International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Advanced Machine Learning did not Surpass Traditional Logistic Regression in First-Trimester Gestational Diabetes Mellitus Prediction: A Retrospective Single-Center Study From Eastern China

Hongyan Ni[1], Jinli Miao[2], Jian Chen[3]

[1]Department of maternity care, PingHu Maternal and Child Health Hospital, Jiaxing, Zhejiang, 314200, People's Republic of China; [2]The Yangtze River Delta Biological Medicine Research and Development Center of Zhejiang Province, Yangtze Delta Region Institution of Tsinghua University, Hangzhou, Zhejiang, 314006, People's Republic of China; [3]Department of internal medicine, PingHu Maternal and Child Health Hospital, Jiaxing, Zhejiang, 314200, People's Republic of China

Correspondence: Jian Chen, Email 15988398470@163.com

Background: Gestational diabetes mellitus (GDM) poses serious health risks to both mothers and fetuses. However, effective tools for identifying GDM are lacking. This study, based on a Chinese cohort, aims to construct and compare the predictive performance of traditional logistic regression (LR) and six advanced machine learning (ML) models, thereby aiding in the early identification and intervention of GDM.

Methods: This retrospective study utilized medical examination data from 956 singleton pregnant women collected between January and December 2023 from ten maternal and child health hospitals in Pinghu City. We employed receiver operating characteristic curves and precision-recall curves to assess the predictive performance of the models. Decision curve analysis (DCA) was used to evaluate clinical utility, while calibration curves and Hosmer-Lemeshow (HL) tests were applied to assess the calibration of each model.

Results: The 956 participants were randomly divided into a training set and a validation set at a 3:1 ratio. We identified 13 features through Spearman correlation analysis and the Boruta algorithm to construct the models. The LR model exhibited the best AUC at 0.787 (0.723–0.85), outperforming the seven other ML models including RF at 0.776 (0.711–0.841). Furthermore, the LR model showed good calibration and clinical utility.

Conclusion: Although ML has tremendous potential, in predicting the occurrence of GDM based on common early pregnancy data, the ML models did not completely outperform the traditional LR model. Simpler, traditional models may be more effective than complex ML approaches.

Keywords: GESTATIONAL diabetes mellitus, logistic regression, machine learning, first trimester, prediction model

Introduction

Gestational diabetes mellitus (GDM) is a metabolic syndrome characterized by abnormal elevations in blood glucose levels during pregnancy. Although it typically resolves after childbirth, GDM poses significant short-term and long-term health risks to both mother and child.[1] The incidence of GDM exhibits substantial variation across different populations and has shown a consistent upward trend.[2] In the United States, the prevalence of GDM increased from 4.6% to 8.2% between 2006 and 2016, representing a 78% relative increase. This rise was particularly pronounced among Hispanic, non-Hispanic Black women, and women of other races/ethnicities compared to non-Hispanic White women.[3] The observed disparities in GDM susceptibility across different racial groups can be attributed to a combination of genetic predisposition, lifestyle factors, and socioeconomic determinants, which contribute to significant variations in incidence

Received: 18 December 2024
Accepted: 19 April 2025
Published: 26 April 2025

rates across regions and ethnicities.[4,5] Similarly, China has experienced a rising prevalence of GDM, influenced by rapid economic development, lifestyle modifications, and changes in fertility policies.[6] The clinical implications of GDM are substantial. For mothers, GDM can increase the risks of hypertension, preeclampsia, and type 2 diabetes. For fetuses, GDM can induce macrosomia, preterm birth, difficult labor, and stillbirth.[7,8]

Current diagnostic protocols typically occur during the second and third trimesters, with no universally accepted gold standard. The International Association of Diabetes and Pregnancy Study Groups (IADPSG) recommends a one-step screening approach involving a 75 g oral glucose tolerance test (OGTT) with measurements at fasting, 1 h, and 2 h intervals.[9] In contrast, the American College of Obstetricians and Gynecologists (ACOG) advocates a two-step Carpenter–Coustan approach, beginning with a non-fasting 50 g OGTT followed by diagnostic 100 g OGTT if initial results exceed threshold values.[10] While both methods are clinically valuable, they are time-intensive and present challenges; notably, the one-step approach may carry a higher risk of false-positive diagnoses compared to the two-step method.[11,12] Emerging evidence indicates that fetal growth abnormalities, particularly excessive growth, may particularly.[13] Furthermore, animal studies demonstrate that insulin treatment following GDM diagnosis in mouse models fails to fully protect offspring from metabolic disorders induced by diet in adulthood.[14] These findings underscore the importance of early intervention, as first-trimester management has been shown to reduce GDM risk and promote optimal fetal development.[15] Therefore, the development of early diagnostic methodologies for GDM represents a critical area of research for improving maternal and fetal outcomes.

Previous studies have identified multiple risk factors associated with GDM onset, including advanced maternal age, pre-pregnancy body mass index (BMI), family history of diabetes, history of macrosomia, and thyroid function.[16–18] In predictive modeling, logistic regression (LR) remains a fundamental statistical approach for disease prediction. Concurrently, machine learning (ML), as advanced artificial intelligence methodologies, are increasingly being applied in disease prediction research. Several investigators have attempted to develop LR and ML models for early GDM prediction.[19–23] However, there remains a paucity of models specifically developed and validated for the Chinese population. In this study, we report the development and validation of traditional LR and six ML models based on a Chinese cohort to, and compare their performance from various aspects.

