Statistics and Applications ISSN 2452-7395 (online)
Volume 18, No. 1, 2020 (New Series), pp 269–280

A Comprehensive Modeling Framework for Estimation and Prediction of COVID19 in India

P. Venkatesan
Former Scientist-F, Department of Statistics, ICMR National Institute for Research in Tuberculosis, Chennai-600 031

Received: 15 June 2020; Revised: 01 July 2020; Accepted: 03 July 2020

Abstract

There has been substantial interest worldwide in understanding the current status of Corona Virus Disease (COVID-19) epidemic and prediction of the future path through the pandemic. Many groups are attempting to provide the description of spread and modeling the transmission processes for short and long term projections. Since the epidemic is in its early stage, there is paucity of data for predicting the future course of the disease. The back-calculation approach is one of the methods used in such a situation. The back-calculation reconstructs the past pattern of the infection and predicts the future number of cases with the present infection curve. Lack of information about incubation distribution, effect of intervention on incubation period and errors in reporting the cases lead to uncertainties associated with modeling. This paper attempts to formulate the problem of estimating future COVID-19 cases as estimation of parameters in a multinomial likelihood with unknown sample size by EM algorithm. Illustrations are provided using reported cases in India and discussed.

Key words: COVID-19; Back-calculation; EM algorithm; Incubation period; Infection density.

1. Introduction

The extensive world-wide spread of COVID-19, which started in late 2019 in China, has become the first modern pandemic in less than six months (Korean Society for Infectious Diseases, 2020; Li *et al*., 2020; Liu *et al*., 2020, Sun and Vibound 2020). Given the need to develop a better understanding of the levels and trends in the epidemic and the limited information on which to base these estimates, the use of modeling approaches can make a valuable contribution has seen in other epidemics (*e.g.* Solomon *et al.*, 2003, Ravanan and Venkatesan 2008). The goal of any modeling exercise is to extract as much information as possible from the available data in order to provide an accurate representation of both the knowledge and uncertainty about the epidemic.

A range of different types of models have been developed and applied to the estimation of epidemics in variety of settings. (One major tradition in modeling infectious diseases like HIV and COVID-19 epidemic is the use of back-calculation of back projection techniques which provides statistical solutions convolution equations relating the number of cases diagnosed over time and incubation period distributions (Anderson, 1988; Venkatesan, 2006; Liu *et al*., 2020; Nishiura, 2020) The objective of this paper is to highlight areas in which further methodological developments are needed given currently available data sources. In general, epidemic modeling is categorized in to four broad categories, but not mutually exclusive ones.

Corresponding Author: P. Venkatesan
E-mail: venkaticmr@gmail.com

(a) *Deterministic models*: In this type of modeling the parameters such as number of susceptible, infected and disease individuals are assumed to deterministic. These models are described by a system of differential or integral equations. The progression of the epidemic is studied using these equations. Many such models were developed in the past for diseases such as HIV/AIDS (Anderson *et al*., 1989; Hyman and Stanley, 1988; Anderson and May, 1992).

(b) *Stochastic models*: Stochastic models assume that some of the key parameters are random variables. It is assumed that is a continuous time stochastic process. The stochastic models are considered to be more realistic than deterministic models and with some special assumptions the results of the deterministic models can be approximated through stochastic models. Several studies showed that stochastic models gave a better interpretation of epidemics than deterministic models (Tan, 2000; Mode *et al*., 1988; Isham, 1991).

(c) *Statistical models*: The statistical models are based on epidemiology of the disease and survey/surveillance data. These models make full use of the available data compared to deterministic and stochastic models. In this type of modeling the disease mechanism and prior information are usually not considered. The back-calculation approach for projection of disease epidemics can be categorized in to this type of modeling (Jewel *et al.*, 1992; Bacchetti *et al*., 1993; Venkatesan, 2006; Ravanan and Venkatesan, 2008; Egan and Hall, 2015).

(d) *State-Space models***:** The state-space models have been introduced by Wu and Tan (1995) for modeling HIV/AIDS epidemic, which takes advantages of both stochastic and statistical models. The state-space models were originally proposed for engineering control and communication. This model was also used for projections and detailed description is given by Tan (2000).

