Appl. Statist. (1996)
45, *No.* 3, *pp.* 275–309

Markov Chains with Measurement Error: Estimating the 'True' Course of a Marker of the Progression of Human Immunodeficiency Virus Disease

By GLEN A. SATTEN†

Centers for Disease Control and Prevention, Atlanta, USA

and IRA M. LONGINI, Jr

Emory University, Atlanta, USA

[*Read before* The Royal Statistical Society *on Wednesday, October 18th, 1995, the President,* Professor A. F. M. Smith, *in the Chair*]

SUMMARY

A Markov chain is a useful way of describing cohort data. Longitudinal observations of a marker of the progression of the human immunodeficiency virus (HIV), such as CD4 cell count, measured on members of a cohort study, can be analysed as a continuous time Markov chain by categorizing the CD4 cell counts into stages. Unfortunately, CD4 cell counts are subject to substantial measurement error and short timescale variability. Thus, fitting a Markov chain to raw CD4 cell count measurements does not determine the transition probabilities for the true or underlying CD4 cell counts; the measurement error results in a process that is too rough. Assuming independent measurement errors, we propose a likelihood-based method for estimating the 'true' or underlying transition probabilities. The Markov structure allows efficient calculation of the likelihood by using hidden Markov model methodology. As an example, we consider CD4 cell count data from 430 HIV-infected participants in the San Francisco Men's Health Study by categorizing the marker data into seven stages; up to 17 observations are available for each individual. We find that including measurement error both produces a significantly better fit and provides a model for CD4 progression that is more biologically reasonable.

Keywords: Acquired immune deficiency syndrome; CD4 cell count; Hidden Markov model; Human immunodeficiency virus disease; Measurement error, San Francisco Men's Health Study

1. Introduction

Infection with human immunodeficiency virus type-1 (HIV-1), the virus that causes acquired immune deficiency syndrome (AIDS), is accompanied by a progressive decline in the CD4 cell count (the number of CD4 cells per microlitre), a type of white blood cell that plays a key role in the functioning of the immune system. This decline can be quantified by a longitudinal study of HIV-1-infected men, such as the San Francisco Men's Health Study (SFMHS), a population-based cohort study begun in mid-1984 (Winkelstein *et al.*, 1987). Modelling these data presents two

†*Address for correspondence*: Division of HIV/AIDS Prevention (E-48), Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, GA 30333, USA.
E-mail: gas0@cidhiv1.em.cdc.gov

statistical challenges. First, the seroconversion time (the natural zero of time in a model of longitudinal CD4 measurements among HIV-1-infected men) is not known for the men who were HIV positive at enrolment. Second, we seek a model that predicts both future CD4 cell counts and the time to the onset of clinical symptoms associated with AIDS; Jewell and Nielsen (1993) showed that constructing a joint model for survival and the time evolution of a variable that also predicts survival can be difficult.

This paper models these data by using staged Markov models which allow for measurement error and misclassification. These models lead to a joint description of the longitudinal CD4 cell counts and the time to onset of clinical AIDS. Various competing models of CD4 decline have been proposed, including log-linear and piecewise log-linear, with random effects correlation structure (DeGruttola *et al.*, 1991; Pawitan and Self, 1993), as well as Markov modelling (Lawless and Yan, 1991; Longini *et al.*, 1991). We have selected a Markov model for these data for several reasons. First, the growth curve approach requires knowledge of the time of infection; if this information is not available, the distribution of infection times must be known or simultaneously estimated (DeGruttola *et al.*, 1991). Second, the growth curve describes only CD4 cell counts, but we are also interested in the time to onset of AIDS; because CD4 is highly predictive of AIDS, the models for CD4 decline and AIDS-free survival time must be interrelated. Although Pawitan and Self (1993) proposed such a model, it requires the simultaneous estimation of parametric distributions of infection time and AIDS incubation times, and the distribution of marker trajectories is conditional on the (future) time of AIDS diagnosis. These problems do not arise when a Markov model is used, because time can be measured from entry to the study, and the model treats the longitudinal CD4 measurements in the same way as progression to AIDS. In addition, the Markov assumption may be reasonably accurate for modelling CD4 cell counts in HIV-infected individuals: Taylor *et al.* (1994) showed that an integrated Ornstein–Uhlenbeck diffusion process fits data from the Multicenter AIDS Cohort Study better than a linear growth curve model does, and studies have shown that a past CD4 cell count does not significantly improve the ability to predict the hazard for developing AIDS given a current CD4 cell count (DeGruttola *et al.*, 1993; Multi-cohort Analysis Project, 1993). Finally, we do not propose to use these models to predict individual CD4 trajectories, but rather to make actuarial projections of the distributions of marker values in a population at some point in time, given information such as a reconstruction of the history of the epidemic in the population.

Unfortunately, CD4 cell count data are subject to substantial measurement error and short timescale fluctuations, due to both imprecise measurement techniques and biological phenomena such as diurnal variation and variation in challenges to the immune system. We present here a novel approach to fitting a Markov chain to noisy data. We propose a hierarchical model comprising an underlying (or hidden) Markov chain and a measurement error model; from this model we determine the marginal distribution of the observed data by using the methodology of hidden Markov chains. Although hidden Markov models have been widely applied in statistics (see for example Rabiner (1989), Albert (1991), Le *et al.* (1992), Collins and Wugalter (1992) and Langeheine (1994)), their application has primarily been in areas in which an observable outcome is assumed to depend on a (hidden) state of the system, but in which a model of these hidden states is of secondary importance

(Besag, 1986; Fredkin and Rice, 1992a, b). In particular, the hidden Markov modelling methodology has not been used to explore situations in which the underlying Markov model is of primary interest, such as disease history modelling.

Section 2 discusses staged Markov models for the SFMHS data and introduces our model for fitting these models to noisy data. Section 2.1 is a brief discussion of the SFMHS data, while Section 2.2 discusses a naïve application of staged Markov models. Section 2.3 contains a description of the model and Section 2.4 discusses conditioning on the initial stage. Section 3.1 discusses the results obtained from our methodology on the behaviour of CD4 cell counts among participants in the SFMHS; Section 3.2 discusses the fit and prediction of a future observation. Section 3.3 discusses the effect of treatment on CD4 cell count progression and time to AIDS. We discuss our results in Section 4.

2. Progression of CD4 Cell Counts in Men Infected with Human Immunodeficiency Virus in San Francisco

2.1. *San Francisco Men's Health Study*

We use data from the SFMHS, a prospective study of a random sample of 1045 homosexual and bisexual men who were 25–54 years old at enrolment (Winkelstein, *et al.*, 1987), which occurred between June 1984 and January 1985. The men were examined approximately every 6 months and tested for HIV type 1 infection. We examine data from the time of enrolment until September 1992. During this time, 430 HIV-infected men had either

(a) at least two examinations with HIV seropositive test and CD4 cell count measurements,
(b) a seronegative test and at least one examination with a seropositive HIV test and CD4 cell count measurement or
(c) at least one examination with a seropositive HIV test with CD4 cell count measurement followed by an AIDS diagnosis.

Of these men, 384 were seropositive at enrolment, and 46 became infected (seroconverted) during the study. HIV-infected men had an average of 8.5 and a maximum of 17 examinations each, and thus a total of 3635 examinations at which CD4 tests were performed or AIDS was first diagnosed. Among seroconverters, the mean interval between the seronegative and seropositive tests was 7.4 months, with values ranging from 5 to 20 months. AIDS was diagnosed in 232 men during the study period.

Some individuals in the SFMHS began antiretroviral treatment with zidovudine in 1986 and pneumocystis pneumonia prophylaxis with pentamidine in 1989. At each examination, the men in the SFMHS were asked to report any use of prophylactic therapy (zidovudine and/or pentamidine) since their last visit. The percentages of HIV-infected men who reported therapy use since their previous examination increased steadily from 6.4% in 1987 to 45% in 1991. Although some studies have shown that therapy slows the progression to lower CD4 cell stages and AIDS (Graham *et al.*, 1991; Longini *et al.*, 1993; Hoover *et al.*, 1994), other recent analyses imply no therapy effect (Hessol *et al.*, 1994; Gauvreau *et al.*, 1994). Note that the SFMHS is an observational study, so that the assignment to treatment regimens was not randomized.

2.2. *Staged Models of CD4 Progression*

We considered a staged model for CD4 cell count in an individual before the onset of clinical AIDS; this model is a generalization of the model considered by Longini *et al.* (1991). We allowed for $d = 6$ CD4-based stages using cut points $\mathbf{A} = (\infty, 900, 700, 500, 350, 200, 0)$. A seventh (absorbing) stage was added for clinical AIDS. The resulting network of stages and allowed transitions is shown in Fig. 1. Unlike Longini *et al.* (1991), we are interested in two types of models: bidirectional, in which individuals can move in either direction between adjacent CD4-based stages, and unidirectional, in which individuals can move only from a stage characterized by a higher CD4 cell count to the adjacent stage characterized by a lower CD4 cell count. In either model, individuals can move directly and irreversibly to AIDS from stages characterized by CD4 cell counts less than 700; these allowable direct transitions to AIDS were determined by examining the observed CD4 stage preceding an AIDS diagnosis. The choice of seven stages follows that of Hethcote *et al.* (1991), who found that the survival distribution for a seven-stage model best agreed with the empirical survival distributions from cohort data. The choice of cut points at CD4 cell counts of 200 and 500 is motivated both by the biology of HIV disease (Centers for Disease Control and Prevention, 1992a; El-Sadr *et al.*, 1994) and by the need for a model that allows us to estimate the proportion of HIV-infected individuals with CD4 cell counts less than these values as a function of time; the remaining cut points are then chosen so that the range of observed CD4 cell counts is broken into stages of approximately equal size.

A summary of the observed transitions between stages in the SFMHS is shown in Table 1, which tabulates the number of times an individual who was observed in stage i was next seen in stage j. For example, 199 times an individual who was observed to be in stage 2 was observed to be in stage 3 at the time of his next visit. In parentheses beneath the number of observed transitions we have indicated the number of these transitions during which the use of treatment (zidovudine and/or pentamidine) was reported. Thus, in seven of the 199 transitions described above, the use of treatment was reported. Treatment information was not recorded at the time of AIDS diagnosis because these diagnoses generally did not occur at regularly scheduled visits. Hence, the treatment status between the last AIDS-free visit and diagnosis was imputed to be the same as the treatment status at the last AIDS-free visit. The use of treatment was reported in 680 of the 3635 transitions observed.

To allow for variation in the time intervals between the successive visits shown in Table 1, and to reduce the number of parameters, we fit a continuous time Markov

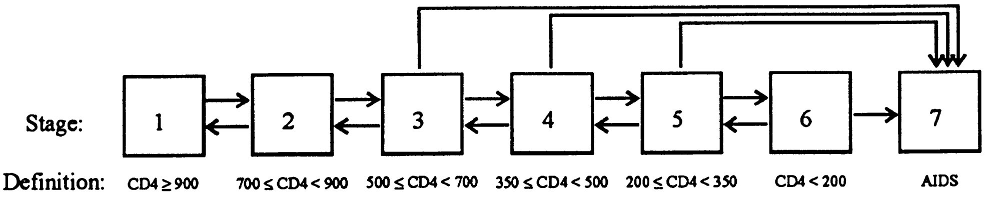

Fig. 1. Network of stages used for modelling the progression to AIDS in HIV-infected individuals: transitions are allowed between adjacent CD4-based stages, and direct transitions to AIDS are allowed from stages characterized by a CD4 cell count less than 700; for the unidirectional model, transitions from stage j to stage j' are not allowed if $j > j'$

TABLE 1
Observed transitions in the SFMHS†

Initial stage	*No. of transitions to the following stages:*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
1	187	104	67	16	3	1	6
	(3)	(0)	(3)	(0)	(0)	(0)	(1‡)
2	89	170	199	44	11	1	11
	(2)	(5)	(7)	(5)	(1)	(1)	(1‡)
3	35	145	377	261	79	10	17
	(2)	(9)	(34)	(20)	(9)	(4)	(1‡)
4	7	27	158	317	221	32	34
	(0)	(8)	(25)	(66)	(39)	(4)	(5‡)
5	1	5	32	121	295	136	48
	(0)	(0)	(4)	(34)	(101)	(52)	(14‡)
6	0	1	4	11	39	197	116
	(0)	(1)	(1)	(3)	(25)	(135)	(55‡)

†Numbers in parentheses are transitions during which the use of treatment (zidovudine and/or pentamidine) was reported.

‡Treatment status for transitions to stage 7 are imputed by assuming that the treatment regimen is unchanged since the most recent pre-AIDS visit.

chain to the serial CD4 cell count data. A continuous time model has fewer parameters than a discrete time model does, because the CD4 cell count must change continuously. Hence, only transitions between adjacent CD4-based stages are required to specify the model. The infinitesimal generator γ for the class of Markov models considered in this paper is

$$\begin{pmatrix} -\gamma_{12} & \gamma_{12} & 0 & 0 & 0 & 0 & 0 \\ \gamma_{21} & -(\gamma_{21}+\gamma_{23}) & \gamma_{23} & 0 & 0 & 0 & 0 \\ 0 & \gamma_{32} & -(\gamma_{32}+\gamma_{34}+\gamma_{37}) & \gamma_{34} & 0 & 0 & \gamma_{37} \\ 0 & 0 & \gamma_{43} & -(\gamma_{43}+\gamma_{45}+\gamma_{47}) & \gamma_{45} & 0 & \gamma_{47} \\ 0 & 0 & 0 & \gamma_{54} & -(\gamma_{54}+\gamma_{56}+\gamma_{57}) & \gamma_{56} & \gamma_{57} \\ 0 & 0 & 0 & 0 & \gamma_{65} & -(\gamma_{65}+\gamma_{67}) & \gamma_{67} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(For the unidirectional model, all elements in the lower half of the matrix are set to 0.)

2.3. *Errors-in-variables Model for Staged Markov Models*

When we tried to fit the two models described above to the SFMHS data, problems arose immediately. First, the backward model cannot be fitted to the data without some adjustment, as transitions from stages characterized by lower CD4 cell counts to stages characterized by higher CD4 cell counts are observed. Second, when we fit the bidirectional model to the observed data, the results, which are summarized in Table 2 under the heading 'naïve' model, are unacceptable in that they do not match clinical intuition: individuals with CD4 cell counts as low as 350 have almost as long a waiting time to AIDS as individuals with CD4 cell counts above 900. The results of the naïve model suggest that the CD4 cell count is not a very effective marker for staging HIV disease. This model's failure is explained by the large number of stages visited before absorption in the naïve model, indicating that many of these

TABLE 2
Mean time to AIDS and number of stages visited before absorption (with 95% confidence intervals), naïve model ($\sigma_1 = 0$) and best bidirectional and unidirectional models ($\sigma_1 = \hat{\sigma}$)†

CD4 stage	*Results for naïve model, $\sigma_1 = 0$*		*Results for best bidirectional model, $\hat{\sigma}_1 = 0.172$*		*Results for best unidirectional model, $\hat{\sigma}_1 = 0.219$*	
	$E(T_a)$	$E(N)$	$E(T_a)$	$E(N)$	$E(T_a)$	$E(N)$
900+	10.1 (9.2, 11.1)	20.7 (18.6, 22.8)	12.0 (11.0, 13.1)	7.5 (6.9, 8.1)	12.6 (11.4, 13.7)	5.6 (5.5, 5.8)
700–900	9.5 (8.6, 10.4)	19.7 (17.6, 21.8)	9.8 (9.0, 10.7)	6.5 (5.9, 7.1)	10.1 (9.3, 11.0)	4.6 (4.5, 4.8)
500–700	8.8 (7.9, 9.7)	17.8 (15.7, 20.0)	8.0 (7.2, 8.7)	5.4 (4.8, 6.0)	7.6 (7.0, 8.2)	3.6 (3.5, 3.8)
350–500	7.6 (6.7, 8.5)	15.1 (13.1, 17.1)	5.8 (5.2, 6.4)	4.1 (3.6, 4.6)	5.3 (4.8, 5.8)	2.7 (2.6, 2.9)
200–350	6.0 (5.2, 6.8)	11.3 (9.4, 13.1)	3.8 (3.3, 4.3)	2.9 (2.5, 3.3)	3.0 (2.7, 3.4)	1.9 (1.8, 2.0)
0–200	3.0 (2.4, 3.6)	5.0 (3.8, 6.2)	1.5 (1.2, 1.9)	1.3 (1.1, 1.5)	1.2 (1.0, 1.4)	1.0 (1.0, 1.0)

†$E(T_a)$, mean time to absorption (AIDS); $E(N)$, mean number of stages visited before absorption (AIDS).

transitions are due to measurement error or short timescale fluctuation (Table 2). Because data on many individuals are censored before an AIDS diagnosis, these transitions shorten the time to AIDS for the earlier stages and lengthen the time to AIDS for the later stages. An alternative way of viewing the failure of the naïve model is that classifying individuals into stages by their observed CD4 cell count is subject to error; this misclassification makes the stages less distinct.

