Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

HIV/AIDS Projection in TamilNadu using back calculation method

P. Venkatesan[1], D. Ramamurthy[2] and N. Sundaram[3]

[1]*Department of Statistics, National Institute for Research in Tuberculosis, ICMR, Chennai – 600031, India*
[2]*Department of Mathematics, Sir Theagaraya College, Chennai – 600 021, India*
[3]*Department 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 & Cascilio, 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,

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

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 back 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_{L+1} represent the number of individuals infected before time T_L who do not become AIDS cases by the time T_L. The problem is to estimate $N=X_1+X_2+….+X_L+X_{L+1}$, 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 $n = \sum_{i=1}^{L} X_i$ 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.

Let the infection times of N individuals be identically and independently distributed with a probability density function I(s, θ). Here θ can be vector valued and I(s, θ) integrated to 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

$$P_j = \int_{T_0}^{T_L} I(s,\theta)\{F(T_j - s) - F(T_{j-1} - s)\}ds, j = 1,2,\dots,L \quad (1)$$

where F(t) is an assumed incubation period distribution with F(t)=0 if T≤ 0. The probability that an individual infected before T_L is not diagnosed before T_L is $P_{L+1} = 1 - \sum_{i=1}^{L} P_i$. 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

$$L(N,\theta) = \frac{N!}{(N-n)!\prod_{j=1}^{L} x_j!}(P_{L+1})^{N-n}\prod_{j=1}^{L} p_j^{x_j}! \quad (2)$$

The interest here is to obtain the estimates for θ, which appear in the definition of P_j and N. Maximum likelihood estimates for N and θ can be obtained by maximizing L (N, θ) simultaneously over N and θ. But this problem is computationally difficult task especially when the number of parameters in I (θ, t) are large. Under certain regularity conditions Sanathanan (1972) hasshown 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, θ) 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.

$$L(N,\theta) = L_1(N, P_{L+1}(\theta))L_2(\theta) \quad (3)$$

Where

$$L_1(N, P_{L+1}(\theta)) = \frac{N!}{n!(N-n)!}[P_{L+1}(\theta)]^{N-n}[1 - P_{L+1}(\theta)]^{n} \quad (4)$$

and $L_2(\theta) = n!\prod_{j=1}^{L}\frac{[q_j(\theta)]^{x_j}}{x_j!}$

with $q_j(\theta) = \frac{P_j(\theta)}{1-P_{L+1}(\theta)} = \frac{P_j(\theta)}{\sum_{i=1}^{L} P_i(\theta)}$ (5)

The conditional likelihood estimates of θ is obtained by maximizing $L_2(\theta)$ and then N is obtained by maximizing L1(N, $P_{L+1}(\theta)$). Sanathanan (1972) has shown that the conditional likelihood estimate of N is $N = [n\{1 - P_{L+1}(\theta)\}^{-1}]$ 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 θ and N. Suppose $(\hat{N}, \hat{\theta})$ is given then estimates for

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Table 1. Reported numbers of AIDS cases in Tamil nadu

Year	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Reported AIDS cases	169	199	210	301	545	801	2391	3964	6484	10725
Year	2003	2004	2005	2006	2007	2008	2009			
Reported AIDS Cases	24667	32423	14394	27392	40841	34295	31159			

Table 2. Parametric values of the incubation period distributions

Incubation Period Distribution	Parameters	Median Years			
		8	10	12	15
Weibull	λ	0.11170	0.089361	0.074467	0.059574
	α	3.2582	3.2582	3.2582	3.2582
Log-Logistic	λ	0.000152423	0.0000059355	0.000027466	0.000010695
	α	4.22654	4.22654	4.22654	4.22654

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

Infection model	Incubation model		Incidence in 2009		Projection of AIDS			
					2010	2011	2012	2013
Exponential	Weibull	1188501	381457	39631	54355	65824	79713	96532
	Log-logistic	1074167	290295	37982	52265	62268	74151	88275
Logistic	Weibull	1338818	412362	36525	51355	62824	76913	90532
	Log-logistic	1374127	420093	38982	52806	65268	75151	83275

expected number of AIDS cases and variances are given by

$\hat{E}(X_j) = \hat{N}\hat{P}_j$ and $\hat{V}(X_j) = \hat{N}\hat{P}_j(1-\hat{P}_j)$. The future number of AIDS cases can be predicted by extrapolating the infection curve V(s,θ) = N I(s,θ). Suppose our interest is to predict the number of AIDS cases X_{AB} in an interval of the calendar time beyond T_L, (T_A,T_B) $T_A > T_L$. 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_L but diagnosed as AIDS Z_{AB} 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

$$\hat{p}_{AB} = \int_{T_0}^{T_L} \mathrm{I}(s,\hat{\theta})\{\mathrm{F}(\mathrm{T_B}-s) - \mathrm{F}(\mathrm{T_A}-s)\}\,ds \quad (6)$$

