J. Commun. Dis. 38 (1) 2006 :40-56

A Comprehensive Back-calculation Framework for Estimation and Prediction of HIV/AIDS in India

P. Venkatesan[*]

ABSTRACT

HIV incubation period is the random time between the HIV-infection and the onset of clinical AIDS. Distribution of this non-negative random variable is known as HIV incubation period distribution. The Back-calculation method reconstructs the past pattern of HIV infection and predicts the future number of AIDS cases with the present infection status. It depends on three important factors: incubation distribution, incidence curve and observed number of AIDS cases over time. This method is very popular and requires less information and assumptions. Lack of information about incubation distribution, the effect of intervention therapy on incubation period, and errors in reported AIDS incidence leads to uncertainties associated with this method. The incubation distribution is assumed to be exactly known in back-calculation methodology. Incubation period of HIV is very long and highly variable within and between cohorts. 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. It calculates the most likely temporal distribution of infected individuals compatible with the number of observed AIDS cases starting from the suitable estimate of the incubation period, derived from the available data. Most of the projections formulated the problem of estimation of future AIDS cases as estimation of parameters in multinomial likelihood with unknown sample size by EM algorithm. In this paper the various approaches for modeling the incubation distribution are compared using real and simulated data under various infection density distributions. The projected minimum AIDS cases in India based on the reported data for the years 2003, 2004, 2005 and 2006 are around 27000, 44000, 70000 and 113000 respectively. The corresponding figures based on the adjusted data are around 67000,100000,150000 and 230000 respectively.

[*] Department of Statistics, Tuberculosis Research Centre (ICMR), Chetpet, Chennai-600 031
venkatesanp@trcchennai.org,

INTRODUCTION

The Acquired Immune Deficiency Syndrome (AIDS) is a devastating disease caused by Human Immune deficiency Virus (HIV) that is transmitted through either sexual, other contacts in which body fluids are exchanged or using infected blood products. Following a few recognized cases among homosexual men in the United States in the year 1981, new cases of AIDS were subsequently reported in a many countries throughout the world. It has now reached a pandemic proportion, as no country in the world is now free from HIV/AIDS. This epidemic ranks as one of the most destructive microbial scourges in human history and has posed a formidable challenge to the biomedical research and public health communities of the world.

The global pandemic of HIV infection comprises many different epidemics, each with its own dynamics e.g., time of introduction, population density, cultural and social issues. Spread of the epidemic has varied considerably between developed and developing countries, depending on the culture as well as other social and behavioral patterns. Incidence rates have been the highest in developing countries where heterosexual transmission is most common.

First case of AIDS was reported in India in the year 1986. Now India's entry into the third phase of the HIV epidemic, as envisaged from the increasing number of HIV infections detected even among housewives and children, signals major AIDS crisis in the offing. The first phase of HIV was recognized when rising trend of HIV prevalence was established among the Commercial Sex Workers (CSW) in 1988 and the professional blood donors in 1989. The second phase started after 1989 when several of the clients of CSW and blood recipients were found to be infected.

BIOLOGICAL ASPECTS OF HIV/AIDS AND ITS TRANSMISSION

AIDS is a condition in which the in built immune mechanism of the human body breaks down completely. The process is gradual but ultimately suppresses the immune ability of the individuals. It is a medically accepted fact that the HIV is the causal agent of AIDS. Those who are affected by AIDS are susceptible to other opportunistic infections like Candida, Mucor Omycosis and Aspergillus etc.

Within a few weeks of its entry into the blood stream of an individual, the HIV sets an insidious and progressive attack on the immune system of the individual by binding itself to one of the T4 helper cells. The viral attachment to the cell is initiated by means of antibodies to the virus. On infecting the T4 cell, the viral RNA is injected into the target cell along with viral transcriptase and viral integrase. The viral reverse transcriptase transcribes the viral RNA into viral DNA and the viral integrase bind the viral DNA with that of the host. The integrated viral DNA may remain latent or in an activated form. In the activated form, genomic RNA and messenger RNA are transcribed from the integrated viral DNA. The regulatory proteins such as *tat* and *rev* are translated from messenger

RNA. These proteins together with the genomic RNA cause the production of new HIV viruses, which bud on the cell wall. These buds accumulate on the cell wall and after a random length of time from the time of infection, the infected cell disintegrates (i.e. the cell undergoes a lyses) releasing a random number of HIV viruses and this process continues indefinitely. This mechanism by which the killing of the T4 cells takes place, the number of free virus in the blood destroys progressively the immune-competence of the host. The viral load (the amount of free HIV in the blood) increases following infection and peaks at the time of sero-conversion (i.e. the time at which detectable levels of antibodies develop in the blood) and falls to a level (called *set-point*) by about 2 years thereafter and then remains at that level throughout the asymptomatic period. A low level of set point corresponds to a low risk of AIDS and a high level of set point corresponds to a high risk of AIDS. The viral load measured at arbitrary base line is a good prognostic marker at all stages of the development of AIDS. On the other hand, the T4 cell counts in the blood decrease rapidly in the initial months following infection and thereafter at a slower rate. After a period of several years, when the level of T4 cells declines to roughly 400 cells/mm^3, the person exhibits symptoms and signs associated with HIV disease. These symptoms and signs do not meet the surveillance definition of AIDS and we say that the person is suffering from AIDS related complex (ARC). The level of T4 cells further continues to decrease to roughly below 200 cells/mm^3 and AIDS defining diseases such as *Kaposi's sarcoma and Pneumocystis carinii pneumonia* occur. At this stage, the patient is said to have developed AIDS.

