



Reporting Delays and the Incidence of AIDS

Jeffrey E. Harris

Journal of the American Statistical Association, Volume 85, Issue 412 (Dec., 1990),
915-924.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28199012%2985%3A412%3C915%3ARDATIO%3E2.0.CO%3B2-1>

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the American Statistical Association is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Journal of the American Statistical Association
©1990 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2002 JSTOR

Reporting Delays and the Incidence of AIDS

JEFFREY E. HARRIS*

It can take several months, and often years, for case reports of acquired immunodeficiency syndrome (AIDS) to be received by the Centers for Disease Control (CDC). As a result, the cumulative number of AIDS cases reported by the CDC at a given date may fall considerably below the actual number thus far diagnosed. Methods are described for estimating both the probability distribution of reporting delays and the actual incidence of AIDS. An estimated 62% of AIDS cases are reported more than 2 months after diagnosis, and 17% are reported with a delay of 3 years or more. An estimated 130,000 AIDS cases were actually diagnosed through March 1989, compared to about 91,000 cases reported by that time. There has been an increase in reporting delays during 1982–1989, as well as significant geographic variation in reporting delays. The actual incidence of AIDS is found to be rising most rapidly in nonurban regions and metropolitan areas with less than one million population. A model of AIDS incidence based upon the incubation of human immunodeficiency virus (HIV) is applied to data on reported AIDS cases among non-drug-using homosexual men. In this group, the estimated incidence of HIV infection peaked at about 8,900 cases per month in early 1983, and the resulting incidence of AIDS will peak at an estimated 3,200 cases per month in early 1993. The latter estimates were very sensitive to the specification of the incubation density for HIV infection.

KEY WORDS: Truncated data; Generalized linear models; Human immunodeficiency virus; Incubation period; Epidemic models.

1. INTRODUCTION

As of March 31, 1989, the U.S. Centers for Disease Control (CDC) had reported 90,990 cases of acquired immunodeficiency syndrome (AIDS). Yet by that date, physicians may have already diagnosed over 130,000 AIDS cases. The difference arises from significant delays in the reporting of AIDS cases to public health authorities.

This article addresses the statistical problem of estimating AIDS reporting delays and thus recovering the actual incidence of the disease. Reporting delays are not the only reason why CDC's surveillance data may fall short of the actual counts. Some cases of AIDS may go undetected permanently, and the official surveillance definition may not include all serious consequences of infection by the human immunodeficiency virus (HIV). These forms of underreporting, which can be viewed as reporting delays of infinite length, were studied elsewhere (General Accounting Office 1989), and will not be the main focus here.

The main statistical problem is one of data truncation and is motivated by Figure 1. For each reported case of AIDS, the CDC records both the date of diagnosis and the date of report. The former is the date on which the attending physician or hospital first identifies an AIDS-associated opportunistic disease in a particular patient. The latter is the date on which the CDC receives the patient's official case report from the local or state health department (Centers for Disease Control 1989). The solid line in Figure 1 shows the distribution by date of diagnosis of 90,616 AIDS cases that were reported by March 31, 1989. (Not shown are 374 AIDS cases that were diagnosed in 1980–1981.) The dashed line shows the counts of 41,649 AIDS cases (46%) that were reported within 2 months of

diagnosis. There is a spurious decline in diagnosed cases during 1988–1989 because many diagnosed cases are not yet reported.

2. BASIC STATISTICAL MODEL

Divide the time axis into intervals of equal length, called "months," which are indexed by the nonnegative integers. Let Y_u be the random variable representing the number of AIDS cases diagnosed in the month t and reported in month $t + u$. Attention is restricted to the months $t = 0, \dots, n$, where the known nonrandom integer n is the most recent month in which AIDS case reports can be received. It is assumed that the reporting delay u ranges from 0 up to a known finite maximum value m , where $m > n$.

The statistical objective is to draw inferences about an unknown vector θ of parameters that characterize the joint probability distribution of the random variables $\{Y_u\}$. The difficulty is that one can observe the realized values y_u of Y_u only when $t + u \leq n$. That is, only those AIDS cases are observable that have been diagnosed *and* reported by month n .

Assume that the random variables $\{Y_u\}$ are independent, and that each Y_u has a Poisson distribution with mean $\Phi_u(\theta)$, where the functional dependence of Φ_u on θ is further specified below. The joint probability function of the data $\{y_u\}$ given the unknown parameters θ is therefore proportional to

$$L(\theta) = \prod_{t=0}^n \prod_{u=0}^{n-t} [\Phi_u(\theta)]^{y_u} \exp[-\Phi_u(\theta)].$$

For each $t = 0, 1, \dots, n$, let the random variable $X_t = \sum_{u=0}^m Y_u$ denote the total number of AIDS cases diagnosed in month t , whether or not they are reported by month n . As a consequence of our Poisson independence assumption, the random variables $\{X_t\}$ are also independent.

* Jeffrey E. Harris is Associate Professor, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, and Clinical Associate, Medical Services, Massachusetts General Hospital, Boston, MA 02114. This work was supported by Grant 13446 from the Robert Wood Johnson Foundation. The opinions and conclusions expressed herein are the author's sole responsibility. An earlier, preliminary version of this article appeared as Harris (1987b). The author acknowledges the valuable criticisms of the Editorial Board and three anonymous referees.

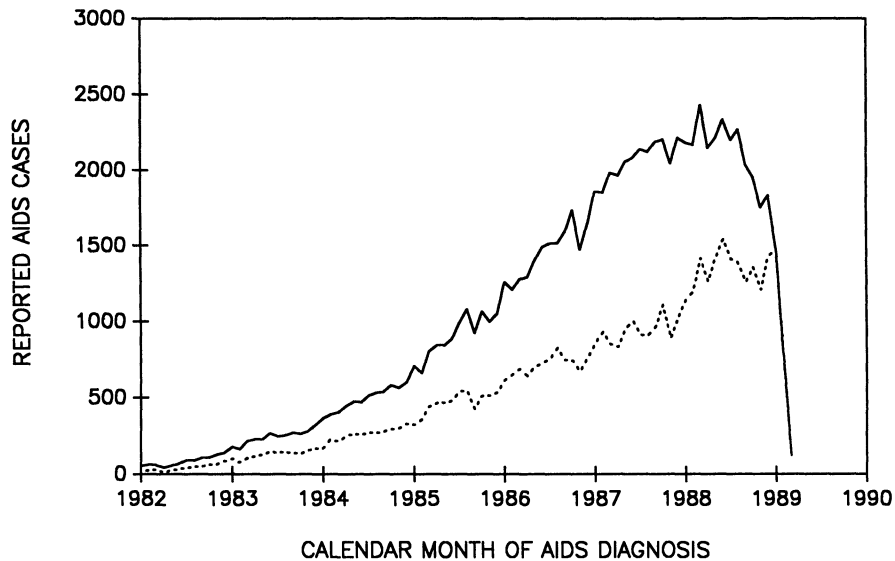


Figure 1. Reported AIDS Cases by Month of Diagnosis: —, Reported by March 31, 1989; ---, Reported Within 2 Months of Diagnosis.

dent Poisson with respective means

$$\Lambda_t(\theta) = \sum_{u=0}^m \Phi_{tu}(\theta),$$

where $\Lambda_t(\theta)$ is interpreted as the “incidence” of AIDS in month t . Moreover, for each t , the conditional distribution of $\{Y_{t0}, \dots, Y_{tm}\}$ given X_t is independent multinomial based upon X_t trials with probabilities $\Pi_{t0}(\theta), \dots, \Pi_{tm}(\theta)$, where

$$\Pi_{tu}(\theta) = \Phi_{tu}(\theta) / \Lambda_t(\theta)$$

is interpreted as the probability that a case of AIDS is reported with delay u , given that it is diagnosed in month t .

