보건행정학회지 : 제 16 권 제 2 호
Korean J. of Health Policy & Administration
Vol. 16, No.2, ,Jun 2006

▧ 원 저 ▧

데이터마이닝 기법을 이용한 당뇨 발생 예측모형 개발

이애경†, 박일수, 강성홍*, 강현철**
국민건강보험 연구센터, 인제대학교 보건관리학과*, 호서대학교 정보통계학과**

<Abstract>

A Development of a Predictive Model Using the Data Mining Technique on Diabetes Mellitus

Ae Kyung Lee†, Il Su Park, Seoung Hong Kang*, Hyn Chul Kang**
National Health Insurance Research Center
*Dept. of Public Health Administration, College of Humanities Social Science, Inje University**
*Dept. of Informational Statistics, College of Natural Sciences, Hoseo University***

As prior studies indicate that chronic diseases are mainly attributed to health behavior, preventive health care rather than treatment for illness needs to improve health status. Since chronic conditions require long-term therapy, health care expenditures to treat chronic diseases have been substantial burden at national level.

In this point of view, this study suggests that the health promotion program should be based on Knowledge Based System. Using Data Mining Technique, we developed a predictive model for preventive healthcare management on diabetes mellitus.

Generally, in the outbreak of diabetes mellitus there is a difference in lifestyle and the risk factors according to gender. So we developed a predictive model in accordance with gender difference and applied the Logistic Regression Model based on Data Mining process. The results of the study were as follow. The lift of the last predictive model was an average 2.23 times(male model : 2.13, female model 2.33) more improved than in the random model in upper 10% group. The health risk factors of diabetes mellitus are

* 접수 : 2005년 9월 6일, 심사완료 : 2006년 4월 3일
† 교신저자 : 이애경, 국민건강보험 연구센터 책임연구원(02-3270-9878, aklee036@nhic.or.kr)

gender, age, a place of residence, blood pressure, glucose, smoking, drinking, exercise rate. On the basis of these factors, we suggest the program of the health promotion .

Key Words : Predictive model, Health promotion, Health risk factor of diabetes mellitus, Data Mining Technique.

Ⅰ. 서 론

최근 들어 만성질환 등의 건강문제는 주로 잘못된 생활습관이 주된 원인으로 밝혀지면서 과거의 사후 치료중심보다는 사전예방관리가 보건의료체계에서 요구되고 있는 실정이며, 특히 만성질환은 완치가 어려울 뿐 아니라 장시간이 소요되어 국민의료비의 증가 등으로 국가경제에도 부정적인 영향을 끼치고 있는 것으로 나타났다. 이에 지선하(2001)는 일반적으로 치료보다는 예방이 급여비 절감에 보다 효과적인 방안이라 설명하였고, 미국의 오하이오주의 관리조직(Managed Care)에서는 관절염환자에 대한 DM(Direct Mail)을 활용한 보건교육을 통해 관절염에 대한 급여비 절감효과를 거두는 등 건강증진사업이 급여비 절감에 기여한다는 다양한 증거를 제시한 바 있다.

건강증진사업은 단순히 질병을 치료하고 예방하는 것에만 국한된 것이 아니고, 적극적인 건강향상을 목적으로 사람들의 건강의식이나 행동변화를 유도하여 건강할 수 있는 잠재력을 기르고 건강위험요인을 조기에 발견함으로써 건강을 유지·증진하고자 하는 것이 주요 목적이다. 그러나 실제 전문인력 중심으로 이러한 목적을 달성하기에는 관리해야 할 대상이 너무 광범위하므로 비용문제가 크게 대두되며, 효과적인 건강증진사업을 추진하기 위해서는 단순한 정보만을 제공하는 것이 아닌 개개인의 특성과 개인의 생활에 연관된 정보를 제공했을 때의 비용 대비 효과성이 고려될 필요가 있다. 이러한 이유로, 최근에는 보험자가 건강증진사업을 보다 효과적으로 추진하기 위한 정보기술의 활용에 대한 검토가 다양하게 이루어지고 있으며, 그 중 고객관계관리기법(Customer Relationship Management : CRM)은 국내외적으로 건강증진사업의 효과적 관리를 지원해 주는 보건의료분야의 신정보기술로 많은 주목받고 있다. CRM기법은 이미 산업계에서 지식경영(Business Intelligence)의 도구로서 널리 활용되고 있으며, 이것은 고객과 관계하는 혁신적인 방법론으로 인식되고 있다.

보건의료분야에 있어서 CRM은 건강에 대한 정보를 필요로 하는 관점에서 고객을 분류하는 것으로, 건강정보가 필요한 계층에게 건강정보를 제공하여 이들이 바람직한 건강행위를 하도록 유도하여 건강증진 및 적절한 치료가 이루어지도록 하는 것이라 할 수 있다. 이와 같은 특성에 기인하여 보건의료분야의 CRM기법의 적용분야는 크게 보건교육(Education

Management), 건강증진(Wellness Management), 질환관리(Disease Management) 그리고 조기검진(Intervention Management)분야 총 4가지로 분류할 수 있다(강성홍, 2003). 이러한 분야에 CRM기법을 적용하기 위해서는 개인에 맞는 건강정보를 감별하는 것이 매우 중요한데, 이미 역학에서 건강위험평가(Health Risk Appraisal)[1]라는 도구를 이용하여 특정 질병에 대한 사망확률을 예측하였다. 건강위험평가(HRA)의 궁극적인 목적은 이를 보건교육에 활용하여 개인의 건강위험요인을 제거하여 건강위험도를 줄이는 것이다. 이러한 건강위험평가 도구는 여러 곳에서 다양하게 이용되고 있는데, 미국의 경우 의료기관뿐 아닌 학교, 군대, 기업체 등에서 약 300여종이 개발되어 사용되어지고 있다. 최근 우리나라에서도 건강보험에 대한 단일보험자인 국민건강보험공단이 국민들에게 다양한 의료이용정보와 건강정보를 인터넷 및 방문서비스를 통하여 제공하고 있는데, 그 중의 하나가 2004년부터 실시하고 있는 건강위험평가(HRA)서비스이다. 그러나 공단이 제공하고 있는 건강위험평가는 건강검진을 받은 우리나라 성인 인구 전체를 대상으로 실시하고 있다는 점에서는 큰 의의가 있으나, 건강나이 산출과 사망확률을 질병에 근거하지 않은 한국인의 평균건강수준에서 비교하여 제시하기 때문에 향후 질병별로 보다 정교한 도구의 개발이 필요하다 할 수 있다.

현재 우리나라는 경제발전, 서구화의 영향에 따른 식습관변화, 체질변화 및 고령화 등으로 인해 만성질환이 증가하고 있다. 이 중에서도 당뇨는 1994년부터 2003년 말까지 10년간 당뇨병으로 한번이상 병·의원을 찾은 국민은 401만 2000명(사망자제외)이며, 이는 우리나라 국민 100명 가운데 8.3명이 당뇨관련 질환을 앓고 있다는 수치이다. 현재의 당뇨병 환자 발생확률을 감안한다면 국내 당뇨병환자는 2015년 553만명, 2030년 722만명(전 국민의 14.4%)에 이를 것으로 추정하고 있다(건강보험심사평가원, 2005). 또한, 이 수치는 당뇨관련 질환으로 의료기관의 방문을 통하여 조사되어진 통계이나 실제로 당뇨를 앓거나 위험에 노출된 국민을 감안한다면, 현재의 당뇨 유병율 8.4% 그 이상이 될 것이라 추측된다. 이러한 점에서 보험자 또는 공공보건기관의 만성질환관리는 보건정책에서 반드시 다뤄져야하는 부분임에는 틀림이 없다고 사료된다.

따라서 본 연구에서는 국민건강보험공단이 보유한 건강검진데이터, 문진데이터, 자격데이터 및 급여데이터를 활용하여 사전예방관리측면에서 관리의 효율성을 높이기 위한 개인별 발생 예측모형을 만성질환인 당뇨중심으로 개발하고, 이를 기초로 개개인별로 예상되는 당뇨 발생확률에 근거를 두고 당뇨 발생 가능환자의 건강관리방안을 제시하고자 한다. 또한 공공기관이 향후 추진해 나가야할 만성질환에 대한 건강증진사업의 방향성을 모색하여 제시하고자 한다.

---

1) 과거의 유사한 특성을 가진 개인들의 정보 즉 역학 자료에 근거해서 위험도를 산출하는 것으로 개인의 생활습관, 가족력, 유전적 특성 등 건강위험요인을 근거로 각 개인의 사망확률을 구하는 것

Ⅱ. 연구방법

1. 분석자료

당뇨 발생 예측모형[2]을 개발하기 위한 분석대상자는 건강검진 및 문진 자료를 근거로 연구 대상은 2000년부터 2003년까지 건강검진을 지속적으로 받은 40세 이상 건강보험가입자로 선정하였다. 단, 당뇨질환의 분류 기준은 건강보험공단의 건강검진업무처리에 있어서의 당뇨 대상자 선정기준[3]과 의료기관에서의 선정기준에 근거하였다. 즉, 국제표준 질병사인분류기준에 근거하여 주진단과 부진단 상병코드(E10~E14, O240~O243, G590, G632, H280, H360, N083)로 진료를 받은 사람, 세계보건기구(WHO)의 당뇨수치 기준에 의한 검진의 진단결과가 당뇨 의심자인 경우, 식전혈당이 121이상인 사람 그리고 과거 당뇨 질환경력이 있는 사람을 당뇨 질환자로 정의하였다.