CD4 cell counts are subject to two sources of variability: measurement error and short-term fluctuations. To measure the CD4 cell count, the total white cell count must be multiplied by the proportion of white cells that are CD4 cells; these two quantities must be determined separately and are each subject to measurement error. In addition, diurnal fluctuations and other short timescale factors of immune system function add variability. To estimate the magnitude of these two effects, Hoover *et al.* (1992) fitted a two-way random effects analysis-of-variance model to the logarithm of CD4 cell count data from 3145 non-HIV-infected men. Because some of these CD4 cell counts were replicated, they could determine the separate effects of measurement error and short timescale variability. The estimated standard error of the laboratory effect was 0.185, whereas that for the short timescale effect was estimated to be 0.075, yielding an estimated standard error for the total effect of 0.260. Because we are interested in modelling the progression of CD4 cell counts on the multiple-month or year timescale, we shall consider the short timescale variability as part of the 'measurement error'.

One approach to the measurement error problem, particularly when the states of the chain correspond to a discretization of a numerically valued variable, is to assume that a Markov model can be fitted to smoothed or adjusted data. This approach was taken by Longini *et al.* (1991), who used a persistence criterion to form a monotonic sequence of stages from data for which the underlying process was assumed monotonic, but in which both forward and backward transitions were observed. Although a smoothing approach is easily implemented, longitudinal data cannot be smoothed unless the correlation structure is known. Thus, if the smoothed data are to be used in a Markov chain, the correlation structure used to smooth the data should properly have the structure of a Markov chain. In addition, determining the proper amount of smoothing when the data consist of many relatively short series of observations is also problematic.

To address the problems presented by the naïve model, we developed an errors-in-variables approach to fitting Markov models to noisy data. We assume that each individual i has a 'true' CD4 cell count $Y_i(t)$ and a corresponding true stage $S_i(t)$ at each time t. At measurement times t_{ik}, $k = 1, \ldots, n_i$, we would like to observe the true CD4 cell count $Y_i(t_k) \equiv Y_{ik}$ and the true stage $S_i(t_k) \equiv S_{ik}$. However, because of measurement error and short timescale variability, we measure only the observed CD4 cell count X_{ik}. We use the term true rather than latent because latent stages often correspond to hypothesized but unobservable states, whereas we assume that the true stages $S(t)$ are defined by the same CD4 ranges shown in Fig. 1. We further assume that the true stages $S(t)$, rather than the discretized observed CD4 cell counts, follow a continuous time Markov chain. As CD4 trajectories may change after the onset of AIDS, we model only CD4 cell counts until AIDS is diagnosed. We assume that AIDS is diagnosed without error, so that, if the true stage S_{ik} corresponds to AIDS, Y_{ik} and X_{ik} each take the non-numerical value 'AIDS'. Finally, we follow the convention that capital letters denote random variables and lower case letters denote their observed values or realizations.

For the first six stages, let the CD4 interval that defines the jth stage be $[l_j, u_j)$. Then an individual is in true stage $S_i(t) = j$ if $Y_i(t) \in [l_j, u_j)$. Given the value of $S_i(t)$, we assume that $Y_i(t)$ is known to a distribution, and that the distribution of Y_{ik} given S_{ik} is independent of $S_{ik'}$, $k' \neq k$. For stages that are defined by the finite range of CD4 cell counts $[l_j, u_j)$ we assume that the stage-specific probability density function of Y_{ik}, denoted $f[y_{ik}|S_{ik} = j]$, is uniform, i.e.

$$f[y_{ik}|S_{ik} = j] = \begin{cases} 1/(l_j - u_j) & l_j \leqslant y_{ik} < u_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Because the first stage in our model is open ended, a non-uniform within-stage distribution must be specified. We chose

$$f[y_{ik}|S_{ik} = 1, \sigma_2] = \begin{cases} \dfrac{2}{\sqrt{(2\pi)} y_{ik} \sigma_2} \exp\left\{ -\dfrac{1}{2} \left(\dfrac{\log y_{ik} - \log 900}{\sigma_2} \right)^2 \right\} & y_{ik} \geqslant 900, \\ 0 & y_{ik} < 900, \end{cases} \tag{2}$$

i.e. the upper half of a log-normal distribution with $\mu = \log 900$. This choice was made so that the integral in equation (6) later could be carried out analytically for the choice of the distribution of X_{ik} given Y_{ik} made below. Finally, if $S_{ik} = 7$ (corresponding to clinical AIDS), we assume that X_{ik} and Y_{ik} take the values AIDS with probability 1.

As the variability in CD4 cell counts is stabilized on the log-scale, we assume that the logarithm of the observed CD4 cell count before AIDS diagnosis was normally distributed about the logarithm of the true CD4 cell count, with standard deviation σ_1, i.e.

$$\log X_{ik} = \log Y_{ik} + \epsilon_{ik}, \qquad \epsilon_{ik} \sim \text{normal}[0, \sigma_1^2],\ \epsilon_{ik} \perp \epsilon_{i'k'} \text{ for } i \neq i' \text{ and } k \neq k'. \tag{3}$$

We denote by $g(x_{ik}|Y_{ik}; \sigma_1)$ the probability density of X_{ik} specified by equation (3).

The model that we are proposing thus has the following hierarchical structure:

$$\mathbf{S}_i \sim \Pr[\mathbf{S}_i|\gamma, \pi] = \Pr[S_{i1}|\pi] \prod_{k=2}^{n_i} \Pr[S_{ik}|S_{i(k-1)}, \gamma], \tag{4a}$$

$$\mathbf{Y}_i|\mathbf{S}_i \sim \prod_{k=1}^{n_i} f[y_{ik}|S_{ik}; \sigma_2], \tag{4b}$$

$$\mathbf{X}_i|\mathbf{Y}_i \sim \prod_{k=1}^{n_i} g[x_{ik}|Y_{ik}; \sigma_1], \tag{4c}$$

where $\mathbf{S}_i$, $\mathbf{Y}_i$ and $\mathbf{X}_i$ denote vectors with kth component S_{ik}, Y_{ik} and X_{ik}, γ is the infinitesimal generator of the Markov chain, π is the vector of parameters which determine the initial stage occupation probabilities and σ_2 is a (possibly zero-dimensional) vector of parameters determining the within-stage distributions.

Equations (4a) and (4b), in conjunction with equations (1) and (2), define a model for the true CD4 cell count $Y(t)$ which is a step function for CD4 cell counts less than 900; CD4 cell counts greater than 900 have a truncated log-normal tail. The heights of the steps are determined by the transition probabilities of the Markov chain. An example of the family of distributions of $Y(t)$ is shown in Fig. 2, which shows the conditional probability of a CD4 cell count 1 year after a CD4 cell count between 700 and 900, conditionally on being AIDS free. We have written equations (4a)–(4c) as a three-level model only for convenience of notation; because $\mathbf{S}_i$ is completely determined by $\mathbf{Y}_i$, we could have written the marginal distribution of $\mathbf{Y}_i$ directly. Finally, we note a connection between our model and growth curve methodology. Both involve an equation like our equation (3); in our method, $Y(t)$ is a realization of a stochastic process, whereas in the growth curve method it is a deterministic function, although possibly with individual-specific parameters.

2.4. *Marginal Likelihood*

Because the true stages $S_i(t)$ and true CD4 cell counts $Y_i(t)$ are not observed, we construct the marginal likelihood of the observed CD4 cell count X_{ik} specified by model (4a)–(4c), at the measurement times of the ith individual. The marginal probability density function for $\mathbf{X}_i$, denoted $p(\mathbf{x}_i|\sigma_1, \sigma_2, \gamma, \pi)$, can be expressed as the integral

$$p[\mathbf{x}_i|\sigma_1, \sigma_2, \gamma, \pi] = \int \mathrm{d}y_1 \ldots \int \mathrm{d}y_{n_i} \prod_{k=1}^{n_i} g[x_{ik}|Y_{ik} = y_k, \sigma_1]\, f[\mathbf{y}|\gamma, \pi, \sigma_2], \tag{5}$$

where $f[\mathbf{y}|\gamma, \pi, \sigma_2]$, the marginal probability density function of $\mathbf{Y}$, is specified in

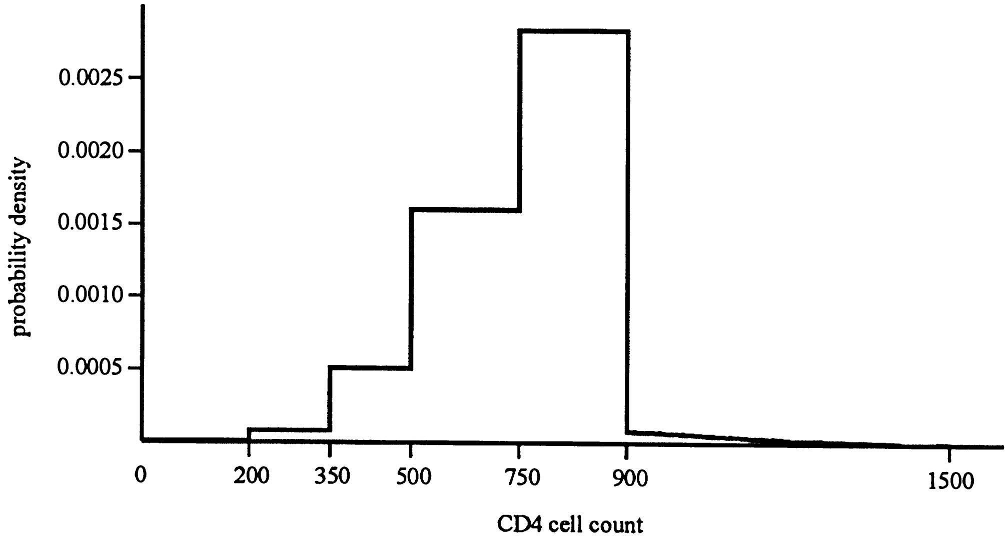

Fig. 2. Conditional distribution of the true CD4 cell count y_k at time t_k given that the true CD4 cell count y_{k-1} at time t_{k-1} 1 year before t_k was between 700 and 900: this distribution is specified in equations (4a) and (4b) and is calculated by using parameters from the best fitting bidirectional model for the SFMHS data

equations (4a) and (4b) . The integrals in equation (5) may be broken into integrals over each of the intervals $[l_j, u_j)$, thereby converting the integrations into sums over stages. By writing

$$p[x_{ik}|S_{ik}=j;\,\sigma_1,\,\sigma_2]=\int_{l_j}^{u_j} g[x_{ik}|Y_{ik}=y,\,\sigma_1]\,f[Y_{ik}=y|S_{ik}=j;\,\sigma_2]\,\mathrm{d}y, \qquad (6)$$

and inserting equation (6) into equation (5), we find that the marginal distribution of $\mathbf{X}_i$ has the form

$$p[\mathbf{x}_i|\sigma_1,\,\sigma_2,\,\gamma,\,\pi]=\sum_{s_1}\sum_{s_2}\ldots\sum_{s_{n_i}}\Pr[\mathbf{X}_i,\,\mathbf{S}_i=\mathbf{s};\,\sigma_1,\,\sigma_2,\,\gamma,\,\pi]. \qquad (7)$$

Without assumptions about the form of the joint distribution of $\mathbf{X}_i$ and $\mathbf{S}_i$, the sum in equation (7) would be over d^n terms, where d is the number of stages; each term in the sum represents one of the possible realizations of $\mathbf{S}_i$. The likelihood is a weighted sum over all possible sample paths of the true process, in which the weight assigned to a given realization of $\mathbf{S}_i$ is the probability that this realization generated the observed data $\mathbf{X}_i$. Generally, this sum rapidly becomes intractable. In our seven-stage process with up to 17 observation times, this sum has over 10^{14} terms. To construct the likelihood for the entire cohort we must evaluate this sum for each individual.

By using the Markov property for $\mathbf{S}_i$ and the independence model for X_{ik} given S_{ik}, we can easily evaluate the sum (7), as Baum *et al.* (1970) first showed. Rewriting equation (7) by passing the sums into the body of the equation, we obtain

$$p[\mathbf{x}_i|\sigma_1, \sigma_2, \gamma, \boldsymbol{\pi}] = \sum_{s_1} p[x_{i1}|S_{i1} = s_1, \sigma_1, \sigma_2]\,\Pr[S_{i1} = s_1|\pi] \sum_{s_2} p[x_{i2}|S_{i2} = s_2,$$

$$\sigma_1, \sigma_2]\,\Pr[S_{i2} = s_2|S_{i1} = s_1, \gamma] \sum_{s_3} \ldots \sum_{s_{n_i}} p[x_{in_i}|S_{in_i}, \sigma_1, \sigma_2]\,\Pr[S_{in}|S_{i(n_i-1)}, \gamma]. \quad (8)$$

Thus, calculating the marginal distribution reduces to matrix multiplication. Specifically, for each individual, define the column vector $V(x_{i1})$ with components V_j and matrices $T^{(k)}(x_{ik})$ with (j, j')th element $T^{(k)}_{jj'}(x_{ik})$ by

$$V_j(x_{i1}) = p[x_{i1}|S_{i1} = j, \sigma_1, \sigma_2]\,\Pr[S_{i1} = j|\boldsymbol{\pi}], \quad (9)$$

$$T^{(k)}_{jj'}(x_{ik}) = p[x_{ik}|S_{ik} = j']\,\Pr[S_{ik} = j'|S_{i(k-1)} = j, \gamma], \qquad k = 2, \ldots, n_i. \quad (10)$$

Then equation (3) can be expressed as

$$p[\mathbf{x}_i|\sigma_1, \sigma_2, \gamma] = \mathbf{V}^{\mathrm{T}}(x_{i1})(T^{(2)}(x_{i2})T^{(3)}(x_{i3}) \ldots T^{(n)}(x_{in_i}))\mathbf{1}, \quad (11)$$

where **1** is the column vector with all elements equal to 1.

2.5. *Accounting for Exact Times of Diagnosis of Acquired Immune Deficiency Syndrome*

The onset of clinical AIDS is usually followed closely by an AIDS diagnosis; because the time of diagnosis of AIDS was known for all men in the SFMHS who had AIDS diagnosed during the study period and who were not lost to follow-up, these AIDS diagnosis times should be treated as exact transition times. Hence, for individuals with AIDS, equation (4a) should be modified to

$$\mathbf{S}_i \sim \Pr[\mathbf{S}_i = \mathbf{s}_i|\gamma, \boldsymbol{\pi}]$$

$$= \Pr[S_{i1} = s_{i1}|\boldsymbol{\pi}] \prod_{k=2}^{n_i-1} \Pr[S_{ik} = s_{ik}|S_{i(k-1)} = s_{i(k-1)}, \gamma] \Bigg(\sum_{j \leqslant 6} \Pr[S_{in_i} = j|S_{i(n_i-1)}, \gamma]\gamma_{j7} \Bigg). \quad (12)$$

Thus, immediately before the AIDS diagnosis, the true stage is unknown (except that it must not be the AIDS stage), but at the next instant a transition to AIDS occurs; the probability of such a jump is proportional to γ_{j7}, where stage 7 is the absorbing stage corresponding to AIDS. The sum over stages in equation (12) can be extended to all stages by noting that $\gamma_{77} \equiv 0$.

