PDF version

Review

Kamal Gholipour,[1,2] Mohammad Asghari-Jafarabadi,[3,4] Shabnam Iezadi,[5] Ali Jannati[1,2] and Sina Keshavarz
[6]

[1]Iranian Center of Excellence in Health Management, School of Management and Medical Informatics, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran. [2]Tabriz Health Services Management Research Center, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran.
[3]

Road Traffic Injury Research Center, Health Management and Safety Promotion Research Institute, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran.
[4]

Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran.
[5]

Social Determinants of Health Research Center, Health Management and Safety Promotion Research Institute, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran (Correspondence to: S. Iezadi: sh_iezadi@yahoo.com).
[6]

Public Health and Preventive Medicine, University of Social Welfare and Rehabilitation Sciences, Tehran, Islamic Republic of Iran.

Abstract

Background: Type 2 diabetes mellitus (T2DM) is a metabolic disease with complex causes, manifestations, complications and management. Understanding the wide range of risk factors for T2DM can facilitate diagnosis, proper classification and cost-effective management of the disease.

Aims: To compare the power of an artificial neural network (ANN) and logistic regression in identifying T2DM risk factors.

Methods: This descriptive and analytical study was conducted in 2013. The study samples were all residents aged 15–64 years of rural and urban areas in East Azerbaijan, Islamic Republic of Iran, who consented to participate (n = 990). The latest data available were collected from the Noncommunicable Disease Surveillance System of East Azerbaijan Province (2007). Data were analysed using SPSS version 19.

Results: Based on multiple logistic regression, age, family history of T2DM and residence were the most important risk factors for T2DM. Based on ANN, age, body mass index and current smoking were most important. To test for generalization, ANN and logistic regression were evaluated using the area under the receiver operating characteristic curve (AUC). The AUC was 0.726 (SE = 0.025) and 0.717 (SE = 0.026) for logistic regression and ANN, respectively (P

Conclusions: The logistic regression model is better than ANN and it is clinically more comprehensible.

Keywords: artificial neural network, diabetes mellitus, multiple regression, risk factors.

Citation: Gholipour K; Asghari-Jafarabadi M; Iezadi S; Jannati A; Keshavarz S. Modelling the prevalence of diabetes mellitus risk factors based on artificial neural network and multiple regression. East Mediterr Health J. 2018;24(8):770–777. https://doi.org/10.26719/emhj.18.012

Received: 08/09/15; accepted: 12/06/17

Copyright © World Health Organization (WHO) 2018. Some rights reserved. This work is available under the CC BY-NC-SA 3.0 IGO license (https://creativecommons.org/licenses/by-nc-sa/3.0/igo).

Introduction

Type 2 diabetes mellitus (T2DM) is a complex metabolic disease with complex causes,

manifestations, complications and management (1,2). The chronic complications of T2DM include accelerated development of cardiovascular disease, end-stage renal failure, blindness and lower limb amputations, which can result in excess morbidity and mortality (3). These chronic complications not only have a major impact on patients and their families but also consume an increasing share of health system resources (4). Three hundred and forty-seven million people worldwide suffer from this serious and costly disease. Diabetes now affects both high- and low-income countries but > 80% of people with diabetes live in low- and middle-income countries (5). However, based on World Health Organization (WHO) reports, diabetes mortality will have doubled between 2005 and 2030 and the prevalence of T2DM is increasing worldwide (3,5,6).

Understanding the wide range of risk factors for T2DM can facilitate diagnosis, proper classification and cost-effective management of the disease (6,7). Recent intervention studies have indicated that T2DM can be prevented or delayed by lifestyle changes in high-risk individuals (2,6). Therefore, identifying such risk factors using the right models is the first stage in successful intervention.

Logistic regression is an efficient and powerful tool to assess independent variable contributions to a binary outcome and it is used to analyse the relationship between 1 or more predictors and a dichotomous outcome (8–10). Simultaneous analysis of multiple explanatory variables and reducing the effect of confounding factors are some important advantages of logistic regression (9). However, its accuracy strongly depends on careful variable selection with satisfaction of basic assumptions, as well as appropriate choice of model-building strategy and validation of results (10). Important considerations when conducting logistic regression include adopting independent variables, ensuring that relevant assumptions are met, and selection of the right modelling strategy (10). Logistic regression can be used to study the factors that predict improvement after an intervention (8).

