STATISTICS IN MEDICINE, VOL. 8, 831–843 (1989)

STATISTICAL ANALYSIS OF THE STAGES OF HIV INFECTION USING A MARKOV MODEL

IRA M. LONGINI Jr., AND W. SCOTT CLARK

Department of Epidemiology and Biostatistics, Division of Biostatistics, Emory University, Atlanta, GA 30322, U.S.A.

ROBERT H. BYERS, JOHN W. WARD AND WILLIAM W. DARROW

AIDS Program, Center for Infectious Diseases, Centers for Disease Control, Atlanta, GA 30333, U.S.A.

GEORGE F. LEMP

AIDS Office, Department of Public Health, 1111 Market Street, San Francisco, CA 94103, U.S.A.

AND

HERBERT W. HETHCOTE

Department of Mathematics, University of Iowa, Iowa City, Iowa 52242, U.S.A.

SUMMARY

We use a staged Markov model to estimate the distribution and mean length of the incubation period for acquired immunodeficiency syndrome (AIDS) from a cohort of 603 human immunodeficiency virus (HIV) infected individuals who have been followed through various stages of infection. The model partitions the infected period into four progressive stages: (1) infected but antibody-negative; (2) antibody-positive but asymptomatic; (3) pre-AIDS symptoms and/or abnormal haematologic indicator; and (4) clinical AIDS. We also model a fifth stage: death due to AIDS. The estimated mean (median) waiting times in each stage of infection are stage 1, 2·2 (1·5) months; stage 2, 52·6 (36·5) months; stage 3, 62·9 (43·6) months; and stage 4, 23·6 (16·3) months. We estimate the mean AIDS incubation period (from infection to development of clinical AIDS) as 9·8 years with a 95 per cent confidence interval of [8·4, 11·2] years. The paper also considers the estimated density function of the AIDS incubation period and the estimated survival functions for individuals in each stage of infection. This work represents one of the most complete statistical descriptions to date of the natural history of HIV infection.

KEY WORDS Acquired immunodeficiency syndrome (AIDS) Human immunodeficiency virus (HIV) Markov model Stages of infection Stochastic process

INTRODUCTION

The recognition that individuals infected with human immunodeficiency virus (HIV) pass through a series of stages, from infected but antibody-negative to acquired immunodeficiency syndrome (AIDS) diagnosis,[1] implies that the mathematical modelling of the infection process is most naturally carried out by the use of a staged model. Such models have had successful use in the description of the progression of individuals through stages of cancer[2–5] and other pathogenic processes.[6–9] Data available on the progression of individuals to AIDS generally are arranged in cohorts of persons who were recently infected or whose serum specimens were found to be

0277–6715/89/070831–13$06.50

Received September 1988
Revised February 1989

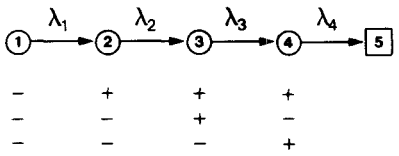

Figure 1. The flows through the four stages of infection and the fifth stage (death). The stages of infection are identified by the measured indicators available for the two cohorts making up the data

positive.[10] One can fit a staged stochastic model to such cohorts with use of maximum likelihood methods[2] and then use the fitted model to estimate the probability density function of the AIDS incubation period, a critical determinant of the dynamics of the HIV epidemic. Both the estimation of the number of infected individuals at previous times from reported AIDS cases and the prediction of the future number of HIV-infected individuals and AIDS cases based on the past number of AIDS cases or on HIV seroconversion curves depend heavily on knowledge of this probability density function.[11–15] In addition, the probability density function of the AIDS incubation period is an integral component of epidemic models used to investigate the dynamics and future trends of the HIV epidemic.[16–19]

In this report, we fit a five-stage, time-homogeneous Markov model to heavily censored data from a cohort of 513 homosexual and bisexual men from the San Francisco area found to be HIV seropositive, 73 individuals known to have received transfusions of HIV-infected blood, and 17 haemophiliacs who received HIV-infected factor VIII. The model partitions the infected period into four progressive stages (Figure 1). The first stage is HIV infection but with antibody-negative status. Stage 2 is antibody-positive status but asymptomatic. The third stage occurs when an individual develops an abnormal haematologic indicator and/or prodromal illnesses (pre-AIDS symptoms), such as persistent generalized lymphadenopathy and oral candidiasis. Stage 4 is clinical AIDS, and we include a fifth stage in the model – death due to AIDS.

Because of the heavy left, right and interval censoring of the data, standard statistical methods do not readily apply to the available data. The staged Markov model that we apply to these data, however, handles the censoring in an efficient and natural fashion. This work represents one of the first attempts to estimate the waiting times for the specific stages of HIV infection.

THE DATA

A random sample of 548 seropositive men was selected from the larger cohort of 6709 homosexual and bisexual men who were enrolled at the San Francisco City Clinic between 1978 and 1980 for studies of hepatitis B.[10] Of these 548 seropositive men, 130 (24 per cent) had been diagnosed with AIDS as of March 1988. Some 494 men (90 per cent) seroconverted, but the time of seroconversion is known only to an interval. Thus, these men may well have been uninfected during the early part of this interval since the time of infection is unknown. For the 54 (10 per cent) men seropositive at the time of blood drawing, we know only that these men were infected some time before. We considered seropositive men not reported to have AIDS as free of AIDS up to January 1987 to allow for reporting delays. There was no useful information on the stages of infection for 35 men who were then excluded from our analysis. Of the 513 men included in this analysis, 130 (25 per cent) had developed AIDS and 76 (58 per cent) of these 130 had died. In addition, for some of these men clinical data concerning stage 3 were not consistently or routinely collected, and after 1984 only data on AIDS status were collected. On the average, follow-up was 7 to 8 years. The average age of these men at the time of their first interview was 30 years with a range of 17 to 57 years. In this paper, we refer to this sample as the San Francisco cohort.

