Communications for Statistical Applications and Methods
2023, Vol. 30, No. 1, 21–35

https://doi.org/10.29220/CSAM.2023.30.1.021
Print ISSN 2287-7843 / Online ISSN 2383-4757

Nomogram for screening the risk of developing metabolic syndrome using naïve Bayesian classifier

Minseok Shin[a], Jeayoung Lee[1,a]

[a]Department of Statistics, Yeungnam University, Korea

Abstract

Metabolic syndrome is a serious disease that can eventually lead to various complications, such as stroke and cardiovascular disease. In this study, we aimed to identify the risk factors related to metabolic syndrome for its prevention and recognition and propose a nomogram that visualizes and predicts the probability of the incidence of metabolic syndrome. We conducted an analysis using data from the Korea National Health and Nutrition Survey (KNHANES VII) and identified 10 risk factors affecting metabolic syndrome by using the Rao–Scott chi-squared test, considering the characteristics of the complex sample. A naïve Bayesian classifier was used to build a nomogram for metabolic syndrome. We then predicted the incidence of metabolic syndrome using the nomogram. Finally, we verified the nomogram using a receiver operating characteristic curve and a calibration plot.

Keywords: metabolic syndrome, naïve Bayesian classifier, nomogram, risk factors, ROC

1. Introduction

Metabolic syndrome is a group of conditions such as obesity, hyperlipidemia, low high-density lipoprotein (HDL) cholesterol, high blood pressure, and hyperglycemia that occur together in an individual owing to chronic metabolic disorders. Reaven first named this condition “Syndrome X” in 1988 (Reaven, 1988), but in 1999, the World Health Organization (WHO) renamed it as “metabolic syndrome.” However, the WHO definition of metabolic syndrome has not been consistently used because of the requirement to measure serum insulin and urinary microalbumin levels (Jung *et al*., 2002; Lee *et al*., 2004). Therefore, we used the metabolic syndrome diagnostic criteria published by the National Cholesterol Education Program (National Cholesterol Education Program, 2001). Metabolic syndrome was diagnosed if three or more of the following criteria were met: Obesity (male ≥90 cm; female ≥85 cm), hyperlipidemia (triglyceride levels ≥150 mg/dL), low HDL cholesterol level (men <40 mg/dL, women <50 mg/dL), high blood pressure (systolic BP ≥130 mmHg and diastolic BP ≥85 mg/dL or currently undergoing drug treatment for hypertension), and hyperglycemia (fasting glucose ≥100 mg/dL). In the United States, 32.8% of men and 36.6% of women aged 20 years or older developed metabolic syndrome in 2012 (Aguilar *et al*., 2015). In Korea, approximately 30.8% of men and 26.3% of women over 20 years of age developed metabolic syndrome in 2013 (Tran *et al*., 2017). Metabolic syndrome is a very serious condition because it can lead to various complications, such as stroke or cardiovascular diseases. Therefore, prevention of metabolic syndrome is important (Yoo *et al*., 2009).

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

Published 31 January 2023 / journal homepage: http://csam.or.kr

The Pearson chi-squared test is mainly used as a statistical analysis method to identify the risk factors for a disease. Logistic regression and Cox proportional hazards models are generally used as statistical models to predict the incidence of a disease. Many studies have been conducted to identify the risk factors for metabolic syndrome (Jung *et al.*, 2002; Tran *et al.*, 2017; Yoo *et al.*, 2009). However, medical workers and those who have less knowledge of statistics have difficulty interpreting and understanding these results. This problem can be overcome by using a nomogram. A nomogram is a tool that can graphically represent numerical relationships between diseases and risk factors without requiring complex calculations (Iasonos *et al.*, 2008; Mozina *et al.*, 2004). Nomograms have also been developed for dyslipidemia and hypertension (Kim *et al.*, 2019; Kim and Lee, 2020).

Although many studies have been conducted to identify the risk factors for metabolic syndrome, this paper focuses on visually describing the risk probability of metabolic syndrome via nomogram. Now, in this study, we used the Rao–Scott chi-squared test—instead of the Pearson chi-squared test—to identify the risk factors of metabolic syndrome. Using data from the Korean National Health and Nutrition Examination Survey (KNHANES), we construct a nomogram that can predict the incidence rates of metabolic syndrome after constructing a naïve Bayesian classifier model (Shin, 2022). The naïve Bayesian classifier model is a machine learning technique applied to classification problems using the Bayesian theorem and has recently been applied widely in the medical field to predict the incidence of diseases. Finally, the constructed nomogram was verified using a receiver operating characteristic (ROC) curve and a calibration plot.

In Section 2, we describe the complex sampling design method and the Rao–Scott chi-squared test. We then explain how to construct and verify a nomogram using the calculated log odds ratio from the naïve Bayesian classifier of complex sampling data. In Section 3, we explain the KNHANES data and present the results of the Rao–Scott chi-squared test. In addition, a nomogram for metabolic syndrome was constructed and verified. In Section 4, we present the conclusions and discussion of this study.

2. Methodology

2.1. Complex sampling design method

Raw data were collected using a simple random sampling method, with each element having the same probability of being selected and having features that were independent of each other. The KNHANES data used in this study were designed to sample representative samples of the Korean population using census data as a sampling frame, and a two-stage stratified cluster sampling method was used. When using complex sample data, one must take into consideration the effects of stratification, clustering, and individual sample weight, which can be corrected for inclusion error, imbalance extraction rate, and non-response error of the target population. This is why we choose the Rao-Scott chi-squared test over the Pearson chi-squared test, as explained in the next section.

2.2. Rao-Scott $\chi^2_{\text{Rao-Scott}}$ Test

To select the risk factors that affect the incidence of metabolic syndrome, it is important to test the independence between the incidence of metabolic syndrome and any risk factors. Generally, to identify the risk factors for metabolic syndrome, the Pearson chi-squared test, which assumes that the frequency of each cell in the contingency table is independent and follows a multinomial distribution, is used. In contrast, the data used in this study are complex data, given the different individual weights for each stratum and cluster. Our data set does not satisfy the assumption that each cell in the table is independent (Rao and Scott, 1981). Therefore, the Rao–Scott chi-squared test, which considers design effects, such as stratification, clustering, and individual sample weight, was used in the study.

