Comment

Ron Brookmeyer

I would like to congratulate the authors on an excellent and comprehensive discussion of backcalculation of HIV-infection rates. Bacchetti, Segal and Jewell (BSJ) develop a framework for backcalculation that incorporates several new attractive features, and they discuss important sources of uncertainty.

The earliest data on the incubation period of HIV infection came from a 1986 study of tranfusion-associated AIDS cases. These data indicated that the incubation period was long and variable. This was alarming because it suggested that the number of diagnosed AIDS cases must be only a fraction of the numbers of individuals who were HIV infected. This was the kernel of the idea behind the backcalculation methodology. Brookmeyer and Gail (1986) introduced and applied the methodology to the U.S. AIDS epidemic based on cases diagnosed through 1985 and concluded that even without accounting "for new infections after 1985 nor very long incubation periods," the cumulative number of AIDS cases would grow by a factor of more than 6 by the end of 1991.

Backcalculation is essentially a deconvolution problem. Using data on the cumulative numbers of AIDS cases $a(t)$, and the incubation distribution F, one tries to glean information about past infection rates $I(s)$ through the convolution equation

$$a(t) = \int_{-\infty}^{t} I(s)\, F(t - s \mid s) ds. \tag{1}$$

The solution of (1) has a long and rich statistical history (O'Sullivan, 1986; Wahba, 1990). Equation (1) arises in a wide range of applications, including geology, meterology, engineering and biomedical applications. An important issue that arises in solving (1) concerns how much structure to impose on $I(s)$. Additional problems associated with applying Equation (1) to the AIDS epidemic are imprecise knowledge of the incubation distribution F and systematic errors in AIDS-incidence data.

1. NONSTATIONARY INCUBATION DISTRIBUTION

Backcalculation analyses must account for changes over calendar time in the incubation distribution (Gail, Rosenberg and Goedert, 1990). For example, failure to account for lengthening incubation periods, perhaps because of new treatments, can lead to underestimation of the numbers of infected individuals. One approach to account for nonstationarity effects is to use a completely external estimate of the nonstationary incubation distribution $F(t|s)$ [the treatment model, Brookmeyer and Liao (1990b); Brookmeyer (1991)].

Alternatively, BSJ propose a methodology for estimating nonstationarity effects in the incubation distribution using AIDS-incidence data and backcalculation methods (Section 3.3). However, it seems to be asking a lot of AIDS-incidence data to provide information about both infection rates and changes in the incubation distribution. A falloff in the growth of AIDS cases could either be explained by the scenario of declining infection rates or a scenario of lengthening incubation periods. Intuitively, I would not expect that AIDS-incidence data alone could distinguish between these two scenarios.

The model proposed by BSJ appears to be "nearly nonidentifiable" in the following sense. Given any $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$, there exists another $\underset{\sim}{\theta}^*$ and $\underset{\sim}{\beta}^*$ that produce the same likelihood, that is, $L_m(\underset{\sim}{\theta}, \underset{\sim}{\beta}) = L_m(\underset{\sim}{\theta}^*, \underset{\sim}{\beta}^*)$. To see this, assume $\beta_j = 0$ for $j < T$ (BSJ used T = January 1986). Now let $\theta_j^* = \theta_j$ for $j < T$, and choose θ_j^* for $j \geq T$ to be any arbitrary values you like. If we set $\beta_j^* = 0$ for $j < T$, and set

$$\beta_j^* = \beta_j + \log\left\{\left(\sum_{i=0}^{j}\theta_i D_{ij}\right)\Big/\left(\sum_{i=0}^{j}\theta_i^* D_{ij}\right)\right\} \tag{2}$$

for $j \geq T$, then $L_m(\underset{\sim}{\theta}, \underset{\sim}{\beta}) = L_m(\underset{\sim}{\theta}^*, \underset{\sim}{\beta}^*)$. Equation (2) was derived by setting $E(Y_j \mid \underset{\sim}{\theta}, \underset{\sim}{\beta}) = E(Y_j \mid \underset{\sim}{\theta}^*, \underset{\sim}{\beta}^*)$. Equation (2) shows that if one wanted to keep $E(Y_j)$ fixed, lower infection rates ($\theta_j^* < \theta_j$) can be compensated by additional shortening of the incubation period ($\beta_j^* > \beta_j$). On the other hand, higher infection rates ($\theta_j^* > \theta_j$) can be compensated by additional lengthening of the incubation periods ($\beta_j^* < \beta_j$) in order to keep $E(Y_j)$ fixed. This result on nonidentifiability [Equation (2)] holds for the general model proposed by BSJ that incorporates seasonal effects (S_j) and reporting delays (R_j) (setting $S_j^* = S_j$ and $R_j^* = R_j$), as well as for a simple model that assumes no seasonal effects ($S_j = 0$) and with the data completely reported ($R_j = 1$).

