Medicine®

OPEN

An artificial neural network prediction model of congenital heart disease based on risk factors

A hospital-based case-control study

Huixia Li, PhD[a,b], Miyang Luo, MD[c], Jianfei Zheng, MD[d], Jiayou Luo, PhD[b,*], Rong Zeng, MPH[e], Na Feng, MPH[a], Qiyun Du, MD[a], Junqun Fang, MPH[a]

Abstract

An artificial neural network (ANN) model was developed to predict the risks of congenital heart disease (CHD) in pregnant women.

This hospital-based case-control study involved 119 CHD cases and 239 controls all recruited from birth defect surveillance hospitals in Hunan Province between July 2013 and June 2014. All subjects were interviewed face-to-face to fill in a questionnaire that covered 36 CHD-related variables. The 358 subjects were randomly divided into a training set and a testing set at the ratio of 85:15. The training set was used to identify the significant predictors of CHD by univariate logistic regression analyses and develop a standard feed-forward back-propagation neural network (BPNN) model for the prediction of CHD. The testing set was used to test and evaluate the performance of the ANN model. Univariate logistic regression analyses were performed on SPSS 18.0. The ANN models were developed on Matlab 7.1.

The univariate logistic regression identified 15 predictors that were significantly associated with CHD, including education level (odds ratio =0.55), gravidity (1.95), parity (2.01), history of abnormal reproduction (2.49), family history of CHD (5.23), maternal chronic disease (4.19), maternal upper respiratory tract infection (2.08), environmental pollution around maternal dwelling place (3.63), maternal exposure to occupational hazards (3.53), maternal mental stress (2.48), paternal chronic disease (4.87), paternal exposure to occupational hazards (2.51), intake of vegetable/fruit (0.45), intake of fish/shrimp/meat/egg (0.59), and intake of milk/soymilk (0.55). After many trials, we selected a 3-layer BPNN model with 15, 12, and 1 neuron in the input, hidden, and output layers, respectively, as the best prediction model. The prediction model has accuracies of 0.91 and 0.86 on the training and testing sets, respectively. The sensitivity, specificity, and Yuden Index on the testing set (training set) are 0.78 (0.83), 0.90 (0.95), and 0.68 (0.78), respectively. The areas under the receiver operating curve on the testing and training sets are 0.87 and 0.97, respectively.

This study suggests that the BPNN model could be used to predict the risk of CHD in individuals. This model should be further improved by large-sample-size research.

Abbreviations: ANN = artificial neural network, AUC = area under the receiver operating curve, BPNN = back-propagation neural network, CHD = congenital heart disease, MSE = average square error, NPV = negative predictive value, OR = odds ratio, PPV = positive predictive value, ROC = receiver operating curves.

Keywords: artificial neural network (ANN), congenital heart disease (CHD), prediction model

1. Introduction

Congenital heart disease (CHD) is the most common congenital malformation and 1 leading cause of infant mortality. The global incidence rate of CHD is 6.8 to 9.0 per 1000 live births.[1–3] Its most common subtypes are atrial septal defect, ventricular septal defect, patent ductus arteriosus, pulmonary stenosis, and tetralogy of Fallot.[4,5] The World Health Organization statistics in 2014 shows that 1.5 million infants are born with CHD in the world each year. The China Birth Defects Prevention Report in 2012 shows that over 130,000 infants in China are born with CHD each year, causing a total economic burden of more than 12.6 billion yuan. CHD can result in long-term disability and need long-term expert medical care, so CHD becomes a major global health problem that may significantly impact individuals, families, and the society.

Many studies in this field are focused on using epidemiological data and/or clinical characteristics that can be used to predict adverse pregnancy outcomes (APOs), such as preterm birth, low birth weight, small-for-gestational-age, large-for-gestational-age, intrauterine fetal demise, and neonatal death.[6–15] Preterm birth can be predicted by many methods (e.g., risk scoring systems and logistic regression models) that are based on epidemiological

Editor: Antonio Palazón-Bru.

This study was supported by the National Natural Science Foundation of China (grants number 81172680).

The authors have no conflicts of interest to disclose.

[a] Department of Child Health Care, Hunan Provincial Maternal and Child Health Care Hospital, [b] Department of Maternal and Child Health, Xiangya School of Public Health, Central South University, Changsha, Hunan Province, China, [c] Department of Epidemiology, Saw Swee Hock School of Public Health, National University of Singapore, Singapore, [d] Department of Emergency and Intensive Care Medicine, The Second Xiangya Hospital, [e] Department of Pharmacy, Xiangya Hospital, Central South University, Changsha, Hunan Province, China.

[*] Correspondence: Jiayou Luo, Department of Maternal and Child Health, Xiangya School of Public Health, Central South University, No. 110, Xiangya Road, Changsha, Hunan Province 410008, China (e-mail: jiayouluo@126.com).

