

# On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion

Ronald B. Geskus

Municipal Health Service  
Division of Public Health and Environment  
Nieuwe Achtergracht 100  
1018 WT Amsterdam  
31.20.555 5524  
rgeskus@ggd.amsterdam.nl

and

Vrije Universiteit  
Faculty of Sciences  
Division of Mathematics and Computer Science  
De Boelelaan 1081<sup>A</sup>  
1081 HV Amsterdam

# Inclusion of prevalent cases in HIV/AIDS natural history studies through a nonparametric, marker-based estimate of time since seroconversion

## SUMMARY

In most cohort studies on HIV infection and AIDS, seroprevalent cases provide a substantial amount of information. Inclusion of these persons in natural history studies requires a fairly unbiased method to estimate their seroconversion distribution. When a cohort-based estimate is not feasible, an alternative is to estimate individual seroconversion distributions, based on marker values at entry. In this paper, a nonparametric, marker-based estimation method is developed.

The method is applied to data from the Amsterdam cohort study on homosexual men. For the seroprevalent cases who entered the study between October 1984 and April 1985, individual seroconversion distributions are estimated based on their first measured CD4 count. In subsequent survival analyses, dates of seroconversion are estimated via conditional mean imputation.

Inclusion of these seroprevalent cases greatly improves the quality of the data. Age at seroconversion is a significant cofactor for disease progression, a result not found when the analysis is restricted to the seroconverters. In order to incorporate the uncertainty in the imputed date of seroconversion, a bootstrap procedure is developed for the computation of p-values and confidence intervals. In our analyses, standard procedures, which ignore the uncertainty in the imputed date of seroconversion, perform almost as well.

Keywords: HIV-infection, AIDS, doubly censored data

## 1. INTRODUCTION

In the majority of the cohort studies on human immunodeficiency virus type one (HIV-1) infection and AIDS, data on the time from seroconversion to AIDS or death are doubly censored. Not only some of the event data are right censored, due to loss to follow-up, death by competing risks or study cutoff, but also the date of seroconversion is hardly ever observed exactly. Depending on study design and entrance criteria, a cohort study may contain prospectively and retrospectively identified seroconverters as well as seroprevalent cases. Prospectively identified seroconverters have a last seronegative and a first seropositive test result, both obtained during follow-up. Seroprevalent cases were already seropositive at their date of entry into the study. Retrospectively identified seroconverters were also seropositive at entry, but an earlier date at which they were known to be seronegative is available, e.g. obtained via a seronegative test result from stored blood samples. Moreover, most cohort studies contain persons who remained seronegative until the end of follow-up. So the date of seroconversion may be left censored, interval censored or right censored, but is hardly ever observed exactly.

In many HIV/AIDS cohort studies, analyses are restricted to seroconverters with a narrow seroconversion interval. However, prevalent cases often constitute a substantial part of the cohort. Moreover, persons who entered seropositively at the start of the study are the ones with the longest follow-up. Inclusion of these persons may considerably decrease variance, but may also increase bias if the uncertainty with respect to the date of seroconversion is not dealt with correctly. Several methods exist for inclusion of prevalent cases and persons with wide seroconversion intervals.<sup>1,2</sup> In a simulation study, conditional mean imputation, using a cohort-based estimate of the seroconversion distribution, hardly led to any bias and considerably decreased variance of the Kaplan-Meier estimator, compared to an analysis restricted to seroconverters with narrow intervals.<sup>2</sup> Although only one specific set of distributional choices of seroconversion date, event time and observation times was used, the amount of censoring and truncation present in the simulated data is fairly large. Therefore, the lack of bias is not due to the dominance of narrow seroconversion intervals and conditional mean imputation may be a good method in many cohort studies with different censoring patterns as well.

Problems occur if no reliable estimate of the cohort seroconversion distribution can be obtained on some part of the time scale of the epidemic. All HIV/AIDS cohort studies started several years after the beginning of the AIDS epidemic, and usually very little or no cohort-based information can be obtained regarding the seroconversion pattern before the start of the study. Even if such information is available, it may originate from a specific subgroup and thus fail to be representative.

Markers of disease progression, like CD4 count and viral load, may provide an alternative approach. They can be used as determinants of stages of disease progression in a Markov model.<sup>3</sup> Data from the Amsterdam cohort study on homosexual men, including prevalent cases, have been analysed in this way.<sup>4</sup> In this paper, we develop another approach. Based on the marker values at entry into the study, individual seroconversion curves are estimated for the prevalent cases, by comparing these values with the marker development of the seroconverters.

In the Amsterdam cohort study on homosexual men, information with respect to seroconversion from the period before the start of the study is available from retrospectively identified

seroconverters, but originates from a specific subgroup of the cohort, and may thus lead to biased estimates. Alternatively, the marker-based method will be used and results obtained via both methods will be compared. Throughout, the time from HIV-1 seroconversion to AIDS, diagnosed according to the 1987 Centers for Disease Control criteria,<sup>5</sup> will be the subject of interest. It will be called the incubation time.

## 2. MARKER-BASED ESTIMATION OF THE SEROCONVERSION DISTRIBUTION

With doubly censored data, two distinct time scales play a role. The date of seroconversion is measured on a calendar time scale, whereas the incubation time is measured on a scale, relatively to the date of seroconversion. Throughout, we let  $N$  be the sample size. As in Geskus,<sup>2</sup> the relevant observable data, excluding the information of covariates, consist of the set

$$\{ (v_i^n, v_i^p, e_i, z_i, \delta_i, \gamma_i, \eta_i) \mid i = 1, \dots, N \}.$$

The date of study entry for person  $i$  is  $e_i$ . The latest date he was known to be seronegative is  $v_i^n$ , and  $v_i^p$  denotes the earliest date he was observed to be seropositive. If  $v_i^n$  is not known, either we do not define  $v_i^n$ , or we let  $v_i^n = \sigma_0$ , with  $\sigma_0$  denoting some common origin of the calendar time scale, like the assumed date of start of the epidemic. If he has not been observed to seroconvert, he is right censored with respect to seroconversion at  $v_i^n$ , and  $v_i^p$  is not defined. We let  $z_i$  denote the date of AIDS diagnosis, if observed ( $\delta_i = 1$ ). Otherwise,  $z_i$  is the date of last visit without AIDS, either for a person with a seropositive test result ( $\gamma_i = 1$ ), or for a person who remained seronegative ( $\eta_i = 1$ ). The variables  $v_i^n$ ,  $v_i^p$ ,  $e_i$  and  $z_i$  are measured in the calendar time scale, whereas  $\delta_i$ ,  $\gamma_i$  and  $\eta_i$  are zero-one indicators, with exactly one of these taking the value one. For most of the persons, more observation times are available, usually in the form of repeated visits, but they do not provide information with respect to the distributions of interest. Hidden behind the observable data is the situation without censoring  $\{(x_i, t_i) \mid i = 1, \dots, N\}$ , with  $x_i$  denoting the date of seroconversion and  $t_i$  denoting the incubation time. We assume the incubation times  $T_1, \dots, T_N$  to be independent random variables with distribution function  $F$  and density  $f$ . In the methods considered in Geskus,<sup>2</sup> one cohort-wide seroconversion distribution  $G$  was assumed. In the present study, we allow this distribution to be different for each individual.

Information with respect to the date of seroconversion is provided by marker values at study entry. For this, we need information on how markers values relate to time since seroconversion. Since marker values may differ by risk group and may be laboratory dependent, such information is best provided by the seroconverters with small seroconversion intervals from the same cohort. The cohort is split up in a reference group, having accurate information with respect to the date of seroconversion, and a group, for which the date of seroconversion is estimated based on the earliest marker values, which typically have been measured at or shortly after the first seropositive test. In the sequel, the first group will be called the reference group, and the second one the prevalent group (although the latter may contain seroconverters with wide intervals).

