STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:441–462 (DOI: 10.1002/sim.1270)

Applications of continuous time hidden Markov models to the study of misclassified disease outcomes

Alexandre Bureau[1,*], Stephen Shiboski[2] and James P. Hughes[3]

[1]*Group in Biostatistics, School of Public Health, University of California, Berkeley, CA 94720, U.S.A.*
[2]*Department of Epidemiology and Biostatistics, University of California, MU 420-W San Francisco, CA 94143, U.S.A.*
[3]*Department of Biostatistics, School of Public Health and Community Medicine, University of Washington, Seattle, WA 98195, U.S.A.*

SUMMARY

Disease progression in prospective clinical and epidemiological studies is often conceptualized in terms of transitions between disease states. Analysis of data from such studies can be complicated by a number of factors, including the presence of individuals in various prevalent disease states and with unknown prior disease history, interval censored observations of state transitions and misclassified measurements of disease states. We present an approach where the disease states are modelled as the hidden states of a continuous time hidden Markov model using the imperfect measurements of the disease state as observations. Covariate effects on transitions between disease states are incorporated using a generalized regression framework. Parameter estimation and inference are based on maximum likelihood methods and rely on an EM algorithm. In addition, techniques for model assessment are proposed. Applications to two binary disease outcomes are presented: the oral lesion hairy leukoplakia in a cohort of HIV infected men and cervical human papillomavirus (HPV) infection in a cohort of young women. Estimated transition rates and misclassification probabilities for the hairy leukoplakia data agree well with clinical observations on the persistence and diagnosis of this lesion, lending credibility to the interpretation of hidden states as representing the actual disease states. By contrast, interpretation of the results for the HPV data are more problematic, illustrating that successful application of the hidden Markov model may be highly dependent on the degree to which the assumptions of the model are satisfied.

KEY WORDS: hidden Markov models; longitudinal data; infectious diseases; misclassification

1. INTRODUCTION

Prospective clinical and epidemiological studies of disease progression often focus on understanding the nature of transitions between defined disease stages or states, and the influence of

*Correspondence to: Alexandre Bureau, Statistical Genetics, Genome Therapeutics Corporation, 100 Beaver Street, Waltham, MA 02453, U.S.A.

Contract/grant sponsor: National Institute of Allergy and Infectious Disease; contract/grant number: R01 AI39932

Received April 2001
Accepted February 2002

observed covariates on these transitions. Examples include presence/absence of human papillomavirus (HPV) [1] infection, presence/absence of intestinal parasites [2], development/remission of oral lesions among immunosuppressed individuals with HIV infection [3], and states of immunosuppression in HIV-infected patients measured by levels of CD4 lymphocytes in plasma [4]. Statistical analysis of data from such studies presents a number of challenges. Observations of disease history are often incomplete, including individuals who are prevalent in a specified disease state at the onset of the study, and incident individuals whose entry and exit times into and out of the state are interval censored. Disease states may recur, leading to multiple and potentially correlated sojourns in those states for each individual. Another common problem is that assessment of the disease state of an individual may be subject to misclassification due to poor performance of a biochemical test, inadequate tissue sample or incorrect visual assessment. This issue arises frequently in social science research, where a number of models allowing for latent outcome states are widely accepted and used [5]. By contrast, epidemiological studies of infectious disease have only recently begun to consider models of this type. Despite recent progress such as the studies cited above, further work in this area is clearly needed.

In this paper we apply continuous-time hidden Markov models to longitudinal measurements of a binary disease outcome. Hidden Markov models (HMM) [6, 7] are an extension of Markov models that provide a way to account for potential misclassification due to the measurement procedure. In this formulation the actual disease outcome is represented by a continuous time two-state homogeneous Markov process. States typically represent presence and absence of the disease or of a particular disease manifestation. Because of the sampling and testing procedure these 'hidden' states cannot be observed directly, and measurements of the disease outcome are subject to error. Our models allow covariates effects on both the hidden process and on the conditional distribution of the observed response given the hidden states using a generalized regression framework similar to that proposed by Hughes [8]. Computation of the estimated variance-covariance matrix of parameter estimates is based on a modified version of the procedure proposed by Hughes [9]. Finally, we introduce diagnostic procedures for evaluating the standard assumptions that both transition intensities and misclassification probabilities are constant over time conditional on covariates. Two applications to data from epidemiological studies of infectious disease outcomes subject to misclassification are presented. The first of these studies investigates the occurrence of hairy leukoplakia, an oral lesion associated with HIV-induced immunosuppression among participants in the San Francisco Men's Health Study [3]. The second examines aspects of cervical human papillomavirus (HPV) infection among young women from a prospective study conducted in San Francisco. Taken together, the results illustrate both the benefits and limitations of the proposed models, and provide valuable insights into implementation and interpretation of hidden Markov models to prospective epidemiologic studies.

The statistical literature contains a number of published applications of Markov models to biological outcomes. Kalbfleisch and Lawless [10] and Gentleman *et al.* [11] have investigated the application of continuous time, discrete state, time homogeneous Markov models to longitudinal disease outcomes which are not subject to misclassification. Continuous time hidden Markov models have been applied in genetic map construction and linkage analysis [12], modelling haematopoiesis in cats [13] and modelling the progression of a marker of human immunodeficiency virus disease [4]. In addition, discrete time versions (excluding covariate effects) have been applied to problems similar to those examined here [1, 2, 14].

Unique aspects of the work presented here include a flexible scheme for incorporation of covariate effects, an approach for computation of approximate standard errors, introduction of several model assessment techniques and application to data from two studies which illustrate both strengths and weaknesses of the HMM approach.

2. METHODS

2.1. Observed data structure

As introduced above, we will focus on data from prospective clinical and epidemiological studies. Subjects enrolled in such studies are typically scheduled for clinic visits at regular intervals following recruitment. However, due to missed and unscheduled appointments the intervals between actual visits often vary in length. At each visit, information on current observed disease status and covariates of interest is collected. The former is assumed to be measured with an unknown degree of error, possibly due to an imperfect diagnostic test. Further, due to the irregular nature of follow-up, observations of the actual time of entry into a disease state are frequently interval censored. For the ith subject in a sample of n, we assume that the following variables are observed at the jth visit:

$$\begin{cases} T_{ij} & \text{chronological time of clinic visit} \\ Y_{ij} & \text{binary disease outcome measurement} \\ \mathbf{Z}_{ij} & 1 \times p \text{ vector of covariates} \end{cases}$$

2.2. Hidden Markov model

As noted above, the observed disease outcome measures Y_{ij} are subject to error. The actual underlying disease status is assumed to be a process evolving in continuous time and is denoted $(X(t),\ t>0)$. This process is unobserved or 'hidden', and will be modelled as a continuous time two-state Markov process with states interpreted as the presence and absence of the disease manifestation. Let Y_1^j and T_1^j denote the sequence from 1 to j of observed disease states and observation times for an individual (the subscript i denoting subjects is omitted). In what follows, all inferences are conditional on the observation times. The Markov assumption for the hidden disease process is given by

$$\begin{aligned} P[X(t_j)|X(t_1),\ldots,X(t_{j-1}),Y_1^{j-1},T_1^j=t_1^j] &= P[X(t_j)|X(t_{j-1}),T_{j-1}^j=t_{j-1}^j] \\ &= P_{x_{j-1},x_j}(t_j-t_{j-1}) \end{aligned} \tag{1}$$

where the quantity P_{x_{j-1},x_j} denotes the transition probability of occupying state x_j at time $T_j=t_j$ given that the process was in state x_{j-1} at t_{j-1}. As indicated by the last equality, the transition probabilities of this process are assumed to be stationary. We also assume that, conditional on the state of the hidden process at time t_j, an observation Y_j is independent of all previous observations and the hidden process prior to time t_j:

$$P[Y_j|X(t_1),\ldots,X(t_j),Y_1^{j-1},T_1^j=t_1^j]=P[Y_j|X(t_j),T_j=t_j]=f(y_j|x_j) \tag{2}$$

When Y is binary, $f(y|x)$ can be interpreted as the probability of correctly or incorrectly classifying the disease state given the true state. Taken together, equations (1) and (2) constitute a hidden Markov model [6]. If the disease status is observed accurately, then $X(t_j)$ and Y_j coincide, and the model reduces to a 'pure' continuous time Markov process. The conditional independence of misclassification probabilities at successive time points is a strong assumption. However, dependence between successive measurements is confounded with dependence between the true disease states. It is not possible to disentangle these two features of the model with misclassified data. Rather, one would need data in which the true disease state has been established more reliably.