Medicine (2017) 96:6(e6090)

Received: 9 June 2016 / Received in final form: 2 January 2017 / Accepted: 18 January 2017

http://dx.doi.org/10.1097/MD.0000000000006090

data,[6,7] biochemical markers (e.g., fetal fibronection,[8] amniotic fluid urocortin-1,[9]), or sonographic parameters (e.g., cervical length[10]). These methods have sensitivity of 0.25 to 0.82, and specificity of about 0.60. Low-birth-weight can be predicted by 2 risk scoring systems based on epidemiological data and clinical characteristics (hemoglobin concentration) that have relatively high sensitivity and specificity.[11,12] Moreover, small-for-gestational-age can be predicted by logistic regression models based on simple maternal demographic factors[13] or second-trimester fetal sonographic parameters (e.g., abdominal circumference and head circumference).[14] These models have similar performances, with sensitivity of 0.52 to 0.73 and specificity of 0.50 to 0.77. Despite the unsatisfactory sensitivity and generally low specificity, these predictions are significant clinical attempts to predict specific APOs and are pivotal in classified management of pregnant women and in prevention of APOs. In all, prediction of APOs is a promising research trend, but the existing models should be further explored and improved.

However, there is rare research about risk prediction of individual congenital malformations including CHD. Previously, by using logistic regression and a decision tree, we established a fetus CHD prediction model based on epidemiological data in early pregnancy, but due to the small sample, we only reported the accuracy while ignoring the sensitivity and specificity.[16] Thus, there is no comprehensive evaluation for the prediction models.

The aim of this study is to develop an effective CHD prediction model using artificial neural networks (ANN) and based on comprehensive epidemiological data. This model can be used as a preliminary screening tool to identify pregnant women who were at high risk of CHD in early pregnancy, and be helpful for prenatal care providers in guiding prenatal management and prevention.

2. Materials and methods

2.1. Subjects

In this hospital-based case-control study, subjects were all recruited from birth defect surveillance hospitals in Hunan Province, China. Mothers who gave birth to CHD infants between July 2013 and June 2014 in these hospitals were involved as cases. CHD was diagnosed by heart specialists. The exclusion criterion were: chromosomal anomalies or other birth defects of known etiology; isolated patent ductus arteriosus or patent foramen ovale in premature infants, or the diameters of pulmonary artery end or patent foramen <3 mm in full-term infants in 24 hours after birth; presence of congenital anomalies other than CHD; refusal or inability to participate in the survey because of mental symptoms, thinking, or memory disorders.

In this hospital-based study, the number of cases was relatively small, and a large number of potential controls were selected from the birth defect surveillance hospitals. To diminish potential risk of bias as much as possible and ensure the statistical power needed to detect an important predictor, we randomly selected those mothers who delivered normal infants without any congenital anomalies at the same hospitals and same time period as controls. Those who could not cooperate with the survey were also excluded from the study.

Informed consent was obtained from all individuals before the interview. Ethical approval was obtained from the Ethics Committee of Xiangya School of Public Health, Central South University. All the procedures in this study conformed to the Declaration of Helsinki.

2.2. Data collection

All subjects were interviewed face-to-face by well-trained obstetricians and gynecologists and asked to fill in a questionnaire. The questionnaire included 36 variables from 5 categories: sociodemographic characteristics, pregnancy history, family history, environmental risk factors, and dietary/lifestyle behaviors during pregnancy (Table 1). The questionnaire was designed by the experts from our research team and modified based on a pilot study.

2.3. Measurements of risk factors

2.3.1. *Sociodemographic characteristics.* Ethnicity was classified into 2 categories: Han and minorities (minorities were the other 55 ethnicities in China except Han). Residence was divided into urban and rural residences. Education level was classified into 3 categories: primary school and below; middle school; college and above. Occupations included farmers, migrant workers, employers/ managers, workers, administrative staff, and housewives or else.

2.3.2. *Pregnancy history.* Maternal pregnancy history consisted of gravidity, parity, and history of abnormal reproduction (stillbirth, spontaneous abortion, or birth defect).

2.3.3. *Family history.* Family history of CHD was defined as 1 or more first relatives of a CHD patient.

Table 1

Data collection from the subjects.

Category	Variable
Sociodemographic characteristics	Maternal age, ethnicity, residence, education level, occupation
Pregnancy history	Gravidity, parity, history of abnormal reproduction
Family history	Family history of CHD
Environmental risk factors	Maternal environmental risk factors: chronic disease, upper respiratory tract infection, reproductive system infection, complications of pregnancy, contraceptive intake, ovulation drugs intake, pets-keeping, folic acid intake, environmental pollution around dwelling place, exposure to occupational hazards, pesticide exposure, mental stress
	Paternal environmental risk factors: chronic disease, exposure to occupational hazards
Dietary and lifestyle behaviors	Maternal dietary behaviors: intake of picked/smoked food, intake of vegetable and fruit, intake of fish/shrimp/meat/egg, and intake of milk/soymilk
	Maternal lifestyle behaviors: smoking, alcohol drinking, betel nuts chewing, and strong tea drinking
	Paternal lifestyle behaviors: smoking, alcohol drinking, betel buts chewing, strong tea drinking, and drug taking

CHD = congenital heart disease.

