External Validation of Multiple Risk Prediction Models for Gestational Diabetes Mellitus

Linjuan Xia[1], Juncheng Lyu[2]*

[1]Department of Public Health, International College, Krirk University, Bangkok, Thailand, 10220.
ORCID iD: https://orcid.org/0009-0000-9309-3602, Email: xialinjuan91@163.com
[2]Department of Public Health, International College, Krirk University, Bangkok, Thailand, 10220.
ORCID iD: https://orcid.org/0000-0003-2070-9503, Email: cheng_china@163.com

Abstract

Objective: To externally validate gestational diabetes mellitus (GDM) prediction models in a Chinese population and assess their performance in early pregnancy risk stratification. **Methods:** Six GDM prediction models identified from a systematic review were externally validated using data from 1,385 pregnant women in tertiary hospital in China. Model performance was evaluated in terms of discrimination (C-statistic), calibration (calibration slope and intercept), and clinical utility (decision curve analysis). **Results:** Among 1,385 women, 661 were diagnosed with GDM. All models showed decreased discrimination compared with the original studies, with area under the curve (AUC) ranging from 0.693 to 0.751. All models underestimated risk in high-risk individuals and most models demonstrated relatively stable net benefit, indicating potential suitability for early pregnancy risk stratification. **Conclusion:** Existing GDM prediction models exhibit variable performance in Chinese populations. Further recalibration and impact assessment are recommended.

Keywords: Gestational Diabetes Mellitus, Prediction Model, External Validation, China.

Introduction

Gestational diabetes mellitus (GDM) is defined as glucose intolerance first diagnosed in the second or third trimester of pregnancy in women without prior diabetes.[1] Its prevalence has risen steadily worldwide. In 2021, the International Diabetes Federation reported that 16.7% of pregnant women aged 20–49 years had hyperglycemia, with 80.3% attributable to GDM; the prevalence in mainland China has reached 14.8%.[2] GDM is associated with increased risks of adverse pregnancy outcomes such as cesarean delivery, preterm birth, macrosomia, and neonatal hypoglycemia,[3,4] as well as long-term complications including postpartum diabetes,[5] cardiovascular disease,[6] obesity[7] and metabolic disorders in offspring.[8] Early identification of women at high risk for GDM enables timely lifestyle interventions, which may reduce the incidence of GDM and its adverse outcomes.[9] Prediction models—statistical combinations of multiple predictors estimating the probability of a specific outcome are increasingly used for early GDM risk assessment.

Although many GDM prediction models have been developed worldwide using diverse statistical and machine learning approaches, their actual clinical impact remains limited.[10] One major reason is the scarcity of external validation, a critical step to assess model performance in new populations before clinical implementation. Without such validation, models often exhibit lower discrimination and calibration in new settings than in their development cohorts, due to factors such as overfitting, omitted predictors, differences in patient characteristics, or variations in diagnostic criteria.[11] External validation not only quantifies a model's generalizability but also informs potential model updating or recalibration to improve accuracy.[10] In the field of GDM, most external validation studies have focused on models developed in non-Chinese populations,[12] with limited systematic evaluation of models for Chinese women. Furthermore, few studies have directly compared the performance of multiple existing models within the same external cohort from China. This gap hampers the selection of reliable and locally applicable tools for early GDM risk stratification in China.

Access This Article Online

Quick Response Code:

Website: www.jnsbm.org

DOI: https://doi.org/10.5281/zenodo.17279482

Address for Correspondence: Department of Public Health, International College, Krirk University, Bangkok, Thailand,10220
Email: cheng_china@163.com

Submitted: 16th August, 2025 **Received:** 29th September, 2025
Accepted: 02nd October, 2025 **Published:** 04th October, 2025

How to Cite This Article: Xia L, Lyu J. External Validation of Multiple Risk Prediction Models for Gestational Diabetes Mellitus. J Nat Sc Biol Med 2025;16(3):44-54

To address these issues, we conducted a systematic review to identify early pregnancy GDM prediction models applicable to Chinese settings and performed external validation of selected models in an independent cohort from mainland China. This study aims to provide empirical evidence on model performance, stability, and applicability, thereby supporting rational model selection and localized implementation in clinical practice.

Literature Review

Definition of External Validation Studies for Prediction Models

According to the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist, prediction model studies can generally be categorized into three types: (1) model development studies without external validation, (2) model development studies including external validation, and (3) studies dedicated solely to the external validation of established models. The preceding section reviewed the first two types of studies in the context of GDM prediction models. These studies primarily construct new prediction models based on population data collected by the research team and perform internal validation using data from homogeneous populations to evaluate model applicability and stability under the same research setting. Some studies further apply external validation using independent cohorts with characteristics different from the development population, aiming to more comprehensively assess model generalizability and applicability.

However, such external validation efforts are often limited, with most studies relying on a single data source, potentially constraining the model's applicability across different ethnicities, regions, or healthcare settings. In the GDM prediction field, several studies have attempted to use data from different cohorts to externally validate established models, thereby exploring their performance in diverse populations and assessing their suitability across geographical regions and healthcare institutions. Moreover, some research has incorporated decision curve analysis (DCA) and recalibration techniques (e.g., intercept adjustment and calibration curve modification) to optimize predictive accuracy, thereby enhancing the model's value in varied clinical contexts.

External Validation of a Single Prediction Model

In an international study, van Leeuwen *et al.*[13] conducted an external validation of the GDM risk scoring system developed by Naylor *et al.*[14] using data from a prospective cohort study. The original scoring system, based on multivariable logistic regression analysis, calculated odds ratios (ORs) for three clinical variables: maternal age, body mass index (BMI), and ethnicity. Based on these variables, each pregnant woman was assigned a clinical risk score, with a maximum possible score of 10 points. Women with scores of 0 or 1 were categorized as low risk, those with scores of 2 or 3 as moderate risk, and those with scores greater than 3 as high risk for GDM.

The external validation cohort comprised 1,266 women, of whom 47 were diagnosed with GDM. The validation results demonstrated moderate discriminatory ability and limited calibration (goodness-of-fit chi-square test: $\chi^2 = 8.89$, $P = 0.06$). Although both discrimination and calibration were modest, the authors found that, compared to universal screening, the score-based selective screening strategy could reduce the number of women requiring screening by 25% while maintaining a similar detection rate. Thus, the model retained some clinical utility. It is worth noting that this validation study was conducted relatively early and employed a selective screening strategy, meaning not all participants underwent GDM testing; consequently, the number of GDM cases was small. Nevertheless, the study provides valuable methodological insights for subsequent research.

