The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.3.329
31(3), 329–342

Comparison of nomogram construction methods using chronic obstructive pulmonary disease

Ju-Hyun Seo[a] · Jea-Young Lee[a,1]

[a]Department of Statistics, Yeungnam University

(Received February 7, 2018; Revised March 26, 2018; Accepted March 28, 2018)

Abstract

Nomogram is a statistical tool that visualizes the risk factors of the disease and then helps to understand the untrained people. This study used risk factors of chronic obstructive pulmonary disease (COPD) and compared with logistic regression model and naïve Bayesian classifier model. Data were analyzed using the Korean National Health and Nutrition Examination Survey 6th (2013–2015). First, we used 6 risk factors about COPD. We constructed nomogram using logistic regression model and naïve Bayesian classifier model. We also compared the nomograms constructed using the two methods to find out which method is more appropriate. The receiver operating characteristic curve and the calibration plot were used to verify each nomograms.

Keywords: chronic obstructive pulmonary disease (COPD), logistic regression model, naïve Bayesian classifier model, nomogram, risk factors

1. 서론

만성 폐쇄성 폐질환(chronic obstructive pulmonary disease; COPD)이란 유해한 입자나 가스의 흡입에 의해 폐에 비정상적인 염증 반응이 일어나면서 점차 기류 제한이 진행되어 폐 기능이 저하되고 이로 인해 호흡곤란을 유발하게 되는 호흡기 질환이다. 세계보건기구(World Health Organization; WHO)에서 만성 폐쇄성 폐질환은 2016년 기준 사망원인 4위로 발표되었고, 최근 통계청의 발표에 의하면 만성 폐쇄성 폐질환을 포함한 호흡계통의 질환 사망률이 인구 10만 명당 54.6명으로 나타났다 (WHO, 2017; Korean Statistical Information Service, 2015). 또한, 우리나라의 경우 40대 이상 기준 폐쇄성 폐질환 유병률은 2009년 10.5%, 2015년 12.3%인 것으로 보고되어 유병 추이가 계속적으로 증가하는 것을 알 수 있다 (Korea Centers for Disease Control and Prevention, 2016). 일반적으로 사람들은 수년 동안 만성 기침이나 가래 생성과 같은 초기 증상을 가볍게 여기다 호흡 곤란이라는 가장 악화된 상황까지 발생하게 되면 그때 병원을 찾아 진료 받는 경향이 있다 (Zieliñski 등, 2001). 만성 폐쇄성 폐질환은 한 번 발병하면 완치가 어려워 지속적인 관리가 필요한 만성적 질환이기 때문에 인식과 예방이 굉장히 중요시된다. 따라서 위험 요인의 인식과 질병 예측에 도움을 줄 수 있는 도구나 방법의 발전이 중요하다. 이를 도울 수 있는 통계적 도구 중 하나가 바로 노모그램이다.

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

노모그램은 분석을 통해 예측한 확률을 시각적으로 설명하기 위해 만들어진 그래프로써 의사결정을 지원하는데 사용된다 (Možina 등, 2004b). 외국에서도 이미 많은 노모그램이 만들어졌고, 국내에서도 위암, 전립선암, 골육종 등의 질병에 대한 노모그램이 개발되었다 (Jun, 2015; Lee와 Chang, 2014; Kim 등, 2014). 노모그램의 가장 큰 장점은 위험 요인을 한눈에 확인할 수 있고, 개개인의 특징을 바탕으로 질병이 발생할 확률을 점수를 통해 바로 예측할 수 있다는 점이다 (Možina 등, 2004b). 일반적으로 노모그램은 로지스틱 회귀모형이나 Cox 비례위험모형을 사용하여 구축되었다. 로지스틱 회귀모형은 임상적 연구에서 개별 환자의 결과가 발생할 가능성에 대한 추정치를 알고자 할 때 사용 한다 (Steyerberg 등, 2000). 의료계에서는 질병 발생에 대하여 위험 인자라 생각하는 요인들의 영향력을 확인하고자 할 때 로지스틱 회귀분석을 사용하고 있다. Seo 등 (2017)에서 COPD의 위험 요인을 선별하여 로지스틱 회귀모형을 얻었고 이를 이용한 노모그램을 제시하였다. 최근 순수 베이지안 분류기 모형을 사용한 노모그램을 구축하는 방법도 소개되었다 (Možina 등, 2004a). 순수 베이지안 분류기 모형은 예측 모형을 구축하는데 가장 간단하면서 강력한 기법 중 하나이다 (Možina 등, 2004a). 베이즈 기법을 이용하여 임의의 위험 요인이 존재할 때 목표변수에 대한 조건부 확률을 계산하여 모형을 만들게 된다. 복잡한 계산 없이 실생활에서도 많이 쓰이는 조건부 확률을 이용하여 비교적 간단하게 결과를 도출한다. Seo와 Lee (2018)에서도 COPD의 위험 요인으로 순수 베이지안 분류기 모형을 이용하여 노모그램을 제시하였다. 이처럼 노모그램을 구축할 때 사용할 수 있는 두 모형의 형태가 차이를 보이기 때문에 각각의 모형을 사용하여 노모그램을 구축할 때 장·단점이 있을 수 있다. 따라서 본 연구에서는 COPD의 노모그램을 로지스틱 회귀모형과 순수 베이지안 분류기 모형으로 각각 구축하였고, 비교를 통해 두 노모그램의 유용성을 살펴보았다.

본 연구는 다음과 같은 순서로 구성되어 있다. 2절에서는 로지스틱 노모그램과 베이지안 노모그램의 구축 방법을 소개한다. 또한 3절에서는 노모그램을 구축하기 위해 사용된 데이터를 설명하고 검증 방법을 소개한다. 4절은 두 모형을 사용하여 구축된 만성 폐쇄성 폐질환의 노모그램을 제시하고 이를 비교하였다. 두 노모그램의 비교를 위하여 왼쪽 정렬 베이지안 노모그램(left-aligned Bayesian nomogram)과 상호작용항을 포함한 로지스틱 노모그램(logistic nomogram with interaction)을 추가로 구축하였다. 마지막으로 5절에서는 분석의 결과와 두 노모그램에 대한 유용성 및 의견을 제시하였다.

