Computer Methods and Programs in Biomedicine (2004) **75**, 95–105

Computer Methods and Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression

Wu Hui-Min[a], **Yen Ming-Fang**[b], **Tony Hsiu-Hsi Chen**[b,*]

[a] *College of Public Health, Institute of Epidemiology, National Taiwan University, Taipei, Taiwan*
[b] *College of Public Health, Institute of Preventive Medicine, National Taiwan University, Taipei, Taiwan, Room 207, No. 19, Hsu-Chou Road, Taipei, Taiwan*

Received 9 October 2003; received in revised form 1 December 2003; accepted 2 December 2003

KEYWORDS
Multi-state model;
Markov model;
Exponential regression model

Summary Writing a computer program for modeling multi-state disease process for cancer or chronic disease is often an arduous and time-consuming task. We have developed a SAS macro program for estimating the transition parameters in such models using SAS IML. The program is very flexible and enables the user to specify homogeneous and non-homogeneous (i.e. Weibull distribution, log–logistic, etc.) Markov models, incorporate covariates using the proportional hazards form, derive transition probabilities, formulate the likelihood function, and calculate the maximum likelihood estimate (MLE) and 95% confidence interval within a SAS subroutine. The program was successfully applied to an example of a three-state disease model for the progression of colorectal cancer from normal (disease free), to adenoma (pre-invasive disease), and finally to invasive carcinoma, with or without adjusting for covariates. This macro program can be generalized to other *k*-state models with *s* covariates.

1. Introduction

Multi-state models of disease progression are useful for describing the natural history of chronic diseases and cancer. A typical three-state model for the early detection of cancer by screening [1] might be defined as follows; normal, pre-clinical screen-detectable phase (PCDP) and clinical phase for early detection of cancer. For a non-malignant chronic disease such as type 2 diabetes [2,3] suitable states would be normal, pre-symptomatic phase, and symptomatic phase. The three-state model can be extended to a five-state model by using tumor attributes or other clinical measures in non-malignant chronic disease. For example, the PCDP phase and clinical phase for the three-state model can be further dichotomized by tumor size ($\geq$2 cm/$<$2 cm) or by node status (node positive/node negative) [4].

To model the transition rates of multi-state disease progression, the continuous-time Markov process has been proposed [5,6]. However, estimating transition parameters pertaining to a multi-state Markov process is not straightforward and requires a sophisticated computation program. There are three key issues to consider when writing the computer program. The first is that the

*Corresponding author. Tel.: +886-2-23587620; fax: +886-2-23587707.
E-mail address: stony@episerv.cph.ntu.edu.tw (T. Hsiu-Hsi Chen).

0169-2607/$ – see front matter © 2004 Published by Elsevier Ireland Ltd.
doi:10.1016/j.cmpb.2003.12.001

formulation of specific Markov models, particularly for non-homogeneous models which have rarely been addressed in previous medical applications. The second is the formulation of the likelihood function from empirical data because there are many potential problems, such as hidden transition (unobservable transition from no disease to pre-clinical phase in the three-state model), truncation and censoring. Finally, the calculation of point estimates and their 95% confidence intervals for relevant parameters is often computationally intensive. Efficient computer programming is therefore a necessity. The requirement to validate the model after fitting also adds to the computational load.

The objective of this study was therefore to develop a computational program for estimating the transition parameters in a multi-state disease process, particularly non-homogeneous models, using SAS IML. The program has been further generalized in a SAS macro program and tested on empirical data on disease progression in colorectal cancer.

2. Model formulation

In principle, our computer program can be applied to any k-state continuous-time Markov process with progressive property (see Fig. 1). For simplicity, a three-state continuous-time Markov processes is demonstrated here. However, the proposed computational program can be generalized to other Markov models without loss of generality.

2.1. Model specification

A general form of k-state continuous-time Markov process model with progressive property is delineated in Fig. 1. Let $X(t)$, the state of an individual at time t, be a random variable with state space $\Omega = \{1,2,3,\ldots,k\}$, where 1 usually represents no disease, and the others represent states of subsequent progression. The transition rate from state 1 to state 2 in the current model is modeled by a non-homogeneous distribution that captures the property of increasing or decreasing transition rate with time denoted as $\lambda_1(t)$. Suitable examples are the Weibull, log—logistic and gamma distributions. In theory, other transition rates between two states can also be modeled with a non-homogeneous distribution. However, the complexity of the algebra increases with number of states. For simplicity, only the transition rate from state 1 to state 2 is modeled with non-homogeneous distribution, the Weibull, and the remaining transition rates given the Markov property by modeling with the exponential distribution. We believe such a simplification is not unreasonable for multi-state disease progression of cancer or chronic diseases.

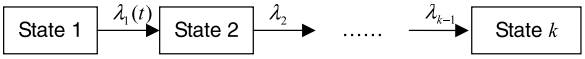

Fig. 1 A k-state progressive Markov model.