APPROACHES FOR HIV/AIDS MODELING

In general, HIV/AIDS epidemic modeling is categorized based on the following four broad approaches but not mutually exclusive ones[1].

Deterministic Models: In this type of modeling the parameters such as number of susceptible individuals, infected individuals and number of AIDS cases are assumed to be deterministic. These models are described by system of differential or integral equations. The progression of the epidemic is studied using these equations. Some of the deterministic models for AIDS epidemic were developed[2-7].

Stochastic Models: Stochastic models assume that some of the key parameters are random variables. It is assumed that the HIV epidemic is a continuous time stochastic process. Stochastic models are considered to be more realistic than deterministic models and with some special assumptions the results of deterministic models can be approximated through stochastic models. But there are studies, which show that stochastic models give a better interpretation of the epidemic than the deterministic models[8-11].

Statistical models: The statistical models are developed based on AIDS epidemiological and survey data. These models make full use of the available data compared to deterministic and stochastic models. But in this type of modeling the mechanism and prior information system is usually not taken into consideration. The back-calculation approach for HIV/AIDS projection can be categorized into this type of modeling[12-14].

State Space Models: Wu and Tan (1995)[15] have introduced the state space models for AIDS epidemic, which takes the advantage of stochastic and statistical models. The state space models were originally proposed[16] for engineering control and communication. This model is also used[10-11,17] in projection and detailed description of the state space models is given by Tan (2000)[1].

PROJECTION OF HIV/AIDS

Projections of HIV/AIDS using the statistical modeling approach are done based on the following three methods :(i). Fitting a model to the incidence of HIV/AIDS and extrapolating the curves into the future[18]. 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.(ii). The next approach is based on modeling the dynamics of the epidemic[19-20]. 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. There is a lot of literature on deterministic and stochastic models for the spread of HIV epidemic[1]. (iii). One of the most popular method used for projection of HIV/AIDS is the back calculation method[21-22]. 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.

Back-calculation Methodology

The back calculation method for short-term projection of AIDS epidemic was introduced[21-22]. The infection rate represents the number of new HIV infections per unit time at calendar time. Sero-conversion refers to infected individuals remain sero-negative until they develop detectable HIV antibodies. The sero-conversion period is usually short and it has a median of 2 to 3 months. The incubation period is the time duration between sero-conversion to AIDS. The incubation periods are highly variable across groups. Several models for incubation period distribution are used to describe the incubation period.

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 HIV infections[23-24]. The time between HIV infection and the diagnosis of AIDS is known as incubation time and it is modeled by a known distribution. Future AIDS incidence can be projected by the estimated infection curve and the known incubation distribution.

Most of the research on the projection of AIDS epidemic is based on the back calculation method[13-14,25]. There are various sources of uncertainties associated with the estimates of AIDS[14,26] based on back-calculation method. Major sources of uncertainties may be due to the inaccuracies of reported AIDS cases, the assumptions about the incidence curve and the incubation distribution. The inaccuracies in the reported AIDS cases may be due to the reporting delays, under diagnosis and underreporting. Reporting delays of AIDS incidence has also been modeled[13,27-28]. It was observed[29] that only a few studies address the problem of underreporting. Various forms of incidence curves and several parametric incubation distributions have been used in the literature. Uncertainties of AIDS incubation time and its effect on back calculation estimates are discussed[30-31].

The methodology to include information from the results of surveillance to improve the accuracy of projection of AIDS cases was given by Bellico and Marschner[32]. Bayesian approaches for AIDS projection have also received a lot of attention[33-34]. Swaminathan et al., (2000)[35] presented the survival experience of a group of HIV patients. Venkatesan (2002)[36] discussed the various mathematical and statistical approaches for projection of HIV/AIDS epidemic. Projections using back calculation method are explored[37]. In this work we have concentrated mainly on understanding the current state of epidemic and forecasting the future path via back calculation method. The broad objective is to provide HIV/AIDS estimates under different model assumptions using surveillance data for India.

Models for Incubation Period

The HIV incubation period is the random time between the HIV infection and the onset of clinical AIDS symptoms. The probability distribution of this non-negative random variable is known as HIV incubation period distribution. Medley et al. (1987)[38] showed that the incubation period of HIV is known to be very long and it is highly variable within and between cohorts.

The analysis of incubation period is very important in AIDS epidemic studies. Incubation period distribution is assumed to be exactly known in back calculation methodology. Several studies [1,6,13-14,26,39-41] have observed that back calculation estimates are very sensitive to the choice of incubation period distribution., It has also been shown that the HIV incubation distribution is significantly affected by age, treatment by antiviral drugs and other opportunistic infections[42-48]. Hence for estimation of the HIV infection and projection of future HIV prevalence and AIDS, it is very important to study the HIV incubation distribution under different conditions.