2. 노모그램(nomogram) 구축

노모그램은 특정한 결과에 대한 가능성을 예측하고자 고안된 통계학적 도구로 의료 분야에서는 특정 질병과 관련된 위험 요인들과 환자의 특성을 바탕으로 구축된다 (Jun, 2015). 또한, 단순한 그래픽 표현을 사용하여 한눈에 확인하기 쉽고, 복잡한 계산 없이 결과에 대한 확률을 예측하는데 사용할 수 있다 (Možina 등, 2004a). 노모그램을 구성하는 요소들은 Points 선, Total points (TP) 선, Probability 선이 있다 (Figure 2.1). Points 선은 각 위험 요인들의 영향력을 점수화하여 나타낸다. TP 선은 환자의 특징에 해당되는 위험 요인의 점수의 누적 합이며, 이 누적 합의 대응되는 확률 값을 나타내는 선이 Probability 선이 된다. 예를 들어, Figure 2.1에서 어떤 환자가 Flow(Yes), Age(85), Clinical size(5), Sex(F)이라면 $100 + 40 + 25 + 0 = 165$점으로 약 85%라는 것을 알 수 있다. 따라서 두 모형을 이용한 노모그램 구축 방법을 소개한다.

2.1. COPD의 로지스틱 노모그램 구축(logistic nomogram construction about COPD)

로지스틱 회귀모형은 종속변수가 범주형인 연구에 대해 다항 회귀 분석의 기술을 확장한 것이다 (Dayton, 1992). 성공/실패, 생존/사망이나 재발/완치와 같이 종속변수가 이분형으로 생성되어 있을 때 사

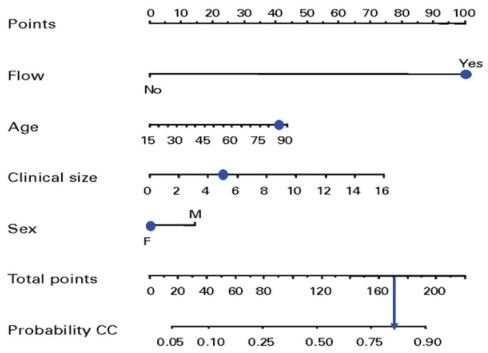


Figure 2.1. Nomogram plot (Iasonos *et al.*, 2008).

용하게 된다. 일반적으로 병리학 연구에서 질병과 연관 있는 위험 요인들을 식별하거나, 임상 연구 자료에서 중요한 요인들을 식별하는 탐색적 분석에 많이 적용 된다 (Lee 등, 2005; Heo과 Lee, 2008).

$$\ln\left(\frac{P(Y=1|X_1,\ldots,X_K)}{1-P(Y=1|X_1,\ldots,X_K)}\right)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k.$$

위의 식이 로지스틱 회귀분석을 통해 얻을 수 있는 회귀모형이다. 이때 β값인 회귀계수를 사용하여 노모그램을 구축한다 (Iasonos 등, 2008; Yang, 2014; Seo 등, 2017).

- Points 선

  먼저 모형에서 각 위험 요인에 대한 회귀계수 값을 이용하여 linear predictor (LP_{ij}) 값을 계산하고, 이를 이용하여 Points 값을 계산한다. Points 값의 범위는 0에서 100 사이의 값으로 변환된다.

$$\mathrm{LP}_{ij}=\beta_i\times X_{ij},$$
$$\mathrm{Point}_{ij}=\frac{\mathrm{LP}_{ij}-\min\limits_{ij}\mathrm{LP}_{ij}}{\max\limits_{j}\mathrm{LP}_{*j}-\min\limits_{j}\mathrm{LP}_{*j}}\times 100.$$

  이때, $i=1,2,\ldots,k$는 위험요인의 수, $j=1,2,\ldots,n_i$는 각 위험 요인 당 범주 수이다. LP_{*j}는 추정된 회귀계수 중 절댓값이 가장 큰 LP_{ij} 값이다.

- TP 선

$$\text{LP값 당 단위 점수(points per unit of linear predictor)}:\ \frac{100}{\max\limits_{j}\mathrm{LP}_{*j}-\min\limits_{j}\mathrm{LP}_{*j}}.$$

  (a) $\mathrm{LP}_{\text{for TP}=0}$: TP가 0의 경우 LP값 (상수값)

  (b) $\mathrm{LP}_{\text{for TP}>0}$: $\ln\left(\frac{P(\text{risk of } Y=1)}{1-P(\text{risk of } Y=1)}\right)$

$$\text{총점수(TP)}=\left(\frac{100}{\max\limits_{j}\mathrm{LP}_{*j}-\min\limits_{j}\mathrm{LP}_{*j}}\right)\times(\text{(b)}-\text{(a)}).$$

  위험 요인에 해당되는 점수를 모두 합하면 TP를 알 수 있지만, 노모그램에서는 임의의 확률에 대응되는 점수를 계산하기 위하여 위의 식과 같이 LP값 당 단위 점수를 계산하여 상수항을 제외한 회귀계수와의 곱으로 총 점수를 얻을 수 있다.

- Probability 선

  TP에 대응되는 확률 값을 나타내는 선이며, 이는 상수값을 제외한 LP 값을 이용하여 회귀모형 식을 P값에 대해 정리하면 얻을 수 있다.

2.2. COPD 베이지안 노모그램 구축(Bayesian nomogram construction about COPD)

순수 베이지안 분류기 모형은 예측 모델을 구축하는데 있어서 가장 간단하면서 강력한 기법 중 하나이다 (Možina 등, 2004a). 속성 값들이 서로 독립이라는 가정 하에 베이즈 정리(Bayes' theory)을 사용하여 임의의 클래스에 대한 확률을 계산하므로 보다 쉽게 사용 가능하다 (Bellazzi와 Zupan, 2008). 이때 클래스 C는 어떤 대상에 대한 목표 범주를 의미하며, 각 속성 $a_1, a_2, \ldots, a_m$은 대상에 대한 특징을 의미한다. 따라서 $P(C)$는 목표 범주가 발생할 확률이며, 이에 따라 $P(\bar{C}) = 1 - P(C)$는 목표 범주가 아닌 $P(\bar{C})$가 발생할 확률을 뜻하게 된다. 속성 값은 $X = a_1, a_2, \ldots, a_m$로 표현된다.