An artificial neural network (ANN) is a nonlinear, computational and complex mathematical model that is constructed to simulate processes of the central nervous system of higher animals, distantly based on the human neuronal structure (11–15). ANNs represents a new method for predictive modelling in medical sciences and they are useful to predict complex, nonlinear and time-dependent relationships. ANNs also can be used when the measures influencing an event are not completely known (14,15). In contrast with traditional statistical techniques, ANNs are capable of automatically resolving these relationships without the need for a priori assumptions about the nature of the interactions between variables. ANNs use data to model and find relationships between factors (11). Another important difference in comparison with traditional statistical methods such as logistic regression is the learning ability of an ANN. A trained network has pooled regulations that are represented by the matrix of the weights between the neurons. This characteristic allows the ANN to forecast cases that have

never been presented to the network before and it is called generalization (16).

When predicting and prioritizing risk factors of T2DM, it is questionable which one of these models is better. To respond to this question, we compared the power of these 2 models in terms of sensitivity, specificity and accuracy. For this purpose, we used receiver operating characteristic (ROC) analysis, which included sensitivity, specificity and accuracy of models to indicate the predictive power of models. The ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied (17). In this study, we investigated the power of logistic regression and ANNs to identify T2DM risk factors and compared them to establish which one was better.

Methods

This descriptive and analytical study was conducted in 2013 to determine the risk factors for T2DM using 2 separate statistical methods. The study sample comprised all residents aged 15–64 years of rural and urban areas of East Azerbaijan Province, Islamic Republic of Iran who were willing to participate. We used a clustered randomized sampling method.

Neighbourhoods and parishes were considered as clusters. In urban settings, a cluster contained 1 or more or parts of a neighbourhood. In rural settings, a cluster contained 1 or more or parts of a village. Cluster heads were selected based on the last digit of the postal code. Each cluster had 20 individuals; 10 males and 10 females living in neighbouring households. From each cluster, we selected 2 men and 2 women from each age group (15–24, 25–34, 35–44, 45–54 and 55–64 years). Every individual in each cluster was selected randomly based on the postal address. We included the nearest right side neighbours to the cluster heads, who were eligible based on age group. Participants gave full informed consent after the study objectives and process were explained. Ethical approval was obtained from the Center for Disease Control of Iran.

T2DM was defined as having a diagnosis or receiving a prescription for antidiabetic drugs. New T2DM was considered if fasting plasma glucose (FPG) level was ≥ 126 mg/dl. Impaired fasting glucose was defined by FPG ≥ 100 mg/dl (5.6 mmol/l) but

At the first stage, during a home visit, health centre staff collected information about sociodemographic, lifestyle and health status through interview. A structured questionnaire (18) was used to explore demographic and ecological characteristics of participants, nutritional status, diabetes risk factors such as high blood pressure and family history of T2DM, and patients' physical activities, based on WHO guidelines.

Anthropometric measurements were conducted by proficient and skilled healthcare staff of Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran. Body height and weight were measured using a portable electronic weighing scale and portable height-measuring instrument. Participants were asked to remove their shoes and any bulky clothing. Waist circumference was measured at the midpoint between the lower part of the lowest rib, and blood pressure was measured with a calibrated sphygmomanometer. The average of 3 measurements, with a mean time of 5 minutes, was used for analysis. Finally, blood samples (10 ml from every participant) were collected in 4 tubes and centrifuged immediately for measurement of FPG (≥ 126 mg/dl), total cholesterol (TC), high-density lipoprotein cholesterol and triglycerides. A cold chain was preserved while transferring blood samples to the Central Reference Laboratory in Tabriz.

ANN modelling and logistic regression were used to analyse the data. Variables associated with T2DM in the univariate analysis were included in multiple logistic regression models. The P values for entry and removal of variables in the logistic regression model were 0.05 and 0.1, respectively. The significant variables in univariate analyses along with confirmatory factors were used to calculate the individual T2DM risk with the ANN. Uncontrolled hypertension, gender and raised TC > 200 mg/dl were considered as confirmatory factors. The variable importance in the logistic regression analysis was calculated based on standardized coefficient (Wald). The data were divided into a training set (67.1%) and test set (32.9%). Automatic architecture selection was used to determine hidden layers. One hidden layer with 7 units was determined. A scaled conjugate gradient option was used to optimize the algorithm. Modelling was continued until the relative error of testing was less than that of training. Description and diagram network structures were used as the network structure.

We compared the importance of T2DM predictors revealed by logistic regression and ANN. For the ANN and logistic regression models, the area under the ROC curve (AUC) was calculated in the test set. ROC curve is a technique for visualizing, organizing and selecting possibly optimal models based on their performance. This technique illustrates the performance of a binary classifier by considering sensitivity, specificity and accuracy of models (17). Data were analysed using SPSS version 19 (IBM Corp., Armonk, NY, USA). P

Results

We used data from 990 participants to identify the T2DM risk factors using 2 separate methods of logistic regression and ANN. Selected risk factors were prioritized and the 2 methods were compared to determine how they differed. Table 1 shows the importance of T2DM predictors according to their priority using the ANN method. Age had the highest score of 0.34, which means that age can predict 34% of T2DM. Raised TC had the lowest score of 0.02. Figure 1 shows the sequence of predictors based on their importance.

