong, Viet Nam, and Singapore were all seeded by visitors to the hotel. At that time, prior to the first global alert issued by WHO on 12 March 2003, no one was aware that a severe new disease, capable of rapidly spreading in hospitals, had emerged. Hospital staff responding to the earliest cases failed to protect themselves from infection as they aggressively fought to save lives. As a result, the disease rapidly spread within hospitals, infecting staff, other patients, and visitors, and then spilled out into the larger community as family members and their close contacts became infected. As the outbreaks grew in size, the number of exported cases rose, with 30 countries and areas eventually reporting cases.

Table 1. Cumulative number of probable cases.
((a) from March 17 2003 to May 24 2003)

date	patients	deaths	recoveries	date	patients	deaths	recoveries
3.17	95	1	–	4.21	1,402	94	436
3.18	123	1	–	4.22	1,434	99	461
3.19	150	5	–	4.23	1,458	105	522
3.20	173	6	–	4.24	1,488	109	567
3.21	203	6	–	4.25	1,510	115	614
3.22	222	7	–	4.26	1,527	121	632
3.24	260	10	–	4.28	1,557	138	710
3.25	286	10	–	4.29	1,572	150	759
3.26	316	10	–	4.30	1,589	157	791
3.27	367	10	–	5.1	1,600	162	834
3.28	425	10	–	5.2	1,611	170	878
3.29	470	10	–	5.3	1,621	179	898
3.31	530	13	–	5.5	1,637	187	930
4.1	685	16	–	5.6	1,646	193	958
4.2	708	16	–	5.7	1,654	204	984
4.3	734	17	–	5.8	1,661	208	1,008
4.4	761	17	–	5.9	1,667	210	1,015
4.5	800	20	–	5.10	1,674	212	1,035
4.7	883	23	–	5.12	1,683	218	1,066
4.8	928	25	–	5.13	1,689	225	1,090
4.9	970	27	–	5.14	1,698	227	1,128
4.10	998	30	154	5.15	1,703	234	1,160
4.11	1,059	32	169	5.16	1,706	238	1,171
4.12	1,108	35	215	5.17	1,710	243	1,191
4.14	1,190	47	229	5.19	1,714	251	1,213
4.15	1,232	56	243	5.20	1,718	253	1,229
4.16	1,268	61	257	5.21	1,719	255	1,237
4.17	1,297	65	272	5.22	1,722	258	1,247
4.18	1,327	69	322	5.23	1,724	260	1,255
4.19	1,358	81	363	5.24	1,724	262	1,266

Table 1. Cumulative number of probable cases.
((b) from May 26 2003 to July 11 2003)

date	patients	deaths	recoveries	date	patients	deaths	recoveries
5.26	1,726	267	1,276	7.4	1,755	298	1,430
5.27	1,728	269	1,285	7.7	1,755	298	1,430
5.28	1,730	270	1,295	7.8	1,755	298	1,430
5.29	1,732	273	1,302	7.9	1,755	298	1,431
5.30	1,736	274	1,304	7.10	1,755	298	1,431
5.31	1,739	278	1,310	7.11	1,755	298	1,433
6.2	1,746	282	1,319				
6.3	1,747	283	1,326				
6.4	1,748	283	1,339				
6.5	1,748	284	1,343				
6.6	1,750	286	1,350				
6.9	1,753	288	1,365				
6.10	1,754	290	1,368				
6.11	1,754	290	1,368				
6.12	1,755	291	1,377				
6.13	1,755	293	1,380				
6.16	1,755	295	1,386				
6.17	1,755	295	1,387				
6.18	1,755	295	1,393				
6.19	1,755	296	1,396				
6.20	1,755	296	1,403				
6.23	1,755	296	1,411				
6.24	1,755	296	1,417				
6.25	1,755	296	1,419				
6.26	1,755	296	1,419				
6.27	1,755	296	1,422				
6.30	1,755	298	1,429				
7.1	1,755	298	1,429				
7.2	1,755	298	1,429				
7.3	1,755	298	1,429				

Table 2. Log-likelihood values in the four probability distribution models. Based on data as of the June 11, 2003, and using the truncated model.

	logistic		log-normal		gamma		Weibull
infected	−6816.40	>	−6817.65	>	−6819.00	>	−6827.28
fatal	−1228.72	>	−1230.46	>	−1230.54	>	−1233.55
cured	−5469.89	>	−5472.55	>	−5475.31	>	−5482.64
total	−13515.0	>	−13520.7	>	−13524.8	>	−13543.5

Table 3. Approximate case fatality ratios and their standard deviations. Based on data as of the December 31, 2003.

