



# Machine Learning for Predicting the 3-Year Risk of Incident Diabetes in Chinese Adults

Yang Wu<sup>1,2,3†</sup>, Haofei Hu<sup>3,4,5†</sup>, Jinlin Cai<sup>1,2,6</sup>, Runtian Chen<sup>1,2,3</sup>, Xin Zuo<sup>7</sup>, Heng Cheng<sup>7</sup> and Dewen Yan<sup>1,2,3\*</sup>

<sup>1</sup> Department of Endocrinology, The First Affiliated Hospital of Shenzhen University, Shenzhen, China, <sup>2</sup> Department of Endocrinology, Shenzhen Second People's Hospital, Shenzhen, China, <sup>3</sup> Shenzhen University Health Science Center, Shenzhen, China, <sup>4</sup> Department of Nephrology, The First Affiliated Hospital of Shenzhen University, Shenzhen, China, <sup>5</sup> Department of Nephrology, Shenzhen Second People's Hospital, Shenzhen, China, <sup>6</sup> Shantou University Medical College, Shantou, China, <sup>7</sup> Department of Endocrinology, The Third People's Hospital of Shenzhen, Shenzhen, China

**Purpose:** We aimed to establish and validate a risk assessment system that combines demographic and clinical variables to predict the 3-year risk of incident diabetes in Chinese adults.

**Methods:** A 3-year cohort study was performed on 15,928 Chinese adults without diabetes at baseline. All participants were randomly divided into a training set ( $n = 7,940$ ) and a validation set ( $n = 7,988$ ). XGBoost method is an effective machine learning technique used to select the most important variables from candidate variables. And we further established a stepwise model based on the predictors chosen by the XGBoost model. The area under the receiver operating characteristic curve (AUC), decision curve and calibration analysis were used to assess discrimination, clinical use and calibration of the model, respectively. The external validation was performed on a cohort of 11,113 Japanese participants.

**Result:** In the training and validation sets, 148 and 145 incident diabetes cases occurred. XGBoost methods selected the 10 most important variables from 15 candidate variables. Fasting plasma glucose (FPG), body mass index (BMI) and age were the top 3 important variables. And we further established a stepwise model and a prediction nomogram. The AUCs of the stepwise model were 0.933 and 0.910 in the training and validation sets, respectively. The Hosmer-Lemeshow test showed a perfect fit between the predicted diabetes risk and the observed diabetes risk ( $p = 0.068$  for the training set,  $p = 0.165$  for the validation set). Decision curve analysis presented the clinical use of the stepwise model and there was a wide range of alternative threshold probability spectrum. And there were almost no the interactions between these predictors (most  $P$ -values for interaction  $>0.05$ ). Furthermore, the AUC for the external validation set was 0.830, and the Hosmer-Lemeshow test for the external validation set showed no statistically significant difference between the predicted diabetes risk and observed diabetes risk ( $P = 0.824$ ).

**Conclusion:** We established and validated a risk assessment system for characterizing the 3-year risk of incident diabetes.

**Keywords:** machine learning, extreme gradient boosting, simple stepwise model, Incident diabetes, risk

## OPEN ACCESS

### Edited by:

Yongcheng He,  
Shenzhen Hengsheng Hospital, China

### Reviewed by:

Chi Chen,  
Guizhou University of Traditional  
Chinese Medicine, China  
Huiqiong Zeng,  
Shenzhen Futian Hospital for  
Rheumatic Diseases, China

### \*Correspondence:

Dewen Yan  
ydeweny@163.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Clinical Diabetes,  
a section of the journal  
Frontiers in Public Health

**Received:** 16 November 2020

**Accepted:** 21 May 2021

**Published:** 29 June 2021

### Citation:

Wu Y, Hu H, Cai J, Chen R, Zuo X,  
Cheng H and Yan D (2021) Machine  
Learning for Predicting the 3-Year Risk  
of Incident Diabetes in Chinese Adults.  
Front. Public Health 9:626331.  
doi: 10.3389/fpubh.2021.626331

## HIGHLIGHTS

- The eXtreme Gradient Boosting system was an effective machine learning technique.
- We establish a risk assessment system for characterizing the 3-year risk of diabetes.
- The external validation showed that our findings were well-generalized.
- Our findings are helpful for identifying individuals at high risk for diabetes.

## INTRODUCTION

The epidemic of diabetes has become a major public health threat across the world. The International Diabetes Federation (IDF) estimated that 451 million adults were suffering from diabetes mellitus worldwide in 2017 and the figure was expected to increase to 693 million by 2045 (1). The prevalence of diabetes among Chinese adults increased from 9.7% in 2007 and to 11.2% in 2017 (2). Diabetes is a debilitating chronic disease with potentially various microvascular and macrovascular complications, such as diabetic kidney disease, diabetic retinopathy, diabetic neuropathy, cardiovascular, and cerebrovascular disease (3–7). Diabetes and its complications have contributed tremendously to the burden of social, financial, and health systems worldwide.

Although diabetes is an irreversible disease, it is largely preventable. Early screening and diagnosis are at the core of effectively preventing diabetes and delaying its progression. Several studies revealed lifestyle modification and pharmacological intervention could reduce the risk of developing diabetes (8, 9). Moreover, for newly diagnosed diabetic patients, intensive lifestyle intervention, early short-term intensive insulin therapy and metabolic surgery can induce long-term glycemic remission without further antidiabetic medication (10–12). Therefore, it is essential to identify individuals at high risk of developing diabetes for diabetes prevention programs.

Machine learning has increasingly been utilized to establish risk prediction models in the field of medicine (13–15). Machine-learning algorithms can be defined as searching through a large number of candidate programs under the guidance of training experience to find a program that optimizes the performance metric (16). Compared with traditional statistical methods, it is mainly applied to iteratively learn the non-linear interactions from a mass of data through computer algorithms (17). Several studies showed that machine learning methods could describe an individual's characteristics and identify individuals at high risk of

diabetes (18–21). A gradient tree boosting method implemented in the eXtreme Gradient Boosting (XGBoost) system is an effective machine learning method that can assemble weak prediction models to establish a more reliable prediction model (22–26). So far, there is no research using the XGBoost method to build diabetes risk prediction models. Therefore, we sought to use the XGBoost method to select the most important variables from candidate variables and further establish and validate a risk assessment system that combines demographic and clinical variables using real-world data from a large cohort of Chinese adults across 32 sites and 11 cities between 2010 and 2016 to predict the 3-year risk of incident diabetes in Chinese adults.

