Frontiers of Nursing • 8(3) • 2021
DOI: 10.2478/fon-2021-0022

sciendo

Frontiers of **Nursing**

Application of machine learning algorithm for predicting gestational diabetes mellitus in early pregnancy†

Original article

Li-Li Wei[a], Yue-Shuai Pan[b], Yan Zhang[a], Kai Chen[b], Hao-Yu Wang[c], Jing-Yuan Wang[d,*]

[a]*Department of Nursing, The Affiliated Hospital of Qingdao University, Qingdao, Shandong 266003, China*
[b]*Intensive Care Unit, The Affiliated Hospital of Qingdao University, Qingdao, Shandong 266003, China*
[c]*Research Institute for Mathematics and Interdisciplinary Sciences, Qingdao University of Science and Technology, Qingdao, Shandong 266061, China*
[d]*Department of Critical Care and Respiratory Medicine, The Affiliated Hospital of Qingdao University, Qingdao, Shandong 266003, China*

Received: 24 November 2020; Accepted: 11 January 2021; Published: 20 September 2021

Abstract: **Objective:** To study the application of a machine learning algorithm for predicting gestational diabetes mellitus (GDM) in early pregnancy.

Methods: This study identified indicators related to GDM through a literature review and expert discussion. Pregnant women who had attended medical institutions for an antenatal examination from November 2017 to August 2018 were selected for analysis, and the collected indicators were retrospectively analyzed. Based on Python, the indicators were classified and modeled using a random forest regression algorithm, and the performance of the prediction model was analyzed.

Results: We obtained 4806 analyzable data from 1625 pregnant women. Among these, 3265 samples with all 67 indicators were used to establish data set F1; 4806 samples with 38 identical indicators were used to establish data set F2. Each of F1 and F2 was used for training the random forest algorithm. The overall predictive accuracy of the F1 model was 93.10%, area under the receiver operating characteristic curve (AUC) was 0.66, and the predictive accuracy of GDM-positive cases was 37.10%. The corresponding values for the F2 model were 88.70%, 0.87, and 79.44%. The results thus showed that the F2 prediction model performed better than the F1 model. To explore the impact of sacrificial indicators on GDM prediction, the F3 data set was established using 3265 samples (F1) with 38 indicators (F2). After training, the overall predictive accuracy of the F3 model was 91.60%, AUC was 0.58, and the predictive accuracy of positive cases was 15.85%.

Conclusions: In this study, a model for predicting GDM with several input variables (e.g., physical examination, past history, personal history, family history, and laboratory indicators) was established using a random forest regression algorithm. The trained prediction model exhibited a good performance and is valuable as a reference for predicting GDM in women at an early stage of pregnancy. In addition, there are certain requirements for the proportions of negative and positive cases in sample data sets when the random forest algorithm is applied to the early prediction of GDM.

Keywords: *early prediction • gestational diabetes mellitus • machine learning algorithm • random forest regression*

© Shanxi Medical Periodical Press.

† This project was supported by the Qingdao Municipal Bureau of Science and Technology (No. 19-6-1-55-nsh).

How to cite this article: Wei LL, Pan YS, Zhang Y, Chen K, Wang HY, Wang JY. Application of machine learning algorithm for predicting gestational diabetes mellitus in early pregnancy. *Front Nurs*. 2021;3:209–222.

* *Corresponding author.*
E-mail: 18661801620@163.com (J. -Y. Wang).

Open Access. © 2021 Wei et al., published by Sciendo. This work is licensed under the Creative Commons Attribution 4.0 License.

1. Introduction

Gestational diabetes mellitus (GDM) refers to abnormal glucose tolerance and persistent high blood glucose concentration during pregnancy. It is a serious threat to maternal and fetal health. GDM not only causes adverse perinatal pregnancy outcomes,[1] such as postpartum hemorrhage, infection, preterm delivery, macrosomia, and neonatal respiratory distress syndrome, but also threatens the long-term health of mothers and infants. Compared with normal pregnant mothers, women with GDM had a 7-fold increased risk of developing type 2 diabetes after delivery.[2] At the same time, the risk of metabolism-related diseases such as obesity and type 2 diabetes in offspring will also increase significantly.[3] With the update of GDM diagnostic criteria, the increase of elderly pregnant women, and lifestyle changes, the global prevalence of GDM increased to 5.40–7.71% and showed a trend of increasing year by year.[4,5] Because of its significant harmfulness and high incidence, GDM has attracted wide attention of researchers all over the world.

Previous studies on high-risk groups of GDM mainly focused on blood glucose control and prevention of complications in the population after the diagnosis of GDM. At present, the diagnosis of GDM needs to be confirmed by an oral glucose tolerance test (OGTT) at 24–28 weeks of pregnancy. However, previous studies have found that before a definitive diagnosis of GDM, the persistent high blood glucose concentration during pregnancy will also have an adverse effect on the outcome of the pregnant woman or fetus.[6] Researchers have put forward the concept of "prediabetes" in diabetes management, which is defined as the blood glucose concentration being higher than normal, but lower than the diagnostic criteria for diabetes.[7] Prediabetes is a high-risk stage of diabetes. At this stage, diabetes can be delayed or prevented if high-risk groups can be identified early enough and given appropriate intervention.[8] This development law is also suitable for the evolution of GDM.

Comprehensive factors, such as genetic factors, environmental factors, and lifestyle, play an important role in the occurrence and development of GDM.[9,10] The American Diabetes Association believes that the main risk factors for GDM include maternal age ≥35 years, Body Mass Index (BMI) ≥25 kg/m^2, family history of diabetes, impaired glucose tolerance, and so on.[11] On the other hand, guidelines issued by the National Institute for Clinical Excellence (NICE) in the UK have indicated that BMI ≥30kg/m^2, macrosomia (≥4.5 kg) delivery history, and high-risk minorities are independent risk factors for GDM.[12] Therefore, there is no consensus about what constitutes GDM high-risk factors, and there are certain limitations in the identification of GDM high-risk populations based on whether or not high-risk factors exist.

In the past 10 years, researchers began to use machine learning algorithms to study the risk factors of some chronic diseases.[13–15] Different machine learning algorithms, such as logistic regression (LR), decision tree (DT), and artificial neural network (ANN), have been applied to predict diseases. These algorithms can assist in the classification of the risk factors, deduce the correlation between attributes, establish a risk prediction model, and predict the occurrence of disease. In contrast, the previous domestic research methods on the risk of GDM are limited to the classical statistical category, and although studies on the application of machine learning algorithms to diagnose and predict GDM are scarce, one such study is mentioned in the successive paragraph.

