예방의학회지 제39권 제5호(2006년 9월)
J Prev Med Public Health 2006;39(5):438-446

대장암 발생 고위험군의 예측모형 개발과 활용

이애경[1], 이상이[1,2], 박일수[1], 김수영[3], 윤태호[4], 정백근[5]

국민건강보험공단 건강보험연구센터[1], 제주대학교 의과대학 의료관리학교실[2], 제주대학교 의과대학 예방의학교실[3],
부산대학교 의과대학 예방의학 및 산업의학교실[4], 경상대학교 의과대학 예방의학교실[5]

Developing the Predictive Model for the Group at High Risk for Colon Cancer

Ae Kyoung Lee[1], Sang-Yi Lee[1,2], Il Soo Park[1], Su Young Kim[3], Tae-Ho Yoon[4], Baek-Geun Jeong[5]

Health Insurance Research Center, National Health Insurance Corporation[1]; Department of Preventive Medicine, College of Medicine, Cheju National University[2]; Department of Health Policy and Management, College of Medicine, Cheju National University[3]; Department of Occupational & Preventive Medicine, College of Medicine, Busan National University[4]; Department of Preventive Medicine, College of Medicine, Gyeongsang National University[5]

Objectives : We developed the predictive model for the incidence of colon cancer by utilizing the health screening data of the National Health Insurance in Korea. We also explored the characteristics of the high risk group for colon cancer.

Methods : The predictive model was used to determine those people who have a high risk for colon cancer within 2 years of their NHI health screening, and we excluded the people who had already been treated for cancer or who were cancer patient. The study population is the insured of the NHI, aged 40 or over and they had undergone health screening from the year 2000 to 2004, according to NHI health screening formula. We performed logistic regression analysis and used SAS Enterprise Miner 4.1.

Results : This study shows that there exists a higher rate of colon cancer in males than females. Also, for the population in their 60s, the incidence rate of colon cancer is much higher by 5.36 times than that for those people in their 40s. Amongst the behavioral factors, heavy drinking is the most important determinant of the colon cancer incidence (7.39 times in males and 21.51 times in females).

Conclusions : Our study confirms that the major influencing factors for the incidence of colon cancer are drinking, lack of exercise, a medical history of colon polypus and a family history of colon cancer. As a result, we can choose the group that is at a high risk for colon cancer and provide customized medical information and selective management services according to their characteristics.

J Prev Med Public Health 2006;39(5):438-446

Key words : Predictive model, Colon cancer, Incidence, Health screening

서 론

암은 우리나라의 사망원인의 1위를 차지하는 질병으로 [1], 2003년 전체 인구 10만 명당 131.1명으로 나타났다. 이는 10년 전(1993년 : 110.6명)보다 22.1% 증가한 수준이다 [2]. 우리나라도 매년 10만 명의 암 환자가 새롭게 발생하고 있으며, 약 6만 명이 암으로 사망하는 것으로 보고되고 있다 [3]. 이처럼 지속적으로 증가하는 암 발생과 사망은 국가적으로 의료비 지출, 인적 자원 손실 등의 사회 · 경제적 손실을 유발하고 있다. 실제로 1999년부터 2003년까지의 암 종류별 진료비 변화추이를 살펴보면, 전체적으로 진료비 규모가 증가하고 있으며, 특히 위암과 대장암에서 증가세가 두드러지고 있다. 2003년의 경우, 위암의 진료비는 2,005억으로 전체 총진료비의 16.32%를 차지하였고, 다음이 대장암(1,597억원 : 13.0%), 폐암 (1,535억원 : 12.5%), 간암 (1,392억원 : 11.3%), 유방암 (1,034억원 : 8.4%)의 순으로 나타났다.

이에 따라 암 발생과 사망을 낮추기 위한 국가적 노력이 보건의료분야의 중요한 정책과제로 대두되고 있다. 우리나라에서는 보건복지부를 중심으로 1996년 암정복 10개년 계획을 수립한 이후, 1999년부터 의료수급자를 대상으로 무료 암 검진사업을 실시(위암, 유방암, 자궁경부암)하였으며, 2004년에서는 기존에 실시하고 있는 위암, 유방암, 자궁경부암, 간암, 대장암을 새롭게 추가하여 5대 암 조기검진사업 체계를 구축하였다. 2005년에는 국민건강보험 가입자 중에서 보험료 하위 50%에 속하는 인구를 대상으로 무료 암 검진을 확대 · 실시하고 있다. 현재, 암 검진사업이 정부의 적극적인 암 관리 정책에 따라 전 국민에게 빠르게 확산되고 있지만, 건강검진의 계층 간 형평성과 사후관리는 아직도 미흡한 실정이다. 특히, 획일적인 대상자 선정 및 검사주기의 적용, 저소득계층이나 차상위계층의 경우 검진 후 치료 보장의 미흡으로 건강검진제도에 대한 필요성을 절감하지 못하는 점 등으로 인한 낮은

접수 : 2006년 3월 28일. 채택 : 2006년 7월 24일
책임저자 : 이상이 (제주도 제주시 아라1동, 전화 : 064-754-3858, 팩스 : 064-725-8593, E-mail : health21@cheju.ac.kr)

건강검진 수검 수준과 같은 문제점들이 노출되고 있다. 2004년 국민건강보험 가입자의 특정 암 수검률은 전체 암 수검 대상 인원(9,216,063명)의 14.67%(1,352,101명)에 그쳤다.

세계보건기구 (WHO)의 보고에 의하면, 의학적인 관점에서 암 발생인구 중 1/3은 식이습관의 변화, 금연, 간염백신 및 운동 등으로 예방이 가능하고, 또한 1/3은 조기 진단만 되면 완치가 가능하기 때문에 잘못된 식이습관을 변화시키고, 금연, 예방접종, 운동 등의 생활양식의 변화 및 조기 진단을 위한 정기적인 건강검진이 필요하다고 권고하고 있다. 이러한 예방관리를 위해서 이미 선진국에서는 일정하게 진전을 보고 있는데, 미국의 보험자는 질환관리 (Disease Management) 사업을 수행함에 있어 많은 비용이 소모됨에 따라 비용절감 차원에서 온라인 질환관리 (Online Disease Management) 기법을 도입하고 있다. 예컨대, Columbia United Providers에서는 데이터마이닝 시스템을 이용하여 천식환자, 당뇨환자에 대한 고위험군을 선별관리하고, Mckesson's Care Enhance에서는 개인별 위험점수를 예측하고, 이를 기준으로 특성별 고위험군을 세분화하여 간호사들이 사례관리를 중점적으로 수행하고 있다 [4]. 이 외에도 수면장애자, 당뇨환자 등에 대해 고위험군 예측모형을 개발하고, 이에 기반한 CRM (customer relationship management)을 도입한 결과, 적절한 시기에 치료와 관리가 가능하여 건강증진에 크게 기여한 것으로 나타났다 [5].

한편, 미국의 의료정보관리기구에서는 의료이용량 예측을 위해 예측모형인 Impact Pro Model을 개발하고, 이를 이용하여 개인별로 향후 1년간의 급여비 및 의료이용량을 예측하여 추정급여비가 높은 자에 대해 적절한 의료이용 방안, 교육 및 사례관리 사업을 실시하고 있다. 즉, 보험가입자의 기본정보, 급여데이터와 검진데이터 등을 이용하여 연간 급여비, 총 재원일수, 외래 이용횟수 등의 의료자원 사용량을 예측하여 CRM을 실시하고 있다.

또한, 고혈압, 당뇨 그리고 심장질환 등에 대해서도 질환별 위험도를 예측하기 위해 데이터마이닝 기법을 이용한 예측모형을 개발하여 활용하고 있다 [6].