The Rao–Scott chi-squared statistic is as follows:

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

where χ^2 is a Pearson chi-squared statistics, and $\widehat{\delta}$ is as follows:

$$\widehat{\delta} = \frac{\sum_i \sum_j \left(1 - \widehat{\pi}_{i+}\widehat{\pi}_{+j}\right)\widehat{d}_{ij} - \sum_i \left(1 - \widehat{\pi}_{i+}\right)\widehat{d}_{i+} - \sum_j \left(1 - \widehat{\pi}_{+j}\right)\widehat{d}_{+j}}{(I-1)(J-1)},$$

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

where $\widehat{\pi}_{ij}$ is the estimated probability of the i^{th} and j^{th} cells, $\widehat{\text{Var}}(\widehat{\pi}_{ij})$ is the estimated variance of $\widehat{\pi}_{ij}$, and n is the number of sample units. $\widehat{d}_{ij}$ is the design effect of $\widehat{\pi}_{ij}$.

2.3. Naïve Bayesian classifier with individual sample weights

The naïve Bayesian classifier is known to be a simple but powerful tool for classification problems. It is based on Bayes' theorem, assuming that the attribute values $(x_1, x_2, \ldots, x_I)$ are independent of each other (Mozina *et al.*, 2004). In this study, our purpose is to predict the occurrence of an event; thus we assumed that the target class is a binomial variable. In addition, we assumed that each x_i is a categorical variable with two or more categories. Given the attribute values $X = (x_1, x_2, \ldots, x_I)$, the odds for $Y = 1$ is defined as follows.

$$\text{Odds} = \frac{P(Y=1|X)}{P(Y=0|X)} = \prod_{i=1}^{I} \frac{P(x_i|Y=1)}{P(x_i|Y=0)} \times \frac{P(Y=1)}{P(Y=0)}.$$

Taking the log on both sides of the above equation, we obtain the following equation:

$$\log \text{Odds} = \ln \frac{P(Y=1|X)}{P(Y=0|X)} = \sum_{i=1}^{I} \ln \frac{P(x_i|Y=1)}{P(x_i|Y=0)} + \ln\left(\frac{P(Y=1)}{P(Y=0)}\right).$$

It can be shown that the $\log \text{it} P(Y = 1|X)$ is comprised of sum of $\ln(P(x_i|Y = 1)/P(x_i|Y = 0))$. Therefore, we obtain the conditional probability of $Y = 1$ given $X = (x_1, x_2, \ldots, x_I)$, which is $P(Y = 1|X)$, as follows:

$$P(Y=1|X) = \frac{1}{1 + \exp\left\{-\sum_{i=1}^{I} \ln\left(P(x_i|Y=1)/P(x_i|Y=0)\right) - \ln\left(P(Y=1)/P(Y=0)\right)\right\}}.$$

Using the above equation, we can get the value of the conditional probability of $Y = 1$ given X if we know $P(x_i \mid Y = 1)$ and $P(Y = 1)$ for all $i = 1, \ldots, I$. As described previously, we propose a method in which $P(x_i \mid Y = 1)$ and $P(Y = 1)$ are calculated using the frequency with individual sample weights for all attribute values x_i. Let x_i be the i^{th} attribute, $j = 1, \ldots, J_i$ be the j^{th} category of x_i, and $k = 0, 1$ be the $Y = k$-class. Let Ω be the set of sample weights for the total population. Then A_{ijk} is subset of Ω classified into the j^{th} category of the i^{th} attribute value and $Y = k$-class, namely the partition of Ω. Let ω_l be the sample weight for l^{th} participant. We can create the $J_i \times 2$ contingency table (Table1).

Table 1: $J_i \times 2$ contingency table for the i^{th}attribute

i^{th} attribute	Class		Total
	$Y = 1$	$Y = 0$	
$x_i = 1$	n_{i11}	n_{i10}	$n_{i1.}$
$x_i = 2$	n_{i21}	n_{i20}	$n_{i2.}$
$x_i = 3$	n_{i31}	n_{i30}	$n_{i3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i = j$	n_{ij1}	n_{ij0}	$n_{ij.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i = J_i$	n_{iJ_i1}	n_{iJ_i0}	$n_{iJ_i.}$
Total	$n_{..1}$	$n_{..0}$	n

In Table1, n_{ijk} is the sum of A_{ijk}, and it refers to the weighted counts of each cell, $n_{ij.}, n_{..k}$ are the row and column marginal total weighted counts, respectively. n is the total number of counts in the Table1, and m is the total number of sample population, where:

$$\begin{aligned} n_{ij.} &= n_{ij1} + n_{ij0}, \\ n_{..k} &= \sum_{j=1}^{J_i} n_{ijk}, \\ n &= \sum_{l=1}^{m} \omega_l. \end{aligned}$$

Let OR (x_i) be (posterior odds)/(prior odds) as the likelihood ratio. When an attribute value x_i is equal to j^{th} category, we calculate OR (x_i) as follows:

$$\begin{aligned} \frac{P(x_i = j|Y = 1)}{P(x_i = j|Y = 0)} &= \frac{P(x_i = j, Y = 1)/P(Y = 1)}{P(x_i = j, Y = 0)/P(Y = 0)}, \\ &= \frac{P(Y = 1|x_i = j)P(x_i = j)/P(Y = 0|x_i = j)}{P(Y = 1)/P(Y = 0)}, \\ &= \frac{(n_{ij1}/n_{ij.})/(n_{ij0}/n_{ij.})}{(n_{..1}/n)/(n_{..0}/n)} = \frac{n_{ij1}/n_{..1}}{n_{ij0}/n_{..0}} = \mathrm{OR}(x_i = j). \end{aligned}$$

2.4. Nomogram construction of naïve Bayesian classifier model

In medical research, many studies have aimed to predict the risk of a disease or death. To predict risk, statistical techniques are used to select risk factors that affect a disease or death and calculate the extent of their effects. However, medical workers and those who have less knowledge of statistics have difficulty interpreting and understanding the results. Therefore, in the present study, we propose a nomogram that calculates the risk of an event as the probability of occurrence (Iasonos *et al.*, 2008; Lee *et al.*, 2009). A nomogram is easy to interpret because the construction method is not complicated, and the composition of the nomogram is simply expressed by a line. There are four components that constitute a nomogram: Point line, risk factor line, probability line, and total point line. We describe the method to construct the nomogram with ln OR (x_i) calculated using the naïve Bayesian classifier model as follows (Iasonos *et al.*, 2008; Park *et al.*, 2018; Yang, 2013).

