Projection of HIV infections in India: An alternative to back-calculation

C. Nagaraja Rao and T. Srivenkataramana*

Short-term projections of HIV/AIDS incidence are critical for assessing future healthcare needs of the society. This article reviews the back-calculation method as a means to reconstruct past pattern of HIV counts and to estimate future AIDS incidence. The major sources of uncertainties and difficulties in application of this method to AIDS data in India are pointed out. A methodology built around HIV seropositivity rate and sizes of different groups is proposed as an alternative to back-calculation. An illustration is provided for Indian data and projections are obtained for a few selected years.

ONE of the greatest challenges faced by the present-day world is the pandemic of AIDS, which is likely to take serious dimensions in the present century. Since the first case of AIDS reported in India in 1986, the spread of HIV is rapid and to all sections of the society across the country. It has now been estimated that one out of 100 Indians is HIV-positive[1]. Prevalence rate of over 60% is noted among sexually transmitted disease (STD) clinic attenders. Some researchers fear that India will soon have a larger number of AIDS cases than any other country in the world. Figure 1 shows cumulative number of reported AIDS counts[2]. The two predominant HIV risk groups in India correspond to heterosexuality and blood transfusion. AIDS has a social stigma and it is surrounded by many confidential issues. Understanding the dynamics of infection and its progression to clinical AIDS, is the only way to check the reckless spread of the virus.

In the study of AIDS, our interest is in understanding the current state and predicting the future path. These quantities are of substantial concern to policy makers, administrators and healthcare systems. Statisticians will have new challenges and a greater role to play in the modelling of the syndrome, developing estimation methods and in collection, analysis and interpretation of AIDS data, which are often incomplete[3]. There are large discrepancies between observed and estimated number of HIV and AIDS cases in India. Given the magnitude of the endemic and the vast geographical area with dense population in India, the projections of AIDS cases are of critical importance for assessing healthcare needs and planning interventions. However, data in India are inadequate and inaccurate for exact assessment of the size and progression of the endemic.

C. Nagaraja Rao is in the Department of Statistics, Vijaya College, Bangalore 560 004, India and T. Srivenkataramana is in the Department of Statistics, Bangalore University, Bangalore 560 056, India.
*For correspondence. (e-mail: nagarajaraoc@hotmail.com)

Estimation methods

Projections of the course of HIV or AIDS have been generally based on the following three methods:

1. Fitting a model to the incidence curve and extrapolating into the future[4,5]. The method is not efficient, as it uses less information and ignores important parameters like incubation period.
2. Modelling the dynamics of the endemic[6]. Models on AIDS depend on critical, but unverifiable assumptions and contain many unknown parameters.
3. Back-calculation: This is the main method used to reconstruct the past pattern of HIV infections and to predict the number of AIDS cases, apart from knowing the present infection status[7–11]. The method depends on three key components, viz. (a) a model for distribution of infection: (b) assumed incubation period distribution and (c) observed counts of AIDS cases over time.

The method may be outlined with the following notations:

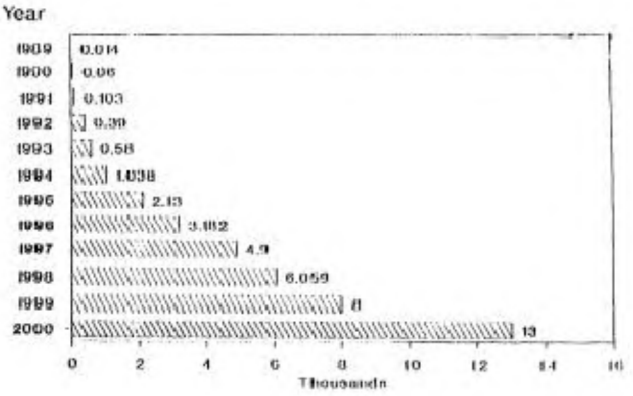
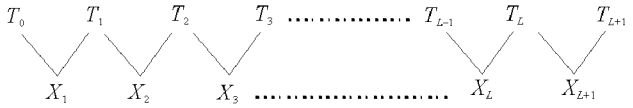


Figure 1. Cumulative number of AIDS cases (reported in India).

T_0: Beginning time of the epidemic (in India, $T_0 = 1986$);
T_0, T_1, T_2, ... T_L: Calendar dates. For example, 1986, 1987, ... 2001;
(T_{j-1}, T_j), $j = 1, 2, 3, \ldots L$: Non-overlapping intervals of time;
X_j: Number of AIDS diagnoses in jth interval;
X_{L+1}: Number infected before T_L, but not yet diagnosed.

