

Nomogram building to predict dyslipidemia using a naïve Bayesian classifier model

Min-Ho Kim^a · Ju-Hyun Seo^a · Jea-Young Lee^{a,1}

^aDepartment of Statistics, Yeungnam University

(Received May 23, 2019; Revised June 18, 2019; Accepted June 20, 2019)

Abstract

Dyslipidemia is a representative chronic disease affecting Koreans that requires continuous management. It is also a known risk factor for cardiovascular disease such as hypertension and diabetes. However, it is difficult to diagnose vascular disease without a medical examination. This study identifies risk factors for the recognition and prevention of dyslipidemia. By integrating them, we construct a statistical instrumental nomogram that can predict the incidence rate while visualizing. Data were from the Korean National Health and Nutrition Examination Survey (KNHANES) for 2013–2016. First, a chi-squared test identified twelve risk factors of dyslipidemia. We used a naïve Bayesian classifier model to construct a nomogram for the dyslipidemia. The constructed nomogram was verified using a receiver operating characteristics curve and calibration plot. Finally, we compared the logistic nomogram previously presented with the Bayesian nomogram proposed in this study.

Keywords: Dyslipidemia, risk factor, naïve Bayesian classifier, nomogram

1. 서론

이상지질혈증은 혈액 중에 총콜레스테롤, LDL콜레스테롤, 중성지방이 많이 증가된 상태이거나 HDL콜레스테롤이 감소된 상태를 뜻하는 질환이다. 이는 고혈압이나 당뇨병과 같이 심혈관계 질환의 위험을 증가시키는 질환 중 하나이고, 대표적인 성인병이자 만성질환으로 알려져 있다 (Van den Berg 등, 2009; Qi 등, 2015). 현대인들에게 잘 알려진 고지혈증, 고콜레스테롤혈증, 고중성지방혈증과 같은 질환은 실제로 이상지질혈증에 속하고 유사한 의미로 통용되고 있다. 혈관 질환은 크게 드러나는 증상이 없어 치료 시기를 놓칠 수 있고, 만성 질환인 만큼 발병하게 되면 완치가 어렵기 때문에 지속적인 관리가 굉장히 중요시 되는 질환이다. World Health Organization (WHO)에서 발표한 바로는 세계적으로 심혈관계질환에 의해 사망한 사람이 2015년 기준 31.1%임을 밝혔다 (WHO, 2015). 우리나라에서는 내분비, 영양 및 대사질환에 의한 사망 비율이 2016년 기준 100,000명 당 21.6명으로 발표된 바 있다 (Korean Statistical Information Service, 2016). 최근 발표된 2016년 국민건강통계에 따르면 고중성지방혈증의 경우 남자 24.2%, 여자 10.8%의 유병률을 보였다. 또한, 고콜레스테롤혈증의 30세 이상 유병률이 남자 19.3%, 여자 20.2%로 2005년 이후 지속적으로 증가하고 있다 (Korea Centers for Disease Control and Prevention, 2016–2018). 현대인의 식습관이나 생활패턴이 변화함에 따라 혈관

This study was submitted as a result of the research year of Yeungnam University.

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

관련 질병의 유병률이 계속적으로 높아져 왔으며, 이에 따라 의료계 종사자 및 개개인들이 이상지질혈증에 대해 인지한 후 예방을 위하여 꾸준한 관리를 하는 것이 중요하다 (Committee for Guidelines for Management of Dyslipidemia, 2015).

이전부터 해외뿐만 아니라 국내에서도 이러한 분석방법을 사용하여 이상지질혈증에 대한 위험 요인을 확인하는 연구를 진행해왔다 (Fukui 등, 2011; Qi 등, 2015; Jeon 등, 2017). 선별된 위험 요인으로는 성별, 나이, 학력, 비만, 흡연여부, 신체활동여부, 고혈압, 당뇨병 등이 있었다. 이처럼 위험 요인을 선별하는데 많은 선행 연구들이 진행되어 왔지만, 실제로 의료계 종사자나 개개인들의 경우에는 분석 결과만 보고서 정확한 해석과 이해를 하기엔 어려움이 많다. 이에 대한 문제점을 보완하기 위해 사용되는 도구가 바로 노모그램(nomogram)이다. 노모그램은 현대 의료에서 진단 및 치료의 효과에 근거를 두는 증거 기반 의료를 지향하는데 있어서 진단 및 예후 예측을 도와주는 그래프이다 (Mozina 등, 2004; Seo, 2019). 임의의 수학적 식을 그래프로 표현함으로써 질병과 관련된 위험 요인의 영향력을 쉽게 알 수 있고, 위험 요인들을 종합하여 질병의 발병 예측을 간단히 계산할 수 있다. 해외에서는 다양한 발병 및 재발 예측을 위한 노모그램이 구축되었고, 국내에서는 위암, 전립선암과 당뇨병에 대한 노모그램도 구축된 바 있다 (Bochner 등, 2006; Brennan 등, 2004; Jun, 2015; Lee와 Chang, 2014; Park과 Lee, 2018). 현재 주로 구축된 노모그램은 암과 같은 심각하고 사망률이 높은 주요 질병들에 대해 많이 구축되어 있으며, 한 번 발병하면 치료가 어렵거나 만성적인 질환에 대해서는 예방과 인지가 가장 중요하지만 아직까지 노모그램 구축이 많이 이루어지지 않아 이에 대한 구축의 필요성을 인지하였다. 질병의 위험 요인을 확인하는데 있어 사용하게 되는 통계적 분석방법은 주로 로지스틱 회귀분석, Cox의 비례위험모형을 많이 사용한다. 하지만, 여기서는 베イズ 기법을 사용하여 보다 간단히 노모그램을 구축할 수 있도록 순수 베이지안 분류기(Naïve Bayesian classifier)를 사용한다. 또한, 이상지질혈증의 위험 요인을 확인하는 연구들은 많이 진행되어 왔으나, 이를 종합하여 개개인의 발병 예측 확률을 도출하는 시도는 진행된 적이 없었다. 따라서 본 연구에서는 여러 질환이 동반될 수 있고 심하면 발병까지 일으킬 수 있는 이상지질혈증에 대하여 위험 요인을 선별하고 이를 종합하여 발병 확률을 예측하는 노모그램을 구축하였다. 본 연구의 2절에서는 위험 요인을 선별하기 위하여 사용한 순수 베이지안 분류기와 이를 시각화하는 노모그램에 대해 소개하고 3절은 이상지질혈증에 대한 위험 요인에 대한 설명과 분석 자료에 대해 소개한다. 4절에서는 자료를 통한 분석 결과를 소개하고 모형을 시각화하는 노모그램을 구축한다. 끝으로 5절에서는 분석의 결과 및 결론에 대해 설명한다.