2.3.4. Environmental risk factors. There were maternal and paternal aspects. Information about exposure to environmental risk factors was collected using the questions with the answer "yes or no." The exposure time of maternal risks was defined as from "6 months before conception" to "the first trimester of pregnancy," while the exposure time of paternal risks was 6 months before conception.

2.3.5. Dietary and lifestyle behaviors. Maternal dietary patterns referred to the dietary behaviors in the first trimester of pregnancy. Each of the 4 maternal dietary behaviors mentioned in Table 1 was classified into 3 scales: ≤ 2, 3 to 5, and ≥ 6 times per week. Maternal lifestyle behaviors and paternal lifestyle behaviors were summarized from the same periods as environmental risk factors. Smoking was defined as smoking any cigarette during pregnancy. Alcohol drinking was defined as drinking any liquor, including beer, wine, and white spirit. Strong tea drinking was defined as more than 200 mL per day on average. Drug taking was defined as taking any drugs, including heroin, cannabis, morphine, cocaine, and ketamine.

2.4. Statistical analysis

The subjects were randomly divided into a training set and a testing set at the ratio of 85:15 (The subjects with missing data >20% were excluded from the study). The training set was used to screen out the predictors using univariate logistic regression and develop ANN prediction models. The testing set was used to test and evaluate the performance of ANN models.

Analysis was undertaken in 3 stages. In the first stage, univariate logistic regression was performed to identify the significant predictors of CHD based on the training set. Continuous variables were sorted into categories to facilitate the risk factor identification. Maternal age was classified into 4 groups: ≤ 24, 25 to 29, 30 to 34, and ≥ 35 years old. Gravidity was divided into 3 groups: 1, 2 to 3, and >3. Parity was sorted into 2 categories: 1, and ≥ 2.

In the second stage, we developed ANN models for the prediction of CHD risk, and inputted the significant predictors selected in the first stage. A standard feed-forward back-propagation neural network (BPNN) was applied due to its relative simplicity and stability.[17,18] In general, a BPNN consists of 3 layers: an input layer that receives information, a hidden layer that processes information, and an output layer that calculates results.[19] BPNN was run with the significant predictors as the input variables and the risk of CHD as the output variable. The numbers of neurons in the input and output layers (marked as N and M, respectively) corresponded to the numbers of significant predictors and output variables, respectively. The number of neurons in the hidden layer (H) was not any actual variable. The optimal H was determined by trial and error, since no authoritative theory is available for such predetermination.[20] The optimal H was determined from the prediction model with the highest sensitivity and specificity. Nevertheless, the trial range of H could be determined as follows: $H=\sqrt{M+N}+a$, where a is a constant ranging from 1 to 10. All data were normalized to the range of 0 to 1. For binary variables, 0 means "No" and 1 means "Yes." Nonbinary variables were normalized as $x'_m=(x_m-x_{min})/(x_{max}-x_{min})$. Continuous log-sigmoid functions were used as the transfer functions of the hidden and output layers. The Levenberg–Marquardt algorithm was used as the training function. The Learngdm algorithm was used as an adaptive learning function. The training parameters such as learning rate and momentum were set at their default values. The networks were trained at a maximum of 100 epochs or until the minimum average square error (MSE) was <0.001.

In the third stage, we assessed the performance of the BPNN model. We calculated its accuracy, sensitivity, specificity, Yuden Index, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating curve (AUC) on both sets. We also plotted receiver operating curves on both sets.

The categorical and continuous variables were both compared between the training set and the testing set by using χ^2 test, Fisher exact test, and t test. Comparisons and univariate logistic regression analyses were all performed on SPSS 18.0 (IBM, Chicago, IL). The BPNN models were developed on Matlab 7.1 (MathWorks, Natick, MA). The significance level was set at $P<0.05$.

3. Results

3.1. Sociodemographic characteristics of the subjects

Initially, 366 subjects (123 cases and 243 controls) were enrolled. Eight subjects with missing data >20% were excluded from the study. Finally, 358 subjects (119 cases and 239 controls) were included, with a valid response rate of 97.8%. The 358 subjects were randomly divided into a training set involving 300 subjects (101 cases and 199 controls) and a testing set involving 58 subjects (18 cases and 40 controls).

The sociodemographic characteristics of the 2 sets are listed in Table 2. The subjects are aged 27.2 ± 4.0 years old (mean ± standard deviation [SD]) in the training set and 27.9 ± 4.3 years old in the testing set. The subjects are predominantly Han and mainly lived in urban areas. The main education level is middle school. The predominant occupations are housewives or else, farmers, and employers/managers. The sociodemographic characteristics are not significantly different between the 2 sets ($P>0.05$).