뿐만 아니라 미국 하버드 대학교에서는 10년의 코호트 자료를 이용한 추적연구를 토대로 암 위험지표 (cancer risk index)를 개발, 이용하여 암 발생 고위험군을 선정하고, 수준별 세분화를 통한 예방관리사업을 수행하고 있다. 개발된 암 위험지표는 약 80% 정도의 암 발생자를 예측해 보이고 있으며, 또한 암 발생 가능성을 사전에 인지하고 생활습관의 올바른 변화를 유도할 경우 평균적으로 암 발생 위험률이 감소하는 것으로 나타나, 이를 활용한 맞춤형 건강증진사업의 적용 가능성을 시사하였다 [7].

이러한 맥락에서 본 연구는 개인별 특성요인을 반영한 차별화된 암 관리사업을 도출하기 위한 목적으로, 최근 들어 생활습관이 서구화됨에 따라 급속히 증가하고 있는 대장암의 발생 예측모형을 개발하고자 하였다. 또한, 본 연구에서는 개발된 대장암 발생 예측모형을 이용하여 고위험군의 특성을 분석하고, 이를 바탕으로 효율적 관리 대상자 선정 방법을 찾고, 향후 암 발생 예측모형의 활용 방안을 제안하고자 한다.

연구 자료 및 방법

1. 연구 자료

본 연구에서의 고위험군이란 과거에 특정질환에 대해 진단 및 치료를 받지 않은 사람이 신체적인 요인, 주위환경, 생활습관, 가족력 등의 다양한 요인에 의해 특정 질환으로 진단 및 치료를 받게 될 확률이 높은 군을 의미한다. 고위험군 예측모형 (predictive modeling)은 보험자(국민건강보험공단)가 보유한 가입자의 각종 정보를 이용하여 개별 가입자의 질병 위험성을 평가하고, 위험 예방을 위한 가입자 관리 프로세스 개발의 기초 자료로 활용하는 제반의 과정을 의미한다.

본 연구에서 개발하고자 하는 대장암 발생 고위험군 예측모형은 국민건강보험 가입자 중 암 검진일 이전에 암으로 진료 받은 경험이 있거나 검진결과 암으로 판정된 자를 제외하고, 검진 받은 일부터 2년 이내에 암이 발생될 가능성이 높은 고위험군을 찾아내는 모형이다.

암 발생 고위험군 예측모형 개발을 위한 자료는 국민건강보험공단의 원천시스템 (Operational Data Store) 및 데이터 웨어하우스 (Data Warehouse)에서 2000년부터 2004년까지의 특정 암 검진 및 문진자료, 1 · 2차 건강검진 및 문진자료, 현물급여자료의 각 연도별 개인급여정보(2005년 5월 현재 지급기준), 상병정보 및 수검자의 자격정보(수검월말 자격)를 이용하였다. 단, 연구대상은 우리나라의 건강검진 대상자 선정기준에 근거하여 2000년부터 2004년 기간에 대장암 검진을 받은 40세 이상 국민건강보험 가입자로 제한하였다.

1) 목표변수(Target variables)

대장암 발생 고위험군의 예측모형을 개발하기 위한 대상자는 국민건강보험 가입자 중에서 검진 받은 시점으로부터 이전 1년 동안 상병 C18~C20으로 치료 받거나, 국립암센터에 암 환자로 등록된 자 그리고 당해 대장암 검진결과가 대장암 의심환자로 종합 판정된 자들은 모두 제외하였다. 따라서 대장암 발생 판정은 대장암 검진을 받은 년도부터 2년 이내에 동일 상병으로 치료 받은 경우로 하였으며, 이에 국립암센터의 암 등록자료, 통계청의 사망자료, 대장암 검진의 종합판정 결과, 그리고 입원건수와 투약일수 정보를 연계하여 대상자를 조정하였다. 즉, 검진 대상자 중에서 국립암센터에 암으로 등록된 자 및 특정 암 검사의 종합판정 결과가 암 질환 의심자인 경우는 우선적으로 암 발생자로 하였고, 또한 암 건강검진에서 정상이 아닌 기타 질환자로 판정된 대상자 중에서 대장암 상병으로 입원 경험이 1회 이상인 경우도 대장암 질환자로 분류하여 대장암 발생자로 정하였다. 단, 1회 입원 경험만 있는 경우에는 외래방문일수 및 투약일수가 2회 이하인 경우는 모두 제외하였다.

따라서 본 연구에서의 대장암 발생률은 국민건강보험공단에서 대장암 건강검진을 받은 자 중에서 앞서 설명한 기준에 의해 대장암이 발견된 자의 비율을 의미한다. 이러한 점에서 일반적으로 인구 10만명당 보고 되는 조발생률과 본 연구에서 정의한 암 발생률 간에는 모집단의 정의

Table 1. Target variable

(Unit : Person, %)

		Male		Female		Total	
		Y=1	Y=0	Y=1	Y=0	Y=1	Y=0
Training data (year)	- 49	25 (24.04)	21,736 (55.96)	16 (31.37)	14,123 (49.95)	41 (26.45)	35,859 (53.43)
	50 ~ 59	39 (37.50)	10,613 (27.32)	17 (33.33)	8,436 (29.84)	56 (36.13)	19,049 (28.38)
	60 -	40 (38.46)	6,496 (16.72)	18 (35.29)	5,714 (20.21)	58 (37.42)	12,210 (18.19)
	Total	104 (100.0)	38,845 (100.0)	51 (100.0)	28,273 (100.0)	155 (100.0)	67,118 (100.0)
Validation data (year)	- 49	10 (22.22)	9,346()	5 (21.74)	6,075 (50.13)	15 (22.06)	15,421 (53.61)
	50 - 59	16 (35.56)	4,444()	7 (30.43)	3,624 (29.91)	23 (33.82)	8,068 (28.05)
	60 -	19 (42.22)	2,859()	11 (47.83)	2,419 (19.96)	30 (44.12)	5,278 (18.35)
	Total	45 (100.0)	16,649()	23 (100.0)	12,118 (100.0)	68 (100.0)	28,767 (100.0)
Test data (year)	- 49	25 (11.96)	31,655()	36 (24.83)	38,464 (45.88)	61 (17.23)	70,119 (45.42)
	50 - 59	60 (28.71)	21,004()	42 (28.97)	26,225 (31.28)	102 (28.81)	47,229 (30.60)
	60 -	124 (59.33)	17,866()	67 (46.21)	19,151 (22.84)	191 (53.95)	37,017 (23.98)
	Total	209 (100.0)	70,525()	145 (100.0)	83,840 (100.0)	354 (100.0)	154,365 (100.0)

Table 2. Input variable

	Variable	Explanation
Socio-demographic characteristics	Sex	Male, Female
	Age	40~49, 50~59, Over 60
	Region	Metropolitan, Urban, Rural
The insured's characteristics	Category of the insured	The employee insured, *The self-employed insured*
	Monthly contribution	≥50,000, 50,00~75,000, 75,000~100,000, >100,000won
Health behavior	Drinking (Frequency)	Never, 2~3(Monthly), 1~2(Weekly),3~4(Weekly), Daily
	Drinking (Amount, Soju)	Half bottle, 1 bottle, 1+1/2 bottle, ≥2 bottle
	Smoking status	Never, Ever smoking Current smoking
	Cigarettes per day	1~9, 10~19, 20~39, ≥40
	Excercise	Never, 1~2(Weekly),3~4(Weekly), 5~6(Weekly), Daily
	Favorite food	*Vegetable, Vegetable&Meat, Meat*
	BMI (BMI : kg/)	Obesity(≥25)
		Overweight(23~25)
		Normal(<23)
Family history	Cancer	Yes or No
Disease history	Colon polyp	Yes or No
	Cholesterol(mg/dl)	Suspicious(≥261), Normal B(231~260), Normal A(≤230)