A separability (or “quasi-independence”) restriction is now imposed on the form of $\Phi_{tu}(\theta)$. Specifically, assume that the parameter vector θ can be partitioned as (α, β) so that

$$\Phi_{tu}(\alpha, \beta) = \Pi_{tu}(\alpha)\Lambda_t(\beta).$$

Thus the reporting delay probabilities Π_{tu} depend solely on α , while the monthly AIDS incidence Λ_t depends solely on β .

For each $t = 0, 1, \dots, n$, let the random variable $Z_t = \sum_{u=0}^{n-t} Y_{tu}$ denote the number of AIDS cases diagnosed in month t and reported by month n . In contrast to the random variables X_t , which are unobservable, the random variables Z_t have observable values $z_t = \sum_{u=0}^{n-t} y_{tu}$. For each t , the conditional distribution of Z_t given X_t is independent binomial based upon X_t trials with success probability $\Omega_t(\alpha)$, where

$$\Omega_t(\alpha) = \sum_{u=0}^{n-t} \Pi_{tu}(\alpha)$$

is the probability that a case of AIDS is reported by month n , given that it is diagnosed in month t . Moreover, for each t the conditional distribution of $\{Y_{t0}, \dots, Y_{t,n-t}\}$ given

Z_t is independent multinomial based upon Z_t trials with probabilities $\Pi_{t0}(\alpha)/\Omega_t(\alpha), \dots, \Pi_{t,n-t}(\alpha)/\Omega_t(\alpha)$, where each $\Pi_{tu}(\alpha)/\Omega_t(\alpha)$ is the conditional probability that a case of AIDS is reported with delay u , given that it is diagnosed in month t and reported by month n .

Accordingly, the likelihood function $L(\theta)$ can be rewritten as

$$L(\alpha, \beta) = L_c(\alpha)L_m(\alpha, \beta),$$

where

$$L_c(\alpha) = \prod_{t=0}^n \prod_{u=0}^{n-t} \left[\frac{\Pi_{tu}(\alpha)}{\Omega_t(\alpha)} \right]^{y_{tu}}$$

and

$$L_m(\alpha, \beta) = \prod_{t=0}^n [\Omega_t(\alpha)\Lambda_t(\beta)]^{z_t} \exp[-\Omega_t(\alpha)\Lambda_t(\beta)].$$

Thus $L(\alpha, \beta)$ is the product of the conditional likelihood $L_c(\alpha)$ (the joint probability function of $\{y_{tu}\}$ given $\{z_t\}$ and α) and the marginal likelihood $L_m(\alpha, \beta)$ (the joint probability function of $\{z_t\}$ given α and β). Statistical inference on α can be based on either the conditional likelihood $L_c(\alpha)$ or the unconditional likelihood $L(\alpha, \beta)$. In general, the use of $L_c(\alpha)$ ignores any information on α that may be contained in the marginal sums $\{z_t\}$. In analyses of AIDS reporting delays, Brookmeyer and Damiano (1989) relied on the conditional likelihood, while Zeger, See, and Diggle (1989) used the unconditional likelihood. Analyses of truncated data on HIV incubation (where Λ is interpreted as the incidence of HIV infection and Π is interpreted as the probability distribution of the time from HIV infection to the diagnosis of AIDS) similarly relied upon the conditional likelihood (Lui et al. 1986; Lagakos, Barraj, and De Gruttola 1988) or the unconditional likelihood (Medley, Anderson, Cox, and Billard 1987; Kalbfleish and Lawless 1989; Wang 1989). As pointed out by Kalbfleish and Lawless (1989), a likelihood of the form $L(\alpha, \beta)$ arises from Poisson sampling in the context of triangular incom-

plete contingency tables (Bishop, Fienberg, and Holland 1975).

The maximum likelihood estimate $\hat{\alpha}$ that is based on the conditional likelihood that $L_c(\alpha)$ alone would solve the first-order conditions

$$\sum_{t=0}^n \sum_{u=0}^{n-t} \left[\frac{y_{tu}}{\Pi_{tu}(\alpha)} - \frac{z_t}{\Omega_t(\alpha)} \right] \frac{d}{d\alpha} \Pi_{tu}(\alpha) = 0, \quad (2.1)$$

where $\hat{\Lambda}_t$ is estimated as $z_t/\Omega_t(\hat{\alpha})$. The maximum likelihood estimates (α^*, β^*) that are based upon the unconditional likelihood $L(\alpha, \beta)$ would solve the first-order conditions

$$\sum_{t=0}^n \sum_{u=0}^{n-t} \left[\frac{y_{tu}}{\Pi_{tu}(\alpha)} - \Lambda_t(\beta) \right] \frac{d}{d\alpha} \Pi_{tu}(\alpha) = 0 \quad (2.2)$$

and

$$\sum_{t=0}^n \left[\frac{z_t}{\Lambda_t(\beta)} - \Omega_t(\alpha) \right] \frac{d}{d\beta} \Lambda_t(\beta) = 0. \quad (2.3)$$

To complete the basic model, we now need to specify the functional forms of $\Pi_{tu}(\alpha)$ and $\Lambda_t(\beta)$. The remainder of this section presents models in which both Π_{tu} and Λ_t are linear functions of categorical variables indexed by t and u . In Sections 3 and 4, alternative models are considered in which Π_{tu} and Λ_t are continuous functions of t and u .

2.1 Categorical Models for Π

To illustrate the analysis of linear categorical models, we consider first the simplest specification for $\Pi_{tu}(\alpha)$. Let α_u be a separate parameter for each $u = 0, \dots, m$, and assume that $\Pi_{tu}(\alpha) = \alpha_u$ for all t and u . In the language of generalized linear models (GLM's) (McCullagh and Nelder 1983),

$$\Pi = U, \quad (2.4)$$

where U is a categorical factor indexed by u .

Under the model (2.4), the conditional likelihood $L_c(\alpha)$ depends only upon $\{\alpha_0, \dots, \alpha_n\}$. Hence the probability of reporting delays beyond n months cannot be identified from the data $\{y_{tu}\}$. Moreover, under (2.4) the conditional likelihood $L_c(\alpha)$ is homogeneous of degree 0 in the arguments $\{\alpha_0, \dots, \alpha_n\}$. Thus $\hat{\alpha}$ can be estimated only up to a proportionality constant. If the constraint $\sum_{u=0}^n \alpha_u = 1 - p$ is imposed, then each α_u can be estimated under the assumption that a known proportion p of cases are reported with a delay of more than n months.