External Validation of Prediction Models using Non-invasive Predictors

With the increasing number of prediction model development studies, more external validation research has begun to focus on validating multiple existing models and comparing their predictive performance, as well as evaluating their clinical utility. One external validation study prospectively recruited a validation cohort of 7,929 Caucasian pregnant women at their first prenatal visit to validate four published prediction models based on clinical characteristics.[15] The cohort was from the metropolitan area of Quebec City, Canada, with 381 women subsequently developing GDM. The four models incorporated risk factors including maternal age, BMI, ethnicity, family history of GDM, previous history of GDM, macrosomia, and adverse obstetric outcomes. These models were originally developed in populations from Canada, Turkey, the Netherlands, and Australia, with specific risk factors and prediction rules described in earlier literature. The area under the receiver operating characteristic curve (AUC) for identifying GDM ranged from 0.668 to 0.756, with performance comparable to that observed in the original studies. The best-performing model included ethnicity, BMI, family history of diabetes, and prior GDM history as variables. For predicting GDM cases requiring insulin therapy, this model achieved a sensitivity of 73%, specificity of 81%, and AUC of 0.824. Given its large sample size of Caucasian pregnant women, the study demonstrated good representativeness, and the four maternal-characteristic–based models showed favorable discrimination, indicating generalizability. The authors also explored the models' potential for early prediction of GDM cases requiring insulin treatment and suggested incorporating newly identified biomarkers to improve predictive performance and reach clinically applicable standards. However, the study assessed only the discriminative ability, which limited the comprehensiveness of model performance evaluation.

Another external validation study addressed this limitation. Researchers systematically searched PubMed up to April 13, 2017, to identify GDM risk prediction models based solely on non-invasive predictors collected in early

pregnancy. Twelve published models were identified and externally validated using data from two Dutch prospective cohort studies (Expect Study I and PRIDE Study).[16] The validation cohort included 5,260 pregnant women, among whom 127 (2.4%) were diagnosed with GDM—72 cases in Expect Study I and 55 in PRIDE Study. The C-statistics of the models ranged from 68% to 75%, with almost all AUCs lower than those reported in the original studies. The model by Nanda *et al*.[17] achieved the highest discriminative performance (AUC = 0.75).

Subgroup analysis in nulliparous women showed only slight decreases in AUC, except for the model by Gabbay-Benziv *et al*.[18], which declined by 0.05. Sensitivity analyses demonstrated similar model performance in the two cohorts. Calibration plots indicated that most models tended to overestimate GDM risk, with the models by Nanda *et al*.[17] and Gabbay-Benziv *et al*.[18] showing the best calibration. Recalibration improved the agreement between predicted probabilities and observed incidence for most models. Decision curve analysis revealed that these models provided positive net benefit compared with treating all or no women as high risk within a risk threshold range of 1%~55%. Further analysis of the Nanda et al. model at different clinically relevant thresholds showed high sensitivity and negative predictive value (NPV) at low thresholds (e.g., 2%), indicating strong ability to exclude low-risk women. However, at high sensitivity levels, many women were falsely classified as high risk (high false-positive rate), and when the threshold exceeded 5%, sensitivity dropped sharply, leading to misclassification of many women who eventually developed GDM as low risk.

External Validation of Prediction Models Combining Non-invasive Predictors

The above studies focused on traditional models using non-invasive predictors only. With growing interest in experimental biomarkers as predictors, some external validation studies have compared the performance of such models. A large, prospective, multicenter cohort study was conducted across 31 midwifery practices and six hospitals in the Netherlands to externally validate all published GDM prediction models available at the time.[19] Women were recruited at their first prenatal visit, and 3,723 participants were included in the final analysis, with 181 (4.9%) developing GDM. Twelve models were validated, with C-statistics ranging from 0.67 to 0.78, indicating moderate variability in discriminative ability. Calibration analysis showed good calibration for Gabbay-Benziv *et al*.[18] and van Leeuwen *et al*.[20]. After recalibration, eight models exhibited good calibration curves closely matching the ideal line, although Nanda *et al*.[17], Pintaudi *et al*.[21], and Shirazian *et al*.[22] tended to over- or underestimate risk in some cases.

AUCs for original and recalibrated models ranged from 0.67 to 0.78, with recalibrated models slightly underperforming compared to their development populations. The highest AUCs were seen in Gabbay-Benziv *et al*.[18], Nanda *et al*.[17], Teede *et al*.[23], and van Leeuwen *et al*.[20] —all including maternal age, BMI, history of GDM, ethnicity, and family history of diabetes. The poorest-performing models had the fewest predictors. For nulliparous women, four models [Gabbay-Benziv *et al*.[18], Nanda *et al*.[17], Naylor *et al*.[14], Teede *et al*.[23] showed lower discrimination than in the overall cohort, whereas other models performed better in nulliparous women. Decision curve analysis for the top four models showed positive net benefit at thresholds between 0% and 40%. This large-sample study, including both low- and high-risk populations across primary and secondary/tertiary care settings, offered good representativeness and comprehensive performance evaluation. However, the high-risk screening strategy used—only performing OGTT in low-risk women if GDM-related symptoms occurred—may have underestimated GDM incidence in this group.

External Validation of Prediction Models Under the IADPSG Diagnostic Criteria

Previous external validation studies did not account for differences in GDM diagnostic criteria. One study evaluated 15 clinical prediction models using the IADPSG 2010 criteria to define GDM in the validation cohort.[24] A total of 1,132 pregnant women were prospectively recruited before 16^{+0} weeks' gestation for risk assessment, including routine laboratory tests. ROC-AUC values ranged from 60.7% to 76.9%, representing moderate to good predictive accuracy. Overall, "propensity score" models (i.e., models calculating continuous GDM probability) outperformed "total score" and "decision tree" models, particularly in discrimination, and exceeded the predictive performance of models using maternal age alone.

Calibration analysis of nine propensity score models with complete prediction rules showed acceptable calibration for Benhalima-1 and Benhalima-2 (2020),[25] and underestimation in van Leeuwen *et al*.[20], Nanda *et al*.[17], Gabbay-Benziv *et al*.[18] and Syngelaki *et al*.[26] models. Poor calibration in most models was attributed to development under older diagnostic criteria, which did not account for the higher current GDM prevalence, thus underestimating risk. Recalibration improved agreement, though van Leeuwen *et al*.[20], Nanda *et al*.[17] and Syngelaki *et al*.[26] still overestimated risk in some scenarios. Random forest analysis indicated that prior GDM history and routine laboratory measures (fasting glucose, HbA1c, triglycerides) were the most important predictors. Benhalima-2 (2020),[25] incorporating these variables, demonstrated the greatest net benefit and best discrimination.

