Original Article

A clinical diagnostic model based on an eXtreme Gradient Boosting algorithm to distinguish type 1 diabetes

Xiaohan Tang[1,2,3], Rui Tang[4], Xingzhi Sun[4], Xiang Yan[1,2,3], Gan Huang[1,2,3], Houde Zhou[1,3,5], Guotong Xie[4], Xia Li[1,2,3], Zhiguang Zhou[1,2,3]

[1]Department of Metabolism and Endocrinology, the Second Xiangya Hospital, Central South University, Changsha, China; [2]Key Laboratory of Diabetes Immunology, Central South University, Ministry of Education, Changsha, China; [3]National Clinical Research Center for Metabolic Diseases, Changsha, China; [4]Department of Intelligent Clinical Decision Support, Ping An Healthcare Technology, Beijing, China; [5]Institute of Metabolism and Endocrinology, Hunan Key Laboratory for Metabolic Bone Diseases, Changsha, China

Contributions: (I) Conception and design: X Tang, X Li; (II) Administrative support: X Yan, G Huang, H Zhou, X Li, Z Zhou; (III) Provision of study materials or patients: X Yan, G Huang, H Zhou, X Li, Z Zhou; (IV) Collection and assembly of data: X Tang, X Yan, G Huang, H Zhou, X Li, Z Zhou; (V) Data analysis and interpretation: R Tang, X Sun, G Xie; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Zhiguang Zhou; Xia Li. Department of Metabolism and Endocrinology, The Second Xiangya Hospital, Central South University, No. 139 Renmin Middle Road, Furong District, Changsha 410011, China. Email: zhouzhiguang@csu.edu.cn; lixia@csu.edu.cn; Guotong Xie. Department of Intelligent Clinical Decision Support, Ping An Healthcare Technology, 9F Building B, Ping An International Financial Center, No. 1-3 Xinyuan South Road, Chaoyang District, Beijing 100027, China. Email: xieguotong@pingan.com.cn.

Background: Accurate classification of type 1 diabetes (T1DM) and type 2 diabetes (T2DM) in the early phase is crucial for individual precision treatment. This study aimed to develop a classification model having fewer and easier to access clinical variables to distinguish T1DM in newly diagnosed diabetes in adults.

Methods: Clinical and laboratory data were collected from 15,206 adults with newly diagnosed diabetes in this cross-sectional study. This cohort represented 20 provinces and 4 municipalities in China. Types of diabetes were determined based on postprandial C-peptide (PCP) level and glutamic acid decarboxylase autoantibody (GADA) titer. We developed multivariable clinical diagnostic models using the eXtreme Gradient Boosting (XGBoost) algorithm. Classification variables included in the final model were based on their scores of importance. Model performance was evaluated by area under the receiver operating characteristic curve (ROC AUC), sensitivity, and specificity. The performance of models with different variable combinations was compared. Calibration intercept and slope were evaluated for the final model.

Results: Among the newly diagnosed diabetes cohort, 1,465 (9.63%) persons had T1DM and 13,741 (90.37%) had T2DM. Body mass index (BMI) contributed the most to the model, followed by age of onset and hemoglobin A1c (HbA1c). Compared with models with other clinical variable combinations, a final model that integrated age of onset, BMI and HbA1c had relatively higher performance. The ROC AUC, sensitivity, and specificity for this model were 0.83 (95% CI, 0.80 to 0.85), 0.77, and 0.76, respectively. The calibration intercept and slope were 0.02 (95% CI, –0.03 to 0.06) and 0.90 (95% CI, 0.79 to 1.02), respectively, which suggested a good calibration performance.

Conclusions: Our classification model that integrated age of onset, BMI, and HbA1c could distinguish T1DM from T2DM, which provides a useful tool in assisting physicians in subtyping and precising treatment in diabetes.

Keywords: Type 1 diabetes (T1DM); type 2 diabetes (T2DM); diagnostic model; eXtreme Gradient Boosting algorithm (XGBoost algorithm)

Submitted Oct 26, 2020. Accepted for publication Jan 18, 2021.
doi: 10.21037/atm-20-7115
View this article at: http://dx.doi.org/10.21037/atm-20-7115

Introduction

Treatment of diabetes mellitus requires accurate discrimination of type 1 diabetes (T1DM) and type 2 diabetes (T2DM) (1). Diabetes classification systems have evolved in recent years. Updated classification criteria recommended in 2019 by the World Health Organization propose subtypes of hybrid diabetes and unclassified diabetes in addition to the well-known T1DM, T2DM, gestational diabetes, and other types of diabetes (2). Ahlqvist *et al.* stratified adult-onset diabetes into five subgroups that had different disease progression and risks of complications (3). Among these five clusters, patients in cluster 1 were denoted as having severe autoimmune diabetes (SAID); cluster 2 was labeled as severe insulin-deficient diabetes (SIDD) that required rapid insulin treatment. In addition to different typing strategies, key clues for the discrimination of T1DM from the majority of patients with T2DM are the presence of diabetes-associated autoantibodies and a deficiency of β cell function characterized by low plasma C-peptide level.

C-peptide is a useful method for assessing β cell function. C-peptides more than 0.3 nmol/L could differentiate insulin-requiring from non-insulin-requiring diabetes (4,5). Our group recently showed that glutamic acid decarboxylase autoantibody (GADA) titer is a valid risk predictor for progression of beta-cell failure in adult patients with autoimmune diabetes (6). About 70% of the patients in high GADA titer (more than 173.5 U/mL) group progressed to beta-cell function failure during the follow-up period. Therefore, C-peptides less than 0.3 nmol/L and GADA titer higher than 173.5 U/mL could be reliable evidence for T1DM identification.