Schematic representation of available data (X_1, ... X_L) used to estimate X_{L+1}:



Let N denote the total number diagnosed up to the year $L + 1$. Then

$$N = \sum_{1}^{L+1} X_j.$$

Here, X_1 to X_L are known from the records; X_{L+1} is to be projected.

The vector of counts of AIDS, $X = [X_1, X_2, \ldots, X_L, X_{L+1}]$ has a multinomial distribution with unknown sample size N and cell probabilities,

$$P = [P_1, P_2, \ldots P_L, P_{L+1}]$$
$$= [P_1, P_2, \ldots P_L, 1 - P^*], \text{ where } P^* = \sum_{I}^{L} P_j.$$

If P_j is the probability that a susceptible individual infected before year T_L is diagnosed in the jth interval, then

$$P_j = \int_{T_0}^{T_j} I(S)[F(T_j - S \mid S) - F(T_{j-1} - S \mid S)]\mathrm{d}s, \qquad (1)$$

where $F(.)$ denotes the incubation distribution, i.e. $F(t) = P_r$ (AIDS diagnosis occurs within time t). This distribution depends on the time of infection S. We set $F(t|S) = 0$ for all $t \leq 0$ and all S.

$I(S)$ is the infection curve which represents the pdf of infection at time S for N individuals. The basic idea of back-calculation is to use external information on $F(.)$ together with observed $[X_1, \ldots, X_L]$ to estimate $I(.)$ through some form of deconvolution. The density $I(S)$ is assumed to come from a parametric family with unknown parameter $\boldsymbol{\theta}$. The multinomial likelihood is maximized to obtain joint estimates of N and $\boldsymbol{\theta}$. An EM algorithm for incomplete data[12] can be used for this purpose. A short-term projection of cumulative AIDS incidence up to the future year T_{L+1} is then given by:

$$\sum_{1}^{L} X_j + \hat{N} \int_{T_0}^{T_L} I(S;\hat{\boldsymbol{\theta}})[F(T_{L+1} - S) - F(T_L - S)]\mathrm{d}S. \qquad (2)$$

This is an MLE and it gives a lower bound for the future AIDS counts under the assumption that no new infection occurs after T_L.

Recent developments

Brookmeyer and Liao[13] extended the back-calculation to provide the stage-specific estimates of the incidence and the prevalence of HIV infections from a two-stage AIDS incubation period. Solomon and Wilson[14] incorporated changes due to treatments such as zudovudine (AZT) and applied the same to Australian AIDS data.

Rosenberg and Gail[15] present a regression approach to back-calculation linear models such as step function or splines of the infection curves. Becker *et al.*[16] suggest non-parametric back-calculation which gives the data more opportunity to determine the shape of the estimated intensity function. Rosenberg[17] extends the method for different age groups from age-specific AIDS incidence, while Marschener[18] suggests the use of time of first positive HIV test as auxiliary, which gives extra information about the incubation period. Computations suggest that such data have the potential to significantly improve HIV incidence estimates. More recently, Bellaco and Marschner[19] used this approach and performed joint analysis of HIV and AIDS surveillance data in back-calculation. This method uses an EM algorithm with generalized additive model smoothing. The estimates obtained in the joint analysis are noted to be more efficient than those based on AIDS surveillance data alone.

Uncertainties and difficulties of back-calculation

Back-calculation methods are subject to a few major systematic uncertainties. These including the following:

Choice of infection density: Step functions, logistic growth and Poisson process are some of the models considered for this function. A standard assumption is that the unobserved times of infection for different individuals are independent. This follows from a stronger assumption that infections occur according to a non-homogeneous Poisson process with intensity $I(.)$. These assumptions may be sometimes violated. The assumed

parametric model for the epidemic density may also be incorrect.

Incubation period: This period for AIDS is random and generally very long. Estimation of this distribution is difficult because the time of infection in the risk groups is usually unknown, except in the blood transfusion-associated cases where the time of infection is retrospectively ascertained from the date of transfusion. Back-calculation is very sensitive to the choice of incubation period. There have been numerous estimates of this period from many sources. Lui *et al.*[20] use a Weibull model which produced a median time of 7.6 years. The widely used estimate obtained from Weibull regression has a median incubation of 10 years[21]. Bacchetti[22] and Bacchetti and Jewell[23] provide a non-parametric estimate of median time to AIDS as between 10 and 11 years. Thus, estimates of mean and median incubation times have lengthened steadily since an early estimate of 4.5 years. This may be the result of the use of (i) drugs like AZT which prolong the life by a few months, and (ii) Weibull model which fits steadily, the increasing hazard functions.

