Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 207268, 12 pages
http://dx.doi.org/10.1155/2014/207268

Research Article

Evaluating the Risk of Metabolic Syndrome Based on an Artificial Intelligence Model

Hui Chen,[1] Shenghua Xiong,[2] and Xuan Ren[1]

[1] *Department of Endocrinology, The Second Hospital of Lanzhou University, Lanzhou 730030, China*
[2] *School of Mathematics and Statistics, Lanzhou University, Tianshui Road 222, Lanzhou, Gansu 730000, China*

Correspondence should be addressed to Hui Chen; 13909313366@163.com

Received 10 March 2014; Accepted 10 April 2014; Published 5 May 2014

Academic Editor: Fuding Xie

Metabolic syndrome is worldwide public health problem and is a serious threat to people's health and lives. Understanding the relationship between metabolic syndrome and the physical symptoms is a difficult and challenging task, and few studies have been performed in this field. It is important to classify adults who are at high risk of metabolic syndrome without having to use a biochemical index and, likewise, it is important to develop technology that has a high economic rate of return to simplify the complexity of this detection. In this paper, an artificial intelligence model was developed to identify adults at risk of metabolic syndrome based on physical signs; this artificial intelligence model achieved more powerful capacity for classification compared to the PCLR (principal component logistic regression) model. A case study was performed based on the physical signs data, without using a biochemical index, that was collected from the staff of Lanzhou Grid Company in Gansu province of China. The results show that the developed artificial intelligence model is an effective classification system for identifying individuals at high risk of metabolic syndrome.

1. Introduction

With the rapid development of world economy and the constant pursuit of a high quality of life, human health has become a main focus of attention. However, a growing number of diagnosed chronic diseases are resulting in a lower quality of life, and the resulting social and economic burden have become a huge obstacle in the pursuit of human progress.

1.1. Cardiovascular Risk Associated with Metabolic Syndrome. Metabolic syndrome (MetS) is a clustering of factors characterized by central obesity, lipid abnormalities, hypertension, impaired glucose metabolism, and insulin resistance, which is associated with an increased risk of type 2 diabetes mellitus (T2DM), cardiovascular disease (CVD), and mortality due to CVD. Other comorbidities include a proinflammatory state, prothrombotic state, nonalcoholic fatty liver disease, and reproductive dysfunction [1, 2]. Insulin resistance, metabolism disorders of the visceral adipose tissue, and inflammatory status are involved in the early stages of MetS.

Insulin resistance, obesity, hypertension, and diabetes are high risk factors for cardiovascular disease. While MetS encompasses all of the above syndromes, its effects on cardiovascular disease are exponential. In 2006, Butler et al. [3] sought to assess the impact of metabolic syndrome on cardiovascular outcomes in 3,031 individuals aged 70 to 79 years and found that metabolic syndrome was independently associated not only with coronary events (CE), myocardial infarction (MI), heart failure (HF), and all-cause hospital stays but also with cardiovascular and coronary mortality. In the same year, Ingelsson et al. [4] studied 2,314 middle-aged men without baseline HF or coronary heart disease. They found that metabolic syndrome was a significant predictor of HF and was independent of established risk factors for HF, including an interim myocardial infarction. A recent longitudinal study reported that people with metabolic syndrome had a 127%, 64%, 48%, and 39% greater risk of diabetes, ischemic heart disease, cardiovascular disease, and stroke, respectively, than normal controls [5]. All of these data imply that individuals with metabolic syndrome are at increased risk of morbidity and mortality from a variety of health conditions.

However, more worrying is that a relatively high prevalence of MetS has become a global phenomenon. Obesity is the pivotal driver of MetS development, and, as obesity levels increase, it will likely result in a parallel rise in the prevalence of metabolic syndrome. Available evidence demonstrates that, in most countries, 20% to 30% of adult population suffers from metabolic syndrome [6]. In the United States, an estimated 20% to 30% of adults suffer from metabolic syndrome. In Europe, the prevalence is between 19.8% and 24%. In some areas of Brazil, the prevalence is as high as 18% to 30%. Other population statistics report that the prevalence for Mexican, American, and Asian populations ranges between 12.4% and 28.5% among males and from 10.7% to 40.5% among females [7]. Not only is this a serious global public health problem but also it will greatly increase the economic burden on society. Healthcare costs for patients with MetS or other diseases were 3.36 times greater than those without MetS [8].

China is now facing the ageing of its population. Changes in lifestyle and longer life expectancy have led to an increased burden of cardiovascular and other chronic diseases, especially metabolic syndrome and a collection of multiple cardiovascular risk factors. A population-based cross-sectional study using the new International Diabetes Federation (IDF) definition of MetS shows that MetS has a higher prevalence in elderly people in Beijing, particularly in women. The prevalence rate in the elderly is as high as 46.3% (34.8% in men, 54.1% in women). Odds ratios (OR) for coronary heart disease (CHD), stroke, peripheral arterial disease (PAD), and CVD in patients who also had MetS were 1.69, 1.58, 1.42, and 1.73, respectively [9].

1.2. Various Definitions of the Metabolic Syndrome. Metabolic syndrome as a constellation of a clustering of cardiovascular risk factors (central obesity, lipid abnormalities, hypertension, impaired glucose metabolism, and insulin resistance) greatly increases the risk of mortality from CVD. Because of the resulting social and economic burden, scholars have been trying to find a more proper definition of the metabolic syndrome, but, for now, numerous debates still exist over disparity in the definitions.

The first widely used MetS criteria were developed by the World Health Organization (WHO) in 1998, with an emphasis on risk factors for type 2 diabetes mellitus [10]. In 2001, the National Cholesterol Education Program-Adult Treatment Panel III (NCEP ATP III) presented a MetS definition which focused on cardiovascular diseases [11]. Finally, in 2005, the International Diabetes Federation (IDF) proposed a new definition of MetS that includes central obesity as a prerequisite and gender- and ethnicity-specific cutoff points for central obesity as measured by waist circumference [12]. The WHO criteria are more complex than the NCEP criteria, because the former contains measurement of plasma insulin levels and microalbuminuria. It seems that the NCEP guidelines are more preferable.

However, Tan et al. [13] confirmed that the NCEP ATP III definition, when applied to Chinese and other Asian populations, underestimates the prevalence of the metabolic syndrome and fails to identify many individuals at risk of future CVD. This is due to the recommended cutoff points for waist circumference. It is inappropriate for Asian populations because Asian people have a higher percentage of body fat, especially abdominal visceral fat, than white people with the same body mass index (BMI) and also because Asians are prone to disorders such as diabetes, dyslipidemia, and hypertension at lower BMI levels than white people.

However, most current studies use WHO- or NCEP ATP III-modified Asian criteria to define MetS. The question remains if there is a more suitable definition for Chinese people. He et al. [9] analyzed the prevalence of MetS based on the IDF and NCEP ATP III criteria to verify the relation between MetS and CVD in a population-based survey of elderly Chinese people in Beijing, China. The study found that the prevalence of MetS defined by the IDF criteria was similar to that in the U.S. population in the same group defined by the NCEP criteria. Additionally, nearly 20% of the subjects met the IDF criteria but not the NCEP criteria; these subjects had significantly increased odds of CHD and stroke.

All of these data indicate that the new IDF criteria are more suitable than the NCEP criteria for screening higher-risk individuals and for estimating the risk of CVD from MetS in the Chinese population.

