working paper
department
of economics

DELAY IN REPORTING ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS)

Jeffrey E. Harris

No. 452 May 1987

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

DELAY IN REPORTING
ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS)

Jeffrey E. Harris

No. 452 May 1987

May 1987

Delay in Reporting

Acquired Immune Deficiency Syndrome (AIDS)

by

Jeffrey E. Harris*

ABSTRACT

As of March 31, 1987, the U.S. Centers for Disease Control had reported 33,350 cases of acquired immune deficiency syndrome. Yet by that date, physicians had actually diagnosed 42,670 cases. The difference arises from significant delays in the reporting of AIDS cases to public health authorities. An estimated 70% of cases are reported two or more months after diagnosis; about 23% are reported seven or more months later; and about 5% take more than three years to come in. Moreover, the probability distribution of delays has been shifting to the right, with the median delay increasing by 0.6 months since mid-1986. From the data on reported cases and the estimated probability distribution of reporting delays, I reconstruct the actual incidence of AIDS from January 1982 through March 1987. The doubling time of the epidemic fell from about 6 months in 1982 to 15-16 months in 1986.

KEYWORDS: empirical distribution; truncated data; EM algorithm; non-parametric models; semi-parametric models; epidemic; incidence.

*Massachusetts Institute of Technology and Massachusetts General Hospital. Address reprint requests to the author at Massachusetts Institute of Technology, Department of Economics, E52-252g, Cambridge, Massachusetts 02139. I gratefully acknowledge the assistance of the AIDS Program at the Centers for Disease Control, especially Dr. W. Meade Morgan and Kathy Winter. The contents this paper are my sole responsibility.

1. INTRODUCTION

As of March 31, 1987, the U.S. Centers for Disease Control (CDC) had reported 33,550 cases of acquired immune deficiency syndrome (AIDS). Yet by that date, I estimate, physicians had actually diagnosed 42,670 cases.

The difference arises from significant delays in the reporting of AIDS cases to public health authorities. Some 9,120 additional persons had already been stricken with the disease, but they were not yet part of the CDC's official tally.

In this paper, I derive the empirical distribution of AIDS reporting delays and test its stationarity. From my results on reporting delays and the data on reported cases, I then estimate the actual incidence of the disease. While CDC reported about 4,500 new AIDS cases during the first calendar quarter of 1987, I find the incidence to be about 5,600.

Reporting delays are not the only reason why CDC's listings may fall short of the actual counts. Some cases of AIDS may never be reported. Doctors may be loath to inform public health authorities about certain patients. Also, the CDC's case definition of AIDS has not included all serious consequences of infection by the human immunodeficiency virus. These forms of underreporting, which can be viewed as reporting delays of infinite length, have been studied elsewhere (Chamberland et al. 1985; CDC 1986ab) and will not be my main focus here.

While researchers have attempted to adjust for reporting delays (Curran et al. 1985; Morgan and Curran 1986), the present paper appears to be the first formal analysis of the problem. Some of this paper's findings have been noted in an earlier report (Harris 1987).

2. THE PROBLEM

Once an individual is stricken with AIDS, the fact of his or her diagnosis is not instantly known to the CDC. Two more events need to take place. First, the attending physician or hospital reports the case to the local or state health department. Second, the health department transmits the information to the CDC.

The first step relies upon a surveillance system that is essentially passive. Although health departments in a few states actively review hospital and clinic records, most merely wait for the reports to come in. The second step entails periodic mailings by the health departments to the CDC. Starting in April 1986, the health departments switched from typewritten case reports to floppy diskettes, which were computer-encoded at the departments. By August 1986, most departments were mailing the diskettes.

For each reported case, the CDC lists both the date of diagnosis and the date of report. Up to March 1983, the date of report meant the time the health department received the information. Thereafter, the reporting date meant the time when the CDC received the data, that is, when both steps had been completed.

Among the 33,350 cases reported through March 1987, 336 were diagnosed during 1981. Of these, only 201 were actually reported during that year. Another 74 were reported in 1982, and 15 were not reported until 1986 or later.

For the 336 cases diagnosed in 1981, the records do not show the specific month of diagnosis. I shall therefore analyze the remaining 33,214 cases--

reported from January 1, 1982 through March 31, 1987-- for which the records do provide both the month and year of diagnosis.

Figure 1 shows the frequency distribution of these cases according to their date of diagnosis. The number of diagnosed cases falls off sharply after October 1986. But this does not mean that the incidence of the AIDS has been falling. Many cases diagnosed in late 1986 or early 1987 may not have been reported by March 1987.

In order to estimate the actual incidence of AIDS, we need to recover the unreported cases, and that requires estimating the distribution of reporting delays. In particular, we need to know the distribution of delays among all diagnosed cases, not just among the ones reported so far. This is because the delays observed for the reported cases constitute a truncated sample from the actual distribution. The question then becomes: What minimum assumptions are required to estimate the distribution of reporting delays?

3. STATISTICAL METHODS

3.1. Notation. Divide the time axis into intervals of equal length, called "periods," indexed by the positive integers. A new case of AIDS diagnosed during period t may not be reported until period t+u, where the non-negative integer u denotes the duration of the reporting delay. For short hand, I use the phrase "at t" to mean "during period t," while "by t" means "at any time up to the end of period t."

Let T be a known, nonrandom positive integer. Among all cases diagnosed by T, we observe only those reported by T. That is, for any case in which $t+u \leq T$, we observe the pair (t,u). But we do not observe even the number of