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 review of some of the parametric models for incubation period is discussed below.

Weibull and Gamma Models: Weibull and gamma models are the most commonly used for many real data applications and in particular for back calculation approach. Between the two, Weibull model is a popular candidate for HIV incubation period because of its nice properties viz., it is proportional hazard as well as accelerated failure time model. The Weibull distribution function is given by

$$F(t) = 1 - e^{-(\lambda t)\alpha} \quad \lambda > 0, \alpha > 0 \text{ and } t > 0 \qquad (1)$$

The earliest studies of Weibull incubation period have been attempted[38,49] and fitted Weibull model to study the incubation period distribution for transfusion associated AID cases

$$F(t) = 1 - e^{0.0243.t\,2.286} \qquad (2)$$

These parameter values correspond to a median incubation period of 4.3 years. The incubation period of patients infected by blood transfusion was studied[38] using Weibull distribution for Children (0-4years) and estimated the parameters as follows; α=1.9390, λ^{α}=0.1566, for others (5-59 years) as α=2.3960, λ^{α}=0.0048. Boldson et al.[50] used gamma, Weibull and log-normal models for incubation time of cohort study from San Francisco AIDS cases. The fitted Weibull model for their data is given by

$$F(t) = 1 - \exp(-0.001296\, t^{2.5}) \qquad (3)$$

The Weibull HIV incubation model used by Anderson et al. (1986) is given by

$$F(t) = 1 - \exp(-0.1190\, t^{1.9974}) \qquad (4)$$

Brookmeyer and Goedert[51] used the Weibull incubation period distributions based on the study of hemophiliacs over 20 years of age. The fitted Weibull model for their data is given by

$$F(t) = 1 - \mathrm{e} -0.0021\, t^{2.516} \qquad (5)$$

This estimate corresponds to a median incubation of 10 years.

Based on 732 HIV-positive hemophiliacs enrolled in Italian registry, Chiarotti et al.[52] estimated the incubation distributions assuming three different parametric models: uniform, uniform in three sub intervals and truncated Weibull under two approaches namely the median and median of three random values. There are altogether six different approaches to estimate the incubation time of individuals. They found that the incubation times obtained using first and third are similar. Therefore they reported only four estimates. The estimates of the four models parameters are α=2.9, 2.4, 2.9, 2.6 and λ^{α}=0.003668, 0.001039, 0.000446, 0.000845. The median incubation time are 13.5, 15, 12.6 and 13.3 respectively.

Munoz and Xu[53] using Multicenter AIDS Cohort Study (MACS) estimated this Weibull model.

$$F(t) = 1 - e^{-0.052087\, t^{1.285347}} \qquad (6)$$

The median incubation period using this model is 7.5 years. Other important studies also used the Weibull model for HIV incubation period [8,39,41,54].

The gamma distribution is another important parametric distribution used to model incubation period of HIV/AIDS. The gamma density function is

(7)

One of the earliest studies[38] that used gamma model for incubation period of HIV estimated parameter for the gamma models are as follows. K=2.669, 2.473 and α=0.911 and 11.001. The parameters estimated from gamma model[50] based on the San Francisco AIDS data gave k = 3.130 and = 5.715 years. Freund and Book[55] fitted gamma model with k = 3 (Erlang form) to the San Francisco AIDS data and the estimate obtained for the parameter σ= 2.660 years.

Log-logistic and Log-normal models: Lawless and Sun[56] used the log-logistic model for HIV incubation period, in addition Chiarotti et al.[52] used log-logistic model and generalized exponential model for their data. The log-normal distribution has been used [50,57] for HIV incubation period. Recent studies[53,58] have shown that lognormal distribution fits well than Weibull model.

The log-logistic distribution function is

$$F(t)=1-[1+(\lambda t)^{\upsilon}]^{-1} \qquad \lambda>0,\ \upsilon>0,\ t>0 \qquad (8)$$

The distribution function of lognormal is

$$F(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right) \qquad (9)$$

where

The distribution function generalized log-logistic is

$$G(t)=\frac{1}{\beta(m_1,m_2)}\int_{0}^{H(t)} x^{m_1-1}(1-x)^{m_2-1}dx \qquad t>0,\ m_1>0,\ m_2>0 \qquad (10)$$

The generalized log-logistic distribution reduces to log-logistic distribution when This has been shown[59-60] that the three parameters generalized log-logistic distribution with fits better than the log-logistic distribution for data on cancer survival analysis. Stacy[61] introduced a generalization of gamma distribution with three parameters. This model is a generalization of many survival distributions. The convolution of exponential distribution has been used as incubation model for HIV.

Longini et al.[62] used a staged Markov model to estimate the distribution and mean length of the incubation period from a cohort study of 603 HIV infected individuals who have been followed through various stages of infection. They used the generalized gamma model to describe the transition probabilities of the Markov model.