분석자료는 건강보험공단의 원천시스템(Source System) 및 데이터웨어하우스(Data Warehouse)의 운영계 데이터저장소(Operational Data Store)와 각 주제별 데이터마트(Data Mart)에서 2000년부터 2003년까지의 건강검진 및 문진자료와 현물급여자료의 각 연도별 개인급여정보(2004년 5월 현재 지급기준)와 상병정보를 이용하였으며, 수검자의 자격정보는 2003년 12월 31일 기준의 자격데이터베이스를 활용하였다.

2. 변수설명

당뇨 발생 예측모형 개발을 위해 1차 건강검진 받은 건강보험 가입자들의 인구사회학적 요인, 생활습관요인, 생물학적요인 및 진료이용량 등이 모두 반영될 수 있도록 설명변수를 고려하였다. 단, 개인의 유전자 정보는 임상자료의 부재로 본 연구에서는 고려하지 않는 것으로 하였다.

첫째, 당뇨 발생 가능한 고위험군의 일반적인 인구사회학적 및 경제적 특성을 보기 위해 성별, 연령, 거주지역, 건강보험료[4]를 독립변수로 사용하였고, 의료이용량 변수로는 총 진료비, 총 내원일수, 총 투약일수를 고려하였다. 둘째, 당뇨질환은 생활습관의 영향(박경수, 2002)도 크게 받기 때문에 건강검진의 문진정보 즉 과거이력, 흡연량, 음주량, 운동량, 식생활습관, 스트레스 정도 등의 생활습관요인을 변수로 선정하였다. 셋째, 건강검진결과를 토대로

2) 본 연구에서의 당뇨 발생의 의미는 진단(검진) 및 진료를 행하였음을 뜻한다.
3) 건강보험공단 건강검진사업의 당뇨수치 기준은 세계보건기구(WHO)에 기준에 준함
4) 소득수준은 건강보험 부과보험료를 대리변수로 이용함.

<표 1> 전체분석모형의 변수 정의

특 성	변 수 명	설 명
종속변수	당뇨 질환 유무	
독립변수		
인구사회학적 특성	성별	남여
	연령	49세이하, 50~59세, 60세 이상
	거주지역	대도시, 중소도시, 농어촌
건강보험 가입자 특성	가입자격	직장가입자, 지역가입자
	보험료	부과보험료
	의료이용량	총 진료비, 총 내원일수, 총 투약일수
건강문진 특성	가족력	당뇨 가족력 있음, 없음
	음주	마시지 않는다, 월2~3회 정도, 일주일에 1~2회 일주일에 3~4회, 거의 매일
	흡연	피우지 않는다, 과거에 피웠으나 지금은 끊었다, 현재도 피운다
	운동시간	안한다, 1~2회, 3~4회, 5~6회, 거의 매일
	과거병력	결핵, 간염, 간장질환, 고혈압, 심장병, 뇌졸중, 당뇨병, 암,
건강검진 특성	비만도(BMI : kg/ m^2)	비만(25이상), 위험체중(23~25미만), 정상체중(23미만)
	수축기혈압 (mmHg)	질환의심(160이상), 정상B(140~159), 정상A(139이하)
	이완기혈압 (mmHg)	질환의심(95이상), 정상B(90~94), 정상A(89이하)
	글루코스 (mg/dl)	질환의심(121이상), 정상B(111~120), 정상A(110이하)
	콜레스테롤 (mg/dl)	질환의심(261이상), 정상B(231~260), 정상A(230이하)

주. 1차 건강검진 관련 문진 및 검진결과 관련 모든 항목을 고려하였으나, 실제 당뇨 발생 관련 영향을 주는 특성요인 중심으로 정리함.

수축기혈압, 이완기혈압, 콜레스테롤, 비만도 등의 생물학적요인들 역시 당뇨 발생의 위험요인으로 고려하였고, 이 외에도 다각적인 발생요인을 찾기 위해 인구학적요인, 기족력, 생활습관요인 그리고 생물학적요인간을 교차변수로 파생시켜 예측모형에 반영하였다(김영식, 2003).

3. 분석방법

본 연구는 CRM 기반의 건강관리를 위해 개인별 당뇨 발생 예측모형을 최신정보기술로

각광받고 있는 데이터마이닝 방법론을 적용하여 개발하고자 한다.

Phase 1(사전단계)에서는 건강관리를 위한 예측모형의 유형을 정하고 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 구축한다. Phase 3(분석단계)에서는 최종 구축된 분석용 데이터마트를 활용하여 다양한 통계기법과 데이터마이닝 기법인 의사결정나무(Decision Tree), 신경망기법(Neural Network), 회귀분석(Linear Regression, Logistic Regression 등) 등을 통한 모델링 작업을 수행하였는데, 개발된 모형의 검증을 위해 본 연구의 모형화 작업에 있어서는 전체 분석자료를 Training Data(70%)와 validation Data(30%)로 분할하여 Training Data을 통해 모형을 생성하고, 만들어진 모형을 Validation Data에 적용시켜 모형생성 시 발생할 수 있는 과적합문제(Overfitting)를 해결하였다. 또한, 다양한 모형생성 알고리즘을 통해 만들어진 모형들 중 가장 좋은 모형을 평가·선정하기 위해 Lift 차트와 ROC 차트 등을 이용하였다. Phase 4(적용단계)에서는 모형생성을 위한 데이터마트와 똑같은 형태로 만들어진 모형적용을 위한 새로운 데이터마트에 생성된 모형을 적용시켜 사전예방 건강관리모형에서는 당뇨에 대해 향 후 2년간 고위험군 발생 확률을 예측하였다. 또한, 예측된 값들의 정확도를 알아보기 위한 검증을 위해 2003년 자료에 적용하였다.

4. 예측모형 개발 프로세스

사전예방관리를 위한 모형은 만성질환이 없는 자가 향후 2년 이내에 만성질환이 걸릴 확률을 구하는 모형으로서, 본 연구에서는 2000년에 건강검진 받은 만 40세 이상 수검자 중 당해 년도에 각각 당뇨 질환으로 진단 및 진료를 받지 않은 대상자들이 2001년에서 2002년까지 당뇨 질환의 진단 및 진료여부에 대해 모델링을 하였다. 그리고 생성된 모형을 2001년 당시 만성질환으로 진단 및 진료를 받지 않은 자에 적용시켜 2002년에서 2003년 2년 동안 각각 당뇨 발생의 상대위험도를 예측하였다. 또한, 예측된 모형 및 결과의 정확도를 파악하기 위해 실제 2003년의 급여데이터 및 검진데이터에서 발견되어진 당뇨 질환의 진단 및 진료여부와 비교·검증하였다(그림 1).

모델링을 위한 분석기간 및 Target의 정의는 <표 2>에 제시된 바와 같고, 본 연구에서는 데이터의 변환, 분석용 데이터셋의 구성 및 통계분석을 위해서는 통계패키지인 SAS 8.2를 활용했고, 데이터마이닝 모델링을 위해서는 SAS의 데이터마이닝 툴인 Enterprise Miner 4.1을 활용했다.

<표 2> 데이터마이닝 프로세스 사전단계 : 사전예방 관리모형

분석주제	분석대상	분석기간 및 Target 정의
사전예방 관리모형 구축 -당뇨발생예측모형	2000년 건강검진 수검자 연령이 만40세 이상이고 당뇨 비진단 및 비진료자	2001년~2002년 당뇨 Target 1 : 당뇨 진단 및 진료자 0 : 당뇨 비진단 및 비진료자

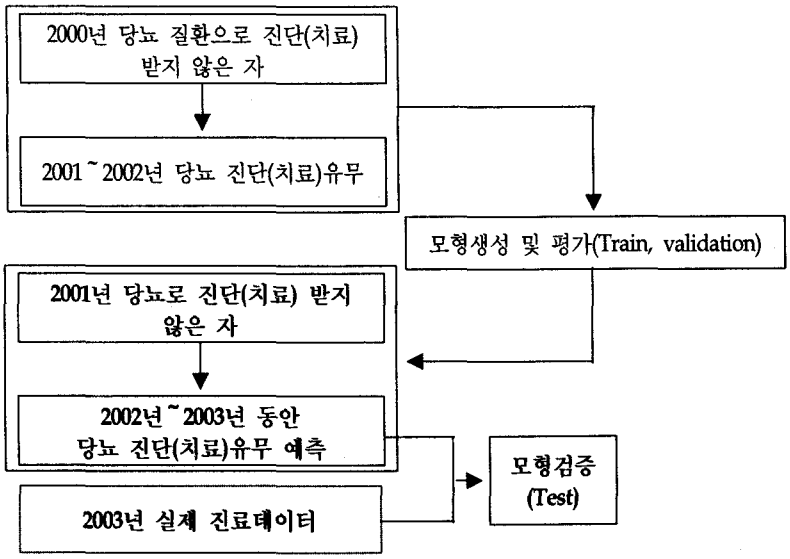


그림 1. 당뇨 발생 예측모형 개발 프로세스

Ⅲ. 데이터마이닝을 이용한 당뇨 발생 예측모형

2000년 건강검진 수검자 5,104,852명 중 건강검진과 진료를 통하여 당뇨질환이 발견(4,137,097명)되지 않고, 이들 중 연령이 만 40세 이상(2000년 검진 당시)이고, 2001년과 2002년에 지속적으로 검진을 받은 사람 386,146명을 대상으로 당뇨 발생 예측모형을 생성하였다.

1. 분석대상자의 일반적 특성

당뇨 발생 예측모형개발을 위한 분석대상자의 일반적 특성은 인구사회학적 특성, 보험가입자격 및 보험료 수준, 진료 서비스 이용량, 건강검진 및 문진결과에 나타난 건강위험요인 등을 보았다.