In summary, metabolic syndrome is a serious threat to human health, causing high morbidity and mortality, greatly reducing people's quality of life and survival beliefs, and causing huge medical expenses to families and society. Therefore, early diagnosis, early treatment, and a reduction in the risk of cardiovascular disease are urgently needed for the metabolic syndrome population. However, all of the diagnostic criteria include a variety of biochemical indices, which make these tests economically unfeasible for some populations, especially in rural areas of developing countries. Thus, an effective screening method has been sought to quickly identify those at high risk of MetS in underdeveloped countries and areas.

1.3. Mathematical Model in the Application of the Metabolic Syndrome. Based on the above global trends, many scholars have conducted research in preventive measures. Hirose et al. [14] have successfully predicted the 6-year incidence of MetS in Japanese male subjects using an artificial neural network (ANN) system and multiple logistic regression (MLR) analysis based on clinical data, including HOMA-IR and serum adiponectin. An artificial neural network is a computational methodology that performs multifactorial analyses and is characterized by self-learning and self-adapting. It has an excellent ability to learn and to generalize from experiences and to describe the highly nonlinear and strongly coupled relationships between multiinput and multioutput variables [15, 16]. Compared with other prediction methods, ANN methods are superior in terms of high data error tolerance, easy adaptability to online measurements, and no need for additional information other than body physical signs data. ANN is trained to "think" like humans by weakening or strengthening interconnected weights that connect its processing elements [17]. ANN has been widely applied

in pattern recognition and classification, prediction, and signal processing, among other applications. The applications for ANN in medicine are growing and include medical decision support, such as identifying and diagnosing cancer, hypertension, type 2 diabetes mellitus, and other diseases [16, 18]. A back propagation neural network (BPNN) is a classic artificial neural network. However, there are almost no studies addressing the use of artificial neural networks for rapid identification of those at high risk of metabolic syndrome without the use of biochemical parameters, particularly in rural residents.

Logistic regression analysis is a statistical method that uses the maximum likelihood to estimate regression coefficients. It does not require variables to obey the equal covariance matrix or residuals to follow a normal distribution, and therefore it has a wide range of applications in epidemiological studies. The logistic regression model requires that a nonlinear function relationship exists between the explanatory variables. However, in many studies, the variables often do not exist alone; there is a certain degree of linear dependence. This phenomenon is called multicollinearity. Multicollinearity often increases the standard error of estimated parameters, reduces the stability of the model, and even leads to the opposite result. Therefore, to reasonably estimate and interpret a regression model, we need to deal with multicollinearity between variables. Principal component analysis (PCA) is a common method to solve the collinearity problem in regression analysis. It is a multivariate technique introduced by Aguilera et al. that explains the variability of a set of variables in terms of a reduced set of uncorrelated linear spans of such variables with maximum variance, known as principal components (PCs) [19]. By principal component transformation, the highly relevant information of the variables is integrated into principal components with low correlation and then replaces the original variables in regression with the principal component.

1.4. The Structure of Paper. The remainder of this paper is organized as follows. The material and data are described in Section 2. The principal component regression and neural network are introduced in Sections 3 and 4, respectively. The ROC figure is illustrated in Section 5. The statistical analysis is explained in Section 6. Section 7 describes the results of the two models. We discuss the model and results in Section 8, and the conclusions are in Section 9. Finally, the limitations of the model are discussed in Section 10.

2. Material and Data Description

ANN and PCLR are used here for the first time to classify those at high risk of metabolic syndrome. The data and indices in the models are described here.

2.1. Study Population. In 2008, 2,107 subjects from a staff of Lanzhou Grid Company in Gansu province of China, with 23–60 years of age at baseline, were selected randomly and invited to participate in a MetS health survey. To be eligible, candidates had to be free of the following conditions: severe kidney, liver disease, cancer, psychological disorders, and infectious diseases. After candidates with severe kidney/liver disease ($n = 13$), cancer ($n = 9$), psychological disorders ($n = 4$), and infectious diseases ($n = 7$) were excluded, 2,074 individuals (male: 1,495, female: 579) who met the criteria were enrolled for analysis.

2.2. Data Acquisition. Anthropometry and biochemical data parameters were collected the morning after overnight fasting by trained physicians using standardized methods.

Height and body weight were measured twice with clothing (heavy clothing removed and 1.0 kg deducted for remaining garments) but without shoes to the nearest 0.1 cm and 0.1 kg, respectively. Body mass index (BMI) was calculated according to the following formula:

$$\text{BMI} = \frac{\text{Bodyweight}\,(\text{kg})}{\text{Height}^2\,(\text{m}^2)}. \tag{1}$$

Waist circumference (WC) was measured twice on standing subjects with a tape measure to the nearest 1 cm in the horizontal plane at the midpoint between the lowest rib and the iliac crest [18]. Hip circumference (HC) was measured twice to the nearest 1 cm at the widest part over the trochanters [20]. Blood pressure (BP) was measured twice from the right arm to the nearest 2 mm/Hg using a mercury sphygmomanometer, with subjects sitting and having relaxed for at least 30 min. BP measurements were taken in 10 min intervals, and mean values were calculated.

The blood specimens were collected using venipuncture after overnight fasting. Plasma glucose was determined using a modified hexokinase enzymatic method. Total cholesterol, triglycerides, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) were measured enzymatically with commercially available reagents. Determination of above indicators was completed using an automatic biochemical analyzer (instrument model: Hitachi automatic analyzer 7600-010).

2.3. Definition of Metabolic Syndrome. The International Diabetes Federation (IDF) criteria, which have been shown to be more suitable for Chinese people, were used in this study. The IDF criteria are as follows [12]: waist circumference ≥90 cm in males and 80 cm in females plus any two of the following criteria: triglyceride levels ≥1.7 mmol/L; HDL-cholesterol levels <1.04 mmol/L in males and <1.30 mmol/L in females; treatment of previously diagnosed hypertension and systolic BP ≥ 130 mmHg or diastolic BP ≥ 85 mmHg; fasting plasma glucose ≥5.6 mmol/L or previously diagnosed type 2 diabetes.

3. Principal Component Logistic Regression

The principal component logistic regression was introduced to establish a classification model. This method consists of two modules: principal component and logistic regression.

3.1. Principal Component. Different than the traditional logistic regression in which the variables are the explanatory variables, principal component logistic regression employs the principal components of the variables as the explanatory variables.

3.1.1. Introduction of Principal Component. Multivariate principal component analysis (PCA) is a multivariate technique to explain a set of correlated variables using a reduced number of uncorrelated variables with maximum variance, called PCs, introduced by Pearson at the beginning of the 20th century and developed by Hötteling in 1933 [21, 22].

PCA is defined from the sample point of view in this paper; namely, it is computed from a sample of observations about a set of variables [19]. Then, given a set of p continuous variables and n observations of such variables, a statistic matrix $X = (x_{ij})_{n\times p}$ in which x_{ij} is the ith observation of the jth variable is derived. The column vectors of such a matrix are denoted by X_1,X_2,...,X_p, where each vector corresponds to a variable.

The sample covariance matrix is denoted by $S = (s_{jk})_{p\times p}$, whose elements are defined in the form of $s_{jk} = (1/(n-1))\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$, where the sample means are computed by $\overline{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$ $(j = 1,\ldots,p)$. Without loss of generality, if the observations are centered; namely, $\overline{x}_1 = \cdots = \overline{x}_p = 0$, the sample covariance matrix can be simplified as $S = (1/(n-1))X'X$.

