Proceedings of the

1999 Health Sciences Simulation Conference

Edited by

James G. Anderson

Meyer Katzper

1999 Western MultiConference

San Francisco, California

January 17–20, 1999

Cathedral Hill Hotel

ISBN: 1-56555-160-5

SPONSORED BY THE SOCIETY FOR COMPUTER SIMULATION INTERNATIONAL

HEALTH SCIENCES SIMULATION

EDITED BY
JAMES G. ANDERSON
MEYER KATZPER

THE SOCIETY FOR COMPUTER SIMULATION INTERNATIONAL

ISBN: 1-56555-160-5

A MONTE CARLO METHOD TO ACCOUNT FOR THE UNCERTAINTY OF THE INCUBATION TIME DISTRIBUTION IN THE ESTIMATION OF HIV INCIDENCE

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

Keywords: HIV/AIDS Epidemic, Backcalculation method, Incubation time distribution, Bootstrap, Variance estimation

ABSTRACT

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 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 affects the precision of the HIV incidence estimates.

Given the AIDS data and the incubation time distribution, we can estimate the HIV incidence via a deconvolution method, called backcalculation. In the backcalculation equations the incubation time distribution is supposed fixed, so that its variability is not considered when assessing the precision of the estimates.

The scope of this work is to investigate the sensitivity of the estimates to variations of the incubation times making use of a Monte Carlo method called the bootstrap.

We compare, through an application to the HIV epidemic in Italy, the precision of the incidence estimates obtained via the standard backcalculation method and via the parametric bootstrap.

The results show that 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 gave great improvement to potential surveillance system in most countries.

Backcalculation methods, originally proposed in HIV estimation by Brookmeyer and Gail (1986, 1988), were among those most appreciated for their simplicity and flexibility. Reliability of estimates obtained by backcalculation 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 backcalculation procedures can be found in Brookmeyer and Gail (1994).

Knowledge of incubation time was indicated as a major problem, being an important source of uncertainty in backcalculation estimates. Most of the uncertainty involved in the estimation of the incubation time distribution 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 backcalculation method and affects the estimation of the precision of HIV incidence.

The bootstrap, introduced by Efron 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 (1997).

In this work we are interested in studying the effect of the epidemiological uncertainty of incubation time distribution to backcalculation estimates. Starting from the parametric model illustrated by Verdecchia and Mariotto (1995), we make use of the parametric bootstrap to resample parameter values from the incubation time distribution and use them in the backcalculation equations. The resulting estimated standard errors of the incidence estimates should incorporate the uncertainty due to the estimation of the incubation distribution.

2. BACKCALCULATION

Backcalculation is a method for estimating past HIV infection rates from AIDS incidence data. The basic idea is to use AIDS incidence counts and an estimate of the incubation time distribution to reconstruct the numbers of individuals who have been previously infected in order to give rise to the observed pattern of AIDS incidence.

The fundamental relation between the expected cumulative number of AIDS cases diagnosed by calendar time t, $\gamma(t)$, the infection rate $\mu(s)$ at calendar time s, and the incubation time distribution $\tau(t)$, is given by the convolution equation

$$\gamma(t) = \int_0^t \mu(s)\tau(t-s)ds. \tag{1}$$

Once we assume τ as fixed, and we know γ from the data, we are able to estimate μ.

Let $Y = (Y_1, \ldots, Y_n)$, where Y_i is the number of AIDS cases diagnosed in the calendar time $[T_{i-1}, T_i)$. It is assumed that individuals become infected according to a Poisson process, whose intensity function $\mu(s;\theta)$ is assumed to come from a parametric family with p unknown parameters θ.

Hence (1) becomes

$$I\!E(Y_i) = \int_{T_0}^{T_i} \mu(s;\theta)[\tau(T_i - s) - \tau(T_{i-1} - s)]ds \tag{2}$$

and we can write down the loglikelihood function $L(\theta, Y)$ for θ, obtain the maximum likelihood estimate $\hat{\theta}$ and an estimate of its covariance matrix

$$\mathrm{cov}(\hat{\theta}) = \left[-\frac{\partial^2 L(\theta;Y)}{\partial\theta_i \partial\theta_j}\right]^{-1}_{\theta=\hat{\theta}}. \tag{3}$$

Since $\mathrm{var}(f(x)) = [f'(x)]^2\mathrm{var}(x)$, from (3) we obtain the asymptotic variance of the HIV incidence estimate.

