JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 4, pp. 685-692, Apr. 2018

건강행위정보기반 고혈압 위험인자 및 예측을 위한 통계분석

허 병 문 · 김 상 엽 · 류 근 호*
충북대학교 데이터베이스/바이오인포매틱스 연구실

Statistical Analysis for Risk Factors and Prediction of Hypertension based on Health Behavior Information

Byeong Mun Heo · Sang Yeob Kim · Keun Ho Ryu*

Database/Bioinformatics Lab, School of Electrical & Computer Engineering, Chungbuk National University, Cheongju, South Korea

[요 약]

본 연구는 통계분석을 이용한 중년 성인의 고혈압 예측모델 개발이 목적이다. 국민건강영양조사자료(2013년-2016년)를 사용하여 통계분석과 예측모델을 개발하였다. 이진 로지스틱 회귀분석으로 통계적 유의한 고혈압 위험인자를 제시하였으며, Wrapper 변수선택기법을 적용한 로지스틱회귀와 나이브베이즈 알고리즘을 이용하여 예측모델을 개발하였다. 통계분석에서 고혈압에 가장 높은 연관성을 갖는 인자는 남성에서 WHtR ($p<0.0001$, OR = 2.0242), 여성에서 AGE($p<0.0001$, OR = 3.9185)로 나타났다. 예측모델의 성능평가에서, 로지스틱 회귀 모델이 남성(AUC = 0.782)과 여성(AUC = 0.858)에서 가장 좋은 예측력을 보였다. 우리의 연구 결과는 고혈압에 대한 대규모 스크리닝 도구를 개발하는데 중요한 정보를 제공하며, 고혈압 연구에 대한 기반정보로 활용할 수 있다.

[Abstract]

The purpose of this study is to develop a prediction model of hypertension in middle-aged adults using Statistical analysis. Statistical analysis and prediction models were developed using the National Health and Nutrition Survey (2013-2016).Binary logistic regression analysis showed statistically significant risk factors for hypertension, and a predictive model was developed using logistic regression and the Naive Bayes algorithm using Wrapper approach technique. In the statistical analysis, WHtR($p<0.0001$, OR = 2.0242) in men and AGE ($p<0.0001$, OR = 3.9185) in women were the most related factors to hypertension. In the performance evaluation of the prediction model, the logistic regression model showed the best predictive power in men (AUC = 0.782) and women (AUC = 0.858). Our findings provide important information for developing large-scale screening tools for hypertension and can be used as the basis for hypertension research.

색인어 : 고혈압, 예측모델, 기계학습, 임상정보, 신체계측

Key word : Hypertension, Prediction model, Machine Learning, Clinical Information, Anthropometry

http://dx.doi.org/10.9728/dcs.2018.19.4.685

Received 30 March 2018; **Revised** 19 April 2018
Accepted 27 April 2018

***Corresponding Author; Keun Ho Ryu**

Tel:
E-mail:

Ⅰ. 서 론

고혈압은 전 세계 질병 부담 중 하나이며, 심혈관질환, 심근경색, 뇌졸중, 그리고, 사망으로 이어질 수 있는 위험인자이다[1, 2].

고혈압 발생의 원인으로는 부적절한 식생활습관, 낮은 신체활동, 과도한 음주, 비만, 등이 있으며[3, 4], 고혈압의 치료방법으로는 항고혈압제 치료와 운동치료가 있다. 운동치료는 수축기혈압(Systolic blood pressure, SBP)과 이완기혈압(Diastolic blood pressure, DBP), 총 콜레스테롤(Total cholesterol, TC), 저밀도 콜레스테롤(Low density lipoprotein cholesterol, LDL-C), 트리글리세이드(Triglyceride, TG) 수치를 낮추며, 고밀도 콜레스테롤(High density lipoprotein cholesterol, HDL-C)을 높이는 효과가 있다[5, 6]. 항고혈압제 치료는 미래의 심혈관질환, 뇌졸중, 그리고 심근경색의 발병률과 사망률을 줄일 수 있다[7, 8].