In multiple logistic regression, after adjusting for other factors, there was a significant association between T2DM and age [odds ratio (OR): 1.05, 95% confidence interval (CI): 1.03–1.08; P Table 2). Considering the results of univariate and multiple logistic regression, age, positive family history of T2DM and residence, by adjusting for body mass index (BMI), gender, uncontrolled hypertension, raised TC, controlled hypertension and current smoking were significant and independent risk factors of T2DM (Tables 2 and 3).

We compared the importance of predictors of T2DM based on ANN and logistic regression modelling. For ANN, age, BMI and current smoking were the 3 most important predictors of T2DM, followed by residence, controlled hypertension, uncontrolled hypertension, family history of T2DM, gender and raised TC (> 200 mg/dl). The estimated errors of testing and training were 11% and 15%, respectively, so the goodness of the model was confirmed. For logistic regression modelling, age, family history of T2DM and residence were the most important predictors of T2DM, followed by current smoking, controlled hypertension, uncontrolled hypertension, BMI, raised TC (> 200 mg/dl) and gender. The latter 5 factors were not significant.

To test the generalization of the results, we evaluated ANN and logistic regression in the test set using AUC values (Table 4). The AUC values were 0.726 (standard error 0.025) and 0.717 (standard error 0.026) for logistic regression and the ANN, respectively. So, the ability of the logistic regression model to predict those with and without T2DM was significantly greater than that of the ANN model (P

Discussion

Comparison of the power of an ANN and logistic regression indicated that the latter is a statistically better predictor. In both methods, age was predicted as the most important risk factor in East Azerbaijan Province. Therefore, we suggest paying attention to aged people in the diagnosis and management of T2DM. According to the results of both models together, people who smoke or live in rural areas and those with a family history of T2DM are more at risk of developing T2DM. Also, the risk may increase with BMI.

Logistic regression is easier than ANN to apply and understand. In contrast, ANN can be applied without assumptions used in logistic regression (such as residual normality, homogeneity of residual variances, residual independence and collinearity). Several studies have shown that ANN models have several advantages over conventional statistical methods (19,20). Such models can rapidly recognize linear patterns, categorical and stepwise linear patterns, nonlinear patterns with threshold impacts, and contingency effects. ANN analyses do not need to be started with a hypothesis or preselected key variables. Therefore, undocumented or quantified potential predictors may be specified if they already exist in the various datasets,

although they may have been neglected in the past (19,20). Logistic regression as a recognized approach is able to predict clinically relevant dichotomous outcomes. It has some advantages over more traditional approaches to analyse such data (e.g., t test and regression), and it is better explored in this context than newer data analysis procedures (e.g., neural nets) (8).

It should be noted that due to the simplicity of interpretation of the variables in the logistic regression model, applying it clinically is more comprehensible. Rahman et al. compared the accuracy of ANN and binary logistic regression models for predicting glucose status (21). They showed a significantly better performance of ANN for detection of impaired glucose tolerance and T2DM patients from disease-free ones (21). Omurlu et al. compared performance of logistic regression and ANN for prediction of albuminuria in T2DM and demonstrated that multilayer perceptron had the highest predictive capability for the presence of albuminuria (22). Zandkarim et al. suggested that logistic regression was more powerful than discriminant analysis for distinguishing T2DM and prediabetes (23). In communities where there is high dependency among case and control groups, recognizing the differences needs stronger methods. Kazemnejad et al. demonstrated that there was no performance difference between models based on logistic regression and ANN in differentiating impaired glucose tolerance and diabetes patients from disease-free patients (24). Zandkarim and Safavi recommended ANNs for medical research in comparison to logistic regression (25).

The present study suggests that statistical analysis of the importance of T2DM risk factors differs using 2 separate models; however, age was the most important predictor in both models. Raised TC and sex had less importance in comparison with other risk factors. Rezaei et al. showed that age, FPG, BMI and mobility variables in their logistic regression model were significant, and FPG, glucose tolerance, BMI and mobility variables indicated the highest predictive power in the neural network model (26). A national survey in 2009 in the Islamic Republic of Iran showed that sex, age and residence were significant predictors of diabetes (27). Logistic regression analysis in a survey in Qatar showed that smoking and family history of DM had a significant association with DM (28).

Our study had 2 major limitations. The first was the small sample size. The second was that a glucose tolerance test was not done and a single FPG test was used. Nevertheless, our study had some strengths, such as inclusion of broad age groups, and it is believed to be the first study to compare 2 models in identifying and prioritizing T2DM risk factors in the Islamic Republic of Iran.