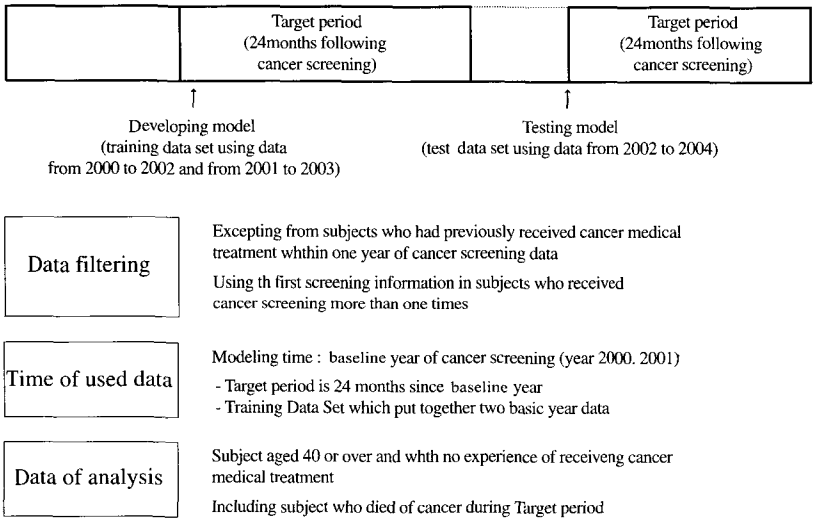

Figure 1. Developing process of predictive model for colon cancer incidence.

가 다르므로 차이가 있음을 밝혀 둔다.

대장암 발생 고위험군 예측모형의 목표변수 는 대장암이 발생된 여부에 따라 각각 y=1 또는 y=0으로 하였다. 그 결과 분석 단계별 사용된 데이터의 성 · 연령별 대장암 발생분포는 Table 1과 같다. 먼저 대장암 발생자는 상대적으로 연령이 높은 군에 밀집되어 있었고, 이는 남녀에서 동일한 분포 형태를 보였다. Training Data에서 성별 대장암 발생분포는 남자가 0.26% (전체 38,845명, 발생자=104명)이고, 여자는 0.18% (전체 28,273명, 발생자 51명)을 보였고, 이러한 성별 분포는 Validation Data에서 남자가 0.27%, 여자는 0.19% 그리고 Test Data에서는 남자 0.29%, 여자는 0.17%로 분포하였다 (Table 1).

2) 입력변수(Input Variable)

대장암 발생에 영향을 주는 요인은 국제암연구소와 미국 국립암협회지에서 발표한 암 발생의 위험요인과 국립암센터에서 권고하고 있는 위험요인들을 고려하였다. 입력변수는 국민건강보험공단의 건강검진자료를 기초로 건강행위(음주, 흡연, 운동, 식생활, 비만), 개인의 과거병력, 가족병력 그리고 건강검진 결과 등의 위험요인이 모두 모형에 반영될 수 있도록 하였다 [8,9]. 또한 수검자의 인구사회학적 특성인 성, 연령, 거주지역 특성이 질환 발생 및 유병률에 영향을 주고 [8,10], 국민건강보험 가입자의 자격(직장가입자 또는 지역가입자, 이하 직역)과 보험료는 건강검진 수검에 영향을 준다는 연구결과에 따라 [11,12], 본 연구는 이러한 요인들도 모형의 입력변수로 포함하였다 (Table 2).

2. 연구 방법

국민건강보험 가입자 중 암 건강검진을 받은 자를 중심으로, SAS Enterprise Miner 4.1을 이용하여 암 발생 예측모형을 개발하였고, 로지스틱 회귀분석을 적용하였다 [13,14]. 분석 데이터는 크게 분석용 (training data), 평가용 (validation data), 검정용 (test data)으로 구분하였고, 분석용과 평가용 데이터는 2000년, 2001년 암 건강검진 대상자를 기준으로 각각 7 : 3의 비율로 분할하여 생성하였다. 검정용 자료는 2002년 암 검진 대상자를 기준으로 분석용 데이터와 동일한 사상으로 구축하였으며, 모형개발은 최신정보기술로 각광받고 있는 데이터 마이닝 방법론을 적용하여 개발하였다 (Figure 1).

Phase I(사전단계)에서는 대장암 발생 고위험군 예측모형 개발의 유형을 정하고

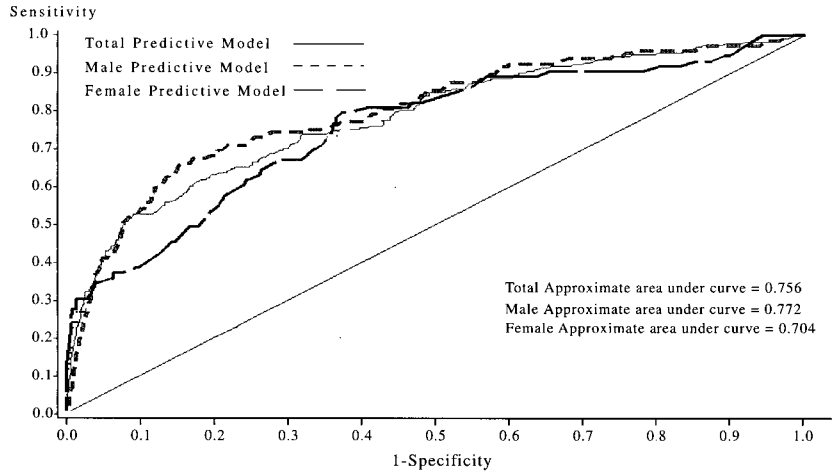

Figure 2. Logistic regression ROC curve of predictive model for colon cancer incidence. (Training data)

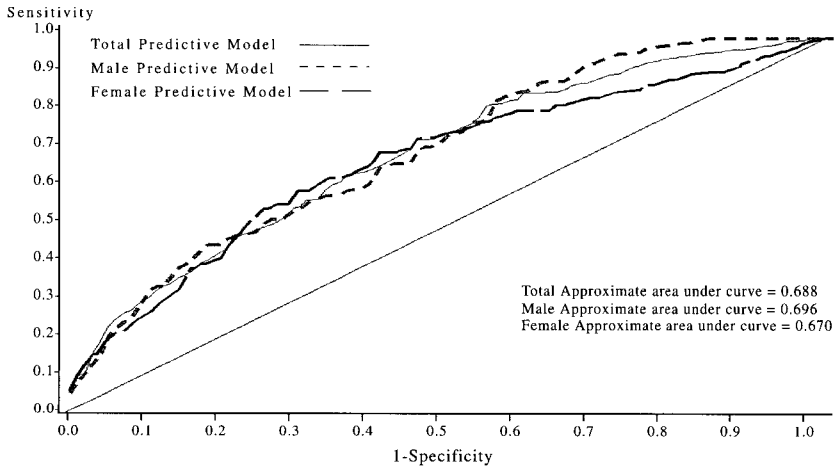

Figure 3. Logistic regression ROC curve of predictive model for colon cancer incidence. (Testing data)

분석대상, 분석기간, 분석주제의 정의, 평가 그리고 예측기간을 정의하였다. Phase 2(데이터단계)에서는 예측모형의 정확성을 높이기 위해 데이터를 탐색 및 데이터의 클린징 작업을 수행하고, 분석대상을 기준으로 분석주제와 방향에 맞게 분석항목(Target 변수 포함)을 분석대상별로 선정하여 유용한 정보를 색출하여 분석용 데이터마트를 구축하였다. Phase 3(분석단계)에서는 최종 구축된 분석용 데이터마트를 활용하여 로지스틱 회귀분석을 통한 모델링 작업을 수행하였는데, 개발된 모형의 검증을 위해 본 연구의 모형화 작업에 있어서는 전체 분석 자료를 분석용(70%)과 평가용(30%)으로 분할하여 분석용(training data)을 통해 모형을 생성하고, 만들어진 모형을 검정용 (test data)에 적용시켜 모형 생성 시 발생할 수 있는 과적합 문제 (overfitting)를 해결하였다. 또한, 다양한 모형생성 알고리즘을 통해 만들어진 모형들 중에서 가장 좋은 모형을 평가 · 선정하기 위해 Lift 차트와 Receiver Operating Characteristic(ROC) curve 분석과 정확도 (accuracy)를 계산하였다. 여기서 Lift는 모형의 효과에 근거한 예측 향상력을 의미하는 것으로 모형을 적용하였을 경우와 적용하지 않았을 경우의 대비 효과를 나타낸다.