## MATERIALS AND METHODS

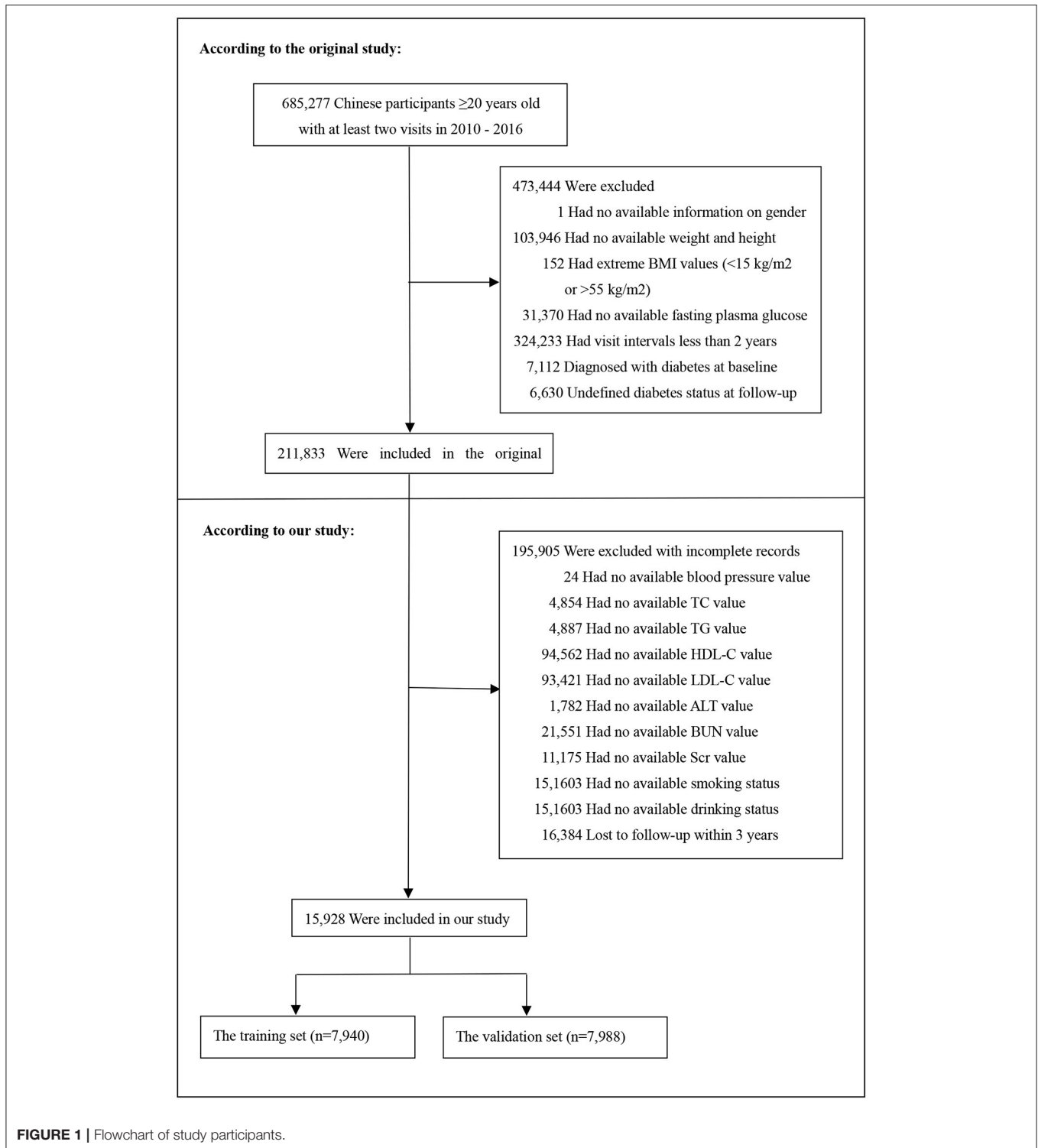
### Study Design and Participants

The data was downloaded from the “DATADRYAD” database ([www.Datadryad.org](http://www.Datadryad.org)), a non-profit computerized database established in China by the Rich Healthcare Group. Its data is available publicly for use. The raw data was provided by Chen et al. (27). The original study recruited a total of 685,277 participants  $\geq 20$  years old with at least two visits from 2010 to 2016 across 32 sites and 11 cities in China.

Baseline demographic and clinical variables were included as follows: age, gender, smoking and drinking status, family history of diabetes, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), fasting plasma glucose (FPG), total cholesterol (TC), triglyceride (TG), low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), serum urea nitrogen (BUN), serum creatinine (Scr), alanine aminotransferase (ALT). The clinical outcome was incident diabetes during a 3-years follow-up. Baseline excluding criteria in the original study included as follows: (1) no available information on weight, height and gender; (2) extreme BMI values ( $<15$  or  $>55$  kg/m<sup>2</sup>); (3) visit intervals  $< 2$  years; (4) no available fasting plasma glucose value; (5) participants diagnosed with diabetes at baseline (participants diagnosed by self-report or diagnosed by a fasting plasma glucose  $\geq 7.0$  mmol/L) and participants with undefined diabetes status at follow-up. A total of 211,833 participants remained after applying exclusion criteria in the original study. In our study, we further excluded participants with incomplete records. To predicting the 3-year risk of incident diabetes, we also excluded participants who lost to follow-up during 3-years follow up and the censored data is excluded (28). **Figure 1** depicted the participants' selection process. Finally, a total of 15,928 subjects (10,313 male and 5,615 female) were included in the present study.

The authors of the original study have waived all copyright and related ownership of the raw data. Therefore, we could use these data for secondary analysis without infringing on the authors' rights. Furthermore, the original study was approved by the Rich Healthcare Group Review Board, and the information was retrieved retrospectively. And the original study was conducted in accordance with the Declaration of Helsinki, so did this secondary research. The data are anonymous, and the requirement for informed consent was waived by the Rich

**Abbreviations:** BMI, Body mass index; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; FPG, Fasting plasma glucose; TC, Total cholesterol; TG, Triglyceride; HDL-C, High-density lipoprotein cholesterol; LDL-C, Low-density lipid cholesterol; ALT, Alanine aminotransferase; BUN, Serum urea nitrogen; Scr, Serum creatinine; Family history, Family history of diabetes; XGBoost, eXtreme Gradient Boosting; SHAP, Shapley Additive exPlanations; SD, Standardized difference; RR, Relative risk; CI, Confidence intervals; PPV, Positive predictive value; NPV, Negative predictive value; PLR, Positive likelihood ratio; NLR, Negative likelihood ratio; DOR, Diagnostic odds ratio; ROC, Receiver operating characteristic; AUC, Area under curve.



Healthcare Group Review Board due to the observational nature of the study, as reported elsewhere (29).

## Variable Measurement

In each visit to the health check center, participants were required to do a personal questionnaire on demographics,

lifestyle, medical history, and family history of chronic disease. And trained staff performed the baseline examination. Weight was measured in light clothing without shoes to the nearest 0.1 kg. The height was accurate to 0.1 cm. BMI was equal to the weight divided by the square of height, which was accurate to 0.1 kg/m<sup>2</sup>. And the staff measured their blood

pressure by a standard mercury sphygmomanometer. Fasting venous blood samples were taken after fasting for at least 10 h each visit. Plasma glucose levels were measured by the glucose oxidase method. The clinical measurements of FPG, TC, TG, LDL-C, HDL-C, BUN, Scr, and ALT were conducted by an autoanalyzer (Beckman 5800).

## Definitions

The definitions of diabetes were fasting blood glucose  $\geq 7.00$  mmol/L and/or self-reported diabetes during follow-up. Patients were censored either at the time of the diagnosis or at the last visit, whichever comes first.

## Statistical Analysis

All eligible participants were randomly assigned to the training set and the validation set. There were 15 candidate baseline variables involving demographic and clinical characteristics. First, we exclude some variables with relatively significant interference based on collinearity screening. Baseline characteristics were described as means  $\pm$  standard deviations (normal distribution) or medians (quartiles) (skewed distribution) for continuous variables and as percentages or frequency for categorical variables. We used two-sample *t*-tests to analyze differences between the training set and validation set for normally distributed continuous variables, Wilcoxon rank-sum tests for non-normally distributed continuous variables, and chi-square tests for categorical variables.

XGBoost is a scalable tree boosting system that can assemble weak prediction models to establish a more reliable prediction model (30). During the training process, it can generate a new decision tree through gradient boosting on the basis of the existing decision trees to better predict the results. Therefore, a risk prediction system consisting of a series of decision trees is formed after training. In the application process, the predicted risk output is the cumulative score of each decision tree, representing the probability of the predicted outcome. XGBoost provides the importance score of each variable, indicating the relative number of times the variable is used to distribute data in all trees. We ranked these variables according to the prediction contribution of each variable. Given the Shapley Additive exPlanations (SHAP) approach can transform the original non-linear XGBoost model to the summation effects of all variable attributions while approximating the output risk for each participant (31). Thus, the SHAP method was used to interpret the results of the XGBoost model. We used Shapley values to construct dependency graphs to capture the actual relationship between diabetes risk and the three variables with the most significant prediction contribution. Additionally, we summarized the specificity, sensitivity, accuracy, negative predictive value (NPV), positive predictive value (PPV), positive likelihood ratio (PLR), and negative likelihood ratio (NLR) of the XGBoost model at different predicted probability.

