Communications for Statistical Applications and Methods
2018, Vol. 25, No. 3, 297–306

https://doi.org/10.29220/CSAM.2018.25.3.297
Print ISSN 2287-7843 / Online ISSN 2383-4757

A novel nomogram of naïve Bayesian model for prevalence of cardiovascular disease

Eun Jin Kang[a], Hyun Ji Kim[a], Jea Young Lee[1,a]

[a]Department of Statistics, Yeungnam University, Korea

Abstract

Cardiovascular disease (CVD) is the leading cause of death worldwide and has a high mortality rate after onset; therefore, the CVD management requires the development of treatment plans and the prediction of prevalence rates. In our study, age, income, education level, marriage status, diabetes, and obesity were identified as risk factors for CVD. Using these 6 factors, we proposed a nomogram based on a naïve Bayesian classifier model for CVD. The attributes for each factor were assigned point values between −100 and 100 by Bayes' theorem, and the negative or positive attributes for CVD were represented to the values. Additionally, the prevalence rate can be calculated even in cases with some missing attribute values. A receiver operation characteristic (ROC) curve and calibration plot verified the nomogram. Consequently, when the attribute values for these risk factors are known, the prevalence rate for CVD can be predicted using the proposed nomogram based on a naïve Bayesian classifier model.

Keywords: Bayes' theorem, calibration plot, cardiovascular diseases, naïve Bayesian classifier model, nomogram, ROC curve

1. Introduction

Cardiovascular disease (CVD) occurs in the heart and major arteries. It is caused by narrowing or clogging of blood vessels and there are few early symptoms. However, it has a high mortality rate after onset. In Europe, Asia, and the Americas, CVD was the leading cause of death, in 2013 (Wilson *et al.*, 2017). According to a report, 3.9 million people die in Europe and 1.8 million people die in the European Union (EU) from CVD, annually (World Health Organization, 2017). In particular, it can be difficult for humans to lead everyday life due to the first outbreak (American College of Sports Medicine, 2013). So patients should pay attention to prevention and recurrence.

Many studies on the risk factors of CVD have been steadily progressing. One of the major risk factors for CVD is obesity. Nowadays, people who prefer fast food are increasing because of their busy daily lives. Therefore, the rate of obesity is gradually increasing. In many studies, it is emphasized that obesity is highly related to CVD (Poirier *et al.*, 2006; Lavie *et al.*, 2009). Another major factor is diabetes. The incidence of diabetes is increasing due to population aging, increased obesity and increased sitting habits, and is emphasized as the cause of CVD (Grundy *et al.*, 1999). In addition, age, depression, smoking, drinking, stress, and eating habits are also considered as risk factors for CVD (Ambrose and Barua, 2004; Britton and McKee, 2000; Dimsdale, 2008). However, these studies focus on one of the risk factors for CVD and research on multiple factors is lacking. There is a study on

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

Published 31 May 2018 / journal homepage: http://csam.or.kr

multiple risk factors of CVD (Bae and Lee, 2016). This study was conducted for adults aged 30 years and older using data from the 2012 to 2014 Korean National Health and Nutrition Examination Survey (KNHANES). The prevalence rate of CVD was 31.16% and risk factors for CVD were socioeconomic status variables including gender, age, income, education and marital status, past smoker, obesity, meal skipping, high-density lipoprotein cholesterol, and waist circumference (Bae and Lee, 2016).

When a risk factor is selected for a certain disease, nomogram which is a visual tool can be constructed to predict the prevalence rate through simple calculation. Nomogram consists of a point line, straight lines for each risk factor, a total point line, and a probability line (Iasonos *et al.*, 2008). Attribute values for each factor in the nomogram are assigned points through the point line and then, prevalence rate can be predicted by adding points of attribute values and finding the probability through total point line, and probability line. Logistic regression model or Cox proportional-hazard model can be used to derive the nomogram (Kawakami *et al.*, 2008; Kattan *et al.*, 1998). However, we have to consider the interaction in the model since factors have complex relationships with each other for disease (Lyssenko *et al.*, 2008). In addition, when using naïve Bayesian model, the prevalence rate is calculated as posterior probability using the independence assumption between explanatory variables (Lee *et al.*, 2009). Therefore, we do not need to estimate regression coefficients in the complex model and it can be obtained using a frequency for each factor (Možina *et al.*, 2004). Also, when we calculate prevalence rate of a personal, it can be calculated in case that the missing value for a risk factor is included. Our study introduces the method of building the nomogram using naïve Bayesian classifier model with these advantages and suggests the nomogram using CVD data from 2013 to 2015 in KNHANES VI. This paper is organized as follows. We explicate data collection and categorize variables related to CVD in Section 2. We estimate prevalence rate based on naïve Bayesian classifier model under Bayes theorem in Section 3. In Section 4, we introduce the building process of a nomogram plot and in Section 5 we apply risk factors for CVD to the nomogram plot using naïve Bayesian classifier model, and evaluate the proposed nomogram plot, using receiver operating characteristic (ROC) curve and calibration plot. Finally, we make conclusion and discussion for the proposed naïve Bayesian nomogram.

