BMJ Open

Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study

Azra Ramezankhani,[1] Esmaeil Hadavandi,[2,3] Omid Pournik,[4] Jamal Shahrabi,[2] Fereidoun Azizi,[5] Farzad Hadaegh[1]

To cite: Ramezankhani A, Hadavandi E, Pournik O, *et al*. Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study. *BMJ Open* 2016;**6**:e013336. doi:10.1136/bmjopen-2016-013336

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-013336).

Received 7 July 2016
Revised 9 September 2016
Accepted 3 October 2016

For numbered affiliations see end of article.

Correspondence to
Dr Farzad Hadaegh;
fzhadaegh@endocrine.ac.ir

ABSTRACT

Objective: The current study was undertaken for use of the decision tree (DT) method for development of different prediction models for incidence of type 2 diabetes (T2D) and for exploring interactions between predictor variables in those models.

Design: Prospective cohort study.

Setting: Tehran Lipid and Glucose Study (TLGS).

Methods: A total of 6647 participants (43.4% men) aged >20 years, without T2D at baselines ((1999–2001) and (2002–2005)), were followed until 2012. 2 series of models (with and without 2-hour postchallenge plasma glucose (2h-PCPG)) were developed using 3 types of DT algorithms. The performances of the models were assessed using sensitivity, specificity, area under the ROC curve (AUC), geometric mean (G-Mean) and F-Measure.

Primary outcome measure: T2D was primary outcome which defined if fasting plasma glucose (FPG) was ≥7 mmol/L or if the 2h-PCPG was ≥11.1 mmol/L or if the participant was taking antidiabetic medication.

Results: During a median follow-up of 9.5 years, 729 new cases of T2D were identified. The Quick Unbiased Efficient Statistical Tree (QUEST) algorithm had the highest sensitivity and G-Mean among all the models for men and women. The models that included 2h-PCPG had sensitivity and G-Mean of (78% and 0.75%) and (78% and 0.78%) for men and women, respectively. Both models achieved good discrimination power with AUC above 0.78. FPG, 2h-PCPG, waist-to-height ratio (WHtR) and mean arterial blood pressure (MAP) were the most important factors to incidence of T2D in both genders. Among men, those with an FPG≤4.9 mmol/L and 2h-PCPG≤7.7 mmol/L had the lowest risk, and those with an FPG>5.3 mmol/L and 2h-PCPG>4.4 mmol/L had the highest risk for T2D incidence. In women, those with an FPG≤5.2 mmol/L and WHtR≤0.55 had the lowest risk, and those with an FPG>5.2 mmol/L and WHtR>0.56 had the highest risk for T2D incidence.

Conclusions: Our study emphasises the utility of DT for exploring interactions between predictor variables.

Strengths and limitations of this study

- We used a large population-based sample for our study.
- The direct measurements of glucose value and anthropometric indices were used rather than self-reported information for predictor variables and outcome.
- Our study proposes a new approach for detecting interactions between predictors.
- There were no data available on the dietary intake among participants.
- External validity did not perform for the derived prediction models.

INTRODUCTION

The prevalence of type 2 diabetes (T2D) mellitus has been increasing rapidly over the past decade. Around 366 million people worldwide had diabetes mellitus in 2011, and this number is expected to reach 552 million by 2030.[1] Several risk factors, such as age, sex, ethnicity, family history, obesity and hypertension, are well documented. However, detecting the precise interaction of these and other risk factors with one another is a complex process that varies both within and across populations.[2–4]

During the past two decades, dozens of prediction models for diabetes have been developed using logistic or Cox regression models,[4 5] while recently a systematic search of those multivariable models has shown that few reported prediction models contain interactions, and it seems that few researchers examine them.[6] There are a number of reasons for not using interactions in traditional statistical methods. First, there are generally many possible predictor variables in medical research which make the task of

variable selection difficult. Hence, traditional statistical methods are poorly suited for this type of multiple comparisons. Second, many clinical variables are not normally distributed and different groups of participants may have markedly different degrees of variations. Third, assessment of interactions, using the traditional regression models, requires prespecification of the interaction terms, for example, in a linear model involving outcome Y, and two predictor variables (x_1 and x_2), the product term x_1x_2 is the common representation of the two-way interaction effect. As the number of variables in the model increases, the number of possible interactions that can be investigated is large and leads to a complicated model that can be difficult to fit and interpret.[6 7] Non-parametric regression has been introduced in 1963 which are another class of simple regression models for explanation and prediction nowadays known as 'recursive partitioning' or 'decision trees' (DT). Many variants and extensions of the tree methods have been published in the past 50 years, which have been widely used in many fields such as machine learning, data mining and pattern recognition.[8 9] Recursive partitioning is a statistical method for exploration of interactions or non-linear relationships among explanatory variables, identification of different subgroups, detection of the most important variables in those subgroups, and finally offering a new way to look at complex data.[8 10 11] Since there will never be enough resources to implement every prevention programme for all target groups, health policymakers prefer interventions that target high-risk groups.[12] Therefore, DT models might be helpful for identifying different groups which allow implementation of specific interventions for each group according to their risk probabilities (low-risk and high-risk groups).

The aim of this study was to develop a series of classification trees for adult men and women based on three commonly used DT algorithms (Classification and Regression Tree (CART), Quick Unbiased Efficient Statistical Tree (QUEST) and commercial version (C5.0)) to gain more information on interactions between factors contributing to the incidence of T2D. We used the Tehran Lipid and Glucose Study (TLGS) database for our analysis.

METHODS

Study population

The TLGS, an ongoing prospective study, has been described in detail elsewhere.[13] Briefly, the baseline study (phase 1) was performed from 1999 to 2001, with follow-ups in three consecutive phases, 2002–2005 (phase 2), 2005–2008 (phase 3), and the last 2009–2012 (phase 4). After the cross-sectional phase (phase 1), participants were assigned to a cohort and a prospective interventional study. For this study, 10 368 participants aged ≥20 years from the first phase were selected and followed from the date of enrolment through phase 4; moreover, in the second phase, 2440 new participants entered and were followed in the next two phases (3 and 4). We excluded participants with prevalent T2D at baseline (n=1376) and those with missing data regarding fasting plasma glucose (FPG) and 2-hour postchallenge plasma glucose (2h-PCPG) (n=1122). Overall, 3663 (35%) participants were lost to follow-up and 729 new cases of T2D were identified by the end of phase 4 (figure 1). The written informed consent was obtained from each participant.

Clinical, anthropometric and laboratory measurements

Information on demographics, education, smoking status, physical activity, and medical and drug history was collected by interview. For women, additional information on reproductive history, menstruation status and interventions to prevent pregnancy was collected using a pretested questionnaire. Anthropometric measures including weight, height and waist circumference (WC) were measured, according to a standard protocol.[14] Body mass index (BMI) was calculated as weight (kg)/height $(m)^2$. Waist-to-hip ratio (WHpR) was calculated as WC/hip circumference and waist-to-height ratio (WHtR) was calculated as WC/height. Systolic and diastolic blood pressure (SBP and DBP, respectively), and blood parameters such as FPG, 2h-PCPG, triglycerides (TGs), total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-c) were measured using previously reported methods.[15] TGs to HDL-c ratio (TG/HDL) ratio was obtained as TG/HDL-c and TC-to-HDL-c (TC/HDL) was calculated as TC/HDL-c.