The marginal probability of observing $\mathbf{x}_i$ in this case has the same form as given in equation (11). The sole change is that $T^{(n_i)}_{jj'}(x_{in_i})$ is replaced by

$$T^{(n_i)}_{jj'}(x_{in_i}) = \gamma_{j'7}\,\Pr[S_{in_i} = j'|S_{i(n_i-1)} = j, \gamma].$$

2.6. *Estimating versus Conditioning on Initial Stage*

The marginal distribution of $\mathbf{X}_i$ depends on the distribution of initial stages. Because S_{i1} is not observed, this distribution must be estimated from the observations X_{i1} and the error model. For individuals who are seropositive at entry to the study, we are interested primarily in the parameters γ that govern the transitions and in the error model parameters σ_1 and σ_2 but not in the initial distribution of true stages. Thus, for these individuals it is desirable to develop the marginal distribution of $(X_{i2}, X_{i3}, \ldots, X_{in})$ conditionally on X_{i1}. For X_{i1} only, assume that instead of specifying $g[x_{i1}|y_{i1}, \sigma_1]$ we specify $\tilde{g}[y_{i1}|x_{i1}, \sigma_1, \sigma_2, \pi]$, the probability density function of y_{i1} given x_{i1}; then the conditional probability density function of $X_{i2}, X_{i3}, \ldots, X_{in}$ given X_{i1}, denoted by $p[x_{i2}, x_{i3}, \ldots, x_{in_i}|x_{i1}, \sigma_1, \sigma_2, \gamma, \pi]$, can be expressed as

$$\begin{aligned} p[x_{i2}, x_{i3}, \ldots, x_{in_i}|x_{i1}, \sigma_1, \sigma_2, \gamma, \pi] = & \sum_{s_1} \Pr[S_{i1} = s_1|x_{i1}, \sigma_1, \sigma_2, \pi] \\ & \times \sum_{s_2} p[x_{i2}|S_{i2} = s_2, \sigma_1, \sigma_2] \Pr[S_{i2} = s_2|S_{i1} = s_1, \gamma] \\ & \times \sum_{s_3} \ldots \sum_{s_{n_i}} p[x_{in_i}|S_{in_i}, \sigma_1, \sigma_2] \Pr[S_{in_i}|S_{i(n_i-1)}, \gamma]. \end{aligned} \tag{13}$$

This sum has the same matrix form as equation (11), except that V is replaced by

$$V_j(x_{i1}) = \Pr[S_{i1} = j|x_{i1}, \sigma_1, \sigma_2, \pi]. \tag{14}$$

Note, however, that $\Pr[S_{i1} = j|x_{i1}, \sigma_1, \sigma_2, \pi]$ depends on the parameter π, since this quantity is related to $p[x_{i1}|S_{i1} = j, \sigma_1, \sigma_2]$ and $P[S_{i1}|\pi]$ by Bayes's theorem. To make the conditional likelihood (13) independent of π, we must choose $P[S_{i1}|\pi]$ or $f[y_{i1}|\pi]$ in the same way as a 'non-informative' Bayes prior distribution. This choice corresponds to assigning the first stage on the basis entirely of the first observation, rather than also including information on the likely initial stages in this cohort as expressed by $P[S_{i1}|\pi]$. Although the assumed non-informative $P[S_{i1}|\pi]$ would not provide as good a fit to the observed marginal distribution of the x_{i1}s, the effect is minimized because the conditional likelihood (13) does not attempt to fit the observed values x_{i1}. Hence, for the first observation we assume that

$$\log Y_{i1} = \log X_{i1} + \epsilon_{i1}, \qquad \epsilon_{i1} \sim \text{normal}[0, \sigma_1^2], \tag{15}$$

corresponding to assuming that the unconditional distribution of Y_{i1} is the non-informative Bayesian prior distribution. The stage occupation probabilities conditional on the observed value of X_{i1} needed in equation (13) can be obtained by integrating between stage boundaries.

The distribution of initial stages among seroconverters (i.e. those individuals who test negatively for HIV at entry to the study but subsequently test positively) *is* of interest. For this group, we assume that the time of seroconversion was the midpoint of the time interval between the last negative and first positive test. Ideally, we would estimate the distribution of initial stages at seroconversion by using a six-bin multinomial distribution. However, because only 46 seroconverters were in the study, we modelled this distribution as a log-normal distribution. In practice, this choice is probably sufficient, as the particular log-normal distribution that we obtained assigns non-negligible mass to only the first three stages.

The overall likelihood is a product of contributions from each individual. Specifically, ordering the data so that the first N_c observations are the seroconverters, the likelihood for observing the data is given by

$$\mathcal{L}=\prod_{i=1}^{N_c}\Pr[X_{i1}, X_{i2}, \ldots, X_{in_i}|\sigma_1, \sigma_2, \gamma, \pi]\prod_{i=N_c+1}^{N}\Pr[X_{i2}, X_{i3}, \ldots, X_{in_i}|X_{i1}, \sigma_1, \sigma_2, \gamma].$$

The likelihood above is maximized with respect to the parameters γ, σ_1, σ_2 and π. For continuous time chains, a standard result is that the transition probability matrix $\Pr[S_{k+1}|S_k]$ is given by the (matrix) exponentiation of $\gamma(t_{k+1}-t_k)$, where γ is the infinitesimal generator of the Markov chain (see for example Karlin and Taylor (1975)). Software to carry out this exponentiation must be able to handle the case of repeated eigenvalues; for this case, we used a computer program developed by Allen (1987). We wrote a Fortran program to carry out the maximization by using the implementation of the downhill simplex method found in Press *et al.* (1986). Confidence intervals for parameters were obtained by inverting a numerical approximation to the matrix of second partial derivatives taken with respect to all parameters in the model, calculated using the estimated parameter values.

3. Results and Interpretation

3.1. *Results of Fitting Measurement Error Models*

The profile likelihood for σ_1 (the standard error of the short-term variability and measurement error) was obtained by maximizing $\mathcal{L}$ with respect to γ, σ_2 and π, with σ_1 held fixed. The profile likelihood for the bidirectional model, expressed as a likelihood ratio statistic, is shown in Fig. 3. Note the well-defined maximizing value,

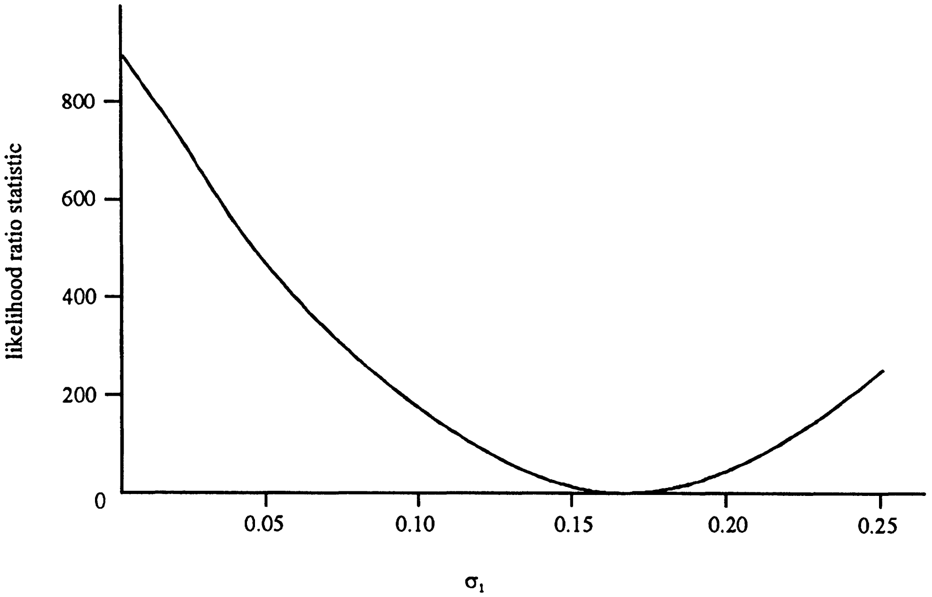

Fig. 3. Profile likelihood for the SFMHS data, as a function of σ_1, the standard error of the measurement error model: the profile likelihood is expressed as a likelihood ratio statistic, i.e. the value of the likelihood at its maximum value has been subtracted and the difference multiplied by −2; the smooth curve shown is a cubic spline interpolation of values of the profile likelihood obtained by varying σ_1 in increments of 0.01, with extra evaluations near the maximum likelihood value

TABLE 3
Elements of the infinitesimal generator for best fitting bidirectional ($\hat{\sigma}_1 = 0.172$, $\hat{\sigma}_2 = 0.231$) and unidirectional ($\hat{\sigma}_1 = 0.219$, $\hat{\sigma}_2 = 0.203$) models†

	Bidirectional model	*Unidirectional model*
γ_{12}	0.0381 (0.0274, 0.0531)	0.0344 (0.0233, 0.0508)
γ_{21}	0.0030 (0.0009, 0.0093)	0
γ_{23}	0.0478 (0.0372, 0.0615)	0.0332 (0.0259, 0.0425)
γ_{32}	0.0087 (0.0044, 0.0170)	0
γ_{34}	0.0399 (0.0334, 0.0478)	0.0314 (0.0262, 0.0378)
γ_{43}	0.0064 (0.0036, 0.0114)	0
γ_{45}	0.0417 (0.0349, 0.0498)	0.0303 (0.0253, 0.0362)
γ_{54}	0.0167 (0.0109, 0.0254)	0
γ_{56}	0.0450 (0.0372, 0.0546)	0.0382 (0.0312, 0.0468)
γ_{65}	0.0071 (0.0034, 0.0148)	0
γ_{37}	0.0016 (0.0006, 0.0042)	0.0015 (0.0005, 0.0044)
γ_{47}	0.0025 (0.0008, 0.0076)	0.0026 (0.0009, 0.0078)
γ_{57}	0.0038 (0.0009, 0.0163)	0.0045 (0.0013, 0.0158)
γ_{67}	0.0647 (0.0540, 0.0775)	0.0686 (0.0569, 0.0827)

†Element γ_{ij} is the rate at which transitions from stage i to stage j occur, in units of months^{-1} (see Section 2.2 for the full form of the infinitesimal generator); 95% confidence intervals are given in parentheses.

which corresponds to a maximum likelihood estimate of $\hat{\sigma}_1 = 0.172$. The maximum likelihood estimates for γ corresponding to the maximum likelihood values of σ_1, σ_2 and π are given in Table 3.

The difference between estimates obtained from our measurement error model and the naïve model, in which no measurement error is assumed, is shown in Table 2. The unacceptable predictions of the naïve model ($\sigma_1 = 0$) regarding the mean time to AIDS (absorption) and the number of stages visited before absorption contrast with those from our best fitting bidirectional model ($\sigma_1 = \hat{\sigma}_1$). Unlike the naïve model, our model confirms that CD4 cell counts are successful in staging HIV disease, with each stage corresponding to approximately 2 years of the incubation period. The number of stages visited is also closer to that expected for a progressive disease such as HIV infection.

To examine the role of measurement error in determining the underlying CD4 trajectories, we show in Fig. 4 the mean number of stages visited before absorption from each of the six CD4-based stages for the bidirectional model. The number of stages visited is a measure of the smoothness of the sample path because it equals one plus the average number of jumps taken before absorption. We also obtained a second measure of the smoothness of sample paths by dividing the mean number of CD4-based stages visited by the mean time to AIDS (absorption) from each of the CD4-based stages, as a function of σ_1. Because this ratio measures the rate at which jumps between stages occur, it is a second measure of the smoothness of the true process $S(t)$. These results, which are not shown, also indicate that increasing σ_1 results in fewer transitions per year. Hence, both the absolute number and the rate of transitions decrease with increasing σ_1, indicating that σ_1 plays the role of a smoothing parameter in our model.

By accounting for measurement error in our model, we can fit a unidirectional Markov chain in which the true CD4 stage decreases monotonically with time,

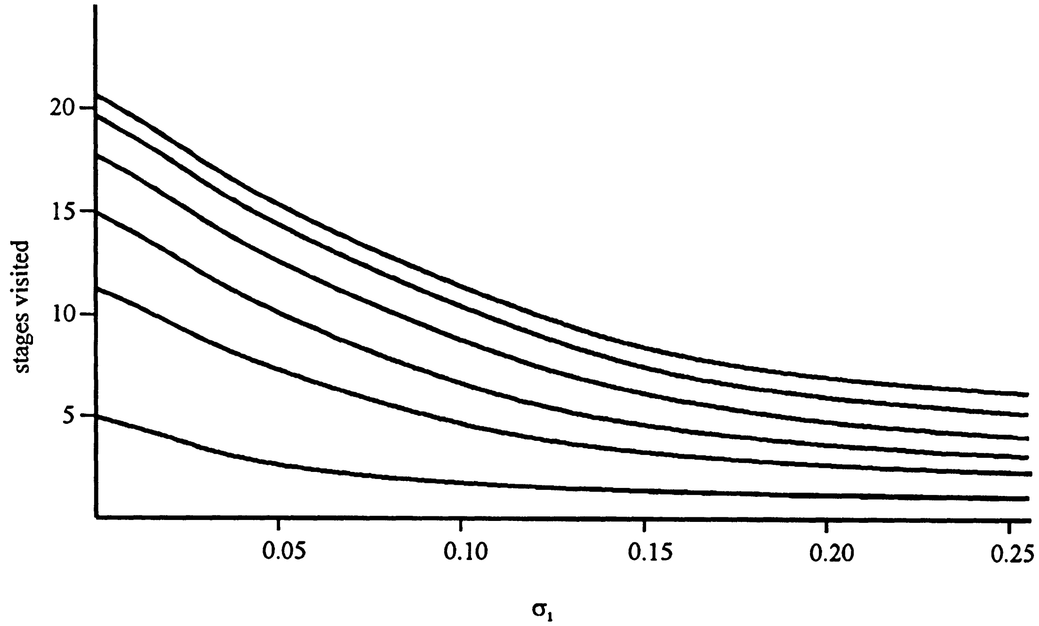

Fig. 4. Expected number of stages visited before absorption, as a function of σ_1, the standard error of the measurement error model: the top curve is calculated conditionally on being in stage 1, the second conditionally on being in stage 2, etc.; each curve is a cubic spline interpolating function of values calculated by varying σ_1 in increments of 0.01

without altering or smoothing the observed data. For the unidirectional model, the maximum likelihood estimate of σ_1 was $\hat{\sigma}_1 = 0.219$, which is substantially larger than the estimate for the bidirectional model. The parameters of the infinitesimal generator for the best-fitting unidirectional model are shown in Table 3, and the mean times to AIDS and mean number of stages visited for this model are shown in Table 2.

Because the unidirectional model is a submodel of the bidirectional model, we can test whether the fit of the bidirectional model is sufficiently better that the unidirectional model must be rejected. The likelihood ratio statistic comparing the best fitting unidirectional model with the best fitting bidirectional model was 209.8 with 5 degrees of freedom; even with the problems associated with inference when parameter values are on boundaries, the bidirectional model provides a significantly better fit.

The observed stage occupation probabilities among the 46 seroconverters at the first visit when they were HIV positive were 0.435, 0.196, 0.217, 0.130, 0.022 and 0.0 in stages 1–6 respectively. In our best fitting bidirectional model, the stage occupation probabilities at the imputed time of seroconversion (the midpoint between the last negative and first positive tests) were 0.356, 0.323, 0.264, 0.054, 0.003 and 0.0, whereas for the unidirectional model they were 0.316, 0.337, 0.288, 0.057, 0.003 and 0.0. These estimates allow inferences to be made about the incubation time from seroconversion to AIDS. For example, the mean time to AIDS is the weighted sum of the mean times to AIDS from each stage shown in Table 2: we obtain 9.9 years for both the bidirectional and the unidirectional models.

