

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. [PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies](#)
2. [PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration](#)

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
<i>Intended use of model:</i>	This study develops an interpretable ML model combining systemic inflammatory indices and traditional clinical markers to predict preterm birth in GDM. Enabling early risk stratification at diagnosis, it facilitates timely interventions for this high-risk population.
Participants including selection criteria and setting:	From August 2019 to August 2024, adults with singleton gestational diabetes mellitus diagnosed by a 75-g OGTT at 24–28 gestational weeks were consecutively screened at Zhoupu Hospital. Patients with pregestational diabetes, multifetal pregnancy, prior preterm birth, incomplete records, or major renal, hepatic, or cardiovascular disease were excluded.
Predictors (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/prohibitions for specialized equipment):	This study collected comprehensive maternal data including demographic characteristics (age, pre-pregnancy BMI), medical history (diabetes family history, smoking/alcohol use, hypertension, parity, uterine curettage, IVF-ET), OGTT results, and hematological indices at GDM diagnosis (complete blood counts and derived inflammatory ratios: NLR, PLR, LMR, SII).
<i>Outcome to be predicted:</i>	Preterm birth

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	PROBAST boxes to complete	Tick as appropriate	Definition for type of prediction model study
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	✓	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	This study
Models of interest	Machine learning
Outcome of interest	Preterm birth

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that “yes” indicates absence of bias. Any signalling question rated as “no” or “probably no” flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as “high”, “low” or “unclear” risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
<p><i>Describe the sources of data and criteria for participant selection:</i></p> <p>This retrospective cohort study was conducted from August 2019 to August 2024, during which 568 patients with GDM were consecutively screened from Shanghai University of Medicine & Health Sciences affiliated Zhoupu Hospital. Participants met the following criteria: (1) age ≥ 18 years; (2) GDM diagnosis confirmed by 75g oral glucose tolerance test (OGTT) at 24-28 weeks' gestation (fasting glucose ≥ 5.1 mmol/L, 1-hour ≥ 10.0 mmol/L, or 2-hour ≥ 8.5 mmol/L); and (3) singleton pregnancy. Exclusion criteria included pre-existing diabetes, multifetal gestation, prior preterm birth, incomplete medical records, and significant comorbidities (renal, hepatic, or cardiovascular diseases). Following strict inclusion/exclusion criteria, 389 GDM patients were enrolled.</p>			
	Dev	Val	
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	Y	Y	
1.2 Were all inclusions and exclusions of participants appropriate?	Y	Y	
Risk of bias introduced by selection of participants	RISK: (low/ high/ unclear)	Low	Low
<p><i>Rationale of bias rating:</i></p> <p>Participants were consecutively enrolled from a well-defined cohort using appropriate inclusion and exclusion criteria, with no evidence of selection based on outcome status.</p>			
B. Applicability			
<p><i>Describe included participants, setting and dates:</i></p> <p>This retrospective cohort study was conducted from August 2019 to August 2024, during which 568 patients with GDM were consecutively screened from Shanghai University of Medicine & Health Sciences affiliated Zhoupu Hospital. Participants met the following criteria: (1) age ≥ 18 years; (2) GDM diagnosis confirmed by 75g oral glucose tolerance test (OGTT) at 24-28 weeks' gestation (fasting glucose ≥ 5.1 mmol/L, 1-hour ≥ 10.0 mmol/L, or 2-hour ≥ 8.5 mmol/L); and (3) singleton pregnancy. Exclusion criteria included pre-existing diabetes, multifetal gestation, prior preterm birth, incomplete medical records, and significant comorbidities (renal, hepatic, or cardiovascular diseases). Following strict inclusion/exclusion criteria, 389 GDM patients were enrolled.</p>			
Concern that the included participants and setting do not match the review question	CONCERN: (low/ high/ unclear)	Low	Low
<p><i>Rationale of applicability rating:</i></p> <p>The study population, clinical setting, and time frame are consistent with the target population defined in the study objective.</p>			