Materials and Methods

Study Population and Data Collection

This retrospective study analyzed medical data from 956 singleton pregnancies between January and December 2023. The data were collected from a network of ten healthcare facilities in Pinghu, China, with Pinghu Maternal and Child Health Hospital serving as the primary coordinating center. The collaborating institutions included: Pinghu Lindai Town Health Center, Pinghu Xindai Town Central Health Center, Pinghu Caoqiao Sub-district Community Health Service Center, Pinghu Zhapu Town Central Health Center, Pinghu Xincang Town Central Health Center, Pinghu Zhongdai Sub-district Community Health Service Center, Pinghu Danghu Sub-district Community Health Service Center, Pinghu Dushangang Town Central Health Center, and Pinghu Guangchen Town Health Center. We retrospectively collected comprehensive clinical variables, including demographic data and laboratory test results. All data, except for OGTT and fasting plasma glucose (FPG) measurements, were obtained through patient interviews and clinical examination before 12th weeks of pregnancy. The inclusion criteria comprised: (1) singleton pregnancy; (2) undergoing a 75 g OGTT or FPG test between the 24th to 28th weeks of pregnancy at our hospitals. We excluded women who had pre-existing diabetes prior to pregnancy.

Diagnosis of GDM

According to the 2010 IADPSG recommendations for GDM diagnosis, a diagnosis can be made if any of the following glucose values are exceeded during a 75 g OGTT conducted in a fasting state between 24–28 weeks of pregnancy: 0 h $\geq$ 5.1 mmol/L, 1 h $\geq$ 10 mmol/L, 2 h $\geq$ 8.5 mmol/L.[24] If an OGTT was not performed, GDM can be directly diagnosed based on the World Health Organization (WHO) 2013 standards, where a mid-pregnancy FPG level $\geq$ 5.1 mmol/L qualifies.[25]

Data Pre-Processing

Missing data were systematically addressed through a rigorous preprocessing protocol. Variables exhibiting missing values exceeding 30% of the total observations, including ferritin, beta-2 microglobulin, insulin resistance (HOMA-IR), and glycated hemoglobin (HbA1c), were excluded from subsequent analyses. For the remaining features with incomplete data, we implemented multiple imputation using a random forest algorithm through the "mice" package (version 3.14.0) in R statistical software. To prevent dimensionality disaster, we assessed multicollinearity among all features by calculating Spearman correlation coefficients, removing redundant features to ensure the stability of subsequent models. Features with absolute correlation coefficients greater than 0.6 and a p-value less than 0.05 were considered significantly associated.

Model Development and Validation

In this study, we performed feature pre-selection on the training set samples using the Boruta algorithm, a feature selection method based on random forests.[26] This algorithm operates by creating shadow features (randomly shuffled copies of the original features) and evaluating feature importance through the random forest algorithm to identify truly significant original features. The pre-selected features were subsequently utilized as input variables for seven predictive models, including LR, eXtreme gradient boosting (XGB), light gradient boosting (LGBM), multi-layer perceptron (MLP), k-nearest neighbors (KNN), random forest (RF), and support vector machine (SVM). For the LR model, we implemented 10-fold cross-validation with 10,000 iterations, resetting the random seed each time to ensure randomness. During each iteration, we calculated the area under the receiver operating characteristic (AUC) curve value for the validation set, retaining the logistic regression model demonstrating the highest AUC. Regarding the 6 machine learning models, we optimized each model by Bayesian optimization and used five-fold cross-validation to select a set of hyperparameters that have the largest area under the curve (AUC) of the subjects' work receiver operating characteristic curve (ROC) in the training set to obtaining optimal performance. These models were then validated on the validation set and the results were compared. Specific hyperparameters are detailed in the Supplementary Material. To ensure robustness, all models underwent five-fold cross-validation on the training and validation sets to ensure robustness. We plotted the receiver operating characteristic (ROC) curve and precision-recall curve (PRC) for each model on the validation set to evaluate predictive performance. Decision curve analysis (DCA) was used to assess clinical utility. Calibration curves and the Hosmer-Lemeshow (HL) validation were used to evaluate the calibration of each model. Additionally, we calculated standard performance metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for all models.

Statistical Analysis

The data from 956 patients were randomly divided into a training set (75%) and a validation set (25%) in a 3:1 ratio. For continuous variables, normality was first assessed. Data following a normal distribution were described using mean ± standard deviation (SD) and compared between groups using independent samples *t*-tests. Non-normally distributed data were described using median and interquartile range and compared using non-parametric tests. For categorical variables, the frequency of each category was calculated, and the percentage of each category within each group was determined (n, %). The chi-square test was then used to compare the distribution differences of categorical variables between groups. P-value < 0.05 was considered statistically significant. All statistical analyses were performed using R software (version 4.3.0) and Python (version 3.8.18).

Results

Population Characteristics

Figure 1A illustrates the missing data distribution for features with less than 30% missing values. We conducted a Spearman correlation analysis to identify and eliminate redundant features. The analysis revealed strong correlations between the following feature pairs: pre-pregnancy BMI and maternal weight, gravidity and abortion, alanine aminotransferase (ALT) and aspartate aminotransferase (AST), total bilirubin and conjugated bilirubin, high-density lipoprotein