The 90 individuals with transfusion and factor-VIII-associated HIV infection are known to have received HIV-infected blood or blood products;[20-22] we refer to this sample as the transfusion cohort, but the reader should keep in mind that the cohort also contains persons with haemophilia. For the transfused individuals used in the analysis, we knew they had received a single transfusion during the period of HIV risk.[20, 22] Fifteen of these 90 individuals (17 per cent) had developed AIDS during the period of follow-up; none had died. In the transfusion cohort, the average age of persons for whom an age was recorded was 54 years, with a range of 12 to 87 years. Since follow-up of these individuals has been for a short period of time, that is an average of 3·3 years, we used the transfusion cohort to estimate the waiting time of individuals for the early stages of infection only.

For HIV staging process considered here, we assumed that infected individuals progress irreversibly through the stages of infection (Figure 1). Thus, for example, once an antibody-positive individual is diagnosed with persistent lymphadenopathy, we classify that individual in stage 3 even if such symptoms do not persist upon subsequent examination; only a few individuals had such a pattern of symptoms. We assigned seropositive individuals with prodromal illness and/or abnormal haematologic indicators to stages 3 or 4 according to the diagnostic criteria of the Centers for Disease Control (CDC)[23,24] and Jaffe *et al.*[10] Since all the observed deaths in the two cohorts occurred among individuals in stage 4, the AIDS stage, the model allowed only that transition to death. The waiting time in stage 1 is the pre-HIV antibody period (which we will refer to as the pre-antibody period), while the waiting time from when an individual enters stage 1 until the individual reaches stage 4 is the AIDS incubation period.

The definitions for stages of HIV infection have changed as knowledge about the biology of the disease has progressed. The information on the individuals in our analysis has been obtained across a ten-year period in which the understanding of all stages of the HIV infection has changed. To categorize accurately individuals from several different sources and time periods, we used these four broad stages; this staging system further ensures the irreversible progression. The present data contain limited staging information and thereby do not allow for analysis that would employ a more sophisticated staging system.

The exact transition times among stages are not available in these data. For example, an antibody-positive individual examined in December 1983 and found free of pre-AIDS symptoms, might, upon re-examination in May 1985, exhibit lymphodenopathy. Thus, we would not know precisely when the individual made the transition from stage 2 to stage 3; we know only that the transition occurred at some time during the 18-month interval. Such a phenomenon has been termed interval censoring.[25] In addition, data may be right censored (that is, at the last observation an individual may still be in one of the infected stages) or left censored (that is, at the time of the first observation an individual may have already been in that stage for an indeterminate amount of time).

STAGED MARKOV MODEL

We modelled the progression of an infected individual through the stages of infection and ultimately to death as a time-homogeneous Markov process in which stages 1 to 4 are transient states, and stage 5 is an absorbing state. The transition intensities are $\lambda_i > 0$, $i = 1,2,3,4$, where $\lambda_i dt + o(dt)$ is the probability that an infected individual in stage i will make a transition to stage $i+1$ in the time interval $(t, t+dt)$, for $t \geqslant 0$. Higher-order terms in dt are $o(dt)$, where $\lim o(dt)/dt \to 0$ as $dt \to 0$. We define the probability that an individual who is in stage i at time t_0 will be in stage $k \geqslant i$, $i = 1, 2, 3, 4, 5$, at time $t_0 + t$ as $p_{ik}(t)$. Standard methods for Markov processes provide explicit formulae for $p_{ik}(t)$.[6] If all the transition intensities are distinct,

that is $\lambda_i \neq \lambda_j$ for all $i \neq j$, then the transition probabilities among the transient states are

$$p_{ik}(t) = (-1)^{k-i}\lambda_i \ldots \lambda_{k-1} \sum_{j=1}^{k} e^{-\lambda_j t} \Big/ \prod_{\substack{l=i \\ l \neq j}}^{k} (\lambda_j - \lambda_l), \quad i = 1, 2, 3, 4;\ k \geqslant i,\ k < 5. \tag{1}$$

The transition probabilities from a transient state i to the absorbing state (death) are

$$p_{i5}(t) = (-1)^{4-i}\lambda_i \ldots \lambda_4 \sum_{j=i}^{4} (1 - e^{-\lambda_j t}) \Big/ \lambda_j \prod_{\substack{l=i \\ l \neq j}}^{4} (\lambda_j - \lambda_l), \quad i = 1, 2, 3, 4. \tag{2}$$

We define T_I as the random variable for the AIDS incubation period. Then the probability density function for T_I is $f_I(t) = \lambda_3 p_{13}(t)$, where $p_{13}(t)$ is given by (1). The cumulative distribution function of T_I, $F_I(t) = Pr(T_I \leqslant t)$, is

$$F_I(t) = \int_0^t f_I(\tau)\mathrm{d}\tau. \tag{3}$$

Then the hazard function, $h_I(t)$, for developing AIDS is

$$h_I(t) = f_I(t)/[1 - F_I(t)], \tag{4}$$

where $h_I(t)\mathrm{d}t$ is the probability that an individual who has been incubating infection up to time t, but who has not yet developed AIDS, will develop AIDS in the next instant. The hazard function of our model is monotonically increasing in t, which agrees with the form of the hazard functions used by other investigators to model the AIDS incubation period.[15, 26–31] The staged model used here further assumes that individuals will eventually develop AIDS, and this assumption is consistent with the findings of Lui, Darrow and Rutherford.[30] It follows from the time-homogeneous assumption that the waiting time in stage i ($i = 1, 2, 3, 4$) is exponentially distributed with a mean of $\mu_i = 1/\lambda_i$ and a median of $\tilde{\mu} = \mu_i \ln 2$. The expected length of the AIDS incubation period is

$$E(T_I) = \int_t^\infty S_I(\tau)\mathrm{d}\tau = \mu_1 + \mu_2 + \mu_3, \tag{5}$$

