ACTUARIAL RESEARCH CLEARING HOUSE
1991 VOL. 1

AIDS-PAC:
Computer Package for Analyzing AIDS Data

I.B. MacNeill[1], *L. Liu*[2], *V.K. Jandhyala*[3] and *Q.P. Duong*[4]

Abstract

AIDS-PAC is a computer package[5] that: (1) adjusts AIDS case counts to account for reporting delay, (2) estimates the size of the HIV infected population, and (3) provides short-term forecasts of the extent of the AIDS epidemic. It is designed to run on a microcomputer based on the 80286 microprocessor with an optional 80287 co-processor and VGA graphics capability. Hardcopy may be obtained using an HP Laserjet III or a QMS800-Lasergrafix printer. The methodology implemented by AIDS-PAC is discussed in this paper. Also, an analysis of the U.S. diagnosed AIDS time series is carried out using AIDS-PAC. The rate of increase in the U.S. HIV infected population is shown to have peaked in 1985/86 and arguments are presented indicating the peak of the AIDS epidemic has been reached now (1990/91) but that the decline will be gradual, resulting in sizeable numbers of AIDS cases through to the millenium. The size of the U.S. HIV infected population as of December 1989 is estimated to be $543,000 \pm 110,000$ (95%). Since the epidemic has peaked and the decline will be gradual, policy decisions for the medum-term regarding health care and insurance may be made on the basis of the present size of the epidemic.

---

[1] Department of Statistical and Actuarial Sciences and Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Canada, N6A 5B9.

[2] Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Canada, N6A 5B9.

[3] Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington, U.S. 99163.

[4] Bureau of Management Consulting, 364 Laurier Avenue West, Ottawa, Canada, K1A 0S5.

[5] A copy of AIDS-PAC on diskette may be obtained by mailing a cheque for US$100.00 to Statistical and Actuarial Services, Western Science Centre, The University of Western Ontario London, Canada, N6A 5B9.

1. Introduction

W.H. McNeill (1976) in his monograph "Plagues and Peoples" speculated that new diseases would emerge from time-to-time and attack human populations. At the time this thought was being expressed, the human immunodeficiency virus (HIV) had begun to infect certain risk groups in Western society but the syndrome, Acquired Immuno-Deficiency Syndrome (AIDS), was not recognized until 1978 when it was discovered among the homosexual population of San Francisco.

The early development of the AIDS epidemic was characterized by exponential growth. Extrapolation of this exponential model lead to apocalyptic forecasts of a calamity that would rival the Black Death, a catastrophe which decimated the population of Europe during the 1300's.

The growth in the incidence of new diagnosed cases of AIDS departed from the exponential mode in early 1984. Duong and MacNeill (1987) identified this departure for the Canadian epidemic, and Jandhyala and MacNeill (1989) have analysed the U.S. data and have estimated early 1984 as the date at which the parameters of the system began to change. Having determined that the exponential hypothesis was no longer tenable, Duong and MacNeill (1987) used information criteria to select, from among a range of growth models, the logistic model as that which best fitted the AIDS series. This model resembles exponential growth in its early stages, but later on it exhibits a decline in its growth rate and ultimately ceases to grow. It is not suitable as a model when the epidemic begins to show marked decline. The model may be fitted to the data and then may be used to provide near-term forecasts which possess good statistical properties.

Although the logistic has epidemiological validity as a model for the growth of epidemics, it should be though of in this context as an empirical model. Non-empirical models are built upon relationships and parameters that have biological and/or physical meaning and they attempt to incorporate information about micro-behaviour for the risk groups. They have been proposed for forecasting the incidence of AIDS; see, for example, Kanouse (1988). However, these models fail for several reasons. First, behaviour among the main risk groups, *viz*, homosexual males and intravenous drug users, is poorly understood and the sizes of these risk groups, although the subject of much speculation, are not known. Second, realistic non-empirical models are characterized by such a large number of parameters that the available data base cannot support an appropriate inferential exercise. For these reasons, the logistic is used in the sequel for near-term forecasts and for smoothing the data provided by such public health agencies as the Centers for Disease Control (CDC) in the U.S.

2. ADJUSTMENTS FOR REPORTING DELAY

Before fitting a model to the time series represented by the number of AIDS cases reported for prior months it is necessary to adjust this series to account for reporting delay. The time required for a diagnosed case to be reported to a central data collection agency (in the U.S. this is the CDC) is highly variable. A small number of cases will be reported within several months, most will be reported within 12 months but some will not be reported for several years. Hence the monthly time series published quarterly by the CDC will not represent the numbers of cases of AIDS diagnosed per month that ultimately will be reported. Therefore it is necessary to adjust the reported series before fitting a model to present and past data, and before extrapolating to the future. Figure 1 illustrates the reporting-delay phenomenon. It contains graphs of the semiannual reports issued by the CDC beginning in January 1987 and ending in July 1990. These curves represent the monthly diagnosed AIDS cases reported to the CDC as of report time. The more recent counts for months in the distant past appear to be close to complete since the graphs lie close together, but serious reporting delay is evident for the more recent months.

MacNeill *et al.* (1990) have developed a methodology for adjusting for reporting delay. Other approaches include those of Morgan and Curran (1986), Healy and Tillett (1988) and Cox and Medley (1989).

We let $D(l,t)$ represent the number of new diagnosed AIDS cases for time period t as reported at time l. In Figure 2, taken from MacNeill *et al.* (1990), $D(l,l-n)$ is the number reported now (l) for n months ago, and $D(l+m,l-n)$ is the number that will be reported in m months time for the same month. The reporting-delay adjustment is

$$f(l,n,m) = \frac{D(l+m,l-n)}{D(l,l-n)} \quad .$$

One seeks $f(l,n,\infty)$ since this represents the reporting-delay adjustment that should be applied to $D(l,l-m)$ to account for all the diagnosed cases that will eventually be reported for the $(l-n)^{\text{th}}$ month. However, one will wish to estimate $f(l,n,\infty)$ now and not wait until m becomes this large.