https://doi.org/10.2147/IJGM.S513064

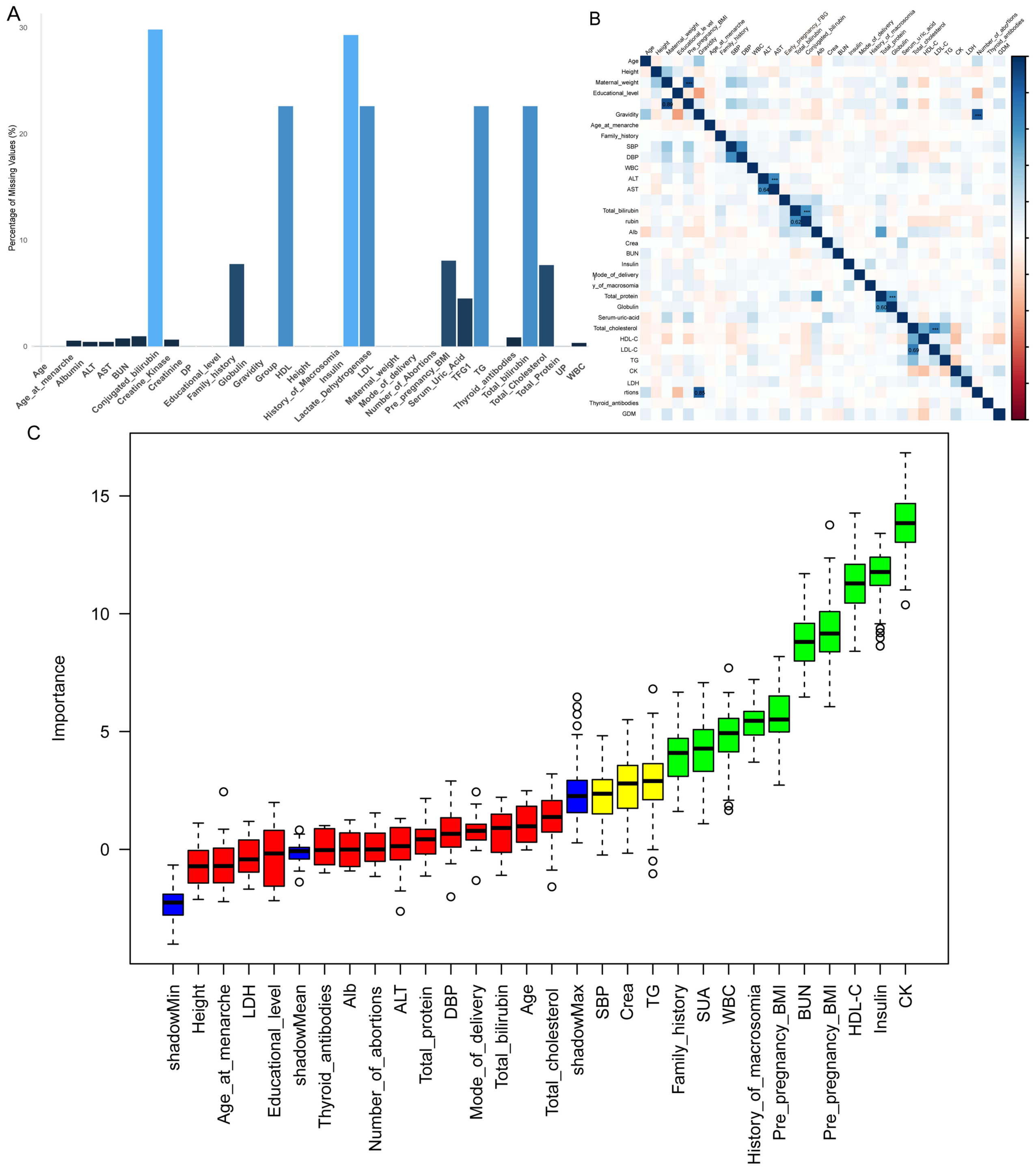

Figure 1 Feature selection workflow for first-trimester GDM prediction models. (**A**) Missing value distribution across candidate clinical parameters. (**B**) Spearman's rank correlation matrix of analyzed biomarkers. (**C**) Boruta algorithm-driven feature importance ranking.

Note: ***Indicates that the correlation P value is < 0.001.

Abbreviations: BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; WBC, white blood cells; ALT, alanine aminotransferase; FBG, fasting blood glucose; Alb, albumin; Crea, creatinine, BUN; SUA, serum uric acid; HDL-C, high density lipoprotein cholesterol; TG, triglycerides; CK, creatine kinase; LDH, lactate dehydrogenase.

cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C), and total protein with globulin (Figure 1B). Based on clinical relevance and missing data patterns, we excluded following redundant features: maternal weight, gravidity, AST, conjugated bilirubin, LDL-C, and globulin as redundant features.

The study population of 956 pregnant women was randomly divided into a training set (N=717) and a validation set (N=239). As shown in Table 1, no statistically significant differences were observed in any clinical factors between the training and validation sets (P > 0.05). Table 2 presents the distribution and comparison of clinical variables between non-GDM and GDM participants in the training set. Compared to the non-GDM group, the GDM group showed significantly higher values for the following parameters: maternal age, pre-pregnancy BMI, systolic blood pressure

Table 1 Clinical Variables Grouped by Training and Validation Dataset

Variables	Training Dataset (n=717)	Validation Dataset (n=239)	P
GDM			0.453
No	471 (65.69%)	164 (68.62%)	
Yes	246 (34.31%)	75 (31.38%)	
Age	28 (26, 31)	29 (26, 32)	0.658
Height	160 (157, 164)	160 (157, 163)	0.700
Pre pregnancy BMI	21.63 (19.88, 23.88)	21.23 (19.78, 24.18)	0.523
Number of abortions	0 (0, 1)	0 (0, 1)	0.901
Age at menarche			0.434
>11	17 (2.37%)	3 (1.26%)	
≤11	700 (97.63%%)	236 (98.74%)	
Educational level			0.965
Primary school	16 (2.23%)	7 (2.93%)	
Middle school	176 (24.55%)	58 (24.27%)	
High school	101 (14.09%)	31 (12.97%)	
Junior college	174 (24.27%)	58 (24.27%)	
Bachelor	248 (34.59%)	85 (35.56%)	
Master	2 (0.28%)	0 (0%)	
Family history			0.984
No	661 (92.19%)	220 (92.05%)	
Hypertension	36 (5.02%)	13 (5.44%)	
Diabetes	9 (1.26%)	3 (1.26%)	
Hypertension & Diabetes	11 (1.53%)	3 (1.26%)	
History of Macrosomia			1.000
No	708 (98.74%)	236 (98.74%)	
Yes	9 (1.26%)	3 (1.26%)	
Insulin			1.000
No	708 (98.74%)	236 (98.74%)	
Yes	9 (1.26%)	3 (1.26%)	
Mode of delivery			0.055
Normal vaginal delivery	368 (51.32%)	138 (57.74%)	
Cesarean section	315 (43.93%)	84 (35.15%)	
Vacuum extraction	18 (2.51%)	5 (2.09%)	
Forceps delivery	15 (2.09%)	11 (4.6%)	
Breech delivery	1 (0.14%)	1 (0.42%)	
Thyroid antibodies			0.911
No	623 (86.89%)	209 (87.45%)	
Yes	94 (13.11%)	30 (12.55%)	
SBP	108 (100, 116)	108 (100, 120)	0.306
DBP	70 (62, 75)	70 (63, 78)	0.137
WBC	8.1 (6.76, 9.23)	8 (6.75, 9.4)	0.720
ALT	14 (10, 20)	14 (10, 20.5)	0.827
Early pregnancy FBG	4.96 (4.6, 5.2)	4.9 (4.6, 5.18)	0.541
Total bilirubin	9.2 (6.9, 12.2)	9.4 (7.25, 11.8)	0.619

