Original Paper

Miscarriage Risk Prediction for Women With Polycystic Ovary Syndrome Who Achieve Clinical Pregnancy After Fresh Embryo Transfer: Multicenter Development, Geographic External Validation, and Prospective Temporal Validation Study

Juncheng Tong[1,2], PhD; Yingying Zhang[3], MD; Yanhui Liu[2,4], PhD; Lifang Liu[4], MSc; Xiaoteng Shang[5], MD; Fanggui Wu[4], BSc; Haoxuan Guo[6], BSc; Sibo Zhang[6], MSc; Kai Zhao[7], MD; Aifa Tang[4,8*]; Mengyuan Qu[4,9*]

[1]Department of Basic Medical Sciences, Shenzhen University, Shenzhen, Guangdong, China

[2]The Third Affiliated Hospital of Shenzhen University, Shenzhen, Guangdong, China

[3]Center for Reproductive Medicine, Third Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

[4]The Reproductive Medical Center, Shenzhen Luohu People's Hospital, The Third Affiliated Hospital of Shenzhen University, Shenzhen, Guangdong, China

[5]Shantou University Medical College, Shantou, Guangdong, China

[6]School of Life Science and Technology, Wuhan Polytechnic University, Wuhan, Hubei, China

[7]Institute of Reproductive Health, Huazhong University of Science and Technology, Wuhan, Hubei, China

[8]Luohu Clinical Institute of Shantou University Medical College, Shantou University Medical College, Shenzhen, China

[9]Luohu Clinical Institute of Shantou University Medical College, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Mengyuan Qu
The Reproductive Medical Center
Shenzhen Luohu People's Hospital, The Third Affiliated Hospital of Shenzhen University
47 Youyi Road
Shenzhen, Guangdong, 518000
China
Phone: 86 18681458948
Email: qumengyuan@qq.com

Abstract

Background: Women with polycystic ovary syndrome (PCOS) who achieve clinical pregnancy following fresh in vitro fertilization (IVF)/intracytoplasmic sperm injection (ICSI) carry an elevated miscarriage risk. However, individualized risk stratification after clinical pregnancy confirmation but before outcome determination remains challenging in clinical practice.

Objective: This study aimed to develop, externally and prospectively validate over time, and interpret a machine learning–based risk stratification model for women with PCOS who had achieved a confirmed clinical pregnancy after fresh IVF/ICSI.

Methods: This multicenter study included 8911 clinical pregnancy IVF/ICSI cycles across development (n=7241; miscarriages: n=973), external validation (n=1170; miscarriages: n=146), and prospective temporal validation (n=500; miscarriages: n=62) cohorts. Prespecified predictors contained no observed missing values; training-fold median imputation was kept only as a leakage-safe fallback. A total of 13 candidate machine learning and traditional statistical algorithms were compared internally using repeated stratified 5-fold cross-validation (5 repeats) within the development cohort; preprocessing and hyperparameter tuning were performed exclusively in training folds. A prelocked Extreme Gradient Boosting (XGBoost) model was carried forward unchanged for external and prospective temporal validation. Cross-fitted logistic recalibration of averaged XGBoost out-of-fold probabilities was evaluated as a secondary development-sample diagnostic analysis. Model performance encompassed discrimination, precision-recall, calibration, threshold-based metrics, decision curve analysis, risk stratification, and Shapley additive explanations–based interpretability.

Results: Miscarriage rates were 13.4% (973/7241), 12.5% (146/1170), and 12.4% (62/500) for the development, external validation, and prospective temporal validation cohorts, respectively. In repeated cross-validation, XGBoost (area under the

XSL•FO
RenderX

receiver operating characteristic curve [AUC] 0.885, 95% CI 0.874-0.896) and LightGBM (AUC 0.882, 95% CI 0.871-0.893) showed comparable discrimination; the paired XGBoost-LightGBM DeLong test was not significant (Holm-adjusted *P*=.07). Raw XGBoost out-of-fold probabilities overestimated absolute risk (mean predicted risk 30.3% vs observed risk 13.4%); cross-fitted recalibration within the development cohort improved calibration but was not propagated to validation cohorts. The locked XGBoost model achieved AUCs of 0.787 (95% CI 0.748-0.827; Brier score 0.099) in external validation and 0.763 (95% CI 0.709-0.816; Brier score 0.104) in prospective temporal validation, showing cross-cohort performance attenuation. Risk-quartile stratification using fixed development-cohort cutoffs showed consistent event enrichment across validation cohorts, but positive predictive values remained modest. Shapley additive explanations analysis identified anti-Müllerian hormone, endometrial thickness on the day of human chorionic gonadotropin administration, and basal luteinizing hormone as leading predictors.

Conclusions: The locked XGBoost model using routine preoutcome variables may support miscarriage risk stratification and counseling for women with PCOS with clinical pregnancy after fresh IVF/ICSI. Given cross-cohort discrimination loss and calibration drift, it should serve only as a risk-ranking tool, not for diagnosis or treatment selection. Local calibration monitoring is mandatory before any real-world clinical deployment.

(J Med Internet Res 2026;28:e101829) doi: 10.2196/101829

KEYWORDS

polycystic ovary syndrome; PCOS; polyendocrine metabolic ovarian syndrome; PMOS; miscarriage; in vitro fertilization/intracytoplasmic sperm injection; IVF/ICSI; prediction model; machine learning; artificial intelligence; AI

Introduction

Polycystic ovary syndrome (PCOS) is the most prevalent endocrine disorder affecting women of reproductive age, with an estimated prevalence of 10% to 15% depending on the diagnostic criteria applied [1,2]. Recently, it has been proposed to rename PCOS as polyendocrine-metabolic ovarian syndrome [3]. The syndrome is characterized by hyperandrogenism, ovulatory dysfunction, polycystic ovarian morphology, and a high prevalence of metabolic comorbidities, including insulin resistance, obesity, and dyslipidemia [4]. International evidence-based guidance emphasizes precise diagnosis, individualized management, and monitoring of reproductive and metabolic risks throughout the lifespan. Mounting clinical evidence has confirmed that women with PCOS carry substantially elevated risks of adverse pregnancy outcomes, particularly miscarriage, following assisted reproductive technologies [5,6].