The function $f(l,n,m)$ satisfies the following equations

$$f(l,n,m_1+m_2) = f(l,n,m_1)f(l+m_1,n+m_1,m_2) \quad ,$$

and

$$f(l,n,m) = \prod_{j=0}^{m-1} f(l+j,n+j,1) \quad .$$

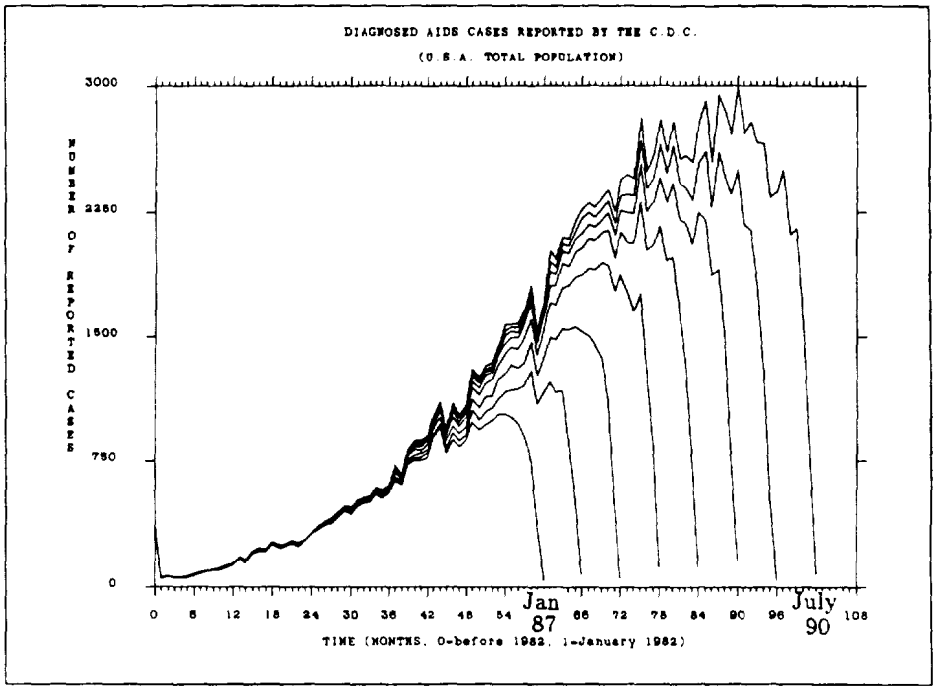

Figure 1. Diagnosed AIDS cases reported by the CDC (total U.S. pop.).

Hence, if one is given the initial condition $f(l, n, 1)$ for all n and l, then one can obtain $f(l, n, m)$ for all l, n, m. This would solve the non-stationary reporting-delay problem since

$$D(\infty, l-n) \simeq f(l, n, \infty) D(l, l-n) \quad .$$

However, estimation of this initial condition places heavy demands on the available data.

If $f(l_1, n, m) = f(l_2, n, m)$ for all reporting times l_1, l_2, then the functional equation reverts to the stationary case represented by (2); namely,

$$f(n, m_1 + m_2) = f(n, m_1) f(n + m_1, m_2) \quad ,$$

which implies that

$$f(n, m) = \prod_{j=0}^{m-1} f(n + j, 1) \quad ,$$

and also that

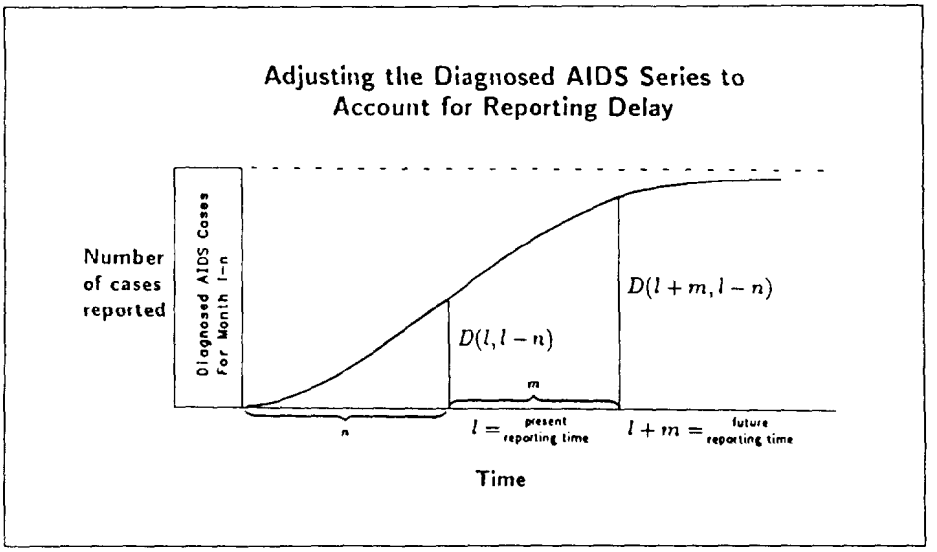

Figure 2. Late-reporting for diagnosed AIDS cases (MacNeill *et al.* 1990).

$$f(n,6m) = \prod_{j=0}^{m-1} f(n+6j,6) \quad . \tag{1}$$

This initial condition is simpler and, provided that reporting efficiency has remained relatively constant, it may be estimated from the data given in Figure 1. An estimation procedure is discussed by MacNeill *et al.* (1990). This procedure followed by the application of (1) yields $f(n,\infty)$, $n = 13,14,\ldots,60$, as given in Figure 3.

