HIV/AIDS Projection in TamilNadu using back calculation method

P. Venkatesan¹, D. Ramamurthy² and N. Sundaram³

¹Department of Statistics, National Institute for Research in Tuberculosis, ICMR, Chennai - 600031, India
²Department of Mathematics, Sir Theagaraya College, Chennai - 600 021, India
³Department of Statistics, Dr. Ambedkar Government Arts College, Chennai - 600 039, India
ramamurthy_phd@rediffmail.com

Abstract
The current prevalence of HIV infection and the corresponding pattern of incidence from the beginning of the epidemic to the present time are mainly estimated by means of back-calculation method. This back-calculation method reconstructs the past pattern of HIV infection and predicts the future number of AIDS cases with the present infection status. The basic data required for back-calculation methodology is the number of AIDS cases over a period of time. TANSACS publishes the reported number of AIDS cases in Tamil Nadu. In this paper, the various approaches for modeling the incubation distribution are compared using real data under various infection density distributions. The projected minimum and maximum AIDS cases in Tamil Nadu, a southern state of India, based on the reported data are 3702712 and 6936047 respectively. These estimates are based on the unadjusted AIDS incidence data. The purpose of this paper is to review the contribution of back-calculation method to our understanding of the AIDS and to summarize and interpret the epidemiological findings.

Keywords: HIV/AIDS, Incubation period, Estimation, Infection distributions, Back calculation.

Introduction
The acquired immunodeficiency syndrome (AIDS) was first recognized in 1981 (CDC, 1981). The etiologic agent, the Human Immunodeficiency Virus (HIV), was discovered in 1984 (Popovic et al., 1984). The first AIDS case in India was detected in 1986 and since then HIV infection has been reported in all states and union territories. As of 2009, about 31159 AIDS cases have been reported in all districts of Tamil Nadu, a southern state of India. Given the magnitude of epidemic, projection of future number of AIDS cases are of critical importance for assessing future health care needs.

Back calculation is one of the most useful methods for obtaining quantitative estimates of HIV prevalence and future AIDS incidence. This method was first proposed by Brookmeyer and Gail (1988). The AIDS incidence data is used in this method to estimate HIV incidence. This method has been used extensively to estimate the HIV infection and to develop short term projection of new AIDS cases.

Back calculation method requires accurate number of reported AIDS cases over each calendar year for several years and an incubation period distribution of AIDS. Using a basic relation among the three quantities, incubation period of AIDS, time of HIV infection and time of diagnosis of AIDS, the cumulative number of AIDS can be estimated as a convolution of HIV incidence density and incubation distribution of AIDS.

This method has been used extensively by many authors to estimate the HIV infection and to develop short term projections of new AIDS cases. (Brookmeyer & Gail, 1988; Bacchetti & Moss, 1989; Brookmeyer, 1991; Mariotti & Casciolo, 1996). Major sources of uncertainties may be due to the in accuracies of reported AIDS cases, the assumption about the infection curve and the incubation distribution. The inaccuracies in the reported AIDS cases may be due to the reporting delays and underreporting. Reporting delays of AIDS incidence has been modeled by Harries (1990); Brookmeyer and Damiano (1989) and Bacchetti et al. (1989) only few studies address the problem of underreporting. Uncertainties of AIDS incubation time and its effect on back calculation estimates can be found in Gigli and Verdecchia (2000). The effect of change of incubation time on the back calculation estimates is given in Dueffic and Castagiola (1999). Bayesian approaches for AIDS projection have also received lot of attention. Tan and Ye (2000) used state space models and generalized Bayesian method. Rao and Venkataramana (2001) explained the limitations of back calculation for the Indian data. Venkatesan (2002, 2006) and Anbupalam et al. (2002) explained the problem of applying back calculation to Indian data.