3.2. Predictors of CHD risk

The 36 variables listed in Table 1 were analyzed by univariate logistic regression. Table 3 shows the 15 significant predictors of CHD risk selected by univariate logistic regression based on the training set (n=300). The following 11 factors are significantly associated with the increased risk of CHD: gravidity (OR=1.95), parity (2.01), history of abnormal reproduction (2.49), family history of CHD (5.23), maternal chronic disease (4.19), maternal upper respiratory tract infection (2.08), environmental pollution around maternal dwelling place (3.63), maternal exposure to occupational hazards (3.53), maternal mental stress (2.48), paternal chronic disease (4.87), and paternal exposure to occupational hazards (2.51). The occurrence of CHD is inversely related to 4 protective factors, including high education level (OR=0.55), intake of vegetable/fruit (0.45), intake of fish/shrimp/meat/egg (0.59), and intake of milk/soymilk (0.55). None of the other 21 variables is significantly associated with CHD.

3.3. BPNN prediction models

BPNN models were built based on the significant CHD risk predictors. The input variables are the 15 significant predictors mentioned above, and the output variable is the binary variable whether an individual gave birth to a CHD infant. The BPNN models each consist of an input layer, a hidden layer, and an

Table 2

Sociodemographic characteristics of the subjects.

Characteristics	Training set (n=300)	Testing set (n=58)	Test statistic	*P* value
Age, y, mean (SD)	27.2 (4.0)	27.9 (4.3)	$t=-1.18$	0.24
Ethnicity (%)				
Han	294 (98.0)	57 (98.3)	—	0.89*
Minorities	6 (2.0)	1 (1.7)		
Residence (%)				
Urban	209 (69.7)	39 (67.2)	$\chi^2=0.13$	0.71
Rural	91 (30.3)	19 (32.8)		
Education level (%)				
Primary school and below	14 (4.7)	2 (3.4)	—	0.72*
Middle school	166 (55.3)	29 (50.0)		
College and above	120 (40.0)	27 (46.6)		
Occupation n (%)				
Farmers	60 (20.0)	9 (15.5)	$\chi^2=3.72$	0.59
Migrant workers	34 (11.3)	3 (5.2)		
Employers/managers	52 (17.3)	13 (22.4)		
Workers	34 (11.3)	8 (13.8)		
Staffs in administrative institutions	51 (17.0)	9 (15.5)		
Housewives or else	69 (23.0)	16 (27.6)		

SD=standard deviation.

* The *P* value was estimated by using Fisher exact test.

output layer. The input and output layers contain 15 and 1 neuron, respectively, corresponding to the numbers of predictors and output variables, respectively. The reference equation mentioned above indicates that the number of neurons in the hidden layer, *H*, ranges from 5 to 14. Therefore, we developed the BPNN model in which *H* increased by 1 from 5 to 14. Finally, a 3-layer BPNN model with 15, 12, and 1 neuron in the input, hidden, and output layers, respectively, was selected as the best prediction model (see Fig. 1).

3.4. Performance of BPNN model

Table 4 shows the performances of the BPNN model on the training and testing sets. The accuracies on training and testing sets are 0.91 and 0.86, respectively. The sensitivity, specificity and Yuden Index on the testing set (training set) are 0.78 (0.83), 0.90 (0.95), and 0.68 (0.78), respectively. To provide more information about the model performances, we also calculated PPV and NPV, which are 0.78 (0.89) and 0.90 (0.92), respectively, on the testing set (training set).

Figure 2 shows the ROC curves for the BPNN model on both sets. The AUCs on the training and testing sets are 0.97 (see Fig. 2A) and 0.87 (see Fig. 2B), respectively.

Thus, the well-trained optimal BPNN model here could successfully predict the individual risk of CHD, with high accuracy and large AUC.

4. Discussion

ANN is ideal for prediction of disease occurrence in individuals, and specifically, it fits a nonlinear correlation between input and output variables until reaching high accuracy. ANN models have been applied to predict the occurrence of hypertension,[19] cardiovascular autonomic dysfunction,[20] coronary artery disease,[21,22] and metabolic syndrome.[23] These studies demonstrate ANN can accurately predict diverse clinical settings and

Table 3

Predictors of congenital heart disease analyzed by univariate logistic regression.

Variable	B	SE	Wald χ^2	*P*	OR	95% CI (OR)	
Education level	−0.60	0.22	7.29	0.01	0.55	0.36	0.85
Gravidity	0.67	0.20	11.77	<0.001	1.95	1.33	2.86
Parity	0.70	0.31	5.14	0.02	2.01	1.10	3.67
History of abnormal reproduction	0.91	0.39	5.50	0.02	2.49	1.16	5.33
Family history of CHD	1.66	0.55	9.14	<0.001	5.23	1.79	15.30
Maternal chronic disease	1.43	0.63	5.26	0.02	4.19	1.23	14.28
Maternal upper respiratory tract infection	0.73	0.52	8.53	<0.001	2.08	1.27	3.41
Environmental pollution around maternal dwelling place	1.29	0.64	4.07	0.04	3.63	1.04	12.71
Maternal exposure to occupational hazards	1.26	0.47	7.27	0.01	3.53	1.41	8.82
Maternal mental stress	0.91	0.37	6.09	0.01	2.48	1.21	5.10
Paternal chronic disease	1.58	0.70	5.09	0.02	4.87	1.23	19.24
Paternal exposure to occupational hazards	0.92	0.41	4.93	0.03	2.51	1.11	5.65
Intake of vegetable/fruit	−0.81	0.27	9.32	<0.001	0.45	0.27	0.75
Intake of fish/shrimp/meat/egg	−0.53	0.25	4.68	0.03	0.59	0.36	0.95
Intake of milk/soymilk	−0.59	0.25	5.70	0.02	0.55	0.34	0.90

