The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.555
33(5), 555–567

Comparison of nomograms designed to predict hypertension with a complex sample

Min Ho Kim[a] · Min Seok Shin[a] · Jea Young Lee[a,1]

[a]Department of Statistics, Yeungnam University

(Received June 3, 2020; Revised July 9, 2020; Accepted August 6, 2020)

Abstract

Hypertension has a steadily increasing incidence rate as well as represents a risk factors for secondary diseases such as cardiovascular disease. Therefore, it is important to predict the incidence rate of the disease. In this study, we constructed nomograms that can predict the incidence rate of hypertension. We use data from the Korean National Health and Nutrition Examination Survey (KNHANES) for 2013–2016. The complex sampling data required the use of a Rao-Scott chi-squared test to identify 10 risk factors for hypertension. Smoking and exercise variables were not statistically significant in the Logistic regression; therefore, eight effects were selected as risk factors for hypertension. Logistic and Bayesian nomograms constructed from the selected risk factors were proposed and compared. The constructed nomograms were then verified using a receiver operating characteristics curve and calibration plot.

Keywords: hypertension, logistic regression, naïve Bayesian classifier, nomogram, risk factor

1. 서론

고혈압은 수축기 혈압 또는 이완기 혈압이 비정상적으로 높은 질병으로 한국에서는 2011년 30세 이상 성인 중 약 28.5% 발병하였고, 65세 이상 성인 중에서는 남성 58.4%, 여성 61.8%가 발병하였다 (Shin 등, 2015). 고혈압은 관상동맥질환, 심부전증, 뇌졸중, 혈관성 치매와 같은 심혈관계 질환으로 발전할 수 있으며 (Van den Berg 등, 2009), 이런 심혈관계질환은 2017년 한국 인구 10만 명당 119.6명으로 한국인 사망 원인 중 2위를 차지하였다 (Nam 등, 2018; Statistics Korea, 2018). 본인이 고혈압임을 인지하지 못하는 환자들이 많고 고혈압이 다른 합병증을 유발하는 질병이므로 인식과 예방이 중요시 된다.

따라서 위험 요인의 인식과 질병 예측에 도움을 줄 수 있는 도구나 방법의 발전이 중요하다. 이를 도울 수 있는 통계적 도구 중 하나가 바로 노모그램(nomogram)이다. 노모그램은 분석을 통해 예측한 확률을 시각적으로 설명하기 위해 만들어진 그래프이다. 노모그램의 가장 큰 장점은 위험 요인을 한 눈에 확인할 수 있고, 개개인의 특징을 바탕으로 질병이 발생할 확률을 점수를 통해 바로 예측할 수 있다는 점이다 (Mozina 등 2004). 이러한 노모그램은 raw data를 이용하여 당뇨와 이상지질혈증에 대해 구축된 바 있다 (Park 등, 2018; Kim 등, 2019). 질병의 위험 요인들을 규명하기 위해 진행된 연구들은 대부분 질병의 발병률을 예측하는 통계적 모형으로 로지스틱 회귀모형이나 Cox 비례위험모형을 많이 사용해왔다.

[1]Corresponding author: Department of Statistics, Yeungnam University, 280 Daehak-Ro, Gyeongsan, Gyeongbuk 38541, Korea. E-mail: Jlee@yu.ac.kr

이처럼 고혈압에서도 위험 요인을 선별하는데 다양한 선행연구가 진행 되었지만 실제로 통계학적 지식이 부족한 비전공자들은 분석 결과만으로 실제 고혈압의 위험 정도를 인지하는데 어려움이 있다. 그래서 본 연구에서는 고혈압의 노모그램을 로지스틱 회귀모형과 순수 베이지안 분류기 모형으로 각각 구축하였고 (Kim, 2020; Kim과 Lee, 2020), 비교를 통해 두 노모그램의 유용성을 분석하였다. 본 논문의 2절에서는 고혈압의 위험요인을 선별하고 위험도를 추정하는 방법인 Rao-Scott chi-squared test와 복합 표본 하에서 로지스틱 회귀모형과 순수 베이지안 분류기 모형을 이용한 노모그램 구축과 검증 방법을 소개한다. 3절에서는 국민건강영양조사 데이터에 대한 설명과 2절에서 설명한 방법을 적용한 고혈압의 발병률을 예측하는 노모그램을 구축하고 검증한다. 마지막 4절에서는 구축한 노모그램에 대한 의견 및 결론을 제시한다.