(*Continued*)

Table 1 (Continued).

Variables	Training Dataset (n=717)	Validation Dataset (n=239)	P
Alb	44.6 (42.4, 46.8)	45 (42.6, 47.6)	0.078
Crea	45 (40, 49.5)	44.9 (40, 49)	0.294
BUN	2.9 (2.4, 3.5)	2.8 (2.38, 3.3)	0.074
Total protein	72.1±5.33	72.8±5.03	0.095
SUA	218 (186.8, 253)	70 (63, 78)	0.331
Total cholesterol	5.0 (4.36, 5.82)	5.1 (4.34, 5.84)	0.946
HDL-C	1.73 (1.4, 2.07)	1.75 (1.42, 2.1)	0.744
TG	1.61 (1.15, 2.22)	1.60 (1.21, 2.19)	0.860
CK	42 (32, 55)	41 (33, 52.5)	0.832
LDH	150 (135, 164)	148 (133, 166.5)	0.724

Abbreviations: GDM, Gestational diabetes mellitus; BMI, Body mass index; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; WBC, White blood cells; ALT, Alanine aminotransferase; FBG, Fasting blood glucose; Alb, Albumin; Crea, Creatinine, BUN; SUA, Serum uric acid; HDL-C, High density lipoprotein cholesterol; TG, Triglycerides; CK, Creatine kinase; LDH, Lactate dehydrogenase.

Table 2 Clinical Variables of Participants With GDM and Non-GDM Participants in the Training Dataset

Variables	Non-GDM (n=471)	GDM (n=246)	P
Age	28 (26, 31)	29 (27, 32)	**0.007**
Height	160 (156, 165)	160 (157.12, 163)	0.484
Pre pregnancy BMI	21.3 (19.53, 23.44)	22.47 (20.57, 24.67)	**<0.001**
Number of abortions	0 (0, 1)	0 (0, 1)	0.594
Age at menarche			1.000
>11	11 (2.34%)	6 (2.44%)	
≤11	460 (97.66%)	240 (97.56%)	
Educational level			0.859
Primary school	12 (2.55%)	4 (1.63%)	
Middle school	114 (24.2%)	62 (25.2%)	
High school	68 (14.44%)	33 (13.41%)	
Junior college	109 (23.14%)	65 (26.42%)	
Bachelor	167 (35.46%)	81 (32.93%)	
Master	1 (0.21%)	1 (0.41%)	
Family history			0.232
No	441 (93.63%)	220 (89.43%)	
Hypertension	20 (4.25%)	16 (6.5%)	
Diabetes	4 (0.85%)	5 (2.03%)	
Hypertension & Diabetes	6 (1.27%)	5 (2.03%)	
History of Macrosomia			**0.001**
No	471 (100%)	237 (96.34%)	
Yes	0 (0%)	9 (3.66%)	
Insulin			**<0.001**
No	471 (100%)	237 (96.34%)	
Yes	0 (0%)	9 (3.66%)	
Mode of delivery			0.201
Normal vaginal delivery	244 (51.8%)	124 (50.41%)	
Cesarean section	204 (43.31%)	111 (45.12%)	
Vacuum extraction	15 (3.18%)	3 (1.22%)	
Forceps delivery	8 (1.7%)	7 (2.85%)	
Breech delivery	0 (0%)	1 (0.41%)	

(*Continued*)

Table 2 (Continued).

Variables	Non-GDM (n=471)	GDM (n=246)	P
Thyroid antibodies			0.861
No	408 (86.62%)	215 (87.4%)	
Yes	63 (13.38%)	31 (12.6%)	
SBP	107 (100, 114)	110 (100, 119)	**0.002**
DBP	70 (62, 75)	70 (64, 75.75)	0.294
WBC	7.9 (6.7, 9.1)	8.28 (6.93, 9.7)	**0.007**
ALT	14 (10, 20)	14 (10, 20.75)	0.969
Early pregnancy FBG	4.9 (4.59, 5.1)	5 (4.7, 5.4)	**<0.001**
Total bilirubin	9.2 (7, 11.95)	9.15 (6.7, 12.6)	0.926
Alb	44.6 (42.35, 47)	44.6 (42.45, 46.7)	0.764
Crea	45.4 (41, 50)	44 (40, 49)	**0.041**
BUN	3 (2.5, 3.5)	2.74 (2.22, 3.45)	**0.001**
Total protein	71.7 (68, 75.75)	72 (69.38, 75.27)	0.362
SUA	215 (183, 248.5)	222 (192.48, 258.5)	**0.009**
Total cholesterol	5.04 (4.39, 5.89)	4.9 (4.27, 5.69)	**0.036**
HDL-C	1.8 (1.46, 2.16)	1.58 (1.31, 1.88)	**<0.001**
TG	1.59 (1.19, 2.2)	1.65 (1.09, 2.27)	0.813
CK	39 (31, 52)	47 (37, 61.75)	**<0.001**
LDH	149 (132, 164)	151 (138, 165)	0.066

Note: Values with P < 0.05 are bolded in the table.