The first-order conditions for the conditional maximum likelihood estimate $\hat{\alpha}$ reduce from (2.1) to

$$\sum_{t=0}^{n-u} \left[\frac{w_u}{\alpha_u} - \frac{z_t}{\Omega_t(\alpha)} \right] = 0 \quad (2.5)$$

for all $u = 0, \dots, n$, where $w_u = \sum_{t=0}^{n-u} y_{tu}$ is the total number of cases reported with a delay of u months, and $\Omega_t(\alpha) = \sum_{u=0}^{n-t} \alpha_u$. Although (2.5) admits a closed-form solution (Kalbfleish and Lawless 1989; Lagakos et al. 1988; Wang 1989), it is useful to show an iterative procedure, analogous to the EM algorithm (Dempster, Laird, and

Rubin 1977; Turnbull 1976), that can be generalized to other models. Consider estimates $\alpha^{(N)}$ obtained at the N th iteration of the algorithm. To determine the estimates $\alpha^{(N+1)}$ at the next stage, we first compute the quantities

$$a_u = w_u / \sum_{t=0}^{n-u} [z_t / \Omega_t(\alpha^{(N)})] \quad (2.6)$$

for each u . Then we normalize the values of a_u to sum to $1 - p$:

$$\alpha_u^{(N+1)} = (1 - p)a_u / \sum_{\mu=0}^n a_\mu. \quad (2.7)$$

Appropriate starting values are $\alpha_u^{(0)} = (1 - p)w_u / \sum_{\mu=0}^n w_\mu$.

In (2.6) and (2.7), one estimates the probability distribution of reporting delays by dividing w_u (the total observed number of cases reported with delay u) by the estimated mean of $\sum_{t=0}^n X_t$ (the total number of AIDS cases that will eventually be reported as having been diagnosed by month n). In essence, we compute the frequency distribution of observed reporting delays from an "augmented" sample (Turnbull 1976), in which each AIDS case that has been reported as diagnosed in month t is to be counted as $1/[(1 - p)\Omega_t]$ cases.

Under the categorical model $\Pi = U$, the unconditional likelihood $L(\alpha, \beta)$ also depends only on $\{\alpha_0, \dots, \alpha_n\}$, but it is not necessarily homogeneous of degree 0 in these parameters. That is, the model of AIDS incidence is now informative about reporting delays. However, if the model Λ_t includes a scale parameter κ [i.e., $\Lambda_t(\kappa, \beta) = \kappa^{-1}\Lambda_t(1, \beta)$] then the unconditional likelihood is homogeneous of degree 0 in the combined arguments $\{\alpha_0, \dots, \alpha_n, \kappa\}$. If a constraint is imposed on $\sum_{u=0}^n \alpha_u$, then the scale parameter κ is identifiable. Otherwise, one cannot distinguish between the scale of the epidemic and the proportion of cases that are reported with a delay exceeding n months.

The first-order condition (2.2) for the unconditional maximum likelihood estimate α^* reduces to

$$\sum_{t=0}^{n-u} \left[\frac{w_u}{\alpha_u} - \Lambda_t(\beta) \right] = 0. \quad (2.8)$$

To compute α^* , we generalize the iterative procedure described in (2.6) and (2.7) as follows. Consider estimates $\alpha^{(N)}$ obtained at the N th iteration. Holding $\alpha^{(N)}$ constant, we first compute the maximum likelihood estimates $\beta^{(N)}$ from the first-order conditions (2.3). Then we compute the quantities

$$a_u = w_u / \sum_{t=0}^{n-u} \Lambda_t(\beta^{(N)}) \quad (2.9)$$

and normalize the $\{a_u\}$ as in (2.7).

The model $\Pi = U$ implies that the reporting delay distribution is stationary. But there are reasons to suspect that it actually has changed over time. Increasing patient volume may have increased reporting delays in the face of fixed resources for state and local health departments. There is wide geographic variation in the methods of case

ascertainment, so reporting delays could have changed as the epidemic spread from its original urban epicenters. The CDC has also periodically revised its method of computer processing of AIDS case reports.

To accommodate possible nonstationarity, we let t_1, \dots, t_k be known integers with $0 = t_1 < \dots < t_k < n$. We let $j(t)$ be the largest value of j such that $t_j \leq t$. As an alternative to (2.4), we assume the existence of separate parameters α_{ju} for each u and each $j = 1, \dots, k$, and specify $\Pi_u(\alpha) = \alpha_{j(t),u}$ for all u and t . In the modeling language of GLM's,

$$\Pi = J(T) * U, \tag{2.10}$$

where $J(T)$ is a function that produces a categorical variable with k levels over the index t , and the interaction operator "*" means that a separate parameter is to be estimated for each level j and each duration u .

The estimation of parameters α_{ju} in the model (2.10) proceeds by a straightforward modification of the iterative procedures given previously. The conditional likelihood $L_c(\alpha)$ is now separately homogeneous of degree 0 in the arguments $\{\alpha_{ju}; u = 0, \dots, n - t_j\}$ for each $j = 1, \dots, k$. Hence we need k restrictions to identify the parameters, one for each interval. A simple set of continuity restrictions is

$$\sum_{u=0}^n \alpha_{1u} = 1 - p;$$

$$\text{and } \sum_{u=0}^{n-t_j} \alpha_{ju} = \sum_{u=0}^{n-t_{j-1}} \alpha_{j-1,u} \text{ for all } j > 1. \tag{2.11}$$

By contrast, the unconditional likelihood $L(\alpha, \beta)$ is now homogeneous in the combined arguments $\{\alpha, \kappa\}$, where κ is a scale parameter of Λ_t . Hence only a single restriction on α or κ is required to identify the parameters.

2.2 Categorical Model for Λ

Assume that there exist separate parameters $\{\beta_0, \dots, \beta_n\}$ such that $\Lambda_t(\beta) = \beta_t$ for all t . In the language of GLM's,

$$\Lambda = T, \tag{2.12}$$

where T is a categorical factor indexed by t .

An immediate consequence of (2.12) is that the estimate $\hat{\alpha}$ based upon the conditional likelihood $L_c(\alpha)$ is identical to the estimate α^* based upon the unconditional likelihood $L(\alpha, \beta)$. This result obtains independent of the functional form of $\Pi_u(\alpha)$. Thus for any arbitrarily fixed value of α , the solution of (2.3) becomes $\beta_t^*(\alpha) = z_t / \Omega_t(\alpha)$. The insertion of $\beta_t^*(\alpha)$ into (2.2) yields an equation identical to (2.1), the first-order conditions for maximizing $L_c(\alpha)$. Put differently, the conditional likelihood $L_c(\alpha)$ is identical to the unconditional profile likelihood $L_p(\alpha) = L(\beta^*(\alpha), \alpha)$. In the special case where the incidence of AIDS in each month t is a separate parameter β_t , and the reporting delay probabilities do not depend upon β , one may just as well condition on $\{z_t\}$, which provide no additional information on α .

