The Statistician (2003)
52, *Part* 2, *pp.* 193–209

Multistate Markov models for disease progression with classification error

Christopher H. Jackson,

Imperial College School of Medicine, London, UK

Linda D. Sharples and Simon G. Thompson,

Medical Research Council Biostatistics Unit, Cambridge, UK

Stephen W. Duffy

Cancer Research UK, London, UK

and Elisabeth Couto

Cancer Research UK, Oxford, UK

[Received February 2002. Revised December 2002]

Summary. Many chronic diseases have a natural interpretation in terms of staged progression. Multistate models based on Markov processes are a well-established method of estimating rates of transition between stages of disease. However, diagnoses of disease stages are sometimes subject to error. The paper presents a general hidden Markov model for simultaneously estimating transition rates and probabilities of stage misclassification. Covariates can be fitted to both the transition rates and the misclassification probabilities. For example, in the study of abdominal aortic aneurysms by ultrasonography, the disease is staged by severity, according to successive ranges of aortic diameter. The model is illustrated on data from a trial of aortic aneurysm screening, in which the screening measurements are subject to error. General purpose software for model implementation has been developed in the form of an R package and is made freely available.

Keywords: Abdominal aortic aneurysm; Disease progression; Hidden Markov models; Multistate models

1. Introduction

When a disease is detected at an early stage, it may be more amenable to treatment. Systematic screening of a population can therefore be an effective way of reducing mortality from a disease. Establishing an adequate screening policy for the disease requires knowledge of its natural history. The type of individuals to screen, and the time of the screen, should be chosen according to the risk of onset of the disease, whereas the intervals between successive screens should be chosen according to the risk of progression. Both risks may vary with characteristics of the patients and the risk of progression may vary with the current stage of the disease.

Address for correspondence: Christopher H. Jackson, Department of Public Health and Epidemiology, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London, W2 1PG, UK.
E-mail: chris.jackson@ic.ac.uk

 0039–0526/03/52193

1.1. Multistate models

Multistate Markov models in continuous time are often used to model the course of diseases. A commonly used model is illustrated in Fig. 1. This represents a series of successively more severe stages of disease and an 'absorbing' state, often death. The patient may advance into or recover from adjacent stages of disease or die at any stage of disease. Observations of the stage $S_i(t)$ are made on a number of individuals i at arbitrary times t, which may vary between individuals. The stages of disease may be modelled as a homogeneous continuous time Markov process. Then the model of Fig. 1 is described by a transition intensity matrix Q

$$Q = \begin{pmatrix} q_{11} & q_{12} & 0 & 0 & \dots & q_{1n} \\ q_{21} & q_{22} & q_{23} & 0 & \dots & q_{2n} \\ 0 & q_{32} & q_{33} & q_{34} & \ddots & q_{3n} \\ 0 & 0 & q_{43} & q_{44} & \ddots & q_{4n} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

whose rows sum to 0, so that the diagonal entries are

$$q_{rr} = -\sum_{s \neq r} q_{rs}.$$

Maximum likelihood estimates for this class of models can be computed from the transition probability matrix $P(t)$, with (r, s) entry

$$p_{rs}(t) = \Pr\{S_i(t+u) = s | S_i(u) = r\}.$$

This depends on the unknown parameters in Q through the Kolmogorov relationship $P(t) = \exp(tQ)$ (Cox and Miller, 1965).

This model, with varying numbers of transient states, has been used in applications such as hepatic cancer (Kay, 1986), diabetic nephropathy (Andersen, 1988), human immunodeficiency virus infection (Longini *et al.*, 1989; Gentleman *et al.*, 1994), chronic post-transplant problems (Klotz and Sharples, 1994), screening for breast cancer (Duffy and Chen, 1995) and diabetic retinopathy (Marshall and Jones, 1995). The disease under study is sometimes assumed to be irreversible, in which case the transition intensities corresponding to recovery are taken as 0.

More general models of this type can be constructed to allow for additional states, representing, for example, periods of treatment, hospital stays or competing causes of death. Computing

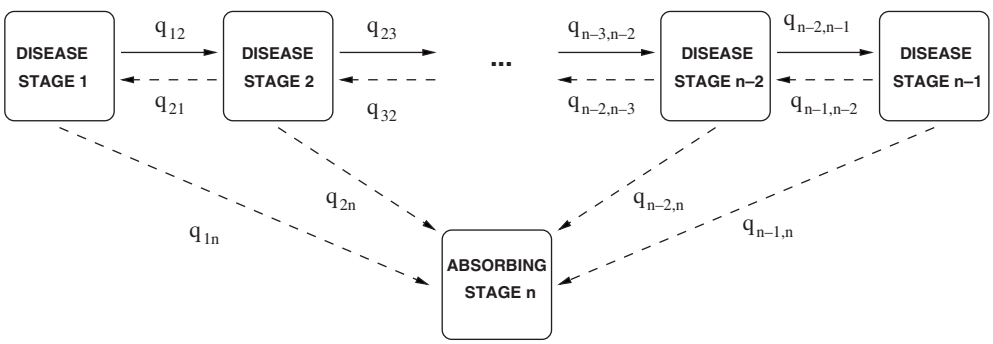

Fig. 1. General model for disease progression

likelihoods is made simpler if the observation times are equally spaced, allowing a discrete time Markov process to be used. See, for example, Aalen *et al.* (1997). The Markov assumption, essentially, that the future of the process depends on the current state, and not on the history of the process, would also be more easy to assess if the exact times of transition between the states are known. However, the assumption of homogeneity of transition rates through time and across individuals can be assessed, for discrete time and continuous time Markov models, by modelling transition rates on observed covariates, as described in Section 2.

1.2. Screening and classification error

The screening process can sometimes be subject to error. Then the Markov disease process $S_i(t)$ is not observed directly, but through realizations $O_i(t)$. The quality of a diagnostic test is often measured by the probabilities that the true and observed states are equal, $\Pr\{O_i(t) = r | S_i(t) = r\}$. Where r represents a 'positive' disease stage, this is the *sensitivity*, or the probability that a true positive result is detected by the test. Where r represents a 'negative' or disease-free stage, this represents the *specificity*, or the probability that, given that the condition of interest is absent, the test produces a negative result.

