Kaohsiung Journal of Medical Sciences (2013) 29, 93–99

Available online at www.sciencedirect.com
SciVerse ScienceDirect
journal homepage: http://www.kjms-online.com

ORIGINAL ARTICLE

Comparison of three data mining models for predicting diabetes or prediabetes by risk factors

Xue-Hui Meng [a], Yi-Xiang Huang [a], Dong-Ping Rao [b], Qiu Zhang [a], Qing Liu [b,*]

[a] Department of Health Service Management, Public Health School of Sun Yat-Sen University, People's Republic of China
[b] Department of Epidemiology, Cancer Center of Sun Yat-Sen University, People's Republic of China

Received 19 December 2011; accepted 12 March 2012
Available online 16 October 2012

KEYWORDS
Community;
Data mining;
Diabetes;
Prediction;
Risk factors

Abstract The purpose of this study was to compare the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or prediabetes using common risk factors. Participants came from two communities in Guangzhou, China; 735 patients confirmed to have diabetes or prediabetes and 752 normal controls were recruited. A standard questionnaire was administered to obtain information on demographic characteristics, family diabetes history, anthropometric measurements and lifestyle risk factors. Then we developed three predictive models using 12 input variables and one output variable from the questionnaire information; we evaluated the three models in terms of their accuracy, sensitivity and specificity. The logistic regression model achieved a classification accuracy of 76.13% with a sensitivity of 79.59% and a specificity of 72.74%. The ANN model reached a classification accuracy of 73.23% with a sensitivity of 82.18% and a specificity of 64.49%; and the decision tree (C5.0) achieved a classification accuracy of 77.87% with a sensitivity of 80.68% and specificity of 75.13%. The decision tree model (C5.0) had the best classification accuracy, followed by the logistic regression model, and the ANN gave the lowest accuracy.

Introduction

Diabetes is a major cause of concern and its prevalence is increasing in China. A national survey conducted in 1994 on individuals aged 25 to 64 years showed that the prevalences of diabetes and impaired glucose tolerance were 2.5% and 3.2%, respectively [1]. In a cross-sectional study in 2000–2001, on individuals aged 35 to 74 years, the

* Corresponding author. Department of Epidemiology, Cancer Center of Sun Yat-Sen University, 651 Dongfeng East Road, Guangzhou 510060, Guangdong Province, People's Republic of China.
E-mail address: liuqingsysu@yahoo.cn (Q. Liu).

1607-551X/$36 Copyright © 2012, Kaohsiung Medical University. Published by Elsevier Taiwan LLC. All rights reserved.
http://dx.doi.org/10.1016/j.kjms.2012.08.016

prevalences were 5.5% and 7.3%, respectively [2]. The China National Diabetes and Metabolic Disorders Study conducted in 2007—2008 reported that the age-standardized prevalences of total diabetes and prediabetes (impaired fasting glucose or impaired glucose tolerance) were 9.7% and 15.5%, respectively [3]. Diabetes results from the interaction between a genetic predisposition and behavioral and environmental risk factors [4]. Although the genetic basis of type 2 diabetes has yet to be identified, there is strong evidence that such modifiable risk factors as obesity and physical inactivity are the main nongenetic determinants of the disease [5]. Lifestyle factors linked to the incidence of diabetes or diabetes-related risk factors include physical activity level, dietary habits, adiposity, alcohol consumption, smoking [6—15], and duration of sleep [16,17]. The World Health Organization (WHO) recommends the development of simple strategies to identify those at risk of diabetes and provide them with early lifestyle interventions [18]. It is very important to establish predictive models using those risk factors for interventions relating to the development of diabetes. Previous studies have suggested that anthropometric measurement and adipocyte size can serve as predictors of diabetes incidence using traditional statistical methods [19—22].

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships that provide a clear and useful result [23,24], and this technique has developed rapidly in recent years. Studies have applied data mining to explore unknown factors and predictive models have been built in the medical field [25—27]. However, relatively little research has considered the use of data mining methods to construct corresponding prediction models for the incidence of diabetes using several common risk factors.