The non-homogeneous and homogeneous parts are expressed in the transition matrix.

$$\begin{array}{c} \\ \text{Previous} \\ \text{State} \end{array} \begin{array}{c} \text{Current State} \\ \begin{array}{cccccc} 1 & 2 & 3 & \cdots & k \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ k \end{array} \left(\begin{array}{c|ccc} -\lambda_1(t) & \lambda_1(t) & 0 & 0 & 0 \\ \hline 0 & & & & \\ 0 & & M & & \\ 0 & & & & \\ 0 & & & & \end{array} \right) \end{array} \tag{1}$$

The transition rate, $\lambda_1(t)$, is defined by

$$\lambda_1(t) = \lim_{\delta t \to 0} \frac{\Pr\{\text{transition } i \to (i+1) \text{ in } [t, (t+\delta t))\,| \text{ in state } i \text{ at time } t\}}{\delta t}$$

The transition time from state 1 to state 2 following the Weibull distribution is denoted as $W(\lambda_{10}, \gamma_1)$. Note that λ_{10} is a scale parameter and γ_1 is a shape parameter. The hazard function for $\lambda_1(t)$ is

$$\lambda_1(t) = \lambda_{10}\gamma_1 t^{\gamma_1 - 1} \tag{2}$$

The elements of the transition matrix with the Markov property, denoted $\boldsymbol{M}$ in Eq. (1), are as follows:

$$\mathbf{M} = \begin{array}{c} \\ \text{Previous} \\ \text{state} \end{array} \begin{array}{c} \text{Current state} \\ \begin{array}{cccccc} 2 & 3 & 4 & \cdots & \text{k-1} & \text{k} \end{array} \\ \begin{array}{c} 2 \\ 3 \\ 4 \\ \vdots \\ k-1 \\ k \end{array} \left[\begin{array}{cccccc} -\lambda_2 & \lambda_2 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_3 & \lambda_3 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{k-1} & \lambda_{k-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{array} \right] \end{array} \tag{3}$$

where λ_i is defined by

$$\lambda_i = \lim_{\delta t \to 0} \frac{\Pr\{\text{transition } i \to (i+1) \text{ in } [t, (t+\delta t)]\,| \text{ in state } i \text{ at time } t\}}{\delta t}$$

2.2. Transition probabilities

Following from Eq. (1), the transition probability of staying in state 1, $P_{11}(t_1, t_2)$, is

$$\begin{aligned} P_{11}(t_1, t_2) &= S_1(t_1, t_2) = \exp\left\{-\int_{t_1}^{t_2} \lambda_1(u - t_1)\, du\right\} \\ &= \exp(-\lambda_{10}(t_2 - t_1)^{\gamma_1}) \end{aligned} \tag{4}$$

where $S_1(t)$ represents the corresponding survival function.

The corresponding probabilities of transition during time interval $[t_1, t_2]$ is the $(k-1)\times(k-1)$ probability matrix, $\boldsymbol{P}^{\boldsymbol{M}}$, where the homogeneous matrix $\boldsymbol{M}$ in which the element of ith row and jth column, denoted by $\boldsymbol{P}_{ij}^{\boldsymbol{M}}(t_1, t_2)$, represents the probability of transition from state i to state j for $i = 2, \ldots, k$ and $j = 2, \ldots, k$. Transition probabilities were calculated by using the forward Kolmogorov equation [7] as follows:

$$\mathrm{d}\boldsymbol{P}^{\boldsymbol{M}}(t) = \boldsymbol{P}^{\boldsymbol{M}}(t)\boldsymbol{Q}$$

Subject to the boundary conditions $\boldsymbol{P}^{\boldsymbol{M}}(0) = \boldsymbol{I}$, the Kolmogorov equation leads to the unique solution $\boldsymbol{P}^{\boldsymbol{M}}(t) = \exp(\boldsymbol{Q}t)$. If $\boldsymbol{Q}$ has unique eigenvalues $v_2, v_3, \ldots, v_k$, denoted as a vector of $\boldsymbol{V}$, and if $\boldsymbol{A}$ is the $(k-1)\times(k-1)$ matrix whose jth column is the right eigenvector for r_j, then the solution is given by

$$\boldsymbol{P}_{2j}^{\boldsymbol{M}}(t) = \boldsymbol{A}\,\mathrm{diag}(\exp(v_2 t), \ldots, \exp(v_k t))\,\boldsymbol{A}^{-1} \tag{5}$$

The probability for an individual progressing from state 1 to state j during $[t_1, t_2]$, is therefore

$$P_{1j}(t_1, t_2) = \int_{t_1}^{t_2} f_1(u) P_{2j}^{\boldsymbol{M}}(t_2 - u)\,\mathrm{d}u \tag{6}$$

where $j = 2, 3, \ldots, k$, $\boldsymbol{P}_{2j}^{\boldsymbol{M}}$ represents the transition probabilities from state 2 to state j derived from Eq. (5), and $f_1(t)$, the probability density function of the Weibull distribution relating to the transition from state 1 to state 2, is written as

$$f_1(t) = \lambda_1(t) S_1(t) = \lambda_{10}\gamma_1 t^{\gamma_1 - 1}\exp(-\lambda_{10} t^{\gamma_1}) \tag{7}$$

From Eqs. (4)–(6), the probability functions for the transition from one state to another state are obtained and denoted as follows:

$$\mathbf{P} = \begin{array}{c} \\ \text{Previous state} \end{array} \begin{array}{c} \begin{array}{ccccc} \multicolumn{5}{c}{\text{Current state}} \\ 1 & 2 & 3 & \cdots & \mathrm{k} \end{array} \\ \begin{bmatrix} P_{11}(t) & P_{12}(t) & P_{13}(t) & \cdots & P_{1k}(t) \\ 0 & P_{22}(t) & P_{23}(t) & \cdots & P_{2k}(t) \\ 0 & 0 & P_{33}(t) & \cdots & P_{3k}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ k \end{array} \tag{8}$$

Note that $\boldsymbol{P}_{ij}(t)$ represents the risk of transition from state i to state j.