We further established three prediction models based on the predictors chosen by the XGBoost model. First, we applied all risk factors selected by the XGBoost method to build a full model. Second, according to the multivariable fractional polynomials (MFP) algorithm, we used the iterative fashion

to determine the significant variables and functional form by backward elimination to establish the MFP model to eliminate the influence of non-linearity and interaction. Third, we conducted a backward step-down selection process based on the Akaike information criterion (AIC) to establish a stepwise model (32). While confirming the statistical significance of the predictor factors, the stepwise logistic regression can achieve local optimal goodness of fit. To assess the discrimination of these risk prediction models, we plotted the receiver operating characteristic (ROC) curve and calculated the area under the ROC curve (AUC) with 95% confidence intervals (CI) for the two sets. Given nomogram is an intuitive graphical prediction model which provides personalized risk predictions for individuals, we further construct the nomogram of the stepwise model. The nomogram is built according to the proportional conversion of each regression coefficient to a 0- to 100-point scale in multiple logistic regression (33). The effect of the variable with the highest  $\beta$  coefficient (absolute value) is assigned 100 points. The point of each variable is added to obtain the total points, which can be converted into the predicted probability of incident diabetes. And we used the Hosmer–Lemeshow test to compare the predicted risk and observed a 3-year incidence of deciles of predicted diabetes risk and we plotted the calibration bar graph of the nomogram for the probability of incident diabetes (34). Besides, we performed decision curve analysis to evaluate the clinical use of the prediction model by quantifying the net benefit at different threshold probabilities: subtracting the proportion of participants with false-positive results from the proportion of participants with true-positive results and then weighing the relative hazards of false positive and false negative results to achieve a net benefit from decision-making (35). And we examined the modifications and interactions between each predictor selected by the stepwise model. In addition, we used a cohort of 11,113 Japanese participants from the NAGALA (NAfd in the Gifu Area, Longitudinal Analysis) database for external validation. The data were also downloaded from the “DATADRYAD” database ([www.Datadryad.org](http://www.Datadryad.org)), shared by Okamura et al. (36) from: Ectopic fat obesity presents the greatest risk for incident type 2 diabetes: a population-based longitudinal study. Dryad Digital Repository. <https://doi.org/10.1038/s41366-018-0076-3>. All results are reported in adherence to the TRIPOD statement (37).

All statistical analyses were performed by the statistical software package R (<http://www.R-project.org>, The R Foundation) and Empower-Stats (<http://www.empowerstats.com>, X&Y Solutions, Inc., Boston, MA). The tests were 2-tailed, and  $P < 0.05$  was taken as statistically significant.

## RESULTS

### Baseline Characteristics of the Study Population

A total of 15,928 eligible participants were included in this study. The mean age of all participants was  $43.33 \pm 12.31$  years old. The male/female ratio was 1.84:1. The mean BMI was  $23.53 \pm 3.30$  Kg/m<sup>2</sup>. The mean FPG was  $4.85 \pm 0.66$  mmol/L. The mean

**TABLE 1** | Baseline characteristics of the training and validation sets.

Characteristic	Training set	Validation set	P-value
Participants	7,940	7,988	
Incident diabetes			0.901
No	7,795 (98.17%)	7,840 (98.15%)	
Yes	145 (1.83%)	148 (1.85%)	
Age (year)	43.43 ± 12.45	43.24 ± 12.17	0.339
Gender			0.595
Male	5,157 (64.95%)	5,156 (64.55%)	
Female	2,783 (35.05%)	2,832 (35.45%)	
BMI (kg/m <sup>2</sup> )	23.51 ± 3.28	23.54 ± 3.32	0.552
SBP (mmHg)	119.90 ± 16.00	119.62 ± 15.77	0.266
DBP (mmHg)	75.12 ± 10.46	75.04 ± 10.38	0.633
FPG (mmol/L)	4.86 ± 0.66	4.84 ± 0.66	0.247
TG (mmol/L)	1.17 (0.80–1.77)	1.16 (0.80–1.75)	0.287
HDL-C (mmol/L)	1.30 ± 0.31	1.30 ± 0.33	0.198
LDL-C (mmol/L)	2.75 ± 0.69	2.75 ± 0.69	0.913
ALT (U/L)	20.00 (14.00–30.00)	20.00 (14.00–30.30)	0.566
BUN (mmol/L)	4.66 ± 1.17	4.67 ± 1.16	0.880
Scr (μmol/L)	72.04 ± 15.07	72.11 ± 15.25	0.767
Smoking status			0.443
Ever/current	1972 (24.84%)	2026 (25.36%)	
Never	5968 (75.16%)	5962 (74.64%)	
Drinking status			0.624
Ever/current	1,544 (19.45%)	1,578 (19.75%)	
Never	6,396 (80.55%)	6,410 (80.25%)	
Family history			0.157
No	7,400 (93.20%)	7489 (93.75%)	
Yes	540 (6.80%)	499 (6.25%)	

Values are n (%) or mean ± SD.

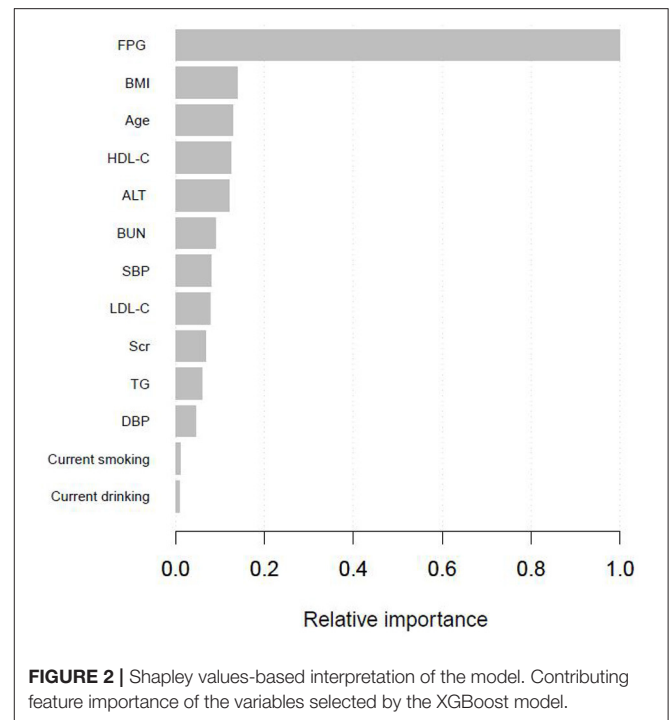
BMI, Body mass index; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; FPG; Fasting plasma glucose; TG, Triglyceride; HDL-C, High density lipoprotein cholesterol; LDL-C, Low density lipid cholesterol; ALT, Alanine aminotransferase; BUN, Blood urea nitrogen; Scr, Serum creatinine; Family history, Family history of diabetes.

HDL-C and LDL-C were  $1.30 \pm 0.32$  and  $2.75 \pm 0.69$  mmol/L, respectively. TC was excluded based on collinearity screening.

**Table 1** compared the baseline characteristics of the training set ( $n = 7,940$ ) and the validation set ( $n = 7,988$ ). After a 3-year follow-up, 148 and 145 incident diabetes cases occurred in the training and validation set, respectively. There were no statistically significant differences in all baseline characteristics and the number of diabetic patients between the two sets (all  $P > 0.05$ ).

## Development of XGBoost Model

**Supplementary Table 1** presented the variables selected by the XGBoost model and the corresponding prediction contributions. The XGBoost model incorporated FPG, BMI, age, HDL-C, ALT, BUN, SBP, LDL-C, Scr, TG, DBP, current smoking, and drinking. The importance score of FPG was 0.5125 and its relative importance was 1.0000, which was the most important variable. The importance score of BMI was 0.0708 and its relative importance was 0.1382, and its prediction contribution



**FIGURE 2** | Shapley values-based interpretation of the model. Contributing feature importance of the variables selected by the XGBoost model.

was only lower than that of FPG. And the importance score of age is 0.0658, ranking third in the prediction contributions. **Figure 2** showed the ranking of the variables based on contributing features. **Supplementary Figure 1** demonstrated the discriminatory ability of the XGBoost model. The AUCs of the XGBoost model in the training set and validation set were 0.977 and 0.920, respectively. Given FPG, BMI and age shared the top 3 contributing features, and we further used the SHAP method to explore the actual relationship between diabetes risk and them (**Supplementary Figure 2**). When FPG < 4.6 mmol/L, the risk of incident diabetes was at a low level. However, when FPG > 4.6 mmol/L, with the increase of FPG, the risk of developing diabetes increased rapidly. And as BMI and age increased, the risk of diabetes gradually increased.

**Supplementary Figure 3** presented the result of the decision curve analysis for the XGBoost model. The results showed that if the personal threshold probability of a participant is 50% (i.e., the participant would opt for diabetes screening if the probability of incident diabetes was <50%), then the net benefit is 0.453 when using the model to decide whether to perform diabetes screening (i.e., oral glucose tolerance test), with added benefit compared to the diabetes screening for all or none participants.

And **Supplementary Table 2** summarized the sensitivity and specificity for predicting incident diabetes at different cutoff values in the XGBoost model. The result showed that although higher cutoff values lead to higher specificity, the sensitivity rapidly dropped to a relatively low level.