고혈압은 주로 비만지표와의 연관성에 대해서 많은 연구들이 진행 되었다. 비만지표들 중에서 허리둘레(Waist circumference, WC)는 카리브와 필리핀 여성에서 고혈압을 예측할 수 있는 가장 높은 비만지표이다[9, 10]. 또한, 홍콩의 중국여성과 호주 남성은 허리와 엉덩이둘레의 비율인 WHR(Waist-to-hip ratio)[11, 12], 홍콩의 중국남성, 타이완 인구는 허리와 신장의 비율을 계산한 WHtR (waist-to-height)[11, 13], 그리고, 중국인구, 필리핀 인구, 그리고 인도 인구는 체중과 신장의 비율을 계산한 BMI(Body mass index)[14]가 고혈압을 예측할 수 있는 가장 높은 비만지표이다. 고혈압과 혈액정보와의 연구에서는 고혈압과 연관성을 갖는 혈액인자는 glucose(GLU), high- sensitivity C-reactive protein(hs-CRP), TG, TC, LDL-C, Hematocrit(HCT), Hemoglobin(HGB) 그리고, HbA1c에서 높은 수치를 나타낸다[15, 16]. 고혈압과 식생활습관에 대한 연구에서 과도한 음주량은 SBP와 DBP의 수치는 같이 증가하는 결과를 가져오며, 20세-59세의 남녀 모두에서 고혈압의 위험을 가져올 수 있다[17].

현재까지 고혈압에 대한 연구는 인구통계학, 신체계측, 혈액샘플, 그리고, 식생활습관의 연관성 또는 위험에 대한 각각의 연구들이 진행되었다. 그러나 다양한 위험인자들을 조합한 고혈압 예측모델은 연구된 바가 없다. 따라서 우리는 통계분석을 기반으로 속성선택기법과 분류 알고리즘을 이용하여 고혈압 예측모델을 제안한다.

제안한 모델은 한국인 중년성인을 대상으로 인구통계학 정보, 식·생활습관 정보, 신체계측 정보, 그리고, 혈액샘플 정보에서 고혈압 위험인자를 통계적 유의성 분석으로 제시한다. 그리고 속성선택기법과 예측 알고리즘을 이용하여 위험인자들을 조합한 고혈압 예측모델을 개발하고, 예측모델들 간의 성능평가를 통하여 최적의 고혈압 예측모델을 제안한다.

Ⅱ. 재료 및 기법

2-1 데이터 수집

이 연구는 2013년에서 2016년도에 국민건강영양조사(Korea national health and nutrition examination survey, KNHANES)에 참여한 40세 이상의 중년성인을 연구대상에 포함하였다. KNHNES는 한국인의 건강과 영양섭취상태에 대한 국가통계를 산출하며, 인구통계학, 신체계측, 혈액샘플 등의 자료를 공개·제공한다.

연구대상은 40세 이상의 성인남녀 17333명에서, 결측치와 고혈압 진단기준 범위에서 벗어난 대상을 제외한 10111명을 연구대상으로 포함하였다. 최종 정상혈압그룹 4602명과 고혈압그룹 5509명으로 분류하였으며, 자세한 사항은 그림 1 데이터전처리에 나타내었다.

2-2 고혈압 정의 및 관련변수 추출

1) 고혈압 정의

정상혈압은 SBP가 120 mmHg 미만, 그리고, DBP는 80 mmHg 미만으로 정의하였다. 고혈압 진단기준은 SBP가 140 mmHg 이상, 또는 DBP가 90 mmHg 이상, 또는 의사진단, 또는 항 고혈압약을 복용하는 참여자로 정의하였다[18, 19].

2) WHtR 산출

이 연구에서는 고혈압과 비만지표와의 관계를 분석하기 위해서 허리둘레와 신장의 비율로 계산하는 WHtR (Waist / height) 을 계산하여 적용하였다.

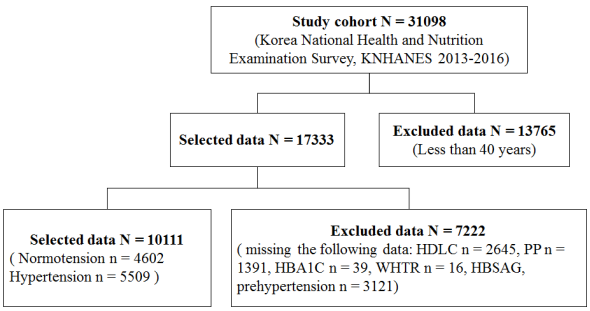

그림 1. 데이터 전처리
Fig. 1. Data preprocessing

2-3 최적의 고혈압 위험인자 조합을 위한 변수선택기법

속성선택기법 중 하나인 Wrapper approach는 속성 선별 시 분류 알고리즘을 적용하여 블랙박스(black box) 검증으로 정확도를 평가하여 최적의 속성을 선택한다 [20]. 이 연구에서는 Logistics regression (LR)과 Naive bayes (NB) 예측 알고리즘을 활용하여 각각의 알고리즘에 적합한 속성들을 선택하였다.