The purpose of this study was to compare multiple prediction models for diabetes incidence based on common risk factors. This study developed three widely used data mining classification models, logistic regression, artificial neural networks (ANNs) and decision tree, along with a 10-fold cross-validation technique. Accuracy, sensitivity and specificity were used to evaluate them.

Materials and methods

Participants

A total of 1487 individuals 20 years of age or older participated in this study. This included 735 volunteers who were confirmed to have diabetes or prediabetes (impaired fasting glucose or impaired glucose tolerance) by a physician according to the 1999 WHO criteria in the past two years and 752 volunteers who were not diabetes or prediabetes patients, and were confirmed as such by physical check in the past two years. These individuals were recruited from the Zhuguang and Liurong communities in Guangzhou, China between July 2007 and December 2008.

Questionnaire

A standard questionnaire, including items of common risk factors of diabetes, was administered to all participants to obtain information on demographic characteristics, family diabetes history, anthropometric measurements and lifestyle risk factors.

Data collection

Demographic characteristics included age, gender, marital status, and level of education. A family history of diabetes was defined as any family member previously having been diagnosed as having diabetes or prediabetes by a physician. Anthropometric measurements were taken on participants who were standing up and wearing light clothes and no shoes. Weight was measured to the nearest 0.1 kg, using a spring balance, and height was measured twice with a standard stadiometer to the nearest 0.5 cm. Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m^2), and a BMI $\geq$25 was defined as overweight. In this study the lifestyle risk factors included the following variables: cigarette smoking (having smoked at least 500 cigarettes in one's life); alcohol drinking (the consumption of at least 100 g of alcohol per week for 1 year or longer); tea and coffee drinking (more than once a week); consumption of beef, pork, mutton, fish, vegetables and fruits (more than three times a week); preference for sweet and salty food in daily life; work stress high, moderate and low, according to the participants' subjective impression; physical activity (participation in moderate or vigorous activity for 30 minutes or more per day for at least 3 days a week); and finally sleep duration (short $\leq$5 hours, normal 5—8 hours, and long $\geq$8 hours).

All staff members successfully completed a training program that familiarized them with both the aims of the study and the specific tools and methods used. At the training sessions, interviewers were given detailed instructions concerning the administration of the study questionnaire.

The study was approved by the Sun Yat-sen University ethics committee and the two communities' ethics committees. Written informed consent was obtained from each participant before data collection.

Statistical analyses

Statistical analyses were performed using SPSS statistical program, version 13.0 for Windows; and data mining prediction models were constructed using SPSS Modeler, version 14.1 (SPSS Inc., Chicago, IL, USA). Descriptive statistical analyses were carried out for all variables, using the Chi-square test to examine differences between proportions with significance value of 0.05. The primary analyses were stratified by demographic characteristics and lifestyle risk factors. Binary logistic regression, back-propagation ANNs and decision tree (5.0) models were then constructed using the training dataset and tested by the testing dataset. The original dataset was randomly divided into two parts, with the training dataset containing about 70% training of the participants (1031 cases), and the testing dataset containing 30% of the participants (456 cases) by the partition node of SPSS Modeler software. The 10-fold cross-validation methods were used to measure the unbiased estimate of the three prediction models for the purposes of comparing their performances.

The dependent variable (output variable) was a binary categorical variable with two categories: 0 and 1, where 0 means normal and 1 means diabetes or prediabetes. The independent variables (input variables) were the 12 risk factors that were statistically significant on the Chi-square test. These were gender, age, marital status, educational level, family history of diabetes, BMI, coffee drinking, physical activity, sleep duration, work stress, consumption of fish, and preference for salty foods. All the results of the descriptive and Chi-square tests are shown in Table 1.