## Construction of the Stepwise Model

We further established three prediction models based on the predictors chosen by the XGBoost model, including the MFP

model, full model and stepwise model. In the training set, AUCs of the MFP model, full model and stepwise model were 0.937, 0.934 and 0.933, respectively. In the validation set, the corresponding AUCs of those models were 0.908, 0.909 and 0.910, respectively (Figure 3, Supplementary Table 3). The AUCs of the three models were relatively close. Given that the stepwise model incorporated fewer risk factors and it was simpler than MFP and full models. Besides, the stepwise model could predict the 3-year diabetes risk relatively well. Therefore, we chose the stepwise model as the optimal risk prediction model for incident diabetes. Table 2 showed the 6 variables were selected by stepwise model, including FPG, BMI, age, HDL-C, ALT, and LDL-C. The results showed FPG, BMI, age, HDL-C and ALT were positively associated with incident diabetes. And participants with relatively high FPG were more likely to develop diabetes [relative risk (RR):11.2812; 95% CI: 8.0798–16.4983]. In contrast, participants with relatively high LDL-C were less likely to develop diabetes (RR, 0.7238; 95% CI: 0.5438–0.9229). We further draw a corresponding nomogram to provide a quantitative and simple tool in predicting the risk of diabetes by using age, BMI, FPG, HDL-C, LDL-C, and ALT (Figure 4). Each variable in the nomogram was assigned a specific point, and the points from each variable value are summed to obtain the total points, which was used to obtain the probability for predicting diabetes. And the algorithm of diabetes risk in stepwise model was logit (risk of incident diabetes) =  $-24.07232 + 0.04191 \cdot \text{age (year)} + 0.15291 \cdot \text{BMI (kg/m}^2) + 2.45073 \cdot \text{FPG (mmol/L)} + 1.14025 \cdot \text{HDL-C (mmol/L)} - 0.32400 \cdot \text{LDL-C (mmol/L)} + 0.00852 \cdot \text{ALT (U/L)}$ .

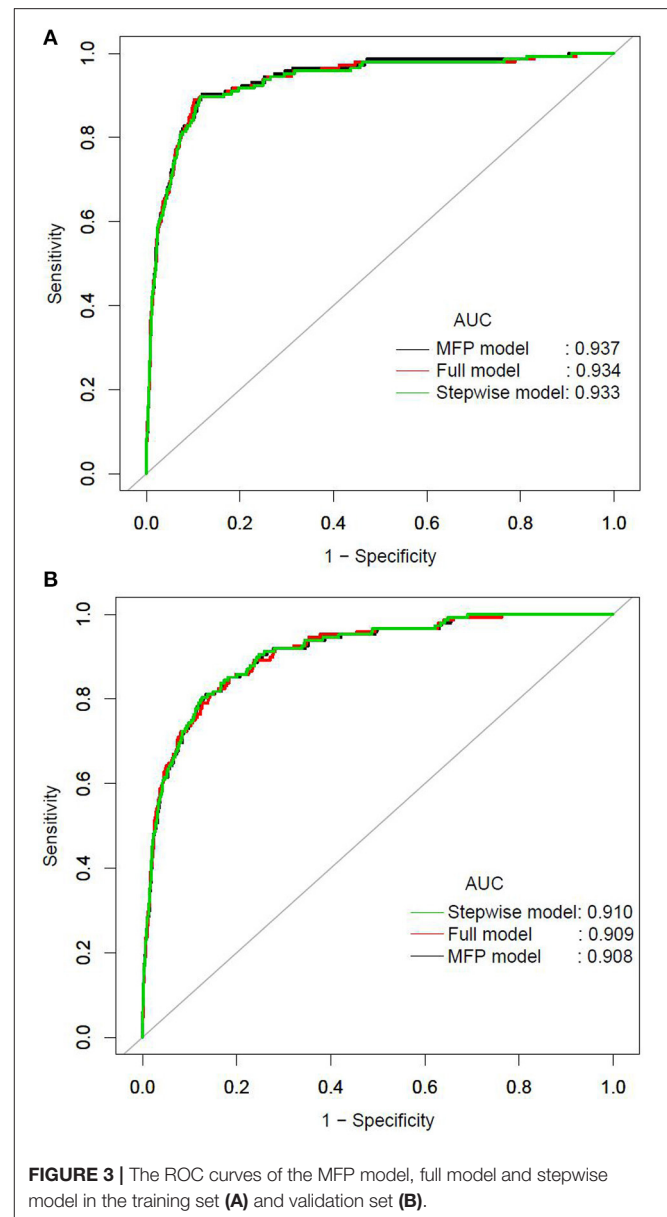
### Performance of the Stepwise Model

The AUCs of the stepwise model were 0.933 and 0.910 in the training and validation sets, respectively (Figure 5). And the result of bootstrap resampling validation (times = 500) confirmed that the prediction performance of the stepwise model in the training cohort was stable (AUC = 0.927) (Supplementary Figure 4). The calibration bar graph of the nomogram for the probability of incident diabetes demonstrated good agreement between observation and prediction both in the training and validation sets (Figure 6). The Hosmer-Lemeshow test indicated that the model was non-significant ( $p = 0.068$  for the training set,  $p = 0.165$  for the validation set), suggesting a perfect fit between the predicted diabetes risk and the observed diabetes risk.

Figure 7 presented the result of decision curve analysis for the stepwise model. The decision curve demonstrated if the threshold probability of a patient was  $>1\%$ , using the XGBoost model to predict incident diabetes was more beneficial than diabetes screening for all or none of the participants. There was a wide range of alternative threshold probability spectrum, which indicated that the stepwise model had significant clinical use.

### Modifications and Interactions Between Each Predictor in the Nomogram

We examined the modifications and interactions between each predictor selected by the stepwise model, including age, BMI, FPG, HDL-C, LDL-C, and ALT. Table 3 showed that almost no



interactions were observed based on our prior specification (most  $P$ -values for interaction  $>0.05$ ), except that BMI and FPG had significant interactions ( $P$ -values for interaction = 0.017).