2-4 통계분석 및 예측 알고리즘

통계분석과 예측모델은 SPSS 20 와 Weka 3.8.1 을 이용하였다. 통계분석은 의료분야 연구에 가장 널리 활용되는 이진 로지스틱 회귀(Binary logistic regression)를 이용하여 고혈압과 다양한 인자들과의 통계적 유의성 분석을 수행하였다. 예측모델은 다양한 예측 알고리즘 중에서 예측성능과 처리속도를 고려하여 LR과 NB 알고리즘을 적용하여 예측모델을 개발하였다.

1) LR

통계적 회귀분석모델로 범주형 종속변수와 하나 이상의 독립변수 사이의 관련성 분석에 쓰인다. 예를 들어 고혈압의 경우, 로지스틱 회귀 모델은 고혈압의 확률로 고혈압을 앓고 있다면, "1", 고혈압이 아니면 "0"로 나타낼 수 있다. 독립변수 $X=(x_1,x_2,...,x_n)$일 때, 조건부확률 계산은 $p(y=1|X)$, $p(y=0|X)$로 나타내며[21], 공식은 다음과 같다.(1)

$$\log\left[\frac{p(x)}{1-p(x)}\right]=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n, \quad (1)$$

2) NB

각 클래스에 속하는 인스턴스의 조건부확률을 계산하고, 학습(training data set) 데이터를 기반으로 해당 인스턴스를 계산 시 가장 높은 조건부확률을 갖는 클래스로 분류한다. 계산식에서 P(c)는 그룹 c가 발생할 빈도이며, P(x)는 특정 개체가 발생할 확률이다. P(c|x)는 개체 x가 그룹 c에 속할 사후확률이며, P(x|c)는 그룹 c인 경우에 특정개체x가 속할 조건부 확률 가능성을 의미하며[21], 공식은 다음과 같다.(2)

$$P(c|x)=\frac{P(x|c)P(c)}{P(x)} \quad (2)$$

2-5 예측모델의 성능평가

예측모델의 평가는 곡선의 아래면적의 계산 값을 고려한 AUC [area under the ROC(receiver operating characteristic) curve]를 이용하였다. AUC는 기준점이 0.5 보다 왼쪽 위로 곡선이 형성될 경우 정확도가 높아지며 1.0일 때 민감도(sensitivity)와 특이도(specificity)는 100% 정확성을 갖는다. 성능평가의 세부적인 설명을 위하여 민감도, 1-특이도를 나타내었다.

Ⅲ. 결과평가

3-1 기본특성 분석

표 1 기본특성분석은 인자들에 대해 남성과 여성의 평균과 표준편차를 나타냈다. 연구에 포함한 남성은 4073명(정상혈압: 1534명, 고혈압: 2539명), 여성은 6038명(정상혈압: 3068명, 고혈압:2970명)으로 분류하였다. 나이는 정상혈압그룹보다 고혈압그룹의 남성(평균 6세)과 여성(평균 12세)에서 높게 나타났다. 흡연량 또한, 고혈압그룹의 남성에서 평균 17.5%와 여성에서 평균 17.1%로 높게 나타났다. 신체계측인자는 HT을 제외한 WT, WC, WHTR, BMI가 고혈압 그룹에서 높게 나타났다.

혈액인자에서는 GLU, HbA1C, TG, AST, ALT, HCV, BUN, CRT, WBC 인자는 고혈압 그룹에서 높은 수치를 보였으며, HCV, CRT인자는 여성보다 남성이 약간 높게 나타났다. 이와 대조적으로 TC, HDLC, HBSAG는 정상혈압그룹에서 높게 나타났다. 교육수준인 EDU는 남성에서 고혈압그룹의 빈도가 높았으며, 여성에서 정상혈압그룹의 빈도가 높았다. 또한, 음주량 인자인 DRINK도 남성에서는 고혈압그룹의 빈도가 높았으며, 여성에서 정상혈압그룹의 빈도가 높았다.

