Contents lists available at ScienceDirect

Diabetes Research and Clinical Practice

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

Brief report

Identification of metabolic syndrome using decision tree analysis

Apilak Worachartcheewan [a], *Chanin Nantasenamat* [a], *Chartchalerm Isarankura-Na-Ayudhya* [a], *Phannee Pidetcha* [b], *Virapong Prachayasittikul* [a,*]

[a] *Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[b] *Center of Medical Laboratory Services, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

ARTICLE INFO

Article history:
Received 11 May 2010
Received in revised form
14 May 2010
Accepted 7 June 2010

Keywords:
Metabolic syndrome
Diabetes
Cardiovascular diseases
Data mining
Decision tree
Health status

ABSTRACT

This study employs decision tree as a decision support system for rapid and automated identification of individuals with metabolic syndrome (MS) among a Thai population. Results demonstrated strong predictivity of the decision tree in classification of individuals with and without MS, displaying an overall accuracy in excess of 99%.

1. Introduction

Metabolic syndrome (MS) can be defined as a group of metabolic abnormalities characterized by central obesity, dyslipidemia, hyperglycemia, and hypertension [1]. MS underlies the increased prevalence of cardiovascular disease and type 2 diabetes worldwide [2,3], therefore, various studies have been performed to evaluate the prevalence of MS in different populations, including USA [4], India [5], Japan [6], China [7], Korea [8], Philippine [9] and Thailand [10]. The criteria used to identify MS in these studies were performed using differing cutoff.

Data mining represents a rapid and easy approach for knowledge extraction from large complex databases [11]. In this respect, data mining has been applied in predicting complications from MS [12] as well as predicting MS status based on dietary and genetic parameters [13]. To develop a rapid and automated approach for the identification of MS, we investigated the prospect of using decision tree analysis for classifying individuals, within a Thai population, as MS and non-MS on the basis of their health parameters.

2. Methods

A data set of 15,365 individuals residing in urban Thailand was obtained from a cross-sectional investigation of those receiv-

* *Corresponding author*. Tel.: +66 2 441 4376; fax: +66 2 441 4380.
E-mail address: mtvpr@mahidol.ac.th (V. Prachayasittikul).
0168-8227/$ – see front matter
doi:10.1016/j.diabres.2010.06.009

ing annual medical check-ups from the Faculty of Medical Technology, Mahidol University in 2007. The details of the study have been described previously [10]. A subset (5, 638 individuals) of the total data set was used in this study by selecting individuals meeting the first IDF criteria as identified by the specific BMI cutoff of $\geq$25 kg/m^2.

Each individual was described by a set of 20 health parameters consisting of 5 physiological parameters: gender, age, systolic blood pressure (SBP), diastolic blood pressure (DBP) and BMI $\geq$25 kg/m^2 as well as 15 blood chemical parameters: fasting plasma glucose (FPG), blood urea nitrogen, creatinine, uric acid, cholesterol, triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, hemoglobin, hematocrit, white blood cell count and platelet count.

Individuals with central obesity plus two or more of the following were classified as MS: BP $\geq$130/85 mm Hg or treatment of previously diagnosed hypertension, FPG $\geq$100 mg/dL or previously diagnosed type 2 diabetes, TG $\geq$150 mg/dL or specific treatment for triglyceride abnormality as well as HDL-C <40 mg/dL in males or <50 mg/dL in females or specific treatment for HDL-C abnormality.

Classification of individuals as having or not having MS was performed using the decision tree classifier based on the J48 algorithm of Weka [11]. The complete health parameters of participants were used as independent variables while their MS status was used as the dependent variable.

Decision tree is a supervised approach that uses a set of if-then rules [11] to classify samples into categories of interest. The algorithm finds the most important independent variable and sets it as the root node, which is followed by bifurcating to the next best variables. The tree flows in a top-down manner from the root node through the internal nodes (the independent variables) and finally to the terminal leaf nodes (the class prediction). The confidence factor was set to 25% as it had been demonstrated to work reasonably well.

Rigorous assessment of the predictive performance was made by separating the data set into three sets: training set, 10-fold cross-validation set and external validation set. The first 2 sets, comprising of 90% of the data set, were used for assessing the internal performance, while the last set, consisting of 10% of the data set, was used for assessing the external performance. Statistical parameters to evaluate the predictive performance of the decision tree include the sensitivity, specificity, accuracy, positive predictive value and negative predictive value [14] and Matthews correlation coefficient [15].

3. Results

As shown in Table 1, the constructed decision tree could classify individuals to their corresponding MS and non-MS groups with accuracy in excess of 99.8% for all three sets of data along with excellent statistical results. The decision tree (as shown in Fig. 1) is comprised of 21 nodes and 22 leaves. Triglyceride level was assigned by the decision tree as the first and most informative node. Furthermore, it was found that 6 (sex, TG, SBP, DBP, FPG and HDL-C) out of 20 independent variables were important for classifying the MS and non-MS populations, while other attributes were implied to be irrelevant for predicting MS. It is observed that a majority of leaves in the left portion of the decision tree was classified as non-MS while most was classified as MS in the right portion.