2.3. Exponential regression model for patient-specific covariates

To model the effect of individual's covariates, say z, on multi-state transitions, the exponential regression model is proposed by treating the scale parameter in the Weibull distribution as a function of an individual's covariates and is expressed by

$$\lambda_{10}^{m} = \lambda_{10}\exp(\beta_{10} z^{m}) \tag{9}$$

where λ_{10} is the scale parameter of Weibull distribution for the transition rate from state 1 to state 2 for the covariate at baseline value, z^m and β_{10} are vectors of covariates and the corresponding regression coefficient for individual m.

For the homogeneous part of the above k-state stochastic model, the effect of patient-specific covariates on multi-state transitions was modeled by the exponential regression model as

$$\lambda_i^m = \lambda_i \exp(\beta_i x^m) \tag{10}$$

for $i = 2, 3, \ldots, k$.

2.4. Likelihood function

Following Eq. (8), we use $P_{ij}(t_1, t_2)$ to represent the transition from state i to state j in a given time interval, $[t_1, t_2]$. Since we attempt to estimate the parameters pertinent to the disease natural history, only data on the first examination are used. This yields k possible observed transitions before the first examination, staying in state 1 (state 1 to state 1), state 1 to state 2, ..., and state 1 to state k. The likelihood function with k states is

$$L = \prod_{j=1}^{k}\prod_{m=1}^{n_j} P_{1j}(t_m)^{\delta_{mj}} \tag{11}$$

where t_m represents the age of the mth individual at first examination, and δ_{mj} is an indicator for the mth individual in state j ($j = 1, 2, \ldots, k$). Taking the logarithm of Eq. (11), the log likelihood function is obtained as

$$\log L = \delta_{mj}\sum_{j=1}^{k}\sum_{m=1}^{n_j} P_{1j}(t_m) \tag{12}$$

2.5. Parameter estimation

Parameters regarding regression coefficients, baseline scale parameters, λ_{10}, shape parameters, γ_1, and the remaining transition parameters, λ_2–λ_{k-1}, were obtained by maximum likelihood estimation (MLE). The corresponding standard errors were also calculated from the inverse of minus the second derivative of the likelihood function given the maximum likelihood estimates. All estimation procedures were performed using SAS IML procedures, SAS Version 8 [8].

2.6. Model validation

The goodness-of-fit of the model will be assessed using Pearson's χ^2-test statistic to determine how well the observed data's empirical distribution function agrees with the posited theoretical distribution function. The statistic is calculated as follows:

$$\chi^2 = \sum_{u=1}^{g} \frac{(O_u - E_u)^2}{E_u}, \tag{13}$$

where g is number of modes (i.e. the number of transition types we observed in the data), O_u is the observed count for the uth mode, and E_u is the expected count for the uth mode. Small probabilities indicate a poor fit.

3. Empirical data for sample runs

The data used in this study were derived from a cohort that consisted of 9021 subjects undergoing first colonoscopic examination at Kaohsiung Medical Center between 1979 and 1998. After receiving colonoscopy, this cohort was classified into three groups, 6522 normal subjects, 1755 adenoma cases (diagnosed at first examination), and 744 colorectal carcinoma (CRC) cases (diagnosed at first examination).

Details of study design were described in full elsewhere [9,10]. In brief, because the clinical attributes associated with progression to adenoma or CRC were recorded in pathological reports and medical charts that were not held on computer, a subset of samples, including 205 normal subjects, 200 adenoma cases, and 100 CRC cases, were randomly selected (Fig. 2). The sampling fractions for normal, adenoma and invasive CRC were denoted as π_1(205/6522), π_2(200/1755), and π_3(100/744).

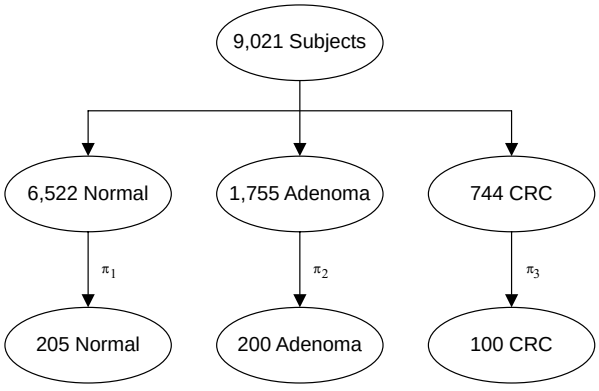

Fig. 2 A non-standard case-cohort design for adenoma and colorectal cancer.