CHD=congenital heart disease, CI=confidence interval, OR=odds ratio.

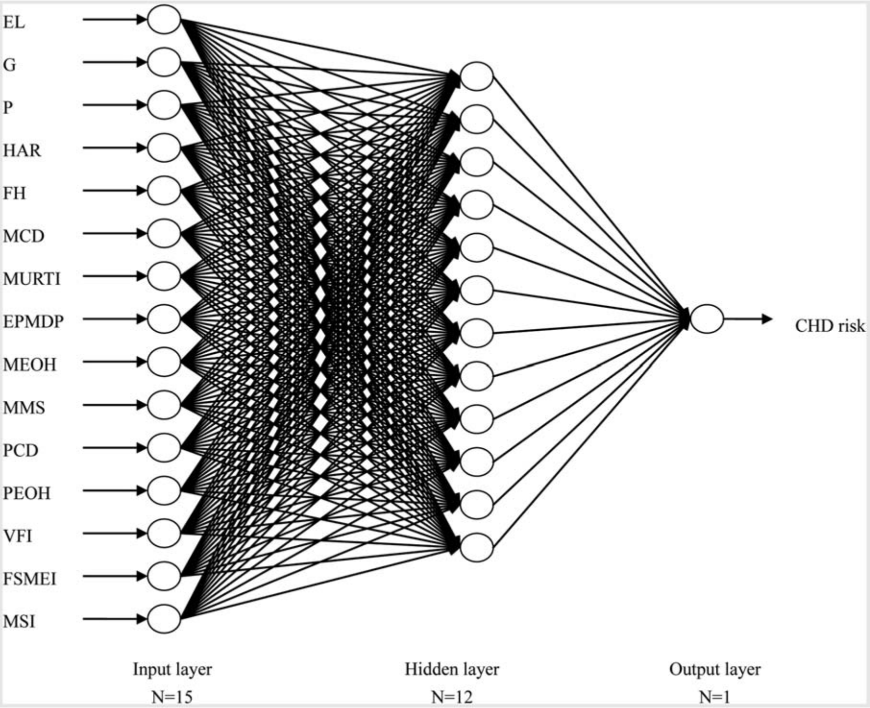

Figure 1. Architecture of back-propagation neural network for predicting congenital heart disease risk. The circles represent neurons, and the lines between circles represent modifiable connections. CHD = congenital heart disease, EL = education level, EPMDP = environmental pollution around maternal dwelling place, FH = family history, FSMEI = fish/shrimp/meat/egg intake, G = gravidity, HAR = history of abnormal reproduction, MCD = maternal chronic disease, MEOH = maternal exposure to occupational hazards, MMS = maternal mental stress, MSI = milk/ soymilk intake, MURTI = maternal upper respiratory tract infection, P = parity, PCD = paternal chronic disease, PEOH = paternal exposure to occupational hazards, VFI = vegetable/fruit intake.

outperforms conventional statistical methods. Using BPNN, we developed a CHD prediction model involving 15 significant predictors, including maternal education level, gravidity, parity, history of abnormal reproduction, family history, maternal chronic disease, maternal upper respiratory tract infection, environmental pollution around maternal dwelling place, maternal exposure to occupational hazards, maternal mental stress, paternal chronic disease, paternal exposure to occupational hazards, intake of vegetable/fruit, fish/shrimp/meat/egg, and milk/soymilk. For the testing set (training set), this model has sensitivity of 0.78 (0.83), specificity of 0.90 (0.95), accuracy of 0.86 (0.91), and Yuden Index of 0.68 (0.78), suggesting that this BPNN model is successful and valuable. This is the first study to develop a high-performance CHD prediction BPNN model based on risk factors.

Table 4

Performances of back-propagation neural network on training and testing sets.

Indicator	Training set	Testing set
Accuracy	0.91	0.86
Sensitivity	0.83	0.78
Specificity	0.95	0.90
Yuden index	0.78	0.68
PPV	0.89	0.78
NPV	0.92	0.90
AUC (95% CI)	0.97 (0.95, 0.99)	0.87 (0.75, 0.98)

AUC = area under the receiver operating curve, NPV = negative predictive value, PPV = positive predictive value.