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

with $\hat{V}(Y_{AB}) = \hat{N}\hat{p}_{AB}(1-\hat{p}_{AB})$. (8)

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

$$\hat{E}(Z_{AB}) = \int_{T_0}^{T_L} \mathrm{V}(s,\hat{\theta})\{\mathrm{F}(\mathrm{T_B}-s) - \mathrm{F}(\mathrm{T_A}-s)\}\,ds \quad (9)$$

An estimate of variance of Z_{AB} is given by

$\hat{V}(Z_{AB}) = \hat{M}\hat{r}_{AB}(1-\hat{r}_{AB})$ (10)

With

$$\hat{r}_{AB} = \int_{T_0}^{T_L} \mathrm{V}(s,\hat{\theta})\{\mathrm{F}(\mathrm{T_B}-s) - \mathrm{F}(\mathrm{T_A}-s)\}\,ds \quad (11)$$

The epidemic density over the interval $[T_I, T_B)$ is H (s,θ) = V (s,θ)/M, with $M = \int_{T_L}^{T_B} I(s,\theta)ds$

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

$\hat{E}(X_{AB}) = \hat{E}(Y_{AB}) + \hat{E}(Z_{AB})$ (12)

with variance $\hat{V}(X_{AB}) = \hat{V}(Y_{AB}) + \hat{V}(Z_{AB})$. (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 the following Table 1.

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

(1) Log - logistic: $V(x,\theta_1,\theta_2) = \theta_1,\theta_2(\theta_1,X)^{\theta_2-1}/\{1+(\theta_2 x)^{\theta_2}\}^2$ (14)

(2) Logistic prevalence:

$$V(x,\theta_1,\theta_2,\theta_3) = \frac{\theta_1\theta_3\exp(\theta_2+\theta_3 x)}{\{1+\exp(\theta_2+\theta_3 x)\}^2} \quad (15)$$

(3) Logistic incidence:

$$V(x,\theta_1,\theta_2,\theta_3) = \frac{\theta_1\exp(\theta_2+\theta_3 x)}{1+\exp(\theta_2+\theta_3 x)} \quad (16)$$

(4) Root exponential:

$$V(x,\theta_1,\theta_2) = \theta_1\exp\{\theta_2(x)^{1/4}\} \quad (17)$$

(5) Exponential: $V(x,\theta_1,\theta_2) = \theta_1\exp(\theta_2 x$ (18)

Taylor (1989) used the infection curves 2 to 5, Brookemeyer and Damiano (1989) and Ding (1995) used the infection curve 1. Note that these infection curves has 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 $F(t) = 1 - \exp(-\lambda t)^{\alpha}$, following Brookmeyer and Goedert (1989), the parameter α was fixed to be 3.2582 and λ was chosen according to the median incubation period of 4, 8, 10, 12 and 15 years. Similarly the log-logistic distribution function is given by, $F(t) = 1 - [1 + (\lambda t)^{\alpha}]^{-1}$, the parameter α has been fixed to be 4.22654 and the parameter λ has been varied corresponding to the median incubation periods of 4, 8, 10, 12 and 15 years (Chiaroti *et al.*,1994).

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

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Table 4. Estimates of HIV incidence and projection of AIDS with median incubation period of 10 years

Infection model	Incubation model		Incidence in 2009		Projection of AIDS			
					2010	2011	2012	2013
Exponential	Weibull	1874775	428347	27841	45543	51801	58913	67000
	Log-logistic	1480165	352212	36516	44736	50064	55884	62264
Logistic	Weibull	14981960	429946	26019	46089	50915	57913	68100
	Log-logistic	1425149	327796	44134	41211	44392	47582	50808

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

Infection model	Incubation model		Incidence in 2009		Projection of AIDS			
					2010	2011	2012	2013
Exponential	Weibull	2600262	728495	22450	36286	36751	37002	37127
	Log-logistic	2325524	514796	74054	41010	43061	44827	46342
Logistic	Weibull	2070157	709301	21762	30144	30516	32431	33324
	Log-logistic	1900969	619729	49239	39894	42289	44550	46702

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

Infection model	Incubation model		Incidence in 2009		Projection of AIDS			
					2010	2011	2012	2013
Exponential	Weibull	3988340	903800	67389	58097	63383	67620	70852
	Log-logistic	3271664	832574	66631	60178	66797	72871	78383
Logistic	Weibull	2814326	930811	19382	21635	24083	28301	30613
	Log-logistic	2842934	865619	66267	37021	37750	38114	38174