The decision tree gave the following criteria for MS classification: TG >149 mg/dL, FPG >99 mg/dL, HDL-C <39.9 mg/dL in males or HDL-C <49.7 in females and SBP was >128 mm Hg, which was in excellent agreement with the IDF criteria: TG >150 mg/dL, FPG >100 mg/dL, HDL-C <40 mg/dL in males or HDL-C <50 in females and SBP >130 mm Hg. It should be noted that such phenomenon further confirms the robustness of the decision tree algorithm for MS classification.

Table 1 – Performance summary of MS prediction using decision tree analysis.

	MS	Non-MS	x_1 [a]	x_2 [b]	x_3 [c]	x_4 [d]	x_5 [e]	x_6 [f]
Training set			100	99.96	99.98	99.96	100	0.9996
MS	2690	1						
Non-MS	0	2377						
10-Fold cross-validation			99.89	99.83	99.86	99.85	99.87	0.9972
MS	2687	4						
Non-MS	3	2374						
External validation			99.67	100	99.82	100	99.63	0.9965
MS	300	0						
Non-MS	1	269						

[a] Sensitivity.
[b] Specificity.
[c] Accuracy.
[d] Positive predictive value.
[e] Negative predictive value [14].
[f] Matthews correlation coefficient [15].

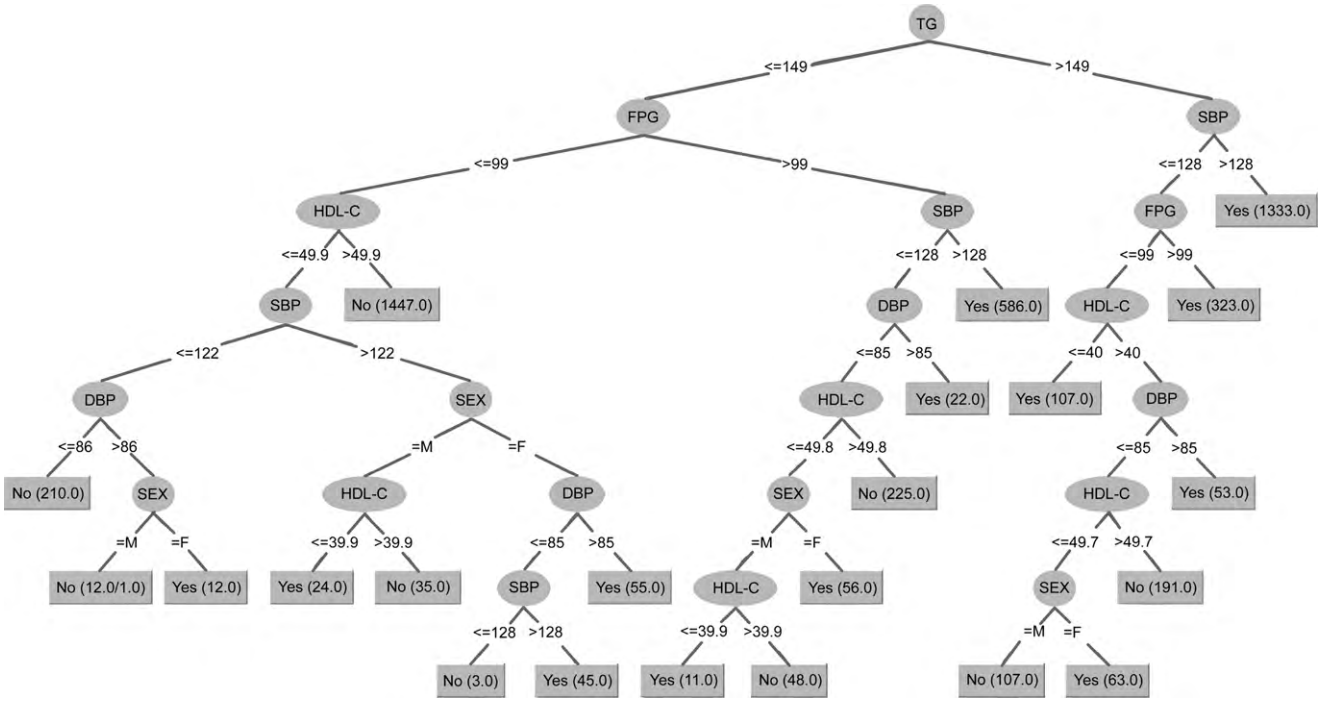

Fig. 1 – A decision tree classification of MS status (yes: MS, no: non-MS). The terminal leaf nodes where classifications are made is assigned by one or two numbers in parentheses where the former and latter describes the number of correctly and incorrectly classified samples, respectively.

