Contents available at Sciverse ScienceDirect

Diabetes Research and Clinical Practice

journal homepage: www.elsevier.com/locate/diabres

Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach

Chongjian Wang [a,b], *Linlin Li* [a], *Ling Wang* [a], *Zhiguang Ping* [a], *Muanda Tsobo Flory* [b], *Gaoshuai Wang* [a], *Yuanlin Xi* [a], *Wenjie Li* [a,*]

[a] *Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, Zhengzhou, Henan, PR China*
[b] *Research Centre, CHU Sainte-Justine, Montreal, QC, Canada*

ARTICLE INFO

Article history:
Received 6 July 2012
Received in revised form
15 October 2012
Accepted 23 January 2013
Published on line 28 February 2013

Keywords:
Type 2 diabetes mellitus
Artificial neural network
Multivariate logistic regression
Classification model

ABSTRACT

Aim: To develop and evaluate an effective classification approach without biochemical parameters to identify those at high risk of T2DM in rural adults.
Methods: A cross-sectional survey was conducted. Of 8640 subjects who met inclusion criteria, 75% ($N_1 = 6480$) were randomly selected to provide training set for constructing artificial neural network (ANN) and multivariate logistic regression (MLR) models. The remaining 25% ($N_2 = 2160$) were assigned to validation set for performance comparisons of the ANN and MLR models. Predictive performance of different models was analyzed by the receiver operating characteristic (ROC) curve using the validation set.
Results: The prevalence rates of T2DM were 8.66% ($n = 561$) and 9.21% ($n = 199$) in training and validation sets, respectively. For ANN model, the sensitivity, specificity, positive and negative predictive value for identifying T2DM were 86.93%, 79.14%, 31.86%, and 98.18%, respectively, while MLR model were only 60.80%, 75.48%, 21.78%, and 94.52%, respectively. Area under the ROC curve (AUC) value for identifying T2DM when using the ANN model was 0.891, showing more accurate predictive performance than the MLR model (AUC = 0.744) ($P = 0.0001$).
Conclusion: The ANN model is an effective classification approach for identifying those at high risk of T2DM based on demographic, lifestyle and anthropometric data.

What is already known on this topic?

Artificial neural network (ANN) has already been used to identify and diagnose cancer, hypertension and other diseases. However, relatively few studies have developed and evaluated an effective classification approach without biochemical parameters to rapidly identify those at high risk of type 2 diabetes mellitus (T2DM) using the ANN model in adult population, particularly in rural residents.

What does this study add to the current literature?

We developed an accurately ANN classification model that can be used to quickly identify those at high risk of T2DM, and screen undiagnosed T2DM patient in rural adults using demographic, lifestyle, and anthropometric data.

How this might change public health policy and clinical practice.

The findings of this study reinforce that the ANN classification model without biochemical parameters may be helpful to rural healthcare practitioners in evaluating the risks of their patients quickly and noninvasively.

1. Introduction

Type 2 diabetes mellitus (T2DM) is a major public health problem causing significant morbidity and mortality both in developed and developing countries [1,2]. Therefore,

* *Corresponding author at*: Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, 100 Kexue Avenue, Zhengzhou, 450001, Henan, PR China. Tel.: +86 371 67781305; fax: +86 371 67781868.
E-mail address: liwj2008@126.com (W. Li).
0168-8227/$ – see front matter
http://dx.doi.org/10.1016/j.diabres.2013.01.023

identifying and diagnosing T2DM patients are very important for both public health policy and clinical practice [3–5]. Blood glucose measurement and continuous monitoring is a standard method for identifying and diagnosing T2DM, but it is not routinely used in resource limited settings, especially in some rural areas of developing countries. A large proportion of people with T2DM remain undetected due to the luck of affordability and availability [6]. Thus, an effective classification approach has been sought to rapidly identify those at high risk of T2DM in resource limited countries and areas.

Multivariate logistic regression (MLR) is often used to identify significant factors that correlated with diseases, and has commonly been used to develop a predictive expression for public health and clinical practice [7,8]. Intelligence algorithms such as artificial neural network (ANN) are machine learning tools and powerful classifiers based on observed behaviors of biological neurons [9,10]. Despite some approaches based on artificial network and machine learning algorithms have been developed and tested against cancer, hypertension and other diseases [11–15], relatively few studies have developed and evaluated an effective classification approaches without biochemical parameters to rapidly identify those at high risk of T2DM using the ANN model in adult population. In addition, data from rural areas are limited and mainly from urban areas and developed countries [16,17]. Therefore, the purpose of the present study was to develop and evaluate an effective classification approach based on demographic, lifestyle and anthropometric parameters to help practitioners to rapidly identify those at high risk of T2DM in rural adults.