The measurement of GADA and C-peptide in every individual with diabetes can be challenging and costly, especially for clinics in developing or undeveloped countries. Thus, efforts have been made to identify T1DM from T2DM from clinical parameters obtained in routine examinations. For instance, the UK Practical Classification Guidelines for Diabetes proposed that age of diagnosis (35 years as the cut off) and insulin treatment could be used to discriminate T1DM and T2DM (7). Fourlanos *et al.* developed a screening tool to identify autoimmune diabetes by integrating the clinical features of age of onset, acute symptoms, body mass index (BMI), and personal or family history of autoimmune disease (8).

Chinese patients with diabetes have somewhat different clinical characteristics compared with other populations. Most new cases of T1DM in China are adults, and Chinese individuals with T2DM are generally less obese than their Caucasian counterparts (9). However, there are no clinical diagnostic models of T1DM in Chinese patients with new-onset diabetes. Thus, in this study, we used the eXtreme Gradient Boosting (XGBoost) algorithm to generate several machine learning models according to the clinical features of participants. The purpose of the algorithm was to identify an optimal diagnostic model to distinguish T1DM from T2DM in adults newly diagnosed with diabetes. We present the following article in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting checklist (available at http://dx.doi.org/10.21037/atm-20-7115).

Methods

Study population

We analyzed data from 18,891 participants with newly diagnosed diabetes from a nationwide, multi-center, cross-sectional survey performed from April 2015 to October 2017. Forty-six tertiary care hospitals were invited from 20 provinces and 4 municipalities, across all 7 geographic regions of China (4 Northeast, 8 North, 3 Northwest, 9 Central, 3 Southwest, 7 South, and 12 East). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Ethics Committees of the Second Xiangya hospitals, Central South University in China approved this study (No. 2014032), and all patients provided written informed consents.

The following factors were inclusion criteria: (I) diagnosis of diabetes that met the World Health Organization 1999 criteria (10); (II) age 18 years and older; (III) diabetes duration less than 1 year; and (IV) outpatients attending clinics in the department of endocrinology. Individuals were excluded if pregnant at the time of diabetes diagnosis, if they had gestational diabetes mellitus, or if they had co-existing acute diseases such as infection or acute myocardial infarction that could affect glucose metabolism. In addition, we excluded 1,844 patients who lacked data on key variables for diabetes classification and 1,841 patients who did not meet the criteria for identifying types of diabetes (shown in model outcome part below). We analyzed the data of the remaining 15,206 patients for model construction (*Figure 1*).

Clinical measurements and data collection

Research nurses at each of the 46 participating hospitals

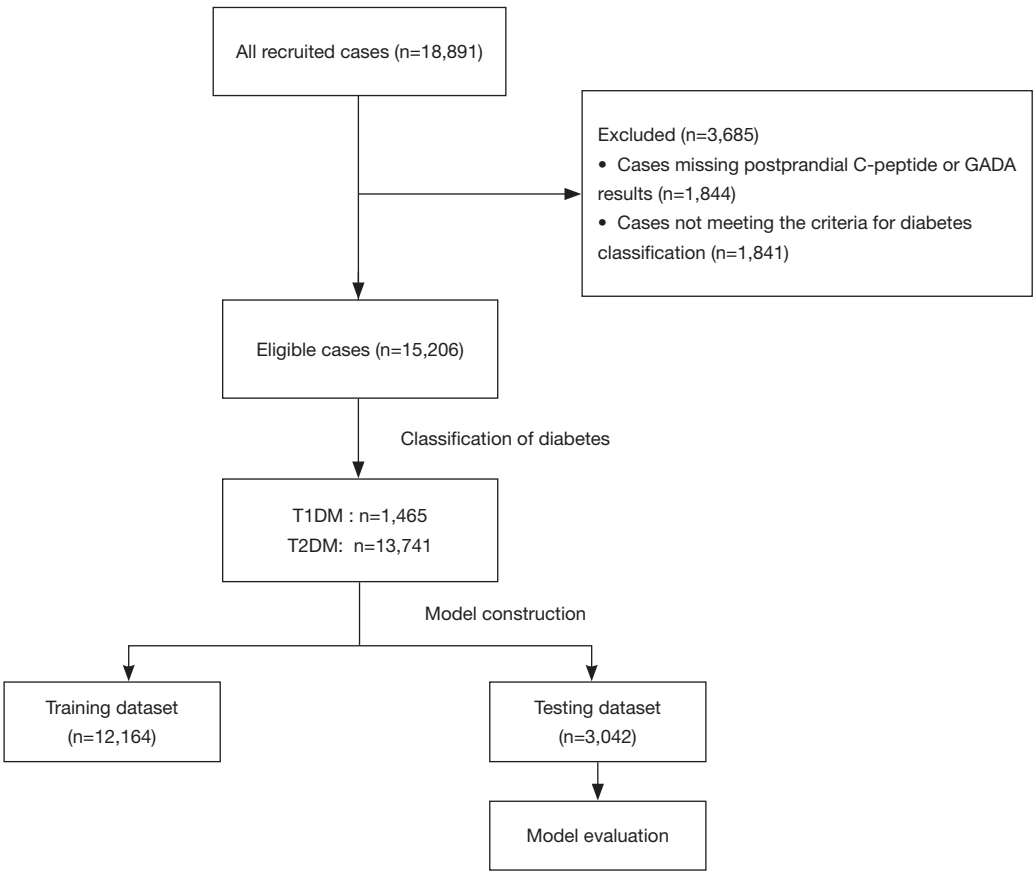

Figure 1 Flow diagram of the study design. GADA, glutamic acid decarboxylase autoantibody; T1DM, type 1 diabetes; T2DM, type 2 diabetes.

participated in a series of training programs to standardize all procedures and methods of data collection. Patients self-reported demographic characteristics (i.e., age and sex), clinical features, and lifestyle risk factors (i.e., exercise habits, diet, smoking, and alcohol consumption). The nurses used standard procedures to measure patient height, weight, waist circumference, hip circumference, and blood pressure.