2. Study population and data management

We used 2013–2015 data from KNHANES to build a nomogram applying a naïve Bayesian classifier model. The data was separated into a training set (n = 5,499) and test set (n = 2,357). In a checkup, diseases were diagnosed in the subjects of each year by a professional research team, and the general characteristics of the subjects were surveyed. The 14 variables we used as risk factors are described as: age (19–39, 40–59, or 60–80), gender (male or female), education level (less than elementary school, middle school, high school, or beyond college), income level ($<$ 25%, 25–50%, 50–75%, or $\geq$ 75% according to the equalized household income per month), and marriage status (married or single). Medical histories of diabetes, renal failure, depression, and rheumatoid arthritis were obtained. Smoking status was divided into current smoker, past smoker, and non-smoker. Frequency of alcohol use was divided into none, $<$ 2/week, 2–3/week, and $>$ 4/week. Subjective stress status was divided into yes or no responses. Obesity status was divided into lower weight, normal, and obesity according to body mass index (BMI). Lower weight corresponds to a BMI less than 18.5, normal to a BMI between 18.5 and 25, and obesity to a BMI 25 or higher. The starvation variable was categorized yes if one of the three daily meals was skipped and no otherwise. One or more of the medical histories is defined as CVD: hypertension, dyslipidemia, stroke, myocardial infarction, and angina pectoris.

3. Estimation of prevalence rate based on a naïve Bayesian classifier model

Posterior probability, $P(c|X)$ can be estimated using a naïve Bayesian classifier model. When $P(c)$ is the prevalence rate for target class c, and, $X = \{a_1, a_2, \ldots, a_m\}$ is attribute values, the naïve Bayesian classifier model is a method applying Bayes' theorem that assumes independence between these attributes. Given X, $P(c|X)$ can be calculated, given attribute values of risk factors through the following procedure.

3.1. $P(c|X)$ thorough Bayes' theorem

In naïve Bayes, the predicted prevalence rate can be calculated by conditional probability. When X is given, the conditional probability of target class c is

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)} = \frac{P(a_1, a_2, \ldots, a_m|c)P(c)}{P(X)} = \frac{\prod_i P(a_i|c)P(c)}{P(X)}. \tag{3.1}$$

In equation (3.1), we use Bayes' theorem for $P(c|X)$, and all attribute values $a_1, a_2, \ldots, a_m$ are assumed to be independent of each other. $P(c)$ is the probability of target class c.

3.2. Odds of conditional probability

According to equation (3.1), $P(\overline{c}|X) = \prod_i P(a_i|\overline{c})P(\overline{c})/P(X)$. Therefore, when X is given, odds of conditional probability of target class c, $P(c|X)$, represents the ratio between the probability of success and the probability of failure.

$$\frac{P(c|X)}{P(\overline{c}|X)} = \frac{\prod_i P(a_i|c)P(c)/P(X)}{\prod_i P(a_i|\overline{c})P(\overline{c})/P(X)} = \frac{P(c)}{P(\overline{c})} \times \prod_i \frac{P(a_i|c)}{P(a_i|\overline{c})}. \tag{3.2}$$

Odds of $P(c|X)$ can be expressed as products of odds of target class c and $\prod_i P(a_i|c)/P(a_i|\overline{c})$ as in equation (3.2) and $P(\overline{c}|X) = 1 - P(c|X)$.

