ORIGINAL ARTICLE

An Application of the Backcalculation Method to Estimate Past HIV Infection Rates in Malaysia

H C Ong, MSc, S H Quah, PhD, H C Low, PhD, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Pulau Pinang

Summary

The method of backcalculation estimates past HIV infection rates from available AIDS incidence data and an estimate of the incubation period. The method is used on the Malaysian data to model the AIDS epidemic because it makes use of the Malaysian AIDS incidence data which is fairly reliable and is more reflective of the trend of the epidemic as compared to the HIV infection rate recorded. An application is made on the monthly AIDS incidence data in Malaysia from January 1995 until August 1996 released by the Ministry of Health, Malaysia using the backcalculation program from Bacchetti et al[1] and the incubation period distribution from Brookmeyer[2] to generate the current HIV infection rate for Malaysia (until August 1996).

Key Words: Backcalculation, AIDS, HIV infection, Modelling, Incubation period distribution

Introduction

There are generally three methods of modelling the acquired immuno deficiency syndrome (AIDS) epidemic. At one extreme is the attempt to fit a function of calendar time such as a polynomial or other mathematically convenient curves to the AIDS incidence curve while the other extreme attempts to model the full dynamics of the transmission of the epidemic in the population providing much insight into the qualitative evolution of the epidemic and identifying the key variables that determine the future number of cases. The method of backcalculation which is intermediate between the first two methods, estimates the past HIV infection rate from the AIDS incidence data and an estimate of the incubation period distribution.

The method of backcalculation requires reliable counts of the number of AIDS cases diagnosed over time and a reliable estimate of the incubation period distribution. The method is popular because it makes use of the AIDS incidence data which represent the most readily available information on the AIDS epidemic as most national AIDS surveillance data systems record only AIDS cases. The incubation period distribution can then be applied to the estimated past HIV infection rate to project future AIDS incidence.

Materials and Methods

The basic convolution equation in backcalculation relates the cumulative number of new cases of AIDS from the time of the onset of epidemic, 0 to t (designated $Z(t)$) and the number of new HIV infections $g(s)$ at each time s since the start of the epidemic ($s=0$) through the incubation period distribution $F(u)$, where u is the time spent between the initial infection and the eventual diagnosis of AIDS. The basic convolution equation is given as:

$$Z(t) = \int_0^t g(s)F(t-s)ds.$$

From the above equation, for an individual to be diagnosed as an AIDS case by calendar time t, he or she must have been infected at some prior time s, and then have an incubation period less than $t - s$. In other words, the backcalculation method uses the above equation together with knowledge of $Z(t)$ obtained from the AIDS cases registries and $F(t)$ obtained from epidemiological studies to give information on past infection rates $g(s)$. If $F(t)$ is known, the above relationship could be inverted to express $g(s)$ for all $0 \leq s \leq t$ as a function of $Z(t)$. In general, a family of values for $g(s)$, $0 \leq s \leq t$, can be constructed which are consistent with a realization of $Z(s), 0 \leq s \leq t$.

Let $z_1, z_2, ., z_n$ be the number of AIDS cases diagnosed in the calendar time interval $[t_{i-1}, t_i)$, $i = 1, 2, ..., n$. It is assumed that individuals become infected according to a point process. Then the expected number of AIDS cases occurring during the time interval $[t_{i-1}, t_i)$ is given by

$$E(z_i) = \int_{t_0}^{t_i} g(s).\ \{\ F(t_i - s) - F(t_{i-1} - s)\ \}\ ds$$

(Brookmeyer and Gail[3])

where $F(t)$ is defined to be 0 for $t \leq 0$. By convention, we shall define calendar time 0 to be the start of the epidemic (that is $g(s) = 0$ prior to that time) and thus $t_0 = 0$.

A lower bound estimate of the expected number of AIDS cases diagnosed in a future calendar interval $[t_{i-1}, t_i)$ is obtained by projecting forward the number of individuals previously infected according to the incubation period distribution. This estimate is given by

$$\int_{0}^{t_i} g(s).\ \{\ F(t_1 - s) - F(t_{i-1} - \mathrm{s})\ \}\ ds$$

(Brookmeyer and Gail[3])

This is a lower bound since it projects only from individuals infected prior to calendar time t_n.

The method of backcalculation is used because it makes use of the AIDS incidence data which is more reflective of the trend of the epidemic. The number of HIV+ cases, on the other hand, is dependent on the test made and is unreliable as a trend. For example, a steep rise in the number of HIV+ cases may be due to the mandatory testing of all intravenous drug users in drug rehabilitation centres and increase in detection through aggressive case finding.

