保健敎育 · 健康增進學會誌 第23卷 第4號(2006. 12) pp.13-28
Journal of Korean Society for Health Education and Promotion, Vol.23, No.4(2006)

고혈압 발생 예측 모형 개발

용왕식* · 박일수* · 강성홍** · 김원중** · 김공현*** · 김광기*** · 박노례****
*국민건강보험공단 · **인제대학교 보건행정학부 · ***인제대학원대학교 · ****인제대학교 보건대학원

〈목 차〉

Ⅰ. 서론

1. 연구의 배경 및 필요성

고혈압은 '소리 없는 살인마(Silent Killer)'로 불릴 정도로 관리를 하지 않을 경우 합병증을 유발하여 중증 질환으로 이행될 가능성이 매우 높은 질환이다. 우리나라는 '2005년 국민건강 · 영양조사' 결과 30세 이상 고혈압의 유병율이 남자가 30.2%, 여자가 25.6%였으며, 김영식 등(2001)의 코호트 연구 결과, 고혈압의 조발생률이 남자에서 1,000명당 1.7명, 여자는 1.29명이었다. 그러나 이 중 의료기관에서 실제로 진단을 받은 사람은 절반에도 미치지 못하며, 진단을 받은 후 치료를 받은 사람도 역시 절반에 미치지 못한다고 추정되고 있다. 또한, 고혈압은 심혈관 질환과 뇌졸중 등 순환기계 질환의 주요 위험요인이다(서일 등, 1997; Beberg 등, 2000). 특히 신생물에 이어 우리나라 전체 사망원인의 25.4%를 차지하고 있는 순환기계통 질환의 주요 원인이 된다(통계청, 2004). 또한 2004년도 건강보험의 질병 소분류별 다발생 순위별 요양급여실적에서 외래의 경우 본태성 고혈압이 1위를 차지하였으며, 요양급여비용이 304억원에 달했다(건강보험심사평가원, 2005). 따라서 고혈압 관리는 개인과 사회, 국가차원에서 중요한 보건문제라고 할 수 있다.

고혈압관리사업은 크게 2가지로 분류할 수 있다. 고혈압의 예방사업과 고혈압의 치료사업으로 분류할 수 있다. 고혈압의 예방사업은 사업대상자가 광범위하므로 고위험군을 선정하여 이들에게 고혈압에 대한 건강위험평가(Health Risk Appraisal, HRA) 정보를 제공하면 보다

교신저자: 용왕식
서울특별시 마포구 염리동 168-9 국민건강보험공단 (우: 121-749)
전화번호: 011-9784-9689, E-mail: youngws1306@hanmail.net

효율적인 고혈압 예방사업이 될 것이다. 즉, 고혈압환자가 아닌 자들에 대해 개인별로 고혈압 발생위험 확률을 예측하여, 이 정보에 근거하여 고혈압 예방사업을 실시할 경우 사업의 효율성이 높아 질 것으로 예측된다.

고혈압 발생위험은 유전자 형태와 생활습관에 의해 주로 영향을 받는데, 아직까지 유전자 형태와 고혈압 발생의 위험확률에 대해서는 실제 보건사업에 사용할 수 있을 정도로 연구가 진행되어 있지 못하기 때문에 현재는 생활습관을 중심으로 고혈압의 발생확률을 예측하여야 할 것이다.

대표적인 생활습관과 고혈압과의 관계를 밝힌 외국의 연구들에 의하면 과체중은 혈압 상승과 관련되어 있으며(Pauliot 등, 1994), 과체중인 고혈압 환자에서 체중감소는 항고혈압제의 혈압강하효과를 향상시켰다(Neaton 등, 1993). 또한 과다한 음주는 고혈압의 중요한 발병요인이며(Stamler 등, 1997), 고혈압 환자에서 항고혈압제에 대한 내성을 유발하였다(Puddey 등, 1992). 신체활동에 있어서는 규칙적인 신체활동이 심혈관계 질환 및 사망률을 감소시키며(Paffenbarger 등, 1993), 좌식생활을 주로 하는 사람들이 활동적인 사람들에 비해 고혈압 발병위험이 20~50% 정도 높았다(Blair 등, 1984). 또한 염분섭취에 대해서는 여러 역학 자료들이 염분섭취와 혈압 사이의 양의 상관관계를 입증하고 있다(Elliott 등, 1996). 여러 임상시험들의 메타분석 결과 일일 75~100mmol 정도의 염분섭취량 감소가 수년에 걸쳐 혈압을 감소시켰다고 하였다(Cutler 등, 1997). 흡연에 관해서는 흡연 직후에 혈압이 오르지만 대부분의 연구에서 흡연자의 혈압은 비흡연자와 비슷하거나 더 낮았다(Friedman 등, 1982).