$$\text{Lift} = \frac{\text{Proportion of Target(=1) in Bin Group}}{\text{Proportion of Target(=1) in Total Group}}$$

한편, ROC curve는 일반적으로 X축(1-특이도)의 값과 Y축의 민감도 값으로 이루어진 곡선을 의미하는 것으로, ROC curve 아래 영역인 AUC(Area Under the ROC Curve) '0.5'는 모델의 예측능력이 없음을 나타내고, '0.8' 이상은 비교적 높은, '1'은 완전한 예측능력을 나타낸다. 본 연구에서는 임계점 (best cut-off point)을 (1 - 민감도)2와 (1 - 특이도)2의 합이 최소화되고 ROC curve에서 (0, 1)에 가장 가까운 지점으로 정하여 정확도 (accuracy)를 계산하여 반영하였다.

마지막으로 Phase 4(적용단계)에서는 모형 생성을 위한 데이터마트와 똑같은 형태로 만들어진 모형 적용을 위한 새로운 데이터마트에 생성된 모형을 적용시켜 도출한 사전예방 건강관리모형에서는 대장암에 대해 향후 2년간 고위험군 발생 확률을 예측하였다. 또한, 예측된 값들의 정확도를 알아보기 위해 실제 데이터(2004년)와 비교 · 검증(external validation)하였다.

연구결과

I. 대장암 발생 예측모형의 개발

1) 모형 평가

대장암 발생 로지스틱 회귀예측모형에 대한 평가는 분석용과 검정용 데이터에서 로지스틱 회귀모형의 ROC 곡선과 Lift에 근거하였다. 그 결과 Figure 2,3에서 보는 바와 같이, ROC 곡선의 밑면적을 나타내는 C-통계량이 분석용과 검정용 데이터 모두에서 0.5보다 큰 값을 보이고 있어, 개발된 대장암 발생 예측모형에 대한 성능과 안정성이 유효함을 보여 주었다. 검정용 데이터인 경우, 개발된 로지스틱 예측모형의 C-통계량은 전체 모형에서 0.688, 남자와 여자 예측모형에서는 각각 0.696과 0.670의 값을 나타내었다.

또한, 누적 리프트 도표 (cumulative lift chart)는 추정된 사후확률 (posterior probability)의 분위수(percentile)에 따른 반응률(%Response)을 도표화 한 것으로, 상위 분위수에 대응되는 리프트가 더 클수록 모형의 성능이 더 우수함을 나타낸다[13-15]. 대장암 예측모형의 사후확률에 대한 상위 분위수에 대응되는 리프트(Lift)가 분석용과 검정용 데이터 모두에서 무작위로 관리하는 경우보다 더 큰 값을 보이고 있는 것으로 나타났다. 특히, 예측된 모형의 상위 10% 수준에서 검정용 리프트 값을 보면, 전체 예측모형의 경우는 2.98, 남자

Table 3. A hitting rate of predictive model for colon cancer incidence according to score groups

Sample	Grade	Total				Male				Female			
		Frequency	CR (%)	RE (%)	Lift	Frequency	CR (%)	RE (%)	Lift	Frequency	CR (%)	RE (%)	Lift
Training data	10%	6579	56.12	1.33	5.73	3894	55.76	1.48	5.57	2803	39.21	0.71	3.96
	20%	13466	65.80	0.75	3.28	7905	70.19	0.92	3.45	5518	50.98	0.47	2.61
	30%	20468	84.83	0.56	2.45	11665	74.03	0.66	2.47	8494	62.74	0.37	2.09
	50%	33642	86.45	0.39	0.86	19602	86.53	0.45	1.71	14047	80.39	0.29	1.62
	100%	67274	100.0	0.23	1.00	38949	100.0	0.26	1.00	28324	100.0	0.18	1.00
Validation data	10%	2,883	45.59	1.07	4.56	1,670	53.33	1.44	5.33	1,172	47.83	0.93	4.96
	20%	5,746	55.88	0.66	2.80	3,377	62.22	0.83	3.08	2,342	60.87	0.60	3.15
	30%	8,655	66.18	0.52	2.04	5,009	75.57	0.67	2.52	3,640	62.17	0.41	2.17
	50%	14,426	80.88	0.37	1.58	8,353	84.44	0.45	1.68	6,045	82.61	0.32	1.67
	100%	28,834	100.0	0.23	1.00	16,694	100.0	0.27	1.00	12,141	100.0	0.18	1.00
Test data	10%	15,109	29.10	0.68	2.98	7,097	28.71	0.85	2.86	8,035	24.83	0.45	2.60
	20%	30,929	42.37	0.48	2.12	14,362	44.98	0.65	2.21	17,235	41.38	0.35	2.02
	30%	46,511	54.80	0.42	1.82	21,011	53.59	0.53	1.80	25,654	59.31	0.34	1.94
	50%	77,150	74.29	0.34	1.49	35,373	74.16	0.44	1.48	41,655	74.48	0.26	1.50
	100%	154,719	100.0	0.23	1.00	70,734	100.0	0.30	1.00	83,985	100.0	0.17	1.00

1) RE(%)=(The number of colon cancer case per group / Total number of health screening per group) × 100
2) CR(%)=(The number of colon cancer case within group / The number of colon cancer in total) × 100
3) Base RE(%)=The number of colon cancer case / Total number of health screening
4) Lift= Incidence rate of colon cancer within group/ Incidence rate of colon cancer in total
5) RE(%) : Response(%), CR(%) : Captured Response(%)

Table 4. Predictive model for colon cancer incidence (Logistic regression)

Variables		Model: total		Model: male		Model: female	
		Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio
	Constant	-5.3837†		-5.5471†		-4.5113†	
Socio-demographic characteristics							
Sex	Male						
	Female	-0.2864†	0.56				
Age	40-49						
	50-59	0.1790	3.03	0.2229	3.32	0.0819	2.02
	Over 60	0.7503†	5.36	0.7545†	5.65	0.5440†	3.22
Region	Rural						
	Urban	0.4336†	3.30	0.3987	3.40	0.2426	2.47
	Metropolitan	0.3277	2.97	0.4291*	3.51	0.4194	2.94
The insured`s characteristics							
Category of the insured	The employee insured						
	The self-employed insured	-0.2979†	0.55	-0.1985*	0.67	-0.2118	0.65
Monthly contribution	100,000 <						
	75,000 ~100,000 won	-0.0355	0.83	-0.1468	0.92	0.0860	0.99
	50,000 ~ 75,000 won	-0.0767	0.80	0.1326	1.21	-0.2429	0.71
	≤ 50,000 won	-0.0315	0.83	0.0781	1.15	0.0699	0.98
Health behavior							
Drinking	Others						
	Over 1 bottle of soju (≥3per week)	1.3186†	13.97	1.0003†	7.39	1.5345†	21.51
Exercise	Yes(over 1~2 per week)						
	Never	0.1317	1.30	0.2052†	1.51		
Family history							
Cancer	Yes						
	No	0.0313	1.06	0.2446†	1.63		
Disease history							
Colon polyp	Yes						
	No	0.5428†	2.96	0.5326*	2.90	0.7632†	4.60

*p<0.1, †p<0.05, ‡p<0.01.

