HIV Prevalence: A Comparison between Canada, the United States, and Mexico

by
B. D. Aggarwala,
Department of Mathematics and Statistics
University of Calgary,
Calgary, Alberta, T2N 1N4, Canada
aggarwal@math.ucalgary.ca

Abstract: We apply the technique of Expectation Maximization (EM) to estimate the HIV prevalence in the three countries in North America namely, U.S.A., Canada and Mexico. We conclude that if we can estimate the recent HIV incidence in a country, then HIV prevalence can be rather accurately estimated. Our results lead us to think that while the official estimates for Canada and Mexico may be fairly accurate, those for the United States are rather uncertain.

Keywords: HIV Prevalence, EM Algorithm, Canada, the United States, Mexico.

2000 Mathematics subject classification: 92D25

1. Introduction:

In the beginning, HIV infection manifests itself in the form of a mild fever, diarrhoea and/or rash which go away after a few weeks. After these initial symptoms, the patient stays asymptomatic for a number of years after which some patients develop symptoms of persistent generalized lymphadenopathy (PGL) and then they may progress to develop AIDS [1]. While an AIDS patient is very sick and goes to a doctor, the symptoms of primary HIV infection are so mild that a patient may simply take a Tylenol and hope that the symptoms will go away. They go away after a few weeks anyway. These facts require only a minor reluctance on the part of a patient for him/her not to report these symptoms to the doctor. His/her financial condition, hesitation to go to a doctor, lack of medical insurance, inaccessibility to the doctor (if the doctor is five miles away in the city, for example) and many other factors may provide this 'minimum amount of hesitation' and the patient may not report these symptoms to a doctor. That a large number of HIV positive patients are unknown to the medical system is borne out by a survey report published by the Center for Disease Control and Prevention (CDC) in Atlanta, Georgia, U.S.A. in 2003 which stated that "the vast majority of young gays and bisexuals with the disease, 90 percent of blacks, 70 percent of Hispanics and 60 percent of whites, did not know that they were HIV positive" [2]. When after a number of years, an HIV positive person comes to the doctor with AIDS, the doctor wonders as to how many people like him/her there are in the community at large, and for how long has this patient been HIV positive. Counting of HIV positive people in a community is, therefore, more of a guesswork than a science [3].

The situation in Canada is no better where the estimates of HIV prevalence

vary from a low of 26000 to a high of 86000 at the end of 2003 [4]. In this paper, we shall present a mathematical model to estimate the HIV prevalence in Canada, in the United States, and in Mexico at the end of 2001 and show that this prevalence is, most likely, much closer to the lower limit than to the upper one for Canada. This shows that the situation in Canada is very good indeed, perhaps the best of all the countries in North America. We suspect that this is because of the Canada Health Act, which act encourages people to go to a doctor if they have a mild case of fever, diarrhoea and/or rash which may turn out to be the early symptoms of HIV positivity. Our mathematical model is based on the method of back calculation which calculates HIV incidence from the data on AIDS incidence. We start to outline the method of back calculation [5].

2. The Back Calculation:

For an HIV positive patient, the time to development of AIDS may be estimated by the so called Weibull distribution. In the beginning, the HIV infection manifests itself as mild fever and/or rash which disappear in a few weeks, after which the person remains asymptomatic for a while, and then, after a few years, typically eight or nine but sometimes less, the person develops AIDS. This situation is typified by the following Weibull distribution [3]

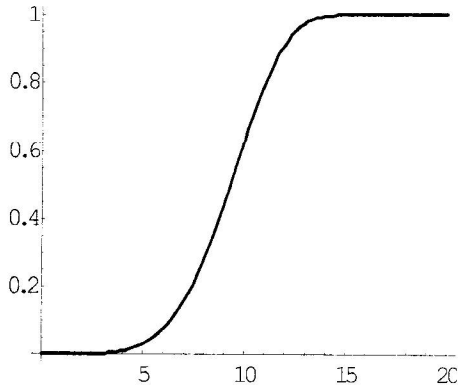

Fig.1: The graph of $F(t) = (1 - \mathrm{Exp}[-(\lambda_1 t)^{\theta_1}])$ **for** $\lambda_1 = .1$ **,** $\theta_1 = 5$**. The unit of t is one year.**

The first few values of F[t] are given by
{F[0],F[1],F[2],F[3],F[4],
F[5],F[6],F[12]}=
{0, 9.99995×10\^-6, 0.000319949, 0.00242705, 0.0101877, 0.0307668, 0.0748136, 0.916951}..... (1)

The graph in Fig. 1 gives the cumulative probability distribution of development to AIDS. The first few values say that out of a hundred patients say, most stay asymptomatic for about four years, after which some patients start developing AIDS, and by about twelve years, most of them have done so. It follows that if g(s) is the number of people who become HIV positive *at* time s, and x(t) is the number of patients who have developed AIDS *by* the time t, then x(t), the total number of patients who are diagnosed with AIDS in the time interval (0, t) is given by the so called Back Calculation Method, namely [5],

$$x(t) = \int_0^t g(s)\,F(t-s)\,ds \ldots\ldots (2).$$

where F(t-s) is the probability that the patient has developed AIDS in t-s years ***or less.*** We shall take g(i) = p[i] for i = 0, 1, 2, 3..... and join these p[i]'s by straight lines, thus making g(s) a spline of order one, and then integrate equation

(2) numerically to estimate the number of AIDS patients at any time t from a given g(s). We used Simpson's method with a sufficiently large number of sub-intervals. Numerical integration of this equation in one particular case gives

x(10) = 0.+ 0.4476556 p[1] + 0.282979 p[2] + 0.15848929 p[3] + 0.0778128 p[4] + 0.0327134 p[5] + 0.0112405 p[6] + .00288404 p[7] + 0.000459792 p[8] + 0.0000299936p[9] + 2.37299*^-7 p[10].....(3)

where x(10) is the number of AIDS cases at the end of year ten. This equation says that, as an example, 28.2979 per cent of the people who got infected in year two developed AIDS by year ten. Similarly for the other coefficients.

