A new approach to estimating AIDS incubation times: results in homosexual infected men

Sylvie Chevret, Dominique Costagliola, Jean-Jacques Lefrere, Alain-Jacques Valleron

Abstract

Study objective—The aim was to develop a new approach for estimating the incubation period of acquired immunodeficiency syndrome (AIDS), based on age distributions.

Design—Incubation period was expressed as the difference between age at time of diagnosis and age at time of contamination. Assuming independence between age at time of infection and incubation period, the age distribution of newly diagnosed AIDS cases is thus the convolution product between the distributions of the age of freshly infected patients and of the incubation times. AIDS incubation time can therefore be estimated from the age distribution of newly HIV infected subjects and newly diagnosed AIDS cases.

Subjects—Subjects were 2220 AIDS cases diagnosed until 1987, reported to the Ministry of Health, France, and 172 subjects discovered to be HIV-1 seropositive during a blood donation in Paris between August 1985 and July 1988. In both groups, the only known risk factor was homosexuality.

Main results—The estimated median incubation period was 3-9 years (90% CI 9-0-10-9 years). Confidence intervals were narrow, even when taking into account the uncertainty in serodetection delay (90% CI 6-7-13-5 years).

Conclusions—The incubation estimate is as accurate as previous estimates based on other models. This technique could therefore be applied to other risk groups.

Methods

The incubation period is defined as the time from HIV infection to AIDS diagnosis. The basic idea on which our technique was built consisted in expressing the incubation period as the difference between the age at the time of AIDS diagnosis and the age at the time of HIV infection. Assuming independence between age at the time of infection and incubation period, the probability density function (pdf) of age at the time of AIDS diagnosis (g) is defined as the convolution product between the pdf of age at the time of HIV infection (f) and the pdf of incubation time (h):

\[ g(a) = (f*h)(a) = \int_0^a f(a-x) h(x) \, dx \]

where \( a \) denotes the age, and \( x \) denotes the incubation time, both expressed in the same units. Therefore, \( h \) can be obtained by deconvolution between \( g \) and \( f \).

Information on the distributions of age at the time of HIV infection and of age at the time of AIDS diagnosis was obtained from the two sets, namely infected patients and AIDS cases.

From 1 August 1985 to 30 June 1988, all blood donors detected as seropositive at the Fondation Nationale de Transfusione Sanguine (Paris, France), whose homosexuality was the only risk factor, were included in this study if they were without symptoms (172 patients). In France, it must be noted that HIV screening can be done free of charge on a voluntary basis in free and anonymous HIV screening centres (one at least in each French department), and that blood donors have been systematically screened since August, 1985.

Ages were known to the nearest year for all individuals. Serum samples were screened for HIV antibody by ELISA with confirmation by western blot, and T helper cell (CD4) count was determined for all of them.

The AIDS cases were all homosexual and bisexual men who were diagnosed between 1 January 1978 and 31 December 1987, and who were reported to the Direction Générale de la Santé (Ministry of Health, France) before 31 March 1989 (2220 patients). Subsequent cases were ignored because of the delay between diagnosis and report.11 It must be noted that the age distribution of AIDS cases was available by intervals of 10 years in length, due to the French legislation at the time.

Assuming the age distribution of both infected patients and AIDS cases to be constant over time, the two observed age distributions of infected men and AIDS cases could define estimates of the
probability density functions \(f\) and \(g\), respectively. The incomplete nature of the observations (use of two independent samples) made estimation of the incubation time an ill posed problem.\textsuperscript{12} In order to obtain untruncated observations and to optimise the deconvolution procedure, considering that such a process is quite sensitive to noise in the input data,\textsuperscript{13} we thereafter fitted each observed age distribution by a parametric distribution. We compared the fits obtained by using three different families of distribution (ie, log-normal, Weibull, and gamma), the parameters of which were estimated by the maximum likelihood method, using the Newton-Raphson algorithm. The form of each age distribution was chosen according to the best adequacy, as assessed by the value of the Pearson \(\chi^2\) test. However, since the data were truncated, these tests were applied over the observed restricted ranges.\textsuperscript{14} For both age distributions, the gamma model was selected:

\[ f_{p,q}(x) = \frac{1}{\Gamma(q)} x^{p-1} e^{-px} \]