3. INCUBATION TIME DISTRIBUTION AND THE BOOTSTRAP

The incubation time is the period between the onset of HIV infection and the diagnosis of AIDS. Incubation periods are extremely variable and some are very long. The incubation time distribution $\tau(t)$ is the probability that an infected individual progresses to AIDS within t years of the time of seroconversion.

In the recent past several studies have been set up for estimating the incubation time distribution via parametric or nonparametric models. The estimated distribution is then included in the backcalculation *as if it were a known variate.*

Let $\tau(t;\psi)$ be the parametric model for the incubation time distribution, which depends upon q parameters ψ, and let $\hat{\psi}$ be the maximum likelihood estimated vector. Equation (2) becomes

$$I\!E(Y_i) = \int_{T_0}^{T_i} \mu(s;\theta)[\tau(T_i - s;\hat{\psi}) - [\tau(T_{i-1} - s;\hat{\psi})]ds, \tag{4}$$

thus representing a completely parametrised model for the expected cumulative number of AIDS cases.

In this formulation it is evident that the HIV incidence μ not only depends on the parameter vector θ of the intensity model, it also depends on the parameter vector ψ of the incubation time model.

Let $\hat{\mu}(t;\theta)$ be the estimated number of HIV cases diagnosed in year t and assume that

$$I\!E(\hat{\mu}(t;\theta)|\psi) = \mu(t;\theta) \tag{5}$$

$$\mathrm{var}(\hat{\mu}(t;\theta)|\psi) = v^2(t;\theta). \tag{6}$$

Having assumed known (and therefore fixed) the contribution of the incubation time in the backcalculation equations, the standard error of the HIV incidence estimates, $v(t;\theta)$, can be interpreted as the statistical error due to the fitting of the model. 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, 1977, p. 36) applies:

$$\begin{aligned}\mathrm{var}(\hat{\mu}) &= \mathrm{var}_\psi[I\!E(\hat{\mu}|\psi)] + I\!E_\psi[\mathrm{var}(\hat{\mu}|\psi)] \\ &= \mathrm{var}_\psi(\mu) + I\!E_\psi(v^2),\end{aligned} \tag{7}$$

following notation of (5) and (6).

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

Let $\hat{\psi}$ be the vector of the maximum likelihood estimates of the parameters of the incubation time distribution and $\hat{C}$ its estimated covariance matrix. The full parametric bootstrap paradigm consists of

(a) resampling a set $T^* = \{t_1^*, \ldots, t_m^*\}$ of incubation times from the distribution $\tau(t;\hat{\psi})$, whose parameter vector ψ is replaced by its maximum likelihood estimate;

(b) estimating the parameter vector $\psi_f^* = \hat{\psi}(t_1^*, \ldots, t_m^*)$ 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 backcalculation equations (4), obtaining the new HIV incidence estimate μ_f^*.

By repeating the resampling of T^* and the computation of μ_f^* B independent times we obtain $\mu_f^{*(1)}, \ldots, \mu_f^{*(B)}$, and the bootstrap approximation of $\mathrm{var}_\psi(\mu)$ is

$$\mathrm{var}^*(\mu_f^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\mu_f^{*(b)} - \bar{\mu_f^*})^2, \qquad (8)$$

where $\bar{\mu_f^*} = (1/B)\sum_{b=1}^{B}\mu_f^{*(b)}$ 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

$$\mathbb{E}^*(v_f^{*^2}) = \sum_{b=1}^{B} v^2(\psi_f^{*(b)})\omega(\psi_f^{*(b)}), \qquad (9)$$

where $v^2(\psi^*) = \mathrm{var}(\mu_f^*|\psi_f^*)$ is the asymptotic variance (6) conditioned on ψ_f^* instead of $\hat{\psi}$ and $\omega(\psi_f^*)$ are normalised weights suitably chosen in order to give more importance to the ψ_f^*'s which are closer to $\hat{\psi}$:

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

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

$$\begin{aligned}\mathrm{var}(\hat{\mu}) \;&\approx\; \frac{1}{B-1}\sum_{b=1}^{B}(\mu_f^{*(b)} - \bar{\mu_f^*})^2 \\ &+\; \sum_{b=1}^{B} v^2(\psi_f^{*(b)})\omega(\psi_f^{*(b)}). \qquad (11)\end{aligned}$$

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 $\mathrm{var}^*(\mu_f^*) = 0$, 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 can be rather complicated, because the incubation time is usually modelled as a multi-state Markov process and therefore the resampling should be implemented from a Markov process and should take into account censored data. Moreover the maximum likelihood estimate $\hat{\psi}$ is usually computed via a nonlinear algorithm and it is quite time-consuming.