Conditional on the true state of the disease at a given time point x_j, the observed binary measurements are Bernoulli trials:

$$f(y_j|x_j=0) = p_0^{y_j}(1-p_0)^{1-y_j}$$
$$f(y_j|x_j=1) = p_1^{y_j}(1-p_1)^{1-y_j}$$

The logarithm of the odds provides a convenient way of parameterizing these binomial distributions, that is, $p_k = \frac{e^{\beta_k}}{1+e^{\beta_k}}$. We consider only applications where a single measurement is taken at each time point, but multiple independent measurements at each time point can easily be accommodated. The variable Y_j becomes the count of the observations in state 1 among the h_j observations at time t_j and follows a binomial distribution. With h_j large enough at all time points it becomes possible to estimate subject specific and time point specific p_0 and p_1 to check the assumption of constant misclassification probability. The two-state continuous time hidden process can be parameterized by the intensity of acquisition of the disease manifestation u and the intensity of clearance v. The intensity matrix Q is constructed from these parameters:

$$Q = \begin{bmatrix} -u & u \\ v & -v \end{bmatrix} \tag{3}$$

A parameterization in term of log-intensities is adopted for likelihood maximization, that is, $u = e^{\theta_u}$ and $v = e^{\theta_v}$. In general the matrix of transition probabilities $P(t)$ over a time interval t is generated as the matrix exponential of a matrix of instantaneous transition intensities Q:

$$P(t) = \exp\{Qt\}$$

For the two-state process, the transition probabilities have the following form:

$$P_{01}(t) = \frac{u}{u+v}(1-e^{-(u+v)t})$$
$$P_{10}(t) = \frac{v}{u+v}(1-e^{-(u+v)t})$$

In some prospective studies, selective entry criteria are used to enrol subjects into the study, so that the initial distribution of the hidden states may not coincide with the stationary distribution of the process. The initial distribution is therefore modelled separately as a binomial

distribution

$$g(x_1) = \pi^{x_1}(1-\pi)^{1-x_1}$$

The initial distribution is also parameterized by the log-odds ($\pi = \frac{e^{\theta_\pi}}{1+e^{\theta_\pi}}$).

2.3. Modelling covariate effects

As discussed above, covariates measuring clinical, behavioural and/or demographic characteristics can affect the intensity of transitions between disease states. The association of covariates with the transitions of the hidden process is modelled by relating the covariates to the transition intensities via a link function. We adopt a multiplicative intensity model for the effect of covariates on transition intensities. Covariates can be fixed or time varying. The latter are handled by assuming that the transition intensity during the interval between two visits depends on the value of the time varying covariates measured at the end of the interval when the outcome is measured, the typical approach used in survival analysis with time varying covariates. The observed value of covariate k on subject i at time t_j will be denoted by z_{ijk}. The expression for the acquisition intensity over the interval $(t_{j-1}, t_j]$ as a function of p covariates then becomes

$$u_{i,j}(z_{ij}) = \exp\left(\theta_{u0} + \sum_{k=1}^{p} \theta_{uk} z_{ijk}\right)$$

The expression for the clearance intensity is similar. The initial distribution of disease states π can also depend on covariates by extending the log-odds parameterization defined above into a logistic model.

The probability distribution of the observations conditional on the hidden state can also be affected by covariates. This corresponds to 'differential misclassification' in the epidemiological literature and may arise in situations where performance of diagnostic procedures vary according to internal characteristics of the test subject or external conditions at the time tests are performed. Here again, covariate effects can be modelled using a link function between the probability that $Y_{ij} = 1$ and a linear function of the covariates. The logit link was chosen here, giving the expression

$$p(x_{ij}, z_{ij}) = \frac{\exp(\beta_x^{\mathrm{T}} z_{ij})}{1 + \exp(\beta_x^{\mathrm{T}} z_{ij})}$$

2.4. Likelihood computations

Based on the assumptions of the hidden Markov model, the likelihood contribution of a sequence of m observations on an individual subject conditional on the observation times and the covariates can be written:

$$\begin{aligned} L(\theta) &= \Pr[Y_1^m | T_1^m = t_1^m, Z_1^m = z_1^m, \theta] \\ &= \sum_{x_1 \ldots x_m} \Pr[Y_1^m, X_1^m | T_1^m = t_1^m, Z_1^m = z_1^m, \theta] \\ &= \sum_{x_1 \ldots x_m} \pi_{x_1|z_1} f(y_1|x_1, z_1) \prod_{j=2}^{m} P_{x_{j-1}, x_j | z_{j-1}}(t_j - t_{j-1}) f(y_j | x_j, z_j) \end{aligned} \tag{4}$$

(Again, the subscript i is omitted.) The overall likelihood is the product of the n individual contributions. Direct evaluation of (4) appears to involve summing over all possible sequences of hidden states. However, under the hidden Markov model assumptions, the summation can be evaluated recursively, using the forward part of the *forward-backward* algorithm of Baum *et al.* [15]. Maximum likelihood estimation of the parameters is carried out with the expectation-maximization (EM) algorithm described in Appendix A1.

Estimates of the variance of the parameter estimates are obtained by inverting the information matrix of the observed data. The method used to compute the information matrix is detailed in Appendix A2. Standard errors of the estimates of functions of the parameters are obtained from the variance-covariance matrix of the parameter estimates using the delta method. Approximate confidence intervals for risk ratios and odds ratios measuring covariate association are constructed on the scale of the parameter estimates and then mapped onto the risk ratio or odds ratio scale.

A software implementation of the hidden Markov model in continuous time with an S-plus user interface has been developed to perform the analysis and is described in reference [16].

2.5. *Goodness-of-fit diagnostic tools and identifiability verification*

Diagnostic methods for HMMs are not well developed. This is partly a reflection of the relative lack of methods for conventional continuous time Markov models, and also because of the complex structure of HMMs. Although the hidden states X_j of these models define a Markov process, the process describing the observed states Y_j is generally not Markov, displaying higher order dependence. Further the parameterization of HMMs does not allow closed-form expressions for transition probabilities of the observed process. Both of these features complicate model assessment. Here we propose some *ad hoc* approaches to evaluating goodness-of-fit and model assumptions, and provide a brief discussion of identifiability issues.

A useful initial step in evaluating whether the addition of the latent structure provides an improved description of a given set of data is to compare the estimated transition probability with that from the corresponding Markov model excluding the hidden layer and to inspect the estimated misclassification probabilities in light of available external knowledge about the measurement process.