Such a study design is called a non-standard case-cohort design (Fig. 2).

In this example data set, we wished to quantify the progression rates from adenoma and invasive carcinoma, and assess the effect of relevant covariates, such as gender, on each progression. To illustrate this, we modeled the natural history of colorectal cancer using a three-state model, incorporating covariates.

3.1. Three-state model without covariates

We model the disease process for colorectal cancer as a three-state continuous-time Markov process (Fig. 3) in which $X(t)$, the state of an individual at time t, is a random variable with a state space $\Omega = \{1, 2, 3\}$, where 1 represents no disease (normal), 2 represents colorectal adenoma, and 3 represents invasive colorectal carcinoma. K is three in this example.

According to Eqs. (1) and (8), the intensity matrix (with transition rates as elements) is

$$\mathrm{Q} = \begin{array}{c} \\ \text{Previous state} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Current state} & & \\ 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \begin{pmatrix} -\lambda_1(t) & \lambda_1(t) & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \tag{14}$$

where λ_1(t) is the annual incidence rate of adenoma, and λ_2 is the annual transition rate from adenoma to cancer. The corresponding transition probability matrix is

$$\mathrm{P} = \begin{array}{c} \\ \text{Previous state} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Current state} & & \\ 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \begin{pmatrix} P_{11}(t) & P_{12}(t) & P_{13}(t) \\ 0 & P_{22}(t) & P_{23}(t) \\ 0 & 0 & P_{33}(t) \end{pmatrix} \end{array} \tag{15}$$

where P_{11}(t) is calculated using Eq. (4). P_{12}(t) and P_{13}(t) are calculated as per Eq. (6) as

$$\begin{aligned} P_{12}(t_1, t_2) &= \int_{t_1}^{t_2} f_1(u) P_{22}^{M}(u, t_2)\, du \\ P_{13}(t_1, t_2) &= \int_{t_1}^{t_2} f_1(u) P_{23}^{M}(u, t_2)\, du \end{aligned} \tag{16}$$

As our study design was based on three subsets of samples, the likelihood function for estimating parameters cannot be formed by direct application of the above transition probabilities. Bayesian revision was used instead to calculate the probability for the selected cases. The conditional

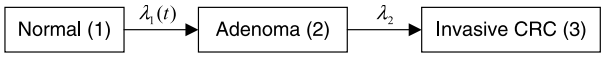

Fig. 3 A three-state Markov model for colorectal cancer.

probability for state j, for $j = 1, 2, 3$, according to whether the sample was selected ($S = 1$) is denoted by

$$\begin{aligned} P_{1_j^*}(t) &= \Pr(\text{state } j \text{ at first examination at age } t | \text{whether to be sampled}) \\ &= \Pr(P_{1j}(t)|S = 1) \\ &= \frac{\Pr(S = 1|P_{1j}(t))P_{1j}(t)}{\sum_{j=1}^{3}\Pr(S = 1|P_{1j}(t))P_{1j}(t)} \\ &= \frac{\pi_j P_{1j}(t)}{\sum_{j=1}^{3}\pi_j P_{1j}(t)} \end{aligned} \tag{17}$$

where π_1, π_2, and π_3 are random sampling fractions for states 1, 2, and 3, respectively. Therefore, according to Eqs. (11) and (12), the likelihood function for such data is

$$L = \prod_{j=1}^{3}\prod_{m=1}^{n_j} P_{1_j^*}(t_m)^{\delta_{mj}} \tag{18}$$

where t_m represents age at which the mth individual was first examined, δ_{mj} is an indicator for the mth individual in state j ($j = 1,2,3$), and the log likelihood function is as follows:

$$\log L = \delta_{mj}\sum_{j=1}^{3}\sum_{m=1}^{n_j} P_{1j}^*(t_m) \tag{19}$$

Table 1 shows the data for the three-state model, where ''AGE'' represents age at first examination, ''S_1'', ''S_2'' and ''S_3'' represent the number of subjects in state 1 (normal), state 2 (adenoma), and state 3 (invasive CRC), respectively, and ''XN'' represents the total number of subjects with the same age at first examination.

3.2. Three-state model incorporating one covariate

In order to assess the effect of gender on multi-state disease progression, we take the covariate, gender, into account using the above three-state Markov model (Fig. 3).

Using Eqs. (2) and (9) was extended to give

$$\lambda_1^m(t) = \lambda_{10}\exp(\beta_{10}\text{Gender})\gamma_1 t^{\gamma_1 - 1} \tag{20}$$

Similarly, using Eqs. (3) and (11) was extended to give

$$\lambda_2^m = \lambda_2\exp(\beta_2\text{Gender}) \tag{21}$$

The likelihood function can be similarly obtained from Eqs. (4)—(6), (8), (17) and (18).

4. Computer program

We developed a computational program for estimating the transition parameters underpinning multi-state disease process using SAS IML language. The program was delineated as follows.

4.1. Three-state model without covariate

(a) Read data

Reading the data shown in Table 1 using matrix form.