2. 통계적 방법

2.1. 순수 베이지안 분류기(Naïve Bayesian classifier)

순수 베이지안 분류기 모델은 예측 모델을 구축하는데 있어서 가장 간단하면서 강력한 기법 중 하나이다 (Mozina 등, 2004). 복잡한 통계 기법을 이용하여 예측 확률을 계산하는 방법과는 다르게, 이 모델은 속성 값들이 서로 독립이라는 가정 하에 베イズ 기법을 사용하여 어떤 클래스에 대한 확률을 계산하여 보다 쉽게 사용할 수 있다 (Han 등, 2012). 이때 클래스 C 는 어떤 대상에 대한 목표 범주를 의미하며, 각 속성 a_1, a_2, \dots, a_m 은 대상에 대한 속성값을 의미한다. 따라서 $P(C)$ 는 목표 범주가 발생할 확률이며 이에 따라 $P(\bar{C}) = 1 - P(C)$ 는 목표 범주가 아닌 \bar{C} 가 발생할 확률을 뜻하게 되고, 속성값은 $X = \{a_1, a_2, \dots, a_m\}$ 로 표현된다. 속성 값이 $X = \{a_1, a_2, \dots, a_m\}$ 일 때, 클래스 C 가 발생할 조건부 확률은 베イズ 정리를 사용하여 계산된다.

$$P(C|X) = \frac{P(X \cap C)}{P(X)} = \frac{P(X|C)P(C)}{P(X)} = \frac{\prod_i P(a_i|C)P(C)}{P(X)}$$

이때, $P(C|X)$ 의 오즈(Odds)는 다음과 같이 계산된다.

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

이는 속성 값이 X 일 때 C 가 일어날 확률을 다른 범주인 \bar{C} 가 일어날 확률로 나눈 값으로, 위의 식에서 양변에 로그(log)값을 취하여 계산하면 결과는 다음과 같이 정리할 수 있다.

$$\begin{aligned} \log \frac{P(C|X)}{P(\bar{C}|X)} &= \log \frac{P(C)}{P(\bar{C})} + \log \prod_i \frac{P(a_i|C)}{P(a_i|\bar{C})} \\ &= \log \frac{P(C)}{P(\bar{C})} + \sum_i \log \frac{P(a_i|C)}{P(a_i|\bar{C})} \\ &= \log \frac{P(C)}{P(\bar{C})} + \sum_i \log \text{OR}(a_i). \end{aligned}$$

이때, $\text{OR}(a_i) = P(a_i|C)/P(a_i|\bar{C}) = \{P(C|a_i)/P(\bar{C}|a_i)\}/\{P(C)/P(\bar{C})\}$ 는 사후오즈와 사전오즈의 비 형태로써 Odds ratio (OR)이라 표기하였다.

따라서, 위의 식을 속성 값이 X 일 때 클래스 C 가 발생할 최종 확률 $P(C|X)$ 에 대하여 정리하면 다음과 같이 계산된다.

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

2.2. 노모그램(nomogram) 구축

노모그램은 알고자 하는 결과에 대한 가능성을 예측하고자 만들어진 통계학적 도구로 환자들의 특징들을 바탕으로 하여 질병에 대한 위험 요인들을 표현한다. 통계적 모형에 직접 대입하여 계산하는 것이 아닌 그래픽으로 표현되어 이해가 훨씬 쉽다 (Mozina 등, 2004; Seo, 2019). 노모그램은 각 속성 값에 할당된 점수를 보여주는 Point 선, 각 속성값의 범주를 점수화 한 Risk factor 선, 일정한 구간으로 나누어져 사건의 확률을 바로 알 수 있게 하는 Probability 선, 그리고 Probability 선에 대응되는 노모그램 점수 합을 나타내는 Total point 선이 있다. 순수 베이시안 분류기를 사용하면 최종 확률 $P(C|X)$ 는 각 속성 값의 $\log \text{OR}(a_i = j)$ 의 합으로 표현되어지기 때문에 이를 이용하여 노모그램을 만들 수 있다.

- Point 선

Point 선은 -100~100점으로 구성한다.

- Risk factor 선

속성 값 $a_i = j$ 에 따라, $\log \text{OR}(a_i = j)$ 는 다음과 같이 계산된다.

$$\log \text{OR}(a_i = j) = \log \frac{P(a_i = j|C)}{P(a_i = j|\bar{C})}.$$

그리고 각 속성의 Point_{ij} 는 $\log \text{OR}(a_i = j)$ 를 이용하여 계산된다.