고혈압과 생활습관 요인에 관한 국내 연구에서는 체중증가에 따라 혈압이 증가된다는 단면연구가 다수이며(문일순 등, 1989; 이강숙 등, 1994; 전종민 등, 1997), 환자대조군 연구에서 과체중인 사람이 고혈압 발생 비교위험도가 3.5배 였다는 결과가 있다(맹광호 등, 1982). 또한 하루에 30g 이상의 알코올 섭취가 고혈압의 유병율을 증가시켰으며(전종민 등, 1997), 음주자의 고혈압 발생 비교위험도가 1.4배였다(Kim 등, 1991). 운동습관에서는 무운동군에서 운동군보다 고혈압 유병율이 더 낮으며(전종민 등, 1997), 운동량과 고혈압과 관련이 없다는 보고가 있다(Kim 등, 1991). 염분섭취량과 고혈압의 관련성에 관한 논문들의 약 50%만이 통계적으로 유의한 결과를 보였다(유재희 등, 1994). 흡연에서는 남자 고혈압 환자의 65%가 흡연을 하였으나(최영환 등, 1996), 비흡연군이 흡연군보다 고혈압이 많거나 차이가 없다는 연구도 있었다(전종민 등, 1997; Kim 등, 1991).

그 동안 수행되었던 고혈압 발생에 관한 연구들이 외국에서 수행되어 우리나라 국민들의 특성을 잘 반영하지 못하고 있으며, 우리나라에서 수행된 연구들은 유병률을 조사하면서 부차적으로 고혈압 환자에서 그 당시에 가지고 있던 요인을 분석한 단면연구가 대부분이어서 시간적인 선행성을 알 수 없어 발병요인으로서의 인과관계를 입증할 수 없다는 한계가 있다(김영식 등, 2001). 또한 연구자료에 있어서도 일개 의료기관을 이용한 환자들이나 일부 지역사회 주민들을 대상으로 하고 있어 자료의 대표성과 연구결과의 일반화에 한계가 있다고 할 수 있다.

국가차원에서 고혈압의 건강위험평가 정보에 근거한 고혈압 예방사업을 실시하기 위해서는

전국적인 규모의 자료를 이용하여 고혈압 발생을 예측할 수 있는 모형을 개발할 필요가 있다.

2. 연구의 목적

본 연구는 생활습관을 중심으로 고혈압의 발생위험 확률을 예측할 수 있는 모형을 개발하는 것이다. 이를 달성하기 위한 구체적인 연구목적은 다음과 같다.

첫째, 고혈압 발생위험 확률을 예측할 수 있는 모형을 개발·평가한다.

둘째, 개발된 모형에 근거하여 고혈압 발생의 요인을 규명한다.

셋째, 개발된 모형을 이용한 고혈압 예방사업 활용방안을 제시한다.

Ⅱ. 연구방법

1. 자료수집

1) 모형개발 및 내적타당도(internal validation) 평가를 위한 데이터

고혈압 발생의 예측 모형을 개발하기 위한 연구대상자는 2000년과 2002년에 건강검진을 모두 받은 사람 중에 1998년부터 1999년 사이에 고혈압 진료를 받은 적이 없으며, 2000년 건강검진에서 혈압이 정상이었던 사람으로 문진표에 고혈압의 병력이 없는 자를 대상으로 하였다. 연구대상자에 대해 국민건강보험공단의 건강검진자료, 건강보험급여자료 그리고 2003년 12월31일 기준의 자격자료[1]를 중심으로 자료를 수집하였고, 수집된 자료 중 인구사회학적 특성을 나타내는 연령, 성별, 자격, 직역, 거주지, 보험료와 임상학적 특성 및 건강행위 특성을 나타내는 신장, 체중, 비만도, 혈압수준, 식전혈당(글루코스)와 콜레스테롤 수준, 고혈압에 대한 가족력, 음주습관, 흡연습관, 운동습관 그리고 진료이용실적을 나타내는 2000년 한 해 동안의 고혈압과 관련된 진료이용을 제외한 진료이용량(투약일수, 입내원일수, 진료비)을 독립변수로 활용하였고, 2002년의 고혈압 판정유무[2]를 종속변수로 활용하였다.

2) 외적타당도(external validation)평가를 위한 데이터

고혈압 발생의 예측모형의 외적타당도(external validation)평가를 위해서 모형개발을 위한 데이터와는 시간적으로 다른 데이터를 수집하였다. 외적타당도 평가를 위한 데이터의 연구대상자는 2001년과 2003년에 건강검진을 모두 받은 사람들 중, 1998년부터 1999년 사이에 고혈압 진료를 받은 적이 없으며, 2001년 건강검진에서 혈압이 정상이었던 사람으로 문진표에 고혈압의 병력이 없는 자를 대상으로 하였다. 또한 이들은 1999년~2000년까지 고혈압 진료를 받은 적이 없는 사람이다. 그리고 데이터 수집항목은 모형개발을 위한 데이터와 동일한 항목으로 하였다.

2. 변수정의

1) 고혈압환자

연구대상자의 고혈압환자에 대한 판단은 건

1) 국민건강보험공단의 자격자료는 건강보험적용인구의 인구사회학적 특성을 나타냄.

2) 본 연구에서 종속변수로서의 고혈압 발생유무는 건강검진결과에 따름.

강검진과 문진자료에서 고혈압 질환의심자, 수축기혈압 160mmHg 이상, 이완기혈압 95mmHg 이상, 과거 고혈압 병력 중 하나라도 해당되는 경우로 하였으며, 고혈압 수치의 기준은 세계보건기구(World Health Organization, WHO)가 정하는 기준을 따랐다(표 1).