표 1. 데이터의 기본특성(범주형)
Table 1. Basic characteristic of features(Category type)

	Men		Women	
Features	Normotension	Hypertension	Normotension	Hypertension
EDU lv.0	130	233	244	535
EDU lv.1	2	0	3	12
EDU lv.2	178	441	385	1080
EDU lv.3	168	415	393	465
EDU lv.4	465	776	1099	594
EDU lv.5	136	138	287	94
EDU lv.6	325	401	568	159
EDU lv.7	130	135	89	31
DRINK lv.0	390	626	1106	1593
DRINK lv.1	311	397	1197	891
DRINK lv.2	295	435	489	299
DRINK lv.3	221	379	166	110
DRINK lv.4	184	391	72	50
DRINK lv.5	133	311	38	27

Values are expressed in frequency; The EDU represents the level of education and is divided into levels from non-education to graduate education; DRINK is divided into 0 to 10 cups of drinking per day

표 2. 데이터의 기본특성(수치형)
Table 2. Basic characteristic of features(Numerical type)

	Men		Women	
	Normotension	Hypertension	Normotension	Hypertension
Features	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
AGE	55.08(11.03)	61.18(11.21)	52.05(9.329)	64.62(10.27)
SMOKE	95.52(153.5)	136.1(180.8)	3.411(26.32)	4.821(39.61)
PP	17.17(1.978)	17.45(2.418)	17.36(1.928)	17.43(2.155)
HT	169.2(6.456)	167.5(6.249)	157.1(5.757)	153.4(5.992)
WT	66.90(9.672)	70.26(10.98)	56.53(7.667)	59.28(9.447)
WC	83.22(7.996)	88.15(8.561)	77.14(7.997)	84.42(9.355)
WHtR	0.491(0.047)	0.526(0.049)	0.491(0.054)	0.550(0.063)
BMI	23.29(2.742)	24.95(3.145)	22.88(2.846)	25.14(3.508)
GLU	102.0(25.81)	111.0(27.40)	95.84(18.10)	107.6(28.38)
HBA1C	5.844(0.889)	6.086(1.009)	5.649(0.615)	6.113(0.928)
TC	191.0(35.33)	183.8(36.59)	194.5(33.87)	194.3(38.79)
HDLC	46.54(11.09)	46.30(11.79)	54.39(12.40)	50.34(12.12)
TG	148.1(114.9)	177.7(149.9)	110.0(75.55)	145.5(97.99)
HBSAG	151.3(955.1)	126.3(903.6)	192.5(1084.)	97.35(725.3)
AST	23.23(11.16)	26.27(16.85)	20.44(7.711)	23.80(12.51)
ALT	23.50(16.55)	26.20(17.24)	17.12(11.00)	21.65(15.46)
HCV	0.108(0.651)	0.150(0.925)	0.118(0.684)	0.155(0.953)
HB	15.01(1.196)	14.98(1.398)	13.01(1.167)	13.24(1.148)
HCT	44.77(3.446)	44.62(3.868)	39.61(3.077)	40.19(3.244)
BUN	15.61(4.154)	16.26(5.960)	13.87(3.867)	15.82(5.060)
CRT	0.957(0.163)	1.029(0.615)	0.711(0.141)	0.760(0.299)
WBC	6.549(2.013)	6.791(1.897)	5.682(1.517)	6.211(1.772)
RBC	4.818(0.421)	4.794(0.464)	4.321(0.317)	4.358(0.386)
BPLT	244.2(61.35)	241.9(62.95)	257.8(61.37)	260.8(65.23)

Values are expressed as the mean and standard deviation (SD); Abbreviations: AGE, Years of age; SMOKE, heavy smoker; PP, Pulse pressure; HT, Height; WT, Weight; WC, Waist circumference; WHTR, Waist-to-height circumference ratio; BMI, Weight divided by height squared; GLU, Glucose; HBA1C, Hemoglobin A1c; TC, Total cholesterol; HDLC, High density lipid Cholesterol; TG, Triglyceride; HBSAG, hepatitis B surface antigen; AST, Aspartate aminotransferase; ALT, Alanine aminotransferase; HCV, anti hepatitis C virus; HB, Hemoglobin; HCT, Hematocrit; BUN, Blood urea nitrogen; CRT, Creatinine; WBC, White blood cell; RBC, Red blood cell; BPLT, Blood platelet

3-2 개별 인자들에 대한 통계적 유의성 및 예측력 분석

표 3은 이진 로지스틱 회귀분석결과로, 고혈압 위험인자에 대해서 남성과 여성 각각의 유의확률(p-value)과 승산비(OR; Odds ratio)를 나타내었다.