Logistic regression is a nonlinear regression method for predicting a categorical dependent variable. Logistic regression was performed to identify risk factors for many diseases using patient characteristics, history, and risk factors. The logistic model formula computes the probability of the selected disease y ($y = 0$ if the subject does not suffer from the disease; otherwise, $y = 1$) as a function of the values of the predictive risk factors. If the individual suffers from this disease, the conditional probability is given by $p(y = 1 \mid X) = p(X)$, and the logistic model formula takes the form: $\log[p(x)/1-p(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ where $X = (x_1, x_2, \ldots, x_k)$ represents the vector of k's risk factors by the logistic regression approach [26]. In order to ensure the selected input variables are same in the three data mining models, this study used an 'entry' method for developing the logistic regression model to ensure that the 12 variables in the model were statistically significant by Chi-square test. The logistic regression model was built based on the training dataset and it was tested using the testing dataset. The training and testing data were saved for further processing by decision tree and neural networks.

ANNs are a popular data mining tool used to construct complicated and nonlinear models [23]. Basically, an ANN model contains three layers: the input layer, the output layer, and the hidden layer, each layer being made up of nodes (neurons) and links. The nodes in input layers are viewed as predicted variables, whereas the nodes in output layers are analyzed as the outcome variables [28]. This study used a popular ANN architecture called a multi-layer perception network MLPN with back-propagation (a supervised learning algorithm), which is arguably the most commonly used and well-studied ANN architecture. MLPNs are feed-forward neural networks trained with the standard back-propagation algorithm and they are known to be a powerful function approximator for prediction and classification problems [24]. The architecture of the MLPN consisted of three layers, an input layer, a hidden layer and an output layer. The input layer contained 20 input neurons with the 12 variables; the hidden layer consisted of 15 hidden nodes; and the output layer consisted of one output neuron. The initial learning rate and momentum for network training were set to 0.3 and 0.9, respectively.

A decision tree is a form for expressing such mappings, and it consists of tests or attribute nodes linked to two or more subtrees and leaves or decision nodes labeled with a class that reflects the decision [29]. Popular decision-tree algorithms include Quinlan's ID3, C4.5, C5.0, and Breiman et al.'s classification and regression trees (CART) [24]. Based on the favorable prediction results we obtained from the preliminary runs, in this study we chose to use C5.0 algorithm as our decision tree method, which is an improved version of the C4.5 and ID3 algorithms [30]. In this study, the tree was built from the training data. It was then heuristically pruned to avoid overfitting the data, which tended to introduce a classification error on the testing data. C5.0 follows the post-pruning approach, which removes branches from a fully grown tree. For each non-leaf node in the tree, the pruning algorithm estimated the expected error rate that would occur if the subtree at that node were pruned. Then the expected error rate occurring if the node was not pruned was estimated using the error rates for each branch, combined by weighting according to the proportion of observations along each branch. If pruning the node led to a greater than expected error rate, then the subtree was kept. Otherwise, it was pruned [31].

This paper used both a confusion matrix to appraise the performance of the three models for incidence of diabetes and three evaluated indices for accuracy, sensitivity and specificity. The classification accuracy measures the proportion of cases correctly classified. Sensitivity measures the fraction of positive cases that are classified as positive. Specificity measures the fraction of negative cases that are classified as negative [32].

- Accuracy = (TP + TN)/(TP + FP + TN + FN)
- Sensitivity = TP/(TP + FN)
- Specificity = TN/(FP + TN)

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. The model with highest the sensitivity, specificity, and accuracy is the best predictive model.

Results

In the present study, 1487 individuals were recruited from two communities; 735 were diabetes or prediabetes patients and 752 were not. The characteristics of participants and Pearson Chi-square test results between two groups are shown in Table 1. Cigarettes smoking ($p = 0.687$), alcohol consumption ($p = 0.058$), tea drinking ($p = 0.591$), beef or pork or mutton consumption ($p = 0.519$), vegetable consumption ($p = 0.190$), fruit consumption ($p = 0.197$), and preference for sweet food ($p = 0.055$) had no statistical significance, and the other 12 factors showed statistically significant differences between the two groups, at a significance level of 0.05.