Wu et al. retrospectively analyzed the data of early pregnancy in southwest China and used the TreeNet method to establish a prediction model for GDM.[16] Due to the limitations of retrospective methods and a lack of data on pregnant women's lifestyle, the prediction accuracy is not high (<65%).

In today's medical industry, the degree of digitization of hospitals is gradually improving. The era of medical big data has come, and is a result of the rapid accumulation of medical data and the continuous increase in data dimensions. Therefore, to achieve accurate identification of GDM high-risk groups, it is necessary to learn from the application of machine learning algorithms in disease diagnosis and prediction, explore, and form a practical, more accurate, and easier-to-operate prediction model.

In this study, pregnant women in the Qingdao area of China were enrolled. Data including results of physical examination, medical history, other genetic and social environment factors, as well as dietary habits, exercise habits, and pre/post-maternal birth weight, were collected by prospective follow-up as independent variables. The GDM prediction model was established by a machine learning algorithm, which provided a theoretical basis for intervention measures for GDM in early pregnancy.

2. Methods

2.1. Participants

Pregnant women who attended medical institutions for an antenatal examination from November 2017 to August 2018 were selected for this study. Inclusion criteria were as follows: positivity on the human chorionic gonadotropin (HCG) test, agreeing to participate, and providing written informed consent. The exclusion criterion was the existence of diabetes before pregnancy.

2.2. Development of indicators related to gestational diabetes

Through a literature review, the following indicators related to gestational diabetes were identified: physical examination results (age, weight, blood pressure, blood type, etc.); medical history (hypertension, diabetes, heart disease, histories of abnormal pregnancy, polycystic ovary syndrome (PCOS), etc.); personal history (birth weight of the pregnant woman at the time of her own birth, her mother's weight at birth, her mother's parity at birth, smoker or nonsmoker, etc.); family history (immediate family suffering from diabetes or not, etc.); specialist examination results (fundal height, abdomen circumference, gravidity, parity, etc.); and laboratory indicators (amniotic fluid index/depth, blood glucose, routine blood tests, routine urine tests, blood biochemical infection indicators, glycated hemoglobin, etc.).

2.2.1. Food frequency questionnaire

Food frequency questionnaire (FFQ) is a widely used tool for assessing dietary intake in nutritional epidemiology.[17] It is used to investigate the long-term patterns of dietary intake in subjects. It is simple to operate and to complete. In this study, food adjustment was carried out based on the dietary intake of residents in Qingdao and the characteristics of their diet during pregnancy were ascertained, including 22 food categories (Table 1). The range of intake frequency was as follows: (1) never; (2) less than once a month; (3) 1–3 times a month; (4) 1–2 times a week; (5) 3–4 times a week; (6) 5–6 times a week; (7) once a day; (8) twice a day; or (9) more than 3 times a day. The range of food intake was as follows: (1) less than 50 g; (2) 100 g; (3) 150 g; (4) 200 g; (5) more than 250 g; and (6) not applicable. Dietary intake = intake frequency × each intake/cycle.

2.2.2. International Physical Activity Questionnaire

The International Physical Activity Questionnaire (IPAQ) short questionnaire was developed by the International Physical Activity Measurement Group to investigate physical activity in the past week. We use this questionnaire as a tool for assessing the physical activity of women during pregnancy.[18,19] It has good reliability and validity and includes 7 items, as indicated by Table 1 (8.1–8.7).

2.3. Data collection

Investigators with identical training conducted a face-to-face questionnaire survey of participants and reviewed the pregnancy-related medical records to obtain the indicators related to GDM.

2.4. Data analyses

After data collection, some indicators were merged or deleted according to the suggestions of obstetric experts and a statistician. Pre-pregnancy BMI was calculated as follows:

BMI = weight (kg)/height $(m)^2$. Pregnancy weight gain (WG) was also calculated as follows: WG (kg) = birth weight − pre-pregnancy weight.

Some laboratory indicators (e.g., amniotic fluid index/depth, routine blood tests, routine urine tests, blood biochemistry, infection indicators, and glycated hemoglobin) were also deleted due to lower acquisition rates among the subjects.

All categorical variables were treated as dichotomous variables (0/1). The output variable was based on whether or not diabetes was diagnosed during 24–28 gestational weeks (yes = 1, no = 0). A total of 67 variables were included in our study as model input according to the suggestions of obstetric experts and a statistician (Table 1).

2.5. Technical roadmap

Based on Python, each model was established using a random forest regression algorithm. Random forest is a combinatorial classifier that is based on statistical theory, with the characteristics of selecting data randomly and feature ranking, and is considered suitable to the feature ranking and selection of medical data.[20]

To systematically train the model and evaluate its accuracy, this study used the train_test_split function to randomly divide the data into 2 categories, namely a 70% training data set and a 30% testing data set. The training data were used to train the model, and the test data were used to validate it, the two of which were performed separately.

It was found that the performance of the prediction models trained using the different sets of indicators or data sets with different ratios of negative/positive cases differed. In this study, 3 datasets (F1, F2, and F3) were established and then compared [see Technical Route (Figure 1)].

2.6. Ethical considerations

The study was approved by the medical ethics committee of the study principal investigator's hospital. All patients were informed about the purposes and methods of the study during the recruitment. Participation was voluntary, and patients could refuse to participate or withdraw at any time. Each participant provided written