고혈압과 가장 높은 연관성을 보인 인자는 남성에서 WHTR($p<0.0001$, OR = 2.0242), 여성은 AGE ($p<0.0001$, OR = 3.9185)로 나타났다.

식생활습관 정보에서는 남성은 흡연량을 나타내는 SMOKE($p<0.0001$, OR = 1.252)가 연관성을 보였으며, 여성은 교육수준인자인 EDU($p<0.0001$, OR = 0.465)가 역의 연관성을 보였다. 비만지표들에서는 WHTR, WC, BMI가 성별과 상관없이 고혈압의 위험인자로 높게 나타났다. 예를 들어, 남성은 WHTR($p<0.0001$, OR = 2.0242), WC($p<0.0001$, OR = 1.7873) 그리고 BMI ($p<0.0001$, OR = 1.7276)순으로 높게 나타났으며, 여성 또한 WHTR($p<0.0001$, OR = 3.1272), WC($p<0.0001$, OR = 2.5083) 그리고 BMI($p<0.0001$, OR = 2.1810)순으로 나타났다. 혈액샘플인자에서 남성은 CRT ($p<0.0001$, OR = 1.430), 여성은 HBA1C($p<0.0001$, OR = 2.308)이 가장 높은 예측인자로 나타났다. 그리고, CRT, GLU, AST, HBA1C, TG, ALT, BUN, WBC는 남성과 여성에서 고혈압과 연관성을 보인 공통인자로 나타났다. 여성에서만 고혈압과 연관성을 보인 인자는 HB, HBSAG, HCT, TC이며, HDLC는 역의 연관성을 보였다. 또한, 교육수준(EDU)과 신장(HT)은 고혈압과 역의 연관성을 보였으며, 남성보다 여성에서 높게 나타났다.

표 3. 고혈압 위험인자의 통계분석
Table 3. Statistical significance of hypertension risk factors

	Men		Women	
Features	p-value	OR	p-value	OR
AGE	<0.0001	1.6783	<0.0001	3.9185
EDU	<0.0001	0.8186	<0.0001	0.4646
DRINK	<0.0001	1.1446	<0.0001	0.7419
SMOKE	<0.0001	1.2519	0.1064	1.0439
PP	0.0002	1.1279	0.1761	1.0355
HEIGHT	<0.0001	0.7694	<0.0001	0.5117
WEIGHT	<0.0001	1.3569	<0.0001	1.3881
WC	<0.0001	1.7873	<0.0001	2.5083
WHTR	<0.0001	2.0242	<0.0001	3.1272
BMI	<0.0001	1.7276	<0.0001	2.1810
GLU	<0.0001	1.4290	<0.0001	2.1732
HBA1C	<0.0001	1.2841	<0.0001	2.3076
TC	<0.0001	0.8277	0.7944	0.9933
HDLC	0.5332	0.9807	<0.0001	0.7133
TG	<0.0001	1.2550	<0.0001	1.7465
HBSAG	0.4173	0.9748	0.0001	0.8977
AST	<0.0001	1.3359	<0.0001	1.6604
ALT	<0.0001	1.1716	<0.0001	1.6196
HCV	0.1406	1.0512	0.0930	1.0472
HB	0.5557	0.9817	<0.0001	1.2192
HCT	0.2289	0.9630	<0.0001	1.2030
BUN	0.0003	1.1324	<0.0001	1.6139
CRT	<0.0001	1.4300	<0.0001	1.6240
WBC	0.0002	1.1273	<0.0001	1.3933
RBC	0.1025	0.9500	<0.0001	1.1122
BPLT	0.2583	0.9651	0.0676	1.0482

Binary logistic regression analysis showed that hypertension risk factors were statistically significant;.Statistical significance was expressed as p-value and odds ratio(OR) for the features