Country	cases	deaths	case fatality ratio (%)	standard deviation (%)	
				$\rho = 0$	$\rho = 1$
Canada	251	43	17.13	2.83	1.53
China	5,327	349	6.55	0.36	0.26
Hong Kong	1,755	299	17.04	1.07	0.58
Taiwan	346	37	10.69	1.85	1.18
Singapore	238	33	13.87	2.58	1.51
Viet Nam	63	5	7.94	3.69	2.55
world-wide	8,096	774	9.56	0.36	0.24

According to [41], 325 cases have been discarded in Taiwan since 11 July 2003 because Laboratory information was insufficient or incomplete for 135 discarded cases, of which 101 died.

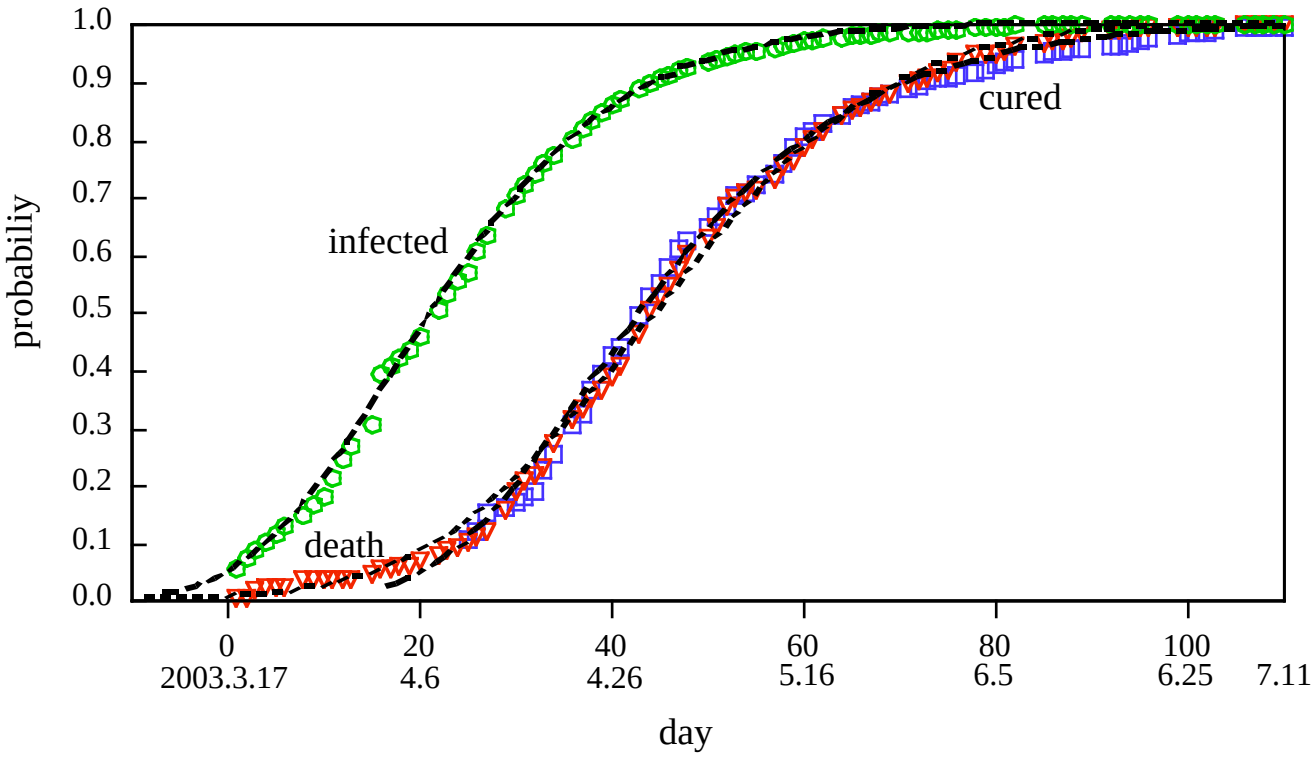

Figure 1. Empirical probability distributions for the patients, deaths, and recoveries, along with the corresponding estimated probability distributions.
circles: infected empirical,
triangles: fatal empirical,
squares: cured empirical.
dashed lines: estimated probability distributions.