Special attention should be paid to the very small coefficients towards the end of equation (3). It says that more than one million people infected in the year ten, will contribute less than one person to the count of AIDS patients in the year ten. It follows that it is particularly difficult to estimate HIV incidence with the help of AIDS incidence from this equation, particularly for later years.

Now let N be the total number of individuals infected up to the end of the year L in Canada. We divide both sides of equation (2) by N and obtain

$$p_j = \int_0^j p(s)\,[F(t_j - s) - F(t_{j-1} - s)]\,ds \;..\; (4a)$$

A similar equation for AIDS deaths gives

$$q_j = \int_0^j p(s)\,[K(t_j - s) - K(t_{j-1} - s)]\,ds \;..\; (4b)$$

where K(s) is the kernel associated with AIDS deaths [3].

Since $p(s) \geq 0$ and

$$\int_0^L p(s)\,ds = 1$$

we may look upon p(s) as the probability density function of HIV infection over the interval [0, L], where we take L to be the year 2001. The coefficients in equation (3) suggest that AIDS cases and AIDS deaths start appearing a few years after HIV positivity has made its appearance in a society, and in a previous paper [3] we have suggested that HIV positivity made its appearance in North America in 1973. We shall continue to do so. The suggested year of 1973 is necessarily very approximate because the early history of HIV is shrouded in mystery [6]. Now in equation (4a), p_j is the probability that an individual infected sometime before the j-th year is diagnosed in the j-th year. The aim of the paper is to estimate p(s), i.e. g(s) for a fixed N.

It may be argued that the values of λ_1 and θ_1 that we have used are not supported by experimental evidence. However, the over riding parameter here appears to be the value of incubation period and not the values of λ_1 and θ_1 individually as shown by the following graph.

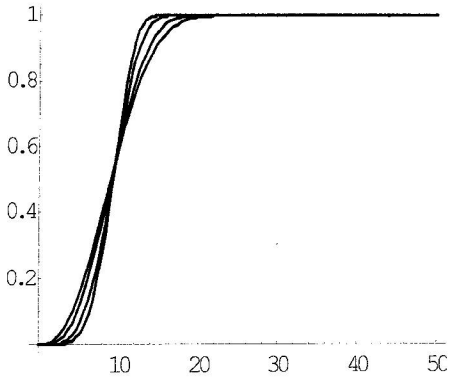

Fig. 2: Graphs of F(t) for θ_1 = 2.5, 3, 4, and 5 keeping the value of the incubation period at 9.18 years.

We shall now develop the expectation maximization (EM) algorithm based on eq. (4).

3. The EM Algorithm: This algorithm is an iterative process and we start by assuming some arbitrary values of p[i]'s. At the m-th stage of the iteration, let these values be denoted by $p^m[i]$. As stated before, we let the total number of people who got infected with the HIV virus in the society by the end of 2001, be denoted by N. We shall assume that all these people will eventually develop AIDS and then die.

We now change the time scale slightly. Since the AIDS cases started appearing only in 1980 in Canada, we shall take this year to be the year one and the years previous to this as the years 0, -1, -2, -3, and so on. Now L = 22. Let the number of people who develop AIDS in the years 1980, 1981, ...2001, *and in later years*, be denoted by $x_1, x_2, \ldots\ldots x_{22}$ and N-xdot, respectively where xdot = $x_1 + x_2 + \ldots\ldots x_{22}$. Similarly, we shall denote the number of AIDS deaths in these years by $y_1, y_2, \ldots\ldots y_{22}$, and N-ydot, where ydot = $y_1+y_2.+\ldots\ldots y_{22}$. We also have $x_0 = x_{-1} = x_{-2} = x_{-3} = x_{-4} = x_{-5} = x_{-6} = 0$. Similarly for the y's. We shall assume that the numbers $x_1, x_2, \ldots x_{22}$ and N-xdot, and the numbers $y_1, y_2, \ldots\ldots y_{22}$ and N - ydot, each have a multinomial distribution with sample size N and cell probabilities p_1 , p_2, ...p_{22}, 1-pdot and q_1, q_2, ...q_{22}, 1-qdot where pdot = $p_1 + p_2 +$ .....$+p_{22}$, and qdot= $q_1 + q_2 +$ .....$+q_{22}$. We take these multinomial distributions to be independent of each other. The likelihood function LF of the joint distribution is given by

LF =

$$\left(\frac{N!}{x_1!x_2!x_3!\ldots\ldots x_{22}!(N-xdot)!}\, p_1^{x_1} p_2^{x_2} p_3^{x_3} \ldots\ldots p_{22}^{x_{22}} (1-pdot)^{N-xdot}\right)$$

$$\left(\frac{N!}{y_1!y_2!y_3!\ldots\ldots y_{22}!(N-ydot)!}\, q_1^{y_1} q_2^{y_2} q_3^{y_3} \ldots\ldots\right.$$

$$\left. q_{22}^{y_{22}} (1-qdot)^{N-xdot}\right) \ldots\ldots\ldots (5a)$$

We shall now maximize this likelihood function LF (hence the name of the algorithm). Since Log is a monotonically increasing function, in order to maximize LF, we may maximize Log(LF). We have

$$Log(LF) = 2Log(N!) - \sum_{i=1}^{22} Log[x_i!] - Log[N-xdot] +$$

$$\sum_{i=1}^{22} x_i \log[p_i] + (N-xdot)\, Log[1-pdot] - \sum_{i=1}^{22} Log[y_i!] -$$

$$Log[N-ydot] + \sum_{i=1}^{22} y_i \log[q_i] + (N-ydot)\, Log[1-qdot]$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5b)$$

Differentiating the right hand side of this equation w.r.t. N, and equating to zero, we get [7]

E[N] = (xdot+ydot)/(pdot+qdot).....(6).

Differentiating equation (5b) w.r.t. x_j gives E[N] =x_j/p_j. Similarly, we get E[N] = y_j/q_j. Summing these equations over j, we again get equation (6).