BSJ jointly estimate infection rates and nonstationarity effects for the U.S. AIDS epidemic. In this case, simultaneous estimation of $\underset{\sim}{\beta}$ and $\underset{\sim}{\theta}$ appears to be possible solely because of the additional smoothness

Ron Brookmeyer is a Professor in the Department of Biostatistics, School of Hygiene and Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, Maryland 21205.

requirements on $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$ induced by the roughness penalties.

BSJ estimated the cumulative numbers of HIV infected individuals by December 1990 to be 897,000 and 1,031,000 for the random sample and treatment incubation models, respectively (Table 1). However, the nonstationary incubation distribution for the random sample model was generally shorter than for the treatment model [see Figure 1 of Bacchetti et al. (1993)]. It is conceivable that infection rates that are higher than those shown in Table 1 for the random sample model together with correspondingly longer incubation periods for the random sample model, could fit the AIDS-incidence data nearly as well, in which case the random sample and treatment incubation models might yield more similar estimates of the cumulative numbers infected. Table 1 becomes difficult to interpret because it is not clear if the varying estimates of the cumulative infections result from the four different external incubation distributions or from numerical instabilities because of near nonidentifiability that arises if nonstationarity effects and infection rates are estimated simultaneously.

Because of these considerations I do not believe that joint estimation of infection rates and nonstationarity effects using only AIDS-incidence data will yield useful estimates. It is simply asking too much of AIDS-incidence data. External estimates of $F(t|s)$ are required.

2. EXTERNAL ESTIMATES OF THE INCUBATION DISTRIBUTION

Backcalculation is sensitive to the assumed incubation distribution. BSJ estimate the incubation distribution from three data sets. However, each of these data sets is relatively small (the numbers of AIDS cases in each data set ranged from 33 to 56 cases). The main differences among the three incubation distributions appear to occur after 9 years following HIV seroconversion [Figure 2 in Bacchetti, Segal and Jewell (1992a)]. However, BSJ censored followup at about 1987 in the analyses of these three cohorts. Thus, because nearly no one was infected prior to 1978, estimates of $F(t)$ beyond 9 years depend on parametric model extrapolation.

Not all studies have found systematic differences in incubation distributions. Mariotto et al. (1992) compared the incubation distribution among male homosexuals and intravenous drug users in Italian cities and did not find a significant difference. Furthermore, the only cofactor (i.e., covariate that modifies the incubation-period distribution) that has been identified is age at infection. The fact that the hepatitis B vaccine trial and random sample data sets yielded different incubation distributions is somewhat surprising since both data sets were composed of homosexual men and were initially part of a larger San Francisco cohort. This statistical finding of different incubation distributions for these two cohorts raises the question of whether there is a plausible biological or clinical explanation. Different incubation distributions could either be explained by a cofactor (e.g., other sexually transmitted diseases) or, alternatively, other study differences such as different intensities of followup of the cohorts for AIDS incidence. If a cofactor is identified, it would be important to replicate the finding in other cohort studies of HIV infection.

Backcalculation should be based on the best available incubation distribution. Analyses that combine information from a number of epidemiological studies should be performed. Meta-analyses along the lines suggested by Harris (1988) are important to perform and could provide invaluable input to backcalculation.

It is important to emphasize that even if the incubation distribution were known exactly, estimates of recent infection rates are very imprecise. Accordingly, it is important to present confidence intervals on the cumulative numbers of HIV-infected individuals.

3. SMOOTHING THE INFECTION CURVE

Some structure has to be imposed on the infection curve $I(s)$. Strongly parametric models can yield very biased estimates of recent infection rates if the parametric assumptions are incorrect. Weakly parametric models such as piecewise constant step functions are flexible yet unsatisfying because they yield unsmoothed and implausible reconstructions of HIV-infection rates. Smoothing through the penalized likelihood is an attractive alternative. The penalized log likelihood is

$$\log L - \frac{\lambda}{2} J,$$

where J measures the roughness of the infection curve and λ is the smoothing parameter.