Inaccuracy in AIDS counts: The diagnosed AIDS counts are generally incomplete owing to non-detection of a large number of cases and reporting delays, causing under-reporting. Karon *et al.*[24] using reported US data have shown that less than 10% is reported in the month of diagnosis, about 50% within 2 months, 85% within one year and 95% within 2 years. There is evidence that the reporting delay varies across geographic regions[25].

Difficulties with Indian data

In India, at present, the following data limitations are observed:

(a) Recording and compilation system is inefficient and inadequate.
(b) Due to the sensitive nature of the syndrome, the responses may not be reliable.
(c) Lack of HIV testing laboratories in rural areas.
(d) Reporting delay and under-reporting of cases.
(e) Estimates of mean or median incubation time for Indian set-up are not yet available and estimates available elsewhere may not be suitable.
(f) Non-recording of the dates of HIV infection, diagnosis and report of AIDS cases.

Thus the use of back-calculation is not practicable in India for the time being.

Present practice in India: The HIV/AIDS sentinel surveillance system in India adopts the following methodology. A cross-sectional survey of two fixed-size samples representative of the high and low risk groups is conducted annually. The blood samples are screened for HIV-positivity by using two test procedures and its trends are monitored over a period of time. The samples for high-risk group are selected from STD clinics and IVDUs, while the samples for low-risk group are drawn from antenatal clinics. These data are used to evaluate seroprevalence in the country. This methodology inflates a local experience to the national level. Also, it ignores the sex, urban–rural divide and age structure. Hence an alternative method is desirable.

Alternative method to estimate HIV infections in India

The back-calculation method needs information on (a) HIV infection density, (b) incubation distribution, and (c) AIDS counts. Here (a) and (b) are parameters and reliable estimates for these are unlikely to be available in several countries, as is the case with India. Further, the data recording system on AIDS counts is inefficient. Thus the back-calculation, though theoretically well-founded, cannot be yet applied in the Indian set-up. As a viable alternative, we outline below a method of projection built around the HIV seropositivity rate, $r(t)$, the motivation being the availability of such information.

Use of seropositivity rate as key parameter

The rate $r(t)$ is the number infected per 1000 high-risk group individuals tested for virus infection. This includes commercial sex workers and their clients, visitors to STD clinics and IVDUs. However, commonly the (15–49) age group is considered as a major risk group. A nationwide estimate of $r(t)$ is difficult to obtain. Alternatively, this is computed by pooling data on seropositivity from different STD clinics in the country[2] and it is taken as an overall estimate $\hat{r}(t)$. Table 1 displays this information for India during the period 1989–2000.

It is common to use $\hat{r}(t)$ for projecting infections among adults in the country[26]. The proposed method for this needs projections of (i) population, and (ii) seropositivity rate. For the former, standard techniques can be adopted and for the latter we can use the past data (e.g. as in Table 1). Figure 2 exhibits the reported values (denoted by 'A') which suggest an exponential trend in $r(t)$. Accordingly, the model $r(t) = \alpha t^{\beta}$ was fitted to give expected values (shown by 'B'). A high correlation of 0.983 between 'A' and 'B' is noted.

It is desirable to validate the fitted model before using it for projection. This may be done applying one of the standard validation techniques (see Montgomery and Peck[27], chapter 3). We use dissection of $r(t)$ series for validation.

Table 1. Seropositivity rate (per thousand) in India

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
$\hat{r}(t)$	4.9	5.2	5.5	11.2	13.5	16.0	21.0	23.0	25.0	26.0	27.0	29.0

Source: Ref. 2.

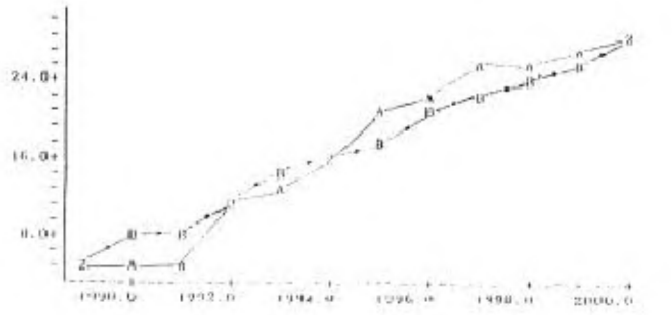


Figure 2. Seropositivity rates observed (A); expected (B).