Dimension	Variables	Input variables
1. Physical examination	1.1 Age	Age
	1.2 BMI	BMI
	1.3 Blood type-A	A
	1.4 Blood type-B	B
	1.5 Blood type-AB	AB
	1.6 Blood type-O	O
	1.7 Systolic blood pressure	SBP
	1.8 Diastolic blood pressure	DBP
2. Past history	2.1 Hypertension	Hypertension
	2.2 Heart disease	Heart disease
	2.3 Thyroid disease	Thyroid disease
	2.4 Gynecological diseases	Gynecological diseases
	2.5 Kidney disease	Kidney disease
	2.6 Congenital spina bifida	Congenital spina bifida
	2.7 Benign tumor	Benign tumor
	2.8 Past surgical history	Past surgical history
	2.9 HPV	HPV
	2.10 Colpomycosis	Colpomycosis
	2.11 PCOS	PCOS
	2.12 Negative reproductive history	Negative reproductive history
	2.13 History of multiple pregnancies	History of multiple pregnancies
	2.14 Regular menstruation	Regular menstruation
3. Personal history	3.1 Body weight at birth	Body weight at birth
	3.2 Mother's weight	Mother's weight
	3.3 Number of births by mother	Number of births by mother
	3.4 Smoking	Smoking
4. Family history	4.1 Family history of diabetes	Family history of diabetes
5. Specialist examination	5.1 Uterine height	Uterine height
	5.2 Abdominal circumference	Abdominal circumference
	5.3 Weight gain	Weight gain
	5.4 Gravidity history	Gravidity history
	5.5 Parity history	Parity history
	5.6 Multiple pregnancies	Multiple pregnancies
	5.7 Fetus number	Fetus number
6. Laboratory indicators	6.1 Fasting glucose	Fasting glucose
	6.2 Hepatitis	Hepatitis
	6.3 HIV	HIV
	6.4 Syphilis	Syphilis
7. Food frequency questionnaire (FFQ)	7.1 Porridge intake	Porridge intake (kg/month)
	7.2 Flour foods intake	Flour foods intake (kg/month)
	7.3 Sweet food Intake	Sweet food Intake (kg/month)
	7.4 Fried food Intake	Fried food Intake (kg/month)
	7.5 Filling food Intake	Filling food Intake (kg/month)
	7.6 Coarse grain intake	Coarse grain intake (kg/month)
	7.7 Potato intake	Potato intake (kg/month)
	7.8 Milk intake	Milk intake (kg/month)
	7.9 Egg intake	Egg intake (kg/month)
	7.10 Red meat intake	Red meat intake (kg/month)
	7.11 Poultry intake	Poultry intake (kg/month)
	7.12 Processed meat intake	Processed meat intake (kg/month)
	7.13 Freshwater fishes intake	Freshwater fishes Intake (kg/month)
	7.14 Seafood intake	Seafood intake (kg/month)
	7.15 Bean products intake	Bean products intake (kg/month)
	7.16 Nuts intake	Nuts intake (kg/month)
	7.17 Dark vegetables intake	Dark vegetables intake (kg/month)
	7.18 Light vegetables intake	Light vegetables intake (kg/month)
	7.19 Mushrooms intake	Mushrooms intake (kg/month)
	7.20 Fruit intake	Fruit intake (kg/month)
	7.21 Sweet drinks intake	Sweet drinks intake (kg/month)
	7.22 Alcohol intake	Alcohol intake (l/month)
8. International Physical Activity Questionnaire (IPAQ)	8.1 Heavy physical activity	Heavy physical activity (min/week)
	8.2 Moderate physical activity	Moderate physical activity (min/week)
	8.3 Light physical activity	Light physical activity (min/week)
	8.4 Walking time	Walking time (min/week)
	8.5 Sedentary time	Sedentary time (min/day)
	8.6 Sleep duration on weekdays	Sleep duration on weekdays (hour/day)
	8.7 Sleep duration in rest days	Sleep duration in rest days (hour/day)

Note: BMI, body mass index

Table 1. Definition of input data variables for model training.

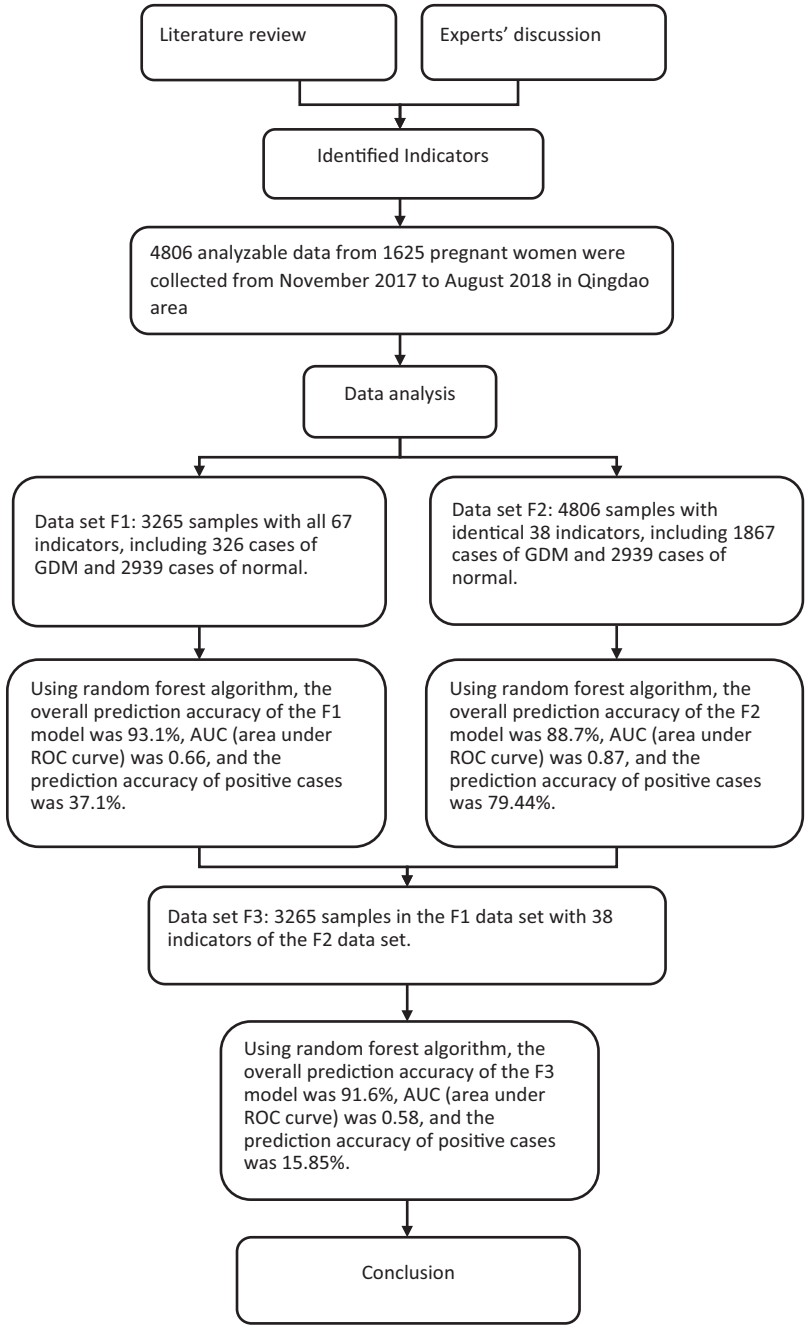


Figure 1. Technical route.

informed consent. The patients' IDs were used instead of their names throughout the study.

3. Results

3.1. Sample data and classification

A total of 4806 analyzable data from 1625 pregnant women were collected. Among these, 3265 samples with all 67 indicators were used to establish data set F1, including 326 cases of GDM and 2939 normal cases. Moreover, 4806 samples with 38 identical indicators were used to establish data set F2, including 1867 cases of GDM and 2939 normal cases.

To explore the impact of sacrificial indicators on GDM prediction, the F3 data set was established using 3265 samples (F1) with 38 indicators (F2), including 326 cases of GDM and 2939 normal cases.