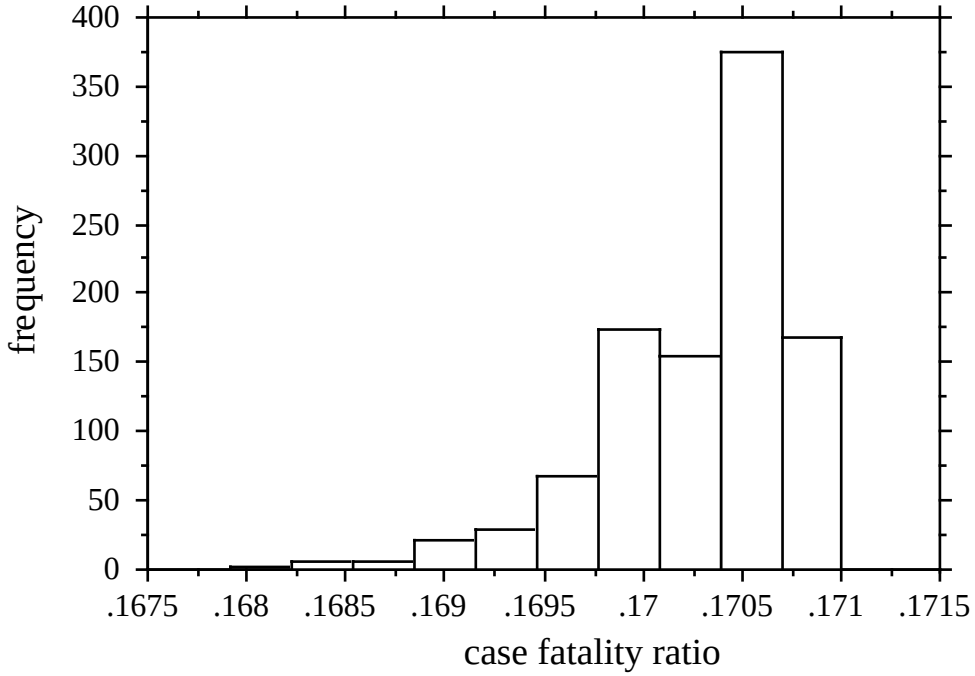

Figure 2. Bootstrapped estimates of the case fatality ratio in the truncated model.
The censoring time is set on July 11, 2003.

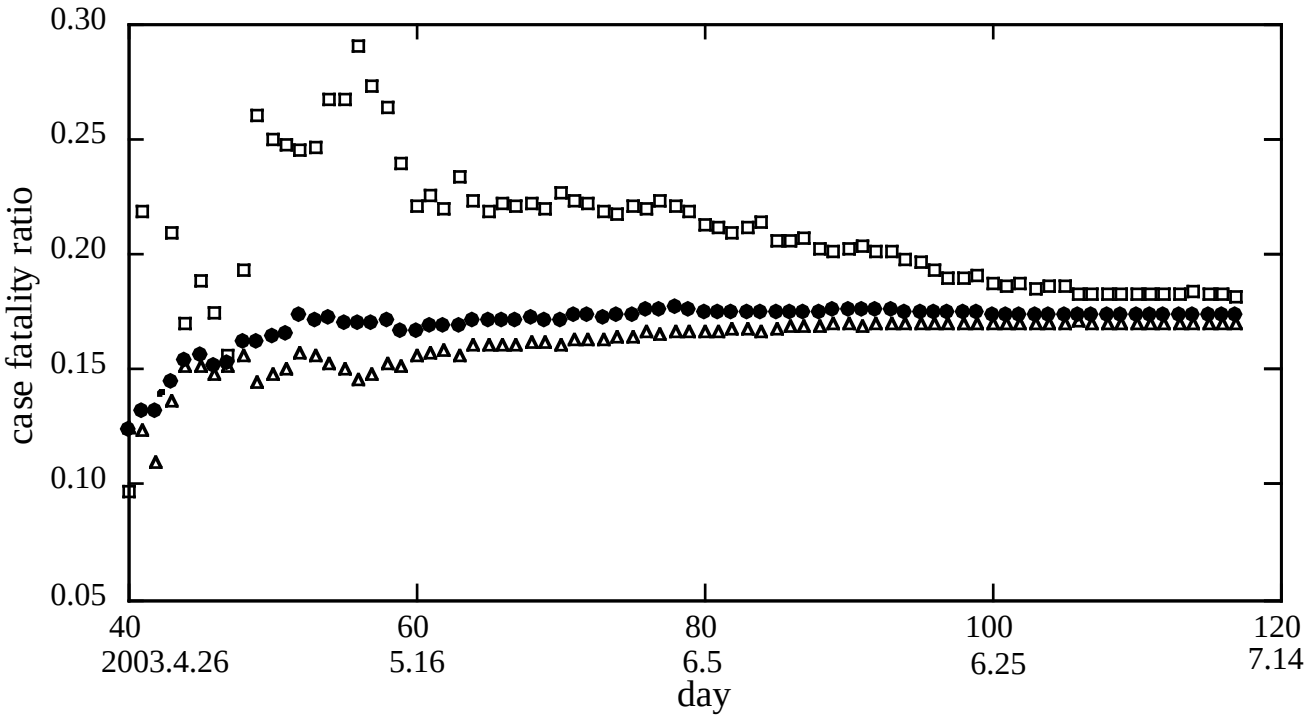

Figure 3. Estimated case fatality ratios.
filled circles: mixed trunsored model using patients, deaths, and recoveries,
triangles: truncated model using patients and deaths,
squares: truncated model using patients and recoveries.