Definition of variables and outcome

Education level was categorised to five levels as illiterate, 1–5 years, 6–12 years, 13–16 years and more than 16 years schooling. Marital status was categorised as single, married, widowed and divorced. A current smoker was defined as a person who smokes cigarettes daily or occasionally. Former smokers were defined as individuals who have smoked daily or occasionally and who had quit smoking. Passive smoking was defined as exposure to secondhand cigarette smoke in the home, at work or in other environments. A family history of premature cardiovascular diseases (CVD) was considered as any experience of fatal or non-fatal myocardial infarction, stroke or sudden cardiac arrest in first-degree relatives, if it occurred before 55 years of age in male relatives and before 65 years of age in female relatives. A history of CVD was defined as previous ischaemic heart disease and/or cerebrovascular accidents. A family history of diabetes (FHD) was defined as having T2D in first-degree relatives. On the basis of their self-reported levels of leisure time physical activity, participants were categorised into two groups in which 'inactive' means those doing exercise or labour less than three times a week or performing activities achieving lower than 600 MET. Mean arterial blood pressure (MAP) was obtained as ([(2×diastolic)+systolic]/3).[16] Pulse pressure was defined as SBP minus DBP. Participants were grouped into two categories based on participating in the lifestyle

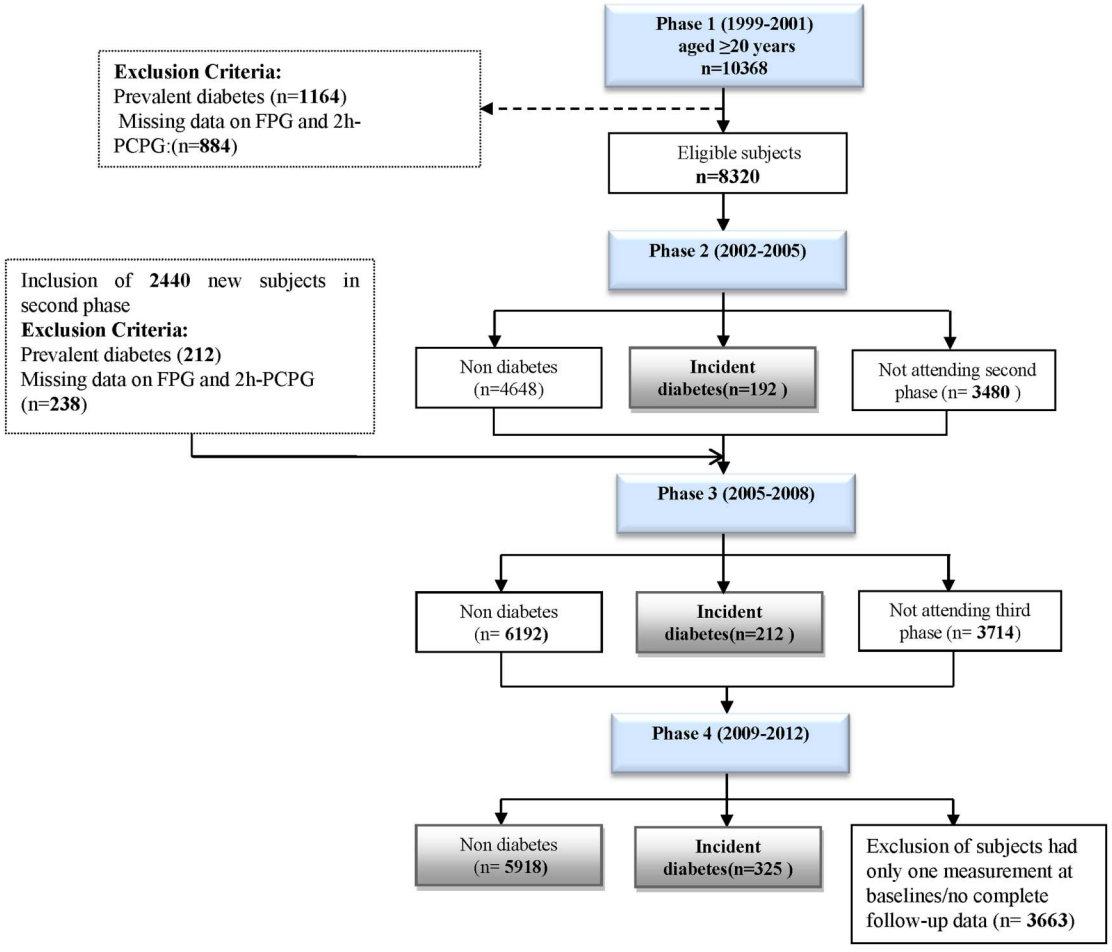

Figure 1 Flow diagram for the selection of study participants in the Tehran Lipid and Glucose Study. 2h-PCPG, 2-hour postchallenge plasma glucose; FPG, fasting plasma glucose.

intervention. Women were categorised into three groups on the basis of their menstruation status: having normal menstrual cycle by taking medication, normal menopause, early menopause because of surgery or other reasons. Women were also categorised to six levels considering pregnancy prevention methods: use of hormonal contraceptive drugs, intrauterine devices (IUDs), using condoms, withdrawal method, tubectomy/vasectomy and not applicable. They were also categorised into two groups based on birth history, a history of hypertension and hyperglycaemia in pregnancy. Incidence of T2D (outcome variable) was defined based on an FPG≥7.0 mmol/L or 2h-PCPG≥11.1 mmol/L or taking antidiabetic medication in all phases of the study.[17] Final data sets consisted of 6647 cases (3762 women) which included 54 and 44 primary predictor variables in women and men, respectively.

Statistical methods

Data preparation

Data were prepared before analysis. Data preparation included: missing data handling, variables selection, defining the train and validation data sets and balancing the train data sets.

Missing data handling

Results of the Little's missing completely at random (MCAR) test[18] on the primary set of predictor variables showed that in men the pattern of missing data was completely at random (MCAR) (p=0.15), but for women missing values were not MCAR (p<0.001).[19] We used single imputation for imputing the missing data. For imputation, all the primary variables were included, except for the outcome variable. Continuous variables were imputed by the CART method,[10] using SPSS modeler (V.14.2.0.3, IBM), and for categorical variables we applied the weighted K-Nearest Neighbor approach using RapidMiner (V.5).[20]

Training and validation data

The entire data sets of men and women were divided into two sets using stratified random sampling: a training set consisted of 70% of the data for model development, and a test or validation set consisted of the remaining (30%) for model validation (internal validation) (figure 2).

Data balancing

Most of the popular classification algorithms such as DT work well when the positive and negative cases are

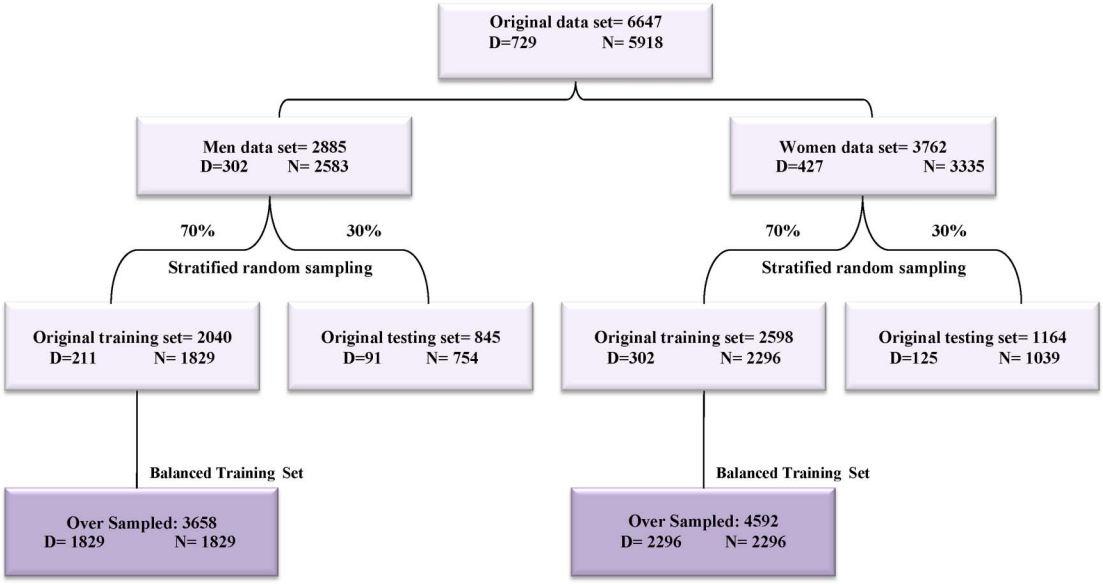

Figure 2 Generation of training and validation data set diagram.

evenly distributed and problems arise when the data set is imbalanced.[21] The class imbalance in medical data occurs when there are many more cases of some classes (majority class or negative) than others (minority class or positive).[22] In such cases, standard classifiers tend to produce high accuracy over that of the majority class.[23] There are a finite number of solutions to handle imbalanced data sets.[21–23] In our previous work, we showed the effectiveness of Synthetic Minority Oversampling Technique (SMOTE) for handling imbalanced data sets.[24] In this study, we balanced two training data sets of men and women using SMOTE as previously reported (figure 2).[24]

Variables selection

Variable or feature selection methods have been used since the 1970s in the fields of statistics and machine learning techniques.[25] Variable selection methods have been shown to be effective in removing redundant and irrelevant variables, improving prediction performance of learning algorithms and reducing the effects of high dimensionality in the data.[26] Therefore, in order to identify the best subset of variables while retaining the predictive power of the original variables, we applied the multivariate filter approach, using correlation-based feature selection and consistency-based feature selection as two evaluation criteria in conjunction with the Best First, Genetic Algorithm as two search strategies.[27] Therefore, four subsets of variables were selected using a combination of the two search strategies and two evaluation criteria. To arrive at the final set, the four subsets were reviewed to choose the variables that were observed at least in two subsets. Variable selection methods were applied on the training data sets after imputation of missing data. We used the Weka toolkit (V.3.2.) for selecting variables.