Laboratory assays

Fasting plasma glucose (FPG), total cholesterol (TC), triglycerides (TGs), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), fasting C-peptide (FCP), and plasma hemoglobin A1c (HbA1c) were assayed by standard methods at the study sites. Postprandial blood samples were tested for 2-h postprandial plasma glucose (PPG) and C-peptide. Serum samples for GADA assays were shipped on ice within 1 day and stored at –80 ℃. The core laboratory (Central South University) performed serum GADA assays by a standardized radioligand assay as reported (11). The assay was assessed in the 2016 Islet Autoantibody Standardization Program (IASP 2016).

Model outcome: identification of T1DM and T2DM

T1DM was defined as postprandial C-peptide (PCP) less than 0.3 nmol/L or GADA titer no less than 173.5 U/mL. T2DM was defined as PCP more than 0.6 nmol/L and GADA titer less than 18.2 U/mL.

Statistical analysis

Continuous variables were expressed as mean [standard deviation (SD)] or median (interquartile range) based on evaluation of normal distribution; categorical variables

were given as number (percent). For analysis of continuous variables, *t*-test or Mann-Whitney test were performed to compare differences between groups where appropriate. Frequency differences were compared using Chi-square test. P<0.05 was considered statistically significant.

Model development using XGBoost

XGBoost is a machine learning method for classification problems (12). XGBoost produces classification models in the form of ensembles of weak classification models, typically decision trees. XGBoost can provide lower bias and better optimize an objective function, compared with traditional linear methods, e.g., logistic regression. We used XGBoost to build a multivariable clinical diagnostic model to identify T1DM and T2DM.

Importance score of variables

For each variable, XGBoost provides an importance score that represents the variable's contribution to predict the class label in the model. The larger the feature score of a variable, the more important is the variable to the model. In our study, we chose the method "gain", which is the average gain of the splits that use the variable, to compute the importance score of each variable.

First, we used all candidate variables to build the initial XGBoost model to assist in distinguishing T1DM and T2DM. Then, we obtained the importance score of each variable and ranked the variables according to those scores. We selected the most important variables from the candidate variables to build the final model intended for actual clinical practice.

Missing values

XGBoost can automatically accommodate missing values, i.e., by using a default direction for the missing values in each tree node (12). If a missing value appears in the validation dataset, it will be handled automatically by following the default direction that is decided in the training phase.

Explanation of classification results

To interpret the model output, we used the SHapley Additive exPlanations (SHAP) method to explain the XGBoost classification results (13). By using the SHAP method, we transformed the XGBoost model to the accumulative effects of all variable attributions on the output probability of diabetes type for each patient. In this manner, the impact of the variables on the outcome from the SHAP transformation can be interpreted easily. A SHAP value of a variable represents its impact on the model output.

Evaluation and comparisons of model performance

We split the dataset randomly into a training set (80%) and a testing set (20%). The training set was used to develop models. For validation, we used the testing set to assess the performance of the models. We computed area under the receiver operating characteristic curve (ROC AUC) of each model. We also computed model performance at different cut-offs for different models. The model performance metrics were sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Youden's index. We developed models with different combinations of variables and selected the final model by comparing the performance of these models. We also performed a 5-fold cross validation to further assess the validity of these models. Basically, we conducted the model development using 4 folds and the model validation using the left fold. Each fold was included in the training set 4 times and in the testing set 1 time.

In order to evaluate the agreement of observed outcomes and predictions of the final model, we discretized the prediction space into 10 bins and draw a calibration curve by plotting T1DM diagnostic predictions on the X-axis and observed frequency of T1DM on the Y-axis and presented the intercept and slope for this curve. An optimal calibration was represented with an intercept of 0 and a slope of 1.

Results

Characteristics of the study participants

Among the 15,206 patients with data for all key variables, 1,465 (9.63%) patients had T1DM and 13,741 (90.37%) had T2DM. The average age of disease onset was 43.7±15.1 and 51.0±12.8 years in the T1DM and T2DM group, respectively. As expected, patients with T1DM were significantly leaner, and they had lower blood pressure and better lipid metabolic parameters but higher serum FPG and HbA1c concentrations. Lifestyle factors liking drinking, diet treatment and physical activity were more frequent in patients with T2DM than in those with T1DM (*Table 1*).

Classification model construction for identifying T1DM and T2DM

An XGBoost classification model was used to identify

Table 1 Characteristics of patients diagnosed with T1DM and T2DM

Variables	T1DM		T2DM		P value
	N	Value	N	Value	
Age of onset (years)	1,465	43.7±15.1	13,741	51.0±12.8	<0.001
Sex					0.227
Female (n, %)	567	38.7%	5,542	40.3%	
Male (n, %)	898	61.30%	8,199	59.67%	
BMI (kg/m^2)	1,416	21.5±3.5	13,385	25.1±3.5	<0.001
FPG (mmol/L)	1,419	9.33 (6.60, 13.43)	13,603	7.87 (6.52, 10.31)	<0.001
PPG (mmol/L)	1,397	17.0±7.1	13,462	15.1±5.5	<0.001
HbA1c (%)	1,417	11.6 (9.0, 13.6)	13,491	8.6 (6.9, 10.9)	<0.001
SBP (mmHg)	1,415	120.5±15.6	13,139	128.2±16.2	<0.001
DBP (mmHg)	1,417	76.2±10.4	13,139	80.4±10.5	<0.001
Waist (cm)	1,320	80.6±10.3	12,478	89.1±10.5	<0.001
TG (mmol/L)	1,419	1.2 (0.8, 1.8)	13,262	1.77 (1.22, 2.75)	<0.001
TC (mmol/L)	1,422	4.6±1.4	13,366	4.8±1.3	<0.001
LDL-C (mmol/L)	1,421	2.7±1.0	13,327	2.9±1.0	<0.001
HDL-C (mmol/L)	1,413	1.2 (1.0, 1.5)	13,205	1.11 (0.94, 1.32)	<0.001
FCP (nmol/L)	1,456	0.14 (0.06, 0.24)	13,576	0.62 (0.43, 0.85)	<0.001
PCP (nmol/L)	1,465	0.22 (0.12, 0.34)	13,741	1.60 (1.09, 2.40)	<0.001
Current smoking (n, %)	448	31.0%	4,084	30.1%	0.485
Current drinking (n, %)	212	14.8%	2,447	18.1%	0.002
Diet treatment (n, %)	687	53.5%	6,240	62.3%	<0.001
Physical activity (n, %)	568	44.2%	5,357	53.5%	<0.001

