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

Wu Hui-Min\textsuperscript{a}, Yen Ming-Fang\textsuperscript{b}, Tony Hsiu-Hsi Chen\textsuperscript{b,}\textsuperscript{*}

\textsuperscript{a} College of Public Health, Institute of Epidemiology, National Taiwan University, Taipei, Taiwan  
\textsuperscript{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.

© 2004 Published by Elsevier Ireland Ltd.

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 \cite{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 \cite{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) \cite{4}.

To model the transition rates of multi-state disease progression, the continuous-time Markov process has been proposed \cite{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
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.

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

\[
\begin{array}{cccccc}
& 1 & 2 & 3 & \cdots & k \\
1 & 0 & 0 & 0 & \cdots & 0 \\
2 & 0 & M & 0 & \cdots & 0 \\
3 & 0 & 0 & M & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
k & 0 & 0 & 0 & \cdots & 0 \\
\end{array}
\]

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

\[
\lambda_1(t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{transition } i \to (i+1) \in [t, t+\Delta t])}{\Delta t}
\]

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

\[
\lambda_1(t) = \lambda_{10} \gamma_1 t^{\gamma_1-1}
\]

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

\[
\begin{array}{cccccc}
& 2 & 3 & 4 & \cdots & k \\
2 & -\lambda_2 & \lambda_1 & 0 & \cdots & 0 \\
3 & 0 & -\lambda_3 & \lambda_2 & \cdots & 0 \\
4 & \vdots & \vdots & \vdots & \ddots & \vdots \\
k & 0 & 0 & 0 & \cdots & -\lambda_k, \lambda_{k-1} \\
\end{array}
\]

where $\lambda_i$ is defined by

\[
\lambda_i = \lim_{\Delta t \to 0} \frac{\Pr(\text{transition } i \to (i+1) \in [t, t+\Delta t])}{\Delta t}
\]

2.2. Transition probabilities

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

\[
P_{11}(t_1, t_2) = S_1(t_1, t_2) = \exp \left\{ -\int_{t_1}^{t_2} \lambda_1(u - t_1) du \right\} = \exp(-\lambda_{10}(t_2 - t_1)^{\gamma_1})
\]
where $S_t(t)$ 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, $P^M$, where the homogeneous matrix $M$ in which the element of $i$th row and $j$th column, denoted by $P_{ij}(t_1, t_2)$, represents the probability of transition from state $i$ to state $j$ for $l = 2, \ldots, k$ and $j = 2, \ldots, k$. Transition probabilities were calculated by using the forward Kolmogorov equation [7] as follows:

$$dP^M(t) = P^M(t)Q$$

Subject to the boundary conditions $P^M(0) = I$, the Kolmogorov equation leads to the unique solution $P^M(t) = \exp(tQ)$. If $Q$ has unique eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$, denoted as a vector of $\lambda$, and if $A$ is the $(k-1) \times (k-1)$ matrix whose $j$th column is the right eigenvector for $\lambda_j$, then the solution is given by

$$P^M(t) = A \exp(\lambda_1 t, \ldots, \lambda_k t) A^{-1}$$

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

$$P_{1j}(t_1, t_2) = \int_{t_1}^{t_2} \exp(tu) P^M(t) A^{-1} du$$

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

$$f_i(t) = \lambda_i (t) S_i(t) = \lambda_i \gamma \exp(-\lambda_i t^\gamma)$$

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

$$P = \begin{bmatrix}
1 & 2 & 3 & \ldots & k \\
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}$$

Note that $P_{ij}(t)$ 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

$$S_{10} = \lambda_0 \exp(\beta_0 z^m)$$

where $\lambda_0$ 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 $\beta_0$ 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

$$S_{10} = \lambda_j \exp(\beta_k m)$$

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

### 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, \ldots, and state 1 to state $k$. The likelihood function with $k$ states is

$$L = \prod_{j=1}^{k} \prod_{m=1}^{n_j} P_{ij}(m)$$

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

$$\log L = \delta_{ij} \sum_{m=1}^{n_j} P_{ij}(m)$$

### 2.5. Parameter estimation

Parameters regarding regression coefficients, baseline scale parameters, $\lambda_1$, shape parameters, $\gamma_1$, and the remaining transition parameters, $\lambda_2, \ldots, \lambda_k$, 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 $\chi^2$-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:

$$\chi^2 = \sum_{u=1}^{g} \frac{(O_u - E_u)^2}{E_u}.$$  \hspace{1cm} (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 $u$th mode, and $E_u$ is the expected count for the $u$th 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 $\pi_1(205/6522), \pi_2(200/1755)$, and $\pi_3(100/744)$.

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

$$Q = \begin{pmatrix} 0 & -\lambda_1(t) & \lambda_2 \\ \lambda_3(t) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$  \hspace{1cm} (14)

where $\lambda_1(t)$ is the annual incidence rate of adenoma, and $\lambda_2$ is the annual transition rate from adenoma to cancer. The corresponding transition probability matrix is

$$P = \begin{pmatrix} P_{11}(t) & P_{12}(t) & P_{13}(t) \\ 0 & P_{22}(t) & P_{23}(t) \\ 0 & 0 & P_{33}(t) \end{pmatrix}.$$  \hspace{1cm} (15)

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

$$P_{12}(t_1, t_2) = \int_{t_1}^{t_2} f_1(u) P_{22}^u(t_2) \, du,$$

$$P_{13}(t_1, t_2) = \int_{t_1}^{t_2} f_1(u) P_{23}^u(t_2) \, du.$$  \hspace{1cm} (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.
SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression

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

\[
P_{ij}^*(t) = \Pr(\text{state } j \text{ at first examination at age } t | \text{whether to be sampled})
\]

\[
= \frac{\Pr(S = 1)}{\sum_{j=1}^{3} \Pr(S = 1)} \pi_j P_{ij}(t)
\]

(17)

where \( \pi_1, \pi_2, \text{ and } \pi_3 \) 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

\[
L = \prod_{j=1}^{3} \prod_{m=1}^{n_j} P_{ij}(t_m)^{X_{nj}}
\]

(18)

where \( t_m \) represents age at which the \( m \)-th individual was first examined, \( X_{nj} \) is an indicator for the \( m \)-th individual in state \( j (j = 1, 2, 3) \), and the log likelihood function is as follows:

\[
\log L = \delta_{nj} \sum_{m=1}^{n_j} P_{ij}(t_m)
\]

(19)

Table 1 shows the data for the three-state model, where "AGE" represents age at first examination, "S1", "S2", and "S3" 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

\[
\lambda_{ij}^m(t) = \lambda_{ij0} \exp(\beta_{ij0} \text{Gender}) \gamma_{ij}
\]

(20)

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

\[
\lambda_{ij}^m = \lambda_{ij2} \exp(\beta_{ij2} \text{Gender})
\]

(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

<table>
<thead>
<tr>
<th>AGE</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>XN</th>
<th>AGE</th>
<th>S1</th>
<th>S2</th>
<th>XN</th>
<th>AGE</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>XN</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>53</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>12</td>
<td>70</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>54</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>71</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>55</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>72</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>56</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>57</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>41</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>58</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>59</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>43</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>60</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>61</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>28</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>62</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>63</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>47</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>64</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>31</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>48</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>65</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>49</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>66</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>67</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>85</td>
</tr>
<tr>
<td>34</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>51</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>68</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>52</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>69</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>
(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 Eq. (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:

$$P_{11} = \exp(-\lambda_1 t_1)$$

$xh[1]$ and $xh[3]$ are variable names of $\lambda_1$ and $\gamma_1$, $t[1]$ 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 $M$ (following Eq. (3)) was calculated using the following statements:

$$Q[1,1] = xh[3]$$

The density function for the transition from normal (state 1) to adenoma (state 2) and the two density functions for the transitions $f_{12}$ and $f_{13}$ were calculated from normal to adenoma as follows:

$$f_{12} = \frac{p11}{p13} \exp(-\gamma_1 t_2)$$

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( )”.

(c) Likelihood function

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

$$\log L = \sum \log(px1) + \sum \log(px2) + \sum \log(px3)$$

The above program is included in a SAS module called _f_logL_, 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_logL_ is the log likelihood function mentioned above. Initial values for $\lambda_0$, $\gamma_1$, and $\gamma_2$ were 0.005, 0.015, and 1, with the corresponding constraints, $10^{-8}$ to 1, $10^{-5}$ to 1, and $0-\infty$. _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.

```
read all var(’S1’ : ’S3’) into num;
read all var[xn] into xn;
read all var[age] into t;
S-1
\lambda_{10} = xh[1];
gamma_{10} = xh[3];
P[1,1] = \exp(-\lambda_{10} t[1]); P[1,2] = \exp(-\lambda_{10} t[1] + \gamma_{10} t[2]); P[1,3] = \exp(-\lambda_{10} t[1]);
S-2
x[2] = t[2];
x[3] = t[1];
S-3
A = teigvec(Q);
V = teigval(Q);
D = diag(exp(v[,1]*(t[m,2]-u))); P = A*D*inv(A);
S-4
f[m,2] = \sum \exp(-\lambda_{10} (u+t[m,2]-u));
f[1] = \lambda_{10} \gamma_{10} \exp(-\lambda_{10} t[1] + \gamma_{10} t[2]) + \exp(-\lambda_{10} t[1] + \gamma_{10} t[2]);
S-5
\lambda_0 = 0.005; \gamma_0 = 0.015; \gamma_2 = 1;
S-6
call nlpnra(rc,xres,’_f_logL_’, h0,optn,con); estimate=xres;
```
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:

$$data t2(drop=freq type);$$

$$set t1;$$

$$z=(O-E) * 2/E;$$

$$proc means sum noprint;$$

$$var z;$$

$$output out=good1 sum=chisquar;$$

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

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

$$Q = J(2, 2, 0);$$

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

where $xh[4]$ and $xh[5]$ are variable names for $\beta_2$ and $\beta_3$ (Eqs. (20) and (21)).

4.3. SAS macro MARKOV

To generalize our SAS program, we therefore developed a SAS macro `MARKOV` to accommodate
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) `cownum` and `cov`, the number of covariates and the declaration of variable names for `s` covariates; (7) `nmode`, number of modes; (8) `like`, the likelihood function. The input SAS data set `data` consists of four numerical variables, `AGE`, age at first examination, `COVs`, variable names for `s` covariates, `Snr`, 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, ..., (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 `sth` 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 `P1-Pk` can be transformed into a new likelihood `PX1-PXk` 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 `λ10`, `λ2`, and `γ1`, are `4.7 × 10^{-3}` (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, `i_k(t)`, from state 1...
to state 2 assuming the Weibull distribution is
\[ \lambda_1(t) = 4.7 \times 10^{-5} \times 2.12 \times t^{(1.12-1)}. \]

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 \( \chi^2 \)-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 \( \lambda_{10} \), \( \gamma_1 \), \( \beta_{10} \), and \( \beta_2 \), are \( 2.6 \times 10^{-5} \) (1.5 \times 10^{-5} to 3.8 \times 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 0.00979 and 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 \( \beta_{10} \) and \( \beta_2 \).

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 Kolmogorov 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 let on, 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 neoplasia (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 y²-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.

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

References