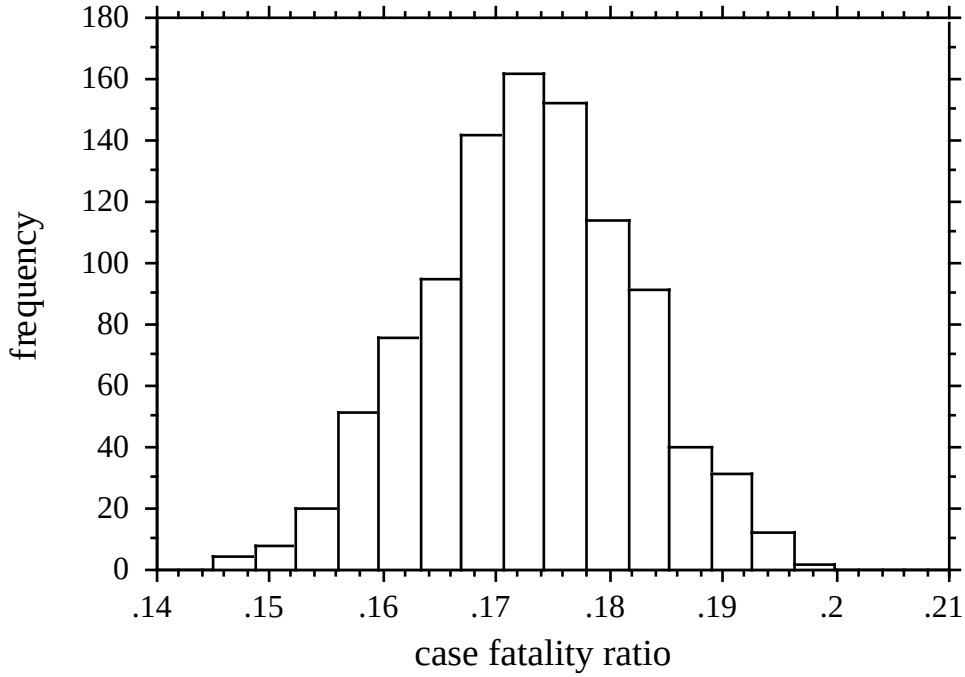

Figure 4. Bootstrapped estimates of the case fatality ratio in the mixed trunsored model.
The censoring time is set on July 11, 2003.