<표 3> 당뇨 분석대상자의 일반적 특성 I

특 성			N	(%)
인구사회학적 특성	연령	60세이상	18,228	(4.72)
		50~59세	102,274	(26.49)
		49세이하	265,644	(68.79)
	성별	남	292,146	(75.66)
		여	94,000	(24.34)
	거주지역	대도시	191,407	(49.57)
		중소도시	169,879	(43.99)
		농어촌	22,881	(5.93)
		None	1,979	(0.51)
보험가입 자격 및 경제적 수준	가입자격	지역	21,835	(5.65)
		공교	2,772	(0.72)
		직장	359,560	(93.12)
		None	1,979	(0.51)
	보험료수준 (4분위수 이용)	상위76%이상	99,290	(25.71)
		상위51%~75%	94,137	(24.38)
		상위26%~50%	103,209	(26.73)
		상위25%이하	89,510	(23.18)
진료이용량	총진료비수준 (4분위수 이용)	상위26%~50%	96,540	(25.00)
		상위51%~75%	96,538	(25.00)
		상위76%이상	96,529	(25.00)
	총내원일수 (4분위수 이용)	상위26%~50%	86,540	(22.41)
		상위51%~75%	96,644	(25.03)
		상위76%이상	95,601	(24.76)
	총투약일수 (4분위수 이용)	상위26%~50%	92,548	(23.97)
		상위51%~75%	95,525	(24.74)
		상위76%이상	95,567	(24.75)
전 체			386,146	(100.0)

주 : None의 의미는 특성치의 결측값을 의미

<표 4> 당뇨 분석대상자의 일반적 특성 II

특		성	N	(%)
건강 위험 요인	검진 결과	비만도 (BMI : kg/)	비만(25이상)	119,473	(30.94)
			위험체중(23~25미만)	111,707	(28.93)
			정상체중(23미만)	154,966	(40.13)
		수축기혈압 (mmHg)	질환의심(160이상)	18,316	(4.74)
			정상B(140~159)	71,163	(18.43)
			정상A(139이하)	296,667	(76.83)
		이완기혈압 (mmHg)	질환의심(95이상)	32,709	(8.47)
			정상B(90~94)	77,936	(20.18)
			정상A(89이하)	275,501	(71.35)
		글루코스 (mg/dl)	질환의심(121이상)	0	(0.00)
			정상B(111~120)	24,982	(6.47)
			정상A(110이하)	361,164	(93.53)
		콜레스테롤 (mg/dl)	질환의심(261이상)	14,559	(3.77)
			정상B(231~260)	41,799	(10.82)
			정상A(230이하)	329,788	(85.41)
		생활습관개선: 흡연	있음	124,372	(48.50)
			없음	132,068	(51.50)
		음주	있음	147,461	(55.08)
			없음	120,270	(44.92)
		운동량	있음	112,900	(42.61)
			없음	152,098	(57.39)
		식습관	있음(규칙적)	158,451	(58.54)
			없음(불규칙)	112,234	(41.46)
	문진 결과	가족력	없음	188,798	(95.21)
			있음	9,497	(4.79)
		음주여부	마시지 않는다	120,270	(44.92)
			월2~3회 정도	53,197	(19.87)
			일주일에 1~2회	62,908	(23.50)
			일주일에 3~4회	22,536	(8.42)
			거의 매일	8,820	(3.29)
		흡연여부	피우지 않는다	100,491	(39.19)
			과거에 피웠으나 지금은 끊었다	31,577	(12.31)
			현재도 피운다	124,372	(48.50)
		운동시간	안한다	152,098	(57.40)
			1~2회	75,015	(28.31)
			3~4회	22,420	(8.46)
			5~6회	5,031	(1.90)
			거의 매일	10,434	(3.93)
전 체				386,146	(100.0)

당뇨 질환관련 분석대상자의 특성 중 성별분포는 남자가 75.66%, 여자가 24.34%였고, 연령은 50세미만이 68.79%, 50대가 26.49%, 60세 이상은 4.72%순으로 분포를 보였다. 보험료 가입자 자격이 직장인 경우가 93.12%으로 가장 많고, 상대적으로 지역가입자는 전체의 5.65%를 나타내었다(표 3). 건강위험요인인 비만도(BMI : kg/m^2)는 비만(25이상)이 30.94%, 위험체중(23~25미만)은 28.93%로 나타나 전반적으로 체중에 대한 관리가 필요로 하고 있음을 보였다. 혈압분포는 수축기혈압(160mmHg 이상 : 질환의심)인 사람이 전체의 4.74%, 이완기혈압(95 mmHg이상 : 질환의심)인 사람은 8.47%였고, 글루코스의 정상B(111~120mg/dl)는 전체의 6.47%이고, 콜레스테롤의 질환의심(261mg/dl 이상)은 전체의 3.77%로 나타났다(표 4). 한편, 문진결과에 의한 흡연 분포는 현재도 피고 있다가 전체의 48.50%로 가장 많았고, 과거에는 피웠다가 지금은 피지 않는 경우가 12.31%를 차지하였다. 또한, 음주 분포는 거의 매일 마신다가 3.29%, 1주일에 1회~4회 정도 마신다가 31.92%로 나타났고, 운동은 안 한다가 전체의 57.40%를 차지하는 것으로 나타났다(표 4).

본 분석 대상자들이 2000년부터 2002년까지 누적 진료내역은 평균적으로 총진료비는 1,148천원이었고, 총 내원일수는 67일 그리고 총 투약일수는 238일로 <표 5>와 같다.

<표 5> 진료 서비스 이용량

	평균	표준편차	최소값	사분위수			최대값
				25%	50%	75%	
총진료비	1,148,474	1,374,930	11,380	401,860	768,080	1,446,610	91,209,230
총내원일수	67	56	3	29	51	86	1,680
총투약일수	238	305	3	56	115	269	5,857

주 : 단위 : 총진료비(원), 총내원일수(일), 총투약일수(일)

2. 당뇨 발생 예측모형

당뇨 질환이 발생될 확률모형을 개발하기 위해, 본 연구에서는 Logistic Regression, Decision Tree, Neural Network을 이용하여 모형을 개발하였고, 데이터마이닝 분야에서 일반적으로 사용되는 여러 가지 모형평가 도구들을 이용하여 모형을 평가한 후 최적모형을 선택하였다.

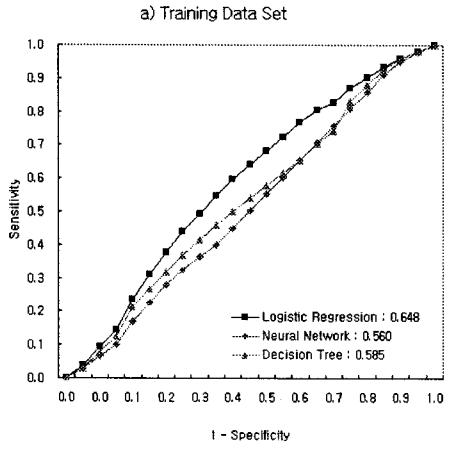
1) 모형평가

일반적으로 최적의 모형을 얻기 위해서는 여러 모형을 비교·평가해야 하고, 이를 통해 하

나의 모형이 선택되면 선택된 모형이 다른 모형에 비해 우수하다는 사실을 입증해야 한다. 모형평가란 앞서 설명한 바와 같이 예측을 위해 만든 모형이 임의의 모형(random model)보다 과연 우수한지, 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측력을 보유하고 있는지를 비교·분석하는 과정을 말한다. 본 연구에서는 당뇨 발생 확률모형으로 선택된 여러 모형들 중에서 최적모형은 ROC 곡선(receiver operating characteristic curves)과 리프트(lift)를 이용하여 선택하였고, 그 결과 Logistic Regression에 의한 모형이 가장 우수한 성능을 가진 것으로 나타났다.

<그림 2>의 ROC 곡선은 절단점(cut-off point)에 따른 민감도(sensitivity)와 특이도(specificity)를 연결한 곡선으로, 일반적으로 45° 직선의 위쪽에 위치하게 되며 ROC 곡선의 밑면적을 나타내는 C-통계량이 클수록 모형의 성능이 더 우수함을 나타낸다(Giudici, 2003; 강현철 외, 2001). 본 연구에서는 Training Data와 Test Data 모두에서 Logistic Regression에 의한 모형의 C-통계량이 가장 크므로 세 모형들 중에서는 Logistic Regression의 성능이 가장 좋은 것으로 나타났다. 또한 Logistic Regression의 경우 Training Data에 비해 Test Data의 C-통계량이 큰 차이가 없으므로 모형의 안정성 측면에서도 비교적 좋은 성능을 보였다.

<그림 3>의 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(Giudici, 2003; 강현철 외, 2001). 본 연구에서는 Training Data와 Test Data 모두에서 Logistic Regression에 의한 모형의 리프트가 상위 분위수에서 전반적으로 더 크므로 Logistic Regression의 성능이 가장 좋은 것으로 나타났다.