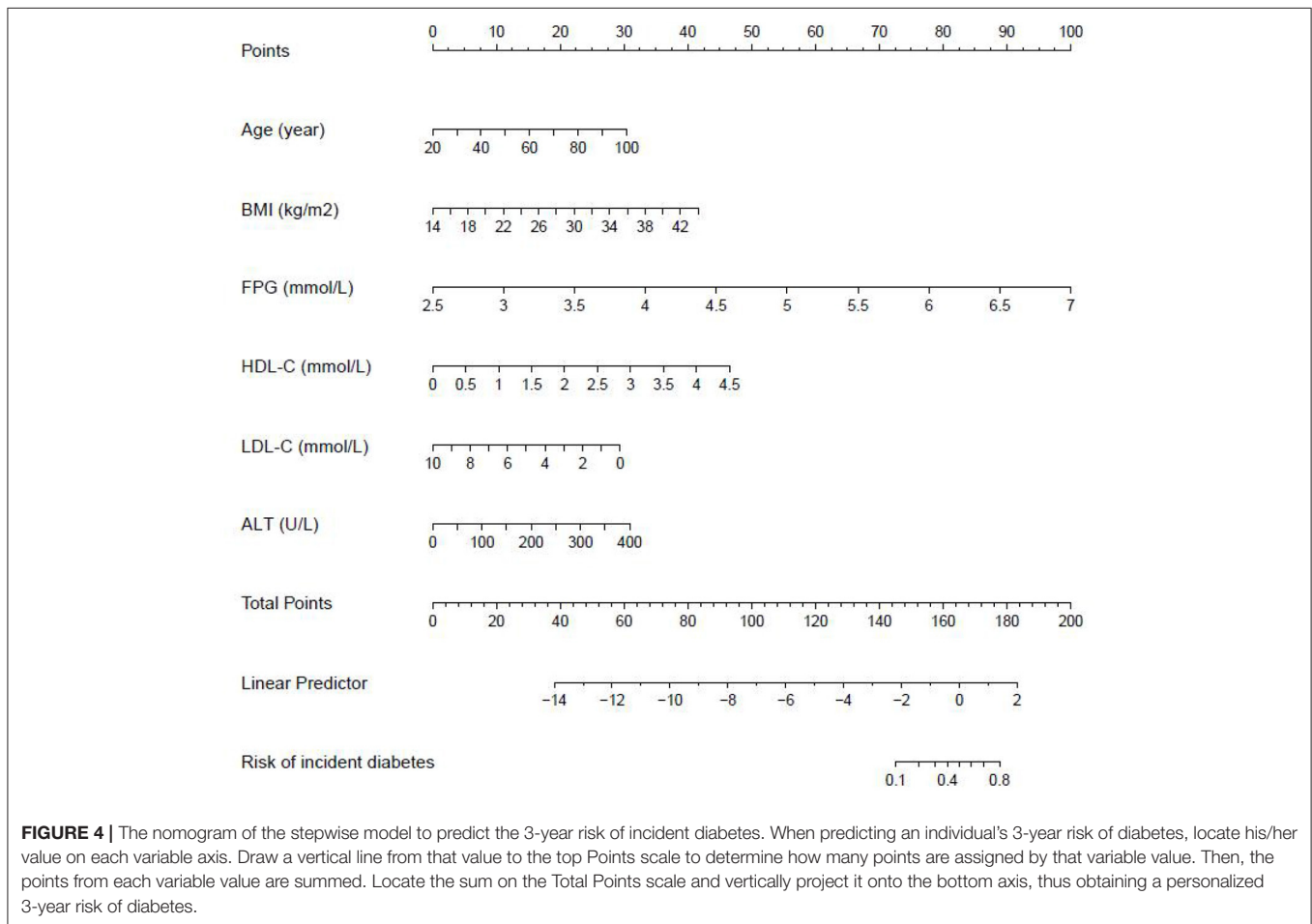
### External Validation

The external validation was performed on a cohort of 11,113 Japanese participants. The AUC for the external validation set was 0.830, which showed good discrimination (Figure 8). And the Hosmer-Lemeshow test for the external validation set showed no statistically significant difference between the predicted diabetes risk and observed diabetes risk, which revealed a perfect fit between the predicted diabetes risk and the observed diabetes risk ( $P = 0.824$ ) (Figure 9). In short, the external validation indicated that the stepwise model was well-generalized.

**TABLE 2** | Variables selected using stepwise logistic regression.

	Beta	Standard error	z-value	RR (95%CI)	P-value
(Intercept)	-24.07232	1.34753	-17.86405	-	-
FPG (mmol/L)	2.45073	0.15763	15.54774	11.2812 (8.0798-16.4983)	0.0000
HDL-C (mmol/L)	1.14025	0.29593	3.85313	3.1101 (1.7651-5.8612)	0.0000
BMI (kg/m <sup>2</sup> )	0.15291	0.03016	5.07010	1.1647 (1.0911-1.2413)	0.0000
Age (year)	0.04191	0.00765	5.47752	1.0427 (1.0276-1.0578)	0.0000
ALT (U/L)	0.00852	0.00335	2.53939	1.0085 (1.0022-1.0146)	0.0060
LDL-C (mmol/L)	-0.32400	0.14526	-2.23050	0.7238 (0.5438-0.9229)	0.0030

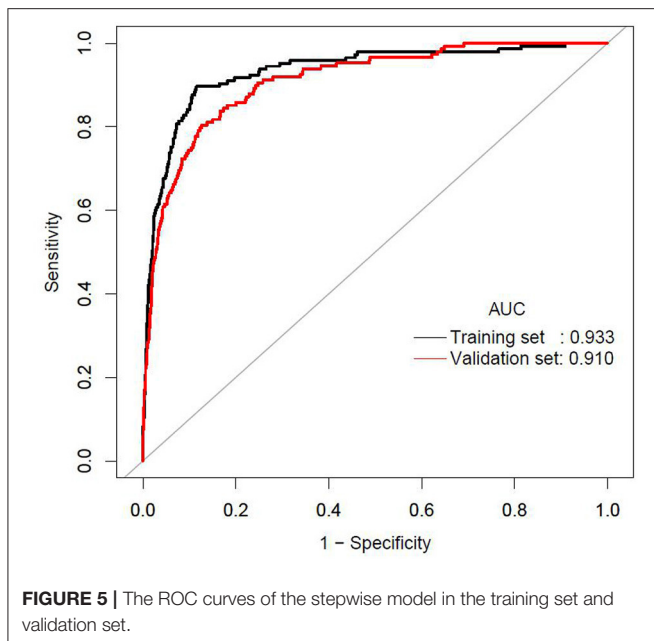
FPG; Fasting plasma glucose; HDL-C, High density lipoprotein cholesterol; BMI, Body mass index; LDL-C, Low density lipid cholesterol; ALT, Alanine aminotransferase; RR, Relative risk; CI, Confidence interval.



## DISCUSSION

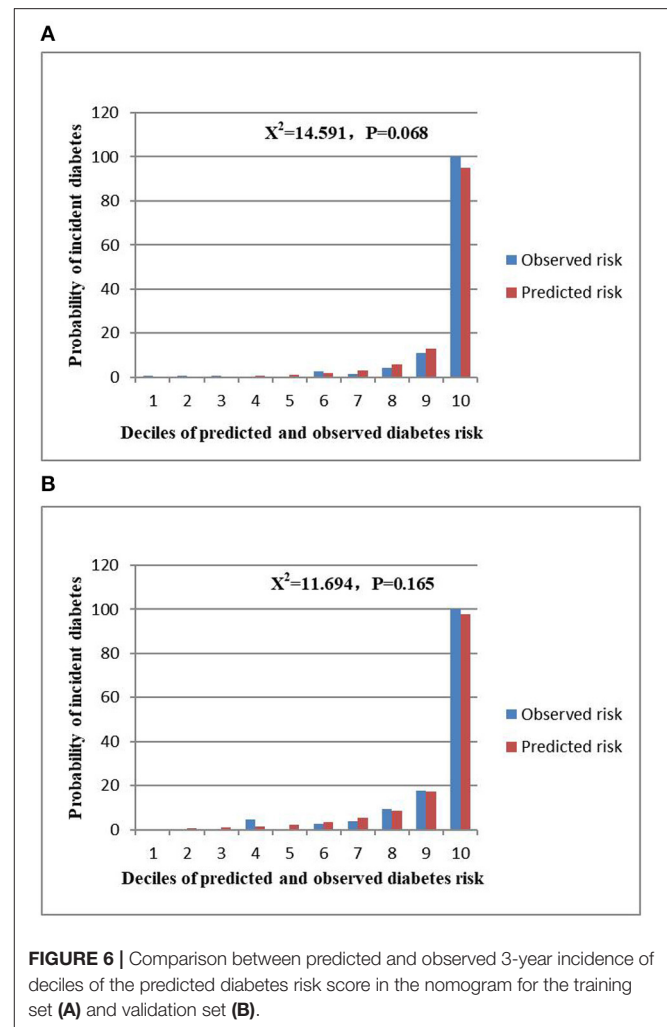
In the present study, we established and validated a risk assessment system for characterizing the 3-year risk of incident diabetes. The XGBoost model incorporated FPG, BMI, Age, HDL-C, ALT, BUN, SBP, LDL-C, Scr, TG, DBP, current smoking, and drinking, of which FPG, BMI and age shared the top three prediction contribution. And we further established a stepwise model and a corresponding prediction nomogram based on the predictors chosen by the XGBoost model. The AUCs of

the stepwise model were 0.933 and 0.910 in the training and validation sets, respectively. The Hosmer-Lemeshow test showed a perfect fit between the predicted diabetes risk and the observed diabetes risk ( $p = 0.068$  for the training set,  $p = 0.165$  for the validation set). Decision curve analysis presented the clinical use of the stepwise model and there was a wide range of alternative threshold probability spectrum. Furthermore, the AUC for the external validation set was 0.830, and the Hosmer-Lemeshow test for the external validation set showed no statistically significant difference between the predicted diabetes risk and observed