3.2. Prediction results of training models for different data sets

A total of 3 prediction models of GDM were established for the different data sets by using a random forest regression algorithm. The results of the prediction models are shown in Table 2. The accuracy curves for the training set and test set of the 3 prediction models are shown in Figures 2–4.

3.3. Prediction performance of the random forest model

Confusion matrixes from the prediction results of the F1, F2, and F3 models are shown in Tables 3–5, respectively.

3.4. Variable weight in models

There are 67 input variables in data set F1, 58 of which were included in the random forest model. The 9 variables that had zero weight are as follows: thyroid disease, kidney disease, congenital spinal bifida, benign tumor, past surgical history, multiple pregnancy history, HIV, syphilis, and human papillomavirus (HPV). The weights of the other variables are shown in Figure 5.

There are 38 input variables in data set F2, 37 of which were enrolled in the random forest model, while 1 variable had zero weight (congenital spinal bifida). Sacrifice indicators are indicators of dietary habits and exercise habits. The weights of the top 20 variables are shown in Figure 6.

The F3 data set was established using 3265 samples (F1) with 38 input variables (F2), 31 of which were included in the random forest model. A total of 7 variables had zero weight, and they are as follows: kidney disease, congenital spinal bifida, benign tumor, past surgical history, HIV, syphilis, and HPV. The other variables are shown in Figure 7.

Data sets	Accuracy for training set	Accuracy for test set	AUC
F1	0.957	0.931	0.66
F2	0.911	0.887	0.87
F3	0.930	0.916	0.58

Note: AUC, area under the receiver operating characteristic curve.

Table 2. Results of prediction models using random forest.

4. Discussion

In the Python environment, we use the random forest algorithm to establish the prediction model, and the model trained by the F2 data set in this study shows the best stability. The accuracy both of training and test sets in the F1, F2, and F3 models tended to converge, and no overfitting occurred (Figures 2–4).

4.1. Best performance of the F2 prediction model

For more information and to obtain the most complete indicator information, 3265 samples with 67 indicators were enrolled in data set F1. The overall prediction accuracy of the F1 model training with the random forest was 93.10% and area under the receiver operating characteristic

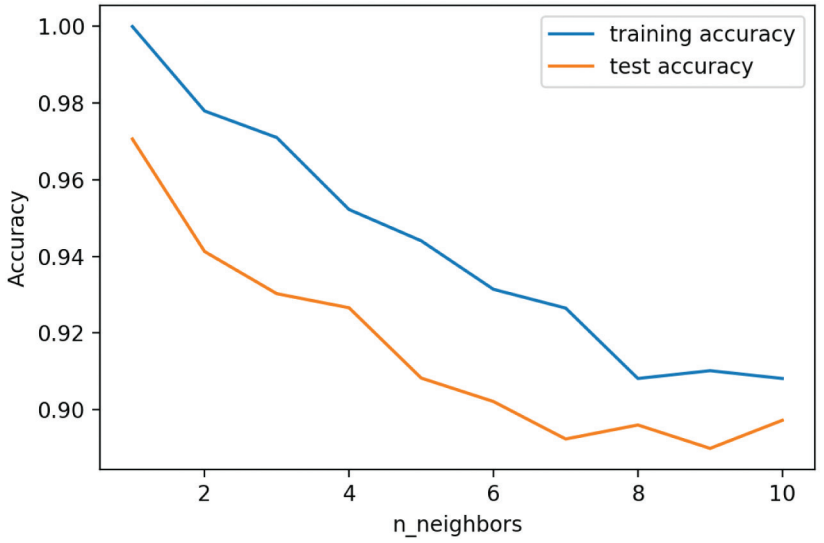


Figure 2. Accuracy curve of the training set and the test set of the F1 model.

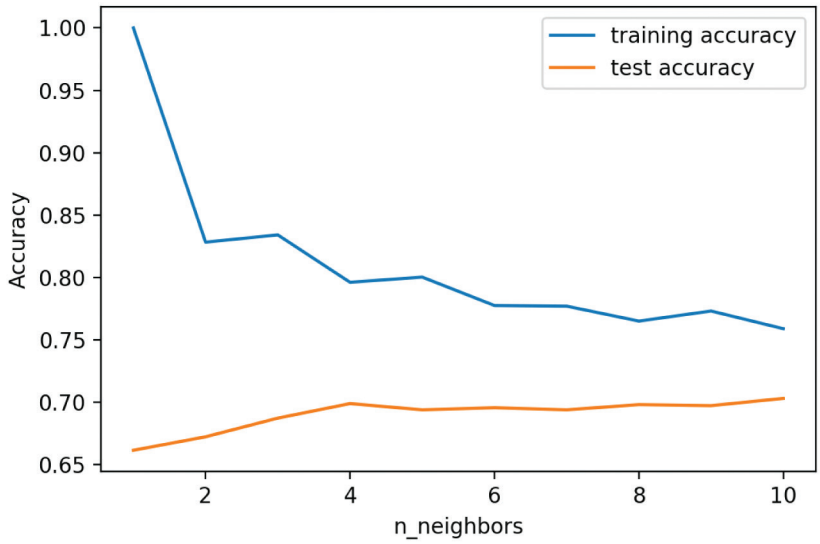


Figure 3. Accuracy curve of the training set and the test set of the F2 model.

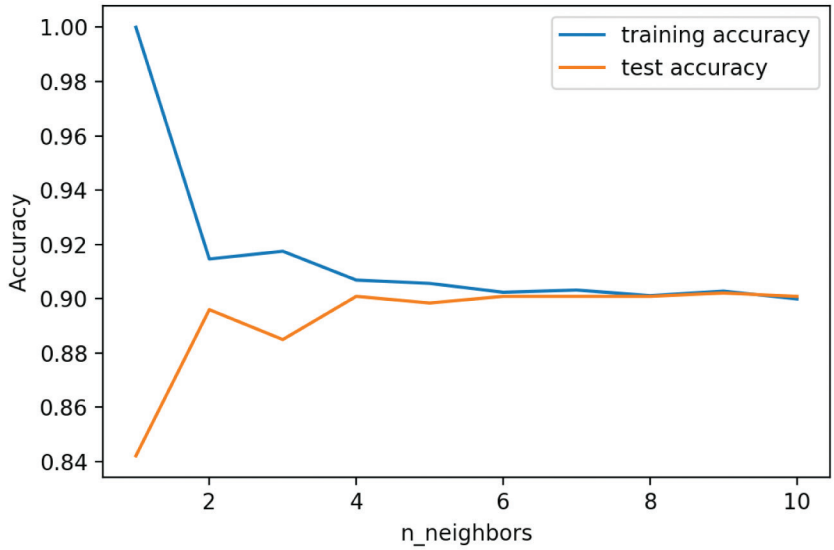


Figure 4. Accuracy curve of the training set and the test set of the F3 model.