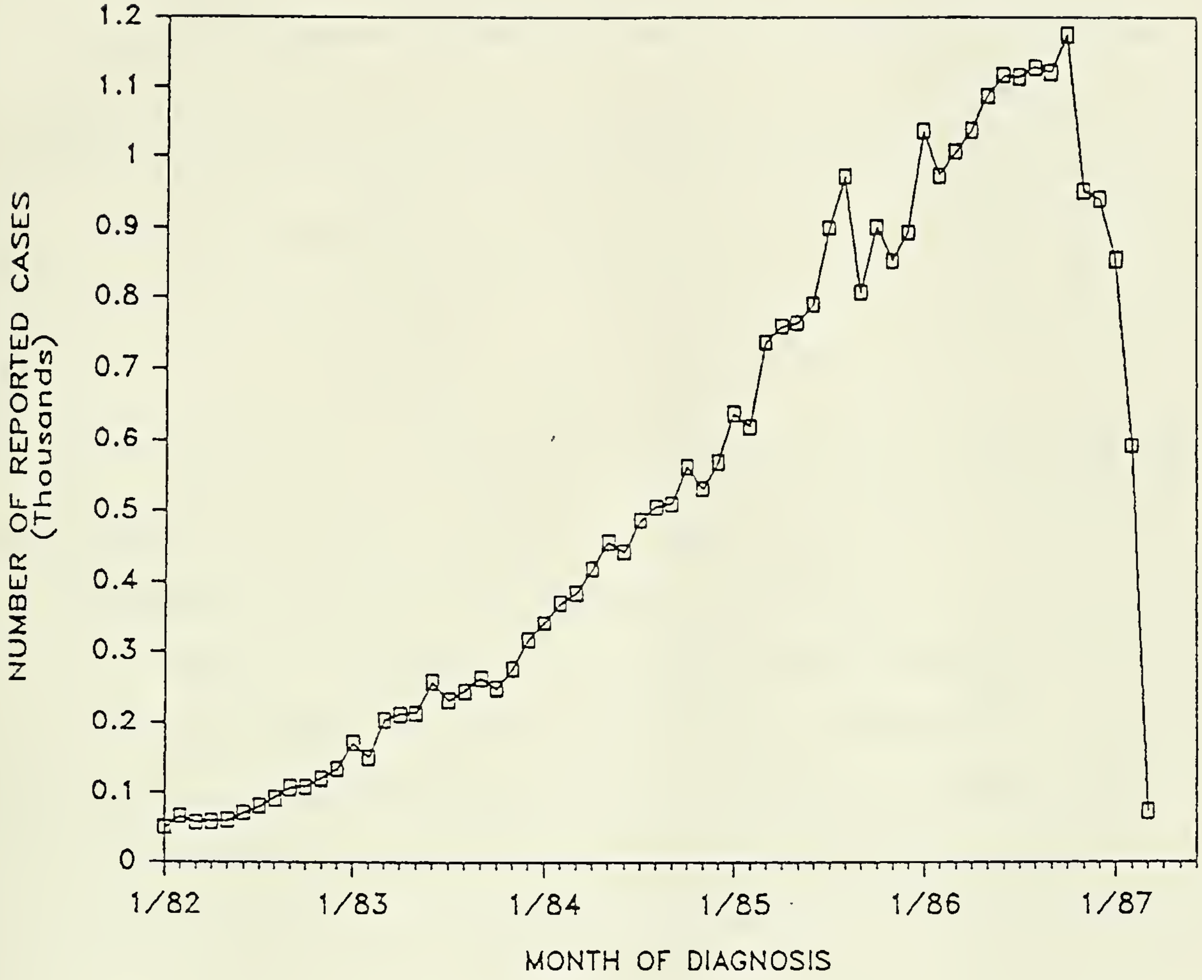

FIGURE 1. Distribution of 33,214 AIDS Cases Reported to the CDC through March 31, 1987 According to Month of Diagnosis. (Not shown in the figure are 336 cases diagnosed during 1981.)

cases for which $t \le T$ but $t+u > T$. From such truncated data, we wish to estimate the actual number of cases diagnosed at each $t \le T$.

Let $y_t(u)$ signify the number of cases diagnosed at t and reported at t+u. Define $y_t = \sum_{u=0}^{T-t} y_t(u)$ as the number of cases diagnosed at t and reported by T. Let $y(u) = \sum_{t=0}^{T-u} y_t(u)$ be the number reported with delay u, and define $\bar{y} = \sum_{t=1}^{T} y_t$ as the total number of reported cases. Let Y denote the set of all $y_t(u)$, and y the set of all y_t.

Let $\pi_t(u)$ denote the probability that a case of AIDS, diagnosed at t, will be reported with delay u. Define $\theta_t = \sum_{u=0}^{T-t} \pi_t(u)$ to be the probability that a case of AIDS, which has been diagnosed at t, will be reported by T. The symbol π will denote all $\pi_t(u)$ and, by implication, all θ_t.

3.2. Basic Model. Denote by x_t the number of cases diagnosed at t, whether or not they are reported by T. The counts x_t are unobserved.

Assumption I (Non-Parametric Model of AIDS Incidence): For all t, the counts x_t of diagnosed cases are independently Poisson distributed, with respective means λ_t. Each λ_t is termed the "incidence at t." Let λ represent the set of all λ_t.

If a case of AIDS is diagnosed at t and reported by T, then it will be reported at t+u with probability $\pi_t(u)/\theta_t$. Therefore, given the marginal sums y_t and the parameters $\pi_t(u)$ and θ_t, the joint distribution of the counts $y_t(u)$ is the product of independent multinomials:

$$g(Y|y,\pi) = \prod_{t=1}^{T} \frac{y_t!}{y_t(0)! \cdots y_t(T-t)!} \prod_{u=0}^{T-t} [\pi_t(u)/\theta_t]^{y_t(u)} \quad .$$

Moreover, given x_t and θ_t for each t, the counts y_t are independently binomially distributed as $b(y_t|x_t,\theta_t)$. By Assumption I, each x_t is Poisson distributed with mean λ_t. Hence, given λ_t and θ_t, each y_t is Poisson distributed with mean $\theta_t\lambda_t$. Given parameters $\pi_t(u)$ and λ_t, the joint distribution of the marginal sums y_t is

$$h(y|\pi,\lambda) = \prod_{t=1}^{T} [\theta_t\lambda_t]^{y_t} \mathit{exp}[-\theta_t\lambda_t]/y_t! \quad .$$

The likelihood of the parameters π and λ is thus the product of expressions $g(Y|y,\pi)$ and $h(y|\pi,\lambda)$. Up to an additive constant, the log-likelihood function is

$$L(\pi,\lambda) = \sum_{t=1}^{T}\sum_{u=0}^{T-t} y_t(u)\mathit{log}(\pi_t(u)/\theta_t) + \sum_{t=1}^{T} [y_t\mathit{log}(\theta_t\lambda_t) - \theta_t\lambda_t] \qquad (1).$$

Now consider the concentrated log-likelihood $L^*(\pi)$. That is, for arbitrarily fixed π, we choose $\lambda = \lambda^*(\pi)$ to maximize $L(\pi,\lambda)$ and then define $L^*(\pi) = L(\pi,\lambda^*(\pi))$. From (1), it is apparent that $\lambda_t{}^*(\pi) = y_t/\theta_t$. Up to an additive constant, the concentrated log-likelihood is therefore

$$L^*(\pi) = \sum_{t=1}^{T}\sum_{u=0}^{T-t} y_t(u)\mathit{log}(\pi_t(u)/\theta_t) \quad .$$

Assumption II (Stationarity of Reporting Delay Distribution): The probability distribution of reporting delays is independent of the date of diagnosis. That is, $\pi_t(u) = \pi(u)$ for all t,u.

