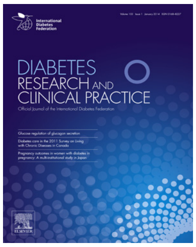
Contents available at ScienceDirect

Diabetes Research and Clinical Practice

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

International Diabetes Federation



Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study

Azra Ramezankhani[a], *Omid Pournik*[b,c], *Jamal Shahrabi*[d], *Davood Khalili*[a,e], *Fereidoun Azizi*[f], *Farzad Hadaegh*[a,*]

[a] *Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Science, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[b] *Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*
[c] *Medical Informatics Research Center, Faculty of Medicine, Mashhad, Iran*
[d] *Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran*
[e] *Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[f] *Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

ARTICLE INFO

Article history:
Received 30 June 2013
Received in revised form
15 April 2014
Accepted 5 July 2014
Available online 18 July 2014

Keywords:
Decision tree
Type 2 diabetes
Prediction model

ABSTRACT

Aims: The aim of this study was to create a prediction model using data mining approach to identify low risk individuals for incidence of type 2 diabetes, using the Tehran Lipid and Glucose Study (TLGS) database.

Methods: For a 6647 population without diabetes, aged ≥20 years, followed for 12 years, a prediction model was developed using classification by the decision tree technique. Seven hundred and twenty-nine (11%) diabetes cases occurred during the follow-up. Predictor variables were selected from demographic characteristics, smoking status, medical and drug history and laboratory measures.

Results: We developed the predictive models by decision tree using 60 input variables and one output variable. The overall classification accuracy was 90.5%, with 31.1% sensitivity, 97.9% specificity; and for the subjects without diabetes, precision and *f*-measure were 92% and 0.95, respectively. The identified variables included fasting plasma glucose, body mass index, triglycerides, mean arterial blood pressure, family history of diabetes, educational level and job status.

Conclusions: In conclusion, decision tree analysis, using routine demographic, clinical, anthropometric and laboratory measurements, created a simple tool to predict individuals at low risk for type 2 diabetes.

* *Corresponding author at*: P.O. Box 19395-4763, Tehran, Iran. Tel.: +98 2122409301x5; fax: +98 2122402463.
E-mail addresses: fzhadaegh@endocrine.ac.ir, farzadhadaegh@gmail.com (F. Hadaegh).

http://dx.doi.org/10.1016/j.diabres.2014.07.003
0168-8227/

1. Introduction

The prevalence of diabetes has been increasing rapidly over the past decade. Around 382 million adults worldwide had diabetes mellitus in 2013, and it is predicted that this number will reach 592 million by 2035 [1]. Diabetes is a costly chronic illness because both of the direct costs for medical services and the indirect cost, due to lost of productivity [2–4]. Diabetes results from interaction between genetic and environmental factors [5,6]. Cohort studies have shown that modification of some lifestyle related risk factors can prevent diabetes [7]. The preventable nature of diabetes led the World Health Organization (WHO) to recommend using simple and practical strategies for early identification of populations at high risk for diabetes, while avoiding the burden of prevention and treatment for the individuals at low risk and for society [7,8]. During the last two decades, many researchers have developed prediction models using risk scoring system [6,7]. There is no preferred diabetes risk score model, because statistical properties, context of use, available data, trade-off between sensitivity and specificity determine which types of models should be used [6]. On the other hand, the false positive and false negative rates in many risk score models raise questions about their utility in clinical and public health practice [7,9].

In any program aimed at the early identification of type 2 diabetes, attention should be paid to avoidable adverse effects of false positive and false negative rates such as psychosocial and economic costs of screening or confirmatory tests, time and other resources needed to conduct screening programs [9]. Physical harm associated with diabetes screening may be considered negligible, but due to involvement of large number of people in screening programs, psychological and social harm could be more substantial [10].

The aim of this study was to develop a predictive model to identify individuals at low risk for type 2 diabetes, and for use as a complementary model in screening and public health programs. This model was developed using data mining which is the process of selecting, exploring and modeling large amounts of data for knowledge discovery of database (KDD) and discovery of unknown patterns or relationships that provide clear and useful results [11,12]. Many data mining techniques have been developed rapidly in the medical field during last years [13–15]. Several studies have used data mining to construct prediction models for the incidence of diabetes [16–22]. In this study we applied data mining to construct predictive models for the ‘‘no occurrence of diabetes type 2’’ based on routine questions, physical examination and biochemical measurement using information collected in the Tehran Lipid and Glucose Study (TLGS). The decision tree technique was selected for building predictive model. The ability of this technique is to model nonlinear relationships and create a set of simple classification rules, which are easier to understand than some other traditional models [23–25].

2. Methods

2.1. *Study population*

The Tehran Lipid and Glucose Study (TLGS), a prospective population-based study has been described in detail elsewhere [26]. Briefly, the baseline study (phase 1) was performed from 1999 to 2001 and was followed in 3 consecutive phases, 2002–2005 (phase 2), 2005–2008 (phase 3), and the last 2009–2012 (phase 4). At baseline, 4751 families including over 15,000 residents, aged ≥3 years, of district 13 of Tehran were selected by cluster random sampling method and followed from the date of enrolment through phase 4. Moreover in the second phase, 3551 new people entered and were followed in the next two phases (3 and 4). In our study subjects aged ≥20 years from the first and second phases were selected. We excluded subjects with prevalent diabetes at baseline and those with missing data regarding fasting and 2-h glucose; finally 10,310 subjects without diabetes remained in first and second phases, which followed in next phases. Overall, 3663 (35%) subjects were lost to follow-up and 729 new cases of type 2 diabetes were identified by the end of the 12 years follow-up (phase 4); in 5918 subjects, diabetes did not occur (see supplementary Fig. S1). Incident type 2 diabetes was defined based on fasting plasma glucose (FPG) ≥126 mg/dl or 2-h post-challenge plasma glucose (2 h PCPG) ≥200 mg/dl or taking anti-diabetic medication [27]. This study has been approved by Ethical Committee of Research Institute for Endocrine Sciences and a subject gave written informed consent.

2.2. *Measurements of variables*

Data included 6647 participants of the TLGS study. Demographic characteristics, including age, gender, marital status, education, smoking status, information on physical activity and medical and drug history were collected by interview using a pretested questionnaire. Anthropometric measures including weight, height, waist, hip and wrist circumference were obtained according to a standard protocol [28]. Systolic and diastolic blood pressures were measured twice on the right arm using a standardized mercury sphygmomanometer, and the mean of two measurements was considered as the subject’s blood pressure. All blood parameters, except for 2 h plasma glucose, were based on fasting blood samples (after 12–14 h overnight fasting). After withdrawal, the blood samples were centrifuged within 30–45 min of collection and kept cool until analysis at the TLGS research laboratory. Fasting and 2 h plasma glucose were determined based on an enzymatic colorimetric method using oxidase kits (Pars Azmoon Inc., Tehran, Iran) with inter- and intra-assay coefficient of variations (CV) less than 2.2%. Triglycerides (TGs), total and HDL-cholesterol were determined as described elsewhere [29,30].

3. Data mining

3.1. *Data understanding*

The most important aspect in data mining is the quality of data because it influences the quality of the results. Assessing