For the AIDS cases, age being grouped at fixed intervals, the likelihood function was expressed as a multinomial probability, regarded as a function of \(p\) and \(q\).\textsuperscript{14}

Finally, considering that the calendar date of infection and the incubation period were independent, a deconvolution procedure, based on Fast Fourier Transforms, was applied to these estimates,\textsuperscript{13} yielding the estimate of \(h\), the probability density function of incubation time.

However, our infected sample only provided age at the time seropositivity was discovered. Since the individual dates of exposure are unknown, we have first neglected the underlying time from HIV infection to HIV seropositive blood donation, and the previous estimate of \(h\) should represent the lower limit of the probability density function of AIDS incubation period. Secondly, a more realistic approach allowed us to take into account the possible delay between time of HIV contamination and time of seropositivity discovery. This delay was expressed in the same way as for the incubation period, ie, as the difference between the age at the time of serodetection and the age at the time of HIV infection. Independence between this delay and the age at the time of infection being supposed, the previous estimate of \(f\), the probability density function of age at the time of HIV infection, was thus corrected by deconvoluting it with the probability density function of this delay. The latter was obtained from the observed distribution of CD4 count, assuming the serodiscovery delay could be expressed as a function of CD4 count, as previously used to infer dates of contamination.\textsuperscript{15-17}

First, to smooth the data, the CD4 count distribution among the HIV seropositive infected subjects was fitted by a gamma distribution, by the maximum likelihood method, using the Newton-Raphson algorithm (fig 1). Then, to infer the delay of serodetection from the observed CD4 count, we applied a linear relationship,\textsuperscript{18} the parameters of which were fixed with respect to the following points: (1) the maximum CD4 value observed in this study defined a fresh infection (ie, without any delay between contamination and discovery of seropositivity), and (2) the observed median CD4 count among the infected sample (median CD4 of 480) corresponded to the median delay reported in Moss's study, that is 2.5 years.\textsuperscript{19}

Thereafter, analyses were reperformed with this corrected estimate of \(f\), yielding an estimate of \(h\) corrected for serodetection delay.

In summary, we obtained two estimates of incubation time, namely the crude estimate and the estimate corrected for serodetection delay. However, such estimates could be related to the uncertainties either about the estimated age distributions or about the supposed median time between HIV contamination and serodetection. To assess the variability of these estimators, we performed sensitivity analyses. Using the Monte-Carlo method, we incorporated the uncertainties about each age distribution by sampling the two parameters \(p\) and \(q\) from two dependent normal distributions, means and standard deviations of which were the maximum likelihood estimates. This yielded confidence intervals for the estimated mean, median and standard deviation of the incubation time distribution. Finally, the uncertainties about the serodetection delay were taken into account in reperforming the analyses with one and four years as two plausible extreme values for the median serodetection delay (fig 1).

![Figure 1](http://jech.bmj.com/) Estimation of the serodetection delay. In this paper, the exact age at the time of infection is unknown and had to be estimated indirectly. Thus the CD4 count has been used to infer this age distribution. The top panel shows the CD4 count distribution among the HIV seropositive subjects, as fitted by a gamma distribution, using Newton-Raphson algorithm (with maximum likelihood estimates at \(p = 1.1655\), and \(q = 4.3585\)). The bottom panel shows the relationship assumed to be linear (see refs 18, 19) which has been used to infer the serodetection delay. The parameters of this relationship were chosen in order (1) that the delay (2.5 years) corresponding to the median CD4 count observed in the study and in Moss's study\textsuperscript{19} was identical and (2) that the maximum CD4 observed value corresponds to a freshly infected subject. However, in order to study the sensitivity of the estimates, two plausible extreme median values (ie, one year and four years) were tested, yielding the bounds shown.
Results

Figure 2 displays the cumulated age distributions of the infected subjects and AIDS cases, both observed and smoothed. Among the 172 infected subjects, the observed mean age was 31-4 years, the median age of 30-0 years, and the standard deviation 7-5 years, while among the 2220 AIDS cases, the observed mean age was 38-1 years, with a median estimated 35-0 years, and the standard deviation 9-6 years. The age distribution of the HIV infected subjects modelled by the gamma distribution \( p = 0.48, q = 15-03 \) did not differ from the observed age distribution. For AIDS cases, the fitted gamma distribution \( p = 0.49, q = 18-51 \) and the observed distribution were roughly similar, although there were some imbalances due to the grouped observed ages.