Empirical estimates of the transition structure of a Markov process provide a useful tool for evaluating parametric alternatives for conventional continuous time Markov models. We propose a similar approach based on applying the Kaplan–Meier estimate to the transition times between a specified pair of observed states defined as the time from an initial observation in the state of origin to the first subsequent observation in the other state. Transition times are right censored when no subsequent observation in the other state is recorded. Because a single individual may contribute multiple times for a given transition, standard methods for inference about such estimates do not apply. However, the empirical estimates can be plotted with estimates obtained via simulation from the HMM fit to the same data. Applying the procedure to every state transitions provides a visual assessment of overall model fit. A complementary approach which also provides an assessment of degree of dependence between successive states is to compare the model prediction $P[Y_{j+1}|Y_1^j, T_1^{j+1}]$ of the observed state at observation $j+1$ conditional on the sequence of past observations to the corresponding empirical proportions. Given sufficient data, the above procedures can also be applied to subsets defined by distinct values of covariates. For the reasons outlined above, observed

lack of fit cannot be easily attributed to particular sources such as non-stationarity of the hidden process.

The identifiability of a model for a given data set is another concern in fitting and interpreting HMMs. An inherent feature of these models is that the hidden states are interchangeable, and any permutation of the hidden states has the same likelihood. Beyond this, identifiability of the model parameters depends on the length and structure of the observed sequences of states. For example, the basic HMM with two hidden states has five parameters: one initial state frequency; two transition intensities and two observations probabilities. In the case where the observations are equally spaced in time, at least some of the sequences must contain three or more observations to obtain eight possible sequences $(000, 001, \ldots)$ and 7 degrees of freedom, enough to fit the model. In the case of unevenly spaced observations, sequences of two observations with at least two different time intervals occurring in the data provide enough degrees of freedom to identify the basic HMM. Addition of covariate effects necessitates the presence of at least two observed sequences long enough to identify the basic model with covariate values differing in at least one time point where the covariates affect a component of the model. In both applications considered below, several hundred sequences had three or more observations and there were large numbers of observations at each level of the chosen covariates, ensuring that the models are identifiable.

3. APPLICATIONS

In this section we apply the modelling techniques described above to data from two prospective studies of infectious disease manifestations. In addition to serving as case studies for application of the methods, we feel that the results provide unique insights into the nature of the underlying disease processes and complement previously published findings. In both applications, we fit hidden Markov models to the basic transition structure and compare the results to empirical and simulated estimates using the diagnostic procedures described in Section 2.5. We also compare estimated transition structure for hidden Markov models with corresponding Markov models not incorporating misclassification probabilities. In addition, we investigate the effects of selected covariates on disease acquisition and clearance intensities. Initial single covariate models include coefficients for effects on both transitions. Models with two covariates are constructed similarly based on variables retained in corresponding single-covariate models. In cases suggested by the supporting literature, we augment the models to include effects of covariates on misclassification probabilities. Inference procedures for evaluating covariate effects are based primarily on likelihood ratio comparison of nested models. Use of these tests rely on the fact that the null distribution of 2 times the log of the likelihood ratio tends to a chi-square distribution as the number of subjects tends to infinity or, as implied by the results of Bickel *et al.* [17], as the length of the observation sequence becomes infinitely long.

3.1. Study of hairy leukoplakia

3.1.1. Description of the subjects and data collection. The first application of the methods described above focuses on the natural history of hairy leukoplakia in HIV-infected men. Hairy leukoplakia (HL) is an oral lesion associated with Epstein-Barr virus infection that occurs in

Table I. Distribution of the number of oral examinations per subject.

Number of examinations	Frequency
1	28
2 to 4	61
5 to 7	67
8 to 10	70
11 or 12	108

individuals with immunosuppression due, for instance, to HIV infection [18]. HL appears as a whitish lesion on the lateral border of the tongue and is usually diagnosed by visual oral examination. The lesion is thought to have prognostic significance for the progression of HIV disease [19]. Following initial diagnosis, HL lesions tend to be fairly persistent. However, spontaneous remission and reappearance may also occur in some patients. In addition, the lesion does respond to treatment with antiviral drugs (for example, Acyclovir). Because of the mode of diagnosis, HL lesions can be missed in routine oral examinations. Hilton *et al.* [20] compared diagnoses made by oral medicine specialists and trained medical assistants and found that the medical assistants detected HL in only 12 of the 40 patients diagnosed with HL by oral medicine clinicians, providing some evidence that lesions may be frequently overlooked in routine examinations. Furthermore, other oral lesions can be misdiagnosed as HL or co-occur with HL (for example, oral candidiasis) leading to false positive or false negative diagnoses. Although other analyses have investigated risk factors for development and remission of HL [21, 3], these have not explicitly accounted for misclassified diagnoses.

Hairy leukoplakia is one of the oral lesions investigated among participants of the San Francisco Men's Health Study (SFMHS) [22] who received oral examination at 6-month intervals from February 1987 to the end of the study in May 1993 [23]. Only the HIV-positive subjects that were still participating in the SFMHS in 1987 are included in the present analysis since hairy leukoplakia is diagnosed almost exclusively in immunosuppressed persons. This includes 291 men who tested positive for HIV at the first oral examination in 1987 and 43 who contracted HIV at a later time. The total number of examinations on all the study subjects is 2500. Table I gives the distribution of the number of oral examinations per participant. Subjects who remained in the study until the end of follow-up received 11 or 12 examinations. Those with a single examination do not contribute to the estimation of transition intensities but their observations provide information for estimating initial prevalence.

3.1.2. Estimates from a Markov and a hidden Markov model. A hidden Markov model allowing for misclassification of the outcome and a Markov model where the recorded diagnosis is taken as correct were fit to the sequences of oral examination results on the subjects of the SFMHS. The initial prevalence estimate were similar under the two models: 0.201 (SE 0.022) for the Markov model and 0.213 (SE 0.026) for the hidden Markov model. The transition intensities for both models are reported in Table II along with the expected time to an event (the inverse of the intensity). The estimates of acquisition and clearance intensities of hairy leukoplakia are 4 to 6 times higher under the Markov model than under the HMM. Thus, accounting for outcome misclassification results in longer observed transition times between states. The estimated probability of a positive diagnosis of HL when the hidden state is 0

Table II. Transition intensities estimates under a Markov and a hidden Markov model for hairy leukoplakia (transition/month).

	Acquisition			Clearance		
	Estimate	Standard error	Expected time	Estimate	Standard error	Expected time
Markov	0.0193	0.0016	51.8 months	0.0837	0.0069	11.9 months
HMM	0.0043	0.0011	233 months	0.0130	0.0044	76.9 months

(interpreted as a false positive rate) is 0.034 (SE 0.006). The probability of a negative diagnosis of HL when the hidden state is 1 (interpreted as a false negative rate) is 0.242 (SE 0.025). False positive diagnoses thus occur with a low but non-negligible frequency while false negatives are common, assuming our interpretation of the hidden process is correct.

The estimated distribution of time to appearance of an HL lesion and time to clearance of the lesion given by a hidden Markov model and by a Kaplan–Meier estimator are plotted on Figure 1. A similar pattern is distinguishable on both plots. The probability of remaining in the same disease state (presence or absence of the lesion) estimated from the observations using the Kaplan–Meier estimator drops at approximately 6 months, corresponding to the time between follow-up examinations. By contrast, the probability estimated from the hidden process does not exhibit a sharp decline. After 6 months, the curves of the probability of remaining in the same state estimated from the observations and from the hidden process of the HMM are roughly parallel. The drop in the empirical probability of remaining free of lesion (Figure 1(a)) is the result of isolated false negative observations which are taken as short periods where the lesion is absent. The even sharper decline in the empirical probability of persistence of a lesion beyond 6 months (Figure 1(b)) is partly due to false positives but also to lesions prematurely declared resolved because of a false negative observation. A replicate of the data set was simulated from the fitted HMM and the Kaplan–Meier estimates of the distributions of time to appearance and time to clearance of HL based on those simulated observations are also plotted on Figure 1. The close agreement between the empirical distributions from the simulated and actual data is taken as a sign of the good fit of the model.