External Validation in Chinese Populations

Most external validation studies have been based on foreign populations, with limited work using Chinese data to validate domestic or foreign models. One study used data from 478 pregnant women to validate four models developed in domestic populations, all multivariable logistic regression models.[27] All four showed moderate discrimination in external validation. Calibration analysis indicated that, except for the model by Rao *et al*.[28], which had good calibration, the others tended to overestimate

or underestimate individual risk. All four models had net benefit within certain risk ranges, with the Rao Jiawei model yielding net benefit at thresholds from 25% to 90%, suggesting potentially greater clinical utility than the others. In summary, Internationally, several well-established GDM prediction models have undergone external validation across diverse countries and ethnicities, with applicability explored in different healthcare systems. Some models developed in Western populations have been validated in multi-center studies and show varying performance across ethnic groups, highlighting the influence of genetic background, lifestyle, and healthcare environments on model generalizability. External validation thus not only assesses model robustness but also identifies key factors influencing applicability and guides further optimization. In contrast, external validation studies in China remain limited, with most research focusing on developing new models from single-population datasets rather than evaluating existing models in Chinese pregnant women. This gap may reflect factors such as a preference for self-developed models, restrictions on multi-center data sharing, and the high demands of independent datasets for external validation. The lack of local external validation may hinder the clinical adoption of GDM prediction models in China.

Methods

Selection of Prediction Models

A systematic search was conducted in both English and Chinese databases using a combination of indexed and free-text terms related to early pregnancy prediction models and GDM. Only studies adopting the International Association of Diabetes and Pregnancy Study Groups (IADPSG) diagnostic criteria, issued in 2010, were included, with a search period from January 2010 to December 2024. Inclusion criteria were: (1) study population comprising pregnant women from mainland China; (2) study type including model development studies (with or without external validation) and independent external validation studies; (3) outcome of interest being GDM diagnosed using the IADPSG criteria;[29] (4) prediction models including at least two predictors with clear model descriptions; and (5) publication in English or Chinese. Exclusion criteria were: (1) pregnant women with other chronic diseases; (2) studies reporting only associations between risk factors and GDM rather than complete prediction models; (3) models containing non-routine clinical predictors (e.g., genetic loci, environmental exposures); (4) predictors measured after 14 weeks of gestation; (5) inaccessible full texts; (6) models developed using machine learning algorithms; and (7) duplicate models.

The risk of bias and applicability of the prediction model studies were assessed using the PROBAST tool,[30] covering four domains: participants, predictors, outcomes, and analysis. Risk of bias was evaluated across all four domains, while applicability was assessed in the first three domains. The risk of bias was categorized as low, high, or unclear. Each domain contains at least two signaling questions, with possible responses of "yes," "probably yes," "no," "probably no," or "no information." The answers to these signaling questions determine the overall risk of bias for each domain. The applicability assessment follows a similar procedure, though it does not involve signaling questions. Any disagreements in the assessment were resolved through consultation with a third reviewer. Based on this systematic evaluation, prediction models suitable for early pregnancy risk assessment and external validation were selected according to the following criteria: (1) methodologically sound study design with low risk of bias; and (2) inclusion of predictors commonly assessed in routine prenatal care, readily obtainable in clinical practice.

External Validation Cohort Data Source

Women who delivered at the Second People's Hospital of Dali City between December 2019 and December 2024 were included as the validation cohort. Inclusion criteria were: (1) singleton pregnancy; (2) complete hospital records. Exclusion criteria were pre-existing diabetes or other metabolic disorders. Regarding the sample size required for external validation, no universally accepted standard exists. However, previous studies recommend including at least 100 events, preferably more than 200,[31] meaning the validation cohort should contain at least 100 women diagnosed with GDM.

Predictors

Trained healthcare personnel extracted data from the hospital electronic medical records system. Maternal information collected included demographics (age, ethnicity, education, residence, occupation, economic status), obstetric history (gravidity, parity, adverse pregnancy history, use of assisted reproductive technology), personal history (smoking, alcohol consumption), past medical history, physical examination (early pregnancy systolic blood pressure [SBP] and diastolic blood pressure [DBP], height, pre-pregnancy weight, BMI), and routine laboratory tests (early pregnancy fasting plasma glucose [FPG], complete blood count, oral glucose tolerance test results).

Statistical Analysis

Descriptive statistics were used for participants' general characteristics, obstetric and medical history. Univariate analyses compared clinical features between the GDM and control groups. The validation cohort data were applied to the selected prediction models using Stata 18.0 to evaluate model performance. Discrimination was assessed by the concordance statistic (C-statistic), i.e., AUC, ranging from 0.5 to 1, with higher values indicating better discriminative ability. Calibration was evaluated using the Stata command pmcalplot to generate calibration plots and calculate intercepts and slopes. Ideally, a perfectly calibrated model has an intercept of 0 and a slope of 1 (the 45° line). A negative intercept indicates risk overestimation, a positive intercept indicates underestimation, and a slope less than 1 suggests potential overfitting. Decision curve analysis was performed to evaluate clinical usefulness

across a range of risk thresholds, with net benefit plotted on the y-axis and threshold probability on the x-axis. The "Treat All" line represents the net benefit if all women are considered GDM cases, decreasing with higher threshold probabilities; the "Treat None" line represents zero risk, with net benefit equal to zero; and the model curve represents net benefit when using the prediction model to guide clinical decisions.

Results

Selection of Prediction Models

A total of 2,996 records was retrieved from the databases, including PubMed (N=467), Cochrane Library (N=41), Embase (N=694), Web of Science (N=558), China National Knowledge Infrastructure (N=257), Wanfang Data (N=242), VIP Database (N=369), and CBM (N=368). After removing 1,296 duplicates, 1,700 records remained for title and abstract screening, of which 107 articles were eligible for full-text review. Ultimately, 48 studies were included. According to the results of the systematic review, all studies exhibited a potentially high overall risk of bias. Therefore, studies with no more than one high-risk domain among the four evaluated domains were selected, considering the predictor variables available in our dataset. Six models were finally chosen for external validation. Their characteristics are summarized in Table 1. Of these, four models presented explicit formulas, where the probability of GDM occurrence was calculated as $P = e^{Lp}/ (1+ e^{Lp})$, and two models used risk scores. The parameters of these models are as follows.

Model 1 (Chen X 2016):[32] Lp = -4.92 + 0.044*age(years) -0.028*height(cm)- 0.309*underweight(if Yes =1,if No =0)+0.362*overweight(if Yes=1,if No=0) + 0.652 *obesity(if Yes=1,if No=0) + 0.326 *DBP≥80mmHg(if Yes=1,if No=0) + 0.501 *family history of diabetes(if Yes=1,if No=0) + 2.894*GDM history(if Yes=1,if No=0) + 1.300 *FPG(mmol/L).