Statistical analysis

Baseline characteristics were compared between participants with and without T2D across men and women. Also, characteristics were compared between followed up versus non-followed up participants. Comparisons were done using Student's t-test and χ^2 with a two-tailed $p<0.05$ being considered significant.

Methods for DT modelling

There are many different algorithms for fitting tree-structured models coming from different communities.[9 28] All the DT algorithms generate a set of classification rules and construct a DT. A tree has three types of nodes: root node, internal node and terminal nodes. Both the root and the internal nodes are partitioned into two nodes in the next layer; however, the terminal nodes do not have offspring nodes. The root node contains the learning sample from which the tree is grown. The basic process of developing a DT includes three elements: the selection of variable for splits the data (splitting criteria), stopping rule to decision of when to stop splitting a node and mark it terminal, and the pruning methods.[8]

To choose the right algorithm for our problem, we applied three types of DT algorithms which are widely used for generating a binary tree: the CART algorithm,[8] QUEST[29] and C5.0.[30] All the DT models were performed using IBM SPSS modeler 14.2.

Model evaluation

Performances of the models were evaluated on the test or validation data sets. In data mining, the classifier is basically evaluated by accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and the area under the curve (AUC). When data are imbalanced, accuracy performs better on the

majority classes (negative cases). The geometric mean (G-Mean), however, indicates the balance between model performance on the negative and positive classes and avoids overfitting to the negative class.[31]

$$G-\text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (1)$$

F-Measure, harmonic mean of PPV (precision) and sensitivity (recall), is another measure that increases proportionally to the increase of precision and recall. A high value of F-Measure indicates that the model performs better on the positive class.[31 32] We chose sensitivity and G-Mean for comparison of the models and select the best one.

$$F-\text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} \times \text{recall}} \quad (2)$$

RESULTS

Missing data analysis showed that about 59% and 70% of primary variables (44 and 54 in men and women, respectively) had at least two missing data. The ranges of missing data were (0.1–6%) and (0.1–5%) in the women and men data set, respectively. Using the variable selection methods, 15 and 20 variables were identified to include in the model building process for men and women, respectively. The percentage of missing data for selected variables has been shown in tables 1 and 2.

Characteristics of participants

Baseline characteristics of the study population are presented in tables 1 and 2. During a median 9.5 years of follow-up (IQR 6.13–10.2 years), T2D developed in 302 men (10%) and 427 women (11%). Comparison of baseline characteristics between the followed and non-followed participants (only for selected variables) is shown in tables 3 and 4. Followed men had higher value for TC/HDL (5.5 vs 5.4), but lower age (41.8 vs 43.4 years). The proportion of individuals with low education levels (≤5 years) was higher in followed men (20.5% vs 26.2%). Followed women had lower value for age (39.6 vs 40.5 years), pulse pressure (39.3 vs 40.2 bpm) and MAP (89.2 vs 89.9 mm Hg). The proportion of illiterate women was lower in followed women (8.1% vs 13.2%).

Model performances

We constructed the DT models using the balanced training data sets with two set of variables: (1) selected variables that included 2h-PCPG, and (2) selected variables without 2h-PCPG. The performance measures for the two types of DT models are shown in tables 5 and 6. Comparison between models 1 and 2 shows that by removing 2h-PCPG from the variables list, the sensitivity for all three models decreases by 5–10% and 2–5% in men and women, respectively. Results show that QUEST has the highest sensitivity and G-Mean among all models for both men and women; therefore, we chose it as the best DT models.

DT analysis in men

Figure 3 depicts the DT for model 1, including the predictor variables and the cut-off points for each predictor. It used four variables (FPG, 2h-PCPG, age and WHtR) for classification and generated seven decision rules; each rule identifies a special subgroup with a certain probability of outcome (positive or negative) for each person belonging to that subgroup. The FPG, located on the top of the tree, was the most important factor in incidence of T2D.

Table 7 shows the seven subgroups identified by the DT of model 1. Each group was specified by a combination of variables that identified a probability for incidence of T2D. For example, group 1 (low risk) consisted of men with an FPG<4.9 mmol/L and 2h-PCPG<7.7 mmol/L who had a 10% probability for incidence of T2D in the study period. Group 7 (high risk) consisted of men with an FPG>5.3 mmol/L and 2h-PCPG>4.4 mmol/L who had a 79% probability for incidence of T2D. The observed risk pattern in each subgroup revealed the interaction between a set of variables; that is, the patterns for group 2 show that in men with an FPG of 4.9–5.3 mmol/L and 2h-PCPG<7.7 mmol/L, risk of incidence depends on the value of WHtR. There was also an interaction between FPG, 2h-PCPG and age such that age >43 years increased the risk of T2D among men who had an FPG>5.3 mmol/L and 2h-PCPG≤4.4 mmol/L (groups 3 and 6). In model 2 (without 2h-PCPG), 9 subgroups were identified. The DT used four variables (FPG, WHtR, MAP and FHD) for classification (table 7). Results showed that FPG was the most important predictor for incidence of T2D; men with an FPG<4.9 mmol/L had a lower risk, but with FPG above 5.3 mmol/L, the risk of incidence depended on the WHtR and MAP.

DT analysis in women

The DT created for women is shown in figure 4. The model used three variables (FPG, 2h-PCPG and WHtR) for identification of seven subgroups (table 8). Group 1 (low risk) consisted of women who had an FPG≤5.2 mmol/L and WHtR≤0.55 (12% probability for incidence of T2D). Group 7 (high risk) consisted of women who had an FPG>5.2 mmol/L and WHtR>0.52 (81% probability for incidence of T2D). The observed patterns in the subgroups show that when the FPG level is <5.2 mmol/L, WHtR and 2h-PCPG are the most important factors in incidence of T2D, whereas if FPG is >5.2 mmol/L, WHtR is the most important factor. Some types of interactions were observed between FPG, WHtR and 2h-PCPG in women; for example, the patterns in groups 4 and 7 show that in women with an FPG>5.2 mmol/L, T2D incidence will increase by about 55% with an WHtR of over 0.52. In model 2 for women, in which we excluded 2h-PCPG from the variables list, nine subgroups were identified using three variables

Table 1 Baseline characteristics of men (TLGS 1999–2012)