또한, 고혈압 환자에 대한 판단기준을 활용하여, 연구대상자들이 2002년 건강검진 결과 중 고혈압 판정유무를 종속변수로서 활용하였다.

<표 1> 고혈압 진단 기준

자료원	고혈압 진단 기준	근거
건강검진 · 문진 자료	○ 고혈압 의심자 ○ 수축기 160mmHg이상/이완기 95mmHg이상 ○ 과거 고혈압 병력	WHO (세계보건기구)

<표 2> 주요 독립변수의 특성

	변수	세부항목
인구사회학적 특성	연령	검진당시 나이
	성별	남, 여
	자격	직장가입자, 지역가입자, 직장피부양자, 지역세대원, 지역비피보세대주
	직역	지역, 공교, 직장
	거주지	대도시, 중소도시, 농어촌
	보험료	2003년 12월 기준, 경제적 수준을 나타내는 대리변수
임상, 건강행위특성	신장	㎝
	체중	㎏
	비만도	BMI : 정상(23미만), 과체중(23~25미만), 위험(25이상)
	혈압최고	수축기 혈압(mmHg) : 정상A(139이하), 정상B(140~159), 질환의심(160이상)
	혈압최저	이완기 혈압(mmHg) : 정상A(89이하), 정상B(90~94), 질환의심(95이상)
	식전혈당	글루코스(㎎/㎗) : 저혈당의심(70미만), 정상A(70~110), 정상B(111~120),질환의심(121이상)
	총콜레스테롤	㎎/㎗ : 정상A(230이하), 정상B(231~260), 질환의심(261이상)
	가족력	가족 중 고혈압 및 기타질환에 걸린 경험
	음주습관	비음주, 절주, 과음
	흡연습관	비흡연, 금연, 흡연
	운동습관	일주일에 운동 횟수
진료실적	투약일수	
	입내원일수	
	입내원진료비	

2) 독립변수의 측정

활용된 주요 독립변수는 표 2와 같으며, 이 연구의 주요 독립변수인 건강행위 특성으로 음주 관련 특성, 흡연관련 특성, 운동습관이 각각 독립적으로, 그리고 상호 교차작용(Interaction)의 조합으로 포함되었다.

3. 분석방법

1) 분석도구

통계분석은 SAS 9.1을 사용하였으며 데이터 마이닝 툴은 SAS사의 Enterprise Miner 4.3을 사용하였다.

2) 모형개발

고혈압 발생을 예측할 수 있는 모형의 개발은 로지스틱회귀분석, 의사결정나무분석, 앙상블 기법을 활용하였고 이를 평가·비교하여 이 중 가장 좋은 모형을 채택하였다. 앙상블(Ensemble) 기법은 여러 모형의 결과를 종합하여 하나의 예측치를 얻고자 하는 경우 사용되는 모형 결합 방법이다. 본 연구에서 사용된 앙상블 모형은 의사결정나무와 로지스틱 회귀모형(Logistic Regression)이 결합된 형태로 만들어졌다. 로지스틱 회귀분석은 단계적 추출방법(Stepwise)을 적용하여 개발되었으며, 의사결정나무분석(Decision Tree)은 CHAID (Chi-squared Automatic Interaction Detection)방법 중 카이제곱 통계량(Chi-Square statistic)의 P값을 활용하여 최적분리를 하고자 했다.

3) 모형평가

고혈압 발생 예측모형에 대한 평가는 내적타당도 검증과 외적타당도 검증을 실시하였다.

(1) 내적타당도(Internal validation)

모형생성을 위해 구축되었던 자료 중 60%는 모형개발용 분석용 자료(Train Data)로 활용하였고, 나머지 40%로서 모형평가 자료(Validation Data)로 활용하여 모형의 내적 타당도 검증을 실시하였다.

(2) 외적타당도(External validation)

외적타당도 검토는 모형개발용 데이터를 이용하여 개발된 모형을 외적타당도 평가를 위한 데이터에 적용시켜 고혈압 발생 확률값을 예측한 후 이들 자료를 2003년의 실제 고혈압 발생에 관한 자료와 비교하여 개발된 모형의 외적 타당도를 검토하였다.

(3) 모형평가지표

개발된 모형의 내적, 외적 타당도 평가와 최종 선정된 모형의 효율성을 평가는 ASE, ROC 곡선 하 면적 및 Lift 지표를 사용하였다.

ASE[3](Average Squared Error, 평균제곱오차)를 산출하는 식은 아래와 같다.

$$ASE = \frac{SSE}{N} = \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{n}$$, (단, N=관측치의 개수) ……(식 1)

ROC곡선(Receiver Operating Characteristic Curves)의 밑면적의 크기가 클수록 개발된 예측모형의 성능이 우수하다고 판단된다. 또한, Lift지표는 추정된 사후확률

3) MSE(Mean Squared Error)=SSE/N-P, (단, N=관측치의 개수, P=독립변수의 개수)

(Posterior Probability)의 분위수(Percentile)에 따른 반응률(%Response)를 나타내는 값으로, 상위 분위수에 대응되는 Lift값이 더 클수록 모형의 성능이 더 우수함을 나타낸다(강현철 등, 2001).