Model 2 (Chen MF 2019):[33] Risk score = 0(if gravidity = 1) + 0.26 (if gravidity = 2) + 0.33 (if gravidity≥3)+ 0(if parity =0) + 0.53(if parity≥ 1) + 0.59 (if delivering macrosomia)+ 0.83(if having family history of diabetes) -0.36 (if pre-pregnancy BMI<18.5kg/m^2) + 0.34(if pre-pregnancy BMI=24~27.9 kg/m^2) + 0.96(if pre-pregnancy BMI≥ 28 kg/m^2 + 2(if with GDM history) + 0.64 (if age= 26~30 years) + 1.03(if age =31~34 years) + 1.39 (if age =35~40 years)+ 1.86 (if age≥ 41years).

Model 3 (Gao S 2020):[34] Risk score=0.0941*age(years) + 0.1278*BMI(kg/m^2) + 0.0093*SBP(mmHg) + 0.6816*Log10(alanine aminotransferase) + 0.5129* family history of history(if Yes =1,if No =0) - 0.0270 * height(cm) -5.7469.

Model 4 (Guo F 2020):[35] Lp = -10.84 + 0.078 * age(years) + 0.119 * pre-pregnancy BMI (kg/m^2) + 0.893 * FPG (mmol/L) + 0.491 * family history of diabetes (if Yes=1, if No=0).

Model 5(Li JJ 2021):[36] Lp = -8.524 + 0.079 * age (<25years = 1,25~29years = 2,30~34years = 3, ≥35years = 4) + 2.160 * pre-pregnancy BMI (<18. 50 kg /m^2 =1,18. 50 ~ <23. 00 kg /m^2 = 2, 23. 00~<25. 00 kg /m^2 = 3, ≥25. 00kg /m^2 = 4) + 1. 191 * family history of diabetes (if Yes=1, if No=0).

Model 6 (Li SH 2024):[37] Lp = - 2. 309 + 0. 692 * age>35(if Yes=1,if No=0) + 0. 894 * BMI>24. 0 kg/m^2 (if Yes=1,if No=0) + 0. 267 * FPG (mmol/L) + 0. 763 * family history of diabetes (if Yes=1, if No=0) + 0. 694 * anemia in early gestation (if Yes=1, if No=0).

Table 1: Characteristics of 6 Models.

Study	Modeling Method	Predictors	Presentation
Chen X 2016	Multivariate logistic regression	Age, pre-pregnancy BMI, SBP≥80mmHg, family history of diabetes, history of GDM, FPG, height	Formula
Chen MF2019	Multivariate logistic regression	History of GDM, age, family history of diabetes, macrosomia, gravidity, parity, pre-pregnancy BMI	Risk score
Gao S 2020	Multivariate logistic regression	Age, BMI, DBP, alanine aminotransferase, family history of diabetes, height	Risk score
Guo F 2020	Multivariate logistic regression	Age, BMI, FPG, family history of diabetes	Formula
Li JJ 2021	Multivariate logistic regression	Age, pre-pregnancy BMI, family history of diabetes	Formula
Li SH 2024	Multivariate logistic regression	Age>35, pre-pregnancy BMI>24. 0 kg/m^2, FPG, family history of diabetes, anemia in early gestation	Formula

Characteristics of the Validation Cohort

A total of 1,385 women from the hospital were included in the validation cohort, aged 18~48 years. The mean height was 1.611± 0.051m, and the mean pre-pregnancy weight was 57.077± 9.187kg. Among them, 661 were diagnosed with GDM and 724 had normal glucose levels. As shown in Table 2, women in the GDM group were significantly older (P<0.05). The GDM group had higher systolic blood pressure, while diastolic blood pressure showed no significant difference (P>0.05). No significant difference in height was observed (P>0.05), but pre-pregnancy weight was significantly higher in the GDM group (P< 0.05). A family history of diabetes was significantly associated with GDM occurrence (P< 0.05). Similarly, a history of GDM and macrosomia in previous pregnancies were significantly related to GDM in the current pregnancy (P < 0.05). Gravidity and parity differed significantly between groups (P< 0.05). Pre-pregnancy BMI showed significant differences, with higher BMI observed in the GDM group (P < 0.05). Moreover, the difference in FPG between the two groups was statistically significant (P < 0.05) and FPG in early pregnancy was higher in the GDM group, but the differences in alanine aminotransferase (ALT) and anemia were not statistically significant (P> 0.05).

Table 2: Comparison of Characteristics between GDM Group and Non-GDM Group.

Variable	Non-GDM (N=724)	GDM (N=661)	$t/c^2/Z$	*P*
Age(years)	30.30±4.280	32.22±4.172	-8.440	<0.001
SBP (mmHg)	113.43±12.767	115.21±12.648	-2.600	0.009
DBP (mmHg)	72.23±10.013	72.38±9.536	-0.288	0.744
Height(m)	1.6103±.05054	1.6126±.05065	-0.824	0.410
Weight (Kg)	55.128±8.2599	59.211±9.6698	-8.472	<0.001
Pre-BMI(Kg/m^2)	21.16±2.844	22.64±3.558	-8.586	<0.001
Gravidity	2.45±1.085	2.69±1.240	-3.268	0.001
Parity	0.83±0.526	0.92±0.512	-2.842	0.004
Macrosomia (Yes/No)	6/718	20/641	9.054	0.003
FPG	4.38±0.387	5.17±0.785	14.512	0.000
ALT	17.02±14.675	19.40±16.744	1.830	0.068
History of DM(Yes/No)	20/704	49/612	15.787	<0.001
GDM history (Yes/No)	11/713	34/627	14.439	<0.001
Anemia (Yes/No)	43/681	29/632	0.717	0.397

Model Performance Discrimination

In the validation cohort of 1,385 women (661 with GDM and 724 without), meeting the sample size requirements for external validation, model formulas or risk scores were applied to the collected data. The discrimination performance of the models in their original studies and in the present external validation is summarized in Table 3, with ROC curves shown in Figure 1. Compared with their performance in the original development cohorts, all models demonstrated lower AUC values after external validation.

Table 3: Predictive Model Discrimination Performance.