DOMAIN 2: Predictors			
A. Risk of Bias			
<p><i>List and describe predictors included in the final model, e.g. definition and timing of assessment:</i></p> <p>The final ML model incorporated seven key variables: maternal age, uterine curettage, and five inflammatory indices (monocyte count, NLR, PLR, LMR, and SII).</p>			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		Y	Y
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	RISK: (low/ high/ unclear)	Low	Low
<p><i>Rationale of bias rating:</i></p> <p>All predictors were clearly defined, consistently measured across participants, assessed without knowledge of outcome data, and available at the intended time of model use; therefore, the risk of bias was considered low.</p>			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	CONCERN: (low/ high/ unclear)	Low	Low
<p><i>Rationale of applicability rating:</i></p> <p>The predictors used in the model were clinically relevant and defined, measured, and timed consistently with the study objective and intended clinical application.</p>			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:			
In this context, this study aims to develop and validate an interpretable ML model that integrates systemic inflammation indices with traditional clinical predictors to identify novel and practical obstetric biomarkers for preterm birth risk assessment in GDM patients. By enabling early risk stratification at the time of GDM diagnosis, this approach will facilitate timely clinical interventions to prevent preterm delivery in this high-risk population.			
		Dev	Val
3.1	Was the outcome determined appropriately?	Y	Y
3.2	Was a pre-specified or standard outcome definition used?	Y	Y
3.3	Were predictors excluded from the outcome definition?	Y	Y
3.4	Was the outcome defined and determined in a similar way for all participants?	Y	Y
3.5	Was the outcome determined without knowledge of predictor information?	Y	Y
3.6	Was the time interval between predictor assessment and outcome determination appropriate?	Y	Y
Risk of bias introduced by the outcome or its determination		RISK: (low/ high/ unclear)	Low Low
Rationale of bias rating: The outcome was predefined using standard criteria, consistently determined for all participants, independent of predictor information, and assessed at an appropriate time interval after predictor measurement; therefore, the risk of bias was judged to be low.			
B. Applicability			
At what time point was the outcome determined:			
At delivery			
If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:			
Not applicable			
Concern that the outcome, its definition, timing or determination do not match the review question		CONCERN: (low/ high/ unclear)	Low Low
Rationale of applicability rating: The outcome definition, timing, and method of determination were consistent with the study objective and intended clinical application, resulting in low concern regarding applicability.			

DOMAIN 4: Analysis			
Risk of Bias			
<p><i>Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor:</i></p> <p>A total of 389 participants were included in the analysis, among whom 53 experienced the outcome. 7 candidate predictors were considered, resulting in an events-per-predictor ratio of approximately 7.6.</p>			
<p><i>Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):</i></p> <p>The prediction model was developed using an XGBoost algorithm with hyperparameter tuning. Feature selection was based on univariate and multivariable logistic regression, without applying arbitrary risk stratification.</p>			
<p><i>Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):</i></p> <p>Internal validation and temporal external validation</p>			
<p><i>Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:</i></p> <p>The model's performance was evaluated using discrimination metrics (AUC-ROC, AUC-PRC, sensitivity, specificity, PPV, NPV, F1 score), calibration measures (brier score, calibration curves), and clinical utility assessment via DCA. The Brier score quantifies prediction accuracy (lower values indicating better calibration), while DCA estimates net clinical benefit.</p>			
<p><i>Describe any participants who were excluded from the analysis:</i></p> <p>No participants were excluded from the analysis after applying eligibility criteria.</p>			
<p><i>Describe missing data on predictors and outcomes as well as methods used for missing data:</i></p> <p>Missing data were handled using multiple imputation.</p>			
		Dev	Val
4.1	Were there a reasonable number of participants with the outcome?	Y	Y
4.2	Were continuous and categorical predictors handled appropriately?	Y	Y
4.3	Were all enrolled participants included in the analysis?	Y	Y
4.4	Were participants with missing data handled appropriately?	Y	Y
4.5	Was selection of predictors based on univariable analysis avoided?	Y	
4.6	Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?	Y	Y
4.7	Were relevant model performance measures evaluated appropriately?	Y	Y
4.8	Were model overfitting and optimism in model performance accounted for?	Y	
4.9	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	Y	
Risk of bias introduced by the analysis		RISK: (low/ high/ unclear)	Low Low
<p><i>Rationale of bias rating:</i></p> <p>The model was developed using appropriate statistical and machine-learning methods with adequate sample size, proper handling of predictors and missing data, internal validation, temporal external validation, and appropriate performance evaluation, with measures taken to address overfitting and optimism. Therefore, the risk of bias in the analysis domain was judged to be low.</p>			

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.

Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a <u>prediction model was developed without any external validation</u> , and it was rated as <u>low risk of bias for all domains</u> , consider downgrading to high risk of bias . Such a model can only be considered as low risk of bias, if the development was based on a very large data set <u>and</u> included some form of internal validation.
High risk of bias	If at least one domain is judged to be at high risk of bias .
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have low concerns regarding applicability .
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have high concerns regarding applicability .
Unclear concerns regarding applicability	If unclear concerns (but no “high concern”) regarding applicability for at least one domain, the prediction model evaluation is judged to have unclear concerns regarding applicability overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	RISK: (low/ high/ unclear)	Low
<i>Summary of sources of potential bias:</i> The overall risk of bias was judged to be low, as all domains were assessed as low risk of bias and the model development included appropriate internal validation.		
Overall judgement of applicability	CONCERN: (low/ high/ unclear)	Low
<i>Summary of applicability concerns:</i> The prediction model evaluation showed low concerns regarding applicability, as the participants, predictors, and outcome were consistent with the review question.		