Let  $N_r$  and  $N_p = N - N_r$  denote the number of persons in the reference group and in the prevalent group. For an individual in the prevalent group, let  $d_i$  denote the earliest calendar time after seroconversion at which the markers of interest have been measured and let  $m_i(d_i)$  denote the corresponding vector of marker values. In our application, estimation is based on

CD4 count only. Therefore, we assume  $m_i(d_i)$  to be one-dimensional. We are interested in estimating  $w_i = d_i - x_i$ , the time elapsed since seroconversion. We introduce the random variable  $W_i := d_i - X_i$ .  $W_i$  can only have values on the interval  $[d_i - v_i^p, d_i - v_i^n]$ , with  $v_i^n = \sigma_0$  for the prevalent cases without a last seronegative test. We assume  $W_i$  to have a distribution depending on the marker value measured at  $d_i$ . So

$$W_i \sim H_i = H(\cdot \mid m_i(d_i), d_i - v_i^p, d_i - v_i^n),$$

and we want to estimate  $H$ .

Markers are highly variable, due to measurement error and short term fluctuations. A better representation of  $m_i(d_i)$ , the true marker value at  $d_i$ , may be obtained by averaging the first marker measurements, as long as they are not too far apart in time. In our application, CD4 count tends to decrease with disease progression. Hence, averaging may be problematic for individuals who show a steady decrease in CD4 count over the first measurements. Therefore, individual CD4 trajectories are smoothed via isotonic regression.<sup>6</sup> Isotonic regression yields the least squares estimate of the “true” CD4 development over time, under the restriction that only downward trends are possible. For individuals with a steady decrease in CD4 count, only the first measurement is used, whereas for individuals with a stable CD4 count, CD4 count is averaged over this stable period. For normally distributed data, the isotonic least-squares estimate is equivalent to the maximum likelihood estimate under the isotonic model.<sup>6</sup>

## 2.1. Estimation methods

For the moment, we assume the marker to be a categorical variable, say  $m(\cdot) \in \{C_1, \dots, C_k\}$ . Suppose  $m_i(d_i) = C_l$  for person  $i$  from the prevalent group. Let  $N_l$  be the number of persons in the reference group with at least one value  $C_l$ . Let  $n_k^l$  be the number of measurements of person  $k$  from the reference group that are equal to  $C_l$  and such that the corresponding times since seroconversion are between  $d_i - v_i^p$  and  $d_i - v_i^n$ . Let  $s_{k1}, s_{k2}, \dots, s_{kn_k^l}$  denote these times since seroconversion. Then an estimate of the cumulative distribution function of  $W_i$  is defined as

$$\hat{H}_i(w) := \frac{1}{N_l} \sum_{k=1}^{N_l} \frac{1}{n_k^l} \sum_{j=1}^{n_k^l} \mathbb{1}_{\{s_{kj} \leq w\}}. \quad (1)$$

Drawing from this weighed empirical distribution function is equivalent to drawing random dates of seroconversion via a two-step procedure. First a person  $k$  who has at least one value  $C_l$  observed on the interval  $[d_i - v_i^p, d_i - v_i^n]$  is drawn from the reference group. In the second step, a time point  $s_{ki}$  is randomly drawn from all time points at which that person has observed marker values  $C_l$ , and we let  $w_i = s_{ki}$ .

We will also consider the version in which each measurement receives equal weight:

$$\hat{H}_i^*(w) := \frac{1}{\sum_{k=1}^{N_l} n_k^l} \sum_{k=1}^{N_l} \sum_{j=1}^{n_k^l} \mathbb{1}_{\{s_{kj} \leq w\}}. \quad (2)$$

Ideally, drawing would be done from all time *periods* at which the persons in the reference group have marker values  $C_i$ . As long as the measurement times are frequent and fairly equidistant, using the measured marker values instead will be almost equivalent. Otherwise, marker values evaluated at equidistant time points, as obtained through linear interpolation or smoothing, may be a better approach.

For markers with a continuous range of values, we use the marker values in the reference group which are in some way closest to  $m_i(d_i)$ . In our application, this distance criterion is based on absolute numbers: we take the  $K$  nearest marker values, for some  $K$  which is held fixed over all persons. These values are chosen before the truncation based on  $d_i - v_i^p$  and  $d_i - v_i^n$  is applied.

After an individual seroconversion distribution has been obtained for each individual in the prevalent group, any of the methods that uses a separate estimate of the seroconversion distribution<sup>2</sup> can be used. Here, the cohort-wide  $g$  is replaced by individual densities  $g_j$ .

Basic to our approach is the assumption that both groups have the same marker development after seroconversion. In our application, the prevalent group mostly consists of persons who seroconverted early in calendar time. So we assume the general shape of marker trajectories not to change over calendar time.

Even if this holds, application of this method involves several problems. The most important one is that no general statistical justification can be provided. As shown in the appendix, one can find situations in which the results are clearly biased. Still it may give fairly unbiased results in practice.

Another bias may arise from the fact that persons with fast disease progression are underrepresented in the prevalent group, because some fast progressors developed AIDS before the start of the cohort study. Among the prospectively identified seroconverters, which generally constitute the main part of the reference group, such persons are not missed. Hence, marker values corresponding with advanced disease progression may be overrepresented in the reference group shortly after seroconversion.

Once estimates of the individual seroconversion distributions have been obtained, the seriousness of these two problems can be ascertained through a simulation procedure. For each person from the prevalent group, a date of seroconversion is drawn from his seroconversion distribution. Next, a person is drawn randomly from the reference group. His marker development after seroconversion is used to obtain a “predicted” marker value for the prevalent case at his actual time of first measurement (using the randomly drawn time of seroconversion as time origin). The distribution of these marker values is compared with the distribution of the measured marker values from the prevalent cases.

Other problems concern the fact that the persons from the reference group do not have marker values over the total relevant time period. First, marker measurements may be unavailable after AIDS diagnosis, and surely stop at death. If the prevalent cases are AIDS-free at the date of first marker measurement, the distribution of time since seroconversion, conditionally on being AIDS-free at the time of first measurement, is considered. Then only marker values in the reference group before AIDS diagnosis are to be used. The same holds for death instead of AIDS. A second, more serious problem may be that persons in the reference group do not have marker measurements over the whole period from seroconversion to AIDS. Measurements start some

time after seroconversion for retrospectively identified seroconverters and may stop before AIDS diagnosis (or death) due to right censoring. Hence, marker values measured shortly after seroconversion or in a much later period may be underrepresented in the reference group. To a certain extent, a correction for these truncation effects with respect to the marker measurements can be incorporated in (1) and (2). It is not possible to make a completely correct modification since the date of AIDS diagnosis for the right censored cases is generally unknown. Moreover, the distribution of the time since seroconversion is completely unspecified for values that are larger than the largest time span between seroconversion and the latest marker measurement among all persons in the reference group. Hence, unless extrapolation is based on rigid parametric assumptions, the method cannot be used if the maximal follow-up of the reference group is shorter than the largest time span between the date of last negative test and first marker measurement from the persons of the prevalent group.

All these problems will be addressed in the application and the discussion.

### 3. APPLICATION

The method from the preceding section will be applied to data on HIV infection and AIDS collected in the Amsterdam cohort study among homosexual men. This study was started in October 1984. Various enrollment criteria have been used over time. Between October 1984 and April 1985, both HIV-seronegative and HIV-seropositive men who were free of AIDS-defining conditions were enrolled (protocol 1). In April 1985, further enrollment was restricted to HIV-seronegative men (protocol 2). From February 1988 onwards, HIV-seropositive men were enrolled anew for participation in an early treatment trial with zidovudine (AZT). No extra effort for recruitment of HIV-negative men has been made since then, and only a limited number entered the study.