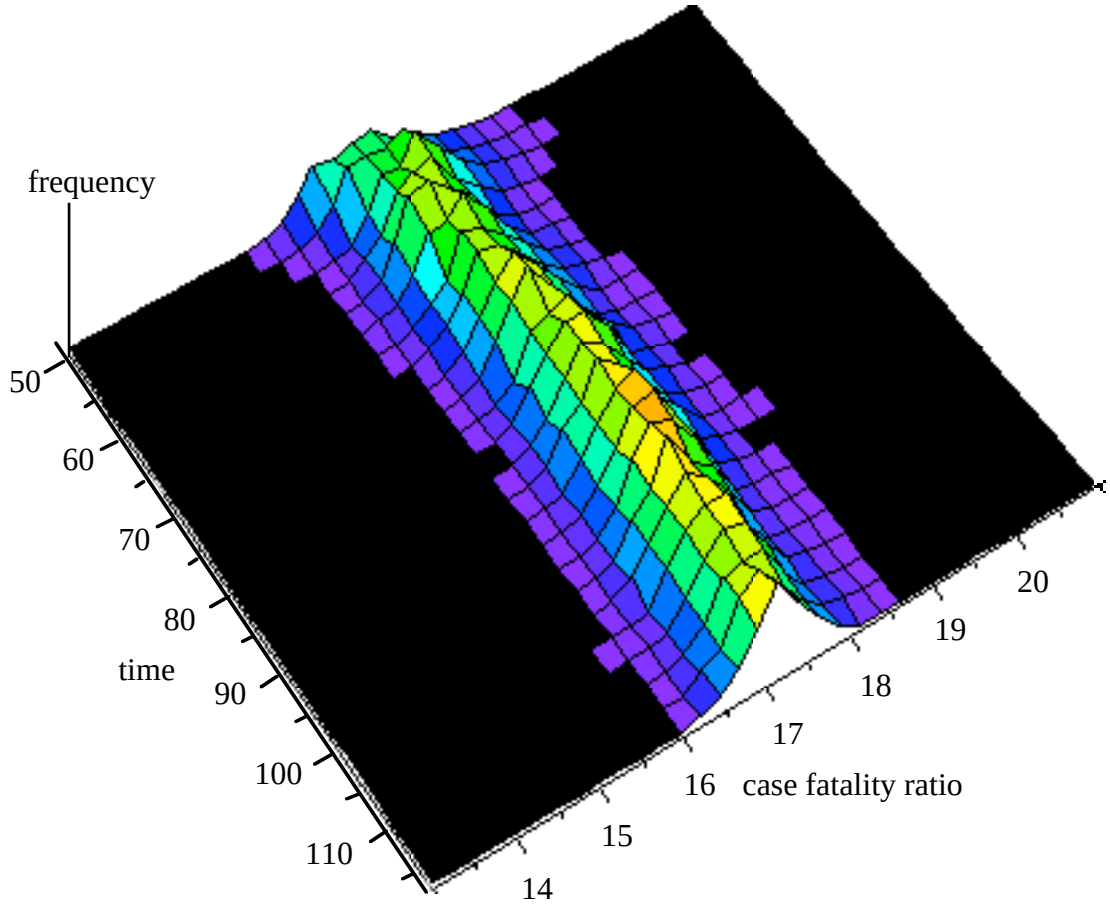

Figure 5 Bootstrapped frequency for the case fatality ratio in the mixed trunsored model.

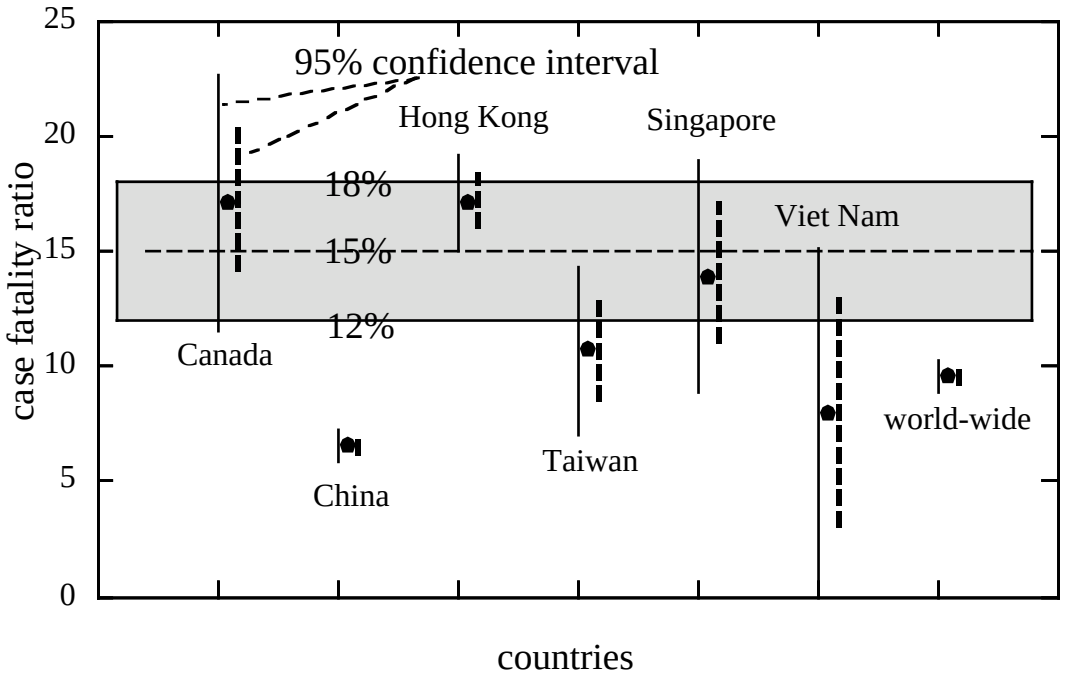

Figure 6. Estimated case fatality ratios and their approximate 95% confidence intervals

Solid line: when correlation coefficiet between numbers of patients and deaths = 0
Dashed line: when correlation coefficiet between numbers of patients and deaths = 1
A band [12,18]% includes points in 95% confidence intervals in Canada, Hong Kong, Taiwan, Singapore, and Viet Nam.