The principal components (PCs) of the sample are defined as orthogonal linear spans with maximum variance of the columns of the matrix X, denoted by $Z_j = XV_j$ $(j = 1,\ldots,p)$. The coefficient vectors $V_1,\ldots,V_p$, which define the PCs, are the eigenvectors of the sample covariance matrix S related to their corresponding eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, which are the variances of the corresponding PCs. Then, the matrix Z whose columns are the sample PCs can be calculated as $Z = XV$, where $V = (v_{jk})_{p\times p}$ is the matrix that contains the eigenvectors of the sample covariance matrix columns.

The sample covariance matrix can be decomposed as $S = V\Delta V'$, where V is orthogonal and $\Delta = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$ so that the matrix of observations can be given by $X = ZV'$.

Based on the PC decomposition, an approximated reconstruction of each original observation can be obtained in terms of a reduced number of PCs as follows:

$$X_j = \sum_{k=1}^{s} Z_k v_{jk}, \quad j = 1,\ldots,p. \tag{2}$$

The percentage of the total variability of the PCs accounting is given by the following:

$$\left[\frac{\sum_{j=1}^{s}\lambda_j}{\sum_{j=1}^{p}\lambda_j}\times 100\right], \quad s \leq p. \tag{3}$$

3.2. Binary Logistic Regression. Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1). This method is used to establish and evaluate the relationship between the binary variable and the explanatory variables.

3.2.1. Logit Transformation. The relationship between a binary dependent variable probability (P) and the continuous predictor (X) in logistic regression is usually an S-shaped curve generated by a Logit transformation function with asymptotes at 0 and 1, which is different than in a linear relationship between a continuous response variable (Y) and continuous predictor (X) within linear regression [23, 24]. The Logit transformation function of probability (P) is defined as follows:

$$\mathrm{Logit}\,(P) = \ln\left(\frac{P}{1-P}\right). \tag{4}$$

The Logit transformation is used to linearize the relationship between P and X, hence modifying the curved nature of the response. Consequently, logistic regression is constructed in the form of a linear function associating Logit(P) with the explanatory continuous variables. If the number of the independent variable X_i is n, then the linear function becomes the following:

$$\mathrm{Logit}\,(P) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n. \tag{5}$$

The direction of the relationship between the continuous X_i predictor and the Logit(P) is determined by the sign of coefficient b_i. To obtain the desired probability P as a function of the explanatory variable, we can change the form of formula (5); thereafter, the Logit(P) can be changed back into the probability scale as follows:

$$P = \frac{1}{1 + e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}}. \tag{6}$$

3.2.2. Maximum Likelihood Estimation (MLE). To calculate the coefficient estimates in binary logistic regression, the maximum likelihood estimation (MLE) procedure is used because it yields the most likely estimates of the unknown coefficients b_i in the binary regression model in which the probability of obtaining the observed set of data is maximized [24, 25]. The likelihood function is defined as follows [24, 25]:

$$L = \prod\left[P_i^{Y_i} \times (1-P_i)^{1-Y_i}\right]. \tag{7}$$

In (7), L is defined as the likelihood function, Y_i is the observed value (0 or 1) of the binary variable, P_i refers to the predicted probability of the case i (using the logistic regression model), and Π is the multiplicative equivalent to the summation sign and means that the function multiplies the values for each case [25]. The likelihood function L ranges from 0 to 1. To simplify the calculation and avoid the typically exceedingly small numbers, the likelihood function L is changed into a log likelihood function ($\ln L$) as follows [24, 25]:

$$\ln L = \sum_{i=1}^{n}\left[(Y_i \times \ln P_i) + (1-Y_i)\times\ln(1-P_i)\right]. \tag{8}$$

Because the range of L is from 0 to 1, $\ln L$ varies from negative infinity to zero. The closer the value of L to 1,

the closer the value of ln L to 0 and the more likely the parameters produce the observed data [25]. In practice, iterative techniques are used to estimate the coefficients b_i in the binary logistic regression model by maximizing ln L.

3.2.3. Goodness of Fit. In binary logistic regression, the Hosmer-Lemeshow test is introduced to assess how well the regression model explains the observed data by comparing the observed and the expected frequencies (by the model). For this, the expected probabilities of the event's occurrence are grouped from lowest to highest [24]. At the beginning, the data are dispersed over a finite number (g *which is not less than* 6) groups containing an equal number of data, with the cases with the lowest expected probabilities being contained in the first group, the next lowest expected probabilities are in the second group, and so on. Then, for the data in each group where $Y_i = 1$ (the event occurred), the estimates of the expected values can be obtained by summing the estimated probabilities over all of the data in a group. For the data in each group where $Y_i = 0$ (the event did not occur), the estimated expected value is obtained by summing over all data in the group, one minus the estimated probability [24]. Finally, a χ^2 test statistic is calculated based on observed and expected frequencies in a $g \times 2$ table, which follows a $(g-2)$ degrees of freedom chi-squared distribution. If the P value is high (>0.05), it means that that the model describes the data well [26].

3.2.4. Test of Significance Using Log Likelihood Values. The question whether a variable is significant or not should be answered using the MLE method after the coefficients b_i in the model are inferred. In another words, does the model that includes the predictor variables tell us more about the response variable than a model which does not include these variables? [24]. If the predicted values are better with the independent variable contained in the model, then the predictor variable in question is considered "significant."

In contrast to the significance test in linear regression in which ANOVA (analysis of variance) is commonly used in the assessment of the significance of the coefficients, in binary logistic regression, a two-stage procedure is employed. First, the G statistic is calculated based on the difference between the log likelihood of the base model (in this case, there are no predictors) and the model including the predictors. The statistic is given by the following:

$$\begin{aligned} G &= -2\ln L_{\text{base model}} - \left(-2\ln L_{\text{with variable}}\right) \\ &= -2\ln\left[\frac{\text{likelihood of base model}}{\text{likelihood with variable}}\right]. \end{aligned} \tag{9}$$

This G statistic follows a 2(DF) chi-square distribution, where the DF denotes the degrees of freedom and is just equal to the number of predictors [25]. If the probability of $[2(\text{DF}) > G] < 0.05$, then it is convincing evidence that the predictor included in the model significantly influences the dependent variable. If there are more predictors, a significant result suggests that at least one of the predictor variables is significantly associated with the response variable.

Second, if the predictor variable in the first step has been identified as significant, the significance of the individual regression coefficient(s) will be estimated by calculating the Wald test statistic, which is equal to the estimated regression coefficient b_i divided by its standard error (SE), and comparing with the standard normal z distribution [26].

3.3. Combination of the Two Methods. The principal components logistic regression (PCLR) as an extension of the principal component regression (PCR) model reduces the dimension of a logistic regression model with continuous covariates and provides an accurate estimation of the model parameters while avoiding multicollinearity. Aguilera et al. [19] proposed the PCLR model to address the collinearity in the logistic model.

In preparation to define the PCLR model, the Logit model is formulated in terms of all the PCs associated with the observations matrix X of the continuous predictor variables. Without loss of generality, the regressors are assumed to be centered, and the probabilities of the Logit model can be expressed in terms of all PCs as follows:

$$\begin{aligned} \pi_i &= \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik}v_{jk}\beta_j\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik}v_{jk}\beta_j\right\}} \\ &= \frac{\exp\left\{\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right\}}{1+\exp\left\{\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right\}}. \end{aligned} \tag{10}$$

z_{ik} $(i = 1, \ldots, n;\ k = 1, \ldots, p)$ is the element of the PCs matrix $Z = XV$ and $\gamma_k = \sum_{j=1}^{p} v_{jk}\beta_j$ $(k = 1, \ldots, p)$. To express the logistic model in matrix form, it can be given in terms of the Logit transformations and the PCs as follows:

$$L = X\beta = ZV'\beta = Z\gamma. \tag{11}$$

Because the model employs all the PCs as covariates, the parameters in the Logit model can be given as follows: $\beta = V\gamma$. Then, based on the invariance property of maximum likelihood estimates, the estimates of parameters can be obtained as follows: $\widehat{\beta} = V\widehat{\gamma}$, where the prediction of the dependent variable is based on the equation $\widehat{Y} = \widehat{\pi}$.