Mixture and Staging Model: One way to accommodate the variation between different groups of the population is by using mixture models. Suppose certain proportions of infected individuals have an incubation period distribution and the remaining proportion have the incubation period distribution. Then the incubation period distribution for the entire population of infected individuals is a mixture of and is given by

$$F(t) = \alpha F_1(t) + (1-\alpha)F_2(t) \qquad 0 < \alpha < 1 \quad (11)$$

Auger et al.[63] considered a mixture of two Weibull densities for the incubation period of Paedictric AIDS cases. The mixture of two Weibull distributions[49] was also used in a study of incubation period distribution of sample individuals drawn from San Francisco AIDS data.

The convolution equation for the incubation period comprising of two stages given by Brookmeyer and Liao (1990) is

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

where

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

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

Suitable changes should be made in the above formulations to account for calendar time of infection. Under staging models the incubation period is considered to be comprised of stages. The progression from time of infection to AIDS was assumed to occur in 3 stages[64]. The stage 1 refers to HIV infection without immunological abnormalities, stage 2 is the development of pre-AIDS disease and stage 3 is the development of clinical AIDS. The incubation time of an individual by definition is the total time spent on stage 1 and stage 2.Therefore different models for these two stages can be assumed. Suppose the times spent on the two stages are not independent, then the time spent on the stage 2 can be conditioned on the time spent on stage 1. Under this case Mariotti and Cascioli[26] have given the survival functions for the second stage.

Change Point and Contact Models: A change point model and contact model for incubation period of HIV was proposed[65]. Suppose the incubation time for an individual is t. It is reasonable to assume that between 0 to t, there is a time point at which the hazard of incubation changes. The point may be the time after infection when the individual realizes the threat of AIDS and seeks some kind of medication. Suppose the hazard before and after the change point is constant, then *h* (t) is given by

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

The survival function of the above change point model is given by

$$S(t) = \exp\{-\int_0^t h(x)dx\}$$

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

The other types of hazards considered were varying hazard and Weibull hazard.

A model for incubation time using the concept of contacts to immune system uses the random contact concept for formulation. Suppose at time t = 0, a member tested for HIV positive for the first time, experiences a random N number of contacts to the immune system before he shows clinical symptoms to AIDS. The number of contacts N experienced by the individual is assumed to follow a Poisson process with parameter λ(>0). We assume that the system undergoes at least one invasion before the individual become AIDS in $(t, t\Delta t)$. Using Laplace transformation and simple algebra we obtain the distribution function for an individual to become AIDS can be obtained as

$$F(t) = \frac{\lambda}{\mu} \int_{0}^{(1-e^{-\mu t})} u(1-u)^{(\lambda/\mu)-1} e^{(\lambda/\mu)u} \; du \quad (17)$$

If λ=μ=, then

$$F(t) = 1 - e^{-\lambda t} e^{1-e^{-\lambda t}} \qquad (18)$$

$$h(t) = \lambda(1 - e^{-\lambda t}) \qquad (19)$$

It can be noted that the hazard rate is increasing function of *t*. The median of the distribution can be obtained numerically by solving (17). The parameters of can be estimated by using the method of maximum likelihood.

BACK-CALCULATION ESTIMATES OF HIV/AIDS IN INDIA

The National AIDS Control Organization (NACO)[66] publishes periodic estimates of HIV/AIDS based on its own projections and also the reports of UNAIDS and WHO. NACO also publishes the reported number of AIDS cases in India as well as in states and union territories in India. In 1986, the reported number of AIDS cases was 6 and in 2002, the number increased to 19324.The estimates published by various other agencies regarding the total number of HIV/AIDS cases in India seemed to be over estimated according to the many researchers. An attempt has been made in this work to give the estimates of the minimum number of people living with HIV/AIDS in India. We have taken 2002 as the base period and estimated the AIDS cases that will be reported to the system in the next four years. The estimated figures are compared with the observed ones for 2003, 2004 and 2005.

The basic data required for back calculation methodology is the number of AIDS cases over a period of time. NACO publishes monthly updates of the reported AIDS cases for the past few years. The monthly updates of the recent years also suffer reporting delay and therefore pooled yearly reported AIDS cases assumed to be more reliable. Moreover the data compiled by NACO is at unequal time intervals. Therefore in the present study only yearly reported AIDS cases were considered for projections.

The starting point of the infection T_0 is taken to be 1980 for India. Ten incubation period distributions are used in the present projection of AIDS cases in India. The estimates of minimum size of the epidemic and future number of AIDS cases are obtained assuming different median incubation periods. For the

Table 1. Estimates of HIV incidence and projection of AIDS under logistic prevalence infection density