Abbreviations: GDM, Gestational diabetes mellitus; BMI, Body mass index; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; WBC, White blood cells; ALT, Alanine aminotransferase; FBG, Fasting blood glucose; Alb, Albumin; Crea, Creatinine, BUN; SUA, Serum uric acid; HDL-C, High density lipoprotein cholesterol; TG, Triglycerides; CK, Creatine kinase; LDH, Lactate dehydrogenase.

(SBP), white blood cells (WBC), first-trimester FPG, serum uric acid (SUA), and creatine kinase (CK), and had more participants with a history of macrosomia and insulin use during pregnancy. Creatinine (Crea), blood urea nitrogen (BUN), total cholesterol, and HDL-C levels were significantly lower than non-GDM group.

Predictive Performance of The Models

After eliminating features with missing data >30% and redundant features, we assessed the importance of the remaining 27 features using the Boruta algorithm. Ultimately, 13 features were deemed relevant (Figure 1C). We developed seven models employing six ML techniques and traditional LR based on the 13 features. All models were adjusted on the training set and internally validated in the validation set. Figure 2 shows the ROC curves of all models in the validation set. The LR model had the highest AUC of 0.787 (0.723–0.85), followed by RF, SVM, XGB, and LGBM, with AUCs of 0.776 (0.711–0.841), 0.771 (0.702–0.84), 0.757 (0.687–0.826), and 0.743 (0.672–0.815), respectively. Both MLP and KNN models did not achieve an AUC above 0.7. The PRC indicated that the LR had the highest area under the PRC (AUPR) of 0.644. The LR model showed higher precision at both low and high recall intervals and comparable precision at medium recall intervals with other ML models (Figure 3A). Among the seven models, the LR model performed well in terms of accuracy, sensitivity, PPV, and NPV (Table 3). However, all seven models had relatively low specificity (0.08 to 0.387), although the LR model was at a relatively higher level (specificity=0.360).

Clinical Utility of The Models

Figure 3B illustrates the net benefit across different thresholds for all models. The threshold range for the KNN model was the narrowest. The threshold range for the LR model was slightly narrower than for LGBM and XGB models but similar to other ML models. The LR model maintained a broad range, approximately from 0.1 to 0.8, and its net benefit was comparable to several ML models.

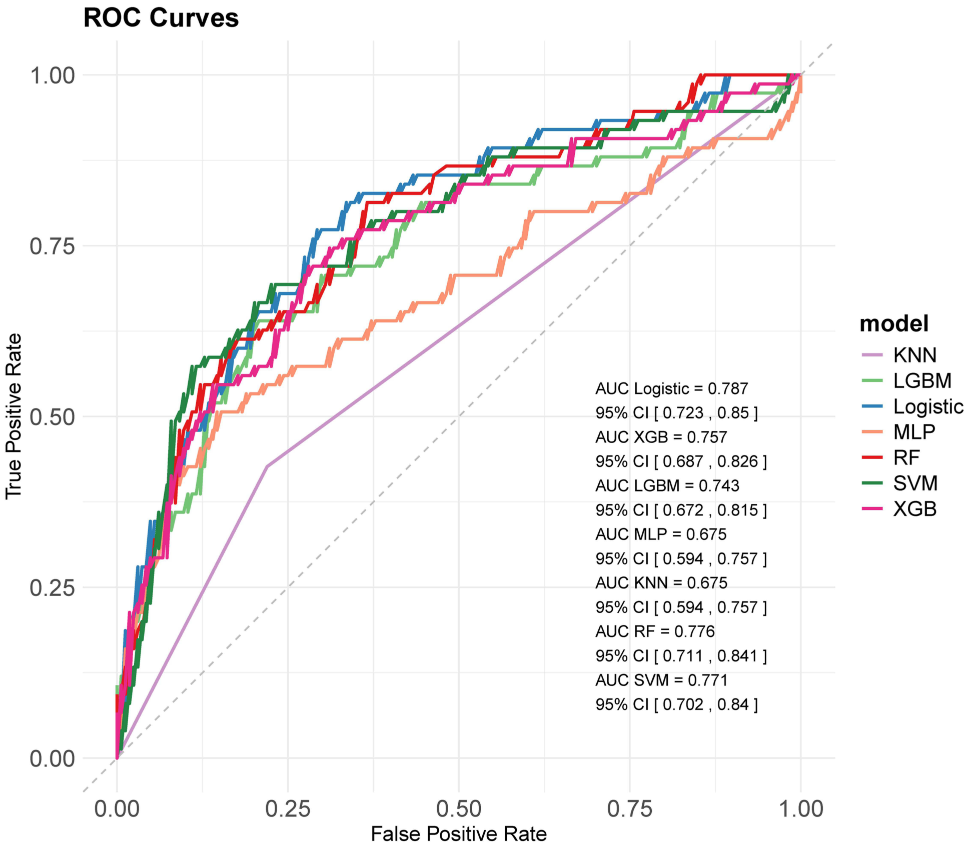

Figure 2 ROC curves of LR and 6 ML algorithms.

Abbreviations: LR, Logistic regression; XGB, eXtreme gradient boosting; LGBM, light gradient boosting; MLP, multilayer perceptron; KNN, k-nearest neighbors; RF, random forest; SVM, support vector machine.