TABLE 4
Stage misclassification probabilities corresponding to $\sigma_1 = 0.172$

True stage	*Probabilities for the following observed stages:*					
	1	*2*	*3*	*4*	*5*	*6*
1	0.80	0.19	0.02	0	0	0
2	0.26	0.50	0.24	0.01	0	0
3	0.02	0.19	0.60	0.18	0	0
4	0	0.01	0.20	0.62	0.17	0
5	0	0	0	0.14	0.75	0.10
6	0	0	0	0	0.06	0.94

An intuitive measure of the magnitude of measurement error is a table giving the probability that an observation is in a particular stage j, given that the true stage is i. Table 4 shows this misclassification matrix. Because the error is assumed to be on the log-scale, misclassification is more likely to occur at higher CD4 cell counts (earlier stages).

The maximum likelihood estimate of σ_2, the variance parameter governing the distribution of true CD4 cell counts in stage 1 (see equation (2)) for the bidirectional model, was 0.231. This value corresponds to a median true CD4 cell count in stage 1 of 1052, with 95% of all true CD4 cell counts in stage 1 less than 1416, whereas the median and 95th percentile of the observed CD4 cell counts greater than or equal to 900 are 1055 and 1554 respectively.

The ability to estimate σ_1, the standard error parameter of the measurement error model, depends on a clear separation of the timescales for the 'noise' and the 'signal'. Our models assume that data have been generated by a process that can be broken into two distinct components, each acting on a separate timescale. The noise evolves on a faster timescale than the true value, which is assumed to follow a Markov chain; call this slower timescale the Markov timescale. In writing equations (4a)–(4c), we specifically assume that the timescale for the noise is sufficiently fast on the Markov timescale that the values of noise at different times are uncorrelated. Any residual correlation in the noise must therefore be accounted for by the Markov chain. Thus, the choice of the bidirectional model suggests that a correlation between successive observations persists on the 6-month or year timescale. As a result, a unidirectional model could be fitted if the measurement error model accounted for these correlations. However, the hidden Markov methodology, which allows easy calculation of the likelihood, would not be applicable in this case.

3.2. *Prediction of Future Observations and Fit of Model*

Having fitted our model to the SFMHS data, we sought to determine how well the model fits the data. We chose as a measure of fit the ability of our model to predict an observation X_{ir} at time t_r, conditionally on those observations made before t_r. Thus, we must calculate the distribution

$$\Pr[X_{ir}|X_{ik}, k=1, \ldots, r-1] = \frac{\Pr[X_{ik}, k=1, \ldots, r]}{\Pr[X_{ik}, k=1, \ldots, r-1]}. \tag{16}$$

TABLE 5
Observed and expected stage occupancies by CD4-based stage, and observed and expected AIDS diagnoses, for 9 years of hypothetical follow-up in the SFMHS cohort†

Stage		*Occupancies (percentages) for the following years of follow-up:*									
		0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	Observed	83.0	51.5	40.8	25.7	15.8	16.0	21.0	11.0	6.5	7.0
		(19.3)	(13.7)	(11.9)	(8.5)	(6.0)	(6.9)	(10.6)	(7.0)	(5.7)	(9.1)
	Expected		72.7	43.3	30.5	16.5	12.7	11.5	10.8	7.7	5.0
			(19.3)	(12.6)	(10.1)	(6.2)	(5.5)	(5.8)	(6.8)	(6.7)	(6.5)
2	Observed	97.0	69.5	51.5	38.3	23.7	25.0	23.5	17.5	12.0	9.0
		(22.6)	(18.4)	(15.0)	(12.7)	(9.0)	(10.8)	(11.9)	(11.1)	(10.4)	(11.7)
	Expected		71.4	54.9	43.2	28.4	21.8	20.2	17.1	11.2	7.0
			(18.9)	(16.0)	(14.4)	(10.8)	(9.4)	(10.2)	(10.9)	(9.8)	(9.1)
3	Observed	139.0	110.0	99.3	72.3	67.0	47.5	37.5	36.5	20.3	15.0
		(32.4)	(29.2)	(28.9)	(24.0)	(25.4)	(20.5)	(18.9)	(23.2)	(17.7)	(19.5)
	Expected		97.4	88.6	73.4	56.0	47.0	41.7	33.1	23.9	13.7
			(25.8)	(25.8)	(24.4)	(21.2)	(20.3)	(21.0)	(21.1)	(20.8)	(17.8)
4	Observed	66.0	77.5	83.3	69.7	57.2	56.5	44.5	36.5	25.3	11.5
		(15.4)	(20.6)	(24.2)	(23.1)	(21.7)	(24.4)	(22.5)	(23.2)	(22.0)	(14.9)
	Expected		72.8	79.9	73.7	64.3	58.2	48.5	36.5	28.0	19.1
			(19.3)	(23.2)	(24.5)	(24.4)	(25.1)	(24.5)	(23.3)	(24.3)	(24.8)
5	Observed	32.0	46.5	45.2	65.5	69.5	50.0	37.0	25.5	27.8	23.5
		(7.5)	(12.3)	(13.1)	(21.8)	(26.3)	(21.6)	(18.7)	(16.2)	(24.2)	(30.5)
	Expected		44.3	53.1	54.0	61.8	58.2	43.8	31.8	23.0	19.4
			(11.7)	(15.4)	(17.9)	(23.4)	(25.1)	(22.1)	(20.2)	(20.0)	(25.2)
6	Observed	12.0	22.0	23.8	29.5	30.8	37.0	34.5	30.0	23.0	11.0
		(2.8)	(5.8)	(6.9)	(9.8)	(11.7)	(15.9)	(17.4)	(19.1)	(20.0)	(14.3)
	Expected		18.5	24.1	26.2	37.1	34.0	32.4	27.8	21.2	12.8
			(4.9)	(7.0)	(8.7)	(14.0)	(14.7)	(16.4)	(17.7)	(18.4)	(16.6)
AIDS	Observed	0.0	20	27	27	21	39	35	32	25	6
	Expected		19.4	24.4	27.3	34.5	35.5	33.0	29.7	23.9	21.5

†Expected values were calculated by using the best bidirectional model.

Both the numerator and the denominator of equation (16) can be efficiently calculated by using the results of Section 2, and each can be accumulated as intermediate steps when calculating $\Pr[X_{ik}, k = 1, \ldots, n_i]$.

Although the true stages S follow a Markov process, note that equation (16) does not reduce to $\Pr[X_{ir}|X_{i(r-1)}]$. However, we can show that

$$\Pr[X_{ir} = x|X_{ik}, k = 1, \ldots, r-1] = \sum_{j_r} \sum_{j_{r-1}} \Pr[X_{ir} = x|S_{ir} = j_r] \\ \times \Pr[S_{ir} = j_r|S_{i(r-1)} = j_{r-1}] \Pr[S_{i(r-1)} = j_{r-1}|X_{ik}, k = 1, \ldots, r-1],$$

i.e. that the role of the earlier observations is to improve the estimate (to a distribution) of the true stage at the last observation before t_r. This true stage is then projected to the time t_r when a prediction is to be made, and the measurement error probabilities are applied to the predicted true stage at t_r to obtain the observed value.

Table 5 illustrates how well prior CD4 cell count observations predict future CD4 cell counts and AIDS diagnoses in the SFMHS, when our best bidirectional model is used. Table 5 is constructed as if observations for each individual began at the same time (year 0). For predicting CD4 cell counts, we used equation (16) at each subsequent observation time for each individual to calculate the probability that this observation falls into each of the six CD4-based stages, conditionally on all previous observations. These probabilities were then grouped by year of follow-up to assess the fit as time progressed. If individuals had multiple CD4 observations in a year, the stage occupation probabilities were averaged. These probabilities were then summed over all individuals with observation times in that year; these results are reported as the expected values in Table 5, with percentages in each of the six stages by year given in parentheses. Similarly, the results corresponding to the observed values in Table 5 represent averaged stage occupations for individuals with multiple observations per year.

For example, suppose that an individual had CD4 cell counts of 1000 and 800, 5 and 11 months after the base-line count respectively. This individual would contribute 0.5 to the number observed in both stage 1 and stage 2 in year 1. Similarly, if the probabilities of the individual being in stages 1–6 at 5 months conditionally on the value at the base-line were 0.75, 0.25, 0, 0, 0 and 0 respectively, and if the probabilities of the individual being in stages 1–6 at 11 months conditionally on the values at the base-line and 5 months were 0.25, 0.50, 0.25, 0, 0 and 0 respectively, then the contribution of this individual to the expected numbers in stages 1–6 in year 1 would be 0.5, 0.375, 0.125, 0, 0 and 0 respectively. The agreement between the observed and expected stage occupation probabilities indicates that our model can predict progression of CD4 cell counts among men in the SFMHS, at least at the population level. In particular, there is little evidence of a systematic trend over the 9-year period shown in the parameters governing the Markov chain. Some lack of fit in year 1 presumably is accounted for by our use of the conditional model of Section 2.5.

To assess our model's success in predicting AIDS diagnoses among individuals at risk, we compared the observed and expected number of AIDS diagnoses in each year. The expected number of AIDS diagnoses was calculated conditionally only on AIDS diagnosis and CD4 information available at the beginning of the year, by using the expression

$$\sum_{i=1}^{n} \Pr[X_i(t+1) = \text{AIDS}|X_{ik}, k = 1, \ldots, r_i(t)] \\ - \Pr[X_i(t) = \text{AIDS}|X_{ik}, k = 1, \ldots, r_i(t)], \qquad t = 0, \ldots, 8,$$

where $r_i(t)$ is the index of the last wave that has occurred by time t for the ith individual, and t is measured in years. The agreement between the observed and expected diagnoses is shown in the last two lines of Table 5; the overestimation of AIDS cases in the last year may be explained by delays in the reporting of AIDS cases.

3.3. *Effect of Treatment with Zidovudine and/or Pentamidine*

To determine the effect of treatment (zidovudine and/or pentamidine) on disease progression, we fit several models in which the infinitesimal generator for the underlying Markov chain, γ, was allowed to depend on treatment. We assumed that, in the presence of treatment, the rate of transitions from stage j to j' would be $\lambda_{jj'}\gamma_{jj'}$. The absence of a treatment effect on transitions from stages j to j' corresponds to $\lambda_{jj'} = 1$, whereas a value of $\lambda_{jj'} > 1$ or $\lambda_{jj'} < 1$ would correspond respectively to treatment accelerating or decelerating transitions between stages j and j'. In earlier analysis of the SFMHS, Longini *et al.* (1993) used a model that was similar to our unidirectional model, but allowed transitions to AIDS only from stage 6, and used data that had been 'smoothed' into a monotonic sequence of stages by application of a persistence criterion. This analysis showed that treatment as defined above slowed progression to AIDS.

For both bidirectional and unidirectional models, we performed two analyses: the first was an intent-to-treat analysis that applied a treatment effect to all subsequent transitions after the first reported use of treatment, and the second applied a treatment effect only for the first transition at which treatment was reported. The intent-to-treat analysis is appropriate if treatment has a lingering effect, whereas the first-use analysis detects a transient treatment effect. Since treatment information for transitions to AIDS must be imputed, we do not report results from a third analysis in which a treatment effect is applied only to those transitions during which treatment use was reported. In all analyses, possible treatment effects were restricted to those elements of γ governing transitions between stages characterized by a CD4 cell count below 500 or by AIDS. In addition, because treatment status immediately before an AIDS diagnosis is not known, the effect of treatment on transitions to AIDS cannot be estimated in the first-use analysis.

In the intent-to-treat analysis, the only treatment effect parameter whose confidence interval did not include 1 was that for transitions from stage 6 to AIDS. In addition, no other treatment effect parameter was significantly different from 1 when tested in a model in which it was the only treatment effect. When we refitted our models and allowed only a treatment effect from stage 6 to AIDS we found $\hat{\lambda}_{67} = 0.44$ (95% confidence interval 0.31–0.61) for the bidirectional model and $\hat{\lambda}_{67} = 0.42$ (95% confidence interval 0.29–0.59) for the unidirectional model, corresponding to a protective effect of treatment on the rate of progression to AIDS. These results agree with those of Longini *et al.* (1993), who estimated that treatment with zidovudine and/or pentamidine decelerated progression to AIDS by a factor of 0.38 (95% confidence interval 0.27–0.53) for individuals with a CD4 cell count less than 200, although Longini *et al.* (1993) also found a weaker but still significant deceleration of transitions from stage 5 to stage 6 due to treatment. No significant treatment effects were found in the first-use analyses.

4. Discussion

Markov models of the natural history of HIV, especially those based on CD4 cell counts, play a central role in AIDS modelling. They have been used to describe the natural history of HIV infection (Longini *et al.*, 1989, 1991; Lawless and Yan, 1991; Frydman, 1992; Gentleman *et al.*, 1994), to evaluate the effect of covariates such as therapy on stage-specific progression rates (Graham *et al.*, 1991; Longini *et al.*, 1993), to predict the stage-specific course of the HIV epidemic in the USA (Brookmeyer, 1991; Longini *et al.*, 1992; Centers for Disease Control and Prevention, 1992b) and to estimate HIV incidence from cross-sectional surveys (Satten and Longini, 1994). In conjunction with epidemic models, Markov models have been used to provide estimates of transmission probabilities (Longini *et al.*, 1989; Jacquez *et al.*, 1994) and to investigate the dynamics of the HIV epidemic (Hethcote *et al.*, 1991; Jacquez *et al.*, 1988).

Fitting a Markov chain model to observed data is easily accomplished when the observed data allow the state of the system at the time of observation to be unambiguously determined. However, if the observed data are subject to measurement or misclassification error or short timescale noise, then the observed sample paths will exhibit more variability than if the state of the system is known with certainty. As a result, fitting a Markov chain model to noisy data will result in estimated transition probabilities that are too large. An additional problem may arise in chains with an absorbing state, if the data contain many observations that are censored before absorption. In this case, the mean time to absorption may be incorrectly estimated.

Simply smoothing each individual's longitudinal data may not be an acceptable solution because the smoothing mechanism generally will not have a Markov correlation structure, and because smoothing many short series of observations may be problematical. We have shown how a likelihood-based approach can account for measurement error in a completely self-consistent way. Parameters corresponding to smoothing parameters can be identified and estimated by maximum likelihood. In fitting CD4 cell count data from the SFMHS, we found dramatic differences between the behaviour of the Markov chain obtained by ignoring measurement error and that obtained by accounting for measurement error.

Our methodology also addresses some conceptual difficulties in applying Markov chains to continuous-valued data. First, when an observation falls at or near a stage boundary, assigning it to a single stage, as is usually required, may seem unsatisfactory. In our methodology, every stage contributes to the likelihood of each observation in proportion to the probability that the observation could have occurred given that stage. Second, when measurement error is ignored, small fluctuations in observations that are far from stage boundaries do not correspond to transitions, whereas fluctuations of the same magnitude near a stage boundary would be counted as transitions. By considering these short timescale fluctuations as part of the noise, our methodology automatically discounts transitions caused by small fluctuations in values near a stage boundary.

Although data values are never smoothed in our model, we have shown that σ_1, the standard error in the measurement error model, plays the role of a smoothing parameter in that increasing σ_1 results in smoother sample paths for the true process. In data from the SFMHS, these smoother sample paths corresponded to a model for HIV progression that better reflected clinical intuition than did the results obtained by naïvely fitting the observed CD4 cell counts. We hope that the methods presented here will be integrated into on-going HIV and AIDS research.

Acknowledgements

The authors acknowledge Dr David Allen for use of his matrix exponentiation program. We also thank Dr Warren Winkelstein and Dr James Wiley for providing access to the SFMHS data, and Dr W. Scott Clark for data management support.

References

Albert, P. S. (1991) A two-state Markov model for a time series of epileptic seizure counts. *Biometrics*, **47**, 1371–1381.

Allen, D. M. (1987) Computation for compartmental models. In *Computer Science and Statistics: Proc. 19th Symp. Interface* (ed. R. M. Heiberger). Washington DC: American Statistical Association.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Statist.*, **41**, 164–171.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.