3.3. Posterior probability based on a naïve Bayesian classifier

When we take the log of both sides of equation (3.2),

$$\log \frac{P(c|X)}{P(\overline{c}|X)} = \log\left(\frac{P(c)}{P(\overline{c})} \times \prod_i \frac{P(a_i|c)}{P(a_i|\overline{c})}\right). \tag{3.3}$$

Therefore, $P(c|X)$ can be expressed as

$$P(c|X) = \frac{1}{1 + \exp\left(-\log \frac{P(c)}{1-P(c)} - \sum_i \log \frac{P(a_i|c)}{P(a_i|\overline{c})}\right)} = \frac{1}{1 + \exp\left(-\log \frac{P(c)}{1-P(c)} - \sum_i \log \mathrm{LR}(a_i)\right)}. \tag{3.4}$$

Here, $P(a_i|c)/P(a_i|\overline{c})$ is represented as $\mathrm{LR}(a_i)$ in equation (3.4) (Morrison, 1969)

$$\frac{P(a_i|c)}{P(a_i|\overline{c})} = \frac{P(c|a_i)P(a_i)/P(c)}{P(\overline{c}|a_i)P(a_i)/P(\overline{c})} = \frac{P(c|a_i)/P(\overline{c}|a_i)}{P(c)/P(\overline{c})} = \frac{\text{posterior odds}}{\text{prior odds}} = \mathrm{LR}(a_i). \tag{3.5}$$

Therefore, equation (3.4) is equation for probability of a person's prevalence rate. $P(c)$ is the probability that the target class c occurs and is determined by the number of events occurring during the study period. So, we can see that the probability varies depending on the value of $\mathrm{LR}(a_i)$. However, it can be a bit cumbersome to find the probability of target class c by substituting each value. Therefore, we can obtain the probability of target class c by simply adding points for the attribute values of each factor using a nomogram plot.

4. Nomogram plot construction by using naïve Bayesian classifier model

Modern medicine is advancing; however, it is still important to find and treat CVD early. The nomogram plot is a visualization technique showing construction for naïve Bayesian classifier model and it can be used to predict prevalence using the status of several factors. It consists of a point line, straight lines for each risk factor, a total point line, and a probability line. Now, attribute value is specified as a_{ij} ($i = 1, \ldots, m$: the number of factor, $j = 1, \ldots, n_i$: the number of the attribute value for the i^{th} factor). When an individual knows the attribute values for risk factors, the process to predict the prevalence rate using the naïve Bayesian nomogram is as follows.

4.1. Calculate the point values for each risk factor

The point value corresponding to each attribute can be obtained by the following equation.

$$\text{Point value}_{ij} = \frac{\log \text{LR}(a_{ij})}{\max |\log \text{LR}(a_{ij})|} \times 100. \tag{4.1}$$

After all $\text{LR}(a_{ij})$ values are calculated corresponding to all attribute values, the largest absolute $\log \text{LR}(a_{ij})$ value is the most influential attribute for target class c, and it can be seen that this value enters denominator in equation (4.1). The attribute value is assigned with –100 or 100 points.

4.2. Calculate total point values

Through (4.1), when knowing the point values for all attributes, each patient can calculate the total point values for that attributes.

$$\text{Total point values} = \sum_i \text{Point value}_{ij} = \sum_{ij} \left(\frac{\log \text{LR}(a_{ij})}{\max |\log \text{LR}(a_{ij})|} \times 100 \right). \tag{4.2}$$

The total point values are the sum of corresponding m attribute values for risk factors.

4.3. Prevalence rate predicted through naïve Bayesian classifier model

In equation (3.4), $\sum_i \log \text{LR}(a_{ij})$ is represented by equation (4.2) as follows

$$P(c|X) = \frac{1}{1 + \exp\left(-\text{logit}P(c) - \max |\log \text{LR}(a_{ij})| \times \text{total point values}\right)}. \tag{4.3}$$

In equation (4.3), $\text{LR}(a_{ij})$ can be obtained by substituting $\max |\log \text{LR}(a_{ij})| \times$ total point values and through these equation, we can predict the prevalence rate of the disease by adding the point values in naïve Bayesian nomogram if we know the attribute values of each risk factor.