Seronegative participants were asked to return for interview, clinical examination and blood sampling every three months until 1988, and every six months thereafter. Over the whole period of follow-up, seropositive participants have been seen on average every three months. At the date of analysis (January 1st, 1997), the study contained 568 participants having a seropositive test result available. Only persons who entered the study seropositively before 1986 or who seroconverted during follow-up are included (protocols 1 and 2). Because of the availability of treatment after 1988, which may influence the decision to take part in the study, the remaining group may constitute a biased sample. Seroprevalent cases without any follow-up information (33 in total) have also been excluded from our analyses. In figure 1, the seroconversion intervals of the remaining 333 persons used in the analysis are depicted graphically. For 151 persons, both a last seronegative as well as a first seropositive test result are available. Twenty-four of them have their last seronegative test result identified retrospectively. They provide information with respect to seroconversion on the period before the start of the cohort study. Except one, they were originally enrolled for a hepatitis B vaccine trial, which was started between November 1980 and December 1981.<sup>7</sup> Since all participants in this trial had to test negative for serological markers of hepatitis B virus infection at entry, they are probably not representative for all prevalent cases with respect to sexual behaviour. A cohort-based method for estimation of the seroconversion distribution<sup>2</sup> may yield biased results. Therefore, the marker-based approach, using CD4 count,

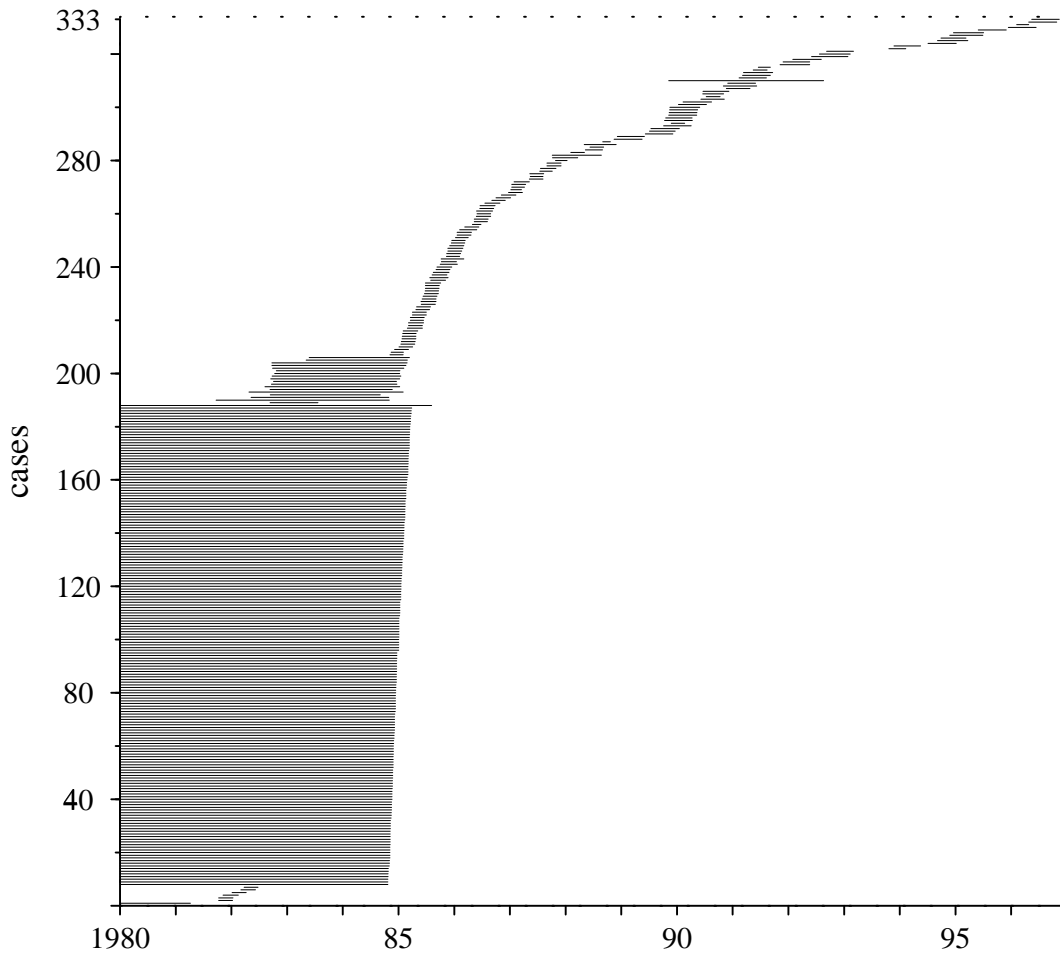


Figure 1: Seroconversion intervals of the persons used in the analysis. Solid lines denote the period between each person’s last seronegative test, or 1980 if no seronegative test is available, and first seropositive test. The vertical ordering is based upon the midpoint date between each person’s last seronegative test (or 1980) and first seropositive test.

is used as alternative method.

We let the reference group consist of the 133 seroconverters whose width of seroconversion interval is less than one year. Since administration of highly active antiretroviral therapy (HAART) influences CD4 trajectories, CD4 measurements from the reference group are used until the earliest date of administration of HAART within this group (November 5th, 1996). The remaining 200 persons constitute the prevalent group. CD4 measurements are available for 198 persons, although two of them only after AIDS diagnosis (three days and one year after diagnosis, respectively). The majority (196 out of 198) has the first CD4 measurement within one year after the first HIV-seropositive test. The exceptions are one prevalent case with a retrospectively identified seropositive test in 1981 and the person with the first CD4 measurement one year after his AIDS diagnosis.

A date of start of the epidemic  $\sigma_0$  has to be assumed. The earliest HIV-positive test result in the Amsterdam hepatitis B vaccine study among homosexual men dates back to December 1980. The influence of both January 1st, 1978 and January 1st, 1980 as choice of  $\sigma_0$  will be investigated. If 1978 is chosen, the time difference between either  $\sigma_0$  or  $v_i^n$  and the first CD4 measurement is at most 8.73 years, with all except three persons having this time interval less than 7.72 years. Using 1980, these numbers are two years smaller. In order to correct for biases due to right censoring, only reference data from persons who seroconverted before February 17th, 1991 are used in case of 1980 (November 5th, 1996 minus 5.72 years), and two years earlier in case of 1978. Also subtracting four persons in the reference group without any CD4 data, amounts to using CD4 data from 104 seroconverters with 2752 values for  $\sigma_0 = 1980$  and 85 persons with 2258 values for  $\sigma_0 = 1978$ .

All computations with respect to CD4 count are per  $\mu L$  and are transformed to a square root scale, since this transformation yields the most symmetric distribution on visual inspection. Values from the reference group have been made equidistant through linear interpolation, taking a gridsize of 0.125 years. Through this procedure, the total number of values increased to 4579 (1980) and 3805 (1978), respectively. The midpoint between the last seronegative and first seropositive test is taken as date of seroconversion. In figure 2, all values within 8.73 years after seroconversion for the reference group are depicted graphically, as well as the CD4 count at first measurement for the prevalent cases, as obtained via isotonic regression. For each prevalent case, the 150 nearest CD4 values from the reference group over the range from 0 to 8.73 years are used to derive a distribution of time since seroconversion, which is subsequently truncated based on  $v_i^n$  (or  $\sigma_0$ ), and  $v_i^p$ . For the two seroprevalent cases without CD4 measurements, a seroconversion distribution is obtained by averaging the other 198 individual seroconversion distributions. Their reason for missing CD4 values is not related to fast disease progression.