(a) Point line: The point line consists of −100 to 100 points.

(b) Risk Factor line: The $\ln \mathrm{OR}(x_i = j)$ values obtained by fitting in the naïve Bayesian classifier model are used to calculated Point_{ij} for each j^{th} category of the i^{th} risk factor.

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

(c) Probability line: We constructed a probability line divided from 0 to 1 into 10 sections of 0.1.

(d) Total Points line: The total point can be expressed as a cumulative sum of Point_{ij}.

$$\begin{aligned}\text{Total Point} &= \frac{100}{\max_{i,j} |\ln \mathrm{OR}\,(x_i = j)|} \times \sum_{i,j} \ln \mathrm{OR}\,(x_i = j)\\ &= \frac{100}{\max_{i,j} |\ln \mathrm{OR}\,(x_i = j)|}\left(-\ln\left(\frac{1}{P\,(Y = 1 \mid X = x)} - 1\right) - \ln \frac{P\,(Y = 1)}{1 - P\,(Y = 1)}\right).\end{aligned}$$

Then, the total point line can be constructed to substitute $P\,(Y = 1|X = x)$ for the value of the probability line.

2.5. Left-aligned method of nomogram for naïve Bayesian classifier model

The results of applying the left-aligned method using $\ln \mathrm{OR}\,(x_i = j)$ of the Naïve Bayesian classifier are given below. This method has the advantage of being easy to compare with the logistic nomogram because the score ranges from 0 to 100.

(a) Point line: The point line consists of 0-100 points.

(b) Risk Factor line: The $\ln \mathrm{OR}(x_i = j)$ values obtained by fitting in the naïve Bayesian classifier model are used to calculated Point_{ij} for each j^{th} category of the i^{th} risk factor.

$$\mathrm{Point}_{ij} = \frac{\ln \mathrm{OR}\,(x_i = j) - \min_j \ln \mathrm{OR}\,(x_i = j)}{\max_j \ln \mathrm{OR}\,(x_i = j) - \min_j \ln \mathrm{OR}\,(x_i = j)} \times 100.$$

(c) Probability line: We constructed a probability line divided from 0 to 1 into 10 sections of 0.1.

(d) Total Points line: The total point can be expressed as a cumulative sum of Point_{ij}.

$$\begin{aligned}\text{Total point} = \sum_{ij} \mathrm{Point}_{ij} &= \frac{100}{\max_j \ln \mathrm{OR}\,(x_i = j) - \min_j \ln \mathrm{OR}\,(x_i = j)}\\ &\times \sum_{i,j}\left(\ln \mathrm{OR}\,(x_i = j) - \min_j \ln \mathrm{OR}\,(x_i = j)\right).\end{aligned}$$

To construct the total point line, we express Total Point as follows:

$$\begin{aligned}\text{Total point} = &\frac{100}{\max_j \ln \mathrm{OR}\,(x_i = j) - \min_j \ln \mathrm{OR}\,(x_i = j)}\\ &\times\left(-\ln\left(\frac{1}{P\,(Y = 1|X = x)} - 1\right) - \log \mathrm{it} P\,(Y = 1) - \sum_{i,j} \min_j \ln \mathrm{OR}\,(x_i)\right).\end{aligned}$$

The total point line can be constructed to substitute $P(Y = 1|X = x)$ for the value of the probability line.

Table 2: Rao-Scott chi-squared test result for 10 risk factors associated metabolic syndrome

Variable	Level	Metabolic (%)	Non-metabolic (%)	χ^2_{R-S}	p-value
BMI	BMI<25	3,138,741(12.1)	22,748,807(87.9)	2128.3904	<.0001
	25≤BMI<30	5,365,781(45.1)	6,542,466(54.9)		
	30≤BMI	1,458,791(66.4)	737,094(33.6)		
Marry	Yes	9,003,536(29.0)	22,056,181(71.0)	309.0303	<.0001
	No	959,776(10.7)	7,972,187(89.3)		
Employ	Yes	6,182,962(23.4)	20,207,289(76.6)	29.727	<.0001
	No	3,780,350(27.8)	9,821,079(72.2)		
Education	Low	3,844,170(41.6)	5,393,007(58.4)	942.6711	<.0001
	High	6,119,142(19.9)	24,635,361(80.1)		
Age	20-34	815,925(8.2)	9,186,929(91.8)	1006.4028	<.0001
	35-64	6,192,295(26.4)	17,245,747(73.6)		
	65+	2,955,092(45.1)	3,595,692(54.9)		
Stroke	Yes	283,695(54.2)	240,097(45.8)	91.1563	<.0001
	No	9,679,617(24.5)	29,788,271(75.5)		
Sex	Man	5,684,391(28.3)	14,390,899(71.7)	87.496	<.0001
	Woman	4,278,921(21.5)	15,637,469(78.5)		
Income	Lowest	2,875,053(28.4)	7,264,837(71.6)	38.88	<.0001
	Low	2,571,634(25.5)	7,519,515(74.5)		
	High	2,367,455(23.7)	7,637,714(76.3)		
	Highest	2,149,170(22.0)	7,606,302(78.0)		
Smoke	Non	4,878,365(21.7)	17,595,872(78.3)	88.4268	<.0001
	Past	2,569,051(29.2)	6,241,132(70.8)		
	Current	2,515,896(28.9)	6,191,363(71.1)		
Family history	Yes	1,516,774(27.7)	3,966,295(72.3)	6.1727	0.0130
	No	8,446,538(24.5)	26,062,073(75.5)		

