STATISTICS IN MEDICINE
Statist. Med. **19**, 175–189 (2000)

UNCERTAINTY OF AIDS INCUBATION TIME AND ITS EFFECTS ON BACK-CALCULATION ESTIMATES

ANNA GIGLI[1]* AND ARDUINO VERDECCHIA[2]

[1] *Istituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, 00161 Roma, Italy*
[2] *Laboratorio di Epidemiologia e Biostatistica, Istituto Superiore di Sanita, Viale Regina Elena 299, 00161 Roma, Italy*

SUMMARY

Incubation time is the period from the onset of HIV infection to AIDS. The distribution of the incubation time is one of the main parameters of the back-calculation method for the estimation of incidence of HIV infection. Because of the long and variable incubation time, the assessment of its distribution is uncertain and this uncertainty spreads through the back-calculation method and affects the estimation of the precision of incidence of HIV infection. We propose a method to investigate the sensitivity of the estimates to variations of the incubation times, with particular regard to the covariate AGE in the modelling of the incubation period, making use of the parametric bootstrap. An application to the HIV epidemic in Italy is presented. The amplification of the uncertainty of the HIV incidence estimates resulting from the implementation of our proposed method tends to concentrate around the earlier periods of the epidemic, corresponding to the right tail of the incubation time distribution, which is very sensitive to small perturbations.

1. INTRODUCTION

Modelling HIV infection incidence greatly improved the potential surveillance systems in most countries. Back-calculation methods, originally proposed in HIV estimation by Brookmeyer and Gail,[1,2] were among those most appreciated for their simplicity and flexibility. Reliability of estimates obtained by back-calculation from AIDS counts and incubation time distribution have been studied in several countries, with reference to the quality of data available and the completeness of AIDS notifications. A review of sources of uncertainty affecting back-calculation procedures can be found in Brookmeyer and Gail.[3]

Knowledge of incubation time was indicated as a major problem, being an important source of uncertainty in back-calculation estimates. Most of the uncertainty involved in the estimation of the incubation time distribution parameters is due to the rather short observation period available (usually 10–15 years) with respect to 8–12 year estimated median times from HIV to AIDS. This uncertainty spreads through the back-calculation method and affects the estimation of the precision of HIV incidence.

* Correspondence to: Anna Gigli, Istituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, 00161 Roma, Italy. E-mail: GIGLI@IAC.RM.CNR.IT

Contract/grant sponsor: Ministero della Sanita, Istituto Superiore di Sanita, IX Progetto AIDS, 1996

CCC 0277–6715/2000/020175–15$17.50

Received October 1997
Accepted March 1999

Age at onset of HIV infection was indicated to be the most important determinant of the progression of infected people to AIDS (see Mariotto *et al.*[4]). Age at infection is an important issue to consider in the estimation of the HIV infection for more general public health implications. It varies by risk category and by sex, and it can also vary by calendar time, as the epidemic goes on. Also, incubation time varies accordingly.

Estimates of the incubation time distribution were derived by cohort studies in several countries (see, among others, Bacchetti and Moss,[5] Giesecke *et al.*,[6] Blaxhult *et al.*[7] and Darby *et al.*[8]), involving different case inclusion criteria, different uncertainty about the knowledge of the exact time of infection, different lengths of follow-up time, and were found to be quite consistent between each other.

Back-calculation methods able to accommodate age at seroconversion as a covariate were developed in non-parametric form (see Rosenberg[9] and Becker and Marschner[10]) and in parametric form (Verdecchia and Mariotto[11] and Verdecchia *et al.*[12]).

The bootstrap, introduced by Efron[13] in 1979, is a computer-intensive method to obtain standard errors, confidence intervals, and other measures of uncertainty in many problems where analytical calculations are not feasible. For a recent review of the bootstrap methods see Davison and Hinkley.[14]

In this work we present a method for studying the effect of the epidemiological uncertainty of incubation time distribution to back-calculation estimates. Starting from the parametric model illustrated by Verdecchia and Mariotto,[11] we make use of the parametric bootstrap to resample parameter values from the incubation time distribution and use them in the back-calculation equations. The resulting estimated standard errors of the incidence estimates should incorporate the uncertainty due to the estimation of the incubation distribution.

Moreover we investigate the specific role of age by checking whether the additional uncertainty included in the back-calculation of HIV infections by considering age-specific incubation times is balanced by greater improvements in the quality and the interpretation of the results.

In the next section we describe the problem by means of an example in the situation of the epidemic curve in Italy. In Section 3 we illustrate the methods applied, with special attention to the non-standard application of the bootstrap. The results are discussed in Section 4.

2. THE EPIDEMIC CURVE IN ITALY

The HIV/AIDS epidemic in Italy, as in most Mediterranean countries, was characterized by there being a great majority of intravenous drug users (IDU) among AIDS cases and HIV infected people as revealed by epidemiological surveys and surveillance systems. Notification to the Italian National Registry of AIDS cases (RAIDS) held by the Istituto Superiore di Sanita', Rome, is mandatory for any diagnosis of AIDS occurring in Italy and reports on AIDS surveillance are published quarterly.

AIDS counts by sex and risk category are individually available for the period 1983–1994. Data are classified into seven risk categories: intravenous drug users (IDU) males and females; men who have sex with men (MSWM); males and females by heterosexual contact transmission (HST); total males and total females.

An estimate of the incubation time distribution is available from data of the Italian Seroconversion Cohort Study by using the method described by Mariotto *et al.*[4] to more recent data.