Chen *et al.* (1996) described a hidden Markov model for breast cancer screening, which simultaneously estimated progression rates between grades of cancer and the sensitivity, specificity and positive predictive value of the screening test. A generalized estimating equation method was used for estimation. Kirby and Spiegelhalter (1994) described a similar model, with false positive and negative rates and progression parameters estimated simultaneously, in the context of cervical smear tests. Bayesian posterior estimates were produced by using the Gibbs sampler. Kirby and Spiegelhalter (1994) presented three-state Markov models, representing the absence of disease, preclinical detectable disease and clinical disease. Misclassification was only modelled between the disease-free and preclinical stages. Stages of disease are often related to levels of a continuous biomarker. For example, Satten and Longini (1996) and Guihenneuc-Jouyaux *et al.* (2000) used hidden Markov models for the progression between stages of HIV infection, monitored by serial measurements of CD4 cell counts. The observation $O_i(t)$ was taken to be the continuous biomarker, rather than a discrete stage of disease, with some distribution dependent on $S_i(t)$. Jackson and Sharples (2002) applied three- and four-state hidden Markov models to functional decline after lung transplantation, using both discrete and continuous responses.

This paper builds on these methods by presenting a general framework for staged Markov models with misclassification, introduced in Section 2. Any form of transition matrix is permitted, with misclassifications between any pair of states. Explanatory variables for both the rates of transition between underlying stages and the probabilities of misclassification can be modelled by using generalized regressions. Section 3 introduces an application to screening for abnormal growth of abdominal aortic aneurysms. The specific models that are used in the application are described in Section 4, followed by a presentation and discussion of the results.

2. General hidden Markov model

A general model for disease progression and diagnosis error can be defined as follows. Let i index I individuals, let j index the m_i observation times for each individual i and let n be the number of states. Suppose that $S_{ij} = S_i(t_{ij})$ represents the true underlying state of individual i at time t_{ij}, and O_{ij} represents the corresponding observed state.

The S_{ij} for individual i are assumed to be realizations from a multistate process $S_i(t)$ with intensity matrix Q. The (r, s) entry of Q represents the instantaneous hazard of progression to

stage s, conditionally on occupying stage r:

$$q_{rs} = q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} [\Pr\{S_i(t + \delta t) = s | S_i(t) = r, \mathcal{F}_t\}/\delta t],$$

where $\mathcal{F}_t$ is the observation history of the process up to the time preceding t. If the process $S_i(t)$ is assumed to be *time homogeneous* and *Markov*, then $q_{rs}(t, \mathcal{F}_t)$ is independent of t and $\mathcal{F}_t$. Often the *mean sojourn time* in a transient state r is also of interest—in a unidirectional progression model this is estimated as $1/q_{ii}$, the mean time until leaving the state. The O_{ij} are generated conditionally on S_{ij} according to a *misclassification matrix* E. This is an $n \times n$ matrix, whose (r, s) entry is

$$e_{rs} = \Pr\{O(t_{ij}) = s | S(t_{ij}) = r\},$$

which we first assume to be independent of time t. Analogously to the entries of Q, some of the e_{rs} may be fixed to reflect knowledge of the diagnosis process. For example, the probability of misclassification may be negligibly small for non-adjacent states of disease.

This is an example of a *hidden Markov model*. The progression through underlying states and the observation process of the underlying states are governed by separate models. This class of model is commonly used in areas such as speech and signal processing (Juang and Rabiner, 1991) and the analysis of biological sequence data (Durbin *et al.*, 1998). In engineering and biological sequencing applications, the Markov process usually evolves over an equally spaced discrete 'time' space. Therefore most of the theory of hidden Markov model estimation was developed for discrete time models. Hidden Markov models have less frequently been used in medicine, where continuous time processes are often more suitable. A disease process evolves in continuous time, and patients are often monitored at irregular and differing intervals.

2.1. Covariates

Explanatory variables can be included at each level of the model through generalized regressions. A proportional hazards model can be used to relate the transition intensities $q_{rs}(t)$ at time t to covariates $z(t)$ at that time,

$$q_{rs}\{t, z(t)\} = q_{rs} \exp\{\beta_{rs}^{\mathrm{T}} z(t)\}.$$

The new Q is then used to calculate likelihoods. If the covariates $z(t)$ are time dependent, the Markov chain transition probability $p_{rs}(t_2 - t_1)$ between states r and s from time t_1 to time t_2 is replaced by

$$p_{rs}\{t_2 - t_1, z(t_1)\}$$

although this requires that the value of the covariate is known at every observation time t_1. Sometimes covariates are observed at different times to the main response, e.g. recurrent disease events or other biological markers. In some of these cases it could be assumed that the covariate is a step function which remains constant between its observation times. Even if the value of the covariate is known at all times, e.g. the age of a patient, then the transition intensities between times t_1 and t_2 are still assumed to depend only on the covariate at time t_1.

Similarly, to investigate explanatory variables $w(t)$ for the probability of misclassification e_{rs} for each pair of stages r and s, a logistic model can be used:

$$\log\left\{\frac{e_{rs}(t)}{1 - e_{rs}(t)}\right\} = \gamma_{rs}^{\mathrm{T}} w(t). \tag{1}$$

2.2. Maximum likelihood estimation

A type of EM algorithm known as the *Baum–Welch* or *forward–backward* algorithm (Baum and Petrie, 1966; Baum *et al.*, 1970) is commonly used for hidden Markov model estimation in discrete time applications. See, for example, Durbin *et al.* (1998) and Albert (1999). A generalization of this algorithm to continuous time was described by Bureau *et al.* (2000).

Macdonald and Zucchini (1997), pages 77–79, described a direct method of calculating likelihoods in discrete or continuous time based on matrix products. Similar methods were presented by Lindsey (1999), page 73, and Guttorp (1995). Satten and Longini (1996) used this method to calculate likelihoods for a hidden Markov model in continuous time with observations of a continuous marker generated conditionally on underlying discrete states. The matrix product method is now illustrated for the misclassification model. However, it can be generalized to any form of data generated conditionally on states of a hidden Markov process.

Patient i's contribution to the likelihood is

$$\begin{aligned} L_i &= \Pr(O_{i1},\ldots,O_{im_i}) \\ &= \sum \Pr(O_{i1},\ldots,O_{im_i}|S_{i1},\ldots,S_{im_i}) \Pr(S_{i1},\ldots,S_{im_i}) \end{aligned} \tag{2}$$

where the sum is taken over all possible paths of underlying states $S_{i1},\ldots,S_{im_i}$. Assume that the observed states are conditionally independent given the values of the underlying states. Also assume the Markov property

$$\Pr(S_{ij}|S_{i,j-1},\ldots,S_{i1}) = \Pr(S_{ij}|S_{i,j-1}).$$

Then the contribution L_i can be written as a product of matrices, as follows. To derive this matrix product, decompose the overall sum in equation (2) into sums over each underlying stage. The sum is accumulated over the unknown first state, the unknown second state and so on until the unknown final state:

$$\begin{aligned} L_i = &\sum_{S_{i1}} \Pr(O_{i1}|S_{i1}) \Pr(S_{i1}) \sum_{S_{i2}} \Pr(O_{i2}|S_{i2}) \Pr(S_{i2}|S_{i1}) \sum_{S_{i3}} \Pr(O_{i3}|S_{i3}) \Pr(S_{i3}|S_{i2}) \\ &\ldots \sum_{S_{im_i}} \Pr(O_{im_i}|S_{im_i}) \Pr(S_{im_i}|S_{in_{i-1}}) \end{aligned} \tag{3}$$

where $\Pr(O_{ij}|S_{ij})$ is the misclassification probability $e_{S_{ij}O_{ij}}$. $\Pr(S_{i,j+1}|S_{ij})$ is the $(S_{ij}, S_{i,j+1})$ entry of the Markov chain transition matrix $P(t)$ evaluated at $t = t_{i,j+1} - t_{ij}$. Let $\mathbf{f}$ be the vector of initial stage occupation probabilities $\Pr(S_{i1})$, and let $\mathbf{1}$ be a column vector consisting of 1s. For $j = 2,\ldots,m_i$ let T_{ij} be the $n \times n$ matrix with (r,s) entry

$$e_{sO_{ij}}\, p_{rs}(t_{ij} - t_{i,j-1}).$$

Then subject i's likelihood contribution is

$$L_i = \mathbf{f} T_{i2} T_{i3},\ldots, T_{im_i}\mathbf{1}. \tag{4}$$

If $S(t_j) = D$ is an absorbing state such as death, measured without error, whose entry time is known to within 1 time unit, then the contribution to the likelihood is summed over the unknown state on the day before entry. Assuming a time unit of days, the (r,s) entry of T_{ij} is then

$$p_{rs}(t_j - 1 - t_{j-1})\; p_{s,D}(1).$$

Numerical methods must be used to maximize the likelihood, e.g. a quasi-Newton algorithm as described by Dennis and Schnabel (1983), which does not require the specification of the derivatives of the function to be maximized. Approximate standard errors of the parameter estimates

can be obtained from the inverse of the matrix of second derivatives (Hessian) of the maximized log-likelihood, which can be estimated by using finite differences. Standard errors for functions of parameters, such as mean sojourn times derived from transition rates, can be estimated by using the delta method. The delta method expands a differentiable function of a random variable about its mean, usually with a first-order Taylor approximation, and then takes the variance. For example, for a random vector X and differentiable function g, an approximation to the covariance matrix of $g(X)$ is given by

$$\mathrm{cov}\{g(X)\} \approx g'(\mu)\Sigma g'(\mu)^{\mathrm{T}}$$

where μ is an estimate of the expectation of X and Σ is an estimate of the covariance matrix of X. Oehlert (1992) discussed regularity conditions under which such approximations are accurate. Alternatively, to verify the accuracy of the approximations, standard errors may be estimated by bootstrap resampling methods.

A package of functions for the free statistical software R (`http://www.r-project.org`) was written to fit multistate Markov models in continuous time with or without classification error, using these methods. These are capable of estimating any form of Markov transition intensity matrix and misclassification matrix, with any number of covariates on either of these matrices. Brief details of how this program is run are given in Appendix A.

High dimensional optimization will produce flat likelihood surfaces if there is insufficient information in the data, leading to failure of convergence. In such cases, simplified models might be constructed with some unstable parameters fixed at values obtained from prior information. In applications to disease progression, adjacent states can often be combined to simplify the model. Some adjustments to the numerical methods may also aid convergence for weakly identifiable models; these are described briefly in Appendix A.

3. Application to screening for abdominal aortic aneurysms

Abdominal aortic aneurysms are common in elderly men. Ruptured aneurysms are responsible for 2.1% of all deaths in men and 0.75% of deaths in women over the age of 65 years in England and Wales (Office for National Statistics, 1995). The main biological marker for the risk of aneurysm rupture is the aortic diameter (Darling, 1970; Szilagyi *et al.*, 1972; Vardulaki *et al.*, 1998) measured by using ultrasonography. The only known remedial action is surgical repair. After rupture, death is very rapid if emergency surgery is not carried out quickly. Even if the patient reaches hospital alive and receives surgery, emergency repair often has a poor outcome. This has led to the hypothesis that elective repair of asymptomatic aneurysms detected by screening may be effective. A randomized controlled trial, the Multi-centre Aneurysm Screening Study (MASS), has recently been conducted in the UK to assess the benefits of screening for aortic aneurysms by ultrasonography (Multicentre Aneurysm Screening Group, 2002).

Establishing an adequate screening policy requires studying the rates of progression of aortic diameter. The intervals between screens should be chosen according to the risk of progression. This risk may vary with the current stage of the disease or with characteristics of the patients. Multistate modelling allows an estimation of the rates of progression. Disease stages in this case can be defined by successive intervals of aortic diameter. The stages used in this application are as follows: stage 1, less than 30 mm; stage 2, 30–44 mm; stage 3, 45–54 mm; stage 4, 55 mm and above. These are the stages that are often used to determine the time to the next screen, e.g. 1 year and 3 months for states 2 and 3 in the MASS trial. Patients are considered for surgery when in stage 4 or following an increase of 10 mm or more per year.

We study data from 156 men from a smaller preclinical trial based in Chichester, UK (Scott *et al.*, 1995). These men all had more than one screen, on account of becoming aneurysmal (entering stage 2). The remaining trial population appeared normal (stage 1) at the initial screen and were not offered additional screens; thus they are not included in our study. Therefore, to estimate progression rates from stage 1, we included data from an additional set of 683 men who were aneurysm free on entry and were rescreened every 2 years. These were studied alongside the Chichester trial cohort, to determine prospectively the rate of incidence of aneurysm in a normal population. Thus we have screening data from a total of 839 individuals.

Otherwise the screening policy was the same as in the MASS trial, except that surgery was offered for a diameter of 60 mm or above or following an increase of 10 mm or more per year. This is a 'doctor's care' screening policy, under the definitions of Grüger *et al.* (1991). Under such a scheme the sampling times are not informative and the bias is minimal. The number of screens ranged between 2 and 20 with a median of 4, and the ages at screen ranged between 65 and 85 years with a median of 69 years.

Exact aortic diameters are not available from the 683 men who remained aneurysm free and were screened every 2 years. Also for the remaining 156 men the diameter was frequently rounded to the nearest 5 mm. Therefore it would be difficult to estimate the distribution of aortic diameters. A model based on observed stages might be a reasonable approximation. Table 1 shows the frequencies of the transitions between observed stages at consecutive pairs of times. Examples of observed state sequences for five patients from the 156 who progressed above stage 1 are given in Table 2. Note that, although the screening intervals are nominally 3 or 12 months, the actual observation times are considerably less regular, owing to the constraints of clinical practice.