2. Back-Calculation Methodology

Brookmeyer and Gail (1986, 1988) introduced back calculation method for short-term projection of HIV/AIDS epidemic. This method uses a form of infection curve, either parametric or non-parametric, for the number of past HIV infections or equivalently a density function for infections as noted by Ding (1995, 1996). The time between infection and the diagnosis of disease is known as incubation time and it is modelled by a known distribution. Many distributions are used for the incubation curve depending on the length (Lawless and Sun, 1992; Ravanan and Venkatesan, 2008; Venkatesan *et al*., 2012). The next section presents some of the useful distributions for modeling the COVID-19 infection curve. The formulation of back-calculation for discrete and continuous cases are considered here.

2.1. Discrete back-calculation formulation

The number of reported COVID-19 cases is available during the calendar time T_0 to T_L. Here T_0 denotes the start of the epidemic and T_L denotes the time up to which the data on reported COVID-19 cases are available. The back-calculation method to reconstruct the COVID-19 infection and projection of future COVID-19 cases can be described in the following sections.

Let X_j denotes the number of COVID-19 cases reported in the interval $[T_{j-1}, T_j), j = 1,2,..., L$. Let $N = \sum_{i=1}^{L} X_i$, then $(X_1, X_2, ... X_L)$ can be assumed to follow multinomial distribution $(N, p_1, p_2, ..., p_L)$. Here p_j denotes the probability that a person infected at time T_0 is diagnosed with COVID-19 in the j^{th} interval. This probability p_j may be estimated using the equation

$$p_j = \frac{1}{N} \sum_{i=1}^{j} I_j [F(j+1-i) - F(j-i)] \tag{1}$$

where $F(t)$ denotes the discrete time formulation of incubation period distribution and I_j denotes the number of infected individuals at the beginning of the j^{th} interval.

Let $$f_{j-I} = F(j+1-i) - F(j-i) \tag{2}$$

then equation (1) can be modified as

$$p_j = \frac{1}{N} \sum_{i=1}^{j} I_j f_{j-i} \tag{3}$$

If p_j's values are known then Np_j denotes the expected number of COVID-19 cases in the interval $[T_{j-1}, T_j)$. Estimation of p_j is done by using various approaches. Let us consider the multinomial likelihood method of estimation of p_j, where a form of infection curve is given by $I_j = I_j(\theta_1, \theta_2, \ldots, \theta_k)$ which is assumed to be known except the k parameters. Therefore, p_j is a function of k unknown parameters, assuming $F(t)$ is completely specified. The unknown parameters p_j can be obtained using the multinomial likelihood as

$$L = \frac{N!}{x_1! x_2! ... x_L!} p_1^{x1} p_2^{x2} ... p_L^{xL} \tag{4}$$

Then $$\log(L) = N! - \sum_{i=1}^{L} \log x_i! + \sum_{i=1}^{L} x_i \log p_i \tag{5}$$

Fisher's scoring algorithm can be used to estimate the unknown parameters θ_i's and hence p_j can be estimated. The above formulation has been used by Taylor (1989). Future COVID-19 cases in the k^{th} time point following T_L can be obtained using the equation

$$\hat{X}_{T_{L+k}} = \sum_{j=k+1}^{T_L} I_j [F_{T_{L+1+k-j}} - F_{T_{L+k-j}}]$$

$$= \sum_{j=k+1}^{T_L} I_j f_{T_{L+k-j}} \tag{6}$$

where $\hat{X}_{T_{L+K}}$ is the minimum number of COVID-19 cases in the interval T_{L+K}.

2.2. Continuous time formulation of back-calculation

In the discrete time formulation the incubation time was treated as a discrete random variable. If the incubation time is treated as a continuous random variable, then the probability of infection in the j^{th} interval given in equation (1) can be rewritten as

$$p_j = \frac{1}{N} \int_{T_0}^{T_j} I(\theta,t)[F(T_j - t) - F(T_{j-1} - t)]\ dt \tag{7}$$

Now $I(\theta,t)$ is assumed to be a smooth function of t. Brookmeyer and Gail (1986) modified (7) by assuming $D(\theta,t)$ to be the density function of infection times of N individuals. Therefore $\int_{T_0}^{T_L} D(\theta,t)\ dt = 1$ and the equation (7) can be written as

$$p_j = \int_{T_0}^{T_j} D(\theta,t)[[F(T_j - t) - F(T_{j-1} - t)]\ dt \tag{8}$$

Thus a model for infection curve is $I(\theta,t)$ and a model for infection density are related by $I(\theta,t) = ND(\theta,t)$ where N = $\int_{T_0}^{T_L} I(\theta,t)dt$.