3.1.3. Association between covariates and the hairy leukoplakia process. A number of factors measured in the SFMHS may potentially be associated with acquisition or clearance of HL, or both. This set of covariates includes:

(i) presence of oral candidiasis, another type of oral lesions (binary present/absent);
(ii) AIDS diagnosis (binary yes/no);
(iii) CD4 lymphocyte count at present examination and difference between last and present examination;
(iv) current cigarette smoking (binary yes/no);
(v) age at entry into the study;
(vi) treatment with Acyclovir since last examination (binary yes/no);
(vii) treatment with anti-fungal agent since last examination (binary yes/no);
(viii) treatment with AZT since last examination (binary yes/no).

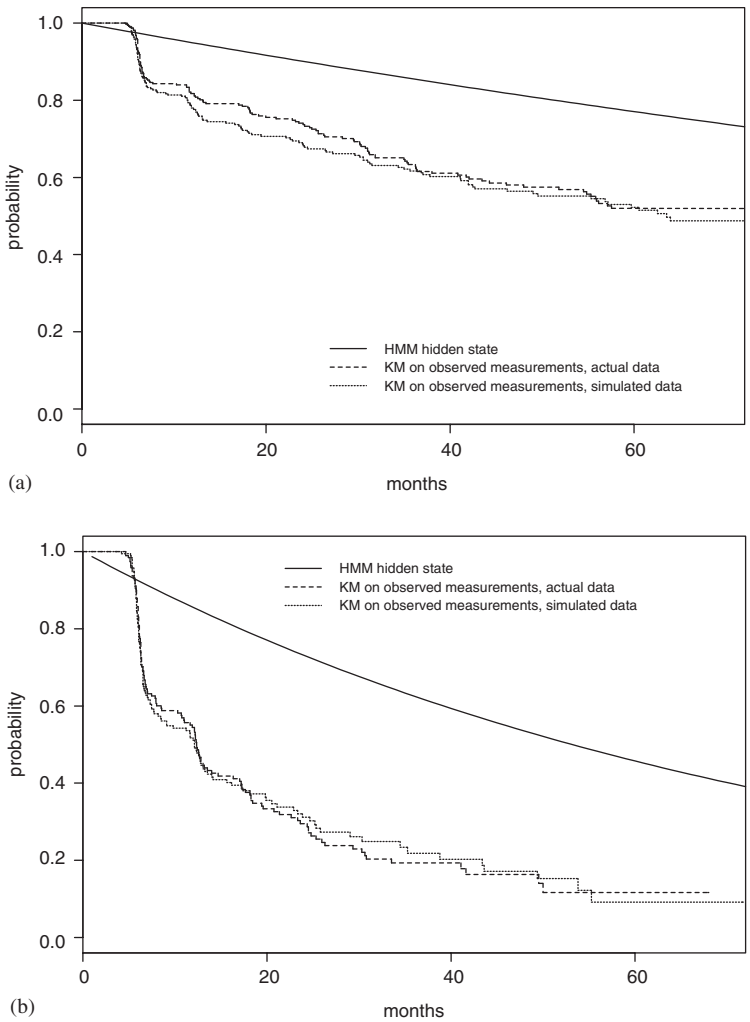

Figure 1. Distribution of (*a*) time to appearance and (*b*) time to clearance of hairy leukoplakia lesions.

Additionally, some covariates may affect the ability to detect hairy leukoplakia. Among the variables measured in the SFMHS, presence of oral candidiasis and treatment with Acyclovir are the only two that were suspected to have an effect on the misclassification probabilities. Table III presents results from models including the above covariates singly. Inferences based on confidence intervals and likelihood ratio tests generally agree at the nominal 5 per cent significance level, except in two cases in which the likelihood ratio test led to rejection while the corresponding confidence intervals for the risk ratio included the null value (one).

The presence of oral candidiasis is strongly associated with acquisition of HL. Also, the probability of a false positive diagnosis of HL is higher when oral candidiasis is present, suggesting that oral candidiasis may be incorrectly diagnosed as HL. Other covariates associated

Table III. Covariate effect estimates for models with a single covariate; study of hairy leukoplakia.

	Acquisition		Clearance	
	Risk ratio	95% confidence interval	Risk ratio	95% confidence interval
Oral candidiasis	19.80	[8.16, 48.07]	1.70	[0.55, 5.25]
AIDS diagnosis	2.11	[1.28, 3.50]	2.78	[1.61, 4.81]
100 units CD4	0.66	[0.51, 0.84]	0.74	[0.48, 1.14]
Smoking	3.94	[1.40,11.05]	1.35	[0.39, 3.08]
Acyclovir	3.17	[0.37, 27.2]	7.84	[0.56,109.14]
	False positive		False negative	
	Odds ratio	95% confidence interval	Odds ratio	95% confidence interval
Oral candidiasis	4.10	[1.81,12.70]	0.83	[0.42, 1.63]

with increased risk of HL acquisition include smoking, reduced CD4 lymphocyte count and a prior AIDS diagnosis. The latter two covariates were also positively associated with HL clearance. This result is counter-intuitive given that both variables are associated with increased immunosuppression, and that HL lesions generally resolve in individuals with improved immune status (for example, those undergoing antiretroviral therapy). A possible explanation of these findings is that individuals with these covariate values are also more likely to have received treatment. We investigate the question further below by fitting a model with those two covariates included. Acyclovir is a drug that may be used at high doses in the treatment of hairy leukoplakia and does indeed seem to be associated with a higher clearance rate. The relatively high variability of the risk ratio estimate is due to incompleteness in the reporting of Acyclovir treatment. (Approximately 40 per cent of the observations were omitted due to incomplete treatment information.) Acyclovir treatment was not found to affect misclassification probabilities. Finally, the per cent difference in CD4 cell count between the previous and the current examination was not strongly associated with either acquisition or clearance of HL.

To investigate the impact of ignoring potential misclassification of HL diagnosis on model estimates and associated inferences, the above covariates were included in conventional Markov models where the HL diagnoses were assumed to represent the true HL status. Resulting parameter estimates had smaller asymptotic variances and were attenuated compared to estimates from the corresponding hidden Markov models (results not shown). A similar effect on the odds ratio is seen when independent binary disease observations are misclassified.

Table IV presents parameter estimates and corresponding confidence intervals from two selected models including multiple covariates. The first is an extension of the previous model with current CD4 lymphocyte count and includes a time-dependent binary indicator of AIDS diagnosis. For acquisition of HL, the results (based on likelihood ratio tests) indicate that the risk ratio for CD4 is significantly lower than one, but not the corresponding ratio for AIDS diagnosis. For clearance of HL the risk ratio for AIDS diagnosis is significantly above one while the ratio for CD4 is not. The observation that AIDS diagnosis and not CD4 lymphocyte count is positively associated with clearance of HL is in agreement with the suggestion that individuals with AIDS are treated more aggressively for HL. The model could be simplified so that acquisition intensity depends only on CD4 lymphocyte counts and clearance intensity

Table IV. Covariate effect estimates for two covariate models; study of hairy leukoplakia.