4. Models

It is supposed that the growth of an aneurysm is an irreversible process in the absence of surgical intervention. Therefore a plausible Markov transition intensity matrix is

$$Q = \begin{pmatrix} -q_{12} & q_{12} & 0 & 0 \\ 0 & -q_{23} & q_{23} & 0 \\ 0 & 0 & -q_{34} & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

indicating progression through each pair of consecutive states and no escape from stage 4. However, the measurement of aortic diameter is subject to error. Apparent transitions from higher

Table 1. Frequencies of transitions between stages of aortic diameter

Initial stage	*Aortic diameter (mm)*	*Frequencies for transitions to the following stages:*			
		1	*2*	*3*	*4*
1	0–29	1877	66	2	0
2	30–44	48	309	48	1
3	45–54	2	18	159	32
4	55	1	0	18	82

Table 2. Examples of progression through aneurysm states for five patients

Results for patient 1		*Results for patient 2*		*Results for patient 3*		*Results for patient 4*		*Results for patient 5*	
Age (years)	*State*	*Age (years)*	*State*	*Age (years)*	*State*	*Age (years)*	*State*	*Age (years)*	*State*
74.8	2	71.2	2	70.8	2	74.6	2	73.2	2
75.8	2	72.3	1	71.8	2	75.6	1	74.2	2
77.1	2	73.4	1	72.9	2	76.6	1	75.3	3
78.1	2	74.5	1	73.9	3	77.6	1	75.5	1
79.1	2	75.5	1	74.2	3	78.6	1	75.6	1
79.5	2	76.5	1	74.4	3	80.6	1	76.4	2
80.5	2	77.4	2	74.6	3				
81.0	2			74.9	3				
81.5	2			74.9	3				
				75.0	3				

to lower states are often seen. Fitting a progression-only multistate model with transition matrix Q and no misclassification would require that these reverse transitions be removed or smoothed out in some way. This can be done by recoding all reverse transitions as remaining in the lower of the two states, as reported previously (Couto *et al.*, 2002). The resulting *no-misclassification* model was compared with a hidden Markov model, which estimates state misclassification rates. For the hidden model, assume that the unobserved true states follow a Markov process with transition matrix Q, and that the observed states are generated from the latent states through a misclassification probability matrix:

$$E = \begin{pmatrix} 1-e_{12} & e_{12} & 0 & 0 \\ e_{21} & 1-e_{21}-e_{23} & e_{23} & 0 \\ 0 & e_{32} & 1-e_{32}-e_{34} & e_{34} \\ 0 & 0 & e_{43} & 1-e_{43} \end{pmatrix}.$$

4.1. Initial state assumptions

Recall that the likelihood for the hidden Markov model (Section 2) requires that at some time origin the true state occupancy is either known or can be modelled. One way of doing this would be to assume that everyone is aneurysm free at some common age. Simple models without misclassification were fitted to the data, assuming in turn that everyone was aneurysm free at ages 40, 50 and 60 years, and finally with no aneurysm-free assumption.

To fit the misclassification model, an estimate was taken from the literature. Grimshaw and Thompson (1995) studied aortic screening data from 10 061 men aged 60–75 years from general practices. The aortic diameters were summarized for the population and for specific ages. From their results, it was assumed that the probabilities of occupying stages 1, 2, 3 and 4 at age 60 years are 98.4%, 1.3%, 0.25% and 0.05% respectively.

4.2. Effect of age at screening

The transition intensities were modelled as a function of the age of the patient at screening, using a proportional hazards model as described in Section 2. Let $z(t)$ be the patient's age at time t minus 60 years, and let $q_{rs}\{z(t)\}$ be the corresponding (r, s) entry of the transition intensity matrix. We compare three models: one assuming different age effects β_{12}, β_{23} and β_{34} for each of the three rates q_{12}, q_{23} and q_{34} respectively (equation (5)), one assuming a common age effect

β on the rates (equation (6)) and a model with no age effects:

$$q_{rs}\{z(t)\} = q_{rs}\exp\{\beta_{rs}\, z(t)\}, \tag{5}$$

$$q_{rs}\{z(t)\} = q_{rs}\exp\{\beta\, z(t)\}. \tag{6}$$

We also considered, separately, models which expressed the misclassification probabilities e_{rs} as a function of age at screening, as in equation (1).

5. Results from aneurysm screening study

The three unknown transition rates, the six unknown misclassification probabilities and any age effects are estimated by maximum likelihood, using the matrix product (equation (4)) and the general R software described in Appendix A.

Table 3 presents the estimated transition rates and mean sojourn times from the models assuming no misclassification. The estimates are compared between models assuming a common aneurysm-free age of 40, 50 and 60 years, and for a model with no common aneurysm-free assumption. These results show that, in general, the estimated time spent in each stage decreases as the assumed aneurysm-free age increases. Therefore the model is sensitive to this assumption. It was similarly found that the equivalent hidden Markov model was sensitive to this assumption. Adding an initial stage 1 at a fixed age adds new transitions from stage 1 to the stage at first screen: specifically 651, 25, 3 and 4 from stage 1 to stages 1, 2, 3 and 4 respectively. The influence of the new transitions seems to depend on the assumed durations of these new transitions.

The likelihoods for the three misclassification models are compared in Table 4. There is a significant reduction in the -2 log-likelihood values when an age effect is included, and when the effects are extended to be different for each transition rate. The estimates and standard errors in Table 4 suggest that the only stage where age has a significant effect on the outgoing transition is stage 1. The model suggests that older men have a greater risk of initial onset of abnormal aortic diameter, but similar rates of subsequent growth of aneurysms to those of younger men.