2.3 Estimates for Categorical Model:

$$\Pi = J(T) * U; \Lambda = T$$

Figure 2 compares the counts of reported cases with estimated counts of diagnosed AIDS cases. The former correspond to the data on z_t and are reproduced from Figure 1. The latter correspond to the estimates β_t^* under models (2.10) and (2.12). As Figure 2 shows, a significant fraction of AIDS cases had not been reported by March 31, 1989, even among those diagnosed in 1987. For example, 1,856 AIDS cases were reported as diagnosed in January 1987. Yet $\Omega_t(\alpha^*)$ for that month was 0.790 (standard error of 0.0034), and this gave an estimated incidence $\beta_t^* = z_t / \Omega_t(\alpha^*)$ of 2,350. [The standard errors (SE's) were computed from the information matrix of $L_c(\alpha)$.] Figure

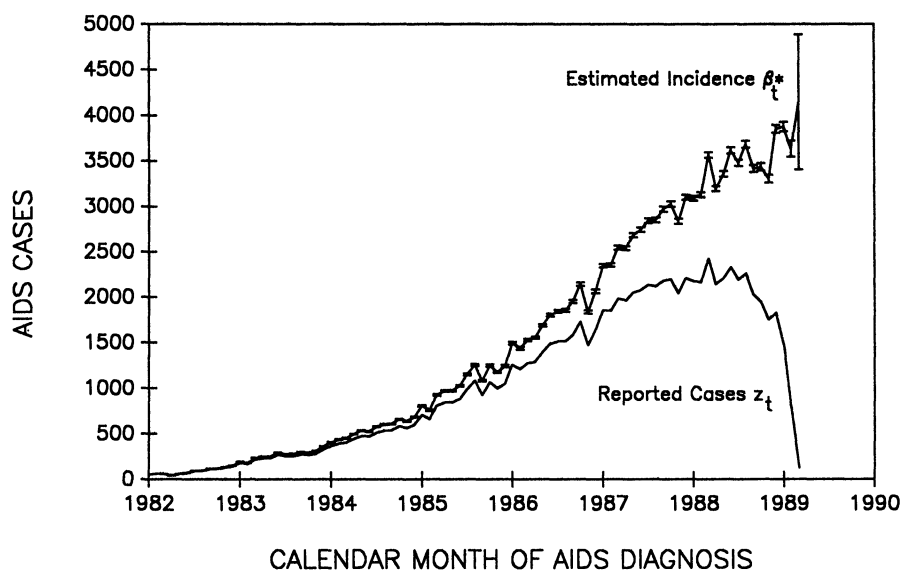


Figure 2. Reported AIDS Cases, z_t , and Estimated Incidence, β_t , by Month of Diagnosis, t . The error bars are 95% confidence intervals of β_t .

2 reports sampling errors of the estimates β_i^* , and does not include the prediction uncertainty arising from the Poisson distribution of the counts X_t .

In the estimation of the model $\Pi = J(T) * U$, the observation period was partitioned into $k = 4$ intervals: (1) January 1982–March 1983, during which the CDC encoded the date of case report as the date of receipt by the state or local health department; (2) April 1983–July 1985, when the date of report was changed to reflect the date received by the CDC; (3) August 1985–September 1987, the period during which the CDC’s 1985 revisions in the AIDS surveillance definition were in effect; and (4) October 1987–March 1989, during which the most recent 1987 revisions in case definition were in effect. Moreover, it was assumed that $p = 0$ in (2.11); that is, all diagnosed cases are fully reported by $n = 86$ months.

Significant nonstationarity in the reporting delay distribution was found. [The log likelihood of the model $\Pi = J(T) * U$, which contained 131 additional parameters, exceeded that of the model $\Pi = U$ by 1145.7. By way of comparison, the critical value of $\chi^2(131)$ for rejecting $\Pi = U$ at the 1% significance level was 171.] Figure 3 compares the estimated reporting delay distribution during the first period (that is, α_{1u}^*) with that of the last period (that is, α_{4u}^*). The estimated proportions of cases reported within 2 months of diagnosis (that is, $\alpha_{j0}^* + \alpha_{j1}^* + \alpha_{j2}^*$) fell from 47% (SE of 1.2%) in the first period to 38% (SE of 0.3%) in the last period. The estimated probability distribution had a very long right-hand tail, with an estimated 16.8% of cases reported after a delay of 3 years.

3. MIXED CATEGORICAL/CONTINUOUS-TIME MODELS OF REPORTING DELAY, $\Pi_{tu}(\alpha)$

In this section, the data are further cross-classified by geographic region. Let Y_{tur} be a random variable denoting the number of cases in region $r = 0, 1, \dots, l$ that are

diagnosed in month t and reported in month $t + u$. Assume that each Y_{tur} is independent Poisson with mean $\Pi_{tur}(\alpha)\Lambda_r(\beta)$, where Λ_r is the incidence of AIDS in month t and region r , and Π_{tur} is the probability that an AIDS case, diagnosed in region r during month t , is reported in month $t + u$. The observed data consist of counts y_{tur} , classified in three ways. Define $z_{tr} = \sum_{u=0}^{n-t} y_{tur}$.

Retain a linear categorical model for AIDS incidence. That is, in a manner analogous to (2.12), assume that $\Lambda_r(\beta) = \beta_{tr}$ for all t and r . In the language of GLM’s,

$$\Lambda = T * R, \tag{3.1}$$

where T and R are categorical factors indexed by t and r .

With respect to Π_{tur} , partition the parameter vector α as (η, γ, δ) , and write

$$\log \Pi_{tur}(\eta, \gamma, \delta) = \eta_{j(t),u} + \gamma_{j(t),r} + \delta_{j(t),r}u$$

for all t, u , and r . Since only the contrasts $(\gamma_{jr} - \gamma_{j'r'})$ and $(\delta_{jr} - \delta_{j'r'})$ can be identified, it is assumed that $\gamma_{j0} = \delta_{j0} = 0$, which means that $\log \Pi_{tu0}(\eta, \gamma, \delta) = \eta_{j(t),u}$. In the language of GLM’s,

$$\log \Pi = J(T) * (U + R + R \cdot u), \tag{3.2}$$

where u is a continuous covariate and the “ \cdot ” operator means that a separate slope term δ_{jr} for u is estimated for each j and each r .

The mixed categorical–continuous model (3.2) requires that for any two diagnosis months t and t' satisfying $j(t) = j(t') = j$, and for any region $r > 0$, the Laplace transforms of Π_{tur} and $\Pi_{t'ur}$, denoted by $\Pi_{tr}^*(s)$ and $\Pi_{t'r}^*(s)$, respectively, satisfy the relation $\Pi_{tr}^*(s) = \Pi_{t'r}^*(s + \delta_{jr}) / \Pi_{t'r}^*(\delta_{jr})$. Such a restriction is satisfied by the class of compound Poisson distributions $\Pi_{tur} = \int [\xi^u \exp(-\xi) / u!] f_{jr}(\xi) d\xi$, where $f_{jr}(\xi)$ is the mixing density of the Poisson parameter ξ . In that case, the quantities $\exp(\delta_{jr})$ operate as scale factors on the mixing densities. Thus if $j(t) = j(t')$,

