Hyeoun-Ae Park*

A Multistate Model for Coronary Heart Disease

—An Application to Different Prevention Strategies—

Ⅰ. Introduction

Coronary Health Disease (CHD) morbidity and mortality rates in the advanced nations rose until 1970's and have been declining since then. Between 1968 and 1982 there was approximately a 30% reduction in deaths from CHD in the U. S. A[1]. These trends may be associated with changes in eating and smoking habits over these time periods. The changes in medical care which occurred during this period of time also may have played a part in this reduction. Goldman and Cook[2] reviewed the literature regarding the effect of various potential explanations for the reduction in CHD mortality between 1968 and 1976. They estimated that changes in life style, specifically the reduction in serum cholesterol levels and cigarette smoking, accounted for more than half of the reduction in ischemic heart disease mortality. In comparison, medical intervention, with coronary care units and the medical treatment of clinical ischemic heart disease and hypertension, accounted for about 40% of the decline.

*Senior Researcher, Korea Institute for Population and Health.

[1] U.S. Bureau of the Census, *Statistical Abstract of the United States, 1986(106th Edition)*, Washington D.C., U.S. Bereau of the Census, 1985.

[2] Goldman, L. and E. F. Cook, "The decline in ischemic heart disease mortality rates: An analysis of the comparative effects of medical interventions and changes in lifestyle", *Annals of Internal Medicine*, Vol. 101, 1984, pp.825~836.

An alternative method to gain further insight into disease mechanism by describing the evolution of disease processes over a lengthy period of time and through various disease stages is to construct a mathematical probabilisitic model. Sacks and Chiang[3] proposed a stochastic model for the study of CHD to describe the transitions from the healthy state and the state of having non-fatal CHD to death from CHD or other causes. However, they assumed that the probabilities of changes in states were independent of patient risk factors and also time-invarient. Several biostatistical techniques are available to relate the outcome of a chronic disease to individual patient characteristics. For example, in a stochastic compartment model, one (a) specifies a set of discrete health states, (b) specifies functions to describe transitions between those states, and (c) fits functions to time-specific morbidity and mortality data. The disease process of CHD is modeled as a stochastic compartment model (a) to study the contribution of individual risk factors to the changes in CHD morbidity and mortality rates and (b) to test the effectiveness and efficiency of different risk factor intervention strategies.

The discrete-time stochastic model presented in this paper pays attention on the association between physiological risk factors and CHD morbidity as well as mortality trends. Transition probabilities are modeled as a polychotmous logistic function of risk factors. A series of separate simple logistic regression analyses proposed by Begg and Gray[4] are performed as a replacement for estimating polychotomous logistic regression parameters. Data from the Finnish North Karelia Project[5] is used to test the model.

The model's goodness of fit is tested by comparing the number of events expected in deciles of estimated 6-year risk with the number of cases observed for each of the endpoints. Monte Carlo simulation, sequentially applying computer-generated ran-

[3] Sacks, S. T. and C. L. Chiang, "A Transition-Probability Model for the Study of Chronic Diseases", *Mathematical Biosciences,* Vol.34, 1977, pp.325~346.

[4] Begg, C. B. and R. Gray, "Calculation of Polychotomous Logistic Regression Parameters Using Individualized Regressions", *Biometrika,* Vol.71, No.1, 1984, pp.11~18.