Projections of HIV/AIDS
Projections of HIV/AIDS using the statistical modeling approach are done based on the following three methods: (1) Fitting a model to the incidence of HIV/AIDS and extrapolating the curves into the future. The estimates obtained using this method depends on the mathematical function used and hence some function can produce anomalous results. This method is also less efficient as this does not include important information on the epidemic like incubation period, infection density and nature of the spread of the epidemic. (2) The next approach is based on modeling the dynamics of the epidemic. This approach requires certain knowledge about mixing pattern of HIV individual with probabilities of infection per contact, size of high risk behavior group,
probabilities of infection through blood product, needle sharing etc. In developing countries like India knowledge about these key parameters is incomplete. Also stochastic modeling of the epidemic demands many parameters, which are generally difficult to estimate due to limitation of appropriate data especially in the Indian context. (3) One of the most popular methods used for projection of HIV/AIDS is the back calculation method. This method is used to reconstruct the past pattern of HIV infection and to predict the future number of AIDS cases, apart from knowing the present infection status. This method depends on three important factors namely, the incubation period distribution, incidence curves and the observed number of AIDS cases over a time period.

There are also uncertainties associated with this approach because lack of certain information about incubation period distribution, the effect of intervention therapy on incubation period and errors in reported AIDS incidence. However calculation method is very popular, as it requires few information and assumptions and thus easy to apply.

**Method of back calculation**

The back calculation method for short-term projection of AIDS incidence has been formulated by Brookmeyer and Gail (1988) as a problem of likelihood estimation of multinomial parameters with unknown sample size. Let us assume that the numbers of reported AIDS cases are available during the calendar time $T_0$ to $T_L$. Here $T_0$ denotes the start of the epidemic in a certain region, the time point $T_L$ represents the time up to which reliable data on the AIDS is available. For example, in the present study $T_0$ is taken to be 1990 and $T_L$ = 2009. Let $X_j$ denote the number of reported AIDS cases in the interval $(T_{j-1}, T_j)$, $j = 1, 2, ..., L$. Here $X_{j+1}$ represent the number of individuals infected before time $T_j$ who do not become AIDS cases by the time $T_j$. The problem is to estimate $N=X_1+X_{j+1}+...+X_L+X_{L+1}$, the total number of infections before the time $T_L$. This number $N$ is the minimum size of the AIDS epidemic, because even if the infections after the year number $N$ is the minimum size of the AIDS epidemic, because even if the infections after the year $T_L$ could be prevented, the cumulative number of AIDS cases would eventually reach $N$. The minimum size is the sum of all cases already diagnosed called $n = \sum_{j=1}^{L} X_j$ 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.

Let the infection times of $N$ individuals be identically and independently distributed with a probability density function $I(s, \theta)$. Here $\theta$ can be vector valued and $I(s, \theta)$ integrated over the interval $[T_0, T_L]$. Then the probability that a susceptible individual infected before $T_L$ is diagnosed in the $j$th interval is given by

\[ P_j = \int_{T_{j-1}}^{T_j} I(s, \theta) \left[ F(T_j - s) - F(T_{j-1} - s) \right] ds, j = 1, 2, ..., L \] (1)

where $F(t)$ is an assumed incubation period distribution with $F(0)=0$ if $T \leq 0$. The probability that an individual infected before $T_L$ is not diagnosed before $T_j$ is $P_{j+1} = 1 - \sum_{i=1}^{L} P_i$. Hence $X = (X_1, X_2, ..., X_L, X_{L+1})$ can be assumed to be multinomial with cell probabilities $(P_1, P_2, ..., P_L, P_{L+1})$ and unknown sample size $N$.

The likelihood function with $N$ as an unknown parameter can be written as

\[ L(N, \theta) = \frac{N!}{(N-n)!} \prod_{j=1}^{L} x_j! \] (2)