and the variance of T_I is

$$\mathrm{Var}(T_I) = \mu_1^2 + \mu_2^2 + \mu_3^2. \tag{6}$$

We define T_i as the random variable for the time to death from stage i, $i = 1, 2, 3, 4$. Then the survival function for individuals in stage i is $S_{T_i}(t) = Pr(T_i > t) = 1 - p_{i5}(t)$, where $p_{i5}(t)$ is given in (2). The mean time to death and its variance from state i are, respectively,

$$E(T_i) = \sum_{j=i}^{4} \mu_j, \tag{7}$$

$$\mathrm{Var}(T_i) = \sum_{j=i}^{4} \mu_j^2, \qquad i = 1, 2, 3. \tag{8}$$

ESTIMATION OF PARAMETERS

We estimate the parameters by formulating the likelihood function on each individual's passage through the stages of infection. Let j ($j = 1, 2, \ldots, n$) denote the index for each of the n individuals in the cohort for whom staging information is available. Let $\tau_{j0} = 0, \tau_{j1}, \ldots, \tau_{jm_j}$ represent the

Table I. Number of individuals contributing to the likelihood function for each possible transition

Transition	Number of individuals	Data source
$1 \rightarrow 1$	16	Transfusion*
$1 \rightarrow 2$	70	Transfusion*
$1 \rightarrow 3$	17	Transfusion*
$1 \rightarrow 4$	2	Transfusion*
$2 \rightarrow 2$	88	San Francisco
$2 \rightarrow 3$	43	San Francisco
$2 \rightarrow$ (2 or 3)†	151	San Francisco
$2 \rightarrow 4$	31	San Francisco
$3 \rightarrow 3$	149	San Francisco
$3 \rightarrow 4$	58	San Francisco
$4 \rightarrow 4$	54	San Francisco
$4 \rightarrow 5$	76	San Francisco
Total	755	

* Blood transfusion or factor VIII.

† Within the San Francisco cohort, 151 HIV+ individuals were known to have not developed AIDS as of January 1987, but their symptom status (that is, whether they were in stage 2 or 3) remains unknown. These 151 subjects still contribute to the likelihood function in the form: $1-[p_{24}(t)+p_{25}(t)]$.

times at which we observe individual j to be in states $y_{j0}, y_{j1}, \ldots, y_{jm_j}$, respectively, where y_{j0} is the first stage in which we observe the individual. Since an assumption of the Markov process is that it is time-homogeneous,[2] the contribution that the jth individual makes to the likelihood function is

$$L_j(\lambda)=\prod_{k=0}^{m_j-1} p_{y_{jk}y_{jk+1}}(\tau_{jk+1}-\tau_{jk}), \tag{9}$$

where $p_{ik}(t)$ comes from (1) and (2), and $\lambda=(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The likelihood function over the n individuals is

$$L(\lambda)=\prod_{j=1}^{n} L_j(\lambda). \tag{10}$$

Additional contributions to the likelihood function are made by individuals for whom staging information is not directly available, as we indicate in the second footnote to Table I.

For the transfusion cohort, we used the known time of exposure to the infecting blood or blood products as the starting time for each individual in stage 1. For the San Francisco cohort, however, the time of infection was unknown, so we used the time of the first seropositive blood reading as the starting time for each individual in stage 2. We then formulated separate likelihood functions for the transfusion cohort to estimate λ_1, and, for the San Francisco cohort, to estimate λ_2, λ_3 and λ_4. We found maximum likelihood estimates (MLE) of the parameters λ by numerically maximizing the natural logarithm of the likelihood function. We accomplished this with the derivative-free, pseudo-Gauss–Newton algorithm in the BMDP statistical package.[32] This algorithm also provides the asymptotic variance-covariance matrix of the MLEs, $\hat{\lambda}$.

Table II. Estimated parameters λ and mean and median waiting times in each stage of infection based on the staged Markov model

Stage i	Parameter estimate $\hat{\lambda}_i \pm$ one std error, months^{-1}	Mean waiting time $\hat{\mu}_i$, months (years)	Median waiting time $\hat{\tilde{\mu}}_i$, months (years)
1	0·4571 ± 0·1381	2·2 (0·2)	1·5 (0·1)
2	0·0190 ± 0·0022	52·6 (4·4)	36·5 (3·0)
3	0·0159 ± 0·0018	62·9 (5·2)	43·6 (3·6)
4	0·0424 ± 0·0044	23·6 (2·0)	16·3 (1·4)