The data are considered inadequate to carry out an appropriate estimation of $f(n,\infty)$, $n = 1,2,\ldots,12$. Hence we apply the adjustment procedure to the data in Figure 1 with the last twelve observations truncated; the truncated data appear in Figure 4.

Adjustment, using $f(n,\infty)$, of the AIDS data for January 1982 to June 1989 as reported July 1990 and graphed in Figure 4 yields the various adjusted series graphed in Figure 5.

A logistic curve was fitted to this adjusted data and extrapolated forward to June 1990; this curve is superimposed on the adjusted series in Figure 6.

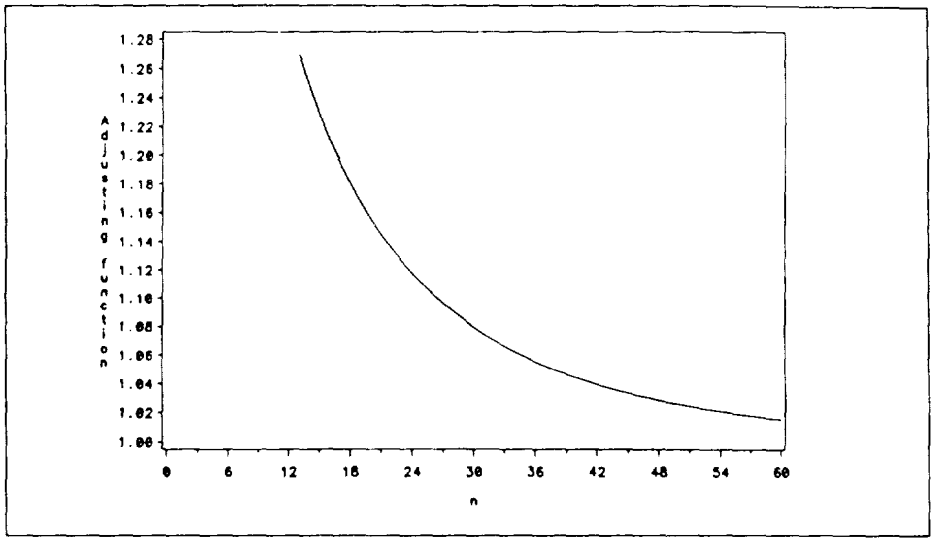

Figure 3. Reporting-delay function for AIDS, $f(n, \infty)$, $n = 13, 14, \ldots, 60$.

3. Incubation Time

One of the unusual features of the HIV infection is its long incubation time. In the earlier part of the 1980's estimates placed the mean incubation time at approximately four years. However, as the epidemic has progressed, and as more information has become available, this mean-time estimate was lengthened by Panjer (1987) to 6.5 years based on a prospective study by Brodt *et al.* (1986), to 8 years based on the same study by Cowell and Hoskins (1987) and more recently to 10 years by Kalbfleisch and Lawless (1988). In a recent study Longini *et al* (1989) modelled the phases of the Walter Reed staging method by independent negative exponentials which results in a distribution that is approximately Gamma with parameters $\alpha = 2$ and $\beta = 5$; that is

$$P(t) \simeq \frac{te^{\frac{t}{5}}}{25} \quad , \qquad t > 0 \tag{2}$$

where t is measured in years. This distribution has a mean of 10 years but a modal value at $t = 5$ years. This is the incubation time distribution used in the sequel. Figure 7 contains the graph of the incubation time distribution given by (2).

An early misconception concerned the proportion of the infected population that would ultimately become AIDS victims. Because of an early lack

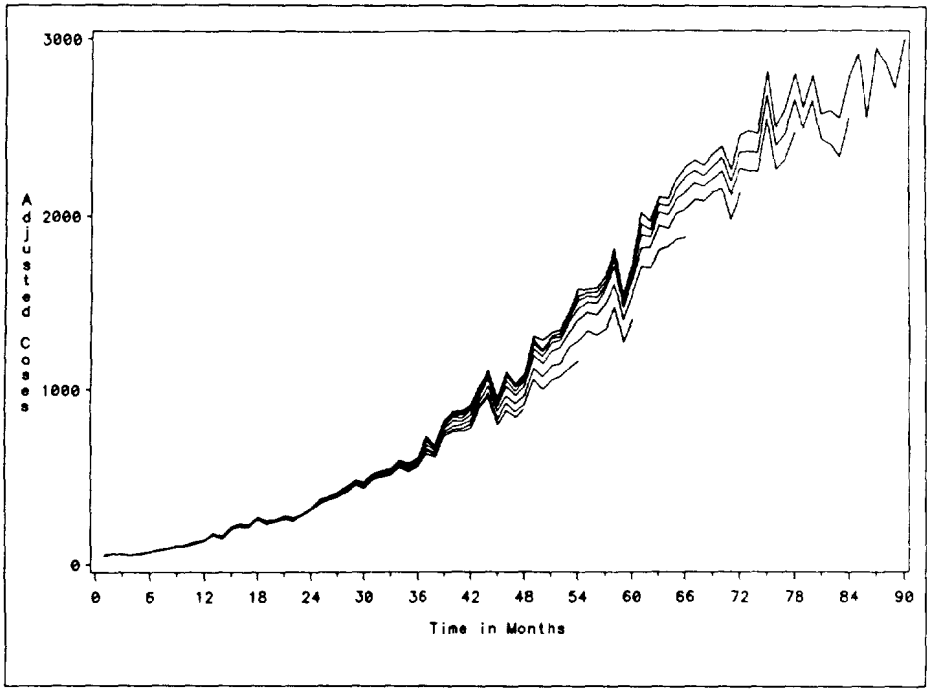

Figure 4. Truncated diagnosed cases (see Figure 1).

of understanding of the incubation time distribution, it was variously speculated that this proportion ranged form 0.1 to 0.5. Better understanding of the incubation time distribution now suggests that all of those infected will become AIDS cases provided they do not first fall victim to other risks.