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

분모는 모든 속성의 $\log \text{OR}(a_i = j)$ 의 절대값 중에서 가장 큰 속성의 값을 나타내며 이는 가장 영향력 있는 속성 값을 의미한다.

- Probability 선

Probability 선은 0에서 1까지 0.1 단위로 나누어 구성한 후 노모그램 하단에 배치한다.

- Total point 선

위의 식에 따라 계산한 $Point_{ij}$ 값을 이용하여 해당되는 범주의 Point 값을 합하면 Total point 값을 얻을 수 있게 된다. 하지만, 노모그램을 표현하기 위해서 0과 1 사이에 해당하는 확률에 대응하는 Total point 값을 계산해야 한다. 그 과정은 다음과 같다.

Total point는 모든 $Point_{ij}$ 의 합이므로 다음과 같이 계산된다.

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

순수 베이저안 분류기 모형 식을 $\sum_{i,j} \log \text{OR}(a_i = j)$ 에 대해 정리한 후 대입하면 아래와 같다.

$$\text{Total point} = \frac{100}{\max_{i,j} |\log \text{OR}(a_i = j)|} \times \left(-\log \left(\frac{1}{P(Y=1|X=x)} - 1 \right) - \log \frac{P(Y=1)}{1-P(Y=1)} \right).$$

그 후, $P(Y=1|X=x)$ 에 해당 Probability 선의 값을 대입하면 Total point 선을 구성할 수 있다.

2.3. 노모그램의 검증(validation about nomogram)

Receiver operating characteristic (ROC)의 곡선 아래의 면적(area under the ROC curve; AUC)과 Calibration plot을 통해서 순수 베이저안 분류기를 사용한 노모그램을 검증한다.

① ROC curve

ROC 곡선은 진단법의 정확성을 비교하는 방법으로 널리 사용되는 방법 중 하나이다 (Akobeong, 2007). X축에는 ‘민감도’, Y축에는 ‘1 - 특이도’로 하여 그래프를 그리고, 이때 대각선을 기준으로 곡선 아래 면적이 0.5로써 이 대각선보다 ROC 곡선이 위에 있을 경우 좋은 모형이라 예측하게 된다. 이 값은 0.5부터 1사이로 갖게 되며, 1에 가까워 질수록 곡선도 넓어지고 예측한 모형이 좋다는 것을 검증할 수 있다. 따라서 본 연구에서는 노모그램의 모형을 검증하기 위해 ROC 곡선을 사용하였다.

② Calibration plot

Calibration plot은 노모그램을 통해서 계산된 예측 확률과 실제로 관찰된 확률이 얼마나 일치하는지를 확인하는 방법이다 (D’Agostino 등, 2001). 도구를 통하여 예측된 확률과 실제로 관찰된 확률이 정확하게 일치할 경우에는 45° 각도의 선이 그려지게 되기 때문에 예측 확률과 실제 확률이 비슷할수록 45° 각도 선에 가깝게 표현된다 (Iasonos 등, 2008). 이에 따라 본 연구에서도 노모그램으로 예측한 확률과 실제 발병 확률이 얼마나 일치하는지 검증하기 위해 Calibration plot을 사용하였다.

3. 연구 자료 특성

국민건강영양조사 6기(2013-2015)와 7기 1차년도 (2016)이다 (Korean Centers for Disease Control and Prevention, 2013-2016)를 사용하여 연구를 진행하였다. 국민건강영양조사는 국민의 건강수준, 건강행태, 식품 및 영양섭취 실태에 대한 국가 및 시도 단위의 대표성과 신뢰성을 갖춘 통계를 산출하고, 이를 통하여 보건정책의 기초자료로 활용되고 있다. 6기 자료는 총 22,948명이 조사 대상자였고, 본 연구에서는 20세 이상 기준으로 의사진단 변수, 약 복용 변수와 혈액검사 수치에 대한 결측값을 가진 대

상은 제외하되, 그 이외의 분석에서 사용되는 변수의 결측값은 최빈값 또는 평균값을 통해 보정을 함으로써 14,295명을 이용하여 분석을 진행하였다. 7기 1차년도 또한 총 8,150명이 조사 대상자였으며 앞서 동일한 기준에 따라 5,609명의 조사참여자를 이용하였다. 본 연구에서는 전향적 연구의 형태로써 먼저 모집된 6기 자료로써 노모그램을 구축하였고, 이후 모집된 7기 자료는 연구 결과를 검증하는데 사용되었다. 분석에 사용된 프로그램은 SAS 9.4와 IBM SPSS Statistics 23을 사용하였다.