Actual value	Total	Predicted value		Prediction accuracy (%)
		0	1	
0	735	735	0	100
1	82	56	26	31.70
In all	817			93.15

Table 3. Confusion matrix from prediction results of GDM incidence in the F1 model.

Actual value	Total	Predicted value		Prediction accuracy (%)
		0	1	
0	735	695	40	94.56
1	467	96	371	79.44
In all	1202			88.69

Table 4. Confusion matrix from prediction results of GDM incidence in the F2 model.

curve (AUC) was 0.66 (Table 2), indicating the model's good performance. However, the confusion matrix (Table 3) showed that the high accuracy was derived from 735 negative cases for which the prediction was correct in all cases; in contrast, the prediction was correct in only 26 out of 82 positive cases, with too many false-negative results.

The low accuracy of 93.10% and the low recognition rate of the model for positive cases may have been related to the large disparity in the numbers of GDM-positive/negative cases (3269/293) in the F1 model.

To utilize all of the available data for the analysis, 4806 samples with 38 indicators were enrolled in data set F2. The numbers of positive/negative cases of GDM increased (1867/2939), but some indicators were excluded, mainly those related to dietary and exercise habits (Table 1). The prediction accuracy of the F2 model was 88.70% and AUC was 0.87, indicating stable and reliable performance. The confusion matrix (Table 4) showed that the prediction was correct in 695 of 735 negative cases, while it was correct in 371 of 467 positive cases. The overall accuracy was thus 88.69%, indicating that the F2 model was more stable and reliable than the F1 model.

To explore the impact of sacrificial indicators on GDM prediction, the F3 data set was established using 3265 samples (F1) with 38 indicators (F2). The overall prediction accuracy of the F3 model was 91.60% and AUC was 0.58. The confusion matrix (Table 5) showed that the prediction was accurate in all 735 negative cases, but in only 13 out of 82 positive cases. The prediction accuracy of positive cases was only 15.85%, which was worse than those of both the F1 and F2 models. This

Actual value	Total	Predicted value		Prediction accuracy (%)
		0	1	
0	735	735	0	100
1	82	69	13	15.85
In all	817			91.55

Table 5. Confusion matrix from prediction results of GDM incidence in the F3 model.

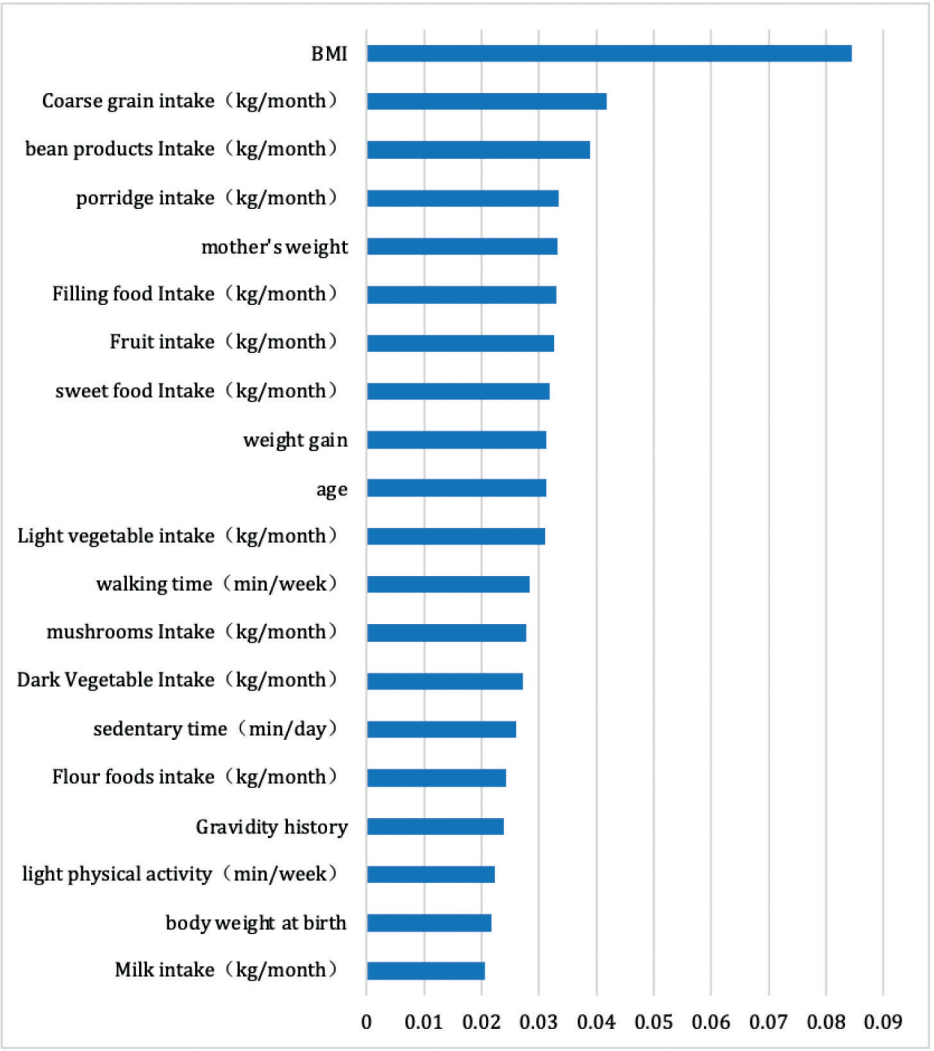


Figure 5. Weighted ranking in the F1 model (top 20).