The definitive diagnostic criteria for diseases indicative of AIDS used in the Malaysian data are found in the Ministry of Health's[4] publication "Plan of Action for Prevention and Control of AIDS" released in May 1988. This definition is similar to the CDC's 1987 definition and the WHO definition and has been in use in Malaysia since 1988. The effects of treatment using AZT on the data is negligible because it has been available since 1987. The under reporting rate is assumed to be 10% (around 90% reported) throughout and the reporting delay, which is around 2 to 4 weeks, from the various districts to the AIDS section, Ministry of Health is assumed to be negligible.

One of the ways of dealing with the problem of delay reporting is by overlooking the most recently reported AIDS cases as was done by Brookmeyer and Gail [5]. However, omitting the most recently reported AIDS data will cause considerable loss of information and thereby resulting in an inability to detect any recent changes in the AIDS epidemic concerning whether there is an increase or a decline in the epidemic. Curran et al[6] assumed a constant delay over time and then projected future cases with a polynomial fitted model while Bacchetti et al[1] assumed a constant under reporting and delay reporting of 10% for most of the time till the then most recent AIDS data. The impact of delay reporting in Malaysia is not as great because of the size of the country as compared to the size of the United States of America. In other cases, the effects of reporting is assumed minimal for the case of Australia [7].

The monthly AIDS data as provided by the Ministry of Health, Malaysia is shown in Table I. However only monthly AIDS data from January 1995 to August 1996 were available. Prior to January 1995, only the cumulative AIDS incidence data till December 1994 is fed into the backcalculation program of Bacchetti et al [1].

A backcalculation program in Fortran from Bacchetti et al [1] is used on the Malaysian data. The program is based on the incubation period distribution from Brookmeyer[2]. The basis for using the incubation period distribution from Brookmeyer[2], which is based mainly on homosexuals, on the Malaysian data which are

Table I
Monthly AIDS data and cumulative HIV incidence recorded and estimated for Malaysia from January 1995 to August 1996

Month	AIDS cases	Cumulative HIV recorded	Cumulative HIV estimated
Cumulative till 12/94	72		
1/95	11	11175	13034.89
2/95	5	11570	13898.18
3/95	13	11868	14811.90
4/95	10	12122	15778.92
5/95	7	12368	16802.25
6/95	10	12660	17885.13
7/95	14	12991	19030.99
8/95	21	13315	20243.47
9/95	10	13642	21526.41
10/95	12	13945	22883.89
11/95	12	14226	24320.24
12/95	17	14418	25840.04
1/96	21	14824	27448.13
2/96	21	15110	29149.65
3/96	15	15471	30950.00
4/96	10	15785	32854.95
5/96	18	16029	34870.56
6/96	32	16349	30003.26
7/96	28	16774	39259.86
8/96	25	17241	41647.54

mainly intravenous drug users is because the incubation period distribution of the two cohorts are similar[8]. Mariotto et al[8] also reported the mode of transmission did appear to have a small but not statistically significant influence on the interval between the diagnosis of AIDS related conditions and AIDS. Even though backcalculation is known to be sensitive to the incubation period distribution, it is usually assumed that adequate data are available for accurate estimation and that incubation does not vary across population (Brookmeyer [2] ; Rosenberg et al [9]).

The backcalculation program used is simplified so as not to include monthly effects as compared to Bacchetti et al[1]'s because in Malaysia, we do not have the pronounced effects of the seasons. There are only 20 monthly AIDS incidences recorded in the Malaysian data as prior to December 1994, there are no consistent monthly record available in Malaysia of AIDS cases. Thus we make the most of what is available to extract the monthly trend from the data available from January 1995 until August 1996. The last column in Table I show the estimated cumulative HIV cases generated by the backcalculation program. Solomon and Wilson [7] used 5 and 6 observed annual AIDS cases for Victoria and New South Wales respectively to estimate the number of HIV+ cases. They estimated the number of individuals infected with HIV by the end of 1988 in New South Wales to lie in the range from 7,000 to 11,000 and the number infected in Victoria to lie in the range from 2,400 to 4,500.

Results

The current and past HIV infected cases are generated from the monthly AIDS incidence data and the incubation period distribution[10]. The estimated cumulative HIV+ cases generated is shown in Figure 1 and is compared with the cumulative total recorded by the Ministry of Health, Malaysia from January 1995 to August 1996 in Figure 2.

Figure 1: Shows the whole model of the estimated cumulative number of HIV infection for each month from the beginning of the epidemic until the current estimate (August 1996). On the other hand, the objective of Figure 2 is to show a comparison between the cumulative HIV cases recorded by the Ministry of Health, Malaysia and the estimated cumulative number of HIV cases obtained by the backcalculation program. The sharp difference between the cumulative number of estimated HIV infections and the recorded HIV infections could be due to delay reporting or under reporting of AIDS incidences in more recent months. The estimated HIV+ incidences generated is fairly accurate since the reported figures in developed countries are generally multiplied by two (to include unreliable cases) while in under developed/poor countries the reported figure is multiplied by 10. Malaysia is somewhere in the middle.