2.6. Nomogram validation

We used an ROC curve and calibration plot to test the accuracy of the nomogram (Akobeong, 2007; Cook, 2008). In the ROC curve, sensitivity was plotted on the y-axis and 1-specificity on the x-axis. The AUC of the diagonal line is 0.5, which is a good model when the ROC curve is well above this diagonal. A calibration plot is a method to determine how closely the actual probabilities correspond to the predicted probabilities calculated using the nomogram. If the predicted probability is the same as the actual probability, a 45° centerline is drawn. The closer the line is to the 45° angle line, the closer the predicted probability to the actual probability (D'Agostino *et al*., 2001; Iasonos *et al*., 2008). Therefore, we validated the nomogram with R^2, a goodness-of-fit indicator of the regression line between the predicted and actual probabilities.

3. Applications

3.1. Materials

The data used in this study were obtained from the KNHANES 2016–2018 database (Korea Centers for Disease Control and Prevention ,2018). This study was conducted in subjects aged 20 years or older. In three years, 5,072 of the 24,269 individuals under 20 years of age were excluded. In addition, 1,613 individuals who did not participate in the health or examination surveys were excluded. The sample population included 17,584 individuals. Missing values were replaced with the mode. The data were then divided into a training set and test set at an 8:2 ratio. The model was fitted using the training set, and the predictive power of the model was verified using the test set. In addition, we compared the complex sample population and the actual Korean population surveyed in the 2018 cen-

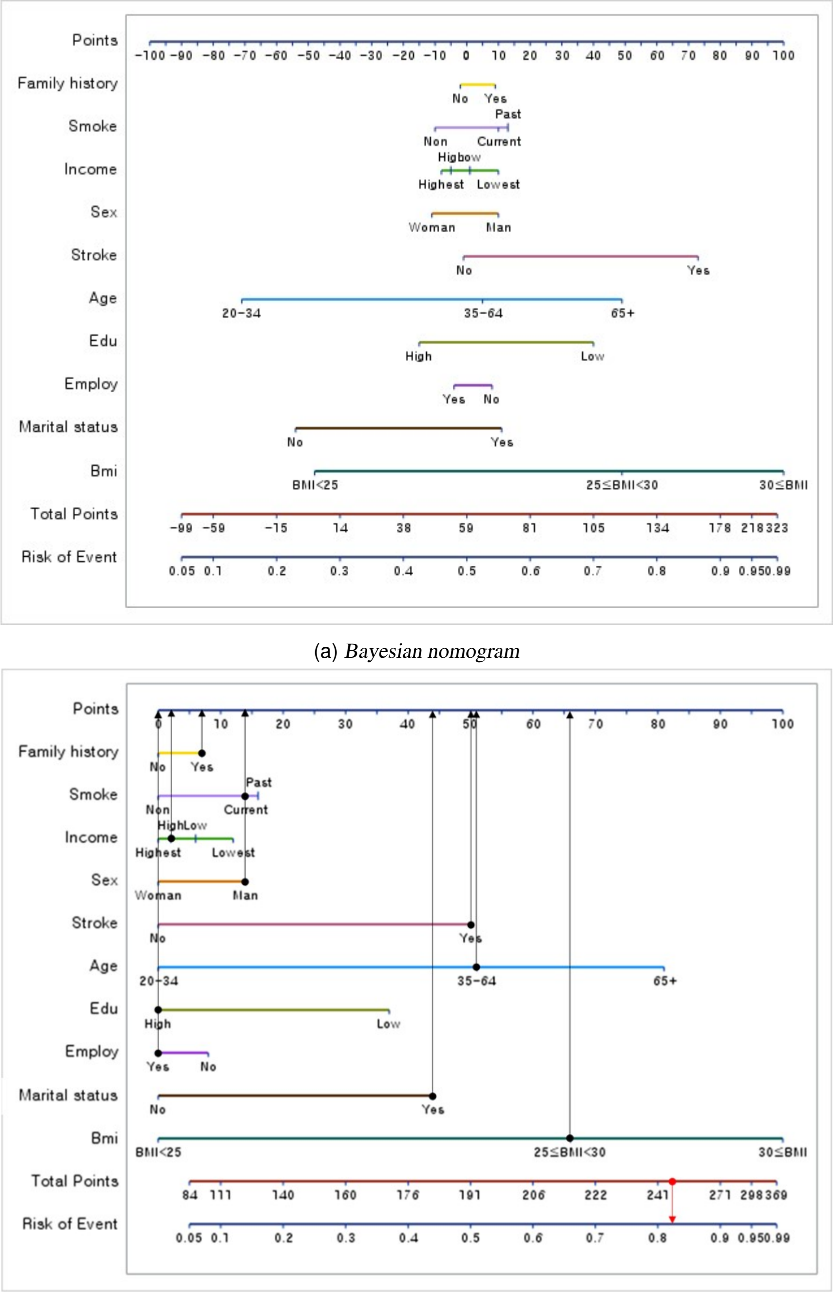
Table 3: The results of naïve Bayesian classifier model for metabolic syndrome and nomogram points