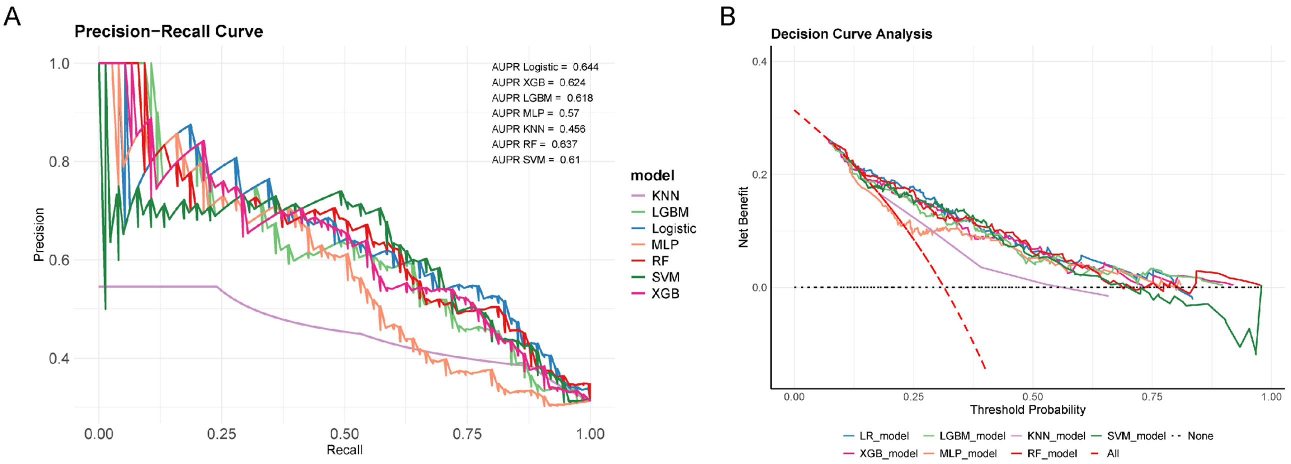

Figure 3 Comparative performance evaluation of prediction models through precision-recall and clinical utility analysis. (**A**) Precision-Recall Curve (PRC) comparison across models. (**B**) Decision Curve Analysis (DCA) quantifying net clinical benefit.

Calibration of The Models

Calibration curves demonstrated the consistency between model-predicted probabilities and actual outcomes. As shown in Figure S1, XGB and RF models were closest to the ideal curve, followed by the LR model. The p-values for HL validation for the LR, XGB, LGBM, and MLP models were all >0.05, specifically 0.361, 0.794, 0.196, and 0.107 (Table 3). These results indicate that these four models provide predictions that align well with actual scenarios and are suitable for clinical decision-making.

Table 3 Performances of Various Prediction Tools Predicting GDM

Models	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	HL P- value
Logistic	0.75	0.93	0.36	0.76	0.71	0.79	**0.36**
XGB	0.74	0.93	0.35	0.76	0.68	0.76	**0.79**
LGBM	0.74	0.91	0.36	0.76	0.64	0.74	**0.20**
MLP	0.76	0.93	0.39	0.77	0.71	0.68	**0.11**
KNN	0.70	0.91	0.24	0.72	0.55	0.60	<0.001
RF	0.71	1.00	0.08	0.70	1.00	0.78	<0.001
SVM	0.73	0.95	0.24	0.73	0.69	0.77	0.01

Notes: HL P-value: Results from the Hosmer-Lemeshow goodness-of-fit test, indicating agreement between predicted probabilities and observed outcomes across decile groups (P>0.05 suggests adequate model calibration with non-significant deviations between groups). Values with HL P > 0.05 are bolded in the table.

Abbreviations: GDM, Gestational diabetes mellitus; PPV, Positive predictive value; NPV, Negative predictive value; AUC, Area under the ROC curve; HL, Hosmer-Lemeshow.

Discussion

In this study, we evaluated the performance of traditional LR and 6 advanced ML models in early prediction of GDM. Surprisingly, LR exhibited the highest AUC, and performed comparably to the ML models in terms of calibration and clinical utility.

According to the widely applied IADPSG recommendation for the 75 g OGTT diagnostic method, the prevalence of GDM in mainland China ranges from 5.12% to 33.30%,[1] a proportion substantially higher than that in Europe (3.8–7.8%) and Africa (approximately 14.0%).[27,28] This highlights the importance of developing simple yet accurate early prediction tools for GDM in the Chinese population to facilitate early intervention and treatment. Although some studies have utilized biochemical parameters, metabolomics, or proteomics to construct prediction models, demonstrating very high predictive accuracy (AUC 0.985–0.998),[29–31] these models rely on costly tests based on unconventional GDM risk factors and complex biomarkers. Despite their potential application value, the high cost and specialized nature of these tests limit their widespread use in current clinical practice. Therefore, developing prediction tools based on routine and readily obtainable GDM risk factors would be more clinically practical.

We employed the Boruta algorithm to select statistically important features for model construction from a pool of candidate features. Ultimately, the 13 features selected for our model is based on are routine clinical data and demographic characteristics. Interestingly, CK, typically used to detect muscle damage and malnutrition, showed the highest importance in the Boruta analysis. Some studies have found associations between blood CK levels and obesity, insulin resistance, diabetes, and heart disease,[32,33] suggesting that metabolic disorders, including GDM, may indirectly affect CK levels. Age is commonly considered a risk factor for GDM. However, in this study, despite statistically significant differences in age between the non-GDM and GDM populations, age did not play a significant role in model construction. A study targeting Chinese pregnant women found that age ≥ 35 was an independent risk factor for GDM (OR: 1.15, 95% CI: 1.05–1.26).[34] This discrepancy could be attributed to the small sample size of older pregnant women (aged 35 and above) in our dataset, which may not have provided sufficient statistical power to significantly impact the model. HbA1c, although not a preferred method for diagnosing GDM, is an important indicator for assessing long-term glucose control in pregnant women and is considered an effective predictor of GDM.[35] Unfortunately, due to economic factors and other reasons, the missing value rate of HbA1c exceeded 30% in this study and it was therefore excluded from the analysis.