For women with PCOS who undergo in vitro fertilization (IVF)/intracytoplasmic sperm injection (ICSI) and achieve clinical pregnancy, miscarriage remains a devastating complication. Population-based and cohort studies have demonstrated that the miscarriage rate among women with PCOS is approximately 1.5-fold higher than that among women without PCOS [7]. Miscarriage in this population is a multifactorial event driven by interactions among endocrine disturbances, metabolic abnormalities, impaired endometrial receptivity, suboptimal embryo quality, and altered luteal support [8-10]. Despite this well-established association, clinicians still lack reliable tools for individualized risk stratification before pregnancy outcomes are ascertained. Traditional assessment relies on single factors, such as age, BMI, or previous miscarriage history, which cannot capture the complex interplay of these determinants and therefore fail to support individualized counseling or to guide targeted preventive strategies.

Machine learning has emerged as a useful approach for predicting clinical outcomes using high-dimensional tabular data. Among modern algorithms, Extreme Gradient Boosting (XGBoost), light gradient boosting machine (LightGBM), CatBoost, TabNet, and related methods are frequently applied to structured clinical data because they can model nonlinear relationships and complex interactions. Several studies have reported that machine learning can predict miscarriage in IVF embryo transfer pregnancies and may outperform traditional logistic regression models in some settings [11,12]. For a prediction model to be clinically useful, however, discrimination alone is insufficient; calibration, external transportability, transparent reporting, and prospective evaluation are also required. Model interpretability is equally critical for clinical adoption, as clinicians require insight into which features drive individual risk predictions. Transparent reporting following guidelines such as TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for AI) is also essential to minimize bias and support clinical translation [13].

Accordingly, the primary objective of this study was to develop and validate a machine learning model for estimating individual miscarriage risk among women with PCOS after confirmed clinical pregnancy following fresh IVF/ICSI. Secondary objectives included comparative evaluation of multiple candidate prediction algorithms, assessment of model calibration, risk-ranking performance, threshold-dependent metrics, and cross-cohort transportability. Exploratory analyses encompassed early- and late-miscarriage subtype models and their corresponding Shapley additive explanations (SHAP) interpretability analyses. This tool is intended for reproductive medicine clinicians for risk communication and follow-up planning only after clinical pregnancy confirmation and before miscarriage outcome ascertainment; it is not designed for diagnostic purposes, treatment selection, or automated clinical interventions.

XSL•FO
RenderX

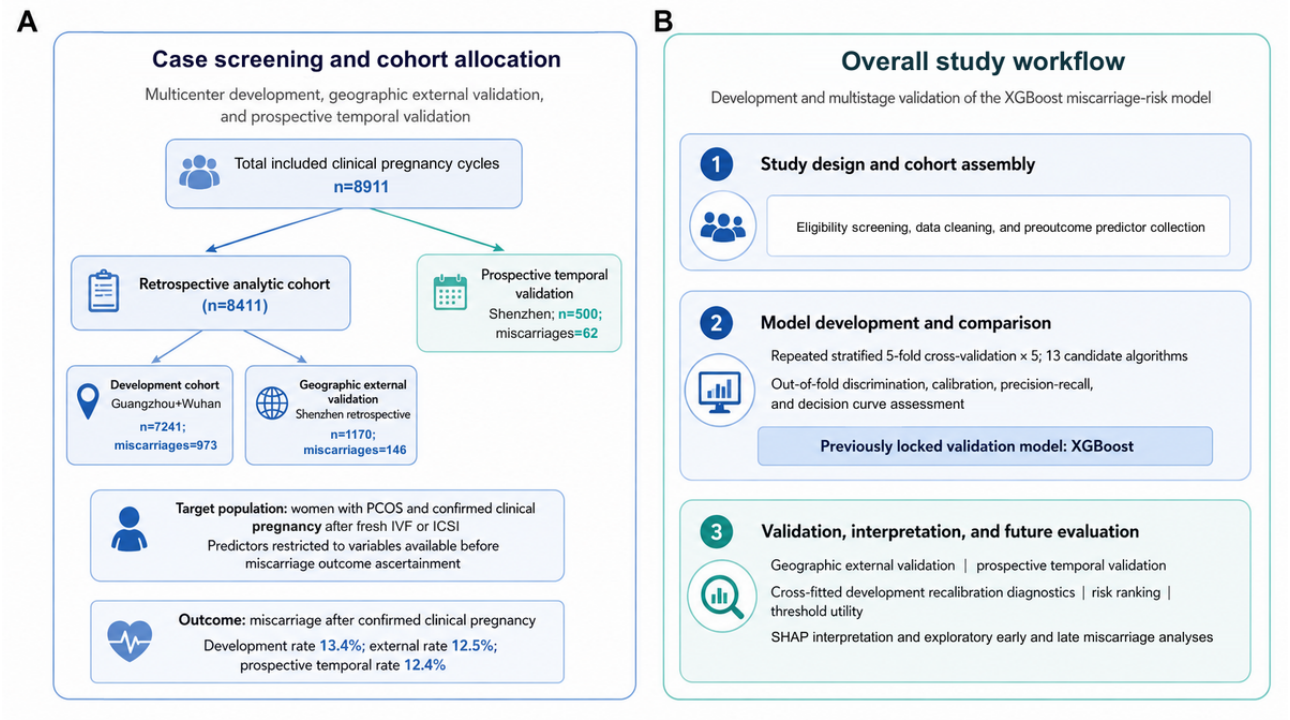
Methods

Study Design and Cohort Allocation

We conducted a multicenter prediction model development and validation study among women with PCOS who achieved clinical pregnancy following fresh IVF/ICSI. Data were collected from tertiary reproductive medicine centers in Guangzhou, Wuhan, and Shenzhen, China. Retrospective data were collected between January 1, 2016, and December 31, 2024. The development cohort combined Guangzhou and Wuhan data (n=7241). Shenzhen retrospective data formed the geographic external validation cohort (n=1170). A separate prospective temporal validation cohort was enrolled consecutively at the same Shenzhen center from January 1, 2025, to October 1, 2025, with a fixed sample size of 500 and no selective exclusion; this cohort was not an independent geographically separated external validation. The prelocked XGBoost model was applied without re-estimation, and all outcomes were ascertained after enrollment completion to ensure an unbiased prospective temporal validation.