속성 값이 $X = a_1, a_2, \ldots, a_m$일 때, 클래스 C가 발생할 조건부 확률 $P(C|X)$와 $P(C|X)$의 오즈(odds)는 다음과 같이 계산된다.

$$P(C|X) = \frac{P(X \cap C)}{P(X)} = \frac{P(X|C)P(C)}{P(X)} = \frac{\prod_i P(a_i|C)P(C)}{P(X)},$$
$$\text{Odds} = \frac{P(C|X)}{P(\bar{C}|X)} = \frac{P(C)}{P(\bar{C})} \times \prod_i \frac{P(a_i|C)}{P(a_i|\bar{C})}.$$

위의 두 식을 이용하면 속성 값이 X일 때 클래스 C가 발생할 최종 확률값 $P(C|X)$에 대하여 정리하면 다음과 같이 계산된다.

$$P(C|X) = \frac{1}{1 + \exp\left(-\log\frac{P(c)}{P(\bar{c})} - \sum_i \log \text{LR}(a_i)\right)}.$$

순수 베이지안 분류기 모형에서 $P(C|X)$ 결과적으로 $\log(\text{LR}_i)$의 합으로 나타날 수 있기 때문에 이를 이용하여 노모그램을 구축한다 (Možina 등, 2004a; Seo와 Lee, 2018).

- 각 속성의 Points 선

  속성 값 a_{ij}에 따라, $\log \text{LR}(a_i)$는 다음과 같이 계산되고, 각 속성의 Points_{ij}는 $\log \text{LR}(a_{ij})$를 이용하여 계산된다. 점수 범위는 -100~100점으로 된다.

$$\log \text{LR}(a_{ij}) = \log \frac{P(a_{ij}|C)}{P(a_{ij}|\bar{C})},$$
$$\text{Points}_{ij} = \frac{\log \text{LR}(a_{ij})}{\max_{ij}(|\log \text{LR}(a_{ij})|) \times 100}.$$

  Points_{ij}에서 분모는 모든 속성의 로그 우도비의 절대값 중에서 가장 큰 속성의 값을 나타내며 이는 가장 영향력 있는 속성값을 의미한다. 분자는 i번째 속성에서 j번째 범주의 로그 우도비를 나타낸다.

- TP 선 및 Probability 선

  TP는 Points_{ij}의 합으로 $\sum_i \text{Points}_{ij}$이 된다. 하지만, 노모그램을 표현하기 위해서 임의의 확률에 해당하는 TP를 선으로 나타내야 하므로 아래의 식을 통하여 표현할 수 있다.

  먼저 각 속성의 범주 안에서 최소 우도비 값들의 합으로 최소 확률값($\min P(C|X)$), 최대 우도비 값들의 합으로 최대 확률값($\max P(C|X)$)을 계산할 있다. 최소 확률값과 최대 확률값, 임의의 확률값을 통하여 이에 대응하는 TP 선을 표현하기 위해 $\log \text{LR}(a_{ij})$값 당 단위 점수(points per unit of $\log \text{LR}(a_{ij})$)와 속성 값들의 합을 이용한다.

(a) $\log \mathrm{LR}(a_{ij})$값 당 단위 점수(points per unit of $\log \mathrm{LR}(a_{ij})$) : $\frac{100}{\max\limits_{ij}(|\log \mathrm{LR}(a_{ij})|}$

(b) 속성 값(a_{ij})들의 합 : $\sum_i \log \mathrm{LR}(a_{ij}) = \log\left(\frac{P(C|X)}{P(\bar{C}|X)}\right) - \log\left(\frac{P(C)}{P(\bar{C})}\right)$

$$\text{총 점수 (TP)} = (a) \times (b) = \frac{100}{\max\limits_{ij}(|\log \mathrm{LR}(a_{ij})|} \times \left(\log\left(\frac{P(C|X)}{P(\bar{C}|X)}\right) - \log\left(\frac{P(C)}{P(\bar{C})}\right)\right)$$

위의 두 식을 이용하여 TP를 계산할 수 있다. 결과적으로 TP의 식을 $\log \mathrm{LR}(a_{ij})$에 대하여 정리할 수 있고, 이는 속성 값이 X일 때 클래스 C가 발생할 최종 확률값 $P(C|X)$는 다음과 같이 정리할 수 있다.

$$P(C|X) = \frac{1}{1 + \exp\left\{-\log\frac{P(C)}{P(\bar{C})} - \frac{\text{Total points}}{\text{Points per unit of } \log \mathrm{LR}(a_{ij})}\right\}}.$$

추가적으로 로지스틱 회귀분석과 비교를 위하여 점수 범위를 $0 \sim 100$점으로 하는 왼쪽 정렬 베이지안 노모그램이 구축 가능하다 (Demsar 등, 2013). 위의 식에서 계산했던 Points_{ij}와 $\log \mathrm{LR}(a_{ij})$값 당 단위 점수는 아래와 같이 계산하여 최종 확률값 $P(C|X)$를 구할 수 있다. 이때의 $\log \mathrm{LR}(a_{ij})$는 범주 간의 편차가 가장 큰 위험 요인을 의미한다.

$$\text{Points}_{ij} = \frac{\log \mathrm{LR}(a_{ij}) - \min\limits_{ij}(\log \mathrm{LR}(a_{ij}))}{\max\limits_{j}(\log \mathrm{LR}(a_{*j})) - \min\limits_{j}(\log \mathrm{LR}(a_{*j}))} \times 100,$$

$$\text{Points per unit of } \log \mathrm{LR}(a_{ij}) = \frac{100}{\max\limits_{ij}(\log \mathrm{LR}(a_{ij})) - \min\limits_{ij}(\log \mathrm{LR}(a_{ij}))}.$$