We found that advanced ML models did not outperform the traditional LR model in the early prediction of GDM, a finding consistent with studies on other diseases such as acute kidney injury, traumatic brain injury, and major chronic diseases.[36–39] However, some studies comparing LR and ML models for predicting GDM have reported that ML models can enhance predictive performance, a result that contradicts ours.[19,20,23] This discrepancy may be attributed to the complexity (or lack thereof) of the data used. Traditional LR offers advantages such as computational efficiency and ease of interpretation, while ML models excel at handling complex nonlinear relationships but require substantial data for training and are sensitive to parameter tuning and model selection. Our results indicate that ML models do not always surpass the simpler, traditional LR models. The performance of ML models can vary significantly depending on the context and requires further investigation. Our developed LR model demonstrated excellent accuracy and sensitivity, with a sensitivity of 0.933, indicating its effectiveness in identifying GDM cases. However, its specificity was only 0.360, indicating a high rate of false positives and a need for improvement in excluding non-GDM cases. Similar outcomes are commonly observed in predictive and prognostic models,[40,41] which can lead to decreased trust in the models and potentially their abandonment. Further optimization in larger sample sizes is needed to ensure the accuracy and practicality of predictions.

Predictive models for GDM are frequently published, yet many studies often lack comprehensive model evaluation, typically focusing primarily on overall performance (AUC). The strength of this study lies its comparison of not only the performance of LR and six ML models but also in its comprehensively assessment of their performance and applicability in early GDM prediction. Moreover, we focused not only on AUC but also compared calibration, clinical efficacy, and other metrics such as accuracy, sensitivity, specificity, PPV, and NPV, providing a detailed view of each model's strengths and weaknesses. However, the limitations of this study are also noteworthy. First, although the sample size of 956 might be adequate for traditional statistical methods, it might not be sufficient to fully leverage the potential of ML techniques. Therefore, the limited sample size could lead to inadequate model training, potentially affecting the reliability of the results and their applicability to a broader population. Additionally, the study only underwent internal validation and has not yet been externally or independently validated.

In conclusion, we utilized 13 readily available early pregnancy clinical data points as predictors to construct one traditional LR model and 6 ML models for predicting GDM. The results suggest that based on common clinical data, ML models may not always outperform to the classic LR model. Nevertheless, significant challenges remain for clinical application, particularly due to unresolved issues with low specificity.

Abbreviations

GDM, Gestational diabetes mellitus; LR, Logistic regression; ML, Machine learning; IADPSG, International Association of Diabetes and Pregnancy Study Groups; OGTT, Oral glucose tolerance test; ACOG, American College of Obstetricians and Gynecologists; BMI, body mass index; FPG, Fasting plasma glucose; WHO, World Health Organization; HbA1c, Glycated hemoglobin; XGB, eXtreme gradient boosting; LGBM, Light gradient boosting; MLP, Multi-layer perceptron; KNN, K-nearest neighbors; RF, Random forest; SVM, Support vector machine; ROC, Receiver operating characteristic; AUC, Area under the ROC; PRC, Precision-recall curve; DCA, Decision curve analysis; HL, Hosmer-Lemeshow; PPV, Positive predictive value; NPV, Negative predictive value.

Ethical Approval

The study ethical approval was obtained from the institutional review board of Pinghu Maternal and Child Health Hospital (NO.202410) and from the ethical committee of each participating sites in accordance with the ethical principles of the Declaration of Helsinki. This study involved only pre-existing, anonymous clinical data and did not require informed consent.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

There is no funding to report.

Disclosure

The authors declare no competing interests.