The E-M algorithm in this case rests on the observation that the complete data sufficient statistics for this case is the number of people who got infected in any given year i and diagnosed in any later year j, and also the number of people who got infected in the year i and died in any later year j [7]. As an example, in equation (3), the term containing p[8] in (x[10] - x[9]) is the number of people who got infected in the year eight and got diagnosed in the year ten. We denote this number by x_{ij} for {i, j} = {8, 10} and the corresponding probability by the symbol p_{ij}, and assume that the numbers x_{ij} have a multinomial distribution with

sample size N and cell probabilities p_{ij}. Similarly for the numbers y_{ij} and the probabilities q_{ij}. We have

P_{ij} = Coefficient of p[i] in p_j..... (7a),

and similarly

q_{ij} = Coefficient of p[i] in q_j.....(7b),

Notice that x(j) is different from x_j. Notice also that $j \geq i$ in as much as a person infected in the year i cannot be diagnosed (or die) before the year i. We also notice that while i runs from 1 to L, j runs from 1 to L+1 where L is the year 2001, and where we take L+1 to be infinity [7]. This is because a large number of people infected till 2001 will be diagnosed (or die) after 2001. We have now discretized the problem and are calling L as the year 2001, but this should be no cause for confusion. Following Brookmeyer and Gail [7], we write

$E[x_{ij}] = x_j (p_{ij}/p_j)$.....(A1)

and similarly

$E[y_{ij}] = y_j (q_{ij}/q_j)$.....(A2)

Where p_{ij} and p_j are computed from equations (7a) and (4a), and q_{ij} and q_j are computed from equations (7b) and (4b) respectively, using the current estimates of the numbers p[i]. We notice that j runs from 1 to L+1 and, in equations (A1) and (A2), for j = L+1, we write x_j = N - xdot, y_j = N - ydot, p_j = 1- pdot and q_j = 1- qdot. We now need to maximize the log likelihood function, a function like the one given in equation [5], but with x_{ij} and p_{ij}, instead of x_j and p_j, and with y_{ij} and q_{ij}, instead of y_j and q_j, for a fixed N and conditional on x_{ij} and y_{ij}, so that we need to differentiate the quantity

$$\Sigma_{j=1}^{23} \Sigma_{i=1}^{22} x_{ij} \log[p_{ij}] + \Sigma_{j=1}^{23} \Sigma_{i=1}^{22} y_{ij} \log[q_{ij}]$$

w.r.t. p_{ij} and q_{ij}. We now divide all the parameters p[i] by N and call the new parameters by the names pp[i] so that the function p(s) is related to the parameters pp[i] in the same manner as the function g(s) is to the parameters p[i]. Since the coefficients of p[i] in equation (3) are functions of F[t], and are therefore known, we only need to differentiate

$$\sum_{j=1}^{23} \sum_{i=1}^{22} x_{ij} \, Log\,(p[i])$$

or for a constant N, the function

$$\sum_{j=1}^{23} \sum_{i=1}^{22} x_{ij} \, Log\,(pp[i])$$

w.r.t. pp[i], and similarly the function

$$\sum_{j=1}^{23} \sum_{i=1}^{22} y_{ij} \, Log\,(pp[i])$$

w.r.t. pp[i] (this is because we consider the quantities p_{ij} and q_{ij} to be independent).

We write

$$z_{1i} = \sum_{j=1}^{23} x_{ij},$$

$$z_{2i} = \sum_{j=1}^{23} y_{ij}$$

and notice that $\Sigma_{i=1}^{22} z_{1i} = \Sigma_{i=1}^{22} z_{2i} = N$, and that $k_1 pp[1] + k_2 pp[2] + k_3 pp[3] + \ldots\ldots\ldots + k_{22} pp[22] = 1$, where $k_1 = k_2 = k_3 = \ldots\ldots\ldots k_{21} = 1$ and $k_{22} = .5$, so

that we obtain $z_{1i}/(k_i\, pp[i]) = z_L/(k_L\, pp[L])$, or

$$z_{1i} = (z_L/(k_L\, pp[L]))\,(k_i\, pp[i]) \ldots\ldots (9)$$

Summing up again w.r.t. i, from 1 to L, we get

$pp[i] = Z_{1i}/(Nk_i)$.....(B1)

and similarly

$pp[i] = Z_{2i}/(Nk_i)$.....(B2),

so that we finally write

$pp[i] = (Z_{1i}+Z_{2i})/(2Nk_i)$.....(B),

which is our next approximation to pp[i]. Multiplication by N gives us the next approximation to p[i] which is $p^{m+1}[i]$. We now cycle from (B) to (A1) and (A2) and back. Keep in mind that the quantity pp[i] is obtained from p[i] by simply dividing by N. We shall now apply this algorithm to the particular cases of Canada, U.S.A. and Mexico.

4. Estimates for Canada: We start with the observation that the number of new reported HIV positive cases in Canada were 2791, 2512, 2333, 2233, 2111, 2174, 2495, 2498, and 2529 in the years 1996 to 2004 respectively [8]. Thus the average number of HIV positive cases were about 2359 over the past six years preceding 2001, about 2272 over the past five years preceding 2001, and are rising somewhat lately.

We considered the reported numbers of both AIDS cases and AIDS deaths to be reliable and based our calculations on both these numbers. As explained in [3], just as the incubation time of HIV (time it takes for an HIV positive person to develop AIDS) can be modeled by a Weibull distribution F(t) with parameters λ_1 and θ_1 as in the previous section, so can the survival time of AIDS patients (amount of time an AIDS patient lives after developing AIDS), be modeled by G(t), which is also a Weibull distribution $G(t) = (1-\mathrm{Exp}[-(\lambda_2 t)^{\theta_2}])$, but with different parameters, λ_2 and θ_2. This gives $K(t) = \int_0^1 tG((1-w)t)(F'(wt))\,dw$ as the kernel associated with AIDS deaths in equation (4b). Also, we take into account the fact that there have been two significant inventions in this field, one when AZT was discovered in mid eighties, and the other one when highly active antiretroviral therapy (HAART) came along in mid nineties. As in [3], we assume that the first HIV positive case occurred in North America in 1973 and divide the time from 1973 to 2001 into three parts with $0 \le t \le t_1$, $t_1 \le t \le t_2$ and $t \ge t_2$ with $t_t = 14$ and $t_2 = 24$ thus putting these path breaking inventions in 1986 and in 1996 respectively. In these three time intervals, we take the values of λ and θ as $\{\lambda_1, \theta_1, \lambda_2, \theta_2\} = \{0.1, 5, 0.45, 5\}$, $\{\lambda_3, \theta_3, \lambda_4, \theta_4\} = \{0.09, 4, .3, 4\}$, and $\{\lambda_5, \theta_5, \lambda_6, \theta_6\} = \{.06, 2, .06, 2\}$, thus taking the values of incubation times and survival times in these three periods as (9.18169, 2.04047), (10.0711, 3.02134) and (14.7701, 14.7701) respectively. These values of incubation periods and survival periods are rather arbitrary, but we feel that they represent realistic values considering the fact that HAART has made a very significant contribution to the area of HIV/AIDS, and that today an HIV positive person may live with the aid of HAART for as long as he would live otherwise without HIV. The corresponding model is given in [3] and will not be reproduced here.