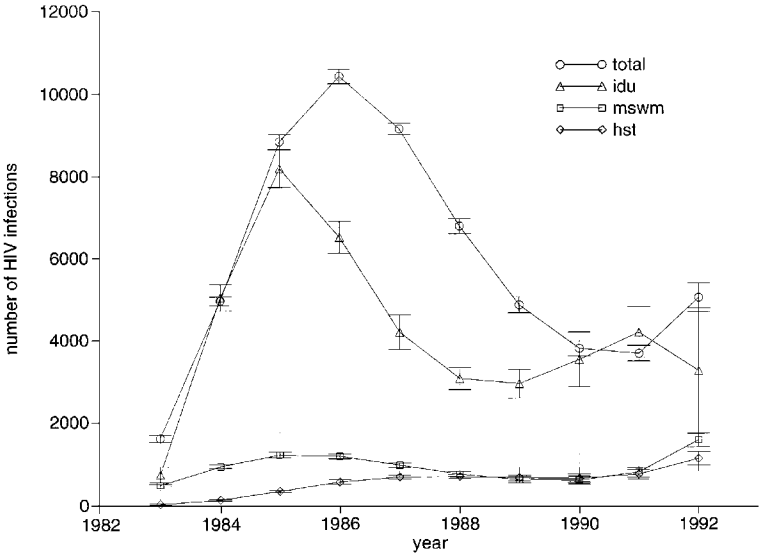

Figure 1. Plot of the HIV incidence and asymptotic standard errors for the subgroups of the male population

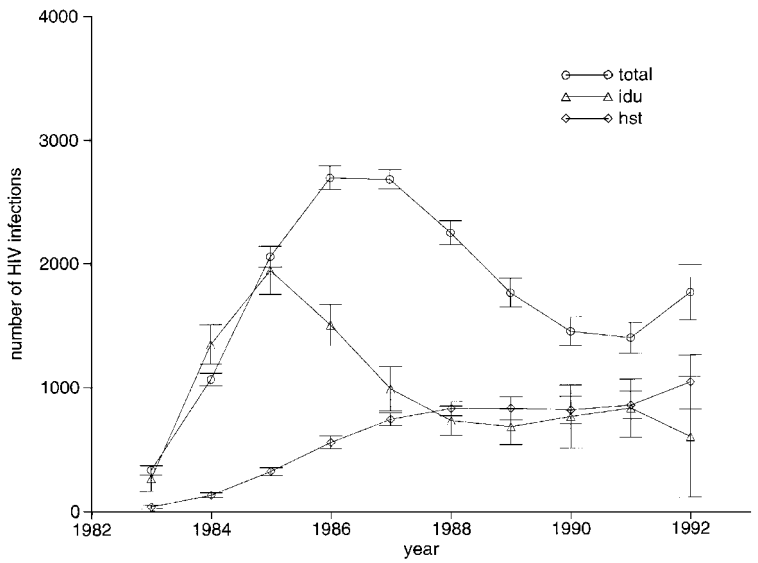

Figure 2. Plot of the HIV incidence and asymptotic standard errors for the subgroups of the female population

National estimates and projections of HIV epidemics have been given in Italy by using a back-calculation approach which incorporates age at onset of HIV infection and its influence on the incubation time distribution, susceptible population and competitive non-AIDS mortality[11]. Figures 1 and 2 report the estimated HIV epidemic curves by sex and risk category. The

spread of HIV infections started earlier among IDU with the epidemic curve peaking in 1985, both for males and females. HIV incidence among MSWM also peaked in 1985, at a much lower level than for IDU, with the epidemic curve not showing an clear decline in recent years. The incidence of HIV infections attributed to HST started later as secondary epidemics, possibly still increasing in recent years.

Asymptotic confidence intervals show that most of the uncertainty of the estimates is concentrated in recent years near data truncation, as occurs with any back-calculation procedure. Based on asymptotic confidence intervals, which approximately includes the statistical uncertainty deriving from the maximum likelihood estimation process, estimates for early years of the epidemic, near the year of peak, are apparently stable and highly reliable.

The incubation time distribution we used was assumed as fixed in the estimation process. To take into consideration the additional epidemiological uncertainty of the sample incubation time distribution estimate is the principal object of this work.

3. METHODS

3.1. The incubation time

We consider here an estimate for the incubation time distribution from Mariotto *et al.*,[4] as based on a three-state Markov process. The transition time from seroconversion to ARC (AIDS-related complex) and from ARC to AIDS are assumed to have a Weibull distribution with hazard functions, respectively:

$$\begin{aligned} h_1(t; z) &= \rho\lambda_1(\lambda_1 t)^{\rho-1}\exp(z\beta) \\ h_2(t; z) &= \rho\lambda_2(\lambda_2 t)^{\rho-1}\exp(z\alpha) \end{aligned} \tag{1}$$

where z is a vector of covariates and $\psi = (\lambda_1, \lambda_2, \rho, \beta, \alpha)$ the parameter vector to be estimated. Mariotto *et al.*[4] notice that a different ρ could have been considered for the two transitions, allowing the possibility of different rates of evolution. However, this hypothesis did not seem essential on the basis of the actual knowledge of the disease, apart from complicating the formulae.

The model was fitted to data from a multi-centre cohort study of 898 individuals at risk of HIV infection who seroconverted between 1982 and 1990 (Verdecchia and Mariotto[11]). The best fit is obtained for a model that includes the categorical covariate AGE at seroconversion, grouped in the categories [16, 24], [25, 34], > 35. This covariate was found significant in the first transition only, that is, $\beta = (0, \beta_1, \beta_2)$, $\alpha = (0, 0, 0)$ and z is a dummy variable that indicates the age group. The values of the maximum likelihood estimates of the parameters are reported in Table I, together with their standard errors (in brackets). The estimated correlation matrix is shown in Table II.

