Estimation and prediction system for multi-state disease process: application to analysis of organized screening regime

Chi-Ming Chang MSc,[1] Wen-Chou Lin MSc,[2] Hsu-Sung Kuo PhD,[3] Ming-Fang Yen MSc[4] and
Tony Hsiu-Hsi Chen PhD[5]

[1]Director, Information Management Office, Center for Disease Control, Department of Health, Taipei, Taiwan
[2]Systems Analytst, Information Management Office, Center for Disease Control, Department of Health, Taipei, Taiwan
[3]Director, Center for Disease Control, Department of Health, Taipei, Taiwan
[4]Assistant Professor, Institute of Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
[5]Professor, Institute of Preventive Medicine, National Taiwan University, Taipei, Taiwan

Keywords
computer-aided system, Markov process, maximum likelihood estimation, SAS system

Correspondence
Professor Tony Hsiu-Hsi Chen
Institute of Preventive Medicine
College of Public Health
National Taiwan University
Room 521, No. 17
Suchow Road
Taipei 100
Taiwan
E-mail: chenlin@ntu.edu.tw

Accepted for publication: 13 February 2006

doi:10.1111/j.1365-2753.2006.00765.x

Abstract

Rationale, aims and objectives The disease progression of cancer and non-malignant chronic disease often involve a multi-state transition. However, estimation of parameters and prediction regarding the multi-state disease process are complex. This study aimed to develop an estimation and prediction system with a computer-assisted software using SAS/SCL as a platform to predict the risk of any outcome arising from the underlying multi-state process with or without the incorporation of individual characteristics.

Method The computer-aided system is first constructed following the theoretical framework of stochastic process. The functions provided in this software include model specification, formulation of likelihood function, parameter estimation, model validation and model prediction. An example of breast cancer screening for a high-risk group in Taiwan was used to demonstrate the usefulness of this software.

Results The natural history of breast cancer of a three-state disease process has been demonstrated. Two suspected risk factors, late age at first full-term pregnancy and obesity, were considered by the form of the proportional hazard model. Formulation of intensity matrix, likelihood function, assignment of initial values, and parameter constraint and estimation were successfully demonstrated in model specification. Model validation suggested a good fit of the constructed model. The application of model prediction enables one to project the effectiveness of organized screening by different inter-screening intervals from a policy level or from an individual basis.

Conclusions A computer-aided estimation and prediction system for multi-state disease process was developed and demonstrated. This system can be applied to data with the property of multi-state transitions in association with events or disease.

Introduction

The disease progression of cancer and non-malignant chronic disease often involves a multi-state transition. Tumour carcinogenesis for the development of cancer may start from normal, pass through occult and pre-clinical screen-detectable phase (PCDP), and surface to clinical phase with overt symptoms [1]. For chronic disease like diabetic retinopathy, patients with type 2 diabetes progress from diabetes without diabetic retinopathy, baseline diabetic retinopathy, progressive diabetic retinopathy, and eventually to blindness. To interrupt a disease process, a variety of decisions regarding primary intervention or secondary intervention programme were compared in order to identify an optimal strategy. Screening regimes for cancer with available tools are typical examples of intervention programme for interrupting the multi-state disease process in the realm of public health. However, while the screening regime, including breast cancer and colorectal cancer, has been demonstrated to be effective in mortality reduction in patients with breast cancer, several issues still remained. To begin with, one would still need to determinate the inter-screening interval on the ground of economic consideration. Due to the requirement of long-term follow-up, one cannot answer such a question using longitudinal follow-up data. Outcome prediction by different inter-screening regimes is therefore needed. Second, individual-centred screening programme has been addressed in the face of occurrence of false negative cases in organized screening regime. The development of such an individualized screening regime is required. To solve these two problems, one needs the natural history of multi-state disease progression.

To quantify the multi-state disease progression can further predict the outcomes by different inter-screening intervals and individual characteristics. However, the procedure of disease prediction in the multi-state transition is very complex, including the formulation of stochastic model, estimation of parameters, calculating the mean sojourn time, and the calculation of transition probabilities by different decisions.

The aim of this study was therefore to develop a computer-assisted software using SAS/SCL as a platform to predict the outcome spawned from multi-state disease progression. The functions provided in this software include model specification, formulation of likelihood function, parameter estimation, model validation and model prediction. This software was further applied to project the effectiveness of cancer screening by different inter-screening intervals, and by a constellation of individual characteristics.

Methods and materials

Figure 1 shows the infrastructure and process of computer-assisted system pertaining to multi-state disease prediction on the basis of Markov process. It consists of four components, including data management, model specification and parameter estimation, model validation and model prediction. Each component is detailed as follows.

Data management

Although data management is not relevant to multi-state disease prediction, providing some specific functions in relation to data management can facilitate the process of estimation and disease prediction. This part was developed and detailed in a previous study designed to develop a computer prediction system [2]. In brief, functions of our data management include updating data, creating new variables, merging different data sources, appending new record, and creating centring variable by subtracting mean value from the original variable to reduce multi-colinearity while polynomial variables were considered in the regression model. Various sampling designs were also provided for the cross-validation of model. Data processing for data aggregation was provided to transform the raw data.

Model specification

The second part is pertaining to model specification using Markov process.

Following the traditional definition of Markov process from previous studies [3–5], let x(t) denote a continuous-time stochastic process with countable state space denoted as $\Omega = \{1, \ldots k\}$

The instantaneous transition rate from state *i* to state *j* can be expressed by an intensity λ_{ij}

$$\lambda_{ij} = \lim_{\Delta t \to 0} \Pr(X(t+\Delta t) = j \mid X(t) = i) \tag{1}$$

This intensity λ_{ij} can be expressed by a specific matrix form depending on the disease process. Our software development can accommodate any form of intensity matrix.

However, for simplicity, we illustrated our prediction system with a three-state model and a five-state model associated with the development of cancer or chronic disease. We assume a progressive disease process that is diagrammed below.

$$\text{Normal} \xrightarrow{\lambda_{12}} \text{PCDP} \xrightarrow{\lambda_{23}} \text{clinical phase.}$$

An intensity matrix Q expresses such a three-state Markov model,

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} -\lambda_{12} & \lambda_{12} & 0 \\ 0 & -\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 \end{pmatrix} \end{array} \tag{2}$$

Thus, any disease process can be formulated by the intensity matrix form as above.

Intensity matrix is converted to transition probabilities according to the forward Kolmogorov equation

$$P(t) = \mathrm{A} \times \mathrm{diag}(\exp(\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3)) \times \mathrm{A}^{-1}, \tag{3}$$

where d_1, d_2 and d_3 represent eigenvalues solved from **Q** by the delta method [6]. A is the corresponding eigenvector of **Q**.

For the three-state model, the transition probabilities during time t in matrix form are expressed as

$$P = \text{Current} \begin{array}{r|ccc|} & \text{Normal} & \overset{\text{Next}}{\text{PCDP}} & \begin{array}{c}\text{Clinical}\\ \text{Phase}\end{array} \\ & (1) & (2) & (3) \\ \text{Normal (1)} & P_{11}(t) & P_{12}(t) & P_{13}(t) \\ \text{PCDP (2)} & P_{21}(t) = 0 & P_{22}(t) & P_{23}(t) \\ \text{Clinical Phase (3)} & P_{31}(t) = 0 & P_{32}(t) = 0 & P_{33}(t) = 1 \end{array} \tag{4}$$

The transition probability is subject to the constraint of $\sum_{i,j \in 3} P_{ij} = 1$.

Clinical interpretation of these transition probabilities is that an individual in the normal state at time t_1 may progress to PCDP or surface to clinical phase at time t_2. Both transition probabilities during time $t(t_2 - t_1)$ are therefore expressed by P_{12}(t) and P_{13}(t).

The transition probability of P_{11}(t) represents the probability of staying in the normal state during time t. A similar interpretation is applied to other transition probabilities.

The effect of covariates of interest on intensity, λ_{ij}, is modelled by the following proportional hazards regression form,

$$\lambda_{ij} = \lambda_{0 \cdot ij} \exp(\mathrm{X}^T \beta), \tag{5}$$

where X is a vector of covariate and β is its corresponding coefficients, $\lambda_{0 \cdot ij}$ is the baseline intensity.

Likelihood function

Given transition probabilities in Eqn 4, the specific likelihood function can be developed to estimate the parameters of intensity matrix, such as λ_{12} and λ_{23}. Our software system is rather flexible in formulating the log-likelihood function for relevant empirical data (see below). Suppose we have n records of empirical transitions, the log-likelihood function is formulated as follows.