% reported for all categorical variables. Data are presented as mean ± SD or median with upper and lower quartiles based on evaluation of normal distribution. Differences between T1DM and T2DM were compared using *t*-test or Mann-Whitney test for continuous variables where appropriate and chi-square test for categorical variables. T1DM, type 1 diabetes; T2DM, type 2 diabetes; BMI, body mass index; FPG, fasting plasma glucose; PPG, postprandial plasma glucose; HbA1c, hemoglobin A1c; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglyceride; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; FCP, fasting C-peptide; PCP, postprandial C-peptide.

T1DM and T2DM. We used 13 clinical features to distinguish patients. The features were ordered by the importance scores of each value in this model (*Figure 2*). BMI contributed the most to the model, followed by age of onset, HbA1c, waist dimension, TG concentration, and FPG.

We used SHAP values, which referred to the impact of one variable on model output, to explain the classification results. *Figure 3* presented how the contribution of an individual variable on the model output was affected by its value. The y position of the dot was the impact of each variable on the diagnosis of T1DM, which can be also interpreted as an increase (above 0) or decrease (below 0) of probability of being diagnosed as T1DM. From *Figure 3A*, the SHAP value decreased as the age of onset grew and the age of onset was about 40 years when SHAP value was 0, which indicated that an age of onset less than 40 increased the probability of having T1DM (*Figure 3A*). Other risk factors that increased the probability of being diagnosed with T1DM included a BMI lower than 23 kg/m^2

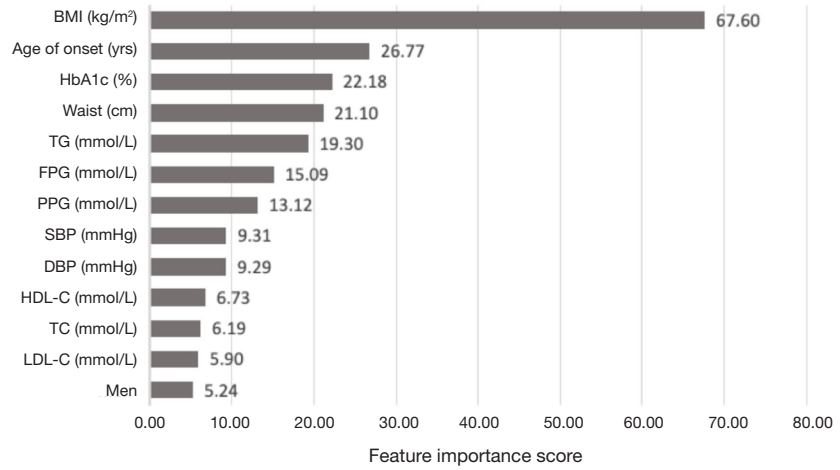

Figure 2 Feature importance scores of clinical variables. BMI, body mass index; HbA1c, hemoglobin A1c; TG, triglyceride; FPG, fasting plasma glucose; PPG, postprandial plasma glucose; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol.