The interest here is to obtain the estimates for $\theta$, which appear in the definition of $P_j$ and $N$. Maximum likelihood estimates for $\theta$ can be obtained by maximizing $L(N, \theta)$ simultaneously over $N$ and $\theta$. But this problem is computationally difficult task especially when the number of parameters in $I(\theta, t)$ are large. Under certain regularity conditions Sanathanan (1972) has shown that the estimates obtained by maximizing the conditional likelihood and unconditional likelihood are both consistent and asymptotically normal. However Brookmeyer and Gail (1988) have suggested the use of EM algorithm for maximization of full likelihood when $I(s, \theta)$ is a step function. Ding (1996) has given a simpler regularity conditions and has given the application with back calculation estimates. It is known that the EM algorithm is computationally intensive and slowly convergent. Ding (1995) has considered the conditional likelihood approach for the data sets used by Brookmeyer and Gail (1988) and Rosenberg and Gail (1991) and has observed the estimates obtained in these two papers and the conditional likelihood estimated are comparable. Moreover the simulation study by Ding (1996) has shown that the coverage probabilities associated with the confidence intervals based on normal approximation for the parameters are very closed to the true values. Hence in this paper conditional likelihood approach is used for estimation of the parameters. The conditional likelihood approach can be described as follows.

\[ L(N, \theta) = L_1(N, \theta) L_2(\theta) \] (3)

Where

\[ L_1(N, \theta) = \frac{N!}{n! (N-n)!} \prod_{j=1}^{L} \left[ q_j (\theta) x_j \right]^{n_j} \left[ 1 - q_j (\theta) \right]^{n-n_j} \] (4)

and

\[ L_2(\theta) = n! \prod_{j=1}^{L} \left[ q_j (\theta) x_j \right]^{n_j} \] (5)

with $q_j (\theta) = \frac{P_j (\theta)}{1-P_j (\theta)} = \frac{F_j (\theta)}{1-F_j (\theta)}$

The conditional likelihood estimates of $\theta$ is obtained by maximizing $L_2(\theta)$ and then $N$ is obtained by maximizing $L_1(N, \theta)$. Sanathanan (1972) has shown that the conditional likelihood estimate of $N$ is $N = \left[ n! \left[ 1 - q_j (\theta) \right]^{-1} \right]$ where $[x]$ denotes the greatest integer less than or equal to $x$. Ding (1995) has given explicit expression for the confidence intervals for the parameters $\theta$ and $N$. Suppose $(\bar{\theta}, \bar{N})$ is given then estimates for

\[ \bar{N} \pm z_{\alpha/2} \frac{SE(\bar{N})}{\sqrt{n}} \]
expected number of AIDS cases and variances are given by

\[ \hat{E}(X) = \hat{N} \hat{P} \] and \[ \hat{V}(X) = \hat{N} \hat{P}(1 - \hat{P}) \text{.} \]

The future number of AIDS cases can be predicted by extrapolating the infection curve \( V(s, \theta) = N I(s, \theta) \). Suppose our interest is to predict the number of AIDS cases \( X_{AB} \) in an interval of the calendar time beyond \( T_L \). \( (T_L, T_B) T_A, T_B \). This number will be the sum of number AIDS cases, \( Y_{AB} \) who were infected before the time \( T_L \) by diagnosed in this interval and the individuals who are infected after \( T_B \) but diagnosed as AIDS in the interval \( \{T_A, T_B\} \). Then the probability of diagnosing AIDS in the interval \( \{T_A, T_B\} \) infected before \( T_L \), is given by

\[ \hat{P}_{AB} = \int_{T_0}^{T_L} I(s, \theta) [F(T_B - s) - F(T_A - s)] \, ds \] \hspace{1cm} (6)

Therefore an estimate for \( Y_{AB} \) is \( \hat{E}(Y_{AB}) = \hat{N} \hat{P}_{AB} \) (7)

with \[ \hat{V}(Y_{AB}) = \hat{N} \hat{P}_{AB} (1 - \hat{P}_{AB}) \] \hspace{1cm} (8)

Similarly an estimate for \( Z_{AB} \), by extrapolating the infection curve \( V(s, \theta) \) beyond \( T_L \) can be obtained as

\[ \hat{F}(Z_{AB}) = \int_{T_0}^{T_L} V(s, \theta) [F(T_B - s) - F(T_A - s)] \, ds \] \hspace{1cm} (9)

An estimate of variance of \( Z_{AB} \) is given by

\[ \hat{V}(Z_{AB}) = \hat{N} \hat{P}_{AB} (1 - \hat{P}_{AB}) \] \hspace{1cm} (10)