Ⅲ. 연구 결과

1. 연구대상자의 인구사회학적 특성

연구대상자는 2,664,946명 이었으며 연령별 분포는 39세 이하 34.9%, 40~49세 34.7% 등이었으며, 직역별 분포는 직장 59.5%, 공교 28.1% 등의 순이었다.

<표 3> 연구 대상자의 인구사회학적 특성

(단위 : 명, %)

특 성		2000년	
		대상자수	비율
연령	60세 이상	255,673	9.6
	50세~59세	552,442	20.7
	40세~49세	925,499	34.7
	39세 이하	931,332	34.9
성별	남	1,889,989	70.9
	여	774,957	29.1
거주지	대도시	1,194,232	44.8
	중소도시	1,182,006	44.4
	농어촌	276,791	10.4
	없음	11,917	0.5
직역	지역	316,646	11.9
	공교	748,760	28.1
	직장	1,587,624	59.5
	없음	11,916	0.5
보험료	Ⅳ	727,171	27.3
	Ⅲ	653,189	24.5
	Ⅱ	619,420	23.2
	Ⅰ	665,166	25.0
전 체		2,664,946	100.0

주) 월기준 보험료
Ⅰ. 상위 25%미만 : 31,610원 미만, Ⅱ. 상위26%~50%미만 : 31,610원 이상 50,820원 미만
Ⅲ. 상위50%~75%미만 : 50,820원 이상 69,930원 미만, Ⅳ. 상위75%이상 : 69,930원 이상

2. 고혈압 발생 예측모형

1) 고혈압 발생 예측모형 개발 및 평가

(1) 모형개발 및 내적 타당도(Internal validation)

2000년 검진 당시 고혈압으로 진단 받지 않은 사람들이 2년 뒤인 2002년에 고혈압이 발생할 확률에 대한 모형을 분석용 자료(Training Data)를 이용하여 데이터마이닝 알고리즘인 로지스틱 회귀모형, 의사결정나무, 앙상블 알고리즘으로 개발한 후, 평가용 자료(Validation Data) 이용하여 각 모형에 대한 내적 타당도를 평가·비교하였다. 그 결과 모형의 예측정확도를 나타내는 평균제곱오차의 제곱근(Root Average Squared Error)은 로지스틱 회귀모형이 0.2364(분석용 자료: 0.2363)로 가장 낮게 나타나서 세 가지 모형 중 가장 우수한 모형으로 평가되었다.

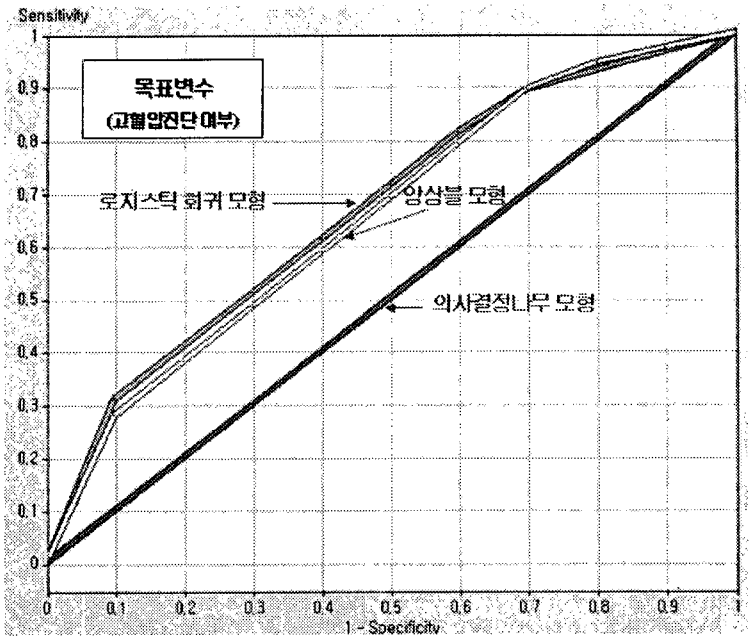
또한, ROC곡선(Receive Operating Characteristic Curve)을 통하여 데이터마이닝 알고리즘으로 개발된 각 모형들의 성능을 비교·파악하였다. 그 결과 로지스틱 회귀모형이 다른 알고리즘에 의해 개발된 모형보다 ROC곡선 하면적이 가장 크며, 또한, 대부분의 '1-특이도'의 상황에서 세 가지 모형 중 로지스틱 회귀모형의 민감도가 가장 높다(그림 1).

(2) 외적타당도(External validation)

외적타당도 평가를 평균제곱오차의 제곱근(Root ASE) 및 오분류율(Misclassification)를 이용하여 비교한 결과, 모형개발 및 내적타당도 평가에서 개발된 로지스틱 회귀모형이 가장 우수한 모형으로 평가되었다. 그 이유는 모형을 평가하는 지표인 평균제곱오차의 제곱근이 3가지 알고리즘 중 로지스틱회귀모형이 가장 낮았으며, 오분류율은 세 모형이 거의 비슷한 것으로 나타났기 때문이다.