이상지질혈증의 진단기준은 한국지질·동맥경화학회에서 발표한 가지 4기준과 추가적인 기준 2가지를 고려하여 이 중 하나라도 해당되면 이상지질혈증이라 진단하였다. 1) 총콜레스테롤(total cholesterol) $\geq 240\text{mg/dL}$, 2) 중성지방(triglyceride) $\geq 200\text{mg/dL}$, 3) LDL 콜레스테롤 $\geq 160\text{mg/dL}$, 4) HDL 콜레스테롤 $< 40\text{mg/dL}$, 5) 의사로부터 이상지질혈증 진단을 받았을 경우, 6) 이상지질혈증 치료약물 복용할 경우 (The Korean Society of Lipid and Atherosclerosis, 2018). 이상지질혈증의 위험 요인으로는 선행 연구들을 참고하여 성별, 나이, 소득, 학력, 결혼여부, 비만, 신체활동, 음주, 흡연, 당뇨병, 고혈압, 심근경색 및 협심증으로 선별하였다. 나이는 20-39세, 40-55세, 56-69세, 70세 이상의 4개 범주로, 소득은 월평균 가구소득에 따라 $< 25\%$, 25-50%, 50-75%, $\geq 75\%$ 로 나누었다. 학력은 초졸 이하, 중졸, 고졸, 대졸 이상으로 나누었으며, 비만의 경우 BMI 수치에 따라 18.5 미만일 경우 저체중, 18.5 이상 25 미만일 경우 정상, 25 이상일 경우 비만으로 범주를 나누었다. 신체활동은 걷기일수가 주 5일 이상이며 걷기지속시간이 1회 30분 이상일 경우 신체활동군으로 분류하였고 이외의 조사참여자들은 비신체활동군으로 분류하였다. 흡연은 비흡연, 과거흡연, 현재흡연으로 나누었다 (Committee for Guidelines for Management of Dyslipidemia, 2015). 음주는 일평균 음주량으로써 선행연구의 분류에 따라 비음주군, 가벼운 음주군, 중등도 음주군, 고도음주군으로 분류하였다 (Jeon 등, 2017). 당뇨병의 경우에는 공복혈당이 126mg/dL 이상이거나 혈당강하제 또는 인슐린을 사용할 경우를 기준으로 하였고, 고혈압은 수축기 혈압이 140mmHg , 이완기 혈압이 90mmHg 이상이거나 고혈압 약 복용 중일 경우로 기준하였다. 마지막으로 심근경색 및 협심증 (심장질환)은 의사에 의해 진단받았을 경우를 기준으로 분류하였다. 추가적으로 이상지질혈증의 경우 개인의 식습관에 따라 많은 영향을 받는 질환이지만, 국민건강영양조사 중 영양 조사는 24시간 회상 조사로 제시하여 개개인의 지속적인 식습관을 유추하기에는 무리가 있을 것으로 판단하였다.

4. 적용 결과

본 연구에서는 이상지질혈증의 여부와 위험 요인간의 관련성을 알기 위해 교차분석을 실시하였다. 결과를 바탕으로 순수 베이저안 분류기를 통해 모형을 얻었으며, 최종적으로 모형을 시각화하기 위하여 노모그램을 구축하였다. 마지막으로 도구에 대한 검증을 위해 ROC 곡선과 Calibration plot을 사용하였다.

4.1. 교차분석(Chi-squared test)에 의한 위험요인 선별

먼저 본 연구에서 이상지질혈증의 위험 요인으로 선별된 변수에 대하여 이상지질혈증의 발병 여부와 독립적으로 차이가 있는지 알기 위하여 교차분석을 진행하였다 (Table 4.1). 총 12개의 위험 요인에 대하여 유의확률은 모두 0.000으로 유의한 것을 확인할 수 있었다. 또한 검정통계량 값을 통해서 나이, 비만, 당뇨병, 학력, 결혼여부 등의 순으로 큰 차이를 보이는 것을 알 수 있다.

4.2. 순수 베이저안 분류기를 이용한 이상지질혈증의 노모그램 구축

최근 선행연구들에 따라 선별된 위험 요인들로는 성별, 나이, 교육수준, 비만상태, 신체활동상태, 흡연, 당뇨병, 고혈압, 심장질환 등이었다. 따라서 위의 교차분석에서 분석한 12개의 위험요인 중 성별, 나이,

Table 4.1. Chi-squared test about 12 risk factors

Risk factors		Dyslipidemia(%)	Non-dyslipidemia(%)	χ^2	<i>p</i> -value
Sex	Female	2984(31.7)	5183(68.3)	272.4	0.000
	Male	2938(45.9)	3190(54.1)		
Age	20-39	897(23.2)	3082(76.8)	916.2	0.000
	40-55	1740(41.1)	2696(58.9)		
	56-69	2079(55.7)	1655(44.3)		
	≥ 70	1206(56.9)	940(43.1)		
Income	< 25%	1413(51.7)	1205(48.3)	147.2	0.000
	25-50%	1546(39.3)	2078(60.7)		
	50-75%	1441(35.2)	2484(64.8)		
	≥ 75%	1522(35.8)	2606(64.2)		
Education	≤Elementary school	1639(57.9)	1212(42.1)	548.7	0.000
	Middle school	846(52.5)	779(47.5)		
	High school	1718(39.7)	2440(60.3)		
	≥University	1719(29.7)	3942(70.3)		
Marriage status	Single	486(23.5)	1597(76.5)	320.2	0.000
	Marriage	5436(42.9)	6776(57.1)		
Obesity status	Lower weight	72(10.5)	480(89.5)	824.3	0.000
	Normal	3151(31.8)	5862(68.2)		
	Obesity	2699(55.8)	2031(44.2)		
Exercise status	Non-physical activity	3761(40.8)	4981(59.2)	27.9	0.000
	Physical activity	2161(35.7)	3392(64.3)		
Alcohol status	No drink	2633(42.0)	3199(58.0)	85.9	0.000
	Low drink	2492(34.7)	4142(65.3)		
	Middle drink	624(45.0)	778(55.0)		
	High drink	173(42.9)	254(57.1)		
Smoking	Non-smoker	3286(33.2)	5443(66.8)	220.0	0.000
	Past smoker	1314(44.8)	1544(55.2)		
	Current smoker	1322(47.2)	1386(52.8)		
Diabetes	No	4623(34.9)	7902(65.1)	699.2	0.000
	Yes	1299(73.6)	471(26.4)		
Hypertension	No	3585(32.7)	7013(67.3)	248.1	0.000
	Yes	2337(62.9)	1360(37.1)		
Heart disease	No	5654(38.2)	8260(61.8)	108.7	0.000
	Yes	268(69.1)	113(30.9)		