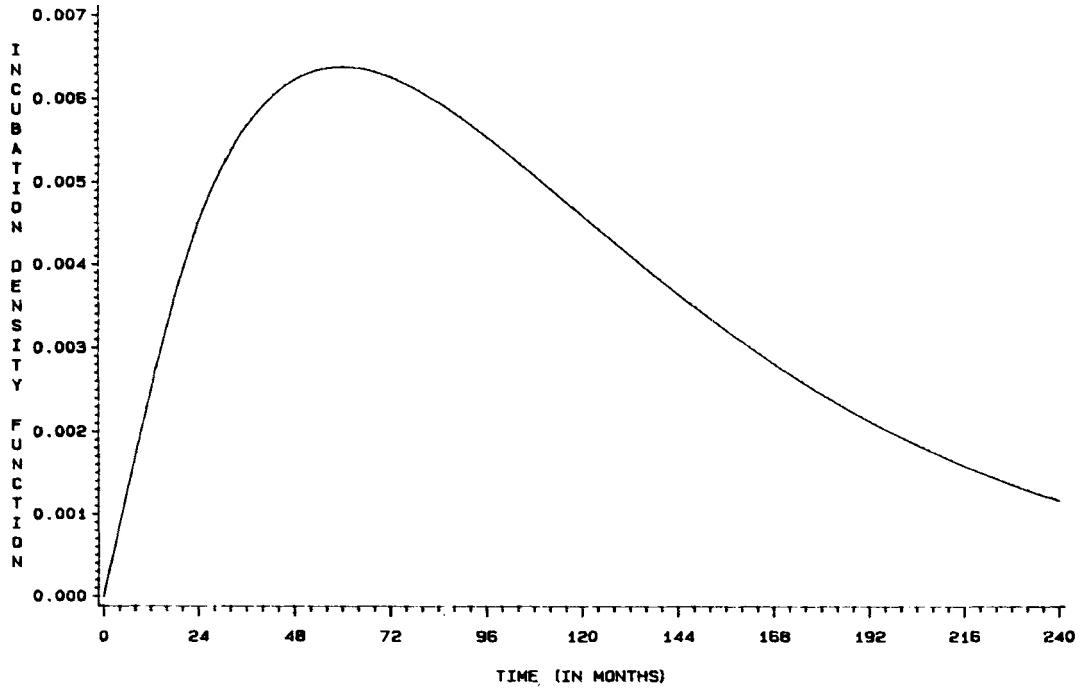

Figure 2. The estimated density function of the AIDS incubation period is

$$\hat{f}_1(t) = \hat{\lambda}_3 \hat{p}_{13}(t) = \hat{\lambda}_1 \hat{\lambda}_2 \hat{\lambda}_3 \sum_{j=1}^{3} e^{-\hat{\lambda}_j t} \Big/ \prod_{\substack{l=1 \\ l \neq j}}^{3} (\hat{\lambda}_j - \hat{\lambda}_l),$$

where $\hat{\lambda}_1 = 0{\cdot}4571$, $\hat{\lambda}_2 = 0{\cdot}0190$, $\hat{\lambda}_3 = 0{\cdot}0159$ from Table II. The mean and median of the AIDS incubation period are 117·7 months and 99·0 months, respectively

RESULTS

Using the ML procedure, we fitted the model to the infection histories of individuals described above. For each of the possible transitions in the model, the number of subjects and the cohort from which they came are given in Table I. There were 755 contributions to the likelihood function from these 603 individuals. The estimated parameters and mean and median waiting times in each stage appear in Table II. The estimated mean AIDS incubation period from (5) is 9·8 years (117·7 months) with a 95 per cent confidence interval of [8·4, 11·2] years. Based on the model, the estimated median AIDS incubation period is 8·3 years (99·0 months). Figures 2 and 3, respectively,

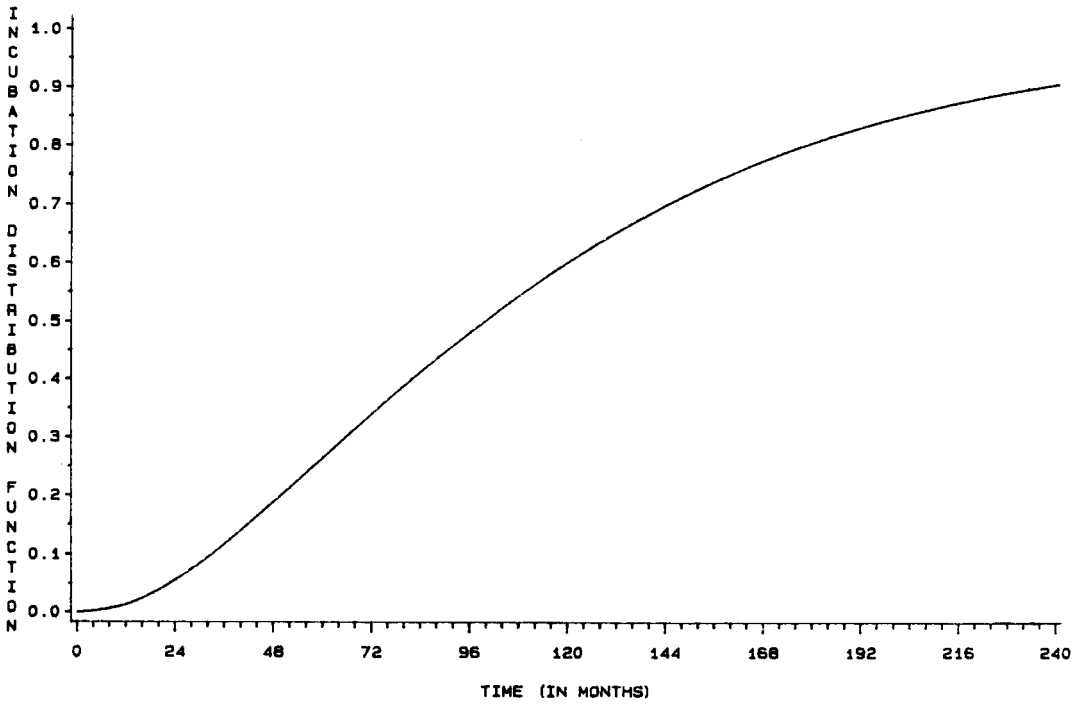

Figure 3. The estimated cumulative distribution function of the AIDS incubation period is

$$\hat{F}_1(t) = \int_0^t \hat{f}_1(\tau)\,d\tau = \hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_3 \sum_{j=1}^{3} (1-e^{-\hat{\lambda}_j t}) \Big/ \hat{\lambda}_j \prod_{\substack{l=1 \\ l\neq j}}^{3} (\hat{\lambda}_j - \hat{\lambda}_l),$$

where $\hat{\lambda}_1 = 0{\cdot}4571$, $\hat{\lambda}_2 = 0{\cdot}0190$, $\hat{\lambda}_3 = 0{\cdot}0159$ from Table II

Table III. Estimated mean and median survival times and five-year survival proportion from each stage of infection based on the staged Markov model

Stage i	Mean survival time $\hat{E}(T_i)$ months (years)	Median survival time, months (years)	Cohort five-year survival proportion
1	141·3 (11·8)	123·6 (10·3)	0·854
2	139·1 (11·6)	121·4 (10·1)	0·843
3	86·5 (7·2)	69·3 (5·8)	0·569
4	23·6 (2·0)	16·3 (1·4)	0·079

show the estimated probability density and cumulative distribution functions. Based on these functions, the probability that a newly infected individual will have developed AIDS within five years of infection is 0·27. From Table II, the estimated mean pre-antibody period is 2·2 months (9·5 weeks) with a 95 per cent confidence interval of [0·9, 3·5] months ([3.9, 15·2] weeks). We estimate the median as 1·5 months (6·5 weeks).