<표 4> 개발된 모형의 내적 타당도 및 외적 타당도 평가

모형	평균오차의 제곱근(Root ASE)			오분류율(Misclassification)		
	2000~2002		2001~2003 검증용 (외적타당도)	2000~2002		2001~2003 검증용 (외적타당도)
	분석용 (60%)	평가용 (40%) (내적 타당도)		분석용 (60%)	평가용 (40%) (내적 타당도)	
로지스틱 회귀모형	0.2363	0.2364	0.2475	0.0625	0.0625	0.0693
의사결정나무 모형	0.2419	0.2419	0.2537	0.0624	0.0624	0.0692
앙상블 모형	0.2375	0.2376	0.2489	0.0624	0.0624	0.0692



<그림 1> 고혈압 발생 예측모형 ROC도표(2000~2002년)

2) 고혈압 발생 예측을 위한 로지스틱 회귀모형의 성능 평가

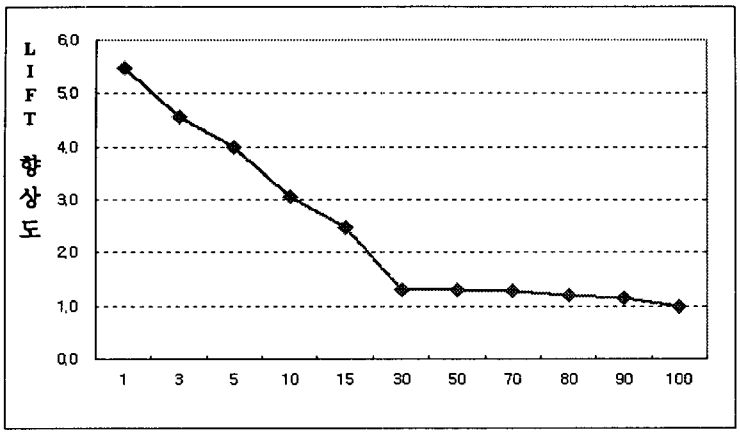
내적타당도 평가와 외적타당도 평가에서 로지스틱 회귀모형이 가장 우수한 모형으로 판명됨에 따라 로지스틱 회귀모형을 고혈압 발생을 예측하는 모형으로 채택하였다. 채택된 모형을 사용하여 고혈압의 고위험군을 예측할 경우 임의의 모형(Random Model)으로 선택하는 경우 보다 효율성을 평가하기 위해 향상도 지표(Lift)를 이용하여 평가하였다.

로지스틱 회귀모형으로 개발(Training Data(60%): 2000~2002년)된 고혈압 발생예측을 적용하였을 때 사후확률이 상위 1%에 해당되는 향상도(Lift)가 5.48로서, 임의의 모형보다 5.48배 향상된 것으로 나타났다. 따라서 이 모형을 적용할 때 상위 1%에 해당하는 경우에서는 고혈압 발생 확률이 5.48배 높아진다는 것을 알 수 있다. 또한, 개발된 모형에 대한 평가(Test Data: 2001~2003년)를 실시한 결과, 개발된 모형을 통해 나타난 향상도 들이 전체적으로 평가용 자료에서도 비슷한 값들이 도출되어 개발된 모형이 모형의 타당도 및 안정성 측면에서도 적절한 모형임을 알 수 있었다(표 5, 그림 2).

<표 5> 고혈압 발생 예측모형 이익도표(Logistic Regression)

	상위%	고혈압 발생 (명)	전체 대상자	예측력 (%Response)	모형적용효과 (%Captured Response)	향상도 (Lift)
모형구축 (Train) (2000~2002)	1	5,449	15,945	34.17	5.48	5.48
	3	13,659	47,868	28.53	13.73	4.57
	5	19,836	79,720	24.88	19.93	3.99
	10	30,293	159,481	18.99	30.44	3.04
	15	36,968	239,191	15.46	37.15	2.48
	30	82,897	1,010,287	8.21	83.30	1.31
	50	83,601	1,020,814	8.19	84.01	1.31
	70	89,122	1,116,131	7.98	89.56	1.28
	80	95,071	1,275,706	7.45	95.54	1.19
	90	97,198	1,371,609	7.09	97.68	1.14
	100	99,511	1,594,589	6.24	100.00	1.00
모형검증 (Test Ⅱ) (2001~2003)	1	7,978	22,382	35.64	5.16	5.17
	3	20,813	67,333	30.91	13.46	4.48
	5	30,528	112,275	27.19	19.75	3.94
	10	47,111	224,192	21.01	30.48	3.05
	15	57,903	335,637	17.25	37.46	2.50
	30	133,828	1,545,948	8.66	86.57	1.26
	50	133,940	1,547,400	8.66	86.65	1.26
	70	135,373	1,569,572	8.62	87.57	1.25
	80	146,838	1,793,398	8.19	94.99	1.19
	90	152,505	2,017,912	7.56	98.66	1.10
	100	154,581	2,242,172	6.89	100.00	1.00

주 : 1) %Response=(해당 집단에서의 고혈압 발생자수/해당 집단에서 전체 대상자의 수)×100
2) %Captured Response=(해당 집단에서의 고혈압 발생자수/전체 고혈압 발생자 수)×100
3) Lift=해당 집단에서의 고혈압 발견율/전체 고혈압 발견율



<그림 2> 고혈압 발생 예측모형 향상도표(Lift Chart : 2000~2002년)

3. 고혈압 발생의 주요 요인

본 연구 결과에 근거한 고혈압 발생의 주요 위험요인은 표 6과 같다.