Figure 3 represents the estimated survival distribution function of the incubation period to AIDS for the three age groups. The three curves differ as the number of years after seroconversion increases. The highest curve represents the youngest age group and corresponds to the baseline curve. The other two curves represent the middle and oldest age groups, respectively.

Table I. Maximum likelihood estimates of the transition time parameters and their standard errors

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\rho}$	$\hat{\beta}_1$	$\hat{\beta}_2$
$7{\cdot}510 \times 10^{-3}$	$2{\cdot}339 \times 10^{-3}$	1·50	0·446	0·892
$(1{\cdot}846 \times 10^{-3})$	$(0{\cdot}368 \times 10^{-3})$	(0·184)	(0·431)	(0·491)

Table II. Maximum likelihood estimate of the correlation matrix

	λ_1	λ_2	ρ	β_1	β_2
λ_1	1	− 0·059	0·45	− 0·647	− 0·561
λ_2	− 0·059	1	− 0·35	− 0·031	− 0·086
ρ	0·45	− 0·35	1	0·126	0·094
β_1	− 0·647	− 0·031	0·126	1	0·557
β_2	− 0·561	− 0·086	0·094	0·557	1

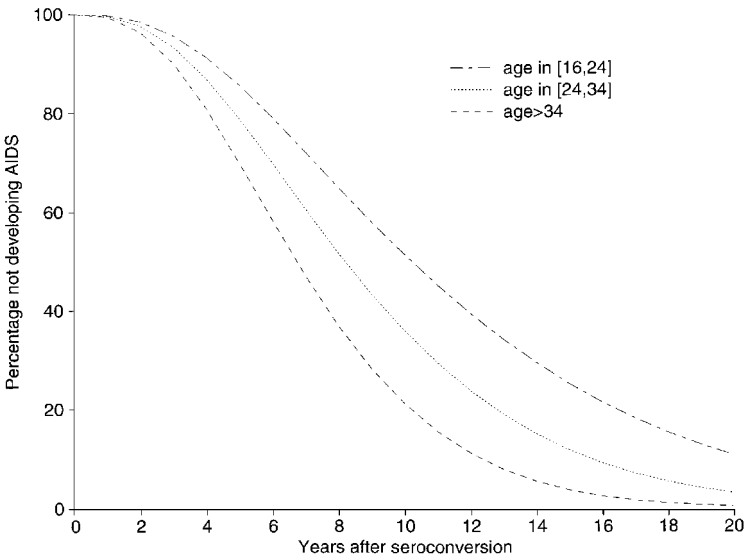

Figure 3. Estimated survival distribution function of the incubation period to AIDS

3.2. Modelling HIV incidence via back-calculation

In this paper we make use of the model illustrated in Verdecchia and Mariotto,[11] with a modification that takes into account the variability of the incubation time parameters. We briefly recall here the method essentials.

For year t and individuals of age i the hazard of developing AIDS, γ_{it}, is obtained from the reported AIDS cases. The hazard of becoming infected with HIV, μ_{it}, is to be estimated, together

with the unknown proportion of HIV infected individuals, v_{it}. Once infected we assume an irreversible progression to AIDS. The density function $\tau_{jt(i-j)}$ of the time to AIDS at age i for individuals infected at age j and year t is usually estimated from follow-up studies on infected subjects and therefore it is assumed known. For individuals born in year $p = t - i$, the observed probability of developing AIDS and the proportion of HIV infected people are given by convolution equations

$$\gamma_{it} = \sum_{j=0}^{i} [1 - v_{j(p+j)}] \mu_{j(p+j)} \tau_{j(p+j)(i-j)} \sigma_{ij(p+j)} \tag{2}$$

and

$$v_{it} = \sum_{j=0}^{i-1} [1 - v_{j(p+j)}] \mu_{j(p+j)} \sigma_{ij(p+j)} \tag{3}$$

where σ_{ijt} represents the probability of surviving the extra risk of death associated with HIV infection at age i for infected non-AIDS individuals, who were infected since age j and year t, which depends on both the incubation time distribution and the competing mortality.

We assume that the observed number Y_{it} of AIDS cases diagnosed in year t at age i is Poisson distributed with mean $\mathbb{E}(Y_{it}) = N_{it} g_{it}(\theta)$, where N_{it} is the AIDS-free population and $g_{it}(\theta)$ is a function of the parameter vector θ which is described below. We call $\hat{\mu}_{it}(\theta)$ the (unobserved) number of HIV cases diagnosed in year t at age i and assume that

$$\mathbb{E}(\hat{\mu}_{it}(\theta)|\psi) = \mu_{it}(\theta) \tag{4}$$

$$\mathrm{var}(\hat{\mu}_{it}(\theta)|\psi) = v_{it}^2(\theta) \tag{5}$$

where ψ is the vector of the transition time parameters described in Section 3.1.

The HIV incidence function is assumed to belong to a family of polynomial functions of age, year and birth cohort (= year − age) on the logistic scale

$$\log\left(\frac{\mu_{it}}{1 - \mu_{it}}\right) = \mathrm{const} + \sum_{k=1}^{K_1} a_k \,(\mathrm{AGE})^k + \sum_{k=1}^{K_2} b_k \,(\mathrm{YEAR})^k + \sum_{k=2}^{K_3} c_k \,(\mathrm{COHORT})^k. \tag{6}$$

We call $\theta = (\mathrm{const}, a_1, \ldots, a_{K_1}, b_1, \ldots, b_{K_2}, c_2, \ldots, c_{K_3})$ the vector of incidence parameters.