4. SIZE OF THE INFECTED POPULATION

The size of the HIV$^+$ population is an aspect of the AIDS epidemic that is subject to considerable speculation. Much of this speculation has been groundless and some has bordered on the hysterical. Estimates of the size of the worldwide seropositive population as high as 100,000,000 have been issued by the World Health Organization; these estimates were later reduced by a factor of 10. Reports from the U.S. Surgeon General based on the Coolfont report (1986) contained estimates that in 1985 between one and 1.5 million citizens in the U.S. would test seropositive; these estimates were based on small samples of the general population. Difficulties in obtaining reliable sample survey data in this area make these estimates highly

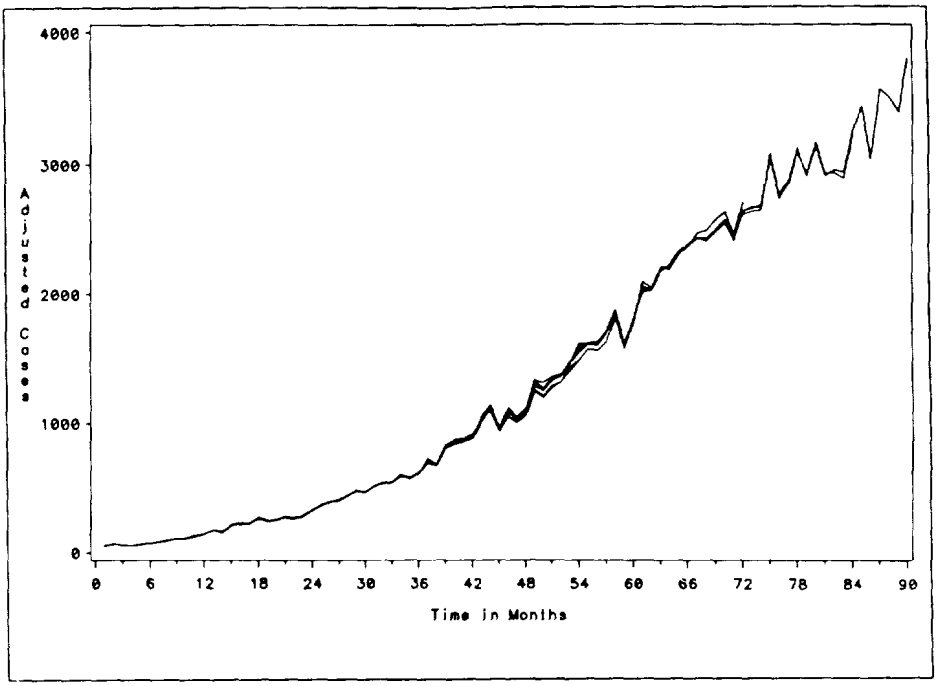

Figure 5. AIDS series adjusted for reporting-delay.

speculative. More recent reports(1989) from the Surgeon General's office estimated the HIV^+ population at approximately one million. This was a substantial reduction, especially since this estimate is for a date 3 years later than the original.

Since each AIDS case is a consequence of an earlier HIV infection, it is possible to use the incubation time distribution to relate the size of the infected population to the number of diagnosed AIDS cases. This is analogous to estimating the size of the underwater portion of an iceberg by observing the volume of ice above the water; one divides the volume of the tip by the known ratio of ice above to that below the water level. Thus, if one uses the following notation,

$D(t)$ = rate of diagnosis of AIDS at time t,

$N(t)$ = rate of infection recruitment at time t,

$P(t)$ = probability density of the incubation time distribution at time t,

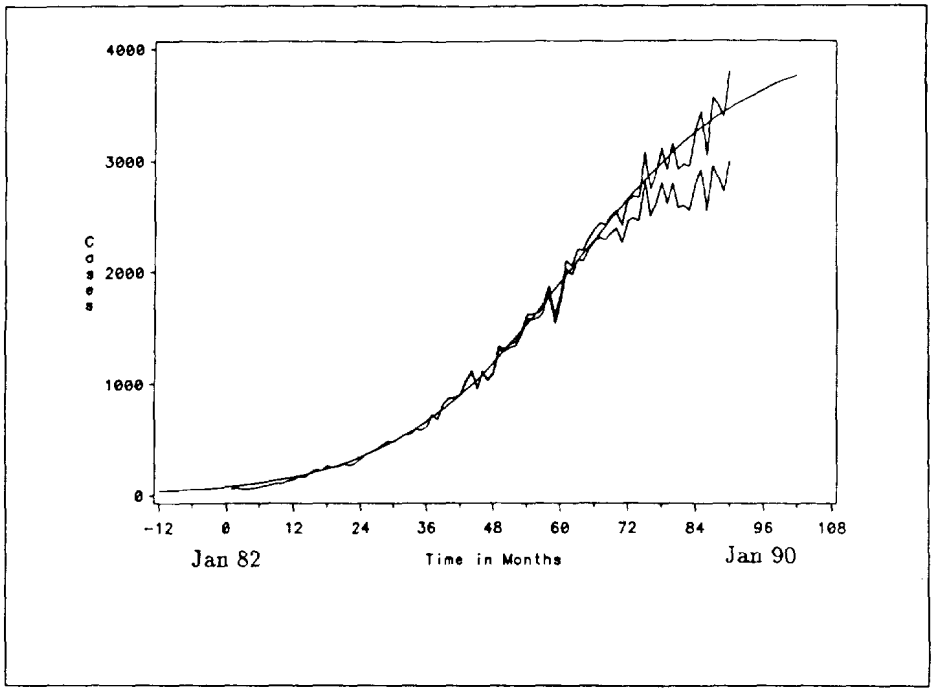

Figure 6. Logistic fit to adjusted AIDS series.

then one can show the functional relationship between the variables under consideration to be contained in the following integral equation:

$$D(t) = \int_0^t P(t-s)N(s)\,ds. \tag{3}$$

This equation forms the basis for so-called back-casting. See Brookmeyer and Gail (1988) for an alternative approach to that taken here.