Table 1 Data for the three-state model

AGE	S_1	S_2	S_3	XN	AGE	S_1	S_2	S_3	XN	AGE	S_1	S_2	S_3	XN	AGE	S_1	S_2	S_3	XN
16	1	0	0	1	36	3	1	3	7	53	3	6	3	12	70	5	1	6	12
18	1	0	0	1	37	2	0	0	2	54	4	1	0	5	71	3	3	1	7
19	2	1	0	3	38	1	0	0	1	55	2	1	5	8	72	1	4	5	10
21	2	0	0	2	39	6	1	0	7	56	5	1	2	8	73	2	2	2	6
23	1	0	0	1	40	6	2	1	9	57	3	2	2	7	74	2	3	3	8
24	1	0	0	1	41	4	2	1	7	58	3	0	3	6	75	5	3	3	11
25	2	0	0	2	42	6	0	1	7	59	5	2	4	11	76	1	3	2	6
26	1	0	1	2	43	7	1	0	8	60	6	3	4	13	77	2	0	1	3
27	1	0	0	1	44	5	1	2	8	61	3	2	0	5	78	0	0	2	2
28	5	1	1	7	45	6	0	0	6	62	4	3	2	9	79	1	1	2	4
29	2	0	0	2	46	2	1	0	3	63	3	2	1	6	80	1	2	1	4
30	3	1	2	6	47	1	0	2	3	64	3	4	5	12	81	0	1	0	1
31	4	1	0	5	48	2	1	0	3	65	3	4	1	8	82	0	0	1	1
32	5	0	1	6	49	5	3	0	8	66	5	4	4	13	84	1	1	0	2
33	6	0	0	6	50	3	6	3	12	67	1	3	3	7	85	1	0	0	1
34	10	0	2	12	51	5	3	2	10	68	5	2	7	14					
35	1	0	0	1	52	5	2	1	8	69	5	2	2	9					

```
read all var(''S1'' :''S3'' ) into num;
read all var{xn} into xn;
read all var{age} into tt;
```

(b) Transition probabilities

The following programs show the corresponding transition probabilities, P_{11}, P_{12}, and P_{13}, with Weibull distribution. For clarity, the following equations were labeled by S-*X*, where *X* was derived from a series of Eqs. (4)–(19) listed above. The corresponding statement for Eq. (4), the transition probability staying at state 1 (P_{11}) from birth (t_1) to the first examination (t_2), was

```
lamda10=xh[1];
gamma1=xh[3]
P11=exp(-lamda1* ((t[2]
    -t[1])** gamma1));
```
S-(1)

xh [1] and xh [3] are variable names of λ_{10} and γ_1, t[1] is set as 0 (age at birth) and t[2] represents age at first examination. In addition to the non-homogeneous part, the homogeneous transition rate matrix $\boldsymbol{M}$ (following Eq. (3)) was calculated using the following statements:

```
Q=J(2,2,0);
Q[1,1]=-xh[2];
Q[1,2]=xh[2];
```
S-(2)

Note that xh[2] is the variable name of λ_2. Then the homogeneous transition probability $\boldsymbol{P}_{2j}^{M}(u, t_2)$ can be calculated from Eq. (5).

```
A=teigvec(Q);
V=teigval(Q);
D=diag(exp(v[,1]#(t[m,2]-u)));
P=A* D* inv(A);
```
S-(3)

t[m, 2] represents age at first examination for the mth individual.

The density function for the transition from state 1 to state 2 in Eq. (7) was written as follows:

```
f1=lamda1* gamma1* (u** (gamma1-1))
   * exp(-lamda1* (u** gamma1));
```
S-(4)

The two density functions for the transitions from normal (state 1) to adenoma (state 2) and invasive CRC (state 3) were

```
f12=f1* P[1,1];
f13=f1* P[1,2];
```

According to Eq. (16), the transition probabilities $P_{12}(t_1, t_2)$ and $P_{13}(t_1, t_2)$ can be calculated by the integration of the above equations using the subroutine of SAS language ''quad( )''.

```
Call quad(P12,''f12'' ,t[m,]);
Call quad(P13,''f13'' ,t[m,]);
```
S-(5)

t[m,] is the range from 0 to age for integration.

The conditional probability, $P_{1j}^{*}(t)$, with sampling fractions, as in Eq. (17) are calculated as

```
pi1=205/6522;
pi2=200/1755;
pi3=100/744;
px1=p11* pi1/(pi1* p11
  +pi2* p12 + pi3* p13);
px2=p12* pi2/(pi1* p11
  +pi2* p12+pi3* p13);
px3=p13* pi3/(pi1* p11
  +pi2* p12+pi3* p13);
```
S-(6)

px_1, px_2, and px_3 are variable names for P_{1j}^{*}.

Note that P_{i1}–P_{i3} will be set to 1 if data are based on the full longitudinal data rather than sampling design.

(c) Likelihood function

Following Eq. (19), the log likelihood functions were programmed as follows:

```
sum=sum+num[m,1]* log(px1/
  (px1+px2+px3))+
num[m,2]* log(px2/
  (px1+px2+px3))+
num[m,3]* log(px3/
  (px1+px2+px3));
```
S-(7)

The above program is included in a SAS module called *f_log L*, and *sum* is the summation of the log likelihood functions.