It must be emphasized that estimates of the recent infection rates depend strongly on both the amount of smoothing (λ) and the form of the roughness-penalty function J. Because of the long incubation period, there is essentially no information in the AIDS-incidence data about infection rates in the most recent year or two. Thus, the estimates of θ in the most recent past would approximately minimize J because $\log L$ is nearly unaffected by changes in recent infection rates. BSJ use the penalty

$$J = \sum [\log\theta_i - 2\log(\theta_{i+1}) + \log(\theta_{i+2})]^2.$$

It follows that estimates of recent infection rates are approximately given by

$$\hat{\theta}_{K+j} \approx \hat{\theta}_K e^{\alpha j}, \tag{3}$$

where $\alpha = \log[\hat{\theta}_K/\hat{\theta}_{K-1}]$ and K is sufficiently large. Thus, the most recent infection rates are estimated by extrapolating an exponential function.

Another choice of the penalty function is

$$J = \sum[\theta_i - 2\theta_{i+1} + \theta_{i+2}]^2.$$

Then, estimates of recent infection rates are approximately

$$\hat{\theta}_{K+j} \approx \hat{\theta}_K + j\delta, \tag{4}$$

where $\delta = [\hat{\theta}_K - \hat{\theta}_{K-1}]$. Thus the most recent infection rates are estimated by extrapolating a linear function.

The piecewise constant step function model for $I(s)$ that was used in the early work on backcalculation assumes that infection rates are constant over intervals. Simulation studies of Rosenberg, Gail and Pee (1991) suggest choosing a last step of 4 to 4.5 years in length. Recent infection rates under this model are estimated by

$$\hat{\theta}_{K+j} = \hat{\theta}_K. \tag{5}$$

Estimates of recent infection rates obtained by backcalculation are essentially extrapolations of trends in $I(s)$. Equations (3) through (5) are different examples of mathematical functions that have been used for such extrapolations and result from different choices of the roughness penalties or parametric assumptions on $I(s)$. Estimates of recent infection rates based on backcalculation are highly dependent on the degree of smoothing λ, the penalty J and the parametric model for $I(s)$.

Appreciable improvements in our ability to reconstruct infection rates may come, not from alternative smoothing procedures or parametric models but rather from obtaining empirical data on recent infection rates.

4. FUTURE PROSPECTS FOR FORECASTING AND RECONSTRUCTING THE AIDS EPIDEMIC

Early in the AIDS epidemic, the only reliable data for monitoring the epidemic was AIDS-incidence data. Since the development of the HIV antibody test in the mid-1980s, numerous surveys of HIV seroprevalence have been conducted. Infection rates have also been directly estimated in several cohorts. Our ability to reconstruct infection rates may drastically improve by incorporating external information about recent infection rates and HIV seroprevalence derived from cohort studies and cross-sectional surveys.

There is considerable underreporting of AIDS cases to national and regional AIDS surveillance registries in developing countries, especially in Africa. Projections of the course of the epidemic in developing countries must rely more on HIV seroprevalence and seroincidence surveys than on AIDS-incidence data. While U.S. AIDS-incidence data are relatively complete, more reliable assessments of the scope of the epidemic may be obtained by considering HIV-seroprevalence and HIV-seroincidence data as well. For example, extensive HIV-seroprevalence surveys among childbearing women are extraordinarily useful for forecasting the future numbers of pediatric AIDS cases. Statistical approaches that combine data from multiple sources (e.g., AIDS-incidence data, HIV-seroprevalence and -seroincidence surveys, incubation distributions) are promising and may considerably improve the accuracy of assessments of the scope of the epidemic.

Comment: Assessing Uncertainty in Backprojection

John B. Carlin and Andrew Gelman

Bacchetti, Segal and Jewell are to be congratulated for providing not only a comprehensive review of an important problem in applied statistics but also for introducing a number of new ideas that should have a practical impact on understanding the course of the HIV epidemic. On a semantic detail, we wonder why the authors (and others) have adopted the term "backcalculation," rather than "backprojection," which seems to carry a more appropriate connotation of uncertain inference (as well as being shorter!).

The authors rightly emphasize the sensitivity of backprojection estimates to assumptions about the incubation distribution, but they seem strangely reluc-

John B. Carlin is Deputy Head, Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Melbourne, Victoria 3052, Australia. Andrew Gelman is Assistant Professor, Department of Statistics, University of California, Berkeley, California 94720.