We substitute the incubation time distribution given by (2) into (3) to obtain

$$D(t) = \frac{1}{25}\int_0^t (t-s)e^{-\frac{t-s}{5}}N(s)\,ds$$

This is equivalent to

$$D(t)e^{\frac{t}{5}} = \frac{1}{25}\int_0^t (t-s)e^{\frac{s}{5}}N(s)\,ds$$

which upon differentiation becomes

$$\frac{d}{dt}\left\{D(t)e^{\frac{t}{5}}\right\} = \frac{1}{25}\int_0^t e^{\frac{s}{5}}N(s)\,ds$$

and

$$\frac{d^2}{dt}\left\{D(t)e^{\frac{t}{5}}\right\} = \frac{1}{25}e^{\frac{t}{5}}N(s) \quad .$$

Therefore

$$N(t) = D(t) + 10D'(t) + 25D''(t) \quad . \tag{4}$$

It should be noted here that if $D(t)$ is not smooth then its first and second derivatives will inject a great deal of irregularity into $N(t)$.

If the number of pre-AIDS infections at time t is denoted by $NP(t)$, then

$$NP(t) = N(t) - D(t) = 10D'(t) + 25D''(t) \quad .$$

If $D(t)$ is modelled by the logistic function, a function of 3 parameters $M, D(0), k$, then

$$D(t) = \frac{D(0)M}{D(0) + \{M - D(0)\}e^{-Mkt}} \tag{5}$$

Then

$$N(t) = D(t) + D^2(t)h(t)\,(10 - 25Mk) + 50D^3(t)h^2(t)$$

where

$$h(t) = k\left(\frac{M}{D(0)} - 1\right)e^{-Mkt}$$

If the total number of pre-AIDS infections for the interval $[-s, t]$ is denoted by $T[s, t]$ then

$$\begin{aligned} T(s,t) &= \int_{-s}^{t} NP(t)\,dt \\ &= 10\{D(t) - D(-s)\} + 25\{D'(t) - D'(-s)\} \end{aligned}$$

We notice that by integrating we reduce the order of the derivatives by one and hence obtain a smoother function.

However, if we ignore the intermediate step of computing equation (4) we may be unaware of important details regarding $N(t)$. For example, inappropriate modelling may result in negative estimates for $N(t)$ over certain time periods; this of course is not physically possible and one should seek

the cause of this anomaly rather than ignore it by considering only the integrated version which conceals the problem. Also, details regarding peaking of $N(t)$ will not be obvious from a casual observance of $T(s,t)$.

If the function $D(t)$ is characterized by an m-vector of parameters θ ($m = 3$ for the logistic), which is estimated by $\hat{\theta}$, then $T(s,t)$ is characterized by these same parameters. If we let

$$\Delta T(s,t \mid \theta) = \left(\frac{\partial T(s,t \mid \theta)}{\partial \theta_1}, \ldots, \frac{\partial T(s,t \mid \theta)}{\partial \theta_m} \right)'$$

and let Σ denote the variance/covariance matrix of $\hat{\theta}$, then

$$\text{Var}\,\{T(s,t)\} \simeq \Delta T'(s,t \mid \hat{\theta})' \Sigma \Delta T(s,t \mid \hat{\theta})$$

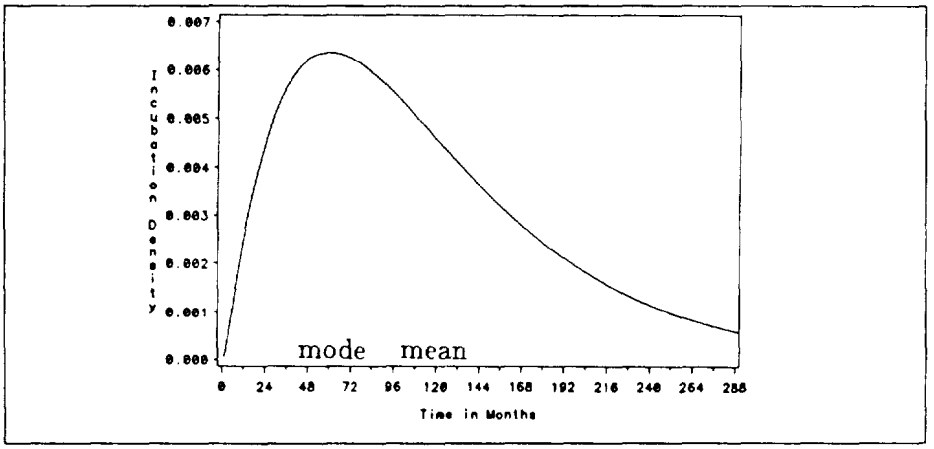

Figure 7. Incubation time distribution with a mean of 10 years.

Figure 8 contains graphs of (4) and (5) superimposed on the adjusted diagnosed case series. The striking feature of this graph is the rapid rise in the number of infections until 1985/86 and the equally precipitous decline during subsequent years.