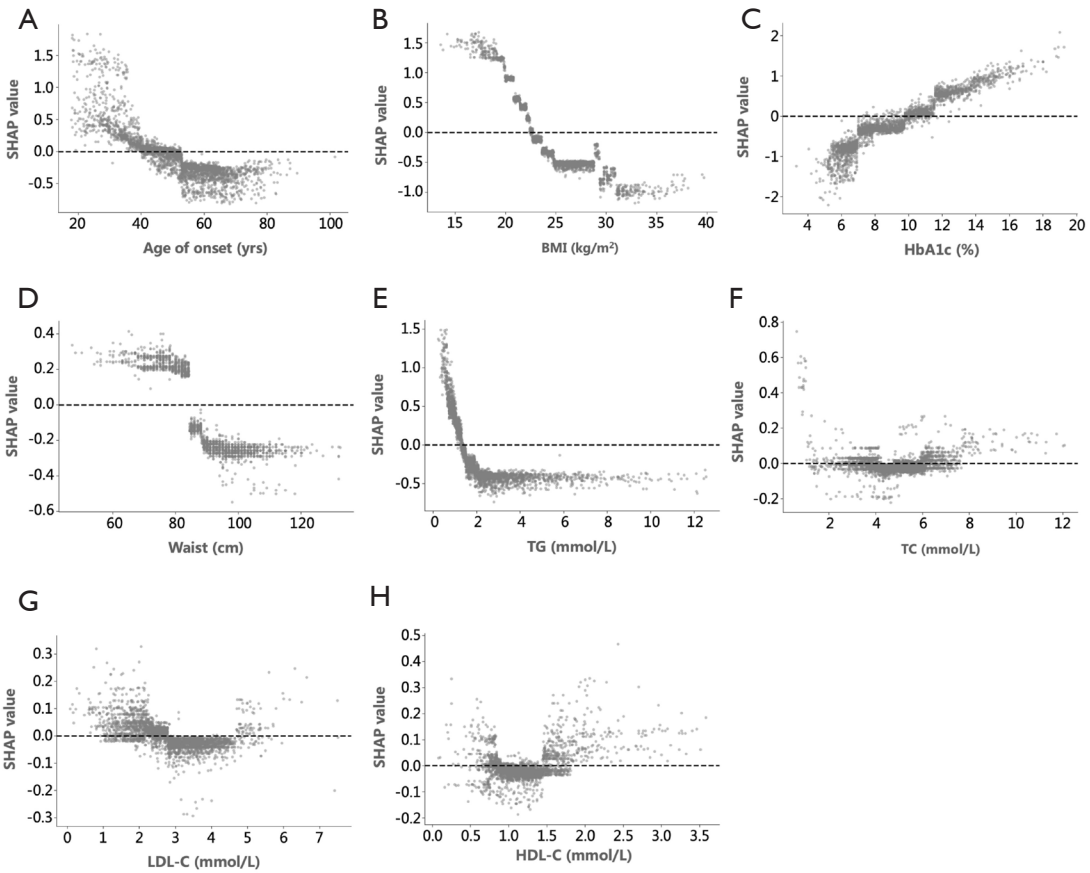

Figure 3 Impact of (A) age at onset, (B) BMI, (C) HbA1c, (D) waist, (E) TG, (F) TC, (G) LDL-C, (H) HDL-C on SHAP value for test dataset only. SHAP value represents the impact of each variable on the model output (diagnosis of T1DM in this model). SHAP, Shapley Additive exPlanations; BMI, body mass index; HbA1c, hemoglobin A1c; TG, triglyceride; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; T1DM, type 1 diabetes.

(*Figure 3B*), HbA1c higher than 7% (*Figure 3C*), and waist circumference lower than 85 cm (*Figure 3D*). Among four lipid parameters, only TG influenced the probability of being identified as T1DM. Patients with TG less than about 1.8 mmol/L (*Figure 3E*) had a higher probability of being identified as T1DM, although there was no obvious association between the level of TC, LDL-C, and HDL-C (*Figure 3F,G,H*) and the probability for identifying as T1DM.

Classification model evaluation for identifying T1DM and T2DM

To make the model concise and practical, we attempted to limit the number of input variables without significantly losing model performance. Because BMI and age of onset were easy to acquire and they ranked high by feature importance score, we included them in the final model. We set the input variables as the combination of BMI, age of onset, and one other variable, and built the candidate models accordingly. We also built models with all 13 variables and a model with only BMI and age of onset. We presented the corresponding cut-off, specificity, and sensitivity, when the Youden's index was the highest. Then, we compared the ROC AUC, sensitivity, specificity, PPV, NPV, and Youden's index for all the models. Finally, to further assess the validity of these models, a 5-fold cross validation was performed and the average ROC AUC, sensitivity, specificity, PPV, and NPV for each model were evaluated.

We ranked the model performance results according to their ROC AUC. The ROC AUC of the model with all 13 features was 0.86 (95% CI, 0.83 to 0.88) and the sensitivity and specificity of this model were both 0.78 (*Table 2*). The ROC AUC of the model that used only BMI and age of onset was 0.80 (95% CI, 0.77 to 0.83) and the sensitivity, specificity, PPV, and NPV were 0.61, 0.86, 0.32, and 0.95.

When we used only three variables to build models, the highest ROC AUC model was a combination of BMI, age of onset, and HbA1c. The ROC AUC, sensitivity, specificity, PPV, and NPV of this model were 0.83 (95% CI, 0.80 to 0.85), 0.77, 0.76, 0.25, and 0.97, respectively. This model was followed by a model composed of BMI, age at onset, and FPG, a model composed with BMI, age at onset, and TG, and a model with BMI, age of onset, and PPG; the ROC AUCs for all these models were from 0.81 (95% CI, 0.78 to 0.84) to 0.82 (95% CI, 0.78 to 0.84) (*Table 2*). All these models had similar discriminatory performances when cross validation was performed (Table S1).

Calibration of the final diagnostic model

We chose the model composed of age of onset, BMI, and HbA1c as the final model since it had the highest discriminatory performance among all models with less variables in *Table 2*.

Calibration evaluation was assessed for the final model in the testing dataset. A calibration curve for the final model was plotted and the intercept and slope for this curve were 0.02 (95% CI, –0.03 to 0.06) and 0.90 (95% CI, 0.79 to 1.02), respectively, suggesting that this model did well at identifying T1DM cases (*Figure 4*).

Online classification tool

Based on the final model composed of age of onset, BMI, and HbA1c, we designed an online classification tool that can compute the probability of T1DM or T2DM based on the input values (available at http://cdss.pingan.com:8082/diabetes/index.html).

Discussion

We established a diagnostic model of high performance to identify patients with newly diagnosed diabetes as likely to have T1DM.

The model was composed of age of onset, BMI, and HbA1c parameters, all easily accessed data. These parameters are associated with beta cell function and disease progression in T1DM. Early age at diagnosis of T1DM was associated with more rapid decline in beta-cell function (14,15). The association between BMI and T1DM progression is conflicting. Some studies showed that patients with higher BMI had better beta cell function at diagnosis and after follow-up (16-18). However, other studies showed that BMI might be a risk factor for developing T1DM and for accelerating T1DM progression (19-21). Better HbA1c levels were associated with higher C-peptide concentrations at diagnosis and could predict residual beta cell function and long-term diabetes control after follow-up (22-24).

Our study has multiple strengths. Previous studies on distinguishing T1DM from T2DM did not focus on patients with new-onset diabetes, whereas the participants in our study had short disease durations of less than 1 year (25-28). Patients with T1DM can obtain timely insulin treatment if they are diagnosed early and correctly.