$$\sum_{i=1}^{n} \log[\mathrm{L}_i(\text{parameters}, t)]. \tag{6}$$

L_i(.) represents individual likelihood function

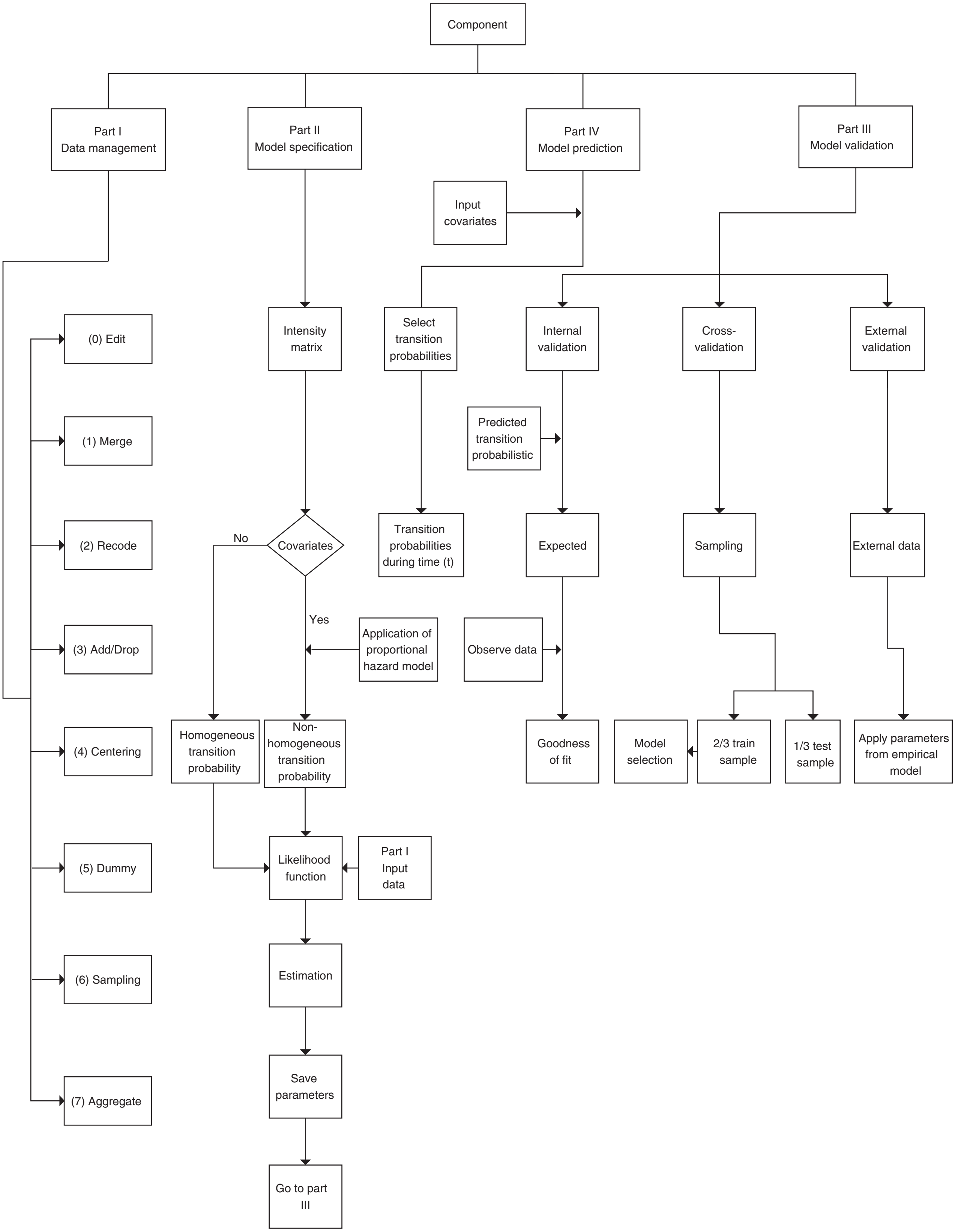


Figure 1 Infrastructure and process of computer-assisted system for multi-state disease prediction.

Empirical data, likelihood function and parameter estimation will be developed and demonstrated in the application section using organized screening data.

Model validation

The fitted model was validated by three levels: internal validation, cross-validation and external validation. The first-level validation was assessed by comparing the predicted number with the observed number. The goodness fit of test with Pearson chi-square was used to assess whether the difference was statistically significant or not.

The second level of validation, cross-validation, was to divide dataset into two parts, at random, 2/3 train sample and 1/3 test sample. It should be noted that as the current study focused on the multi-state outcome, receiver operating characteristic (ROC) curves corresponding to number of states should be calculated and assessed.

The third level of validation was to apply parameters of the current model to predict multi-state outcomes based on external dataset. Numbers of predicted outcomes were compared with those of observed outcomes.

Model prediction

Transition probabilities using the parameters estimated in this study are used to predict the transition probability from state *i* to state *j* given a series of known covariates during time *t*.

Software development

Figure 2 shows main menu for four components mentioned above. Figure 3 specifies two options, deterministic and stochastic models. The stochastic model needs to estimate transition parameter from empirical data.

Data for illustration

In order to illustrate the details of our computer-assisted software system, an example of breast cancer screening for a high-risk group in Taiwan was used to demonstrate the usefulness of this software. Details of the study design for this high-risk group screening project, Taiwan Multicentre Cancer Screening, have been published elsewhere [7]. In brief, this was a multi-centre project to identify early cancer for a population at high risk via different screening tools. Here the high-risk group of breast cancer is identified as female relatives of breast cancer cases, including mothers, daughters, sisters and grandmothers. Since 1994, relatives over 35 years old of breast cancer index cases from 12 hospitals were invited to annual screening by a combination of mammography, ultrasound and physical examination. Those with either one positive result on the above three tests were further confirmed by histological biopsy. Once an individual is affirmed as a breast cancer case, she will be properly treated with medical care. Up to 1997, a total of 4280 women received their first screen, and their age ranged from 34 to 86 years (with an average age of 46.67 years, SD = 9.00). However, it should be noted that about 75% of the subjects were under 50 years old; 50 breast cancer cases were found among them. A total of 1584 women received the second screening 1 year after their first screen, and eight cases were found. No interval cases were obtained in this study.

In this example, we selected two suspected risk factors for demonstration, including late age at first full-term pregnancy and obesity. These factors are referred to as covariates and abbreviated as AP1AGEGP and BMIGP. The definitions of these covariates are described as follows:

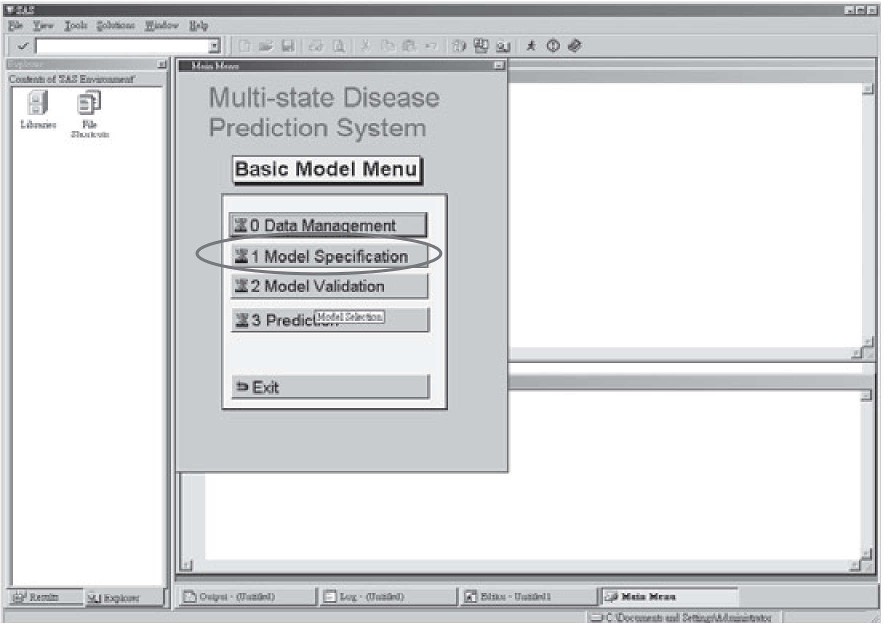


Figure 2 Main menu of the computer-aided system for multi-state process.

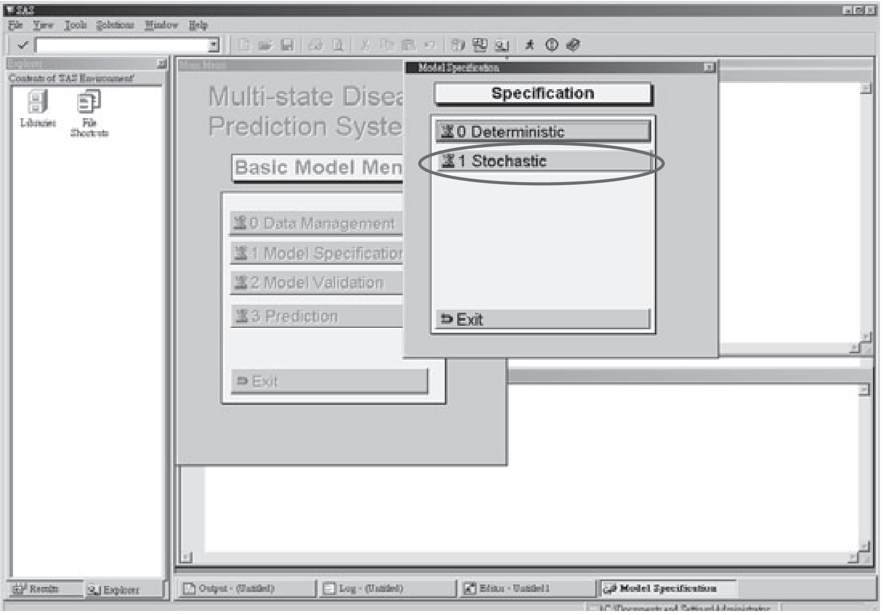


Figure 3 Menu for model specification.

Table 1 Transition type, transition probabilities and number of transition applied for the three-state Markov model, breast cancer screening data from Taiwan Multicentre Cancer Screening

Screening round and detection mode	Transition type	Transition probability	Number
First screen			
Prevalent case	Normal→PCDP	$\frac{p_{12}(0,t_1)}{p_{11}(0,t_1)+p_{12}(0,t_1)}$	50
Normal	Normal→Normal	$\frac{p_{11}(0,t_1)}{p_{11}(0,t_1)+p_{12}(0,t_1)}$	4230
Second screen			
Incident case	Normal→PCDP	$p_{12}(t_1,t_2)$	8
Normal	Normal→Normal	$p_{11}(t_1,t_2)$	1576

PCDP, pre-clinical screen-detectable phase.

$$\text{AP1AGEGP}=\begin{cases}1 & \text{if age at first full-term pregnancy is older} \\ & \text{than 25 years} \\ 0 & \text{otherwise}\end{cases}$$

$$\text{BMIGP}=\begin{cases}1 & \text{if one's body mass index } \left(\text{kg m}^{-2}\right) \text{ is greater} \\ & \text{than 23} \\ 0 & \text{otherwise}\end{cases}$$

Throughout the following text, the positive cases with respective to AP1AGEGP and BMIGP are defined as AP1AGEGP = 1 and BMIGP = 1, respectively, otherwise negative cases. Data on round of screen, transition type, transition probability, and number of transition are listed in Table 1.