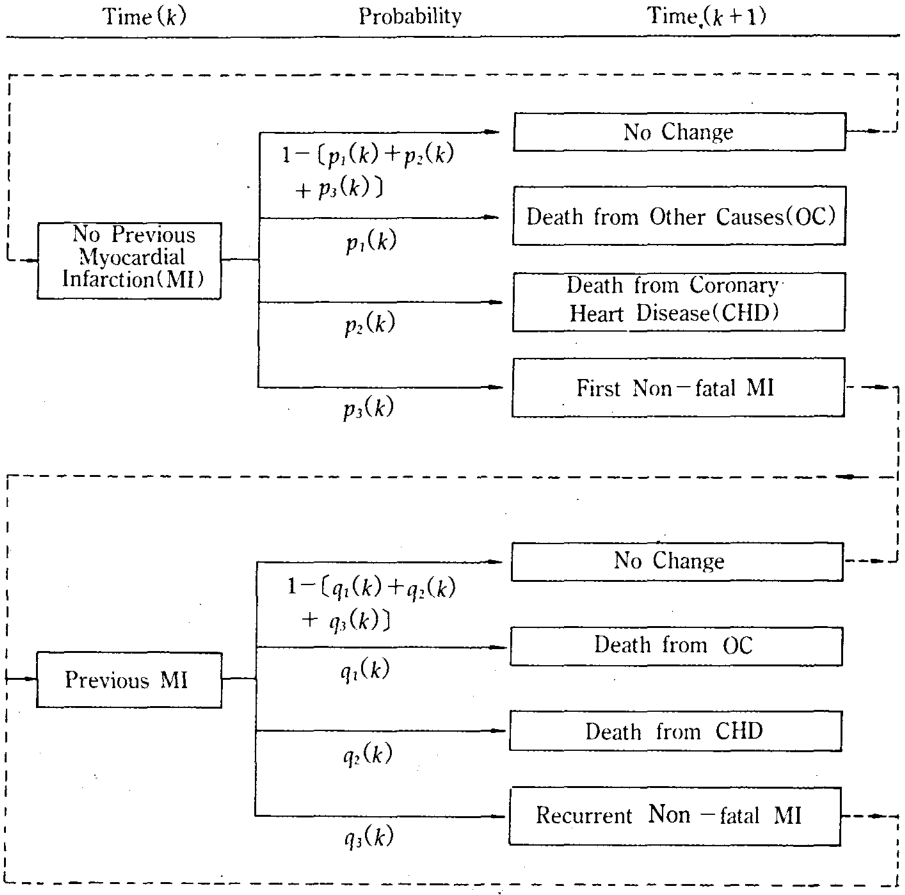
[5] Puska, P., J. Tuomilehto, J. Salonen and et al., *The North Karelia Project. Evaluation of a Comprehensive Community Programme for Control of Cardiovascular Diseases from 1972~1977 in* North *Karelia,* Finland, Copenhagen, World Health Organization, 1981.

dom numbers to determine transitions, is used to study the projected effects of various preventive intervention strategies, i. e., reducing the risk factor levels.

II. Model

The dynamics of CHD in a cohort population can be illustrated by a model as shown in Fig. 1, where the transition probabilities [$p(k)$'s and $q(k)$'s] define the proportion of the population making the transition from one state to another during a specified period of time k. These probabilities depend on history of prior MI and risk factor levels.

Figure 1. A Model for Coronary Heart Disease



This model is designed to investigate CHD morbidity and mortality rates from CHD as well as non-CHD causes. There are two initial states: free of MI and already affected with MI during the kth time interval. Healthy men free of MI of the first branch either stay healthy or proceed to one of three states: two death states and one illness state. The death states include death from CHD and death from other causes. The illness state is experiencing a non-fatal MI. The survivors of a first non-fatal MI are moved to the 'previous MI' pool and included in the group at risk when the next time interval begins. The second branch starts with those who have a history of MI. These men may encounter the same possible events as the first branch, but their transition probabilities are different. Survivors of recurrent non-fatal MI reenter the 'previous MI' pool when the next time interval starts.

Probabilities $p_1(k)$, $p_2(k)$ and $p_3(k)$ are defined as the likelihood that a healthy individual will die from non-CHD causes, die from CHD, or suffer a non-fatal MI, respectively during the kth time interval. The probability of no change from healthy status, say $p_0(k)$, is 1 minus the sum of the three probabilities of having any events. The q probabilities are similarly defined except that they refer to individuals who have a history of non-fatal MI prior to the kth time interval.