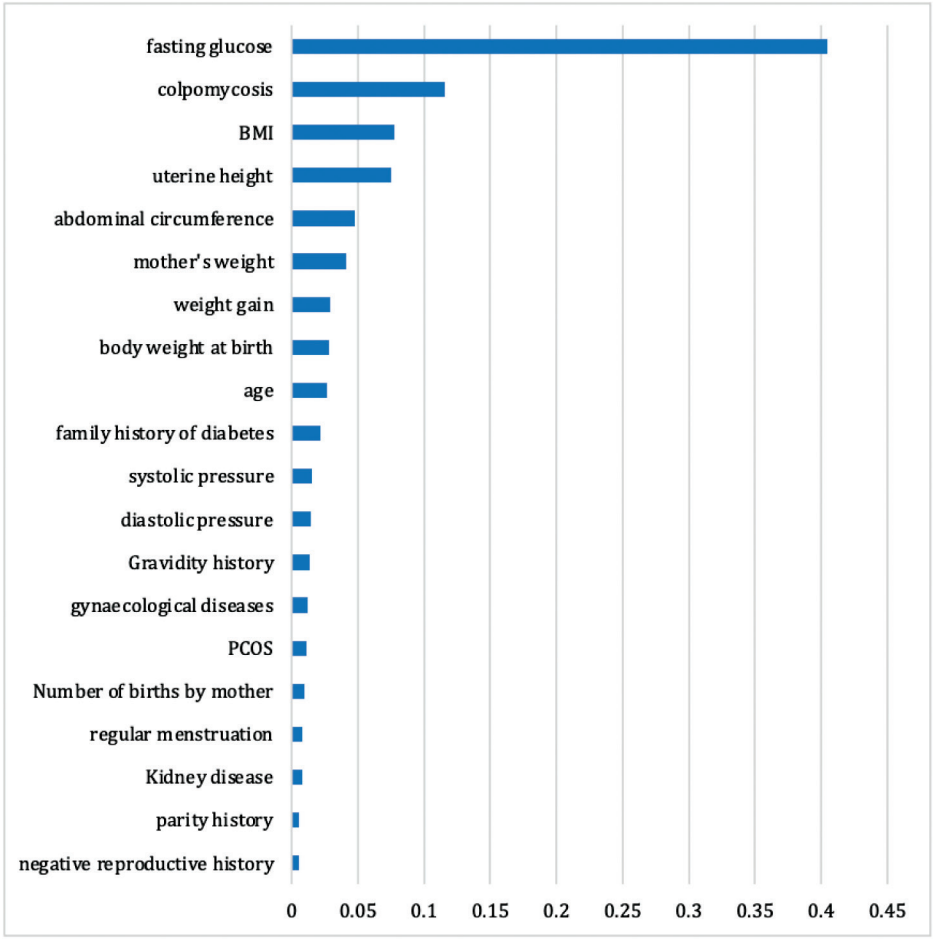


Figure 6. Weighted ranking in the F2 model (top 20).

indicated that the sacrifice indicators related to dietary and exercise habits would reduce the predictive performance of the model.

4.2. Comparison of variables captured by different models

A total of 58 variables obtained from the F1 model were correlated with the incidence of GDM, the top 20 of which were as follows in this model: BMI, coarse grain intake, bean product intake, porridge intake, mother's weight, filling food intake, fruit intake, sweet food intake, weight gain, age, light vegetable intake, walking time (min/week), mushroom intake, dark vegetable intake, sedentary time (min/day), flour food intake, gravidity history, light physical activity, body weight at birth, and milk intake (Figure 5).

Moreover, 37 variables obtained from the F2 model were correlated with the incidence of GDM, the top 20 of which were as follows: fasting glucose, colpomycosis, BMI, uterine height, abdominal circumference, mother's weight, weight gain, body weight at birth, age, family history of diabetes, systolic pressure, diastolic pressure, gravidity history, gynecological diseases, PCOS, number of births by mother, regular menstruation, kidney disease, parity history, and negative reproductive history.

A total of 31 variables obtained from the F3 model were correlated with the incidence of GDM, the top 20 of which were as follows: BMI, age, weight gain, mother's weight, abdominal circumference, fasting glucose, body weight at birth, systolic pressure, diastolic pressure, uterine height, gravidity history, parity history, number of births by mother, family history of diabetes, blood type B, blood type A, regular menstruation, negative reproductive history, blood type AB, and hepatitis.

Among these top 20 variables in the 3 models, there were 6 overlapping variables (BMI, mother's weight, weight gain, age and gravidity history, and body weight at birth), which also had similar rankings.

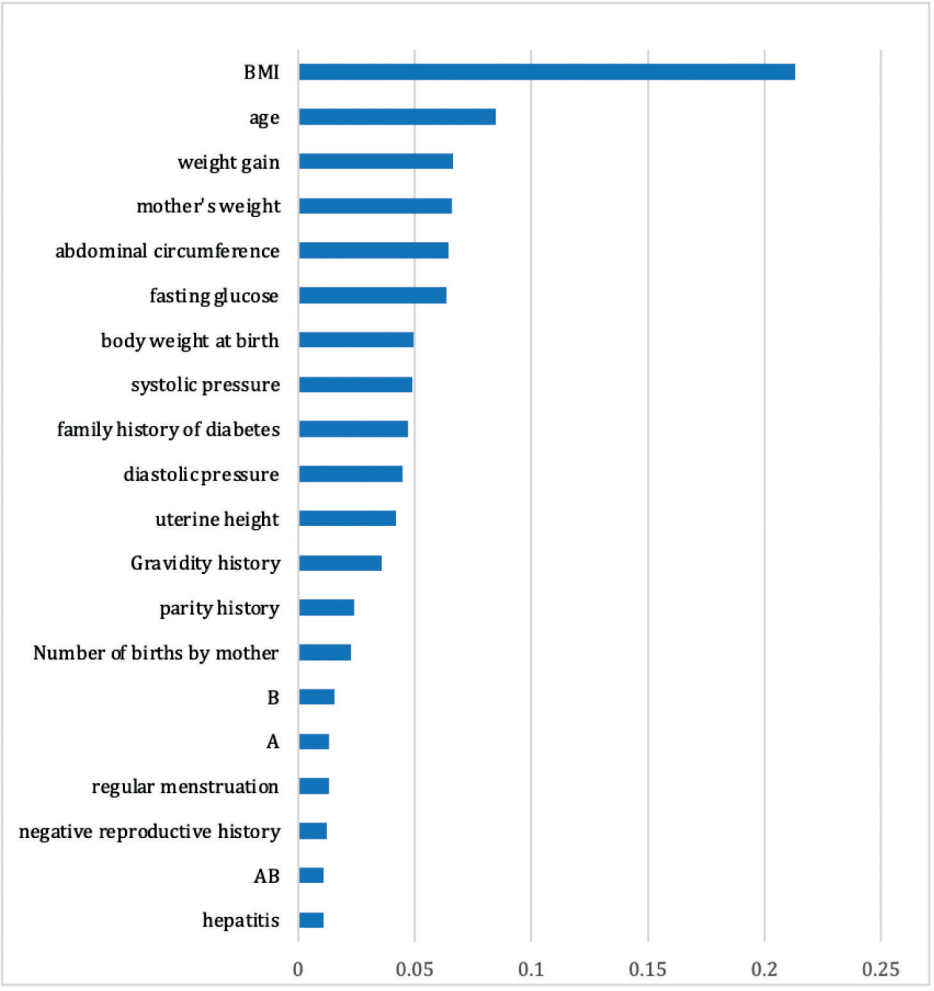


Figure 7. Weighted ranking in the F3 model (top 20).