With

\[ \hat{F}_{AB} = \int_{T_0}^{T_L} V(s, \theta) [F(T_B - s) - F(T_A - s)] \, ds \] \hspace{1cm} (11)

The epidemic density over the interval \( [T_A, T_B] \) is \( H(s, \theta) = V(s, \theta) / M \), with \( M = \int_{T_0}^{T_B} I(s, \theta) \, ds \)

Therefore estimates of the total expected number of AIDS cases are given of

\[ \hat{E}(X) = \hat{E}(Y_{AB}) + \hat{E}(Z_{AB}) \] \hspace{1cm} (12)

with variance \[ \hat{V}(X) = \hat{V}(Y_{AB}) + \hat{V}(Z_{AB}) \] \hspace{1cm} (13)

**Application of back calculation to Tamil Nadu AIDS data**

It is of interest to apply back calculation methods to reported AIDS cases in Tamil Nadu. The National AIDS Control Organization NACO provides the monthly updates of reported number of AIDS cases in all over India. For the present study, number of AIDS cases during the period 1993 to 2009 has been obtained from Tamil Nadu AIDS Control Society. Yearly reported AIDS cases are given in Table 1.

Out of the following five infection curves for infection density exponential and logistic were used in this work.

1. Log - logistic: \( V(x, \theta_1, \theta_2, \theta_3) = \theta_1, T_2 (\theta_1, x) x^{\theta_2-1} / (1 + (\theta_1, x)^{\theta_3})^2 \) \hspace{1cm} (14)
2. Logistic prevalence: \( V(x, \theta_1, \theta_2, \theta_3) = \theta_1, x \exp(\theta_2 + \theta_3 x) (1 + \exp(\theta_2 + \theta_3 x))^2 \) \hspace{1cm} (15)
3. Logistic incidence: \( V(x, \theta_1, \theta_2, \theta_3) = \theta_1, x \exp(\theta_2 + \theta_3 x) (1 + \exp(\theta_2 + \theta_3 x)) \) \hspace{1cm} (16)
4. Root exponential: \( V(x, \theta_1, \theta_2, \theta_3) = \theta_1, x \exp(\theta_2 + \theta_3 x) \) \hspace{1cm} (17)
5. Exponential: \( V(x, \theta_1, \theta_2, \theta_3) = \theta_1, x \exp(\theta_2 x) \) \hspace{1cm} (18)

Taylor (1989) used the infection curves 2 to 5, Brookmeyer and Damiano (1989) and Ding (1995) used the infection curve 1. Note that these infection curves have to be suitably normalized so that they integrate to 1 between the calendar times \( T_0 \) and \( T_L \) and they represent the infection densities.

The incubation distribution was taken to be Weibull and Log-logistic. The Weibull distribution function is given by \( f(x) = \frac{\alpha}{\lambda} \left(\frac{x}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\lambda}\right)^\alpha\right) \)

Table 2. Parametric values of the incubation period distributions

<table>
<thead>
<tr>
<th>Incubation Period Distribution</th>
<th>Parameters</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>( \lambda )</td>
<td>0.11170</td>
<td>0.089361</td>
<td>0.074467</td>
<td>0.059574</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>3.2582</td>
<td>3.2582</td>
<td>3.2582</td>
<td>3.2582</td>
</tr>
<tr>
<td>Log-Logistic</td>
<td>( \lambda )</td>
<td>0.000152423</td>
<td>0.0000059355</td>
<td>0.000027466</td>
<td>0.000010695</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>4.22654</td>
<td>4.22654</td>
<td>4.22654</td>
<td>4.22654</td>
</tr>
</tbody>
</table>

Table 3. Estimates of HIV incidence and projection of AIDS with median incubation period of 8 years

<table>
<thead>
<tr>
<th>Infection model</th>
<th>Incubation model</th>
<th>Incidence in 2009</th>
<th>Projection of AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Weibull</td>
<td>381457</td>
<td>54355</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>39631</td>
<td>65824</td>
</tr>
<tr>
<td>Logistic</td>
<td>Weibull</td>
<td>37892</td>
<td>52265</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>38275</td>
<td>62268</td>
</tr>
</tbody>
</table>