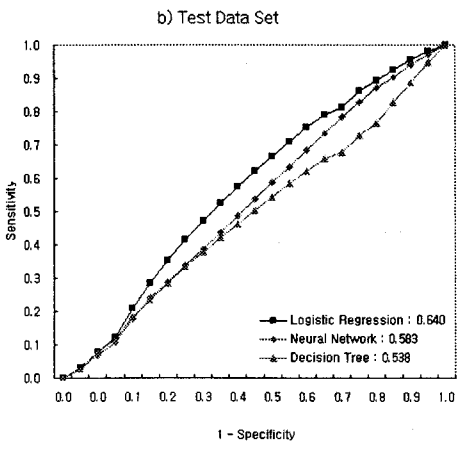


그림 2. 당뇨 발생 예측모형의 ROC곡선(Receiver Operating Characteristic Curves)

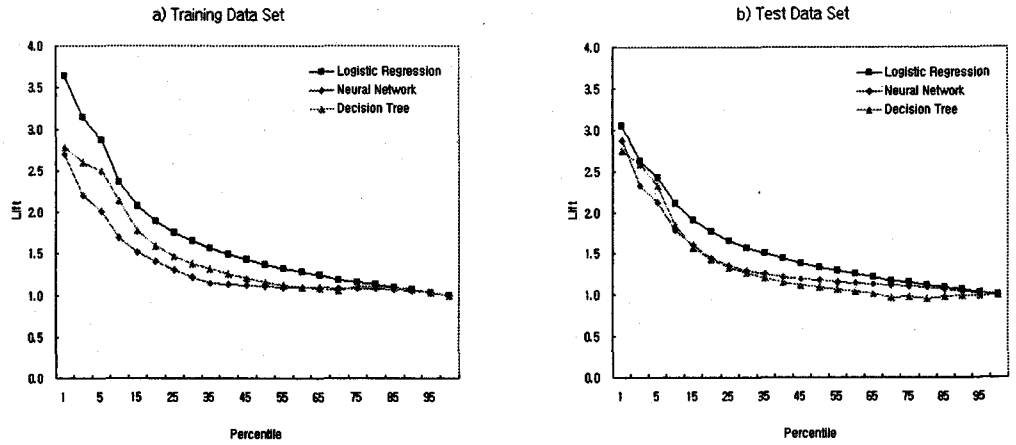


그림 3. 당뇨 발생 예측모형의 누적 리프트 도표(Cumulative Lift Chart)

<표 6> 분위수별 당뇨 발생 예측모형의 정확도(Test Data Set : Logistic Regerssion 기준)

분위수	당뇨 질환여부			리프트		정확도		
Percentile	Total	Y=1	Y=0	%Response	Lift	Accuracy	Sensitivity	Specificity
1	3,745	1,018	2,727	27.18%	3.64	0.921	0.036	0.992
3	11,236	2,639	8,597	23.49%	3.14	0.909	0.094	0.975
5	18,726	4,024	14,702	21.49%	2.88	0.897	0.144	0.958
10	**37,455**	**6,597**	**30,858**	**17.61%**	**2.36**	**0.860**	**0.236**	**0.911**
15	56,184	8,710	47,474	15.50%	2.07	0.822	0.311	0.863
20	**74,911**	**10,597**	**64,314**	**14.15%**	**1.89**	**0.782**	**0.379**	**0.814**
25	93,626	12,293	81,333	13.13%	1.76	0.741	0.439	0.765
30	**112,360**	**13,860**	**98,500**	**12.34%**	**1.65**	**0.699**	**0.495**	**0.716**
35	131,101	15,289	115,812	11.66%	1.56	0.657	0.546	0.666
40	149,822	16,685	133,137	11.14%	1.49	0.614	0.596	0.616
45	168,512	17,949	150,563	10.65%	1.43	0.571	0.641	0.566
50	187,282	19,137	168,145	10.22%	1.37	0.527	0.684	0.515
55	206,003	20,285	185,718	9.85%	1.32	0.484	0.725	0.464
60	224,734	21,491	203,243	9.56%	1.28	0.440	0.768	0.414
65	243,443	22,537	220,906	9.26%	1.24	0.396	0.805	0.363
70	262,188	23,134	239,054	8.82%	1.18	0.349	0.827	0.310
75	281,076	24,374	256,702	8.67%	1.16	0.305	0.871	0.259
80	299,614	25,254	274,360	8.43%	1.13	0.260	0.902	0.208
85	318,580	26,098	292,482	8.19%	1.10	0.214	0.932	0.156
90	337,079	26,863	310,216	7.97%	1.07	0.169	0.960	0.105
95	355,866	27,447	328,419	7.71%	1.03	0.122	0.981	0.052
100	374,555	27,990	346,565	7.47%	1.00	0.075	1.000	0.000

주 : 1) Response(%) = (해당 Percentile에서 당뇨 질환의 빈도/해당 Percentile에서 전체 빈도)×100
2) Lift = (해당 Percentile에서 당뇨 질환의 비율/전체에서 당뇨 질환의 비율)

한편, <표 6>는 Test Data에 대한 Logistic Regression 모형의 각 분위수별 정확도 통계량들을 제시한 것이다. <표 6>에서 볼 수 있듯이, 예컨대 Logistic Regression 모형에 의한 사후확률의 상위 10%에 해당되는 리프트 값이 임의의 모형(random model)의 경우에 비해 2.36배이므로 당뇨 발생 확률모형으로 개발된 본 연구모형을 실제로 사용할 때 큰 효과를 볼 수 있음을 알 수 있다(상위 10%의 절단점에서 정확도는 86% 정도이다).

<표 7> Score대별 당뇨 발생 예측모형의 적중률(Logistic Regerssion 기준)

구분	표본	등급	남자				여자			
			건수	Base Response (%)	Response (%)	Lift	건수	Base Response(%)	Response (%)	Lift
2000	Train	10%	27,588	14.13	28.36	2.01	8,975	9.06	19.73	2.18
		20%	55,176	14.13	20.01	1.42	17,949	9.06	12.40	1.37
		30%	82,764	14.13	16.66	1.18	26,925	9.06	10.74	1.19
2001	Test	10%	28,446	8.19	18.46	2.25	9,010	5.22	12.89	2.47
		20%	56,891	8.19	11.11	1.36	18,020	5.22	7.47	1.43
		30%	85,337	8.19	9.24	1.13	27,030	5.22	6.37	1.22

주 : 1) Response(%) = (해당 Percentile에서 Target 변수의 특정범주 빈도/ 해당 Percentile에서 전체빈도)×100

2) Base Response(%)는 전체 Target 변수의 특정범주의 구성비

3) Lift = (해당 Percentile에서 Target변수의 특정범주빈도/ 전체 Target 변수의 특정범주빈도)

2) 당뇨 발생에 영향을 미치는 특성 요인

만성질환인 당뇨 발생에 영향을 미치는 특성 요인들을 살펴보기 로지스틱 회귀모형을 이용하였고, 고려된 위험요인으로는 인구사회학적인 특성, 진료 이용량 그리고 건강검진 및 문진 결과 등이고 그 결과 <표 8>, <표 9> 그리고 <표 10>과 같다. 예측모형은 발생에 영향을 주는 여러 요인들은 성별에 따라 다른 행태를 보이므로, 본 연구에서는 당뇨 발생 예측모형을 전체 그리고 남자와 여자로 구분하여 개발하였다.

(1) 당뇨 발생 예측모형(전체)

먼저 인구사회학적 특성요인으로 고려된 성별에서는 남성이 여성보다 당뇨 발생 가능성은 상대적으로 더 높은 것으로 나타났고, 연령별로는 연령이 많을수록 당뇨 발생 가능성이 높게 나타났다. 즉 연령대가 60세 이상 그리고 50~59세인 사람들은 40~49세인 사람들보다 당뇨

발생 가능성은 상대적으로 각각 1.445배, 1.314배 높은 것으로 나타났다. 한편 개인이 거주하는 지역규모별 특성에서는 대도시 및 중소도시에 거주하는 사람이 농어촌에 거주하는 사람들보다 당뇨 발생 가능성은 미미하나 높은 것으로 보였고, 진료이용량 특성으로 고려된 총 투약일수 역시 투약일수가 많을수록 당뇨 질환 발생 가능성이 큰 것으로 나타났으나 그 정도는 역시 미미한 수준이다. 건강검진 및 문진결과로부터 나타나는 개인별 건강위험요인에 따른 당뇨질환의 발생에 영향을 미치는 특성들은 수축기혈압, 콜레스테롤, 비만도, 글루코스, 검진판정결과, 흡연 및 음주 그리고 가족력 등이다.

먼저 수축기혈압이 질환의심(160mmHg 이상)이거나 정상B(140mmHg~159mmHg)인 사람은 정상A(139mmHg 이하)인 사람에 비해 각각 당뇨 발생 위험도가 각각 1.153배, 1.108배정도 높았으며, 콜레스테롤은 질환의심(121mmHg 이상)자인 경우는 정상A(230mmHg 이하)인에 비하여 1.219배 높게 나타났다. 체중과 관련된 비만도는 비만이거나 위험체중인 사람이 정상체중인 사람보다 당뇨 발생 위험도가 각각 1.223배, 1.016배 높았고, 검진의 판정결과가 정상B로 당뇨 관리를 받을 필요가 있는 사람음 그렇지 않은 사람보다 당뇨로 진료를 받게 되는 상대위험도가 1.323배, 고혈압 질환이 있는 사람은 그렇지 않은 사람보다 당뇨 발생 위험도가 1.167배 높게 나타났다. 또한, 문진결과에서 나타난 개인별 가족력 및 생활습관의 경우, 가족력이 있는 사람은 없는 사람보다 당뇨 발생 상대적 위험도는 1.382배 높았다. 그리고 생활습관 중 1주일에 1회 이상 음주하는 남자는 그렇지 않은 남자 및 여자들에 비해 당뇨 발생위험도는 1.085배 높았으며, 육식을 좋아하는 사람 그리고 흡연을 하는 사람은 그렇지 않은 사람들에 비해 당뇨 발생 상대적 위험도가 각각 1.129배, 1.192배 높은 것으로 나타났다 (표 8).