	Acquisition		Clearance	
	Risk ratio	95% confidence interval	Risk ratio	95% confidence interval
100 units CD4	0.68	[0.51,0.92]	1.03	[0.75,1.44]
AIDS	1.09	[0.62,1.89]	2.19	[1.24,3.85]
100 units CD4	0.65	[0.50,0.85]	0.70	[0.41,1.21]
Smoking	3.39	[1.53,7.47]	1.69	[0.58,4.92]

depends only on AIDS diagnosis. The second model examines the potential confounding effect of HIV disease stage (as measured by current CD4 cell count) on the association between acquisition of HL and a binary indicator of current smoking status. Once adjusted for CD4 cell count, the relative acquisition intensity between smokers and non-smokers is somewhat attenuated, but the increase in log-likelihood from the effect of smoking on acquisition of HL remains substantial.

3.1.4. Predictions based on past observations. The empirical distribution of HL diagnoses at visits $2,3$ and 4 conditional on possible sequences of diagnoses at visits $1,2$ and 3 respectively, $P[Y_{j+1}|Y_1^j, T_1^{j+1}]$, $j=1,2,3$, were calculated from the sequences of observations, both marginally and stratified by level of CD4 count, a covariate associated with acquisition and clearance of HL lesions. Intervals between visits are approximately 6 months. To ensure that the conditional probability estimates are based on sequences with intervals of comparable length, only sequences with visits spaced by less than 8 months were included. Probabilities of the next HL diagnosis based on past observations were also estimated from the marginal HMM and from an HMM with the CD4 strata included as covariates of the hidden process. Probabilities were averaged over identical sequences of past diagnoses. The empirical and predicted probabilities for sequences occurring at least 8 times in the data set are reported in Table V, along with χ^2 statistics computed based on observed and predicted counts. Although not known to follow the nominal distribution, these statistics provide a useful means of comparing models.

The predicted probabilities are generally in good agreement with the empirical probabilities both with and without stratification. The predicted probabilities from the HMM with CD4 strata as covariate of the hidden process follow the changes in observed transitions between the CD4 strata. The χ^2 statistic values within strata for predictions based on the first observation are low and close to the χ^2 value for the basic model. Some features of the observed sequences of diagnoses reveal second-order dependence. For the marginal frequencies, the probability of a positive diagnosis at the third visit given a negative diagnosis at the second visit is higher if the first diagnosis was positive (25 per cent) than if it was negative (4 per cent). The same is true if the second diagnosis was instead positive (81 per cent versus 43 per cent). The HMM predictions for the third observation present a similar pattern.

3.2. Study of HPV infection

3.2.1. Description of the data. A second application of HMMs to recurrent disease outcomes is provided by a study of the natural history of HPV infection in young women conducted in

Table V. Empirical and predicted distributions of HL diagnoses conditional on the sequence of past observations (0 = negative, 1 = positive). Degrees of freedom (k) for χ^2_k statistics are based on the number of cells in the parent subtable ($k+1$).

		No stratification			CD4 count $\leqslant 200$			$200 <$ CD4 count $\leqslant 500$			CD4 count > 500		
Past sequence	Next observation	Empirical		Predicted Frequency	Empirical		Predicted frequency	Empirical		Predicted frequency	Empirical		Predicted frequency
(Y_1, Y_2, Y_3)		Count	frequency		Count	Frequency		Count	Frequency		Count	Frequency	
(1)	0	18	0.32	0.385	10	0.37	0.420	6	0.21	0.364	5	0.45	0.327
	1	39	0.68	0.615	17	0.63	0.580	23	0.79	0.636	6	0.55	0.673
	Total	57			27			29			11		
(0)	0	207	0.92	0.908	45	0.87	0.862	101	0.92	0.905	100	0.96	0.927
	1	18	0.08	0.092	7	0.13	0.138	9	0.08	0.095	4	0.04	0.073
	Total	225			52			110			104		
χ^2_3			1.54			0.28			3.31			2.65	
(1,1)	0	6	0.19	0.304	2	0.25	0.376	4	0.25	0.309			
	1	26	0.81	0.696	6	0.75	0.624	12	0.75	0.691			
	Total	32			8			16					
(1,0)	0	12	0.75	0.621	7	0.78	0.678						
	1	4	0.25	0.379	2	0.22	0.322						
	Total	16			9								
(0,1)	0	8	0.57	0.514									
	1	6	0.43	0.486									
	Total	14											
(0,0)	0	184	0.96	0.934	23	0.82	0.865	59	0.97	0.926	76	0.97	0.956
	1	8	0.04	0.066	5	0.18	0.135	2	0.03	0.074	2	0.03	0.044
	Total	192			28			61			78		
χ^2_7			5.21										
(1,1,1)	0	8	0.31	0.301									
	1	18	0.69	0.699									
	Total	26											
(0,0,0)	0	153	0.96	0.940									
	1	6	0.04	0.060									
	Total	159											

Table VI. Distribution of the number of visits per subject.

Number of examinations	Frequency
1	59
2 to 4	155
5 to 7	156
8 to 10	209
11 to 13	66
14 to 18	18

San Francisco [1]. The human papillomaviruses are a widespread family of viruses that cause a variety of manifestations in human epithelial tissue. Over 70 types of HPV (20–30 of which are genital) have been identified, several of which are known to be sexually transmitted and associated with squamous cell carcinomas of the anus, cervix, penis and vulva. Cancer of the cervix is the most common of these, with 12000 to 14000 new cases reported annually in the U.S., resulting in approximately 7000 deaths. Prevalence of HPV in the genital tract of females between the ages of 15 and 50 is thought to be between 5 per cent and 20 per cent, and HPV DNA has been detected in over 70 per cent of cases of cervical cancer examined [24, 25]. Commonly used methods of testing for HPV are based on detecting viral DNA in samples of tissue, and may have appreciable false negative and false positive rates, due both to properties of the tests and difficulties in obtaining representative samples. Major goals of prospective epidemiological studies of HPV natural history include understanding factors associated with acquisition and clearance of particular viral types, and the relationship between HPV infection history and subsequent progression to cervical dysplasia. Here we focus on the first of these goals.

The study group consists of 895 women who were between 13 and 22 years old at recruitment. Enrolment started in 1990 and is ongoing. Women are scheduled for clinic visits every four months. At each visit, behavioural questionnaires are completed and a clinical examination including Pap smear and collection of cervical cell specimens for DNA testing is conducted. The actual intervals between visits are highly variable due to missed and supplementary visits. In addition, different DNA testing methods have been used over different periods of the study. The present analysis is limited to measurements made with the most recent and most reliable testing method involving PCR amplification of the viral DNA and hybridization with HPV type-specific probes. Results obtained with this test are available for 4400 visits made by 663 women.

Table VI gives the distribution of the number of visits per women included in the analysis. The median length of follow-up was 37 months (range 0 to 82 months). The 59 subjects with only one visit contributed only to the estimate of the initial distribution of HPV infection. The analyses presented below focus on the two most prevalent HPV types or group of types observed in the cohort: types 16 and 18. These types are thought to be associated with an increased risk of cervical dysplasia and cervical cancer.

3.2.2. Estimates from a hidden Markov model. The estimated initial probabilities of the hidden state interpreted as the infection state is 0.189 (SE 0.021) for HPV type 16 and 0.077 (SE 0.018) for HPV type 18. The estimated transition intensities for each process and their

Table VII. Transition intensities estimates for hidden Markov model (transition/month).