4.2.1. Influence of pre-pregnancy BMI, pregnancy weight gain, age, number of pregnancies/ adverse pregnancy history

The influence of pre-pregnancy BMI, pregnancy weight gain, age, and the number of pregnancies/ adverse pregnancy history on GDM was consistent with the results of previous studies. Previous studies showed that overweight before pregnancy, weight gain during pregnancy, and old age were independent risk factors for GDM. The risk of GDM in the obese was also reported to be much higher than that in individuals of normal weight.[21] Even if the weight of pregnant women during 2 pregnancies was within the normal range, excessive weight gain during pregnancy (weight before the second birth minus weight before the first birth) remains an independent risk factor for GDM.[22] The risk of GDM was also found to be 1.74 times higher in pregnant women with a weight gain of ≥0.41 kg per week than that in pregnant women with a weight gain of ≤0.27 kg per week.[23] The rate of weight gain in early pregnancy had a greater impact on the occurrence of GDM. Older gestational age was also shown to be linked to a higher risk of GDM,[24] and the prevalence of GDM in 35-year-old pregnant women was reported to be 2.06 times higher than that in 25-year-old pregnant women.[25] Consistent with these results, the variables of BMI, weight gain, and age were shown to be of particular importance for predicting GDM in all models.

Studies have shown that a negative reproductive history such as history of spontaneous abortion, embryo sterilization, and preterm birth also increased the risk of GDM, which was 3 times higher than that of other pregnant women.[26] Consistent with previous studies, the variables of parity history and negative reproductive history were shown to be of particular importance for predicting GDM in all models.

4.2.2. Influence of birth weight and mothers' weight on the incidence of GDM

The influence of birth weight and mothers' weight on the incidence of GDM is consistent with the research hypothesis. In terms of genetic factors, there is a tendency for GDM to be inherited more from the mother than from the father, and there is a significant increase in the risk of metabolic-related diseases in offspring whose mother has been diagnosed with GDM. Therefore, when initiating this study, we proposed that body weight at birth and mother's weight might be highly correlated with the incidence of GDM during pregnancy. Consistent with this, the variables of body weight at birth and mother's weight were shown to be of particular importance for predicting GDM in all models, which is a major novel finding of this study.

4.2.3. Influence of diet and exercise variables on the incidence of GDM

The predictive performance of the F3 model was worse than that of the F1 model, indicating that sacrifice indicators related to dietary and exercise habits would reduce the predictive performance of the model.

The most influential indicators related to diet in the F1 model were as follows: coarse grain intake, bean product intake, porridge intake, filling food intake, fruit intake, sweet food intake, light vegetable intake, mushroom intake, dark vegetable intake, flour-based food intake, and milk intake. The most influential indicators related to exercise in the F2 model were as follows: walking time (min/week), sedentary time (min/day), and light physical activity.

Bao et al. followed up 21,079 pregnant women, 847 of whom were diagnosed with GDM.[27] They found that a greater number of times at which pregnant women consumed fried food every week was associated with a higher risk of GDM. Specifically, the risk of GDM was 2.18 times higher in pregnant women who consumed fried food more than 7 times a week than that in pregnant women who consumed it less than once a week. In this study, fried food intake ranked 55 among the 58 variables of the F2 model, which may be related to the different geographical origins of the data or the small sample size. It is necessary to further expand the sample size to improve the predictive capacity of the model.

Meng et al. found that excessive fruit intake (≥250 g/d) increased the risk of GDM, and daily exercise time of ≥30 min was a protective factor against GDM.[28] Pons et al. also reported that a sedentary lifestyle increases the risk of GDM.[29] Consistent with this, the variables of fruit intake, walking time, and sedentary time were shown to be of particular importance for predicting GDM in all models.

4.2.4. Other variables in prediction models

In the F2 and F3 models, fasting blood glucose showed high weight, which was consistent with our clinical experience. Consistent with previous studies, the risk of GDM increased with increasing waist circumference, although the variables of uterine height and abdominal circumference were affected by pregnant women's basic waist circumference, fetal growth rate, and amniotic fluid volume.[30,31] Hypertension, a family history of diabetes, and a history of PCOS, which were identified as strong predictors of GDM in all models, are also listed as high-risk factors for GDM in clinical practice guidelines.[32]

Some other variables (e.g., colpomycosis, kidney disease, number of births by the mother, regular menstruation, blood type, and hepatitis) that consistently ranked in the top 20 most influential factors were also found to be linked to GDM in this study. However, the effects of these variables on the incidence of GDM are still unclear due to the lack of relevant reports and the small sample size in this study; thus, further research on this issue is required.

5. Conclusions

If we take 67 indicators as input variables, it can be concluded that 3265 pieces of data would be obtained. The number of cases positive/negative for GDM was 326/2939. The predictive accuracy of the F1 model for positive cases was low. It is necessary to further expand the sample size of positive cases to improve the predictive performance of the model.

Based on Python, to acquire a more accurate model, a total of 4806 samples with 38 indicators were selected by the random forest algorithm, and the rate of GDM-positive/negative cases was increased to 1867/2939. The F2 model exhibited stable and reliable performance and high predictive accuracy, enabling its use for the prediction of GDM in early pregnancy.

The sacrifice of indicators related to dietary and exercise habits reduces the predictive performance of the model. The variables BMI, weight gain, age, parity history, negative reproductive history, fasting glucose, uterine height, abdominal circumference, hypertension, family history of diabetes, and PCOS were identified to be strongly predictive of GDM in all models, which is consistent with previous studies.

Consistent with the hypothesis of this study, the variables of body weight at birth and mother's weight were identified to be strongly predictive of GDM in all models. These 2 indicators can be analyzed in a future study to

further verify their potential to predict the emergence of GDM at an early stage.

Some other variables (e.g., colpomycosis, kidney disease, number of births by the mother, regular menstruation, blood type, and hepatitis) that consistently ranked in the top 20 most influential factors were also found to be linked to GDM in this study However, the effects of these variables on the incidence of GDM are still unclear due to a lack of relevant reports; thus, further research on this issue is required.

In the future, the F2 prediction model will be applied to clinical practice to screen for extremely high-risk cases and carry out early interventions. Early warning and intervention can markedly decrease the occurrence of GDM, saving health education and intervention costs. We plan to continue the cohort study in the future, including more variables such as the dietary and exercise habits of pregnant women. We also plan to include more samples to improve the predictive performance of the F2 model.

Limitations

The prediction model developed in this study has certain limitations in terms of its accuracy and applicability for various reasons. For example, uterine height, abdominal circumference, fasting blood sugar, and other indicators may vary among different examination times. In addition, owing to the fewer indicators of dietary and exercise habits of pregnant women in GDM-positive cases, the predictive accuracy of the F1 model for positive cases was low. Finally, the samples were only derived from Qingdao, but there may be regional differences in the data.