Proportional hazard model form based on Eqn 5 is expressed by the following form:

$$\lambda_{12}=\lambda_{0\cdot 12}\exp(\beta_{11}\text{AP1AGEGP}+\beta_{12}\text{BMIGP})$$

$$\lambda_{12}=\lambda_{0\cdot 23}\exp(\beta_{22}\text{AP1AGEGP}+\beta_{22}\text{BMIGP})$$

Stochastic model

Figure 4 displays the menu for model specification, including the formulation of intensity matrix, likelihood function, assignment of initial values, parameter constraint and estimation. Figure 5 displays how to build up intensity matrix Q as in Eqn 2. The middle panel lists all variables used in the following Markov process, including those for transition type and transition time. The left panel in Fig. 5 is tailored for forming intensity matrix of interest. For the three-state model, three vectors were denoted as ($-h_1$, h_1, 0) for the first row, (0, $-h_2$, h_2) for the second row, and (0, 0, 0) for the third row.

Note that while number of state, say 3, was defined, the program automatically yields a consecutive number from 1 to 3 in this example.

The function of the right panel is to select the variable pertaining to transition type and transition time. In our illustration, transition types were expressed by 'type' with the following labels: 11 = staying in state 1 (normal); 12 = state 1 to state 2 (PCDP); 13 = state 1 to state 3 (clinical phase); 22 = stay in state 2; and 23 = state 2 to state 3. Transition time encoded in the transition probabilities is coded as 't'.

The third and fourth lines in the right panel also provide two functions for the model validation (see below), aggregation of same transition type and total counts of each transition type. The effect of covariates, illustrated by AP1AGEGP and BMIGP for age at first full-term pregnancy and body mass index (BMI), respectively, on each transition state are also selected in the bottom line in the right panel. The estimates regarding sensitivity and specificity, two measurement errors observed in cancer or chronic disease screening are also considered in the model. The bottom panel in

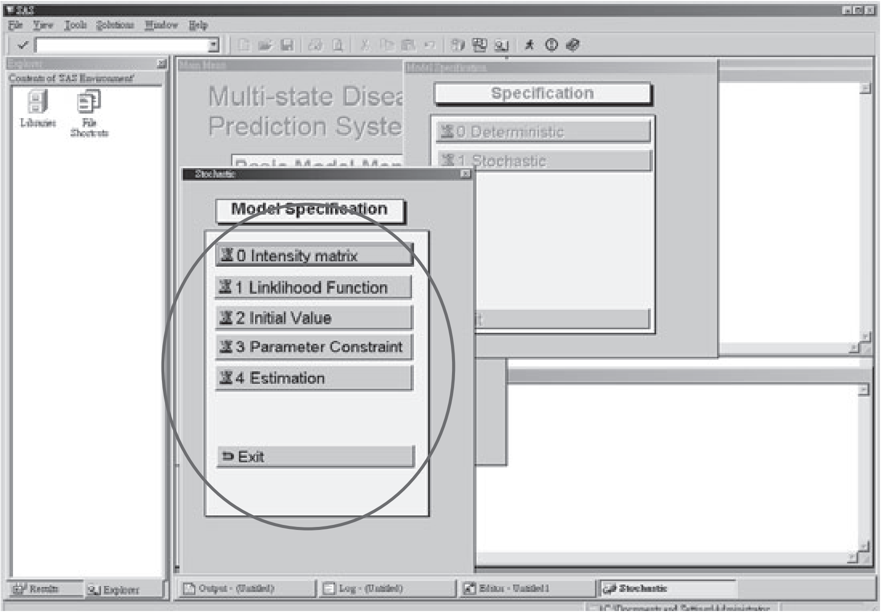
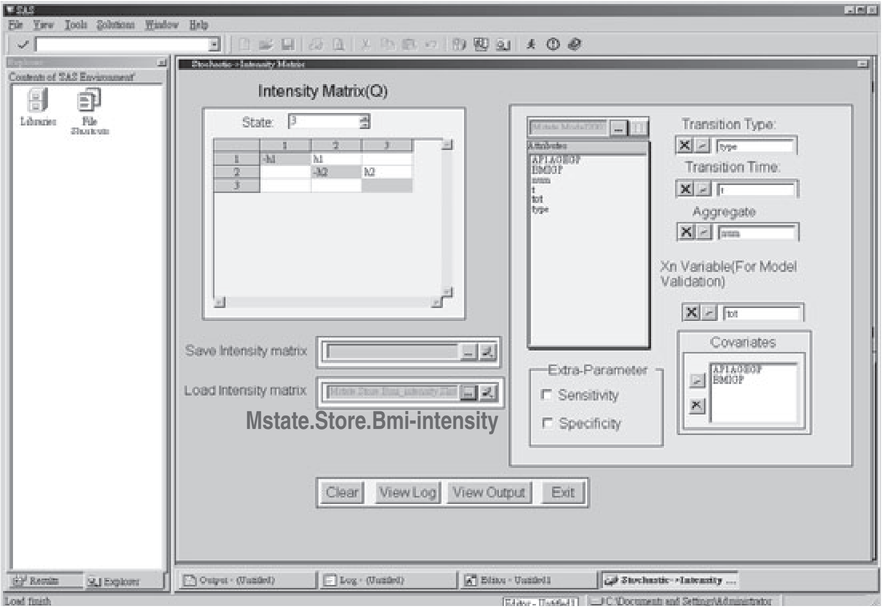


Figure 4 Menu for stochastic model specification.



State: 3

	1	2	3
1	-h1	h1	
2		-h2	h2
3			

Attributes: APIAOBOF, BMIGP, sum, t, tot, type

Transition Type: type

Transition Time: t

Aggregate: sum

Xn Variable(For Model Validation): tot

Covariates: APIAOBOF, BMIGP

Save Intensity matrix

Load Intensity matrix

Mstate.Store.Bmi-intensity

Figure 5 Tailor for forming model specification.

Fig. 5 also provides a series of functions for saving and re-loading previously used models.

Figure 6 provides the template for writing a likelihood function. The default algorithm is

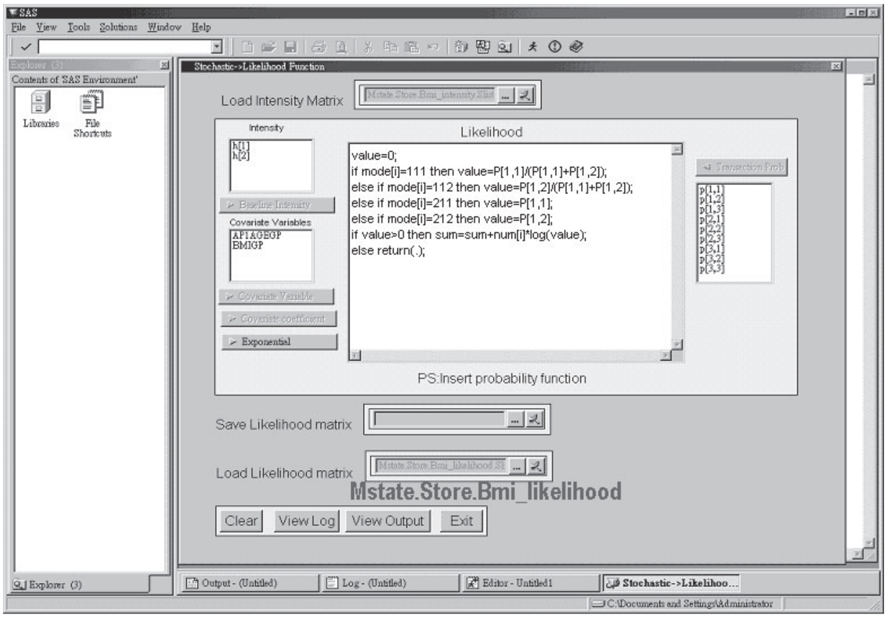
If mode(transition state) = " "

Then sum = sum + log (user-define transition probability)

These statements correspond to the summation of log-likelihood function in Eqn 4. The likelihood function for data on the first and second screens is illustrated as follows.

```
value = 0;
if mode[i] = 111 then value = P[1,1]/(P[1,1] + P[1,2]);
else if mode[i] = 112 then value = P[1,2]/(P[1,1] + P[1,2]);
else if mode[i] = 211 then value = P[1,1];
else if mode[i] = 212 then value = P[1,2];
if value > 0 then sum = sum + num[i]*log(value);
else return(.);
```

The corresponding statement is shown in Fig. 6. The numbers of '111' and '112' stand for normal and prevalent cases at the first screen. The numbers of '211' and '212' stand for normal and screen-detected cases at the second screen.