Having assumed known (and therefore fixed) the contribution to the incubation time in the back-calculation equations (2) and (3), the standard error of the HIV incidence estimates, v_{it}, can be interpreted as the statistical error due to the fitting of the model.

3.3. The parametric bootstrap

Because of the long and variable incubation times and the lack of observation in the tail of the distribution (the data were collected from 1982 to 1990, and for some age groups we are not yet observing the second half of their distribution), the assessment of $\tau_{jt(i-j)}$, the density function of developing AIDS after $i - j$ years from seroconversion for an individual infected at age j and time t, is uncertain and this uncertainty must reflect into the estimation of the precision of HIV incidence estimates. In the standard back-calculation method the asymptotic variance of the HIV incidence is computed by inverting the Fisher information matrix related to the incidence parameters in (6).

However, a fuller formulation of $\mathrm{var}(\hat{\mu})$ should take into account sources of variation related to both μ and τ. Therefore the following variance decomposition formula (Bickel and Doksum,[15]

p. 36) applies:

$$\mathrm{var}(\hat{\mu}) = \mathrm{var}_{\psi}[\mathbb{E}(\hat{\mu}|\psi)] + \mathbb{E}_{\psi}[\mathrm{var}(\hat{\mu}|\psi)] = \mathrm{var}_{\psi}(\mu) + \mathbb{E}_{\psi}(v^2) \tag{7}$$

following notation of (4) and (5).

An approximation to the two summands in (7) can be found by bootstrapping the incubation time distribution.

Let $\hat{\psi} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\rho}, \hat{\beta}_1, \hat{\beta}_2)$ be the vector of the maximum likelihood estimates of the parameters of the transition time distribution. Its estimated covariance matrix $\hat{C}$ is obtained from the estimated correlation matrix of Table II and the estimated variances of the components of $\hat{\psi}$ of Table I.

The full parametric bootstrap paradigm consists of:

(a) resampling a set $T^* = \{t_1, \ldots, t^*_{898}\}$ of incubation times from a distribution $F_{\hat{\psi}}$, whose hazard functions are described by (1), and the parameter vector ψ is replaced by its maximum likelihood estimate;
(b) estimating the parameter vector $\psi^*_f = \hat{\psi}(t^*_1, \ldots, t^*_{898})$ using the same estimation procedure implemented to obtain $\hat{\psi}$, applied to the new data set T^*;
(c) plugging the new incubation time estimates ψ^*_f into the algorithm that estimates the HIV incidence via back-calculation, obtaining $\mu^*_f = \mu(\psi^*_f, \hat{\theta})$, where $\hat{\theta}$ is the vector of the maximum likelihood estimates of the incidence model (6).

By repeating the resampling of T^* and the computation of μ^*_f independently B times we obtain $\mu^{*(1)}_f, \ldots, \mu^{*(B)}_f$, and the bootstrap estimate of $\mathrm{var}_{\psi}(\mu)$ is

$$\mathrm{var}^*(\mu^*_f) = \frac{1}{B-1}\sum_{b=1}^{B}(\mu^{*(b)}_f - \bar{\mu}^*_f)^2 \tag{8}$$

where $\bar{\mu}^*_f = (1/B)\sum_{b=1}^{B}\mu^{*(b)}$ is the average of the B bootstrap incidences. This approximates the first summand of (7).

The second summand of (7), $\mathbb{E}_{\psi}(v^2)$, can be approximated by

$$\mathbb{E}^*(v^{*2}_f) = \sum_{b=1}^{B} v^2(\psi^{*(b)}_f)\omega(\psi^{*(b)}_f) \tag{9}$$

where $v^2(\psi^*) = \mathrm{var}(\mu^*_f|\psi^*_f)$ is the asymptotic variance (5) conditioned on ψ^*_f instead of $\hat{\psi}$ and $\omega(\psi^*_f)$ are normalized weights suitably chosen in order to give more importance to the ψ^*_f's which are closer to $\hat{\psi}$:

$$\omega(\psi^*_f) = \frac{\prod_{i=1}^{5}\exp\left\{-\frac{1}{2}\left[\frac{\psi^*_{f,i} - \hat{\psi}_i}{\sqrt{(\mathrm{var}\hat{\psi}_i)}}\right]^2\right\}}{\sum_{b=1}^{B}\prod_{i=1}^{5}\exp\left\{-\frac{1}{2}\left[\frac{\psi^{*(b)}_{f,i} - \hat{\psi}_i}{\sqrt{(\mathrm{var}\hat{\psi}_i)}}\right]^2\right\}} \tag{10}$$

From (7), (8) and (9) we obtain an estimate of the overall variance of the HIV incidence

$$\mathrm{var}(\hat{\mu}) \approx \frac{1}{B-1}\sum_{b=1}^{B}(\mu^{*(b)}_f - \bar{\mu}^*_f)^2 + \sum_{b=1}^{B} v^2(\psi^{*(b)}_f)\omega(\psi^{*(b)}_f) \tag{11}$$

Notice that if we were not to employ any bootstrap resampling, that is $\psi_f^{*(b)} = \hat{\psi}$ and $\mu_f^{*(b)} = \hat{\mu}$ for $b = 1, \ldots, B$, from (8) we would have had $\text{var}^*(\mu_f^*) = 0$, from (10) $\omega(\psi_f^{*(b)}) = 1/B$ and therefore (11) would have become $\frac{1}{B}\sum_{b=1}^{B} v^2(\hat{\psi}) = v^2$, as expected.