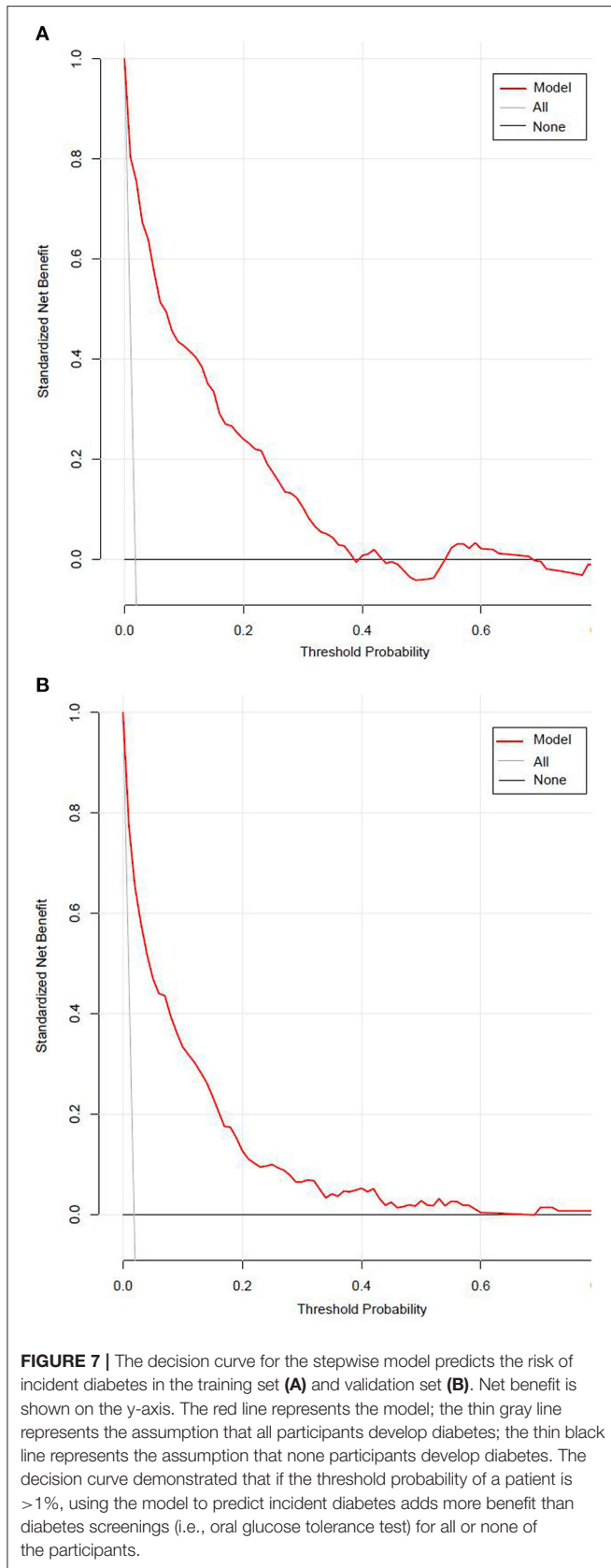
diabetes risk ( $P = 0.824$ ). Therefore, the external validation indicated that the stepwise model was well-generalized.

Machine learning is a collection of data analysis techniques, which aims to establish prediction models that improve with experience and it is becoming an important part of modern medical research (13). It holds promise to enable computers to assist humans to analyze large and complex data sets (14). So far, researchers have developed a variety of machine learning algorithms, including decision trees, kernel machines, neural networks, support vector machines, logistic regression, Bayesian classifiers, ensemble learning, multilayer perceptron, and so on (38–45). Machine learning has unique advantages, including scalability and flexibility, making it applicable to various tasks, such as classification, risk stratification, diagnosis and survival predictions (46). Besides, it handles large multidimensional sets of time-to-event data without the need for assumptions of normality of distributions, linearity of risk prediction, and overfitting of models (47). As yet, machine learning techniques have been applied to a broad range of areas within diabetes, some of which are used to build risk prediction models for incident diabetes (20, 21, 48–52). As a novel machine learning method, XGBoost has become well-established in the machine learning community and gained a positive reputation through numerous machine learning challenges (53). The XGBoost algorithm can automatically handle missing data by adding a default direction for the missing values in each tree node (54). XGBoost has higher calculating speed and accuracy based on the principle of gradient boosting (30). Evidence showed that XGBoost's performance was significantly better than traditional statistical methods (24, 55, 56). To our knowledge, the XGBoost method has not been applied to develop a diabetes risk prediction model. In addition, in those studies using machine learning techniques to predict the risk of diabetes, researchers mainly focused on comparing



various machine learning methods but did not extend the results of machine learning to clinical applications. And AUCs of those models were between 0.580 and 0.925 (20, 21, 48–52). However, this was the first study that used XGBoost method to evaluate the importance of variables and characterize the 3-year risk of incident diabetes among Chinese adults. Notably, we ranked the variables according to the prediction contribution of each selected variable. Furthermore, we used the SHAP method to capture the actual relationship between diabetes risk and the three variables with the largest predicted contribution. Moreover, we developed a simple stepwise model and constructing a corresponding nomogram based on the XGBoost model. And we performed the Hosmer-Lemeshow test to explore the difference between the predicted diabetes risk and the observed diabetes risk. And we did the decision curve analysis to explore the clinical use of the stepwise model, and there was a wide range of alternative threshold probability spectrum. Moreover, we examined the modifications and interactions between each predictor selected by the stepwise model. Furthermore, we used a cohort of 11,113 Japanese participants as the external validation set to explore the reliability and generalizability of the stepwise model.

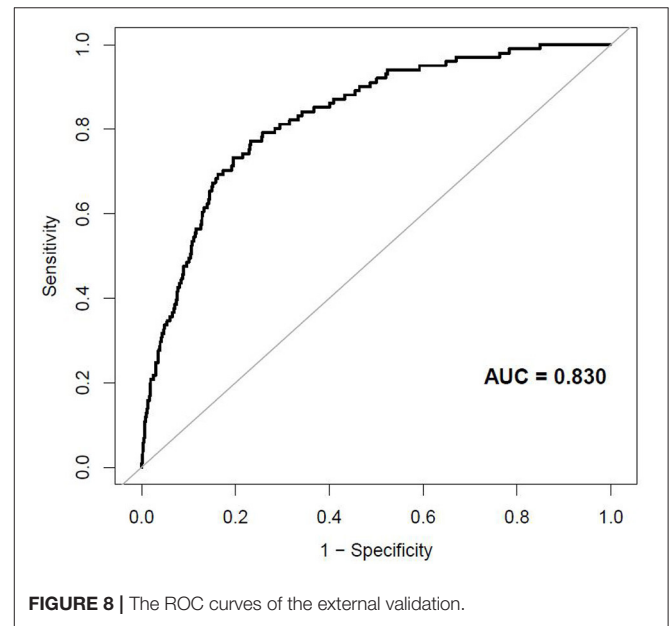




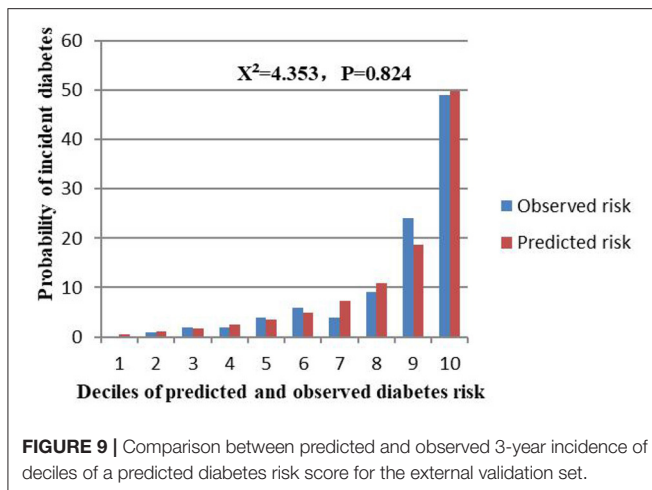
**TABLE 3 |** Modifications and interactions between each predictor selected by the stepwise model.

Predictor	Modifier	HR (95%CI)	P for interaction
Age	BMI	0.997 (0.994, 1.001)	0.186
Age	FPG	0.980 (0.958, 1.002)	0.077
Age	ALT	1.000 (0.999, 1.000)	0.824
Age	HDL-C	1.015 (0.969, 1.064)	0.524
Age	LDL-C	0.996 (0.974, 1.018)	0.699
ALT	FPG	1.001 (0.991, 1.011)	0.902
ALT	BMI	1.000 (0.999, 1.002)	0.627
ALT	HDL-C	0.999 (0.979, 1.019)	0.896
ALT	LDL-C	0.994 (0.986, 1.002)	0.148
BMI	FPG	0.904 (0.832, 0.982)	0.017
BMI	HDL-C	0.978 (0.840, 1.139)	0.776
BMI	LDL-C	1.001 (0.923, 1.086)	0.979
FPG	HDL-C	1.903 (0.692, 5.233)	0.213
FPG	LDL-C	1.034 (0.643, 1.665)	0.889
HDL-C	LDL-C	1.268 (0.560, 2.872)	0.569

FPG; Fasting plasma glucose; HDL-C, High density lipoprotein cholesterol; BMI, Body mass index; LDL-C, Low density lipid cholesterol; ALT, Alanine aminotransferase; HR, Hazard Ratio; CI, Confidence interval.