Hence we now work in the formulation of p_j as given in equation (8). The parameter in p_j can be estimated using the Fisher's scoring algorithm assuming a multinomial likelihood. Brookmeyer and Gail (1988) formulated the problem of estimation of future cases in short interval of time as the problem of estimation of parameters in multinomial likelihood with unknown sample size and the method as explained in in the next section.

2.3. EM algorithm approach

The Expectation-Maximization (EM) algorithm was first proposed by Dempster *et al*. (1977) for the analysis of incomplete data. The algorithm is formulated as follows:

X_{L+1} denote the number of individuals infected before the time T_L who have not become COVID-19 cases by time T_L. The problem is to estimate the total number of infections before the time T_L. This number N is the minimum size of the COVID-19 epidemic, because even if the infections after the time T_L could be prevented, the cumulative number of COVID-19 cases would eventually reach N. The minimum size is the sum of all cases already diagnosed, called $n = \sum_{i=1}^{L} X_i$ and all the susceptible individuals infected before T_L but not yet diagnosed, called $X_{L+1}=N-n$. It can be noted that, in this formulation both N and X_{L+1} are unknown. Therefore an estimate of the minimum cumulative incidence of COVID-19 that can be anticipated in some future time point T_{L+1} is

$$n + \hat{N}\hat{P}_{L+1} = n + \hat{N}\int_{T_0}^{T_L} I(\theta,t)[F(T_{L+1}-t) - F(T_L - t)]\,dt \tag{9}$$

where $\hat{N}$ is the estimate of N and $\hat{P}_{L+1}$ is the probability of becoming COVID -19 in the future interval $[T_L, T_{L+1})$.

Assuming $I(\theta,t)$ as step function, Brookmeyer and Gail (1988) gave the following EM algorithm for the estimation of the parameter.

Suppose $I(\theta,t) = \theta_i$ for t in $[T_{r_{i-1}}, T_{r_i})$ $i = 1,2,...L.\, T_{r_i}$ denote the time point defining the i^{th} step. Let X_{ij} denote the number of COVID-19 cases who were infected in the i^{th} step $[T_{r_{i-1}}, T_{r_i})$ and diagnosed in the j^{th} interval $[T_{r_{i-1}}, T_{r_i})$. Note that $T_{r_{i-1}} \le T_j$, since X_{ij} is not defined if $T_{r_{i-1}} > T_j$.

For a fixed N,

$$\hat{X}_{ij} = X_j(\hat{P}_{ij}/\hat{P}_j) \tag{10}$$

where $\hat{P}_{ij}$ is the estimated probability that individual infected in the i^{th} step is diagnosed as COVID-19 in the j^{th} interval. These estimates are obtained using the current estimate of the θ values at the m^{th} iteration *i.e.*, $\hat{\theta}^{(m)}$, then the updated estimates are obtained using the equation

$$\hat{\theta}_i^{(m)} = \sum_{j=1}^{L+1} \hat{X}_{ij} / N\delta_i \tag{11}$$

where δ_i is the width of the i^{th} step. The numeration in equation (11) is an estimate of the number of individuals N infected during the i^{th} step. Further detail of the algorithm for a step function $I(\theta,t)$ is given by Brookmeyer and Gail (1988).

3. Statistical Models for Incubation Period

The incubation period models are similar to survival models based on non-negative random variables and can be fitted using either parametric or semi-parametric approach. A detailed description can be found in Lawless (2011). Here we restrict our attention to only parametric models for incubation period as described in our earlier work (Ravanan and Venkatesan, 2008). The common distributions used for the incubation distribution are given in Table 1 and the infection densities used for prevalence are given in Table 2. Two other important distributions used are the staging model mode and change point model which are described below:

3.1. Staging model

Under staging models the incubation period is considered to be comprised of stages (Brookmeyer and Liao 1990). Different models for these two stages can be assumed. Let $h_1(t)$ and $h_2(t)$ denote the hazard functions of the two stages. The convolution equation for the incubation period comprising of these two stages

$$F(t) = \int_0^t f_1(u)\, F_2(t-u)du \tag{12}$$

where $$f_1(u) = h_1(u)\exp\{-\int_0^u h_1(s)ds\} \tag{13}$$

and $$F_2(u) = 1-\exp\{-\int_0^u h_2(s)ds\} \tag{14}$$

Suitable changes should be made in the above formulations to account for calendar time of infection

Table 1: Incubation period distributions

Model	Distribution Function
Weibull	$F(t) = 1-e^{-(\lambda t)^{\alpha}}$ $\quad \lambda>0, \alpha>0,\ t>0$
Gamma	$F(t) = \frac{1}{\sigma\Gamma(k)}\int_0^t \left(\frac{x}{\sigma}\right)^{k-1} \exp(-\frac{x}{\sigma})\ dx \quad t>0,\ \sigma>0,\ k>0$
Lognormal	$F(t) = \Phi\frac{(\log t-\mu)}{\sigma},\ \Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx$ $t>0,\ -\infty<\mu<\infty,\ \sigma>0$
Log-logistic	$F(t) = 1-[1+(\lambda t)^{\upsilon}]^{-1} \quad \lambda>0,\ \upsilon>0,\ t>0$
Gen. Exponential	$F(t) = 1-[1-\exp\{-t\lambda\}]^{\upsilon} \quad t>0,\ \lambda>0,\ \upsilon>0$
Gen. Log-logistic	$G(t) = \frac{1}{\beta(m_1,m_2)}\int_0^{H(t)} x^{m_1-1}(1-x)^{m_2-1}dx;\ t>0, m_1>0, m_2>0$ $H(t) = \frac{1}{1+\exp\{-[\log(t)-\mu]/\tau\}} \quad t>0,\ -\infty<\mu<\infty,\ \tau>0$
Gen. Gamma	$F(t) = \int_0^t\left[\sum_{j=1}^{k} B_{1k}(j)\exp(-\lambda_j x)\right]dx$ $B_{1k}(j) = (\prod_{i=1}^{k}\lambda_i)/\prod_{\substack{i=1\\ i\neq j}}^{k}(\lambda_i-\lambda_j),\quad \lambda_j = 1/\sigma_j$
Mixed Weibull	$F(t) = pF_1(t)+(1-p)F_2(t) \quad 0<p<1$ $F_1(t) = 1-e^{-(\lambda_1 t)^{\alpha 1}} \quad \lambda_1>0, \alpha_1>0,\ t>0$ $F_2(t) = 1-e^{-(\lambda_2 t)^{\alpha 2}} \quad \lambda_2>0, \alpha_2>0,\ t>0$

3.2. Change point model

Estimates of the population parameters are obtained in the case of subpopulations are exponentially distributed and sampling is censored at a predetermined test termination time was first introduced by Mendenhall and Hader (1958). The change point model considered here is briefly presented below.

Suppose the hazard before and after the change point is constant, then $h(t)$ is given by

$$h(t) = \begin{cases} \alpha & t \le \tau \\ \beta & t > \tau \end{cases} \tag{15}$$

The distribution function is given by

$$F(t) = \begin{cases} 1 - e^{-\alpha t} & t \le \tau \\ 1 - e^{-\alpha\tau}\, e^{-\beta(t-\tau)} & t > \tau \end{cases} \tag{16}$$

The median of the incubation period is given by

$$M = \begin{cases} \log 2 / \alpha & M \le \tau \\ (2e^{-\tau\alpha} - 1)/(e^{\tau(\beta-\alpha)}) & M > \tau \end{cases} \tag{17}$$