### 3.1. Results

The quality of the four estimates (weighed/unweighed and 1978/1980 as start of the epidemic) is determined by a simulation study as described in section 2.1. Using ten simulation runs ( $r = 10$ ), the empirical cumulative distribution functions (ecdf) of predicted CD4 ( $\hat{F}_p$ ) and measured CD4 ( $F_m$ ) are compared. The choice of method is based on two measures, related to mean squared error and bias:

$$\sum_{r=1}^{10} \sum_{j=1}^5 [\hat{F}_m^r(200 \cdot j) - F_m(200 \cdot j)]^2 \quad \text{and} \quad \sum_{r=1}^{10} \sum_{j=1}^5 [\hat{F}_m^r(200 \cdot j) - F_m(200 \cdot j)].$$

The weighed ecdf with 1980 and the unweighed ecdf with 1978 performed best and showed close agreement with  $F_m$  upon visual inspection.

Using the individual seroconversion distributions for the persons in the prevalent group, and assuming a uniform seroconversion distribution for the persons in the reference group, conditional mean imputation, multiple imputation as well as the likelihood based method (methods EXP, RAN and LIK as described in Geskus<sup>2</sup>) can be used to obtain an estimate of the incubation time distribution. In figure 3, survival curves based on the two best performing estimates of individual seroconversion distributions and the one based on the cohort-based estimate of the

seroconversion distribution<sup>2</sup> are plotted, along with the survival curve when the analysis is restricted to the seroconverters with a seroconversion interval of width less than one year. All estimates are based on conditional mean imputation of a date of seroconversion.

We see that the assumed date of start of the epidemic has little influence on the estimate of the incubation time distribution when the marker-based seroconversion distributions are used. Using the cohort-based seroconversion estimates, this difference is negligible (not shown). The estimate based on the cohort-wide seroconversion distribution is fairly similar to the estimate using the marker-based approach.

For the marker-based approach, we also compared the two other methods. Results obtained via the likelihood method (LIK) were almost identical to the results from method EXP. When 1980 is assumed as date of start of the epidemic, the survival curve obtained via multiple impu-

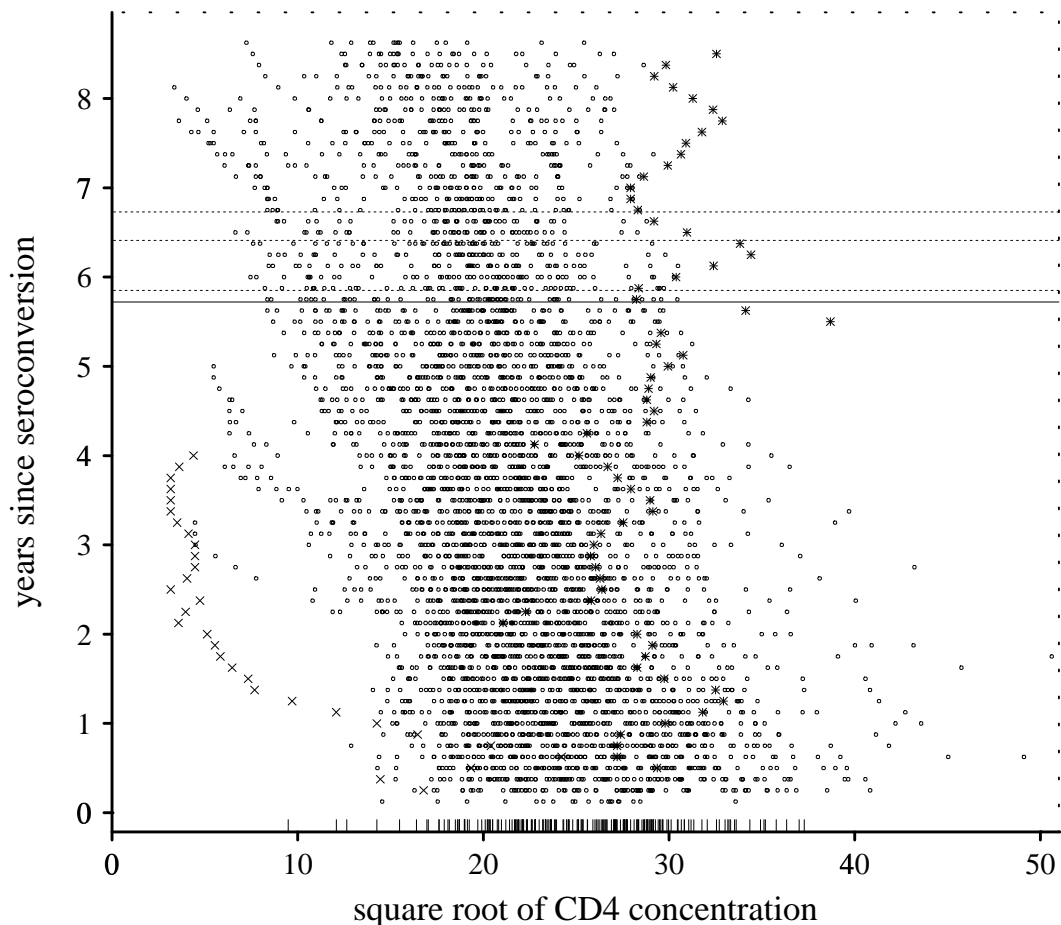


Figure 2: Square root of CD4 counts per  $\mu L$  from the reference group. Values are made equidistant through linear interpolation. The solid horizontal line is at 5.72 years after seroconversion. Assuming 1980 as start of the epidemic, all but three seroprevalent cases have  $d_i - \sigma_0 < 5.72$ . For the remaining three persons, these values are represented by the dotted lines. CD4 numbers at entry for the prevalent cases are given by the vertical tick marks on the x-axis.  $\times$  and  $*$  depict two extreme individual CD4 trajectories.

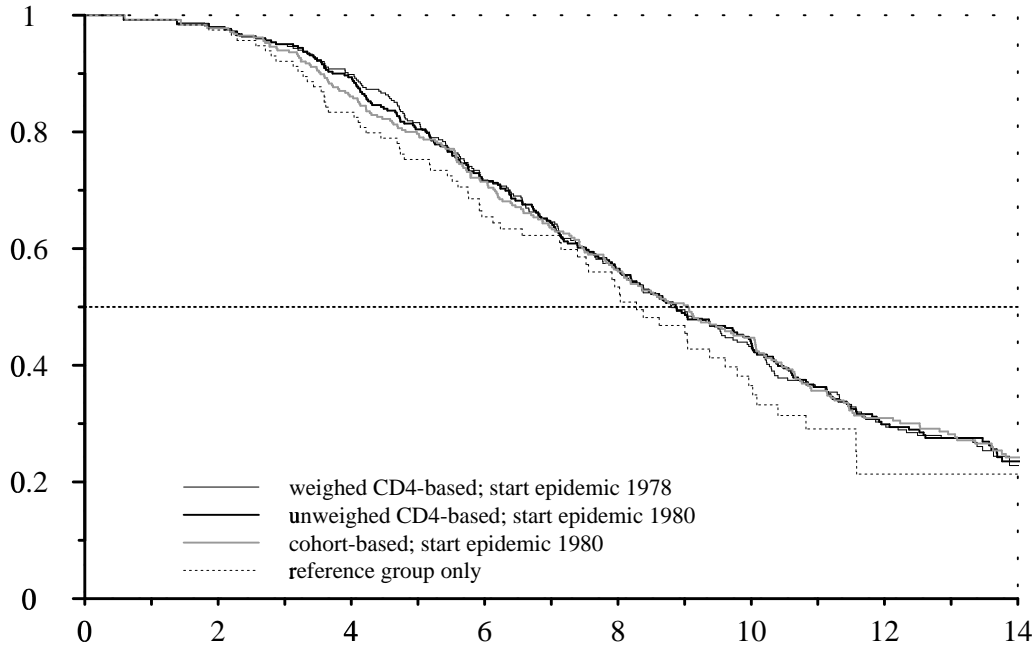


Figure 3: Estimate of incubation time distribution via conditional mean imputation.

tation (RAN, average of 10 sets of imputed values) is slightly lower. For 1978, the difference is more pronounced, with multiple imputation yielding the lowest curve. This is in correspondence with results found in Geskus.<sup>2</sup> All further survival analyses in this paper use conditional mean imputation, based on the individual marker-based estimates of the seroconversion distributions.