The sensitivity analysis performed for this research is presented in Table 2, where each variable is placed in order of its relative importance. The standardization regression coefficients denoted the independent variables importance in the logistic regression model. Sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and output of an ANN model [24]. In the C5.0 decision tree model, to determine the order in which variables must be chosen to split the data, this model used entropy-based information gain as a measure to allow comparison of the variables across a few scales meaning we could show the relative importance of each variable.

The three models were evaluated based on accuracy, sensitivity and specificity, and the detailed predictions produced from the training and testing datasets are presented in the form of confusion matrices. A confusion matrix is a matrix representation of classification results. Table 3

Table 1 Characteristics of the participants and statistical analyses in diabetes or prediabetes group and normal group.

Characteristics	Diabetes or prediabetes	Normal	Total	Pearson	*p*-value
	$n = 735$	$n = 752$	$n = 1487$	Chi-square test	
Gender					
Men	284 (38.6%)	363 (48.3%)	647 (43.5%)	14.03	<0.001
Women	451 (61.4%)	389 (51.7%)	840 (56.5%)		
Age					
20—39 years old	13 (1.8%)	143 (19.0%)	156 (10.5%)	341.33	<0.001
40—59 years old	175 (23.8%)	395 (52.5%)	570 (38.3%)		
60—79 years old	455 (61.9%)	189 (25.1%)	644 (43.3%)		
80 years and older	92 (12.5%)	25 (3.4%)	117 (7.9%)		
Marital status					
Unmarried	16 (2.2%)	58 (7.7%)	74 (5.0%)	100.42	<0.001
Married	506 (68.8%)	603 (80.2%)	1109 (74.6%)		
Widowed	185 (25.2%)	57 (7.6%)	242 (16.3%)		
Others	28 (3.8%)	34 (4.5%)	62 (4.1%)		
Education level					
Less than 9 years	520 (70.7%)	318 (42.3%)	838 (56.4%)	122.57	<0.001
Between 9 and12 years	154 (21.0%)	304 (40.4%)	458 (30.8%)		
More than 12 years	61 (8.3%)	130 (17.3%)	191 (12.8%)		
Family history of diabetes					
No	534 (72.7%)	725 (96.4%)	1259 (84.7%)	161.59	<0.001
Yes	201 (27.3%)	27 (3.6%)	228 (15.3%)		
Body mass index					
<25	608 (82.7%)	681 (90.6%)	1289 (86.7%)	19.78	<0.001
≥25	127 (17.3%)	71 (9.4%)	198 (13.3%)		
Smoking					
Less than 500 cigarettes in one's life	581 (79.0%)	588 (78.2%)	1169 (78.6%)	0.16	0.687
At least 500 cigarettes in one's life	154 (21.0%)	164 (21.8%)	318 (21.4%)		
Alcohol consumption					
Less than 100 g alcohol a week	676 (92.0%)	670 (89.1%)	1346 (90.5%)	3.59	0.058
At least 100 g alcohol a week	59 (8.0%)	82 (10.9%)	141 (9.5%)		
Drinking tea					
Less than once a week	257 (35.0%)	253 (33.6%)	510 (34.3%)	0.29	0.591
At least once a week	478 (65.0%)	499 (66.4%)	977 (65.7%)		
Drinking coffee					
Less than once a week	719 (97.8%)	721 (95.9%)	1440 (96.8%)	4.60	0.032
At least once a week	16 (2.2%)	31 (4.1%)	47 (3.2%)		
Physical activity					
Less than 30 minutes a day or 3 days a week	455 (61.9%)	565 (75.1%)	1020 (68.6%)	30.19	<0.001
30 minutes or more a day on at least 3 days a week	280 (38.1%)	187 (24.9%)	467 (31.4%)		
Duration of sleep					
Less than 5 hours a day	26 (3.5%)	6 (0.8%)	32 (2.2%)	26.94	<0.001
Between 5 and 8 hours a day	423 (57.6%)	373 (49.6%)	796 (53.5%)		
More than 8 hours a day	286 (38.9%)	373 (49.6%)	659 (44.3%)		
Work stress					
Low	675 (91.8%)	557 (74.1%)	1232 (82.9%)	83.04	<0.001
Moderate	52 (7.1%)	176 (23.4%)	228 (15.3%)		
High	8 (1.1%)	19 (2.5%)	27 (1.8%)		
Eating beef or pork or mutton					
Less than 3 times a week	323 (43.9%)	318 (42.3%)	641 (43.1%)	0.42	0.519
At least 3 times a week	412 (56.1%)	434 (57.7%)	846 (56.9%)		
Eating fish					
Less than 3 times a week	321 (43.7%)	373 (49.6%)	694 (46.7%)	5.25	0.022
At least 3 times a week	414 (56.3%)	379 (50.4%)	793 (53.3%)		
Eating vegetables					
Less than 3 times a week	268 (36.5%)	299 (39.8%)	567 (38.1%)	1.71	0.190
At least 3 times a week	467 (63.5%)	453 (60.2%)	920 (61.9%)		