Variables	Diabetes n=302	No diabetes n=2583	Total population n=2885	p Value*	Percentage of missing data†
Age (years)	48.5 (13.3)	41.1 (13.5)	41.8 (13.7)	<0.001	0
2-hour postchallenge plasma glucose(mmol/L)	7.3 (2.0)	5.5 (1.5)	5.7 (1.7)	<0.001	0
Fasting plasma glucose (mmol/L)	5.6 (0.6)	5.0 (0.5)	5.0 (0.5)	<0.001	0
Wrist circumference (cm)	18 (0.9)	17.6 (0.9)	17.6 (0.9)	<0.001	1.2%
Waist circumferences (cm)	95.1 (11.0)	88.3 (10.7)	89.0 (10.9)	<0.001	1.2%
BMI (kg/m^2)	27.9 (4.1)	25.5 (3.8)	25.8 (3.9)	<0.001	1.2%
Waist-to-hip ratio	0.95 (0.1)	0.91 (0.1)	0.90 (0.1)	<0.001	1.2%
Waist-to-height ratio	0.56 (0.1)	0.51 (0.1)	0.50 (0.1)	<0.001	1.2%
Triglyceride (mmol/L)	2.7 (2.1)	1.9 (1.3)	2.1 (1.4)	<0.001	0
Total cholesterol (mmol/L)	5.6 (1.1)	5.2 (1.1)	5.2 (1.1)	<0.001	0
HDL cholesterol (mmol/L)	0.9 (0.2)	0.9 (0.2)	0.9 (0.2)	0.196	0
Triglyceride-to-HDL ratio	5.1 (4.3)	7.4 (8.4)	5.3 (4.9)	<0.001	0
Cholesterol-to-HDL ratio	6.1 (1.9)	5.5 (1.6)	5.5 (1.7)	<0.001	0.1%
Estimated glomerular filtration rate (mL/min/ 1.73 m^2)	63.2 (11.2)	67.6 (11.0)	110.9 (67.1)	<0.001	0
Systolic blood pressure (mm Hg)	126.2 (17.5)	117.4 (15.7)	118.3 (16.1)	<0.001	1.4%
Diastolic blood pressure (mm Hg)	81.6 (11.6)	76.4 (10.3)	76.9 (10.5)	<0.001	1.4%
Pulse pressure (mm Hg)	44.6 (12.3)	40.9 (11.8)	41.3 (11.9)	<0.001	1.4%
Mean arterial blood pressure (mm Hg)	96.3 (12.4)	90.0 (10.9)	90.7 (11.3)	<0.001	1.5%
Total length of stay in the city (years)	40.7 (14.1)	34.3 (12.6)	35.0 (12.9)	<0.001	0.3%
Goitre size					
Grade 0	226 (74.8)	1872 (72.5)	2098 (72.7)	0.511	0
Grade 1	55 (18.2)	483 (18.7)	538 (18.6)		
Grade 2	21 (7.0)	228 (8.8)	249 (8.6)		
History of hospitalisation until now	165 (54.6)	1260 (48.7)	1425 (49.4)	0.059	0
Family history of diabetes in first-degree relatives	108 (36.1)	611 (23.9)	719 (25.1)	<0.001	0.9%
Family history of premature cardiovascular diseases in male relatives	24 (7.9)	212 (8.2)	236 (8.2)	0.876	0.9%
Family history of premature cardiovascular diseases in female relatives	34 (11.3)	179 (6.9)	213 (7.4)	0.006	0.9%
Former cigarette smoking	55 (18.2)	385 (14.9)	440 (15.2)	0.220	0
Current cigarette smoking	76 (25.1)	652 (25.2)	728 (25.2)	0.989	0
Exposed to secondhand smoke at home or at work	83 (27.5)	827 (32.0)	910 (31.5)	0.1	0
Education					0
Level 1 (illiterate)	13 (4.3)	80 (3.1)	93 (3.2)	<0.001	
Level 2 (≤5 years)	90 (29.8)	409 (15.8)	499 (17.3)		
Level 3 (6–12 years)	164 (54.3)	1573 (60.9)	1737 (60.2)		
Level 4 (13–16 years)	28 (9.3)	455 (17.6)	483 (16.7)		
Level 5 (>16 years)	7 (2.3)	66 (2.6)	73 (2.5)		
Marital status					0
Single	17 (5.6)	488 (18.9)	505 (17.5)	<0.001	
Married	282 (93.4)	2081 (80.6)	2363 (81.9)		
Divorced	0 (0)	8 (0.3)	8 (0.3)		
Widowed	3 (1.0)	6 (0.2)	9 (0.3)		
Physical activity levels					
Inactive‡	215 (74.7)	1744 (71.2)	1959 (71.5)	0.217	5%
Use of the ACE inhibitors	6 (1.9)	24 (0.9)	30 (1.1)	0.134	0
Use of blood lipid-lowering drugs	11 (3.6)	37 (1.4)	48 (1.7)	0.005	0
Use of antihypertensive drugs	21 (6.9)	76 (2.9)	97 (3.4)	<0.001	0
Use of aspirin	55 (19.0)	257 (11.2)	312 (11.1)	<0.001	2.6%
Participating in the lifestyle intervention group	100 (33.1)	821 (31.8)	921 (31.9)	0.777	0

Figures are either mean (SD) or N (%) for continuously and categorically distributed variables, respectively, in the data set with no missing values (after imputation of missing values). Data were collected from the TLGS study between 1999 and 2012.
*Continuous and categorical variables were compared by Student's t-test and χ^2, respectively.
†Per cent of missing data in the original data set.
‡Doing exercise or labour less than three times a week or performing activities achieving a lower than 600 MET.
BMI, body mass index; HDL, high-density lipoprotein; TLGS, Tehran Lipid and Glucose Study.

Table 2 Baseline characteristics of women (TLGS 1999–2012)

Variables	Diabetes n=427	No diabetes n=3335	Total population n=3762	p Value*	Percentage of missing data†
Age (years)	47.1 (11.7)	38.7 (12.1)	39.6 (12.3)	<0.001	0
2-hour postchallenge plasma glucose(mmol/L)	7.7 (1.8)	5.8 (1.4)	6.0 (1.5)	<0.001	0
Fasting plasma glucose (mmol/L)	5.5 (0.6)	4.9 (0.5)	4.9 (0.5)	<0.001	0
Wrist circumference (cm)	16.5 (1.0)	15.9 (0.99)	15.9 (1.0)	<0.001	2.5%
Waist circumferences (cm)	95.9 (11.2)	85.5 (11.6)	86.6 (12.0)	<0.001	3.8%
BMI (kg/m^2)	30.5 (4.9)	26.9 (4.5)	27.3 (4.7)	<0.001	3.2%
Waist-to-hip ratio	0.88 (0.1)	0.82 (0.1)	0.83 (0.1)	<0.001	3.8%
Waist-to-height ratio	0.61 (0.1)	0.54 (0.1)	0.5 (0.1)	<0.001	3.8%
Triglyceride (mmol/L)	2.3 (1.3)	1.6 (0.9)	1.7 (1.0)	<0.001	0
Total cholesterol (mmol/L)	5.8 (1.3)	5.3 (1.2)	5.3 (1.2)	<0.001	0
HDL cholesterol (mmol/L)	1.1 (0.3)	1.2 (0.3)	1.1 (0.3)	<0.001	0
Triglyceride-to-HDL ratio	5.2 (3.7)	3.5 (2.7)	3.7 (2.9)	<0.001	0.1%
Cholesterol-to-HDL ratio	5.5 (1.7)	4.7 (1.5)	4.8 (1.6)	<0.001	0.1%
Estimated glomerular filtration rate (mL/min/1.73 m^2)	59.4 (9.9)	63.8 (10.9)	63.3 (10.8)	<0.001	0.1%
Systolic blood pressure (mm Hg)	126.4 (19.8)	114.1 (16.4)	115.4 (17.2)	<0.001	1.2%
Diastolic blood pressure (mm Hg)	81.7 (10.5)	75.4 (10.1)	76.1 (10.3)	<0.001	1.2%
Pulse pressure (mm Hg)	44.6 (14.9)	38.6 (11.5)	39.2 (12.1)	<0.001	1.2%
Mean arterial blood pressure (mm Hg)	96.6 (12.6)	88.3 (11.3)	89.2 (11.7)	<0.001	1.2%
Total length of stay in the city (years)	47.1 (11.7)	38.7 (12.1)	33.1 (12.3)	<0.001	0.3%
Goiter size					0
Grade 0	342 (80.1)	2443 (73.3)	2785 (74)	0.009	
Grade 1	53 (12.4)	530 (15.9)	583 (15.5)		
Grade 2	394 (10.5)	362 (10.9)	32 (7.5)		
History of hospitalisation until now	373 (87.4)	2768 (83.0)	3141 (83.5)	0.022	0
Family history of diabetes in first-degree relatives	174 (40.7)	838 (25.1)	1012 (26.9)	<0.001	1%
Family history of premature cardiovascular diseases in male relatives	27 (6.4)	260 (7.9)	287 (7.7)	0.284	0
Family history of premature cardiovascular diseases in female relatives	55 (13.0)	298 (9.0)	353 (9.5)	0.008	0
Former cigarette smoking	6 (1.4)	43 (1.3)	49 (1.3)	0.747	3.8%
Current cigarette smoking	11 (2.6)	91 (2.7)	102 (2.7)	0.075	1.1%
Exposed to secondhand smoke at home or at work	81 (19.0)	699 (21.0)	780 (20.7)	0.340	0
Education					0
Level 1 (illiterate)	68 (15.9)	236 (7.1)	304 (8.1)	<0.001	
Level 2 (≤5 years)	168 (39.3)	717 (21.5)	885 (23.5)		
Level 3 (6–12 years)	159 (37.2)	1980 (59.4)	2139 (56.9)		
Level 4 (13–16 years)	30 (7.0)	385 (11.5)	415 (11.0)		
Level 5 (>16 years)	2 (0.5)	17 (0.5)	19 (0.5)		
Marital status					0
Single	11 (2.6)	353 (10.6)	364 (9.7)	<0.001	
Married	372 (87.1)	2777 (83.3)	3149 (83.7)		
Divorced	8 (1.9)	40 (1.2)	48 (1.3)		
Widowed	36 (8.4)	165 (4.9)	201 (5.3)		
Physical activity levels					6%
Inactive‡	311 (72.8)	2387 (71.6)	2698 (71.7)	0.608	
Use of the ACE inhibitors	12 (2.8)	29 (0.9)	41 (1.1)	<0.001	0
Use of blood lipid-lowering drugs	26 (6.1)	75 (2.3)	101 (2.7)	<0.001	0
Use of antihypertensive drugs	61 (14.3)	166 (5.0%)	227 (6.1)	<0.001	0
Use of aspirin	66 (15.5)	265 (7.9)	331 (8.8)	<0.001	2.5%
Participating in the lifestyle intervention group	183 (42.9)	1472 (44.1)	1655 (44.0)	0.036	0
Menstruation status					0
Normal menstrual cycle	235 (55.0)	2578 (77.3)	2813 (74.8)	<0.001	
Normal menopause	120 (28.1)	478 (14.3)	598 (15.9)		
Early menopause	72 (16.9)	279 (8.3)	351 (9.3)		