(2) 당뇨 발생 예측모형(남자)

분석대상에서 여자를 제외한 남자들의 특성을 보다 구체적으로 살펴본 결과, 연령대에서는 50~59세 연령군이 40~49세 연령군보다 당뇨 발생 가능성이 상대적으로 1.190배 높은 것으로 나타났다. 한편 경제적 수준의 대리변수로 고려된 보험료는 전체모형에서는 당뇨 발생가능성에 유의하지 않은 요인이었으나, 보험료 부담수준이 높을수록 당뇨질환의 발생 가능성이 더 적게 나타났다. 그리고 개인이 거주하는 지역규모별 특성에서는 대도시에 거주하는 사람이 농어촌에 거주하는 사람들보다 당뇨로 진료를 받게 되는 가능성이 1.133배 높은 것으로 설명되었다. 그러나 총 투약일수는 앞의 전체모형에서와 같이 총투약일수가 많을수록 당뇨 발생가능성이 높은 것으로 나타났으나, 그 수준은 미미한 정도이다.

<표 8> 당뇨 발생 예측모형(전체)

특성 요인			추정회귀계수	상대위험도
절편			-0.9837	
인구사회학적 특성		성별	남자	1.000
			여자	-0.2136*** 0.652
		연령그룹	40~49세	1.000
			50~59세	0.0594*** 1.314
			60세 이상	0.1544*** 1.445
		거주지역	농어촌	1.000
			중소도시	-0.0111 1.029
			대도시	0.0506*** 1.094
진료이용량		총투약일수		0.000507*** 1.001
건강] 위험 요인	검진 결과	수축기 혈압 (mmHg)	정상A(139이하)	1.000
			정상B(140~159)	0.0212* 1.108
			질환의심(160이상)	0.0605*** 1.153
		콜레스테롤 (mg/dl)	정상A(230이하)	1.000
			정상B(231~260)	-0.0140 1.081
			질환의심(261이상)	0.1059*** 1.219
		비만도 (BMI : kg/m^2)	정상체중	1.000
			위험체중	-0.0563*** 1.016
			비만	0.1288*** 1.223
		복합요인 D (글루코스 mg/dl)	기 타, 기타 : 정상B, 정상체중	1.000
			정상B, 위험체중	0.1246*** 2.141
			정상B, 비만	0.2769*** 2.493
				0.2351*** 2.391
		판정관리 정상B (당뇨관리)	필요없음	1.000
			필요있음	0.1401*** 1.323
		고혈압 질환	없음	1.000
			있음	0.0774*** 1.167
	문진 결과	당뇨가족력	없음	1.000
			있음	0.1619*** 1.382
		복합요인 C (음주습관)	남,여, 적당(음주무, 월2~3회)	1.000
			남, 과음(1회이상/일주일)	0.0410*** 1.085
		육식	무	1.000
			유	0.0608*** 1.129
		흡연	무	1.000
			유	0.0879*** 1.192

주 : 1) 비만도: 정상체중(23미만), 위험체중(23~25미만), 비만(25이상)
2) 글로코스 정상A(70~110), 정상B(111~120)
3) 복합요인 C, D는 2개의 건강위험요인을 고려한 변수(C는 음주습관과 성별 ; D는 글루코스와 비만)
4) 유의확률 * : P<0.1, ** : p<0.05, *** : p<0.01

건강검진 및 문진의 결과로 나타나는 개인별 건강위험요인의 경우, 먼저 콜레스테롤이 질환의심(261mg/dl 이상)인 사람은 정상A(230mg/dl 이하)인 사람들에 비하여 당뇨 발생 상대위험도는 1.177배 높은 것으로 나타났다. 또한 검진판정결과 정상이나 당뇨관리가 요구되는 경우, 고혈압 질환이 있는 사람은 그렇지 않은 경우보다 각각 1.340배, 1.194배 당뇨 발생 상대위험도가 높은 것으로 나타났다. 한편, 문진결과에서 나타난 개인별 생활습관 특성에서는 음주량 그리고 흡연량이 많으면 많을수록 당뇨에 대한 상대위험도는 미비하나 높게 나타났다(표9).

(3) 당뇨 발생 예측모형(여자)

당뇨 발생 예측모형을 개발함에 있어, 여자들만의 특성을 보다 구체적으로 살펴보고자 <표 10>에서 같이 당뇨 발생 예측모형을 추정하여 보았다. 그 결과 여자의 경우는 연령이 60세 이상이 40~49세 보다 당뇨 발생의 상대적 위험도가 1.665배 높은 반면에, 남자의 경우는 50~59세 연령군이 당뇨 발생위험도가 유의하게 높은 것으로 나타났다. 경제적 수준의 대리변수로 고려된 보험료는 남자의 경우와 마찬가지로 보험료 부담수준이 높을수록 당뇨 발생 가능성은 더 적은 것으로 나타났다. 그러나 개인이 거주하는 거주지 특성은 유의하지 않은 것으로 나타났다. 총 투약일수는 앞의 전체모형에서와 같이 총투약일수가 많을수록 당뇨 발생가능성은 높은 것으로 나타났고, 그 수준은 극히 미미한 정도였다.

여자의 경우에 건강검진 및 문진의 결과로 나타나는 개인별 건강위험요인의 특성을 살펴보면, 먼저 콜레스테롤 수치가 질환의심(261mg/dl 이상)인 사람은 정상A(230mg/dl 이하)인 사람들에 비하여 당뇨 발생 상대위험도가 1.316배 높았다. 그리고 검진판정결과가 정상이나 비만관리가 요구되는 경우, 고혈압 질환이 있는 사람은 그렇지 않은 경우보다 각각 1.173배, 1.221배 당뇨 발생위험도가 높게 나타났다.

한편, 문진결과에서 나타난 개인별 생활습관 특성에서는 당뇨 가족력이 있거나, 흡연을 하는 경우는 당뇨 발생 상대적 위험도가 각각 1.369배, 1.339배 높은 것으로 나타났다.

(4) 2005년 이내 당뇨 발생 예측

본 연구에서 개발한 당뇨 발생 예측모형을 2003년 검진 또는 진료결과 당뇨가 없었던 자에 적용시켜, 향후 2년 뒤인 2005년 이내에 당뇨 발생 확률을 추정된 회귀모형식을 이용하여 개인별로 예측하였다.

<표 9> 당뇨 발생 예측모형(남자)

특성 요인			추정회귀계수	상대위험도
절편			-0.8225	
인구사회학적 특성	연령그룹	40~49세		1.000
		50~59세	0.0744***	1.190
		60세 이상	0.0253	1.133
	보험료수준	상위 25%미만	0.0526***	1.000
		상위 25~50%미만	-0.0631***	0.856
		상위 50~75%미만	-0.1980***	0.762
		상위 75%이상		0.666
	거주지역	농어촌		1.000
		중소도시	-0.00190	1.061
		대도시	0.0633***	1.133
진료이용량	총투약일수		0.000515***	1.001
건강 위험 요인	검진 결과	혈압	정상혈압	1.000
			경계역혈압	0.0465***	1.107
			고혈압	0.00846	1.065
		콜레스테롤	정상A(230이하)		1.000
			정상B(231~260)	-0.00225	1.081
			질환의심(261이상)	0.0826*****	1.177
		복합요인 D (글루코스 : mg/dl)	정상A, 기타		1.000
			정상B, 정상체중	0.1061***	2.045
			정상B, 위험체중	0.1906***	2.226
			정상B, 비만	0.3128***	2.515
		판정관리정상B (당뇨관리)	필요없음		1.000
			필요있음	0.1464***	1.340
		고혈압질환	없음		1.000
			있음	0.0888***	1.194
		복합요인 E (흡연관리)	흡연관리양호, 기 타		1.000
			흡연관리개선, 비만관리양호	-0.0182	1.071
			흡연관리개선, 비만관리개선	0.1048***	1.211
	문진 결과	타질환 과거력	없음		1.000
			있음	0.1122***	1.252
		당뇨 가족력	없음		1.000
			있음	0.1799***	1.433
		음주량		0.000934	1.001
		흡연량		0.00351***	1.004

주 : 1) 보험료 상위 25% : 21,670원, 상위 50% : 41,960원, 상위 75% : 69,930원
2) 비만도 : 정상체중(23미만), 위험체중(23~25미만), 비만(25이상)
3) 혈압 : 정상혈압(수축기혈압 정상A(139이하)이고 이완기혈압 정상A(89이하)), 경계역 혈압(수축기혈압 정상B(140~159이하)이거나 이완기혈압 정상B(90~94이하))
4) 복합요인 D, E는 2개의 건강위험요인을 고려한 변수(D는 글루코스와 비만; E는 흡연관리개선 여부와 비만)
5) 유의확률 * : P<0.1, ** : p<0.05, *** : p<0.01

<표 10> 당뇨 발생 예측모형(여자)

특성요인				추정회귀계수	상대위험도
		절편		-1.1673***	
인구사회학적 특성		연령그룹	40~49세		1.000
			50~59세	0.0404	1.371
			60세 이상	0.2346***	1.665
		보험료수준	상위 25%미만		1.000
			상위 25~50%미만	0.0115	0.941
			상위 50~75%미만	0.0288	0.957
			상위 75%이상	-0.1131*	0.830
진료이용량		총투약일수		0.000488***	1.000
건강위험요인	검진결과	혈압 (mmHg)	정상혈압		1.000
			경계역혈압	0.0450*	1.141
			고혈압	0.0423	1.138
		콜레스테롤 (mg/dl)	정상A(230이하)		1.000
			정상B(231~260)	-0.0206	1.112
			질환의심(261이상)	0.1476***	1.316
		복합요인 D (글루코스 : mg/dl)	정상A, 기타		1.000
			정상B, 정상체중	0.0306	2.241
			정상B, 위험체중	0.2753***	2.863
			정상B, 비만	0.4706***	3.480
		관정관리정상B (비만관리)	필요없음		1.000
			필요있음	0.0796***	1.173
		당뇨질환	없음		1.000
			있음	0.1000***	1.221
	문진결과	당뇨가족력	없음		1.000
			있음	0.1569***	1.369
		흡연	무		1.000
			유	0.1461**	1.339