To start the E-M algorithm in this case, we start by taking p[i]'s as all equal numbers. We noticed that as we cycled from equations (A) to (B) and back, there was no convergence. So, we adopted the following scenario. We noticed that with successive iterations, the difference between N and the right hand side of equation (6) oscillated from positive to negative (or vice-versa) and back. We tried the above iteration for N = 35000, 40000, and 50000, and in each case, we looked for those values of p[i]'s when the difference between the calculated values of AIDS cases plus AIDS deaths and the actual value of this number (as reported to Health Canada) was as small as we could make it. The results of our investigation are given in Fig. 3

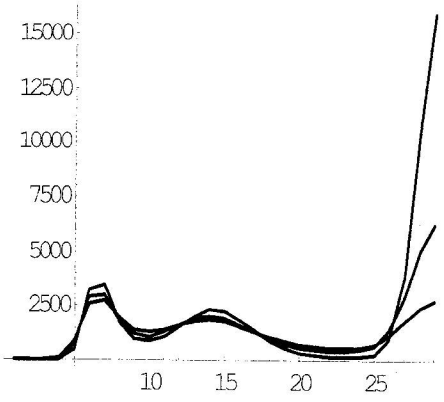

Fig 3: Graphs of p[i] 's from i = 1 to L for N = 35000, 40000 and 50000 in ascending order (on the right). The year 1973 is the year one.

Notice that for large N (50000 in our case), the agreement (i.e. approximate satisfaction of eq.(6)) is obtained by simply pushing a large number of infections to the right where they have an insignificant contribution to the number of AIDS cases and AIDS deaths, while the values of p[i]'s are not very different from other cases in the beginning. For N = 35000, the average incidence turns out to be 1581 over the past 6 years (preceding 2001) and 1775 over the previous five years while for N = 40000, the corresponding figures are 2784 and 3245. Considering that the reported numbers are 2359 and 2272 and assuming that there is no underreporting (because of Canada Health Act), we conclude that the figure of N = 35000 is a bit too low and N = 40000 is a bit too high. The truth lies somewhere in between. We report the results for N = 37000 for which the average incidence turns out to be 2054 over the previous six years (preceding 2001) and 2350 over the previous five. We consider these numbers to be sufficiently close to the actual numbers. Of course, if the HIV incidence has been much larger than what has been reported to Health Canada, then perhaps N = 40000 or even higher may be the right value of N, but we doubt very much that it is more than 40,000. Taking into account the number of AIDS deaths, which was 12813 at the end of 2001, this leaves an HIV prevalence of somewhere between 24,000 and 27,000 which is towards the lower end of the UNAIDS estimate of somewhere between 26000 and 86000. We would also doubt the published estimate of Health Canada, which is 56000 in 2002 [9].

To get another estimate, the author wrote to Health Canada asking them as to how many people in Canada were being treated with ARV's. Because of Canada Health Act, this number should be very close to the number of HIV positive people in Canada. Their reply was [10] "For the number taking ARVs, we only have a rough guesstimate that approx. 20,000 HIV+ people in Canada are on

ARVs". This reply is very much in line with our present estimate.

As indicated above, we stopped the iteration in each case when the difference between the calculated number of AIDS cases plus the AIDS deaths and the actual ones (as reported to Health Canada) was sufficiently close (i.e. equation (6) was satisfied as best as we could). For N = 37000, the difference between these two numbers is mapped in Fig. 4 and we report the results when the curve in this figure crosses the x-axis after about 200 iterations. Of course, in a discrete number of iterations, we do not expect the two numbers to exactly coincide. Notice that the curve in Fig. 4 also crosses the x-axis after a couple of iterations but we do not report the results at this point because we feel that the assumed initial flat HIV incidence is still settling in, and also because the results of no single iteration were sufficiently close to the x-axis in the beginning. It should be kept in mind that the results here are isolated points which have been joined together. For the case we have reported, these two numbers (divided by pdot plus qdot) were 37000 (which is the assumed value of N) and 36999.2

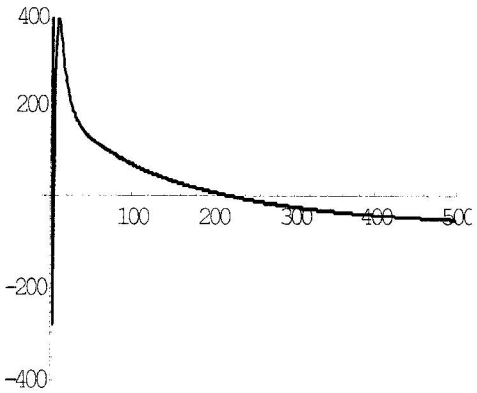

Fig.4: Difference between the calculated number of AIDS cases plus the AIDS deaths and the actual ones at the end of 2001 for Canada for N = 37000 for the first five hundred iterations. We have given the results for the case when the curve in this figure crosses the x-axis after about 200 iterations.