The risk of the selected endpoint $(d=1,\cdots,3)$ for an individual i at the kth time interval can be written as

$$p_{di}(k)=\frac{1}{1+exp\{-f[\underset{\sim}{\beta}_d(k),\ \underset{\sim}{x}_i]\}}$$

where $f[\underset{\sim}{\beta}_d(k),\ \underset{\sim}{x}_i]=\beta_{d1}(k)*x_{i1}+\beta_{d2}(k)*x_{i2}+,\cdots,+\beta_{dp}(k)*x_{ip}$,

$\underset{\sim}{\beta}_d(k)=(\beta_{d1}(k),\cdots,\beta_{dp}(k))$ is a vector of unknow parameters,

and $\underset{\sim}{x}_i=(x_{i1},\cdots,x_{ip})$ are possible risk factors with x_{i1} an indicator for the constant term.

III. Methods

There are three distinct transition and a 'no change endpoints for each of the two branches of the model. When risk factors are incorporated into the model to affect the transition probabilities to more than two endpoints from the intial state, a polychotomous logistic regression analysis may be used. However, only a limited number of software packages are available for the polychotomous logistic regression analyses. It also presents some computational limitations, such as the inability to accommodate sparse data. For these reasons, the individualized logistic regression technique proposed by Begg and Gray is utilized. In this technique, each endpoint is individually compared with the 'no change' baseline state using simple dichotomous logistic regression models. It has been shown that the maximum likelihood estimators (MLE) of individualized regression coefficients are asymptotically unbiased estimators of polychotomous logistic regression parameters. The relative efficiencies of MLE using the individualized model as compared with the polychotomous model are observed to be generally high. Suppose we have p-1 covariates, x_{ij} $(j=2,\cdots,p)$, with x_{i1} an indicator for the constant term, for n cases $(i=1,\cdots,n)$. By suppressing the index for time unit, let $\underset{\sim}{\beta}_d=(\beta_{d1},\cdots,\beta_{dp})$ be the individual logistic regression parameter vector for the tranistion probabilities from baseline category to the dth category. Here, denote 'no change' as the baseline category, death form non-CHD causes as the first category, CHD death as the second category, and non-fatal MI as the third category. Let z_{di} be an indicator variable which takes the value 1 if the ith individual belongs to the dth category and 0 otherwise.

Let $\phi_{di}=\mathrm{pr}(z_{di}=1/\underset{\sim}{x}_i) \qquad (d=0,\cdots,\mathrm{T}),$

where

$$\phi_{di}=\frac{1}{1+\exp(-\underset{\sim}{\beta}_d'\underset{\sim}{x}_i).}$$

The polychotomous model is based on the assumption that

$$\log(\phi_{di}/\phi_{oi})=\underset{\sim}{\beta}_d'\,\underset{\sim}{x}_i \qquad (d=1,\cdots,\mathrm{T})\text{-----------}(1)$$

where $\underset{\sim}{\beta}_d=(\beta_{d1},\cdots,\beta_{dp})$ is a p-vector of unknown parameters. An individual logistic regression comparing category d with the normal category, denoted category 0, would have a model of the form

$$\log(\theta_{di}/\theta_{oi})=\underset{\sim}{r}_d'\underset{\sim}{x}_i \qquad (d=1,\cdots,\mathrm{T}]\text{----------}(2)$$

where $\theta_{di} = \mathrm{pr}(z_{di} = 1 / \underset{\sim}{x}_i, z_{oi} + z_{di} = 1)$

$\theta_{oi} = \mathrm{pr}(z_{oi} = 1 / \underset{\sim}{x}_i, z_{oi} + z_{di} = 1)$

$= 1 - \theta_{di},$

and where $\underset{\sim}{r}_d$ is a p-vector of unknown parameters. It can be verified that the two models are parametically equivalent, that is, $\underset{\sim}{r}_d = \underset{\sim}{\beta}_d (d = 1, \cdots, \mathrm{T})$.
Using Bayes' theorem,

$$\theta_{di} = \phi_{di} / (\phi_{oi} + \phi_{di}),$$

therefore,

$$\phi_{di} / \phi_{oi} = \theta_{di} / (1 - \theta_{di}).$$

Consequently, if the individualized method is adopted and T separate logistic analyses are performed using (2) for $d = 1, \cdots, \mathrm{T}$, the resulting parameter estimates, denoted by $\hat{r}_1, \cdots, \hat{r}_\mathrm{T}$ may be substituted in (1) to obtain predicted probability estimates. Moreover, if maximun likelihood estimation is employed, the estimates, $\underset{\sim}{\hat{r}}_d$ will be asymptotically unbiased.