Table 3. Aneurysm growth models without misclassification, assuming different common aneurysm-free ages†

Aneurysm-free age (years)	*Transition rates (per year)*		*Mean sojourn times (years)*	
40	$\hat{q}_{12}$	0.00997 (0.00065/0.00069)	$\hat{T}_1$	100 (6.5/6.9)
	$\hat{q}_{23}$	0.0380 (0.0044/0.0052)	$\hat{T}_2$	26.3 (3.1/3.5)
	$\hat{q}_{34}$	0.327 (0.063/0.082)	$\hat{T}_3$	3.05 (0.59/0.89)
50	$\hat{q}_{12}$	0.0147 (0.000956/0.000973)	$\hat{T}_1$	68.2 (4.4/4.5)
	$\hat{q}_{23}$	0.0521 (0.0060/0.0071)	$\hat{T}_2$	19.2 (2.2/2.6)
	$\hat{q}_{34}$	0.360 (0.064/0.080)	$\hat{T}_3$	2.78 (0.50/0.61)
60	$\hat{q}_{12}$	0.0277 (0.0018/0.0017)	$\hat{T}_1$	36.1 (2.4/2.1)
	$\hat{q}_{23}$	0.0831 (0.0093/0.0105)	$\hat{T}_2$	12.0 (1.35/1.47)
	$\hat{q}_{34}$	0.440 (0.070/0.077)	$\hat{T}_3$	2.27 (0.36/0.42)
No assumption	$\hat{q}_{12}$	0.0177 (0.0021/0.0021)	$\hat{T}_1$	56.6 (6.9/6.9)
	$\hat{q}_{23}$	0.156 (0.022/0.0022)	$\hat{T}_2$	6.42 (0.99/0.99)
	$\hat{q}_{34}$	0.493 (0.087/0.087)	$\hat{T}_3$	2.030 (0.41/0.41)

†Estimates, followed by standard errors (Hessian standard error/bootstrap standard error) given in parentheses.

Table 4. Parameter estimates, followed by standard errors (Hessian standard error/bootstrap standard error) in parentheses, from the hidden Markov models for aneurysm growth

Parameter	*Results for the following models:*		
	No age effect	*Common age effect*	*Different age effects*
−2 log-likelihood	2542	2484	2446
Number of parameters	9	10	12
Transition rates per year (at age 60 years)			
$\hat{q}_{12}$	0.0218 (0.0018/0.0019)	0.00784 (0.0017/0.0021)	0.0036 (0.0017/0.0020)
$\hat{q}_{23}$	0.0630 (0.0085/0.0081)	0.00117 (0.0052/0.0093)	0.088 (0.014/0.015)
$\hat{q}_{34}$	0.257 (0.054/0.049)	0.0475 (0.029/0.067)	0.260 (0.084/0.099)
Covariate effects for age			
		$\hat{\beta}$ 0.107 (0.014/0.015)	$\hat{\beta}_{12}$ 0.176 (0.019/0.018)
			$\hat{\beta}_{23}$ −0.0292 (0.029/0.032)
			$\hat{\beta}_{34}$ 0.022 (0.042/0.049)
Misclassification probabilities			
$\hat{e}_{12}$	0.0094 (0.0028/0.0034)	0.0082 (0.0024/0.0035)	0.0099 (0.0027/0.0032)
$\hat{e}_{21}$	0.0863 (0.022/0.032)	0.0976 (0.020/0.035)	0.0937 (0.019/0.029)
$\hat{e}_{23}$	0.0332 (0.0101/0.0139)	0.0357 (0.0097/0.013)	0.0326 (0.010/0.014)
$\hat{e}_{32}$	0.0396 (0.021/0.028)	0.0337 (0.018/0.031)	0.0416 (0.023/0.028)
$\hat{e}_{34}$	0.0560 (0.021/0.027)	0.0637 (0.022/0.028)	0.0552 (0.021/0.029)
$\hat{e}_{43}$	0.0879 (0.038/0.054)	0.0856 (0.037/0.057)	0.0902 (0.038/0.057)

The misclassification rates shown in Table 4 are not sensitive to the three different models for the transition rates. These suggest that the highest aortic diameter stage 4 is misdiagnosed as stage 3 about 9% of the time, and a diameter over 30 mm (stage 2) is misdiagnosed as normal (stage 1) about 9% of the time. These 'false negative' results seem to be more common than the corresponding 'false positive' results.

The misclassification model assumed a 98.4% probability of stage 1 at age 60 years. Although for this reason it is not directly comparable with the models without misclassification, it might be expected to be most similar to the model with the assumption of no aneurysms at age 60 years. Compare the transition rates from the misclassification model with no age effects with the rates from Table 3. The rates of 0.02 per year (mean sojourn time in stage 1 of 46 years) and 0.06 per year (mean sojourn time in stage 2 of 16 years) are similar to the rates from the 'age 60 years' model, but the transition rate of 0.26 per year (mean sojourn time in stage 3 of 4 years) from stage 3 is a slower rate than given by any of the models without misclassification. This might reflect the substantial measurement error in distinguishing between stages 3 and 4.

The estimated age effects in Table 4 are substantial. The assumed common age effect corresponds to an estimated hazard ratio of 2.9 per 10-year increase in age beyond 60 years. However, it appears that the age effect largely operates on the disease onset transition rate q_{12}, whereas the estimated age effects for the other transitions are close to 0.

Patients in the MASS trial are considered for elective repair surgery to prevent aneurysmal rupture when they enter stage 4, with an aortic diameter of 55 mm or more. Therefore an important quantity in determining the screening intervals for different men is the probability of truly being in stage 4 at some fixed time in the future, given truly being in an earlier stage at a certain age. Using the estimated transition rates and age effects, a selection of these probabilities (which are conditional on survival of the patient) are illustrated in Fig. 2. The risk of progression from a