5. Application: CVD nomogram plot using naïve Bayesian classifier model

Now, we apply nomogram using naïve Bayesian classifier model to CVD data. 2013 to 2015 data was used in KNHANES VI, and the total were 7,856. The prevalence rate of CVD was 27.5%, the prevalence rate is 3.1% from 19 to 39 years old, 23.3% from 40 to 59 years old, and 60.2% from 60 to 80 years old (Table 1). Thus, it increases rapidly with age. Statistical analysis was conducted using Statistical Package for the Social Sciences (SPSS) Version 23.0. Table 1 is a result of Chi-square test for training data set. All variables, including sex, age, income, education level, marriage status,

Table 1: Characteristics of the study population

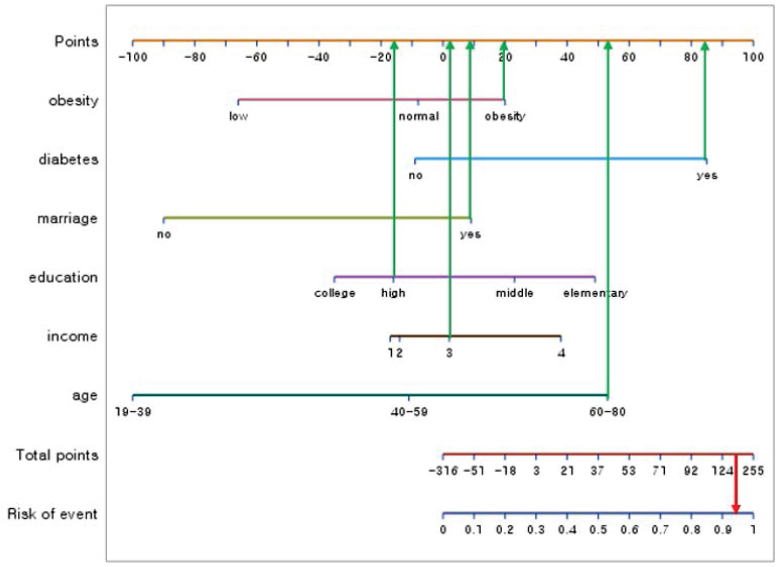
Characteristic		Cardiovascular disease		χ^2 (df)	p-value
		Yes n (%)	No n (%)		
Sex	male	1114 (30.6)	2529 (69.4)	14.9 (1)	0.0001
	female	1221 (26.7)	3349 (73.3)		
Age	19–39	78 (3.1)	2414 (96.9)	2069.8 (2)	< 0.0001
	40–59	749 (23.3)	2468 (76.7)		
	60–80	1508 (60.2)	996 (39.8)		
Income	<25%	716 (51.0)	689 (49.0)	472.2 (3)	< 0.0001
	25–50%	619 (29.5)	1479 (70.5)		
	50–75%	503 (21.7)	1814 (78.3)		
	>75%	497 (20.8)	1896 (79.2)		
Education level	elementary school	970 (57.6)	714 (42.4)	1136.9 (3)	< 0.0001
	middle school	366 (41.7)	512 (58.3)		
	high school	611 (21.2)	2272 (78.8)		
	college	388 (14.0)	2380 (86.0)		
Marriage status	married	2282 (33.1)	4615 (66.9)	458.6 (1)	< 0.0001
	single	53 (4.0)	1263 (96.0)		
Diabetes	yes	508 (77.2)	150 (22.8)	836.3 (1)	< 0.0001
	no	1827 (24.2)	5728 (75.8)		
Renal failure	yes	26 (72.2)	10 (27.8)	34.1 (1)	< 0.0001
	no	2309 (28.2)	5868 (71.8)		
Depression	yes	161 (42.5)	218 (57.0)	38.5 (1)	< 0.0001
	no	2174 (27.8)	5660 (72.2)		
Rheumatoid arthritis	yes	67 (48.9)	70 (51.1)	28.7 (1)	< 0.0001
	no	2268 (28.1)	5808 (71.9)		
Smoking status	current smoker	355 (22.5)	1223 (77.5)	97.4 (2)	< 0.0001
	past smoker	670 (37.1)	1138 (62.9)		
	non-smoker	1310 (27.1)	3517 (72.9)		
Alcohol status	none	647 (40.4)	956 (59.6)	178.1 (3)	< 0.0001
	<2/week	1115 (23.8)	3571 (76.2)		
	2-3/week	356 (27.1)	958 (72.9)		
	>4/week	217 (35.6)	393 (64.4)		
Stress	yes	486 (24.5)	1500 (75.5)	20.2 (1)	< 0.0001
	no	1849 (29.7)	4378 (70.3)		
Obesity status	lower weight	25 (7.1)	329 (92.9)	281.2 (2)	< 0.0001
	normal	1279 (24.3)	3974 (75.7)		
	obesity	1031 (39.6)	1575 (60.4)		
Starvation	yes	520 (20.1)	2062 (79.9)	127.2 (1)	< 0.0001
	no	1815 (32.2)	3816 (67.8)		