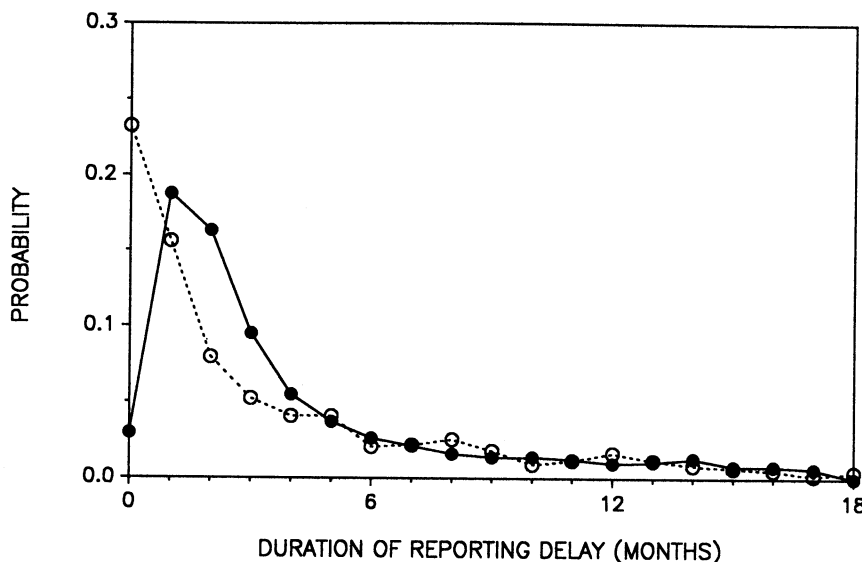


Figure 3. Estimated Reporting Delay Distributions Π_{tu} for Cases Diagnosed January 1982–March 1983 ($j = 1$, open circles) and Cases Diagnosed October 1987–March 1989 ($j = 4$, closed circles).

then the respective mixing densities of Π_{tur} and $\Pi_{t'u_0}$ are related by $f_{jr}(\xi) = f_{j0}(\xi \exp(\delta_{jr}))$.

Under the categorical model $\Lambda = T * R$, the data $\{z_{tr}\}$ are not informative about the parameters $\alpha = (\eta, \gamma, \delta)$. Accordingly, as in Section 2.2, one may work with the conditional (profile) likelihood

$$L_c(\alpha) = \prod_{r=0}^l \prod_{t=0}^n \prod_{u=0}^{n-t} \left[\frac{\Pi_{tur}(\alpha)}{\Omega_{tr}(\alpha)} \right]^{y_{tur}}$$

where $\Omega_{tr}(\alpha) = \sum_{u=0}^{n-t} \Pi_{tur}(\alpha)$. This likelihood depends only on η and δ . However, once η and δ are estimated, the auxiliary parameters γ_{jr} can be identified by restrictions on Π_{tur} that are analogous to (2.11):

$$\sum_{u=0}^n \exp(\eta_{1u} + \gamma_{1r} + \delta_{1r}u) = 1 - p_r$$

for all $r = 1, \dots, l$ and known $\{p_r; r = 1, \dots, l\}$, and

$$\sum_{u=0}^{n-t_j} \exp(\eta_{ju} + \gamma_{jr} + \delta_{jr}u) = \sum_{u=0}^{n-t_j} \exp(\eta_{j-1,u} + \gamma_{j-1,r} + \delta_{j-1,r}u) \quad (3.3)$$

for all $r = 1, \dots, l$ and $j = 2, \dots, k$.

A suitable modification of (2.1) yields the following first-order conditions for maximization of L_c with respect to η and δ :

$$\sum_{r=0}^l \sum_{t=0}^n \sum_{u=0}^{n-t} \left(y_{tur} - \frac{z_{tr}}{\Omega_{tr}} \Pi_{tur} \right) \frac{d}{d\eta} \log \Pi_{tur} = 0 \quad (3.4)$$

and

$$\sum_{r=0}^l \sum_{t=0}^n \sum_{u=0}^{n-t} \left(y_{tur} - \frac{z_{tr}}{\Omega_{tr}} \Pi_{tur} \right) \frac{d}{d\delta} \log \Pi_{tur} = 0, \quad (3.5)$$

where Π_{tur} and Ω_{tr} depend upon (η, γ, δ) . The first-order conditions (3.4) and (3.5), along with the constraints (3.3),

form the basis of an iterative procedure comparable to that given in Section 2. In particular, the conditions (3.4) can be rewritten as a set of equations that are linear in the quantities $\exp(\eta_{ju})$ for each j and u , which can then be solved subject to the normalization that $\sum_{u=0}^n \exp(\eta_{ju}) = 1 - p_0$. The conditions (3.5) can be rewritten as a set of polynomial equations of order n in the quantities $\exp(\delta_{jr})$ for each j and $r > 1$. In each polynomial equation, the zero-order term is negative, while all other coefficients are positive, and therefore there is a unique real root.

Figure 4 shows some results from the model of (3.1) and (3.2). The CDC identifies the region of residence of each reported nonpediatric AIDS victim who lives in a Metropolitan Statistical Area (MSA) with more than one million population (Centers for Disease Control 1989). Six regions ($r = 0, \dots, 5$) are identified: Northeast, Central, West, South, and Mid-Atlantic, as well as a Residual Category of patients not living in large MSA's. Prior to September 1987 ($j = 1, 2, 3$), the Western region ($r = 2$) and the Residual Category ($r = 5$) were found to have significantly longer reporting delays. After September 1987 ($j = 4$), the Southern region ($r = 4$) and Residual Category ($r = 5$) had longer reporting delays. For the Residual Category, the estimates of δ_{j5}^* (relative to the Northeast region, for which $r = 0$) were $\delta_{15}^* = 0.015$ (SE of 0.002) for January 1982–March 1983, $\delta_{25}^* = 0.027$ (SE of 0.001) for April 1983–July 1985, $\delta_{35}^* = 0.027$ (SE of 0.001) for August 1985–September 1987, and $\delta_{45}^* = 0.021$ (SE of 0.002) for October 1987–March 1989. The estimate $\Omega_{tr}(\eta^*, \gamma^*, \delta^*)$ for January 1989 was 0.455 for the Northeast ($r = 0$) region versus 0.331 for AIDS cases outside large MSA's ($r = 5$).

The estimated differences in reporting delays are reflected in the estimated incidence rates $\Lambda_t(\beta^*) = \beta_{tr}^*$ for the Northeast MSA's and the Residual Category, shown in Figure 4 on a logarithmic scale. The current growth rate of the epidemic is considerably higher among patients who did not reside in a large MSA. The epidemic doubling

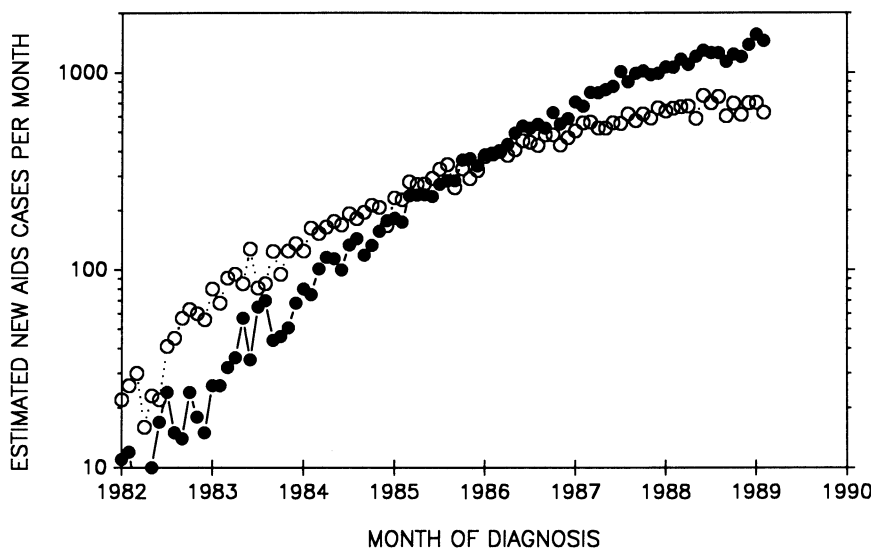


Figure 4. Estimated Counts of Diagnosed AIDS Cases (logarithmic scale) for Residents of Large MSA's in the Northeast Census Region (open circles) and for Residents of Small MSA's or Nonurban Areas Throughout the U.S. (closed circles).

time is about 5 months for the former group and about 2 years for the latter. Higher epidemic growth rates were also seen for estimated counts for large MSA's in the Central, South, and Mid-Atlantic regions, as compared to the Northeast and West. These results suggest that the epidemic has diffused out of the major urban centers in which it began.