Variable	Level	$P(x_i = j \mid \text{Metabolic})$	$P(x_i = j \mid \text{Non-metabolic})$	$\ln \text{OR}(x_i = j)$	Point
BMI	BMI<25	0.31	0.76	−0.888	−48
	25≤BMI<30	0.54	0.22	0.919	49
	30≤BMI	0.14	0.02	1.857	100
Marry	Yes	0.90	0.73	0.208	11
	No	0.10	0.27	−0.995	−54
Employ	Yes	0.62	0.67	−0.082	−4
	No	0.38	0.33	0.151	8
Education	Low	0.37	0.18	0.737	40
	High	0.63	0.82	−0.269	−15
Age	20-34	0.08	0.31	−1.316	−71
	35-64	0.62	0.57	0.087	5
	65+	0.29	0.12	0.910	49
Stroke	Yes	0.03	0.01	1.357	73
	No	0.97	0.99	−0.020	−1
Sex	Man	0.58	0.48	0.179	10
	Woman	0.42	0.52	−0.201	−11
Income	Lowest	0.29	0.24	0.188	10
	Low	0.26	0.25	0.028	1
	High	0.24	0.26	−0.086	−5
	Highest	0.22	0.25	−0.151	−8
Smoke	Non	0.49	0.58	−0.186	−10
	Past	0.26	0.20	0.240	13
	Current	0.26	0.21	0.188	10
Family history	Yes	0.16	0.13	0.169	9
	No	0.84	0.87	−0.028	−2

sus (Korean Statistical Information Service, 2018). The weighted sample population was 39,991,680. According to the 2018 census from Statistic Korea, the total number of Koreans over 20 years of age was 40,762,796. The difference between the complex sample population and the actual Korean population was 771,116 (1.9% of the actual population). The weighted sample and actual Korean population were similar.

The criteria for the diagnosis of metabolic syndrome include three or more factors, including abdominal obesity, hyperlipidemia, low HDL cholesterol, hypertension, and hyperglycemia. We used the metabolic syndrome diagnostic criteria published by the National Cholesterol Education Program (NCEP ATP-III) in 2001. However, among the five criteria for diagnosis, the criteria for abdominal obesity were applied to the Korean Society for the Study of Obesity in 2014 to reflect the characteristics of Koreans. In this study, 4,850 of 17,584 people received a diagnosis of metabolic syndrome.

A total of 10 risk factors were selected in several prior studies that had an important effect on the incidence of diabetes, dyslipidemia, and hypertension (Chung *et al.*, 2018; Kim and Lee, 2020; Kshirsagar *et al.*, 2010; Seo *et al.*, 2020). The risk factors were BMI, marital status, employment status, education, age, stroke, sex, income, smoking, and family history. BMI was categorized into three groups: < 25, ≥ 25, < 30, and ≥ 30. Those with education levels below high school were classified as 'low', and those above high school as 'high'. Age was categorized as 20 to 34, 35 to 64, and 65 or older. Participants were categorized into four groups based on their income quantile: Lowest, low, high, and highest. They were also categorized into three groups based on their smoking habits: Present smoking, past smoking, and nonsmoking. The other variables were categorized as yes or no.



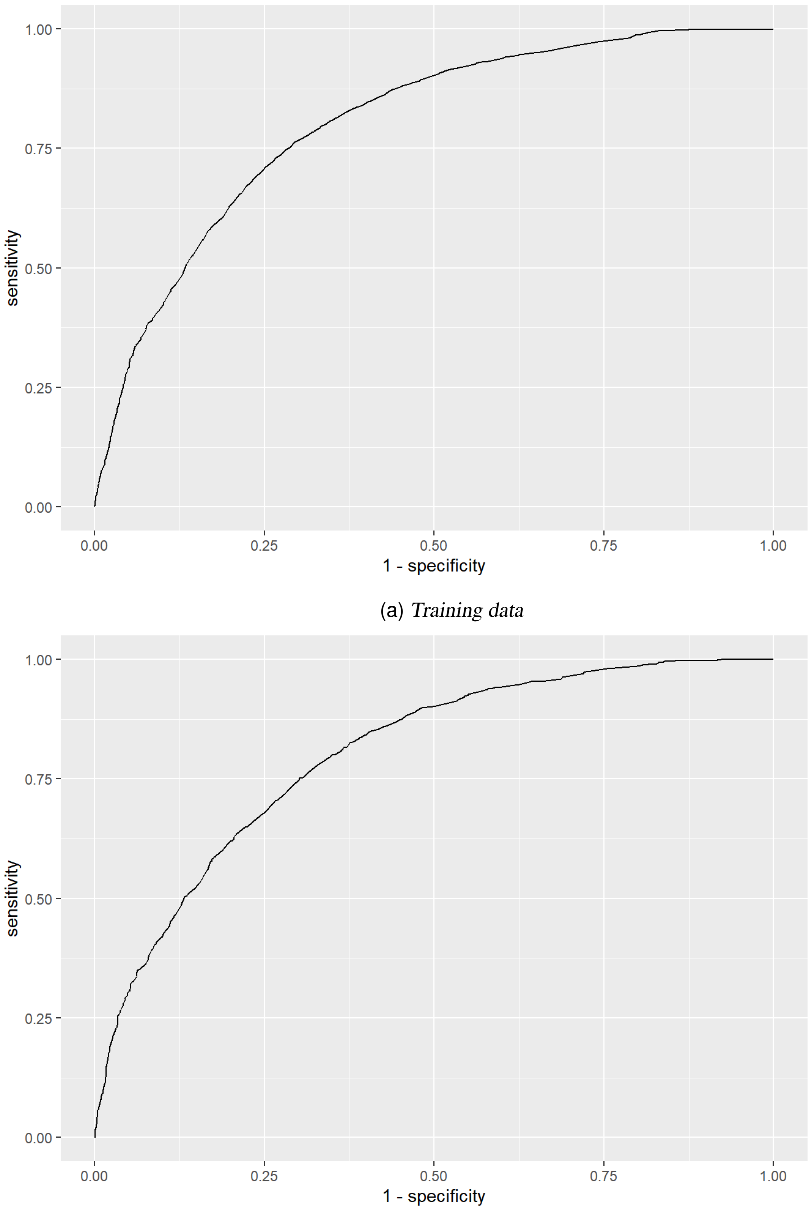
(a) *Bayesian nomogram*

(b) *Left-aligned Bayesian nomogram*

Figure 1: *Nomogram for metabolic syndrome using naïve Bayesian classifier model.*

3.2. Rao-Scott chi-squared test results

We used the Rao–Scott chi-squared test to select the risk factors related to metabolic syndrome. Table 2 presents the weighted frequency and the results of the Rao–Scott chi-squared test. As shown in Table