Since the sum of all possible transition probabilities must be equal to 1, normalization is achieved as follows:

$$\theta_{oi} + \sum_{d=1}^{\mathrm{T}} \theta_{di} = 1$$

where $\theta_{di} = \theta_{oi} * \exp(\underset{\sim}{r}_d' \underset{\sim}{x}_i)$ from (2) $(d = 1, \cdots, \mathrm{T})$.

Thus,

$$\theta_{oi}(1 + \sum_{d=1}^{\mathrm{T}} \exp(\underset{\sim}{r}_d' \ \underset{\sim}{x}_i)) = 1$$

Therefore,

$$\theta_{oi} = \frac{1}{1 + \sum_{d=1}^{\mathrm{T}} \exp(\underset{\sim}{r}_d' \ \underset{\sim}{x}_i)}$$

θ_{di} is computed in turn by using θ_{oi}

$\theta_{di} = \theta_{oi} * \exp(\underset{\sim}{r}_d' \underset{\sim}{x}_i) \ (d = 1, \cdots, \mathrm{T}).$

The miximum likelihood estimator, denoted as $\underset{\sim}{\hat{r}}_d$; and its individual variance may be obtained from standard output form simple logistic regression packages. The transition probability of the selected endpoint $(d = 1,2,3)$ for an ith individual at the kth time interval can be written as

$$p_{di}(k)=p_{oi}(k) * \exp(\underset{\sim}{\hat{r}_d}'\underset{\sim}{x_i}) \quad d=1,2,3 \text{ --------} (3)$$

$$p_{oi}(k) = \frac{1}{1+\sum_{d=1}^{T}\exp(\underset{\sim}{\hat{r}_d}(k)'\underset{\sim}{x_i})}$$

where $\underset{\sim}{\hat{r}_d}(k)=[\hat{r}_{di}(k),\cdots,\hat{r}_{dp}(k)]$ is a vector of unknown parameter, and $\underset{\sim}{x_i}=(x_{i1},\cdots,x_{ip})$ are possible risk factors with x_{i1} an indicator for the constant term.

Each transition probability are not only a function of risk factors, but vary over time period. To incorporate this time dependency into the model, year-specific crude transition probabilities are examined first. The sample data are then aggregated for each time period if the crude year-specific transition probabilities are shown to be at the same level. For example, in the six-year follow-up study for the North Karelia project, we found that the transition probabilities could be divided into three periods of two years each, within which the probabilities are at about the same level. These observations have led us to properly aggregated data for the construction of likelihood functions. In practical applications, using yearly data for the estimation of year-specific regression coefficients tends to provide less significant coefficients and thus less conclusive inferences about the effects of covariates. This is due to the fact that there are smaller number of events or occurrences than the aggregated data. To check the validity of the model, the observed number of events were compared with the expected number of events by deciles of estimated 6-year risk in each event category separately.

Two strategies are employed using Monte Carlo simulations to study the potential benefits of preventing MI incidence or CHD death by lowering serum cholesterol levels. The first strategy concentrates on the high risk group, for example, reducing the cholesterol level of people in the top decile of the cholesterol distribution to 180 mg/dl. The second strategy reduces the cholesterol level of the entire population by a certain amount. The percent reduction in MI incidence, CHD mortality rates, or all cause mortality rates for each stratgy are computed and compared.

This paper utilized simulation software tools from the Resource for Simulation of Stochastic Micropopulation Models located at University of Minnesota and funded by the NIH Biomedical Research Technology Program. The purpose of the Resource is to advance the use of Monte Carlo simulation of structured population for biomedical research. The emphasis is on development of simulation and analysis methods and software, with application to epidemiological research studies. Stochastic micropopulation models allow a variety of hypotheses about spread and control of disease to be tested, and provide information on the effects of population structure and random variation. Core and collaborative researches, training and dissemination activities are being occurred in the areas of chronic, genetic and infectious diseases.