The estimates of reporting delays derived from (3.2) revealed significant interactions between date of diagnosis (T) and geographic region (R). Such a conclusion differs from that of Brookmeyer and Damiano (1989), who found the effect of T to be largely explained by the effect of R . However, these authors relied upon the more restricted, noninteractive model $\log \Pi = T + U + R$.

4. CONTINUOUS-TIME MODEL OF AIDS INCIDENCE $\Lambda_t(\beta)$

The specification $\Lambda = T$, used in Sections 2 and 3, permits one to rely upon the conditional likelihood $L_c(\alpha)$ and thus to ignore any information on AIDS incidence that is contained in the marginal sums $\{z_i\}$. An alternative specification for $\Lambda_t(\beta)$ may yield more precise estimates of both Λ and Π , but with the important proviso that one must specify the right model. In an earlier analysis of AIDS reporting delays, Harris (1987b) used the model $\log \Lambda = t + t^2$, where t is interpreted as a continuous variable. Studies of HIV incubation (Wang 1989; Kalbfleish and Lawless 1989; Medley et al. 1987) have used variants of $\log \Lambda = t$ for HIV incidence. These simple exponential models are useful for illustrative purposes, but there is a serious question about their accuracy.

To model AIDS incidence in the 1980s, one needs to extend the time axis back to the mid-1970s, when the spread of HIV infection is thought to have begun. Define the continuous time variable s , where $s = 0$ corresponds to the approximate start of the HIV epidemic. Let $t = s - \tau$, so that the first AIDS diagnoses were recorded at $s = \tau$, that is, at $t = 0$. Then assume

$$\Lambda_t(\beta) = \int_0^{t+\tau} h(s; \beta) f(t + \tau - s) ds. \quad (4.1)$$

Here $h(s; \beta)$ is the incidence of HIV infection at date s , while $f(\cdot)$ is the density of incubation times from infection to disease, which is assumed to be known a priori and independent of HIV incidence. The estimation of β thus entails inversion of the integral equation (4.1), a deconvolution procedure that has been termed "working backwards" or "back calculation" by AIDS researchers (Harris 1987a; Gail and Brookmeyer 1988).

The inversion of (4.1) is known to be an ill-posed problem (Mendelsohn and Rice 1982; Tikhonov and Arsenin 1977; McMahan, Maxwell, and Shepherd 1986). In fact, when there are no restrictions on the form of $h(s)$, the solution can display very wide, high-frequency oscillations. Accordingly, assume that $h(s)$ is a continuous piecewise linear spline function with known nodes at $s_0 < s_1 < \dots < s_l < s_{l+1}$. That is, for $s_i \leq s \leq s_{i+1}$,

$$h(s; \beta) = \beta_i \frac{s_{i+1} - s}{s_{i+1} - s_i} + \beta_{i+1} \frac{s - s_i}{s_{i+1} - s_i}, \quad (4.2)$$

while $h(s; \beta) = 0$ for $s < s_0$ and $s > s_{l+1}$. Now define

$$F_{ii} = \int_{t+\tau-s_{i+1}}^{t+\tau-s_i} f(s) ds \quad \text{and} \quad G_{ii} = \int_{t+\tau-s_{i+1}}^{t+\tau-s_i} sf(s) ds,$$

where it is understood that $f(s) = 0$ for $s < 0$. Inserting (4.2) into (4.1) yields

$$\Lambda_t(\beta) = \sum_{i=0}^l \left[\beta_i \frac{G_{ii} - (t + \tau - s_{i+1})F_{ii}}{s_{i+1} - s_i} + \beta_{i+1} \frac{(t + \tau - s_i)F_{ii} - G_{ii}}{s_{i+1} - s_i} \right],$$

which can be rewritten as $\Lambda_t(\beta) = \sum_{i=0}^{l+1} c_{ii}\beta_i$. In the language of GLM's,

$$\Lambda = C_0(t) + \dots + C_{l+1}(t), \quad (4.3)$$

where each $C_i(t)$ is a function that produces a continuous covariate from the continuous variable t . The linear spline (4.2) is not the only class of functions for which (4.3) obtains.

The first-order condition (2.3) for the maximum likelihood estimate of β is now

$$\sum_{i=0}^n \frac{c_{ii}}{\Lambda_t(\beta)} \left[z_i - \Omega_t(\alpha) \sum_{i=0}^{l+1} c_{ii}\beta_i \right] = 0 \quad (4.4)$$

for all i . If A is a matrix with typical element $a_{ii} = \Omega_t(\alpha)c_{ii}$ and W is the matrix with typical element $c_{ii}/\Lambda_t(\beta)$, then (4.4) becomes $W'(z - A\beta) = 0$, where W depends on β and A depends on α and β . Given α , this system can be solved by iteratively reweighted least squares.

The model ($\Pi = J(T) * U$; $\Lambda = \sum_i C_i$) was estimated for 52,816 AIDS cases among adult homosexual and bisexual men with no intravenous drug use or no other risk factors for AIDS (e.g., hemophilia, transfusion with infected blood), who were reported during January 1981–March 1989. Studies of HIV serology among selected cohorts of homosexual and bisexual men suggest that the incidence of new HIV infection rose from the late 1970s until 1982–1984, after which incidence rates declined (Stevens et al. 1986; Centers for Disease Control 1987a). To incorporate this prior information, we set β_0 equal to 0 at $s_0 =$ January 1976 and β_{l+1} equal to 0 at $s_{l+1} =$ January 1988. The intermediate nodal points were $s_1 =$ January 1980, $s_2 =$ January 1981, $s_3 =$ January 1982, $s_4 =$ January 1983, and $s_5 =$ January 1984.

For the incubation density, the Weibull was used:

$$f(s) = \omega \rho^\omega s^{\omega-1} \exp[-(\rho s)^\omega].$$

Denoting $v_i = [\rho(t + \tau - s_i)]^\omega$, one then has $F_{ii} = \exp(-v_i) - \exp(-v_{i+1})$ and $G_{ii} = I[v_i, (1 + \omega)/\omega]/\rho - I[v_{i+1}, (1 + \omega)/\omega]/\rho$, where $I[x, b] = \int_0^x \xi^{b-1} \exp(-\xi) d\xi$ is the incomplete gamma function. In particular, $\omega = 2.516$ and $\rho = 7.18 \times 10^{-3}$ per month were taken from Brookmeyer and Goedert (1989), which predicts that 25% of HIV-infected persons will have AIDS within 7 years and 50% will have AIDS within 10 years.