The calculated incidence rate for N = 37000 is given in Fig. 5

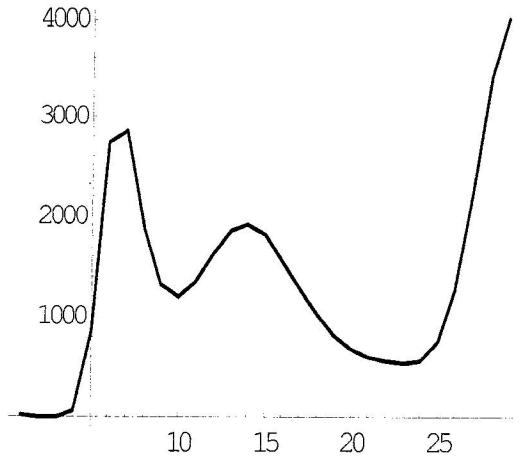

Fig 5: HIV incidence rate according to our model for N = 37000. The year 1973 is the year one. Notice the sharp rise in the infection rate in the late seventies and early eighties and then again in late nineties.

The agreement between the actual cumulative number of AIDS cases, as reported to Health Canada, and the ones calculated according to this model is given in Fig. 6.

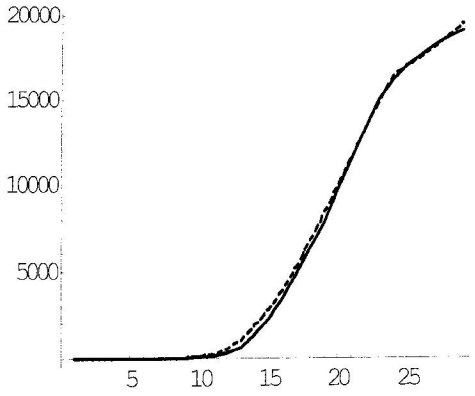

Fig 6: Comparison of actual cumulative AIDS cases (solid line)

with the calculated ones in Canada. The year 1973 is the year one.

The agreement between the actual cumulative number of AIDS deaths, as reported to Health Canada, and the ones calculated according to this model is given in Fig. 7.

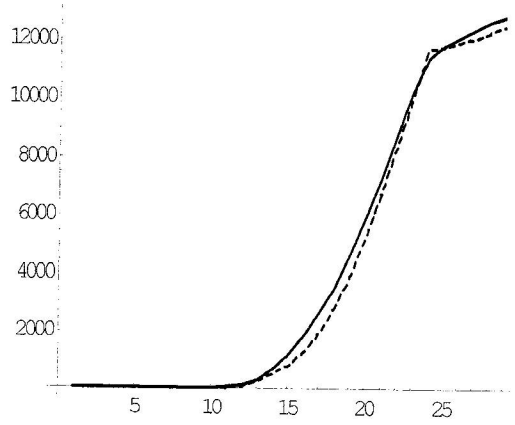

Fig 7: Comparison of the actual cumulative AIDS deaths (solid line) with the corresponding calculated numbers in Canada. The year 1973 is the year one.

We also compare the actual yearly AIDS deaths with the calculated ones in Fig. 8. The AIDS deaths after 2001 were calculated by assuming that all the p[i]'s after that year are zero [7].

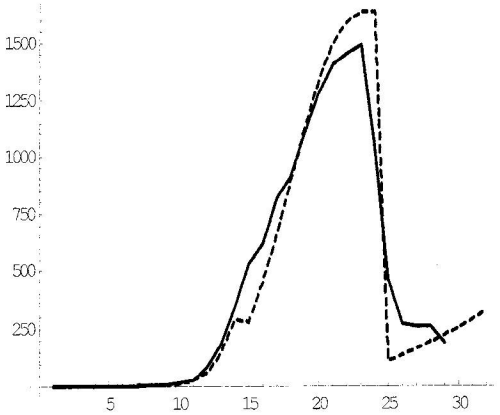

Fig. 8: Comparison of actual AIDS deaths (solid line) with the calculated ones in Canada on an yearly basis according to our model. The year 1973 is the year one.

It would be interesting to know whether our results would have been very different from each other if we would have based our analysis on AIDS cases only or on AIDS deaths only. We compare z_{1i} and z_{2i} in Fig.9.

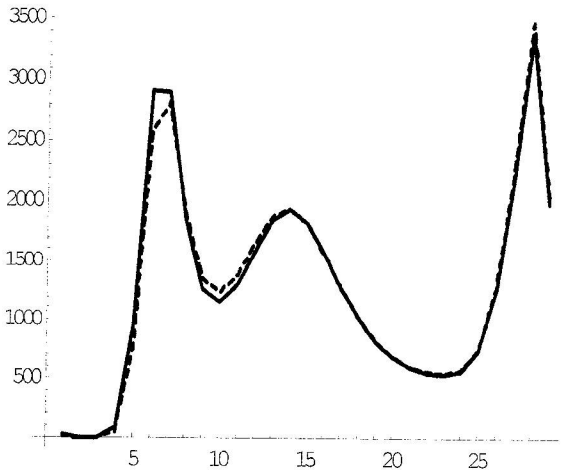

Fig.9: Comparison of z_{1i} and z_{2i} (solid line) with i going from 1 to L for N = 37000. The year one is 1973. The year L is 2001.

5. Estimates for the United States: In this case, the analysis is exactly the same as for Canada, and we give only the results. We took the same values of λ's and θ's as in the case of Canada, and we considered three different cases with N = 2000000, 2300000 and with N = 2500000. Taking into account AIDS deaths (481351), this accounts for HIV prevalence of slightly over 1.5, 1.8 and two million people at the end of 2001. As in [3], the AIDS cases and AIDS deaths were supposed to start one year earlier in this case than in the case of Canada.

Once again, there was no convergence and we stopped the iteration when the

difference between the calculated number of AIDS cases plus the AIDS deaths and the actual ones (as reported to CDC) was as close as we could make it. For N = 2000000, the difference between these two numbers is mapped in Fig. 10 and we report the results when the curve in this figure crosses the x-axis after about forty five iterations. Of course, in a discrete number of iterations, we do not expect the two numbers to exactly coincide. In Fig.10 (as in Fig.3), the results are isolated points which have been joined together. For the case we have reported, these two numbers (divided by pdot plus qdot) were 2.0 million (which is the assumed value of N) and 2.0 million (correct to six figures). The figure also suggests rather strongly that the iteration does not converge in this case.