All 7241 development records were evaluated using repeated stratified 5-fold cross-validation with 5 repeats. Candidate models were compared using identical outer folds. Preprocessing and tuning were performed exclusively within corresponding training portions to prevent information leakage, and performance was summarized from record-level averaged out-of-fold predictions. The XGBoost model had been prelocked before both validation assessments and applied unchanged. The study flowchart is shown in Figure 1.

Figure 1. Study cohort screening and analytic workflow for miscarriage risk prediction in women with polycystic ovary syndrome (PCOS) after confirmed clinical pregnancy following fresh in vitro fertilization (IVF)/intracytoplasmic sperm injection (ICSI). (A) Allocation of 8911 clinical pregnancy cycles to development, geographic external validation, and prospective temporal validation cohorts. (B) Analytic workflow: repeated stratified 5-fold cross-validation with 5 repeats for internal model comparison, application of the prelocked Extreme Gradient Boosting (XGBoost) model (unchanged) to validation cohorts, development-only cross-fitted recalibration diagnostics, risk ranking and threshold utility analyses, and exploratory Shapley additive explanations (SHAP)–based interpretation. See Multimedia Appendix 2 for a higher-resolution version of this figure.



Study Population

All women were diagnosed with PCOS according to the Rotterdam 2003 criteria. Clinical pregnancy was defined as an intrauterine gestational sac confirmed by transvaginal ultrasound 4 to 5 weeks after embryo transfer. Biochemical pregnancies were excluded because the intended prediction time was after clinical pregnancy confirmation; consequently, the target outcome was miscarriage after ultrasound-confirmed clinical pregnancy rather than all posttransfer pregnancy loss. This restriction created a selected postconception cohort and limits generalizability to women who have already achieved a clinical pregnancy. Inclusion criteria were (1) PCOS diagnosed using the Rotterdam 2003 criteria, (2) aged 20 to 45 years and undergoing IVF/ICSI embryo transfer, (3) confirmed clinical pregnancy, and (4) complete outcome data. Exclusion criteria were (1) ectopic pregnancy or hydatidiform mole; (2) adenomyosis, uterine malformations, or moderate to severe intrauterine adhesions; (3) antiphospholipid syndrome or inherited thrombophilia; (4) pregnancy termination for medical or social reasons; (5) preimplantation genetic testing cycles; and (6) missing outcome data or missing variables that could not be handled under the prespecified preprocessing workflow.

The primary outcome was miscarriage vs no miscarriage after a confirmed clinical pregnancy. Early miscarriage was defined as pregnancy loss before 12 weeks of gestation, and late miscarriage as pregnancy loss from 12 to 28 weeks of gestation.

XSL•FO
RenderX

Candidate Predictors and Leakage Control

All prespecified model predictors were selected on clinical grounds and restricted to variables fully ascertained before outcome determination to prevent data leakage. Final harmonized predictors entering the locked prediction model included female age, male age, female BMI, basal luteinizing hormone (LH), prolactin (PRL), testosterone, anti-Müllerian hormone (AMH), human chorionic gonadotropin (hCG)-day endometrial thickness (EMT), total transferred embryo count, high-quality embryo count, and embryo transfer stage. Additional clinical variables were collected for descriptive baseline characterization only but were excluded from the final locked modeling matrix. These included follicle-stimulating hormone (FSH), antral follicle count (AFC), infertility duration, prior miscarriage history, multiple hCG-day endocrine markers, and number of oocytes retrieved. These variables were not incorporated into the prediction model because they could not be harmonized consistently across all 3 study cohorts (development, geographic external validation, and prospective temporal validation). Baseline between-group comparisons are for cohort descriptive purposes only and do not represent predictor selection steps for the machine learning model.

Variables that were unavailable before pregnancy outcome ascertainment or directly related to the outcome were not used as predictors, including biochemical pregnancy, clinical pregnancy, pregnancy outcome, and miscarriage timing. Administrative variables and center labels were also excluded to avoid data leakage and minimize institutional confounding.

Preprocessing

Before modeling, records with missing outcome data or nonimputable core eligibility information were excluded before predictor preprocessing. No observed missing values existed among the 11 prespecified predictors within the harmonized development cohort dataset. Training-fold–derived median imputation was implemented as a leakage-safe fallback safeguard within the pipeline but was never executed for the final analysis dataset. Continuous variables were standardized using the training-fold mean and SD. The embryo stage was one-hot encoded following the prespecified validation protocol reference category. All preprocessing parameters fitted within each training fold were propagated unchanged to held-out test folds, the geographic external validation cohort, and the prospective temporal validation cohort. The outcome exhibited moderate class imbalance. For XGBoost, the scale_pos_weight parameter was derived from the negative-to-positive event ratio within each training fold. No oversampling, undersampling, or synthetic data augmentation was used to retain real-world population distributions for clinical generalizability. Skewed continuous variables were not log-transformed, as tree-based algorithms are robust to monotonic transformations. Full model building, hyperparameter, and software environment specifications are documented in Table S1 in Multimedia Appendix 1 [11,12,14]; predictor availability, missing-value handling, preprocessing, and variable inclusion details are provided in Table S2 in Multimedia Appendix 1 [11,12,14].