It will prove convenient to define $\delta(v) = \sum_{u=v}^{\infty} \pi(u)$, the right-hand tail of the reporting delay distribution. If we permit π to be a defective distribution, then the tail $\delta(v)$ equals the probability of finite reporting delays of v or more periods plus the probability that a case may never be reported. The concentrated log-likelihood function is now

$$L^*(\pi) \quad = \quad \sum_{u=0}^{T-1} y(u)\, log\, \pi(u) \quad - \quad \sum_{t=1}^{T} y_t\, log\, \theta_t \tag{2},$$

which is homogeneous of degree zero in the arguments $\pi(0),\ldots\pi(T-1)$. That is, from Assumptions I and II alone, we can identify the probabilities $\pi(0),\ldots,\pi(T-1)$ only up to a proportionality constant. To solve this problem, we could impose a parametric form of the entire distribution $\pi(u)$. Instead, I shall assume that we have prior information on $\delta(T)$, the proportion of diagnosed cases that will go unreported for T or more periods.

Constrained maximization of $L^*(\pi)$ in (2) can be achieved by the following iterative procedure, analogous to the EM algorithm (Dempster, Laird and Rubin 1977). Consider estimates $\pi^{(N)}(u)$ obtained at the N^{th} stage in the iteration and define $\theta_t^{(N)} = \sum_{u=0}^{T-t} \pi^{(N)}(u)$ for each t. Given $\theta_t^{(N)}$, the maximum likelihood estimate of λ_t is $\lambda_t^{(N)} = y_t/\theta_t^{(N)}$. To obtain $\pi^{(N+1)}(u)$ at the $N+1^{st}$ stage, we first compute the quantities

$$p(u) \;=\; y(u)/\sum_{t=1}^{T-u} \lambda_t^{(N)} \qquad (3),$$

for each u, and then normalize the values of p(u) in (3) to sum to 1-δ(T):

$$\pi^{(N+1)}(u) \;=\; [1-\delta(T)]p(u) \,/\, \sum_{\mu=0}^{T-1} p(\mu) \qquad (4).$$

An appropriate starting value is $\pi^{(1)}(u) = (1-\delta(T))y(u)/\bar{y}$.

3.3. Variants of the Basic Model. Consider the following alternative to Assumption I.

Assumption IA (Parametric Model of AIDS Incidence): For all t, the counts x_t are independently Poisson distributed with respective means $f(t,\beta)/\alpha$, where α is a scale parameter and β is a vector of other parameters.

Conditional upon θ_t, α and β, the counts y_t are now independently Poisson distributed with means $\theta_t f(t,\beta)/\alpha$. Under Assumptions IA and II, the log-likelihood function becomes

$$L(\pi,\alpha,\beta) \;=\; \sum_{u=0}^{T-1} y(u)\,log\,[\pi(u)/\alpha] + \sum_{t=1}^{T} y_t\,log\,f(t,\beta) - \sum_{u=0}^{T-1} [\pi(u)/\alpha] \sum_{t=1}^{T-u} f(t,\beta) \qquad (5).$$

L(π,α,β) is homogeneous of degree zero in the arguments $\alpha,\pi(0),\ldots\pi(T-1)$. Hence, we still need an identifying restriction on either the scale α of the epidemic or the proportion δ(T) of cases reported with delays of T or more.

When we have prior information on δ(T), we can maximize L(π,α,β) by an interative procedure analogous to (3) and (4). Consider estimates $\pi^{(N)}(u)$ obtained at the Nth stage of the iteration. Given $\pi^{(N)}(u)$, we estimate $\alpha^{(N)}$ and $\beta^{(N)}$ by maximization of L($\pi^{(N)},\alpha,\beta$) with respect to α and β. We then obtain the N+1st values of $\pi(u)$ by computing $p(u) = y(u)\alpha^{(N)}/\sum_{t=0}^{T-u} f(t,\beta^{(N)})$ for each u, and then (given $\delta(T)$) applying the normalization (4).

Let T' be a known positive integer for which T' < T, and consider the following alternative to Assumption II.

Assumption IIA (Non-Stationarity of Reporting Delay Distribution): All cases of AIDS diagnosed by T' have an identical probability distribution $\pi(u)$ of reporting delays. Those cases diagnosed after T' also have an identical, but possibly different distribution $\pi'(u)$ of reporting delays.

Under Assumptions I and IIA, we obtain a concentrated log-likelihood function that is a generalization of (2). In that case, $L(\pi,\pi')$ is homogeneous of degree zero in the arguments $\pi(0),\ldots,\pi(T-1)$ and separately in the arguments $\pi'(0),\ldots,\pi'(T-T'-1)$. Hence, we need two restrictions to identify the parameters: one on $\delta(T)$, the right-hand tail of π; and another on $\delta'(T-T')$, the right-hand tail of π'.

Alternatively, under Assumptions IA and IIA, we obtain a log-likelihood function that is a generalization of (5). In that case, $L(\pi,\pi',\alpha,\beta)$ is homogeneous of degree zero in the combined arguments $\pi(0),\ldots,\pi(T-1)$, $\pi'(0),\ldots,\pi'(T-T'-1)$, and α. Only a single restriction (such as on α, $\delta(T)$ or $\delta'(T-T')$) is sufficient to identify the parameters.