Ⅳ. Results

The dataset used for this study consists of 3,022 healthy men, free of MI initially, aged 40-59 in 1972. This cohort was defined as part of the baseline evaluation of the North Karelia Project. It includes risk factors levels such as serum cholesterol and diastolic blood pressure measured in 1972 and the dates and types of follow-up events: MIs or death between 1972 and 1977.

Table 1 shows the summary of descriptive statistics of the risk factors. The yearly morbidity rates as well as mortality rates for both CHD and other causes are presented in Table 2. Using Table 2, data for each two-year period were aggregated as described in the previous section to estimate individualized logistic regression parameters. Normalized regression coefficients and standard errors are presented in Table 3. Both risk factors have significant effects on the incidence of MI and CHD death but not on the death from other causes as expected. However, cholesterol clearly has a greater effect than diastolic blood pressure. The transition probabilities to various states during

Table 1. Descriptive Statistics of Risk Factors in the North Karelia Dataset

	Mean± s. d.	Range*	75 Percentile
Cholesterol	274± 49	108,486	303
Diastolic BP	94± 12	48,148	100

* Range values are for the minimum and maximum values.

Table 2. Numbers of Events Occurred Between 1972 and 1977

Year	NO MI				Previous MI			
	At Risk*	New MI	CHD Death	Other Death	At Risk*	Non-fatal MI	CHD Death	Other Death
1972	3022	19(.6%)+	8(.3%)	20(.7%)	0	0	0	0
1973	2975	27(.9%)	16(.5%)	19(.6%)	19	1	2	0
1974	2913	31(1.1%)	14(.5%)	27(.9%)	44	5	3	1
1975	2841	26(.9%)	9(.3%)	23(.8%)	71	1	4	3
1976	2783	13(.5%)	16(.5%)	20(.7%)	90	3	2	1
1977	2734	18(.7%)	12(.4%)	33(1.2%)	100	1	0	3

* Number of people at risk at the beginning of the year

+ The percentage is obtained as the ratio of number of occurrences to the people at risk

Table 3. Normalized Logistic Regression Coefficients and Standard Errors

State	Year	Intercept	Cholesterol	Diastolic BP
New MI	72~73	−4.93	.40(.14)*	−.25(.15)
	74~75	−4.66	.35(.12)*	−.04(.13)
	76~77	−5.50	.64(.15)*	.40(.17)*
CHD Death	72~73	−5.81	.36(.19)+	.62(.17)*
	74~75	−5.61	.38(.19)*	.21(.20)
	76~77	−5.47	.49(.17)*	.34(.18)+
Other Death	72~73	−5.02	−.01(.16)	−.11(.17)
	74~75	−4.82	−.17(.15)	.42(.13)*
	76~77	−4.64	.03(.14)	.15(.13)

* Significant at 0.05 level

+ Significant at .10 level

each time period for each individual could be determined from equation (3) and used to generate events in the Monte Carlo simulations.

Table 4 shows the comparisons between observed and expected number of events by deciles of 6-year risk scores. Chi-squared statistics, as compared with the critical value at $\alpha=0.05$ and 7 degrees of freedom, i. e. 14.07, indicate that the proposed model fits the observed data significantly.