From a practical point of view the implementation of the full parametric bootstrap paradigm is rather complicated, because the incubation time is modelled as a three-state Markov process and its distribution function F_ψ is a convolution of two distributions. Therefore the resampling should be implemented form a Markov process and should take into account censored data. Moreover the maximum likelihood estimate $\hat{\psi}$ is computed via a non-linear algorithm and it is quite time-consuming.

We chose a simplified paradigm, which consists of resampling the vector ψ^* from a normal distribution with mean vector $\hat{\psi}$ and covariance matrix $\hat{C}$. Since the maximum likelihood estimate $\hat{\psi}$ is asymptotically normally distributed, and so are the vectors $\psi_f^{*(1)}, \ldots, \psi_f^{*(B)}$ obtained from the full paradigm, we can say that the variability of ψ_f^* and ψ^* is of the same order. The simulation above will remain the same, but ψ^* will replace ψ_f^*.

3.4. Technical aspects of the implementation

The bootstrap approach is now widely spread in the statistical world for its flexibility and independence from model assumptions. In the classical setting of bootstrap estimation we have some known function of the sample, let us call it $h(x)$, such as the sample mean, median or a more complex function, and we are interested in assessing some kind of error of $\hat{h}(x)$, an estimate of $h(x)$. Usually we solve the problem by Monte Carlo simulations: we repeatedly resample from the original sample, obtain replicates of $h^* = h(x^*)$ and estimate the sampling error of the replicates.

In our situation, however, we resample the parameters of a distribution, the incubation time distribution, and compute the variance of the HIV incidence function, which is linked to this distribution through a convolution equation (2). Because of this non-standard application of the bootstrap method, greater care must be paid to the implementation aspects, which are dealt with more extensively in Gigli and Verdecchia.[16]

We recall here that the link between the parameters ψ of the incubation time distribution and the HIV incidence estimate μ cannot be explicitly expressed. Formally it is given by the back-calculation equations, which can be considered a kind of 'black box', and all we know is that the relationship between $\tau(\psi)$ and $\mu(\theta)$ is continuous. However, we need to check whether the relationship is also montonic, that is whether a large value of the input vector ψ^* corresponds to a large value of the output variable μ^*.

Figure 4 illustrates a plot summarizing two characteristics that are linked to the input and the output values, respectively: the bootstrap median incubation time of a given age class and the bootstrap HIV prevalence for a given year, which is defined as the cumulative HIV incidence minus the AIDS cases and deaths.

Obviously the correspondence between the two features cannot be bijective, as is evident from the figure, because different sets of bootstrapped parameters of the incubation time distribution can refer to the same median incubation time. Also, with increasing median incubation times, sensitivity of the estimates is enhanced and scattering of the points increases. The only purpose of this scatter plot is to ensure that a short incubation period corresponds to a small HIV prevalence and a longer incubation period corresponds to a larger HIV prevalence.

Because of the structure of the transition time distributions (1), we need not resample the five parameters λ_1, λ_2, ρ, β_1, β_2, at the same time; by bootstrapping the baseline parameters λ_1, λ_2,

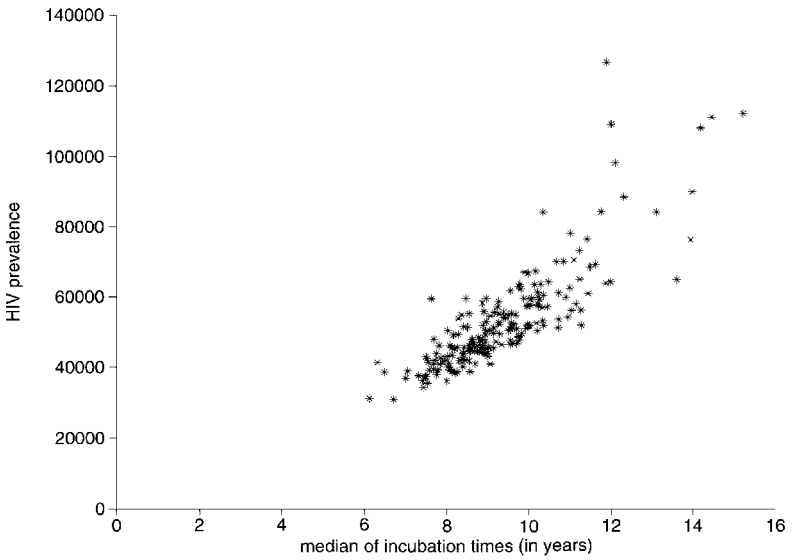

Figure 4. Plot of HIV prevalence for the age group [25, 34) and the year 1992 versus median incubation times

ρ and the AGE parameters β_1, β_2 separately we can isolate the covariate AGE at seroconversion and investigate its contribution towards the assessment of the precision of the HIV incidence estimates.

With reference to Figure 3, the different role played by the two subgroups of baseline and AGE parameters is quite evident; if we fix $\hat{\beta}_1$ and $\hat{\beta}_2$ and bootstrap only the parameters λ_1, λ_2, ρ we obtain a shift of the baseline curve, while the distance between the curves remains the same. If we fix $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\rho}$ and bootstrap β_1 and β_2, the baseline curve does not move and the other two curves move closer or further from it.

In the next section we will refer to the *baseline bootstrap* and the *age bootstrap*, respectively, while a bootstrapping of the five parameters is the *total bootstrap*.