3. 분석자료 및 검증

3.1. 분석자료와 특징

본 연구에서는 국민건강영양조사(Korean National Health and Nutrition Examination Survey; KNHANES) 6기(2013–2015) 자료를 사용하여 분석을 진행하였다. 국민건강영양조사는 국민의 건강수준, 건강행태, 식품 및 영양섭취 실태에 대한 국가 및 시도 단위의 대표성과 신뢰성을 갖춘 통계를 산출하여 보건정책의 기초자료로 활용하고자 하는 목적으로 시행되었다. 먼저, 만성 폐쇄성 폐질환의 판단은 폐기능 검사 결과를 이용하여 확인하였다. 폐기능 검사에서는 1초간 노력성 호기량(forced expiratory volume; FEV_1), 노력성 폐활량(forced vital capacity; FVC) 등의 수치를 얻을 수 있다. 이를 활용하여 Global Initiative for Chronic Obstructive Lung Disease (GOLD)의 지침에 따라 $\mathrm{FEV}_1/\mathrm{FVC} < 0.7$일 경우 만성 폐쇄성 폐질환이라 판단한다. Seo 등 (2017)에서는 총 16개의 만성 폐쇄성 폐질환 위험 요인으로 교차분석을 진행하여 12개의 요인이 유의하다는 결과를 얻었고, 후진제거법을 통한 로지스틱 회귀분석에서 6개의 위험 요인(성별(sex), 나이(age), 교육수준(education), 흡연 여부(smoking), 결핵(tuberculosis), 천식(asthma))이 모형에 포함되었다. 따라서 본 연구에서는 6개의 위험 요인을 사용하여 분석을 진행하였다. 나이는 65세 미만, 65세 이상으로 나누고, 교육 수준은 중졸 이하, 고졸, 대졸 이상으로 구분하였다. 그리고 흡연 여부는 일생 동안 5갑(100개비) 미만을 피웠거나 피운 적 없는 사람을 비흡연자, 5갑(100개피) 이상 핀 사람을 흡연자로 나누었으며, 마지막으로 동반질환(천식, 결핵)은 '이전에 특정 질병을 의사에게 진단 받은 적이 있는가?'라는 질문 대한 답변으로 질병의 유무를 판단하였다.

조사 참여자는 총 22,948명 중 40세 이상의 폐기능 검사를 실시한 응답자 12,225명 이었으며 이 중 결측값이 포함된 응답자는 제외하여 최종적으로 8,258명을 활용하여 분석을 진행하였다. 또한, 노모그램을 구축한 뒤 이를 검증하기 위해 자료를 무작위로 7 : 3의 비율로 나누어 training data ($n = 5781$)는 모형을 만들어 노모그램을 구축하는데 사용하였고, test data ($n = 2477$)는 검증하는데 사용하였다.

3.2. 노모그램에 대한 검증

구축한 노모그램에 대한 검증은 receiver operating characteristic (ROC) curve의 area under curve (AUC)와 calibration plot을 통하여 이루어 진다.

① ROC curve의 AUC

ROC 곡선의 AUC는 진단 및 예후 모델의 임상적 유용성을 평가하기 위해 사용된다 (Cook, 2008). 그래프는 '민감도'를 수직축, '1 − 특이도'를 수평축으로 하여 선을 그려내고, ROC 곡선이 대각선 위쪽으로 많은 자리를 위치하게 될수록 좋은 성능을 가진 모형이라 판단할 수 있다. AUC가 곡선의 면적을 의미하며 예측 모형의 성능을 측정하는 값으로 사용되고 이 값은 0.5와 1 사이에 존재하여 면적이 넓을수록 값이 1에 가까워지기 때문에 이를 예측 모형의 성능이 좋다고 할 수 있다.

② Calibration plot

Calibration plot은 노모그램으로 예측한 확률과 실제로 관찰된 확률이 얼마나 일치하는가를 확인하는 것이다 (D'Agostino 등, 2001; Nam과 D'Agostino, 2002). 노모그램을 통해 확인한 예측 확률과 실제 관찰된 확률이 정확할 때 45° 각도로 선이 그려지게 되기 때문에, 분석에 의해서 그려진 선이 45° 각도 선에 가까울수록 정확한 예측력을 보인다고 할 수 있다 (Iasonos 등, 2008). 따라서 노모그램에서 보여지는 예측 발병률과 실제 발병률이 얼마나 정확한지를 검증하기 위해 calibration plot을 사용하였다.

4. COPD에 대한 분석 결과

로지스틱 회귀모형과 순수 베이지안 분류기 모형을 이용하여 두 노모그램을 구축하였다. 또한, 두 모형을 비교하기 위해 점수 범위를 수정한 왼쪽 정렬 베이지안 노모그램과 상호작용의 확인을 위한 상호작용항이 포함된 로지스틱 노모그램도 새롭게 구축하여 비교를 통해 유용성을 확인하였다.

4.1. COPD의 로지스틱 노모그램(nomogram using logistic regression model)

Table 4.1이 6개의 위험 요인을 이용하여 로지스틱 회귀분석을 실시한 결과이다. 로지스틱 회귀모형의 적합성을 검정하는 Hosmer-Lemeshow 적합도 검정은 유의확률 0.376으로 모형이 적합하였다. 그리고 회귀모형의 회귀계수를 바탕으로 노모그램을 구축하였다 (Figure 4.1). 먼저 천식의 선이 가장 길어서 COPD의 발병 영향 큰 영향을 미친다고 할 수 있다. 또한, 나이, 성별, 교육수준, 결핵, 흡연 여부 순으로 발병에 영향을 미치는 것을 알 수 있었다. 예를 들어, Figure 4.1에서 표시된 파란색 점들을 통해서 70세 남성이 고졸이고 흡연자이며 천식이 있다면 총 점수는 293점이고 이는 COPD가 일어날 확률이 81%임을 알 수 있다.

4.2. COPD의 베이지안 노모그램(nomogram using naïve Bayesian classifier model)

동일하게 선별된 6가지의 위험 요인을 토대로 순수 베이지안 분류기 모형을 계산하였다 (Table 4.2). 이때 $\log \mathrm{LR}(a_{ij})$값을 모두 점수화하여 그에 해당되는 예측 확률까지 계산하여 노모그램으로 구축하였다

Table 4.1. Multiple logistic regression analysis result using the 6 risk factors

Risk factors		Coefficient	Odd ratio	95% CI	p-value
Sex	Men	1.327	3.771	2.883–4.933	0.000
	Women	·	1	·	
Asthma	No	·	1	·	0.000
	Yes	1.975	7.205	5.013–10.356	
Tuberculosis	No	·	1	·	0.000
	Yes	0.670	1.953	1.428–2.671	
Smoking	No	·	1	·	0.000
	Yes	0.627	1.872	1.459–2.400	
Education	$\leq$ Middle	0.719	2.053	1.614–2.611	0.000
	High school	0.392	1.480	1.156–1.895	0.002
	$\geq$ University	·	1	·	
Age	< 65	·	1	·	0.000
	≥ 65	1.372	3.942	3.297–4.713	

p-value of Hosmer and Lemshow goodness-of-fit test is 0.376.