Table 2 Performance of models with different combinations of variables

Features	ROC AUC (95% CI)	Cut-off (%)	Youden's index	Sensitivity	Specificity	PPV	NPV
All 13 features	0.86 (0.83, 0.88)	7	0.56	0.78	0.78	0.27	0.97
BMI + age of onset + HbA1c	0.83 (0.80, 0.85)	8	0.52	0.77	0.76	0.25	0.97
BMI + age of onset + FPG	0.82 (0.78, 0.84)	9	0.49	0.71	0.77	0.25	0.96
BMI + age of onset + TG	0.81 (0.78, 0.84)	8	0.48	0.73	0.75	0.23	0.96
BMI + age of onset + PPG	0.81 (0.78, 0.84)	10	0.49	0.69	0.80	0.27	0.96
BMI + age of onset + men	0.80 (0.77, 0.83)	14	0.48	0.60	0.88	0.34	0.95
BMI + age of onset + TC	0.80 (0.77, 0.83)	10	0.48	0.67	0.81	0.27	0.96
BMI + age of onset + HDL-C	0.80 (0.77, 0.83)	12	0.46	0.61	0.85	0.30	0.95
BMI + age of onset + DBP	0.80 (0.77, 0.83)	13	0.46	0.61	0.86	0.31	0.95
BMI + age of onset + LDL-C	0.80 (0.77, 0.83)	14	0.47	0.60	0.87	0.33	0.95
BMI + age of onset + SBP	0.80 (0.77, 0.83)	9	0.47	0.69	0.77	0.24	0.96
BMI + age of onset	0.80 (0.77, 0.83)	13	0.47	0.61	0.86	0.32	0.95
BMI + age of onset + waist	0.79 (0.76, 0.82)	10	0.46	0.67	0.79	0.25	0.96

ROC AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; BMI, body mass index; HbA1c, hemoglobin A1c; FPG, fasting plasma glucose; TG, triglyceride; PPG, postprandial plasma glucose; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; DBP, diastolic blood pressure; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure.

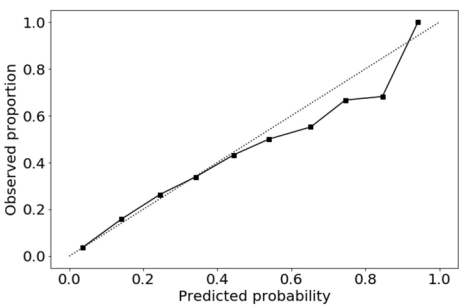

Figure 4 Calibration curve of the model composed of BMI, age of onset and HbA1c. BMI, body mass index; HbA1c, hemoglobin A1c.

Additionally, diagnosis criteria in our study were based on C-peptide and GADA concentrations. GADA present at early preclinical stages of T1DM and the titer of GADA are important for treatment decision and disease progression in patients with latent autoimmune diabetes in adults (LADA) (29-33). Patients with LADA are defined by adult age of onset, insulin independence for at least 6 months after diagnosis, and positivity for islet autoantibodies. They present a similar clinical characteristic to T2DM at the onset of disease and require insulin treatment in the early phase. In our study, GADA-positive patients with relatively good beta cell function were also included as T1DM; thus, our algorithm could also identify individuals with LADA.

Our study had limitations. We did not include patients with gestational diabetes mellitus or maturity-onset diabetes of the young. Patients with gestational diabetes are relatively easy to identify, and often they are diagnosed in an obstetrical department instead of an endocrinology department. Patients with maturity-onset diabetes usually do not have beta cell failure or GADA positivity; thus, there is a low possibility for misdiagnosis of these individuals as having T1DM. The possibility of being diagnosed with T1DM could be overestimated due to the suppression of beta cell function when beta cells are chronically exposed to hyperglycemia, an overlap C-peptides concentration with some subjects with T2DM (34). However, our goal is to decrease the misdiagnosis rate of T1DM by identifying cases who are likely to be diagnosed with T1DM and treat them with insulin timely. By providing beta cell repose in the early phase of the disease, insulin treatment may be beneficial for patients misdiagnosed with T1DM.

In conclusion, we show that a diagnostic model that integrates age of onset, BMI, and HbA1c could distinguish T1DM from T2DM among adult patients with newly

diagnosed diabetes, which provide a useful tool in assisting physicians in subtyping and precising treatment in diabetes.