Table 2: Infection densities

Model	Infection Density
Logistic Prevalence	$D_1(\theta,t) = \frac{1}{k_1} I_1(\theta,t) = \frac{1}{k_1} \frac{\theta_1\theta_3 e^{\theta_2+\theta_3 t}}{(1+e^{\theta_2+\theta_3 t})^2}$
Logistic Incidence	$D_2(\theta,t) = \frac{1}{k_2} I_2(\theta,t) = \frac{1}{k_2} \frac{\theta_1 e^{\theta_2+\theta_3 t}}{1+e^{\theta_2+\theta_3 t}}$
Double Exponential incidence	$D_3(\theta,t) = \frac{1}{k_3} I_3(\theta,t) = \frac{1}{k_3} \theta_1 e^{-e^{\theta_2+\theta_3 t}}$
Log-logistic incidence	$D_4(\theta,t) = \frac{1}{k_4} I_4(\theta,t) = \frac{1}{k_4} \theta_1\theta_2(\theta_1 t)^{\theta_2 - 1}/[1+(\theta_1 t)^{\theta_2}$
Exponential incidence	$D_5(\theta,t) = \frac{1}{k_5} I_5(\theta,t) = \frac{1}{k_5} \theta_1 e^{\theta_2 t}$
Root exponential incidence	$D_6(\theta,t) = \frac{1}{k_6} I_6(\theta,t) = \frac{1}{k_6} \theta_1 e^{\theta_2 t^{1/4}}$ $k_i = \int_{T_0}^{T_L} I_i(\theta,t)\, dt \quad ; \quad i = 1,2,\ldots,6.$

4. An Illustration Using Indian Data

The basic data required for back-calculation methodology is the number of COVID-19 cases over a period of time (Brookmeyer and Gail, 1986, 1988, 1990; Ding, 1995, 1996). The Ministry of Health publishes daily updates of the reported COVID-19 cases for the past few

months. The updates of the recent days also suffer reporting delays and under reporting and therefore pooled weekly reported cases may be more reliable. In this illustration only weekly reported cases were considered. The period is from 1st March 2020 to 30th May 2020 (13 weeks) (https://www.coronatracker.com/country/india/). It is also reported that level of under reporting may vary from 50-90%. For this work, it is assumed that the level of under reporting is around 90% in early March 2020 and gradually decreased to 50% exponentially in the end of May 2020. The exponential decay model

$$P(t) = 0.\,90\, e^{-0.05t} \tag{18}$$

gives a better approximation of the above assumption. The upward adjustments for the weeks are carried out. Table 3 gives the actual number of reported and adjusted COVID-19 cases along with 3-week moving averages. The reported cases are smoothed using a three week moving averages as a first step. The linear and quadratic models are fitted to find the best fit liner model for the trend for the moving average cases which serves as a bench mark for comparisons. The results are given in Table 4. From the table it is seen that the quadratic trend model seems to be a better fit for moving averages and the corresponding model is

$$\text{Cases} = 1114.9 - 1205.7\ \mathit{Time} + 373.2\ \mathit{Time}^2 \tag{19}$$

Table 3: Weekly confirmed COVID-19 cases in India

Week	Weekly Confirmed COVI-19 Cases	3week Moving Average	Cumulative COVID-19 Cases	Adjusted weekly Cases
March 1-7	31	-	31	63
March 8-14	50	104	81	93
March 15-21	231	294	312	419
March 22-29	601	995	913	1067
March29-April 4	2154	2407	3067	3741
April 5-11	4467	4628	7534	7581
April 12-18	7263	7460	14797	12105
April 19-25	10650	10319	25447	17404
April 26-May 2	13044	15193	38491	20913
May 3-9	21886	20924	60377	34446
May 10-16	27842	30168	88219	43040
May17-23	40775	39797	128994	61948
May24-30	50404	-	179398	75300

Table 4: Trend lines based on the moving averages

Trend	Variable	B	Se(B)	Z	Sig	R^2
Linear	Constant	–8742.3	2500.8	–3.496	0.01	
	Time	3333.0	368.7	9.039	8.24e-05	0.889
Quadratic	Constant	879.0	1314.0	1.505	0.163	
	Time	–1107.6	519.2	–4.655	0.0009	0.984
	$Time^2$	370.1	49.9	7.413	7.54e-05	

4.1. Estimation of parameters

Based on the availability of data, the starting point of the epidemic T_0 is taken as March 2020. The incubation distributions discussed in the previous section are used in this section to illustrate projection of COVID-19 in India. The estimates of minimum size of the epidemic and future COVID-19 cases are obtained assuming a median incubation period of two weeks. For the incubation period models Weibull, gamma, log-logistic, log-normal and generalized exponential distribution prior estimates of their parameters are obtained methods described in Venkatesan (2006). All these models have two parameters and one parameter is fixed based on the estimates reported (Table 1). The other parameter was determined such that the median distribution period is known. The parameters of the generalized log-logistic, generalized gamma, mixed Weibull and change point models are not available. The parameters of these models are decided based on the simulation study as described in our earlier work (Ravanan and Venkatesan, 2009). For the infection density, the exponential, root exponential, double exponential, logistic and log-logistic are commonly used (Table 2). In this work only logistic density incidence based projections are given for illustrative purpose. The projections based on logistic infection density under various incubation period distributions are presented in Table 5.