Load Intensity Matrix: Mstate.Store.Bmi_intensity.Slist

Intensity:
h[1]
h[2]

Covariate Variables:
AP1AGEGP
BMIGP

Likelihood

```
value=0;
if mode[i]=111 then value=P[1,1]/(P[1,1]+P[1,2]);
else if mode[i]=112 then value=P[1,2]/(P[1,1]+P[1,2]);
else if mode[i]=211 then value=P[1,1];
else if mode[i]=212 then value=P[1,2];
if value>0 then sum=sum+num[i]*log(value);
else return(.);
```

p[1,1]
p[1,2]
p[1,3]
p[2,1]
p[2,2]
p[2,3]
p[3,1]
p[3,2]
p[3,3]

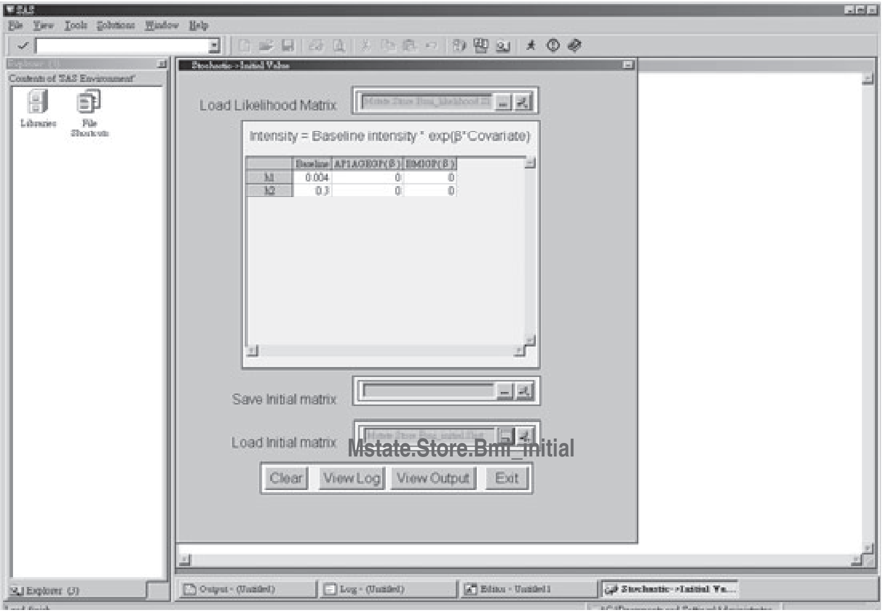
PS:Insert probability function

Save Likelihood matrix

Load Likelihood matrix: Mstate.Store.Bmi_likelihood.Slist

Mstate.Store.Bmi_likelihood

Figure 6 Formulation of the likelihood function.



Load Likelihood Matrix: Mstate.Store.Bmi_likelihood.Slist

Intensity = Baseline intensity * exp(β*Covariate)

	Baseline	AP1AGEGP(β)	BMIGP(β)
h1	0.004	0	0
h2	0.3	0	0

Save Initial matrix

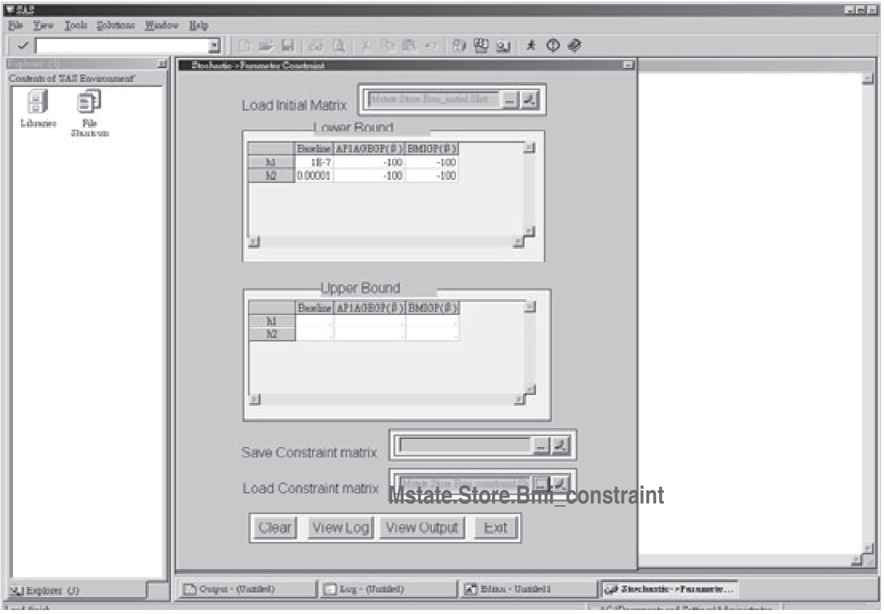
Load Initial matrix: Mstate.Store.Bmi_initial.Slist

Mstate.Store.Bmi_initial

Figure 7 Tailor for initial values of baseline intensity and regression coefficients.

The left panel in Fig. 6 includes two window menu for selecting intensity and relevant covariates. The right panel in Fig. 6 also provides all possible transition probabilities for the formulation of the likelihood function. Figure 7 shows the initial values of baseline intensity (0.004 and 0.3 for onset of PCDP and the transition from the PCDP phase to clinical phase) and regression coefficients (0 used in every occasion). Figure 8 shows upper and lower bounds for the estimated transition parameters. Figure 9 shows the estimated results. The predicted transition probabilities with time by different characteristics are also displayed in Fig. 10.

The user calculated the matrix of transition probabilities given time t and a combination of covariates. In this example, there are four combinations of age at first full-term pregnancy (1: late age; 0: early age for age at full-term pregnancy) and level of BMI (1: overweight; 0: normal), including 11 overweight + late age, 10 overweight + early age, 01 normal BMI + late age, and 00 normal BMI + early age.



Load Initial Matrix

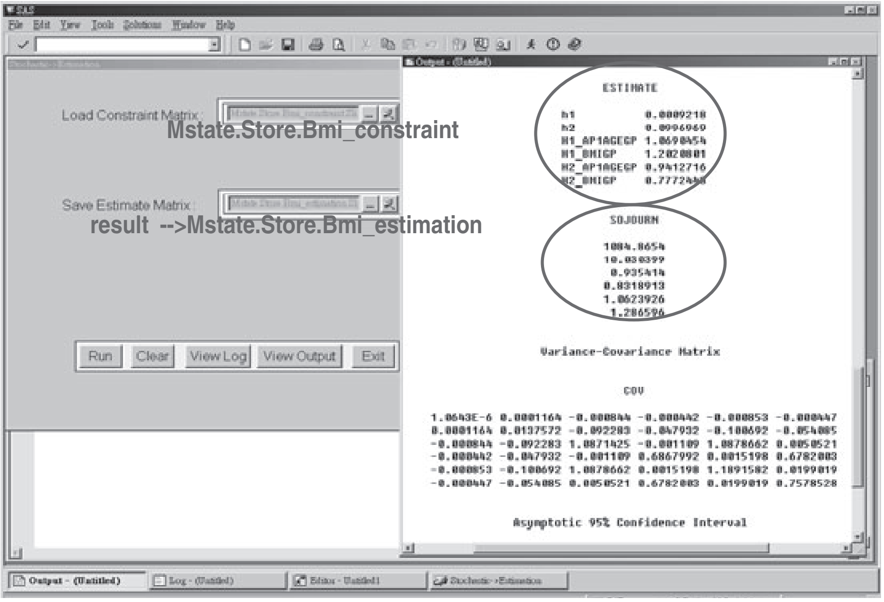
	Baseline	AP1AGEGP(β)	BMIGP(β)
h1	1E-7	-100	-100
h2	0.00001	-100	-100

	Baseline	AP1AGEGP(β)	BMIGP(β)
h1			
h2			

Save Constraint matrix

Load Constraint matrix Mstate.Store.Bmi_constraint

Figure 8 Tailor for upper and lower bounds for the estimated transition parameters.



Load Constraint Matrix : Mstate.Store.Bmi_constraint

Save Estimate Matrix : result -->Mstate.Store.Bmi_estimation

```
                 ESTIMATE

          h1          0.0009218
          h2          0.0996960
          H1_AP1AGEGP 1.0690454
          H1_BMIGP    1.2020801
          H2_AP1AGEGP 0.9412716
          H2_BMIGP    0.7772444

                 SOJOURN

                1004.8654
                10.030399
                 0.935414
                0.8318913
                1.0623926
                 1.286596


        Variance-Covariance Matrix


                   COV

1.0643E-6 0.0001164 -0.000844 -0.000442 -0.000853 -0.000447
0.0001164 0.0137572 -0.092283 -0.047932 -0.100692 -0.054085
-0.000844 -0.092283 1.0871425 -0.001109 1.0878662 0.0050521
-0.000442 -0.047932 -0.001109 0.6867992 0.0015198 0.6782083
-0.000853 -0.100692 1.0878662 0.0015198 1.1891582 0.0199019
-0.000447 -0.054085 0.0050521 0.6782083 0.0199019 0.7578528


       Asymptotic 95% Confidence Interval
```

Figure 9 Show the estimated results.

Deterministic model

The deterministic model (Fig. 11) is tailored for the user when the transition parameters have been already known.

Figures 12 and 13 show the three parameters of the three-state model by two age groups, ≤50 and >50 years, respectively. Figures 14 and 15 yield the predicted probability and curves.

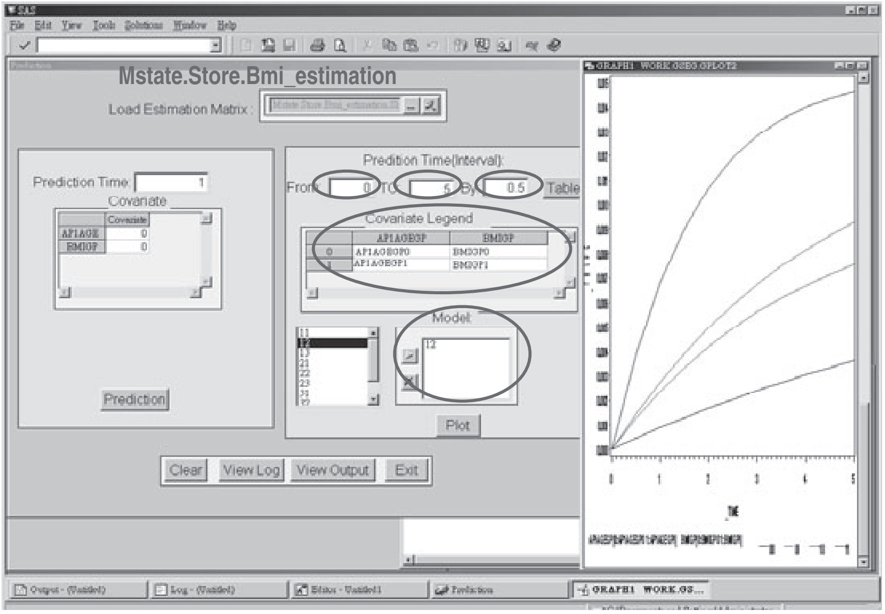
Model validation

Figures 16 and 17 show how goodness of fit for interval and external model validation was conducted. The tot[i] variable represents the total number of attendants at time i. The expected cases, fitted[i], were calculated by multiplying tot[i] by transition probabilities. The formula is similar to the likelihood function used for estimating transition parameters. Figure 18 shows the results of observed and expected values. The chi-square and *P*-value of goodness of fit was shown in Fig. 19.