주 : 1) 보험료 상위 25% : 21,670원, 상위 50% : 41,960원, 상위 75% : 69,930원
2) 비만도 : 정상체중(23미만), 위험체중(23~25미만), 비만(25이상)
3) 혈압 : 정상혈압(수축기혈압 정상A(139이하)이고 이완기혈압 정상A(89이하)), 경계역 혈압(수축기혈압 정상B(140~159이하)이거나 이완기혈압 정상B(90~94이하))
4) 복합요인 D는 2개의 건강위험요인을 고려한 변수(D는 글루코스와 비만요인)
5) 유의확률 * : P<0.1, ** : p<0.05, *** : p<0.01

a) 추정된 예측모형식 : 전체

□□□□□□□= -0.9837 - 0.2136×성별(남자) + 0.0594×연령(50대) + 0.1544×연령(60대이상) - 0.0111×거주지(중소도) + 0.0506×거주지(대도시) + 0.000507×총투약일수 + 0.0212×수축기혈압(정상B) + 0.0605×수축기혈압(질환의심) - 0.0140×콜레스테롤(정상B) + 0.1059×콜레스테롤(질환의심) - 0.0563×비만도(위험체중) + 0.1288×비만도(비만) + 0.1246×복합요인D(정상, 정상체중) + 0.2769×복합요인D(정상, 위험체중) + 0.2351×복합요인D(정상, 비만) + 0.1401×판정결과정상B(당뇨관리) + 0.0774×고혈압질환자 + 0.1619×당뇨가족력(있음) + 0.0410×복합요인C(남자, 주 1회 이상) + 0.0608×육식 + 0.0879×흡연

b) 추정된 예측모형식(남자)

□□□□□□□ = -0.8225 + 0.0744×연령(50대) + 0.0253×연령(60대이상) + 0.0526×보험료수준(상위25%-50%미만) - 0.0631×보험료수준(상위50%-75%미만) - 0.1980×보험료수준(상위75 이상) - 0.0019×거주지(중소도시) + 0.0633×거주지(대도시) + 0.000515×총투약일수 + 0.0465×혈압(경계역혈압) + 0.00846×혈압(고혈압) - 0.0025×콜레스테롤(정상B) + 0.0826×콜레스테롤(질환의심) + 0.1061×복합요인D(정상, 정상체중) + 0.1906×복합요인D(정상, 위험체중) + 0.3128×복합요인D(정상, 비만) + 0.1464×판정결과정상B(당뇨관리) + 0.0888×고혈압질환유 - 0.0182×복합요인E(흡연관리개선, 비만관리양호) + 0.1048×복합요인E(흡연관리개선, 비만관리개선) + 0.1122×타질환 가족력(유) + 0.1799×당뇨 가족력(유) + 0.000934×음주량 + 0.00351×흡연량

c) 추정된 예측모형식(여자)

□□□□□□□=-1.1673 +0.0404×연령(50대) + 0.2364×연령(60대이상) + 0.0115×보험료수준(상위25%-50%미만) + 0.0288×보험료수준(상위50%-75%미만) - 0.1131×보험료수준(상위 75% 이상) + 0.000488×투약일수 + 0.0450×혈압(경계역혈압) + 0.0423×혈압(고혈압) - 0.0206×콜레스테롤(정상B) + 0.1476×콜레스테롤(질환의심) + 0.0306×복합요인D + 0.2573×복합요인D + 0.4706×복합요인D + 0.0706×복합요인D + 0.1×고혈압질환(유) + 0.1569×당뇨 가족력(유) + 0.1461×흡연(유)

추정된 예측모형식을 이용하여 2003년 검진 또는 진료를 통해서 당뇨질환이 없었던 자에 적용시킨 결과, 향후 2년 뒤인 2005년 이내에 당뇨 발생위험도가 높은 고위험군 집단(상위 10%)과 저위험군 집단(하위 10%)의 인구학적 특성, 진료 이용량 및 건강위험요소 등을 비교·분석하였다(표 11). 당뇨 고위험군과 저위험군의 특성 차이를 보면, 당뇨 발생위험도가 큰 상위 10%인 고위험군 집단에서는 남성이 76.68%로 여성(23.32%)보다 월등히 많이 분포하고

<표 11> 당뇨 발생 예측모형을 통한 고위험군과 저위험군의 특성 비교

특성			고위험군 (상위 10%)		저위험군 (하위10%)		계	
			N	(%)	N	(%)	N	(%)
인구사회학적 특성	성별	남	201,432	(76.68)	42,156	(16.04)	1,379,386	(52.51)
		여	61,268	(23.32)	220,705	(83.96)	1,247,583	(47.49)
	나이그룹	40~49세	66,372	(25.27)	166,445	(63.32)	1,231,196	(46.87)
		50~59세	89,689	(34.14)	65,564	(24.94)	717,183	(27.30)
		60세 이상	106,639	(40.59)	30,852	(11.74)	678,590	(25.83)
	지역특성	대도시	127,712	(48.62)	97,784	(37.20)	1,175,656	(44.75)
		중소도시	97,454	(37.10)	119,334	(45.40)	1,071,402	(40.78)
		농어촌	37,534	(14.29)	45,743	(17.40)	379,907	(14.46)
경제적 수준	보험료수준	25%이하	62,464	(23.78)	58,485	(22.25)	644,989	(24.55)
		26~50%이하	62,904	(23.95)	62,844	(23.91)	650,629	(24.77)
		51~75%이하	68,391	(26.03)	61,987	(23.58)	657,106	(25.01)
		76%이상	68,941	(26.24)	79,545	(30.26)	674,245	(25.67)
진료이용량	총진료비	25%이하	94,273	(35.89)	45,715	(17.39)	656,737	(25.00)
		26~50%이하	67,815	(25.81)	69,702	(26.52)	656,734	(25.00)
		51~75%이하	50,806	(19.34)	76,704	(29.18)	656,734	(25.00)
		76%이상	49,806	(18.96)	70,740	(26.91)	656,764	(25.00)
	총내원일수	25%이하	83,930	(31.95)	51,640	(19.65)	642,095	(24.44)
		26~50%이하	71,708	(27.30)	68,609	(26.10)	659,616	(25.11)
		51~75%이하	55,979	(21.31)	73,586	(27.99)	664,969	(25.31)
		76%이상	51,083	(19.45)	69,206	(26.26)	660,289	(25.14)
	총투약일수	25%이하	113,829	(43.33)	31,906	(12.14)	654,597	(24.92)
		26~50%이하	54,061	(20.58)	79,502	(30.24)	658,027	(25.05)
		51~75%이하	32,350	(12.31)	56,835	(21.62)	457,481	(17.41)
		76%이상	62,460	(23.78)	94,618	(36.00)	856,864	(32.62)
			262,700	(100.0)	262,861	(100.0)	2,626,969	(100.0)

주 : ( )내의 값은 전체 N에 대한 구성비

(계속)

특성			고위험군 (상위 10%)		저위험군 (하위10%)		계	
			N	(%)	N	(%)	N	(%)
건강 위험 요인	검진 결과	비만도 (kg/ m^2) 정상	3,424	(1.30)	0	(0.00)	69,467	(2.64)
		위험체중	44,569	(16.97)	0	(0.00)	957,675	(36.46)
		비만	214,707	(81.73)	262,861	(100.0)	1,599,827	(60.90)
		최고혈압 (mmHg) 정상A	139,138	(52.96)	228,074	(86.77)	1,962,731	(74.74)
		정상B	87,496	(33.31)	30,958	(11.78)	516,433	(19.66)
		질환의심	36,066	(13.73)	3,829	(1.46)	147,805	(5.63)
		최저혈압 (mmHg) 정상A	152,704	(58.13)	219,063	(83.34)	1,975,408	(75.20)
		정상B	62,371	(23.74)	36,559	(13.91)	441,209	(16.80)
		질환의심	47,625	(18.13)	7,239	(2.75)	210,352	(8.01)
		글루코스 (mg/dl) 정상A	101,046	(38.46)	262,861	(100.0)	2457,232	(93.54)
		정상B	161,654	(61.54)	0	(0.00)	169,737	(6.46)
		콜레스테롤 (mg/dl) 정상A	191,184	(72.78)	226,068	(86.00)	2149,966	(81.84)
		정상B	45,818	(17.44)	30,778	(11.71)	339,320	(12.92)
		질환의심	25,698	(9.78)	6,015	(2.29)	137,683	(5.24)
	문진 결과	당뇨가족력 없다	214,697	(90.76)	230,196	(96.84)	2,237,812	(94.51)
		있다	21,857	(9.24)	7,514	(3.16)	129,995	(5.49)
		음주습관 월 2~3회	113,757	(44.04)	6192,129	(74.54)	1478,654	(57.28)
		주1~2회	36,853	(14.27)	40,270	(15.62)	401,156	(15.54)
		주3~4회	58,448	(22.63)	18,731	(7.27)	422,269	(16.36)
		주5~6회	29,482	(11.41)	4,385	(1.70)	172,815	(6.69)
		거의 매일	19,756	(7.65)	2,238	(0.87)	106,763	(4.14)
전체			262,700	(100.0)	262,861	(100.0)	2,626,969	(100.0)

주 : ( )내의 값은 전체 N에 대한 구성비

(계속)