Brookmeyer, R. (1991) Reconstruction and future trends of the AIDS epidemic in the United States. *Science*, **253**, 37–42.

Centers for Disease Control and Prevention (1992a) 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Morb. Mort. Wkly Rep.*, **41** no. RR-17, 1–19.

———(1992b) AIDS case projection and estimates of HIV-infected immunosuppressed persons in the United States, 1992–1994. *Morb. Mort. Wkly Rep.*, **41**, no. RR-18, 1–27.

Collins, L. M. and Wugalter, S. E. (1992) Latent class models for stage-sequential dynamic latent variables. *Multivar. Behav. Res.*, **27**, 131–157.

DeGruttola, V., Lange, N. and Dafni, U. (1991) Modeling the progression of HIV infection. *J. Am. Statist. Ass.*, **86**, 569–577.

DeGruttola, V., Wulfsohn, M., Fischl, M. A. and Tsiatis, A. (1993) Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *J. AIDS*, **6**, 359–365.

El-Sadr, W., Oleske, J. M., Agins, B. D., Bauman, K. A., Brosgart, C. L., Brown, G. M., Geaga, J. V., Greenspan, D., Hein, K., Holzemer, W. L., Jackson, R. E., Lindsay, M. K., Makadon, H. J., Moon, M. W., Rappoport, C. A., Scott, G., Shervington, W. W., Shulman, L. C. and Wofsy, C. B. (1994) *Evaluation and Management of Early HIV Infection: Clinical Practice Guideline No. 7.* Rockville: Agency for Health Care Policy and Research.

Fredkin, D. R. and Rice, J. A. (1992a) Bayesian restoration of single-channel patch clamp recordings. *Biometrics*, **48**, 427–448.

———(1992b) Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. R. Soc. Lond.* B, **249**, 125–132.

Frydman, H. (1992) A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to Aids. *J. R. Statist. Soc.* B, **54**, 853–866.

Gauvreau, K., DeGruttola, V., Pagano, M. and Bellocco, R. (1994) The effect of covariates on the interval between infection with HIV and the development of AIDS. *Statist. Med.*, **13**, 2021–2030.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C. and Yan, P. (1994) Multi-stage Markov models for analyzing incomplete disease history data with illustrations for HIV disease. *Statist. Med.*, **13**, 805–821.

Graham, N. M. H., Zeger, S. L., Park, L. P., Phair, J. P., Detels, R., Vermund, S. H., Ho, M., Saah, A. J. and Multicenter AIDS Cohort Study (1991) Effects of zidovudine and *pneumocystis carinii* pneumonia prophylaxis on progression of HIV-1 infection to AIDS. *Lancet*, **338**, 265–269.

Hessol, N. A., Koblin, B. A., van Griensven, G. J. P., Bacchetti, P., Liu, J. Y., Stevens, C. E., Coutinho, R. A., Buchbinder, S. P. and Katz, M. H. (1994) Progression of HIV-1 infection among homosexual men in hepatitis B vaccine trial cohorts in Amsterdam, New York City, and San Francisco. *Am. J. Epidem.*, **139**, 1077–1087.

Hethcote, H. W., Van Ark, J. W. and Longini, I. M. (1991) A simulation model of AIDS in San Francisco: I, Model formulation and parameter estimation. *Math. Biosci.*, **106**, 203–222.

Hoover, D. R., Graham, N. M. H., Chen, B., Taylor, J. M. G., Phair, J., Zhou, S. Y. J. and Muñoz, A. (1992) Effect of CD4+ cell count measurement variability on staging HIV-1 infection. *J. AIDS*, **5**, 794–802.

Hoover, D. R., Muñoz, A., He, Y., Taylor, J. M. G., Kingsley, L., Chmiel, J. and Saah, A. (1994) Estimating effectiveness of self-selected interventions on population AIDS incubation. *Statist. Med.*, **13**, 2127–2139.

Jacquez, J. A., Koopman, J. S., Simon, C. P. and Longini, I. M. (1994) The role of primary infection in the epidemics of HIV infection in gay cohorts. *J. AIDS*, **7**, 1169–1184.

Jacquez, J. A., Simon, C. P. and Koopman, J. S. (1988) Modeling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.*, **92**, 119–199.

Jewell, N. P. and Nielsen, J. P. (1993) A framework for consistent prediction rules based on markers. *Biometrika*, **80**, 153–164.

Karlin, S. and Taylor, H. M. (1975) *A First Course in Stochastic Processes*, 2nd edn. San Diego: Academic Press.

Langeheine, R. (1994) Latent variables Markov models. In *Latent Variables Analysis* (eds A. von Eye and C. C. Clogg), pp. 373–395. Thousand Oaks: Sage.

Lawless, J. F. and Yan, P. (1991) Some statistical methods for followup studies of disease with intermittent monitoring. In *Multiple Comparisons, Selection, and Applications in Biometry* (ed. F. M. Hoppe), pp. 427–446. New York: Dekker.

Le, N. D., Leroux, B. G. and Puterman, M. (1992) Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, **48**, 317–323.

Longini, I. M., Byers, R. H., Hessol, N. A. and Tan, W. Y. (1992) Estimating the stage-specific numbers of HIV infection using a Markov model and back-calculation. *Statist. Med.*, **11**, 831–843.

Longini, I. M., Clark, W. S., Gardner, L. I. and Brundage, J. F. (1991) The dynamics of CD4 + T-lymphocyte decline in HIV-infected individuals: a Markov modeling approach. *J. AIDS*, **4**, 1141–1147.

Longini, I. M., Clark, W. S., Haber, M. and Horsburgh, R. (1989) The stages of HIV infection: waiting times and infection transmission probabilities. *Lect. Notes Biomath.*, **83**, 112–137.

Longini, I. M., Clark, W. S. and Karon, J. (1993) The effect of routine use of therapy on the clinical course of HIV infection in a population-based cohort. *Am. J. Epidem.*, **137**, 1229–1240.

Multi-cohort Analysis Project (1993) Markers as time-dependent covariates in relative risk regression. *Statist. Med.*, **12**, 2087–2098.

Pawitan, Y. and Self, S. (1993) Modeling disease marker processes in AIDS. *J. Am. Statist. Ass.*, **88**, 719–726.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1986) *Numerical Recipes: the Art of Scientific Computing*. Cambridge: Cambridge University Press.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Satten, G. A. and Longini, I. M. (1994) Estimation of incidence of HIV infection using cross-sectional marker surveys. *Biometrics*, **50**, 675–688.

Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994) A stochastic model for analysis of longitudinal AIDS data. *J. Am. Statist. Ass.*, **89**, 727–736.

Winkelstein, Jr, W., Lyman, D. M., Padian, N., Grant, R., Samuel, M., Wiley, J. A., Anderson, R. E., Lang, W., Riggs, J. and Levy, J. A. (1987) Sexual practices and risk of infection by the human immunodeficiency virus. *J. Am. Med. Ass.*, **257**, 321–325.

Discussion of the Paper by Satten and Longini

Caroline A. Sabin (Royal Free Hospital School of Medicine, London): I feel honoured to be in a position to propose the vote of thanks to the speaker for presenting this excellent and stimulating paper, and especially to thank him for coming such a distance! As there are many in the audience who I am sure will wish to discuss the statistical methods involved, I would like to focus my comments on the clinical application of the results.

The issue of variability in the measurement of CD4 counts is not new. Many researchers have already described variability in the CD4 count and the effect that this may have on the clinical interpretation of results of studies involving the CD4 count. Researchers into acquired immune deficiency syndrome (AIDS) are, in general, very aware of this variability. Indeed, clinicians themselves rarely make clinical decisions on the basis of a single CD4 count, preferring to wait for a confirmatory measurement. I believe that this awareness is partly due to a general acceptance of the important role that statistics plays within the field of research on the human immunodeficiency virus (HIV) and also partly due to the publication of some well-written, well-cited and clear papers which discuss the issue of variability in specialist AIDS journals. However, there is a tendency for researchers (and by that term I include statisticians and epidemiologists) either to ignore the issue of variability, perhaps making a brief mention of it in the discussion section of their papers, or to refuse to use the CD4 count at all on the basis that if it is variable then it can be of no use. This, I feel, results from a general misconception about the effect of variability and a lack of knowledge about how findings from these statistical papers can be translated into and applied in general patient care in clinical practice.

There is no doubt that the CD4 count is currently the most useful (and best tested) marker of disease stage available to clinicians for assessing the prognosis of patients. This prognostic value has been shown despite the presence of variability in the count. However, there are always individuals who do not fit the usual mould—either by developing AIDS at relatively high CD4 counts or by remaining AIDS free for long periods of time despite having low or almost zero CD4 counts. This can partly be explained by the fact that by measuring CD4 cells we may simply be measuring a consequence of the infection, rather than directly measuring virus levels themselves. It may also be partly due to the fact that the development of AIDS is itself almost a random process, depending much on the chance that an HIV-infected individual encounters a particular infectious pathogen. However, some of these discrepancies may also be explained by the fact that the measured CD4 count is quite a variable marker. When we measure the CD4 count we may not be measuring the true state of the person's immune system or even how well that immune system is functioning. It is possible now to measure HIV ribonucleic acid (RNA) levels directly, but the cost of doing this is prohibitively expensive, so for the time being at least it is unlikely that we shall move across to RNA levels for routine monitoring and the CD4 count will remain the measurement of choice in the near future. Consequently, papers, such as this by Satten and Longini, are invaluable for describing the effect of variability and for introducing statistical methods which take account of it.

So, how can the findings of this paper be applied in the clinical setting? The authors themselves say in the last sentence of the paper 'We hope that the methods presented here will be integrated into on-going HIV and AIDS research' but shy away from suggesting ways in which the results may be applied. I would like to suggest a couple of areas in which it is thought that variability is an issue, in the hope that these may be addressed in the future.

One area in which variability in the CD4 count may have important implications is in the field of clinical trials for anti-HIV therapies. As the incubation period of HIV can be very long, trials of new anti-HIV therapies tend to follow patients for long periods of time. To shorten the time period of these trials, researchers usually focus on patients who are perhaps in more advanced stages of the disease, and this is often achieved by having an entry criterion which is based on CD4 counts. For example, trials may only recruit patients with CD4 counts less than 200 cells mm^{-3}. As has already been suggested by Michael Hughes at the International Society of Clinical Biostatistics meeting in Barcelona in 1995, trial results may therefore be driven by a small number of clinical events in individuals with very low underlying CD4 counts, even though their observed CD4 counts are relatively high. It might be of interest therefore to be able to restrict entry to a clinical trial to a more homogeneous group of patients with a well-defined range of underlying CD4 counts, so that results can be more easily interpreted.

A second area in which variability in the CD4 count can have undesirable results is when studying the pathogenesis of HIV. The role of the CD4 count in HIV pathogenesis has largely been established through biological research and also by estimating its prognostic role for the development of AIDS. Other potential laboratory markers are compared with this standard marker. If the potential laboratory marker provides additional prognostic information to that which is provided by the CD4 count, then it is heralded as a new and important marker. Ultimately, somebody somewhere will conjure up a suitable hypothesis which provides the marker with a pathogenic role. However, this 'independent' prognostic ability may simply be a result of measuring the CD4 count imperfectly. If the two markers are correlated then this may be a case of residual confounding by the CD4 count.

So the issue of measurement variability in the paper presented here is very important. This is the first time, I believe, that variability in the count has been directly incorporated into a model to assess the incubation period of HIV infection. The results are consistent with many other studies which have also tried to assess the incubation period and found it to be around 9–12 years—even more in some cases. One of the good aspects is that it is a bidirectional model. In the past, some researchers have had concerns about a unidirectional model, feeling that perhaps it is not very realistic—especially now perhaps with the availability of combination therapy which may have a long-lasting positive effect on the CD4 count. However, I still have some doubts that the results can be easily extended to other areas of HIV research—after all, estimating the incubation period is only a fairly small part of that. Satten and Longini's paper suggests that, given a patient's CD4 history, it should be possible to predict what his CD4 count will be in the future. But how do we do this? Perhaps clinicians should be given a simple guide on how to incorporate this.

If I have one further concern with many of the papers which discuss the issue of variability in CD4 counts, it is that we tend to make the assumption that the underlying CD4 count tends to follow some smooth curve and that any deviations from this represent random 'noise' which is not important.

However, I do not feel that we are ready to state this yet. Our knowledge of HIV pathogenesis, although increasing at a rapid rate, is still relatively patchy. It may be that these deviations from a smooth curve represent times at which the patient is genuinely at increased risk, even if only temporarily. If so, by smoothing them out we may lose important information.

So, to conclude, I would like to repeat my thanks to the authors for presenting this interesting paper. I believe that the results have provided a useful example of how statistical methods can be incorporated into epidemiological research, and my hope for the future is that this will be incorporated into more areas and that we can have more collaboration.

Deborah Ashby (University of Liverpool): There are two fundamental aspects to the model proposed for these data: the Markov chain assumption and the specification and estimation of the measurement error. I shall voice reservations about each of these aspects, and then I shall make some comments about how they interact with each other.

From a medical perspective, the categorization needed for the Markov chain modelling has echoes of the pragmatic need to dichotomize. For CD4 count, there is a critical difference between operational guidelines for management of patients, which focus on levels such as counts of 500 and 200, and fundamental understanding of the process where such boundaries are almost irrelevant.

From a statistical perspective, it is useful to distinguish the data display and exploration phase of analysis from the formal modelling and inference. Table 1 is very effective at displaying the data, and in helping to suggest potential models, but, to me, the most natural way of formally modelling a continuous measurement is as a continuous variable, not by categorizing it arbitrarily. Fig. 2, for example, is implausible as the conditional distribution of the true CD4 count at a given time point. As the interval width tends to 0 the graph should become smoother, until it becomes a continuous function. I only find a Markov model an intuitive model of the process in so far as it approximates this ideal specification. However, the question is to what extent it nevertheless serves as a useful modelling device, giving sensible answers to sensible questions.

The measurement error for CD4 count certainly needs to be taken seriously: this analysis gives an estimate on the log-scale of 0.172, which translates to a multiplicative error of about 20%. In contrast, the Hoover analysis gives an estimated error of about 30%. I have reservations about where the information comes from, as well as how it is modelled.

An estimate from any statistical analysis depends on assumptions or data, or the two in combination. In general, there are three possible sources of information on measurement error, these being an external validation study, an internal comparison of a subset of the observation against a standard comparator, or internal replication of at least a subset of the measurements. In work by my doctoral student, Seyed Mehri, on psychiatric data, such estimates come from a validation of the scoring instruments against an experienced psychiatrist's diagnosis. In work by another doctoral student, Mark Hirst, on survival from cancer, such estimates will come from a detailed consideration of the cancer registration process, and from special audit studies involving double abstraction of case notes. In the Hoover study, cited here, some of the CD4 counts were replicated.

In Satten and Longini's study, there must surely be scope for talking to people in the laboratory. We could use the information that the CD4 count is calculated from the white cell count and from an estimated proportion that are CD4 cells, although this will complicate the distributional assumptions. Alternatively, we could obtain estimates of measurement error directly from repeat measurement studies, or even use estimates from Hoover *et al.* (1992). The assumption of independent errors also needs verification: apart from begging the question of residual within-patient variation, laboratory measurements can be subject to long-term drift (Broughton *et al.*, 1986).

In the absence of additional information, we are left resting entirely on modelling assumptions. As far as I can see, the estimation of measurement error comes entirely from the assumption that the observed transitions in Table 1 that are more than one cell different are highly likely to be due to measurement error. The formal model is specified through the infinitesimal generator γ, which is then estimated from data observed at approximately 6-monthly intervals.

This is where my concerns about the Markov assumption, and in particular the arbitrary specification of the categories, return to haunt me. I wonder to what extent the estimated measurement error is determined by the choice of the number of categories relative to the typical observation interval. The categories are fairly broad: in passing, I note that the classes widths are of the same kind of order of magnitude as the measurement error that we are trying to retrieve. I wonder whether a finer categorization might give a larger estimated measurement error, and conversely whether a cruder categorization might yield a smaller estimate.