diabetes, renal failure, depression, rheumatoid arthritis, smoking status, alcohol status, stress, obesity, and starvation were significant at the 5% significance level and we choose 6 factors (age, income, education, marriage, diabetes, and obesity) as a risk factors for CVD according to χ^2 and p-value (Kang, 2018).

In Table 2, each likelihood ratio is obtained through equation (3.5) and the attribute value having the largest absolute $\log \mathrm{LR}(a_{ij})$ is age: 19–39. Since it has a negative value, the point value is assigned –100 points. The remaining point values are calculated by only $\log \mathrm{LR}(a_{ij})$ in equation (4.1). Of attribute values, attributes having negative points are age: 19–39 and 40–59, income: 50–75% and >75%, education level: high school and college, marriage status: single, and obesity: lower weight. They are attribute values that lower prevalence rate for CVD. However, age: 60–80, income: <25%, education level: elementary school and middle school, diabetes: yes, obesity status: obesity have positive point values and they increase the prevalence rate for CVD. There are also attribute values

Table 2: Log likelihood ratio and point values for risk factors

Risk factor	Attribute value	Likelihood ratio	Point value
Age	19–39	0.0813	−100
	40–59	0.7640	−11
	60–80	3.8114	53
Income	< 25%	2.6160	38
	25–50%	1.0536	2
	50–75%	0.6980	−14
	> 75%	0.6599	−17
Education level	elementary school	3.4199	49
	middle school	1.7995	23
	high school	0.6770	−16
	college	0.4104	−35
Marriage status	married	1.2448	9
	single	0.1056	−90
Diabetes	yes	8.5254	85
	no	0.8029	−9
Obesity status	lower weight	0.1913	−66
	normal	0.8102	−8
	obesity	1.6479	20



income: 1 (: >75%), 2 (: 50-75%), 3 (: 25-50%), 4 (: <25%)

Figure 1: *Proposed naïve Bayesian nomogram for cardiovascular disease.*

that are assigned a point close to zero (income: 25–50%, obesity: normal, marriage status: married, and diabetes: no). These have little impact on the prevalence rate.

The CVD nomogram plot was drawn using the point values for each attribute obtained from Table 2 (Figure 1). The composition of the nomogram is point line, 6 predictor lines, total point line, and probability line and it graphically represents the numerical relationships between CVD and 6 risk factors (age, income, education level, marriage status, diabetes, and obesity status). A nomogram plot is obtained from naïve Bayesian classifier model and each patient receives a point values for each