Fig. 1. Weibull Model for exponential incidence

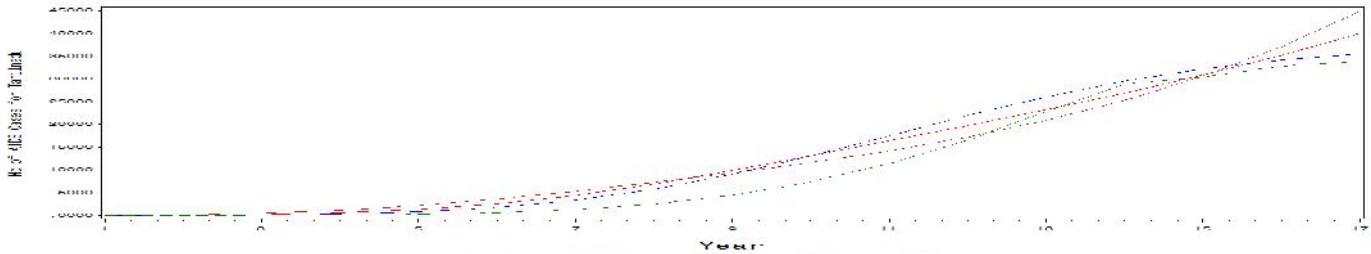


Fig. 2. Log-logistic model for exponential incidence

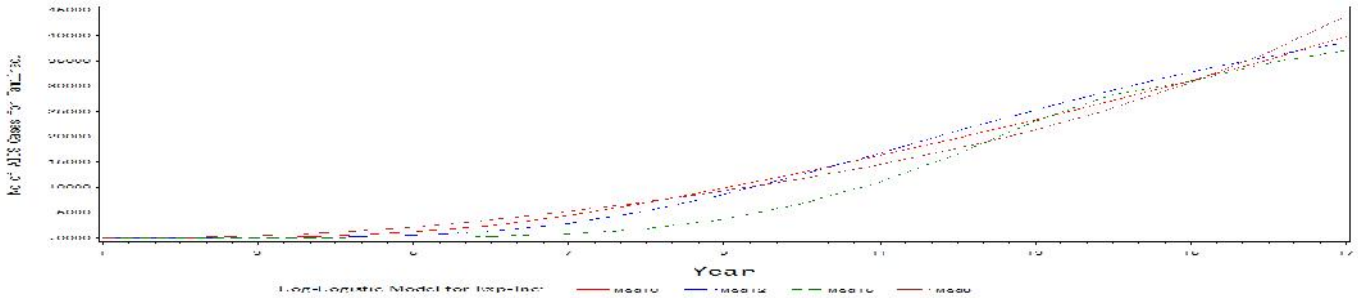
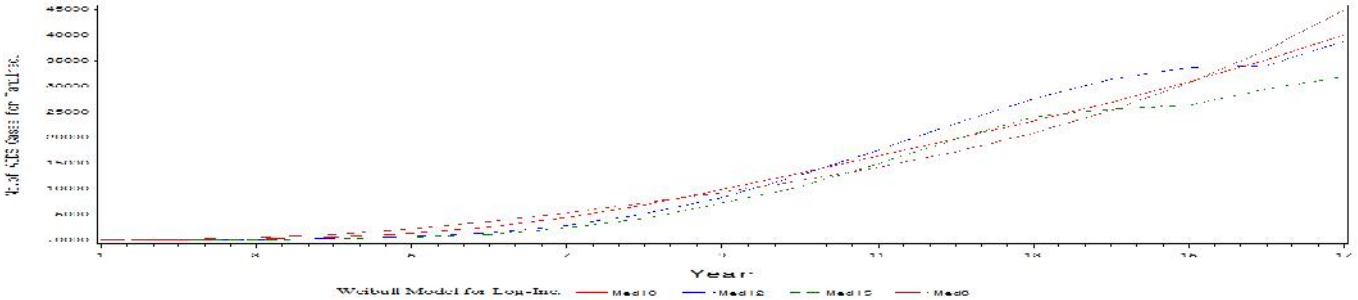
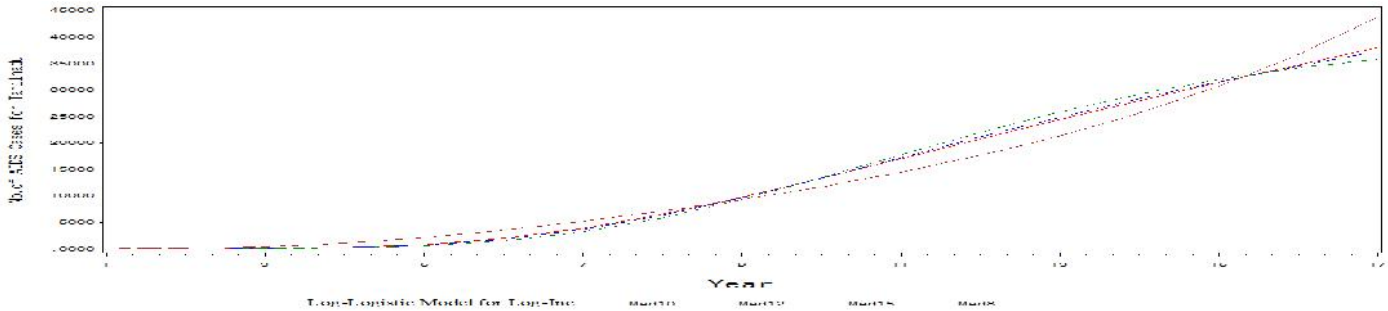