Acknowledgments

We would like to thank all of the patients and investigators involved at the 46 participating centers of National Clinical Research Center for Metabolic Diseases. The authors thank AiMi Academic Services for English language editing and review services. The members of National Clinical Research Center for Metabolic Diseases (investigators and hospitals); Linong Ji, Xueyao Han, Ling Chen, Xiaoling Chen, Peking University People's Hospital; Lixin Guo, Xiaofan Jia, Shan Ding, Beijing Hospital; Xinhua Xiao, Cuijuan Qi, Xiaojing Wang, Peking Union Medical College Hospital; Zhongyan Shan, Yaxin Lai, Zhuo Zhang, the First Hospital of China Medical University; Yu Liu, Yan Cheng, Hanqing Cai, the Second Hospital of Jilin University; Yadong Sun, Yan Ma, Haiying Wang, People's Hospital of Jilin Province; Yiming Li, Chaoyun Zhang, Shuo Zhang, Hua Shan Hospital, Fudan University; Tao Yang, Hao Dai, Mei Zhang, the First Affiliated Hospital with Nanjing Medical University; Liyong Yang, Peiwen Wu, Xiaofang Yan, the First Affiliated Hospital of Fujian Medical University; Yangang Wang, Fang Wang, Hong Chen, the Affiliated Hospital of Qingdao University; Qifu Li, Rong Li, the First Affiliated Hospital of Chongqing Medical University; Qiuhe Ji, Li Wang, Xiangyang Liu, Xijing Hospital, Fourth Military Medical University; Jing Liu, Suhong Wei, Gansu Provincial Hospital; Yun Zhu, Rui Ma, the First Affiliated Hospital of Xinjiang Medical University; Gebo Wen, Xinhua Xiao, Jianping Qin, the First Affiliated Hospital of University of South China; Jian Kuang, Yan Lin, Guangdong General Hospital; Shaoda Lin, Kun Lin, the First Affiliated Hospital of Shantou University Medical College; Xiaohong Niu, Li Li, Heji Hospital Affiliated to Changzhi Medical College; Shuoming Luo, the Second Xiangya Hospital of Central South University; Huibiao Quan, Leweihua Lin, Hainan General Hospital; Hongyu Kuang, Weihua Wu, the First Affiliated Hospital of Harbin Medical University; Yuling He, the First Affiliated Hospital of Guangxi Medical University; Xiaoyan Chen, Yuyu Tan, the First Affiliated Hospital of Guangzhou Medical University; Ling He, Guangzhou First People's Hospital; Chao Zheng, the Second Affiliated Hospital of Wenzhou Medical University; Jianying Liu, Zhifang Yang, the First Affiliated Hospital of Nanchang University; Xiaoyang Lai, the Second Affiliated Hospital of Nanchang University; Ling Hu, Yan Zhu, Ying Hu, the Third Affiliated Hospital of Nanchang University; Xuqing Li, Henan Provincial People's Hospital; Hong Li, Yushan Xu, the First Affiliated Hospital of Kunming Medical University; Heng Su, Yang Ou, the First People's Hospital of Yunnan Province; Jianping Wang, the Second Hospital University of South China; Changqing Luo, Xiaoyue Wang, the First People's Hospital of Yueyang; Zhiming Deng, Shenglian Gan, the First People's Hospital of Changde City; Zhaohui Mo, Ping Jin, Honghui He, the Third Xiangya Hospital of Central South University; Qiuxia Huang, Dongguan People's Hospital; Fang Wang, Heping Hospital Affiliated to Changzhi Medical College; Yi Zhang, Zhenzhen Hong, First Hospital of Quanzhou Affiliated to Fujian Medical University; Yuezhong Ren, Pengfei Shan, the Second Affiliated Hospital of Zhejiang University School of Medicine; Caifeng Yan, Hui Zhang, Northern Jiangsu People's Hospital; Zhiwen Liu, Shanghai Xuhui District Central Hospital; Meibiao Zhang, the First People's Hospital of Huaihua; Ming Liu, Heting Wang, Tianjin Medical University General Hospital; Hongwei Jiang, Liujun Fu, the First Affiliated Hospital of the Henan University of Science and Technology; Hui Fang, Tangshan Gongren Hospital; Hui Sun, the Affiliated Hospital of Inner Mongolia Medical University.

Funding: This work was supported by the National Science and Technology Infrastructure Program (2013BAI09B12, 2015BAI12B13) and the National Key R&D Program of China (2016YFC1305000, 2017YFC1309604).

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-7115

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-7115

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-7115). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki

(as revised in 2013). The study was approved by Ethics Committees of the Second Xiangya hospitals, Central South University in China (No. 2014032), and informed consent was taken from all individual participants.

References

1. Leslie RD, Pozzilli P. Type I diabetes masquerading as type II diabetes. Possible implications for prevention and treatment. Diabetes Care 1994;17:1214-9.
2. World Health Organization. Classification of diabetes mellitus. 2019. Available online: https://www.who.int/publications/i/item/classification-of-diabetes-mellitus
3. Ahlqvist E, Storm P, Karajamaki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 2018;6:361-9.
4. Buzzetti R, Tuomi T, Mauricio D, et al. Management of latent autoimmune diabetes in adults: a consensus statement from an international expert panel. Diabetes 2020;69:2037-47.
5. Jones AG, Hattersley AT. The clinical utility of C-peptide measurement in the care of patients with diabetes. Diabet Med 2013;30:803-17.
6. Li X, Chen Y, Xie Y, et al. Decline pattern of beta-cell function in adult-onset latent autoimmune diabetes: an 8-year prospective study. J Clin Endocrinol Metab 2020;105:dgaa205.
7. Royal College of General Practitioners. Coding, classification and diagnosis of diabetes A review of the coding, classification and diagnosis of diabetes in primary care in England with recommendations for improvement. 2011.
8. Fourlanos S, Perry C, Stein MS, et al. A clinical screening tool identifies autoimmune diabetes in adults. Diabetes Care 2006;29:970-5.
9. Weng J, Zhou Z, Guo L, et al. Incidence of type 1 diabetes in China, 2010-13: population based study. BMJ 2018;360:j5295.
10. WHO/ADA. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 1997;20:1183-97.
11. Zhou Z, Xiang Y, Ji L, et al. Frequency, immunogenetics, and clinical characteristics of latent autoimmune diabetes in China (LADA China study): a nationwide, multicenter, clinic-based cross-sectional study. Diabetes 2013;62:543-50.
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016:785-94.
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017:4765-74.
14. Barker A, Lauria A, Schloot N, et al. Age-dependent decline of beta-cell function in type 1 diabetes after diagnosis: a multi-centre longitudinal study. Diabetes Obes Metab 2014;16:262-7.
15. Mortensen HB, Swift PG, Holl RW, et al. Multinational study in children and adolescents with newly diagnosed type 1 diabetes: association of age, ketoacidosis, HLA status, and autoantibodies on residual beta-cell function and glycemic control 12 months after diagnosis. Pediatr Diabetes 2010;11:218-26.
16. Redondo MJ, Rodriguez LM, Escalante M, et al. Beta cell function and BMI in ethnically diverse children with newly diagnosed autoimmune type 1 diabetes. Pediatr Diabetes 2012;13:564-71.
17. Yu HW, Lee YJ, Cho WI, et al. Preserved C-peptide levels in overweight or obese compared with underweight children upon diagnosis of type 1 diabetes mellitus. Ann Pediatr Endocrinol Metab 2015;20:92-7.
18. Gong S, Wu C, Zhong T, et al. Complicated curve association of body weight at diagnosis with C-peptide in children and adults with new-onset type 1 diabetes. Diabetes Metab Res Rev 2020;36:e3285.
19. Hyppönen E, Virtanen SM, Kenward MG, et al. Obesity, increased linear growth, and risk of type 1 diabetes in children. Diabetes Care 2000;23:1755-60.
20. Lauria A, Barker A, Schloot N, et al. BMI is an important driver of beta-cell loss in type 1 diabetes upon diagnosis in 10 to 18-year-old children. Eur J Endocrinol 2015;172:107-13.
21. Knerr I, Wolf J, Reinehr T, et al. The 'accelerator hypothesis': relationship between weight, height, body mass index and age at diagnosis in a large cohort of 9,248