Figure 9 is a graph of $T(s,t)$ for the period from January 1979 to the present. The total size of the pre-AIDS population at the end of the 1980's was $543,000 \pm 110,000$ (.95). This is a large number but roughly one third that given in the 1986 Coolfont report.

It can be noted from Figure 8 that the peak of the HIV infections occurred in 1985/86. Furthermore, it can be observed that the modal value

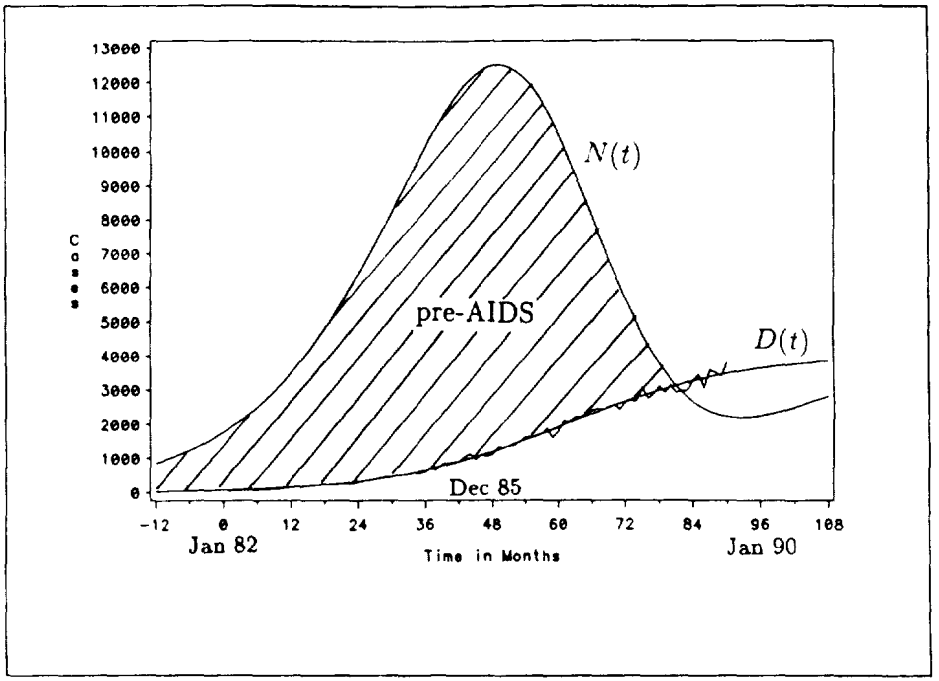

Figure 8. Rate of HIV infection recruitment, $N(t)$, and rate of diagnosis of AIDS, $D(t)$.

of the incubation time distribution is at 5 years. A consequence of these points is that the peak of the diagnosed AIDS cases must occur in 1990/91. Although the peak of the AIDS epidemic is now, due to the long incubation time, the decline in the epidemic will be slow and there will still be significant numbers of AIDS cases diagnosed at the end of the millenium due to the substantial numbers of HIV infections contracted during the middle of the 1980's.

Because the peak of the epidemic is now, and because the decline will be quite slow, it is possible to make policy decisions for the medum-term regarding health care and insurance on the basis of the present level of the epidemic.

5. AIDS Forecasts

What does the future hold regarding the number of AIDS cases in the U.S.? Several scenarios were explored by MacNeill *et al* (1990) by

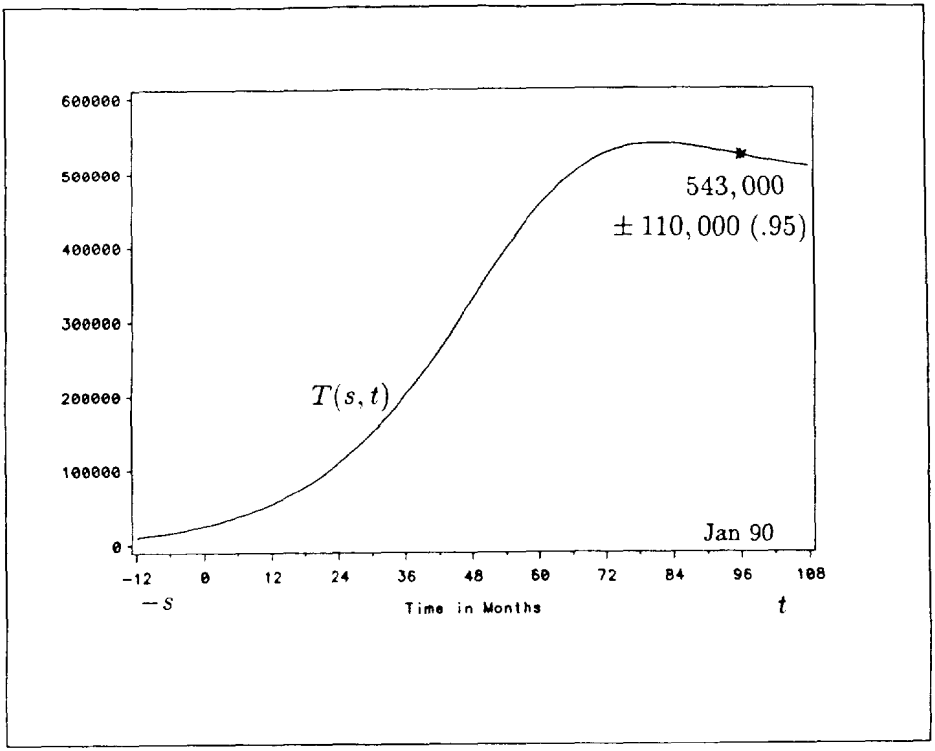

Figure 9. Total size of the pre-AIDS population–January 1979 to the present.

extrapolation of $N(t)$, the HIV infections function; Figure 10 contains the three different extrapolations they considered.