Assumption IIA is only the simplest case of non-stationarity. In principle, we could partition the time axis into more than two intervals, with boundaries T', T", etc., and specify a different reporting delay distribution (π, π', π'', etc.) for each interval. If we continue to maintain Assumption I, then we will require a separate identifying restriction on each of the corresponding tail probabilities $\delta(T)$, $\delta'(T-T')$, $\delta''(T-T'')$, etc. In particular, in the computations reported below, I shall assume that $\delta(T) = 0$; and further that the tails of successive distributions are "matching," that is, $\delta'(T-T') = \delta(T-T')$, $\delta''(T-T'') = \delta'(T-T'')$, etc. In practice, this means that we first compute the estimates $\hat{\pi}(0),\ldots,\hat{\pi}(T-1)$ under the restriction that $\delta(T) = 0$. We then compute $\hat{\pi}'(0),\ldots,\hat{\pi}'(T-T')$ under the restriction $\delta'(T-T') = \hat{\pi}(T-T') + \cdots + \hat{\pi}(T-1)$; then estimate $\hat{\pi}''(0),\ldots,\hat{\pi}''(T-T'')$ under the restriction $\delta''(T-T'') = \hat{\pi}'(T-T'') + \cdots + \hat{\pi}'(T-T'+1) + \delta'(T-T')$; and so forth.

Under Assumptions IA and IIA, however, we still require only a single identifying restriction. In particular, in the results below, I shall assume that *min* $\{\delta(T),\delta'(T-T'),\delta''(T-T''),\ldots\} = 0$.

3.4. Remarks. In the basic model (Assumptions I and II), the concentrated log-likelihood $L^*(\pi)$ in (2) has T unknown parameters. Alternatively, under Assumptions IA and II, the full log-likelihood $L(\pi,\alpha,\beta)$ in (5) entails at least T+2 unknown parameters, and under Assumptions IA and IIA, the generalization $L(\pi,\pi',\alpha,\beta)$ entails at least 2T+T'+2 parameters. In each case, the maximum likelihood estimates of the parameters are consistent and asymtotically efficient as the number of reported cases $\bar{y}$ grows large, provided that the counts y_t grow faster than T.

Under Assumption I, we have posited what amounts to a null model of the AIDS epidemic. Hence, we can estimate the reporting delay parameters π (at least up to a proportionality factor) from the concentrated log-likelihood $L^*(\pi)$ in (2). Under Assumption IA, by contrast, the parametric model of AIDS incidence is informative about the reporting delay distribution. In that case, the log-likelihood $L(\pi,\alpha,\beta)$ in (5) cannot always be concentrated in a simple way, and the delay distribution π and the incidence model $f(t,\beta)/\alpha$ thus have to be estimated jointly.

Even when we have a specific model for AIDS incidence, the function $L^*(\pi)$ can still be interpreted as a partial likelihood in the sense of Cox (1975). Suppose that each count x_t has unspecified probability distribution $k(x_t|t,\Phi)$, which depends on the set Φ of parameters and which is not necessarily Poisson. The log-likelihood function can be written

$$L(\pi,\Phi) = L^*(\pi) + \sum_{t=1}^{T} \log \left[\sum_{x=0}^{\infty} b(y_t|x,\theta_t)\cdot k(x|t,\Phi)\right] ,$$

where $b(y_t|x,\theta_t)$ is the binomial distribution. Even if we were informed about the AIDS epidemic model $k(x_t|t,\Phi)$, the dimensionality of Φ could be so large that we might want to treat Φ essentially as a set of nuisance parameters and estimate π from $L^*(\pi)$ alone.

4. RESULTS

4.1. Non-Parametric Model of AIDS Incidence. Figure 2 compares the distribution of reported cases with the estimated incidence of AIDS. The curve denoted Reported Cases, reproduced from Figure 1, corresponds to the counts y_t. The curve denoted Estimated Incidence corresponds to the estimates of λ_t under Assumption I, where we posit no parametric model of the epidemic.

As Figure 2 shows, a significant fraction of AIDS cases had not been reported by March 31, 1987, even among those diagnosed in early 1986. For example, 1,037 AIDS cases were reported as diagnosed in January 1986. Yet