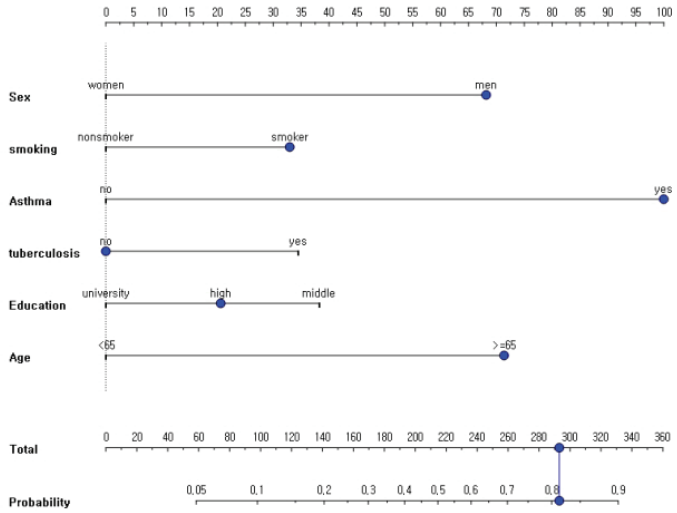


Figure 4.1. Logistic nomogram about chronic obstructive pulmonary disease.

(Figure 4.2). 천식이 양수 중 가장 큰 값을 가지고, 선의 길이가 가장 길기 때문에 COPD의 발병에 가장 영향을 많이 미친다고 할 수 있다. 또한, 음수 중 가장 큰 값은 여성일 때로, 이는 발병에 가장 영향을 덜 미친다 할 수 있다. 예를 들어, Figure 4.2에서 표시된 파란색 점들을 통해서 70세 남성이 고졸이고 흡연자이며 천식이 있다면 총 점수는 230점이고 이는 COPD가 발생할 확률이 85%임을 할 수 있다.

4.3. 두 노모그램의 비교

COPD의 6개 위험 요인을 사용하여 구축된 로지스틱 노모그램과 베이지안 노모그램을 비교해보았다. Figure 4.1에서 로지스틱 노모그램의 점수 범위가 0∼100점이며, Figure 4.2와 같이 베이지안 노모그램의 점수 범위는 −100∼100점으로 범위가 다르게 나타나기 때문에 직접적으로 두 노모그램을 비교하기

Table 4.2. Naïve Bayesian classifier model about COPD

Risk factors		P(a_{ij}|COPD)	P(a_{ij}|Non-COPD)	LR(a_{ij})	log LR(a_{ij})
Sex	Men	0.75	0.39	1.90	0.64
	Women	0.25	0.61	0.42	-0.87
Asthma	No	0.91	0.98	0.92	0.08
	Yes	0.09	0.02	4.70	1.55
Tuberculosis	No	0.90	0.96	0.94	-0.06
	Yes	0.10	0.04	2.41	0.88
Smoking	No	0.32	0.65	0.50	-0.70
	Yes	0.68	0.35	1.93	0.66
Education	$\leq$ Middle	0.55	0.39	1.42	0.35
	High school	0.28	0.34	0.82	-0.20
	$\geq$ University	0.17	0.37	0.62	-0.48
Age	<65	0.42	0.77	0.55	-0.60
	≥ 65	0.64	0.23	2.78	1.02

COPD = chronic obstructive pulmonary disease.

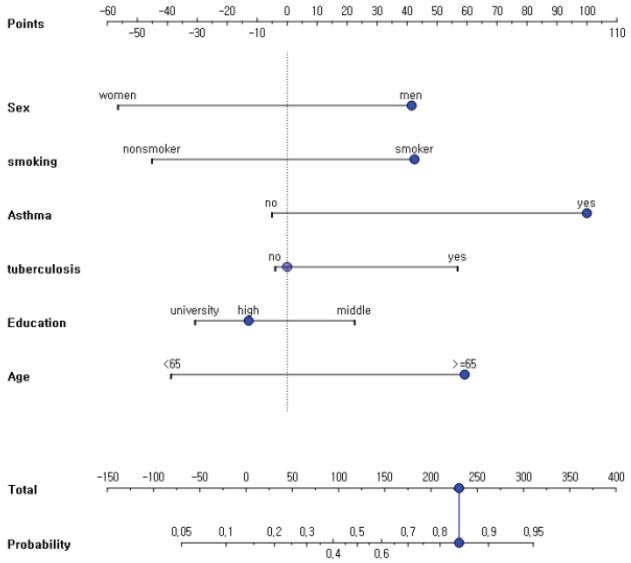
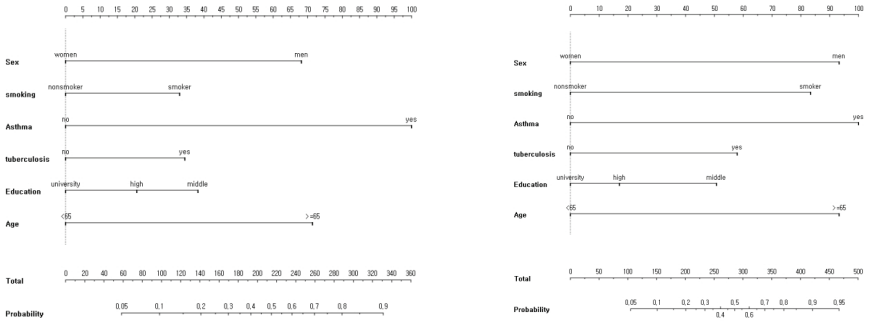


Figure 4.2. Bayesian nomogram about chronic obstructive pulmonary disease.

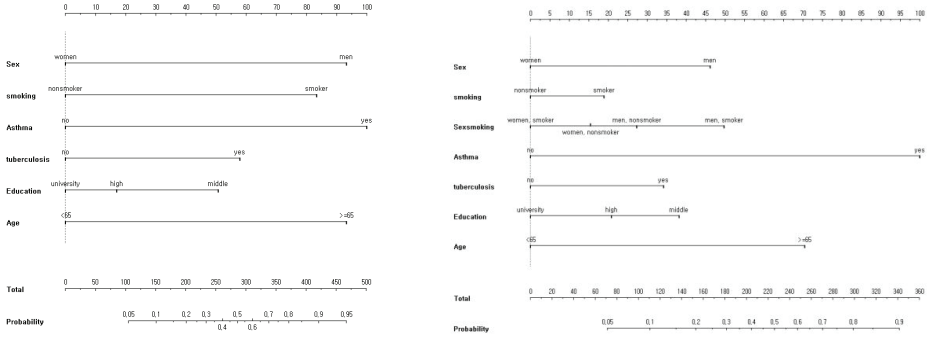
엔 무리가 있다.