(a) *Training data*

(b) *Test data*

Figure 2: *ROC curve of Bayesian nomogram with complex sample.*

2, the prevalence of metabolic syndrome increased with increasing BMI. For example, the incidence of metabolic syndrome was 12.1% for BMI less than 25, 45.1% for BMI ≥25 and < 30, and 66.4% for BMI ≥30. In addition, the older the individual, the higher the incidence of metabolic syndrome. The incidence rate was 8.2% in the age group 20–34 years, 26.4% in the age group 35–64 years, and

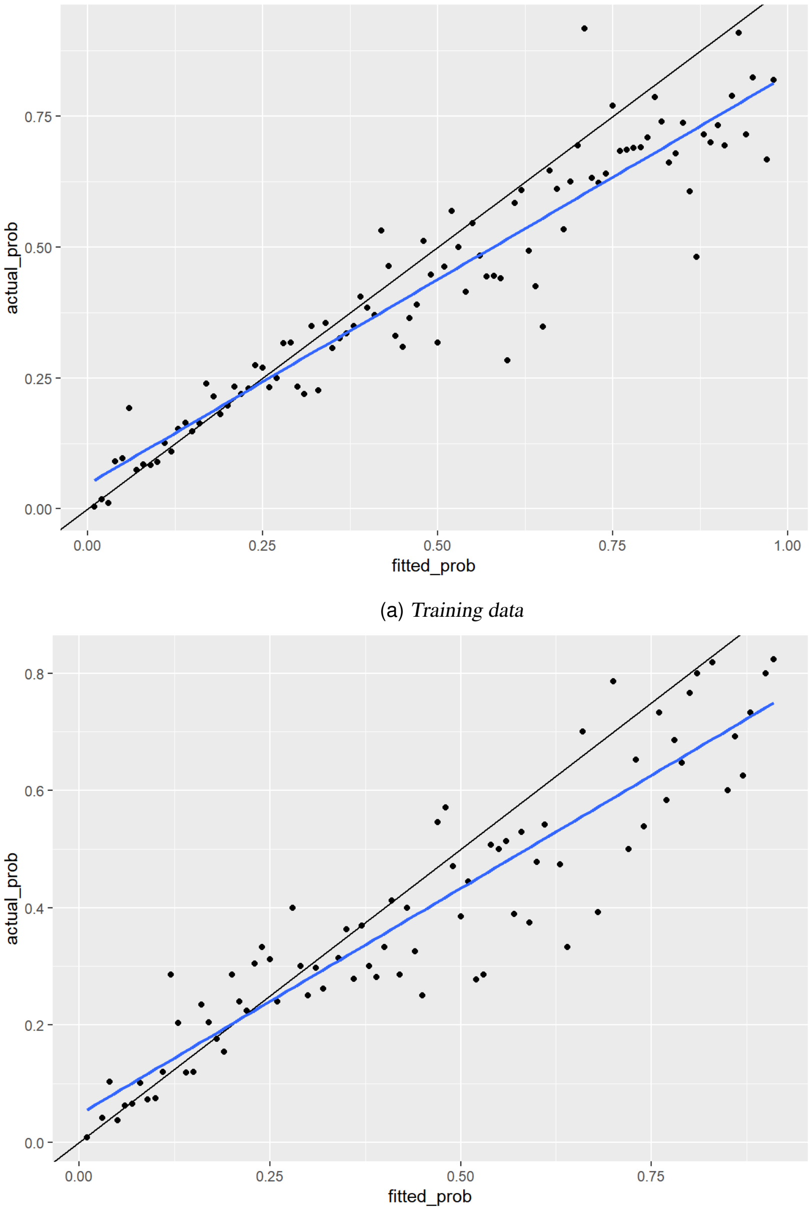
45.1% in the age group 65 years or older. Moreover, 41.6% of the people whose education level was low and 19.9% of those whose education level was high received a diagnosis of metabolic syndrome. The incidence was 29.0% for married people and 10.7% for unmarried people. Moreover, 54.2% of those who received a diagnosis of stroke and 24.5% of those who did not receive a diagnosis of stroke developed metabolic syndrome. Furthermore, the incidence rate was 21.7% in the nonsmoking group, 29.2% in the past smoking group, and 28.9% in the smoking group. Men and women had metabolic syndrome incidence rates of 28.3% and 21.5%, respectively. The lower the income, the higher the incidence of metabolic syndrome. Additionally, individuals with metabolic syndrome were generally unemployed and had a family history of metabolic syndrome. The results of the Rao–Scott chi-squared test (χ^2_{R-S} test) showed that all factors were statistically significant at 0.05. Therefore, 10 risk factors were found to be important for the incidence of metabolic syndrome.

3.3. Nomogram for metabolic syndrome using naïve Bayesian classifier model with complex sample

The values of ln OR ($x_i = j$) for the j^{th} category of attribute value x_i calculated from the naïve Bayesian classifier model and the nomogram points are presented in Table 3. The Bayesian nomogram for metabolic syndrome is shown in Figure 1(a), and the nomogram for which the statistical method introduced in Section 2.5 was used is shown in Figure 1(b). The point range is set to -100 to 100 in Figure 1(a) and 0 to 100 in Figure 1(b). As can be seen from Figure 1(b), the risk factor with the longest point range is BMI (range 0–100), indicating that BMI is the most influential factor in the occurrence of metabolic syndrome. In addition, the incidence of metabolic syndrome increases with age. Stroke had the highest impact on BMI and age. People with a history of stroke had higher scores than those without a history of stroke; therefore, the former were more likely to have metabolic syndrome. In addition, marital status highly influences the incidence of metabolic syndrome, followed by education, smoking, sex, income, employment, and family history. Both nomograms showed that BMI had the greatest impact on metabolic syndrome. The left-aligned Bayesian nomogram is easy to compare with the logistic nomogram because they both have the same point range. The nomogram of the naïve Bayesian classifier model shows the risk factors that affect the incidence rate of metabolic syndrome at a glance. For example, if a 50-year-old male with a BMI of 25 was married, employed, a college graduate, had a history of stroke, belonged to a high-income level, was a smoker, and had a family history of metabolic syndrome, then his nomogram points would be as follows: 66 points on the BMI line, 44 points on the Marital status line, 0 points on the Employ line, 0 points on the Education line, 51 points on the Age line, 50 points on the Stroke line, 14 points on the Sex line, 2 points on the Income line, 14 points on the Smoke line, and 7 points on the Family history line. The sum of the points was 248. Therefore, the probability of metabolic syndrome corresponding to 248 points is 83%.