The authors give plausible answers on the average time from seroconversion to the development of acquired immune deficiency syndrome and on how this average time varies according to the initial CD4 count. They also make some useful observations on the effect of treatment. These conclusions would be more secure if the authors reported sensitivity analyses which looked at the dependence of the measurement error estimate on the number of categories used, and the timing of observations, and more importantly the robustness of the main conclusions to varying the assumptions on the measurement error.

Despite my criticisms, the analysis addresses real, pertinent questions, using data that are clearly well understood by the authors. To this extent the paper is a serious example of applied statistics, and it gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Sylvia Richardson (Institut National de la Santé et de la Recherche Médicale, Paris) **and C. Guihenneuc-Jouyaux** (Université Paris V): We would like to congratulate the authors warmly on their interesting paper and to concur with them that taking into account measurement error and short-term variability when analysing marker data to characterize the underlying evolution of a progressive disease is of prime importance. Motivated by the same application and in collaboration with the authors, we have developed a Bayesian framework for estimating the transition rates of a hidden Markov process.

The model can be represented by the directed acyclic graph in Fig. 5 where ij denotes the jth follow-up of individual i. To define the 'measurement error' model, we relate directly the observed CD4 counts X_{ij} to the true stage S_{ij} on the log-scale by

$$P(\log X_{ij}|S_{ij} = k, \sigma) \sim N(\mu_k, \sigma^2) \qquad k = 1, \ldots, 6,$$

where the set of μ_ks is chosen to span the range of CD4 counts and σ (possibly also indexed by k) is to be estimated.

Thus we do not introduce the notion of a 'true CD4 cell count', which calls for continuous stochastic process modelling, but base our discretization on the μ_ks. Similarly to Satten and Longini, the number of stages and the values of the μ_ks can be inspired by clinical knowledge—for example $\{\mu_k, k = 2, \ldots, 5\}$ could be taken to correspond to the class centre of the intervals (on a log-scale) defined by Satten and Longini, with suitable modifications for the first and last stage. Alternatively the discretization can be viewed as a semiparametric approximation. Note that σ^2 encompasses the variability modelled by σ_1^2 and σ_2^2 by the authors.

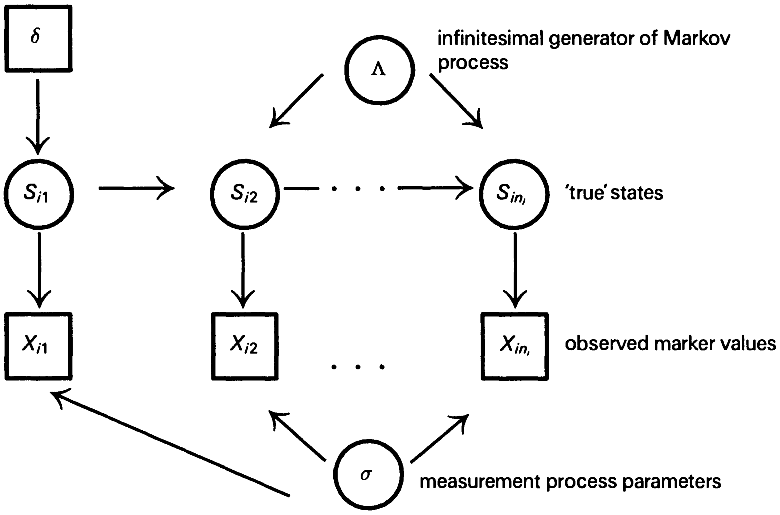

Fig. 5. Directed acyclic graph corresponding to the hidden Markov model

The full probabilistic structure of the model is completed by assuming a uniform prior distribution for the transition rates γ_{ij} on bounded intervals, weakly informative gamma distributions for the precision $(\sigma^2)^{-1}$ and a first-stage prior distribution.

We have implemented this model by using Markov chain Monte Carlo algorithms for simulating the joint posterior distribution of all the parameters. Sampling from the full conditional distribution is straightforward for all parameters except the transition rates where we used a Metropolis step.

By inferring from the joint posterior distribution of all parameters, fluctuations of measurement process parameters are coherently propagated on the estimation of the underlying transition rates, which is not the case when using profile likelihoods for fitting the measurement error parameters. The model is flexible and in particular allows σ^2 to differ between subgroups of patients and/or different stages. It can be extended to investigate treatment effect or the effect of other covariates.

Further developments will include a hierarchical model on the location of the μs as well as a prior model on the number of stages, to relax some of the arbitrariness of the discretization.

Frank Hansford-Miller (Canterbury): My series on the medical history of Western Australia from colonization in 1829 until 1870 (Hansford-Miller, 1996) is relevant. 1829 was when cholera migrated from India, through Russia, to hit Sunderland, in north-east England, in October 1831, then to become a deadly scourge for mankind.

There is a great similarity in the last century's march of cholera to the modern spread of acquired immune deficiency syndrome (AIDS) whether or not due to human immunodeficiency virus type 1, 'the virus that causes AIDS', as the authors tell us, although some dispute this assertion. In Africa, or wherever the disease originated, a similar situation in world communications has developed to enable it to spread. With cholera in the 1830s it was the growth of shipping, the colonization of Australia and other distant places, the expansion of world trade and the industrial revolution that dispersed mankind across the face of the earth. Today it is the worldwide expansion of air transport that has done the same thing on an even more massive scale, and instituted a new flux of human beings throughout the world. And just as we then had the migration of cholera so now we have the spread of AIDS as a new disease to be faced by humanity everywhere.

The discussion here highlights our ignorance, but this, also, echoes the 19th century until the bacteriological theory of disease causation became accepted following the work of Pasteur, Koch, Lister and others. Medicine was then in a similar state of ignorance, bacteria had not been discovered, and the cause of diseases such as cholera was unknown. There were terrible arguments—was it miasma, effluvia, climate?—no one knew. The medical profession was heated and divided until science came to the rescue with the likes of Pasteur. It is the same today with AIDS, and the arguments that we have, and the conflicting theories, show this. It is viruses now not bacteria that are the problem, and this is surely where the next breakthrough must come.

Cholera and other diseases such as smallpox, whooping-cough, measles, scarlatina, typhus and plague were then sought to be contained by the ancient method of quarantine. All ports had their lazarettos. When AIDS burst on to the scene I was surprised that quarantine was not seriously considered, and I wrote a book about the problem (Hansford-Miller, 1994). I spoke about this on Sky television recently.

J. F. Lawless (University of Waterloo): The use of staged continuous time Markov processes to model marker processes and a terminal event is attractive, especially when individuals are observed intermittently and may be in various disease stages when first seen. Satten and Longini's paper is a welcome advance, because biological markers are, as they note, subject to substantial short-term fluctuations and to measurement error and because the 'naïve' application of Markov models has some shortcomings. I have several comments.

(a) It should be stressed that the Markov chain models considered are time homogeneous; this, along with the Markov assumption, allows the time of infection to be ignored for subjects that are seropositive on entry to the study. Frequently time homogeneity is too strong an assumption; Kalbfleisch and Lawless (1989) and Gentleman *et al.* (1994) discuss diagnostic tests.

(b) The naïve model and model (4a)–(4c) are based on different data. The former involves only the observed stages whereas the latter involves the observed (i.e. measured) CD4 counts X_{ik}. The model for the observed stages (i.e. those based on the X_{ik}) that arises from model (4a)–(4c) is not first order Markov; the failing of the naïve model is thus a failing of the first-order time homogeneous Markov model for the stages defined in the paper.

(c) Model (4a)–(4c) may be more succinctly described as a hidden Markov model for which the distribution of log X_{ik}, given that $S_{ik}=j$, is a convolution of two independent random variables, one $N(0, \sigma_1^2)$ and one log-uniform(l_j, u_j). Presumably other distributions, e.g. log $X_{ik} \sim N(a_j, \sigma^2)$, with $a_j=(l_j u_j)^{1/2}$, would also work well. Have the authors tried other models?

(d) It is tempting to note that $\hat{\sigma}_1$s for the second and third models in Table 2 are close to the estimated standard error from the CD4 measurement study mentioned in Section 2.3. (The standard error there should apparently be 0.20 instead of 0.26, if the component standard errors are 0.075 and 0.185.) Incidentally, is Fig. 3 correct? It implies very narrow confidence limits for σ_1 and portrays σ_1-values far outside any plausible range. It should also be noted that, contrary to a remark in Section 3.1, the case $\sigma_1=0$ does not yield the naïve model.

(e) Although models of the type fitted in the paper provided a reasonable fit to mean times to AIDS and observed CD4 transition counts, as illustrated in Tables 2 and 5, measured individual CD4 sample paths generated by such models do not closely resemble those observed in patients, and sample paths for the 'true' CD4 count are not especially plausible. The authors note that their objective is not to model individual marker paths, but this nevertheless mutes the appeal of the models somewhat.

Ping Yan (Health Canada, Ottawa): The attractiveness of modelling disease history by a multistate process through laboratory markers (quantitative) jointly with clinical diagnosis (qualitative) is to utilize the *first passage time distribution* for predicting the onset of clinical symptoms in the light of marker values. It seems weak to use staged CD4 models for predicting the distribution of future CD4 counts and estimating the 'true' course of the marker, since they depend on how well one can model the markers' sample paths. Continuous state models might be preferable.

The ability to estimate σ_1 in terms of separation of the timescales for the 'noise' and the 'signal' deserves more elaboration. Intuitively, the 'noisy' data X_i might not provide much information for σ_1 and the likelihood surface would have been 'flat' near $\hat{\sigma}_1$, unlike in Fig. 3. Perhaps the assumption $y_{ik}|S_{ik}=j \sim U[l_j, u_j)$ also plays a role in enhancing this ability. Incidentally, the confidence limits for $\hat{\gamma}_{ij}$ in Table 3 are calculated at the point estimates $(\hat{\sigma}_1, \hat{\sigma}_2)$. Should not the variance of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ be taken into account?

We used a two-state progressive Markov model (with a single known parameter), a set of computer-generated true markers and 100 simulated samples of noisy data to mimic the situation in the paper. Each of the samples has 100 individuals. Conditional distributions for the true markers given the states are defined following equations (2) and (1). The error distribution to generate the noisy data was the same as that in expression (3). We used $\sigma_1=0.219$ and $\sigma_2=0.203$ from the paper as estimated from the one-directional model. The purpose is to investigate the behaviour of the estimates in repeated samples on the basis of the naïve model and on the proposed models with measurement error.

(a) With the naïve model, we obtained 100 point estimates from simulated samples scattered over a wide range. The 95% confidence intervals based on likelihood ratio statistics provide a coverage rate of about 85%. Those which failed to cover the value of the true parameter were biased towards a direction in agreement with 'fitting a Markov chain model to noisy data will result in estimated transition probabilities that are too large'. This might be valid in general circumstances, not only for the San Francisco Men's Health Study cohort data.

(b) Using model (4a)–(4c) to reanalyse the simulated data yielded point estimates more concentrated around the true parameter. The confidence limits, taking into account the uncertainties introduced by measurement errors, provided a coverage rate of more than 95%. These are desirable and promising features that one would hope from model (4a)–(4c). The models might be useful for analysing longitudinal data with similar measurement errors. However, detailed studies on statistical properties of these estimates are needed.

P. V. Bertrand (University of Birmingham):

Derivation of the transition probabilities and proof that they are not independent

I would like to describe how the matrix of transition probabilities can be derived from the known measurement errors of CD4 counts and from their prior distribution in the study population. Then I prove that the sequence of transitions between the stages of the disease is not independent of each transition.

Derivation of the transition probabilities

Most measurements of medical substances have a coefficient of variation of between 10% and 20%, i.e. a patient with a true CD4 count of 600 would, owing to measurement errors, record as at stage 2 in about 10% of times, or as at stage 3 in 80% of times or as at stage 4 in 10% of times. More precise estimates could be made given full knowledge of measurement errors and of the prior distribution of true values in the population.

From such knowledge the matrix of transition probabilities can easily be deduced in the steady state. In the non-steady state the additional knowledge of the trend over time of the true values of CD4 counts in the study population is required. This can be inferred from the observed values in the treated population and from the known measurement variation of CD4 counts.

The sequence of transitions is not independent

Whereas the sequence of CD4 values is likely to be independent of each value the sequence of transitions is not.

Example

Suppose that at a certain point in time the random variation in the reported CD4 value produces an extreme result so that a misclassification error occurs. Then the next observation on that patient is likely to be closer to its true value. So an erroneous transition in one direction at one time point is likely to be followed by a transition in the other direction at the next time point, i.e. whereas the CD4 data are not themselves autocorrelated the transitions will be and will not be independent of each other.

Proof

To prove this, assume that the CD4 values are independent of each other. For simplicity assume that they have a constant mean and standard deviation. A sequence of independent CD4 values, X_1, X_2, X_3, . . ., will be observed at the time points 1, 2, 3, ...

Between time points 1 and 2 a transition between stages depends on the value of $X_1 - X_2$. But between time points 2 and 3 a transition between stages depends on the value of $X_2 - X_3$. Obviously the values of $X_1 - X_2$ and $X_2 - X_3$ are negatively correlated with a correlation coefficient of -0.5. Each possible transition depends on the value of the difference between a consecutive pair of CD4 values. This is not independent of the transition decided by the difference between the previous consecutive pair of CD4 values. Thus when the true CD4 value remains constant the sequence of transitions observed will not be independent and will be negatively correlated. The same principles apply for the non-steady state.

Conclusion

Many of the results of the paper depend on the assumption of independence of transitions. This is here shown to be invalid.

The following contributions were received in writing after the meeting.

Odd O. Aalen (University of Oslo): Longini, Satten and coworkers have put considerable effort into developing Markovian models for the course of marker processes like the CD4 cell count. There is no doubt that this is a fruitful idea. From a statistical point of view markers are undoubtedly stochastic processes and should be modelled and understood as such. Markovian models ought to be much more widely applied in epidemiology and medical statistics.

One possible objection that may be raised in this case is to the authors' use of a model with discrete state space, when a model with a continuous state space (i.e. a diffusion process) seems to correspond better to the actual data. Furthermore, the definition of individual stages by dividing the CD4 count into intervals of 200 or 150 seems rather arbitrary. By the way, do we really need seven stages? Fewer stages would decrease the basic problem of measurement error, and it is my experience that we may still obtain a good fit to estimated incubation distributions.

The more specific aim of the present paper is to include measurement error and to distinguish this from the 'genuine' development of the process. Although this is done in a novel and elegant way, it may run the risk that the estimated model is too dependent on the specific mathematical structure imposed. Perhaps we should rather estimate measurement error by separate studies with replication of the CD4 measurements, and then introduce this estimate into the Markov model. The idea of suitable replication at different levels is ordinarily considered basic to distinguish separate sources of variation. There is also the more basic problem of what is really meant by measurement error, which may be decomposed into several components. This is just briefly touched on by the authors.

In spite of these objections I believe in Markov modelling. It has a close connection to the interesting field of phase-type models in probability theory. What I find especially appealing and fruitful about it is the possibility of extending the state space to include various other events that might happen to the individual, e.g. diagnosis of HIV infection or the introduction of treatment. It is not necessary that only states describing the progression of HIV disease, or CD4 count, should be included in the Markov model. Extended Markov models are at present being developed by a group working on AIDS predictions for England and Wales.

I congratulate the authors on their interesting further development of Markov model methodology.

Carlo Berzuini (University of Pavia) **and Cecilia Gobbi** (Ospedale San Raffaele, Milan): Our first comment on this interesting paper pertains to the choice between the Markov methods proposed by the authors and methods based on a growth curve representation of marker evolution. One advantage of the latter is that they do not require discretization of continuously valued marker variables; the price for this will often be the need to use Markov chain Monte Carlo methods of inference. A more general approach uses dynamic (autoregressive) models in which the time evolution of the marker and that of the hazard of failure (AIDS) in each subject are jointly modelled as interrelated stochastic processes (Berzuini and Larizza, 1996). Perhaps an absence of strong parametric assumptions in this case attenuates problems due to unknown individual times of infection.