교육수준, 비만상태, 신체활동상태, 흡연, 당뇨병, 고혈압, 심장질환으로 9개의 위험 요인 사용하여 순수 베이지안 분류기의 모형을 얻었다. 각 위험 요인의 빈도를 이용하여 직접 $\log \text{OR}(a_i = j)$ 을 계산하였다 (Table 4.2). 이를 토대로 점수화하여 이상지질혈증의 발병률을 예측하는 노모그램을 구축하였다 (Figure 4.1). $\log \text{OR}(a_i = j)$ 의 절대값이 클수록 점수의 절대값 또한 커지는 경향을 보이고 있으며, 당뇨병, 심장질환, 고혈압, 교육수준이 초졸 이하일 경우, 비만 등의 순으로 양의 점수가 높았다. 또한, 저체중일 경우, 20-39세, 교육수준이 대졸 이상일 경우, 고혈압이 없을 경우 등의 순으로 음의 점수가 높았다. 순수 베이지안 분류기의 결과를 그래픽으로 보여주는 노모그램을 이용하면 해당 질병에 어떤 위험 요인들이 있는지, 얼마나 발병에 영향을 미치는지를 한 눈에 확인 가능하다. 예를 들어, Figure 4.2를 통해서 고졸의 55세 남성이 비만이며, 신체활동을 하지 않고, 과거에 흡연했으며 심근경색 및 협

Table 4.2. Naive Bayesian classifier model results about dyslipidemia

Risk factors		$P(a_i = j Yes)$	$P(a_i = j No)$	$\log OR(a_i = j)$	p -value
Sex	Female	0.496	0.619	-0.21	-13
	Male	0.503	0.381	0.26	17
Age	20-39	0.151	0.368	-0.89	-57
	40-55	0.294	0.322	-0.09	-6
	56-69	0.351	0.198	0.57	37
	≥70	0.204	0.112	0.60	38
Education	≤Elementary school	0.277	0.145	0.65	42
	Middle school	0.143	0.093	0.43	28
	High school	0.290	0.291	0.00	0
	≥University	0.290	0.471	-0.48	-31
Obesity status	Lower weight	0.012	0.057	-1.55	-100
	Normal	0.532	0.700	-0.27	-18
	Obesity	0.456	0.243	0.63	41
Exercise status	Non-physical activity	0.635	0.595	0.07	4
	Physical activity	0.365	0.405	-0.10	-7
Smoking	Non-smoker	0.555	0.650	-0.16	-10
	Past smoker	0.222	0.184	0.18	12
	Current smoker	0.223	0.166	0.30	19
Diabetes	No	0.781	0.944	-0.19	-12
	Yes	0.219	0.056	1.36	88
Hypertension	No	0.605	0.838	-0.32	-21
	Yes	0.395	0.162	0.89	57
Heart disease	No	0.955	0.987	-0.03	-2
	Yes	0.045	0.013	1.21	78

심증과 고혈압이 있을 경우 총 점수는 191점으로 이상지질혈증의 발병 예측 확률은 약 93%라는 것을 알 수 있다.

4.3. 구축된 노모그램의 검증

위와 같이 순수 베이저안 분류기를 사용한 이상지질혈증의 노모그램을 검증하기 위해 ROC 곡선과 Calibration plot을 사용하였다.

① ROC curve (Area under ROC curve)

국민건강영양조사 6기(2013-2015) 자료를 사용해서 노모그램을 구축하였고, 7기 제1차년도(2016)를 통하여 구축한 노모그램을 검증하는데 사용하였다. 각 자료들의 ROC 곡선을 그렸을 때 결과는 다음과 같다 (Figure 4.3). Training data (2013-2015)의 ROC curve에서 AUC 값은 0.737이었고, Test data (2016)의 ROC curve에서 AUC 값은 0.715이었다. 이에 따라 구축된 노모그램이 굉장히 유용하다는 것을 검증할 수 있었다.

① Calibration plot

두 번째로 노모그램을 검증한 방법은 예측 확률과 실제 확률을 비교할 수 있는 Calibration plot을 사용하였다. 먼저 6기 자료를 사용하여 예측 확률들을 동일한 확률끼리 그룹지어 총 100개의 그룹으로 나누어졌고 이에 대한 그래프는 Figure 4.4(a)와 같았다. 또한 결정계수(R^2)가 0.815이었기

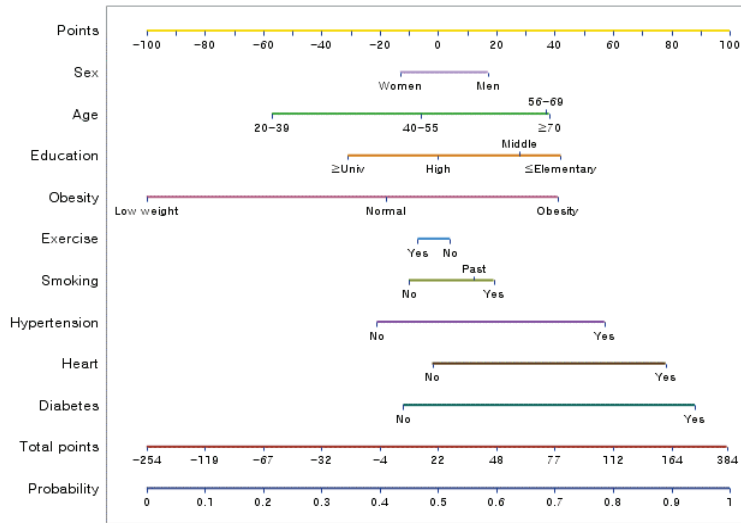


Figure 4.1. Bayesian nomogram for dyslipidemia.