2. Methodology

2.1. Rao-Scott chi-squared test

고혈압의 위험요인을 선정할 때 실제로 고혈압 유무에 따라 위험요인의 영향이 있는지 밝히는 과정은 필수적이다. 일반적으로 Pearson chi-squared test를 사용하여 고혈압의 위험요인을 선별한다.

Pearson chi-squared test는 고혈압과 위험 요소로 구성된 분할표에서 각 셀에 들어갈 빈도가 독립이라는 가정이 필요하다. 그러나 본 연구에서 사용된 데이터는 2단계 층화집락추출법을 사용했으므로, 분할표의 각 셀이 독립이라는 가정을 만족하지 못한다. 이 때 Pearson chi-squared test를 사용하면 검정통계량 값이 과도하게 커지므로 귀무가설을 쉽게 기각한다고 알려져 있다 (Rao와 Scott, 1981; Sung, 2012). 따라서 층, 집락, 가중치 등 설계 효과를 고려한 Rao-Scott chi-squared 통계량을 사용한다. Rao-Scott chi-squared 통계량은 다음과 같다.

$$\chi^2_{\text{Rao-Scott}} = \frac{\chi^2}{\hat{\delta}},$$

여기서 분자 χ^2은 Pearson chi-squared 통계량이고,

$$\hat{\delta} = \left[\sum_i \sum_j (1 - \hat{\pi}_{i+}\hat{\pi}_{+j})\, \hat{d}_{ij} - \sum_i (1 - \hat{\pi}_{i+})\, \hat{d}_{i+} - \sum_j (1 - \hat{\pi}_{+j})\, \hat{d}_{+j} \right] \Big/ (I-1)(J-1),$$

$$\hat{d}_{ij} = \frac{\widehat{\text{Var}}(\hat{\pi}_{ij})}{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})/n}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

여기서 $\hat{\pi}_{ij}$는 추정된 i번째 행, j번째 셀의 확률이고, $\widehat{\text{Var}}(\hat{\pi}_{ij})$는 $\hat{\pi}_{ij}$의 추정된 분산, n은 표본의 수이다. 그리고 $\hat{d}_{ij}$는 $\hat{\pi}_{ij}$의 설계 효과 추정치이다.

2.2. Nomogram construction method

의료 분야에서 질병이나 사망과 관련된 위험 요인을 선별하고, 질병 발생률을 예측하는 연구들이 활발하게 진행되고 있다. 위험을 예측하기 위해 질병 또는 사망에 영향을 주는 위험인자를 선별하고, 어느 정도 영향을 주는지 계산을 하는 통계적 기법들을 사용한다. 하지만 비 전공자들이 통계적 결과 만으로 위험률을 계산한다는 것은 다소 어려움이 있다. 따라서 복잡한 계산 없이 한 눈에 여러 위험요인들에 의한 질병이나 사망의 위험률을 알 수 있는 노모그램을 제시한다 (Lee 등, 2009; Iasonos 등, 2008). 노모그램의 구성요소로는 4가지가 있다. 위에서부터 Point 선, Risk Factor 선, Probability 선, Total Point 선이 있다. 질병의 발생 확률은 질병과 관련된 요인들의 Risk Factor 선으로부터 얻은 점수의 합으로 Total Point를 구하고 이에 대응하는 확률을 Probability 선에서 도출함으로써 예측할 수 있다.

2.2.1. Nomogram construction of logistic regression model 로지스틱 회귀모형의 결과를 이용해 노모그램을 구축하는 방법은 다음과 같다 (Iasonos 등, 2008; Park 등, 2018).

- point 선

  Point 선은 0점에서 100점으로 구성된다.

- Risk factor 선

  로지스틱 회귀모형으로부터 도출된 회귀계수 β_{ij}값으로 LP_{ij}값을 계산한다. 독립변수 X가 범주형 변수이고 j개의 범주를 가지는 경우 $j-1$개의 가변수(dummy variable)를 갖는다. 이 때 기준 범주의 회귀계수는 0이다.

$$\mathrm{LP}_{ij} = \beta_{ij} \times X_{ij},$$
$$\mathrm{Point}_{ij} = \frac{\mathrm{LP}_{ij} - \min_j \mathrm{LP}_{ij}}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \times 100,$$

  여기서 β_{ij}는 i번째 위험요인의 j번째 범주의 회귀계수 값, X_{ij}는 i번째 위험요인의 j번째 범주의 속성값을 나타낸다. LP_{*j}는 추정된 회귀계수의 편차가 가장 큰 위험요인의 LP값을 나타낸다.