The individual seroconversion distributions can be combined to derive a cohort-wide seroconversion distribution for the 748 persons who entered the study under the same protocol 1, i.e. between October 1984 and April 1985. For the persons who entered seropositively, all individual seroconversion distributions are averaged. For the persons who entered seronegatively, the Kaplan-Meier estimator is used with seroconversion as endpoint. Both curves are weighed by the number of cases in each group. The resulting cumulative seroconversion distribution as well as the seroconversion density are depicted in figure 4. The density is obtained via kernel smoothing with the Epanechnikov kernel, using a boundary kernel<sup>8</sup> at the assumed date of start of the epidemic.

Although the influence of the choice of  $\sigma_0$  on the incubation time estimate is little, we pay some more attention to the assumption with respect to the date of start of the epidemic. In 1982, the first Dutch AIDS case was diagnosed in a homosexual man in Amsterdam.<sup>9</sup> Using our estimates of the incubation time distribution and of the seroconversion distribution for the persons in protocol 1, the probability that the first AIDS case occurs in 1982 or later can be computed. This probability depends on the size  $N$  of the homosexual population at risk of HIV

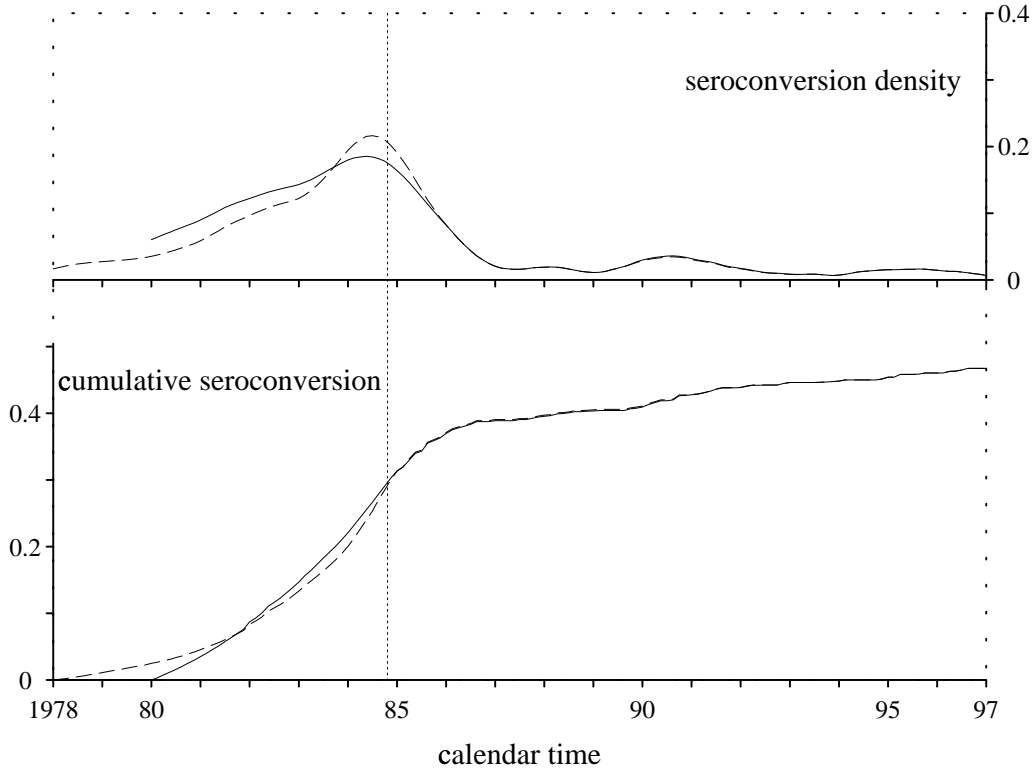


Figure 4: Seroconversion cumulative distribution function and density for the persons who entered the study between October 1984 and April 1985. For the density, kernel smoothing is used with a bandwidth of one year. The vertical dotted line denotes the start of the cohort study.

infection, as represented by the cohort. This probability is computed as

$$\left(1 - \int_{\sigma_0}^{1982} g(s) F(1982 - s) ds\right)^N.$$

The problem is that  $N$  is unknown. In figure 5, the probability is plotted as a function of risk group size. In order to obtain some idea of the amount of uncertainty of these curves, the same probability is computed using the upper (1978) and lower (1980) pointwise confidence limits of the incubation time estimates. Using the upper confidence limit in case of 1980 results in a probability to seroconvert after 1982 of value one, and is therefore not shown. Based on a cross-sectional survey held in 1989, the size of the male homosexual population in Amsterdam has been estimated to be around 21,500 (95% confidence interval: 17,000-26,000).<sup>10</sup> Since the cohort participants were at higher risk of HIV infection,<sup>10</sup>  $N$  is smaller than 21,500. At the other end, already at  $N = 2000$ , the probability of the first AIDS case to have occurred before 1982 is below 0.05 when 1978 is assumed. Although it is possible that some AIDS cases before 1982 have been missed, we think 1980 is more likely than 1978. Hence in the sequel we assume  $\sigma_0 = 1980$ .

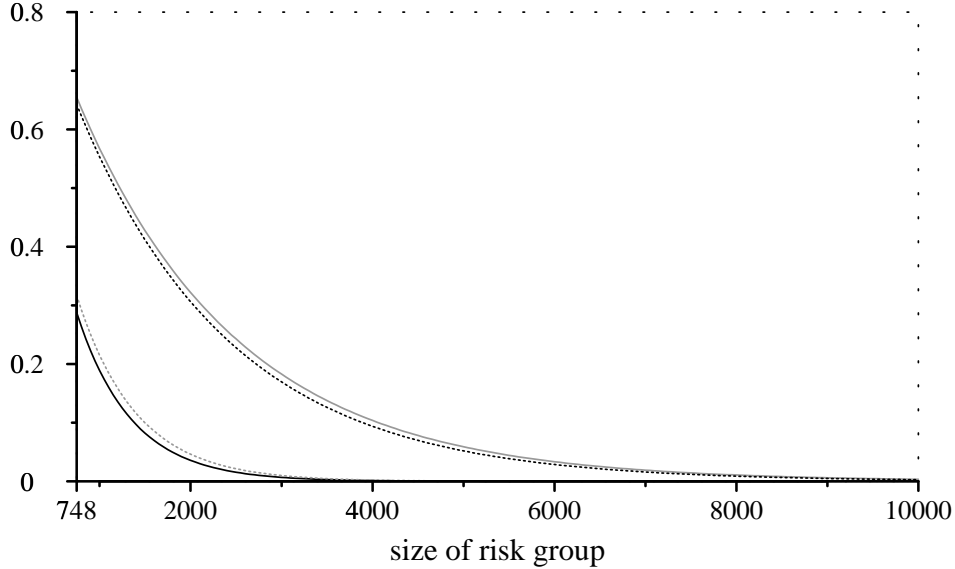


Figure 5: Estimate of the probability that the first AIDS case occurred in 1982 or later, as a function of the size of the risk group. Solid line: epidemic started in 1978; dashed line: epidemic started in 1980. Grey lines are based on upper (1978) and lower (1980) pointwise confidence limits of incubation time estimates.

### 3.1.1. Confidence intervals

In the survival analyses, confidence intervals and p-values, as obtained through standard methods assuming the estimated dates of seroconversion to be the true ones, may be too small. Therefore, we also computed confidence intervals using the following bootstrap procedure.

The bootstrap approximation to the incubation time distribution is our estimate of the incubation time  $\hat{F}$ . The censoring time distribution is approximated by the Kaplan-Meier estimate  $\hat{G}$  of the time from seroconversion to loss to follow-up. In this estimate, all persons who are right-censored with respect to AIDS diagnosis after January 1st, 1996 are considered still to be in follow-up and are therefore considered as right censored with respect to the censoring distribution. They are censored at  $\tau = 1/1/1997$  instead. Dates of seroconversion and dates of cohort entry are drawn from the individual persons.