So far, the existing APO prediction models based on epidemiological data and/or clinical characteristics are mainly risk scoring models and conventional statistical models. With logistic regression and based on maternal demographic factors (maternal age, race, education, marital status, parity, prenatal care initiations, and smoking), Tan et al[6] built preterm birth models for prediction of singleton, twin, and triplet pregnancies, and got the sensitivity of 0.25, 0.65, and 0.64, respectively, and the specificity of 0.94, 0.57, and 0.54, respectively. Based on 6 variables (weight gain in the mother during pregnancy, intake of proteins in diet, history of preterm birth, history of low birth weight, maternal anemia, and passive smoking), Singh et al[11] developed a weighted risk score model for the prediction of low birth weight and obtained the sensitivity of 0.72 and specificity of 0.64. In this study, we established a relatively stable BPNN model for CHD prediction using comprehensive epidemiological data from maternal and paternal factors (e.g., common sociodemographic characteristics, pregnancy history, family history, environmental risk factors, dietary, and lifestyle factors) and for training and testing sets, and achieved sensitivity of 0.83 and 0.78, specificity of 0.95 and 0.90, accuracy of 0.91 and 0.86. Our prediction model outperforms the above methods, which indicates the superiority and rationality of BPNN models in solving complex nonlinear relationships.

The new model based on epidemiological data can be used as a preliminary screening tool to identify pregnant women at high risk of CHD in early pregnancy. The model-predicted risk probability helps prenatal care providers to guide prenatal management and prevent high-risk pregnant women. The predictors included in the model are common and available in prenatal routine practice, and this model might be applicable to

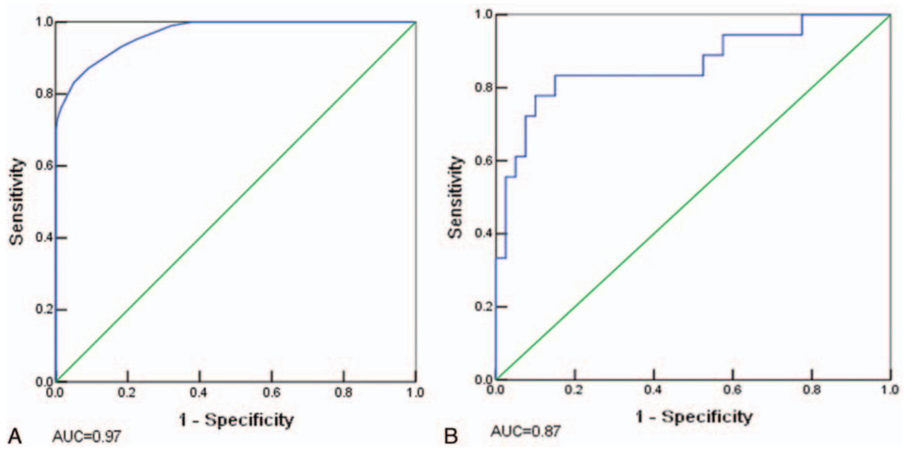

Figure 2. Receiver operating characteristic (ROC) curves for the back-propagation neural network on training and testing sets. A, ROC curve and corresponding area under the receiver operating curve (AUC) on the training sets. B, ROC curve and corresponding AUC on the testing set.

the general population. The new model could be saved as a program in the computer. After a clinician inputs the 15 predictors of a pregnant woman into the program, the computer automatically calculates the probability of CHD. The new model may help clinicians to identify pregnant women at high-risk CHD, who should be considered with high priority for prenatal counseling and diagnosis. Thus, more-expensive and complicated prenatal diagnosis technology (e.g., fetal echocardiography, chromosome karyotype analysis, and gene detection) can be more efficient for high-risk pregnant women. Furthermore, the new model can also be used to prevent the causes of CHD. For example, by applying presumed data, women of childbearing age could estimate the CHD risk of their future babies. The model would help them to reduce the exposure to environmental risk factors and conduct healthy dietary and lifestyle behaviors throughout the pregnant course. This is the fundament of CHD primary prevention.

In our study, the BPNN predictors seem reasonable as most of them are reportedly associated with CHD. The risk factors of CHD reported in previous epidemiological studies include maternal low education level,[24,25] family history of CHD,[26–28] maternal diseases (e.g., upper respiratory tract infection, hyperhomocysteinaemia, phenylketonuria, diabetes mellitus, hypertension, thyroid disorders, obesity),[25,29–31] maternal and paternal exposures to occupational/environmental risks,[32–34] and maternal mental stress.[25] Most predictors selected by logistic regression in our study are consistent with previous reports. However, unlike other findings,[26,31,35–37] we do not find maternal age, maternal medication exposure, complications of pregnancy, or parental lifestyle (alcohol drinking, smoking) as significant risk factors of CHD. This inconsistency may result from the small sample size and low exposure rates of those investigated factors in our study. We also find that intake of vegetable/fruit, intake of fish/shrimp/meat/egg, and intake of milk/soymilk are all protective factors of CHD. Thus, the CHD risk can be alleviated by reducing the exposure to environmental risk factors and appropriately increasing intakes of vegetable/fruit, fish/shrimp/ meat/egg, and milk/soymilk during pregnancy.