Application to organized cancer screening

Structure of organized screening

We applied the multi-state disease prediction system mentioned above to develop a subsequent computer-aided system for analysis



Mstate.Store.Bmi_estimation

Load Estimation Matrix : Mstate.Store.Bmi_estimation

Prediction Time: 1

Covariate

	Covariate
AP1AGE	0
BMIGP	0

Predition Time(Interval):

From 0 TO 5 By 0.5

Covariate Legend

	AP1AGEGP	BMIGP
0	AP1AGEGP0	BMIGP0
1	AP1AGEGP1	BMIGP1

Model: 12

Figure 10 Predicted transition probabilities with time by different characteristics.

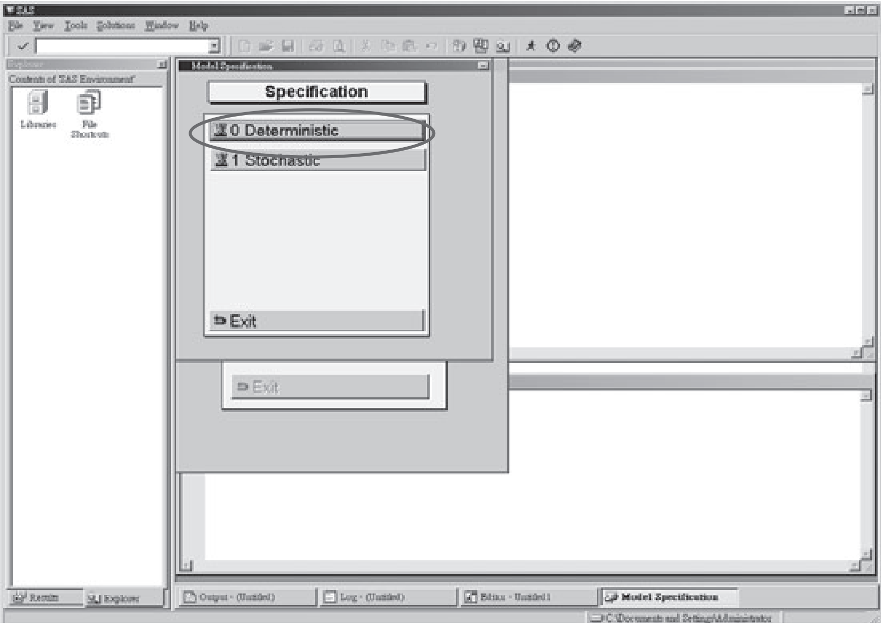


Figure 11 Menu for the deterministic model.

of organized screening, particularly for the determination of inter-screening interval. The structure of organized screening is illustrated in Fig. 20.

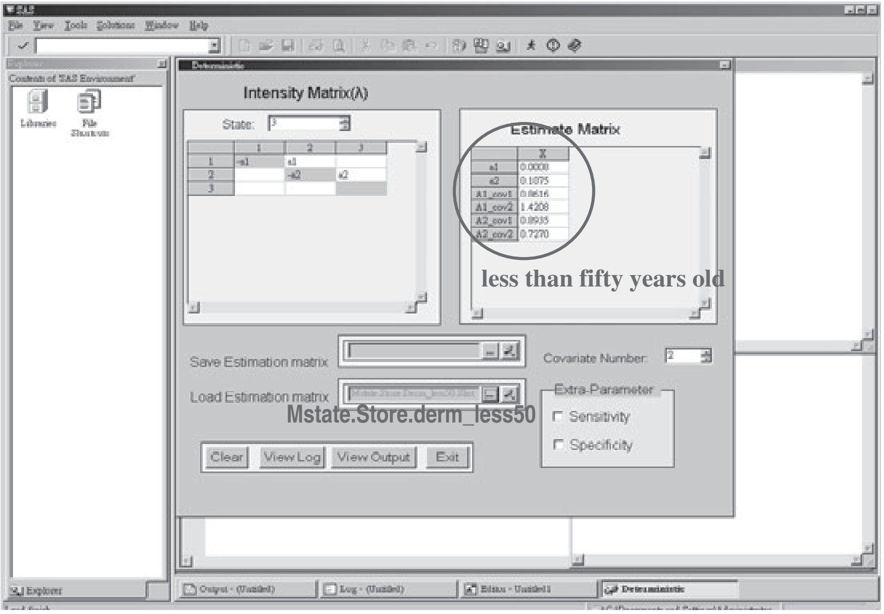
In the organized screening, the target population was first invited to screen. The parameter of attendance rate determines number of attendants. Those who refuse to attend the screening are called the refuser group, whose outcome follows the disease natural history determined by relevant transition parameters from either the stochastic or deterministic model.

Among attendants, the combination of transition probabilities, sensitivity and specificity yields four possible outcomes: true negative cases, false negative cases, true positive cases (screen-detected cases) and false positive cases.

This forms a cycle of organized screening. Next cycle will follow the same flow chart. It should be noted that as the characteristics of the first screen are usually different from the repeated screen. Analysis of organized screen is usually divided by the first screen and the repeated screen.

Hypothetical cohort and prediction

Suppose we have 10 000 women with first-degree relatives afflicted with breast cancer. The distribution of age (73.65% aged



Intensity Matrix(λ)

State: 3

	1	2	3
1	-a1	a1	
2		-a2	a2
3			

Estimate Matrix

	X
a1	0.0008
a2	0.1075
A1_cov1	0.8616
A1_cov2	1.4208
A2_cov1	0.8935
A2_cov2	0.7270

less than fifty years old

Save Estimation matrix:

Load Estimation matrix: Mstate.Store.derm_less50

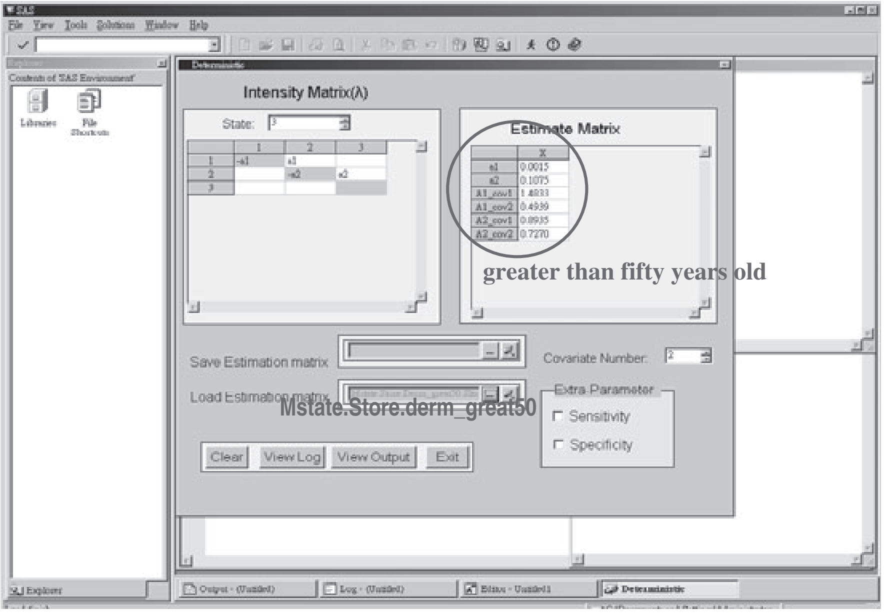
Covariate Number: 2

Extra-Parameter

□ Sensitivity

□ Specificity

Figure 12 Three parameters of the three-state model for young women (<=50 years old).



Intensity Matrix(λ)

State: 3

	1	2	3
1	-a1	a1	
2		-a2	a2
3			

Estimate Matrix

	X
a1	0.0015
a2	0.1075
A1_cov1	1.4833
A1_cov2	0.4939
A2_cov1	0.8935
A2_cov2	0.7270

greater than fifty years old

Save Estimation matrix:

Load Estimation matrix: Mstate.Store.derm_great50

Covariate Number: 2

Extra-Parameter

□ Sensitivity

□ Specificity

Figure 13 Three parameters of the three-state model for old women (>50 years old).

≤50 years, and 26.35% aged >50 years) at first full-term pregnancy, and BMI follows the underlying population illustrated above.

Suppose the screen period covers 6 years, and numbers of screens were 7, 4, 3 and 2 for annual, biennial, triennial and six-yearly screening regimes. Assume the estimates of sensitivity and specificity using mammography are 80% and 90%.