와 여자의 경우는 각각 2.86, 2.60을 보여, 대장암 발생 예측모형을 활용하여 고위험군을 관리할 경우, 전체 대장암 수검자 대비 각각 2.98배, 2.86배 그리고 2.60배의 효율을 기대할 수 있는 것으로 나타났다(Table 3).

2) 예측모형

개인별 인구사회학적 특성, 건강행위, 개인의 과거병력, 가족병력 그리고 검진결과 등의 요인들이 대장암 발생에 미치는 영향력 정도를 보고자, 로지스틱 회귀분석을 실시하였다.

대장암 발생 로지스틱 회귀모형의 추정된 결과는 Table 4와 같다. 먼저, 대장암 발생에 영향을 미치는 인구사회학적 요인으로 성별의 경우는 여자가 남자보다 대장암 발생 가능성이 더 낮은 것으로 나타났다. 연령에서는 40대보다 60대 이상의 경우가 대장암 발생 위험도가 5.36배 더 높은 것으로 나타났고, 그 중 남자는 5.65배, 여자는 3.22배 정도 대장암 발생 위험도가 높은 것으로 나타났다. 거주 지역별로는 농어촌 거주자들보다는 도시 거주자들이 대장암 발생 가능성이 높은 것으로 나타났으며, 특히 남자의 경우는 대도시 거주자가 농어촌 거주자보다 대장암 발생 위험도가 3.51배 더 높은 것으로 나타났다. 단, 검진자의 사회경제적 수준을 나타내는 보험료의 수준은 보험료를 적게 납부할수록 대장암 발생 가능성이 낮은 관계를 보이고 있으나, 통계적으로 유의하지는 않았다. 한편, 개인의 건강행위에 따른 위험요인으로, 음주습관에서 소주 1병 이상을 주 3회 이상 마시는 경우는 그렇지 않은 수검자들보다 대장암 발생 위험도가 13.97배 더 높았고, 특히 남자는 7.39배, 여자는 21.51배나 더 높은 것으로 나타났다. 그리고 운동습관에서 운동을 전혀 하지 않는 경우에는 운동을 주기적으로 하는 경우보다 남자의 경우는 대장암 발생 위험도가 1.51배 더 높은 것으로 나타났다. 이 외에 대장암 발생의 위험요인으로 알려진 개인의 과거병력 중 대장용종이 있는 경우 또는 있었던 경우는 그러한 경우가 없는 경우보다 대장암 발생 위험도가 2.96배 높았으며, 그 중 남자는 2.9배, 여자는 4.6배 더 높은 것으로 나타났다. 더욱이 가족 중에 암 병력이 있을 경우는 그렇지 않은 경우보다 대장암 발생 위험도가 1.63배 더 높은 것으로 나타났다.

2. 대장암 발생 고위험군의 특성

효과적인 암 관리사업을 추진하기 위해서는 개인별 특성에 따른 암 관리사업 방법의 마련, 건강정보의 생성 및 제공 등이

Table 5. Comparison of characteristics of high risk group with low risk group according to predictive model for colon cancer incidence

Factor		High risk group (%) (Upper 10%)	Low risk group (%) (Lower 10%)	Sub total (%)
Demographic characteristics				
Age	Less than 50	3,339 (22.10)	14,690 (95.22)	70,180 (45.36)
	50-59	7,372 (48.79)	620 (4.02)	47,331 (30.59)
	Over 60	4,398 (29.11)	118 (0.76)	37,208 (24.05)
Gender	Male	13,174 (87.19)	10,009 (64.88)	70,734 (45.72)
	Female	1,935 (12.81)	5,419 (35.12)	83,985 (54.28)
Region	Metropolitan	7,443 (49.26)	7,484 (48.52)	77,785 (50.28)
	Urban	7,237 (47.90)	4,737 (30.71)	67,299 (43.50)
	Rural	429 (2.84)	3,204 (20.77)	9,630 (6.22)
The insured's characteristics				
Category of the insured	The self-employed insured	5,339 (35.34)	13,123 (85.06)	67,829 (43.84)
	The employee insured	9,770 (64.66)	2,305 (14.94)	86,890 (56.16)
Monthly contribution	≤ 50,000 won	4,118 (27.32)	5,828 (37.83)	49,344 (31.96)
	50,000 ~ 75,000 won	2,863 (18.99)	5,489 (35.63)	33,811 (21.90)
	75,000 ~100,000 won	2,270 (15.06)	2,744 (17.81)	24,175 (15.66)
	≥100,000 won	5,822 (38.63)	1,345 (8.73)	47,061 (30.48)
Health risk factor				
Life-style				
Food	Meat	1,270 (8.41)	725 (4.70)	6,671 (4.31)
	Vegetable, Vegetable&Meat	13,839 (91.59)	14,703 (95.30)	148,048 (95.69)
Duration of smoking	No response	6,034 (39.94)	10,089 (65.48)	112,402 (72.71)
	< 5 year	499 (3.30)	574 (3.73)	4,202 (2.72)
	5 ~ 9 year	554 (3.67)	672 (4.36)	3,957 (2.56)
	10 ~ 19 year	1,831 (12.12)	1,845 (11.97)	10,840 (7.01)
	20 ~ 29 year	2,807 (18.58)	1,878 (12.19)	13,207 (8.54)
	>30	3,384 (22.40)	350 (2.27)	9,978 (6.45)
Smoking status	No response	105 (0.69)	266 (1.73)	3,139 (2.03)
	Never	5,680 (37.59)	9,771 (63.42)	108,932 (70.47)
	Ever smoking	3,190 (21.11)	2,172 (14.10)	16,764 (10.84)
	Current	6,134 (40.60)	3,199 (20.76)	25,751 (16.66)
Cigarettes per day	No response	9,004 (59.59)	12,217 (79.29)	128,962 (83.42)
	1 ~ 9	1,159 (7.67)	634 (4.11)	6,521 (4.22)
	10 ~ 19	3,328 (22.03)	1,761 (11.43)	13,310 (8.61)
	20 ~ 39	1,540 (10.19)	742 (4.82)	5,471 (3.54)
	40 <	78 (0.52)	54 (0.35)	322 (0.21)
Drinking (Frequency)	No response	17 (0.11)	162 (1.05)	1,884 (1.22)
	Never	492 (3.26)	9,044 (58.70)	90,321 (58.43)
	2 ~ 3(Monthly)	73 (0.48)	4,156 (26.97)	23,852 (15.43)
	1 ~ 2(Weekly)	7,833 (51.84)	1,328 (8.62)	22,545 (14.58)
	3 ~ 4(Weekly)	4,169 (27.59)	415 (2.69)	10,137 (6.56)
	Daily	2,525 (16.71)	303 (1.97)	5,847 (3.78)
Drinking (Amount)	No response	480 (3.18)	8,892 (57.71)	89,295 (57.76)
	Half bottle	135 (0.89)	4,392 (28.50)	30,106 (19.48)
	1 bottle	10,102 (66.86)	1,559 (10.12)	24,405 (15.79)
	1+ 1/2 bottle	3,023 (20.01)	371 (2.41)	7,449 (4.82)
	2 <	1,369 (9.06)	194 (1.26)	3,331 (2.15)
Exercise	No	8,023 (53.10)	5,905 (38.27)	84,863 (54.85)
	Yes	7,086 (46.90)	9,523 (61.73)	69,856 (45.15)

필수적이다. 여기서는 개인별 특성을 반영한 차별화된 관리 프로세스를 유도하기 위해, 개발된 예측모형을 검정용 데이터에 적용하여 2002년 암 검진을 받은 자가 2년 뒤인 2004년 이내에 해당 암으로 발견될 확률이 높은 상위 10%와 반대로 발견확률이 낮은 하위 10%를 각각 고위험군과 저위험군이라 정의하고, 집단의 특성을 비교 · 분석하였다 (Table 5). 단, 2002년 수검자 중에서 과거에 검진 또는 진료를 통해 대장암에 노출되지 않은 수검자만을 고려하였다.