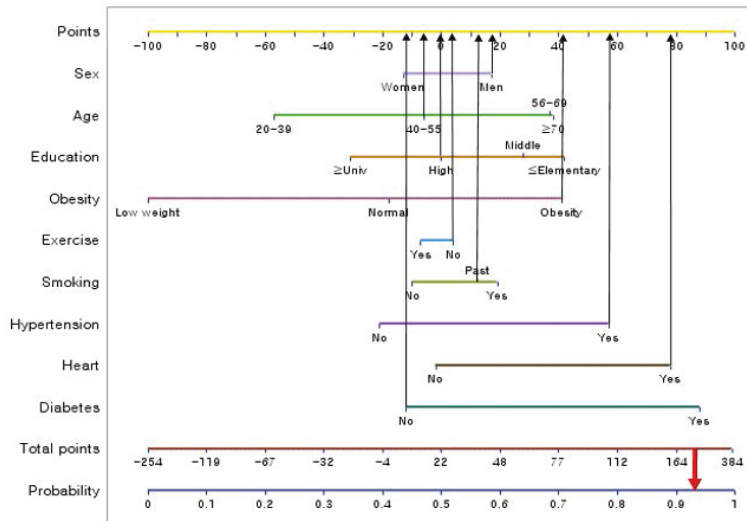


Figure 4.2. Application for one case using Bayesian nomogram.

때문에 굉장히 예측이 잘 되었음을 확인할 수 있었다. 마찬가지로 7기 자료를 사용한 calibration plot에 대해서도 그룹이 98개로 나누어져 Figure 4.4(b)와 같이 그려졌고, 결정계수(R^2)가 0.803로 나왔기 때문에 본 연구에서 구축된 노모그램이 유용하다는 것을 검증할 수 있었다.

5. 결론 및 토의

본 연구에서는 심혈관계 질환의 위험 요인이라 알려진 질병 중 하나인 이상지질혈증에 대하여 발병과 관련된 위험 요인을 확인하였고, 예측 발병 확률을 계산할 수 있는 노모그램을 구축하였다. 사용한 분

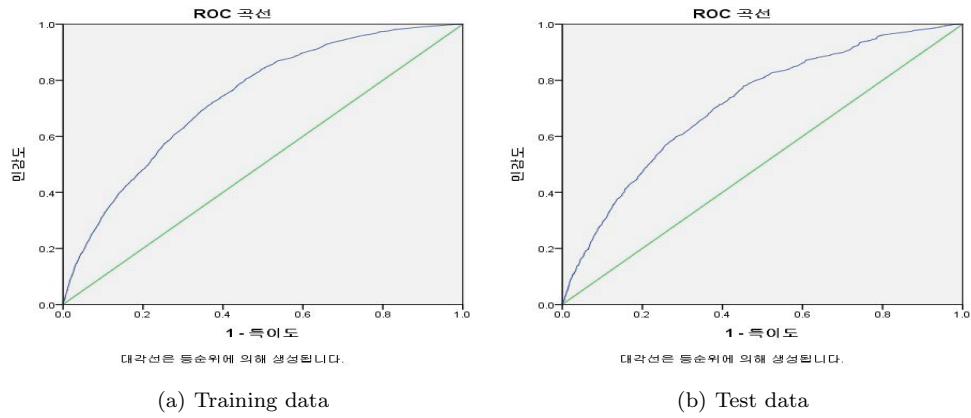


Figure 4.3. ROC curve for the constructed Bayesian nomogram. ROC = receiver operating characteristic.

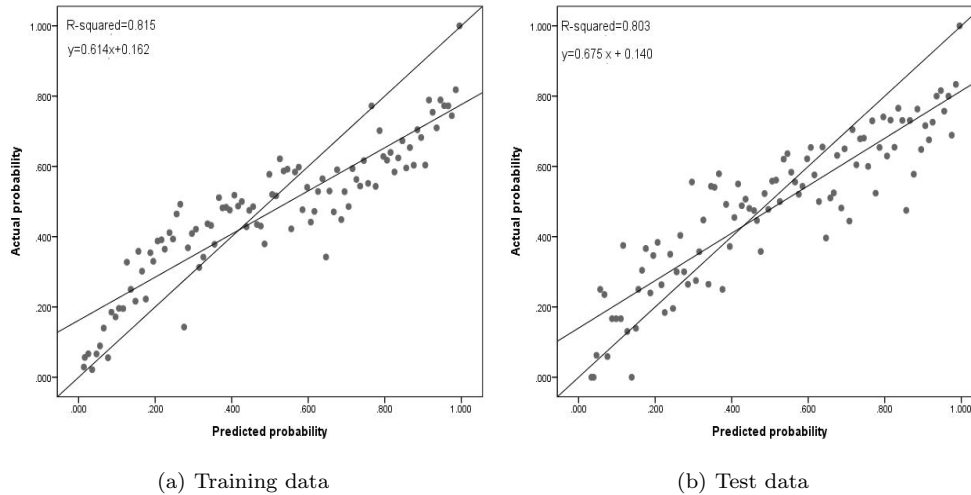


Figure 4.4. Calibration plot for the constructed Bayesian nomogram.