Our specific experience developed within a study on paediatric HIV infection (Berzuini and Gobbi, 1995). We took the 'true' square-root CD4 count for child i at age t, $Y_i(t)$, to follow a linear growth pattern

$$Y_i(t) = \alpha_i + \beta_i t,$$

with a bivariate normal population prior on the unknown parameters (α_i, β_i). The instantaneous hazard of onset of AIDS for child i at age t, $\lambda_i(t)$, was taken to depend on past CD4 history according to the equation

$$\log \lambda_i(t) = \theta_0 + \theta_1 Y_i(t) + \theta_2 \beta_i. \tag{17}$$

The model was fitted by Gibbs sampling using the program BUGS (Thomas *et al.*, 1992).

We take the above model quite seriously for individual prediction. Would the authors consider the use of a Markov model for the same purpose? It is very important to assess the validity of the model forecasts. For example, when a given subject i is aged t and is still AIDS free, we may use the model to estimate the probability p_i that this subject develops AIDS within T years, conditionally on information about previously observed subjects and on observations made on subject i in $(0, t)$. If we do so, within T years we shall be able to assess how good this forecast was. A measure Q_i of predictive performance in this case is $\log p_i$ if the subject did develop AIDS before age $t + T$; otherwise it is $\log(1 - p_i)$. Faced with a sequence of forecasts and outcomes on different subjects, we can take the cumulative performance score

$$Q = \sum_i Q_i$$

as a basis for assessing the validity of the model entertained, along the lines set down by Seillier-Moiseiwitsch and Dawid (1993). For more 'robust' forecasts, we may even fix the θ-parameters in equation (17) to the values θ^* which retrospectively maximize Q. Interestingly, we found that both Q and the optimal θ^* are much dependent on forecasting age t and forecasting horizon T. This is acknowledged by the 'adaptive' approaches to forecasting that we are currently exploring.

Anna Gigli (Istituto per le Applicazioni del Calcolo, Rome) **and Arduino Verdecchia** (Istituto Superiore di Sanitá, Rome): First we wish to thank the authors for another stimulating and useful paper. Researchers active in the field of AIDS modelling know how crucial and influential the assessment of the incubation period is.

The approach based on CD4 counts which has been proposed by the authors for several years has anticipated the method adopted in the USA for identifying AIDS cases. Backcalculation modelling to estimate the incidence and prevalence of HIV infection requires homogeneity of criteria adopted in the definition of the end points, i.e. AIDS cases and incubation periods. In Europe the definition of AIDS is still based on a clinical approach (Centers for Disease Control and Prevention, 1993). However, these two approaches are only different ways of measuring something and are not to be contrasted. In our proposed model for estimating the incidence and prevalence of HIV infections in Italy (Verdecchia *et al.*, 1994) the age at onset of HIV infection has a relevant role, which is incorporated by the covariate AGE in modelling the incubation period. The role of AGE is somehow related to the role of the CD4 counts, for the number of CD4 counts usually decreases as the age of an individual increases. Indeed, in our experience, the expected incubation time from seroconversion to AIDS is very similar to that presented by Satten and Longini in Table 2: Mariotto *et al.* (1992), in a multicentre cohort study of 420 individuals who seroconverted between 1982 and 1990, found an expected incubation time of 10.3 years for individuals aged 25–34 years, with a 95% confidence interval of (7.7, 12.9). These figures are comparable with the 9.8 years estimated by Satten and Longini, although their 95% confidence interval is narrower (this is partially due to the larger size of their sample).

We are currently studying the role played by the uncertainty related to AGE in the modelling of the incubation period, by means of a Monte Carlo method, the parametric bootstrap. Such uncertainty spreads through the backcalculation model and affects the estimation of the precision of incidence and prevalence of HIV infection. From preliminary results we obtained an increase up to seven times the standard errors of incidence and prevalence estimates previously obtained via backcalculation (see Verdecchia and Mariotto (1995)).

W. R. Gilks and D. DeAngelis (Medical Research Council Biostatistics Unit, Cambridge):

Prediction

The value of the hidden Markov model for making predictions is that it provides a framework for utilizing all past CD4 counts for an individual, instead of just the most recent measurement. Table 5 suggests that the bidirectional version of the model is performing reasonably well for predictions about 1 year ahead, although validation on an independent data set, with both short- and long-range predictions, would be more convincing. It would be useful to see how the model compares with a simpler approach to prediction using only the most recent CD4 count.

Inference

Whether genuine increases in CD4 count can be sustained for more than a few months is fundamental to understanding the disease process. The better fit of the bidirectional model suggests that such reverses can be sustained, but we believe that the assumptions in the model are too strong to draw this conclusion safely. In particular, it seems unrealistic to assume that the time spent in the current stage does not affect progression to the next stage. Any consequent lack of fit in the unidirectional model could give rise to the apparent superiority of the bidirectional model.

The following nonparametric approach might be used to assess whether increases in true CD4 counts can be sustained for more than a few months. Let c denote a positive constant. Let n be the number of non-overlapping sequences of observed CD4 counts x_{ij}, $x_{i,j+1}$, $x_{i,j+2}$, $x_{i,j+3}$ in the data set, in which $x_{i,j+1} + c < x_{i,j+2}$. Let m be the number of these n in which $x_{ij} < x_{i,j+3}$. If long-term increases in true CD4 counts are impossible then occurrences of $x_{i,j+1} + c < x_{i,j+2}$ and $x_{ij} < x_{i,j+3}$ must be due to independent measurement errors, and consequently

$$p = \mathrm{pr}(x_{ij} < x_{i,j+3} | x_{i,j+1} + c < x_{i,j+2}) \leqslant 0.5.$$

However, if some occurrences of $x_{i,j+1} + c < x_{i,j+2}$ reflect a long-term increase in the true CD4 count, then p will be increased, perhaps to a value above 0.5. Testing the null hypothesis that $p \leqslant 0.5$, by referring m to a binomial(n, 0.5) distribution, would then provide a basis for inference about genuine long-term increases in CD4 counts. The constant c should be chosen to maximize power. Although this approach will lack power compared with a parametric approach, it has the merit of avoiding strong assumptions. Conceivably, more powerful nonparametric tests of this sort might be devised.

Michael A. Newton (University of Wisconsin, Madison): A curious feature of the proposed hidden Markov model (HMM) struck me as I read this very interesting paper. Rather than deriving the stages S from the underlying true CD4 counts Y, the authors build Y from S by supposing that S is a Markov chain and the true counts are conditionally independent given S. A more natural view, perhaps, treats S as a discretized version of the more primitive process Y. Intuitively, such discretization amounts to loss of information, and hence to an increased dependence on the past. So viewed, how can the stages exhibit only Markovian dependence? This problem has been studied. A Markov chain Y is said to be *lumpable* (relative to a certain discretization) if the induced chain S is a Markov chain for any initial distribution (Kemeny and Snell (1976), p. 124). Equivalently, the transition probability from state $Y(t) = y$ in one stage to all states in another stage cannot depend on y. Lumpability seems to be a very rigid assumption, especially if different stage boundaries are allowed.

The problem of HMM comparison is ripe for further study. For example, do we know that the likelihood ratio between unidirectional and bidirectional models is distributed approximately as χ^2? Are there simple diagnostics for model fit? In a model checking problem, Newton *et al.* (1995) observed a property of a certain HMM that may provide a useful diagnostic tool. Consider a stationary process Y, and data points $X(t)$ arising conditionally independently given Y, with conditional distribution determined by $Y(t)$ but not t (as in most HMMs), and constructed so that $E\{X(t)|Y(t)\} = Y(t)$. Whereas the marginal variance of $X(t)$ exceeds that of $Y(t)$, it is readily shown that covariances of Y are identical with covariances of X under a simple independence property of Y—specifically that $Y(s)$ be independent of the increment $Y(t) - E\{Y(t)|Y(s)\}$ for times $s \neq t$. Under this condition, the data provide direct information about the covariance of the unobserved process. As part of model checking, we may add the fitted autocovariance function to a scatterplot of all pairs $\{t_i - t_j, \{X(t_i) - \bar{X}\}\{X(t_j) - \bar{X}\}\}$. The autocovariance function is defined for stationary processes, and so this check is possible for the bidirectional model conditioned on the absence of transfer to acquired immune deficiency syndrome, but not for the unidirectional model.

Frank van de Pol (Statistics Netherlands, Voorburg): For longitudinal HIV data Satten and Longini have constructed a new measurement error correction model, which takes the numerical nature of the data into account. In the social sciences their model might be used as a new starting point for the study of income in poverty research or for the study of backsliders in criminology.

Parts of their approach could be useful for the closely related latent Markov model for categorical data (Lazarsfeld and Henry, 1968). The EM algorithm, which can be used to estimate this model for categorical data, will be much quicker with long time series when the result which they attribute to Baum *et al.* (1970) is used. It can be used straightforwardly in the E-step, and there should also be ways to use it in the M-step.

Although this model is a step forward, it is not the end of development. I can see three directions for new methodological research: firstly, the possibility to incorporate exogenous variables, which has already been mentioned by the authors. This is partly a matter of combining a model formula, algorithms and software practically.

Secondly, are the results sensitive for the assumption of a transition rate (called an infinitesimal generator by Satten and Longini), which does not depend on states before the last (i.e. the Markov assumption)? Perhaps people who had a lower than average transition rate before will also have a low transition rate later (and vice versa), i.e. not only observable exogenous variables could be added to the model but also unobserved heterogeneity. Such a specification might also better explain the small number of six people in the category 'AIDS' after 9 years, compared with the present model which expects 21 people (Table 5).

Thirdly, the present approach does not provide a likelihood ratio type of test for the model against a saturated model of the data, i.e. the nine-dimensional cross-table of CD4 stage at nine time points. One could pursue the development of a more efficient (Baum-type) algorithm for the latent Markov model for categorical data. Fitting the HIV data to such a model would have the advantage that the assumption of a latent (or hidden) Markov chain could be tested by using bootstrap methodology (Langeheine *et al.*, 1995). Such a test is useful. A latent Markov chain without parameters for exogenous variables or for heterogeneity does not fit many social science data sets with a large number of time points.

Finally did the authors have any reason for not using the latent Markov model, apart from the problem that the algorithms that are at present available cannot handle so many time points? They mention in Section 1 that this model has been used with other objectives, but this is no reason to reject it.

Mark R. Segal (University of California, San Francisco): Satten and Longini present a comprehensive treatment of stage modelling of longitudinal CD4 count data and the attendant necessity to accommodate errors in variables. Interestingly, were CD4 counts better behaved and measurement error inconsequential, it is unlikely that a staging scheme would be adopted: the largely arbitrary stratification of a continuous marker would not afford a useful clinical staging system. The authors' motivation derives not from the practicalities of the staging scheme but rather from difficulties associated with growth curve models and the accuracy of the Markov assumption.

One difficulty ascribed to growth curve modelling is the need to know or estimate infection times. For the San Francisco Men's Health Study (SFMHS) much is known about the HIV infection density (Bacchetti, 1990). These infection rates have been used in modelling CD4 progression (DeGruttola *et al.*, 1991; Lange *et al.*, 1992a; Vittinghoff *et al.*, 1994). The issue can be circumvented by using first differences for modelling derivatives (Taylor *et al.*, 1994; Vittinghoff *et al.*, 1994). Although the staging approach avoids infection times, it requires specification of the true CD4 density in the open-ended first stage. Satten and Longini's choice (truncated log-normal) is apparently for convenience.

The other difficulty cited is the mandated interrelated modelling of CD4 decline and AIDS-free survival. Such modelling has been approached in various ways (DeGruttola and Tu, 1994; Tsiatis *et al.*, 1995; Faucett and Thomas, 1995). These methods explicitly accommodate measurement error, are readily generalized to handle additional markers or covariates and provide estimates of such variables' effect on developing AIDS. It is unclear how such extensions would be developed within the staging framework.

The Markov assumption itself seems apt. However, it is surprising how slowly the correlation between raw CD4 counts decays over time; see Lange *et al.* (1992a) for SFMHS results. Satten and Longini deem the bidirectional chain with measurement error successful in staging HIV disease with each stage representing approximately 2 years of the incubation period. This implies a piecewise linear model for CD4 decline with the slope for the later (smaller range) stages being attenuated compared with the early stages. But, there is an accelerated rate of CD4 decline before progressing to AIDS (Kiuchi *et al.*, 1995). Another component of model validation is the agreement between observed and expected stage occupation probabilities. Yet, in calibrating this agreement, it is important to be mindful of the number of parameters fitted and that the predictions are at the population level. In addition to reporting delays explaining the deficit of observed AIDS cases there is the possibility of informative censoring whereby individuals with rapidly deteriorating immune function withdraw from the study before progressing to AIDS.

Finally, a CD4 count below 200 now constitutes AIDS. This means that the staging scheme will require modification to reconcile CD4 measurement error, bidirectionality and AIDS as an absorbing state.

P. J. Solomon (University of Adelaide): Satten and Longini propose an intriguing use of hidden Markov chains. Although interesting, the paper raises the question: does it help?

Overall, the authors' results on time to CD4 stages, including AIDS, are in substantial agreement with previously published results about the AIDS incubation period. A mean time to AIDS of 10 years, with treatment, is generally thought to be about right. The estimates of treatment effects in Section 3.3 are also in accord with clinical trial results and other work on HIV and AIDS; see Solomon (1996) for a review.

The authors' results also appear to provide evidence that it is appropriate to 'smooth' directly CD4 cell counts which could then be fitted as covariates in survival models. The evidence lies in the role of increasing σ_1 in reducing the number of stages visited (i.e. its role as a 'smoothing' parameter), the observed smooth trend in shorter time to AIDS with declining CD4 cell counts (Table 2) and the evidence for serial correlation from fitting the bidirectional model. Examining the sensitivity of their results to the choice of CD4 stage cutpoints would be helpful in confirming that the observed effects are measuring what we think they are measuring.

There are difficulties in interpreting survival models with markers of disease progression included as evolutionary or time-dependent covariates, as Satten and Longini note. However, survival models are straightforward to fit within the counting process framework implemented in S-PLUS (and SAS) which readily accommodates complex censoring, truncation and filtering patterns, time-dependent covariates and multivariate data (see, for instance, Ripley and Solomon (1995)).

One simple way to estimate time to AIDS or other CD4 stage for the San Francisco Men's Health Study (SFMHS) data is to summarize (possibly transformed) individual CD4 cell count trajectories by an intercept or starting CD4 value x_0, a measure of serial correlation $\tilde{\rho}$, or trend, and a within-individual variance component $\tilde{\sigma}^2$, then to fit these as covariates in a survival model with hazard

$$h\{t;\, Z(\;)\} = h_0(t)\, \exp\{\beta_0 x_0 + \beta_\rho \tilde{\rho} + \beta_\sigma \tilde{\sigma} + \gamma^{\mathrm{T}} X(t)\}.$$

$X(t)$ represents additional covariates of interest, such as other temporal trends, including changes in the definition of AIDS and age (neither of which is incorporated in the authors' model), seroconversion and treatment effects; the covariates may be fixed, random or time dependent.

The degree of smoothing and parameter fitting will depend on the number of observations for each individual. There are ways of estimating variance components for panel data and the within-individual variability could none-the-less be incorporated directly as a random effect. My graduate student, Ms Colleen Hunt, has found that the assumed transformation model for 'stabilizing' the variance in the SFMHS data is valid for the earlier disease stages (i.e. higher CD4 cell counts) but may not be for disease stages closer to AIDS. It may be helpful to account for such features in analysis.

Jeremy M. G. Taylor (University of California, Los Angeles): The authors present an excellent approach to overcoming the problems caused by measurement error in continuous time Markov chains. However, it seems to me that some of the problems that they so ingeniously solve are a consequence of their decision to use staged Markov models. CD4 counts are measured on a continuous scale, and thus it seems unnatural to discretize these data into a small number of stages. I would be interested to know how the authors would adapt their approach, if at all, to handle a much larger number of stages, thus in essence regarding the change in CD4 count as a continuous process.