Conclusion

Comparison of the power of ANN and logistic regression models indicated that the latter is

better than ANN and is clinically more comprehensible. Logistic regression can provide coefficients such as probability ratio to express the impact of each independent variable on the model and it is better to be used in medicine. However, we should bear in mind that ANNs can easily be used and analysed. It is possible to enter a large number of variables into an ANN and there is no need for assumptions such as normality. Thus, if there is no assumption, we recommend using an ANN model. Our results also showed that age, BMI, family history of T2DM, current smoking and residence are the most important predictors of T2DM in East Azerbaijan Province. A comprehensive programme of diagnosis and management of T2DM, as well as providing consultation for high-risk individuals, which is based on prioritizing people, can be an appropriate initiative to decrease the prevalence of T2DM in East Azerbaijan.

Acknowledgements

We would like to acknowledge the co-operation of health centre staff, in East Azerbaijan Province.

Funding: Student Research Committee, Tabriz University of Medical Sciences, Tabriz, Islamic Republic of Iran.

Competing interests: None declared.

Modélisation de la prévalence des facteurs de risque du diabète sucré sur la base d’un réseau de neurones artificiels et d’une régression multiple

Résumé

Contexte : Le diabète sucré de type 2 est une maladie métabolique dont les causes, les manifestations, les complications et la prise en charge sont complexes. La compréhension de la vaste palette de facteurs de risque de ce type de diabète peut faciliter le diagnostic, l’établissement de la classification et améliorer la prise en charge de la maladie avec un bon rapport coût-efficacité.

Objectif : Comparer la performance d’un réseau de neurones artificiels (RNA) et de la régression logistique dans l’identification des facteurs de risque du diabète sucré de type 2.

Méthodes : La présente étude descriptive et analytique a été menée en 2013. Tous les sujets de l'étude résidaient dans des zones urbaines ou rurales de la partie orientale de l'Azerbaïdjan et de la République islamique d'Iran ; ils avaient entre 15 et 64 ans et tous avaient consenti à participer à l'étude (n = 990). Les données les plus récentes ont été recueillies par l'intermédiaire du Système de surveillance des maladies non transmissibles de la Province orientale de l'Azerbaïdjan (2007). Elles ont été analysées à l'aide du logiciel SPSS (version 19).

Résultats : Pour la régression logistique, l'âge, les antécédents familiaux de diabète sucré de type 2 et le lieu de résidence se sont avérés être les facteurs de risque les plus importants. En ce qui concerne le RNA, l'âge, l'indice de masse corporelle et le tabagisme étaient les facteurs de risque de diabète sucré de type 2 les plus importants. Afin d'établir un test servant de base à une généralisation, le RNA et la régression logistique ont été évalués en utilisant la zone située sous la courbe ROC (caractéristique du fonctionnement du récepteur). La ROC était à 0,726 (erreur-type= 0,025) et 0,717 (erreur-type= 0,026) pour la régression logistique et le RNA respectivement (p

Conclusion : Le modèle de régression logistique est meilleur que celui du RNA et cliniquement plus compréhensible.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

WHO EMRO | Modelling the prevalence of diabetes mellitus risk factors based on artificial neural network

جميع مَنْ السكان الذين تتراوح أعمارهم بين 15 و64 سنة في المناطق الريفية والحضرية في
أذربيجان الشرقية، جمهورية إيران الإسلامية، والذين أخذنا موافقتهم على المشاركة في
الدراسة (وعددهم 990). وقد جمعنا أحدث البيانات المتاحة من نظام ترصُّد الأمراض غير
السارية في مقاطعة أذربيجان الشرقية (2007). كما قمنا بتحليل البيانات باستخدام نظام
SPSS ، الإصدار 19.

: استناداً إلى التَّحَوُّف اللوجستي المتعدد، اتضح أن عوامل الخطر الأكثرالنتائ ج
أهمية في الإصابة بالسكري من النمط 2 هي العمر والتاريخ العائلي للإصابة بالسكري من
النمط 2 ومكان السكن . بينما اتضح استنادًا إلى شبكة عصبية اصطناعية أن العمر ومنسب
كتلة الجسم والتدخين الحالي هي عوامل الخطر الأكثر أهمية . ومن أجل اختبار التعميم،
قَيَّمنا الشبكة العصبية الاصطناعية والتَّحَوُّف اللوجستي باستخدام المنطقة
الواقعة تحت المنحنى المميز للمُتَقَبِّل العامل فوجدنا أنه كان 0.726 (SE = 0.025) ، في
التَّحَوُّف اللوجستي SE = 0.025)) وأنه كان SE = 0.026)) 0.717 في الشبكة العصبية
الاصطناعية

SE = 0.026)) ، وكان (p

10 / 10