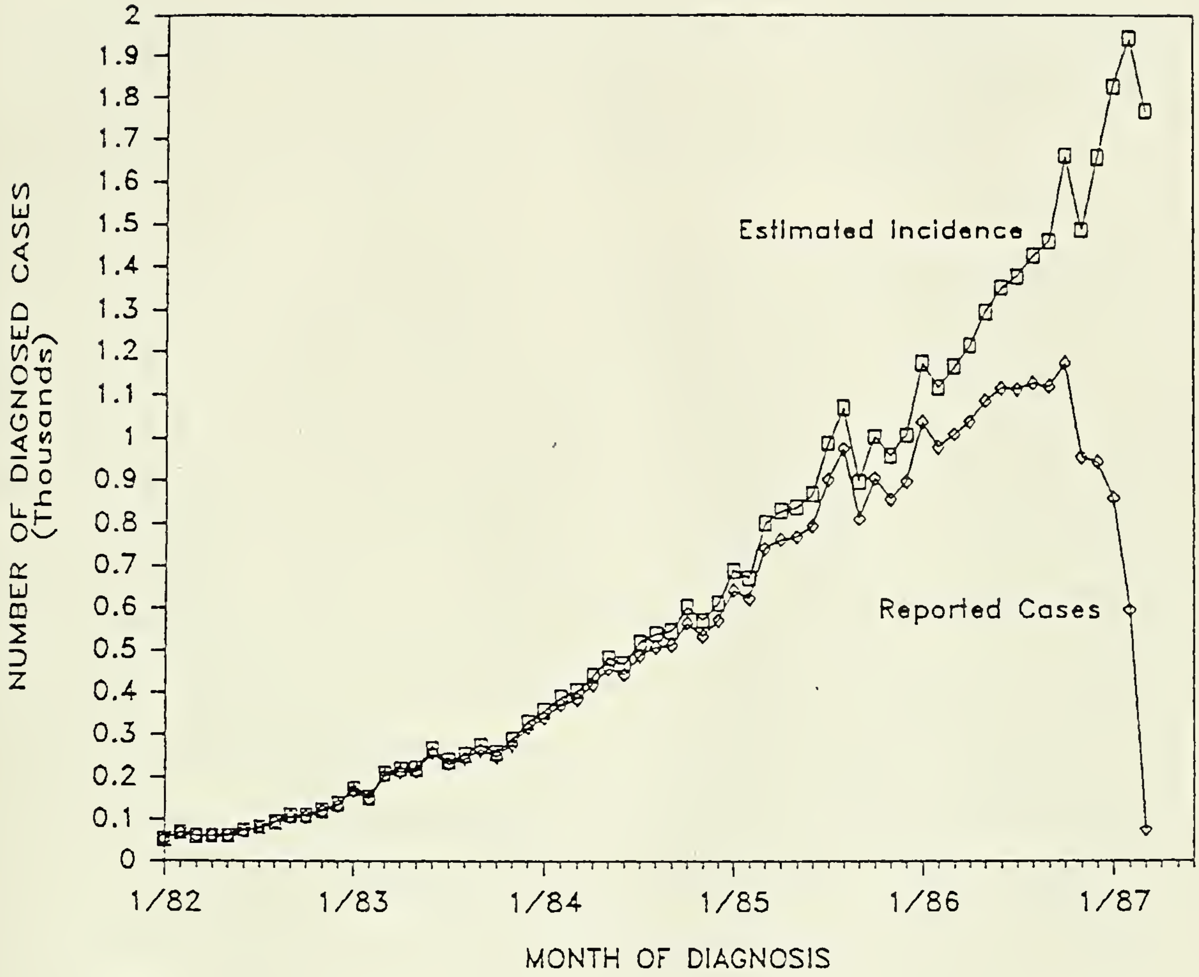

FIGURE 2. Estimated Incidence of AIDS Compared to Number of Diagnosed Cases Reported through March 31, 1987. (The curve of "Reported Cases" is duplicated from Figure 1. The "Estimated Incidence" curve was estimated under Assumptions I, IIA.)

Figure 2 gives an estimated incidence of 1,175 in that month, with the approximate lower and upper 95% confidence bounds (not given in the Figure) of 1,120 and 1,242. Similarly, 855 AIDS cases were reported as diagnosed in January 1987, yet the estimated incidence is 1,829 with 95% confidence bounds of 1,730 and 1,959. (I computed the confidence bounds by the bootstrap method of Efron (1979), where repeated draws were made from the empirical distribution of the counts $y_t(u)$.)

In the computations of Figure 2, I allowed for non-stationarity in the reporting delay distribution π (that is, Assumption IIA). In doing so, I partitioned the observation period (January 1982-March 1987) into four intervals: (1) January 1982-March 1983, when the encoded date of report was actually the date of receipt by the health department; (2) April 1983-March 1986, when the date of report was changed to the date received by CDC; (3) April-August 1986, when the health departments switched to computer-encoded diskettes; and (4) September 1986-March 1987, when the current reporting system was in place. Numbering successive months from January 1982 to March 1987 as $t = 1,\ldots,63$, we thus have $T' = 15$, $T'' = 51$ and $T''' = 56$. There are four potentially different reporting delay distributions, π, π', π'' and π''', identified by restrictions on $\delta(63)$, $\delta'(48)$, $\delta''(12)$ and $\delta'''(7)$. The "matching tails" restrictions, in particular, mean that $\delta(63) = 0$, $\delta'(48) = \delta(48)$, $\delta''(12) = \delta'(12)$ and $\delta'''(7) = \delta''(7)$.

Significant non-stationarity in the reporting delay distribution was found. The estimated proportions of cases reported within the same month were: $\hat{\pi}(0) = 0.287$; $\hat{\pi}'(0) = 0.059$; $\hat{\pi}''(0) = 0.088$; and $\hat{\pi}'''(0) = 0.041$. Estimates of the proportion of cases reported in the same or the subsequent month (that is, $\pi(0)+\pi(1)$) were respectively: 0.491, 0.350, 0.367, and 0.305. Allowing for $\pi \neq \pi'$ (but retaining the restrictions $\pi' = \pi'' = \pi'''$) added 48 parameters but increased the log-likelihood by 422.0 ($P < 0.0001$ by the chi-squared test). Allowing for $\pi \neq \pi'$ and $\pi' \neq \pi''$ (but retaining $\pi'' = \pi'''$) added 12 more parameters but increased the log-likelihood by 14.8 ($P < 0.005$). The completely unconstrained model added 7 more parameters with a further log-likelihood increase of 68.6 ($P < 0.0001$).

Figure 3 shows the estimated distribution $\hat{\pi}'$ for cases diagnosed during April 1982-March 1986 (interval 2). The distribution fits neither a Poisson nor a negative binomial. Up to about 18 months, $\hat{\pi}'$ approximately follows a Pareto rule (that is, the probability of reporting delays in excess of u months is approximately proportional to $u^{-0.85}$).

Figure 4 shows the cumulative number of diagnosed and reported AIDS cases by calendar quarter, based upon the results given in Figure 2. A total of 33,350 cases had been reported by March 31, 1987 (including those reported in 1981). Yet by that date, an estimated 42,670 had been diagnosed (95% confidence bounds 41,736 and 44,399). While the CDC had reported 4,523 new cases during the first quarter of 1987, I estimate that 5,542 were actually diagnosed (95% confidence bounds 5,180 and 6,044).

4.2. Parametric Models of the Epidemic. The estimated incidence curve in Figure 2 is not exponential. The doubling time of the epidemic, which appears to have been about 6 months in 1982, fell to about 9-10 months in 1984 and 15-16 months in 1986. While a subexponential epidemic may be plausible, the validity of the doubling-time estimates hinges on the "matching tails" restrictions on $\delta(T)$, $\delta'(T-T')$, $\delta''(T-T'')$ and $\delta'''(T-T''')$. Since