In practice, extensive collinearity exists in the explanatory variables, influencing the construction of the model. To address this collinearity in the estimation of the original parameters, the PCLR model is introduced by taking the covariates of the Logit model as a reduced set of PCs of the original predictors.

Matrices Z and V can be split in boxes as follows:

$$\begin{aligned} Z &= \left(\begin{array}{cccc|ccc} 1 & z_{11} & \cdots & z_{1s} & z_{1s+1} & \cdots & z_{1p} \\ 1 & z_{21} & \cdots & z_{2s} & z_{2s+1} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{n1} & \cdots & z_{ns} & z_{ns+1} & \cdots & z_{np} \end{array}\right) \\ &= \left(Z_{(s)} \middle| Z_{(r)}\right), \quad (r = p - s), \end{aligned} \tag{12}$$

$$V = \left(\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & v_{11} & \cdots & v_{1s} & v_{1s+1} & \cdots & v_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & v_{p1} & \cdots & v_{ps} & v_{ps+1} & \cdots & v_{pp} \end{array} \right) = \left(V_{(s)} \middle| V_{(r)} \right). \tag{13}$$

Because $Z = XV$, we can obtain $Z_{(s)} = XV_{(s)}$ and $Z_{(r)} = XV_{(r)}$, and then the original parameters can be expressed as follows:

$$\beta = V\gamma = V_{(s)}\gamma_{(s)} + V_{(r)}\gamma_{(r)}, \tag{14}$$

where $\gamma = (\gamma_0 \ \ \gamma_1 \ \ \cdots \ \ \gamma_s | \ \ \gamma_{s+1} \ \ \cdots \ \ \gamma_p)' = (\gamma'_{(s)} | \gamma'_{(r)})'$.

Taking (14) into (11), the Logit model containing all the PCs can be decomposed as $L = Z\gamma = Z_{(s)}\gamma_{(s)} + Z_{(r)}\gamma_{(r)}$. Then the PCLR model containing s PCs (PCLR(s)) is obtained by getting rid of the r last PCs in the last equation, so the dependent variables can be obtained as follows:

$$y_i = \pi_{i(s)} + \varepsilon_{i(s)}, \tag{15}$$

where

$$\pi_{i(s)} = \frac{\exp\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}{1 + \exp\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}, \quad i = 1, \ldots, n. \tag{16}$$

Based on the vector of Logit transformations $L_{(s)} = (l_{1(s)}, \ldots, l_{n(s)})$ with components $l_{i(s)} = \ln(\pi_{i(s)}/(1 - \pi_{i(s)}))$, the model can be expressed in matrix form as follows:

$$L_{(s)} = Z_{(s)}\gamma_{(s)} = XV_{(s)}\gamma_{(s)} = X\beta_{(s)}. \tag{17}$$

Therefore, a reconstruction of the original parameters can be given by $\beta_{(s)} = V_{(s)}\gamma_{(s)}$. In the new model, the parameters of the PCLR model just contain the first s PCs as covariates. An estimation of the original parameters β can be obtained using the MLE method as follows:

$$\widehat{\beta}_{(s)} = V_{(s)}\widehat{\gamma}_{(s)}. \tag{18}$$

The estimation $\widehat{\beta}$ can be improved if there is multicollinearity in the original variables [19]. In other words, PCLR employs the first s PCs except for the original predictor variables to explain the binary dependent variable.

4. Back Propagation Neural Network

As a typical artificial intelligent model, the back propagation (BP) neural network is one type of neural network with wide application. In this paper, a BP neural network is introduced to establish the classification model of the MetS.

4.1. The Structure of BP Neural Network. The topology of the BP neural network construction is shown in Figure 1. This BP neural network consists of three layers: the input layer to accept the input data, the hidden layer, and the output layer to output the result of the network. The number of nodes in the input layer is equal to the number of explanatory variables. In designing the hidden layer, the Hecht-Nelson method is employed, and $2i + 1$ was chosen as the number of nodes in the hidden layer; i is equal to the number of input nodes in input layer [27].

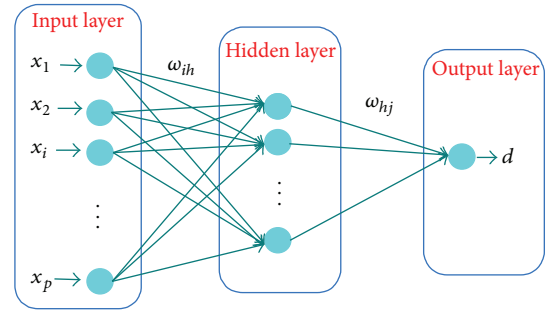


Figure 1: The topology of BP neural network. Note: $x_i, i = 1, 2, \ldots, p$ is the ith input variable of neural network, $\omega_{ih}, h = 1, 2, \ldots, q$ is the weight from the ith node in the input layer to hth node in the hidden layer, ω_{hj}, $j = 1$ is the weight from the hth node in the hidden layer to the ith node in the output layer, and d is the output of the neural network.

4.1.1. Data Standardization. Data standardization or normalization is the first step to use the neural network. Because differences in the dimensions and numeric range of the predictor variables usually exist, the appropriate processing should be carried out to transform the raw data before the network computes. In this paper, a method of normalization is adopted.

The input vectors and output vector are normalized according to the formula below, and then the final input and output data to train the BP neural network are obtained as follows:

$$x'_i = \frac{x_i - \min(x_i, \ldots, x_n)}{\max(x_i, \ldots, x_n) - \min(x_i, \ldots, x_n)}, \tag{19}$$

where x_i is the ith sample value of a variable.

4.1.2. Network's Training. The network should be trained before being used to forecast the dependent variable value. First, the network's structure should be initialized according to the rule of the BP algorithm in Section 4.1.3. Then, the data used to train the network is introduced into the network. Finally, the BP algorithm is employed to obtain the final parameters in the network. In the process of training the network, the number of nodes in the hidden layer is not easily determined; in the previous study, this was still an open question. In this paper, a trial and error method is employed to confirm the number of nodes.

4.1.3. Description of BP Algorithm. As a type of back propagation learning algorithm, BP algorithm is used in the BP neural network to train the network to perform efficiently to calculate the final parameters. In practice, the algorithm is departed into two parts: network training and network

testing. The steps of BP algorithm are described as follows [28].

Step 1 (initialization of the network). The parameters such as the net structure, layer numbers, number of nodes in each layer, input vector(X, D), weight ω_{ih} between the input layer and the hidden layer, weight ω_{hj} between the hidden layer and the output layer, learning rate η, momentum coefficient α, greatest acceptable MSE of the network eps, and other parameters are initialized in this step.

Step 2. Select a pattern and pass forward to calculate the hidden layer output as follows:

$$\text{net}_h^k = \sum_{i=1}^{p} \left(\omega_{ih}^k x_i^k - a_i^k\right), \qquad O_h^k = f\left(\text{net}_h^k\right). \tag{20}$$

net_h is denoted as the input of the hth node in the hidden layer, O_h is the output of the hth node in the hidden layer, a_i is the threshold from the ith node in the input layer to the hth node in the hidden layer, $f(x) = 1/(1+e^{-x})$, ^{k} means the kth iteration, and p is the number of nodes in input layer.