그래서 비교를 위하여 로지스틱 노모그램과 점수 범위가 0∼100점으로 동일한 왼쪽 정렬 베이지안 노모그램을 구축하였다 (Figure 4.3(b)). 왼쪽 정렬된 베이지안 노모그램의 위험 요인들은 대체로 로지스틱 노모그램에서보다 높은 점수를 얻었으며, 그 중 성별이 약 26점, 흡연 여부가 약 50점 더 높았다. 이처럼 같은 점수 범위일 경우 위험 요인을 통해 얻게 되는 점수의 차가 존재하기 때문에 베이지안 노모그램은 조건부확률의 계산에 따라 상호작용을 포함한 결과를 보여준다 (Možina 등, 2004a). 따라서 점수의 변화가 가장 컸던 성별과 흡연 여부의 상호작용항을 고려하여 로지스틱 회귀모형을 구하고, 이를 노모그램으로 구축했다 (Figure 4.4(b)).



(a) Logistic nomogram (b) Left-aligned Bayesian nomogram

Figure 4.3. Comparison of the chronic obstructive pulmonary disease nomograms.



(a) Left-aligned Bayesian nomogram (b) Logistic nomogram with interaction

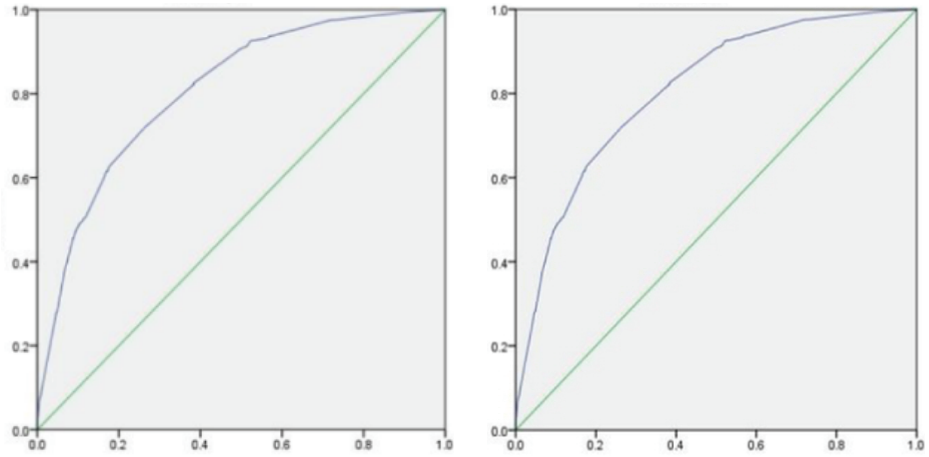
Figure 4.4. Comparison of the chronic obstructive pulmonary disease nomograms with interaction.

위의 비교에서 왼쪽 정렬 베이지안 노모그램과 상호작용항이 포함된 로지스틱 노모그램이 유사하게 나타남을 알 수 있었다. 먼저 두 노모그램에서 천식, 결핵, 교육수준, 나이는 점수가 비슷하게 나타났고, 성별이 남성이며 흡연자의 경우 왼쪽 정렬 베이지안 노모그램은 176.62점, 상호작용항이 포함된 로지스틱 노모그램은 114.67점을 얻었다. 이는 상호작용항이 없을 때 보다 포함되었을 때 점수가 증가함을 확인할 수 있었다. 따라서, 베이지안 노모그램이 상호작용을 포함한 결과를 보여주며, 특정 위함 요인들간의 상호작용이 존재한다면 이를 고려한 로지스틱 노모그램을 표현하는 것이 결과 해석에 효과적이라는 것을 알 수 있었다 (Možina 등, 2004a).

4.4. 노모그램 검증

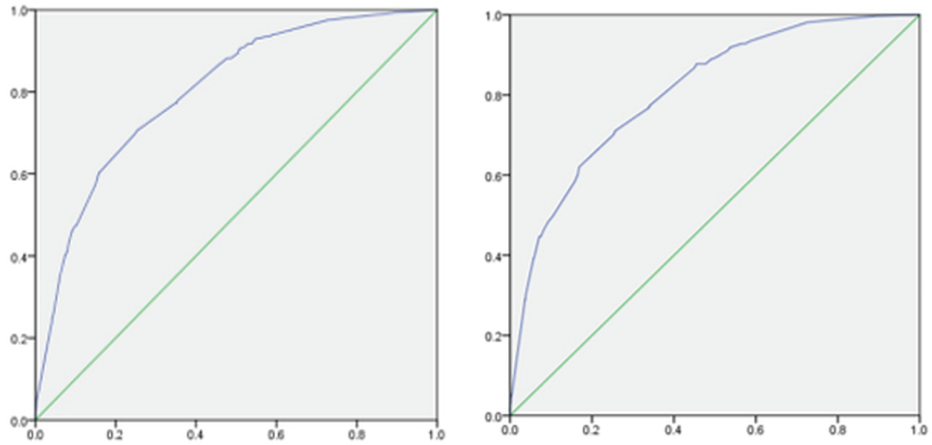
① Discrimination

ROC 곡선의 AUC는 진단 및 예후 모델의 임상적 유용성을 평가하기 위해 사용된다 (Cook, 2008). 그래프는 '민감도'를 수직축, '1 − 특이도'를 수평축으로 하여 선을 그려내고, ROC 곡선이 대각선 위쪽으로 많은 자리를 위치하게 될수록 좋은 성능을 가진 모형이라 판단할 수 있다. Figure 4.5과 Figure 4.6을 보면 각 노모그램의 Training data와 Test data를 이용하여 그려진 ROC 곡선이 제시되어 있다. 먼저 로지스틱 노모그램에서 Training data의 AUC는 0.807 (p = 0.000)이며, Test data의 AUC는



(a) Training data (b) Test data

Figure 4.5. Receiver operating characteristic curve of logistic nomogram.