개발된 고혈압 발생 예측모형은 연구모형에서 제시된 인구사회학적 특성, 임상학적 특성, 건강행위 특성, 진료이용량 및 각 특성의 교호작용을 고려하여 개발되었다. 그 결과 인구학적 특성인 성별, 연령과 임상학적 특성인 혈압, 고혈압 가족력, 그리고 건강행위 특성에서 음주량과 흡연여부 및 기타 복합요인이 가장 유의한 변수로 나타났다. 다른 특성요인들은 고혈압 발생에 통계적으로 유의한 영향을 미치지 못하여 주요 요인변수에는 포함되지 않았다.

개발된 모형결과에 의하면, 여자가 남자보다 고혈압 발생 확률이 0.834배 낮았으며, 연령이 높을수록 고혈압 발생 확률이 높았다. 60세 이상의 경우 40세 미만인 사람보다 2년 안에 고혈압 발생 확률이 4.628배 높은 것으로 나타났다. 임상학적 특성인 비만도에 대해서는 비만인 사람이 정상인 사람보다 고혈압 발생 확률이 2.103배 높은 것으로 나타났으며, 수축기 혈압 및 이완기 혈압 모두 정상A인 경우보다 정상B인 경우가 고혈압 발생 확률이 높은 것으로 나타났다. 식전혈당을 나타내는 글루코스 또한 식전혈당의 양이 클수록 고혈압 발생 확률이 높아지는 것으로 나타났다. 건강위험요인인 고혈압 가족력이 있는 사람(1.512배)이 없는 사람보다 음주량이 많은 사람이 적은 사람보다 고혈압 발생 확률이 높았다.

<표 6> 고혈압 발생모형(Logistic Regression)

특 성 요 인			추정회귀계수	상대진단 예측도
		절편	-1.8203***	
인구사회학적 특성	성별	남자		1
		여자	-0.0908***	0.834
	연령	40세 미만		1
		40대	-0.2271***	1.760
		50대	0.2792***	2.919
		60대 이상	0.7400***	4.628
임상학적 특성	비만도 (BMI)	정상		1
		위험체중	0.0000	.
		비만	0.3717***	2.103
	수축기 혈압 mmHg	정상A		1
		정상B	0.3568***	2.041
	이완기 혈압 mmHg	정상A		1
		정상B	0.1914***	1.466
	글루코스 (mg/dl)	정상A		1
		정상B		1.065
		질환의심	0.0344**	1.086
	가족력	무	0.0143	1
		유	0.2067***	1.512
건강행위특성	음주량	비음주		1
		월2~3회	-0.0802***	1.037
		일주일 1~2회	0.0576***	1.190
		일주일 3회이상	0.1391***	1.291
	흡연여부	무		1
		유	-0.0218***	0.957
진료이용량 (2000년)	투약일수	상위75%이상		1
		상위75%미만	-0.2204***	0.644

→ 다음 페이지에 계속

	특 성 요 인		추정회귀계수	상대진단 예측도
기타	복합요인(1)	음주습관양호		1
		음주습관개선권고*가족력무	0.0403***	1.101
		음주습관개선권고*가족력유	0.0159***	1.075
	복합요인(2)	혈압(정상)*비만도(정상)		1
		혈압(정상)*비만도(비만)	-0.3313***	1.032
		혈압(경계역)*비만도(정상)	0.4195***	2.186
		혈압(경계역)*비만도(위험)	0.2158***	1.783
		혈압(경계역)*비만도(비만)	0.0585	1.523
	복합요인(3)	글루코스(정상A)*비만도(정상)		1
		글루코스(정상A)*비만도(위험)	0.0060	1.109
		글루코스(정상A)*비만도(비만) 글루코스(정상B, 의심)*비만도(비만, 위험, 정상)	0.0917***	1.209

주 : 1) 투약일수 : 상위75미만(21일미만), 상위75이상(21일 이상)
2) *** : p<0.01

음주습관과 가족력에 대한 복합요인이 고혈압 발생에 미치는 영향을 살펴본 결과 건강검진 시 의사에 의해 음주습관이 양호하다고 판단되어진 경우에 비해, 음주습관에 대한 개선을 권고 받았고, 고혈압 가족력이 있는 경우가 고혈압 발생 확률이 1.075배 높았다. 혈압이 정상(수축기 혈압 139mmHg이하, 이완기 혈압 89mmHg이하)이고 비만도를 나타내는 BMI지수가 정상(23미만)인 사람보다 혈압이 경계역(수축기혈압 140mmHg이상~159이하 이완기혈압 90mmHg이상~94mmHg미만)이고, BMI지수가 비만(25이상)인 사람이 고혈압 발생 확률이 약 1.523배 이상 높았다.

4. 개발된 모형을 활용한 고혈압 예방사업 방안

본 연구에서는 인구 사회학적 특성, 임상학적 특성, 건강행위 특성 그리고 진료실적으로 집단화(Grouping)하여 고혈압 발생확률을 예측하였다. 개발된 모형을 활용한 고혈압 예방사업 방안을 다음과 같이 제시할 수 있다.