Figure 2 shows the simulated percent reduction of CHD mortality in a six-year period with various strategies using the North Karelia dataset. The solid line depicts the simulated outcome when the cholesterol levels of various proportions of high risk people are reduced to 180mg/dl. For example, if the quartile at greatest risk reduce their cholesterol level to 180mg/dl using drugs, over six-year period there

Table 4. Observed (o) Number of Events and Expected (e) Number of Events by Decile of Estimated 6 Year Risk (3022 North Karelia Men Aged 40~59)

Decile	Non-Fatal MI		CHD Death		Other Causes of Death	
	e	o	e	o	e	o
1	6.4	10	2.5	1	11.0	12
2	8.0	9	3.5	6	11.8	10
3	9.1	5	4.2	4	12.4	13
4	10.2	9	4.9	5	12.9	12
5	11.2	9	5.7	7	13.4	12
6	12.5	10	6.5	6	13.9	10
7	13.9	14	7.6	9	14.5	13
8	15.7	12	9.0	6	15.2	21
9	18.8	24	11.5	7	16.5	21
10	29.0	32	20.0	24	20.4	18
Total	134.8	134	75.4	75	142.0	142
Chi-square	7.70		6.93		5.59	

would be a 34% reduction in CHD deaths. The dotted line shows the outcome for population intervention. For instance, if every individual reduces their cholesterol level by 30mg/dl, there will be a 24% reduction in CHD death. These finding are consistent

Figure 2. Projected Changes in CHD Deaths with Cholesterol Intervention in North Karelia Men Aged 40~59

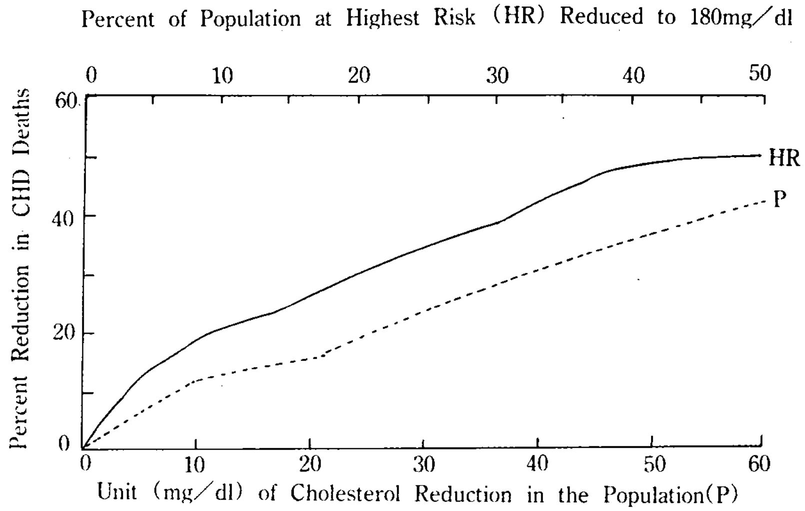


Figure 3. Projected Changes in New MI with Cholesterol Intervention in North Karelia Men Aged 40~59

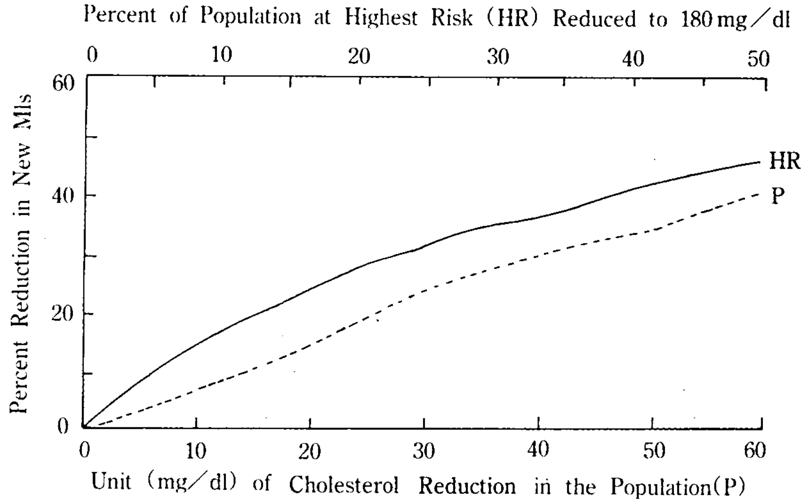
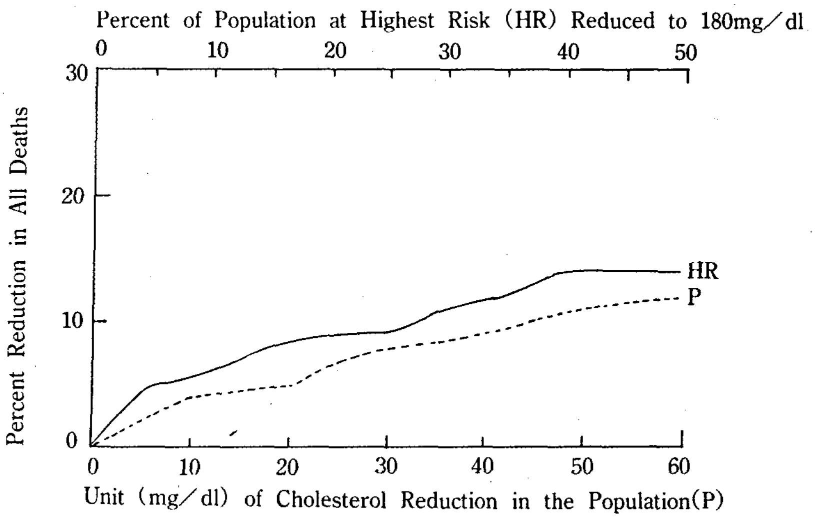