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\text{Total Point} = \frac{100}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \sum_i \sum_j \left(\mathrm{LP}_{ij} - \min_j \mathrm{LP}_{ij} \right).$$

  이제 위 Probability 선의 각 값에 대응하는 Total point 값을 구하기 위해 로지스틱 회귀모형을 $\sum_{i,j} \mathrm{LP}_{ij}$에 대해 정리한 뒤, 위 식에 대입하면 다음과 같은 식이 도출된다.

$$\text{Total Point} = \frac{100}{\max_j \mathrm{LP}_{*j} - \min_j \mathrm{LP}_{*j}} \left(\log \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} - \beta_0 - \sum_i \sum_j \min_j \mathrm{LP}_{ij} \right).$$

  그 뒤 $P(Y=1|X=x)$에 Probability 선의 값을 대입하여 Total point 선을 구축한다.

2.2.2. Nomogram construction of naïve Bayesian classifier model

- point 선

  Point 선은 -100점에서 100점으로 구성된다.

- Risk factor 선

  순수 베이지안 분류기 모형으로 얻어진 $\log \mathrm{OR}(a_i = j)$를 이용해 Point_{ij}를 계산하면 다음과 같다.

$$\mathrm{Point}_{ij} = \frac{\log \mathrm{OR}(a_i = j)}{\max_{i,j} |\log \mathrm{OR}(a_i = j)|} \times 100.$$

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\begin{aligned}\text{Total Point} &= \frac{100}{\max_{i,j}|\log \text{OR}(a_i = j)|}\sum_i\sum_j(\log \text{OR}(a_i = j)) \\ &= \frac{100}{\max_{i,j}|\log \text{OR}(a_i = j)|} \times \left(-\log\left(\frac{1}{P(Y=1|X=x)}-1\right) - \log\frac{P(Y=1)}{1-P(Y=1)}\right).\end{aligned}$$

  그 뒤 $P(Y=1|X=x)$에 Probability 선의 값을 대입하여 Total point 선을 구축한다.

2.2.3. Left-aligned method of nomogram for naïve Bayesian classifier model 순수 베이지안 분류기 모형은 점수가 −100∼100점이므로 left-aligned 방법을 적용한다. 이는 점수가 0∼100점이므로 로지스틱 노모그램과 비교하기 쉽다.

- point 선

  Point 선은 0점에서 100점으로 구성된다.

- Risk factor 선

  순수 베이지안 분류기 모형에서 적합시켜 도출된 $\log \text{OR}(a_i = j)$값으로 각 위험요인의 범주 별 Point_{ij}를 계산한 후 Point 선에 맞추어 정렬한다.

$$\text{Point}_{ij} = \frac{\log \text{OR}(a_i = j) - \min_{i,j}\log \text{OR}(a_i = j)}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \times 100.$$

- Probability 선

  Probability 선은 0에서 1까지 확률을 적절한 기준으로 분할해 구간을 만든다.

- Total point 선

  Total point는 각 위험요소의 Point_{ij}들의 총합이다.

$$\begin{aligned}\text{Total Point} = &\frac{100}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \\ &\times \sum_{i,j}\left(\log \text{OR}(a_i = j) - \min_{i,j}\log \text{OR}(a_i = j)\right).\end{aligned}$$

  이제 위 Probability 선의 각 값에 대응되는 Total point 값을 구하기 위해 순수 베이지안 분류기 모형을 $\sum_{i,j}\log \text{OR}(a_i = j)$에 대해 정리한 뒤, 위 식에 대입하면 아래와 같은 식이 도출된다.

$$\begin{aligned}\text{Total Point} = &\frac{100}{\max_j \log \text{OR}(a_* = j) - \min_j \log \text{OR}(a_* = j)} \\ &\times \left(-\log\left(\frac{1}{P(Y=1|X=x)}-1\right) - \text{logit}P(Y=1) - \sum_{i,j}\min_{i,j}\log \text{OR}(a_i = j)\right).\end{aligned}$$

  그 뒤 $P(Y=1|X=x)$에 probability 선의 값을 대입하여 Total point 선을 구축한다.