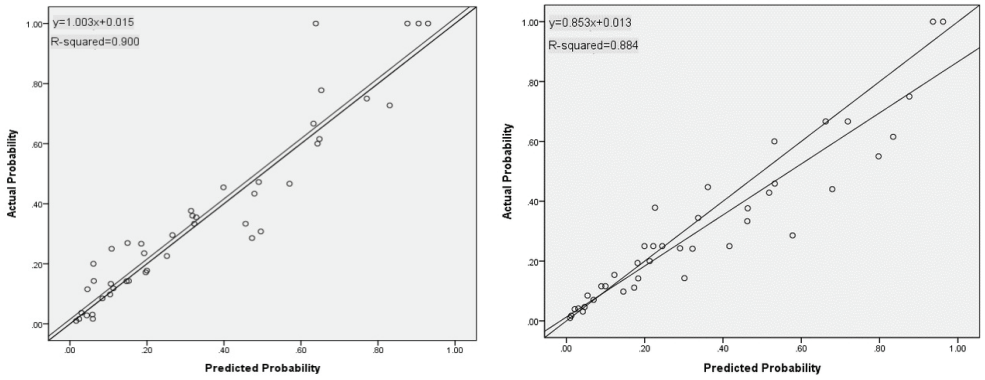
(a) Training data (b) Test data

Figure 4.6. Receiver operating characteristic curve of Bayesian nomogram.

0.811 ($p = 0.000$)로 통계적으로 유의하였다. 베이지안 노모그램에서도 Training data의 AUC는 0.802 ($p = 0.000$)이며, Test data의 AUC는 0.809 ($p = 0.000$)로 통계적으로 유의하였다.

② Calibration plot

Calibration plot은 노모그램으로 예측한 확률과 실제로 관찰된 확률이 얼마나 일치하는가를 확인하는 것이다 (D'Agostino 등, 2001; Nam과 D'Agostino, 2002). 노모그램을 통해 확인한 예측 확률과 실제 관찰된 확률이 정확할 때 45° 각도로 선이 그려지게 되기 때문에, 분석에 의해서 그려진 선이 45° 각도 선에 가까울수록 정확한 예측력을 보인다고 할 수 있다 (Iasonos 등, 2008). Figure 4.7에서 training data를 이용한 각 노모그램의 Calibration plot을 제시하였다. 로지스틱 회귀모형은 44개의 그룹으로 환자들을 나누어 그려진 Calibration plot의 결정계수(R^2)는 0.900이고, 베이지안 노모그램은 40개의



(a) Logistic nomogram (b) Bayesian nomogram

Figure 4.7. Calibration plot.

그룹으로 환자들을 나누어 그린 Calibration plot의 R^2는 0.884로 추정된 직선이 그려졌다. 따라서, 두 노모그램 모두 COPD의 발병률을 예측하는데 적합하다고 판단하였다. 추가적으로, Test data를 이용한 Calibration plot의 R^2는 로지스틱 노모그램이 0.824, 베이지안 노모그램이 0.843으로 적합한 수치를 보였으므로 이는 생략하였다.

5. 결론 및 토의

본 논문에서는 만성 폐쇄성 폐질환의 위험 요인을 이용하여 로지스틱 회귀모형과 순수 베이지안 분류기 모형을 얻었고 이를 시각화하는 통계적 도구인 노모그램을 구축하였다. 또한, 두 노모그램의 유용성을 비교하였다. 분석 자료는 국민건강영양조사 6기(2013–2015) 자료를 사용하여 Training data ($n = 5781$)는 노모그램을 구축하는데 사용하고, Test data ($n = 2477$)은 이를 검증하는데 사용하였다. 위험 요인 선별은 선행 연구를 바탕으로 하여 6개의 위험 요인 (성별, 나이, 교육수준, 흡연여부, 결핵, 천식)으로 분석을 진행하였다 (Seo 등, 2017). 먼저 로지스틱 회귀분석을 실시하여 6개의 위험 요인이 포함된 회귀모형을 얻었다. 이때 모형의 회귀 계수를 이용하여 로지스틱 노모그램을 구축하였다. 천식이 있을 때 100점을 할당 받기 때문에 이는 COPD의 발병에 가장 큰 영향을 준다는 것을 의미한다. 그 이후 나이, 성별, 교육수준, 결핵, 흡연여부의 순서로 발병에 영향을 준다는 것을 시각적으로 확인할 수 있다. 마찬가지로 6개의 위험 요인으로 순수 베이지안 분류기 모형을 얻었고, 이에 따라 베이지안 노모그램을 구축하였다. 먼저 양의 점수를 가장 높게 받은 요인의 범주는 천식의 Yes로 COPD의 발병에 가장 큰 영향을 미친다고 할 수 있다. 반대로 음의 점수가 가장 높은 범주는 성별의 여성이며 이는 COPD의 발병에 가장 영향을 덜 미친다 할 수 있다. 앞서 구축한 로지스틱 회귀모형과 베이지안 회귀모형을 비교해 보았다.

먼저 두 노모그램의 점수 범위가 다르기 때문에 직접적으로 비교를 하기엔 무리가 있어 점수 범위를 0∼100점으로 수정한 왼쪽 정렬 베이지안 노모그램과 로지스틱 노모그램을 비교하였다. 이때, 로지스틱 노모그램보다 왼쪽 정렬 베이지안 노모그램의 위험 요인들 점수가 전체적으로 증가한 것을 확인하였다. 이 결과는 조건부확률을 이용한 베이지안 분류기 모형으로 인하여 위험 요인들 간의 상호작용이 고려되어 나타난 결과임을 알 수 있다 (Možina 등, 2004a). 따라서 로지스틱 노모그램에서도 점수 차이가 가장 많이 났던 성별·흡연여부의 상호작용항을 포함하여 왼쪽 정렬 베이지안 노모그램과 비교를 진행하였

다. 성별과 흡연 여부의 점수를 고려하여 노모그램을 비교하였을 때, 일반 로지스틱 노모그램의 점수보다 상호작용항이 포함됨으로써 점수가 증가하는 모습을 보였다. 따라서 왼쪽 정렬 베이지안 노모그램이 상호작용을 포함한다는 것을 확인하였고, 특정 위험 요인들 간의 상호작용을 표현하기 위해서는 상호작용항을 포함한 로지스틱 노모그램이 더욱 효과적이라는 것을 알 수 있었다. 이러한 비교를 통해서 베이지안 노모그램은 조건부확률을 통해서 계산을 하여 상호작용을 포함하기 때문에 수치적인 점수 확인은 어렵지만, 로지스틱 노모그램의 경우 특정 상호작용항을 Points 선으로 표현 가능했다. 또한 로지스틱 노모그램은 회귀계수의 overfitting 방지를 위해 coefficient shrinkage를 해야 할 수도 있지만 통계 분석 프로그램을 통해 결과를 쉽게 얻을 수 있고, 베이지안 노모그램은 조건부확률을 이용하기 때문에 직접 계산한다는 번거로움이 존재한다.