Acknowledgments

We thank Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

Ethical approval

The study was approved by the medical ethics committee of the study principal investigator's hospital. All patients were informed about the purposes and methods of the study during the recruitment. Participation was voluntary, and patients could refuse to participate or withdraw at any time. Each participant provided written informed consent. The patients' IDs were used instead of their names throughout the study.

Conflicts of interest

All contributing authors declare that no conflicts of interest exist.

References

1. Obstetrics Subgroup, Chinese Society of Obstetrics and Gynecology, Chinese Medical Association, Group of Pregnancy with Diabetes Mellitus, Chinese Society of Perinatal Medicine, Chinese Medical Association. Diagnosis and therapy guideline of pregnancy with diabetes mellitus. *Chin J Obstet Gynecol*. 2014;49:561–569 (in Chinese).
2. Moon JH, Kwak SH, Jang HC. Prevention of type 2 diabetes mellitus in women with previous gestational diabetes mellitus. *Korean J Intern Med*. 2017;32:26–41.
3. Avagliano L, Massa V, Terraneo L, et al. Gestational diabetes affects fetal autophagy. *Placenta*. 2017;55:90–93.
4. Eades CE, Cameron DM, Evans JMM. Prevalence of gestational diabetes mellitus in Europe: A meta-analysis. *Diabetes Res Clin Pract*. 2017;129:173–181.
5. Wu JF, Wu SQ, Liu Z, Xu H, Zhang L. Influence of new diagnostic criteria for gestational diabetes mellitus on the outcome of perinatal infants. J Pub Health Prev Med.2015;26:121-124 (in Chinese).
6. Farrar D, Simmonds M, Griffin S, et al. The identification and treatment of women with hyperglycaemia in pregnancy: an analysis of individual participant data, systematic reviews, meta-analyses and an economic evaluation. *Health Technol Assess*. 2016;20:1–348.
7. Wilson ML. Prediabetes: beyond the Borderline. *Nurs Clin North Am*. 2017;52:665–677.
8. Zand A, Ibrahim K, Patham B. Prediabetes: Why should we care?. *Methodist Debakey Cardiovasc J*. 2018;14:289–297.
9. Booth GL, Luo J, Park AL, et al. Influence of environmental temperature on risk of gestational diabetes. CMAJ. 2017;189:E682–E689.
10. Xu X, Zhu XW, Jiang YM, et al. Study on the incidence and risk factors of gestational diabetes in 2748 hospitalized pregnant women. Acta Universitatis Medicinalis Nanjing (Natural Science). 2015;35:695-698 (in Chinese).
11. Sahu L, Satyakala R, Rani R. Comparison of the American Diabetes Association and World Health

Organization criteria for gestational diabetes mellitus and the outcomes of pregnancy. Obstet Med. 2009;2:149–153.
12. National Collaborating Centre for Women's and Children's Health. Diabetes in pregnancy: management of diabetes and its complications from preconception to the postnatal period. https://www.nice.org.uk/guidance/ng3. Accessed September7, 2019.
13. Esmaily H, Tayefi M, Doosti H, et al. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *J Res Health Sci*. 2018;18:e00412.
14. DuBrava S, Mardekian J, Sadosky A, et al. Using random forest models to identify correlates of a diabetic peripheral neuropathy diagnosis from Electronic Health Record Data. *Pain Med*. 2017;18:107–115.
15. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.
16. Wu BH, Huang HY, Yao Q, Deng R, Li H, Liang T. The Application of Big Data and Artificial Intelligence Methods in Prediction of GDM. *Chin J Health Inf Manag*. 2017;14:832-837 (in Chinese).
17. McGowan CA, Curran S, McAuliffe FM. Relative validity of a food frequency questionnaire to assess nutrient intake in pregnant women. *J Hum Nutr Diet*. 2014;27 Suppl 2:167–174.
18. Li Q. Effects of cognitive viewpoint, dietary habits and physical activity on blood glucose in pregnant women in Guangdong: Southern Medical University; 2013 (in Chinese).
19. Ogawa K, Jwa SC, Kobayashi M, Morisaki N, Sago H, Fujiwara T. Validation of a food frequency questionnaire for Japanese pregnant women with and without nausea and vomiting in early pregnancy. *J Epidemiol*. 2017;27:201–208.
20. Rigatti SJ. Random forest. *J InsurMed (New York, NY)*. 2017;47:31–39.
21. Feng R, Shao H, Zhu H, Xie J. Meta analysis of risk factors for gestational diabetes mellitus. *Matern Child Health Care China*. 2014;29:2824–2827.
22. Sorbye LM, Skjaerven R, Klungsoyr K, Morken NH. Gestational diabetes mellitus and interpregnancy weight change: a population-based cohort study. PLoS Med. 2017;14:e1002367.
23. Moore Simas TA, Waring ME, Callaghan K, et al. Weight gain in early pregnancy and risk of gestational diabetes mellitus among Latinas. *Diabetes Metab*. 2019;45:26–31.
24. Collier A, Abraham EC, Armstrong J, Godwin J, Monteath K, Lindsay R. Reported prevalence of gestational diabetes in Scotland: the relationship with obesity, age, socioeconomic status, smoking and macrosomia, and how many are we missing? *J Diabetes Investig*. 2017;8:161–167.
25. Wang S, Yang HX. Analysis of the effect of risk factors at gestational diabetes mellitus. *Chin J Obstet Gynecol*. 2014;49:321–324 (in Chinese).
26. Zhang X. Analysis of the factors influencing the incidence of gestational diabetes mellitus and adverse pregnancy outcome Qinhuangdao: North China University of Science and Technology; 2016 (in Chinese).
27. Bao W, Tobias DK, Olsen SF, Zhang C. Pre-pregnancy fried food consumption and the risk of gestational diabetes mellitus: a prospective cohort study. Diabetologia. 2014;57:2485–2491.
28. Meng LP, Xiao LF, Geng LR, Wang CP. High risk scoring for prediction of gestational diabetes mellitus. *J Nurs Sci*. 2015;30:7–10 (in Chinese).
29. McIntyre HD, Catalano P, Zhang C, Desoye G, Mathiesen ER, Damn P. Gestational diabetes mellitus. *Nat Rev Dis Primers*. 2019;5:47.
30. Han Q, Shao P, Leng J, et al. Interactions between general and central obesity in predicting gestational diabetes mellitus in Chinese pregnant women: a prospective population-based study in Tianjin, China. *J Diabetes*. 2018;10:59–67.
31. Bolognani CV, Reis LBSM, de Souza SS, Dias A, Rudge MVC, Calderon IMP. Waist circumference in predicting gestational diabetes mellitus. *J Matern Fetal Neonatal Med*. 2014;27:943–948.
32. ACOG Practice Bulletin No. 190: Gestational Diabetes Mellitus. *Obstet Gynecol*. 2018;131: e49–e64.