Sample Size Adequacy

Because we used a retrospective multicenter dataset, a post hoc sample size adequacy assessment was conducted using the framework described by Riley et al [15] for binary prediction models. With 17 modeling parameters from 11 prespecified predictors (Table S2C in Multimedia Appendix 1 [11,12,14]), an expected miscarriage prevalence of 13.4%, a target shrinkage of 0.90, and a conservative C-statistic of 0.759 [11], *pmsampsize* estimated a minimum requirement of 1404 participants and 189 events. Our development cohort (7241 cycles, 973 miscarriages; 57.2 events per modeling parameter) exceeded this requirement.

Candidate Models and Internal Model Comparison

The following candidate algorithms were evaluated: logistic regression, least absolute shrinkage and selection operator (LASSO) logistic regression, k-nearest neighbors, decision tree, random forest, support vector machine with radial basis kernel, gradient boosting machine, neural network, naive Bayes, XGBoost, LightGBM, CatBoost, and TabNet. Repeated cross-validation was used to characterize relative algorithm performance; this comparison was not applied to reselect or modify the prelocked model used for external validation.

Validation and Performance Assessment

Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC) with 95% CIs, along with average precision for precision-recall performance. Calibration was evaluated using Brier score, expected calibration error (ECE), calibration intercept, calibration slope, calibration-in-the-large (CITL), and decile-level calibration plots. Calibration-related metrics were derived from model-predicted risks vs observed miscarriage outcomes. Internal model comparison used record-level averaged out-of-fold probabilities. Paired AUC differences between XGBoost and comparator models were assessed using DeLong tests with Holm adjustment. For the focused XGBoost-LightGBM comparison, cycle-level bootstrap resampling (1000 replicates) was used to compare Brier score, ECE, absolute calibration-slope deviation from 1, and absolute CITL, with Holm adjustment. A total of 6 prespecified fixed thresholds (0.05, 0.10, 0.15, 0.20, 0.25, and 0.30) were used to calculate sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F_1-score, and net benefit from decision curve analysis (DCA). The Youden index–derived threshold was reported descriptively and was not treated as a clinical cutoff. Risk stratification was assessed using predicted-risk groups; fixed cut points were derived once from prelocked internal reference predictions and applied unchanged to the external and prospective temporal validation cohorts. No recalibration or model updating was applied to the locked model during validation. As a secondary diagnostic analysis on the development sample, repeated cross-fitted logistic recalibration was performed; the resulting recalibrated probabilities were only used for secondary calibration, threshold utility, risk ranking, and subgroup analyses.

Model Interpretation

SHAP analysis was used solely for model interpretation and feature attribution. SHAP importance for overall miscarriage

XSL•FO
RenderX

was assessed for the locked XGBoost model. Separate early- and late-miscarriage subtype models and their SHAP analyses were designated exploratory. SHAP values describe the fitted model's contributions and do not establish biological causality, optimal clinical ranges, or treatment effects.

Statistical Analysis

Continuous variables are presented as median (IQR), and categorical variables as counts and percentages. All statistical tests were 2-sided. $P<.05$ was defined as the threshold for statistical significance for primary clinical comparisons; multiplicity-adjusted P values are reported for paired model comparisons. All analyses were performed using R (version 4.5.2; R Foundation for Statistical Computing).

Ethical Considerations

This study was performed in accordance with the Declaration of Helsinki. Ethics approval was obtained from Shenzhen Luohu People's Hospital (2026-LHQRMYY-KYLL-036), the Third Affiliated Hospital of Guangzhou Medical University (2024012), and Tongji Medical College (S032). Informed consent was waived by all ethics committees due to the retrospective, anonymized, and noninterventional nature of this study. No personally identifiable information was used or disclosed.

Results

Cohort Composition and Baseline Characteristics

A total of 8911 clinical pregnancy cycles with PCOS were included across development, external validation, and prospective temporal validation stages (Figure 1). The development cohort comprised 7241 cycles from Guangzhou and Wuhan (miscarriages: n=973, 13.4%), the geographic external validation cohort included 1170 cycles (miscarriages: n=146, 12.5%), and the prospective temporal validation cohort contained 500 cycles (miscarriages: n=62, 12.4%). Baseline demographics, endocrine profiles, and embryo transfer characteristics are summarized in Table 1. Compared with the nonmiscarriage groups, patients with PCOS who experienced miscarriage were older and had higher BMI, higher basal LH levels, higher basal FSH levels, higher testosterone levels, higher hCG-day progesterone levels, higher rates of previous miscarriage, lower AMH and AFC, fewer retrieved oocytes and high-quality embryos transferred, thinner EMT on hCG day, and a higher proportion of day 3 embryo transfers. Early miscarriage accounted for approximately 82.7% (977 /1181) of all pregnancy losses across cohorts.

XSL•FO
RenderX

Table 1. Baseline characteristics of women with polycystic ovary syndrome with confirmed clinical pregnancy, stratified by cohort and miscarriage status (N=8911)[a].