Suppose the attendance rate is 80%, of 8000 attendant at the first screen, number of screen-detected cases (PCDP_1), true negative case (Normal_1), false positive cases (FP_1), and false negative cases (FN_1) are calculated by the following formula:

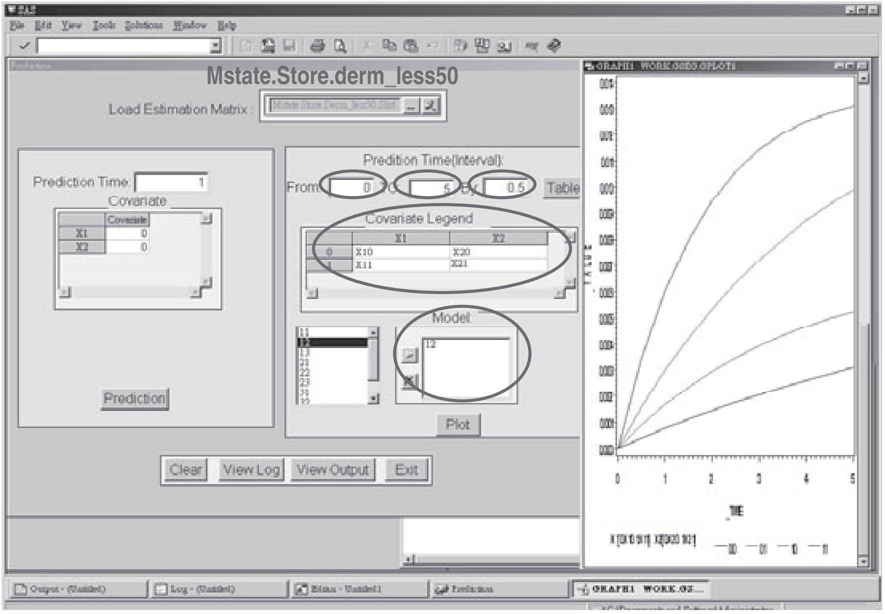
$$\mathrm{PCDP}_1 = \frac{p_{12}(\mathrm{age}) \times S}{p_{11}(\mathrm{age}) + p_{12}(\mathrm{age})} \times 8000$$

$$\mathrm{Normal}_1 = \frac{p_{11}(\mathrm{age}) \times SP}{p_{11}(\mathrm{age}) + p_{12}(\mathrm{age})} \times 8000$$

$$\mathrm{FP}_1 = \frac{p_{11}(\mathrm{age}) \times (1 - SP)}{p_{11}(\mathrm{age}) + p_{12}(\mathrm{age})} \times 8000$$

$$\mathrm{FN}_1 = \frac{p_{12}(\mathrm{age}) \times (1 - S)}{p_{11}(\mathrm{age}) + p_{11}(\mathrm{age})} \times 8000$$

$$\text{Interval case}_1 = \mathrm{FN}_1 \times p_{23}(\mathrm{x}) + \mathrm{FP}_1 \times p_{13}(\mathrm{x}) + \mathrm{Normal}_1 \times p_{13}(\mathrm{x})$$



Mstate.Store.derm_less50

Load Estimation Matrix: Mstate.Store.Derm_less50.Ilist

Prediction Time: 1

	Covariate
X1	0
X2	0

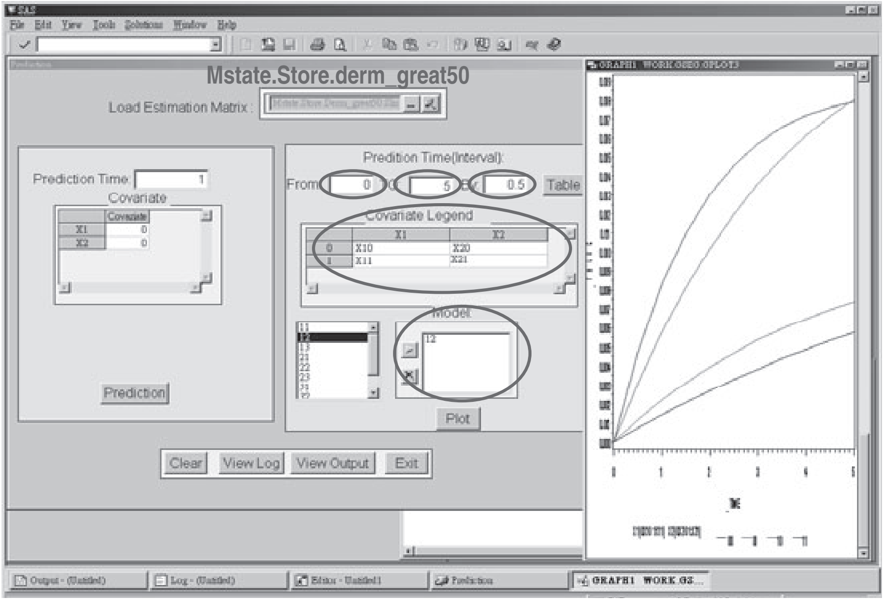
Predition Time(Interval): From 0 To 5 By 0.5

Covariate Legend

	X1	X2
0	X10	X20
1	X11	X21

Model: 12

Figure 14 Predicted probability and curves for young women (<=50 year old).



Mstate.Store.derm_great50

Load Estimation Matrix: Mstate.Store.Derm_great50.Ilist

Prediction Time: 1

	Covariate
X1	0
X2	0

Predition Time(Interval): From 0 To 5 By 0.5

Covariate Legend

	X1	X2
0	X10	X20
1	X11	X21

Model: 12

Figure 15 Predicted probability and curves for old women (>50 year old).

where age represents the age when subject attending the first screen, x is for screening interval, SP for specificity, and S for sensitivity.

Note that P_{1j}(age) (i = 1,2,3) is calculated by the application of prediction system of multi-state disease process with 40 and 60 of prediction time for age ≤ 50 years and age > 50 years, respectively. The reason of using conditional probability in the formula mentioned above is that clinical cases defined as state 3 are excluded from the eligible population. The numbers of screen-detected cases ($\mathrm{PCDP_j}$), true negative cases ($\mathrm{Normal_j}$), false positive cases ($\mathrm{FP_j}$) and false negative case ($\mathrm{FN_j}$) at jth screen, and interval cases after *j*th screen are also calculated by the following equations:

$$\mathrm{PCDP_j} = (\mathrm{Normal_{j-1}} + \mathrm{FP_{j-1}}) \times p_{12}(\mathrm{x}) \times S + FN_{\mathrm{j-1}} \times p_{22}(\mathrm{x})$$

$$\mathrm{Normal_j} = (\mathrm{Normal_{j-1}} + \mathrm{FP_{j-1}}) \times p_{11}(\mathrm{x}) \times SP$$

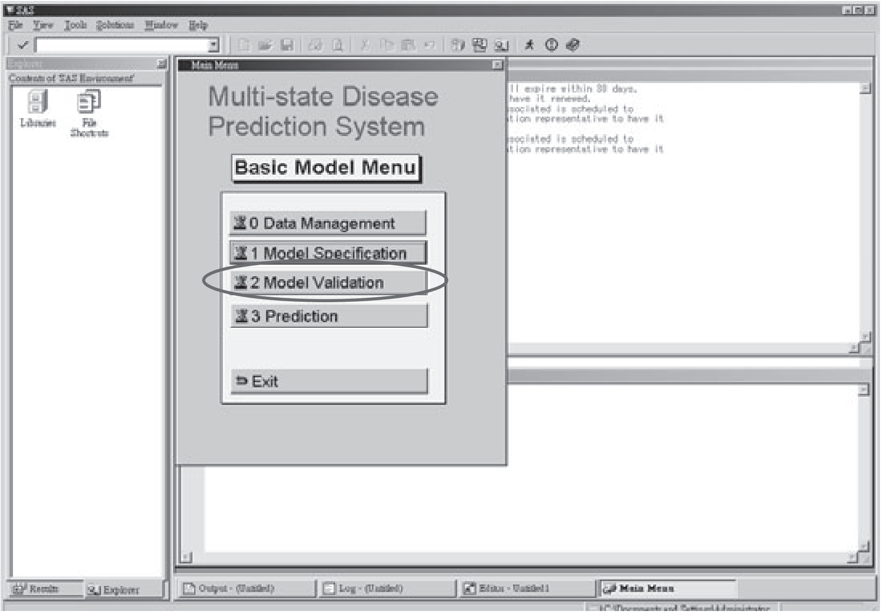
$$\mathrm{FP_j} = (\mathrm{Normal_{j-1}} + \mathrm{FP_{j-1}}) \times p_{11}(\mathrm{x}) \times (1 - SP)$$

$$\mathrm{FN_j} = (\mathrm{Normal_{j-1}} + \mathrm{FP_{j-1}}) \times p_{12}(\mathrm{x}) \times (1 - S)$$

$$\text{Interval case}_\mathrm{j} = \mathrm{FN_j} \times p_{23}(\mathrm{x}) + \mathrm{FP_j} \times p_{13}(\mathrm{x}) + \mathrm{Normal_j} \times p_{23}(\mathrm{x})$$

where the transition probabilities were also derived from transition probability.

The number of breast cancer cases arising from the refuser (RF) group in the 6 years is given by



```
ll expire within 30 days.
have it renewed.
ssociated is scheduled to
tion representative to have it

ssociated is scheduled to
tion representative to have it
```

Figure 16 Model validation component of main menu.