HPV type	Acquisition			Clearance		
	Estimate	Standard error	Expected time	Estimate	Standard error	Expected time
16	0.0037	0.0010	270 months	0.0569	0.0077	17.6 months
18	0.0020	0.0006	500 months	0.0875	0.0250	11.4 months

Table VIII. Probability of observation given the hidden state for type-specific PCR test data.

HPV type	$P[\text{negative}\|\text{state}=1]$		$P[\text{positive}\|\text{state}=0]$	
	Estimate	Standard error	Estimate	Standard error
16	0.1879	0.0427	0.0032	0.0030
18	0.3325	0.1167	0.0000	—

inverses, the expected times to an event, are reported in Table VII. Results indicate that in addition to being more frequent initially, HPV type 16 has a higher infection intensity and lower clearance intensity than type 18. By contrast, infections with type 18 appear to be of shorter duration than infections by HPV type 16. Estimates of the distribution of observations conditional on the hidden states for each HPV type are presented in Table VIII. Under the interpretation that $P[\text{negative}|\text{state}=1]$ is the false negative rate of the test, and $P[\text{positive}|\text{state}=0]$ the false positive rate, the present results are in agreement with the claim that PCR tests are very specific. However, sensitivity is lower than might be expected, especially for the type 18 test with a false negative rate of almost one-third.

Kaplan–Meier and hidden Markov model estimates of the distribution of the time to infection and time to clearance for HPV type 16 are plotted in Figure 2. Both Kaplan–Meier estimates drop sharply after four months, which is the typical interval between visits. The distribution of time to clearance estimated from the HMM (Figure 2(b)) agrees quite well with the Kaplan–Meier estimate, which is surprising given the seemingly large estimated proportion of false negatives in Table VIII. Similarly, the apparent difference between estimates of these distributions for time to infection (Figure 2(a)) does not appear to be in accordance with the very low estimated false-positive rate in the table. The distributions of time to infection and time to clearance estimated on a replicate of the data set as described in Section 2.5 are also shown in Figure 2. Taken together, these results differ from those in the analyses of the hairy leukoplakia data, and call into question the interpretation of the observed process as a misclassified realization of a hidden process representing actual HPV status. The agreement between the actual and simulated distributions of the lengths of positivity and negativity is good, as in the HL example. Results for type 18 are similar to those for type 16 and are not presented here.

We investigated the effects of a variety of covariates on transition probabilities for both type 16 and 18. The results (not presented in tabular form) indicate that infection and clearance rates of HPV type 16 tend to decrease as the length of sexual activity prior to entry into the study increases, and that report of sexual contact with a new partner in the interval

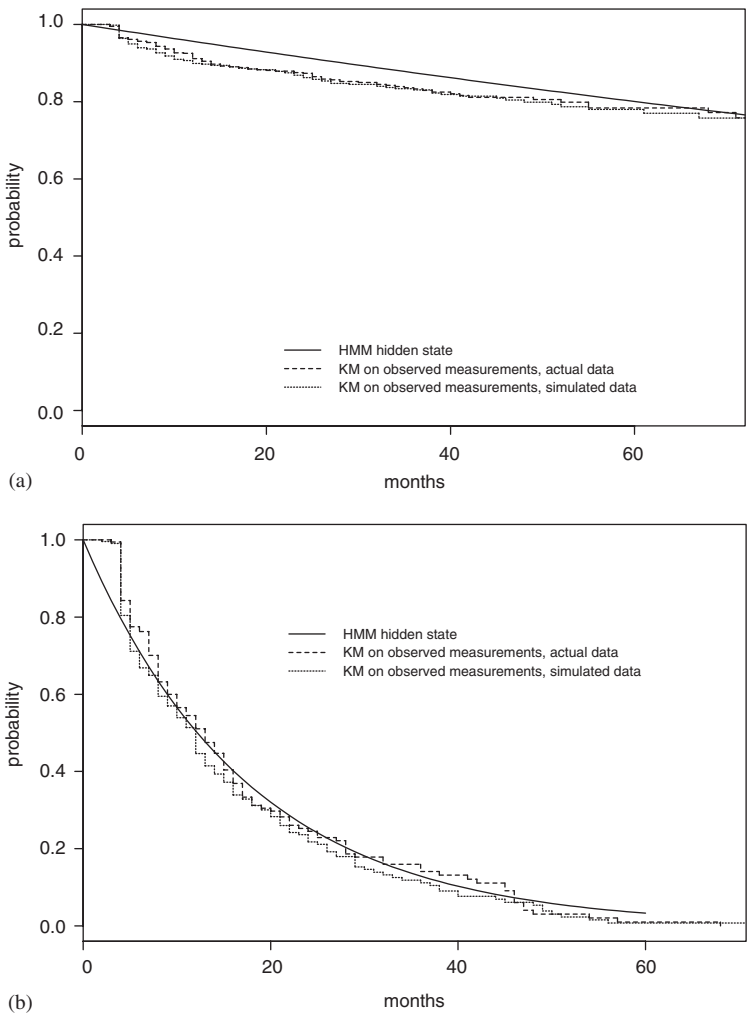

Figure 2. Distribution of (*a*) time to infection (*b*) and time to clearance of HPV type 16.

preceding the current visit is associated with an increase in the infection intensity of both type 16 and 18.

3.2.3. Predictions based on past observations. Empirical and predicted distributions of observed HPV status conditional on previous observations were calculated for HPV type 16 using methods described in Section 3.1.4. Sequences of visits that are less than 6 months apart were used for estimation. Empirical and predicted probabilities are reported in Table IX. The agreement between predicted and empirical results ranges from excellent to poor. Close matches are obtained for past sequences such as $Y_1 = 1$ and $Y_1 = 0,\ Y_2 = 0$, while for the sequence $Y_1 = 0$ the predictions are far from the empirical values. The χ^2 statistic for

Table IX. Empirical and predicted distributions of HPV 16 status conditional on the sequence of past observations (0 = negative, 1 = positive).

Past sequence	Next observation	Empirical		Predicted frequencies
(Y_1, Y_2)		Count	Frequencies	
(1)	0	18	0.43	0.448
	1	24	0.57	0.552
	Total	42		
(0)	0	294	0.964	0.917
	1	11	0.036	0.083
	Total	305		
χ^2_3			8.89	
(1,1)	0	6	0.35	0.456
	1	11	0.65	0.544
	Total	17		
(1,0)	0	11	0.73	0.819
	1	4	0.27	0.181
	Total	15		
(0,1)	0	1	0.14	0.434
	1	6	0.86	0.566
	Total	7		
(0,0)	0	205	0.962	0.972
	1	8	0.038	0.028
	Total	213		
χ^2_7			4.60	

predictions given a single past observation (computed as described in Section 3.1.4) is much higher than for the model fitted to the hairy leukoplakia data while for predictions based on two observations it is close to the value for the hairy leukoplakia data. However, since the number of occurrences of three out of four possible sequences of the first two observations is small in the HPV data, the variability of the observed counts is higher, limiting the precision of comparison of observations to predictions in investigating second-order dependence for HPV.

4. DISCUSSION

Longitudinal observations of recurring disease outcomes are frequently characterized by interval censored observations, misclassification of measured outcomes and unknown disease history of individuals prior to enrolment. Hidden Markov models have a number of features that accommodate these difficulties. The Markov and conditional independence assumptions lead to a relatively simple model structure which easily accommodates the irregular observations of repeated outcomes typical in many studies, and allows simultaneous investigation of all possible transitions between disease states. In addition, with the modifications for inclusion of covariates and estimation of variances proposed here, inference for a variety of covariate effects can be based on techniques familiar from generalized linear models.