(d) Parameter estimation

Maximum likelihood estimates was obtained by using the iterative Newton–Raphson method (using subroutine *nlpnra( )*), where *f_log L* is the log likelihood function mentioned above. Initial values for λ_{10}, λ_2, and γ_1, were 0.005, 0.015, and 1, with the corresponding constraints, 10^{-8} to 1, 10^{-5} to 1, and 0–∞. *Optn*[1] = 1 gives maximum likelihood estimates and *optn*[2] = 2 provides details of the iteration process. *Estimate* is the vector of MLEs of parameters.

```
h0={0.005 0.015 1};
con={1.e-8 1.e-5 0, 1 1 .};
optn={1 2};
call nlpnra(rc,xres,''f_logL'',
  h0,optn,con);
estimate=xres';
```

The SAS subroutine *nlpfdd( )*, in IML, was also applied to calculate the standard errors of estimates and 95% CIs. The Hessian matrix is named as *hes2*. The variance and covariance were calculated by taking minus the inverse of the Hessian matrix. The program is written as follows:

```
call nlpfdd(f,g,hes2,''f_logL'',
  estimate);
cov=-inv(hes2);
norqua=probit(1-0.05/2);
stderr=sqrt(vecdiag(cov));
low=estimate-norqua* stderr;
up=estimate+norqua* stderr;
```

Low and *up* are the corresponding 95% confidence intervals.

(e) Model validation

Substituting the parameter estimates into the equation of transition probabilities enables one to calculate the expected values.

```
expect[i,1]=xn[i]* (px1);
expect[i,2]=xn[i]* (px2);
expect[i,3]=xn[i]* (px3);
create exp from expect[colname
  ={''exp1'' ''exp2'' ''exp3'' }];
append from expect;
```

Note that ''px_1''–''px_3'' are transition probabilities calculated by estimated parameters; three variables, *exp1*, *exp2*, and *exp3* represent the expected values of states 1, 2, and 3, respectively, and were stored in a data set named as *exp*.

The comparison between the expected and the observed results was programmed as follows:

```
data good;
merge dataset exp;
observed=s1; expected=exp1;
  state=1; output;
observed=s2; expected=exp2;
  state=2; output;
observed=s3; expected=exp3;
  state=3; output;
drop s1-s3 exp1-exp3;
proc means noprint;
var observed expected;
class state;
output out=t1 sum(observed)
  =O sum(expected)=E;
```

The above statements yield the observed, *O*, and the expected number, *E*, for each mode. Following Eq. (13), the chi-square values are calculated as follows:

```
data t2(drop=_freq_ _type_);
set t1;
z=(O-E)* * 2/E;
proc means sum noprint;
var z;
output out=good1 sum=chisquar;
```

S-(8)

where ''chisquare'' is the estimated chi-square value.

```
data good1(drop=_freq_ _type_);
set good1;
n_mode=3; n_para=3;
df=n_mode-n_para;
p_value=1-probchi(chisquar,df);
```

where ''n_mode'' is the number of modes including three transition types, from normal to normal, from normal to adenoma, and from normal to invasive cancer; ''n_para'' is the number of parameters. The *P*-value given the estimated chi-square was also calculated as ''p-value''.

4.2. Three-state model with covariates

The program for the three-state model with one covariate is similar to the above model without covariate. Following Eqs. (9) and (10), we have

```
lamda1=xh[1]* exp(xh[4]
  * (gender=1));
```

S-(9)

The homogeneous transition rate matrix, ***M***, according to Eqs. (10) and (17), is written as follows:

```
Q=J(2, 2, 0);
Q[1, 1]=-xh[2]* exp(xh[5]
  * (gender=1));
Q[1, 2]=xh[2]* exp(xh[5]
  * (gender=1));
```

S-(10)

xh[4] and xh[5] are variable names for β_{10} and β_2 (Eqs. (20) and (21)).

The transition probabilities $P_{ij}(t)$, log likelihood function, and parameter estimates are derived in a similar way.

4.3. SAS macro MARKOV

To generalize our SAS program, we therefore developed a SAS macro *MARKOV* to accommodate