Characteristic	Development cohort (Guangzhou+Wuhan; n=7241)			External validation cohort (Shenzhen; n=1170)			Prospective temporal validation cohort (n=500)		
	No miscarriage (n=6268)	Miscarriage (n=973)	*P* value	No miscarriage (n=1024)	Miscarriage (n=146)	*P* value	No miscarriage (n=438)	Miscarriage (n=62)	*P* value
Demographic characteristics									
Female age (years), median (IQR)	31 (28-33)	32 (28-34)	.048	31 (28-33)	33 (30-35)	<.01	31 (28-33)	33 (30-35)	<.01
Male age (years), median (IQR)	34 (31-38)	34 (32-37)	.82	33 (31-38)	35 (32-38)	<.01	34 (31-38)	34 (33-38)	.91
Female BMI (kg/m²), median (IQR)	22.1 (20.0-24.5)	22.7 (20.6-25.0)	<.01	21.5 (19.9-23.7)	22.7 (20.6-25.4)	<.01	21.9 (20.3-24.2)	22.8 (20.9-25.4)	.04
Infertility duration (years), median (IQR)	3.2 (2.0-5.0)	3.6 (2.1-5.5)	<.01	3.0 (2.0-4.8)	3.8 (2.5-5.9)	<.01	3.1 (2.0-5.0)	3.7 (2.3-5.6)	.07
Primary infertility, n (%)	4463 (71.2)	641 (65.9)	<.01	743 (72.6)	92 (63.0)	.02	314 (71.7)	39 (62.9)	.16
Previous miscarriage history, n (%)	620 (9.9)	174 (17.9)	<.01	101 (9.9)	28 (19.2)	<.01	44 (10.0)	13 (21.0)	.01
Endocrine and ovarian reserve characteristics, median (IQR)									
Basal FSH[b] (IU/L)	6.7 (5.3-8.1)	7.1 (5.7-8.8)	<.01	6.6 (5.2-7.9)	7.4 (6.1-9.0)	<.01	6.7 (5.4-8.2)	7.5 (6.0-9.1)	<.01
Basal LH[c] (IU/L)	4.09 (2.74-6.28)	4.71 (3.50-6.81)	<.01	4.00 (2.76-6.02)	4.80 (3.40-6.90)	<.01	4.12 (2.77-6.44)	4.90 (3.20-7.00)	.04
Basal E2[d] (pg/mL)	43.6 (31.5-58.8)	45.1 (32.8-60.7)	.03	43.2 (31.2-58.5)	45.6 (33.0-61.2)	.09	43.5 (31.4-58.7)	46.0 (33.1-62.4)	.18
PRL[e] (ng/mL)	14.63 (10.45-20.66)	14.54 (10.81-19.61)	.70	16.18 (10.97-21.57)	15.50 (10.80-21.10)	.48	16.40 (11.45-21.24)	15.80 (11.00-20.80)	.51
Testosterone (ng/mL)	0.35 (0.26-0.46)	0.39 (0.29-0.52)	<.01	0.35 (0.27-0.47)	0.41 (0.30-0.53)	<.01	0.34 (0.25-0.47)	0.40 (0.30-0.54)	.01
AMH[f] (ng/mL)	5.83 (4.20-8.16)	5.16 (3.80-6.90)	<.01	6.02 (4.22-8.47)	5.20 (3.70-6.90)	<.01	6.10 (4.27-8.12)	5.05 (3.60-6.80)	<.01
AFC[g] (n)	14 (10-20)	12 (8-17)	<.01	15 (10-21)	11 (8-16)	<.01	14 (10-20)	11 (8-16)	<.01
Controlled ovarian stimulation characteristics, median (IQR)									
E2 on hCG[h] day (pg/mL)	2860 (1920-4300)	2580 (1710-3890)	<.01	3000 (2010-4520)	2550 (1650-3900)	<.01	2850 (1880-4250)	2480 (1580-3720)	.09
FSH on hCG day (IU/L)	10.2 (7.6-13.5)	10.1 (7.5-13.4)	.21	10.2 (7.6-13.5)	10.0 (7.3-13.2)	.61	10.3 (7.7-13.6)	10.0 (7.4-13.2)	.61
LH on hCG day (IU/L)	1.74 (1.05-2.82)	1.92 (1.14-3.10)	<.01	1.68 (1.01-2.70)	1.95 (1.18-3.22)	.04	1.72 (1.04-2.76)	2.03 (1.20-3.30)	.13
Progesterone level on hCG day (ng/mL)	0.87 (0.58-1.21)	1.05 (0.70-1.42)	<.01	0.85 (0.57-1.18)	1.10 (0.73-1.48)	<0.01	0.86 (0.57-1.20)	1.12 (0.75-1.53)	<.01
Endometrial thickness on hCG day (mm)	10.4 (9.0-12.0)	10.2 (9.0-11.5)	.04	10.5 (9.0-12.0)	10.1 (8.8-11.6)	.03	10.7 (9.0-12.0)	10.0 (8.8-11.5)	.01
Number of oocytes retrieved	11 (7-16)	9 (6-14)	<.01	12 (8-17)	9 (5-13)	<.01	11 (7-16)	8 (5-13)	<.01

XSL•FO
RenderX

Characteristic	Development cohort (Guangzhou+Wuhan; n=7241)			External validation cohort (Shenzhen; n=1170)			Prospective temporal validation cohort (n=500)		
	No miscarriage (n=6268)	Miscarriage (n=973)	*P* value	No miscarriage (n=1024)	Miscarriage (n=146)	*P* value	No miscarriage (n=438)	Miscarriage (n=62)	*P* value
Fertilization method, n (%)			.08			.33			.60
IVF[i]	4450 (71.0)	660 (67.8)		754 (73.6)	99 (67.8)		318 (72.6)	42 (67.7)	
ICSI[j]	1780 (28.4)	309 (31.8)		265 (25.9)	46 (31.5)		118 (26.9)	20 (32.3)	
IVF/ICSI	38 (0.6)	4 (0.4)		5 (0.5)	1 (0.7)		2 (0.5)	0 (0.0)	
Embryo transfer and embryology characteristics, median (IQR)									
Number of transferred embryos	2 (1-2)	2 (1-2)	.87	2 (1-2)	2 (1-2)	.91	2 (1-2)	2 (1-2)	.90
Number of high-quality embryos transferred	1 (1-2)	1 (1-1)	<.01	1 (1-2)	1 (1-1)	.20	1 (1-2)	1 (1-1)	.04
Embryo stage, n (%)			<.01			<.01			.01
D3[k]	4835 (77.1)	846 (86.9)		717 (70.0)	129 (88.4)		329 (75.1)	57 (91.9)	
D5[l]	1295 (20.7)	117 (12.0)		290 (28.3)	15 (10.3)		103 (23.5)	5 (8.1)	
Other or mixed	138 (2.2)	10 (1.0)		17 (1.7)	2 (1.4)		6 (1.4)	0 (0.0)	
Overall miscarriage rate, n (%)									
Miscarriage rate	—[m]	973 (13.4)	—	—	146 (12.5)	—	—	62 (12.4)	—
Early miscarriage	—	804 (82.6)	—	—	121 (82.9)	—	—	52 (83.9)	—
Late miscarriage	—	169 (17.4)	—	—	25 (17.1)	—	—	10 (16.1)	—