Figure 4. Projected Changes in All Deaths with Cholesterol Intervention in North Karelia Men Aged 40~59



with the findings of Kottke et al.'s study[6]. Figures 3 and 4 show the predicted percent reduction of MI incidence rates and all-cause mortality rates for various intervention strategies. A more detailed description of the relative effectiveness of high-risk versus population-based intervention is discussed in Park's work[7].

V. Conclusion

The advantages of the modeling system developed in this study are as follows: This modeling system is a multi-state and discrete-time one. Thus, it provides the opportunity to study the dynamic aspects of the disease progress over a period of

6) Kottke, T. E., P. Puska, J. T. Salonen, J. Tuomilehto and A. Nissinen, "Projected Effects of High-risk Versus Population-based Prevention Strategies in Coronary Heart Disease", *American Journal of epidemiology,* Vol.121, No.5, 1985, pp.679~703.

7) Park, H., *Simulation of a Population-based Model of Coronary Heart Disease Morbidity and Mortality.* Ph. D. Thesis in Biometry and Health Information Systems, University of Minnesota, Minneapolis, MN., U.S., Sep. 1989.

time. Also the changes of event rates for each time interval can be reflected to predict the event rates for each time interval as well as the entire study period.

The transition probabilites among states are dependent upon individual characteristics. Thus, it allows one to study the effect of risk factors on the trends of CHD incidence and prevalence as well as mortality rates, and the influence of a risk factor reduction of CHD morbidity and mortality. This model also may help policy makers understand the potential population impact and benefits of public health strategies which affect the probability of disease or death by changing the life style of the high risk group of the entire population.

Competing risks are incorporated in the model. Thus, mortality and morbidity of CHD as well as mortality from other causes can be studied at the same time. This allows the investigator to look at what happens to the mortality from other causes or total mortality rates when studying the effect of the intervention strategies on CHD mortality and morbidity rates. In order to simulate the probabilistic elements of CHD, this modeling system utilizes a Monte Carlo method to generate the events. In the Monte Carlo technique, the events are generated by the use of a random number generator and the opportunity to study the variability of the simulated event rates for each set of replications.

However, many tasks remain to be accomplished to expand and improve on the current modeling system, for example, it is necessary to develop more states in the evolution of MI such as emergency medical services, various levels of hospitalization and other CHD symptoms to Simulate and examine health service utilization and outcomes. It is possible to investigate some alternative functions for generating transition probabilities such as proportional hazards function, accelerated failure model, or power transformation to discriminate additive and multiplicative nature of risk factors.

References

Begg, C. B. and R. Gray, "Calculation of polychotomous logistic regression parameters using individualized regressions", *Biometrika,* Vol. 71, No.1, 1984, pp.11~18.

Goldman, L. and E. F. Cook, "The decline in ischemic heart disease mortality rates: An analysis of the comparative effects of medical interventions and changes in lifestyle", *Annals of Internal Medicine,* Vol.101, 1984, pp.825~836.