Equation (3) was used to forecast the AIDS series, $D(t)$. Forecasts of AIDS incidence as found by MacNeill *et al* (1990) are given in Figure 11. Extrapolation A is compatible with continued logistic growth in $D(t)$. Extrapolations B and C are more in keeping with the dynamics of the $N(t)$ series and suggest a decline in the number of diagnosed AIDS cases beginning in 1990/91.

6. DISCUSSION

Several reasons may be put forth to explain the post-1986 precipitous decline in the U.S. HIV infection rate. First, the effect of education among the high risk groups has resulted in less risky behaviour on their part.

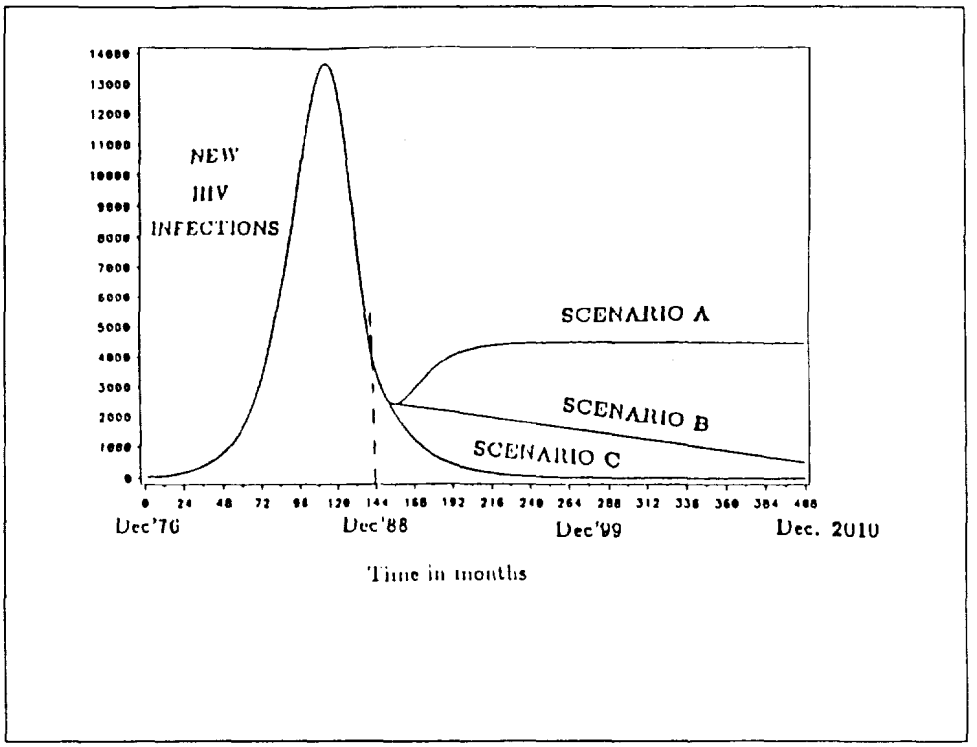

Figure 10. Scenarios for future HIV infection rates in the U.S.A.

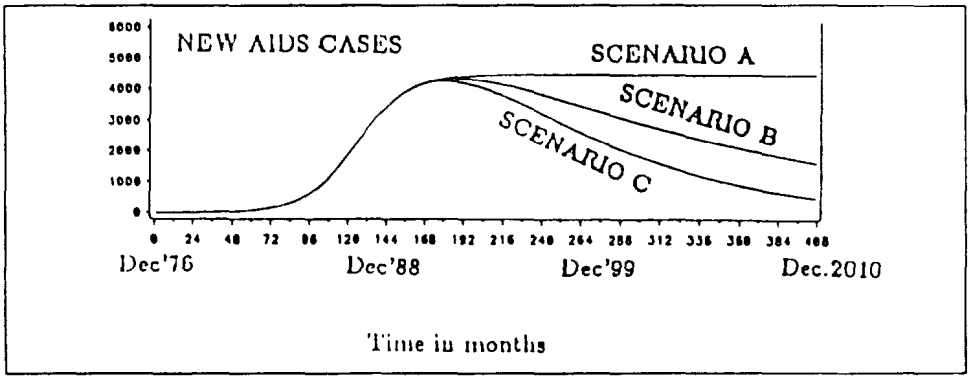

Figure 10. New AIDS cases with three scenarios for the future (also see Figure 10).

However, because of the lengthy incubation time, education probably had a minimal impact on the shape of the $N(t)$ function in the pre-1985/86 period. The second, and most plausible explanation for the rapid growth and decline is a saturation effect among those who put themselves most at risk during the period immediately prior to the time when general awareness of the disease emerged. That is, exponential-type growth occurred at first, but after the infection spread to a large part of this group there was little room left for continued growth.

The public health education programs and the broad dissemination of information about AIDS that occurred during the last decade are likely to have their greatest impact on the HIV infection rate during the present decade. If these programs have been effective, then scenario C will be more likely; if the programs have not been effective, then scenario A will be more likely. Evidence is given by Johnson (1988), McKusick *et al.* (1985), Martin (1987), Winkelstein *et al.* (1987) and others to the effect that among the homosexual population in the U.S., education has resulted in substantial behavioural change. Hence, the weight of that evidence points to scenario C as the most likely of the scenarios discussed. The effect of education upon intra-venous drug users is more problematical.

We have assumed that 100% of the HIV infections will become AIDS cases. If this is not true but rather only a proportion p $(0 < p < 1)$ of those infected will become AIDS cases, then the levels of infection derived in this paper should be multiplied by p^{-1} to obtain the actual levels of infection. If one is interested only in the number of pre-AIDS infections then the above estimates remain valid.