Table III gives the mean and median survival times from each stage of infection along with the proportion surviving after five years. For those individuals who enter the AIDS state, that is stage 4, median survival time is 16·3 months, and the proportion who survive at least five years is 0·079. For newly infected individuals, those in stage 1, median survival time is 10·3 years; 0·85 of

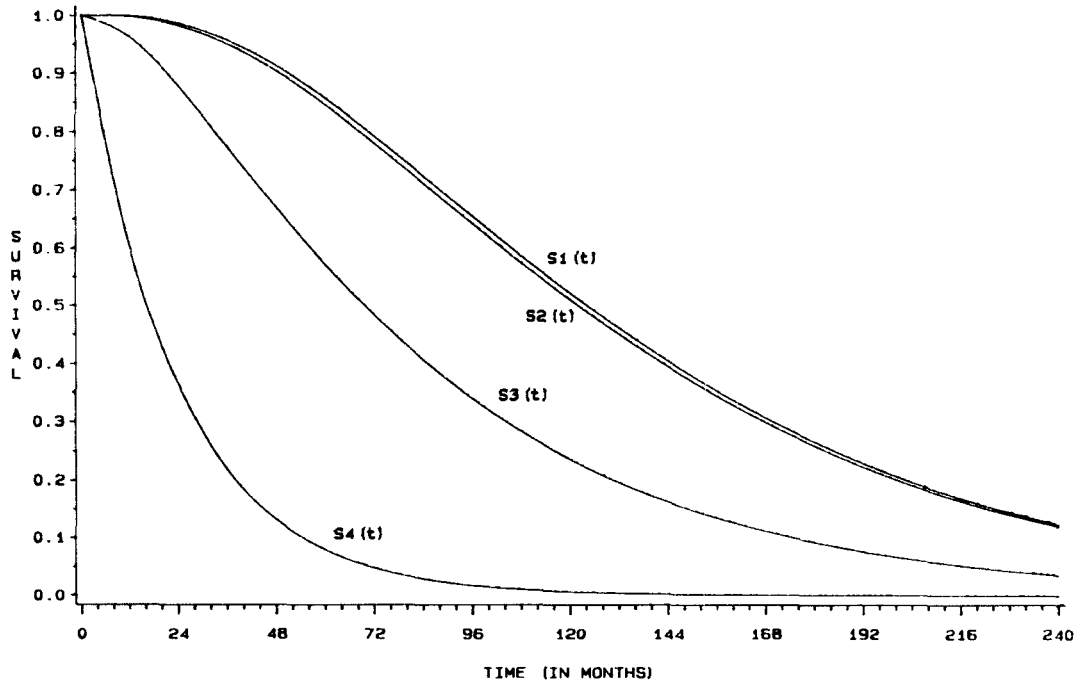

Figure 4. The estimated survival functions from each of the four stages of infection are $\hat{S}_{T_i}(t) = 1 - \hat{p}_{i5}(t)$, $i = 1, 2, 3, 4$, where $\hat{p}_{ik}(t)$ is found by evaluating (2) at the estimates $\hat{\lambda}$ given in Table II.

such individuals survive five years or more. The estimated survival functions, $\hat{S}_{T_i}(t) = 1 - \hat{p}_{i5}(t)$, from each stage of infection appear in Figure 4.

DISCUSSION

Our estimated mean AIDS incubation period of 9·8 years for primarily sexually-infected homosexual and bisexual men and transfusion-infected individuals, is a bit longer than that of 7·97 years found for transfusion-infected individuals aged 5 to 59 by Medley *et al.*[26, 27] Lu *et al.*[28] estimated a mean incubation period of 4·5 years with a 90 per cent confidence interval of [2·6, 14·2] years. The data sets used by Medley *et al.* and Lui *et al.* were subject to length-biased sampling since all individuals were ascertained because they had an AIDS diagnosis. Thus, these cohorts included only those individuals who had relatively short incubation periods. Although both analyses attempted to adjust for such a source of bias, the corrective measures were indirect and may have achieved only partial success. In addition, other sources of bias may have occurred due to the increasing hazard function of the Weibull distribution[33, 34] used in both analyses. In the analysis presented here, with use of the staged Markov model, we partitioned the incubation period into three stages where we assumed each stage had a constant hazard function. This renders the analysis less subject to those forms of bias that affect estimation based on models that do not have constant hazard functions.[33, 34] Ascertainment of individuals used in our analysis occurred only because of their known infection and, thus, length-biased sampling was probably not a

problem either. If, however, our assumption of constant hazard functions within stages was violated, than our analysis would be subject to model mis-specification biases.

Because the transition rates are estimated from two different sources, the model assumes a similar pattern of disease progression among individuals infected from sexual and blood-borne sources. The lack of stage 1 information in the San Francisco cohort and the relatively short follow-up in the transfusion cohort allows λ_2 to be the only transition rate which can be separately estimated for each cohort. The transfusion cohort contains 49 additional transitions from stage 2 that are not included in Table I because these transitions were not used in estimating λ_1. When we use these additional data, we estimate transition intensity for stage 2 to be $\hat{\lambda}_2 = 0{\cdot}0198 \pm 0{\cdot}0032$, which is not significantly different (statistically) from that of the San Francisco cohort ($\hat{\lambda}_2 = 0{\cdot}0190 \pm 0{\cdot}0022$, from Table II). Thus, the average waiting time in stage 2 is the same for individuals with sexual or blood-borne exposure to HIV. Such a comparison of the transition intensity for stage 3 is not possible at present since few individuals in the transfusion cohort have developed AIDS. Therefore, we do not have sufficient data to conclude that the average length of the AIDS incubation period is different for sexual and blood-borne HIV infections.