Continued

Table 2 Continued

Variables	Diabetes n=427	No diabetes n=3335	Total population n=3762	p Value*	Percentage of missing data†
Pregnancy prevention methods					0
Use of hormonal contraceptive drugs	19 (4.4)	243 (7.3)	262 (7.0)	<0.001	
IUDs	6 (1.4)	197 (5.9)	203 (5.4)		
Using condoms	10 (2.3)	181 (5.4)	191 (5.1)		
Withdrawal method/tubectomy/vasectomy	381 (89.2)	2361 (70.8)	2742 (72.9)		
Not applicable	11 (2.6)	353 (10.6)	364 (9.7)		
Having a birth history	388 (90.9)	2720 (81.6)	3108 (82.6)	<0.001	
Current status of pregnancy	2 (0.5)	48 (1.4)	50 (1.3)	<0.001	0
History of hypertension in pregnancy	35 (8.2)	189 (5.7)	224 (6.0)	<0.001	0
History of hyperglycaemia in pregnancy	9 (2.1)	19 (0.6)	28 (0.7)	<0.001	0

Figures are either mean (SD) or N (%) for continuously and categorically distributed variables, respectively, in the data set with no missing values (after imputation of missing values). Data were collected from the TLGS study between 1999 and 2012.
*Continuous and categorical variables were compared by Student's t-test and χ^2, respectively.
†Per cent of missing data in the original data set.
‡Doing exercise or labour less than three times a week or performing activities achieving a lower than 600 MET.
BMI, body mass index; HDL, high-density lipoprotein; IUDs, intrauterine devices; TLGS, Tehran Lipid and Glucose Study.

Table 3 Baseline characteristics of followed up and non-followed up men (TLGS 1999–2012)

Selected variables	Followed up (n=2885)	Non-followed up (n=1592)	p Value*
Age (year)	41.8 (13.7)†	43.4 (16.2)	0.001
2-hour postchallenge plasma glucose (mmol/L)	5.7 (1.7)	5.6 (1.7)	0.42
Fasting plasma glucose (mmol/L)	5.0 (0.5)	5.0 (0.5)	0.10
Wrist circumference (cm)	17.6 (0.9)	17.6 (0.9)	0.06
Waist-to-hip ratio	0.91 (0.07)	0.91 (0.07)	0.19
Waist-to-height ratio	0.52 (0.06)	0.52 (0.07)	0.24
Cholesterol-to-HDL ratio	5.5 (1.7)	5.4 (1.7)	0.014
Mean arterial blood pressure (mm Hg)	90.7 (11.3)	91.4 (12.7)	0.07
History of hospitalisation until now	1425 (49.4)	813 (51.1)	0.28
Family history of diabetes in first-degree relatives	724 (25.1)	360 (22.6)	0.07
Exposed to secondhand smoke at home or at work	910 (31.5)	461 (29.0)	0.07
Goitre size			
Grade 0	2085 (72.3)	1155 (72.6)	0.75
Grade 1	545 (18.9)	307 (19.3)	
Grade 2	255 (8.8)	128 (8.1)	
Use of aspirin during the last month	321 (11.1)	194 (12.1)	0.33
Education			
Level 1 (illiterate)	93 (3.2)	103 (6.5)	0.001
Level 2 (≤5 years)	498 (17.3)	314 (19.7)	
Level 3 (6-12 years)	1738 (60.2)	884 (55.5)	
Level 4 (13-16 years)	483 (16.7)	255 (16.0)	
Level 5 (>16 years)	73 (2.5)	36 (2.3)	
Participating in the lifestyle intervention group	1230 (42.6)	740 (46.5)	0.01

*Continuous and categorical variables were compared by Student's t-test and χ^2, respectively.
†Figures are either mean (SD) or N (%) for continuously and categorically distributed variables, respectively, in the data set with no missing values (after imputation of missing values). Data were collected from the TLGS study between 1999 and 2012.
HDL, high-density lipoprotein; TLGS, Tehran Lipid and Glucose Study.

(FPG, 2h-PCPG and MAP). This model had a lower sensitivity than model 1. Different interactions were found by this model; that is, when FPG is >5.2 mmol/L, WHtR≥0.56 is the most important risk factor for T2D, whereas when FPG is <5.2 mmol/L, WHtR and MAP play an important role in T2D incidence.

DISCUSSION

In this study, we used the three types of DT-based methods to provide insight into the factors that have an important role in the incidence of T2D and how these factors might interact to reveal specific subgroups. We used the more established and widely available algorithms to select the one with the best

Table 4 Baseline characteristics of followed up and non-followed up women (TLGS 1999–2012)

Selected variables	Followed up (n=3762)	Non-followed up (n=2071)	p Value*
Age (years)	39.6 (12.3)†	40.5 (15.1)	0.03
2-hour postchallenge plasma glucose (mmol/L)	6.0 (1.5)	6.0 (1.6)	0.87
Fasting plasma glucose (mmol/L)	4.9 (0.5)	4.9 (0.5)	0.16
Wrist circumference (cm)	15.9 (1.03)	15.9 (1.07)	0.56
BMI (kg/m^2)	27.4 (4.7)	27.4 (5.3)	0.84
Waist-to-hip ratio	0.8 (0.08)	0.8 (0.08)	0.06
Waist-to-height ratio	0.5 (0.08)	0.5 (0.08)	0.09
Cholesterol-to-HDL ratio	4.8 (1.6)	4.8 (1.6)	0.93
Triglyceride-to-HDL ratio	3.7 (2.9)	3.7 (3.7)	0.47
Pulse pressure (bpm)	39.3 (12.1)	40.2 (13.7)	0.01
Mean arterial blood pressure(mm Hg)	89.2 (11.8)	89.9 (12.7)	0.04
Glomerular filtration rate (mL/min/1.73 m^2)	63.3 (10.9)	63.2 (12.1)	0.9
Total length of stay in the city (years)	33.1 (13.2)	33.4 (15.5)	0.48
Goitre size			
Grade 0	2785 (74.0)	1551 (74.9)	0.74
Grade 1	583 (15.5)	316 (15.2)	
Grade 2	394 (10.5)	204 (9.8)	
Use of the ACE inhibitors	33 (0.9)	29 (1.4)	0.06
Current status of pregnancy	50 (1.3)	23 (1.1)	
Use of aspirin	331 (8.8)	178 (8.6)	
Educational level			
Level 1 (illiterate)	304 (8.1)	273 (13.2)	0.001
Level 2 (≤5 years)	885 (23.5)	456 (22.0)	
Level 3 (6–12 years)	2139 (56.9)	1099 (53.1)	
Level 4 (13–16 years)	415 (11.0)	236 (11.4)	
Level 5 (>16 years)	19 (0.5)	7 (0.3)	
Family history of premature cardiovascular diseases in male relatives	353 (9.4)	201 (9.7)	0.63
Family history of diabetes in first-degree relatives	1012 (26.9)	533 (25.8)	0.06

*Continuous and categorical variables were compared by Student's t-test and χ^2, respectively.
†Figures are either mean (SD) or N (%) for continuously and categorically distributed variables, respectively, in the data set with no missing values (after imputation of missing values). Data were collected from the TLGS study between 1999 and 2012.
BMI, body mass index; HDL, high-density lipoprotein; TLGS, Tehran Lipid and Glucose Study.

performance. Considering sensitivity and G-Mean, QUEST had the best performance in both men and women data sets. Although our study focused on exploration of interactions, DT models can be used for predicting the 9 years risk of developing T2D. Also, it is possible to identify who needs more or different treatments if we take interactions into account.