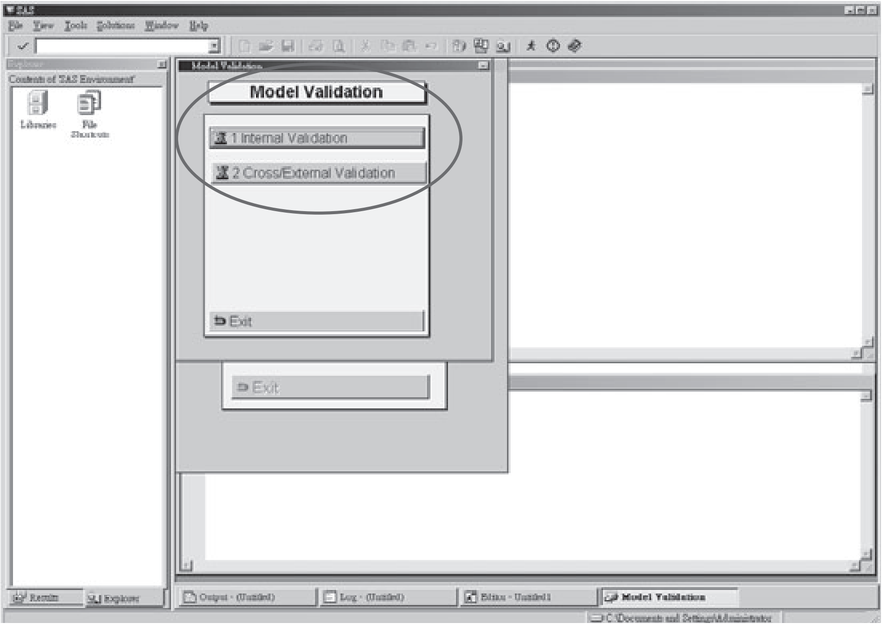
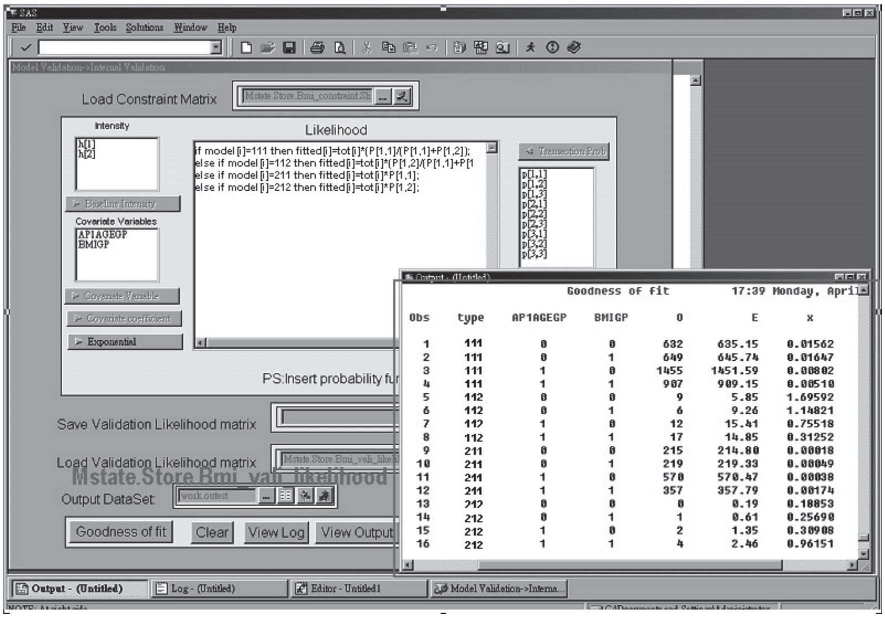


Figure 17 Menu for internal and external model validation.

$$\mathrm{RF}=\frac{p_{11}(\mathrm{age})\times p_{13}(6)+p_{12}(\mathrm{age})\times p_{23}(6)}{p_{11}(\mathrm{age})+p_{12}(\mathrm{age})}\times 2000$$

Table 2 lists screen-detected cases at the first screen and the incident screen, interval cancers, and refuser by different inter-screening intervals in accordance with the formula indicated above. It can be seen that the later age at first full-term pregnancy or the more the obese, the higher the risk for breast cancer. In Table 2, of 10 000 women, the number of predicted breast cancer increase from 105 in non-obese women with early age at first full-term pregnancy to 766 in obese women with late age at first full-term pregnancy during the 6 years of study. Within the identical risk group, the longer the inter-screening interval, the more the number of interval cancers and the less the incident screen-detected cases. Table 2 also provides the reference for the determination of individualized screening policy. This can be demonstrated by the incidence rate of interval cancer as a percentage of the expected incidence rate (I/E ratio). The expected incidence rate is often available from registry or primary data. In this example, this underlying incidence rate can be easily calculated from the study group during the 6-year period. The I/E ratio in the highest risk group (obese and late age at first full-term pregnancy) increased from 20% for annual screening regime to 51% for six-yearly screening regime, whereas the corresponding figure for the lowest risk group (non-obese and late age at first full-term pregnancy) is increased from 4% to 15%. This suggests that annual



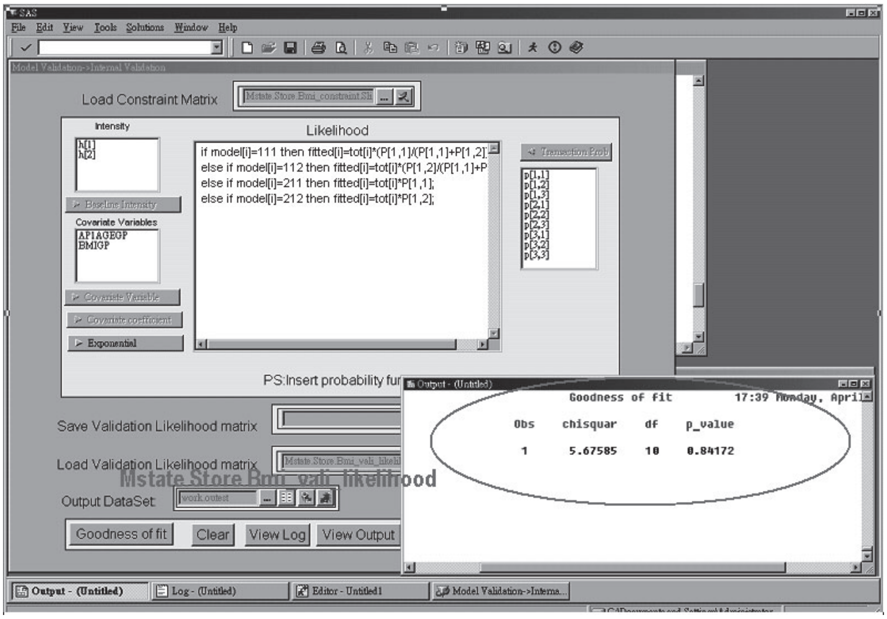
```
if model [i]=111 then fitted[i]=tot[i]*(P[1,1]/(P[1,1]+P[1,2]);
else if model [i]=112 then fitted[i]=tot[i]*(P[1,2]/(P[1,1]+P[1
else if model [i]=211 then fitted[i]=tot[i]*P[1,1];
else if model [i]=212 then fitted[i]=tot[i]*P[1,2];
```

```
                    Goodness of fit        17:39 Monday, April

Obs   type   AP1AGEGP   BMIGP      O         E         x

  1    111       0        0       632     635.15   0.01562
  2    111       0        1       649     645.74   0.01647
  3    111       1        0      1455    1451.59   0.00802
  4    111       1        1       907     909.15   0.00510
  5    112       0        0         9       5.85   1.69592
  6    112       0        1         6       9.26   1.14821
  7    112       1        0        12      15.41   0.75518
  8    112       1        1        17      14.85   0.31252
  9    211       0        0       215     214.80   0.00018
 10    211       0        1       219     219.33   0.00049
 11    211       1        0       570     570.47   0.00038
 12    211       1        1       357     357.79   0.00174
 13    212       0        0         0       0.19   0.18853
 14    212       0        1         1       0.61   0.25690
 15    212       1        0         2       1.35   0.30908
 16    212       1        1         4       2.46   0.96151
```

Figure 18 Results of observed and expected values.



```
if model[i]=111 then fitted[i]=tot[i]*(P[1,1]/(P[1,1]+P[1,2]
else if model[i]=112 then fitted[i]=tot[i]*(P[1,2]/(P[1,1]+P
else if model[i]=211 then fitted[i]=tot[i]*P[1,1];
else if model[i]=212 then fitted[i]=tot[i]*P[1,2];
```

```
                 Goodness of fit        17:39 Monday, April

        Obs    chisquar    df    p_value

          1     5.67585    10    0.84172
```

Figure 19 Result of the chi-square and p-value of goodness of fit.

screening regime may be necessary for the high-risk group, but three-yearly regime may be sufficient for the low-risk group.

Discussion

Using SAS/SCL as a platform, the present study developed a computer-assisted system for estimating parameters of multi-state disease process in the first stage and calculated the predicted probabilities for various transitions between states at the second stage. Individual prediction by incorporating individual characteristics is also carried out by exponential Markov regression model. Based on the predicted probabilities, this system has been successfully applied to evaluation of the efficacy of organized screening.

By using the concept of stochastic model, the components of this system consist of the formulation of the intensity matrix, the derivation of transition probabilities, application of transition probabilities to write the likelihood function in the light of data with multi-state property, model validation, and the matrix of predicted transition probabilities at specific time. The unique character of this system lies in the fact that the application of stochastic model to estimation and prediction is a step-by-step procedure, including (1) estimating transition parameters based on empirical data first; (2) assessing whether the model is adequate; and (3)