k-state disease natural history with *s* covariates. The SAS macro *MARKOV* has eight components, (1) *data*, the SAS data set to be analyzed; (2) *k*, the number of states; (3) *dist*, the distribution for non-homogeneous transition from state 1 to state 2; (4) *init*, the initial values of each relevant transition parameter; (5) *upcon* and *lowcon*, the values of upper constraints and lower constraints on each relevant transition parameters; (6) *covnum* and *cov*, the number of covariates and the declaration of variable names for *s* covariates; (7) *n_mode*, number of modes; (8) *like*, the likelihood function. The input SAS data set *data* consists of four numerical variables, *AGE*, age at first examination, COV_s, variable names for *s* covariates, S_n, number of subjects in state *n*, *XN* represent the total number of subjects with the same age at first examination and with the same covariate status. Types of distribution, *dist*, for time to state 2 include *exponent*, *weibull*, *llogist*, and *gamma* for exponential distribution, Weibull distribution, log—logistic distribution, and gamma distribution, respectively. *Initial*, *upcon*, and *lowcon*, are three vectors containing a series of initial values, upper constraints, and lower constraints, which the first $(k-1)$ columns are transition rates λ_k where $k = 1,2,\ldots,(k-1)$, relating to each state; column *k* is the second parameter, γ, of the Weibull, log—logistic or gamma distributions; the remaining columns contain the regression coefficients relating to the *s*th covariate on $k-1$ transition rates. Further modifications to the likelihood function were needed because the likelihood function for estimating parameters cannot always be formed by direct application of the above transition probabilities, the original likelihood P_1—P_k can be transformed into a new likelihood PX_1—PX_k by macro the *like*.

The above SAS programs of macro *MARKOV* are available at website http://211.20.120.19/sas_program. The macro program was also applied to two examples mentioned above (see Appendix A).

5. Result of sample runs

Fig. 4 displays the SAS output for the three-state Markov model. The estimates and the corresponding 95% confidence intervals for λ_{10}, λ_2, and γ_1, are 4.7×10^{-5} (1.9×10^{-5} to 7.4×10^{-5}), 0.038 (0.028—0.048), and 2.12 (1.97—2.27), respectively. Therefore, the transition rate, $\lambda_1(t)$, from state 1

(Part of the output omitted)

Optimization Results

Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function
1	X1	0.000046870	-412.726097
2	X2	0.038424	-0.004908
3	X3	2.120660	-0.004569

Value of Objective Function = -388.2442317

Asymptotic 95% Confidence Interval

	ESTIMATE	CI	
	0.0000469	0.0000194	0.0000744
	0.0384244	0.0284976	0.0483511
	2.1206603	1.9740588	2.2672619

Goodness of fit

Obs	state	O	E
1	1	205	205.337
2	2	94	94.115
3	3	100	99.548

obs	chisquar	dist	n_para	df	p_value
1	.	WEIBULL	3	0	.

Fig. 4 Analysis of colorectal cancer data set: output of the SAS program for the three-state Markov model.

(Part of the output omitted)

Optimization Results

Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function
1	X1	0.000026440	-706.49038
2	X2	0.035678	0.02671
3	X3	2.185053	0.05974
4	X4	0.610417	0.00876
5	X5	0.100327	0.0005385

Value of Objective Function = -386.8599338

Asymptotic 95% Confidence Interval

ESTIMATE	CI	
0.000026	0.000015	0.000038
0.035678	0.021095	0.050261
2.185053	2.063565	2.3065404
0.610417	0.241606	0.979230
0.100327	-0.423344	0.623998

Goodness of fit

Obs	chisquar	dist	n_para	df	p_value
1	0.009967	WEIBULL	5	1	0.92047

Fig. 5 Analysis of colorectal cancer data set: output of the SAS program for the three-state Markov model with one covariate.

to state 2 assuming the Weibull distribution is

$$\lambda_1(t) = 4.7 \times 10^{-5} \times 2.12 \times t^{(2.12-1)}.$$

The second part shows goodness-of-fit of the model. The observed number is close to the expected number but because the model is saturated, with 0 degrees of freedom, the model cannot be tested with Pearson's χ^2-statistics.

Fig. 5 shows the output for the three-state Markov model with one covariate, gender. The estimates and the corresponding 95% confidence intervals for λ_{10}, λ_2, γ_1, β_{10}, and β_2, are 2.6×10^{-5} (1.5×10^{-5} to 3.8×10^{-5}), 0.036 (0.02—0.05), 2.19 (2.06—2.31), 0.61 (0.24—0.98), and 0.10 (—0.42—0.62), respectively. The chi-square statistic with one degree of freedom is 0.00997and the *P*-value is 0.92047. The results show perfect model fit. The hazard ratios for the effects of gender on annual incidence rate of adenoma and the annual transition rate from adenoma to cancer are 1.84 (1.27—2.66) and 1.11 (0.65—1.87), as calculated by exponentiating of β_{10} and β_2.

6. Discussion

A series of computer programs using IML language, from non-macro to macro program, for a multi-state disease progression model was developed. Several features render the program easy to use and flexible in modeling the natural history of a various cancers or chronic diseases with multi-state transitions.

First, the application of different distributions enables our model to allow the transition rates to vary with time. This characteristic dispenses with the assumption of constant hazards which has been made in the majority of previous studies and is useful in epidemiology because the annual incidence rate for the onset of the first state of disease rarely satisfies this criterion. The current finding, that the annual incidence rate of adenoma, from the three-state model, increases with age is an illustration of this. The Weibull or gamma distributions with increasing hazard rates were therefore considered. Our SAS macro program provides a series of survival distributions.

Secondly, calculation of transition probabilities using the spectral method with the forward Kolomogorov Eq. (S-3) also enables one to avoid the complexity of symbolic mathematical algebra. One can extend the program into a *k*-state model with ease by only making a few modifications on the transition rate matrix. Therefore, although only a three-state model was demonstrated in the text, the proposed program can be extended easily to any progressive *k*-state model.