A number of other investigators have estimated the 'AIDS incubation period' as the waiting time from stage 2 to the entering of stage 4. DeGruttola and Mayer[29] fitted a Weibull distribution to data for men from the San Francisco cohort observed to seroconvert (to an interval); the estimated mean waiting time was 9·05 years. This is relatively close to our estimate of 9·6 years (that is $\hat{\mu}_2 + \hat{\mu}_3$, Table II). Lui, Darrow and Rutherford[30] fitted a Weibull distribution to data for 84 men from the San Francisco cohort observed to seroconvert within a one-year period. Their estimate of the mean waiting time from the first seropositive blood, that is, the early part of stage 2 to AIDS diagnosis, was 7·8 years, with a 90 per cent confidence interval of [4·2, 15·0] years. Harris[31] combined cohorts of HIV-infected individuals from the San Francisco cohort, transfusion recipients, and adults with haemophilia. He estimated the mean AIDS incubation period as 9·8 years, identical to our estimate.

Recently, Lemp *et al.*[15] and Hessol *et al.*[35] examined 359 men from the hepatitis B vaccine trial who had seroconverted within a two-year period. These men represent a special subset of the San Francisco City Clinic Cohort Study. Lemp *et al.*[15] defined the AIDS incubation period as the waiting time from the midpoint between the date of the last negative and the first positive blood specimen and the date of AIDS diagnosis. Using the Kaplan–Meier estimator,[36] they estimated the median AIDS incubation period as 10·8 years, somewhat longer than our estimate of 8·3 years. In addition, Hessol *et al.*[35] estimated that 0·15 of newly infected individuals will develop AIDS within five years, compared with our estimate of 0·27. The cohort used by these investigators exhibited less censoring than that used in our analysis; this allowed them to employ standard statistical methods to estimate the AIDS incubation period. Their cohort contains some of the individuals used in our sample plus additional individuals that we did not use. The differences in the results of the two analyses are due to the differences in the two cohorts used, not due to the methods employed.

Statistical estimates of the mean and median of the pre-antibody period, that is the waiting time in stage 1, have not been previously available. Clinical studies suggest an pre-antibody period somewhere between 6 and 16 weeks,[21, 37–41] a range close to the 95 per cent confidence interval reported here. It could, however, possibly be as long as 6 to 14 months.[42] These clinical studies were difficult to carry out and were based on small numbers of individuals. C. R. Horsburgh has examined the findings from a number of those clinical studies in which the exposed individuals had at least one negative antibody test after exposure to HIV (personal communication). He concludes that the median of the upper bound on the pre-antibody period is 90 days, a figure above our estimated median of 45·5 days.

Stage 1 may be a critical period for HIV transmission because individuals are thought to have higher concentrations of HIV antigen in their blood[43, 44] and are presumably more infectious to others through sexual or blood-related contact at a time when they are to have detectable antibodies. Such individuals may continue to have sexual relations, give blood, or possibly share needles (if they are IV drug users) without the knowledge that they could transmit HIV to others. With regard to blood transfusions, Ward *et al.*[45] have estimated that 26 per million blood transfusions were infected by individuals who gave blood while in stage 1 or by those in a later stage of infection but with no antibodies detected (assuming a sensitivity of 99 per cent for the antibody test). With regard to the former source of infection, they assumed that the pre-antibody period had a fixed-length interval of eight weeks, a figure slightly shorter than our estimated mean of 9·5 weeks.

Some investigators have hypothesized that a proportion p of HIV-infected individuals will eventually develop AIDS, while the remainder will not. There is little hope, however, of estimating this proportion from our data, given the degree of right censoring for AIDS (76 per cent). Lui, Darrow and Rutherford[30] estimated p from a cohort that exhibited 75 per cent right censoring for AIDS. Their best estimate was $\hat{p}=0{\cdot}99$ with a 90 per cent confidence interval of [0·38, 1·00]. Because of this lack of precision, we chose to assume in our model that all HIV-infected individuals would eventually develop AIDS (if they did not die first from some other competing cause of death).

Our estimated median survival time of 16·3 months is slightly longer than the previous estimate from San Francisco of 12·2 months[46] or the estimate from New York City of 11·4 months.[47] Nevertheless, we included our estimates of survival to show their relative contribution to the overall staging process modelled; we do not intend that they be used as the 'best estimates' of survival following the diagnosis of AIDS.

As mentioned above, the data on the progression of individuals through the stages of disease exhibit a high degree of left, right and interval censoring, and this makes it impossible to specify exact transition times for individuals and necessitates the use of a time-homogeneous model. Thus, it is impossible to specify a goodness-of-fit test or to examine critically the appropriateness of the time-homogeneity assumption. We could drop this assumption if we let the transition intensities have the form $\lambda_i(t)=\lambda_i\theta_i(t)$, where $\lambda_i(t)dt+o(dt)$ is the probability that an infected individual who has been in stage i for t units of time will make a transition to stage $i+1$ in the time interval $(t, t+dt)$, for $t\geqslant 0$, $i=1, 2, 3, 4$. We would have to specify the form of $\theta_i(t)$. We could not, however, fit such a process effectively to interval-censored data except in the case of a very simple step function, such as $\theta_i(t)=1$ for $0\leqslant t\leqslant t_1$, and $\theta_i(t)=\phi>1$ for $t_1>t$. As more data that exhibit less censoring become available, then we can use time-dependent staged models effectively in their analysis. Another possible approach to data analysis is to use a semi-Markov model.[9] The heavy censoring in the present data, however, would make non-parametric estimation of the waiting-time distribution for each stage difficult. Such non-parametric estimates, if obtained, could be examined to see if the waiting-time distributions within stages were really exponentially distributed, that is, constant hazard functions. Such an analysis may only be possible if we obtained data for which the dates of transitions among stages are better known than they are for the data we used.