Assuming first that the delay between HIV contamination and seropositivity discovery was negligible, the estimated mean incubation period was 6-6 years (90% CI 5-7-7-8 years), the estimated median was 6-1 years (90% CI 5-3-7-1 years), and the estimated standard deviation (SD) was 3-9 years (90% CI 3-5-4-3 years).

Secondly when we took into account the serodetection delay, the estimated mean incubation time was 10-4 years (90% CI 9-4-11-6 years), the estimated median became 9-9 years (90% CI 9-0-10-0 years), and the SD was estimated at 3-9 years (90% CI 3-4-4-5 years). Based on this latter distribution, we inferred that nine years after HIV contamination, \( 42\% \) (90% CI 32%-51%) of infected homosexual men will have developed AIDS.

Finally, we examined the sensitivity to the median serodetection delay, the assumed value ranging from one year to four years. The estimates of the mean incubation time then ranged from 8-1 years (assuming a median delay of one year) to 13-0 years (assuming a median delay of four years), with a lower 90%, confidence limit of 7-1 years and an upper limit of 14-1 years, and the estimates of the median incubation time ranged from 7-6 years to 12-0 years, with a lower 90% confidence limit of 6-7 years and an upper limit of 13-5 years (table I).

Discussion

We presented a new model for estimating the incubation period of AIDS, based on data broadly available, i.e., the age distribution of infected men and AIDS patients. Bacchetti24 similarly estimated the distribution of incubation period by comparing the estimated distribution of infection and the observed pattern of diagnosis. However, our method was applied to age distributions, considering that they were more likely to be available from routine examination than dates of contamination.

This model was applied to two different data-sets. Seropositive individuals were recruited among blood donors, and even though donor’s age is limited by the French legislation, its range (from 18 to 65 years old) is large enough to cover the majority of the homosexual infected population. Furthermore, data were fitted to a gamma distribution, so that the truncated tails of the observed age distribution were smoothed. The AIDS cases were formed by the entire population of the AIDS cases diagnosed and reported by the Direction Générale de la Santé during a fixed time period, so that the delay between the diagnosis and the report of the syndrome could be neglected in the analysis.11 Consequently, we used data through 1 April 1989 to obtain nearly all of the AIDS cases for analysis at the reference date of 31 December 1987. Moreover, while some cases might be undeclared, this is likely to be independent of age and should not have any impact on the estimate. Finally, this analysis employed AIDS data before 1988 so that estimates of the incubation distribution may not have been affected by any treatment.25

Several assumptions have been made in building the model and in applying it to these data. First, independence between age at the time of HIV infection and AIDS incubation period was assumed for the homosexual infected population. Indeed, the most commonly reported difference in incubation times concerns comparison between children and adults.4,26,27 Fewer studies have focused on the influence of age on incubation time among HIV infected
adults\(^3\) 4 \(19\) 26–29 some of which reported shorter median incubation times among older than younger adults. Nevertheless, most of these studies\(^4\) 26 \(27\) did not concern the homosexual population, but involved either homosexuals\(^8\) 26 \(27\) or transfusion recipients.\(^4\) Furthermore, the potential effect of age on HIV incubation time among the latter exposure group is still a matter of controversy since subsequent studies did not succeed in demonstrating such a dependence.\(^3\) 29 Among homosexual men, age has been reported as a significant factor for progression to AIDS,\(^9\) but we argue that this could have been biased considering that time to AIDS progression was assessed from date of the first HIV antibody seroconversion, without incorporating the delay from infection.\(^20\) Moreover, in the reports with the longest follow up concerning homosexual men, age was not found to be a cofactor for incubation times.\(^7\) 28 Finally, we argue that the method could easily be extended to account for any age association with incubation time by introducing separate integrals, although stable estimates would require a larger sample of seropositive subjects.