Two sets of variables were used for DT development. In model (1), we used selected variables which included 2h-PCPG, and in model (2), we excluded the 2h-PCPG from the variables list. Results of QUEST showed that although four similar predictors had the highest power both in men and women, they had different interaction patterns in the two genders; for instance, women with WHtR≤0.52 had a lower risk (26%) for T2D even with an FPG level of above 5.2 mmol/L. However, in men, the results showed that when FPG is >5.3 mmol/L, there was still 56% of T2D risk, even with a lowering of WHtR to below 0.45. A systematic review of existing evidences has shown that the mean of suggested cut-off values for WHtR in men and women, respectively, was 0.52 and 0.53 for incidence of T2D.[33] However, the results of this study showed that the recommended cut-off of 0.52 for WHtR is not a safe value for decreasing the risk of T2D among men,since significant risk of T2D was observed among men with WHtR≤0.45, as we pointed out above. Therefore, men with WHtR below 0.52 should not be given false assurances about their risk of incident T2D if their FPG level is >5.3 mmol/L.

A review of current studies shows that being aged >40 years is a risk factor fordeveloping T2D.[34]Theresults of our study show that age ≥43 years is a risk factor for men who have an FPG level >5.3 mmol/L. Results from this study confirm previous findings about the FPG cut-off point, obtained using traditional methods; additionally, we found the FPG cut-off point for men and women separately. For instance, two published studies of TLGS have shown that individuals with FPG levels <5.1 mmol/L are very unlikely to develop T2D during 6 and 9 years follow-up.[35 36] This study shows that among men with an FPG level <4.9 mmol/L, there is only 14% risk for T2D incidence within about 9 years. Another interesting finding of our study was the important role

Table 5 Performances of the decision tree models for men (Tehran Lipid and Glucose Study 1999–2012)

Types of decision tree models	Performance measures	C5.0	QUEST	CART
Models (1)	Sensitivity	72%	78%	74%
	Specificity	70%	72%	72%
	PPV	23%	25%	24%
	NPV	95%	96%	96%
	Accuracy	71%	73%	72%
	F-Measure	0.35	0.38	0.36
	G-Mean	0.71	0.75	0.73
	AUC	0.80	0.78	0.79
Models (2)	Sensitivity	67%	68%	64%
	Specificity	73%	78%	74%
	PPV	23%	27%	23%
	NPV	95%	95%	94%
	Accuracy	73%	77%	73%
	F-Measure	0.34	0.39	0.34
	G-Mean	0.70	0.73	0.69
	AUC	0.74	0.77	0.78

Model (1) was developed based on 15 variables which included 2h-PCPG.
Model (2) was developed based on 14 variables (2h-PCPG was excluded).
F-Measure: Harmonic mean between PPV and sensitivity, F-Measure=2 (sensitivity×PPV)/(sensitivity+PPV).
G-Mean of sensitivity and specificity, G=$\sqrt{}$sensitivity×specificity.
2h-PCPG, 2-hour postchallenge plasma glucose; AUC, area under the curve; CART, Classification and Regression Tree; G-Mean, geometric mean; NPV, negative predictive value; PPV, positive predictive value; QUEST, Quick Unbiased Efficient Statistical Tree.

Table 6 Performances of the decision tree models for women (Tehran Lipid and Glucose Study 1999–2012)

Types of decision tree models	Performance measures	C5.0	QUEST	CART
Models (1)	Sensitivity	75%	78%	70%
	Specificity	78%	78%	79%
	PPV	29%	30%	29%
	NPV	96%	97%	96%
	Accuracy	78%	78%	78%
	F-Measure	0.42	0.43	0.41
	G-Mean	0.76	0.78	0.74
	AUC	0.81	0.81	0.81
Models (2)	Sensitivity	73%	73%	66%
	Specificity	77%	78%	78%
	PPV	28%	29%	27%
	NPV	96%	96%	95%
	Accuracy	77%	78%	77%
	F-Measure	0.40	0.42	0.38
	G-Mean	0.75	0.75	0.72
	AUC	0.80	0.81	0.81

Model (1) was developed based on 20 variables which included 2h-PCPG.
Model (2) was developed based on 19 variables (2h-PCPG was excluded).
F-Measure: Harmonic mean between PPV and sensitivity, F-Measure=2 (sensitivity×PPV)/(sensitivity+PPV).
G-Mean of sensitivity and specificity, G=$\sqrt{}$sensitivity×specificity.
2h-PCPG, 2-hour postchallenge plasma glucose; AUC, area under the curve; CART, Classification and Regression Tree; G-Mean, geometric mean; NPV, negative predictive value; PPV, positive predictive value; QUEST, Quick Unbiased Efficient Statistical Tree.

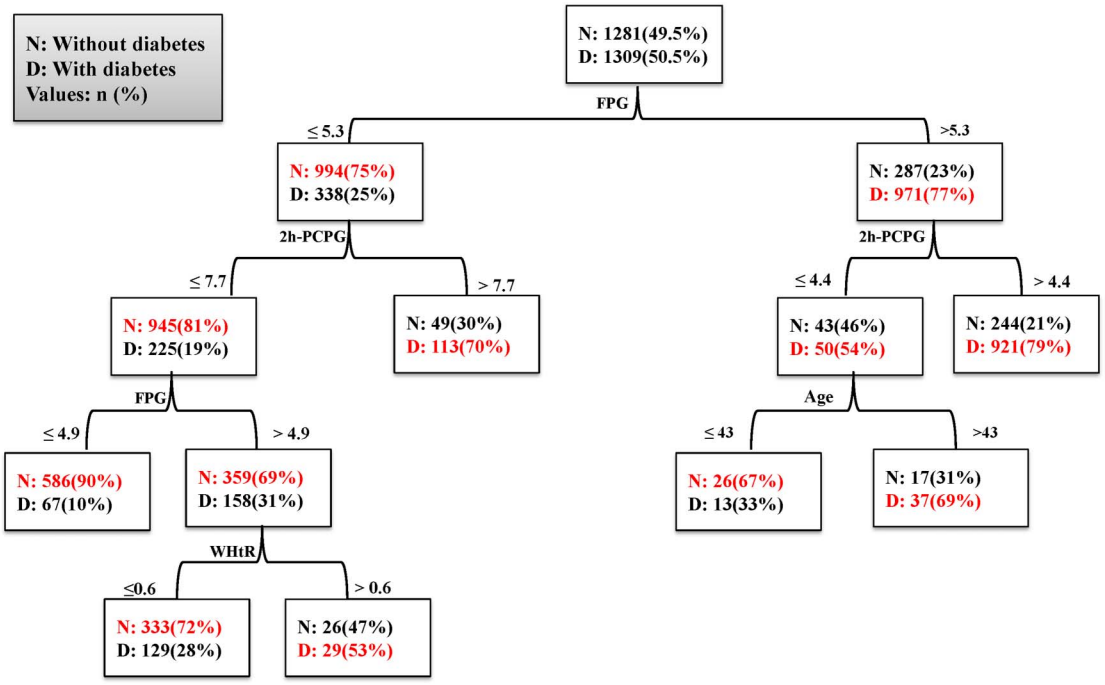

Figure 3 Decision tree for model 1 in men. Performance measures: sensitivity: 78%, specificity: 72%, G-Mean: 0.75, AUC: 0.78. 2h-PCPG, 2-hour postchallenge plasma glucose; AUC, area under the curve; FPG, fasting plasma glucose; G-Mean, geometric mean; WHtR, waist-to-height ratio.