The BPNN model developed here has high specificity but relatively low sensitivity. The sensitivities on training and testing sets are 0.83 and 0.78, respectively. The low sensitivity may be attributed to 2 reasons. First, except for family history of CHD, the remaining 14 significant predictors are not specific indicators of CHD, but are common environmental risk factors for a variety of birth defects, such as neural tube defects, orofacial clefts, renal malformations, congenital hydrocephalus, and congenital clubfoot. Second, not all of risk factors of CHD were included as the predictors. Due to the small sample size and low exposure rates of some investigated factors, only a small number of significant risk factors were identified by logistic regression. As the occurrence of CHD may be affected by multiple unknown factors, the BPNN model should be updated continuously. Therefore, further large-sample-size research is needed to identify the specific predictors of CHD (e.g., CHD-associated biological markers or genes) and to improve the model sensitivity.

This study has several limitations. First, we developed the BPNN model using epidemiological data, mainly including family history and environmental factors, but did not consider relevant laboratory data such as biochemical indicators and genetic factors (CHD-associated genes). The epidemiological data was collected using a case-control study, which was a retrospective observational study and susceptible to bias. The CHD cases and controls were only a sample of the source population, so there might be potential selection bias, which could be seen as 1 “incoherence” of the cases and controls with respect to the corresponding population at risk.[38] The method of data collection in our study was based on self-report by the subjects (recall of past events), which inevitably led to recall bias in the data. In addition, measurement bias might also exist in the CHD case ascertainment. Second, most of factors measured were dichotomous variables rather than continuous variables, without considering dose response relationship between exposure levels of these risk factors and CHD, which may hide their true relationships with CHD. Third, the training and testing samples were all from the same population. The predictive performance of the new model was not validated in other populations, and its generalizability could not be correctly determined. Fourth, the model could not be expressed by specific equations due to the complexity of ANN and the weak explanatory of their weights. The application of this model was not as simple and convenient as nomogram models for the clinicians,[39] since it relies on a computer and a specific program.

5. Conclusions

Despite the limitations, this is the first study using BPNN to estimate the CHD risk for pregnant women. With the new BPNN model, we can identify pregnant women at high-risk CHD in early pregnancy, and the model-predicted risk probability is helpful for prenatal care providers in guiding prenatal management and prevention of high-risk pregnant women.

Acknowledgments

The authors express their great gratitude to Professor Xianru Liu from School of Information Science and Engineering, Central South University for her artificial neural network technology support. The authors thank all the birth defects surveillance hospitals of Hunan Province for supplying CHD cases and controls to the study.