Then the output layer's result is calculated as follows:

$$\begin{gathered}\text{net}_j^k = \sum_{h=1}^{q} \left(\omega_{hj}^k O_h^k - b_i^k\right), \qquad O_j^k = \varphi\left(\text{net}_j^k\right) \\ y_j^k = O_j^k.\end{gathered} \tag{21}$$

net_j is denoted as the input of the jth node in the output layer. Here, $j = 1$, O_j is the output of the jth node in the output layer, b_i is the threshold from the hth node in the hidden layer to the jth node in the hidden layer, $\varphi(x) = 1/(1 + e^{-x})$ or $\varphi(x) = x$, and q is the number of nodes in hidden layer.

Step 3. Pass backward to calculate the neuron error beginning from the output layer as follows:

$$\delta_j^k = \left(d_j^k - O_j^k\right)\left(O_j^k\right), \qquad \Delta\omega_{hj}^k = \eta\delta_j^k O_h^k. \tag{22}$$

δ_j is the error of the jth node in output layer and $\Delta\omega_{hj}$ is the correction to ω_{hj}.

Then, calculate the hidden layer's error as follows:

$$\delta_h^k = \left(\sum_{j=1}^{r} \delta_j^k \omega_{hj}^k\right)\left(O_h^k\right), \qquad \Delta\omega_{ih}^k = \eta\delta_h^k x_i^k, \tag{23}$$

where δ_h is the error of the hth node in the output layer and $\Delta\omega_{ih}$ is the correction to ω_{hj}.

Step 4. Update the weights as follows:

$$\begin{aligned}\omega_{hj}^{k+1} &= \omega_{hj}^k + \Delta\omega_{hj}^k + \alpha\Delta\omega_{hj}^{k-1}, \\ \omega_{ih}^{k+1} &= \omega_{ih}^k + \Delta\omega_{ih}^k + \alpha\Delta\omega_{ih}^{k-1}.\end{aligned} \tag{24}$$

Table 1: Calculation of the sensitivity and specificity for a specific cut-off point of the predicted probability P.

Model result	Reality		
	Metabolic syndrome	No MetS	Total cases
Metabolic syndrome	a (true positive)	b (false positive)	$a + b$
No MetS	c (false negative)	d (true negative)	$c + d$
Total cases	$a + c$	$b + d$	

Sensitivity = $a/(a + c)$ = true positive rate.
Specificity = $d/(b + d)$ = true negative rate.
l − specificity = $1 - [d/(b + d)] = b/(b + d)$ = false positive rate.
"a" cases have a $P >$ cut-off point, which have an observed feather as well.
"b" cases have a $P >$ cut-off point, although no feather was observed in reality.
"c" cases have a $P <$ cut-off point, although feather was observed in reality.
"d" cases have a $P <$ cut-off point, and for these cases no feather was observed.

Step 5. Calculate the network's output error as follows:

$$\varepsilon^k = \frac{1}{r}\sum_{j=1}^{n}\left(y_j^k - d_j^k\right)^2. \tag{25}$$

If this value is greater than eps, then go back to Step 2 or the algorithm ends. After the above steps, the network will be trained to arrive at the preset accuracy. Once tested and assessed by a pile of patterns, the network can be in use [29].

5. The ROC Figure

The receiver operating characteristic (ROC) curve is commonly shown and discussed in reference handbooks on logistic regression [23, 24] and in the study of medical statistics. AUROC (the area under the ROC curve) provides a measure of the model's ability to correctly discriminate cases into proper category. The ROC curve is drawn based on a unique pair of values for sensitivity and specificity computed from the predicted (modeled) probability (P) under a range of cut-off points. For every cut-off point of the predicted probability P, a table is obtained as described in Table 1.

A unique pair of values for sensitivity and specificity is obtained for a series of cut-off points. Subsequently, the ROC curve is obtained by plotting the sensitivity (true positive rate) against one minus the specificity (the false positive rate) [23, 24]. The best prediction method performance would result in a point in the upper left corner of this plot, meaning 100% sensitivity and 100% specificity. If the ROC curve of a model lies in the upper left-hand quadrant of the graph, then the model has good discriminative ability, while if it lies along the 45°diagonal, then the model's discriminating ability is no better than chance at will [26]. The AUROC is commonly used for the validation of the predicted probabilities. To interpret the AUROC, Hosmer and Lemeshow [24] propose the general rule shown in Table 2.

The closer the curve is to the upper left corner of the graph, the better the predictive ability of the model is [30].

Table 2: The general assessment rule of AUROC.

Range of AUROC	Discrimination ability
AUROC = 0.5	No discrimination
0.7 ≤ AUROC < 0.8	Acceptable
0.8 ≤ AUROC < 0.9	Excellent
AUROC ≥ 0.9	Outstanding

6. Statistical Analysis

Data were represented as the mean ±SD and as percentages. One-way ANOVA was applied to evaluate the comparability of the training and validation data sets. The ANN and PCLR models were performed with the probability of MetS. The area under the receiver operating characteristic (ROC) curve was applied to measure the discrimination of the models using the validation data set. The PCLR model and ROC curve analysis were constructed by SPSS 17.0 and MATLAB 2010b (MathWorks Institute, USA). The ANN model was performed with MATLAB 2010b (MathWorks Institute, USA). P values of less than 0.05 were considered to be statistically significant.

7. Results

Two thousand and seventy-four individuals were included in the training and validation sets (mean age: 46.93 years and 47.06 years, resp.). The characteristics of all subjects are shown in Table 3. The prevalence rate of MetS was 23.0% ($n = 334$) and 20.0% ($n = 124$), respectively. Variables representing gender, age, BMI, WC, HC, WHR, SBP, and DBP had no significant differences between the training and validation sets ($P > 0.05$), which means that the two sets of data are comparable.

7.1. Classification Models and Predictive Performance. In this paper, we compare two methods in the evaluation of the risk of metabolic syndrome. The data flow in this paper is shown in Figure 2.

We split the data into two parts: the training data (70% of the data) and the testing data (30% of the data).

The response variable is MS, and the predictive variables are Sex, Age, BMI, WC, HC, WH, SBP, and DBP.

7.1.1. Principal Component Logistic Regression Classification Model. In this part, the data are organized in the form of [Sex, Age, BMI, WC, HC, WHR, SBP, and DBP], where every vector contains the sample data of the variable feather.

The training matrix contains the following data: $X_{\text{training}} = [\text{SEX AGE BMI WC HC WHR SBP DBP}]_{\text{training}}$. The testing matrix contains the following data: $X_{\text{testing}} = [\text{SEX AGE BMI WC HC WHR SBP DBP}]_{\text{testing}}$.

Step 1 (principal analysis). During the principal analysis, the main principles are obtained as follows.