Diabetes can cause various complications, bring severe physical and psychological distress to patients, and bring a huge burden to the healthcare system. And it tends to be undiagnosed due to the lack of specific symptoms. However, screening for diabetes through oral glucose tolerance test may increase the yield and economic efficiency of screening (57). Our results made up for this deficiency, which helps identify individuals with a high risk of developing diabetes and avoiding the costs and efforts of prevention and treatment in low-risk groups.



Identifying key factors has great clinical significance in the risk assessment of incident diabetes. FPG was the most important risk predictor in our study. Impaired fasting plasma glucose is one of the diagnostic criteria for diabetes. Researchers found compared with those with impaired fasting blood glucose, people with normal fasting blood glucose have a significantly lower risk of developing diabetes (4.0 vs. 11.3%) (58). BMI had the second-largest predicted contribution. The original research showed for every 1 kg/m<sup>2</sup> increase in BMI among Chinese adults, the risk of diabetes increases by 23% (27). Multiple studies have demonstrated overweight or obesity was related to the risk of diabetes (59, 60). Evidence showed obesity, dyslipidemia, abnormal hepatocellular function, and diabetes usually coexist in the same subject and have common pathological mediators (inflammation, metabolic disorders, insulin resistance and intestinal flora imbalance, etc.) (61–63). The prevalence of diabetes markedly increases with age (64). The aging of pancreatic  $\beta$  cells can lead to decreased glucose sensitivity and insulin secretion defects (65). Therefore, the application of these risk predictors in our models is well-founded.

There are some strengths of our study, as follows: (1) As a large-scale multicenter study, our models can be well-applied to the Chinese population. (2) This was the first study that used the XGBoost method to characterize the 3-year risk of incident diabetes. (3) We presented the predicted contribution of each variable selected by the XGBoost model and sorted them in the form of a bar chart. (4) We developed a simple stepwise model based on the XGBoost model and constructed a corresponding nomogram to provide a personalized risk assessment tool. (5) We examined the modifications and interactions between each predictor selected by the stepwise model. (6) We used a cohort of Japanese participants as the external validation set to explore the reliability and generalizability of the stepwise model. (7) Since this was a retrospective cohort study, it could decrease the risk of selection bias and observation bias.

However, there are still some potential limitations. First, the variables we extracted were limited and lacked information about other diabetes risk factors, such as glycated glycosylated hemoglobin, serum insulin and C-peptide concentration. Second, due to the original study design, we cannot distinguish the

types of diabetes mellitus. Considering type 2 diabetes mellitus is the most common kind of diabetes, accounting for over 90% of diabetes cases (66), our findings represent type 2 diabetes mellitus. Third, the researchers did not perform a 2-h oral glucose tolerance test. Thus, our diagnostic criteria for diabetes mellitus may have missed some diabetic patients. However, it is not feasible to perform an oral glucose tolerance test on all participants in such a large-scale cohort study. Fourth, there are too many missing values of variables in the original data, and multiple imputations to replace missing values were not feasible. Therefore, we excluded participants with incomplete records for a complete case study.

## CONCLUSION

We established and validated a risk assessment system for characterizing the 3-year risk of incident diabetes, which showed outstanding performance. And FPG, BMI and age shared the top three prediction contributions. We also constructed a prediction nomogram to provide a personalized risk assessment tool for developing diabetes.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Rich Healthcare Group Review Board, and the information was retrieved retrospectively. The data are anonymous, and the requirement for informed consent was waived by the Rich Healthcare Group Review Board due to the observational nature of the study, as reported elsewhere.

## AUTHOR CONTRIBUTIONS

YW and HH conceived and designed the research and drafted the manuscript. JC and RC did statistical analysis. XZ and HC took part in the discussion. DY revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This study was supported in part by Discipline Construction Ability Enhancement Project of Shenzhen Municipal Health Commission (SZXJ2017031).

## ACKNOWLEDGMENTS

The authors thank the funder and all study participants.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2021.626331/full#supplementary-material>

## REFERENCES

1. Cho NH, Shaw JE, Karuranga S, Huang Y, Da RFJ, Ohlrogge AW, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract.* (2018) 138:271–81. doi: 10.1016/j.diabres.2018.02.023
2. Li Y, Teng D, Shi X, Qin G, Qin Y, Quan H, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. *BMJ.* (2020) 369:m997. doi: 10.1136/bmj.m997
3. Thomas MC, Brownlee M, Susztak K, Sharma K, Jandeleit-Dahm KA, Zoungas S, et al. Diabetic kidney disease. *Nat Rev Dis Primers.* (2015) 1:15018. doi: 10.1038/nrdp.2015.70
4. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* (2010) 376:124–36. doi: 10.1016/S0140-6736(09)62124-3
5. Feldman EL, Callaghan BC, Pop-Busui R, Zochodne DW, Wright DE, Bennett DL, et al. Diabetic neuropathy. *Nat Rev Dis Primers.* (2019) 5:42. doi: 10.1038/s41572-019-0092-1
6. Zelniker TA, Wiviott SD, Raz I, Im K, Goodrich EL, Bonaca MP, et al. SGLT2 inhibitors for primary and secondary prevention of cardiovascular and renal outcomes in type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. *Lancet.* (2019) 393:31–9. doi: 10.1016/S0140-6736(18)32590-X
7. Hardigan T, Ward R, Ergul A. Cerebrovascular complications of diabetes: focus on cognitive dysfunction. *Clin Sci.* (2016) 130:1807–22. doi: 10.1042/CS20160397
8. Li Y, Wang DD, Ley SH, Vasanti M, Howard AG, He Y, et al. Time trends of dietary and lifestyle factors and their potential impact on diabetes burden in China. *Diabetes Care.* (2017) 40:1685–94. doi: 10.2337/dc17-0571
9. Le Roux CW, Astrup A, Fujioka K, Greenway F, Lau D, Van Gaal L, et al. 3 years of liraglutide versus placebo for type 2 diabetes risk reduction and weight management in individuals with prediabetes: a randomised, double-blind trial. *Lancet.* (2017) 389:1399–409. doi: 10.1016/S0140-6736(17)30069-7
10. Gregg EW, Chen H, Wagenknecht LE, Clark JM, Delahanty LM, Bantle J, et al. Association of an intensive lifestyle intervention with remission of type 2 diabetes. *JAMA.* (2012) 308:2489–96. doi: 10.1001/jama.2012.67929
11. Brito JP, Montori VM, Davis AM. Metabolic surgery in the treatment algorithm for type 2 diabetes: a joint statement by international diabetes organizations. *JAMA.* (2017) 317:635–6. doi: 10.1001/jama.2016.20563
12. Shi X, Shi Y, Chen N, Lin M, Su W, Zhang H, et al. Effect of exenatide after short-time intensive insulin therapy on glycaemic remission maintenance in type 2 diabetes patients: a randomized controlled trial. *Sci Rep.* (2017) 7:2383. doi: 10.1038/s41598-017-02631-1
13. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. *Cell.* (2018) 173:1581–92. doi: 10.1016/j.cell.2018.05.015
14. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* (2015) 16:321–32. doi: 10.1038/nrg3920
15. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol.* (2018) 14:91–118. doi: 10.1146/annurev-clinpsy-032816-045037
16. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science.* (2015) 349:255–60. doi: 10.1126/science.aaa8415
17. Lagani V, Koumakis L, Chiarugi F, Lakasing E, Tsamardinos I. A systematic review of predictive risk models for diabetes complications based on large scale clinical studies. *J Diabetes Complicat.* (2013) 27:407–13. doi: 10.1016/j.jdiacomp.2012.11.003
18. Xiong XL, Zhang RX, Bi Y, Zhou WH, Yu Y, Zhu DL. Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults. *Curr Med Sci.* (2019) 39:582–8. doi: 10.1007/s11596-019-2077-4
19. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. *Front Genet.* (2018) 9:515. doi: 10.3389/fgene.2018.00515
20. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci Rep.* (2019) 9:13805. doi: 10.1038/s41598-019-49563-6
21. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Sci Rep.* (2020) 10:4406. doi: 10.1038/s41598-020-61123-x
22. Zhao X, Wu S, Fang N, Sun X, Fan J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief Bioinform.* (2020) 21:1581–595. doi: 10.1093/bib/bbz096
23. Polano M, Chierici M, Dal Bo M, Gentilini D, Di Cintio F, Baboci L, et al. A pan-cancer approach to predict responsiveness to immune checkpoint inhibitors by machine learning. *Cancers.* (2019) 11:1562. doi: 10.3390/cancers11101562
24. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit Care.* (2019) 23:112. doi: 10.1186/s13054-019-2411-z
25. Wang K, Zuo P, Liu Y, Zhang M, Zhao X, Xie S, et al. Clinical and laboratory predictors of in-hospital mortality in patients with COVID-19: a cohort study in Wuhan, China. *Clin Infect Dis.* (2020) 71:2079–88. doi: 10.2139/ssrn.3546115
26. Tahmassebi A, Wengert GJ, Helbich TH, Bago-Horvath Z, Alaei S, Bartsch R, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol.* (2019) 54:110–7. doi: 10.1097/RLI.0000000000000518
27. Chen Y, Zhang XP, Yuan J, Cai B, Wang XL, Wu XL, et al. Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study. *BMJ Open.* (2018) 8:e21768. doi: 10.1136/bmjopen-2018-021768
28. Wang K, Gong M, Xie S, Zhang M, Zheng H, Zhao X, et al. Nomogram prediction for the 3-year risk of type 2 diabetes in healthy mainland China residents. *EPMA J.* (2019) 10:227–37. doi: 10.1007/s13167-019-00181-2
29. Geleris J, Sun Y, Platt J, Zucker J, Baldwin M, Hripcsak G, et al. Observational study of hydroxychloroquine in hospitalized patients with Covid-19. *N Engl J Med.* (2020) 382:2411–8. doi: 10.1056/NEJMoa2012410
30. 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*. San Francisco, CA: ACM (2016). p. 785–94.
31. Lundberg S, Lee S. A unified approach to interpreting model predictions. Advances in neural information processing systems. Long Beach, CA: Neural Information Processing Systems (2017). p. 4765–74.
32. Collignon O, Monnez J. Clustering of the values of a response variable and simultaneous covariate selection using a stepwise algorithm. *Appl. Math.* (2016) 7:1639–48. doi: 10.4236/am.2016.715141
33. Lei Z, Li J, Wu D, Xia Y, Wang Q, Si A, et al. Nomogram for preoperative estimation of microvascular invasion risk in hepatitis B virus-related hepatocellular carcinoma within the milan criteria. *JAMA Surg.* (2016) 151:356–63. doi: 10.1001/jamasurg.2015.4257
34. Sun F, Tao Q, Zhan S. An accurate risk score for estimation 5-year risk of type 2 diabetes based on a health screening population in Taiwan. *Diabetes Res Clin Pract.* (2009) 85:228–34. doi: 10.1016/j.diabres.2009.05.005
35. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA.* (2015) 313:409–10. doi: 10.1001/jama.2015.37
36. Okamura T, Hashimoto Y, Hamaguchi M, Obora A, Kojima T, Fukui M. Ectopic fat obesity presents the greatest risk for incident type 2 diabetes: a population-based longitudinal study. *Int J Obes.* (2019) 43:139–48. doi: 10.1038/s41366-018-0076-3
37. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* (2015) 350:g7594. doi: 10.1136/bmj.g7594
38. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. *Lancet Diabetes Endocrinol.* (2014) 2:19–29. doi: 10.1016/S2213-8587(13)70103-7
39. Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol.* (2019) 29:R231–6. doi: 10.1016/j.cub.2019.02.034