Table 5: Projection of COVID-19 prevalence under logistic infection density and total expected confirmed cases (Median incubation = 2 weeks)

Incubation period Model	***Projection of COVID-19 cases ('000)***		
	Up to June 6	**Up to June 13**	**Up to June 20**
Weibull	240.5	318.8	408.1
Gamma	242.8	321.9	411.8
Log-logistic	248.5	327.7	419.2
Log-normal	244.8	327.0	419.7
Gen. Exponential	249.6	328.7	420.9
Gen. Log-logistic	248.0	326.9	418.4
Gen. Gamma	249.3	327.1	417.8
Mixed Weibull	241.5	322.1	414.4
Change point	247.7	326.1	418.1
Quadratic	246.9	324.8	413.8
Observed	236.2	321.6	411.8

From Table 5 we see that the projections obtained for the next three weeks under different models do not differ widely. This may be due to the behaviour of the epidemic in the early stage. However, the projections based on Weibull, gamma, lognormal and mixed Weibull are close the observed cases. The quadratic model also gives results close to the observed cases. The projections based on exponentially adjusted cases for under reporting resulted higher cases and are not reported here. We also considered the other infection densities for incidence given

in Table 2. They resulted in higher cases than the logistic prevalence. Hence only the results pertaining to the logistic prevalence infection density are given to illustrate the use of the models. After obtaining sufficient data in the infection curve, the comparisons will provide valid estimates.

5. Discussion

There has been research showing that on average, each infected person spreads the infection to more than two persons. Therefore the majority of the population is at risk of infection if no intervention measures were undertaken. The true size of the COVID- 19 epidemic remains unknown, as a significant proportion of infected individuals only exhibit mild symptoms or are even asymptomatic. Timely assessment of the evolving epidemic size is crucial for resource allocation and plan strategies. In this article, we used the back-calculation algorithm to obtain a lower bound estimate of the numbers of COVID-19 infected confirmed cases in India using the available data. Since the data source is limited and suffers from under reporting, under diagnosis and delay in reporting, adjustments are needed before making any modeling and projections.

One of the critical issues in infectious disease epidemiology is that the time of infection event is seldom directly observable. For this reason, the time of infection needs to be statistically estimated, employing a back-calculation method. It is observed that the short-term projection of three weeks do not vary much across various incubation period distributions. Further the estimates vary widely for different infection densities. The projected COVID-19 cases for three weeks under Weibull, gamma, lognormal and mixed Weibull are similar and close to the confirmed cases. One reason could be that they are related models and the do not differ in the initial stages. We also considered projections under the logistic, exponential double exponential and root exponential infection densities with varying median incubation periods. But the estimates vary significantly particularly under exponential infection density. Since the infection curve is at its early sage, we have reported only the logistic infection density estimates as an illustration. Once sufficient size data is available, the comparisons are reliable. This paper provides a methodology based on the back-calculation for short-term projections which ae widely used in diseases like HIV/AIDS.

References

Anderson, R. M. (1988). The role of mathematical models in the study of HIV transmission and the epidemiology of AIDS. *AIDS*, **1**, 241-246.

Anderson, R. M., Blythe, S. P., Gupta, S. and Konings, E. (1989). The transmission dynamics of the human immunodeficiency virus type I in the male homosexual community in the United Kingdom: The influence of changes in sexual behaviour. *Philosophical Transactions of Royal Soc*iety, **B325**, 45-98.

Anderson, R. M. and May, R. M. (1992). Understanding the AIDS epidemic. *Scientific Amer*ican, **266**, 58-66.

Bacchetti, P., Segal, M. and Jewell, N. P. (1993). Back-calculation of HIV infection rates, *Statistical Science*, **8**, 82-119.

Brookmeyer, R. and Gail, M. H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *Lancet*, **2**, 1320-1322.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, **83**, 301-308.