첫째, 국가의 고혈압 건강정책수립을 위한 정보자료를 다차원적인 측면에서 추출, 활용이 가능하다.

데이터마이닝 프로세스에 의하여 로지스틱 회귀모형으로 만들어진 고혈압 발생 예측모형이므로 특성별로 추출된 결과를 국가 고혈압 정책에 활용할 수 있다. 다만, 특성별 정밀분석을 위해서는 데이터수집, 정제, 추출 등을 위한 시스템구축이 선행되어야 한다.

둘째, 고혈압 발생가능 위험성의 사전예측정보를 생산하여 고혈압 발생을 줄여 나가는 개인별 맞춤형 정보제공자료 제공이 가능하다. 이

모형은 신체계측, 임상자료, 건강행태 등 건강위험평가시스템을 연계한 개인별 맞춤형 정보를 생산하여 SMS, DM(Direct Mail)등을 통하여 제공하고 행태변화가 이루어지도록 지속적으로 관리할 수 있다.

셋째, 국민건강보험법에 의한 건강검진자를 관리하고 있는 국민건강보험공단에서 사전관리 시스템을 구축·운용할 수 있다 다만, 이 시스템을 이용하기 위해서는 문진정보의 일관성, 생애주기별 개인정보 통합시스템이 구축되어야 한다. 국민건강보험공단에서는 건강검진제도 개선을 위하여 효과적인 프로그램 설계, 프로그램의 품질유지, 가입자의 수검율 향상과 함께 사후관리를 강화해야 할 것이다.

넷째, 신뢰성 있는 고혈압 관련 통계지표 생산으로 국가 또는 지방자치단체의 보건교육 등 공중보건학적 자료로 활용할 수 있다. 예컨대 각 개인이 건강위험요인을 줄이거나 제거할 때 입게 될 시간적, 인적, 경제적 효용을 추정할 수 있다. 또한 보건교육이나 홍보 등을 실시할 때 질병예방의 강조점을 어디에 두어야 하는지를 시사 하는 중요한 실증적 자료가 될 것이다.

Ⅳ. 결 론

본 연구는 국민건강보험공단에서 실시하고 있는 가입자들에 대한 건강검진 자료, 요양급여 자료 및 자격자료를 이용하여, 국가 차원에서 주요 관리대상 질환으로 주목받고 있는 고혈압의 발생을 예측하는 모형을 도출함으로써 향후 우리나라 고혈압관리체계의 기초자료를 제공하기 위해 수행되었다.

데이터마이닝 프로세스에 의하여 로지스틱 회귀모형으로 만들어진 고혈압 발생 예측모형은 고혈압 발생 예측확률 상위 1%에서 예측정확도가 모형을 적용하지 않을 경우에 비해 5.48배 높았다. 모형연구결과에 의하면, 인구학적 특성에서 여자보다는 남자가, 연령대가 높을수록 고혈압으로 발생할 확률이 높았다. 진료이용량 특성에서는 고혈압과 관련된 진료가 아닐지라도 다른 질병과 관련된 진료를 받는 것도 고혈압으로 발생할 확률이 높았다. 이는 타 질환의 발생도 고혈압 발생에 큰 영향을 주고 있음을 나타내고 있다. 또한, 임상학적 특성에서는 고혈압 가족력이 있고, 고혈압진단 이전의 혈압이 경계역에 가까울수록 고혈압 발생 확률이 높았고, 건강행위 특성에서는 음주량이 많을수록 고혈압 발생 확률이 높은 것으로 나타났다.

이 연구는 자료의 특성상 연구 모형에 포함된 변수들이 이론적 틀을 통해 도출되지 않고 이미 구축된 자료에 포함되어 있는 변수들만을 이용하였다는 제한점을 가지나, 연구 결과가 기존의 연구결과들과 유사하며, 현실적으로 이용가능한 자료를 가지고 설명력 있는 모형을 도출하였다는 점에서 의의를 가진다고 판단된다.

또한 기존의 연구들이 소규모의 일개 의료기관이나 지역에 국한된 자료를 사용함으로서 연구 결과의 대표성과 일반화에 한계를 가졌던 점 그리고 기본적인 통계분석을 통한 연구결과의 도출이라는 점과 비교할 때 이 연구는 전국규모의 수년간 축적된 자료를 최신 정보기술인 데이터마이닝을 활용함으로써 고혈압의 발병확률을 예측함으로써 우리나라 고혈압관리시스템 구축에 기여할 것으로 보인다.

참고 문헌

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

건강보험심사평가원. 2004년도 건강보험심사통계 지표. 건강보험심사평가원, 2005.

김영식. 한국인 고혈압과 당뇨병의 발병요인 규명을 위한 코호트 연구. 보건복지부 보건의료 기술연구개발사업 보고서. 2001.

맹광호, 박정일. 한국 도시 직장인의 고혈압과 체중을 관찰한 사례-비교군 관찰 연구. 대한의학협회지 1982;25(10):939-42.

문일순, 박성림, 박몽하, 황의정, 홍명호, 김순덕. 정기건강검진을 통하여 나타난 일부 도시지역 성인의 과체중과 그 영향. 가정의학회지 1989;10(8):20-30.