Model	Original Study AUC (95%CI)	Validation Cohort AUC (95%CI)
Chen X 2016	0.722 (0.696~0.747)	0.676 (0.630~0.723)
Chen MF 2019	0.659 (0.63~0.688)	0.621 (0.582~0.661)
Gao S 2020	0.710 (0.680~ 0.741)	0.707 (0.665~0.749)
Guo F 2020	0.69 (0.67~0.72)	0.663 (0.616~0.709)
Li JJ 2021	0. 870 (0. 756~0. 985)	0.693 (0.650~0.735)
Li SH 2024	0.82 (0. 76~0.89)	0.646 (0.599~0.694)

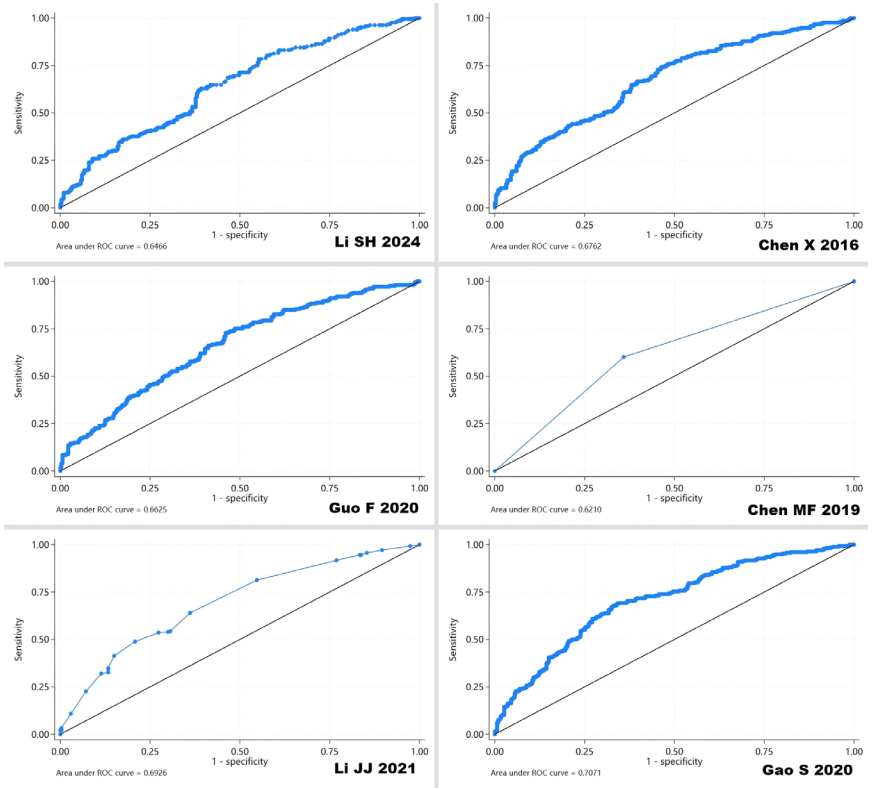

Figure 1: ROC Curves for 6 Models.

Calibration

Calibration plots for the six prediction models are shown in Figure 2. All models indicating potential underestimation of individual risk. Among them, the Li SH model had a calibration-in-the-large (CITL) of 0.232, closer to 0, suggesting smaller overall bias than the other models.

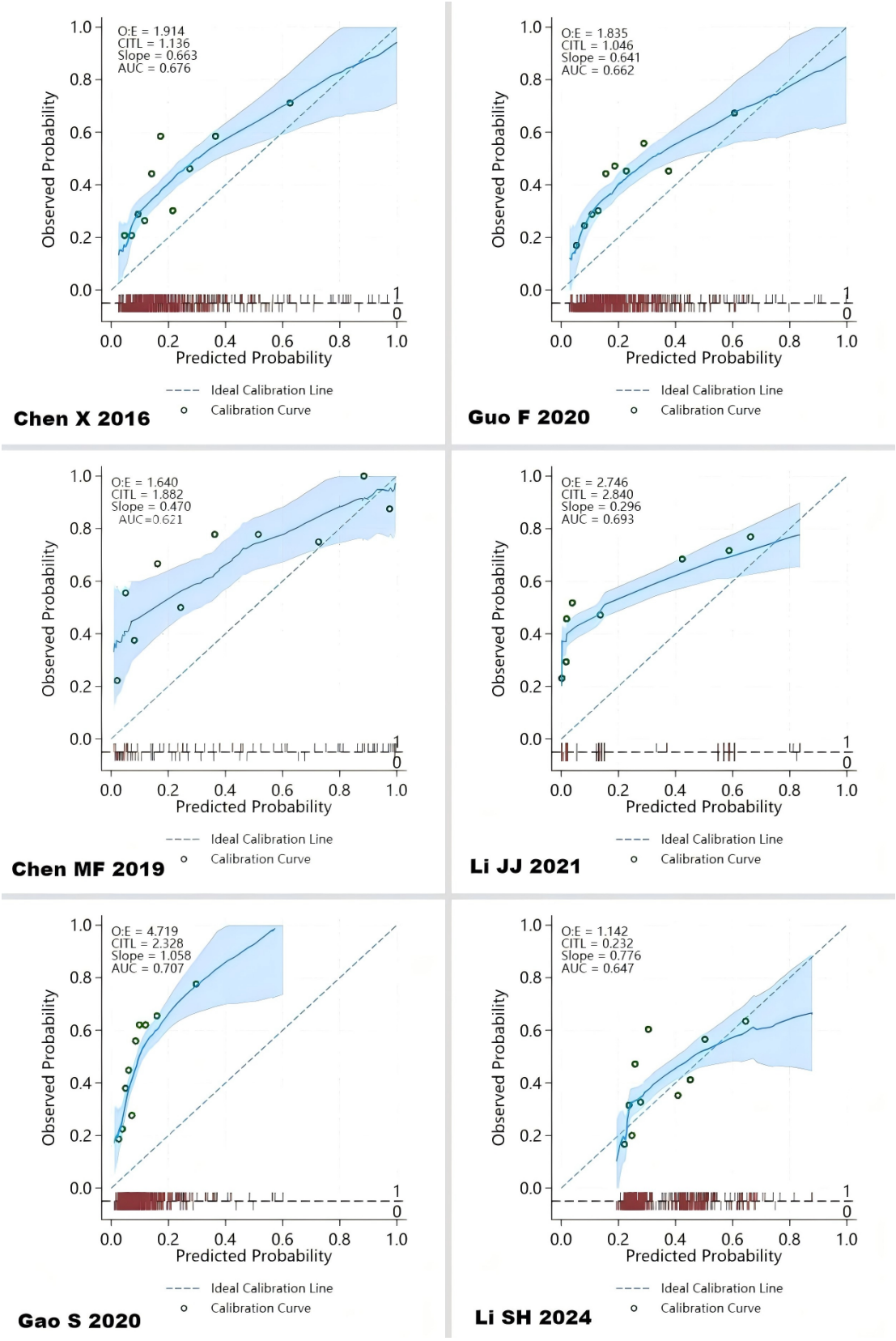

Figure 2: Calibration Curves for 6 Models.