Despite their appeal, the attractive features of HMMs derive from assumptions that may be inappropriate in some applications and must be checked thoroughly before conclusions can be considered valid. For example, the constant intensity assumption may not hold in practice. We propose some diagnostic techniques which allow fit of the models to be assessed, and provide indirect information on violation of these key assumptions. In particular, comparison of empirical and model predicted distributions of a next observation based on the past observations indicates whether the HMM captures higher order dependencies in the observation sequences. A good fit of the model estimates to those empirical probabilities is however not sufficient to validate the assumptions of the model. Agreement between external estimates of misclassification probabilities and model estimates is an important confirmation of the interpretation of the parameter estimates.

Of the two case studies considered here, the analysis of hairy leukoplakia presented in Section 3.1 provides a good example of a successful application of hidden Markov models. The estimated transition rates support clinical observations that hairy leukoplakia tends to persist for long periods of time, and estimated misclassification probabilities correspond well with documented misclassification of diagnoses. In addition, observed covariate effects are consistent with and expand the findings of a previous analysis restricted to development of a first lesion among subjects free of the condition at baseline [23]. New findings include observed associations between clearance of hairy leukoplakia and indicators of current Acyclovir treatment and diagnosis of AIDS, the effect of the latter probably reflecting unmeasured changes in therapy.

Hidden Markov models fitted to the HPV infection data from the San Francisco study (Section 3.2) were more difficult to interpret. Although the model was sufficiently flexible to provide a reasonably good fit to the distribution of observed states occupancy times, predictions of subsequent outcomes conditional on past values were poor in some cases. In addition, estimates of transition probabilities for the hidden states do not seem to support the interpretation of the observed process as a misclassified realization of the hidden process. This may be due to time variation in the transition intensities, possibly the result of progressive immunity to HPV, or to unobserved factors influencing persistence of infection. Fitting a hidden Markov model with constant intensities in that situation could lead to intensity estimates that are too high. An alternative explanation which cannot be ruled out is fluctuation in the false negative rate over time as a function of the concentration of virus. During prolonged periods of low virus concentration, high false negative rates would tend to produce long false negative sequences that an HMM with constant misclassification probabilities would fit as true negative. Despite these apparent failures of the model, the analysis yields important information about the dependence in serial HPV infection tests, confirms the importance of several covariates, and suggests modifications which might result in a more appropriate (and substantially more complex) modelling approach.

There are a number of avenues for further research in applications of hidden Markov models to studies of infectious diseases. Several disease outcomes may be measured on the same patients, for example infection by different types of HPV. These outcomes are likely to be dependent. Extensions of the hidden Markov model to multiple interacting hidden processes could prove useful to the study of the relation between distinct disease outcomes. In addition, modifications to allow time varying intensities and more sophisticated diagnostic techniques would be extremely valuable.

APPENDIX

A1. EM algorithm for hidden Markov models

Although the EM algorithm was popularized by the work of Dempster *et al.* [26], an EM algorithm for hidden Markov models had previously been described in Baum *et al.* [15]. To apply EM to the hidden Markov model of Section 2.2, the observations Y are augmented with the realizations of the hidden process X to form the 'complete data'. Assuming now that the X are observed, the log-likelihood of the 'complete data' for one subject (dropping the subscript i) can be written as

$$\ell^{c}(\theta) = \log \pi_{x_1|z_1} + \sum_{j=2}^{m} \log P_{x_{j-1},x_j|z_{j-1}}(t_j - t_{j-1}) + \sum_{j=1}^{m} \log f(y_j|x_j, z_j) \tag{A1}$$

The idea behind the EM algorithm is that $\ell(\theta)$ is easier to maximize than the log-likelihood itself. Since in reality the X are unobserved, (A1) is replaced by its expectation over X with respect to the current parameter values θ', conditional on the observed data (Y, T and Z) to define the expected 'complete data' log-likelihood

$$\begin{aligned}\mathscr{L}(\theta'|\theta) = &\sum_{s} v_1(s) \log \pi_{s|z_1} + \sum_{j=2}^{m} \sum_{s_1,s_2} w_j(s_1,s_2) \log P_{s_1,s_2|z_{j-1}}(t_j - t_{j-1}) \\ &+ \sum_{j=1}^{m} \sum_{s} v_j(s) \log f(y_j \mid s, z_j)\end{aligned} \tag{A2}$$

where

$$\begin{aligned} v_j(s) &= \Pr(X_j = s \mid Y_1^m, T_1^m, Z_1^m, \theta) \\ w_j(s_1,s_2) &= \Pr(X_{j-1} = s_1, X_j = s_2 \mid Y_1^m, T_1^m, Z_1^m, \theta)\end{aligned}$$

Baum *et al.* [15] and Dempster *et al.* [26] have shown that maximizing $\mathscr{L}(\theta'|\theta)$ also maximizes the (observed data) log-likelihood. The EM algorithm alternates between two steps to maximize (A2): (i) in the E-step, $\mathscr{L}(\theta'|\theta)$ is computed at the current parameter values θ – this amounts to computing the weights $v(s)$ and $w(s_1,s_2)$; (ii) in the M-step, $\mathscr{L}(\theta'|\theta)$ is maximized with respect to θ' to yield updated parameter estimates. Iterating these two steps guarantees convergence to a (local) maximum of $\mathscr{L}(\theta'|\theta)$ and of the log-likelihood:

E step. The weights $v_k(s)$ and $w_k(s_1,s_2)$ are computed using the *forward-backward* algorithm of Baum *et al.* [15]

$$\begin{aligned} v_j(s) &= \alpha_j(s)\beta_j(s) \\ w_j(s_1,s_2) &= \alpha_{j-1}(s_1) f(y_j \mid s_1, z_j) P_{s_1,s_2|z_{j-1}}(t_j - t_{j-1})\beta_j(s_2)\end{aligned}$$

where the forward and backward probabilities, α and β, are computed recursively as

$$\begin{aligned} \alpha_1(s) &= \pi_{s|z_1} f(y_1 \mid s, z_1) \\ \alpha_j(s) &= \sum_{r} \alpha_{j-1}(r) P_{r,s|z_{j-1}}(t_k - t_{k-1}) f(y_j \mid s, z_j)\end{aligned}$$

$$\beta_m(s) = 1$$
$$\beta_j(s) = \sum_r P_{s,r|z_j}(t_{j+1} - t_j) f(y_{j+1} \mid r, z_{j+1}) \beta_{j+1}(r)$$

M step. The parameters of the terms of $\mathscr{L}(\theta' | \theta)$ relating to the hidden process are distinct from those of the term relating to the distribution of the observations given the hidden states. Maximization with respect to each set of parameters can therefore be carried separately. Let $\theta = (\theta_Y, \theta_S)$, the parameters of the distribution of the observations and of the hidden process, respectively. Then the function

$$\mathscr{L}_1(\theta'_Y \mid \theta) = \sum_{j=2}^{m} \sum_s v_j(s) \log f(y_s \mid s, z_j, \theta'_Y)$$

is maximized with respect to θ'_Y and

$$\mathscr{L}_2(\theta'_S \mid \theta) = \sum_s v_j(s) \log \pi_{s|z_1}(\theta'_S) + \sum_{j=2}^{m} \sum_{s_1, s_2} w_j(s_1, s_2) \log P_{s_1, s_2 | z_{j-1}}(t_j - t_{j-1}, \theta'_S)$$

is maximized with respect to θ'_S by the gradient descent method.