Our model yields quite reliable results for the median AIDS incubation period, but it is no more able to predict the long-term behaviour of the AIDS incubation period distribution than any other available method. Such predictions will depend on additional data, especially those concerning the right tail of the distribution (that is, all HIV-infected individuals may not develop AIDS). Nonetheless, we believe that the time-homogeneous Markov model that we employ is appropriate for the available data.

ACKNOWLEDGEMENTS

We thank J. M. Karon, W. M. Morgan, H. W. Jaffe and C. R. Horsburgh of the AIDS program at the Centers for Disease Control (CDC) for their helpful comments, and G. W. Rutherford III, Alan Lifson and the City Clinic Cohort Study staff of the AIDS office, San Francisco Department of Health, for their assistance. We also thank the two anonymous referees for their constructive criticism. This research was partially supported by contract 200-07-0515 from the CDC and by NIH Grant 1-RO1-AI22877.

REFERENCES

1. Redfield, R. R., Wright D. and Tramont, E. 'The Walter Reed staging classification for HTLV-III/LAV infection'. *New England Journal of Medicine*, **314**, 131–132 (1986).
2. Kay, R. 'A Markov Model for analysing cancer markers and disease states in survival studies', *Biometrics*, **42**, 855–865 (1986).
3. Moolgavkar, S. H. 'The multistage theory of carcinogenesis and the age distribution of cancer in man', *Journal of the National Cancer Institute*, **61**, 49–52 (1978).
4. Armitage, P. and Doll, R. 'Stochastic models for carcinogenesis', Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 19–38 (1961).
5. Gail, M. H. 'Evaluating serial cancer marker studies in patients at risk of recurrent disease', *Biometrics*, **37**, 67–78 (1981).
6. Chiang, C. L. *An Introduction to Stochastic Processes and Their Applications*, 2nd edn, Krieger, New York, 1980.
7. Fix, E. and Neyman, N. J. 'A simple stochastic model of recovery, relapse, death and loss of patients', *Human Biology*, **23**, 205–241 (1951).
8. Kalbfleisch, J. D., Lawless, L. F. and Vollmer, N. M. 'Estimation in Markov models from aggregate data', *Biometrics*, **39**, 907–919 (1983).
9. Lagakos, S. W., Sommer, C. J. and Zelen, M. 'Semi-Markov models for partially censored data', *Biometrika*, **65**, 311–317 (1978).
10. Jaffe, H. W., Darrow, W. W., Echenberg, D. F., O'Malley, P. M., Getchell, J. P., Kalyanaraman, V. S., Byers, R. H., Drennan, D. P., Braff, E. H., Curran, J. W. and Francis, D. P. 'The acquired immunodeficiency syndrome in a cohort of homosexual men: a six-year follow-up study', *Annals of Internal Medicine*, **103**, 210–214 (1985).
11. Hyman, J. M. and Stanley, E. A. 'Using mathematical models to understand the AIDS epidemic', *Mathematical Biosciences*, **90**, 415–473 (1988).
12. Brookmeyer, R. and Gail, M. H. 'The minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States', *Lancet*, **ii**, 1320–1322 (1986).
13. Coolfont Report, *Public Health Report*, **101**, 341–348 (1986).
14. Morgan, W. M. and Curran, J. W. 'Acquired immunodeficiency syndrome: current and future trends', *Public Health Reports*, **101**, 459–465 (1986).
15. Lemp, G. F., Payne, S. F., Rutherford, G. W., Hessol, N. A., Winkelstein, W., Chen, R. T., Wiley, J. A., Moss, A. R., Chaisson, R. E., Feigal, D. W. and Werdegar, D. 'Projections of AIDS morbidity and mortality in San Francisco using epidemic models', Fourth International Conference on AIDS, Stockholm (abstract) (1988).
16. Anderson, R. M., May, R. M., Medley, G. F. and Johnson, A. M. 'A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS', *Journal of Mathematical Applications in Medicine and Biology*, **3**, 229–263 (1986).
17. Anderson, R. M. and May, R. M. 'The invasion, persistence and spread of infectious diseases within animal and plant communities', *Philosophical Transactions of the Royal Society of London*, **B314**, 533–570 (1986).
18. May, R. M. and Anderson, R. M. 'Transmission dynamics of HIV infection', *Nature (London)*, **326**, 137–142 (1986).
19. Dietz, K. and Hadeler, K. P. 'Epidemiological models for sexually transmitted diseases', *Journal of Mathematical Biology*, **26**, 2–25 (1988).
20. Ward, J. W., Deppe, D. A., Samson, S., Perkins, H., Holland, P., Fernando, L., Feorino, P. M., Thompson, P., Kleinman, S. and Allen, J. R. 'Risk of human immunodeficiency virus infection blood