The main principal in training data matrix is as follows:

$$\begin{aligned}
\text{PC}_1 &= -0.137\text{Sex} + 0.089\text{Age} + 0.231\text{BMI} + 0.251\text{WC} \\
&\quad + 0.201\text{HC} + 0.181\text{WHR} + 0.175\text{SBP} + 0.165\text{DBP} \\
\text{PC}_2 &= 0.194\text{Sex} + 0.439\text{Age} - 0.130\text{BMI} - 0.238\text{WC} \\
&\quad - 0.077\text{HC} - 0.271\text{WHR} + 0.496\text{SBP} + 0.335\text{DBP} \\
\text{PC}_3 &= 0.546\text{Sex} - 0.075\text{Age} + 0.338\text{BMI} + 0.048\text{WC} \\
&\quad + 0.549\text{HC} - 0.436\text{WHR} - 0.090\text{SBP} - 0.146\text{DBP}.
\end{aligned} \tag{26}$$

The main principle in testing data matrix is as follows:

$$\begin{aligned}
\text{PC}_1 &= -0.142\text{Sex} + 0.093\text{Age} + 0.230\text{BMI} + 0.253\text{WC} \\
&\quad + 0.192\text{HC} + 0.192\text{WHR} + 0.178\text{SBP} + 0.160\text{DBP} \\
\text{PC}_2 &= 0.045\text{Sex} + 0.402\text{Age} - 0.192\text{BMI} - 0.255\text{WC} \\
&\quad - 0.184\text{HC} - 0.199\text{WHR} + 0.494\text{SBP} + 0.394\text{DBP} \\
\text{PC}_3 &= 0.543\text{Sex} + 0.241\text{Age} + 0.278\text{BMI} + 0.001\text{WC} \\
&\quad + 0.513\text{HC} - 0.429\text{WHR} + 0.046\text{SBP} - 0.207\text{DBP}.
\end{aligned} \tag{27}$$

Step 2. Logistic regression based on the main principle in the training data matrix and MS.

SPSS17.0 was used for the logistic regression with the following result:

$$\text{Logit}(P) = -1.809 + 1.722\text{PC}_1 + 0.276\text{PC}_2 + 0.403\text{PC}_3. \tag{28}$$

P represents the probability of MS, PC_1, PC_2, and PC_3 to be the main principle. Table 4 shows the significance of the parameters in the logistic regression model.

From the table, every PC is significant to the response variable. Model (28) is considered as the desired PCLR model.

Step 3. Predict MetS using the main principal components.

The main principal components in testing data matrix were used as the input of model (28), and the predicted results of MS were obtained. The ROC and discrimination result of the PCLR model are shown in Figure 2.

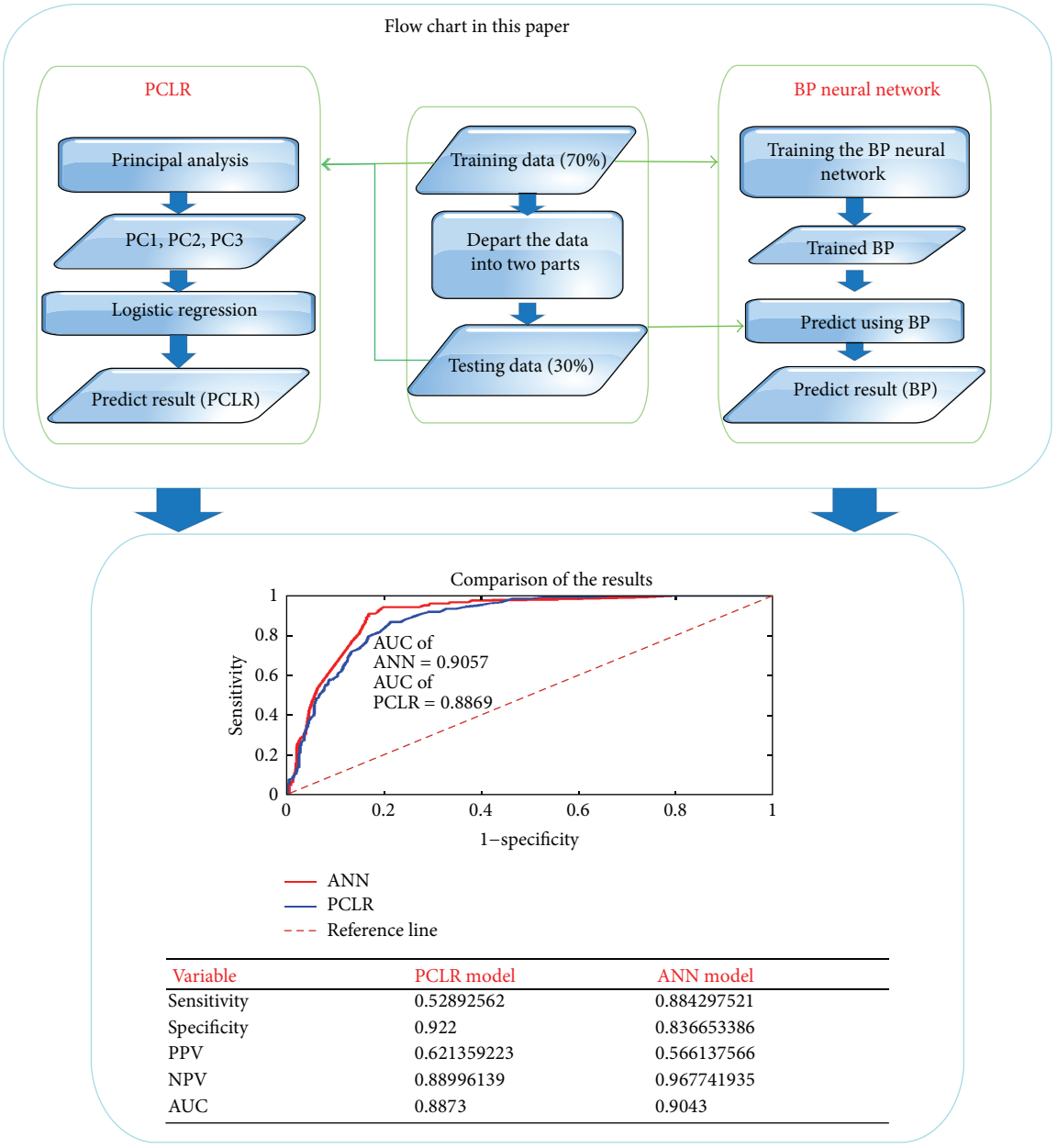
7.1.2. Back Propagation Neural Network Classification Model. Here, there are 8 input layer nodes because there are 8 explanatory variables (Sex, Age, BMI, WC, HC, WHR, SBP, and DBP) as the input of the BP neural network. The number of the hidden nodes of the BPANN is set in the range from 5 to 17 (2 times the number of input nodes). The number of nodes in the output layer is set to be 1. The structure of the BPANN we designed is shown in Figure 3. Here, MS is referred to MetS.

Step 1 (data normalization). Here, the following formula is used to normalize the data:

$$x_i' = \frac{x_i - \min(x_i, \ldots, x_n)}{\max(x_i, \ldots, x_n) - \min(x_i, \ldots, x_n)}. \tag{29}$$

Table 3: Comparison of baseline characteristics between training and validation sets.

Variables	Participants		P value
	Training set (N_1 = 1453)	Validation set (N_2 = 621)	
Sex	1.45 (0.4979)	1.42 (0.4937)	0.1514
Age	46.93 (12.2517)	47.06 (13.0362)	0.8294
BMI	23.78 (3.4992)	23.54 (3.4497)	0.1490
WC	84.69 (9.40180	84.55 (9.4059)	0.7627
HC	89.52 (6.2202)	89.30 (5.9638)	0.4457
WHR	0.95 (0.0726)	0.95 (0.0754)	0.8499
SBP	122.13 (20.0570)	121.57 (20.3817)	0.5663
DBP	78.39 (10.5885)	77.80 (10.9112)	0.2442
MS	0.23 (0.4231)	0.20 (0.3976)	0.0645



Variable	PCLR model	ANN model
Sensitivity	0.52892562	0.884297521
Specificity	0.922	0.836653386
PPV	0.621359223	0.566137566
NPV	0.88996139	0.967741935
AUC	0.8873	0.9043

Figure 2: Flow chart in this paper.