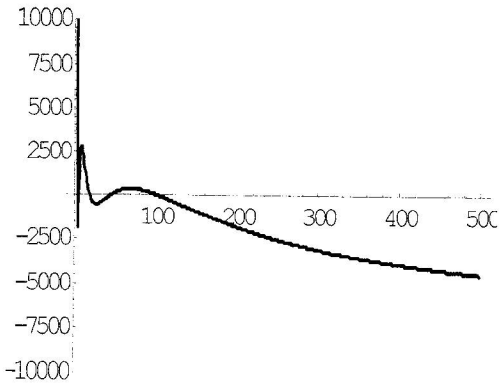

Fig. 10: Difference between the calculated number of AIDS cases plus the AIDS deaths and the actual ones at the end of 2001 for the United States for N = 2000000 for the first five hundred iterations. We have reported the results for the case when the curve in this figure crosses the x-axis after about forty five iterations.

The yearly HIV incidence rate is given in Fig.11

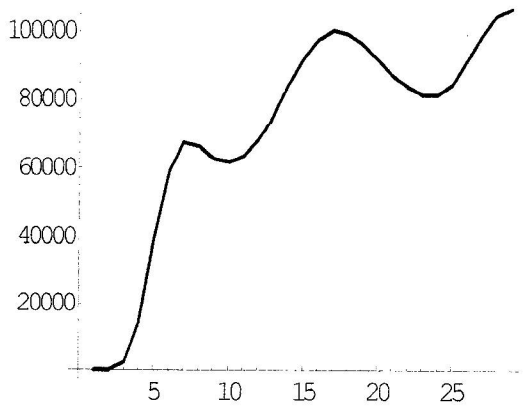

Fig. 11: HIV incidence rate in the United States according to our model for N=2000000. The year 1973 is the year one.

The agreement between the actual cumulative number of AIDS cases, as reported to CDC, and the ones calculated according to this model is given in Fig. 12.

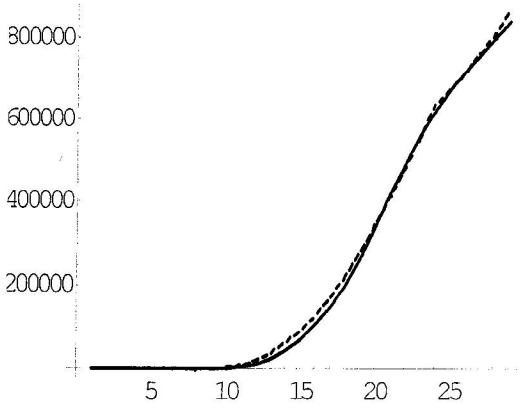

Fig. 12: Comparison of actual cumulative AIDS cases (solid line) with the calculated ones. The year 1973 is the year one.

We also compare the actual cumulative AIDS deaths with the calculated ones in Fig. 13.

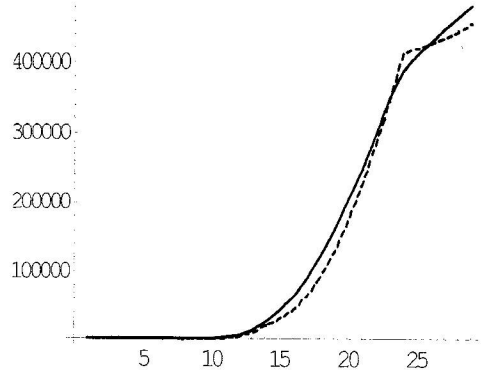

Fig. 13: Comparison of actual cumulative AIDS deaths (solid line) with the calculated ones according to our model. The year 1973 is the year one.

We also compare the actual yearly AIDS deaths with the calculated ones in Fig. 14. The AIDS deaths after 2001 were calculated by assuming that all the p[i]'s after that year are zero [7].

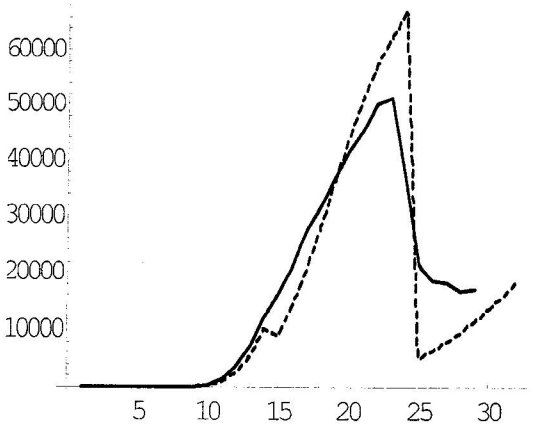

Fig. 14: Comparison of actual AIDS deaths (solid line) with the calculated ones in the United States on an annual basis according to our model. The year 1973 is the year one. The AIDS deaths after 2001 were calculated by assuming that all the p[i]'s after that year are zero.

Finally, the values of z_{1i} and z_{2i} are compared in Fig.15.

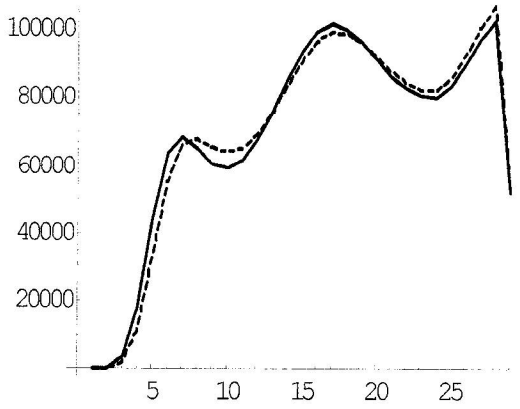

Fig. 15: Comparison of z_{1i} and z_{2i} with i going from 1 to L for N = 2000000. The year one is 1973. The year L is 2001.

We also considered N = 2300000, and N = 2500000 and we report the corresponding HIV incidence in these two cases in Fig 16. The agreement of cumulative calculated number of AIDS cases and AIDS deaths with the actual ones was approximately the same as in the case reported above.