German and Austrian children with type 1 diabetes mellitus. Diabetologia 2005;48:2501-4.
22. Szypowska A, Groele L, Wysocka-Mincewicz M, et al. Factors associated with preservation of C-peptide levels at the diagnosis of type 1 diabetes. J Diabetes Complications 2018;32:570-4.
23. Grönberg A, Espes D, Carlsson PO. Better HbA1c during the first years after diagnosis of type 1 diabetes is associated with residual C peptide 10 years later. BMJ Open Diabetes Res Care 2020;8:e000819.
24. Wolnik B, Orlowska-Kunikowska E, Blaszkowska M, et al. The phenomenon of HbA1c stability and the risk of hypoglycemia in long-standing type 1 diabetes. Diabetes Res Clin Pract 2019;152:96-102.
25. Chi GC, Li X, Tartof SY, et al. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. BMJ Open Diabetes Res Care 2019;7:e000547.
26. Klompas M, Eggleston E, McVetta J, et al. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 2013;36:914-21.
27. Lo-Ciganic W, Zgibor JC, Ruppert K, et al. Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. J Diabetes Sci Technol 2011;5:486-93.
28. Lynam A, McDonald T, Hill A, et al. Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18-50 years. BMJ Open 2019;9:e031586.
29. Taplin CE, Barker JM. Autoantibodies in type 1 diabetes. Autoimmunity 2008;41:11-8.
30. Brooks-Worrell B, Gersuk VH, Greenbaum C, et al. Intermolecular antigen spreading occurs during the preclinical period of human type 1 diabetes. J Immunol 2001;166:5265-70.
31. Zampetti S, Campagna G, Tiberti C, et al. High GADA titer increases the risk of insulin requirement in LADA patients: a 7-year follow-up (NIRAD study 7). Eur J Endocrinol 2014;171:697-704.
32. Kasuga A, Maruyama T, Nakamoto S, et al. High-titer autoantibodies against glutamic acid decarboxylase plus autoantibodies against insulin and IA-2 predicts insulin requirement in adult diabetic patients. J Autoimmun 1999;12:131-5.
33. Li X, Liu L, Xiang Y, et al. Response to comment on Liu et al. Latent autoimmune diabetes in adults with low-titer GAD antibodies: similar disease progression with type 2 diabetes: a nationwide, multicenter prospective study (LADA China Study 3). Diabetes Care 2015;38:16-21. Diabetes Care 2015;38:e44.
34. Kaneto H. Pancreatic beta-cell glucose toxicity in type 2 diabetes mellitus. Curr Diabetes Rev 2015;11:2-6.

Cite this article as: Tang X, Tang R, Sun X, Yan X, Huang G, Zhou H, Xie G, Li X, Zhou Z. A clinical diagnostic model based on an eXtreme Gradient Boosting algorithm to distinguish type 1 diabetes. Ann Transl Med 2021;9(5):409. doi: 10.21037/atm-20-7115

Supplementary

Table S1 Performance of models with different combinations of variables using a 5-fold cross validation

Features	ROC AUC (95% CI)	Youden's index	Sensitivity	Specificity	PPV	NPV
All 13 features	0.86 (0.85, 0.87)	0.57	0.78	0.79	0.29	0.97
BMI + age of onset + HbA1c	0.83 (0.83, 0.83)	0.50	0.76	0.75	0.24	0.97
BMI + age of onset + FPG	0.81 (0.81, 0.82)	0.47	0.74	0.73	0.23	0.96
BMI + age of onset + TG	0.81 (0.79, 0.82)	0.47	0.73	0.74	0.23	0.96
BMI + age of onset + PPG	0.81 (0.80, 0.81)	0.46	0.73	0.73	0.22	0.96
BMI + age of onset + men	0.79 (0.78, 0.80)	0.43	0.71	0.72	0.21	0.96
BMI + age of onset + TC	0.79 (0.78, 0.80)	0.43	0.70	0.73	0.22	0.96
BMI + age of onset + HDL-C	0.79 (0.78, 0.80)	0.44	0.70	0.73	0.22	0.96
BMI + age of onset + DBP	0.79 (0.79, 0.80)	0.44	0.70	0.74	0.22	0.96
BMI + age of onset + LDL-C	0.79 (0.78, 0.80)	0.43	0.71	0.73	0.22	0.96
BMI + age of onset + SBP	0.80 (0.79, 0.80)	0.45	0.71	0.74	0.23	0.96
BMI + age of onset	0.79 (0.79, 0.80)	0.44	0.70	0.73	0.22	0.96
BMI + age of onset + waist	0.79 (0.79, 0.79)	0.44	0.71	0.73	0.22	0.96

Cut-off value was set to 0.08 for these models. ROC AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; BMI, body mass index; HbA1c, hemoglobin A1c; FPG, fasting plasma glucose; TG, triglyceride; PPG, postprandial plasma glucose; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; DBP, diastolic blood pressure; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure.

http://dx.doi.org/10.21037/atm-20-7115