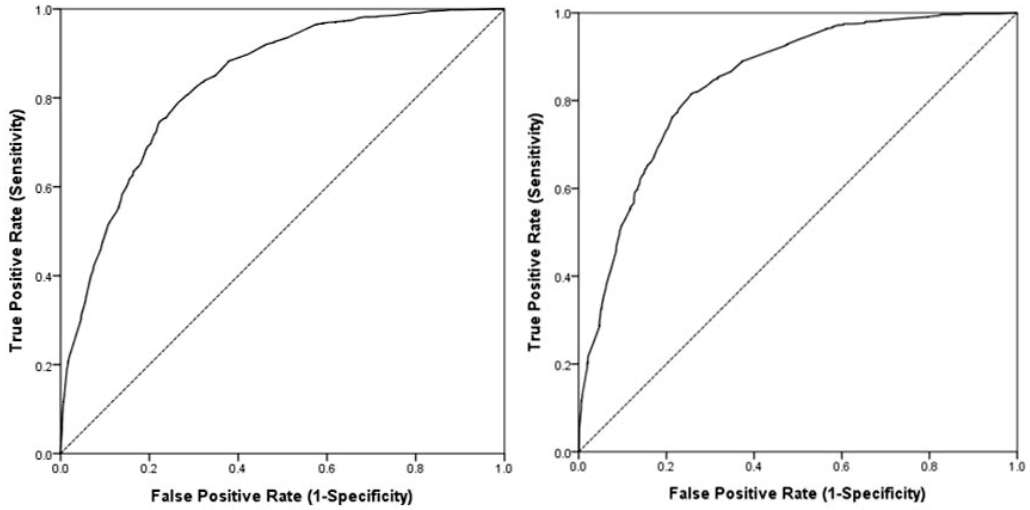


Figure 2: *Receiver operation characteristic curve for cardiovascular disease.*

factor. We can then get the predicted prevalence rate by adding all of those point values, find the total points corresponding to the total point line, and find the prevalence rate corresponding to the probability line. We can also calculate the prevalence rate by assigning a 0 point, even if we do not know the attribute value for a risk factor. In Figure 2 the assigned point values are: 19–39 in age is –100 points, 40–59 in age –11 points, 60–80 in age is 53 points, < 25% in income is 38 points, 25–50% in income is 2 points, 50–75% in income is –14 points, ≥ 75% in income is –17 points, elementary school in education level is 49 points, middle school in education level is 23 points, high school in education level is –16 points, college in education level is –35 points, married in marriage status is 9 points, single is –90, yes in diabetes is 85 points, no is –9, lower weight in obesity status is –66 points, normal is –8, and obesity is 20 points. The higher point values in each factor, the more important factor for CVD. Age has the longest line and it is the most influential factor for the prevalence of CVD according to age. Diabetes are assigned the largest point value. For example, a patient (age: 68 years old, income: 25–50%, education level: high school, marriage status: married, diabetes: yes, obesity status: obesity) has total point values of about 153, and prevalence rate of CVD is about 95%. In this case, the risk of having CVD is high, the patient needs to establish a treatment plan for CVD.

To verify the discrimination of the nomogram plot, an ROC curve was drawn in Figure 2. The left figure in Figure 2 is an ROC curve using the training set and the area under the ROC curve (AUC) is 0.836 ($p < 0.001$). The right of the figure in Figure 2 is an ROC curve using the test set and the AUC is 0.845 ($p < 0.001$). Therefore, we show a statistically significant determination and the nomogram was sufficiently powerful.

Figure 3 is a calibration plot for CVD and it is presented for verification for nomogram which is obtained for CVD. The groups with similar probability were grouped into 59 groups according to the predicted probabilities for nomogram, and the average value and the observation probability of each group were compared. The dotted line represents the straight line of $y = x$, and the straight line, the regression model is $y = 0.15 + 0.59 * x$, represents the regression line of the point values. The decision coefficient (R^2) is 0.693 and it reflects the calibration of the model.