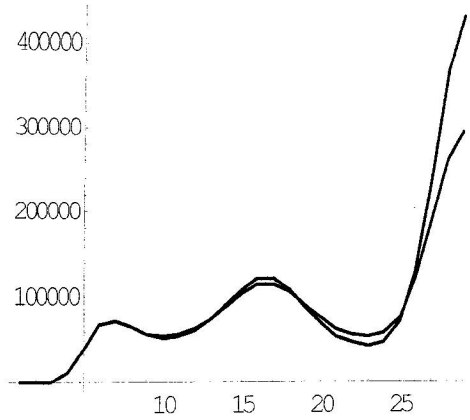

Fig. 16: HIV incidence rate in the United States according to our model for N=2300000, and for N=2500000. The curve for N = 2500000 is higher

on the right. The year 1973 is the year one.

Notice the similarity in the HIV incidence in the two cases of N = 2300000 and N = 2500000. Both of them indicate peaks in the early eighties and in early nineties. Once again, for the case of N = 2500000, the 'additional' HIV infections are pushed to the end (towards 2001), so that they have little effect on AIDS incidence and/or AIDS deaths. Comparing the three cases given above, we would conclude that N = 2000000, 2300000, or 2500000, depending upon whether, in the late nineties, the HIV incidence was approximately 100000, 150000, or more than 200000 per year. Of course, if this incidence rate was lower, then a lower value of N would be more appropriate. While the official numbers (as estimated by CDC) in this case are approximately 40000 per year, we would think that if "90 percent of blacks, 70 percent of Hispanics and 60 percent of whites, did not know that they were HIV positive", then the actual HIV prevalence in this case is anybody's guess.

We should like to point out that in Canada, where the situation should be comparable, the number of AIDS cases reported averaged 605.4 per year during the five years preceding 2001. If we assume HIV prevalence in Canada in 2001 to be 26000, and take into consideration the number of AIDS people still alive, we come up with the conclusion that 3.1043 per cent of the HIV positive people who do not have AIDS contracted AIDS per year during these years. If we apply the same ratio to the average number of AIDS cases reported in the United States (43412) during the five years preceding 2001, we come up with a figure of 1.758 million HIV positive people living in the United States at the end of 2001, suggesting that the correct value of N in this case should be about 2.24 million.

That the actual number of HIV positive people in the United States may be considerably more than the official number is again suggested by the following statement.

"In Washington, D.C., nearly 10,000, or one in about every fifty people, have AIDS and there are an unknown, but even higher number with HIV. D.C. also has the highest rate of new AIDS cases in the country-12 times the national average- and has more people living with AIDS than all but nine states.

By all counts, it is an epidemic, and the statistics rival a number of African countries. To make matters worse, the problem is growing even though the city has spent $500 million over the past eight years on medical care, HIV testing, counselling and other services." [12].

6. Estimates for Mexico: In Mexico, UNAIDS estimates of HIV prevalence vary from a low of 78000 to a high of 260000 with a mean estimate of approximately 160000 at the end of 2003 [11]. In this case, we based our calculations on AIDS cases only. Reliable numbers of AIDS deaths were not available to us. We shall, therefore, cycle between equations (A1) and (B1). We shall now give the results for Mexico.

The first cases of AIDS were diagnosed in Mexico in 1983. However, as pointed out before, it takes more than five years after HIV positivity has made its appearance in a society, for AIDS cases to start showing up. We shall, therefore

assume that HIV positivity made its appearance in Mexico in 1978. The actual number of AIDS cases reported [11] is given in Fig.18 (solid line), where it is compared with our calculations. We notice from Fig. 18, that the ARV's started making a difference in Mexico in the early nineties, and that the situation has not improved much after that. We therefore divide the time from 1978 onwards into two parts, $0 \le t \le t_1$ and $t \ge t_1$ where we took $t_1 = 14$. The values of λ_1, λ_2, θ_1, and θ_2 were taken to be equal to .1, .095, 5 and 5 respectively thus giving the incubation period of 9.18 and 9.665 respectively. Once again, these values are rather arbitrary, but we think that they represent realistic numbers in the present context.

We tried three values of N, namely N = 130,000, N = 150,000, and N=200,000. HIV incidence in each case is given in Fig. 17.

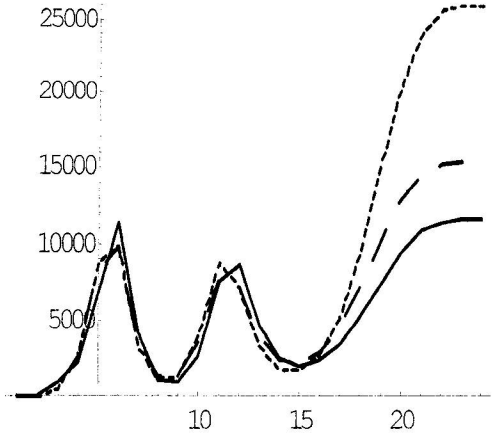

Fig. 17: HIV incidence in Mexico for N=130,000, N=150,000 and N=200,000. The corresponding curves are higher on the right. The year one is 1978.

We notice the same behaviour in this case as in the other two cases. HIV incidence is rather independent of N in the beginning, and for higher values of N, 'additional' cases are pushed towards the end where they make little difference in the number of AIDS cases. It follows that if we know HIV incidence in recent years with some certainty, then current HIV prevalence can be estimated rather accurately. If we consider the recent HIV incidence in Mexico at about 20,000 per year, and HIV deaths up to 2001 to be about 30000 (there were approximately 5000 AIDS deaths in 2004 alone), then N = 160,000 is approximately the correct value of N and HIV prevalence in Mexico at the end of 2001 should be about 130000. This compares with UNAIDS estimate of 160000 at the end of 2003.

Comparison of yearly and cumulative actual AIDS cases with calculated ones for different values of N is given in Fig.18 and in Fig.19.