4. RESULTS

The variation in the parameters of the incubation time distribution caused by the bootstrap resampling introduces a significant variation in the estimated median incubation times m^*.

Figure 5 represents the histograms of 200 values of m^* for each of the three age groups.

The range of values of the three bootstrap medians is very high (they take values between 1 and 24 years, and over), but they tend to concentrate around the corresponding maximum likelihood estimates $\hat{m}$.

It is interesting to notice that the three histograms have a slightly log-normal shape, with a longer right tail.

Tables III–IX report for each of the seven subgroups the bootstrap standard errors of the HIV incidence estimates obtained via the back-calculation method (Verdecchia and Mariotto[11]): SE* is the square root of var(μ^*), as obtained in (11), $\%B$ represents the percentage of the variance explained by the bootstrap, and it is computed as the ratio between the bootstrap variance given

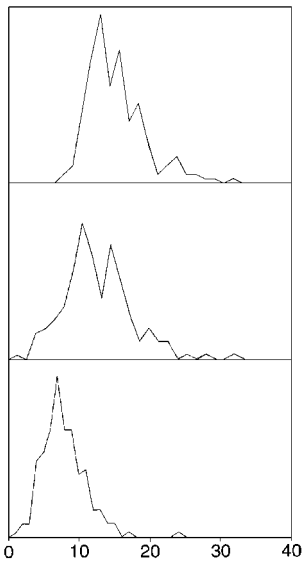

Figure 5. Histograms of the median incubation times over 200 bootstrap replications for the age group [16, 24] (top), age group [25, 34] (middle), age group $\geqslant 35$ (bottom)

by (8) and the overall variance given by (11). It is a useful indicator of how the extra uncertainty added into the model through varying the incubation times affects the overall variance estimate. The bootstrap standard errors are split into three components: total, baseline and age, as described in Section 3.4. In the tables they are indicated as SE^*_{TOT}, SE^*_{BASE}, SE^*_{AGE}, respectively.

The results clearly show two different behaviours; for some subgroups the bootstrap standard error is only a little greater (never more than twice) than the asymptotic standard error, for others the increment is substantial (between 2 and 7 times).

The incubation times are assumed constant throughout the subgroups, provided that age is taken into account (Mariotto *et al.*[4]) and as generally found in the literature, therefore do not help to explain different results obtained in the subgroups.

The increase of the standard errors is then linked to the subgroup's sample size: Tables VI–IX give the results for the subgroups of small size and the corresponding asymptotic standard error is already large (10–30 per cent of the estimated HIV incidence). In these situations the introduction of the extra uncertainty into the model does only slightly increase the error in the incidence estimate.

On the other hand, Tables III and IV are related to subgroups of larger size, whose HIV incidence is estimated with a small asymptotic standard error (2–10 per cent of the estimated HIV incidence). In these cases the bootstrap approach introduces a substantial increase of the standard error, which is located mainly in the very early periods of the epidemic, as we can see from the $\%B$ column.

An intermediate behaviour is shown in Table V (male IDUs), which describes the results related to a subgroup of fairly large size, whose asymptotic standard error is already quite large; in this case the bootstrap procedure yields an increase of the standard error between 2 and 4 times the asymptotic one, and this increase is situated in the early periods of the epidemic.

Table III. HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of all males

Year	$\hat{\mu}$	$\widetilde{\text{SE}}$	SE^*_{TOT}	$\%B$	$\text{SE}^*_{\text{BASE}}$	$\%B$	SE^*_{AGE}	$\%B$
1983	1617	83	331	72	326	71	182	52
1984	4950	105	720	83	676	81	465	76
1985	8829	176	1050	82	905	79	704	74
1986	10418	181	1081	82	911	78	641	71
1987	9131	145	868	81	769	78	396	61
1988	6767	182	627	69	591	66	258	27
1989	4840	196	449	54	439	51	235	15
1990	3772	184	334	42	329	39	218	14
1991	3657	190	289	30	280	27	225	13
1992	5012	342	465	20	433	15	420	15

Table IV. HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of all females

Year	$\hat{\mu}$	$\widetilde{\text{SE}}$	SE^*_{TOT}	$\%B$	$\text{SE}^*_{\text{BASE}}$	$\%B$	SE^*_{AGE}	$\%B$
1983	329	37	101	57	85	52	47	17
1984	1055	51	213	71	173	67	91	43
1985	2048	83	309	70	246	65	152	45
1986	2686	95	339	69	274	63	170	43
1987	2674	78	327	72	273	68	139	41
1988	2241	99	300	63	262	61	127	20
1989	1752	118	261	51	237	48	131	9
1990	1437	117	220	41	205	39	127	6
1991	1384	123	202	32	191	30	134	5
1992	1750	225	306	17	292	16	238	3

Figure 1 suggests a possible explanation; the estimated epidemic curve shows a clear peak followed by a rather flat trend, as if we were in an endemic situation. In such cases it is difficult to model the overall trend through a polynomial. As a result the estimated curve may not fit the data too well and therefore the standard errors tend to be larger.

The $\%B$ column suggests that the bootstrap introduces a correction in the scarce uncertainty of the more remote estimates, while in the more recent estimates it does not add much uncertainty.

Regarding the partial bootstrap mentioned in Section 3.4, the Tables III–IX show a common feature: the uncertainty measured by the bootstrap is mainly concentrated in the baseline bootstrap ($\text{SE}^*_{\text{BASE}}$), whilst the age bootstrap (SE^*_{AGE}) is only slightly larger than its asymptotic counterpart ($\widetilde{\text{SE}}$). This means that the role played by the covariate AGE, besides being strongly significant, does not add extra cost in the model in terms of an increment of the error of the estimates.