Fig. 3. Weibull model for logistic incidence



Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

Fig. 4. Log-logistic 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

1. Anbupalam T, Ravanan R and Venkatesan P (2002) Backcalculation of HIV/AIDS in Tamilnadu: In Biostatistical aspects of Health and Epidemiology (Edition: Pandey, Pradeep Mishra and Uttam Singh), Department of Biostatistics. *Sanjay Gandhi Postgraduate Instit. Medic. Res.,* Lucknow, India. pp: 232-243.
2. Brookmeyer R (1991) Reconstruction and future trends of the AIDS epidemic in the United States. *Sci.*253, 37-42.
3. Brookmeyer R and Gail MH (1988) A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *J. American Statis. Assoc.*83, 301-08.
4. Brookmeyer R and Damiano A (1989) Statistical methods for short-term projections of AIDS incidence. *Statis.Medic.*8, 23-34.
5. Brookmeyer R and Goedert JJ (1989) Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*. 45, 325-335.
6. Bacchetti P and Moss AR (1989) Incubation period of AIDS in San Francisco. *Nature.* 338, 251-253.
7. Centers for Disease Control (1981) Pneumocystis pneumonia-Los Angeles. Morbidity and Mortality. *Weekly Report.* 30, 250-252.
8. Chiarotti F, Palombi M, Schinaia N, Ghirardini A and Bellocco R (1994) Median time from seroconversion to AIDS in Italian HIV positive hemophiliacs, different parametric estimates. *Statis. Medic*. 13, 163-175.
9. Ding Y (1995) Computing backcalculation estimates of AIDS epidemic. *Statis. Medic.*14, 1505-1512.
10. DingY (1996) On the asymptotic normality of multinomial population size estimates with application to the back calculation epidemic of AIDS. *Biometrika.*83, 695-699.
11. Dueffic S and Costagliola D (1999) Is the incubation time changing? A backcalculation approach. *Statis. Medic.* 18, 1031-1047.
12. Gigli A and Verdecchia (2000) Uncertainties of AIDS incubation time and its effect on backcalculation estimates. *Statis. Medic.*19, 175-189.
13. Mariotti S and Cascilio R (1996) Sources of uncertainty in estimating HIV infection rates by

Indian Journal of Science and Technology Vol. 5 No. 8 (August 2012) ISSN: 0974- 6846

backcalculation: Application to Italian data. *Statis. Medic.* 15, 2669-2687.

14. Rao CN and Srivenkataramana T (2001) Projection of HIV infections in India: An alternative to backcalculation. *Curr. Sci.* 81, 1302-1307.
15. Rosenberg PSand Gail MH (1991) Back calculation of flexible linear models of the human immunodeficiency virus infection curve. *Appl. Statis.* 40, 269-282.
16. Sanathanan L (1972) Estimating the size of a multinomial population. *Annals Maths. Statis.* 43,142-152.
17. Tan WY and Ye ZZ (2000) Estimation of HIV infection and HIV incubation via state space models. *Maths. Biosci.*167, 31-50.
18. Taylor JMG (1989) Models for the HIV infection and AIDS epidemic in the United States. *Statist. Medic.* 8, 45-58.
19. Venkatesan P (2002) Methods for projection of HIV/AIDS epidemic: Epidemiology, Health and population (Edition: Anil Kumar).*Proc. 18th Ann. Conf. Ind. Soc. Medic. Statis*. pp: 143-155.
20. Venkatesan P (2006) A comprehensive back calculation framework for estimation and prediction of HIV/AIDS. *Indi. J. Communicable Disease.* 38(1), 40-56.