[a]Baseline between-group *P* values are for descriptive purposes only and not adjusted for multiplicity.

[b]FSH: follicle-stimulating hormone.

[c]LH: luteinizing hormone.

[d]E2: estradiol.

[e]PRL: prolactin.

[f]AMH: anti-Müllerian hormone.

[g]AFC: antral follicle count.

[h]hCG: human chorionic gonadotropin.

[i]IVF: in vitro fertilization.

[j]ICSI: intracytoplasmic sperm injection.

[k]D3: day 3.

[l]D5: day 5.

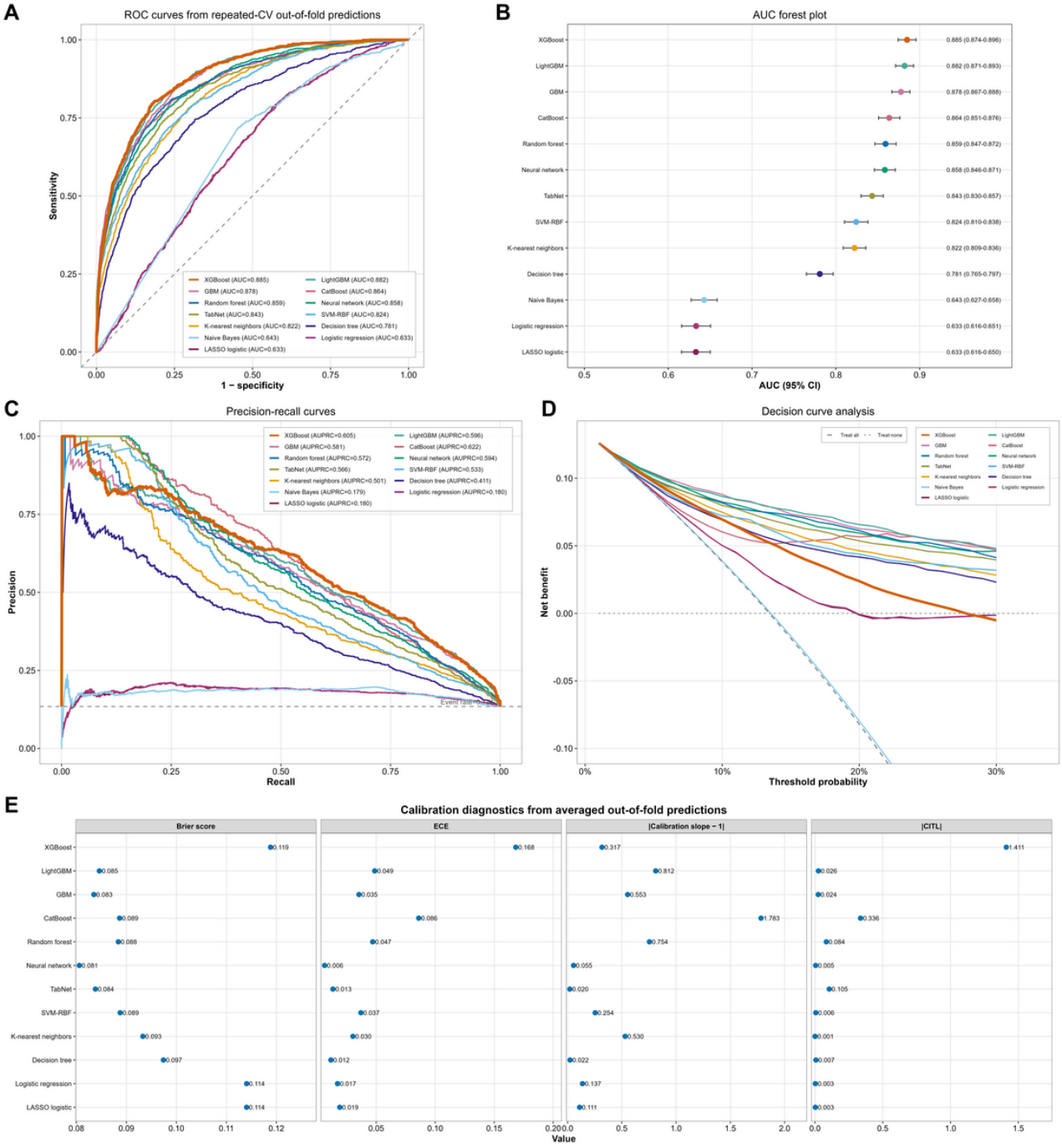
[m]Not applicable.