Incubation Model	$\hat{N}$	Incidence in 2002	*Projection of AIDS*			
			2003	2004	2005	2006
When median incubation= 8 years						
Weibull	4243202	280156	26795	43327	70060	113286
Gamma	3023853	160377	26841	43433	70282	113728
Log-logistic	4141256	301222	26797	43332	70069	113302
Log-normal	4378305	249908	26816	43375	70159	113482
Generalized Exponential	4446099	146690	26846	43444	70304	113771
Generalized Log-logistic	4052072	306432	26680	43035	69410	111944
Generalized Gamma	3624674	146772	26840	43432	70279	113721
Mixed Weibull	4219371	280259	26789	43321	70050	113280
Change Point	4287475	279968	26780	43325	70065	113289
Immune Invasion	2917156	175218	26829	43406	70226	113616
When median incubation= 10 years						
Weibull	4681218	444462	26745	43265	69925	113013
Gamma	3814500	255189	26824	43394	70199	113561
Log-logistic	4104232	488436	26766	43258	69911	112987
Log-normal	4250104	385640	26753	43326	70055	113273
Generalized Exponential	4613632	251003	26826	43398	70206	113576
Generalized Log-logistic	4706203	550776	26676	43023	69383	111885
Generalized Gamma	4074022	199972	26832	43412	70237	113638
Mixed Weibull	4730853	443742	26731	43242	69927	113007
Change Point	4732736	442925	26725	43235	69919	113020
Immune Invasion	3912703	255182	26816	43374	70158	113479
When median incubation= 12 years						
Weibull	5069775	657874	26750	43216	69820	112799
Gamma	4219708	405246	26804	43345	70095	113352
Log-logistic	4630203	750856	26736	43186	69757	112675
Log-normal	4964605	570339	26774	43275	69945	113052
Generalized Exponential	5412150	451076	26795	43325	70052	113266
Generalized Log-logistic	5144634	883083	26674	43018	69369	111856
Generalized Gamma	4708975	259962	26825	43396	70203	113568
Mixed Weibull	5070029	653421	26732	43208	69815	112810
Change Point	4908890	660570	26724	43251	69803	112893
Immune Invasion	4253183	353892	26805	43348	70099	113362
When median incubation= 15 years						
Weibull	5293098	1093251	26729	43165	69704	112560
Gamma	4362512	812829	26763	43249	69888	112935
Log-logistic	4447016	1320676	26698	43091	69550	112254
Log-normal	5210566	965345	26741	43197	69777	112713
Generalized Exponential	6078422	1185040	26714	43130	69634	112423
Generalized Log-logistic	6585507	1585456	26672	43013	69359	111834
Generalized Gamma	5215866	365798	26817	43376	70160	113483
Mixed Weibull	5246183	1081218	26732	43171	69716	112585
Change Point	5323427	1105845	26748	43210	69689	112645
Immune Invasion	4661770	534254	26793	43319	70037	113235

Table 2. Estimates of HIV incidence and projection of AIDS under logistic prevalence infection density (adjusted for under-reporting and delay in reporting)

Incubation Model	$\hat{N}$	Incidence in 2002	*Projection of AIDS*			
			2003	2004	2005	2006
When median incubation= 8 years						
Weibull	4580854	561695	66495	100801	152786	231785
Gamma	4042568	356453	66735	101312	153795	233433
Log-logistic	4752540	598002	66504	100819	152832	231671
Log-normal	4505245	512988	66601	101031	153241	232415
Generalized Exponential	3952652	330758	66756	101368	153902	233646
Generalized Log-logistic	4589124	545150	65874	99305	149699	225625
Generalized Gamma	3952681	323395	66732	101302	153772	233428
Mixed Weibull	4253141	562326	66510	100789	152764	231574
Change Point	4650115	561549	66502	100795	152799	231602
Immune Invasion	4101854	375978	66677	101172	153521	232975
When median incubation= 10 years						
Weibull	4852649	867059	66356	100475	152148	230404
Gamma	4953246	546248	66644	101102	153374	232676
Log-logistic	4858625	935054	66341	100446	152125	230375
Log-normal	4751689	765549	66495	100786	152751	231516
Generalized Exponential	4152645	538489	66623	101115	153416	232748
Generalized Log-logistic	4925448	975477	65818	99242	149575	225354
Generalized Gamma	4251006	434145	66685	101202	153566	233025
Mixed Weibull	4556245	867121	66355	100486	152164	230417
Change Point	4854662	864987	66341	100457	152149	230429
Immune Invasion	4526488	536745	66601	101004	153198	232317
When median incubation= 12 years						
Weibull	5145678	1258796	66244	100228	151654	229441
Gamma	5215674	834486	66526	100855	152898	231755
Log-logistic	5043259	1395541	66195	100129	151442	229072
Log-normal	4921486	1099115	66374	100527	152255	230605
Generalized Exponential	4658415	912247	66487	100741	152702	231411
Generalized Log-logistic	5428503	1560546	65822	99218	149514	225243
Generalized Gamma	4628457	558025	66641	101105	153399	232712
Mixed Weibull	4856477	1253741	66258	100242	151674	229486
Change Point	4958115	1261146	66205	100266	151408	229454
Immune Invasion	4726489	733543	66533	100872	152914	231798
When median incubation= 15 years						
Weibull	5891562	2051215	66124	99965	151089	228348
Gamma	5945188	1576427	66318	100389	151933	229987
Log-logistic	5518246	2370985	65987	99662	150521	227314
Log-normal	5258995	1788556	66218	100759	151524	229216
Generalized Exponential	5348926	2146419	66074	99847	150897	228005
Generalized Log-logistic	6654897	2796984	65818	99201	149451	225142
Generalized Gamma	4854981	775325	66602	101014	153187	232311
Mixed Weibull	5624189	2037019	66134	99987	151132	228438
Change Point	5548050	2065756	66111	99956	151021	228315
Immune Invasion	5218227	1091589	66465	100714	152605	231204

incubation period models Weibull, Gamma, log-logistic, log-normal and generalized exponential distribution prior estimates of their parameters were available. All these models have only two parameters and therefore one parameter was fixed based on the estimates available in the literature. The parameters of the generalized log-logistic, generalized gamma, mixed Weibull, change point and immune invasion level models were not available in the literature. Therefore parameters of these models were decided based on a simulation study.