Figure 5 shows the estimated incidence of HIV infection, $h(s; \beta^*)$, which peaked at about 8,900 cases per

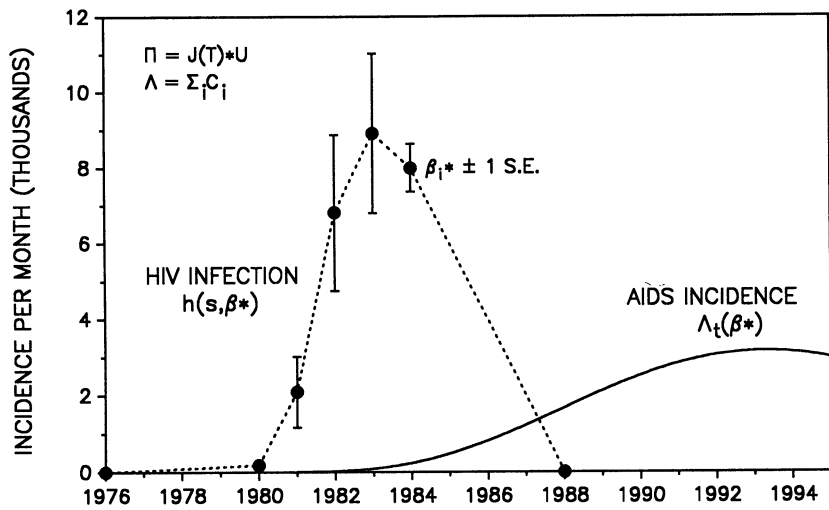


Figure 5. Estimated HIV Infections, $h(s; \beta)$, and Predicted AIDS Incidence $\Lambda_t(\beta)$ From the Back-Calculation Model for Non-Drug-Using Homosexual Men.

month (SE of 2,100 per month) in early 1983. The cumulative HIV incidence (the area under the dashed curve) was 459,000 cases (SE of 41,700). The solid line in the figure shows the predicted incidence of AIDS, $\Lambda_t(\beta^*)$, which I have extended beyond the 1981–1989 sample period. The predicted incidence of AIDS peaks in early 1993 at about 3,150 new cases per month (SE of 290 per month). While the estimates of β_i exhibited large standard errors, the pairs $(\beta_i^*, \beta_{i+1}^*)$ exhibited highly negative correlations in the range -0.85 to -0.94 , so that the estimates of cumulative HIV infection and predicted AIDS incidence showed less uncertainty.

Figure 6 compares the estimated incidence of AIDS, $\Lambda_t(\beta^*)$, for the continuous-time model [$\Pi = J(T) * U$; $\Lambda = \sum_i C_i$] (solid line) with corresponding estimates $\Lambda_t(\hat{\beta})$ for the categorical model [$\Pi = J(T) * U$; $\Lambda = T$] (closed circles). The dotted lines are the 95% confidence bounds around the continuous-time model. (These bounds are not

prediction intervals; they do not reflect the Poisson uncertainty in the counts X_i .) Also shown in the figure are the counts z_i of reported cases.

The categorical estimates $\Lambda_t(\hat{\beta})$ of AIDS incidence fall entirely within the confidence limits of the categorical estimates $\Lambda_t(\beta^*)$. Beginning in late 1987, however, the categorical estimates lie mostly below the continuous-time model. This deviation is reflected in the differences between $\hat{\alpha}_{4u}$ and α_{4u}^* . Thus for the categorical model, the estimated probability of case report within 2 months of diagnosis (that is, $\hat{\alpha}_{40} + \hat{\alpha}_{41} + \hat{\alpha}_{42}$) was 41.5%, while for the continuous-time back-calculation model, the estimate $\alpha_{40}^* + \alpha_{41}^* + \alpha_{42}^*$ was 38.2%.

If the model $\Lambda = \sum_i C_i$ is accurate, then there appear to have been greater reporting delays than are captured by the categorical model $\Lambda = T$, which does not use the information contained in the marginal sums $\{z_i\}$. On the other hand, the back-calculation model may be inaccurate.

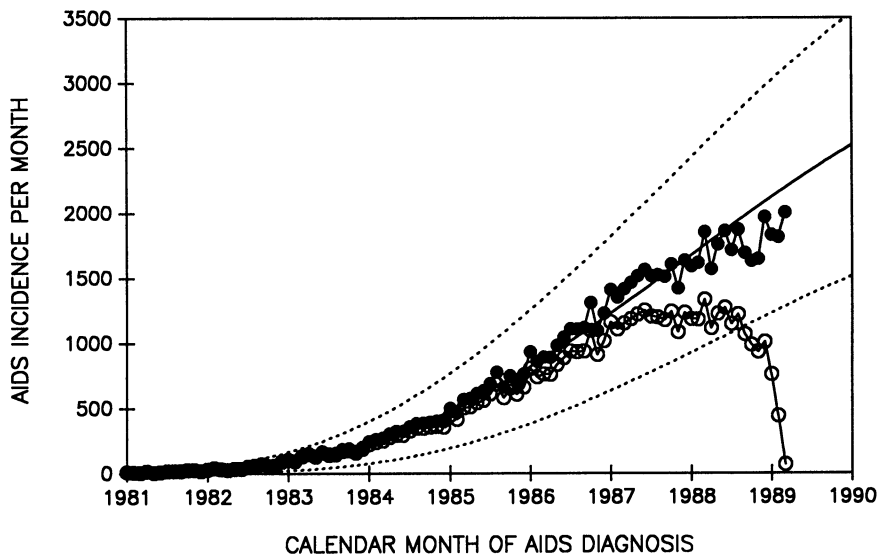


Figure 6. AIDS Case Reports and Estimated AIDS Incidence Among Non-Drug-Using Homosexual Men. The predicted values of Λ_t from the model $\Lambda = \sum_i C_i(t)$ are given by the solid line, with 95% confidence bounds given by the dotted lines. The closed circles correspond to the predicted values of Λ_t from the categorical model $\Lambda = T$. The open circles correspond to case reports z_i .

To test the latter possibility, we varied the locations of the nodes $\{s_i\}$ in the 1980–1984 range, but this had little effect on the results. For example, when s_5 was moved from January 1984 to July 1984, the estimated HIV incidence still peaked in early 1984 at $s_4 = 9,700$ per month, while the estimate of β_5 was 6,700 per month, and that of cumulative HIV infection was 455,800 cases. However, the estimates showed more sensitivity to the specification of $f(s)$, the HIV incubation curve, which is not known with great precision (Harris 1988). For example, specifying $\omega = 2.571$ and $\rho = 9.545 \times 10^{-3}$ per month (Lui, Darrow, and Rutherford 1988), which has a median incubation of 7.5 years, gave an estimated incidence of HIV infection with a first peak at 7,700 cases per month in mid-1982, which then fell to 3,800 per month in early 1984, but rose again to 4,900 per month in mid-1985. The estimate of cumulative HIV infection was 333,000 (SE of 20,000). The predicted incidence of AIDS during 1984–1987 exceeded that shown in Figure 6, and it peaked earlier, at 2,950 cases per month in mid-1991.