Clinical Utility

Decision curve analysis results (Figures 3~8) showed that the Gao S model consistently yielded net benefits lower than the "Treat All" strategy, suggesting relatively limited clinical utility. The other five models demonstrated net benefit at threshold probabilities >0.4, indicating potential value for clinical decision-making. The decision curves for the Chen X, Chen MF, and Guo F models declined more gradually, reflecting relatively stable clinical performance.

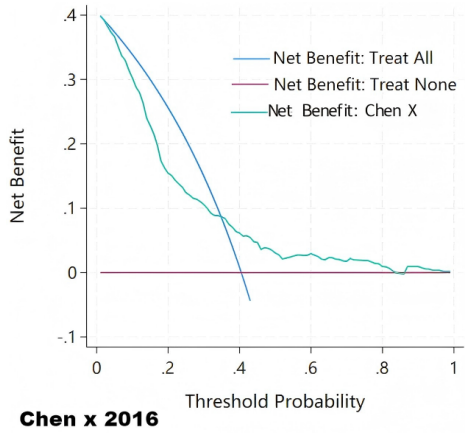

Figure 3: DCA Result for the Chen X Model.

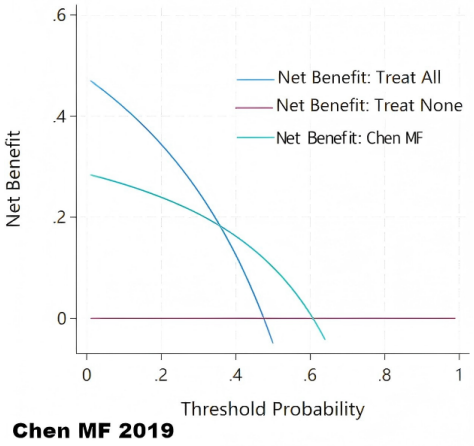

Figure 4: DCA Result for Chen MF Model.

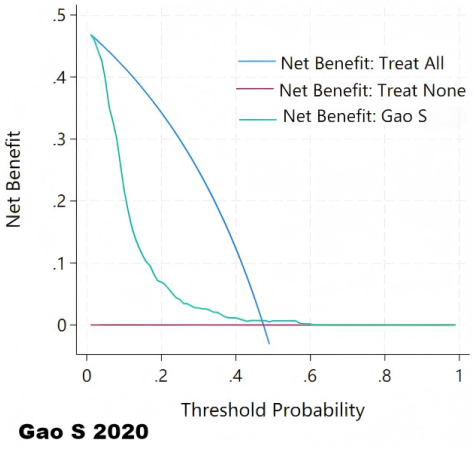

Figure 5: DCA Result of Gao S Model.

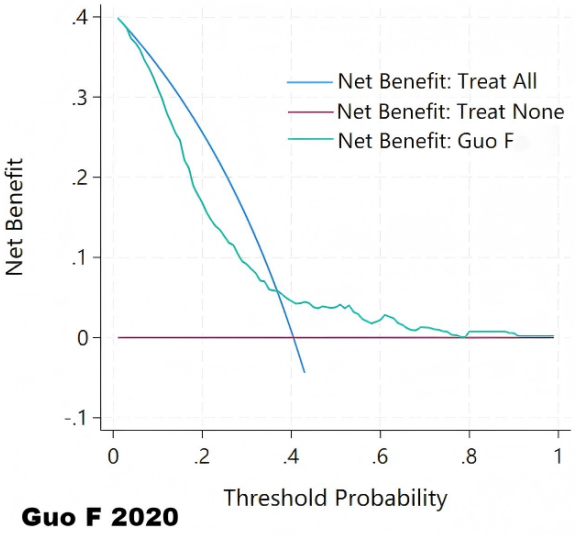

Figure 6: DCA Result of the Guo F Model.

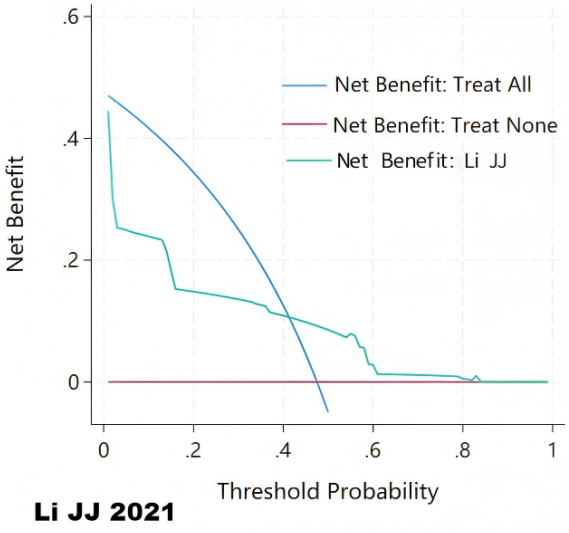

Figure 7: DCA Result of Li JJ Model.

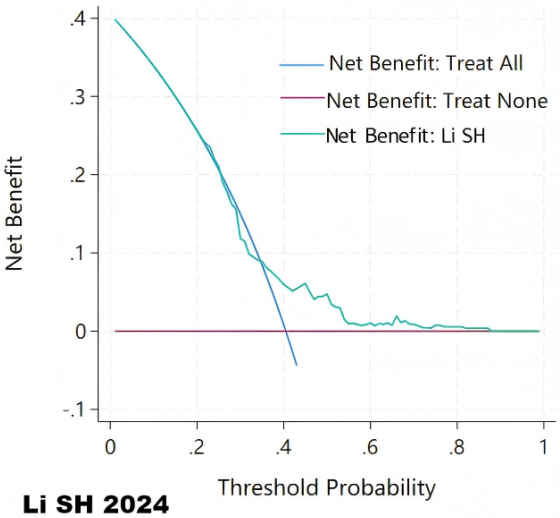

Figure 8: DCA Result of the Li SH Model.

Discussion

Key Findings

External validation is a crucial step to evaluate the robustness and generalizability of prediction models across different populations and real-world clinical settings.[10] In this study, six GDM prediction models with relatively low risk of bias and compatible predictor variables were selected from a systematic review of 48 studies for external validation. When applied to an independent cohort, all models showed a decrease in AUC compared with the original studies; however, discrimination remained acceptable, with the highest AUC reaching 0.751. These results indicate that the models can reasonably distinguish high- from low-risk individuals in early pregnancy. Domestic studies evaluating GDM prediction models reported AUCs ranging from 0.719 to 0.759,[27] consistent with our findings. Notably, some models with high original AUCs, such as the Li JJ model (original AUC = 0.870), decreased to 0.693 in our cohort, suggesting limited stability across broader populations.