Projection formula

The rate $r(t)$ is applied to the size of the risk group ($N(t)$), to give the number of infections ($M(t)$). Then we have

$$M(t) = N(t).r(t). \tag{3}$$

No standard technique to estimate $N(t)$ is available at present. In India, this is further complicated by the large population with high density. An upper bound on $N(t)$ is clearly the set of *all* adults. Denote the size of this group by $N_U(t)$. This leads to the estimated upper bound $\hat{M}_U(t)$ given by:

$$\hat{M}_U(t) = N_U(t) \cdot \hat{r}(t). \tag{4}$$

It is clear that the gap between $N(t)$ and $N_U(t)$ reduces with increasing infection risk. A fine tuning of this method can be effected by considering sex and urban–rural divide for the following reasons:

(i) The percentage with risk behaviour is higher for males compared to females due to the differences in their nature of activities and sexual mobility. As a result, men and women are not exposed to equal risks.

(ii) The rural population is less exposed to HIV infection compared to its urban counterpart and as a consequence the urban incidence is higher. The Experts' Group[28] has also noted these differentials. This motivates the stratification as in Table 2.

Applying the structure (4) separately to each stratum, we get

Table 2. Stratification

Stratum no. (h)	Composition	Estimated size of the risk group, $\hat{N}_h(t)$
1	Urban-Male (UM)	$\hat{N}_1(t)$
2	Urban-Female (UF)	$\hat{N}_2(t)$
3	Rural-Male (RM)	$\hat{N}_3(t)$
4	Rural-Female (RF)	$\hat{N}_4(t)$

$$\hat{M}_{Uh}(t) = \hat{N}_{Uh}(t)\hat{r}_h(t); \quad h = 1,2,3,4. \tag{5}$$

An overall estimate $\hat{M}_U(t)$ is now obtained by summing over the strata. This approach needs estimation of $r(t)$ stratum-wise, which implies that the data recording system in STD clinics is to be suitably designed. Otherwise, if a common $r(t)$ is used for the strata, a ratio of 2:1 may be applied to split $\hat{M}_U(t)$ between men and women and a ratio of 3:1 for an urban–rural division, according to the recommendation of the Experts' Group.

State-wise projection

In a diverse country like India, a state-wise projection of infections is highly desirable for the following reasons:

(a) Infections are not uniformly distributed over the states.

(b) There is a large variation in the HIV-related population characteristics among the states.

(c) Data are available state-wise.

(d) Separate projection allows effective implementation of intervention strategies, depending on the local conditions.

(e) An aggregation over states produces a national estimate as a corollary.

The state-wise strata sizes of the population can be obtained from census reports. Though the adults are at a greater risk of HIV infection, the exposure may not be the same for all the ages in the range (15–49) years. Thus a stratification by age or marital status may further improve the projection.

Validation of projected $r(t)$

The HIV seropositivity data for 1989–1997 (Table 1) exhibit an exponential trend (Figure 1). Hence a model

Table 3. Projection of $r(t)$ values and their validation

Base period	Year of projection	Projected value	95% Projection interval	Observed value
1989–1997	1998	23.0	(21.1, 42.7)	26.0
1989–1998	1999	32.5	(23.2, 45.5)	27.0
1989–1999	2000	34.5	(24.6, 48.3)	29.0
1989–2000	2001	36.5	(26.1, 51.0)	–

Table 4. HIV projections in India (figures in lakhs)

Year	Projected adult population	Projected $r(t)$	Number HIV infected			Percentage infected among adults
			M	F	T	
2000	5112	34.4*	117	58	175	3.4
2001	5260	37.78	130	68	198	3.7
2002	5413	41.10	149	75	224	4.1
2008	6231	61.60	256	128	384	6.2
2012	6626	75.70	334	167	501	7.6
2016	6913	90.30	416	209	625	9.1

*Reported $r(t)$ for the year 2000 was 27 and the total HIV-infected was 138 lakhs.

of the form $r(t) = \alpha t^{\beta}$ is fitted for projection purpose. Dissection of time series is used for validating the fitted model. This is done by first considering data for a base period of 8 years (1989–1997) and obtaining predicted value as well as a 95% confidence interval for the next year. This procedure is repeated by extending the base by a year at a time, till 2001. It is noted that the predicted interval includes the corresponding observed value in all the cases. This validates the projections. The details are in Table 3.

Illustration

The proposed projection method is applied to the Indian data on $r(t)$ with the 1991 census population in the background. Country level projections are made for selected years, beginning with 2000. The results are summarized in Table 4.

The Indian data show adult HIV incidence of 3–4 per cent, which is quite alarming. This also agrees with the projections made by a few other agencies.