Are the authors willing to interpret biologically their finding that a bidirectional model fits better than a unidirectional model? Do HIV-infected subjects really show improvements in their 'true' CD4 values? Some of the slow timescale variation in CD4 count could be due to changes in reagents and equipment used to measure lymphocyte subsets, as has been found by others (Vittinghoff *et al.*, 1994) in their analysis of data from the San Francisco Men's Health Study.

Other researchers have used square-root (Lange *et al.*, 1992b; Vittinghoff *et al.*, 1994) and fourth-root (Taylor *et al.*, 1994) transformations, instead of logarithms, to achieve homogeneity of measurement error variance for CD4 counts. Are the substantive conclusions, such as the effect of treatment, robust to the choice of transformation?

The **authors** replied later, in writing, as follows.

We thank the discussants for their many interesting comments and will address some of their individual concerns later. We begin with a general discussion of our decision to use a staged rather than a continuous model which was questioned by several discussants. Because our model has two interpretations (as a staged model, and as a growth curve type of model with step function distributions for the underlying probability distribution function (PDF) of CD4 cell counts), this decision can be examined in either light. A continuous Markov model would correspond to a diffusion process; in general, the determination of the PDF of future values conditionally on the current value requires solving a partial differential equation that does not have a closed form in the general case. The class of tractable models of this type corresponds roughly to linear stochastic differential equations, featuring deterministic progression with additive noise. Adding measurement error to obtain a marginal likelihood analogous to equation (5) would require integrating these (possibly numerically calculated) PDFs at each measurement time. Although the Markov structure would simplify this operation in a manner similar to the simplification of expression (8), carrying out the n_i-fold integral would be extremely difficult. Of course, the integrals could be approximated by finite sums; the resulting approximation would then be very similar to our model. We find it interesting that most of the discussants find the strong assumptions of a log-linear decline in CD4 with additive noise preferable to the much more flexible and non-linear model that we have presented.

If we consider the growth curve interpretation of our model (see the end of Section 2.3 and Fig. 2), then the role of stages is somewhat diminished. Although Dr Ashby is technically correct that the step function model for the distribution of true CD4 counts is 'implausible', it none-the-less is a very flexible model, and as such may capture the essence of the true, continuous distribution. In this interpretation, the major role of the number of stages is to determine the number of steps in the approximation. Because the true CD4 count is never observed, the step function approximation that we have used may be adequate when integrated against the error distribution (4c). Most discussants argue for more stages for added flexibility; interestingly, Professor Aalen suspects that we have already added too many.

The parameter σ_1 measures the extent to which measurement error causes the true and observed CD4 counts to differ. However, some caution must be exercised in comparing our estimated σ_1 with published values, because extra variability is accounted for in our finite width stages. Specifically, the distribution of true CD4 counts immediately after a true value y should be nearly a degenerate distribution about y; in our model, this distribution would be the uniform distribution over the stage containing y. Although the finite width of the stages limits our ability to describe the short time behaviour of the true CD4 count accurately, in our application (and in most acquired immune deficiency syndrome (AIDS) cohort studies) data are typically spaced by at least 6 months. Additionally, we want a model which describes the long timescale behaviour. However, the 'correct' number of stages necessary should be determined to some extent by the spacing between waves. Thus, the value of σ_1 that we obtain represents a trade-off between the amount of measurement error that would be obtained in replicate observations and the interval width; we would expect the value of σ_1 to vary with the number of stages.

Several discussants commented on the desirability of obtaining replicate measurements to make an independent estimate of error. We agree that this is desirable; however, even if multiple aliquots are obtained from each subject at a single visit, obtaining replicate samples a few hours, days or weeks apart, as would be necessary to estimate the variability on the day-to-week timescale, is impossible in many studies. We agree with Professor Aalen that our estimates of what is signal and what is noise depends on the mathematical structure of the model; however, in the absence of replicates, we feel that it is still useful to ask what amount of 'measurement error' would produce the data leading to the best possible trade-off between model fit and misclassification.

Finally, several discussants wondered about the effect of unmeasured heterogeneity on our estimated model. Incorporating such heterogeneity falls into the general category of frailty modelling. In our case, elements of the infinitesimal generator γ would be multiplied by the random variable Z that follows some distribution. The parameters of this distribution would be jointly estimated with the other model parameters. Such a model is thought to be necessary to describe the progression of human immunodeficiency virus (HIV) disease in children; however, the situation in adults is less clear. Although cofactors explaining some of the variability of the incubation period have been proposed, it has not been convincingly established that long-term survivors constitute more than the tail of a single distribution. In any event, adding frailty to our model is likely to be difficult; see Aalen (1988) for details of the case where Z is continuous and no measurement error is assumed.

Although we appreciate the comments of all the discussants, for brevity we respond only to those individual comments that directly relate to some aspect of our work.

The effect of measurement is very large; we assure Professor Lawless that Fig. 3 is correct. This leads to a fairly narrow 95% confidence interval for σ_1 (0.162, 0.182) which was inadvertently left out of the paper. To clarify our statement concluding Section 2.6, all confidence intervals reported account for the variability of *all* parameters in the model, *including* σ_1 and σ_2. The profile likelihood in Fig. 3 is for display only; all confidence interval statements were made using the full likelihood.

Although time homogeneity may be violated, we feel that the massive measurement error effect must be accounted for first; the diagnostic tests referred to by Professor Lawless presumably do not account for this effect. Professor Lawless also questions whether the model (4a)–(4c) is first order Markov; the hidden Markov model for the true stages *is* first order Markov, whereas the marginal distribution of the observed CD4 counts (or observed stages) is not Markov, as indicated in the discussion after equation (16). Professor Lawless points out a minor inconsistency in that we claimed that our naïve model reduced to a Markov model fit to the observed stage data; in fact, the naïve model is based on the same data as those of equations (4a)–(4c) and hence also fits the observed CD4 data. To produce a family of models that uses only stage information, we integrated expressions (5) and (6) over the stage boundaries for the observed CD4 as well as the true CD4 counts. The estimated infinitesimal generator as a function of σ_1 (and hence mean times to AIDS and number of stages visited) using this model is very similar to that presented in our paper; the only major difference is the scale in Fig. 3, where the likelihood ratio statistic comparing $\sigma_1 = 0$ and $\sigma_1 = \hat{\sigma}_1$ is 435 (1 degree of freedom) for the model using only stage information. Professor Lawless also claims that the sample paths for the 'true' CD4 count are not especially plausible because they do not resemble those observed in patients. In fact, working with simulated data, we find that the observed paths for our model do resemble the sort of data collected from real patients, whereas the 'true' sample paths are meant to correspond to averaged, time-smoothed CD4 values that are rarely available in clinical studies. Finally, Professor Lawless is correct in identifying an error in our citation of Hoover's work: Hoover *et al.* (1992) estimated the standard error of the laboratory effect to be 0.185 and the standard error of the total effect to be 0.260. Thus, by subtraction, the standard error of the short timescale effect would be 0.183.

Dr Taylor asks whether we are willing to interpret the significantly better fit of the bidirectional model as indicative of an increase in true CD4 values in individuals with HIV disease. It seems plausible that some individuals will have sample paths in which short-term increases in these values occur; however, the expected CD4 function is monotone decreasing, even in the bidirectional model. Further, we agree with Dr Gilks and Dr DeAngelis that the success of the bidirectional model may be due to a true CD4 process that is actually semi-Markov; we are currently exploring such models.

Professor Newton proposes an extremely interesting technique for assessing model fit in stationary hidden Markov models. Development in this area is particularly important, because model validation is difficult when an underlying unobservable process is assumed. Unfortunately, it requires longer series of observations than are available in the San Francisco Men's Health Study (SFMHS); in particular, we would have to observe individuals who are AIDS free sufficiently long for their stage occupation distribution to approximate the stationary distribution. In our particular application, the hazard for AIDS is so strong that we do not have the opportunity to observe steady state behaviour among the AIDS-free population.

Professor Berzuini and Dr Gobbi compare their approach to modelling CD4 trajectories in children with our treatment of the San Francisco Men's Health Study (SFMHS). The difference in analytical options for data where the time of infection for all study participants is known (as with children) and data where most participants are infected before the study begins should not be underestimated. In a prevalent cohort such as the SFMHS, it appears that there are only two valid options: either the infection time must be imputed in some way, or the Markov assumption must be made. This added uncertainty also affects the quality of prediction which can be made using the model. As a result, we said in our paper that we expect our model to be of most use for epidemiological rather than individual prediction. That said, we would welcome a comparison of the quality of prediction by our model and the growth curve model at the individual level, especially in the light of the results of Taylor *et al.* (1994) on the superior fit of an integrated Ornstein–Uhlenbeck process when compared with the growth curve model. The prediction algorithm suggested by Professor Berzuini and Dr Gobbi is similar to our comparison of observed and expected AIDS cases in the last two rows of Table 5; the agreement between observed and expected appears to be satisfactory.

Dr Segal asks how an additional marker could be included in our model. A two-dimensional array of stages is possible, with one axis corresponding to each marker. He also notes that, in the USA, an observed CD4 count of less than 200 now warrants an AIDS diagnosis, and he suggests that this change in the AIDS surveillance case definition may require changes to our staging system. Since AIDS diagnosis by CD4 can only occur when a CD4 test is performed we would prefer to keep the staging system unchanged, and to add a model to predict when testing occurs. An overall incubation distribution can then be obtained by averaging over the testing model. In this context, we note the work that Professor Aalen referred to in his contribution, which incorporates other events involved in the natural history of HIV infection into the state space of the Markov model.

We are pleased that Dr Richardson and Dr Guihenneuc-Jouyaux are approaching the problem of fitting hidden Markov chains to data by using a Bayesian framework with Markov chain Monte Carlo algorithms. We look forward to seeing their results, especially those concerning identification of the number of stages.

Dr Gilks and Dr DeAngelis would like to compare the prediction from our model with that from a simpler model based on only the last CD4 count. Presumably this prediction should be done using the last 'true' CD4 value; otherwise, this comparison is with our naïve model. In the absence of replicates, the true CD4 count can only be determined as a function of the entire observed CD4 trajectory (e.g. through smoothing). In fact, our prediction comes fairly close to their goal: the discussion after equation (16) shows that the role of prior CD4 counts is to improve the estimate of the current CD4 value. We agree that a nonparametric test for determining whether backward transitions truly occur would be desirable.

Dr Verdecchia points out that age has been shown to affect the duration of the incubation period. In models such as ours, age at study entry could be considered as an explanatory variable, but a proper model with age as a time-dependent effect would lead to a non-homogeneous Markov model. Such a model would be considerably more difficult numerically. In any event, the age range in the SFMHS is very narrow; the interquartile range among HIV-infected men is only 9 years.

Finally, we are very pleased with the results of Dr Yang's small simulation study and look forward to future contributions from him in this area.

References in the Discussion

Aalen, O. O. (1988) Dynamic description of a Markov chain with random time scale. *Math. Scient.*, **13**, 90–103.

Bacchetti, P. (1990) Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *J. Am. Statist. Ass.*, **88**, 1002–1008.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Statist.*, **41**, 164–171.

Berzuini, C. and Gobbi, C. (1995) Bayesian graphical models of the natural history of HIV infection. In *Probabilistic Reasoning and Bayesian Belief Networks* (ed. A. Gammerman). Henley: Waller.

Berzuini, C. and Larizza, C. (1996) A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *IEEE Trans. Pattn Anal. Mach. Intell.*, **18**, 109–123.

Broughton, P., Holder, R. and Ashby, D. (1986) Long-term trends in biochemical data obtained from two population surveys. *Ann. Clin. Biochem.*, **23**, 474–486.

Centers for Disease Control and Prevention (1993) Impact of the expanded AIDS surveillance case definition on AIDS case reporting: US, first quarter, 1993. *Morb. Mort. Wkly Rep.*, **42**, 308–310.

DeGruttola, V., Lange, N. and Dafni, U. (1991) Modeling the progression of HIV infection. *J. Am. Statist. Ass.*, **86**, 569–577.

DeGruttola, V. and Tu, X. M. (1994) Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

Faucett, C. L. and Thomas, D. C. (1995) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statist. Med.*, to be published.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C. and Yan, P. (1994) Multi-stage Markov models for analyzing incomplete disease history data with illustrations for HIV disease. *Statist. Med.*, **13**, 805–821.

Hansford-Miller, F. (1994) *A.I.D.S. and Quarantine.* Canterbury: Abcado.

——— (1996) *A History of Medicine in Western Australia 1829–1870*, vols 1–12. Yanchep: Abcado.

Hoover, D. R., Graham, N. M. H., Chen, B., Taylor, J. M. G., Phair, J., Zhou, S. Y. J. and Muñoz, A. (1992) Effect of CD4+ cell count measurement variability on staging HIV-1 infection. *J. AIDS*, **5**, 794–802.

Kalbfleisch, J. D. and Lawless, J. F. (1989) Some statistical methods for panel life history data. In *Proc. Statistics Canada Symp. Analysis of Data in Time*, pp. 185–192. Ottawa: Statistics Canada.

Kemeny, J. G. and Snell, J. L. (1976) *Finite Markov Chains.* New York: Springer.

Kiuchi, A. S., Hartigan, J. A., Holford, T. R., Rubinstein, P. and Stevens, C. E. (1995) Change points in the series of T4 counts prior to AIDS. *Biometrics*, **51**, 236–248.

Lange, N., Carlin, B. P. and Gelfand, A. E. (1992a) Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4+ counts. *J. Am. Statist. Ass.*, **87**, 628–631.

——— (1992b) Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *J. Am. Statist. Ass.*, **87**, 615–626.

Langeheine, R., Pannekoek, J. and van de Pol, F. (1995) Bootstrapping log-linear and latent class models. Submitted to *Sociol. Meth. Res.*

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis.* Boston: Houghton-Mifflin.

Mariotto, A. B., Mariotti, S., Pezzotti, P., Rezza, G. and Verdecchia, A. (1992) Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: a comparison with male homosexuals. *Am. J. Epidem.*, **135**, 428–437.

Newton, M. A., Guttorp, P., Catlin, S. Assunção, R. and Abkowitz, J. L. (1995) Stochastic modeling of early hematopoiesis. *J. Am. Statist. Ass.*, **90**, 1146–1155.

Ripley, B. D. and Solomon, P. J. (1995) Correspondence on statistical models for prevalent cohort data. *Biometrics*, **51**, 373–375.

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993) On testing the validity of sequential probability forecasts. *J. Am. Statist. Ass.*, **88**, 355–359.

Solomon, P. J. (1996) AIDS: modelling and predicting. In *Models for Infectious Human Diseases: Their Structure and Relation to Data* (eds V. Isham and G. Medley). Cambridge: Cambridge University Press.

Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994) A stochastic model for analysis of longitudinal AIDS data. *J. Am. Statist. Ass.*, **89**, 727–736.

Thomas, A., Spiegelhalter, D. J. and Gilks, W. R. (1992) BUGS: a program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 837–842. Oxford: Clarendon.

Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. (1995) Modelling the relationship of survival to longitudinal data measured with error: application to survival and CD4 counts in patients with AIDS. *J. Am. Statist. Ass.*, **90**, 27–37.

Verdecchia, A. and Mariotto, A. B. (1995) A backcalculation method to estimate the age and period HIV infection intensity, considering the susceptible population. *Statist. Med.*, **14**, 1513–1530.

Verdecchia, A., Mariotto, A. B., Capocaccia, R. and Mariotti, S. (1994) An age and period reconstruction of the HIV epidemic in Italy. *Int. J. Epidem.*, **23**, 1027–1039.

Vittinghoff, E., Malani, H. M. and Jewell, N. P. (1994) Estimating patterns of CD4 lymphocyte decline using data from a prevalent cohort of HIV infected individuals. *Statist. Med.*, **13**, 1101–1118.