4. Discussion

A study by Helminen et al. [16] indicated that general practitioners fail to accurately identify MS in patients and that patients were unaware of having MS. Such findings necessitates a critical reconsideration of MS detection. Many previous studies have confirmed that decision tree analyses are effective tools for clinical data analysis [13,17,18]. In this report, we have studied the usefulness of decision tree for automated MS classification. Results suggested that TG level, which was the root node, was most informative for MS identification. This is in accordance with the findings by Lemieux et al. [19] that TG is an important parameter that can be used as a first screening phenotype for identifying a subgroup of individuals likely to be characterized with a cluster of MS features. Results indicated that the combination of TG + BP, FPG + BP and TG + BP + FPG were strong predictors of MS, which correlated well with our previous study [10] and the study of Lee et al. [20]. Interestingly, decision tree analysis was able to discover the same criteria used by IDF in classifying MS.

In conclusion, this study demonstrated the practical usefulness of decision tree analysis as a decision support system for automatic detection of MS from an urban Thai population. The findings in this study revealed that decision tree analyses can be successfully employed to identify individuals with MS. The parameters that are closely associated with risk of MS include TG, SBP and DBP, FPG and HDL-C.

Acknowledgments

We gratefully acknowledge financial support from the annual budget grant of Mahidol University (2551–2555). C.N. gratefully acknowledges support from the Young Scholars Research Fellowship from the Thailand Research Fund (Grant No. MRG5080450), and the Science and Technology Research Grant from the Thailand Toray Science Foundation. A.W. is thankful for a scholarship from the Royal Golden Jubilee Ph.D. program of the Thailand Research Fund under supervision of V.P.

Conflict of interest

There are no conflicts of interest.

REFERENCES

[1] Babu A, Fogelfeld L. Metabolic syndrome and prediabetes. Dis Mon 2006;52:55–144.
[2] World Health Organization, Diabetes; 2008. Available from http://www.who.int/mediacentre/factsheets/fs312/en/index.html [cited 15 January 2009].
[3] World Health Organization, Cardiovascular diseases; 2007. Available from http://www.who.int/mediacentre/factsheets/fs317/en/index.html [cited 15 January 2009].

[4] Ford ES, Giles WH, Mokdad AH. Increasing prevalence of the metabolic syndrome among U.S. adults. Diabetes Care 2004;27:2444–9.
[5] Ramachandran A, Snehalatha C, Satyavani K, Sivasankari S, Vijay V. Metabolic syndrome in urban Asian Indian adults—a population study using modified ATP III criteria. Diabetes Res Clin Pract 2003;60:199–204.
[6] Arai H, Yamamoto A, Matsuzawa Y, Saito Y, Yamada N, Oikawa S, et al. Prevalence of metabolic syndrome in the general Japan population in 2000. J Atheroscler Thromb 2006;13:202–8.
[7] Li Y, Yang X, Zhai F, Kok FJ, Zho W, Piao J, et al. Prevalence of the metabolic syndrome in Chinese adolescents. Br J Nutr 2008;99:565–70.
[8] Park HS, Lee SY, Kim SM, Han JH, Kim DJ. Prevalence of the metabolic syndrome among Korean adults according to the criteria of the International Diabetes Federation. Diabetes Care 2006;20:933–4.
[9] Morales DD, Punzalan FE, Paz-Pacheco E, Sy RG, Duante CA. Metabolic syndrome in the Philippine general population: prevalence and risk for atherosclerotic cardiovascular disease and diabetes mellitus. Diab Vasc Dis Res 2008;5: 36–43.
[10] Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Lower BMI cutoff for assessing the prevalence of metabolic syndrome in Thai population. Acta Diabetol; doi:10.1007/s00592-009-0137-0.
[11] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., San Francisco: Morgan Kaufmann; 2005.
[12] Wojtusiak J, Michalski RS, Simanivanh T, Baranova AV. Towards application of rule learning to the meta-analysis of clinical data: an example of the metabolic syndrome. Int J Med Inform 2009;78:e104–111.
[13] Szabo de Edelenyi F, Goumidi L, Bertrais S, Phillips C, Macmanus R, Roche H, et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters using Random Forest. Genes Nutr 2008;3:173–6.
[14] Kuo WJ, Chang RF, Chen DR, Lee CC. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Res Treat 2001;66:51–7.
[15] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[16] Helminen EE, Mantyselka P, Nykanen I, Kumpusalo E. Far from easy and accurate—detection of metabolic syndrome by general practitioners. BMC Fam Pract 2009;10:76.
[17] Firouzi F, Rashidi M, Hashemi S, Kangavari M, Bahari A, Daryani NE, et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. Eur J Gastroenterol Hepatol 2007;19:1075–81.
[18] Quentin-Trautvetter J, Devos P, Duhamel A, Beuscart R. Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France. Stud Health Technol Inform 2002;90:557–61.
[19] Lemieux I, Poirier P, Bergeron J, Almeras N, Lamarche B, Cantin B, et al. Hypertriglyceridemic waist: a useful screening phenotype in preventive cardiology? Can J Cardiol 2007;23(Suppl B):23B–31B.
[20] Lee CM, Huxley RR, Woodward M, Zimmet P, Shaw J, Cho NH, et al. The metabolic syndrome identifies a heterogeneous group of metabolic component combinations in the Asia-Pacific region. Diabetes Res Clin Pract 2008;81:377–80.