The backcalculation estimates were obtained using the conditional likelihood approach for the multinomial likelihood with unknown sample size. Although EM algorithm[22] can also be used, it has been observed[24] that both of the approaches yield almost similar estimates. Also the conditional likelihood approach converges at much faster rate and is easy to implement in a computer algorithm. SAS IML programming was done for the actual computation of the estimates. Non-linear optimization routines were extensively used for maximization of the conditional likelihood. The calculations were repeated under logistic incidence, double exponential and root exponential for the different median incubation period distributions. The estimates of the minimum number of HIV cases (), HIV incidence for 2002 and short-term projection for next 4 years for four median times under logistic prevalence are presented in the Table 1.

Adjustments for Reporting Delays and Under-Reporting:

According to WHO Global HIV/AIDS Survey[67], the level of under reporting in the South-Eastern Asian Countries including India is about 80%. The various small studies conducted in different parts of India reported the level of under-reporting between 50% - 95 %. Vandal and Remis[68] used 75% upward adjustment for projections for Canada surveillance data. In this work we assume 90% under-reporting during 1986-87 and it decreases to 60% in 2001-02 in an exponential decay form. The upward adjustments for different years are carried out. The reporting delays were modeled by many authors. In this work we have used the approach proposed[69]. The over-dispersed Poisson regression model with modifications as adopted[68] for the reporting delays was used and the adjusted figures are used for recalculating the estimates under the above model assumptions, median incubation distributions and infection curves. If the adjusted AIDS counts are taken to the upper bound of the under reporting and delay in reporting, the figure in the estimates give the upper limit of the number of cases expected. The adjusted estimates of the minimum number of HIV cases (), HIV incidence for 2002 and short-term projection for next 4 years for four median times under logistic prevalence are presented in the Table 2.

DISCUSSION

It is generally observed that the short term projected AIDS cases do not vary much across various infection densities and incubation period distributions. 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 a 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 log-logistic and root exponential infection densities are less compared to the estimates obtained using the other four infection densities. The projected AIDS estimates obtained using the four infection densities logistic prevalence, logistic incidence, exponential and double exponential seems to be more plausible compared to the estimates obtained using the other two models since the observed AIDS cases in the year 2000 were nearly 20000. Hence based on the four infection densities and 10 incubation period distributions with four possible median incubation periods (several possible combinations for the parameters of these distributions), we infer that the projected AIDS cases in India for the years 2003, 2004, 2005 and 2006 will be approximately around 27000, 43500, 70000 and 113500 respectively. It is to be noted that these estimates are based on the unadjusted AIDS incidence data. These estimates may not be the correct number of AIDS cases that may develop in India during these periods. The exact number of AIDS cases that may develop in India 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 2002 is calculated based on the relationship between the infection density and the infection or incidence 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 2002. The minimum size of the epidemic denotes the total number of people who may ultimately become AIDS cases even if the new infections are currently stopped. Hence this figure can be taken to be the number of people living with HIV/AIDS in India and the adjusted AIDS counts can be taken to the upper bound. For all combinations of eight infection densities and ten incubation period distributions, the increase as median incubation period increases.

One limitation of this study is that the various staging models incorporating effect of therapy, different risk groups and other methods for reporting delays and underreporting were not considered due to non-availability and non-accessibility of the necessary data. AIDS incidence data among all cases reported up to Dec. 2002 and diagnosed before Dec. 2002 were adjusted for reporting delays and underreporting and the results were compared. Back-calculation was performed on the adjusted data using flexible models and conditional maximum likelihood estimation method. The models used were time variant, which allowed for testing availability and therapy effects on disease progression. The HIV incidence estimates on the cumulative incidences ascertained through simulation account for variability associated with adjustments and random components of the models. The estimates obtained are lower than the estimates given by other international bodies even at extreme model assumption. Data used consists of predominantly adult population for back-calculation. There is an increase in trend among the children. The non-parametric backcalculation approach is an alternative which uses mild assumptions for projections.

REFERENCES

1. Tan WY. Stochastic modeling of AIDS epidemiology and HIV pathogenesis. World Scientific publication, Singapore. 2000

2. Anderson RM. The role of mathematical models in the study of HIV transmission and the epidemiology of AIDS. *AIDS* 1988; 1: 241-246.

3. Hyman JM and Stanley EA. Using mathematical models to understand the AIDS epidemic. *Math. Biosciences* 1988; 90: 415-474.