Table 1 (*continued*)

Characteristics	Diabetes or prediabetes $n = 735$	Normal $n = 752$	Total $n = 1487$	Pearson Chi-square test	*p*-value
Eating fruits					
Less than 3 times a week	348 (47.3%)	331 (44.0%)	679 (45.7%)	1.66	0.197
At least 3 times a week	387 (52.7%)	421 (56.0%)	808 (54.3%)		
Preference for sweet food					
No	25 (3.4%)	41 (5.5%)	66 (4.4%)	3.69	0.055
Yes	710 (96.6%)	711 (94.5%)	1421 (95.6%)		
Preference for salty food					
No	68 (9.3%)	39 (5.2%)	107 (7.2%)	9.20	0.002
Yes	667 (90.7%)	713 (94.8%)	1380 (92.8%)		

The numbers were the observed count and the proportion (%) in the categorical variable of each group.

shows the results in a tabular format. The present study found that the training dataset logistic regression model achieved a classification accuracy of 75.95% with a sensitivity of 79.68% and a specificity of 72.40%; the ANN model achieved a classification accuracy of 73.52%, with a sensitivity of 83.47% and a specificity of 64.08%. However, the decision tree (C5.0) performed best among the three evaluated models. The decision tree had a classification accuracy of 78.27%, with a sensitivity of 81.87% and a specificity of 74.86%. In the testing dataset, the logistic regression model achieved a classification accuracy of 76.54% with a sensitivity of 79.40% and a specificity of 73.54%. The ANN model gave a classification accuracy of 72.59%, with a sensitivity of 79.40% and a specificity of 65.47%, and the decision tree achieved a classification accuracy of 76.97%, with a sensitivity of 78.11% and a specificity of 75.78%. In the whole dataset, the accuracy, sensitivity and specificity of logistic regression model were 76.13%, 79.59% and 72.74%, respectively; in the ANN model they were 73.23%, 82.18% and 64.49%, respectively; and in the decision tree model they were 77.87%, 80.68% and 75.13%, respectively, which was the best result among three models.

Discussion

Assessing and building predictive models for diabetes using common risk factors is important in the rapidly growing economy and changing lifestyles of China. As shown in Table 1, 12 variables were associated with incidence of diabetes at a significance level of 0.05. Older age, family history of diabetes, BMI, and preference for salty food were positively related, while education level and drinking coffee were negatively related to the presence of diabetes. Previous surveys of risk factors for diabetes reported similar results [33,34], and several large-scale trials have demonstrated the benefits of the prevention of diabetes with lifestyle interventions [5,35,36]. However, the data from this study suggested that physical activity and extent of stress related to work were also positively associated with diabetes. One of the reasons is the difference in age between the two groups; as shown in Table 1, the proportion of people aged 60 years and older in the diabetes and normal groups are 74.3% and 25.5%, respectively, and age is a risk factor for diabetes. The majority of older people in China are retired, so they have more time to participate in physical activity and they lack work stress. Smoking, drinking alcohol, and eating vegetables and fruit on a daily basis were not significantly associated with diabetes, which is not in agreement with the results of some previous studies [15,37–39].