References

[1] Abid D, Elloumi A, Abid L, et al. Congenital heart disease in 37,294 births in Tunisia: birth prevalence and mortality rate. Cardiol Young 2014;24:866–71.
[2] Yu ZB, Xi YY, Ding WX, et al. Congenital heart disease in a Chinese hospital: pre- and postnatal detection, incidence, clinical characteristics and outcomes. Pediatr Int 2011;53:1059–65.
[3] Dadvand P, Rankin J, Shirley MD, et al. Descriptive epidemiology of congenital heart disease in Northern England. Paediatr Perinat Epidemiol 2009;23:58–65.
[4] Rahman F, Salman M, Akhter N, et al. Pattern of congenital heart diseases. MMJ 2012;21:246–50.
[5] Yeh SJ, Chen HC, Lu CW, et al. Prevalence, mortality, and the disease burden of pediatric congenital heart disease in Taiwan. Pediatr Neonatol 2013;54:113–8.
[6] Tan H, Wen SW, Chen XK, et al. Early prediction of preterm birth for singleton, twin, and triplet pregnancies. Eur J Obstet Gynecol Reprod Biol 2007;131:132–7.
[7] Goyal NK, Hall ES, Greenberg JM, et al. Risk prediction for adverse pregnancy outcomes in a medicaid population. J Womens Health (Larchmt) 2015;24:681–8.
[8] Ruiz RJ, Fullerton J, Brown CE. The utility of fFN for the prediction of preterm birth in twin gestations. J Obstet Gynecol Neonatal Nurs 2004;33:446–54.
[9] Karaer A, Celik E, Celik O, et al. Amniotic fluid urocortin-1 concentrations for the prediction of preterm delivery. J Obstet Gynaecol Res 2013;39:1236–41.
[10] Jung EY, Park JW, Ryu A, et al. Prediction of impending preterm delivery based on sonographic cervical length and different cytokine levels in cervicovaginal fluid in preterm labor. J Obstet Gynaecol Res 2016;42: 158–65.
[11] Singh A, Arya S, Chellani H, et al. Prediction model for low birth weight and its validation. Indian J Pediatr 2014;81:24–8.
[12] Metgud C, Naik V, Mallapur M. Prediction of low birth weight using modified Indian council of medical research antenatal scoring method. J Matern Fetal Neonatal Med 2013;26:1812–5.
[13] Wen SW, Tan H, Yang Q, et al. Prediction of small for gestational age by logistic regression in twins. Aust N Z J Obstet Gynaecol 2005;45: 399–404.
[14] Seravalli V, Block-Abraham DM, Turan OM, et al. Second-trimester prediction of delivery of a small-for-gestational-age neonate: integrating sequential Doppler information, fetal biometry, and maternal characteristics. Prenat Diagn 2014;34:1037–43.
[15] Rossi A, Vogrig E, Ganzitti L, et al. Prediction of large-for-gestation neonates with first-trimester maternal serum PAPP-A. Minerva Ginecol 2014;66:443–7.
[16] Zhou LB, Zheng L, Luo JY, et al. [Risk prediction model of perinatal congenital heart disease] (article in Chinese). Zhonghua Liu Xing Bing Xue Za Zhi 2008;29:1251–4.
[17] Grossi E. How artificial intelligence tools can be used to assess individual patient risk in cardiovascular disease: problems with the current methods. BMC Cardiovasc Disord 2006;6:20.
[18] Montie JE, Wei JT. Artificial neural networks for prostate carcinoma risk assessment: an overview. Cancer 2000;88:2655–60.
[19] Huang SQ, Xu YH, Yue L, et al. Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a Chinese area. Hypertens Res 2010;33:722–6.
[20] Tang ZH, Liu J, Zeng F, et al. Comparison of prediction model for cardiovascular autonomic dysfunction using artificial neural network and logistic regression analysis. PLoS One 2013;8:e70571.
[21] Harrison RF, Kennedy RL. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med 2005;46:431–9.
[22] Colak MC, Colak C, Kocatürk H, et al. Predicting coronary artery disease using different artificial neural network models. Anadolu Kardiyol Derg 2008;8:249–54.
[23] Hirose H, Takayama T, Hozawa S, et al. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. Comput Biol Med 2011;41:1051–6.
[24] Kuciene R, Dulskiene V. Maternal socioeconomic and lifestyle factors during pregnancy and the risk of congenital heart defects. Medicina (Kaunas) 2009;45:904–9.
[25] Liu SW, Liu JX, Tang J, et al. Environmental risk factors for congenital heart disease in the Shandong Peninsula, China: a hospital-based case-control study. J Epidemiol 2009;19:122–30.
[26] Fung A, Manlhiot C, Naik S, et al. Impact of prenatal risk factors on congenital heart disease in the current era. J Am Heart Assoc 2013;2: e000064.
[27] Wang X, Li P, Chen S, et al. Influence of genes and the environment in familial congenital heart defects. Mol Med Rep 2014;9:695–700.
[28] Liu X, Liu G, Wang P, et al. Prevalence of congenital heart disease and its related risk indicators among 90796 Chinese infants aged less than 6 months in Tianjin. Int J Epidemiol 2015;44:884–93.
[29] Verkleij-Hagoort AC, Verlinde M, Ursem NT, et al. Maternal hyperhomocysteinaemia is a risk factor for congenital heart disease. BJOG 2006;113:1412–8.
[30] Kuciene R, Dulskiene V. Selected environmental risk factors and congenital heart defects. Medicina (Kaunas) 2008;44:827–32.
[31] Liu S, Joseph KS, Lisonkova S, et al. Association between maternal chronic conditions and congenital heart defects: a population-based cohort study. Circulation 2013;128:583–9.
[32] Cresci M, Foffa I, Ait-Ali L, et al. Maternal and paternal environmental risk factors, metabolizing GSTM1 and GSTT1 polymorphisms, and congenital heart disease. Am J Cardiol 2011;108:1625–31.
[33] Gorini F, Chiappa E, Gargani L, et al. Potential effects of environmental chemical contamination in congenital heart disease. Pediatr Cardiol 2014;35:559–68.
[34] Cresci M, Foffa I, Ait-Ali L, et al. Maternal environmental exposure, infant GSTP1 polymorphism, and risk of isolated congenital heart disease. Pediatr Cardiol 2013;34:281–5.
[35] Silva KP, Rocha LA, Leslie AT, et al. Newborns with congenital heart diseases: epidemiological data from a single reference center in Brazil. J Prenat Med 2014;8:11–6.
[36] Ul Haq F, Jalil F, Hashmi S, et al. Risk factors predisposing to congenital heart defects. Ann Pediatr Cardiol 2011;4:117–21.
[37] Karatza AA, Giannakopoulos I, Dassios TG, et al. Periconceptional tobacco smoking and isolated congenital heart defects in the neonatal period. Int J Cardiol 2011;148:295–9.
[38] Kopec JA, Esdaile JM. Bias in case-control studies. A review. J Epidemiol Community Health 1990;44:179–86.
[39] Gotto GT, Yu C, Bernstein M, et al. Development of a nomogram model predicting current bone scan positivity in patients treated with androgen-deprivation therapy for prostate cancer. Front Oncol 2014;4:296.