40. Allalou A, Nalla A, Prentice KJ, Liu Y, Zhang M, Dai FF, et al. A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes. *Diabetes*. (2016) 65:2529–39. doi: 10.2337/db15-1720
41. Li Y, Zhang T. Deep neural mapping support vector machines. *Neural Netw*. (2017) 93:185–94. doi: 10.1016/j.neunet.2017.05.010
42. Song H, Thiagarajan JJ, Sattigeri P, Spanias A. Optimizing kernel machines using deep learning. *IEEE Trans Neural Netw Learn Syst*. (2018) 29:5528–40. doi: 10.1109/TNNLS.2018.2804895
43. Wiggins M, Saad A, Litt B, Vachtsevanos G. Evolving a Bayesian classifier for ECG-based age classification in medical applications. *Appl Soft Comput*. (2008) 8:599–608. doi: 10.1016/j.asoc.2007.03.009
44. Mao S, Lin W, Jiao L, Gou S, Chen JW. End-to-end ensemble learning by exploiting the correlation between individuals and weights. *IEEE Trans Cybern*. (2021) 51:2835–46. doi: 10.1109/TCYB.2019.2931071
45. Tang J, Deng C, Huang GB. Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst*. (2016) 27:809–21. doi: 10.1109/TNNLS.2015.2424995
46. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. (2019) 20:e262–73. doi: 10.1016/S1470-2045(19)30149-4
47. Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. *Diabetes Metab Res Rev*. (2020) 36:e3252. doi: 10.1002/dmrr.3252
48. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open*. (2013) 3:e002457. doi: 10.1136/bmjopen-2012-002457
49. Cho SB, Kim SC, Chung MG. Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Sci Rep*. (2019) 9:3329. doi: 10.1038/s41598-019-40058-y
50. Lee BJ, Kim JY. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE J Biomed Health Inform*. (2016) 20:39–46. doi: 10.1109/JBHI.2015.2396520
51. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford Exercise Testing (FIT) project. *PLoS ONE*. (2017) 12:e179805. doi: 10.1371/journal.pone.0179805
52. Ji X, Tong W, Liu Z, Shi T. Five-feature model for developing the classifier for synergistic vs. antagonistic drug combinations built by XGBoost. *Front Genet*. (2019) 10:600. doi: 10.3389/fgene.2019.00600
53. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis*. (2019) 74:300–9. doi: 10.1053/j.ajkd.2019.02.016
54. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J Transl Med*. (2019) 17:119. doi: 10.1186/s12967-019-1860-0
55. Khemasuwan D, Sorensen J, Griffin DC. Predictive variables for failure in administration of intrapleural tissue plasminogen activator/deoxyribonuclease in patients with complicated parapneumonic effusions/empyema. *Chest*. (2018) 154:550–6. doi: 10.1016/j.chest.2018.01.037
56. Selph S, Dana T, Blazina I, Bougatsos C, Patel H, Chou R. Screening for type 2 diabetes mellitus: a systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med*. (2015) 162:765–76. doi: 10.7326/M14-2221
57. Nichols GA, Hillier TA, Brown JB. Progression from newly acquired impaired fasting glucose to type 2 diabetes. *Diabetes Care*. (2007) 30:228–33. doi: 10.2337/dc06-1392
58. Arnlov J, Sundstrom J, Ingelsson E, Lind L. Impact of BMI and the metabolic syndrome on the risk of diabetes in middle-aged men. *Diabetes Care*. (2011) 34:61–5. doi: 10.2337/dc10-0955
59. Tirosh A, Shai I, Afek A, Dubnov-Raz G, Ayalon N, Gordon B, et al. Adolescent BMI trajectory and risk of diabetes versus coronary disease. *N Engl J Med*. (2011) 364:1315–25. doi: 10.1056/NEJMoa1006992
60. Vozarova B, Stefan N, Lindsay RS, Saremi A, Pratley RE, Bogardus C, et al. High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes*. (2002) 51:1889–95. doi: 10.2337/diabetes.51.6.1889
61. Meijnikman AS, Gerdes VE, Nieuwdorp M, Herrema H. Evaluating causality of gut microbiota in obesity and diabetes in humans. *Endocr Rev*. (2018) 39:133–53. doi: 10.1210/er.2017-00192
62. Wu L, Parhofer KG. Diabetic dyslipidemia. *Metabolism*. (2014) 63:1469–79. doi: 10.1016/j.metabol.2014.08.010
63. Qiao Q, Hu G, Tuomilehto J, Nakagami T, Balkau B, Borch-Johnsen K, et al. Age- and sex-specific prevalence of diabetes and impaired glucose regulation in 11 Asian cohorts. *Diabetes Care*. (2003) 26:1770–80. doi: 10.2337/diacare.26.6.1770
64. Chang AM, Halter JB. Aging and insulin secretion. *Am J Physiol Endocrinol Metab*. (2003) 284:E7–12. doi: 10.1152/ajpendo.00366.2002
65. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. (2018) 14:88–98. doi: 10.1038/nrendo.2017.151
66. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care*. (2019) 42:2298–306. doi: 10.2337/dc19-0587

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Wu, Hu, Cai, Chen, Zuo, Cheng and Yan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.