식자료는 국민건강영양조사 6기(2013–2015), 7기 제1차년도(2016)로 사용하면서 6기 자료는 노모그램을 구축하는데 사용하였고, 이후 조사된 7기 제1차년도 자료를 사용해서 노모그램의 검증을 진행하였다. 이상지질혈증의 위험 요인으로 성별, 나이, 소득, 학력, 결혼여부, 비만, 신체활동, 음주, 흡연, 당뇨병, 고혈압, 심근경색 및 협심증으로 12가지를 선별하여 진행하였다. 이상지질혈증의 발병 여부와 위험 요인들의 차이가 있는지 확인하기 위하여 교차분석을 진행하였고, 모든 위험 요인들이 유의하였다. 이전 선행 연구들과 비교하였을 때 이상지질혈증의 위험 요인이 흡사함을 알 수 있었다. 이러한 결과를 이용하여 선행 연구에서 많이 언급되던 9개의 위험 요인(성별, 나이, 교육수준, 비만상태, 신체활동상태, 흡연, 당뇨병, 고혈압, 심장질환)을 순수 베이지안 분류기에 대입하여 결과를 얻었다. 순수 베이지안 분류기를 이용한 노모그램이 Figure 4.1과 같았다 양의 점수로써 당뇨병, 심장질환, 고혈압 등의 순으로 점수가 높았고, 음의 점수는 저체중, 20–39세, 대졸 이상 등의 순으로 높았다. 이전에 구축되었던 상호작용이 포함된 로지스틱 노모그램과 비교해보았을 때, 점수의 범위가 -100~100 point로 긍정적/부정적 영향도 모두 고려하면서 분류기 만들기가 용이하고 위험 요인의 정보가 없더라도 충분히 예측 도구로써

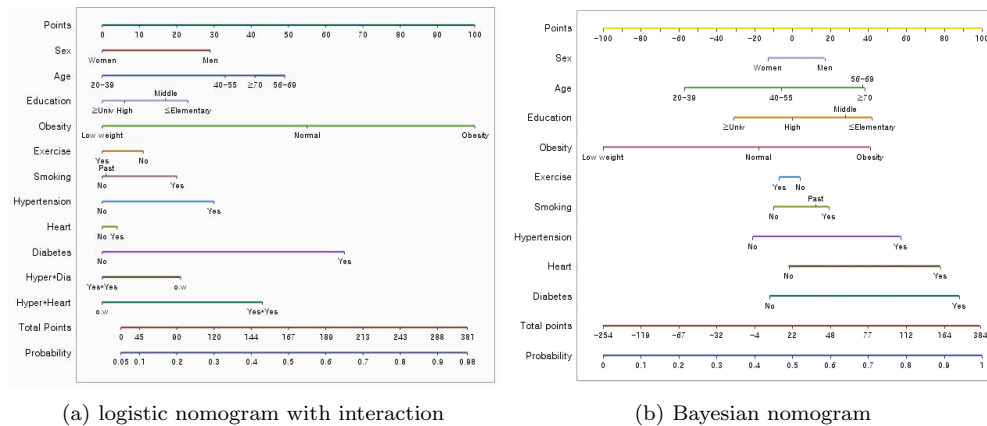


Figure 5.1. Comparison of two dyslipidemia's nomograms.

의 활용성이 높음을 알 수 있었다. 추가적으로 베이저안 노모그램이 유용한지 확인하기 위하여 ROC 곡선과 Calibration plot을 사용하여 검증하였다. 먼저 ROC 곡선을 Training data와 Test data를 사용하여 그려보았을 때, 각각 AUC 값이 0.737, 0.715으로 높은 수치를 확인하였다. 또한, 예측 확률과 실제 확률을 비교하는 Calibration plot을 그려보았을 때, Training data의 결정계수(R^2)는 0.815이었고 Test data의 결정계수(R^2)는 0.803였으며 두 그래프 모두 직선이 45° 각도에 비슷하게 그려졌다. 이에 따라 본 연구에서 구축한 이상지질혈증의 노모그램이 유용한 것을 확인하였다.

추가적으로 Figure 5.1에서는 상호작용 변수가 포함된 로지스틱 노모그램과 순수 베이저안 분류기를 이용한 노모그램을 비교해보았다 (Seo와 Lee, 2018; Seo, 2019). 먼저, 로지스틱 노모그램의 경우 점수의 범위가 0~100 point이며, 베이저안 노모그램은 -100~100 point이다. 이러한 이유는 로지스틱 회귀모형을 얻을 때 범주형 변수에 대해서 기준범주를 두고 만들어지기 때문에 기준범주에 0을 부여하게 된다. 따라서, 로지스틱 노모그램은 기준범주보다 다른 범주에 속할 때의 발병여부를 표현하고 있지만, 위험 요인에 대한 모든 정보를 필요로 하고 있다. 베이저안 노모그램의 경우 각 범주마다 모두 점수 환산하면서 위험 요인에 대한 정보가 없을 경우 0점을 부여할 수 있다. 또한, 로지스틱 회귀분석의 경우 모든 위험요인들 간의 상호작용을 고려하여 나타내기엔 무리가 있으며, 일부의 상호작용만 Figure 5.1(a)와 같이 표현해낼 수 있다. 반면에 순수 베이저안 분류기의 경우 계산 상에서 조건부 확률을 이용하여 계산하기 때문에 어느 정도 상호작용을 고려하여 결과를 나타내준다.

본 연구에서는 우리나라 국민에 대한 건강 행태를 파악할 수 있도록 국민건강영양조사의 가장 최근 자료인 6기(2013-2015), 7기 제 1차년도(2017)를 사용하여 분석을 진행하였다. 이를 통해 이상지질혈증의 위험 요인을 확인하면서 발병 예측 확률을 그래프로 함께 확인할 수 있는 노모그램을 순수 베이저안 분류기를 이용하여 구축하였다. 추가적으로 로지스틱 회귀모형을 이용한 노모그램과 비교하며 훨씬 효율적이라는 것을 확인할 수 있었다. 혈관질환은 피검사를 하지 않는 이상 쉽게 알아차릴 수 있는 증상이 없기 때문에 의료계 종사자나 개개인들이 자신의 특징을 토대로 자가진단을 통해 예방할 수 있도록 스스로 관리가 가능할 것으로 생각된다.