The distribution of the bootstrap approximation is obtained through simulation. At each iteration, the following steps are performed:

- Sample with replacement  $N = 333$  persons from the cohort.
  - a. **Generation of dates of seroconversion.** If person  $k$  belongs to the prevalent group, we draw a random date of seroconversion  $x_k^*$ , based on his individual seroconversion curve  $g_k$ . Otherwise, we draw a random date of seroconversion from a uniform distribution between the date of last seronegative test and the date of first seropositive test.
  - b. **Generation of incubation times and censoring times.** An incubation time  $t_k^*$  is

drawn from  $\hat{F}$ , a censoring time  $c_k^*$  is drawn from  $\hat{G}$ . We let the event time be the minimum of  $x_k^* + t_k^*$ ,  $x_k^* + c_k^*$  and  $\tau$ . The incubation time is observed if  $x_k^* + t_k^*$  is the smallest of these numbers.

c. **Truncation.** If  $x_k^* + t_k^*$  is smaller than his date of entry, this person is excluded from the estimation procedure.

- **Estimation procedure (seroconversion date).** If person  $k$  belongs to the prevalent group, a date of seroconversion  $s_k^*$  is imputed as the expected date based on his seroconversion curve. Otherwise,  $s_k^*$  is the midpoint between the date of last negative and first positive test.
- **Estimation procedure (incubation time).** The statistic of interest is computed, based on the difference between  $\min\{x_k^* + t_k^*, x_k^* + c_k^*, \tau\}$  and  $s_k^*$ . Truncation times are based on the dates of study entry.

We used 999 bootstrap simulations. Confidence intervals are computed at each year, from years one to thirteen. The results are in figure 6. Results from both the percentile and the basic approach<sup>11</sup> are given. We see that the bootstrap procedure does not result in wider confidence

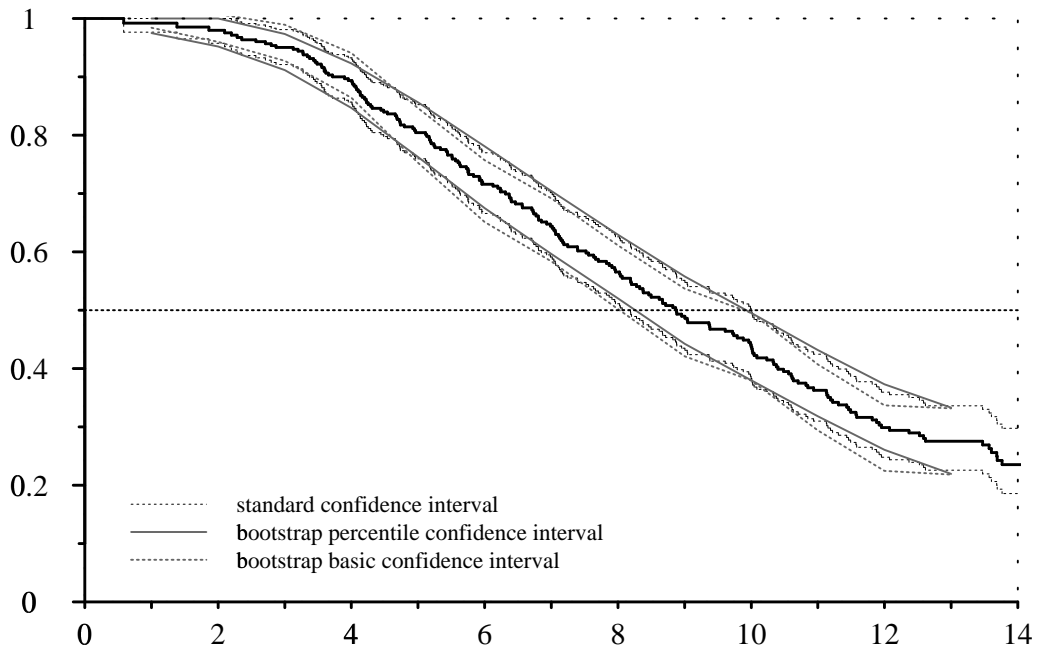


Figure 6: Estimate of incubation time distribution via conditional mean imputation and confidence intervals obtained by ignoring the uncertainty in the date of seroconversion (standard) and by the percentile and the basic bootstrap approximations.

intervals. However, the uncertainty in the individual seroconversion curves has not been incorporated in our procedure. Moreover, it is based on resampling with the same sample size as the

original data, which may fail asymptotically. If asymptotic failure occurs, the amount of incorrectness of the finite sample behaviour remains to be established. Note that the lack of widening of the confidence intervals is in correspondence with results found in a simulation study.<sup>2</sup> An alternative may be to use subsamples and suitably normalise the statistic.<sup>12</sup>

### 3.1.2. Inclusion of covariates

Inclusion of the seroprevalent cases offers the opportunity better to investigate the influence of covariates on the incubation time. Older age at seroconversion has frequently been found to accelerate time to AIDS. However, this result is not found in the Amsterdam cohort study among homosexual men if the analysis is restricted to seroconverters (relative hazard 0.85 per ten-year period, p-value 0.33). The presence of a difference in incubation time between persons who seroconverted early in calendar time and those who seroconverted later can be investigated as well, since a large group of persons is included who seroconverted before the start of the cohort study.

Although the bootstrap confidence bands obtained in the previous subsection hardly differ from the standard ones, covariates like date of seroconversion and age at seroconversion may induce larger differences since they need to be estimated. Used as a covariate, date of seroconversion is not estimated correctly through conditional mean imputation: 94 out of the 205 cases (46%) who were seropositive at entry between October 1984 and April 1985 have their imputed date in 1983, whereas our estimate of the seroconversion pattern gives a probability of 0.23 to seroconvert in this year. Randomly imputed seroconversion dates far better reflect the seroconversion pattern, but yield a biased estimate of the incubation time distribution.<sup>2</sup> Except for one person, all seroprevalent cases have their date of entry before April 1985. Therefore, date of seroconversion is dichotomised into the periods before and after April 1st, 1985. The estimated ages at seroconversion, obtained via conditional mean imputation and via random imputation of a date of seroconversion, do not show any difference in distribution (results not shown). Therefore, this covariate is used as a continuous one.

Note that in general non-significant effects as obtained using standard confidence intervals and p-values remain non-significant if the bootstrap method is used. In a univariate analysis, the log-rank test statistic for a difference in incubation time between persons seroconverting before and after April 1st, 1985 has the value 5.78. Under the standard chi-squared distribution with one degree of freedom, the corresponding p-value is 0.0162. Age at seroconversion is fitted linearly. No deviation from the proportional hazards assumption is found. Moreover, martingale residuals and addition of restricted cubic splines<sup>13</sup> do not show an effect of nonlinear terms. The relative hazard of age at seroconversion per ten-year period is 1.28, with standard 95% confidence interval (CI) [1.05, 1.55]. If both covariates are fitted in a proportional hazards model, the relative hazard of the early seroconverters compared to the late ones is 0.71, with standard 95% CI [0.508, 0.984] and p-value 0.040. The relative hazard of age at seroconversion is 1.24 per ten-year period, with 95% CI [1.02, 1.50] and p-value 0.031.

We slightly adapt the bootstrap procedure from the previous section. The estimates of the relative hazard parameters  $(\hat{\theta}_1, \hat{\theta}_2)$  and baseline distribution function  $\hat{F}_0$  are used as input values. For each person  $k$  with drawn date of seroconversion  $x_k^*$ , we use his date of birth  $b_k$  to obtain his

age at seroconversion  $x_k^* - b_k$ . An incubation time is generated according to the survivor function

$$[1 - \hat{F}_0]^{\exp\{\hat{\theta}_1 (x_k^* - b_k) + \hat{\theta}_2 1_{\{x_k^* \leq 1985.248\}}\}}.$$

The confidence intervals widen slightly. For date of seroconversion, the bootstrap basic CI is [0.488, 0.965], the percentile CI is [0.518, 1.025]; for age at seroconversion, these intervals are [1.003, 1.54] and [1.002, 1.53] respectively.