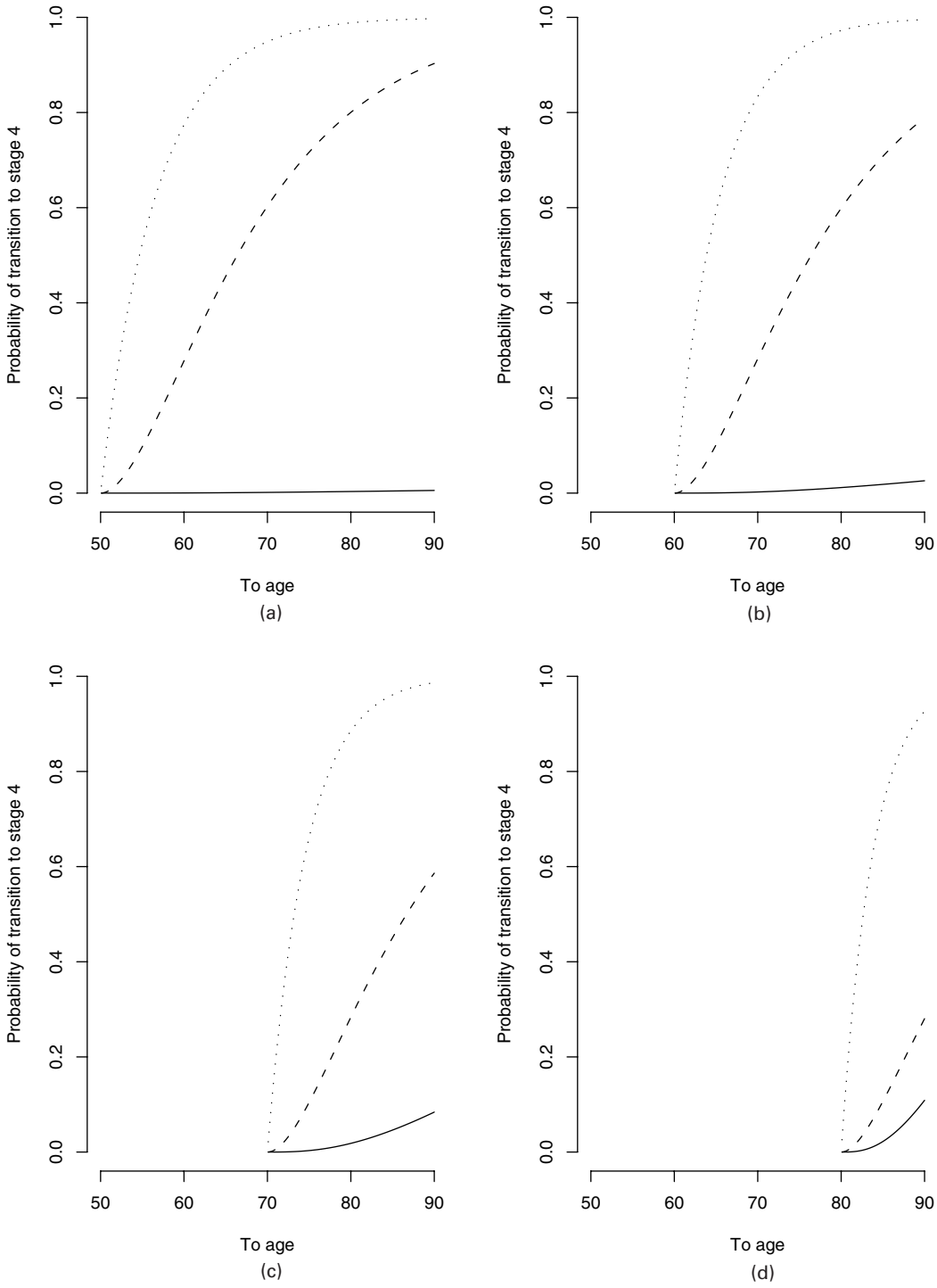

Fig. 2. Estimated transition probabilities into the final stage 4 (———, from stage 1; - - - - - - -, from stage 2; · · · · · · · , from stage 3): (a) from age 50 years; (b) from age 60 years; (c) from age 70 years; (d) from age 80 years

normal aortic diameter (stage 1) is low for all ages but becomes non-negligible after age 80 years. Given an aortic diameter of 30–44 mm (stage 2), the 5-year risk of moving to a diameter greater than 55 mm is around 11%, and the 10-year risk is around 28% for all ages. The progression rate from stage 3 to stage 4 is high for all ages, with 10-year transition probabilities between 82% and 91%.

The models illustrated also assumed that the misclassification probabilities were constant. Variables which might influence the probabilities of misclassification include age at screen, as was discovered in a comparable application to breast cancer screening (Chen *et al.*, 1996). Thus the misclassification model was also fitted with the age at screening as a covariate on the misclassification probabilities, using the logistic model (Section 2). No notable effect of age was detected.

5.1. Standard error estimation

Standard errors were also estimated for all the models by using the bootstrap. For the non-misclassification models, bootstrap samples were drawn from the set of transitions between pairs of states at consecutive times. For the misclassification models, consecutive state pairs within patients are not independent, so we sampled I patients with replacement from the entire set of I patient series. The models were fitted repeatedly to 1000 samples and the empirical standard deviations of the parameter estimates were calculated. These are presented in Tables 3 and 4 alongside the standard errors estimated from the inverse Hessian at the maximized log-likelihood. In most cases, the bootstrap standard errors are larger than the Hessian standard errors, often by 20–30% or greater. The bootstrap estimates are expected to be more accurate in this case. The estimate of covariance from the Hessian is a lower bound for the covariance matrix of the asymptotic normal distribution of the maximum likelihood estimator. Using this bound as an asymptotic approximation may be inaccurate for complex models, such as these, with irregular likelihood surfaces. See, for example, Cox and Hinkley (1974), for further discussion of these approximations and the necessary regularity conditions.

6. Discussion

The main advance of this paper is the presentation of a general model, with associated software, for misclassified Markov processes in continuous time. It builds on previous work in cancer screening by Chen *et al.* (1996) and Kirby and Spiegelhalter (1994) by generalizing the model to any number of states and any form of transition matrix, with covariates on both transition and misclassification rates. The model was used to estimate the sensitivity and specificity of ultrasonography screening for three successive grades of aortic aneurysm, while progression rates between latent stages were estimated simultaneously.

A frequently encountered issue in screening for chronic disorders is whether a single screen is sufficient, and when this screen should take place. The Chichester trial investigators (Scott *et al.*, 2001; Couto *et al.*, 2002), concluded that screening once for abdominal aortic aneurysm at age 65 years can identify a large population at low risk of rupture who can be ruled out of further surveillance. In this paper, progression rates were estimated for the same cohort. These rates support the view that it is not necessary to screen men with normal aortic diameters repeatedly. In addition, men with greater aortic diameters should have successively shorter follow-up intervals, as they have a higher probability of reaching stage 4, when surgery is performed to prevent rupture of the aneurysm. The results of the larger MASS trial (Multicentre Aneurysm Screening Group, 2002) also supported the idea of a national screening policy based on a single screen in men at age 65 years.

The states in the application were related to levels of a continuous marker. If the available marker information had been sufficiently detailed, then the model could be extended to make use of it. A distribution for the continuous marker, conditionally on the underlying stages, could be specified (Satten and Longini, 1996). Misclassification will be more likely if the true marker level is closer to a stage boundary, and therefore a continuous model should give a more accurate indication of progression rates. The more detailed data from the MASS trial might be used to apply such a model to aortic aneurysm screening.

This application has shown the importance of specifying the initial conditions of a Markov process. For models without misclassification, all individuals were assumed to be disease free at a common age. The transition rates, even the rates between stages 2–3 and 3–4, were sensitive to the assumed disease-free age. For the misclassification model, the underlying stage occupancy probabilities must be either specified or estimated at the initial time.

The assumption of time homogeneity of the Markov process was assessed by modelling age at screen as an explanatory variable for transition rates. Age was found to influence the disease onset rate, but not further rates of progression. However, the method used has the disadvantage of assuming that the transition intensities are constant between observation times. In the aneurysm application, this might affect the modelling of the transition rate from age 60 years to first screen, which would have greatest impact on the stage 1 to stage 2 intensity. However, screen times are reasonably close together after age 65 years. The models also assume homogeneity with respect to patient covariates other than age, which we did not model. Omitted covariates will presumably have the effect of increasing estimated misclassification probabilities.