Brookmeyer, R. and Liao, J. (1990). Statistical modeling of the AIDS spread for forecasting health care need. *Biometrics*, **46**, 1151-1163.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* **B39,** 1-3.

Ding, Y. (1995). Computing back-calculation estimates of AIDS epidemic. *Statistics in Medicine*, **14**, 1505-1512.

Ding, Y. (1996). On the asymptotic normality of multinomial population size estimates with application to the back-calculation epidemic of AIDS. *Biometrika*, **83**, 695-699.

Egan, J. R. and Hall, I. M. (2015). A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. *Journal of the Royal Society Interface*, **12**, 20150096, 1-14.

https://www.coronatracker.com/country/india/ (Accessed as on 30/05/2020)

Hyman, J. M. and Stanley, E. A. (1988). Using mathematical models to understand the AIDS epidemic. *Mathematical Biosciences*, **90**, 415-474.

Isham, V. (1991). Assessing the variability of stochastic epidemic. *Mathematical Biosciences*, **107**, 209-224.

Jewell, N. P., Dietz, K. and Farewell, V. T. (1992). *AIDS Epidemiology: Methodological issues*. Birkhauser, Basel.

Korean Society of Infectious Diseases/Korean Society of Pediatric Infectious Diseases/ Korean Society of Epidemiology/Korean Societyfor Antimicrobial Therapy/Korean Society for Healthcare-associated Infection Control and Prevention/Korea Centers for Disease Control and Prevention, (2020) Report on the Epidemiological Features of Coronavirus Disease 2019 (COVID-19) Outbreak in the Republic of Korea from January 19 to March 2, 2020. *Journal of Korean Medical Sciences*, **35**, e112.

Lawless, J. F. (2011). *Statistical Models and Method for Life Time Data*. John-Wiley.

Lawless, J. F. and Sun, J. (1992). A comprehensive back-calculation framework for the estimation and prediction of AIDS cases. In: *AIDS Epidemiology: Methodological Issues*, (Eds. Jewell, N. P., Dietz, K. and Farewell, V. T.), 81-104.

Li, Y., Wang, B., Peng, R., Zhou, C., Zhan, Y., Liu, Z., Jiang, X. and Zhao, B. (2020). Mathematical Modeling and Epidemic Prediction of COVID-19 and Its Significance to Epidemic Prevention and Control Measures. *Annals of Infectious Disease and Epidemiology*, **5** (**1052**), 1-9.

Liu, Y., Qin, J., Fan, Y., Zhou, Y., Follmann, D. A. and Huang, C. Y. (2020). Infection Density and Epidemic Size of COVID-19 in China outside the Hubei province. https://doi.org/10.1101/2020.04.23.20074708

Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504–520.

Mode, C. J., Gollwitzer, H. E. and Hermann, N. (1988). A methodological study of a stochastic model of an AIDS epidemic. *Mathematical Biosciences*, **92**, 201-229.

Nishiura, H. (2020). Back-calculating the Incidence of Infection with COVID-19 on the Diamond Princess. *Journal of Clinical Medicine*, **9**, 1-4.

Ravanan, R. and Venkatesan, P. (2008). Some new approaches for modeling the incubation period of HIV/AIDS epidemic. *International Journal of Computer, Mathematical Sciences and Applications,* **2**, 223-237.

Ravanan, R. and Venkatesan, P. (2009). A simulation study on uncertainties associated with Back-calculation methodology. *International Journal on Information Sciences and Computing,* **3**, 47-52.

Solomon, P. J. and Wilson, S. R. (1990). Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics*, **46**, 1165-1170.

Sun, K., Chen, J. and Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *Lancet Digit Health*, **2**, e201-e208.

Tan, W. Y. (2000). *Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis*. World Scientific Publication, Singapore.

Venkatesan, P. (2006). A Comprehensive back-calculation frame work for estimation and projection of HIV/AIDS in India. *Journal of Communicable Diseases*, **38**, 40-56.

Venkatesan, P., Ramamurthy, D. and Sundaram, N. (2012). HIV/AIDS projection for Tamil Nadu using back calculation method. *Indian Journal of Science and Technology*, **5**, 3157-3162.

Wu, H. and Tan, W. Y. (1995). Modeling the HIV epidemic: A state space approach. In: *American Statistical Association 1995 Proceeding of the Epidemiology Section*, ASA, Alexdria, VA, 66-71.