Table V. Bootstrap HIV incidence standard errors for HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of male IDU

Year	$\hat{\mu}$	$\widetilde{\mathrm{SE}}$	$\mathrm{SE}^*_{\mathrm{TOT}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{BASE}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{AGE}}$	$\%B$
1983	742	199	551	50	491	52	230	15
1984	5028	315	1302	64	978	59	463	33
1985	8181	459	1649	68	1119	61	693	37
1986	6497	387	1386	67	1057	62	640	42
1987	4183	424	1003	54	879	53	547	25
1988	3051	271	641	50	603	50	328	18
1989	2926	346	578	22	549	23	353	2
1990	3505	667	791	10	727	10	633	1
1991	4171	615	918	8	828	6	601	2
1992	3233	1520	1324	6	1280	5	1406	1

Table VI. Bootstrap HIV incidence standard errors for HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of female IDU

Year	$\hat{\mu}$	$\widetilde{\mathrm{SE}}$	$\mathrm{SE}^*_{\mathrm{TOT}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{BASE}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{AGE}}$	$\%B$
1983	263	104	151	21	140	23	103	4
1984	1338	160	283	29	258	28	177	4
1985	1937	194	308	30	272	27	196	6
1986	1493	167	262	28	245	25	193	7
1987	978	177	229	16	223	17	188	5
1988	721	117	157	14	155	17	126	5
1989	670	143	173	6	178	6	165	1
1990	751	252	273	2	265	2	255	0
1991	817	235	274	2	265	3	235	1
1992	585	486	501	1	492	1	474	0

Table VII. Bootstrap HIV incidence standard errors for HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of men who have sex with men

Year	$\hat{\mu}$	$\widetilde{\mathrm{SE}}$	$\mathrm{SE}^*_{\mathrm{TOT}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{BASE}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{AGE}}$	$\%B$
1983	486	34	60	33	58	32	45	29
1984	935	49	84	30	83	29	70	34
1985	1204	71	102	25	101	24	92	26
1986	1169	56	95	32	92	31	84	24
1987	954	48	87	35	84	33	65	18
1988	738	63	90	22	88	20	69	6
1989	609	77	96	13	94	11	78	2
1990	597	91	106	8	104	7	87	1
1991	787	116	132	5	131	4	111	1
1992	1561	174	206	4	206	4	185	1

Table VIII. Bootstrap HIV incidence standard errors for HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of males by heterosexual transmission

Year	$\hat{\mu}$	$\widetilde{\mathrm{SE}}$	$\mathrm{SE}^*_{\mathrm{TOT}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{BASE}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{AGE}}$	$\%B$
1983	33	13	17	4	15	3	15	1
1984	135	21	29	15	26	9	26	9
1985	333	28	47	24	40	17	40	23
1986	549	51	74	18	66	14	68	16
1987	671	50	75	22	70	18	66	18
1988	683	44	69	22	66	21	57	16
1989	650	69	89	9	86	9	78	5
1990	642	94	113	4	109	4	102	2
1991	739	111	134	3	128	3	123	2
1992	1112	160	200	4	198	3	183	4

Table IX. Bootstrap HIV incidence standard errors, HIV incidence estimate with asymptotic and bootstrap standard errors and percentage of the variance explained by the bootstrap for the subgroup of females by heterosexual transmission

Year	$\hat{\mu}$	$\widetilde{\mathrm{SE}}$	$\mathrm{SE}^*_{\mathrm{TOT}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{BASE}}$	$\%B$	$\mathrm{SE}^*_{\mathrm{AGE}}$	$\%B$
1983	34	13	18	12	16	9	14	2
1984	128	20	34	32	28	23	23	8
1985	315	32	62	38	50	28	39	15
1986	547	54	92	34	76	26	64	13
1987	733	51	99	41	83	33	65	17
1988	817	56	105	37	92	32	68	14
1989	818	92	129	20	119	18	100	5
1990	803	111	145	14	136	13	119	3
1991	841	112	151	13	141	12	124	4
1992	1024	218	274	7	260	5	232	3

5. DISCUSSION

The uncertainty reflected on the back-calculation estimates by the limited epidemiological knowledge of the incubation time distribution has been evaluated in various settings.

We have used a non-standard bootstrap procedure. Application of the bootstrap in connection with HIV estimation is not new (Rosenberg and Gail[17] and Mariotti and Cascioli[18]), but to our knowledge this is the first time that the bootstrap method is used to study how the variability of the parameters of the incubation time distribution affects the precision of the HIV incidence estimates.

The peculiarity of the proposed method lies in the resampling of a characteristic that is not explicitly linked to the feature of interest, as explained in Section 3.4. In dealing with this situation we have made a number of simplifying assumptions:

(i) the use of the normal distribution to generate the bootstrap sample, which is justified by asymptotic reasoning, since the maximum likelihood estimates are asymptotically normal;

(ii) the choice of the number of bootstrap replications, which is a compromise between the accuracy of the bootstrap estimates and the cost of a single simulation (about half minute) on a Digital Alpha 3000/300 LX workstation;
(iii) in the choice of the weight function (10) we have assumed that the elements of ψ are independent and hence the weight function is a product of five uninormal distributions, without taking into consideration the correlation between the elements of ψ.