Discussion

The available data on HIV infections in India represent an incomplete description of the virus spread phenomenon, which is on the whole relatively poorly understood and we should aim to bring as much knowledge as possible to bear on the problem. Back-calculation is a theoretically well-accepted method to estimate number of HIV infections from AIDS incidence in a particular population. Since the essential elements of this method are not known in countries like India, suitable modifications in the methodology are required. One such method based on seropositivity rate and population structure of the region is suggested as a simple and viable alternative. This method is made more effective by incorporating information on age, sex and urban–rural status of the individual. The application of the procedure needs computation of seropositivity rates for each group. The exposure to HIV infection may not be the same for all the ages in the range 15–49 years. Hence a further stratification by age or marital status may be useful.

In epidemiology, the standard errors for rates (or derivatives of functions) and ratios of rates are generally based on the assumption that the counts like HIV positive have either binomial or Poisson distribution. In a large group, the individual chances of acquiring infection in a short period are quite small, so that there is some plausibility to the assumption that acquiring infection is like a series of Bernoulli trials, where the total infection can be approximated by the Poisson distribution. However, when estimating seropositivity rate, we are not dealing with perfectly homogeneous group of people and the rates are seldom completely stable over the relevant time periods, so that the Bernoulli assumption will be violated to some extent. One consequence of this is a greater variance associated with the counts than would be the case with binomial and Poisson distributions. This is sometimes dealt with by using models incorporating 'over dispersion'[29]. In India, the common

practice has been to obtain the seropositivity rate by applying a sample proportion to the entire adult population. This may be improved by using the modifications suggested in the paper. Further, a 15–20% inflation may be allowed on the estimates as is usually done in the case of under-reporting[30], in view of the uncertainties of estimating a rate.

1. NACO Study, *The Times of India*, 20 September 1988.
2. National AIDS Control Organization reports: HIV surveillance in India, 1994–2000.
3. Rao, C. N. and Srivenkataramana, T., *ISMS Bull.*, March 1988.
4. Healey and Tillet, J. R., *Stat. Soc. A*, 1988, **151**.
5. Ziger, S. L. and Diggle, P. J., *Stat. Med.*, 1989, **8**.
6. Bailey, N. T. J., *The Mathematical Theory of Infectious Diseases*, 1975.
7. Brookmeyer, R. and Gail, M. H., *Lancet*, 1986, **2**.
8. Brookmeyer, R. and Gail, M. H., *J. Am. Stat. Assoc.*, 1988, **83**.
9. Brookmeyer, R. and Damiano, A., *Stat. Med.*, 1989, **8**.
10. Jewell, N. P., *Stat. Med.*, 1990, **9**.
11. Bachetti, P., Segal, M. R. and Jewell, N. P., *Stat. Sci.*, 1993, **8**.
12. Dempster, A. P., Laird, N. M. and Rubin, D. B., *JRSS-B*, 1977, **39**.
13. Brookmeyer, R. and Liao, J., *Biometrics*, 1989, **46**.
14. Solomon, H. and Wilson, *Biometrics*, 1990, **46**.
15. Rosenberg, P. S. and Gail, M. H., *Appl. Stat.*, 1991, **40**.
16. Becker, N. G., Watson, L. G. and Carlin, J. B., *Stat. Med.*, 1991, **10**.
17. Rosenberg, P. S., *Stat. Med.*, 1994, **13**.
18. Marschner, I. C., *Stat. Med.*, 1994, **13**.
19. Belleco, R. and Marschner, C. Ian, *Stat. Med.*, 2000, **19**.
20. Lui, K. J., Darrow, and Rutterford, G. W., *Science*, 1988, **240**.
21. Brookmeyer, R. and Goedard, J. J., *Biometrics*, 1989, **45**.
22. Bacchetti, P., *JASA*, 1990, **85**.
23. Bacchetti, P., Segal, R. and Jewell, N. P., *Stat. Sci.*, 1993, **8**.
24. Karon, J. M., Devine, O. J. and Morgan, W. M., *Statistical Approaches to AIDS*, Springer, NY, 1989.
25. Brookmeyer, R. and Damiano, A., *Stat. Med.*, 1989, **8**.
26. Shivlal, *Indian J. Publ. Health*, 1995, **34**.
27. Montgomery, D. C. and Pechk, E. A., *Introduction to Linear Regression*, John Wiley and Sons, 1982.
28. The Experts' Group at MAP Meeting, Kuala Lumpur, 1999.
29. McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, Chapman and Hall, 1989, 2nd edn.
30. Rosenberg, P. S. *et al.*, *J. Am. Med. Assoc.*, 1992, **268**.

ACKNOWLEDGEMENT. We are grateful to the referee for the constructive comments, which have led to substantial improvement of the contents.

Received 13 July 2000; revised accepted 3 October 2001