2.3. Nomogram validation method

노모그램을 구축한 뒤, 노모그램의 정확성을 검증하기 위해 receiver operating characteristics (ROC) curve와 calibration plot을 사용하였다 (Akobeng, 2007; Cook, 2008). ROC curve는 X축은 1 − Specificity, Y축은 Sensitivity로 구성되어 있으며, curve 아래 면적의 넓이(area under curve; AUC)는 예측 정확도의 지표로 사용된다. 다른 도구로써 Calibration plot을 이용하는데, X축은 예측확률, Y축은 실제확률로 이루어져 있다. 만약 예측확률이 실제 확률과 가깝다면, Calibration plot은 45° 각도의 선에 가깝게 그려진다 (D'Agostino 등, 2001). 그리고 예측확률과 실제 확률간의 회귀직선의 적합도 지표인 R^2으로 노모그램을 검증한다. 모든 분석은 R software 3.6.1을 사용했고, 노모그램을 구축하는 툴은 SAS 9.4를 사용하였다 (SAS Institute Inc., Cary, NC, USA).

3. Applications

3.1. Complex sample materials

본 연구에 사용된 데이터는 국민건강영양조사(Korean National Health and Nutrition Examination Survey; KNHANES) 2013∼2016년도 자료이다. 국민건강영양조사는 표본의 대표성 및 추정의 정확성 향상을 위해 복합 표본 설계 방식인 2-staged stratified cluster sampling method을 사용하였다 (Korea Centers for Disease Control and Prevention, 2016). 고혈압의 진단 기준은 수축기 혈압이 140mmHg 이상이거나 이완기 혈압이 90mmHg 이상이거나 고혈압 약을 복용 중이거나 의사의 진단을 받은 사람으로 선정하였다.

사용된 위험요인은 총 10개로 고혈압 발병에 중요한 영향을 미치는 여러 선행 연구에서 선정하였다 (Kshirsagar 등, 2010). 위험요인은 나이, 성별, 흡연 상태, BMI, 고혈압 가족력, 당뇨병, 음주 유무, 운동 유무, 뇌졸중, 이상지질혈증이다. 나이는 20세 이상 44세 이하, 45세 이상 64세 이하, 65세 이상으로 범주화 하였다. 흡연 상태는 현재 흡연을 하는 그룹(present), 과거에 흡연을 했던 그룹(past), 그리고 흡연을 한적이 없는 그룹(no)으로 나누었다. BMI는 25 미만을 정상(normal), 25 이상 30 미만을 과체중(overweight), 그리고 30 이상은 비만(obese)으로 범주화 하였다. 고혈압 가족력 유무는 부모 및 형제 중 누구 한 명이라도 고혈압이 있을 경우를 yes로, 그렇지 않으면 no로 범주화 시켰다. 당뇨병 유무는 공복 혈당이 126mg/dL 이상이거나 의사 진단을 받았거나 혈당 강하제를 복용 중이거나 인슐린 주사를 투여 받은 사람을 yes로, 그렇지 않으면 no로 범주화 하였다. 음주 유무는 음주 경험이 없으면 no, 그렇지 않으면 yes로 범주화 하였다. 운동 유무는 일주일에 적어도 한번은 걷기 또는 근력 운동을 하면 yes로, 그렇지 않으면 no로 범주화 하였다. 뇌졸중 유무는 의사 진단 여부에 따라 범주화 하였다. 이상지질혈증 유무는 의사 진단을 받았거나, 총 콜레스테롤이 240mg/dL 이상이거나 콜레스테롤 강하제를 복용하거나 HDL콜레스테롤이 40mg/dL 미만이거나 고중성지방이 200mg/dL 이상인 경우 중 하나라도 해당되면 yes로, 그렇지 않으면 no로 범주화 하였다.

조사대상자는 20세 이상 성인 총 24,095명 중 건강설문조사에 참여하지 않은 1,727명을 제외한 22,368명이였다. 한 개인이 가지고 있는 결측치의 경우, 수치형 자료는 평균으로, 범주형 자료는 결측치 처리하려는 변수와 가장 관련이 깊은 2개의 변수를 chi-squared test를 통해 선정하고, 그 2개의 변수를 기준으로 그룹화 했을 때 결측치 변수의 최빈값을 구해 값을 대체하였습니다. 그 후 모형의 예측력을 판단하기 위해 데이터를 무작위로 7 : 3의 비율로 나누어 Training data ($n = 15659$)는 모형을 만들어 노모그램을 구축하는데 사용하였고 Test data ($n = 6709$)는 검증하는데 사용하였다.