노모그램은 비전공자가 이해하기에 어려운 통계적 분석 방법들을 그래프로 표현함으로써 의료계 종사자뿐만 아니라 일반인들에게도 큰 도움이 될 수 있다. 그리고 질병의 위험 요인을 한눈에 알 수 있고 개개인의 특징에 따라 점수를 계산하여 예측 확률을 직접 확인할 수 있다. 또한, 특정 위험 요인들 간의 상호작용항을 노모그램에 나타어야 할 경우 로지스틱 노모그램으로 구축하고, 모든 위험 요인들 간의 상호작용이 고려된 노모그램을 구축할 경우에 베이지안 노모그램을 구축할 수 있다. 따라서, 연구에 사용되는 데이터의 특징과 분석 결과를 고려하여 모형을 선택한 뒤 노모그램을 구축하는 것이 효율적일 것이다.

References

Bellazzi, R. and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines, *International Journal of Medical Informatics*, **77**, 81–97.

Cook, N. R. (2008). Satistical evaluation of prognostic versus diagnostic models: beyond the ROC curve, *Clinical Chemistry*, **54**, 17–23.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P., and CHD Risk Prediction Group (2001). Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation, *Journal of the American Medical Association*, **286**, 180–187.

Dayton, C. M. (1992). Logistic regression analysis, *Stat*, 474–574.

Demšar, J., Curk, T., Erjavec, A., *et al.* (2013). Orange: data mining toolbox in Python, *Journal of Machine Learning Research*, **14**, 2349–2353.

Heo, M. H. and Lee, Y. G. (2008). *Data Mining Modeling and Example*, Hannarae, Seoul.

Iasonos, A., Schrag, D., Raj, G. V., and Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis, *Journal of Clinical Oncology*, **26**, 1364–1370.

Jun, H. J. (2015). *Establishment of a nomogram to predict the prognosis of metastatic or recurrent gastric cancer patients* (Master's thesis), Yonsei University, Seoul.

Kim, S. H., Shin, K. H., Kim, H. Y., Cho, Y. J., Noh, J. K., Suh, J. S., and Yang, W. I. (2014). Postoperative nomogram to predict the probability of metastasis in Enneking stage IIB extremity osteosarcoma, *BMC Cancer*, **14**, 666.

Korea Centers for Disease Control and Prevention (2016). Korea Health Statistics 2015: Korea National Health and Nutrition Examination Survey (KNHANES VI-3), Cheongju, from: https://knhanes.cdc.go.kr/knhanes/sub04/sub04_03.do?classType=7

Korean Statistical Information Service (2015). Cause of Death, from: http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B34E01&conn_path=I2

Lee, J. W., Park, M. R., and Yu, H. N. (2005). *Statistical Method for Bioscience Research*, Freedom Academy, Seoul.

Lee, S. C. and Chang, M. C. (2014). Development and validation of web-based nomogram to predict postoperative invasive component in ductal carcinoma in situ at core needle breast biopsy, *Healthcare Informatics Research*, **20**, 152–156.

Možina, M., Demšar, J., Kattan, M., and Zupan, B. (2004a). Nomogram for visualization of naïve Bayesian vlassifier, *Knowledge Discovery in Databases: PKDD 2004*, 337–348.

Možina, M., Demšar, J., Smrke, D., and Zupan, B. (2004b). Nomograms for naïve Bayesian classifiers and how can they help in medical data analysis. In *Proceedings of MEDINFO 2004*, 1762.

Nam, B. H. and D'Agostino, R. B. (2002). Discrimination index, the area under the ROC curve, Huber-Carol C., Balakrishnan N., Nikulin M.S., Mesbah M. (eds), In *Goodness-of-Fit Tests and Model Validity* (pp. 267–279), Birkhauser, Boston.

Seo, J. H. and Lee, J. Y. (2018). Novel nomogram based on risk factors of chronic obstructive pulmonary disease (COPD) using a naive Bayesian classifier model, *Communications in Statistics - Simulation and Computation*, To summitted.

Seo, J. H., Oh, D. Y., Park, Y. S., and Lee, J. Y. (2017). Build the nomogram by risk factors of chronic obstructive pulmonary disease (COPD), *The Korean Journal of Applied Statistics*, **30**, 591–602.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., and Habbema, J. D. (2000). Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets, *Statistics in Medicine*, **19**, 1059–1079.

World Health Organization (2017). The top 10 cause of deat, from: http://who.int/mediacentre/factsheets/fs310/en/

Yang, D. (2014). Build prognostic nomograms for risk assessment using SAS. In *Proceedings of SAS Global Forum 2013*, from: http://support.sas.com/resources/papers/proceedings13/264-2013.pdf.

Zieliński, J., Bednarek, M., and Know the Age of Your Lung Study Group (2001). Early detection of COPD in a high-risk population using spirometric screening, *Chest*, **119**, 731–736.

만성 폐쇄성 폐질환을 이용한 노모그램 구축과 비교

서주현[a] · 이제영[a,1]

[a]영남대학교 통계학과

(2018년 2월 7일 접수, 2018년 3월 26일 수정, 2018년 3월 28일 채택)

요 약

노모그램은 질병의 위험 요인과 예측 확률을 쉽게 이해할 수 있도록 시각적으로 표현하는 통계적 도구이다. 본 논문은 만성 폐쇄성 폐질환(chronic obstructive pulmonary disease)의 위험 요인을 이용하여 로지스틱 회귀모형과 순수 베이지안 분류기 모형의 노모그램을 구축하고 이를 비교하였다. 분석 데이터는 국민건강영양조사 6기(2013–2015)를 이용하여 진행하였다. 총 6개의 위험 요인을 이용하였다. 그리고 로지스틱 회귀모형, 순수 베이지안 분류기 모형과 각각의 구축 방법을 이용하여 만성 폐쇄성 폐질환의 노모그램을 제시하였다. 또한, 구축된 두 노모그램을 비교하여 유용성을 살펴보았다. 마지막으로 ROC curve와 Calibration plot을 통하여 각 노모그램을 검증하였다.

주요용어: 만성 폐쇄성 폐질환, 로지스틱 회귀모형, 순수 베이지안 분류기 모형, 노모그램, 위험 요인

[1]교신저자: (38541) 경북 경산시 대학로 280, 영남대학교 통계학과. E-mail: jlee@yu.ac.kr