Results

The results based on the conditional likelihood approach to the multinomial likelihood are summarized in Table 2 to 6. The results show wide variability of the...
Table 4. Estimates of HIV incidence and projection of AIDS with median incubation period of 10 years

<table>
<thead>
<tr>
<th>Infection model</th>
<th>Incubation model</th>
<th>Incidence in 2009</th>
<th>Projection of AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010</td>
</tr>
<tr>
<td>Exponential</td>
<td>Weibull</td>
<td>1874775</td>
<td>428347</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>1480165</td>
<td>352212</td>
</tr>
<tr>
<td>Logistic</td>
<td>Weibull</td>
<td>14981960</td>
<td>429946</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>1425149</td>
<td>327796</td>
</tr>
</tbody>
</table>

Table 5. Estimates of HIV incidence and projection of AIDS with median incubation period of 12 years

<table>
<thead>
<tr>
<th>Infection model</th>
<th>Incubation model</th>
<th>Incidence in 2009</th>
<th>Projection of AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010</td>
</tr>
<tr>
<td>Exponential</td>
<td>Weibull</td>
<td>2600262</td>
<td>728495</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>2325524</td>
<td>514796</td>
</tr>
<tr>
<td>Logistic</td>
<td>Weibull</td>
<td>2070157</td>
<td>709301</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>1900969</td>
<td>619729</td>
</tr>
</tbody>
</table>

Table 6. Estimates of HIV incidence and projection of AIDS with median incubation period of 15 years

<table>
<thead>
<tr>
<th>Infection model</th>
<th>Incubation model</th>
<th>Incidence in 2009</th>
<th>Projection of AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2010</td>
</tr>
<tr>
<td>Exponential</td>
<td>Weibull</td>
<td>3988340</td>
<td>903800</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>3271664</td>
<td>832574</td>
</tr>
<tr>
<td>Logistic</td>
<td>Weibull</td>
<td>2814326</td>
<td>938011</td>
</tr>
<tr>
<td></td>
<td>Log-logistic</td>
<td>2842934</td>
<td>865619</td>
</tr>
</tbody>
</table>

Fig. 1. Weibull Model for exponential incidence

Fig. 2. Log-logistic model for exponential incidence

Fig. 3. Weibull model for logistic incidence
minimum size of the cumulative number of AIDS cases that may ultimately develop even if further infection beyond 2009 is stopped. The estimated AIDS cases for the future show a consistent pattern. But if the present trend of reporting by TANSACS continues the estimates given here is the expected number of AIDS cases in the near future. Hence the projected AIDS cases given here can be taken to be the expected number of cases that are likely to be reported to TANSACS, if not the actual AIDS cases that may develop in Tamil nadu. The comparison across the incubation period distribution reveals that the data is very insensitive to the change in the incubation pattern. It can be noted that the reported AIDS cases in Tamil nadu suffers both underreporting and reporting delays (Fig. 1-4).

Discussion

It is generally observed that the short term projected AIDS cases do not vary much across various infection densities and incubation period distribution. But the minimum size of the epidemic and HIV incidences are highly variable across the infection densities and incubation period distributions. The projected AIDS cases within an infection density across various incubation distributions are found to be very stable. But across the infection densities the variation is observed to be high. The projected AIDS cases using the logistic infection density are less compared to the estimates obtained using the exponential infection density. These estimates may not be the correct number of AIDS cases that may develop in Tamil nadu during these periods. The exact number of AIDS cases that may develop in Tamil nadu will be certainly higher than these figures and hence these figures can be taken to be a lower bound for possible number of AIDS cases. The HIV incidence reported for the year 2009 is calculated based on the relationship between the infection density and the infection curves. These figures are found to be highly variable and are not smoothed estimates. Hence these figures cannot be taken as exact number of HIV incidence in the year 2009. For all combinations of two infection densities and two incubation period distributions, the increase as median incubation period increases.

References