We propose 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^*.

4. PRACTICAL 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 throughout the convolution equation (4). Because of this nonstandard application of the bootstrap method greater care must be paid to the implementation aspects.

As an application of our proposed method we have used the parametric model suggested by Verdecchia and Mariotto (1995). It consists of a logistic model for the incidence function $\mu(t;\theta)$, with covariates *age at AIDS diagnosis*, *year of AIDS diagnosis* and *birth cohort*, and a set of two independent Weibull distributions for the transition rates of the incubation time distribution, with covariate *age at seroconversion*.

The first issue is the choice of the number of bootstrap simulations to perform. In the literature it is suggested that for moment estimation 50–200 simula-

tions should be enough (Efron and Tibshirani, 1993). We performed 20 independent bootstrap experiments, with the number of simulations varying from 50 to 1,000, and measured the distance between the original and the bootstrap parameters. The results are very unstable at first (between 50 to 150 simulations) and tend to stabilise at around 1,000 simulations, but at 200 simulations the obscillations are much reduced. This seems a suitable compromise between the accuracy of the estimates and the computational costs involved (for each simulation we have to resample the incubation time distribution, find the maximum likelihood estimates and solve a nonlinear equation). More details can be found in Gigli and Verdecchia (1997).

Another issue to be tackled involves the link between the simulated parameters of the incubation time distribution τ and the HIV incidence estimates μ, which cannot be explicitely expressed. Formally it is given by the backcalculation equations, which can be considered a kind of "black box", and all we know is that the relationship between $\tau(t;\psi)$ and $\mu(t;\theta)$ is continuous. However we need to check whether the relationship is also monotonic, that is whether to a large value of the input vector ψ^* corresponds a large value of the output variable μ^*.

Figure 1 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.

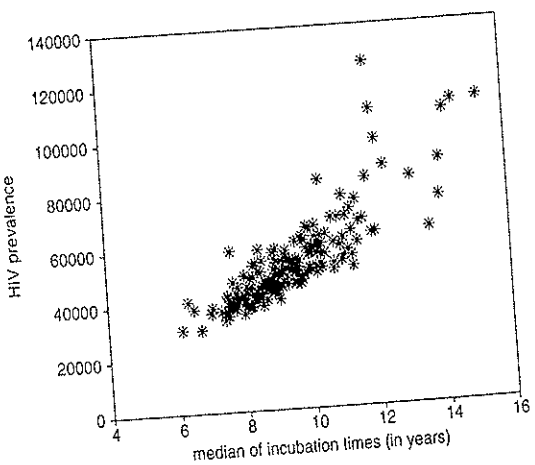

Figure 1: *Plot of HIV prevalence for the age group [25,34) and the year 1992 versus median incubation times*

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 scatterplot 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.

5. RESULTS AND DISCUSSION

We have implemented the method illustrated in the previous sections to data related to the Italian epidemic. The data consist of the AIDS reported cases from 1983 to 1994 and are grouped in 7 categories: intravenous drug users (IVDU) males and females, males who have sex with men (MWSM), eterosexual contact males and females, and finally the overall male and female AIDS cases. The data are described in more details by Verdecchia and Mariotto (1995).

Here we will concentrate on discussing the behaviour of the HIV incidence standard errors, which we expect to increase because of the introduction of the extra uncertainty in the backcalculation equations, caused by the variation in the incubation time distribution.

The results are illustrated in more details in Gigli and Verdecchia (1997).

In tables 1 and 2 we describe the results corresponding to two subgroups: the overall male AIDS cases (excluding the blood recipients and the vertically infected cases), and the female IVDU's. The tables report the HIV incidence estimates $\hat{\mu}$, their asymptotic standard errors s.e. obtained via the classical backcalculation method (Verdecchia and Mariotto, 1995), the bootstrap standard errors s.e.* obtained from our proposed method as square root of (11), and the percentage of the variance explained by the bootstrap, $\%B$, which is computed as the ratio between the bootstrap variance given by (8) and the overall variance given by (11). The latter is a useful indicator of how the extra uncertainty added into the model through varying the incubation times affects the overall variance estimate.