References

1. Hod M, Kapur A, Sacks DA, et al. The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: a pragmatic guide for diagnosis, management, and care. *Int J Gynaecol Obstet*. 2015;131 Suppl 3:S173–211. doi:10.1016/s0020-7292(15)30033-3
2. Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabet Res Clin Pract*. 2018;138:271–281. doi:10.1016/j.diabres.2018.02.023
3. Zhou T, Du S, Sun D, et al. Prevalence and trends in gestational diabetes mellitus among women in the United States, 2006-2017: a population-based study. *Front Endocrinol*. 2022;13:868094. doi:10.3389/fendo.2022.868094
4. Xiang AH. Diabetes in pregnancy for mothers and offspring: reflection on 30 years of clinical and translational research: the 2022 norbert freinkel award lecture. *Diabetes Care*. 2023;46(3):482–489. doi:10.2337/dci22-0055
5. Yuen L, Wong VW, Simmons D. Ethnic Disparities in Gestational Diabetes. *Curr Diab Rep*. 2018;18(9):68. doi:10.1007/s11892-018-1040-2
6. Gao C, Sun X, Lu L, Liu F, Yuan J. Prevalence of gestational diabetes mellitus in mainland China: a systematic review and meta-analysis. *J Diabetes Investig*. 2019;10(1):154–162. doi:10.1111/jdi.12854
7. Jovanovic L, Pettitt DJ. Gestational diabetes mellitus. *JAMA*. 2001;286(20):2516–2518. doi:10.1001/jama.286.20.2516
8. Mayo Clinic. Gestational diabetes. Available from: https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339. Accessed June 10, 2024.
9. Metzger BE, Gabbe SG, Persson B, et al. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. *Diabetes Care*. 2010;33(3):676–682. doi:10.2337/dc09-1848
10. O'Sullivan JB, Mahan CM, Charles D, Dandrow RV. Screening criteria for high-risk gestational diabetic patients. *Am J Obstet Gynecol*. 1973;116 (7):895–900. doi:10.1016/s0002-9378(16)33833-9
11. Hillier TA, Pedula KL, Ogasawara KK, et al. A pragmatic, randomized clinical trial of gestational diabetes screening. *N Engl J Med*. 2021;384 (10):895–904. doi:10.1056/NEJMoa2026028
12. Nakanishi S, Aoki S, Kasai J, et al. High probability of false-positive gestational diabetes mellitus diagnosis during early pregnancy. *BMJ Open Diabetes Res Care*. 2020;8(1):e001234. doi:10.1136/bmjdrc-2020-001234
13. Sovio U, Murphy HR, Smith GC. Accelerated fetal growth prior to diagnosis of gestational diabetes mellitus: a prospective cohort study of nulliparous women. *Diabetes Care*. 2016;39(6):982–987. doi:10.2337/dc16-0160
14. Zhu H, Chen B, Cheng Y, et al. Insulin therapy for gestational diabetes mellitus does not fully protect offspring from diet-induced metabolic disorders. *Diabetes*. 2019;68(4):696–708. doi:10.2337/db18-1151
15. Juan J, Prevalence YH. Prevention, and lifestyle intervention of gestational Diabetes mellitus in China. *Int J Environ Res Public Health*. 2020;17 (24):9517. doi:10.3390/ijerph17249517
16. Li F, Hu Y, Zeng J, et al. Analysis of risk factors related to gestational diabetes mellitus. *Taiwan J Obstet Gynecol*. 2020;59(5):718–722. doi:10.1016/j.tjog.2020.07.016
17. Rottenstreich M, Rotem R, Reichman O, et al. Previous non-diabetic pregnancy with a macrosomic infant - Is it a risk factor for subsequent gestational diabetes mellitus? *Diabet Res Clin Pract*. 2020;168:108364. doi:10.1016/j.diabres.2020.108364
18. Li X, Zuo J, Li YH, Tang YP, Bao YR, Ying H. Association between thyroid function and risk of gestational diabetes mellitus in assisted pregnancies: a retrospective cohort study. *Diabet Res Clin Pract*. 2021;171:108590. doi:10.1016/j.diabres.2020.108590
19. Cubillos G, Monckeberg M, Plaza A, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. *BMC Pregnancy Childbirth*. 2023;23(1):469. doi:10.1186/s12884-023-05766-4
20. Hu X, Hu X, Yu Y, Wang J. Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm. *Front Endocrinol*. 2023;14:1105062. doi:10.3389/fendo.2023.1105062
21. Man B, Schwartz A, Pugach O, Xia Y, Gerber B. A clinical diabetes risk prediction model for prediabetic women with prior gestational diabetes. *PLoS One*. 2021;16(6):e0252501. doi:10.1371/journal.pone.0252501
22. Watanabe M, Eguchi A, Sakurai K, Yamamoto M, Mori C. Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan environment and children's study. *Sci Rep*. 2023;13(1):17419. doi:10.1038/s41598-023-44313-1
23. Belsti Y, Moran L, Du L, et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *Int J Med Inform*. 2023;179:105228. doi:10.1016/j.ijmedinf.2023.105228
24. Metzger BE, Gabbe SG, Persson B, et al. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy: response to Weinert. *Diabetes Care*. 2010;33(7):e98–e98. doi:10.2337/dc10-0719
25. Dunne F, Newman C, Alvarez-Iglesias A, et al. Early metformin in gestational diabetes: a randomized clinical trial. *JAMA*. 2023;330(16):1547–1556. doi:10.1001/jama.2023.19869
26. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Software*. 2010;36:1–13. doi:10.18637/jss.v036.i11
27. Mwanri AW, Kinabo J, Ramaiya K, Feskens EJ. Gestational diabetes mellitus in sub-Saharan Africa: systematic review and metaregression on prevalence and risk factors. *Trop Med Int Health*. 2015;20(8):983–1002. doi:10.1111/tmi.12521
28. Eades CE, Cameron DM, Evans JMM. Prevalence of gestational diabetes mellitus in Europe: a meta-analysis. *Diabet Res Clin Pract*. 2017;129:173–181. doi:10.1016/j.diabres.2017.03.030
29. Xiong Y, Lin L, Chen Y, et al. Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques. *J Matern Fetal Neonatal Med*. 2022;35(13):2457–2463. doi:10.1080/14767058.2020.1786517
30. Koos BJ, Gornbein JA. Early pregnancy metabolites predict gestational diabetes mellitus: implications for fetal programming. *Am J Obstet Gynecol*. 2021;224(2):215.e1–215.e7. doi:10.1016/j.ajog.2020.07.050
31. Zhao D, Shen L, Wei Y, et al. Identification of candidate biomarkers for the prediction of gestational diabetes mellitus in the early stages of pregnancy using iTRAQ quantitative proteomics. *Proteomics Clin Appl*. 2017;11(7–8). doi:10.1002/prca.201600152
32. Soleimani E, Ardekani AM, Fayyazishishavan E, Farhangi MA. The interactive relationship of dietary choline and betaine with physical activity on circulating creatine kinase (CK), metabolic and glycemic markers, and anthropometric characteristics in physically active young individuals. *BMC Endocr Disord*. 2023;23(1):158. doi:10.1186/s12902-023-01413-3

33. Frank M, Finsterer J. Creatine kinase elevation, lactacidemia, and metabolic myopathy in adult patients with diabetes mellitus. *Endocr Pract*. 2012;18(3):387–393. doi:10.4158/ep11316.Or
34. Shuang W, Huixia Y. Analysis of the effect of risk factors at gestational diabetes mellitus. *Zhonghua Fu Chan Ke Za Zhi*. 2014;49(5):321–324.
35. Gupta Y, Simmons D. Value of early pregnancy HbA(1c) to predict gestational diabetes. *Lancet Diab Endocrinol*. 2024;12(8):505–507. doi:10.1016/s2213-8587(24)00160-8
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004
37. Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol*. 2020;122:95–107. doi:10.1016/j.jclinepi.2020.03.005
38. Nusinovici S, Tham YC, Chak Yan MY, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol*. 2020;122:56–69. doi:10.1016/j.jclinepi.2020.03.002
39. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. *Int J Med Inform*. 2021;151:104484. doi:10.1016/j.ijmedinf.2021.104484
40. Satici C, Demirkol MA, Sargin Altunok E, et al. Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19. *Int J Infect Dis*. 2020;98:84–89. doi:10.1016/j.ijid.2020.06.038
41. Martin A, Bauer V, Datta A, et al. Development and validation of an asthma exacerbation prediction model using electronic health record (EHR) data. *J Asthma*. 2020;57(12):1339–1346. doi:10.1080/02770903.2019.1648505