Table 4: The significance of the parameters in the logistic regression model.

Variable	OR (95% CI)	*P* value
PC_1	5.596 (4.552–6.880)	0.0000
PC_2	1.318 (1.147–1.514)	0.0000
PC_3	1.497 (1.291–1.734)	0.0000

Step 2 (training the neural network). Here, the normalized training data [SEX AGE BMI WC HC WHR SBP AND DBP] (every node corresponds to one node in the input layer) and training data MS are used as the output of the neural network to train the network. The network with the largest sensitivity was chosen as desired network in the case study, and the number of hidden nodes in neural network was confirmed to be 5.

Step 3 (predict MetS using predictive variables). The normalized testing data [SEX AGE BMI WC HC WHR SBP AND DBP] was used as the input of the desired network, and the network was used to obtain the predicted result. The ROC and discrimination result of the BP neural network model are shown in Figure 2.

7.2. Comparison of the Two Methods. The ROC of the two methods is shown in the Figure 2; from the figure, the AUROC of the BPANN is larger than PCLR. Thus, the BPANN has a better predictive ability than PCLR. Sensitivity, specificity, PPV, NPV and AUROC are depicted in Table 5.

From Table 5, BPANN has a higher predictive accuracy than PCLR, with a larger sensitivity value (0.884297521) for BPANN than for PCLR (0.52892562).

8. Discussion

Neural network models have been widely used in a variety of clinical medicine settings, such as cancer patient survival estimation, medical prognosis, predicting mortality, and evaluation of quality of life [31–34]. However, to our knowledge, this is the first study to establish and evaluate an effective quantitative model without biochemical parameters to predict individuals at high risk of MetS using the ANN model.

Neural networks have played a key role in medical decision making because they are effective in multifactorial analysis. They can consider many factors at the same time by combining and recombining the factors in different ways to provide appropriate decision supportive tools for prediction, classification, function fitting, and diagnostic tasks [16]. Model sensitivity and specificity are quite important when testing whether a model can accurately recognize positive and negative outcomes [35]. A successful model has both high sensitivity and high specificity [36]. In the current study, the results of the predictive performance showed that the ANN model had a higher predictive rate for identifying true positive or negative patients from undiagnosed MetS patients because it had sufficient sensitivity (88.42%) and specificity (83.66%) compared to the PCLR model (52.89% and 92.2%).

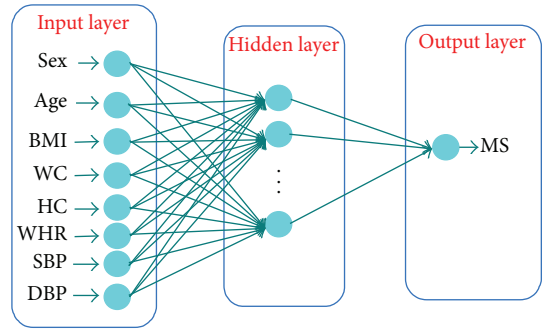


Figure 3: The structure of the designed BPANN.

Table 5: The result comparison of the two models.

Variable	ANN model	PCLR model
Sensitivity	0.884297521	0.52892562
Specificity	0.836653386	0.922
PPV	0.566137566	0.621359223
NPV	0.967741935	0.88996139
AUC	0.9043	0.8873

The ROC (receiver operating characteristic) curve is a valuable tool that plays a central role in the evaluation of the performance of a classification rule, especially in medical diagnoses [37]. The area under the ROC curve (AUC) can be defined as the probability of the classifier ranking a randomly chosen positive example higher than a randomly chosen negative example, and higher AUC values can be interpreted as having higher predictive accuracy [38, 39]. Our study used AUC values for performance comparisons of two prediction models. AUC values (AUC = 0.9043) obtained by the ANN model for identifying MetS were superior to values obtained by the PCLR model (AUC = 0.8873), which means that the ANN model had a higher predictive accuracy compared to the PCLR model.

Thus, the above comparisons confirm that the sensitivity, specificity, and the area under the ROC curve of the ANN model were significantly higher than those of the PCLR model. We can infer that the ANN model outperformed the PCLR model for screening those at high risk of MetS.

The practicality of the indicators is critical when evaluating whether a model can accurately distinguish between positive and negative results. Our study verified the feasibility of the predictors of risk of MetS. The input variables contained eight parameters significantly associated with MetS: age, gender, BMI, WC, HC, WHR, SBP, and DBP. Compared to Hirose et al. [14], all of the predictors without biochemical parameters, such as insulin resistance index and serum adiponectin, were more readily and quickly obtained in a population through data that is routinely collected in general practice or through epidemiology survey data.

MetS is thought to be a driver of the modern epidemics of CVD and has become a significant public health challenge around the world. According to the IDF criteria, the prevalence of MetS is 44% for men and 21% for women, and the morbidity rates will likely increase as people age

[40]. A systematic review and meta-analysis have confirmed that the MetS is associated with a 2-fold increase in risk of CVD, CVD mortality, myocardial infarction, and stroke, and a 1.5-fold increase in risk of all-cause mortality [41]. Therefore, the need for a low cost, fast, and effective predictive method for MetS is particularly urgent. Our ANN model identified an important gap in the literature. We suggest that healthcare workers use the ANN model as a forecasting tool for identifying patients who are at high risk of metabolic syndrome and cardiovascular events. There is an urgent need to use the predictive expressions developed here; this model will help physicians motivate individuals at annual health check-ups to change their lifestyles and diets and implement prevention and treatment strategies to reduce the prevalence of the metabolic syndrome and its associated cardiovascular risk.

9. Conclusions

The prevalence of MetS as a worldwide public health problem with high morbidity, high mortality, and high cost highlights the urgency of efforts to identify and modify risk factors for MetS. Convincing evidence has shown that, in developing countries, the incidences of metabolic syndrome are still rising and China is no exception. For early prevention and treatment to reduce the risk of CVD, stroke, myocardial infarction, and other diseases, it is important to develop effective public health strategies to quickly identify those at high risk of MetS in underdeveloped countries and areas. In this paper, we first use BPANN model and principal component analysis to forecast the high risk populations of MetS based on physical data without using biochemical parameters. From the experiment, we found that the principal component analysis is more scientific than the traditional logistic regression analysis, and the BPANN model is an effective classification approach for identifying those at high risk of metabolic syndrome; this system could help to reduce the social and medical burden of MetS in China.

Our Contribution. This is the first introduction of the BP artificial neural network in the study of MetS. To solve the commonly existing multicollinearity in the medical statistical data, the principal component analysis was employed to estimate the statistical model with satisfactory results.

10. Limitations

Although we took the lead in research and development of the ANN model without biochemical indicators for predicting an individual's risk of metabolic syndrome, study limitations should be considered. The main limitation to our study is that we did not include lifestyle (smoking, drinking, physical activity status, and fat intake); this is because of the lack of this clinical information, and we believe these data could be easily incorporated once available. Additionally, evaluations of the different models were based on a cross-sectional survey without long-term follow-up data.

The subjects, who were all from a Lanzhou electric power company, were limited ethnically and geographically. In spite of this, the ANN model based on a large population-based study was reliable and effective to screen undiagnosed metabolic syndrome patients.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Funding

This project was approved and funded by Gansu Provincial Sci. & Tech. Department.

References

[1] M. M. Takahashi, E. P. De Oliveira, A. L. R. De Carvalho et al., "Metabolic syndrome and dietary components are associated with coronary artery disease risk score in free-living adults: a cross-sectional study," *Diabetology and Metabolic Syndrome*, vol. 3, no. 1, article 7, 2011.