A2. Computation of the information matrix of the observed data

The computation of the information of the observed data is based on the result derived by Louis [27] for the EM algorithm that expresses the information of the observed data as a function of the complete data observed information and the complete data score function. Let Y denote all the sequences of observations, and X all the sequences of realizations of the hidden process. All probability computations are implicitly conditional on the observation times T and the vector of covariates Z, but we drop them from the notation. Louis's result may be written

$$I_Y(\theta) = E[I_{(X,Y)}(\theta) | Y] - [E[S_{(X,Y)}(\theta) S^{\mathrm{T}}_{(X,Y)}(\theta) | Y] - E[S_{(X,Y)}(\theta) | Y] E[S_{(X,Y)}(\theta) | Y]^{\mathrm{T}}] \qquad \text{(A3)}$$

where $I_Y(\theta)$ is the information of the observed data, $I_{(X,Y)}(\theta)$ is the negative of the matrix of second derivatives of the complete data log-likelihood, and $S_{(X,Y)}(\theta)$ is the vector of first derivatives of the complete data log-likelihood. All expectations are taken with respect to the distribution $P[X | Y, \theta]$.

Hughes [9] derived the specific form of the terms in (A3) for a discrete-time hidden Markov model. Conditional on the observation times, the same expressions can be used for the continuous-time case. The computation of the second derivatives of the transition matrix over a given time interval with respect to the parameters of the model becomes, however, more difficult. Expressions for these derivatives in terms of the second derivatives of the intensity matrix were taken from Kosorok and Chao [28]. Evaluation of the second term of (A3) requires summing the values of the first derivatives over the joint distributions of all pairs of observations in a sequence $P[X_t, X_s | Y]$, a computationally demanding operation. Lystig and Hughes [29] recently extended the forward-backward algorithm to compute $I_Y(\theta)$ directly, an approach that both simplifies and speeds-up the computation.

ACKNOWLEDGEMENTS

We would like to thank Dr C. Shiboski (University of California, San Francisco, Oral AIDS Center; National Institute of Dental Research grant PO1 DE07946) and Dr B. Moscicki (University of California, San Francisco, Division of Adolescent Medicine; National Cancer Institute grant RO1 CA 51323-07) for their co-operation and permission to use data from their studies. Dr Bureau's work was supported in part by a scholarship from Fonds pour la Recherche et la Formation de Chercheurs (Québec).

REFERENCES

1. Moscicki A-B, Shiboski S, Broering J, Powell K, Clayton L, Jay N, Darragh TM, Brescia R, Kanowitz S, Miller SB, Stone J, Hanson E, Palefsky J. The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women. *Journal of Pediatrics* 1998; **132**(2):277–284.
2. Nagelkerke NJD, Chunge RN, Kinoti SN. Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 1990; **9**:1211–1219.
3. Shiboski CH, Hilton JF, Greenspan D, Westenhouse JL, Derish P, Vranizan K, Lifson AR, Canchola A, Katz MH, Cohen JB. Human immunodeficiency virus-related oral manifestations and gender: a longitudinal analysis. *Archives of Internal Medicine* 1996; **156**:2249–2254.
4. Satten GA, Longini IM. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics* 1996; **45**(3):275–295.
5. Clogg CC. Latent class models: recent developments and prospects for the future, In *Handbook of Statistical Modeling in the Social Sciences*, Arminger G, Clogg CC, Sobel ME (eds). Plenum: New York, 1995; 311–359.
6. MacDonald IL, Zucchini W. *Hidden Markov Models and Other Models for Discrete-valued Time Series*. Chapman and Hall: New York, 1997.
7. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 1989; **77**(2):257–285.
8. Hughes JP. *A Class of Stochastic Models for Relating Synoptic Atmospheric Patterns to Local Hydrologic Phenomena*. PhD thesis, Department of Statistics, University of Washington, 1993.
9. Hughes JP. Computing the observed information in the hidden Markov model using the EM algorithm. *Statistics & Probability Letters* 1997; **32**:107–114.
10. Kalbfleisch JD, Lawless J. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association* 1985; **80**:863–871.
11. Gentleman R, Lawless J, Lindsey JC, Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine* 1994; **13**:805–821.
12. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences* 1987; **84**:2363–2367.
13. Newton MA, Guttorp P, Catlin S, Assunção R, Abkowitz JL. Stochastic modeling of early hematopoiesis. *Journal of the American Statistical Association* 1995; **90**(432):1146–1155.
14. Kirby AG, Spiegelhalter DJ. Modeling the precursors of cervical cancer, In *Case Studies in Biometry*, Lange N, Ryan L, Billard L, Brillinger D, Conquest L, Greenhouse J (eds). Wiley: New York, 1994; 359–383.
15. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* 1970; **41**(1):164–171.
16. Bureau A, Hughes JP, Shiboski S. An S-PLUS implementation of hidden Markov models in continuous time. *Journal of Computational and Graphical Statistics* 2000; **9**(4):621–632.
17. Bickel PJ, Ritov Y, Rydén T. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics* 1998; **26**(4):1614–1635.
18. Greenspan J, Greenspan D, Lennette ET, Abrams DI, Conant MA, Petersen V, Freese UK. Replication of Epstein-Barr virus within the epithelial cells of oral 'hairy' leukoplakia, an AIDS-associated lesion. *New England Journal of Medicine* 1985; **313**:1564–1571.
19. Katz M, Greenspan D, Westenhouse J, Hessol NA, Buchbinder SP, Lifson AR, Shiboski S, Osmond D, Moss A, Samuel M *et al.* Progression to AIDS in HIV-infected homosexual and bisexual men with hairy leukoplakia and oral candidiasis: results from 3 epidemiologic cohorts. *AIDS* 1992; **6**:95–100.
20. Hilton JF, Alves M, Anastos K, Canchola AJ, Cohen M, Delapenha R, Greenspan D, Levine A, MacPhail LA, Micci SJ, Mulligan R, Navazesh M, Phelan J, Tsaknis P. Accuracy of diagnoses of HIV-related oral lesions by medical clinicians. Findings from the Women's Interagency HIV Study. *Community Dentistry and Oral Epidemiology* 2001; **29**:362–372.
21. Hilton JF, Donegan E, Katz MH, Canchola AJ, Fusaro RE, Greenspan D, Greenspan JS. Development of oral lesions in human immunodeficiency virus-infected transfusion recipients and hemophiliacs. *American Journal of Epidemiology* 1997; **145**(2):164–173.

22. Winkelstein WJ, Lyman DM, Padian N, Grant R, Samuel M, Wiley JA, Anderson RE, Lang W, Riggs J, Levy JA. Sexual practices and risk of infection by the human immunodeficiency virus. The San Francisco Men's Health Study. *Journal of the American Medical Association* 1987; **257**:326–330.
23. Shiboski, CH, Neuhaus JM *et al.* The effect of receptive oral sex and smoking on the incidence of hairy leukoplakia in HIV-positive gay men. *Journal of AIDS* 1999; **21**:236–242.
24. Koutsky LA, Galloway DA, Holmes KK. Epidemiology of genital human papillomavirus infection. *Epidemiologic Reviews* 1988; **10**:122–163.
25. Schiffman MH. Epidemiology of cervical human papillomavirus infections. *Current Topics in Microbiology and Immunology* 1994; **186**:55–81.
26. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1977; **39**(1):1–38.
27. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1982; **44**(2):226–233.
28. Kosorok MR, Chao W-H. The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association* 1996; **91**(434):807–817.
29. Lystig TC, Hughes JP. Baum's algorithm extended: Likelihood calculations for hidden Markov Models. *Journal of Computational and Graphical Statistics* 2002 (in press).