A logistic regression model is most widely used when the prediction of disease or health status is of interest. In recent years, there has been growing interest in data mining techniques to construct and compare predictive models. Many previous studies have employed predictive models for

Table 2 The importance of the 12 input variables in three models.

Order[a]	Logistic regression	Artificial neural network (B-P)	Decision tree (C5.0)
1	Age	Age	Age
2	Family history of diabetes	Family history of diabetes	Education level
3	Marital status	Duration of sleep	Family history of diabetes
4	Education level	Preference for salty food	Marital status
5	Work stress	Marital status	Preference for salty food
6	Duration of sleep	Education level	Drinking coffee
7	Physical activity	Work stress	Duration of sleep
8	Preference for salty food	Physical activity	Body mass index
9	Gender	Drinking coffee	Work stress
10	Eating fish	Gender	Eating fish
11	Drinking coffee	Body mass index	Physical activity
12	Body mass index	Eating fish	Gender

[a] The order according to importance, from the most to the least important.

Table 3 The performance of three predictive models for diabetes incidence.

True classification in each dataset		Logistic regression		Artificial neural network		Decision tree C5.0	
		+[b]	−[c]	+[b]	−[c]	+[b]	−[c]
Training dataset							
True[a]	+	400	102	419	83	411	91
	−	146	383	190	339	133	396
Accuracy (%)		75.95		73.52		78.27	
Sensitivity (%)		79.68		83.47		81.87	
Specificity (%)		72.40		64.08		74.86	
Testing dataset							
True[a]	+	185	48	185	48	182	51
	−	59	164	77	146	54	169
Accuracy (%)		76.54		72.59		76.97	
Sensitivity (%)		79.40		79.40		78.11	
Specificity (%)		73.54		65.47		75.78	
Total dataset							
True[a]	+	585	150	604	131	593	142
	−	205	547	267	485	187	565
Accuracy (%)		76.13		73.23		77.87	
Sensitivity (%)		79.59		82.18		80.68	
Specificity (%)		72.74		64.49		75.13	

[a] "True +" and "True −" denotes the count of true positive and true negative.
[b] Denotes the count of predictive positive.
[c] Denotes the count of predictive negative.

disease incidence, prognosis [28,31,40] and hospital charges. The ANN model was reported to perform better than decision tree model [41—43], and the decision tree model was reported to perform better than logistic regression [44]. The results of this study indicated that the C5.0 decision tree is the best predictor, with 77.87% accuracy on the whole dataset. The logistic regression model came out second, with 76.13% accuracy. The ANNs model was poorest of the three models, with 73.23% accuracy. From the sensitivity analysis, as shown in Table 2, age, family history of diabetes, BMI and preference for salty food played an important role in the incidence of diabetes in our study communities, and the results provide evidence for the prevention of diabetes through community interventions.

As we show here, based on certain predictive attributes models can be developed that accurately predict the outcome of the incidence of a disease and its risk factors. These predictive models can be valuable tools in medicine [24]. However, there are areas of concern in the development of predictive models:

- the model should include all clinically relevant data;
- the model should be tested on an independent sample; and
- the model must make sense to the medical personnel who use supposed to implement it.

It has been shown that not all predictive models constructed using data mining techniques satisfy these requirements [45].

This study has limitations of feasibility and resource constraints. On one hand, the 1487 individuals were recruited only from two communities in Guangzhou, China. If more information from diabetes patients and normal controls was collected from every region of China, the sample would be more representative. On the other hand, the individuals in our study were diagnosed by a physician or during physical check in the past two years by self-reporting. If all the subjects underwent an oral glucose-tolerance test, and fasting and 2-hour glucose levels were measured to identify diabetes and prediabetes, the result would be more reliable.