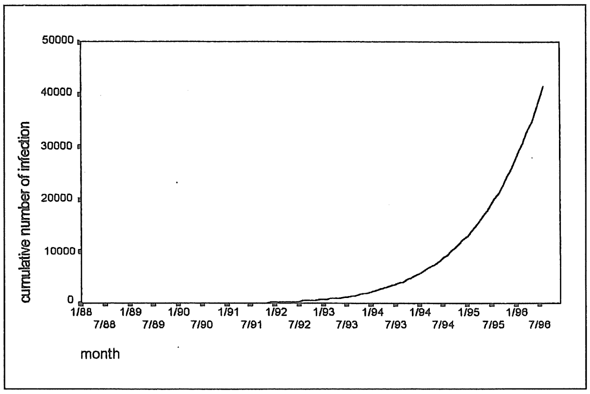

Fig. 1: Estimated cumulative number of HIV infection for each month from the backcalculation program used on the Malaysian data

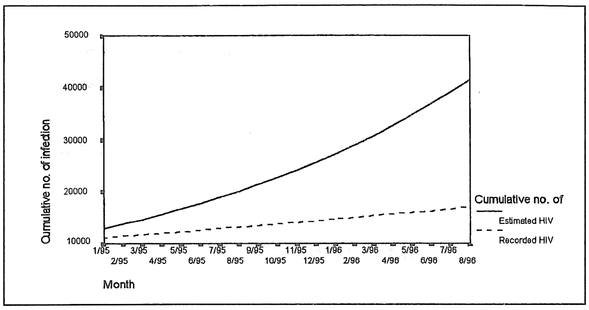

Fig. 2: A comparison between the recorded and estimated cumulative number of HIV infections in Malaysia

Beside the uncertainty resulting from inaccurate knowledge of the inputs of AIDS data, the estimates from the method of backcalculation have some inherent stochastic uncertainty. The confidence bounds in Table II reflect only the uncertainty that would be present if we actually knew the incubation and reporting inputs and the smoothness weights.

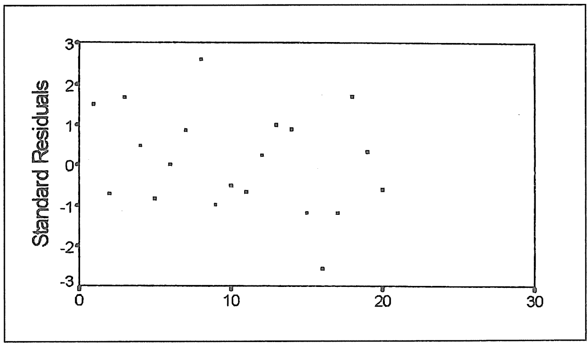

Fig. 3: Standardized residuals for the fit to the Malaysian AIDS cases from the backcalculation model

Table II
Pointwise 95% lower and upper confidence bounds of the estimated HIV+ incidences

Month	Cumulative HIV estimated	Lower bound	Upper bound
Cumulative till 12/94			
1/95	13034.89	11726.83	14342.95
2/95	13898.18	12586.29	15210.07
3/95	14811.90	13500.01	16123.79
4/95	15778.92	14467.03	17090.80
5/95	16802.25	15490.36	18114.13
6/95	17885.13	16573.25	19197.02
7/95	19030.99	17719.11	20342.88
8/95	20243.47	18931.58	21555.35
9/95	21526.41	20214.52	22838.29
10/95	22883.89	21572.00	24195.77
11/95	24320.24	23008.36	25632.13
12/95	25840.04	24528.16	27151.93
1/96	27448.13	26136.25	28760.02
2/96	29149.65	27837.76	30461.53
3/96	30950.00	29638.12	32261.89
4/96	32854.95	31543.06	34166.84
5/96	34870.56	33558.67	36182.45
6/96	30003.26	35691.38	38315.15
7/96	39259.86	37947.97	40571.74
8/96	41647.54	40335.66	42959.43

Table III summarizes the model fit for the Malaysian AIDS data and Figure 3 shows a fairly good fit has been obtained and that little heterocedasticity is present in the standardized residuals obtained for the 20 months.

Discussion

The error or accuracy involved is primarily due to the inadequate monthly incidence data available. However, an obvious result from the generated data is that the HIV/AIDS epidemic in Malaysia is in its early stages with its rapid increase in the number of infected. In comparison, there is a slowdown in the increase of the number of infected cases in the US and UK where the AIDS epidemic had begun earlier (Bacchetti et al [1] ; Brookmeyer [2] ; Chariotti [11]).