4. Jager JC and Ruittenberg EJ. *Statistical Analysis and Mathematical Modeling of AIDS*, Oxford University Press, Oxford. 1988

5. Wilkie AD. An actuarial model for AIDS. *Journal of Royal Statistical Society,* 1988; *Series A,* 151: 35-39.

6. Hethcote HW, Van Ark JW and Longini IM. A simulation model of AIDS in San Francisco: I. Model formulation and parameter estimation. *Math. Biosciences 1991;* 106: 203-222.

7. Anderson RM and May RM. Understanding the AIDS epidemic. *Scientific Amer* 1992; 266: 58-66.

8. Mode CJ, Gollwitzer HE and Hermann N. A methodological study of a stochastic model of an AIDS epidemic. *Math.* Biosciences 1988; 92: 201-229.

9. Isham V. Assessing the variability of stochastic epidemic. Math. Biosciences 1991; 107: 209-224.

10. Tan WY and Xiang ZH. A state space model of HIV pathogenesis under treatment by anti-viral drugs in HIV infected individuals. *Math. Biosciences* 1999;156: 69-94.

11. Tan WY and Xiang ZH. State Space Models for the HIV pathogenesis. In Mathematical Models in Medicine and Health Science. (Eds: Horn, M.A., Simonett, G. and Webb, G.), Vanderbilt University Press, Nashville, TN , 1998. 351-368.

12. Jewell NP, Dietz K and Farewell VT. AIDS Epidemiology: Methodological issues. Birkhauser, Basel. 1992)

13. Bacchetti P, Segal M and Jewell NP. Backcalculation of HIV infection rates, *Statistical Science* 1993; 8: 82-119.

14. Brookmeyer R and Gail MH. *AIDS epidemiology: A Quantitative Approach.* Oxford University Press, Oxford. 1994

15. Wu H and Tan WY. Modeling the HIV epidemic: A state space approach. In: " ASA 1995 Proc- the Epidemiology Section". ASA, Alexdria, VA: 1995, 66-71.

16. Kalman RE . A new approach to linear filter and prediction problems. *J Basic Eng 1960;* 82: 35-45.

17. Cazelles B and Chau NP. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. Math. Bioscience 1997; 140: 131-154.

18. Healy MJR and Tillett HE. Short-term extrapolation of the AIDS epidemic. J Royal Stat Soc, Series *A* 1988; 151: 50-61

19. Anderson RM, Medley GF, May RM and Johnson AM . A preliminary study of the transmission dynamics of the human immunodeficiency (HIV), the causative agent of AIDS, IMA J Math Appl Med and Biol 1986; 3: 229-263.

20. Isham V. Mathematical modeling of the transmission dynamics of HIV infection and AIDS: A review. . J Royal Stat Soc,, *Series A 1988;* 151: 5-30.

21. Brookmeyer R and Gail MH. Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *Lancet 1986;* 2: 1320-1322.

22. Brookmeyer R and Gail MH. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. J Amer Stat Asso *1988;* 83: 301-308.

23. Ding Y. Computing backcalculation estimates of AIDS epidemic. Stat Med 1995, 14: 1505-1512.

24. Ding Y. On the asymptotic normality of multinomial population size estimates with application to the backcalculation epidemic of AIDS. Biometrika 1996; 83: 695-699.

25. Brookmeyer R. AIDS, Epidemics and Statistics. *Biometrics* 1996; 52: 781-796

26. Mariotti S and Cascilio R. Sources of uncertainty in estimating HIV infection rates by backcalculation: Application to Italian data. Stat Med 1996; 15: 2669-2687.

27. Brookmeyer R and Damiano A. Statistical methods for short-term projections of AIDS incidence. Stat Med 1989; 8: 23-34.

28. Harris JE. Reporting delays and incidence of AIDS. J Am Stat Asso 1990; 85: 915-924.

29. Evans BG and McCormick A. Completeness of reporting of acquired immune deficiency syndrome by clinicians. J Royal Stat Soc, *Series A 1994;* 157: 105-114.

30. Dueffic S and Costagliola D. Is the incubation time changing? A backcalculation approach. Stat Med 1999; 18: 1031-1047.

31. Gigli A and Verdecchia A. Uncertainty of AIDS incubation time and its effect on backcalculation estimates. Stat Med 2000; 19: 175-189.

32. Bellico R and Marschner IC. Joint analysis of HIV and AIDS surveillance data in backcalculation. Stat Med 2000; 20:2017-2033.

33. Liao J and Brookmeyer R. Empirical Bayes approach to smoothing in backcalculation of HIV infection rates. Biometrics 1995; 51: 579-588

34. De Angelis D, Gilks WR and Day NE. Bayesian projection of the AIDS epidemic. Appl Stat 1998; 47: 449-498.

35. Swaminathan S, Ramachandran R, Baskaran C et al. Risk of development of Tuberculosis in HIV infected patients. Inter J Tuber Lung Dis 2000; 4: 832-844.

36. Venkatesan P. Methods of projection of HIV/AIDS epidemic; In *Epidemiology, Health and Population*, (Eds; Anil Kumar), 2002;143-155.