Thirdly, the incorporation of exponential regression models, also lets one to elucidate the influence of different patient-specific covariates on each progression rate. In our example of gender and colorectal cancer, the only significant effect of gender is on the transition between normal to adenoma. Males have almost a two-fold increase in risk of adenoma. The natural history of cervical intraepithelial neoplasm (CIN) and its relationship with Human Papillomavirus (HPV) is another example. Modeling the effect of patient-specific covariates on multi-state disease natural history can also throw light on the role of each covariate on each transition. Besides, in our program, one can incorporate several covariates simultaneously and the covariates can also be continuous variables. In addition, another advantage of using exponential regression models is to estimate the hazard ratio with which the epidemiologist is familiar, by taking exponentials of the regression coefficients of the covariates. Hence, the interpretation of the results is convenient and meaningful. Fourthly, the corresponding standard errors and 95% CIs of the relevant parameters can be easily calculated by applying the SAS IML subroutine *nlpfdd( )*. Hence, hypothesis testing can be carried out without difficulty.

Finally, for economy a two-stage sampling design is used in our example. Thus, the likelihood of such design becomes complicated due to sampling. Our purpose is not to show the advantage of two-stage sampling design but to elaborate how the likelihood functions for different empirical data can be easily accommodated using our SAS macro program. Besides the two-stage sampling design, our program can be easily adapted to other empirical data such as truncation, censoring, and other properties.

Furthermore, the program also includes model validation by using Pearson's χ^2-test to assess whether the model is well fitted.

Markov process has been widely used in modeling the transition rates of multi-state disease progression, however, estimation of transition parameters is usually a stumbling block. We developed flexible SAS non-macro and macro computer programs for multi-state disease progression Markov models by using SAS IML language. The macro program can be generalized to other *k*-state models with *s* covariates.

Appendix A. Illustration using two examples of colorectal cancer screening

Example 1. Three-state model using Weibull distribution with no covariate.

```
%main(data=c.state3, k=3,
  dist=weibull, covnum=0,n_mode=3,
init=0.005 0.015 1, upcon=1 1 .,
  lowcon=1.e-8 1.e-5 0,
like=
pi1=205/6522;
pi2=200/1755;
pi3=100/744;
px1=p1* pi1/(pi1* p1+pi2* p2+pi3* p3);
px2=p2* pi2/(pi1* p1+pi2* p2+pi3* p3);
px3=p3* pi3/(pi1* p1+pi2* p2+pi3* p3);
);
```

Example 2. Three-state model using Weibull distribution with one covariate, gender.

```
%main(data=c.state3c, k=3,
  dist=weibull, covnum=1,
  cov=cov1=2-gender;,
n_mode=6, init=0.005 0.015 1 0.02 0.04,
lowcon=1.e-8 1.e-5 0 0 0, upcon=1 1 . . .,
like=
pi1=205/6522;
pi2=200/1755;
pi3=100/744;
px1=p1* pi1/(pi1* p1+pi2* p2+pi3* p3);
px2=p2* pi2/(pi1* p1+pi2* p2+pi3* p3);
px3=p3* pi3/(pi1* p1+pi2* p2+pi3* p3);
);
```

References

[1] N.E. Day, S.D. Walter, Simplified models of screening for chronic disease: estimation procedures from mass screening programmes, Biometrics 40 (1984) 1–14.
[2] The CDC Diabetes Cost-Effectiveness Study Group, The Cost-Effectiveness of Screening for Type 2 Diabetes, vol. 280, 1998, pp. 1757–1763.
[3] H.S. Kuo, H.J. Chang, P. Chou, L. Teng, T.H.H. Chen, A Markov chain model to assess the efficacy of screening for

non-insulin dependent diabetes mellitus (NIDDM), Int. J. Epidemiol. 28 (1999) 233—240.
[4] T.H.H. Chen, M.F. Yen, M.S. Lai, Estimation of sojourn time in chronic disease screening without data on interval cases, Biometrics 56 (2000) 167—172.
[5] S.W. Duffy, L. Tabar, N.E. Day, Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry and exit from the preclinical detectable phase, Stat. Med. 14 (1995) 1531—1543.
[6] H.H. Chen, S.W. Duffy, N.E. Day, Markov chain models for progression of breast cancer. Part I. Tumor attributes and the preclinical screen-detectable phase, J. Epidemiol. Biostat. 2 (1997) 9—23.
[7] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, Wiley, New York, 1980.
[8] SAS Institute Inc., SAS Language: Reference, Version 8, Cary, NC, 1999.
[9] C.D. Chen, M.F. Yen, W.M. Wang, J.M. Wong, T.H.H. Chen, A case-cohort study for the disease natural history of adenoma-carcinoma and de novo carcinoma and surveillance of colon and rectum after polypectomy: implication for efficacy of colonoscopy, Br. J. Cancer 88 (2003) 235—243.
[10] T.H.H. Chen, M.F. Yen, M.N. Shiu, T.H. Tung, H.M. Wu, Stochastic model for non-standard case-cohort design, Stat. Med., in press.