특성			고위험군 (상위 10%)		저위험군 (하위10%)		계	
			N	(%)	N	(%)	N	(%)
건강 위험 요인	문진 결과	음주량	소주 반병 이하	54,804	(36.76)	48,427	(65.96)	531,777 (45.47)
			소주 한병	65,662	(44.05)	19,181	(26.12)	455,442 (38.94)
			소주 1병 반	18,052	(12.11)	4,052	(5.52)	122,290 (10.46)
			소주 2병 이상	10,558	(7.08)	1,754	(2.39)	59,933 (5.12)
		흡연	비흡연	138,137	(53.59)	238,176	(93.28)	1730,691 (67.36)
			현재금연	32,422	(12.58)	14,318	(5.61)	243,815 (9.49)
			흡연	87,222	(33.84)	2,842	(1.11)	594,750 (23.15)
		흡연량	반갑미만	28,427	(27.71)	6,111	(56.70)	210,808 (29.52)
			반갑이상~한갑미만	52,266	(50.95)	3,546	(32.90)	360,715 (50.50)
			한갑이상~두갑미만	19,935	(19.43)	942	(8.74)	131,400 (18.40)
			두갑이상	1,936	(1.89)	148	(1.37)	10,999 (1.54)
		흡연기간	5년 미만	6,944	(5.94)	3,810	(22.72)	59,341 (7.23)
			5~9년	7,475	(6.39)	3,901	(23.26)	66,864 (8.15)
			10~19년	26,398	(22.56)	5,788	(34.51)	229,634 (27.98)
			20~29년	34,832	(29.77)	2,860	(17.05)	262,971 (32.05)
			30년 이상	41,350	(35.34)	411	(2.45)	201,757 (24.59)
		운동	안한다	137,716	(53.70)	160,419	(62.89)	1,471,018 (57.49)
			1~2회	64,566	(25.17)	50,756	(19.90)	605,320 (23.66)
			3~4회	25,864	(10.08)	21,819	(8.55)	246,651 (9.64)
			5~6회	7,131	(2.78)	5,987	(2.35)	65,793 (2.57)
			거의 매일	21,193	(8.26)	16,112	(6.32)	170,032 (6.64)
		운동시간	30분미만	19,699	(7.68)	23,858	(9.29)	214,756 (8.36)
			30분이상~1시간미만	75,732	(29.52)	85,512	(33.28)	830,198 (32.33)
			1시간이상~2시간미만	134,254	(52.33)	122,657	(47.74)	1,263,321 (49.19)
			2시간이상	26,850	(10.47)	24,877	(9.68)	259,724 (10.11)
전체				262,700	(100.0)	262,861	(100.0)	2,626,969 (100.0)

주 : ( )내의 값은 전체 N에 대한 구성비

있는 반면, 하위 10%인 저위험군 집단에서는 그 반대인 여성(83.96%)이 남성(16.04%)보다 월등히 많은 것으로 나타났다. 이러한 특성은 남성이 여성보다 당뇨 발생 가능성이 상대적으로 높음을 말해준다. 연령대별로는 고위험군내에 고연령층이 저연령층보다 많이 분포하고 있고, 저위험군내에서는 저연령층이 집중되어있는 것으로 나타났다. 또한 고위험군의 35.89%가 전체 총진료비 분포의 25%수준에 해당되는 반면, 저위험군에서는 17.39%가 전체 총진료비 분포의 25%수준인 것으로 보아 고위험군이 저위험군보다 의료서비스를 덜 받고 있음을 알 수 있다.

한편 건강위험척도인 비만도, 혈압, 글루코스, 콜레스테롤 관련하여 그 수치가 높아 건강관리를 요하는 사람들의 분포는 저위험군보다 고위험군에 보다 밀집되어 있었다. 건강행위에 대해서는 고위험군은 저위험군보다 음주는 주 3회 이상(고위험군 : 41.69%, 저위험군 : 9.84%), 1회 음주량은 소주 한병 이상(고위험군 : 63.24%, 저위험군 : 34.03%), 과거 및 현재 흡연자(고위험군 : 46.42%, 저위험군 : 6.72%) 등 좋지 못한 건강행위습관을 일정범주이상 행하는 집단인 것으로 나타났다.

V. 고 찰

본 연구는 향후 건강증진사업의 지식기반시스템구축의 일환으로 국민건강보험공단의 건강검진자료 및 급여정보를 활용하여, 사전예방관리 측면에서의 당뇨 발생 예측모형을 데이터마이닝 기법을 적용하여 개발하였다. 성별에 따라 생활습관 및 위험요인의 특성이 다소 차이가 있음을 고려하여 당뇨 발생 예측모형은 남자·여자 그리고 전체로 구분하여 개발하였다. 모형개발은 데이터마이닝 프로세스에 의하였고, 최종 예측모형은 데이터마이닝의 로지스틱 회귀모형을 적용하였다. 그 결과 최종예측모형의 예측 향상력은 임의의 모형(Random Model)보다 발생 확률분포의 상위 10%에서 전체 2.36(남자 : 2.25, 여자 : 2.47)배정도 향상된 것으로 평가되었다(표 6, 표 7). 이는 관리대상군의 특성을 고려한 관리접근방식을 시스템화하므로서, 관리 인원 대비 관리의 효율성을 높일 수 있는 가능성을 시사하는 바이다.

당뇨 발생 예측모형으로부터 나타난 일반대상자의 당뇨 발생 가능 요인은 성별, 연령, 거주지, 개인별 혈압, 콜레스테롤, 글루코스, 비만도(BMI), 흡연량, 음주량, 운동량 등으로 나타났다. 이러한 요인들 중 여자들보다는 남자들이 질환의 발생위험도가 상대적으로 높은 것으로 나타났으며, 연령에서는 40대를 기준으로 50대 남자의 경우는 60대의 남자보다 그리고 여자의 경우는 60대가 50대보다 당뇨 발생 위험도가 높은 것으로 나타났다. 거주 지역별로는 대도시 사람들이 중소도시 사람들보다 상대적으로 당뇨 발생 위험도가 높은 것으로 나타났다.

또한 개인별 건강위험요인에 따라 질환의 발생위험도는 남녀 모두 혈압이 경계역에 있는

경우가 정상혈압에 있는 사람들보다 각각 1.107배, 1.141배 높게 나타났으며, 콜레스테롤은 여자 질환의심자인 경우가 남자 질환의심자의 경우보다 발생 위험도가 높게 나타났다. 그리고 비만인 사람의 경우 글루코스가 정상B에 해당하면, 저체중이고 글루코스가 정상A인 사람보다 당뇨 발생 위험도가 약 2.5배 높게 나타났으며, 여자의 경우는 약 3.5배가량 높은 것으로 나타났다. 이는 우리나라 울산광역시 중구 주민을 대상으로 강성홍 등(2004)이 개발한 당뇨 발생 예측모형의 결과와 유사하였다. 강성홍 등은 당뇨 발생 가능요인으로 개인별 생활습관특성을 고려하지 않았지만, 본 연구에서는 고혈압 질환이 있고, 당뇨 관련 가족력이 있고, 육식, 흡연 특성을 고려한 경우, 그렇지 않은 사람들보다 각각 1.167(남자 1.194, 여자 1.221)배, 1.382(남자 1.433, 여자 1.369)배, 1.129배, 1.192배 정도가 당뇨로 질환이 발생할 상대적 위험도가 높다는 것을 알 수 있었다.

한편, 당뇨 발생 모형을 이용한 고위험군에 속하는 대상자를 선별하여 집중관리 시 예상되는 진료비 효과[5]는 최소 약 7억원 정도로 나타났다(표 12). 이러한 차이는 건강증진사업을 추진함에 있어 효과적인 대상자 선정에 의한 관리의 효율성이 사업의 목표 달성에 크게 영향을 줄 수 있음을 말해준다.

<표 12> 당뇨 발생 가능자 사전관리에 따른 진료비 절감

		관리 성공률(상위10%)			Random 관리
		100%	80%	60%	100%
당뇨	관리대상자	6,595명	5,276명	3,957명	2,797명
	진료비 절감액	13억원	10억원	7억원	5억원

주 : 1) 전체 374,550명 중 당뇨 유병자는 27,978명(7.47%)이고 당뇨발생확률분포의 상위 10% 37,455명에 해당되는 모델 적중률(Test Data set)은 17.61%(표 6참조).

2) 1인당 당뇨 평균연간 진료비 : 199,106원 (연간 총당뇨진료비/당뇨 환자수, 2001년 건강보험통계연보)

3) 6,595명=당뇨발생확률분포의 상위 10% 사전관리대상자 x 적중률 = 37,455명 x 17.61%

박경수(2002) 연구를 보면, 세계보건기구에서도 당뇨병의 예방의 중요성을 강조하면서 국가단위의 정책과 프로그램을 수립하여 당뇨병의 역학조사, 지역사회에서의 당뇨병 일차예방 프로그램개발, 이미 발병한 당뇨병과 합병증의 이차적 예방 프로그램 개발, 당뇨병의 예방과 관리를 위한 인적자원의 양성, 당뇨병의 연구 및 정보교류의 강화 등의 전략을 세워야 한다

5) 여기서 말하는 진료비 절감액이란 2001년 건강한 사람이 2년 이내에 당뇨 발생 가능성이 높은 상위 10%를 사전관리하므로서, 당뇨 발생 후에 지출되는 연간 총진료비를 의미함.

고 권장하고 있다. 이러한 점에서 개인별 맞춤형 건강 및 의료 정보를 시스템적으로 지원해 줄 수 있는 건강관리시스템 및 발생 예측시스템 등은 건강증진사업에 있어서 중요한 의미가 있다.

Ⅵ. 결 론

건강증진사업은 단순히 질병을 치료하고 예방하는 것에만 국한된 것이 아니고 적극적인 건강향상을 목적으로 사람들의 건강의식이나 행동변화를 유도하여 건강할 수 있는 잠재력을 기르고 건강위험요인을 조기에 발견함으로써 건강을 유지·증진하고자 하는 것이 주요 목적이다. 이러한 관점에서 본 연구는 만성질환 중 당뇨 발생의 중요한 위험요인들을 근거로 사전예방을 위한 예측모형을 개발하였다.