The amplification of the uncertainty of the HIV incidence estimates resulting from the implementation of the parametric bootstrap is not so dramatic as expected. We found the major increase concentrated in the subgroups of larger size, where the asymptotic standard errors were too small. In particular, for the overall males and females subgroups, the epidemic curve has a clear peak at around the years 1985–1986, and the corresponding asymptotic standard errors of the HIV incidence are very small (probably unrealistically so).

By introducing the extra uncertainty in the model via the bootstrap, we obtain an increase in the standard errors, especially in the early years of the epidemic and around the peak. This result is not surprising and it can be explained by the structure of the back-calculation equations: estimates for the early periods of the HIV epidemic correspond to the right tail of the incubation time distribution, which is very sensitive to small perturbations. Hence the bootstrap tends to concentrate its contribution towards the beginning of the epidemic, as shown by the $\%B$ values. In the late periods of the epidemic the asymptotic estimates obtained from the back-calculation are sufficiently large to describe the errors.

Finally, the effect of age at infection time in influencing the progression towards AIDS is expected to enhance the uncertainty of estimates of the incubation time distribution, as it requires further data stratification and additional parameters to be estimated on the same data. Moreover, for particular age strata, that is, very young people under 25 with 12·5 year median time, the estimates might be affected by rather large variability.

Notwithstanding, using age-dependent incubation time distribution in the back-calculation of HIV infection incidence does not add much uncertainty to the estimates, being mostly concentrated on the baseline distribution. Considering age in back-calculation applications is supported by our findings, also with reference to the great public health importance of this characterization of HIV infection epidemics.

With the proposed methodology we have used a specific back-calculation method and a specific estimate of the incubation time distribution applied to the Italian epidemic. However, the incubation estimate is obtained from data which are comparable to other international studies and widely used in the Italian estimates, the back-calculation equations are commonly used and the proposed bootstrap method is rather general and therefore applicable in different contexts.

ACKNOWLEDGEMENTS

We are grateful to Professor Anthony Davison for helpful methodological advice and to Dr. Angela Mariotto for providing updates of the estimates of the incubation time distribution parameters. We also acknowledge useful comments following the presentation of an early stage of this work at the 'Workshop on statistical models and inference for the prediction of the AIDS epidemic', IAC, Rome, June 1996. The work was partially supported by a grant from the Ministero della Sanita, Istituto Superiore di Sanita, IX Progetto AIDS, 1996.

REFERENCES

1. Brookmeyer, R. and Gail, M. H. 'The minimum size of the AIDS epidemic in the United States', *Lancet*, **2**, 1320–1322 (1986).
2. 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*, **83**, 301–308 (1988).
3. Brookmeyer, R. and Gail, M. H. *AIDS Epidemiology: a Quantitative Approach*, Oxford University Press, Oxford, 1994.
4. Mariotto, A. B., Mariotti, S., Pezzotti, P., Rezza, G. and Verdecchia, A. 'Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: a comparison with male homosexuals', *American Journal of Epidemiology*, **135**, 428–437 (1992).
5. Bacchetti, P. and Moss, A. R. 'Incubation period of AIDS in San Francisco', *Nature*, **338**, 251–253 (1989).
6. Giesecke, J., Scalia Tomba, G., Berglund, O., Berntorp, E., Schulman, S. and Stigendal, L. 'Incidence of symptoms and AIDS in 146 Swedish haemophiliacs and blood transfusion recipients infected with human immunodeficiency virus', *British Medical Journal*, **297**, 99–102 (1988).
7. Blaxhult, A., Granath, F., Lidman, K. and Giesecke, J. 'The influence of age on the latency period of AIDS in people infected by HIV through blood transfusion', *AIDS*, **4**, 125–129 (1990).
8. Darby, S. C., Doll, R., Thakrar, B., Rizza, C. R. and Cox, D. R. 'Time from infection with HIV onset of AIDS in patients with hemophilus in the U.K.', *Statistics in Medicine*, **9**, 681–689 (1990).
9. Rosenberg, P. S. 'Backcalculation models of age-specific HIV incidence rates', *Statistics in Medicine*, **13**, 1975–1990 (1994).
10. Becker, N. and Marschner, I. C. 'A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data', *Biometrika*, **80**, 165–178 (1993).
11. Verdecchia, A. and Mariotto, A. B. 'A back-calculation method to estimate the age and period HIV infection intensity, considering the susceptible population', *Statistics in Medicine*, **14**, 1513–1530 (1995).
12. Verdecchia, A., Mariotto, A. B., Capocaccia, R. and Mariotti, S. 'An age and period reconstruction of the HIV epidemic in Italy', *International Journal of Epidemiology*, **23**, 1027–1039 (1994).
13. Efron, B. 'Bootstrap methods: another look at the jackknife', *Annals of Statistics*, **7**, 1–26 (1979).
14. Davison, A. C. and Hinkley, D. *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997.
15. Bickel, P. J. and Doksum, K. A. *Mathematical Statistics*, Holden-Day, Oakland, 1977.
16. Gigli, A. and Verdechia, A. 'Influence of incubation time on the uncertainty of the back-calculation estimates: an application to the HIV epidemic in Italy', *Quaderni IAC*, **12**, 1–28 (1997).
17. Rosenberg, P. S. and Gail, M. H. 'Uncertainty in estimates of HIV prevalence derived by back-calculation', *Annals of Epidemiology*, **1**, 105–115 (1991).
18. Mariotti, S. and Cascioli, R. 'Sources of uncertainty in estimating HIV infection rates by back-calculation: an application to Italian data', *Statistics in Medicine*, **15**, 2669–2687 (1996).