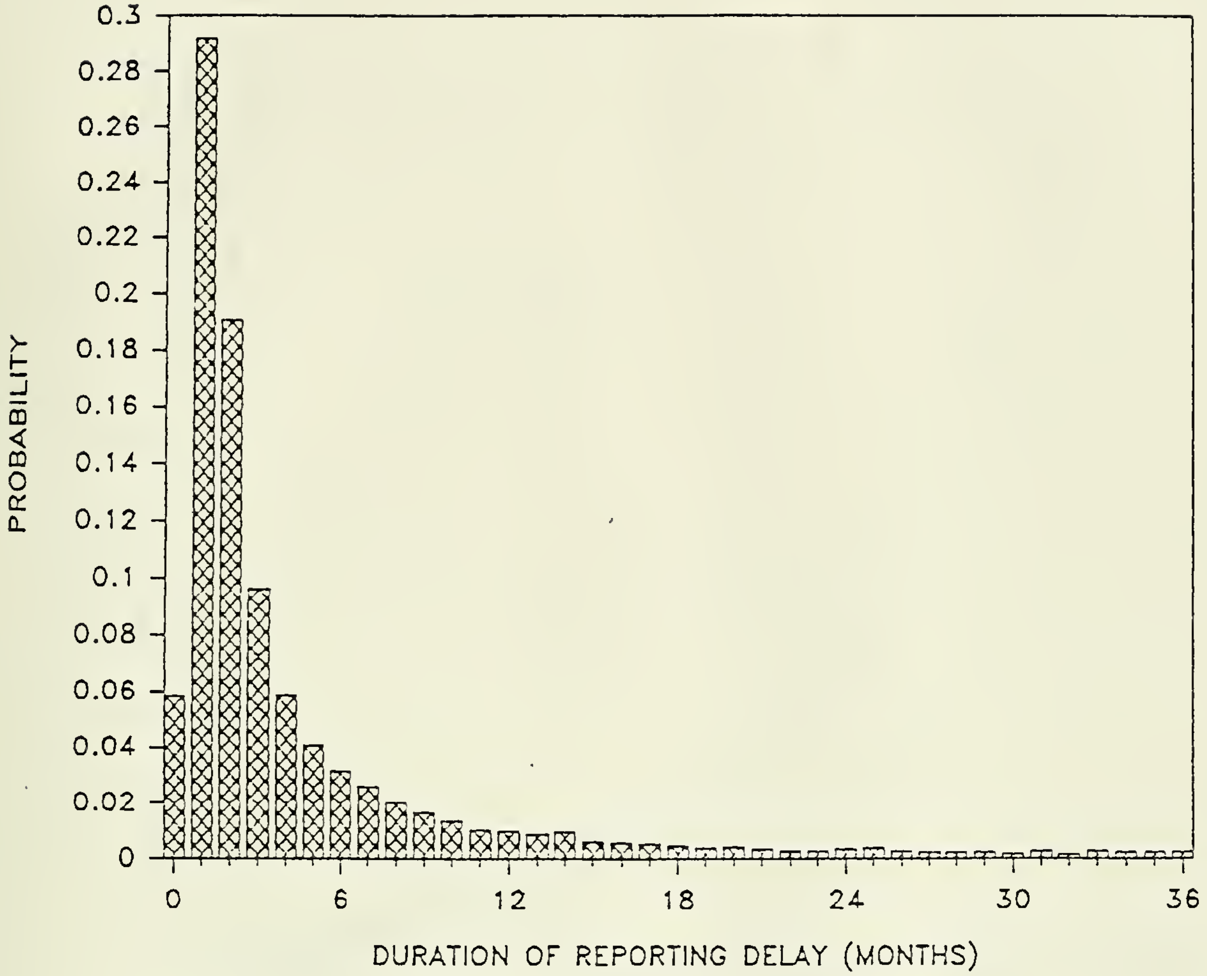

FIGURE 3. Estimated Probability Distribution of Reporting Delays for AIDS Cases Diagnosed During April 1983-March 1986. (The estimated probability of delay in excess of 36 months is 0.048.)

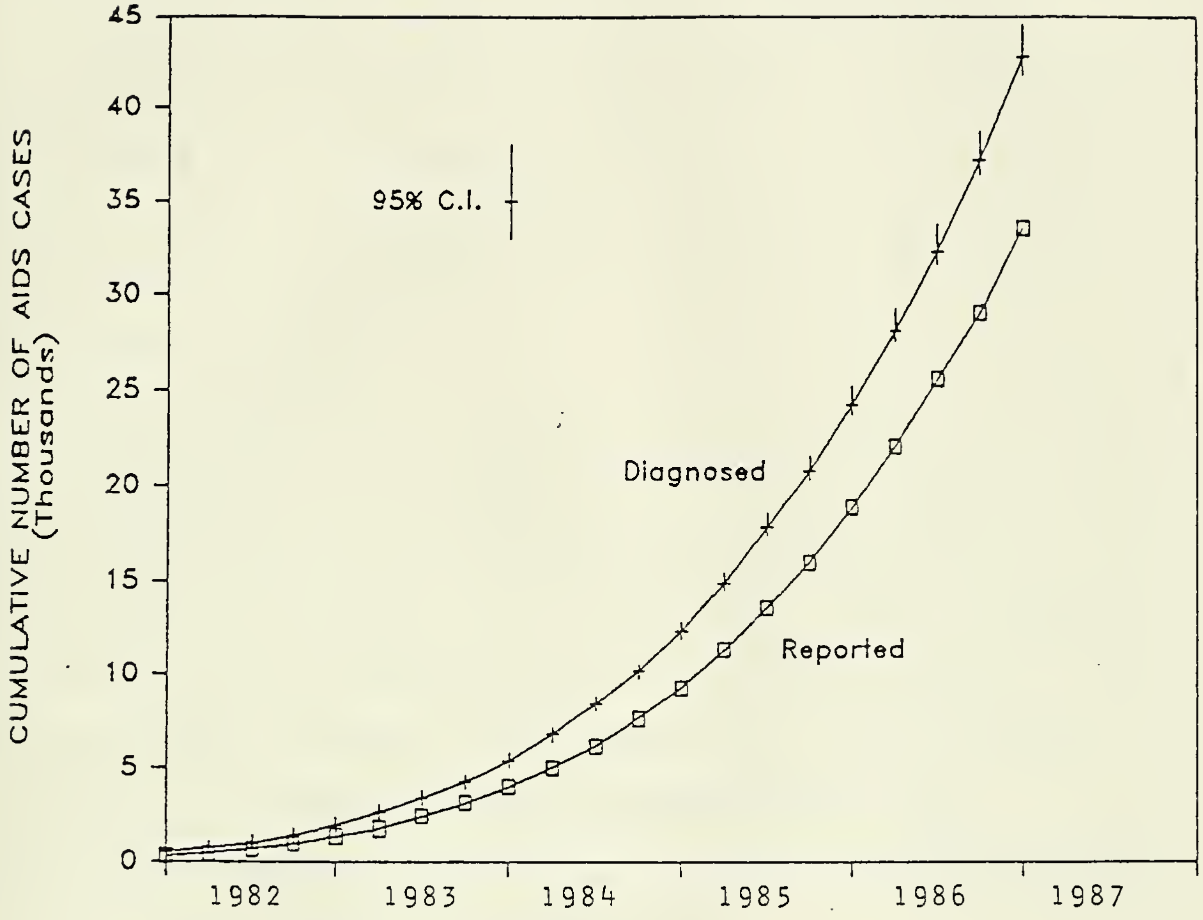

FIGURE 4. Estimated Cumulative Incidence of AIDS Compared to Cumulative Reported Cases, First Quarter of 1982 through First Quarter of 1987. (Approximate 95 percent confidence intervals are given for the cumulative incidence.)

these restrictions remain untested, the confidence intervals in Figure 4 understate the degree of uncertainty in the estimated size of the epidemic.