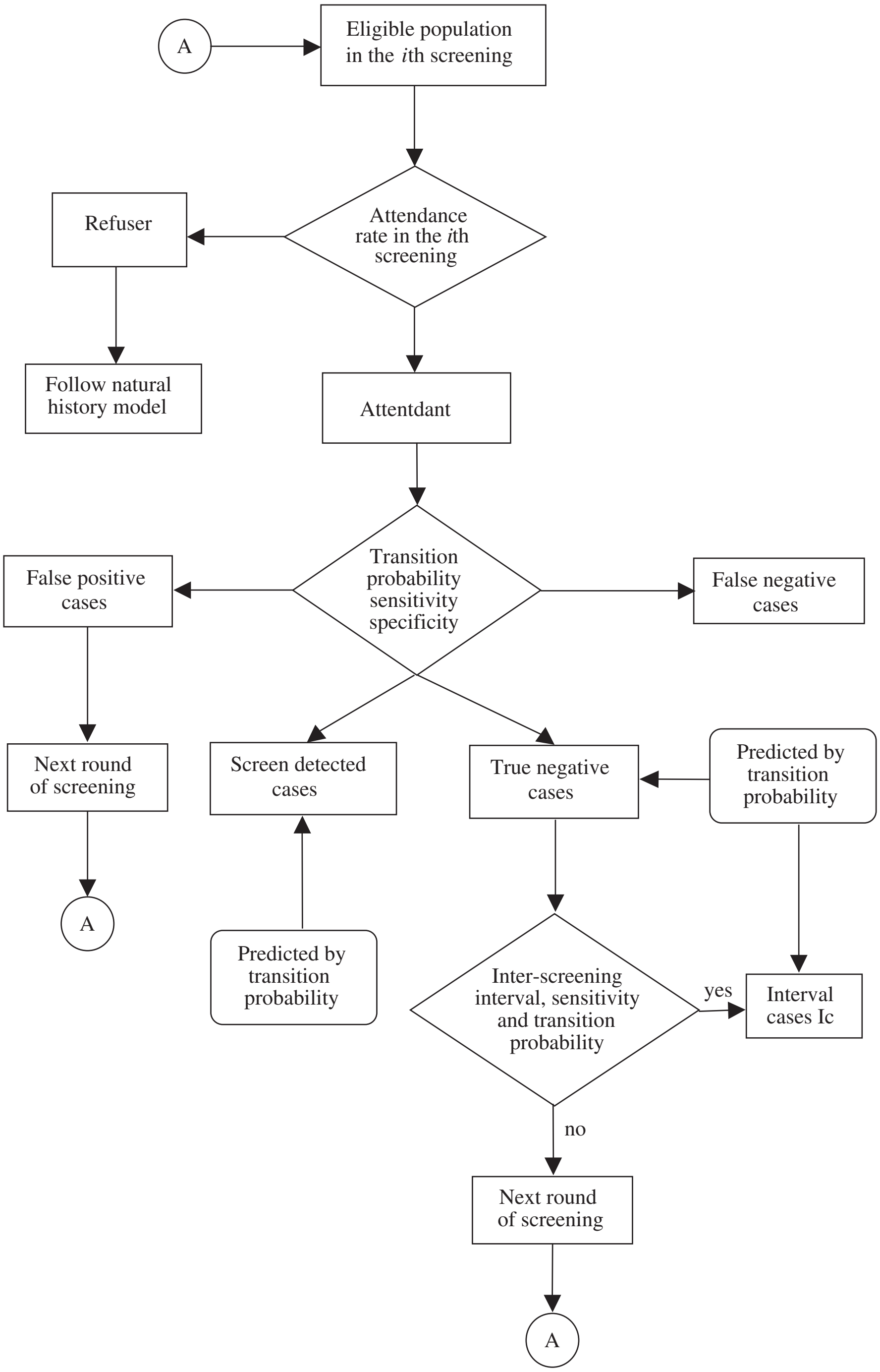

A
Eligible population in the *i*th screening
Refuser
Attendance rate in the *i*th screening
Follow natural history model
Attentdant
False positive cases
Transition probability sensitivity specificity
False negative cases
Next round of screening
Screen detected cases
True negative cases
Predicted by transition probability
A
Predicted by transition probability
Inter-screening interval, sensitivity and transition probability
yes
Interval cases Ic
no
Next round of screening
A

Figure 20 Structure of organized screening.

calculating the predicted value. In addition to the stochastic model, the system also includes the component of deterministic model for prediction without estimating parameters.

The usefulness of this model can be seen from several respects. First, the system enables one to estimate transition parameters dispensing with intractable programming language. This system is also efficient for users who are adept at writing estimation program because log view function provided in this system enables these users to debug or modify the program with ease. Second, our estimation and prediction system can be applied to predicting the risk or the probability of being one of multi-state outcomes arising from the event or disease history characterized by a multi-state disease process such as multi-state carcinogenesis in animal studies [8], the progression of cancer or chronic disease in screening [9], and five-stage behaviour change regarding transtheoretical model applied to smoking cessation [10]. Third, the function for prediction plays an important role in calculating risk of multi-state transitions by different individual characteristics. This is very important for medical consultation with clients. In our example of breast cancer screening, different risk groups have different risks

Table 2 Predictive breast cancer case of different detection modes by inter-screening interval risk groups and age

	Screening regime			
	Annual	Biennial	Triennial	Six-yearly
Age ≤ 50 years				
AP = 0 BMI = 0 N = 1167.2 NN = 291.8				
FS	6.8523	6.8523	6.8523	6.8523
IS	6.5344	5.9126	5.3784	4.1689
IC	0.5518	1.0163	1.4092	2.2747
RF	1.3830	1.3830	1.3830	1.3830
Total	15.3214	15.1642	15.0228	14.6789
AP = 0 BMI = 1 N = 816.8 NN = 204.2				
FS	9.7305	9.7305	9.7305	9.7305
IS	15.2012	12.8771	11.0817	7.6185
IC	2.5956	4.5429	6.0355	8.9032
RF	4.0179	4.0179	4.0179	4.0179
Total	31.5452	31.1683	30.8655	30.2700
AP = 1 BMI = 0 N = 2633.6 NN = 658.4				
FS	15.1864	15.1864	15.1864	15.1864
IS	27.0319	22.4372	19.0018	12.6602
IC	5.4404	9.3712	12.2951	17.7051
RF	7.4379	7.4379	7.4379	7.4379
Total	55.0965	54.4326	53.9211	52.9895
AP = 1 BMI = 1 N = 1275.2 NN = 318.8				
FS	14.7167	14.7167	14.7167	14.7167
IS	42.0946	30.9660	24.0008	13.6171
IC	17.8811	28.1576	34.6241	44.4452
RF	14.6494	14.6494	14.6494	14.6494
Total	89.3416	88.4896	87.9910	87.4283
Age > 50 years				
AP = 0 BMI = 0 N = 523.2 NN = 130.8				
FS	5.8304	5.8304	5.8304	5.8304
IS	5.4715	4.9508	4.5031	3.4886
IC	0.4629	0.8528	1.1827	1.9104
RF	1.1704	1.1704	1.1704	1.1704
Total	12.9353	12.8044	12.6867	12.3999
AP = 0 BMI = 1 N = 788.0 NN = 197.0				
FS	6.9673	6.9673	6.9673	6.9673
IS	10.9440	9.2714	7.9797	5.4894
IC	1.8678	3.2685	4.3419	6.4036
RF	2.8841	2.8841	2.8841	2.8841
Total	22.6632	22.3912	22.1729	21.7444
AP = 1 BMI = 0 N = 376.8 NN = 94.2				
FS	7.5863	7.5863	7.5863	7.5863
IS	13.1225	10.8872	9.2128	6.1148
IC	2.6475	4.5640	5.9916	8.6363
RF	3.6635	3.6635	3.6635	3.6635
Total	27.0198	26.7010	26.4543	26.0010
AP = 1 BMI = 1 N = 419.2 NN = 104.8				
FS	6.6851	6.6851	6.6851	6.6851
IS	18.8607	13.8670	10.7391	6.0721
IC	8.0202	12.6343	15.5393	19.9536
RF	6.5956	6.5956	6.5956	6.5956
Total	40.1616	39.7819	39.5590	39.3063

AP, 1 if age at first full-term pregnancy is older than 25 years; 0 otherwise; body mass index (BMI), 1 if one's BMI (kg m^{-2}) is greater than 23; 0 otherwise.

N, number of attendance; NN, number of non-attendants; FS, first screen; IS, incident screens; IC, interval cancers; RF, refuser.

for breast cancer. Our proportional hazard form can easily incorporate a constellation of risk factors into the multi-state model. Factors responsible for different transitions between states may have different clinical implications. In the three-state model for breast cancer, initiator factors may be crucial for onset of breast cancer, and promoter may be pivotal in the progression from PCDP to clinical phase.

There is one limitation in this system. Non-homogeneous property was not included in the current system. For example, the Weibull distribution for accommodating constant, increasing or decreasing hazard may be needed as demonstrated in the previous study [4]. This should be incorporated in the system in the next version of software. In addition, the most possible barrier for using this system may be related to the formulation of the likelihood function. As the likelihood functions vary with data, it is difficult to render the formulation of likelihood function become a standard procedure. This can be solved by applying our prediction system to develop a series of specific programs in response to different applications.

In conclusion, a computer-aided stochastic and deterministic estimation and prediction system for multi-state disease process was developed. This system is useful for predicting individual risks. The predicted system can be applied to different data with the property of multi-state transition.

References

1. Walter, S. D. & Day, N. E. (1983) Estimation of the duration of a preclinical disease state using screening data. *American Journal of Epidemiology*, 118, 865–886.
2. Chang, C. M., Kuo, H. S., Chang, S. H., Chang, H. J., Liou, D. M. & Chen, T. H. H. (2005) Development and evaluation of a computer-aided disease prediction application software with SAS component language. *Journal of Evaluation in Clinical Practice*, 11, 139–159.
3. Chen, T. H. H., Kuo, H. S., Yen, M. F., Lai, M. S., Tabar, L. & Duffy, S. W. (2000) Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics*, 56, 167–172.
4. Hsieh, H. J., Chen, T. H. H. & Chang, S. H. (2002) Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine*, 21, 3369–3382.
5. Wu, H. M., Yen, M. F. & Chen, T. H. H. (2004) SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression. *Computer Methods and Programs in Biomedicine*, 75, 95–105.
6. Kalbfleisch, J. D. & Lawless, J. F. (1985) The analysis of panel data under a Markov assumption. *Journal of American Statistics Association*, 80, 863–871.
7. Lai, M. S., Yen, M. F., Kuo, H. S., Kong, S. L., Chen, T. H. H. & Duffy, S. W. (1998) Efficacy of breast-cancer screening for female relatives of breast-cancer-index cases: Taiwan multicentre cancer screening (TAMCAS). *International Journal of Cancer*, 78, 21–26.
8. McKnight, B. & Crowley, J. (1984) Tests for differences in tumor incidence based on animal carcinogenesis experiments. *Journal of American Statistics Association*, 79, 639–648.
9. Duffy, S. W., Chen, T. H. H., Tabar, L. & Day, N. E. (1995) Estimation of mean sojourn time in breast cancer screening using a Markov Chain Model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*, 14, 1531–1543.
10. Glanz, K., Rimer, B. K. & Lewis, F. M. (2002) Health Behavior and Health Education Theory: Research and Practice. San Francisco, CA: Jossey-Bass.