Internal Model Comparison

Among 13 candidate models evaluated using repeated stratified 5-fold cross-validation with 5 repeats in all 7241 development records, XGBoost had the highest AUC (0.885, 95% CI 0.874-0.896; area under the precision-recall curve [AUPRC] 0.605), followed by LightGBM (AUC 0.882, 95% CI 0.871-0.893; AUPRC 0.596) and gradient boosting machine (AUC 0.878, 95% CI 0.867-0.888; AUPRC 0.581; Figure 2; Table 2; Table S3 in Multimedia Appendix 1 [11,12,14]). The XGBoost-LightGBM AUC difference was small and not significant (Holm-adjusted paired DeLong *P*=.07; Tables S4 and S5 in Multimedia Appendix 1 [11,12,14]). Calibration performance differed between XGBoost and LightGBM across bootstrap comparisons. The mean net benefit from thresholds 0.05 to 0.30 was higher for LightGBM and gradient boosting machine than for XGBoost. Multicriteria performance, bootstrap stability, prediction agreement, and focused XGBoost-LightGBM contrasts are shown in Figures S1 to S4 in Multimedia Appendix 3. The leading algorithms should therefore be interpreted as closely related in terms of discrimination, without a single model dominating all performance domains.

XSL•FO
RenderX

Figure 2. Comparative internal performance of 13 candidate models based on repeated cross-validation (CV). (A) Receiver operating characteristic (ROC) curves from record-level averaged out-of-fold predictions. (B) Area under the ROC curve (AUC) estimates with 95% CIs. (C) Precision-recall curves with corresponding area under the precision-recall curve (AUPRC) values; the dashed horizontal line represents development cohort event prevalence. (D) Decision curve analysis for all 13 candidate models, with treat-all and treat-none reference strategies. (E) Calibration diagnostics for raw averaged out-of-fold probabilities used in the primary comparison. Lower values indicate better performance for Brier score, expected calibration error (ECE), absolute deviation of calibration slope from 1, and absolute calibration-in-the-large (CITL). Naive Bayes was omitted from panel E because its extreme calibration and probability accuracy values compressed the visual scale; complete results are reported in Table 2. GBM: gradient boosting machine; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; SVM-RBF, support vector machine with radial basis function; XGBoost: Extreme Gradient Boosting. See Multimedia Appendix 2 for a higher-resolution version of this figure.



XSL•FO
RenderX

Table 2. Internal model comparison performance of 13 candidate machine learning models for miscarriage prediction after confirmed clinical pregnancy in women with polycystic ovary syndrome[a].

Model	AUC[b] (95% CI)	AUPRC[c]	Brier score	ECE[d]	Calibration slope	CITL[e]	Descriptive Youden threshold	Sensitivity	Specificity	PPV[f]	NPV[g]	Net benefit[h] 0.05-0.30
XGBoost[i]	0.885 (0.874-0.896)	0.605	0.119	0.168	1.317	−1.411	0.463	0.801	0.817	0.404	0.963	0.038
LightGBM[j]	0.882 (0.871-0.893)	0.596	0.085	0.049	1.812	0.026	0.169	0.788	0.812	0.394	0.961	0.072
GBM[k]	0.878 (0.867-0.888)	0.581	0.083	0.035	1.553	0.024	0.139	0.831	0.759	0.349	0.967	0.071
CatBoost	0.864 (0.851-0.876)	0.622	0.089	0.086	2.783	−0.336	0.214	0.735	0.829	0.400	0.953	0.059
Random forest	0.859 (0.847-0.872)	0.572	0.088	0.047	1.754	0.084	0.151	0.789	0.782	0.359	0.960	0.067
Neural network	0.858 (0.846-0.871)	0.594	0.081	0.006	1.055	−0.005	0.146	0.775	0.779	0.353	0.957	0.067
TabNet	0.843 (0.830-0.857)	0.566	0.084	0.013	1.020	−0.105	0.175	0.740	0.787	0.351	0.951	0.063
SVM-RBF[l]	0.824 (0.810-0.838)	0.533	0.089	0.037	1.254	−0.006	0.124	0.719	0.769	0.326	0.946	0.054
K-nearest neighbors	0.822 (0.809-0.836)	0.501	0.093	0.030	1.530	−0.001	0.149	0.781	0.705	0.292	0.954	0.056
Decision tree	0.781 (0.765-0.797)	0.411	0.097	0.012	0.978	0.007	0.103	0.746	0.683	0.268	0.945	0.051
Naive Bayes	0.643 (0.627-0.658)	0.179	0.835	0.837	0.148	−15.209	1.000	0.717	0.547	0.197	0.926	−0.056
Logistic regression	0.633 (0.616-0.651)	0.180	0.114	0.017	0.863	−0.003	0.115	0.806	0.411	0.175	0.932	0.022
LASSO[m] logistic	0.633 (0.616-0.650)	0.180	0.114	0.019	0.889	−0.003	0.119	0.783	0.437	0.178	0.928	0.022

[a]Repeated stratified 5-fold cross-validation with 5 repeats was performed in the full development cohort (n=7241; miscarriages: n=973; nonevents: n=6268). Five outer-test predictions were averaged for each record. Models are ordered by AUC. The Youden threshold refers to raw averaged out-of-fold probabilities, which is a descriptive statistical operating point and is not a universal clinical cutoff.

[b]AUC: area under the receiver operating characteristic curve.

[c]AUPRC: area under the precision-recall curve.

[d]ECE: expected calibration error.

[e]CITL: calibration-in-the-large.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]Net benefit was calculated as the optimal Youden threshold for each model.

[i]XGBoost: Extreme Gradient Boosting.

[j]LightGBM: light gradient boosting machine.

[k]GBM: gradient boosting machine.