Kottke, T. E., P. Puska, J. T. Salonen, J. Tuomilehto and A. Nissinen, "Projected effects of high-risk versus population-based prevention strategies in coronary heart disease", *American Journal of Epidemiology,* Vol.121, No.5, 1985, pp.697~703.

Park, H., *Simulation of a Population-based Model of Coronary Heart Disease Morbidity and Mortality.* Ph. D Thesis in Biometry and Health Information Systems, University of Minnestota, Minneapolis, MN., U.S., Sep. 1989.

Puska, P., J. Tuomilehto, J. Salonen and et al., *The North Karelia Project. Evaluation of a Comprehensive Community Programme for Control of Cardiovascular Diseases from 1972~77 in North Karelia,* Finland, Copenhagen, World Health Organization, 1981.

Sacks, S. T. and C. L. Chiang, "A transition-probability model for the study of chronic diseases", *Mathematical Biosciences,* Vol.34, 1977, pp.325~346.

U.S. Bureau of the Census, *Statistical Abstract of the United States, 1986(106th Edition),* Washington D. C., U. S. Bereau of the Census, 1985.

〈國文抄錄〉

관상동맥질환의 진행과정 모형화
-질병예방에 응용-

박현애*

본 연구의 목적은 관상동맥질환의 위험요인과 가능한 질병상태를 로지스틱 회귀분석을 이용하여 이산, 다중상태의 확률과정으로 모형화한 후 여러가지 예방책을 시뮬레이션을 통해 예측하는 데 있다. 관상동맥질환의 진행과정을 모형화하는데 사용된 질병상태는 건강한 경우, 관상동맥질환을 앓는 경우, 관상동맥질환으로 사망하는 경우, 관상동맥질환이 아닌 다른 원인으로 사망하는 경우 등이다. 질병상태간의 추이확률(Transition Probabilities)은 성별, 나이, 관상동맥질환의 이환여부, 콜레스테롤, 혈압등 위험요인에 따라 달라진다. 이들 중 관상동맥질환의 이환여부는 모형설정시 두개의 분지를 도입함으로 통제하였으며, 성별, 나이는 40세에서 59세 사이의 남자 자료를 사용하여 통제하였다

본 연구를 위해 핀란드 North Karelia 프로젝트에서 얻은 자료를 사용하였으며 이 자료의 특징은 코호트로서 위험요인과 6년간의 관상동맥질환의 질병상태변화에 따른 기록을 포함하고 있는 점이다. 로지스틱 회귀분석의 계수추정은 최대우도추정법(Maximun Likelihood Method)을 적용하였으며, 적합도 검증(Goodness of Fit Test)은 위험도의 10분위 방법(Deciles of Risk Approach)을 사용하였다. 모형의 수행평가(Evaluation of Model Performance)는 회귀분석의 계수추정시 사용한 자료의 재치환(Resubstitution)방법을 사용하였다. 시뮬레이션을 통해 관상동맥질환의 예방시 고위험집단(High Risk Group)과 전체인구를 대상으로 했을때 관상동맥질환의 이환율과 사망율의 감소를 추정하였다.

연구결과 로지스틱 회귀분석을 통해 관상동맥질환의 이환율과 사망율을 예측하는데 콜레스테롤과 이완기 혈압이 유의한 변수였으며 콜레스테롤이 이완기 혈압보다는 높은 설명력을 보였다. 적합도 검증을 통해 콜레스테롤과 이완기 혈압으로 이루어진 로지스틱 모형을 사용한 시뮬레시션에서 얻은 관상동맥질환이 이환자와

[*] 韓國人口保健研究院 責任研究員.

사망자수는 실제로 관찰된 결과와 통계학적으로 일치하였다. 또한 관상동맥질환의 유병율과 사망율을 낮추려면 전체인구를 대상으로 하는 것이 고위험집단을 중심으로 예방대책을 수립하는것 보다 더 효율적인 것으로 나타났다.