For instance, the "matching tail" assumption meant that $\delta'(48) = 0.032$, $\delta''(12) = 0.146$, and $\delta'''(7) = 0.233$. The last restriction means that among cases diagnosed during September 1986-March 1987, 76.7 percent would be the reported within 6 months of diagnosis. But if we arbitrarily changed $\delta'''(7)$ to 0.5, then the estimated incidence in the first quarter of 1987 would jump from 5,600 to 8,500 cases, while the total number of diagnosed cases would stand at 49,000.

As a means of validating the results of Figures 2, 3 and 4, I tested a flexible parametric model of the epidemic (Assumption IA). Specifically, I assumed that the counts x_t were independently Poisson distributed with respective means equal to $exp\ [\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3]$. Such a functional form was used not for any theoretical appeal, but simply as a means of smoothly approximating the path of the epidemic thus far. The resulting estimates of π, π', π'' and π''' (and the corresponding tails) were virtually identical to those in Section 4.1. The fitted incidence model was $exp\ [3.723 + 0.118t - 1.44\times10^{-3}t^2 + 8.28\times10^{-6}t^3]$.

Figure 5 compares the non-parametric model of AIDS incidence to a strictly exponential model. In contrast to earlier figures, the ordinate has a logarithmic scale. The individual data points reflect the non-parametric estimates, along with their approximate 95% confidence intervals. The fitted exponential model has an estimated slope of 0.0492, which means a doubling time of 14 months.

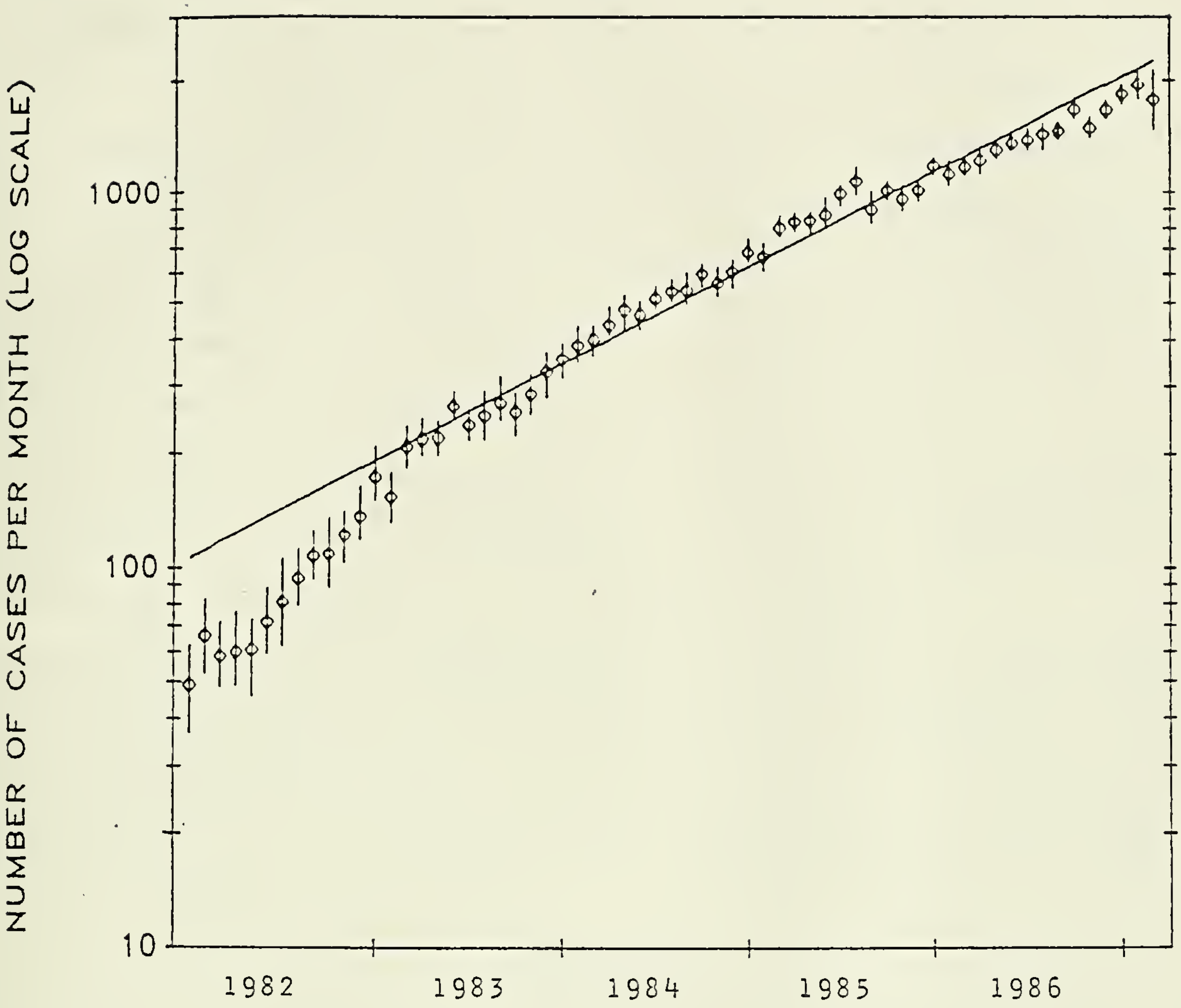

FIGURE 5. Estimated Incidence Under a Non-Parametric Model and an Exponential Model of the AIDS Epidemic. (The incidence data are plotted on a logarithmic scale. Approximate 95 percent confidence intervals are given for the non-parametric estimates.)