To obtain estimates of the age distribution of the newly infected population and AIDS cases, it was next assumed that age distributions were constant over time, in both the infected population and the AIDS population. None of the demographic studies reported among the homosexual infected population have shown the age distribution to vary significantly over a 10 year period.\(^31\) Moreover, the mean age at the time of either seroconversion or AIDS diagnosis observed in our data did not vary significantly from 1985 to 1987 (table II).

\[
\text{Table II Observed mean age of seropositive blood donors and AIDS cases, according to the date of either seroconversion for the former, or AIDS diagnosis for the latter. The estimates are rounded by the 95% upper and lower confidence limits.}
\]

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean (95% CI)</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985 (n=52)</td>
<td>35.1 (31.1–32.4)</td>
<td>31</td>
</tr>
<tr>
<td>1986 (n=74)</td>
<td>30.7 (30.2–32.1)</td>
<td>28</td>
</tr>
<tr>
<td>1987 (n=40)</td>
<td>32.1 (30.0–34.2)</td>
<td>31</td>
</tr>
<tr>
<td>1988 (n=6)</td>
<td>34.5 (29.7–39.3)</td>
<td>35</td>
</tr>
</tbody>
</table>

\*95% confidence interval

Regardless of the application, questions of the model sensitivity to input feature values are obviously present, so that the modelling of either age distributions and seroconversion delay must be discussed. We smoothed the age distributions of both seropositive men and AIDS cases, using a gamma distribution, a distribution flexible enough to represent a wide range of practical situations.\(^4\) For the latter, some discrepancies were found between the observed and the estimated distributions, which we explained by the unavailability of individual ages due to the French legislation at the time. However, we considered that the essential feature of fitting parametric models was estimation rather than testing hypotheses, and that we only wished to obtain a continuous distribution giving age probabilities as close as possible to the observed proportions. Therefore we judged these estimates to be acceptable.

The lack of availability of the distribution of age at the time of HIV infection led us impute the unknown delay from seroconversion to seroconversion for the seropositive subjects. At the time of serodiagnosis, the level of CD4 cells in the blood of all these individuals was measured. Therefore, we constructed a model in which the time elapsed since the initial infection was expressed as a linear function of CD4, considering the well known downward trend of CD4 count following the initial HIV infection.\(^15\) 19 Other parametric models have already been used for such an inference, either the stochastic model\(^13\) or the Weibull model,\(^17\) as well as parametric growth curves.\(^16\) However, these models suppose several follow up visits, each including a measure of CD4 level, that were not available in this study. The rough linear relationship used in this study led us to assume that 50% of the infected population have been infected within 2-5 years (fig 1). Nevertheless, a sensitivity study was performed, examining two extreme plausible values for this median delay, ie, one year and four years. As expected, the estimates were modified, the mean ranging from a lower limit of 7-1 years to an upper limit of 14-1 years, while the median ranged from 6-7 years to 13-9 years, respectively. However, we argue this is one major advantage of this method, provided confidence intervals are narrow as compared to the previous estimates among homosexual infected men (table I). Indeed, some of these studies have not reported confidence intervals for their estimates\(^17\) 20 21 or have reported an upper limit as being infinite.\(^23\)

CONCLUSION

This method could now be applied to other risk groups, provided that the two age distributions of the infected population and of new AIDS cases were known and that the natural history of AIDS has not been modified by treatment. It could bring still more accurate results if subsamples of newly infected subjects could be used, avoiding the estimation of the underlying delay of serodiagnosis.

This work was partly supported by ANRS grant 90 182 (D Costagliola). We wish to thank Professor Guy Thomas and Professor Claude Chaang for helpful comments; and Dr Anne Laporte for making available data on French AIDS diagnosed cases.