[2] M.-A. Cornier, D. Dabelea, T. L. Hernandez et al., "The metabolic syndrome," *Endocrine Reviews*, vol. 29, no. 7, pp. 777–822, 2008.

[3] J. Butler, N. Rodondi, Y. Zhu et al., "Metabolic syndrome and the risk of cardiovascular disease in older adults," *Journal of the American College of Cardiology*, vol. 47, no. 8, pp. 1595–1602, 2006.

[4] E. Ingelsson, J. Ärnlöv, L. Lind, and J. Sundström, "Metabolic syndrome and risk for heart failure in middle-aged men," *Heart*, vol. 92, no. 10, pp. 1409–1413, 2006.

[5] Y.-H. Khang, S.-I. Cho, and H.-R. Kim, "Risks for cardiovascular disease, stroke, ischaemic heart disease, and diabetes mellitus associated with the metabolic syndrome using the new harmonised definition: findings from nationally representative longitudinal data from an Asian population," *Atherosclerosis*, vol. 213, no. 2, pp. 579–585, 2010.

[6] S. M. Grundy, "Metabolic syndrome pandemic," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 28, no. 4, pp. 629–636, 2008.

[7] M. P. C. Leitão and I. S. Martins, "Prevalence and factors associated with metabolic syndrome in users of primary healthcare units in São Paulo-SP, Brazil," *Revista Da Associação Médica Brasileira*, vol. 58, no. 1, pp. 60–69, 2012.

[8] A. B. Schultz and D. W. Edington, "Analysis of the association between metabolic syndrome and disease in a workplace population over time," *Value in Health*, vol. 13, no. 2, pp. 258–264, 2010.

[9] Y. He, B. Jiang, J. Wang et al., "Prevalence of the metabolic syndrome and its relation to cardiovascular disease in an elderly Chinese population," *Journal of the American College of Cardiology*, vol. 47, no. 8, pp. 1588–1594, 2006.

[10] K. G. M. M. Alberti and P. Zimmet, "Definition diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus, Provisional report of a WHO consultation," *Diabetic Medicine*, vol. 15, no. 7, pp. 5397–5553, 1998.

[11] J. I. Cleeman, "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)," *Journal of the American Medical Association*, vol. 285, no. 19, pp. 2486–2497, 2001.

[12] K. G. M. M. Alberti and P. Zimmet, "The metabolic syndrome—a new worldwide definition," *The Lancet*, vol. 366, no. 9491, pp. 1059–1062, 2005.

[13] C.-E. Tan, S. Ma, D. Wai, S.-K. Chew, and E.-S. Tai, "Can we apply the National Cholesterol Education Program Adult Treatment Panel definition of the metabolic syndrome to Asians?" *Diabetes Care*, vol. 27, no. 5, pp. 1182–1186, 2004.

[14] H. Hirose, T. Takayama, S. Hozawa, T. Hibi, and I. Saito, "Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin," *Computers in Biology and Medicine*, vol. 41, no. 11, pp. 1051–1056, 2011.

[15] D. Yu, Y. Qing, Z. Jianxun, and D. Jun, "An artificial neural network approach to the predictive modeling of tensile force during renal suturing," *Annals of Biomedical Engineering*, vol. 41, no. 4, pp. 786–794, 2013.

[16] J. E. Dayhoff and J. M. DeLeo, "Artificial neural networks: opening the black box," *Cancer*, vol. 91, no. 8, pp. 1615–1635, 2001.

[17] L. Wang and K. Fu, *Artificial Neural Networks*, Wiley Online Library, 2008.

[18] C. Wang, L. Li, L. Wang, Z. Ping, M. T. Flory, G. Wang et al., "Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach," *Diabetes Research and Clinical Practice*, vol. 100, no. 1, pp. 111–118, 2013.

[19] A. M. Aguilera, M. Escabias, and M. J. Valderrama, "Using principal components for estimating logistic regression with high-dimensional multicollinear data," *Computational Statistics and Data Analysis*, vol. 50, no. 8, pp. 1905–1924, 2006.

[20] M. E. J. Lean, T. S. Han, and P. Deurenberg, "Predicting body composition by densitometry from simple anthropometric measurements," *American Journal of Clinical Nutrition*, vol. 63, no. 1, pp. 4–14, 1996.

[21] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[22] H. Hötteling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[23] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*, Sage, 2009.

[24] D. W. Hosmer Jr. and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, 2004.

[25] F. C. Pampel, *Logistic Regression: A Primer*, Sage, 2000.

[26] A. Petrie and C. Sabin, *Medical Statistics at A glance*, John Wiley & Sons, 2009.

[27] Z.-H. Guo, J. Wu, H.-Y. Lu, and J.-Z. Wang, "A case study on a hybrid wind speed forecasting method using BP neural network," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1048–1056, 2011.

[28] S. Kumar, *Neural Networks: A classroom Approach*, Tata McGraw-Hill Education, 2004.

[29] H. Xiaorui and L. Changchuan, "A preliminary study on targets association algorithm of radar and AIS using BP neural network," *Procedia Engineering*, vol. 15, pp. 1441–1445, 2011.

[30] B. Peeters, R. Dewil, and I. Y. Smets, "Improved process control of an industrial sludge centrifuge-dryer installation through binary logistic regression modeling of the fouling issues," *Journal of Process Control*, vol. 22, no. 7, pp. 1387–1396, 2012.

[31] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell et al., "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, no. 4, pp. 857–862, 1997.

[32] L. Ohno-Machado, "A comparison of Cox proportional hazards and artificial neural network models for medical prognosis," *Computers in Biology and Medicine*, vol. 27, no. 1, pp. 55–65, 1997.

[33] C.-C. Lin, Y.-K. Ou, S.-H. Chen, Y.-C. Liu, and J. Lin, "Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture," *Injury*, vol. 41, no. 8, pp. 869–873, 2010.

[34] M. R. N. Rao, G. R. Sridhar, K. Madhu, and A. A. Rao, "A clinical decision support system using multi-layer perceptron neural network to predict quality of life in diabetes," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 4, no. 1, pp. 57–59, 2010.

[35] W.-H. Ho, K.-T. Lee, H.-Y. Chen, T.-W. Ho, and H.-C. Chiu, "Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network," *PLoS ONE*, vol. 7, no. 1, Article ID e29179, 2012.

[36] H. K. Walker, W. D. Hall, and J. W. Hurst, *The Oral Cavity and Associated Structures—Clinical Methods*, The History, Physical, and Laboratory Examinations: Butterworths, 1990.

[37] P. Martínez-Camblor, C. Carleos, and N. Corral, "General nonparametric ROC curve comparison," *Journal of the Korean Statistical Society*, vol. 42, no. 1, pp. 71–81, 2013.

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[39] W.-S. Ke, Y. Hwang, and E. Lin, "Pharmacogenomics of drug efficacy in the interferon treatment of chronic hepatitis C using classification algorithms," *Advances and Applications in Bioinformatics and Chemistry*, vol. 3, no. 1, pp. 39–44, 2010.

[40] G. Nilsson, P. Hedberg, T. Jonason et al., "Waist circumference alone predicts insulin resistance as good as the metabolic syndrome in elderly women," *European Journal of Internal Medicine*, vol. 19, no. 7, pp. 520–526, 2008.

[41] S. Mottillo, K. B. Filion, J. Genest et al., "The metabolic syndrome and cardiovascular risk: a systematic review and meta-analysis," *Journal of the American College of Cardiology*, vol. 56, no. 14, pp. 1113–1132, 2010.