먼저, 대장암 발생과 관련된 고위험군 대비 저위험군의 특성을 인구사회학적 측면에서 살펴보면, 50대 이상 수검자의 경우가 고위험군(77.90%)에 보다 밀집되어 분포하였고, 반대로 저위험군에서는 40대 이하의 수검자들이 95.22% 이상 분포하는 것으로 나타났다.

한편, 개인의 건강행위 위험요인 중 식습관의 경우, 육식을 선호하는 자의 분포가 저위험군(4.70%)보다 고위험군(8.41%)에 미미하게나마 많았으며, 또한 흡연기간이 20년 이상인 검진자의 분포는 고위험군이 40.98%, 저위험군이 14.46%로 나타나 상대적으로 고위험군에 흡연에 대한 건강행위 개선이 요구되는 검진자들이 많이 분포하고 있는 것으로 나타났다. 음주습관의 경우, 1주일에 3회 이상 음주하는 자가 고위험군에 44.30%, 저위험군에서는 4.66% 분포하고 있었다. 특히, 음주량으로 소주 1병 이상을 마실 경우는 고위험군에서 차지하는 비율이 95.93%로 나타나, 대장암 발생 가능성이 높은 군에서는 전반적으로 음주습관에 대한 개선 및 사전 예방관리를 집중적으로 수행할 필요성이 있음을 보여주었다.

가족병력에서는 간장질환(5.1-4%), 뇌졸중(8.66%), 심장병(3.9-8%) 및 암(19.72%)에 대한 가족력이 있는 사람이 저위험군의 경우보다 고위험군에서 상대적 비율이 높은 것으로 나타났으며, 개인의 과거병력과 검진결과에서는 위 · 십이지장궤양, 위수술, 대장용종, 간장질환, 고혈압, 뇌졸중, 심장병, 당뇨 역시 저위험군보다 고위험군에서 상대적으로 차지하고 있는 비율이 높았다. 특히, 대장용종이 있었을 경우, 고위험군(4.72%)의 비율이 저위험군(0.03%)보다 매우 큰 것으로 나타났다. 또한, 고위험군에서는 건강검진 결과 비만한 사람, 콜레스테롤 수치가 260 ㎎/㎗을 초과하는 자의 비율이 상대적으로 저위험군보다 높은 분포를 보이는 것으로 나타났다.

고 찰

일반적으로 암은 초기에 그 증상이 나타나지 않고 상당히 진행된 이후에야 발견되는 경우가 많음에 따라, 최근 암 예방에 대한 사회적 관심이 크게 증가하고 있다. 그러나 암 예방을 위해서 개인의 과거 건강상태 정보를 근거로 특정 질환이 발생할 가능성이 높은 고위험군 선정 시스템이나, 이를 이용한 고위험군의 특성별로 차별화된 사후관리체계는 아직 마련되어 있지 못한 상태이다.

이러한 점에서 본 연구는 국민건강보험공단이 보유하고 있는 방대한 데이터베이스를 이용하여, 사후의 치료 중심이 아닌 사전적 예방관리를 위한 암 발생 예측모형을 개발하였다는 점에서 의의가 있다 하겠다.

1. 대장암 발생 위험요인 및 고위험군의 특성

국민건강보험공단에서 특정 암 건강검진을 받은 대상자를 중심으로 개발된 대장암 발생 로지스틱 예측모형을 이용하여, 인구사회학적 특성, 건강행위 및 검진결과 등에 따른 상대적 발생위험도를 살펴보았다. 그 결과, 남자보다 여자가 상대적으로 대장암 발생위험도가 높은 것으로 나타났으며, 연령의 경우에는 나이가 많을수록 대장암 발생위험이 높았다. 특히, 60대의 경우는 40대보다 대장암 발생위험도가 5.36배 높은 것으로 나타났다. 대장암은 거주지역 특성에 따라서도 발생분포의 차이가 있을 수 있는데 [16], 본 연구에서도 농어촌지역보다는 도시지역에서 대장암 발생위험도가 높은 것으로 나타났다. 특히, 중소도시의 경우는 농어촌지역보다 대장암 발생위험도가 3.30배 더 높았다.

건강행위 요인에서 음주는 대장암 발생위험에 결정적인 영향을 주고 있었으며, 소주 1병 이상을 주 3회 이상 음주하는 자는 그렇지 않은 자보다 대장암 발생위험도가 남자는 7.39배, 여자는 21.51배 높은 것으로 나타났다. 운동습관은 여자보다는 남자의 경우에서 대장암 발생에 더 큰 영향을 미치고 있었고, 남자의 경우는 운동을 전혀 하지 않는 경우가 운동을 하는 경우보다 상대적으로 대장암 발생위험도가 1.51배 높은 것으로 나타났다. 개인의 과거병력 중에서 대장용종(폴립)이 있는 경우는 남자와 여자 모두에서 대장암 발생위험에 영향을 미치고 있었으며, 남자의 경우는 2.90배, 여자의 경우는 4.60배 대장암 발생위험도가 높았다.

특히, 대장암 발생 예측분포의 상위 10%에 해당되는 고위험군의 특성은 주로 연령이 50대 이상인 남자이고, 도시에 거주하는 직장 및 공교 가입자들이며, 대장암 검진 당시 월 부과보험료가 10만원을 초과하는 경우였다. 개인별 건강행위의 위험요인이라 할 수 있는 식습관의 경우는 육식을 선호하는 대상자 분포가 상대적으로 저위험군 보다는 고위험군에서 많았고, 흡연력이 20년 이상인 자, 음주빈도나 음주량이 많은 자, 운동도 전혀 하지 않는 자

Table 5. Comparison of characteristics of high risk group with low risk group according to predictive model for colon cancer incidence (continued)

Factor		High risk group (%) (Upper 10%)	Low risk group (%) (Lower 10%)	Sub total (%)
Health risk factor				
Family history				
Liver disease	Yes	776 (5.14)	668 (4.33)	6,801 (4.40)
	No	14,333 (94.86)	14,760 (95.67)	147,918 (95.60)
Hypertension	Yes	1,538 (10.18)	1,570 (10.18)	16,256 (10.51)
	No	13,571 (89.82)	13,858 (89.82)	138,463 (89.49)
Stroke	Yes	1,309 (8.66)	1,162 (7.53)	11,470 (7.41)
	No	13,800 (91.34)	14,266 (92.47)	143,249 (92.59)
Heart disease	Yes	602 (3.98)	566 (3.67)	5,918 (3.82)
	No	14,507 (96.02)	14,862 (96.33)	148,801 (96.18)
Diabetes mellitus	Yes	1,227 (8.12)	1,373 (8.90)	12,551 (8.11)
	No	13,882 (91.88)	14,055 (91.10)	142,168 (91.89)
Cancer	Yes	2,979 (19.72)	2,587 (16.77)	27,194 (17.58)
	No	12,130 (80.28)	12,841 (83.23)	127,525 (82.42)
Medical history				
Ulcer of stomach and duodenum	Yes	1,455 (9.63)	1,124 (7.29)	11,727 (7.58)
	No	13,654 (90.37)	14,304 (92.71)	142,992 (92.42)
Operation of stomach	Yes	404 (2.67)	68 (0.44)	1,656 (1.07)
	No	14,705 (97.33)	15,360 (99.56)	153,063 (98.93)
Colon cancer	Yes	713 (4.72)	5 (0.03)	1,811 (1.17)
	No	14,396 (95.28)	15,423 (99.97)	152,908 (98.83)
Disorder of liver and intestine	Yes	485 (3.21)	403 (2.61)	3,563 (2.30)
	No	14,624 (96.79)	15,025 (97.39)	151,156 (97.70)
Hypertension	Yes	1,748 (11.57)	596 (3.86)	13,991 (9.04)
	No	13,361 (88.43)	14,832 (96.14)	140,728 (90.96)
Stroke	Yes	88 (0.58)	55 (0.36)	1,001 (0.65)
	No	15,021 (99.42)	15,373 (99.64)	153,718 (99.35)
Heart disease	Yes	215 (1.42)	97 (0.63)	2,275 (1.47)
	No	14,894 (98.58)	15,331 (99.37)	152,444 (98.53)
Diabetes mellitus	Yes	705 (4.67)	319 (2.07)	5,638 (3.64)
	No	14,404 (95.33)	15,109 (97.93)	149,081 (96.36)
Cancer	Yes	66 (0.44)	74 (0.48)	988 (0.64)
	No	15,043 (99.56)	15,354 (99.52)	153,731 (99.36)
Health index				
BMI*	Obesity	6,029 (39.90)	5,268 (34.16)	54,603 (35.30)
	Normal, overweight	9,080 (60.10)	10,155 (65.84)	100,073 (64.70)
Cholesterol†	Suspicious	1,317 (8.72)	507 (3.29)	9,576 (6.19)
	Normal	13,786 (91.28)	14,916 (96.71)	145,077 (93.81)