The above discussion and analysis have presumed that all AIDS cases will be reported. If the proportion of incidence that is reported is f $(0 < f < 1)$ then the levels of AIDS and HIV infections given above should be multiplied by f^{-1} to obtain true levels. If one is interested only in pre-AIDS infections that become reported AIDS cases then the above analysis remains valid.

Most attention is focused on the present and future course of the AIDS epidemic. However extrapolation backwards in time of the curves in Figure 8 suggests there had to be a significant level of infection back as far as the late 60's for the epidemic to have taken the course it has. Perhaps significant life-style changes had an impact. There may be sociological reasons as to why the epidemic erupted when it did. This suggests that gathering early data on opportunistic infections related to AIDS might point to possible ways to control the spread of the epidemic.

7. Conclusion

If medical science is unsuccessful in discovering cures, vaccines or effective drugs/therapies, then mankind will have to manage with AIDS as it has managed with other epidemics in the past; namely, by developing protective natural immunity. The childhood diseases, such as measles and chicken pox, brought by the Spanish to the New World in the 16th century, had devastating effects upon the native populations of the Americas. However, today, these diseases have no more impact upon their descendants than they do upon the descendants of the Spanish. Several generations have been required to build immunity defences in past; W.H. McNeill (1976) has estimated six generations. Meanwhile, public health education is the main tool for near-term management of the AIDS epidemic.

Acknowledgements

The authors wish to thank D. Naylor for valuable discussion regarding the solution of integral equations. They also wish to thank L.M. Kwarciak and M. Lavdas for valuable support in developing the AIDS-PAC package. This research was supported by a contract from the Ontario Ministry of Industry, Trade and Technology and by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Brodt, H.R., E.B. Helm, A. Joetten, L. Bergmann, A. Kluver and W. Stille (1986), "Spontanverlauf de LAV HTLV-III - Infektion; Verlaufsbeobachtungen bei Personen aus AIDS-Risikogruppen". *Deutsche Medizinische Wochenscrift*, Stuttgart, **III**, 1175–1180.

Brookmeyer, R. and M.H. Gail (1988), "A method for obtaining short term projections and lower bounds on the size of the AIDS epidemic". *Journal of the American Statistical Association* **83**, 301–308.

Cowell, M.J. and W.H. Hoskins (1987), "AIDS, HIV Mortality and Life Insurance, Parts 1 and 2". Society of Actuaries, distributed as a Special Report.

Cox, D.R. and G.F. Medley (1989), "A process of events with notification delay and the forecasting of AIDS". *Philosophical Transactions of the Royal Society of London* **325**, 135–145.

Duong, Q.P. and I.B. MacNeill (1987), "Selection and estimation of growth models with application to forecasting AIDS". Technical Report TR-87-09, Department of Statistical and Actuarial Sciences, The University of Western Ontario.

Healy, M.J.R. and H.E. Tillett (1988), "Short-term extrapolation of the AIDS epidemic". *Journal of the Royal Statistical Society, Series A* 50–61.

Jandhyala, V.K. and I.B. MacNeill (1989), "Change detection methodology for modelling the incidence of AIDS". Technical Report TR-89-01, Department of Statistical and Actuarial Sciences, The University of Western Ontario.

Johnson, A.M. (1988), "Social and behavioural aspects of the HIV epidemic — A review". *Journal of the Royal Statistical Society, Series A* **151**, 99–114.

Kalbfleisch, J.D. and J.F. Lawless (1988), "Inference based on retrospective ascertainment. An analysis of the data on transfusion related AIDS". Technical Report STAT-88-02, Department of Statistics and Actuarial Science, University of Waterloo.

Kanouse, D.E. (1988), "Identifying possible futures for the AIDS epidemic". Paper presented at the eighth International Symposium on Forecasting, Amsterdam, Netherlands, June 12–15.

Longini, I.M., W.S. Clark, R.H. Byers, J.W. Ward, W.W. Darrow, G.F. Lemp and H.W. Hethcote (1989), "Statistical analysis of the stages of HIV infection using a Markov model". *Statistics in Medicine* **8**, 831–843.

MacNeill, I.B. (1989), "The reporting-delay function". Technical Report TR-89-04, Department of Statistical and Actuarial Sciences, The University of Western Ontario.

MacNeill, I.B., Q.P. Duong, V.K. Jandhyala and L. Liu (1990), "Adjustment for reporting-delay and estimation of the size of the HIV infected population in the U.S.A.". In *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time, October 1989*, eds. A.C. Singh and P. Whitridge. (to appear).

Martin, J.L. (1987), "The impact of AIDS on gay male sexual behaviour patterns in New York City". *American Journal of Public Health* **77**, 578–581.

McKusick, M., W. Horstman and J.J. Coates (1985), "AIDS and sexual behaviour reported by gay men in San Franscisco". *American Journal of Public Health* **75**, 493–496.

McNeill, W.H. (1976). *Plagues and Peoples*. Doubleday: New York.

Morgan, W.M and J.W. Curran (1986), "Acquired Immunodeficiency Syndrome: current and future trends". *Public Health Report 101* **5**, 459–465.

Panjer, H.H. (1987), "AIDS: Survival analysis of persons testing HIV^+". Working paper series in Actuarial Science ACTSC87-14, The University of Waterloo.

"Public Health Service Plan for the Prevention and Control of AIDS and the AIDS virus", Report of the Coolfont Planning Conference, U.S. Public Health Service, Washington, 1986, p.1.

Winkelstein, W., M. Samuel, N.S. Padian and J.A. Whiley (1987), "Select sexual practices of San Franscisco heterosexual men and risk of infection by the human immunodeficiency virus". *Journal of American Medical Association* **257**, 1470–1471.