보건복지부. 2005년 국민건강 · 영양조사. 보건복지부, 2006.

서 일 외. 성장기 청소년의 혈압변화와 결정요인. 예방의학회지 1997;30(2):308-26.

유재희, 정해원. 문헌고찰을 통한 한국인의 식염섭취량 추이분석에 관한 연구. 국민보건연구소연구논총 1994;4(1):92-106.

이강숙, 최환석, 신호철, 박정일. 과체중, 고혈당 및 고콜레스테롤혈증에 대한 고혈압의 비교위험도. 가정의학회지 1994;15(12):1147-55.

전종민 외 6명. JNC-5 분류에 의한 한국 성인 환자에서의 고혈압 유병률에 관한 역학적 연구-포천지역 거주 직장인을 대상으로. 대한내과학회지 1997;52(2):209-23.

통계청. 사망원인별 사망자수. 통계청, 2004.

최영환, 채성철, 전재은, 박의현. 고혈압 환자에서 동맥경화증 위험인자. 순환기학회지 1996;26(2):490-99.

Beberg HT et al. Health promotion and cardiovascularrisk factors. The level of knowledge among 510 inpatients of an acute coronary care unit. *Medizinische Klinik* 2000;95(2):75-80.

Blair SN, Goodyear NN, Gibbons LW, Cooper KJ. Physical fitness and incidence of hypertension in healthy normotensive men and women. *JAMA* 1984;487-90.

Cutler JA, Follmann D, Allender PS. Randomized trials of sodium reduction: an overview. *Am J Clin Nutr* 1997;65(2 Suppl):643-651.

Elliott P et al. For the Intersalt Cooperative Research Group. Intersalt revisited: further analysis of 24 hour sodium exertion and blood pressure within and across populations. *BMJ* 1996;312:1249-53.

Friedman GD, Klatsky AL, Siegelaub AB. Alcohol, tabacco, and hypertension. *Hypertension* 1982;4(suppl Ⅲ):143-50.

Kim JS, Jones D, Kim SJ, Hong YP. A study on prevalence and risk factors of hypertension among Koreans (Abstract). Korean Society of Preventive medicine 43rd Meeting, Oct. 31-Nov.2,1991.

Neaton JD et al. For the Treatment of Mild Hypertension Study Research Group. Treatment of Mild Hypertension Study: final results. *JAMA* 1993;270:713-24.

Paffenbarger RS Jr, Hyde RT, Wing AL, Lee IM, Jung DL, Kamper JB. The association of change in physical-activity level and other lifestyle characteristics with mortality among men. *N Eng J Med* 1993;328:538-45.

Pauliot MC et al. Waist circumferenceand abdominal sagittal diameter: best simple anthropometric indexes of abdominal visceral adipose tissue accumulation and related cardiovascular risk in men and women. *Am J Cardiol* 1994;73:460-8.

Puddey IB, Parker M, Beilen LJ, Vandongen R, Masarei JRL. Effects of alcohol and caloric restrictions on blood pressure and serum lipids in overweight men. *Hypertension* 1992;20:533-41.

Stamler J, Caggiula AW, Grandits GA. Chapter 12. Relation of body mass and alcohol, nutrient, fiber, and caffeine intakes to blood pressure in the special intervention and usual care groups in the Multiple Risk Factor Intervention Trial. *Am J Clin Nutr* 1997;65(suppl):338-65.

<ABSTRACT>

Development of Hypertension Predictive Model

Wang-Sik Yong* · Il-Su Park* · Sung-Hong Kang** · Won-Joong Kim** ·
Kong-Hyun Kim*** · Kwang-Kee Kim*** · No-Yai Park****

*Korea National Health Insurance Corporation
**Department of Health Care Administration, Inje University
***Inje Institute of Advanced Studies
****Graduate School of Public Health, Inje University

Objectives: This study used the characteristics of the knowledge discovery and data mining algorithms to develop hypertension predictive model for hypertension management using the Korea National Health Insurance Corporation database(the insureds' screening and health care benefit data).

Methods: This study validated the predictive power of data mining algorithms by comparing the performance of logistic regression, decision tree, and ensemble technique. On the basis of internal and external validation, it was found that the model performance of logistic regression method was the best among the above three techniques.

Results: Major results of logistic regression analysis suggested that the probability of hypertension was:

- lower for the female(compared with the male)(OR=0.834)
- higher for the persons whose ages were 60 or above(compared with below 40)(OR=4.628)
- higher for obese persons(compared with normal persons)(OR= 2.103)
- higher for the persons with high level of glucose(compared with normal persons) (OR=1.086)
- higher for the persons who had family history of hypertension(compared with the persons who had not)(OR=1.512)
- higher for the persons who periodically drank alcohol(compared with the persons who did not)(OR=1.037~1.291)

Conclusions: This study produced several factors affecting the outbreak of hypertension using screening. It is considered to be a contributing factor towards the nation's building of a Hypertension Management System in the near future by bringing forth representative results on the rise and care of hypertension.

Key words : Hypertension, Predictive Model, National Health Insurance Corporation, Data Mining, Logistic Regression