Table 7 Groups identified by decision tree models for men (Tehran Lipid and Glucose Study 1999–2012)

Models	Groups	Definition (rules)	Probability*	Predicted class†
Model (1)	1	FPG≤4.9 and 2h-PCPG≤7.7	0.90	Non-diabetic
	2	4.9<FPG≤5.3 and 2h-PCPG≤7.7 and WHtR≤0.6	0.72	Non-diabetic
	3	FPG>5.3 and 2h-PCPG≤4.4 and age ≤43	0.67	Non-diabetic
	4	4.9<FPG≤5.3 and 2h-PCPG≤7.7 and WHtR>0.6	0.53	Diabetic
	5	FPG≤5.3 and 2h-PCPG>7.7	0.70	Diabetic
	6	FPG>5.3 and 2h-PCPG≤4.4 and age >43	0.68	Diabetic
	7	FPG>5.3 and 2h-PCPG>4.4	0.79	Diabetic
Model (2)	1	FPG≤4.9	0.86	Non-diabetic
	2	4.9<FPG≤5.3 and WHtR≤0.56	0.70	Non-diabetic
	3	4.9<FPG≤5.3 and WHtR>0.56 and FHD='no'	0.58	Non-diabetic
	4	FPG>5.3 and 0.4<WHtR≤0.49and MAP≤92	0.75	Non-diabetic
	5	4.9<FPG≤5.3 and WHtR>0.56 and FHD='yes'	0.78	Diabetic
	6	FPG>5.3 and WHtR≤0.45	0.56	Diabetic
	7	FPG>5.3 and 0.45<WHtR≤0.49 and MAP>92	0.67	Diabetic
	8	FPG>5.3 and≤0.49≤WHtR<0.56	0.74	Diabetic
	9	FPG>5.3 and WHtR>0.56	0.84	Diabetic

Model (1) was developed based on 15 variables which included 2h-PCPG.
Model (2) was developed based on 14 variables (2h-PCPG was excluded).
*The percentage of population in the defined subgroup, which can be interpreted as probability of an outcome.
†Predicted outcome for men who belong to the defined subgroup.
2h-PCPG, 2-hour postchallenge plasma glucose (mmol/L); FHD, family history of diabetes; FPG, fasting plasma glucose (mmol/L); MAP, mean arterial blood pressure (mm Hg); WHtR, waist-to-height ratio.

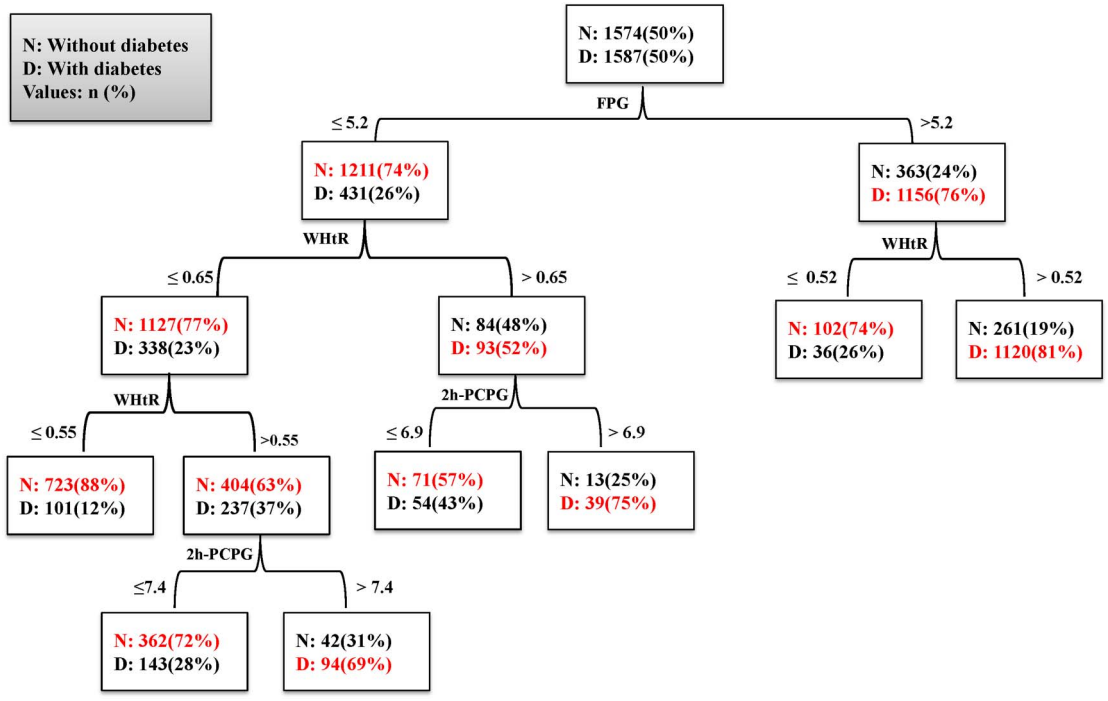

Figure 4 Decision tree for model 1 in women. Performance measures: sensitivity: 78%, specificity: 78%, G-Mean: 0.78, AUC: 0.81. 2h-PCPG, 2-hour postchallenge plasma glucose; AUC, area under the curve; FPG, fasting plasma glucose; G-Mean, geometric mean; WHtR, waist-to-height ratio.

of MAP in incidence of T2D in men and women. There are very few studies assessing the role of MAP in T2D incidence. Based on some previous studies, hypertension has been recognised as a risk factor for incident T2D in various populations.[37] The inter-related pathophysiology of hypertension and T2D is complex and not fully understood.[38] Our study showed that an MAP of ≥92 mm Hg is a risk factor among men with an FPG>5.3 mmol/L even if WHtR is <0.49. In women, an MAP of ≥97 mm Hg is a risk factor when WHtR is >0.66, even if the FPG level is ≤5.2 mmol/L. These results imply that the co-occurrence of a high level of MAP and

Table 8 Groups identified by decision tree models for women (Tehran Lipid and Glucose Study 1999–2012)

Models	Groups	Definition (rules)	Probability*	Predicted class†
Model (1)	1	FPG≤5.2 and WHtR≤0.55	0.88	Non-diabetic
	2	FPG≤5.2 and 0.55<WHtR≤0.66 and 2h-PCPG≤7.4	0.72	Non-diabetic
	3	FPG≤5.2 and WHtR>0.66 and 2h-PCPG≤6.9	0.57	Non-diabetic
	4	FPG>5.2 and WHtR≤0.52	0.74	Non-diabetic
	5	FPG≤5.2 and 0.55<WHtR≤0.66 and 2h-PCPG>7.4	0.69	Diabetic
	6	FPG≤5.2 and WHtR>0.66 and 2h-PCPG>6.9	0.75	Diabetic
	7	FPG>5.2 and WHtR>0.52	0.81	Diabetic
Model (2)	1	FPG≤5.2 and WHtR≤0.55	0.88	Non-diabetic
	2	FPG≤4.9 and 0.55<WHtR≤0.66	0.73	Non-diabetic
	3	4.9<FPG≤5.2 and 0.55<WHtR≤0.66 and MAP≤97	0.64	Non-diabetic
	4	FPG≤5.2 and WHtR>0.66 and MAP≤99	0.59	Non-diabetic
	5	FPG>5.2 and WHtR≤0.52	0.74	Non-diabetic
	6	4.9<FPG≤5.2 and 0.55<WHtR≤0.66 and MAP>97	0.67	Diabetic
	7	FPG≤5.2 and WHtR>0.66 and MAP>99	0.66	Diabetic
	8	FPG>5.2 and WHtR>0.52	0.81	Diabetic
	9	FPG>5.2 and WHtR>0.56	0.84	Diabetic

Model (1) was developed based on 20 variables which included 2h-PCPG.
Model (2) was developed based on 19 variables (2h-PCPG was excluded).
*The percentage of population in the defined subgroup, which can be interpreted as probability of an outcome.
†Predicted outcome for women who belong to the defined subgroup.
2h-PCPG, 2-hour postchallenge plasma glucose (mmol/L); FPG, fasting plasma glucose (mmol/L); MAP, mean arterial blood pressure (mm Hg); WHtR, waist-to-height ratio.

central obesity among women is a risk factor for T2D, whereas in men an increased level of FPG and MAP together is a risk factor for T2D. A simple point score system has recently been developed based on the TLGS database, including SBP, FHD, WHtR, TG/HDL-c and FPG as predictors;[39] continuous variables such as FPG and WHtR were, however, categorised into three or four groups. In other words, the cut-off points were predefined for prediction of T2D. In our study, DT algorithms generated optimal cut-off points for these variables as they relate to the best classification of participants with and without T2D.