#### 4. DISCUSSION

In the Amsterdam cohort study among homosexual men, inclusion of the seroprevalent cases who entered the study during the first six months of enrollment considerably increases the amount of information, with respect to both the number of persons and the total amount of follow-up. The number of persons increases from 133 to 333, whereas the percentage of AIDS cases diagnosed until January 1st, 1997 (used as a crude measure of amount of follow-up) increases from 50 (67 cases) to 62 (208 cases). A requirement for inclusion of prevalent cases is that a date of seroconversion can be estimated which is fairly unbiased at the population level.

We used CD4 count at entry to obtain individual seroconversion distributions for the prevalent cases. The problems involved in the use of this method have been mentioned at the end of section 2.

First, the reference group has to be representative for the prevalent group with respect to marker development. Although only marker measurements have been used from the period before the administration of highly active anti-retroviral therapies (HAART), there may still be some influence of treatment on CD4 trajectories. In the reference group, 11 persons had some anti-HIV treatment administered during the first 5.6 years after seroconversion at a CD4 count higher than  $275/\mu L$  (9 were on zidovudine (AZT) monotherapy, one on *Pneumocystis carinii* pneumonia (PCP) prophylaxis and one on both). Note that only five persons from the prevalent group had a CD4 count at entry smaller than  $275/\mu L$ .

Second, an important issue is that the problem is lacking statistical justification. However, by comparing CD4 counts based on the estimated seroconversion curves with true CD4 counts at first measurement, an ad-hoc justification can be provided. We saw good agreement for two of the combinations of method and assumed date of start of the epidemic chosen.

Third, in general, a date of start of the epidemic has to be chosen. Often, the range of plausible values is not large. For our data, 1981 is too late, whereas we showed that 1978 is likely to be too early. A good choice may be 1980. However, even when 1978 is chosen, differences in incubation time estimates are not large.

Fourth, fast disease progressors are overrepresented in the reference group. By comparing estimated and true CD4 counts at first measurement, we already know that our estimated seroconversion distributions are fairly accurate. However, one can also estimate the number of persons that has been missed as

$$\int_{1980}^{1985} \hat{g}(s) \hat{F}(1985 - s) ds \times \text{number in study} = 0.0128 \times 748 = 10.$$

We apply a correction by randomly throwing away

$$10 \times \frac{\text{number in seroconverter group}}{\text{number prevalent at entry}} = 10 \times \frac{104}{238} = 4$$

persons from the reference group who developed AIDS within five years after seroconversion. Note that they are deleted when the seroconversion distributions are estimated, not in the survival analyses. Results with respect to the incubation time distribution did not change noticeably.

Fifth, persons from the reference group may enter the study some time after seroconversion or may be right censored within the time span between the assumed date of start of the epidemic and the latest date of entry of a prevalent case. We largely corrected for this by excluding the recent seroconverters from the reference group. However, seven more persons were right censored because of loss to follow-up or competing risks. Again, deleting these seven persons did not result in noticeably different survival estimates.

Sixth, when the effect of a covariate on the incubation time is investigated, this covariate should not be associated with marker development. We looked at the influence of age at seroconversion on the incubation time. CD4 count and development after seroconversion may depend on age. In the Amsterdam seroconverter group, however, no significant association was found.<sup>14</sup> Moreover, among the persons who were seropositive at entry, no correlation is found between age at entry and imputed date of seroconversion (spearman rank correlation -0.05; p-value 0.51). Finally, using the cohort-based estimate for the imputation procedure only slightly increased age effects (relative hazard 1.29 per ten years).

The last problem, which does not play a role in our data, is that marker measurements may only be available for a subgroup of the prevalent cases. When the probability to have marker measurements is related to disease progression, a bias will be introduced.

Our methods resembles a marker based approach that has been used before.<sup>15-17</sup> The main difference is that we use a nonparametric estimate of the individual seroconversion distributions. Moreover, we also looked at a method that weighs for multiple observations within the same individual. Instead of incorporating some correction for right-truncation effects in the estimation method, we corrected for this effect by restricting the reference group to persons who had sufficient follow-up. Other differences are in the imputation method and in the bootstrap procedure to obtain confidence intervals. We used conditional mean imputation instead of multiple imputation, since this yielded less biased estimates in a simulation study.<sup>2</sup> Instead of using bootstrap approximations to the distribution of time from seroconversion to first seropositive test and time from first seropositive test to event time (AIDS), we used bootstrap approximations to the seroconversion distribution and the incubation time distribution, which we think is more natural. Moreover, their approximands will generally be dependent, which is not incorporated in the procedure. A shortcoming of our procedure is that it does not correct for the uncertainty in the estimates of the seroconversion distributions. A correction can be included, but we left it out for reasons of computing time. However, we do not think it will substantially widen confidence intervals, since our general experience is that the influence of the uncertainty in the seroconversion distributions on the survival distributions is fairly weak. For example, the choice of the date of start of the epidemic (1978 or 1980) was not very important for the survival estimates.

We investigated the influence of both age at seroconversion and year of seroconversion on the

incubation time. In order to obtain better confidence interval estimates, we developed a bootstrap procedure. Confidence intervals were only slightly wider. Both covariates are about borderline significant. Older age accelerates time to AIDS. Persons seroconverting after April 1985 have faster progression to AIDS. However, this effect was highly dependent on the cutoff date. When August 1985 or later was used, p-values were above 0.25 (standard log-rank test). Dichotomising date of seroconversion into the periods before and after April 1985 is not the correct model. Alternatively, we could have used randomly imputed dates as continuous covariate values in the analysis, while retaining the expected dates as starting points in the incubation time scale.

Because of the above mentioned problems, and the greater complexity of the marker-based model, we strongly favour the use of the cohort-based method if sufficient and representative cohort seroconversion data are available over the total relevant calendar period. Stored blood samples may provide very valuable information. In a simulation study, even a fairly small amount of information with respect to seroconversion pattern for the period before the start of the cohort study was enough to obtain almost unbiased estimates. Because in our application stored blood samples originated from a specific subgroup, we also looked at a marker based approach. Estimates of the incubation time distribution and the influence of the covariates were fairly similar.

Even though biases easily occur, in some situations the marker-based method may provide more accurate seroconversion information, since it uses information with respect to the state of the immune system. Our analysis was based on CD4 count only. CD4 count tends to decrease with disease progression. Still, CD4 count is not very discriminating with respect to time since seroconversion. Using a combination of markers, with viral load as the most important addition, may give more accurate estimates with respect to the individual dates of seroconversion. Note however that the improvement in the estimate of the incubation time distribution is likely to be much smaller, since only estimates at the population level play a role.

Inclusion of the seroprevalent cases who entered the Amsterdam cohort study between October 1984 and April 1985 greatly improves the quality of the data. Moreover, standard procedures for the computation of confidence intervals seem to perform quite well. Therefore, inclusion of these persons is highly recommended in future natural history analyses.

## 5. ACKNOWLEDGEMENTS

This study was supported by grants from the Netherlands Foundation for Preventive Medicine (2823700), on advice of the Dutch Program Committee of AIDS Research in the context of the National AIDS Research Stimulation Program. This study was performed as part of the Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration between the Municipal Health Service, the Academic Medical Centre and the Central Laboratory of the Netherlands Red Cross Blood Transfusion Service, Amsterdam, the Netherlands.