There are several potential sources of error and limitations underlying the assumptions of the

Table III
Actual reported AIDS counts fitted values and standardized residuals for 20 months

Month	Recorded AIDS cases z_i	Estimated AIDS cases z_i	Standardized Residuals $(z_i - z_i) / \sqrt{z_i}$
Cumulative till 12/94	72	77.77	
1/95	11	7.03	1.4973
2/95	5	6.86	-0.7102
3/95	13	8.21	1.6717
4/95	10	8.58	0.4848
5/95	7	9.57	-0.8308
6/95	10	9.99	0.0032
7/95	14	11.13	0.8603
8/95	21	11.99	2.6020
9/95	10	12.49	-0.9875
10/95	12	13.89	-0.5071
11/95	12	14.46	-0.6601
12/95	17	16.06	0.2522
1/96	21	17.25	1.0062
2/96	21	17.33	0.8816
3/96	15	19.88	-1.1723
4/96	10	20.63	-2.5535
5/96	18	22.86	-1.1674
6/96	32	23.71	1.7025
7/96	28	26.24	0.3436
8/96	25	28.09	-0.5830

methodology. Firstly, the parametric model for the HIV infection rate provides no information about future incidence rates as it only attempts to estimate the historical infection rates. It is for this reason that back-calculation is referred to as a method for estimating the minimum size of the epidemic (Brookmeyer and Gail [5]). Secondly, there is little information about the recent infection rate because of the long incubation period. However, short term projections of AIDS are reliable because such projections depend more strongly on the infection rate in the distant past than on the recent ones. The third limitation is that the incubation distribution is not known precisely although it is assumed known. The incubation distribution may be different for different subgroups of infected individuals with age as a cofactor of disease progression. Fourthly, the assumption of independence between the calendar date of infection and incubation period implicit in the convolution equation would be violated if cofactors of disease progression are identified that are more prevalent among those infected earlier (or later) in calendar time. Also, a smaller F needs to be compensated by a larger g in order to fit the cumulative AIDS incidence series. The fifth potential source of limitation is in the inaccuracies in the AIDS incidence data over time. It could be clouded with issues like reporting delays and changes in AIDS definition.

The application of backcalculation to data is useful in several respects. Firstly, the backcalculation method provides a simple conceptual framework for relating the incubation distribution with the AIDS incidence data and the infection rate. Secondly, backcalculation leads to short-term projections of AIDS incidences that are robust to changes in the incubation distribution (Brookmeyer and Gail [12]; Taylor [13]). Thirdly, although backcalculation estimates of cumulative infections are known to be highly sensitive to the choice of F, plausible ranges of estimates of g from backcalculation for the number infected in the United States based on data through mid-1987 were in broad agreement with estimates based on surveys in selected populations (Brookmeyer and Gail [12]; Taylor [13]).

References

1. Bacchetti P, Segal MR, Jewell NP. Backcalculation of HIV infection rates. Statistical Science 1993; 8(2): 82-119.

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

3. Brookmeyer R, Gail MH. AIDS epidemiology: A quantitative approach. Oxford University Press. 1994.

4. Ministry of Health, Malaysia. Plan of action for the prevention and control of AIDS. 1988. Malaysia: 1988:23.

5. Brookmeyer R, Gail MH. Minimum size of the acquired immuno deficiency syndrome (AIDS) in the United States. Lancet 1986; ii:1320-22.

6. Curran JW, Morgan MW, Hardy AM, Jaffe HW, Darrow WW, Dowdle WR. The epidemiology of AIDS: Current status and future prospects. Science. 1985; 229:1352-57.

7. Solomon PJ, Wilson SR. Accommodating changes due to treatment in the method of back projection for estimating HIV infection incidence. Biometrics (Keiding N., ed) 1990; 46:1165-70.

8. Mariotto AB, Mariotti S, Pezzotti P, Rezza G, Verdecchia A. Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: A comparison with malehomosexuals. American Journal of Epidemiology 1992; 135(4):428-37.

9. Rosenberg PS, Gail MH. Backcalculation of flexible linear models of the human immuno deficiency virus infection curve. Journal of the Royal Statistical Society. Series C. 1991; 40: 269-82

10. Ong HC. A study of statistical methods in AIDS modelling. M.Sc. thesis. Universiti Sains Malaysia 1997.

11. Chariotti F, Palombi M, Nicola S, Ghirardini A, Prospero L. Effects of different parametric estimates of seroconversion time on analysis of progression to AIDS among Italian HIV positive haemophiliacs. Statistics in Medicine 1992; 11:591-601.

12. Brookmeyer R, Gail MH. A method for obtaining short-term projection and lower bounds on the size of the AIDS epidemic. Journal of the American Statistical Association 1988; 83(402): 335-65.

13. Taylor JMG. Models for the HIV infection and AIDS epidemic in the United States. Statistics in Medicine 1989; 8:45-58.