Some strengths of this study include a large population-based sample. We used direct measurements of glucose value and anthropometric indices rather than self-reported information for both predictor variables and outcomes. Applying two variable selection methods with two evaluation criteria, missing data imputation and construction of DT models for both genders are other notable strengths. We have described the methodology in detail, allowing medical researchers to perform similar studies in different domains using DT methods.

The limitation of this study is the 35% loss to follow-up rate, although a number of authors have proposed a value of 50–80% as an acceptable level of follow-up rates.[40] In this study, we found statistically but not clinically important differences between the followed versus non-followed population in some baseline variables. The followed men had a higher value for the TC-to-HDL ratio, but lower age. In women, age, pulse pressure and MAP were lower for the followed population. Since these factors were associated with T2D, the results may be biased towards an underestimation of the association between these risk factors such as age and MAP and T2D. Additionally, we did not have data on dietary intake, which is an important factor in T2D studies. Finally, the models need to be validated on an independent population considering the ethnic and racial variations in T2D incidence.

CONCLUSIONS

DT analysis identified different interactions between predictor variables of T2D incidence in men and women. Sensitivity and G-Mean were measured on the validation data and showed acceptable performance of the DT models. Our results showed that WHtR and FPG were important risk factors in women and men, respectively.

Author affiliations
[1]Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Science, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran
[3]Department of Industrial Engineering, Birjand University of Technology, Birjand, Iran
[4]Department of Community Medicine, School of Medicine, Iran University of Medical Sciences, Tehran, Iran
[5]Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Acknowledgements The authors wish to acknowledge Ms Niloofar Shiva for critical editing of English grammar and syntax of the manuscript.

Contributors FA and FH designed the study protocol, and participated in the coordination and management of the study. AR performed the statistical analysis and wrote the manuscript. EH, JS and OP participated in the statistical analysis and interpretation of data. All authors read and approved the final manuscript.

Funding This study was supported by grant number 121 from the National Research Council of the Islamic Republic of Iran.

Disclaimer The funding source had no role in the design, in the collection, analysis and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

Competing interests None declared.

Patient consent Obtained.

Ethics approval This study was approved by the Ethical Committee of the Research Institute for Endocrine Sciences.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Whiting DR, Guariguata L, Weil C, *et al*. IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Res Clin Pract* 2011;94:311–21.
2. Hippisley-Cox J, Coupland C, Robson J, *et al*. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880.
3. Park KS. The search for genetic risk factors of type 2 diabetes mellitus. *Diabetes Metab J* 2011;35:12–22.
4. Noble D, Mathur R, Dent T, *et al*. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011;343:d7163.
5. Abbasi A, Peelen LM, Corpeleijn E, *et al*. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012;345:e5900.
6. Moons KG, Altman DG, Reitsma JB, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
7. Kleinbaum DG, Klein M. *Logistic regression: a self-learning text*. Springer Science & Business Media, 2010.
8. Han J, Kamber M, Pei J. *Data mining: concepts and techniques*. Elsevier, 2011.
9. Loh WY. Fifty years of classification and regression trees. *Int Stat Rev* 2014;82:329–48.
10. Van Buuren S. *Flexible imputation of missing data*. CRC Press, 2012.
11. Zhang H, Singer B. *Recursive partitioning and applications*. Springer Science & Business Media, 2010.
12. Epping-Jordan JE, Galea G, Tukuitonga C, *et al*. Preventing chronic diseases: taking stepwise action. *Lancet* 2005;366:1667–71.
13. Azizi F, Ghanbarian A, Momenan AA, *et al*. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials* 2009;10:5.
14. Azizi F, Rahmani M, Emami H, *et al*. Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1). *Soz Praventivmed* 2002;47:408–26.
15. Harati H, Hadaegh F, Saadat N, *et al*. Population-based incidence of type 2 diabetes and its associated risk factors: results from a six-year cohort study in Iran. *BMC Public Health* 2009;9:186.
16. Franklin SS, Gustin W, Wong ND, *et al*. Hemodynamic patterns of age-related changes in blood pressure: the Framingham Heart Study. *Circulation* 1997;96:308–15.
17. Gavin J, Alberti K, Davidson M, *et al*. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 1997;20:1183–97.
18. Enders CK. *Applied missing data analysis*. Guilford Press, 2010.
19. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2009.
20. Akthar F, Hahne C. RapidMiner 5 Operator Reference (2012). https://rapidminer.com/wp-content/uploads/ 2013/10 /RapidMiner_OperatorReference_en.pdf (accessed 12 Feb 2015).
21. Chawla NV, Lazarevic A, Hall LO, *et al*. SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač N, Gamberger D, Todorovski L, *et al*. *Knowledge discovery in databases: PKDD 2003*. Berlin: Springer, 2003:107–19.
22. Chawla N, Bowyer K, Hall L, *et al*. SMOTE: Synthetic Minority Over-Sampling Technique. *J Artif Intell Res* 2002;16:321–57.
23. López V, Fernández A, Moreno-Torres JG, *et al*. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Syst Appl* 2012;39:6585–608.
24. Ramezankhani A, Pournik O, Shahrabi J, *et al*. The Impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Med Decis Making* 2016;36:137–44.
25. John GH, Kohavi R, Pfleger K. eds. *Irrelevant features and the subset selection problem*. Machine Learning: Proceedings of the Eleventh International Conference. 1994.
26. Liu H, Motoda H. *Computational methods of feature selection*. CRC Press, 2007.
27. Liu H, Motoda H. *Feature selection for knowledge discovery and data mining*. Springer, 1998.
28. Rusch T, Zeileis A. To see the wood for the trees: discussion of "50 years of classification and regression trees". *Int Stat Rev* 2014;82:361–7.
29. Ture M, Tokatli F, Kurt I. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Syst Appl* 2009;36:2017–26.
30. Ville BD. *Decision tree for business intelligence and data mining*. SAS Publishing, 2006.
31. Bekkar M, Djemaa HK, Alitouche TA. Evaluation measures for models assessment over imbalanced data sets. *J Info Eng Appl* 2013;3:27–38.
32. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27:861–74.
33. Ashwell M, Gunn P, Gibson S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. *Obes Rev* 2012;13:275–86.
34. Stevens JW, Khunti K, Harvey R, *et al*. Preventing the progression to type 2 diabetes mellitus in adults at high risk: a systematic review and network meta-analysis of lifestyle, pharmacological and surgical interventions. *Diabetes Res Clin Pract* 2015;107:320–31.
35. Bozorgmanesh M, Hadaegh F, Saadat N, *et al*. Fasting glucose cutoff point: where does the risk terminate? Tehran lipid and glucose study. *Acta Diabetol* 2012;49:341–8.
36. Ramezankhani A, Pournik O, Shahrabi J, *et al*. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. *Diabetes Res Clin Pract* 2014;105:391–8.
37. Hatami M, Hadaegh F, Khalili D, *et al*. Family history of diabetes modifies the effect of blood pressure for incident diabetes in Middle Eastern women: Tehran Lipid and Glucose Study. *J Hum Hypertens* 2012;26:84–90.
38. Cooper-DeHoff RM, Egelund EF, Pepine CJ. Blood pressure lowering in patients with diabetes-one level might not fit all. *Nat Rev Cardiol* 2011;8:42–9.
39. Bozorgmanesh M, Hadaegh F, Ghaffari S, *et al*. A simple risk score effectively predicted type 2 diabetes in Iranian adult population: population-based cohort study. *Eur J Public Health* 2011;21:554–9.
40. Kristman V, Manno M, Côté P. Loss to follow-up in cohort studies: how much is too much? *Eur J Epidemiol* 2004;19:751–60.