본 연구의 당뇨 발생 예측모형은 효율적인 건강증진사업을 위한 몇 가지 업무 프로세스에 활용될 수 있다. 첫째, 발생 예측분포의 사후확률에 기초한 관리대상군의 특성을 파악하고, 이를 유형화 할 수 있다. 당뇨 발생위험도가 큰 상위분포부터 고위험군을 5%, 10%, 15% 순으로 정하고, 관리군별 인구사회적 특성, 진료특성 및 건강위험요소 등을 비교분석하여 유형화 하는데 활용할 수 있다. 둘째, 관리대상군의 특성을 고려한 관리접근방식을 시스템화할 수 있다. 효율적인 관리를 위해서는 관리군의 특성을 고려한 관리인원 대비 관리의 효율성을 높이기 위한 관리접근방식을 차별화되어 관리될 필요가 있다.

셋째, 개인별 맞춤형 교육, 홍보 및 건강의료정보를 제공할 수 있다. 맞춤형태의 교육 및 홍보는 바로 국민의 만족수준을 향상시키고, 이는 국민이 필요로 하는 건강증진사업으로 발전할 수 있는 기반이 된다. 또한 당뇨 발생 예측모형과 같은 사전예방 건강관리모형의 지속적인 개발은 향후 대국민 맞춤형 건강정보제공을 위한 정보인프라구축에 크게 기여할 것이다.

넷째, 건강검진 수검 독려 시 활용가능하다. 건강검진제도가 실시된 이후 지금까지 지역가입자의 수검률은 직장 가입자들에 비해 현격하게 낮은 수준을 보이고 있어, 앞으로 건강보험공단이 검진사업을 추진함에 있어 건강검진 실시 및 계획 그리고 미수검 대상자들에 대한 수검독려대상자 선정기준 등에 활용가능하다.

다섯째, 합리적 의료이용지원을 위한 의료정보를 제공할 수 있다. 본 연구에서 개발된 질환 발생예측모형은 개인들이 자가 진단할 수 있는 프로그램으로 일반화 할 수 있다. 이는 대국민들이 스스로 건강관리를 적절히 할 수 있는 능력을 키워줄 수 있는 방안이며, 궁극적으로는 합리적 의료이용의 정착을 유도할 수 있다.

이러한 사전예방관리시스템으로부터 만성질환(당뇨, 고혈압 등)으로 발생 확률이 높은 고

위험군을 선별하고, 또한 개인별 특성을 고려한 예방관리사업을 할 수 있다. 즉, 건강검진 대상자 중에서 고위험군에 속하는 대상자를 선별하여 반드시 건강검진을 받도록 유도하고, 사전에 질환을 관리하여 예방할 수 있도록 함이다. 이는 단기적으로는 건강검진의 수검률 향상을 가져올 수 있으며, 장기적으로는 현재 의료계가 지향하는 평생건강관리체계의 기틀을 마련할 수 있을 것이라 사료된다.

다만 본 연구는 건강보험공단의 건강검진 정보를 근거로 건강관리모형을 개발함에 따라 분석대상을 40세 이상인 사람으로 제한할 수밖에 없는 한계점이 있으며, 또한 개인별 코호트 자료를 이용하여 진단 이후 2년 이내에 질환의 발생위험도를 예측하는 모형개발을 목적으로 하였기 때문에 매년 건강검진을 받은 대상자만을 이용하였다. 따라서 본 연구에서 개발한 건강관리모형은 연령대별 발생 가능한 개인별 건강위험요인의 특성을 반영한 것이 아니므로, 일반적인 건강관리모형으로 사용하기에는 다소 주의가 필요하다.

끝으로 본 연구에서 개발한 건강관리모형은 임상자료를 활용하지 못해 임상적인 신뢰성이 떨어지는 것에 대해 한계성이 있으며, 추후 의료기관의 EMR(Electronic Medical Record)이 활성화가 되고, 공공기관의 EHR(Electronic Health Record)시스템이 도입이 현실화되면 이 두 시스템간의 자료연계를 통해 보다 신뢰성이 있는 건강관리모형이 구축될 수 있으리라 사료된다.

또한, 본 연구에서 개발된 건강관리모형을 실제 현업에 적용하지 못한 아쉬움이 있으나, 임상정보를 제외한 국민건강보험공단이 보유하고 있는 건강검진 및 문진, 급여 그리고 자격정보를 활용하여, 최근 신정보기술(New Information Technology)로서 각광받고 있는 데이터마이닝 기법으로 건강증진사업의 목적에 부합하는 프로세스를 제안하였다는 데에 큰 의의가 있다 하겠다.

참 고 문 헌

강성홍, 구방본, 김병철 외. 병원경영 정보관리. 고려의학, 2002.

강성홍, 용왕식 외. 건강보험공단의 건강증진시스템 개발, 인제대학교, 2003.

강성홍, 최순호. 데이터마이닝을 이용한 보건소의 건강증진사업의 효율화 방안. 대한의료정보학회지, 제7권, 제 2호, pp.37~48, 2001

강현철. SAS Enterprise Miner를 이용한 데이터마이닝. 자유아카데미, 1999

강현철, 한상태, 최종후, 김은석, 김미경. SAS Enterprise Miner를 이용한 데이터마이닝 - 방법론 및 활용 - . 자유아카데미, 2001.

김영식. 건강생활습관과 만성질환 : 고혈압과 당뇨병의 발병요인 규명을 위한 코호트 연구를

중심으로, , 대한지역사회영향학회 추계학술지, 2003

박경수. 유전체 연구와 당뇨병의 예방, 제30차 종합예술대회, 2002

송미숙. 지역건강담당제 추진을 위한 건강증진 정보관리시스템 개발, 보건복지부 건강증진기금 보고서, 2001.

송태민 외. 지식기반 건강보험정보 데이터베이스 구축 및 활용방안 연구, 한국보건사회연구원, 2002.

이주원. 뇌졸중의 진단명 예측을 위한 전문가시스템 설계, 대한의료정보학회지, 제7권, 제 1호, pp.77~82, 2001

이주열. 평생국민건강관리체계 구축방안에 관한 연구, 건강증진기금보고서, 2000.

용왕식 외. 데이터마이밍을 활용한 의료보험료 부과체계개발, 대한의료정보학회지, 2001

장남식, 홍성완, 장재호. 데이터마이닝 : 성공적인 지식경영을 위한 핵심정보기술, 대청, 2002.

장동인. 실무자를 위한 데이터웨어하우스. 대청, 1999

지선하. 건강증진사업의 경제적 효과와 건강보험 재정. 2003년 한국건강증진학회 추계학술대회, 2003:15~23

정현순. 건강보험 진료비 삭감예방을 위한 사전심사 의사결정시스템 구축. 연세대학교 보건대학원, 석사학위논문, 2002.

최순호 외. 건강증진사업을 위한 CRM 시스템 개발. 울산광역시 중구보건소, 2003

최길림. 병원이용빈도와 진료수익성에 따른 환자군집별 특성과 데이터베이스 마케팅의 활용성. 인제대학교, 박사학위논문, 2001

채준호 외. 데이터마이닝 학술지, 2003

호승희. 데이터마이닝 기법을 활용한 고혈압 관리를 위한 의사결정지원시스템의 개발 및 평가. 아주대학교, 의학과, 2000

홍두호 외. 데이터마이닝 기법을 이용한 DRG 확인심사 대상건 검색방법. 예방의학학회지, 제26권, 제2호, 2003.

2003년도 건강검진 결과분석, 2004, 건강보험공단

건강보험공단 중장기 정보화발전계획 수립 보고서, 2002, LG CNS컨소시엄.

보건복지부. 건강보험 요양기관 부정청구 감시를 위한 Data Mining 기법 적용방안 보도자료, 2004

Dogu Celebi. The Power of Predictive Modeling. Healthcare Informatics, 2003, Vol. 8, pp.56~58

Diana L. Dalley, et al. The Impact of a Health Education Program Targeting Patients with High visit rates in a Managed Care Organization. Am J Health Promot, 2002, Vol 17,

No 2. pp. 101-111

Giudici, P. Applied Data Mining - Statistical Methods for Business and Industory, Wiley, 2003.

Judith H. Hibbard, Ellen Peters. Supporting Informed Customer Health Care Decisions : Data Presentation Approaches the Facilitate the use of information in choice. Annual Review of Public Health, 2003, Vol 24, pp. 413~433

John M. Wilkinson, Paul V. Targonski. Health Promotion in a Changing World : Preparing for the Genomics Revolution. Am J Health Promotion, 2003 Nov-Dec; Vol. 18, No. 2, pp.157~161

Joseph, B., Steven R. S. "Hospital Shopping and Consumer Choice", Journal of Health care Marketing, 1982, Vol 2, No 2.. pp.15~23

Luftman, J. N., Lewis, P. R and Oldach, S. H. "Transforming the Enterprise : The Alignment of Business and Information Technology Strategies", IBM Systems Journal, 1993, Vol. 32, No. 1, pp. 198-221.

Taiki, S., Charles, H., Michael j. O'shea. "Case Study : How to Apply Data Mining Techniques un a healthcare Data Warehouse", Journal of Healthcare Information Management, 2001, Vol. 15, No. 2, pp.155~164

The Promise of Prevention: Reducing health and economic burden of chronic disease. CDC, 2003

Sun Ha Jee, et al. The relationship between modifiable health risks and future medical expenditures: The Korean Medical Insurance Corporation(KMIC) Study . American Journal of Health, 2001, Vol. 15, No. 4. pp.244~255

www.cpm.com