37. Anbupalam T, Ravanan R and Venkatesan P. Backcalculation of HIV/AIDS in Tamilnadu; In Bio statistical Aspects of Health and Epidemiology (Eds; Pandey, C.M., Pradeep Mishra and Uttam Singh), Deptt of Biostatistics, Sanjay Gandhi Postgraduate Institute of Medical Research, Lucknow, India.2002; 232-243.

38. Medley GF, Anderson RM, Cox DR and Billard L Incubation period of AIDS in patients infected via blood transfusion. Nature 1987; 328: 719-721.

39. Kalbfleisch JD and Lawless JF. Inference based on retrospective ascertainment: An analysis of data on transfusion related AIDS. J Am Stat Assoc 1989; 84: 360-372.

40. Jewell NP. Some statistical issues in studies of the epidemiology of AIDS. Stat Med 1990; 9: 1387-1416.

41. Rosenberg PS and Gail MH. Uncertainty in estimates of HIV prevalence derived by back-calculation. Annals Epidemiol 1990; 1: 105-115.

42. Solomon PJ and Wilson SR. Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. Biometrics 1990, 46: 1165-1170.

43. Brookmeyer R. Reconstruction and future trends of the AIDS epidemic in the United States. Science 1991; 253: 37-42.

44. Longini IM, Byers RH, Hessol NA and Tan WY. Estimation of the state specific numbers of HIV infections via a Markov model and backcalculation. Stat Med 1992; 11: 831-843.

45. Rosenberg PS, Gail MH and Carroll RJ. Estimating HIV prevalence and projecting AIDS incidence in the United States: A model that accounts for therapy and changes in the surveillance definition of AIDS. Stat Med 1992; 11: 1633-1655.

46. Rosenberg PS. Backcalculation models of age-specific HIV incidence rates. *Stat Med 1994;* 13: 1975-1990.

47. Becker NG and Marschner IC. A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. Biometrika 1993; 80: 165-178.

48. Tan WY, Tang SC and Lee SR. Characterization of the HIV incubation distribution and some comparative studies. Stat Med 1996; 15: 197-220.

49. Lui KJ, Darrow WW and Rutherford GW. A model-based estimate of the mean incubation period for AIDS in homosexual men. Science 1988; 240: 1333-1335.

50. Boldson JL, Jensen JL, Sogarrd J and Sorensen M. On the incubation time distribution and the Danish AIDS data. J Royal Stat Soc, serias A 1988; 151: 42-43.

51. Brookmeyer R and Goedert JJ. Censoring in an epidemic with an Application to hemophilia-associated AIDS. Biometrics 1989; 45: 325-335.

52. Chiarotti F, Palombi M, Schinaia N, Ghirardini A and Bellocco R. Median

time from seroconversion to AIDS in Italian HIV positive hemophiliacs: different parametric estimates. Stat Med 1994; 13: 163-175.

53. Munoz A and Xu J. Models for the incubation of AIDS and variations according to age and period. Stat Med 1996; 15: 2459-2473.

54. Isham V. Estimation of the incidence of HIV infection. Spec. Phil. Trans. Roy. Soc. Lond. B 1989; 325: 113-121.

55. Freund HP and Book DL. Determination of the spread of HIV from the AIDS incidence history. Math. Biosciences 1990; 98: 227-241.

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

57. Rees M. The sombre view of AIDS. Nature 1987, 326: 343-345

58. Munoz A, Sabin CA and Phillips AN. he incubation period of AIDS. AIDS 1997, 11 (Suppl. A): S69-S76.

59. Singh K and George EO. *Generalized log-logistic model for the survival data*, Technical Report, Dept. of Mathematics, Central Michigan University, Michigan. 1987; 87(3).

60. Singh K, Lee CMS and George EO. On generalized log-logistic model for censored survival data. Biomed J 1988; 30: 843-850.

61. Stacy EW. A generalization of the gamma distribution. Ann Math Stat 1962. 33: 1187-1192.

62. Longini IM, Clark WS, Byers RH, Ward JW, Darrow WW, Lemp GH and Hethcote HW. Statistical analysis of the stages of HIV infection using a Markov model. Stat Med 1989, 8: 831-843.

63. Auger I, Thomas P, Gruttola VD, Morse D, Moore D, Williams R, Truman B and Lawrence CE. Incubation periods for paediatric AIDS patients. Nature 1988; 336: 575-577.

64. Brookmeyer R and Liao J. Statistical modeling of the AIDS spread for forecasting health care need. Biometrics 1990; 46: 1151-1163

65. Ravanan R. Statistical modeling of HIV/AIDS epidemic: A backcalculation approach. PhD thesis, University of Madras, 2004.

66. NACO. *Country Scenario: 2001-2002.* National AIDS Control Organisation, Ministry of Health and Family Welfare, Government of India, Nirman Bhavan, New Delhi 2002.

67. WHO. Report on the Global HIV/AIDS Epidemic. Geneva. 1993.

68. Vandal AC and Remis RS. Backcalculation of the HIV epidemic in Quebee. Fifth Annual Conference on HIV and AIDS Research. J Infec Diseas, (spl.B) 1995;6:351-52.

69. Zeger SL, See LC and Diggle PJ. Statistical methods for monitoring the AIDS epidemic. Stat Med 1989; 8: 3-21.