Conclusion

In summary, we compared three prediction models for diabetes or prediabetes incidence using 12 risk factors. The results indicated that the C5.0decision tree model performed best on classification accuracy. This study may assist future researchers in choosing the optimal predictive models for implementing community lifestyle interventions to decrease the incidence of diabetes.

References

[1] Pan XR, Yang WY, Li GW, Liu J. Prevalence of diabetes and its risk factors in China, 1994. Diabetes Care 1997:1664—9.

[2] Gu D, Reynolds K, Duan X, Xin X, Chen J, Wu X, et al. Prevalence of diabetes and impaired fasting glucose in the Chinese adult population: International Collaborative Study of Cardiovascular Disease in Asia (InterASIA). Diabetologia 2003;46:1190—8.

[3] Yang W, Lu J, Weng J, Jia W, Ji L, Xiao J, et al. Prevalence of diabetes among men and women in China. N Engl J Med 2010; 362:1090—101.

[4] Neel JV. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 1962;14:353—62.

[5] Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, et al. Prevention of type 2

diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 2001;344:1343—50.

[6] Helmrich SP, Ragland DR, Leung RW, Paffenbarger Jr RS. Physical activity and reduced occurrence of non—insulin-dependent diabetes mellitus. N Engl J Med 1991;325:147—52.

[7] Bassuk SS, Manson JE. Epidemiological evidence for the role of physical activity in reducing risk of type 2 diabetes and cardiovascular disease. J Appl Physiol 2005;99:1193—204.

[8] Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med 2001;345:790—7.

[9] Salmeron J, Hu FB, Manson JE, Stampfer MJ, Colditz GA, Rimm EB, et al. Dietary fat intake and risk of type 2 diabetes in women. Am J Clin Nutr 2001;73:1019—26.

[10] Anderson JW, Randles KM, Kendall CW, Jenkins DJ. Carbohydrate and fiber recommendations for individuals with diabetes: a quantitative assessment and meta-analysis of the evidence. J Am Coll Nutr 2004;23:5—17.

[11] Colditz GA, Willett WC, Stampfer MJ, Manson JE, Hennekens CH, Arky RA, et al. Weight as a risk factor for clinical diabetes in women. Am J Epidemiol 1990;132:501—13.

[12] Wang Y, Rimm EB, Stampfer MJ, Willett WC, Hu FB. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. Am J Clin Nutr 2005;81:555—63.

[13] Koppes LL, Dekker JM, Hendriks HF, Bouter LM, Heine RJ. Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. Diabetes Care 2005;28:719—25.

[14] Carlsson S, Hammar N, Grill V. Alcohol consumption and type 2 diabetes: meta-analysis of epidemiological studies indicates a U-shaped relationship. Diabetologia 2005;48:1051—4.

[15] Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. JAMA 2007;298:2654—64.

[16] Gangwisch JE, Heymsfield SB, Boden-Albala B, Buijs RM, Kreier F, Pickering TG, et al. Sleep duration as a risk factor for diabetes incidence in a large US sample. Sleep 2007;30: 1667—73.

[17] Chao CY, Wu JS, Yang YC, Shih CC, Wang RH, Lu FH, et al. Sleep duration is a potential risk factor for newly diagnosed type 2 diabetes mellitus. Metab Clin Exp 2011;60:799—804.

[18] World Health Organization. 2008—2013 action plan for the global strategy for the prevention and control of non-communicable disease. Geneva: WHO; 2008.

[19] Schulze MB, Heidemann C, Schienkiewitz A, Bergmann MM, Hoffmann K, Boeing H. Comparison of anthropometric characteristics in predicting the incidence of type 2 diabetes in the EPIC-Potsdam study. Diabetes Care 2006;29:1921—3.

[20] Nyamdorj R, Qiao Q, Soderberg S, Pitkaniemi JM, Zimmet PZ, Shaw JE, et al. BMI compared with central obesity indicators as a predictor of diabetes incidence in Mauritius. Obesity 2009;17:342—8.