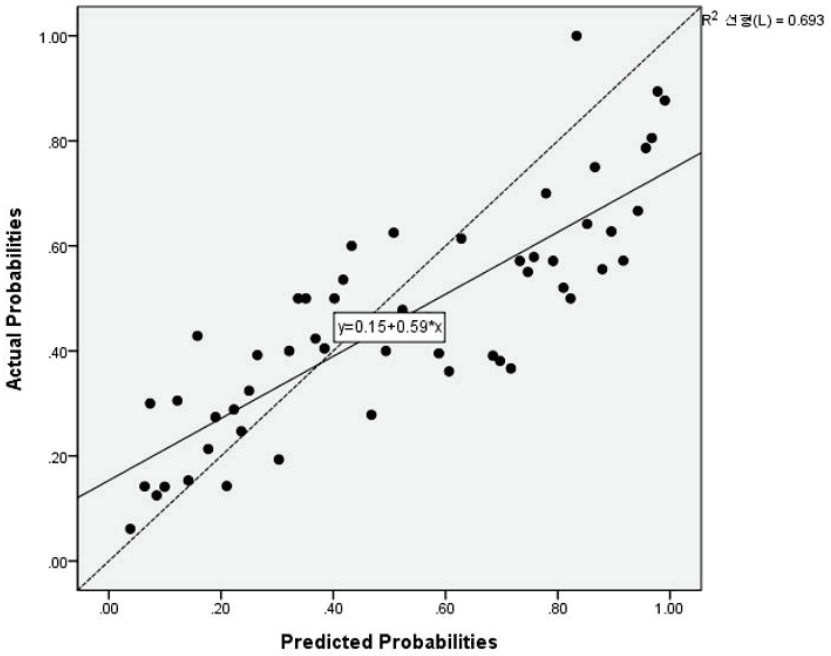
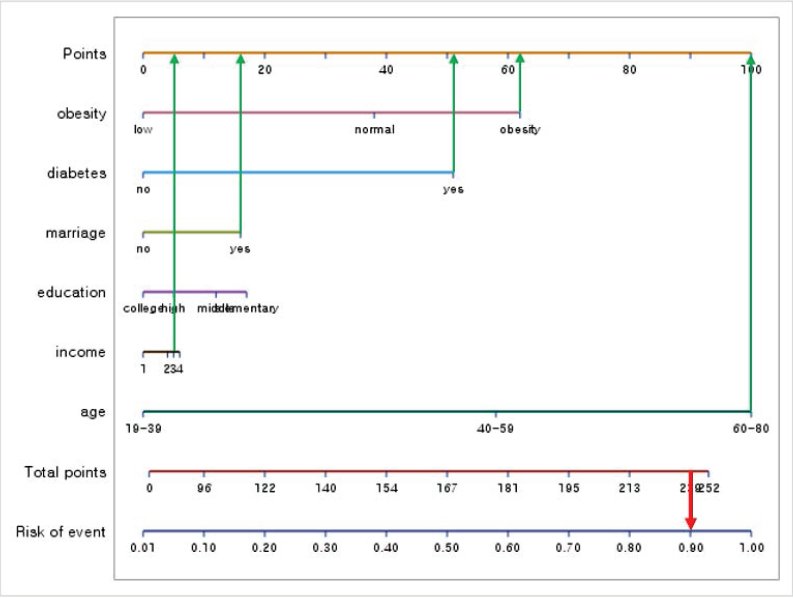


Figure 3: *Calibration plot for proposed nomogram of cardiovascular disease.*

6. Conclusion and discussion

A nomogram is a graphical tool and can be used for anyone to a predict prevalence rate because it is calculated by adding points. It has been developed for many diseases studied in preventive medicine or diagnostic medicine (Ahn *et al.*, 2014; Kattan *et al.*, 1998; Kim *et al.*, 2016). In this study, we introduced nomogram based on naïve Bayesian classifier model and applied it to CVD data from KNHANES 2013–2015. We used risk factors including age, income, education level, marriage status, diabetes, and obesity. In the proposed nomogram, we can compare the importance at a glance. Diabetes is widely known to be associated with CVD (Heart UK, 2015). It was assigned the highest point value and it is the attribute value that influences the highest prevalence rate in the study. Therefore, a medical history of diabetes necessitates greater attention for CVD is necessary. Also, of the factor lines, the longest line in the nomogram was age. The younger, the lower total points regardless of including other risk factors, but the older, the more the effect on the total points. Therefore, older individuals have a greater need to care for the heart and major arteries. Obesity is associated with a higher prevalence rate and it can be seen that the single lowers the prevalence because an unmarried person usually belongs to a young group at marriage status. For discrimination of the proposed nomogram plot, ROC curve and calibration plot were used and we can conclude that it represents a statistically significant determination.

The nomogram for CVD can also be built by a logistic regression model. Figure 4 is a nomogram when applied to the logistic regression model using the selected 6 factors. Diabetes and obesity factors should be considered in relation to the remaining social factors since we only considered the main effect in the logistic regression model, even though the length of the straight line is long. However, it is impossible to calculate it when we have missing values and calculate the prevalence rate using the



income: 1 (: >75%), 2 (: 50-75%), 3 (: 25-50%), 4 (: <25%)

Figure 4: *Nomogram using logistic regression analysis.*

logistic regression because there is no value to replace. Instead, the naïve Bayesian nomogram plot is useful to calculate it by replacing the 0 point without the need to consider overfitting for estimating regression coefficients as well as shrinkage for the problem. The prevalence rate is also calculated as posterior probability using the independence assumption between explanatory variables. However, it is disadvantageous in that it cannot independently identify the influence of each factor, unlike the logistic nomogram (Hosmer and Lemeshow, 2000).