* Body mass index : Normal (less than 23), Overweight (23~25), Obesity (Over 25)
† Cholesterol : Normal A (≥230mg/dl), Normal B (230mg/dl~260mg/dl), Suspicious (>260mg/dl)

역시 저위험군보다는 고위험군에 상대적으로 많이 분포하고 있었다. 한편, 개인의 과거병력이 위 · 십이지장궤양이나 대장용종이 있는 경우, 가족병력 중에서 암 질환자가 있던 경우, 그리고 검진결과 비만이나 콜레스테롤 수치가 질환의심군에 속하는 자들은 상대적으로 저위험군 보다는 고위험군에 많이 분포하고 있었다.

2. 고위험군 특성을 반영한 관리 프로세스의 유도

먼저, 관리의 효율성 제고 측면에서 암 발생 예측확률분포를 이용하여 사업대상군 접근을 달리 할 수 있다. 즉, 검진 후 건강한 사람이 2년 이내에 대장암으로 진단 및 진료 받을 확률이 가장 높은 상위 5~10% 이내의 속하는 대상자들을 고위험군으로 정의하고 집중 관리하며, 상위 10~30% 이하에 속하는 대상자들은 위험군으로서 경고 관리, 기타는 지속적인 모니터링 통해 추이를 관찰하는 방법을 취함으로써 대장암 관리의 효율성을 높일 수 있다.

한편, 관리군별 대상자의 특성을 반영한 프로세스의 유도는 성, 연령, 개인의 생활습관 그리고 과거병력 등을 반영하여, 상위 10%에 해당되는 고위험군에 Figure 4와 같이 4개의 관리 특성을 부여할 수 있다. 즉, 고위험 관리군(1)은 50대 남자, 과거병력(대장용종)이 없으면서 상습적인 음주와 흡연하는 자들로 구성되어 있으므로, 이러한 특성이 반영된 차별화된 관리를

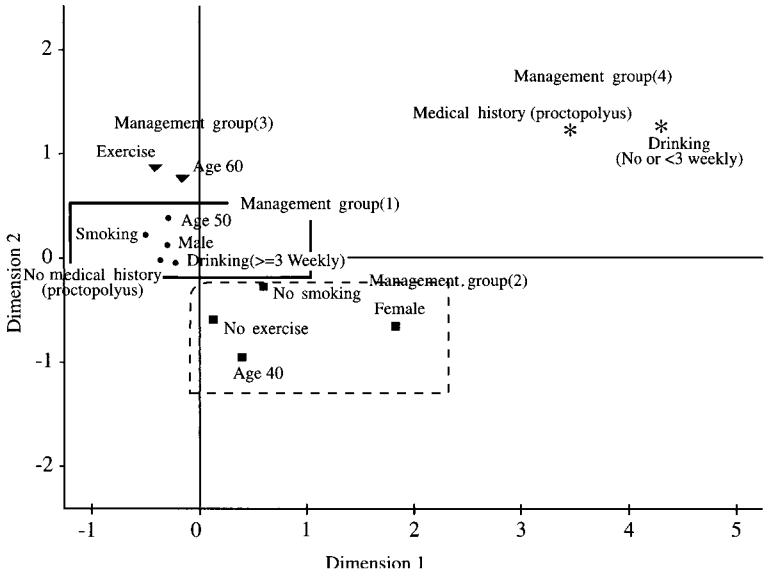

Figure 4. Type of care management group according to the risk factor of colon cancer incidence.

유도할 수 있을 것이다. 고위험 관리군(2)은 40대 여자로 흡연은 전혀 하지 않지만 건강행위 중 운동을 전혀 하지 않는 대상자이므로, 운동관리에 대한 가이드라인 및 관련 건강정보를 제공할 수 있다. 고위험 관리군(3)은 60대 이상의 노인계층으로 주기적으로 운동하는 대상자 특성을 갖고, 고위험 관리군(4)는 개인의 과거병력(대장용종)이 있고 비음주자 또는 간헐적으로 음주하는 대상자이므로 건강한 생활습관을 권장하여 과거의 질환이 재발되지 않도록 하는 차별화된 관리 프로세스를 계획, 수행할 수 있을 것으로 사료된다.

3. 활용 및 발전 방안

국민건강보험공단의 암 건강검진은 정부 등의 노력으로 검진대상이 확대되고는 있으나, 암 검진을 통해 나타난 결과에 대해 보험자로서 수검자에게 제공하는 사전 및 사후 관리서비스는 여전히 부족한 실정이다. 그런데, 건강행위, 가족력 등의 건강위험요인에 대한 관리는 암 예방에 큰 효과를 기대할 수 있어, 관리서비스의 제공은 암 예방차원에서 반드시 필요한 요소라 할 수 있다. 뿐만 아니라, 건강검진사업의 효과를 높이기 위해서는 획일적인 서비스의 제공에서 벗어나, 검진결과를 통해 확인된 수검자의 특성에 맞는 일련의 관리서비스를 제공해야 할 것이다. 이러한 점에서 본 연구에서 인구사회학적 특성, 생활습관, 가족력, 의사의 임상학적 소견 등으로 개발된 암 발생 예측모형을 활용하여 집단을 세분화하고 관리대상자를 선정한다면, 보다 효율적으로 관리대상자가 선정될 것이며, 그 특성에 맞는 관리서비스를 제공할 수 있을 것으로 판단된다.