[21] Jia Z, Zhou Y, Liu X, Wang Y, Zhao X, Wang Y, et al. Comparison of different anthropometric measures as predictors of diabetes incidence in a Chinese population. Diabetes Res Clin Pract 2011;92:265—71.

[22] Lonn M, Mehlig K, Bengtsson C, Lissner L. Adipocyte size predicts incidence of type 2 diabetes in women. Faseb J 2010; 24:326—31.

[23] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77: 81—97.

[24] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113—27.

[25] Chen HY, Chuang CH, Yang YJ, Wu TP. Exploring the risk factors of preterm birth using data mining. Expert Syst Appl 2011;38:5384—7.

[26] Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011; 38:5507—13.

[27] Aslan K, Bozdemir H, Sahin C, Ogulata SN. Can neural network able to estimate the prognosis of epilepsy patients according to risk factors? J Med Syst 2010;34:541—50.

[28] Lee TT, Liu CY, Kuo YH, Mills ME, Fong JG, Huang CY. Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. Int J Med Inform 2011;80:141—50.

[29] Zorman M, Podgorelec V, Kokol P, Peterson M, Sprogar M, Ojstersek M. Finding the right decision tree's induction strategy for a hard real world problem. Int J Med Inform 2001;63: 109—21.

[30] Quinlan JR. Induction of decision trees. Mach Learn 1986;1: 81—106.

[31] Yeh JY, Wu TH, Tsao CW. Using data mining techniques to predict hospitalization of hemodialysis patients. Decis Support Syst 2011;50:439—48.

[32] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3—23.

[33] Ravikumar P, Bhansali A, Ravikiran M, Bhansali S, Walia R, Shanmugasundar G. Prevalence and risk factors of diabetes in a community-based study in North India: the Chandigarh urban diabetes study (CUDS). Diabetes Metab 2011;37: 216—21.

[34] Reis JP, Loria CM, Sorlie PD, Park Y, Hollenbeck A, Schatzkin A. Lifestyle factors and risk for new-onset diabetes: a population-based cohort study. Ann Intern Med 2011;155:292—9.

[35] Li G, Zhang P, Wang J, Gregg EW, Yang W, Gong Q, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing diabetes prevention study: a 20-year follow-up study. Lancet 2008;371:1783—9.

[36] Saaristo T, Moilanen L, Korpi-Hyovalti E, Vanhala M, Saltevo J, Niskanen L. Lifestyle intervention for prevention of type 2 diabetes in primary health care. Diabetes Care 2010;33:2146—51.

[37] Jorgensen L, Joakimsen R, Ahmed L, Stormer J, Jacobsen BK. Smoking is a strong risk factor for non-vertebral fractures in women with diabetes: the Ttomso study. Osteoporosis Int 2011;22:1247—53.

[38] Zhang L, Curhan GC, Hu FB, Rimm EB, Forman JP. Association between passive and active smoking and incident type 2 diabetes in women. Diabetes Care 2011;34:892—7.

[39] Liu C, Yu Z, Li H, Wang J, Sun L, Qi Q, et al. Associations of alcohol consumption with diabetes mellitus and impaired fasting glycemia among middle-aged and elderly Chinese. BMC Public Health 2010;10:713.

[40] Lai CL, Lai CL, Chien SW, Fang K. Identification and validation of predictive factors for glycemic control: neural networks vs. logistic regression. Proceedings of the 2007 WSEAS International Conference on Computer Engineering and Applications, Gold Coast, Australia, January 17—19, 2007;300—5.

[41] Wang J, Li M, Hu Y, Zhu Y. Comparison of hospital charge prediction models for gastric cancer patients: neural network vs. decision tree models. BMC Health Serv Res 2009;9:161.

[42] Kang JO, Chung SH, Suh YM. Prediction of hospital charges for the cancer patients with data mining techniques. J Korean Soc Med Inform 2009:1513—23.

[43] Lee SM, Kang JO, Suh YM. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. J Korean Med Sci 2004;19:677—81.

[44] Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. Int J Med Inform 2001;62:103—11.

[45] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 2001;22:215—31.