3.4. Validation of nomogram for metabolic syndrome

The ROC curve and calibration plot were used for the nomogram verification. The results are shown in Figures 2 and 3, respectively. The ROC curves of the training data and test data are shown in Figure 2. The AUC-values were 0.8029 for the ROC curve of the training data and 0.8011 for the ROC curve of the test data. The calibration plots for the training and test data are shown in Figure 3. The R^2-values for the calibration plot were 0.8914 and 0.8552, respectively. Therefore, it can be concluded that the nomogram has sufficient predictive power.



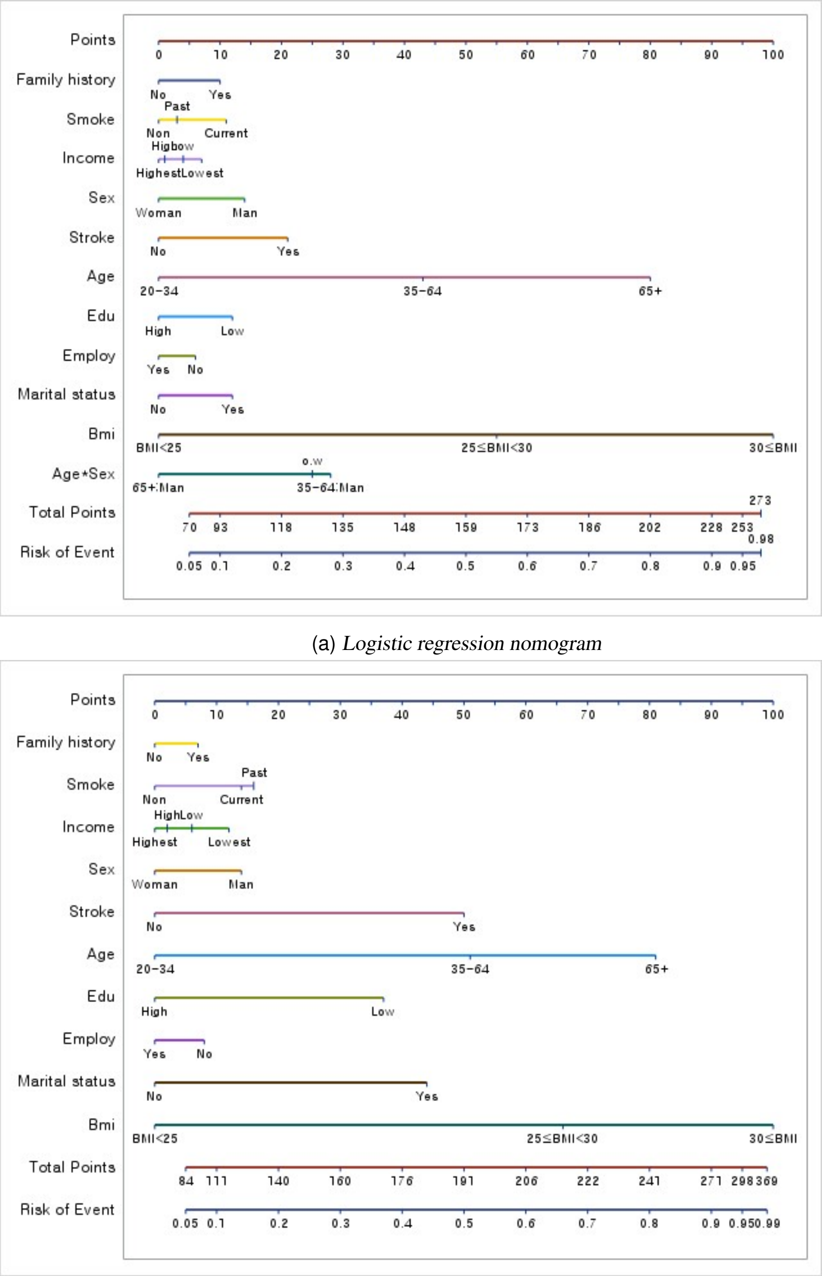
(a) *Training data*

(b) *Test data*

Figure 3: *Calibration plot of Bayesian nomogram with complex sample.*

4. Conclusions and discussions

In this study, the data from the KNHANES (2016–2018), which can identify the health behavior of Koreans, were used, and the number of data points used was 17,584. However, when the individual weights assigned to each stratum and cluster were applied, the complex sample population was



(a) *Logistic regression nomogram*

(b) *Left-aligned Bayesian nomogram*

Figure 4: *Comparison of logistic regression nomogram and Bayesian nomogram.*

39,991,680. This is only 1.9% different from the real population. Thus, complex sample analysis is more reasonable than raw data analysis. Therefore, 10 main effects were selected as the risk factors for metabolic syndrome. The Rao–Scott chi-squared test was used to screen for the risk factors for metabolic syndrome, and all 10 risk factors were statistically significant. Training data were used to build a naïve Bayesian nomogram to predict the probability of the occurrence of metabolic syndrome, and the performance of the nomogram was verified with test data. In the proposed nomogram, the most relevant risk factor for metabolic syndrome was BMI, followed by age, stroke, marital status, and education. The AUC of the ROC curve and R^2 of the calibration plot were used to assess the accuracy of the proposed nomogram. In the training data ROC curve, the AUC was 0.8029, and in the test data ROC curve, it was 0.8011. The R^2-values of the calibration plot were 0.8914 and 0.8552 for the training and test data, respectively. In other words, the proposed nomogram has sufficient reliability.