We have proposed a naïve Bayesian nomogram for CVD using KNHANES data representing Korea for three years. This will also be useful in medical fields because dependencies are taken into consideration through conditional probability and the nomogram can be used to determine the CVD prevalence rate with attribute values. Also, for efficiency of naïve Bayesian nomogram, we are expected for various fields to build and use it in practice for early detection and treatment.

References

Ahn JH, Lee JZ, Chung MK, and Ha HK (2014). Nomogram for prediction of prostate cancer with serum prostate specific antigen less than 10 ng/mL, *Journal of Korean Medical Science*, **29**, 338–342.

Ambrose JA and Barua RS (2004). The pathophysiology of cigarette smoking and cardiovascular disease: an update, *Journal of the American College of Cardiology*, **43**, 1731–1737.

American College of Sports Medicine (2013). *ACSM's Guidelines for Exercise Testing and Prescription*, Lippincott Williams & Wilkins.

Bae Y and Lee K (2016). Risk factors for cardiovascular disease in adults aged 30 years and older, *Journal of The Korean Society of Integrative Medicine*, **4**, 97–107.

Britton A and McKee M (2000). The relation between alcohol and cardiovascular disease in Eastern Europe: explaining the paradox, *Journal of Epidemiology & Community Health*, **54**, 328–332.

Dimsdale JE (2008). Psychological stress and cardiovascular disease, *Journal of the American College of Cardiology*, **51**, 1237–1246.

Grundy SM, Benjamin IJ, Burke GL, *et al.* (1999). Diabetes and cardiovascular disease, *Circulation*, **100**, 1134–1146.

Heart UK (2015). Risk Factors for Cardiovascular Disease (CVD), from: https://heartuk.org.uk/files/uploads /documents/huk_fs_mfsI_riskfactorsforchd_v2.pdf

Hosmer DW and Lemeshow S (2000). Interpretation of the fitted logistic regression mode, Shewhart WA, Wilks SS Eds., *Applied Logistic Regression* (2nd ed), 47–90.

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.

Kang EJ (2018). Development of nomograms based on naïve Bayesian classifier and logistic regression model for predicting the prevalence rate of cardiovascular disease (Master's thesis), Yeungnam University, Gyeongsan.

Kattan MW, Eastham JA, Stapleton AMF, Wheeler TM, and Scardino PT (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer, *Journal of the National Cancer Institute*, **90**, 766–771.

Kawakami S, Numao N, Okubo Y, *et al.* (2008). Development, validation, and head-to-head comparison of logistic regression-based nomograms and artificial neural network models predicting prostate cancer on initial extended biopsy, *European Urology*, **54**, 601–611.

Kim W, Kim KS, and Park RW (2016). Nomogram of naive Bayesian model for recurrence prediction of breast cancer, *Healthcare Informatics Research*, **22**, 89–94.

Lavie CJ, Milani RV, and Ventura HO (2009). Obesity and cardiovascular disease: risk factor, paradox, and impact of weight loss, *Journal of the American College of Cardiology*, **53**, 1925–1932.

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

Lyssenko V, Jonsson A, Almgren P, *et al.* (2008). Clinical risk factors, DNA variants, and the development of type 2 diabetes, *New England Journal of Medicine*, **359**, 2220–2232.

Morrison DG (1969). On the interpretation of discriminant analysis, *Journal of Marketing Research*, **6**, 156–163.

Možina M, Demšar J, Kattan M, and Zupan B (2004). Nomograms for visualization of naive Bayesian classifier. In *Proceeding PKDD '04 Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Pisa, Italy, 337–348.

Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, and Eckel RH (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss, *Arteriosclerosis, Thrombosis, and Vascular Biology*, **26**, 968–976.

Wilson L, Bhatnagar P, and Townsend N (2017). Comparing trends in mortality from cardiovascular disease and cancer in the United Kingdom, 1983–2013: joinpoint regression analysis, *Population Health Metrics*, **15**, 23.

World Health Organization (2017). Cardiovascular diseases (CVDs), from: http://www.who.int/mediacentre/factsheets/fs317/en/Updated May 2017

Received February 5, 2018; Revised March 26, 2018; Accepted March 27, 2018