또한 본 연구의 제한점들을 보완하여 보다 정확한 예측 모형을 개발한다면 맞춤형 건강정보의 제공이 가능하다. 특히, 대장암 발생 예측모형을 통해 고위험군으로 밝혀진 수검자에 대해서는 질병관리 서비스 제공 방안을 제시하고, 추후 검진 및 진료에 대해서는 알림 기능 서비스를 제공하고, 집단별 특성을 반영한 암 예방관리 맞춤형 정보를 제공할 수 있게 될 것이다. 더욱이, 맞춤형 건강정보는 수검자에게만 활용되는 것이 아닌, 의료공급자를 위해서도 그 활용도가 클 것이다. 즉, 각 개인 단위의 맞춤형 건강정보가 구축된다면, 의료공급자에게도 진료나 검진 시 환자의 상태를 파악할 수 있는 의사결정의 근거자료로도 활용될 수 있을 것이다. 단, 개인정보 보호라는 중요한 사항이 걸림돌이 되나, 이러한 문제가 원활히 해결된다면 앞으로 EHR (Electronic Health Record : 전자건강기록) 및 HRA (Health Risk Appraisal : 건강위험평가)의 중요한 컨텐츠 (contents)로 활용될 수도 있을 것이며, 이러한 정보의 피드백 (Feed Back)을 통해 더 가치 있는 맞춤형 건강정보가 생성될 수 있을 것으로 판단된다. 또, 개인별 암 발생 위험지수를 통해 향후 암이 발생될 위험성을 인지하여 지속적으로 스스로 건강관리를 할 수 있도록 유도하고, 보험자로서 국민건강보험공단은 다양한 맞춤형 건강정보의 제공을 통해 수검자들이 건강검진을 지속적으로 받을 수 있도록 유도할 수 있게 된다. 더 나아가 적극적인 홍보를 수행함으로서 현재 건강검진 미수검자가 검진을 받을 수 있도록 적극 유도할 수 있을 것이다. 이러한 일련의 과정을 통해 조기검진 유도를 위한 다양한 프로세스의 개발을 위한 기반이 마련될 수 있을 것이다.

4. 연구의 제한점

본 연구에서 개발된 예측모형은 일반화된 예측모형으로 사용함에 있어 몇 가지 제한점을 갖고 있어 주의를 요한다. 첫째, 국민건강보험공단의 건강검진을 받은 수검자를 중심으로 예측모형을 개발했기 때문에, 우리나라 모든 국민들에게 적용하기에도 다소 부적절할 수 있다. 둘째, 분석대상자의 선택편의가 존재한다. 현재 국민건강보험 가입자의 직역별, 거주지역별 그리고 연령 및 성별로 수검률에 있어 큰 차이를 보이고 있음에 따라, 건강행위, 질환발생 및 유병율 등의 특성에 있어 선택편의가 존재할 가능성이 있다. 셋째, 모형의 대장암 발생 여부에 대한 확인조사가 필요하다. 본 연구는 국립암센터의 암 등록 자료와 통계청의 사망원인 통계자료를 연계하여 모형의 대장암 발생 여부를 정의하고, 이 외에 암 검진의 종합판정 결과, 암 치료 대상자를 포함하였다. 또한 암 검진 후, 기타질환인 사람의 입원내역과 투약일수를 기초로 암 발생의 대장암 발생 여부를 정하였다. 때문에 환자의 실제 대장암 발생을 모르는 상황에서 단지 청구된 상병코드만을 근거로 대장암 발생 여부를 정했다는 면에서 다소의 편의가 존재할 수 있을 것으로 사료된다. 넷째, 암 발생 예측모형은 공단의 특정 암 건강검진 사업에 활용할 목적으로 개발하였기 때문에, 본 연구는 암 발생기간을 검진시점으

로부터 검진주기인 2년으로 하였다. 이는 역학적인 측면에서 암 질환의 특성을 고려해 볼 때, 추적기간이 매우 짧을 수 있다. 이에 Richard와 David는 보건의료분야에서 예측모델에 의한 접근은 기본적으로 모집단에 대한 제한된 가정에 근거하여 추론하기 때문에 의학적인 자료의 분석에서 모델의 예측력이 부족할 수 있으며, 복합적인 상관관계를 보이는 생물학적인 현상에 잘 맞지 않는 경우가 발생할 수 있다고 지적하였다[17].

따라서 향후 연구에서는 이러한 문제를 최소화하기 위해, 무엇보다 낮은 수검률을 50% 이상으로 향상시킬 필요가 있으며, 검진결과와 관련된 자료의 질 관리뿐만 아니라, 검진기관의 정도관리에도 관심을 가질 필요가 있다. 또한, 검진주기와 상관없이 장기 암 연구 코호트 자료를 구축하여 암 발생 예측모형을 개발해 보는 것도 이러한 측면에서 의의가 있다 하겠다. 본 연구는 이러한 한계에도 불구하고 국민건강보험공단의 방대한 데이터베이스를 이용하여 가입자에 대한 적극적인 건강관리사업으로 발전시킬 수 있는 정보기술의 인프라 구축이라는 측면에서 중요한 의의를 갖는다고 할 수 있겠다.

결 론

본 연구에서는 국민건강보험의 건강검진자료를 활용하여 개인의 생활습관, 과거병력 등에 기초하여 비감염성 만성질환인 대장암의 로지스틱 회귀예측모형을 개발하였고, 이를 이용하여 고위험군의 특성을 분석하였다. 그 결과, 대장암의 발생위험에 영향을 미치는 요인으로는 음주, 운동, 개인의 과거병력(대장용종), 가족병력 등이 중요한 요인으로 확인되었다.

따라서 대장암 등 중요 만성질환의 관리와 관련하여 건강검진사업이나 맞춤형 의료정보제공사업을 수행함에 있어, 가입자의 특성 등을 반영한 선별적 관리가 필요하며, 대상 집단을 세분화하고 관리대상자를 선정한다면, 보다 효율적으로 관리대상자가 선정될 것이며, 그 특성에 맞는 관리서비스를 효과적으로 제공할 수 있을 것으로 판단된다. 앞으로, 국민건강보험공단의 건강관리사업 및 가입자지원사업을 보다 효과적으로 수행하기 위해서는 본 연구에서 개발한 예측모형 등을 이용하여 효율적이고 효과적인 질병 관리방안을 도출할 수 있을 것으로 판단된다.

참고문헌

1. Yi JJ, Yoo WK, Kim SY, Kim KK, Yi, SW. Medical expenses by site of cancer and survival time among cancer patients in the last one tear of life. *J Prev Med Public Health* 2005; 38(1): 9-15(Korean)
2. 통계청. 사망원인통계분석; 각 연도
3. Shin HR, Ahn YO, Bae JM, Shin MH, Lee DH, Lee CW, Ohrr H, Ahn DH, Ferlay J, Parkin, DM, Oh DK, Park JG. Cancer incidence in Korea. *Cancer Res Treat* 2003; 34(6): 405-408
4. Baldwin FD. Well-managed care. Predictive modeling helps keep high-risk patients from becoming high-cost patients. *Healthc Inform* 2004; 21(8): 27-28
5. http://www.cpmk.com
6. Dogu Celebi. The power of predictive modeling. *Healthc Inform* 2003; 8: 56-58
7. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, Hunter DJ. Harvard report on cancer prevention volume 4: Harvard cancer risk index. Risk index working group, Harvard center for cancer prevention. *Cancer Causes Control* 2000; 11(6): 477-88
8. Yoo KY, Shin HR. Cancer epidemiology and prevention. *Korean J Epidemiol* 2003; 25(1): 1-15(Korean)
9. http://www.ncc.re.kr
10. 김정순. 역학원론. 신광출판사; 2004
11. Lee DC, Soe I, Lee HR, Kim DK. Factors affecting on the compliance of he health screening program : A study for the insured in a district medical insurance association. *J Korean Acad Fam Med* 1997; 18(7): 739-751(Korean)
12. 이원철 외 8인. 건강검진 · 암검사 수검행태 연구. 국민건강보험공단, 가톨릭대학교 의과대학; 2004
13. 강현철, 한상태, 최종후, 김은석, 김미경. SAS Enterprise Miner를 이용한 데이터마이닝 - 방법론 및 활용. 자유아카데미; 2001
14. 강현철. SAS Enterprise Miner를 이용한 데이터마이닝. 자유아카데미; 1999
15. Giudici, P. Applied Data Mining - Statistical Methods for Business and Industry. Wiley, 2003
16. Kang SH, Choi SH. The efficient ways of health promotion program on public health center using data mining. *J Korean Soc Med Inform* 2001; 7(2): 37-48 (Korean)
17. Richard L. David S. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg* 1997; 63(6): 1635-1643