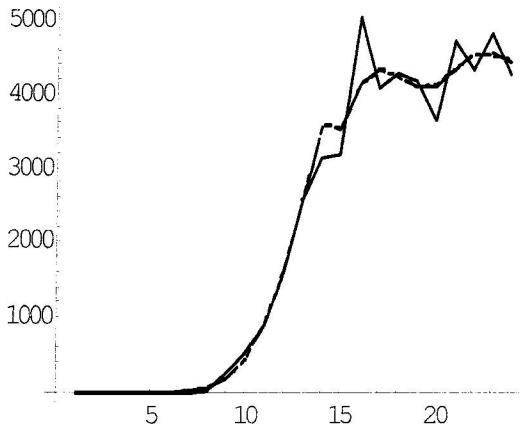

Fig. 18: Comparison of the actual number of yearly AIDS cases (solid line) with the calculated ones for the three different values of N. The results of the three different cases are indistinguishable from each other.

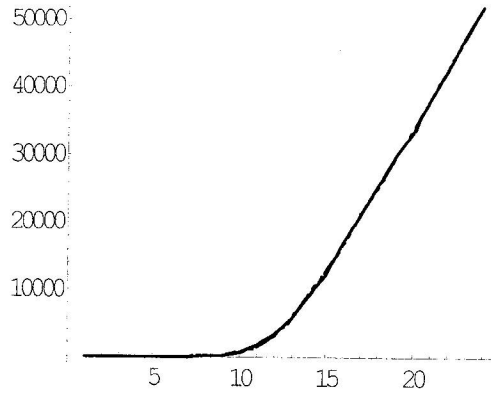

Fig.19: Comparison of the actual number of cumulative AIDS cases (solid line) with the calculated ones for the three different values of N. The results of the three different cases are indistinguishable from the actual ones.

7. Reliability of Procedure: To test the reliability of our procedure, we present here two cases of simulation with this process. In the first case, we assume that p[i] = 100 for each i from i = 1 to i = 29, i.e. all the way from 1973 to 2001 while in the second case we assume that p[i] = 100 i. We calculate the AIDS cases and AIDS deaths based on these p[i]'s, and arbitrarily put the AIDS cases and AIDS deaths equal to zero during the first seven years (as was the situation in the case of Canada above). Then we calculate p[i]'s, i.e. past HIV infections based on these AIDS cases and AIDS deaths. The results are shown in Fig.20 and in Fig.21 for two different cases with N = 2900 and N = 5800 for p[i] = 100 and for N = 43500 and N = 87000 for p[i]=100i. The first values of N represent the results when the total number of HIV infections assumed is simply the sum of all the p[i]'s and the second one when the total number assumed is twice as large. It is clear from the figures that the past infections are fairly accurately calculated in all cases while the 'additional' infections in these cases are simply shifted to the right.

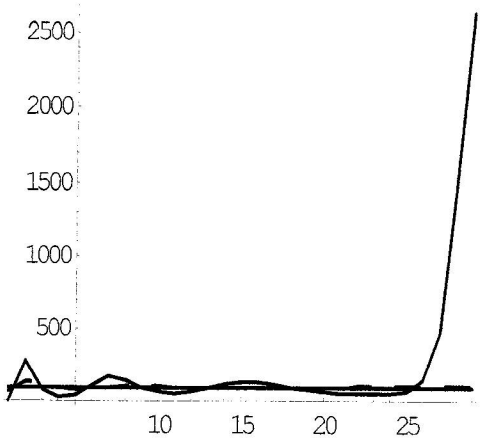

Fig. 20: Simulation of the case p[i] = 100 for all i. The straight line represents the actual values to be reproduced. The dotted line is this reproduction with N = 2900 while the curve rising on the right represents this reproduction with N = 5800.

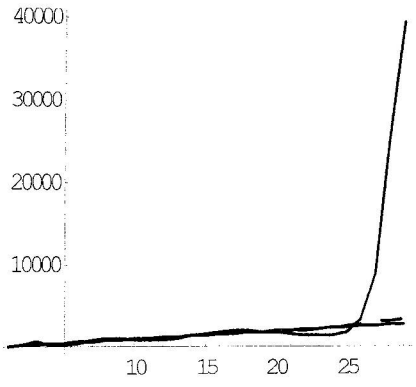

Fig. 21 : Simulation of the case p[i] = 100i for all i. The straight line represents the actual values to be reproduced. The dotted line is this reproduction with N = 43500 while the curve rising on the right represents this reproduction with N = 87000.

8. Conclusion: The Expectation Maximisation technique can give

accurate estimates of HIV prevalence if the HIV incidence in recent years is known with some certainty. It is hoped that, with time, our estimates of recent HIV incidence will improve and that should enable us to obtain reliable estimates of HIV prevalence. With our analysis in this paper, we would conclude that the HIV prevalence in Canada should be towards the lower end of UNAIDS estimates, in Mexico towards the middle, and in the United States, towards the higher end of these estimates, or perhaps, even higher.

9. Acknowledgment: Helpful suggestions from Dr. Rita Aggarwala of the University of Calgary are gratefully acknowledged.

References:

1. Anderson R.M. and May R.M., "Infectious Diseases of Humans" Oxford University Press, (1991).
2. The Financial Times, Special Report, "Business and AIDS", Dec. 1, 2003.
3. Aggarwala, B.D., On estimating HIV Prevalence in Canada and in the United States, Far East Journal Of Applied Mathematics, Vol. 20, No: 3 (Sept. 2005).
4. UNAIDS, Epidemiological Fact Sheets, 2004 Update.
5. Brookmeyer, R. and Gail, M.H., "AIDS Epidemiology, A Quantitative Approach", Oxford University Press, (1994).
6.Smith, R.A., Encyclopaedia of AIDS, Penguin Books, 2001.
7. Brookmeyer, R and Gail, M.H., A method for obtaining short term projections and lower bounds on the size of the AIDS epidemic, Journal of the American Statistical Association, Vol. 83, No: 402, (June 1986), pp. 301-308.
8. Health Canada, HIV/AIDS EPI Updates, May 2005.
9. Geduld J., Gatali M., Remis R. S., Archibald C.P. "Estimates of HIV prevalence and incidence in Canada, 2002" Canada Communicable Disease Report, vol.29-23, Dec.1, 2003, pp. 197-206.
10. Privete communication, Oct. 18, 2005.
11. UNAIDS/WHO Epidemiological Fact Sheet, 2004 update.
12. www.Healthology.com, April 17, 2006.