Under the exponential model, the estimated tails of the reporting delay distributions were: $\delta(63) = 0.331$; $\delta'(48) = 0$; $\delta''(12) = 0.215$; and $\delta'''(7) = 0.328$. An exponential model would thus require 33.1 percent underreporting of AIDS cases during January 1982-March 1983 (interval 1). After August 1986 (interval 4), the proportion of reporting delays in excess of 6 months would be 32.8 percent. For the four intervals, the estimated proportion of cases reported in the same or the subsequent month (that is, $\pi(0)+\pi(1)$) were respectively: 0.330, 0.362, 0.337, and 0.267.

In all models analyzed, the duration of reporting delays was found to be increasing over the course of the epidemic, especially after August 1986 (interval 4). If we convert $\hat{\pi}$, $\hat{\pi}'$, $\hat{\pi}''$ and $\hat{\pi}'''$ into continuous distibutions by linear interpolation, then in the non-parametric case, the estimated median reporting delays (in months) would be, respectively: 1.10 (95% confidence interval 0.92 to 1.49); 1.79 (95% conf. int. 1.72 to 1.91); 1.73 (95% conf. int. 1.64 to 1.84); and 2.33 (95%% conf. int. 2.20 to 2.53).

5. CONCLUSIONS

By March 31, 1987, the CDC had reported 79 percent of all AIDS cases diagnosed by that date. This divergence between reported and incident cases grew larger as the epidemic progressed (Figure 4). If we projected the smoothed incidence model *exp* $[\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3]$ out of sample, and if we assumed that the reporting delay distribution $\hat{\pi}'''$ remained unchanged, then by December 31, 1990, only 285,000 (74 percent) of a cumulative total of 383,000 cases would be reported. While projections based upon purely empirical curve-fitting are hazardous (Curran et al. 1985; Morgan and Curran

1986), there are compelling epidemiological reasons to expect the incidence of AIDS to continue to rise for at least the next five years (Brookmeyer and Gail 1986; May and Anderson 1987; Rees 1987; Harris 1987). Accordingly, the difference between reported and diagnosed cases is likely to grow larger.

I tentatively conclude that the distribution π has been shifting to the right. Since September 1986, about 70 percent of cases remain unreported two or months after diagnosis. This increase in the duration of reporting delays has occurred in spite of (or perhaps as a result of) the partial automation of the case surveillance system. I found the same trend toward increasing reporting delays in separate analyses of AIDS cases in homosexual and bisexual men and in intravenous drug users. The same conclusion applied to separate analyses of AIDS cases first presenting with pneumocystis carinii pneumonia, with Kaposi's sarcoma, and with other conditions. In fact, the changing mix of AIDS cases appears to go against the trend in reporting delays. Cases with Kaposi's sarcoma took significantly longer to report, with 78 percent now going unreported two or more months after diagnosis. Yet they comprise a declining fraction of newly diagnosed cases (from 28 percent of cases diagnosed in the first quarter of 1982 to 14 percent in the first quarter of 1987).

The CDC encoded the dates of diagnosis and case report by calendar month. Accordingly, I modeled the reporting delay phenomenon in discrete time. It is unlikely, however, that a continuous-time model would have yielded markedly different conclusions. In particular, in a continuous-time exponential epidemic with a stationary reporting delay distribution, the discrete proportion $\pi(0)$ would remain time-independent. In a sub-exponential epidemic with stationary reporting delays, $\pi(0)$ would fall. Yet we observed $\pi(0)$ increasing.

There are two untested explanations for the trend toward longer delays. First, doctors and hospitals are taking longer to report cases to the health departments. Second, the health departments are taking longer to send the reports to the CDC. While the latter explanation cannot be excluded from the data at hand, the former deserves our serious attention.

Perhaps increasing case loads have overburdened treating physicians. In the early years of the epidemic, doctors may have had more incentive to report a novel disease. Initially, infectious disease specialists made the diagnosis of AIDS. Now, a different type of physician may be the first contact with an AIDS patient. Successive changes in the official definition of AIDS may have created increasing confusion about which patients were to be reported.

In the non-parametric model, I found the doubling time of the epidemic to have increased from about 6 months in 1982 to 15-16 months in 1986. As Figure 5 suggests, most of the deceleration occurred in 1982. If there was substantial underreporting during that period, the epidemic may not have decelerated as much as it appears. Still, the conclusion that the epidemic is decelerating to some degree appears reasonably robust.

REFERENCES

BROOKMEYER, R. and GAIL, M.H. (1986), "Minimum Size of the Acquired Immunodeficiency Syndrome (AIDS) Epidemic in the United States," *Lancet*, ii, 1320-1322.

CENTERS FOR DISEASE CONTROL (1986a), "Revision of the Case Definition of Acquired Immunodeficiency Syndrome for National Reporting-- United States," Morbidity and Mortality Weekly Report, 34, 721-726, 731-732.

CENTERS FOR DISEASE CONTROL (1986b), "CDC Classification System for HIV Infections," Morbidity and Mortality Weekly Report, 35, 334-339.

CHAMBERLAND, M.E., ALLEN, J.R., and MONROE, J.M. (1985), "Acquired Immunodeficiency Syndrome, New York City: Evaluation of an Active Surveillance System," Journal of the American Medical Association, 254, 383-387.

COX, D.R. (1975), "Partial Likelihood," Biometrika, 62, 269-276.

CURRAN, J.W., MORGAN, W.M., HARDY, A.M., JAFFE, H.W., DARROW, W.W., and DOWDLE, W.R. (1985), "The Epidemiology of AIDS: Current Status and Future Prospects," Science, 229, 1352-1357.

DEMSTER, A.M., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, 39, 1-22.

EFRON, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Annals of Statistics, 7, 1-26.

HARRIS, J.E. (1987), "The AIDS Epidemic: Looking into the 1990s," Technology Review, in press.

MAY, R.M. and ANDERSON, R.M. (1987), "Transmission Dynamics of HIV Infection," Nature, 326, 137-142.

MORGAN, W.M. and CURRAN, J.W. (1986), "Acquired Immunodeficiency Syndrome: Current and Future Trends," Public Health Reports, 101, 459-465.

REES, M. (1987), "The Sombre View of AIDS," Nature, 326, 343-345.