We assumed that the sampling times were not informative, since the time of a screen is nominally chosen according to the previous stage. In fact, the times of observation varied around the proposed 3- and 12-month intervals, which may raise concerns about their informativeness. However, we expect that these variations are caused by the constraints of clinical practice, rather than patient 'self-selection'.

The Markov assumption is strong. Failure of the Markov assumption will probably appear in these models as an inflation of the misclassification probabilities. Different people typically have different aneurysm growth rates, so transition probabilities may depend on their history as well as the most recent stage. This might be assessed by fitting *ad hoc* elaborated models. For example, Weibull instead of exponential distributions might be assumed for the time spent in each state. The transition rates may also be conditioned on characteristics of the history of the process, such as the time spent in the current state, or the number of visits to each state. However, such increases in complexity would need detailed data to ensure identifiability. The process would need to be observed at frequent intervals, with little misclassification, to be able to estimate these characteristics of the true history.

Other assumptions included the parametric form of the effect of age on log-transition-rates. It is possible that linear terms in age may misrepresent the data. Assessing the goodness of fit of this class of models, in addition to simple model comparison using likelihoods, is worth further research. For example, Satten and Longini (1996) assessed the short-term predictive power of a hidden Markov model by forecasting stage occupancy probabilities at individuals' next observations. This class of models is complex and contains assumptions on many different levels which should be checked. It is difficult to distinguish between the appropriateness of the Markov model for the latent stage progression, the model for misclassification and the models for covariates at both levels.

The multistate and hidden Markov models discussed in this paper are widely applicable in many medical screening and monitoring situations. Work is in progress to apply this class of models, using the associated R software, to assessing the accuracy of angiographic screening

and risk factors for progression of coronary artery vasculopathy after heart transplantation, studying cognitive and functional decline in the elderly and assessing time-varying predictors for progression of psoriatic arthritis.

In developing our understanding of disease screening programmes, modelling the progress of the disease through time is important. A commonly encountered problem, however, in applying Markov progression models is the size of the programming task. It is hoped that the provision of the software written for this project will be of assistance to other researchers in the future. The framework presented here allows a model to encompass arbitrarily complex transition matrices, misclassification matrices and covariate effects. However, the warning should be given that very complex models may become close to non-identifiable. If there is insufficient information in the data, then latent transitions and misclassifications will be difficult to distinguish.

Acknowledgements

The authors would like to thank Allan Scott, Hilary Ashton and the Chichester trial investigators for making data available, Neil Walker for assistance with data preparation and Brian Tom and Laura Miller for helpful comments and testing of the R package.

Appendix A: R software—the 'msm' package

A package of functions for the statistical software R was written to fit general multistate Markov models in continuous time with and without misclassification, and covariates on transition and misclassification rates. The package is available under a free licence from the Comprehensive R Archive Network repository of contributed R code, `http://cran.r-project.org`. An example of the way in which this program is run is given briefly here. Full documentation is available with the package. Information about R itself is available from its Web page `http://www.r-project.org`.

Consider, firstly, a model without misclassification. Here we use '>' to indicate the R command prompt. The user specifies a matrix of 0–1 indicators for the allowed transitions whose intensities will be estimated, e.g. 1–2, 2–3 and 3–4, as follows:

```
> qmat <- rbind(c(0, 1, 0, 0),
                c(0, 0, 1, 0),
                c(0, 0, 0, 1),
                c(0, 0, 0, 0))
```

The diagonal of this matrix is ignored. Initial values for the unknown parameters should also be given, e.g.

```
> inits <- c(0.001, 0.03, 0.3)
```

The `msm` function is called to fit the model. Additional arguments `data`, `fromstate`, `tostate` and `timelag` specify the rectangular data frame with rows corresponding to observed transitions. Columns indicate the starting state, end state and time difference. Data can also be given with rows corresponding to observations, in the form 'subject identification, observation time, observed state'.

```
> age40.msm <- msm(data=age40.df, fromto=TRUE, fromstate=from,
                   tostate=to, timelag=dt, qmatrix=qmat, inits=inits)
```

For maximum likelihood estimation, a choice of optimization algorithms is available (e.g. Nelder–Mead or quasi-Newton) each of which can be fine tuned through options to the R `optim` function. For example, a scaling factor may be necessary to normalize the likelihood or individual parameters, to avoid numerical overflow or underflow. The maximum number of iterations, or the tolerance criteria for reporting convergence, may also need to be altered under some circumstances with flat likelihood surfaces. In these applications, the quasi-Newton method was found to give the greatest speed of convergence. The algo-

rithms do not require derivatives of the object function—these are computed by finite differences and used to estimate the Hessian at the maximum. To aid exploratory analysis, any number of parameters can also be fixed at their initial values during the optimization. If the algorithm converges successfully, printed output from the fitted model is as shown in Table 5.

Other summaries that are available include plots of expected survival from each state and tables of observed and expected prevalences of states at each time. To fit multistate models with misclassification, specify `misc=TRUE` in the call to `msm`. The data supplied should then include vectors of subject indicators, observation times and observed states. An additional required argument `ematrix` is a matrix indicating the allowed misclassifications between states. For example, in the aneurysm application we used

```
> ematrix <- rbind(c(0, 1, 0, 0),
                   c(1, 0, 1, 0),
                   c(0, 1, 0, 1),
                   c(0, 0, 1, 0))

> and.msm <- msm(state ~ age, misc=TRUE, subject=ptnum, data=an.df,
                 qmatrix=qmat, ematrix=ematrix, covariates = ~ age60,
                 inits = c(0.02, 0.06, 0.2, 0, 0, 0, 0.02, 0.06,
                 0.03, 0.03, 0.06, 0.08),
                 initprobs = c(0.984375, 0.012625, 0.0025, 0.0005))
```

The matrix `ematrix` indicates that an individual occupying underlying stage 1 may be misclassified as stage 2, stage 2 may be misclassified as stages 1 or 3, stage 3 may be misclassified as stage 2 or 4 and stage 4 may be misclassified as stage 3. The diagonal of this matrix is ignored, since we assume that correct classifications are always possible. The argument `initprobs` is a vector specifying the probabilities of occupying each of the four states at the beginning of the process.