Despite acceptable discrimination, all models exhibited suboptimal calibration and underestimated individual risk. Except for the Gao S model, calibration slopes were <1, indicating overfitting. Evidence from international external validation studies has shown that recalibration can substantially improve agreement between predicted and observed risks.[16,24] Among our models, the Li SH model demonstrated the closest CITL to zero, indicating minimal overall bias and superior calibration. Decision curve analysis confirmed that five models yielded positive net benefit at predicted risk thresholds >0.4, while the Gao S model demonstrated limited clinical utility.

Comparison with International Studies

Several GDM prediction models have undergone external validation in European and Australian populations, highlighting the importance of population-specific assessment. A prospective cohort of 7,929 Caucasian women validated four clinical-feature-based models, with AUCs ranging from 0.668 to 0.756.[15] Similarly, a multicenter Dutch study externally validated 12 published models across primary and tertiary care, reporting C-statistics between 0.67 and 0.78. After recalibration, eight models demonstrated good calibration, and decision curve analysis indicated positive net benefit at predicted risk thresholds of 0~40%.[19] Another prospective cohort, using the IADPSG 2010 diagnostic criteria, evaluated 15 clinical prediction models in 1,132 women before 16^{+0} weeks gestation. ROC-AUC ranged from 0.607 to 0.769, reflecting moderate to good discrimination. Propensity-score-based models outperformed total-score or decision-tree models, particularly in distinguishing GDM cases from non-cases. Calibration analysis revealed miscalibration in older models due to outdated diagnostic criteria, whereas recalibration substantially improved concordance between predicted and observed risks.[24] Random forest analysis identified GDM history and routine laboratory measures (e.g., FPG, HbA1c, triglycerides) as highly influential predictors, supporting the superior performance and net benefit of models incorporating these variables, such as the Benhalima-2 2020 model.

Implications and Recommendations

Our findings confirm that most existing GDM prediction models experience performance degradation when applied to new populations, primarily due to differences in cohort characteristics, single-center development, and lack of robust internal and external validation. Models with acceptable discrimination and net benefit may facilitate individualized, risk-based prenatal care in Chinese populations. Future research should focus on recalibration, model updating, and integration into clinical workflows, while improving interpretability to enhance understanding among clinicians and pregnant women, thereby promoting wider adoption in practice.

Conclusion

This study externally validated six GDM prediction models selected from a systematic review, providing empirical evidence for their real-world applicability and generalizability in a Chinese population. Most existing model development studies carry a high risk of bias, report incomplete calibration information, and lack external validation. Our findings indicate that while discrimination of the validated models remains acceptable, calibration varies, with some models underestimating risk in high-risk individuals. Models validated in this study showed relatively stable net benefit, highlighting their potential for early pregnancy risk stratification and individualized counseling.

However, this study has several limitations. First, models that were not fully reported were excluded from the systematic review, potentially omitting some high-performing prediction models. Second, the external validation cohort was drawn from a single tertiary hospital in one province, limiting generalizability to other regions. Future research should focus on updating and externally validating existing models, conducting model impact studies to assess safety, effectiveness, and cost-effectiveness in clinical practice, and facilitating model adaptation for local populations. Such efforts will support more precise early-pregnancy prediction and management of GDM, ultimately contributing to evidence-based, individualized prenatal care.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

1. ElSayed NA, Aleppo G, Aroda VR, et al. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023. Diabetes Care. 2023; 46(Suppl 1): S19-s40. doi: https://doi.org/10.2337/dc23-s002.
2. 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): 154-62. doi: https://doi.org/10.1111/jdi.12854.

3. Ye W, Luo C, Huang J, Li C, Liu Z, Liu F. Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis. Bmj. 2022; 377: e067946. doi: https://doi.org/10.1136/bmj-2021-067946.
4. Metzger BE, Lowe LP, Dyer AR, et al. Hyperglycemia and adverse pregnancy outcomes. N Engl J Med. 2008; 358(19): 1991-2002. doi: https://doi.org/10.1056/nejmoa0707943.
5. Li Z, Cheng Y, Wang D, et al. Incidence Rate of Type 2 Diabetes Mellitus after Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis of 170,139 Women. J Diabetes Res. 2020; 2020: 3076463. doi: https://doi.org/10.1155/2020/3076463.
6. Kramer CK, Campbell S, Retnakaran R. Gestational diabetes and the risk of cardiovascular disease in women: a systematic review and meta-analysis. Diabetologia. 2019; 62(6): 905-14. doi: https://doi.org/10.1007/s00125-019-4840-2.
7. Mantzorou M, Papandreou D, Pavlidou E, et al. Maternal Gestational Diabetes Is Associated with High Risk of Childhood Overweight and Obesity: A Cross-Sectional Study in Pre-School Children Aged 2-5 Years. Medicina (Kaunas). 2023; 59(3): 455. doi: https://doi.org/10.3390/medicina59030455.
8. Lowe WL, Jr., Scholtens DM, Kuang A, et al. Hyperglycemia and Adverse Pregnancy Outcome Follow-up Study (HAPO FUS): Maternal Gestational Diabetes Mellitus and Childhood Glucose Metabolism. Diabetes Care. 2019; 42(3): 372-80. doi: https://doi.org/10.2337/dc18-1646.
9. Sadiya A, Jakapure V, Shaar G, Adnan R, Tesfa Y. Lifestyle intervention in early pregnancy can prevent gestational diabetes in high-risk pregnant women in the UAE: a randomized controlled trial. BMC Pregnancy Childbirth. 2022; 22(1): 668. doi: https://doi.org/10.1186/s12884-022-04972-w.
10. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021; 14(1): 49-58. doi: https://doi.org/10.1093/ckj/sfaa188.
11. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014; 14: 40. doi: https://doi.org/10.1186/1471-2288-14-40.
12. Huang QF, Hu YC, Wang CK, Huang J, Shen MD, Ren LH. Clinical First-Trimester Prediction Models for Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis. Biol Res Nurs. 2023; 25(2): 185-97. doi: https://doi.org/10.1177/10998004221131993.
13. van Leeuwen M, Opmeer BC, Zweers EJ, et al. External validation of a clinical scoring system for the risk of gestational diabetes mellitus. Diabetes Res Clin Pract. 2009; 85(1): 96-101. doi: https://doi.org/10.1016/j.diabres.2009.04.025.
14. Naylor CD, Sermer M, Chen E, Farine D. Selective screening for gestational diabetes mellitus. Toronto Trihospital Gestational Diabetes Project Investigators. N Engl J Med. 1997; 337(22): 1591-6. doi: https://doi.org/10.1056/nejm199711273372204.
15. Thériault S, Forest JC, Massé J, Giguère Y. Validation of early risk-prediction models for gestational diabetes based on clinical characteristics. Diabetes Res Clin Pract. 2014; 103(3): 419-25. doi: https://doi.org/10.1016/j.diabres.2013.12.009.
16. Meertens LJE, Scheepers HCJ, van Kuijk SMJ, et al. External validation and clinical utility of prognostic prediction models for gestational diabetes mellitus: A prospective cohort study. Acta Obstet Gynecol Scand. 2020; 99(7): 891-900. doi: https://doi.org/10.1111/aogs.13811.
17. Nanda S, Savvidou M, Syngelaki A, Akolekar R, Nicolaides KH. Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. Prenat Diagn. 2011; 31(2): 135-41. doi: https://doi.org/10.1002/pd.2636.
18. Gabbay-Benziv R, Doyle LE, Blitzer M, Baschat AA. First trimester prediction of maternal glycemic status. J Perinat Med. 2015; 43(3): 283-9. doi: https://doi.org/10.1515/jpm-2014-0149.
19. Lamain-de Ruiter M, Kwee A, Naaktgeboren CA, et al. External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. Bmj. 2016; 354: i4338. doi: https://doi.org/10.1136/bmj.i4338.
20. van Leeuwen M, Opmeer BC, Zweers EJ, et al. Estimating the risk of gestational diabetes mellitus: a clinical prediction model based on patient characteristics and medical history. Bjog. 2010; 117(1): 69-75. doi: https://doi.org/10.1111/j.1471-0528.2009.02425.x.
21. Pintaudi B, Di Vieste G, Corrado F, et al. Improvement of selective screening strategy for gestational diabetes through a more accurate definition of high-risk groups. Eur J Endocrinol. 2014; 170(1): 87-93. doi: https://doi.org/10.1530/eje-13-0759.
22. Shirazian N, Emdadi R, Mahboubi M, et al. Screening for gestational diabetes: usefulness of clinical risk factors. Arch Gynecol Obstet. 2009; 280(6): 933-7. doi: https://doi.org/10.1007/s00404-009-1027-y.
23. Teede HJ, Harrison CL, Teh WT, Paul E, Allan CA. Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention. Aust N Z J Obstet Gynaecol. 2011; 51(6): 499-504. doi: https://doi.org/10.1111/j.1479-828x.2011.01356.x.
24. Kotzaeridi G, Blätter J, Eppel D, et al. Performance of early risk assessment tools to predict the later development of gestational diabetes. Eur J Clin Invest. 2021; 51(12): e13630. doi: https://doi.org/10.1111/eci.13630.