We wish to thank Judith Lok, Frits van Griensven, Anneke Krol, Piet Groeneboom and Roel Coutinho for their advice on some aspects of the manuscript. Longhow Lam did a lot of programming work and gave several good suggestions. Furthermore, we wish to thank Nel Albrecht and Ethel Boucher for supplying the data.

## REFERENCES

1. Brookmeyer, R. and Gail, M. H. *AIDS epidemiology: a quantitative approach*, Oxford University Press, 1994.
2. Geskus, R. B. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored, 1999, Submitted for publication.
3. Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F. and Hethcote, H. W. 'Statistical analysis of the stages of HIV infection using a markov model', *Statistics in Medicine*, **8**, 831–843 (1989).
4. Hendriks, J. C. M., Satten, G. A., Longini, I. M., van Druten, H. A. M., Schellekens, P. T. A., Coutinho, R. A. and van Griensven, G. J. P. 'Use of immunological markers and continuous-time markov models to estimate progression of HIV infection in homosexual men', *AIDS*, **10**, 649–656 (1996).
5. Centers for Disease Control 'Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome', *Morbidity and Mortality Weekly Report*, **36**, 1S–15S (1987).
6. Robertson, T., Wright, F. T. and Dykstra, R. L. *Order Restricted Statistical Inference*, Wiley, New York, 1988.
7. Coutinho, R. A., Lelie, P. N., van Lent, P. A., Reerink-Brongers, E. E., Stoutjesdijk, L., Dees, P., Nivard, J., Huisman, J. and Reesink, H. W. 'Efficacy of heat inactivated hepatitis b vaccine in male homosexuals: outcome of a placebo controlled double blind trial', *British Med. Journal*, **286**, 1305–1308 (1983).
8. Gasser, T. and Müller, H. G. Kernel estimation of regression functions, in 'Smoothing Techniques for Curve Estimation', Vol. 757 of *Lecture Notes in Mathematics*, Springer-Verlag, pp. 23–68, 1979.
9. Prummel, M. F., ten Berge, R. J. M., Barrowclough, H. and Cejka, V. 'Kaposi sarcoom en dodelijke opportunistische infecties bij een homoseksuele man met een deficient immuunapparaat', *Ned. Tijdschr. Geneesk.*, **127**, 820–824 (1983).
10. Veugelers, P. J., van Zessen, G., Hendriks, J. C. M., Sandfort, T. G. M., Coutinho, R. A. and van Griensven, G. J. P. 'Estimation of the magnitude of the HIV epidemic among homosexual men: utilization of survey data in predictive models', *European Journal of Epidemiology*, **9**, 436–441 (1993).
11. Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and Their Applications*, Cambridge University Press, 1997.
12. Politis, D. N. and Romano, J. P. 'Large sample confidence regions based on subsamples under minimal conditions', *Ann. Statist.*, **22**, 2031–2050 (1994).
13. Harrell, F. E. 'Design: S functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, and prediction', 1998. Programs available from <http://hesweb1.med.virginia.edu/biostat/s/index.html> or from [http://lib.stat.cmu.edu/\(DOS\)/S](http://lib.stat.cmu.edu/(DOS)/S).

14. Veugelers, P. J., Strathdee, S. A., Kaldor, J. M., Shafer, K. A., Moss, A. R., Schechter, M. T., Schellekens, P. T. A., Coutinho, R. A. and van Griensven, G. J. P. ‘Associations of age, immunosuppression and AIDS among homosexual men in the tricontinental seroconverter study’, *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **14**, 435–441 (1997).
15. Muñoz, A., Wang, M.-C., Bass, S., Taylor, J. M. G., Chmiel, J. S., Kingsley, L., Polk, B. F. and The Multicenter AIDS Cohort Study Group ‘Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men’, *Am. J. Epidemiol.* (1989).
16. Muñoz, A., Carey, V., Taylor, J. M. G., Chmiel, J. S., Kingsley, L., van Raden, M. and Hoover, D. R. ‘Estimation of time since exposure for a prevalent cohort’, *Statistics in Medicine; Wiley (New York)*; (1992).
17. Alcabes, P., Muñoz, A., Vlahov, D. and Friedland, G. H. ‘Incubation period of human immunodeficiency virus’, *Epidemiologic Reviews*, **15**, 303–318 (1993).

## 6. APPENDIX

The unweighed version of the method described has been motivated in the past by saying that ignoring the within-individual dependence structure still yields consistent estimates.<sup>15</sup> Although this may be true when the repeated marker values are regressed on time, this no longer holds when time is regressed on marker values. The following artificial example, in which the weighed and the unweighed approach are equal, yields results that are clearly biased.

Let the marker value distribution at seroconversion be

$$\mathbb{P}\{k\} = 0.2, \quad k = 6, 7, 8, 9, 10.$$

Marker values decrease after seroconversion in a deterministic way with a slope of one unit per year. Suppose an exponential increase of the epidemic, with total number of seroconversions per year given in the rightmost column “total sc” of table 1. Then the marker values at entry (=1984) have the distribution as given in the row “at entry” in table 1 (divided by 310).

Suppose a prevalent case has value 7 at entry. Based on the marker information from the seroconverters, his seroconversion distribution is uniform on the values

$$\{1981, 1982, 1983, 1984\},$$

which is quite different from the true pattern

$$\mathbb{P}\{1981\} = 4/60, \quad \mathbb{P}\{1982\} = 8/60, \quad \mathbb{P}\{1983\} = 16/60 \quad \text{and} \quad \mathbb{P}\{1984\} = 32/60.$$

Only if the true seroconversion distribution were uniform would both answers agree.

Our procedure to check correctness of the estimate can easily be performed on this example. The last row of table 1 (“reconstructed”), divided by 310, is the distribution of marker values at entry, obtained by assuming the estimated (uniform) individual seroconversion distributions to be correct. Results are clearly different from the true distribution of marker values at entry.

| year          | marker values at entry |    |    |    |    |    |    |    |    | total sc |
|---------------|------------------------|----|----|----|----|----|----|----|----|----------|
|               | 2                      | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |          |
| 1980          | 2                      | 2  | 2  | 2  | 2  | 0  | 0  | 0  | 0  | 10       |
| 1981          | 0                      | 4  | 4  | 4  | 4  | 4  | 0  | 0  | 0  | 20       |
| 1982          | 0                      | 0  | 8  | 8  | 8  | 8  | 8  | 0  | 0  | 40       |
| 1983          | 0                      | 0  | 0  | 16 | 16 | 16 | 16 | 16 | 0  | 80       |
| 1984          | 0                      | 0  | 0  | 0  | 32 | 32 | 32 | 32 | 32 | 160      |
| at entry      | 2                      | 6  | 14 | 30 | 62 | 60 | 56 | 48 | 32 | 310      |
| reconstructed | 6                      | 14 | 26 | 42 | 62 | 56 | 48 | 36 | 20 | 310      |

Table 1: Distribution of marker values at entry (1984), stratified by year of seroconversion.

This result can be computed by hand, but is also easily obtained through some matrix multiplications. Let  $M$  denote the  $5 \times 9$  matrix of marker values at entry, stratified by year of seroconversion, as given in table 1. Let the yearly number of seroconversions be denoted by the vector  $x$ ,  $x^T = (10, 20, 40, 80, 160)$ . Let  $M^T$  and  $x^T$  denote matrix and vector transpose. Let “ $*$ ” denote a matrix by vector multiplication given by  $M * x = (a_{ij})$  with  $a_{ij} = m_{ij} \times x_i$ , and let the operator “/” define a similar type of matrix by vector division. The estimate of the seroconversion distribution, conditionally on each marker value at entry, is obtained as  $(M/x)^T / \text{colsum}(M/x)$ .

The last row of table 1 can be obtained via

$$\text{rowsum} \left[ (M/x)^T * \text{colsum} \left\{ [(M/x)^T / \text{colsum}(M/x)] * \text{colsum}(M) \right\} \right].$$