2. Material and methods

2.1. Participants

The study was a population-based cross-sectional survey. All participants from rural district in Henan Province were selected randomly from eligible candidates listed in the residential registration record. To the eligible, candidates had to be stable residents for at least 10 years in the areas aged 35–74 years, and free from the following conditions: (1) severe psychological disorders, physical disabilities, cancer, chronic kidney disease, Alzheimer's disease, or dementia, within 6 months; or (2) currently diagnosed with tuberculosis, acquired immune deficiency syndrome (AIDS), and other infectious diseases. After severe psychological disorders ($n = 2$), cancer ($n = 21$), chronic kidney disease ($n = 76$), physical disabilities ($n = 5$), tuberculosis ($n = 5$), and other infectious diseases ($n = 4$) were excluded, 8688 subjects who met the criteria were enrolled in the study. Of the eligible participants, 48 (0.55%) subjects were excluded because of missing information on plasma glucose ($n = 15$) and other covariates ($n = 33$). Ultimately, a total of 8640 subjects were selected for the present analysis. The procedure of the study was approved by the Zhengzhou University Medical Ethics Committee, and written informed consent was obtained from all participants.

2.2. Data collection

Data were collected by specially trained physicians and public health workers using standardized methods with stringent levels of quality control. Data on demographic characteristics (age, sex, occupation, educational level, and marital status), family history (hypertension, diabetes, heart disease, and stroke), dietary and lifestyle (smoking, drinking, fat, vegetables and fruit intake and physical activity) were obtained by using a standard questionnaire.

Age was classified into four categories: 35–44, 45–54, 55–64, and 65–74 years. Education level was classified as: illiterate, elementary level, secondary school, high school and college. The latter two categories were set as high education level. Marital status was categorized as: married/cohabitation, and unmarried/divorced/widowed. Family history was considered positive if the participants' parents or siblings had a history of T2DM at/or before the baseline examination.

Three-day dietary intake data were collected from each subject using a 24-h diet recall and a 2-day diet record. The daily intake of energy, nutrients, food and food sets for each subject was calculated using the China Food Composition Table [18]. Based on the Chinese Dietary Guidelines [19], recommended vegetable and fruit intake was defined as consuming an average of more than 500 g per day, and high fat intake was defined as consuming an average of more than 25 g per day. Smoking status included current smoker and not current smoker. Participants who currently smoked and/or had smoked at least 100 cigarettes during their lifetime were classified as current smokers if they answered affirmatively to the questions, ''Do you smoke cigarettes now?'' and ''Have you smoked at least 100 cigarettes during your lifetime?'' Self-reported alcohol consumption data were also collected from the questionnaire using the following question: ''Considering all types of alcoholic beverages, how many times during the past 30 days did you drink?'' Physical activity level for each individual was classified as low, moderate, or high based on the International Physical Activity Questionnaire (IPAQ) [20].

Body weight and height were measured twice in light indoor clothing without shoes to the nearest 0.1 kg and 0.1 cm, respectively. Waist circumference (WC) was measured twice at the mid-point between the lowest rib and the iliac crest to the nearest 0.1 cm, after inhalation and exhalation. Central obesity based on WC (male: WC $\geq$ 90 cm; cm; female: WC $\geq$ 80 cm) was defined according to WHO criteria for the Asia-Pacific population [21]. Blood pressure and pulse rate were measured in the sitting position triple by a standardized protocol adapted from procedures recommended by the American Heart Association [22]. Participants were advised to avoid alcohol, cigarettes, coffee, tea, and excessive exercise for at least 30 min prior to having their pulse rate and blood pressure measured. Pulse rate was classified into five categories: under 60, 60–69, 70–79, 80–89, and over 90 bpm. Pulse pressure is the difference between systolic and diastolic blood pressure. The normal range of pulse pressure is from 20 to 60 mmHg, and the abnormal ranges is defined as a pulse pressure higher than 60 mmHg or lower than 20 mmHg according to the Chinese Hypertension Prevention Guide [23].