5. COMMENTS

The current findings need to be interpreted with caution. The estimates of the reporting delay probabilities $\Pi_u(\alpha)$ were conditional upon complete reporting by 7 years after diagnosis. The unconditional distribution, however, is likely to be defective, with a substantial proportion of cases never reported. Increasing underreporting of AIDS cases, in fact, may have contributed to the apparent slowing of disease incidence in large MSA's in the Northeast and West (Fig. 4).

The random variables Y_{it} were assumed to be independent Poisson. As a consequence, the number of AIDS cases X_t in each month followed a discrete-time Poisson process with time-dependent means $\Lambda_t(\beta)$, while the distribution of the counts $\{Y_{it}\}$ given X_t became multinomial with probabilities $\Pi_u(\alpha)$. Moreover, the models $\Pi_u(\alpha)$ and $\Lambda_t(\beta)$ were assumed to have no common parameters. These assumptions, however, may be unwarranted.

AIDS is a consequence of earlier HIV infection, and the incubation period between infection and disease is long and variable. Accordingly, we would not expect the counts X_t of AIDS cases to be independent Poisson. Even if the counts $\{X_t\}$ have some arbitrary joint distribution, our use of the conditional likelihood $L_c(\alpha)$ will produce asymptotically equivalent estimates of α so long as the conditional distribution of $\{Y_{it}\}$ given X_t is independent multinomial for each t . The latter assumption, however, is open to serious challenge. While the most significant reporting delays occur between the time of diagnosis and the date of notification to a state or local health department (Centers for Disease Control 1989), there are also delays at the state and local health departments. Many such departments, in fact, report AIDS cases to the CDC in batches. This lumpiness in the pattern of case reporting is not captured by the independent multinomial assumption.

The results of Section 3 illustrate the point that both

AIDS incidence and AIDS reporting delays may depend on other covariates, such as geographic region. These findings highlight the restrictiveness of the quasi-independence assumption that Λ and Π share no common parameters. More general models of $\Phi_u(\theta)$ need to be explored.

The estimates of HIV incidence $h(s; \beta)$ among non-drug-using homosexual men are qualitatively consistent with serological data on selected cohorts. The reported sampling errors in Figures 5 and 6, it needs to be emphasized, do not reflect uncertainty in the HIV incubation density $f(s)$. In fact, the assumption that $f(s)$ is stationary is challengeable. The composition of the HIV-infected population may have changed over time. The first manifestations of AIDS may now be diagnosed at an earlier stage of HIV infection. Introduction of zidovudine (AZT) and chemoprophylaxis for *Pneumocystis carinii* pneumonia in 1987–1989 (Harris 1990) may have altered the natural history of infection.

In September 1987, the CDC implemented significant revisions in its surveillance definition of AIDS (Centers for Disease Control 1987b). The current analysis permitted the reporting delay distribution Π to change after September 1987 but made no distinction between AIDS cases satisfying the old surveillance definition and those satisfying the new definition only, and the estimates of AIDS incidence are for both types of cases combined. The analysis of AIDS incidence and reporting by type of AIDS-associated disease needs further study.

[Received June 1987. Revised April 1990.]

REFERENCES

- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.
- Brookmeyer, R., and Damiano, A. (1989), "Statistical Methods for Short-Term Projections of AIDS Incidence," *Statistics in Medicine*, 8, 23–34.
- Brookmeyer, R., and Goedert, J. (1989), "Censoring in an Epidemic, With an Application to Hemophilia-Associated AIDS," *Biometrics*, 45, 325–335.
- Centers for Disease Control (1987a), "Human Immunodeficiency Virus Infection in the United States: A Review of Current Knowledge," *Morbidity and Mortality Weekly Report*, 36 (Suppl. 6), 1–48.
- (1987b), "Revision of the CDC Surveillance Case Definition for Acquired Immunodeficiency Syndrome," *Morbidity and Mortality Weekly Report*, 36 (Suppl. 1), 3–15.
- (1989), *AIDS Public Information Data Set*, Atlanta, GA: Centers for Disease Control, Center for Infectious Diseases, Division of HIV/AIDS.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22.
- Gail, M. H., and Brookmeyer, R. (1988), "Methods for Projecting Course of Acquired Immunodeficiency Syndrome Epidemic," *Journal of the National Cancer Institute*, 80, 900–911.
- General Accounting Office (1989), *AIDS Forecasting: Undercount of Cases and Lack of Key Data Weaken Existing Estimates*, Washington, DC: Author.
- Harris, J. E. (1987a), "The AIDS Epidemic: Looking Into the 1990s," *Technology Review*, 90, 59–64.
- (1987b), "Delay in Reporting Acquired Immunodeficiency Syndrome (AIDS)," Working Paper 2278, National Bureau of Economic Research, Cambridge, MA.
- (1988), "The Incubation Period for HIV-1," in *AIDS 1988: AAAS Symposia Papers*, ed. R. Kulstad, Washington, DC: American Association for the Advancement of Science, pp. 64–74.
- (1990), "Improved Short-Term Survival of AIDS Patients Initially

- Diagnosed With Pneumocystis Carinii Pneumonia, 1984 Through 1987," *Journal of the American Medical Association*, 263, 397-402.
- Kalbfleish, J. D., and Lawless, J. F. (1989), "Inference Based on Retrospective Ascertainment. An Analysis of the Data on Transfusion Related AIDS," *Journal of the American Statistical Association*, 84, 360-372.
- Lagakos, S. W., Barraj, L. M., and De Gruttola, V. (1988), "Nonparametric Analysis of Truncated Survival Data, With Application to AIDS," *Biometrika*, 75, 515-523.
- Lui, K.-J., Darrow, W. W., and Rutherford, G. W., III (1988), "A Model-Based Estimate of the Mean Incubation Period for AIDS in Homosexual Men," *Science*, 240, 1333-1335.
- McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman & Hall.
- McMahan, C. A., Maxwell, L. C., and Shepherd, A. P. (1986), "Estimation of the Distribution of Blood Vessel Diameters From the Arteriovenous Passage of Microspheres," *Biometrics*, 42, 371-380.
- Medley, G. F., Anderson, R. M., Cox, D. R., and Billard, L. (1987), "Incubation Period of AIDS in Patients Infected Via Blood Transfusion," *Nature*, 328, 719-721.
- Mendelsohn, J., and Rice, J. (1982), "Deconvolution of Microfluorometric Histograms With B Splines," *Journal of the American Statistical Association*, 77, 748-753.
- Stevens, C. E., Taylor, P. E., Zang, E. A., Morrison, J. M., Harley, E. J., Rodriguez de Cordoba, S., Bacino, C., Ting, R. C. Y., Bodner, A. J., Sarngadharan, M. G., Gallo, R. C., and Rubinstein, P. (1986), "Human T-Cell Lymphotropic Virus Type III Infection in a Cohort of Homosexual Men in New York City," *Journal of the American Medical Association*, 255, 2167-2172.
- Tikhonov, A., and Arsenin, V. (1977), *Solutions of Ill-Posed Problems*, New York: John Wiley.
- Turnbull, B. W. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society, Ser. B*, 38, 290-295.
- Wang, M.-C. (1989), "A Semiparametric Model for Randomly Truncated Data," *Journal of the American Statistical Association*, 84, 742-748.
- Zeger, S. L., See, L., and Diggle, P. J. (1989), "Statistical Methods for Monitoring the AIDS Epidemic," *Statistics in Medicine*, 8, 3-22.