Table 5. Output from a fitted model

> age40.msm

Multi-state Markov models in continuous time

Maximum likelihood estimates:

* Matrix of transition intensities

	Stage 1	Stage 2	Stage 3	Stage 4
Stage 1	−0.009966	0.009966	0.00000	0.0000
Stage 2	0.000000	−0.038009	0.03801	0.0000
Stage 3	0.000000	0.000000	−0.32747	0.3275
Stage 4	0.000000	0.000000	0.00000	0.0000

Corresponding standard errors

	Stage 1	Stage 2	Stage 3	Stage 4
Stage 1	.	0.0006495	0.000000	0.00000
Stage 2	0	.	0.004397	0.00000
Stage 3	0	0.0000000	.	0.06287
Stage 4	0	0.0000000	0.000000	.

* No covariates on transition intensities

* Mean sojourn times in transient states

	mean	SE
Stage 1	100.341	6.5393
Stage 2	26.309	3.0432
Stage 3	3.054	0.5863

−2 ∗ log-likelihood: 2401

Any number of covariate effects on the transition rates or the misclassification probabilities can be specified by using the `covariates` or `misccovariates` arguments to `msm`. The model `and.msm` above is the misclassification model for aneurysm growth assuming different effects on each transition rate of years since 60th birthday (Table 4). Fitting this model took less than 3 min on a personal computer with a 1.3 GHz processor and 11 min on a 500 MHz SunBlade 100 computer.

For reassurance that the true maximum likelihood estimates have been found, models should be run repeatedly starting from different initial values. However, a sensible choice of initial values can be important for unstable models with flat or multimodal likelihoods. For example, the transition rates for a model with misclassification could be initialized at the corresponding estimates for an approximating model without misclassification. Initial values for a model without misclassification could be set by supposing that transitions between stages take place only at the observation times. If we observe n_{rs} transitions from state r to state s, and a total of n_r transitions from state r, then q_{rs}/q_{rr} can be estimated by n_{rs}/n_r. Then, given a total of T_r years spent in state r, the mean sojourn time $1/q_{rr}$ can be estimated as T_r/n_r. Thus, n_{rs}/T_r is a crude estimate of q_{rs}.

Feed-back and suggestions for improvement are welcome from anyone wishing to use the package.

References

Aalen, O. O., Farewell, V. T., DeAngelis, D., Day, N. E. and Gill, O. N. (1997) A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction. *Statist. Med.*, **16**, 2191–2210.

Albert, P. S. (1999) A mover-stayer model for longitudinal marker data. *Biometrics*, **55**, 1252–1257.

Andersen, P. K. (1988) Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statist. Med.*, **7**, 661–670.

Baum, L. E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

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

Bureau, A., Hughes, J. P. and Shiboski, S. C. (2000) An S-Plus implementation of hidden Markov models in continuous time. *J. Comput. Graph. Statist.*, **9**, 621–632.

Chen, H. H., Duffy, S. W. and Tabar, L. (1996) A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Statistician*, **45**, 307–317.

Couto, E., Duffy, S. W., Ashton, H. A., Walker, N. M., Myles, J. P., Scott, R. A. P. and Thompson, S. G. (2002) Probabilities of progression of aortic aneurysms: estimates and implications for screening policy. *J. Med. Screen.*, **9**, 40–42.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Miller, H. D. (1965) *The Theory of Stochastic Processes*. London: Chapman and Hall.

Darling, R. C. (1970) Ruptured arteriosclerotic abdominal aortic aneurysms: a pathologic and clinical study. *Am. J. Surg.*, **119**, 397–401.

Dennis, J. E. and Schnabel, R. B. (1983) *Methods for Unconstrained Optimisation and Nonlinear Equations*. Englewood Cliffs: Prentice Hall.

Duffy, S. W. and Chen, H. H. (1995) Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase. *Statist. Med.*, **14**, 1531–1543.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

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

Grimshaw, G. M. and Thompson, J. M. (1995) The abnormal aorta: a statistical definition and strategy for monitoring change. *Eur. J. Vasc. Endvasc. Surg.*, **10**, 95–100.

Grüger, J., Kay, R. and Schumacher, M. (1991) The validity of inferences based on incomplete observations in disease state models. *Biometrics*, **47**, 595–605.

Guihenneuc-Jouyaux, C., Richardson, S. and Longini, I. M. (2000) Modelling markers of disease progression by a hidden Markov process: application to characterising CD4 cell decline. *Biometrics*, **56**, 733–741.

Guttorp, P. (1995) *Stochastic Modeling of Scientific Data*. London: Chapman and Hall.

Jackson, C. H. and Sharples, L. D. (2002) Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statist. Med.*, **21**, 113–128.

Juang, B. H. and Rabiner, L. R. (1991) Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Kay, R. (1986) A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, **42**, 855–865.

Kirby, A. J. and Spiegelhalter, D. J. (1994) Statistical modelling for the precursors of cervical cancer. In *Case*

Studies in Biometry (eds N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest and J. Greenhouse). New York: Wiley.

Klotz, J. H. and Sharples, L. D. (1994) Estimation for a Markov heart transplant model. *Statistician*, **43**, 431–438.

Lindsey, J. K. (1999) *Models for Repeated Measurements*, 2nd edn. Oxford: Oxford University Press.

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

Macdonald, I. L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman and Hall.

Marshall, G. and Jones, R. H. (1995) Multi-state Markov models and diabetic retinopathy. *Statist. Med.*, **14**, 1975–1983.

Multicentre Aneurysm Screening Group (2002) The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial. *Lancet*, **360**, 1531–1539.

Oehlert, G. W. (1992) A note on the delta method. *Am. Statistn*, **46**, 27–29.

Office for National Statistics (1995) *Mortality Statistics: Cause, England and Wales*. London: Stationery Office.

Satten, G. A. and Longini, Jr, I. M. (1996) Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Appl. Statist.*, **45**, 275–295.

Scott, R. A. P., Vardulaki, K. A., Walker, N. M., Day, N. E., Duffy, S. W. and Ashton, H. A. (2001) The long-term benefits of a single scan for abdominal aortic aneurysm (AAA) at age 65. *Eur. J. Vasc. Endvasc. Surg.*, **21**, 535–540.

Scott, R. A. P., Wilson, N. M., Ashton, H. A. and Kay, D. N. (1995) Influence of screening on the incidence of ruptured abdominal aortic aneurysm—5-year results of a randomized controlled study. *Br. J. Surg.*, **82**, 1066–1070.

Szilagyi, D., Elliot, J. and Smith, R. (1972) Clinical fate of the patient asymptomatic abdominal aortic aneurysm unfit for surgical treatment. *Arch. Surg.*, **104**, 600–606.

Vardulaki, K. A., Prevost, T. C., Walker, N. M., Day, N. E., Wilmink, A. B. M., Quick, C. R. G., Ashton, H. A. and Scott, R. A. P. (1998) Growth rates and risk of rupture of abdominal aortic aneurysms. *Br. J. Surg.*, **85**, 1674–1680.