In addition, we compared two nomograms generated using the logistic regression model and the naïve Bayesian classifier (Shin and Lee, 2021). The nomogram for metabolic syndrome using logistic regression model is shown in Figure 4(a), and the Bayesian nomogram for metabolic syndrome with a left-aligned method is shown in Figure 4(b). Both nomograms have a score range of 0 to 100, with BMI and age being the most influential factors for metabolic syndrome. Stroke, marital status, and education level also had a significant impact on metabolic syndrome. When the individual metabolic syndrome prediction probability is calculated, the logistic regression model should have values for all attributes, but the Bayesian model can provide results even if values for not all attributes are given. Moreover, logistic regression is unreasonable when considering the interactions between all risk factors; thus, a few interactions can be expressed, as shown in Figure 4(a). In contrast, for the naïve Bayesian classifier, the interaction effects among the risk factors were calculated using conditional probabilities. Hence, we show the results considering the interaction effects.

Metabolic syndrome is strongly associated with and likely the cause of an increased rate of stroke and cardiovascular disease as well as a corresponding increase in the death rate in a given population. Therefore, we aimed to develop a nomogram as a predictive tool in screening for metabolic syndrome. It is easy for medical staff or people who do not have statistical knowledge to diagnose the disease through a nomogram. We expect that the awareness and treatment rates of metabolic syndrome will increase, especially in the elderly who are vulnerable to disease, by calculating the risk of metabolic syndrome based on simple information. Therefore, the metabolic syndrome nomogram proposed in this study will assist in establishing future treatment plans in the medical field.

References

Aguilar M, Bhuket T, Torres S, Liu B, and Wong RJ (2015). Prevalence of the metabolic syndrome in the United States, 2003-2012, *JAMA*, **313**, 1973–1974.

Akobeng AK (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves, *Acta Paediatrica*, **96**, 644–647.

Chung SM, Park JC, Moon JS, and Lee JY (2018). Novel nomogram for screening the risk of developing diabetes in a Korean population, *Diabetes Research and Clinical Practice*, **142**, 286–293.

Cook NR (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve, *Clinical Chemistry*, **54**, 17–23.

D'Agostino Sr RB, Grundy S, Sullivan LM, 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.

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

Jung CH, Park JS, Lee WY, and Kim SW (2002). Effects of smoking, alcohol, exercise, level of education, and family history on the metabolic syndrome in Korean adults, *Korean Journal of Medicine*, **63**, 649–659.

Kim MH, Seo JH, and Lee JY (2019) A study on the method of constructing a nomogram for predicting dyslipidemia, *Journal of the Korean Data & Information Science Society*, **30**, 1063–1075.

Kim MH and Lee JY (2020). How to construct a nomogram for hypertension using complex sampling data from Korean adults, *Communications in Statistics-Theory and Methods*, **51**, 2357–2367.

Korea Centers for Disease Control and Prevention (2018). *The Seventh Korea National Health and Nutrition Examination Survey (KNHANES VII)*, Available from: https://knhanes.kdca.go.kr/knhanes/ sub03/sub03_02_05.do

Korean Statistical Information Service (KOSIS) (2018). Census, Statistic Korea, Republic of Korea. Accessed February 2021, Available from: https://kosis.kr/statisticsList/statisticsListIndex.do? menuld=M _01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01&outLink=Y&entrType=#content-group.

Kshirsagar AV, Chiu YL, Bomback AS, August PA, Viera AJ, Colindres RE, and Bang H (2010). A hypertension risk score for middle-aged and older adults, *The Journal of Clinical Hypertension*, **12**, 800–808.

Lee KM, Kim WJ, and Yun SJ (2009). A clinical nomogram construction method using genetic algorithm and naïve Bayesian technique, *Journal of Korean Institute of Intelligent Systems*, **19**, 796–801.

Lee WY, Park JS, Noh SY, Rhee EJ, Kim SW, and Zimmet PZ (2004). Prevalence of the metabolic syndrome among 40,698 Korean metropolitan subjects, *Diabetes Research and Clinical Practice*, **65**, 143–149.

Mozina M, Demˇsar J, Kattan M, and Zupan B (2004). Nomograms for visualization of Naïve Bayesian classifier, In *Proceedings of the Knowledge Discovery in Databases: PKDD 2004*, Italy, 337–348.

National Cholesterol Education Program (2001). Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation and treatment of high blood cholesterol in adults (adult treatment panel III), *JAMA*, **285**, 2486–2497.

Park JC, Kim MH, and Lee JY (2018). Nomogram comparison conducted by logistic regression and naïve Bayesian classifier using Type 2 diabetes mellitus, *The Korean Journal of Applied Statistics*, **31**, 573–585.

Rao JNK and Scott AJ (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit the independence in two-way tables, *Journal of the American Statistical Association*, **76**, 221–230.

Reaven GM (1988). Diabetes, *Role of Insulin Resistance in Human Disease*, **37**, 1595–1607.

Seo JH, Kim HJ, and Lee JY (2020). Nomogram construction to predict dyslipidemia based on a logistic regression analysis, *Journal of Applied Statistics*, **47**, 914–926.

Shin MS and Lee JY (2021). Building a nomogram for metabolic syndrome using logistic regression with a complex sample — 39,991,680 case study, *Healthcare*, **10**, 372.

Shin MS (2022). Proposal of nomogram using logistic and Bayesian technique for metabolic syndrome with complex sample (Master's thesis), Yeungnam University.

Tran BT, Jeong BY, and Oh JK (2017). The prevalence trend of metabolic syndrome and its components and risk factors in Korean adults: Results from the Korean National Health and Nutrition Examination Survey 2008-2013, *BMC Public Health*, **17**, 71.

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

Yoo JS, Jung JI, Park CG, Kang SW, and Ahn JA (2009). Impact of life style characteristics on prevalence risk of metabolic syndrome, *Journal of Korean Academy of Nursing*, **39**, 594–601.

Received January 24, 2022; Revised July 15, 2022; Accepted July 26, 2022