References

- Akobeng, A. K. (2007). Understanding diagnostic tests 3: receiver operating characteristic curves, *Acta Paediatrica*, **96**, 644-647.

- Bochner, B. H., Kattan, M. W., and Vora, K. C. (2006). Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer, *Journal of Clinical Oncology*, **24**, 3967–3972.
- Brennan, M. F., Kattan, M. W., Klimstra, D., and Conlon, K. (2004). Prognostic nomogram for patients undergoing resection for adenocarcinoma of the pancreas, *Annals of Surgery*, **240**, 293.
- Committee for Guidelines for Management of Dyslipidemia (2015). 2015 Korean guidelines for management of dyslipidemia, *Journal of Lipid and Atherosclerosis*, **4**, 61–92.
- D’Agostino Sr, 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, *Jama*, **286**, 180–187.
- Fukui, M., Tanaka, M., Toda, H., Senmaru, T., Sakabe, K., Ushigome, E., Asano, M., Yamazaki, M., Hasegawa, G., Imai, S., and Nakamura, N. (2011). Risk factors for development of diabetes mellitus, hypertension and dyslipidemia, *Diabetes Research and Clinical Practice*, **94**, e15–e18.
- Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed), Elsevier, Amsterdam.
- 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.
- Jeon, M. Y., Choi, W. H., and Seo, Y. M. (2017). Risk factors of dyslipidemia and related factors of medication adherence in Korea Adults: KNHANES 2013–2015, *Journal of Korean Biological Nursing Science*, **19**, 131–140.
- Jun, H. J. (2015). Establishment of a nomogram to predict the prognosis of metastatic or recurrent gastric cancer patients, Yonsei University, Seoul.
- Korea Centers for Disease Control and Prevention (2016–2018). Korea Health Statistics 2016: Korea National Health and Nutrition Examination Survey (KNHANES VII-1), Cheongju. Available from: <https://knhanes.cdc.go.kr>
- Korean Statistical Information Service (2016). Cause of Death. Available from: <http://kosis.kr/>
- 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.
- Mozina, M., Demsar, J., Smrke, D., and Zupan, B. (2004). Nomograms for Naive Bayesian Classifiers and How Can They Help in Medical Data Analysis, *MEDINFO 2004*, 1762.
- Park, J. C. and Lee, J. Y. (2018). How to build nomogram for type 2 diabetes using a naive Bayesian classifier technique, *Journal of Applied Statistics*, 1–13.
- Qi, L., Ding, X., Tang, W., Li, Q., Mao, D., and Wang, Y. (2015). Prevalence and risk factors associated with dyslipidemia in Chongqing, China. *International Journal of Environmental Research and Public Health*, **12**, 13455–13465.
- Seo, J. H. (2019). Nomogram build for predicting the incidence of chronic diseases - dyslipidemia and chronic obstructive pulmonary disease (Master’s thesis), Yeungnam University, Gyeongsan.
- Seo, J. H. and Lee, J. Y. (2018). Nomogram construction to predict dyslipidemia based on logistic regression analysis, submitted: *Journal of Applied Statistics*.
- The Korean Society of Lipid and Atherosclerosis (2018). The Korean Guidelines for Management of Dyslipidemia (4th ed). Available from: <http://www.lipid.or.kr/bbs/?code=care>
- Van den Berg, E., Kloppenborg, R. P., Kessels, R. P., Kappelle, L. J., and Biessels, G. J. (2009). Type 2 diabetes mellitus, hypertension, dyslipidemia and obesity: a systematic comparison of their impact on cognition. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, **1792**, 470–481.
- World Health Organization. Disease burden and mortality estimates [cited 2018 May 16]. Available from: http://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html

순수 베이지안 분류기 모델을 사용하여 이상지질혈증을 예측하는 노모그램 구축

김민호^a · 서주현^a · 이제영^{a,1}

^a영남대학교 통계학과

(2019년 5월 23일 접수, 2019년 6월 18일 수정, 2019년 6월 20일 채택)

요약

이상지질혈증은 한국인의 대표적인 성인병이며 지속적인 관리가 필요한 만성질환이다. 또한 고혈압이나 당뇨병과 함께 심혈관계 질환의 위험 요인으로 잘 알려져 있다. 하지만 혈관 질환은 검사 없이는 질병 판단을 하기 어려운 것이 현실이다. 본 연구에서는 이상지질혈증의 인지와 예방을 위하여 관련된 위험 요인을 확인한다. 이들을 종합하여 시각화하면서 발병률 예측까지 가능한 통계적 도구 노모그램을 구축하였다. 데이터는 국민건강영양조사 6기, 7기 제1차년도 (2013-2016) 데이터를 사용하였다. 분석 순서로는 먼저 이상지질혈증의 총 12가지 위험 요인을 교차분석을 통해 확인하였다. 그리고 순수 베이지안 분류기를 이용하여 이상지질혈증에 대한 모형으로 노모그램을 구축하였다. 구축한 노모그램은 ROC 곡선과 Calibration plot을 사용하여 신뢰성을 검증하였다. 마지막으로 이전에 제시했던 로지스틱 노모그램과 본 연구에서 제안한 베이지안 노모그램을 비교하였다.

주요용어: 이상지질혈증, 위험요인, 순수 베이지안 분류기, 노모그

이 논문은 영남대학교 연구년 결과물로 제출됨.

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