donors who later developed the acquired immunodeficiency syndrome', *Annals of Internal Medicine*, **106**, 61–62 (1987).
21. Simmons, P., Lainson, F. A. L., Cuthbert, R., Steel, C. M., Peutherer, J. F. and Ludlam, C. A. 'HIV antigen and antibody detection: variable responses to infection in the Edinburgh haemophiliac cohort', *British Medical Journal*, **296**, 593–598 (1988).
22. Courouce, A.-M., Bouchardeau, F., Jullien, A.-M., Faucher, V. and Lentzy, M. 'Blood transfusion and human immunodeficiency virus (HIV) antigen', *Annals of Internal Medicine*, **108**, 771–772 (1988).
23. Centers for Disease Control, *Morbidity and Mortality Weekly Report*, **35**, 334–339 (1985).
24. Centers for Disease Control, *Morbidity and Mortality Weekly Report Supplement* **36**, 3s–15s (1987).
25. Peto, R. 'Experimental survival curves for interval-censored data', *Applied Statistics*, **22**, 86–91 (1973).
26. Medley, G. F., Anderson, R. M., Cox, D. R. and Billard, L. 'Incubation period of AIDS in patients infected via blood transfusion', *Nature (London)*, **328**, 719–721 (1987).
27. Medley, G. F., Billard, L., Cox, D. R. and Anderson, R. M. 'The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS)', *Journal of the Royal Statistical Society, Series B*, **233**, 367–377 (1988).
28. Lui, K.-J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W. and Bregman, D. J. 'A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome', *Proceedings of the National Academy of Sciences of the USA*, **83**, 3051–3055 (1986).
29. DeGruttola, V. and Mayer, K. H. 'Assessing and modelling heterosexual spread of the human immunodeficiency virus in the United States', *Review of Infectious Diseases*, **10**, 138–150 (1988).
30. Lui, K.-J., Darrow, W. W. and Rutherford, III, G. W. 'A model-based estimate of the mean incubation period for AIDS in homosexual men', *Science*, **240**, 1333–1335 (1988).
31. Harris, J. E. 'The incubation period for human immunodeficiency virus (HIV)', in Kulstad, R. (ed.) *AIDS 1988: AAAS Symposia Papers*, AAAS, Washington DC, 1988.
32. Ralston, M. 'Derivative-free nonlinear regression', in *BMDP Statistical Software Manual*, 305–329, University of California Press, Berkeley, 1985.
33. Brookmeyer, R. and Gail, M. H. 'Biases in prevalent cohorts', *Biometrics*, **43**, 739–749 (1987).
34. Brookmeyer, R., Gail, M. H. and Polk, B. F. 'The prevalent cohort study and the acquired immunodeficiency syndrome', *American Journal of Epidemiology*, **126**, 14–24 (1987).
35. Hessol, N. A., Rutherford, G. W., Lifson, A. R., O'Malley, P. M., Cannon, L., Doll, L. S., Darrow, W. W., Jaffe, H. W. and Werdegar, D. 'The natural history of HIV infection in a cohort of homosexual and bisexual men: a decade of follow-up', Fourth International Conference on AIDS, Stockholm (abstract) (1988).
36. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
37. Cooper, D. A., Gold, J., Maclean, P., Donovan, B., Finlayson, R., Barnes, T. G., Michelmore, H. M., Brooke, P. and Penny, P. R. 'Acute AIDS retrovirus infection', *Lancet*, **i**, 537–540 (1985).
38. Ho, D. D., Sarngadharan, M. G., Resnick, M. D., DiMarzo-Veronese, F., Rota, T. R. and Hirsh, M. S. 'Primary human T-lymphotropic virus type III infection', *Annals of Internal Medicine*, **103**, 880–883 (1985).
39. Esteban, J. I., Wai-Kuo Shih, J., Tai, C.-C., Bodner, A. J., Kay, J. W. and Alter, H. J. 'Importance of western blot analysis in predicting infectivity of anti-HTLV-III/LAV positive blood', *Lancet*, **ii**, 1083–1086 (1985).
40. Kumar, P., Pearson, J. E., Martin, D. H., Leech, S. H., Buisseret, P. D., Bezbak, H. C., Gonzalez, F. M., Royer, J. R., Streicher, H. Z. and Sazinger, W. C. 'Transmission of human immunodeficiency virus by transplantation of a renal allograft, with development of the acquired immunodeficiency syndrome', *Annals of Internal Medicine*, **106**, 244–245 (1987).
41. Gaines, H., von Sydow, M., Sonnerborg, A., Albert, J., Czajkowski, J., Pehrson, P. O., Chiodi, F., Moberg, L., Fenyo, E. M., Asjo, B. and Forsgren, M. 'Antibody response in primary human immunodeficiency virus infection', *Lancet*, **i**, 1249–1253 (1987).
42. Ranki, A., Valle, S.-L., Krohn, M., Allain, J. P., Franchini, G., Valle, S. L., Antonen, J., Leuther, M. and Krohn, K. 'Long latency precedes overt seroconversion in sexually transmitted human-immunodeficiency-virus infection', *Lancet*, **ii**, 589–593 (1987).
43. Goudsmit, J., deWolf, F., Paul, D. A., Speelman, H., Van Der Noordaa, J., Van Der Helm, H. J., De Wolf, F., Epstein, L. G., Krone, W. J., Wolsters, E. C., Oleske, J. M. and Coutinho, R. A. 'Expression of human

immunodeficiency virus (HIV-Ag) in serum and cerebrospinal fluid during acute and chronic infection', *Lancet*, **ii**, 177–180 (1986).
44. Lange, J. M. A., Paul, D. A., Huisman, H. G., deWolf, F., Van Den Berg, H., Coutinho, R. A., Danner, S. A., Van Der Noordaa, J. and Goudsmit, J. 'Persistent HIV antigenaemia and decline of HIV core antibodies associated with transition to AIDS', *British Medical Journal*, **293**, 1459–1462 (1986).
45. Ward, J. W., Holmberg, S. D., Allen, J. R., Cohn, D. O., Kritchley, S. E., Cleinman, S. H., Lenes, B. A., Ravenholt, O., Davis, J. R., Quinn, M. G. and Jaffe, H. W. 'Transmission of human immunodeficiency virus (HIV) by blood transfusion screened as negative for HIV antibody', *New England Journal of Medicine*, **318**, 473–478 (1988).
46. Lemp., G. F., Barnhart, J. L. and Rutherford, G. W. 'Trends in the length of survival for AIDS cases in San Francisco', Fourth International Conference on AIDS, Stockholm (poster) (1988).
47. Rothernberg, R., Woelfel, M., Stoneburner, R., Milberg, J., Parker, R. and Truman, B. 'Survival with the acquired immunodeficiency syndrome', *New England Journal of Medicine*, **317**, 1297–1302 (1987).