25. Benhalima K, Van Crombrugge P, Moyson C, et al. Estimating the risk of gestational diabetes mellitus based on the 2013 WHO criteria: a prediction model based on clinical and biochemical variables in early pregnancy. Acta Diabetol. 2020; 57(6): 661-71. doi: https://doi.org/10.1007/s00592-019-01469-5.
26. Syngelaki A, Pastides A, Kotecha R, Wright A, Akolekar R, Nicolaides KH. First-Trimester Screening for Gestational Diabetes Mellitus Based on Maternal Characteristics and History. Fetal Diagn Ther. 2015; 38(1): 14-21. doi: https://doi.org/10.1159/000369970.
27. Liu Y, Lu H. External Validation and Comparison of Four Risk Prediction Models for Gestational Diabetes Mellitus. Journal of Hubei University of Medicine. 2023; 42(1): 57-62. doi: https://doi.org/10.13819/j.issn.2096-708X.2023.01.011.
28. Rao J-W, Zhang J-F, Han F-Q, et al. Analysis of Risk Factors of Gestational Diabetes Mellitus and Establishment of Risk Prediction Model. Modern Preventive Medicine. 2022; 49(03): 441-46. Available from: https://d.wanfangdata.com.cn/periodical/xdyfyx202203012.
29. Obstetrics Group of the Obstetrics and Gynecology Branch of the Chinese Medical Association, Pregnancy Complicated with Diabetes Collaborative Group of the Perinatal Medicine Branch of the Chinese Medical Association. [Diagnosis and therapy guideline of pregnancy with diabetes mellitus]. Zhonghua Fu Chan Ke Za Zhi. 2014; 49(8): 561-9. doi: https://doi.org/10.3760/cma.j.issn.0529-567X.2014.08.001.
30. Chen XP, Zhang Y, Zhuang YY, Zhang ZH. PROBAST: a tool for assessing risk of bias in the study of diagnostic or prognostic multi-factorial predictive models. Chinese Journal of Evidence-Based Medicine. 2020; 20(06): 737-44. doi: https://doi.org/10.7507/1672-2531.201910087.
31. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016; 35(2): 214-26. doi: https://doi.org/10.1002/sim.6787.
32. Chen X. The Risk Factors and Risk Assessment Model of Gestational Diabetes Mellitus. MS thesis, Soochow University; 2016. doi: https://doi.org/10.7666/d.D01006872.
33. Chen MF. Early Prediction Model for Gestational Diabetes Mellitus Combining Demographic Characteristics and Clinical Characteristics. MS thesis, Zhejiang University; 2019. doi: https://doi.org/10.27461/d.cnki.gzjdx.2019.001238.
34. Gao S, Leng J, Liu H, et al. Development and validation of an early pregnancy risk score for the prediction of gestational diabetes mellitus in Chinese pregnant women. BMJ Open Diabetes Res Care. 2020; 8(1): e000909. doi: https://doi.org/10.1136/bmjdrc-2019-000909.
35. Guo F, Yang S, Zhang Y, Yang X, Zhang C, Fan J. Nomogram for prediction of gestational diabetes mellitus in urban, Chinese, pregnant women. BMC Pregnancy Childbirth. 2020; 20(1): 43. doi: https://doi.org/10.1186/s12884-019-2703-y.
36. Li JJ, Yang MY, Jiang MH. Establishment and Application Value of an Early Screening Model for Gestational Diabetes Mellitus in Pregnant Women. Maternal and Child Health Care of China. 2021; 36(1): 69-73. doi: https://doi.org/10.19829/j.zgfybj.issn.1001-4411.2021.01.024.
37. Li SH, Yue ZH, Zuo Q, et al. Development and Evaluation of an Early Risk Prediction Model for Gestational Diabetes Mellitus. Chinese Journal of Preventive Medicine. 2024; 25(1): 87-91. doi: https://doi.org/10.16506/j.1009-6639.2024.01.015.