XSL•FO
RenderX

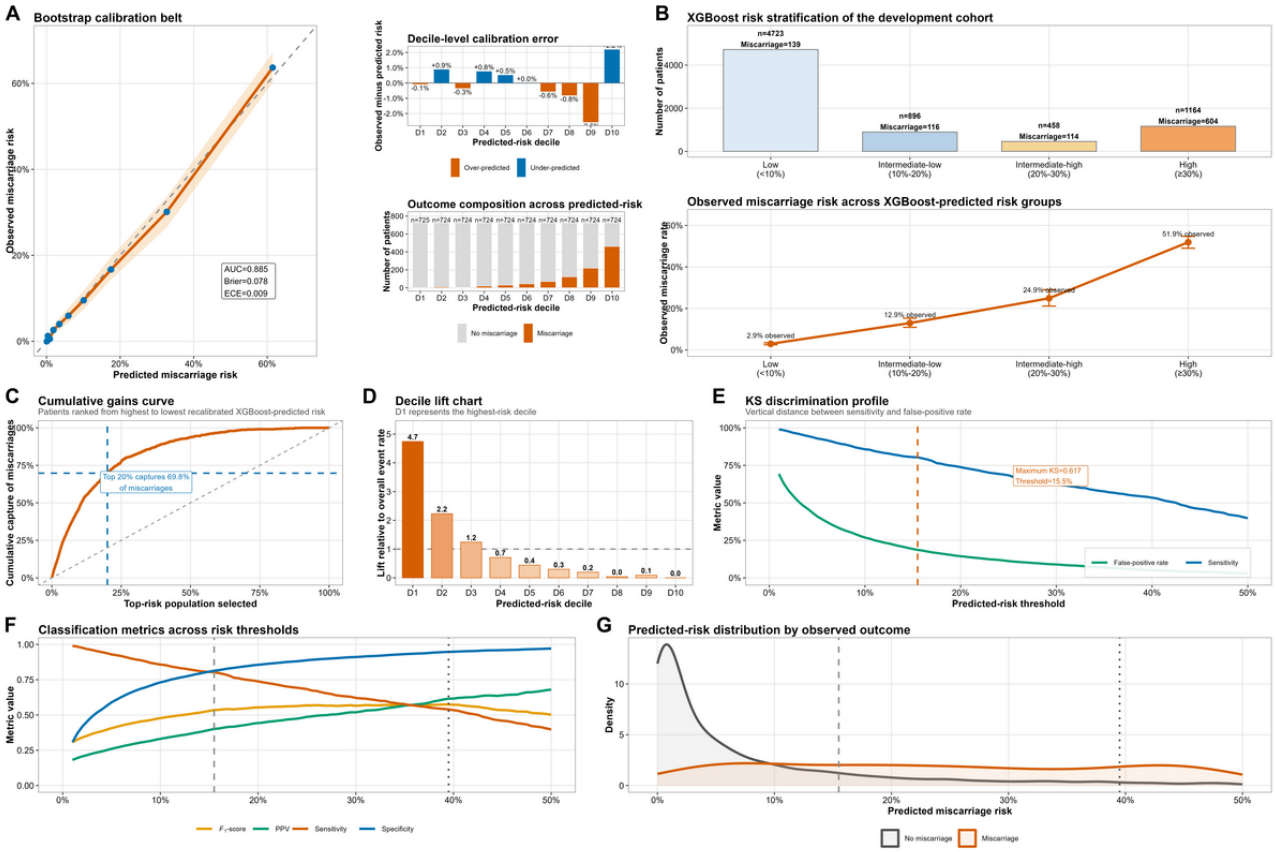
[l]SVM-RBF, support vector machine with radial basis function.

[m]LASSO: least absolute shrinkage and selection operator.

Secondary Development-Sample Recalibration and Utility Analysis

Raw averaged XGBoost out-of-fold probabilities overestimated absolute risk, with a mean predicted probability of 30.3% vs an observed event rate of 13.4%. Repeated 5-fold cross-fitted logistic recalibration preserved discrimination (AUC 0.885 before and after recalibration) while improving Brier score from 0.119 to 0.078 and ECE from 0.168 to 0.009 (Figure 3A); the calibration intercept changed from –1.475 to –0.001, calibration slope from 1.317 to 0.999, and CITL from –1.411 to approximately 0 (Table S6 in Multimedia Appendix 1 [11,12,14]). In descriptive development-sample recalibrated risk groups (<10%, 10% to <20%, 20% to <30%, and ≥30%), observed miscarriage rates were 2.0% (15/733), 13.1% (54/412), 30.9% (55/178), and 56.3% (71/126), respectively (Figure 3B). The top 20% (290/1449) of patients by recalibrated risk captured 69.8% of miscarriages, the highest-risk decile had a lift of approximately 4.0, and the maximum Kolmogorov-Smirnov statistic was 0.564 at a predicted-risk threshold of 13.0% (Figures 3C, 3D, and 3E). Threshold-dependent classification metrics and predicted-risk distributions are shown in Figures 3F and 3G; the development-specific Youden and F_1-optimal operating points (on cross-fitted recalibrated probabilities) were 15.5% and 39.5%, respectively, and were treated as descriptive rather than transferable clinical cutoffs. Figures S5 to S9 in Multimedia Appendix 3 provide additional recalibration, classification error, risk ranking, threshold utility, and subgroup diagnostics. Recalibrated development probabilities were not used to modify either validation cohort.

Figure 3. Secondary development-sample calibration and utility analysis of Extreme Gradient Boosting (XGBoost) using cross-fitted recalibrated out-of-fold probabilities. (A) Bootstrap calibration belt with decile-level calibration error and outcome composition. (B) Descriptive development-cohort risk groups (<10%, 10% to <20%, 20% to <30%, and ≥30%) with observed miscarriage rates. (C) Cumulative gains curve. (D) Decile lift chart; D1 denotes the highest-risk decile. (E) Kolmogorov-Smirnov (KS) discrimination profile. (F) Sensitivity, specificity, positive predictive value (PPV), and F1-score across predicted-risk thresholds, annotated with development-sample Youden (15.5%) and F1-optimal (39.5%) operating points; the 15.5% cutoff refers to cross-fitted recalibrated probabilities. (G) Predicted-risk distribution stratified by observed outcome. These development-sample thresholds and risk-group definitions are descriptive and were not applied to either the geographic external or prospective temporal validation cohorts. AUC: area under the receiver operating characteristic curve; ECE: expected calibration error. See Multimedia Appendix 2 for a higher-resolution version of this figure.



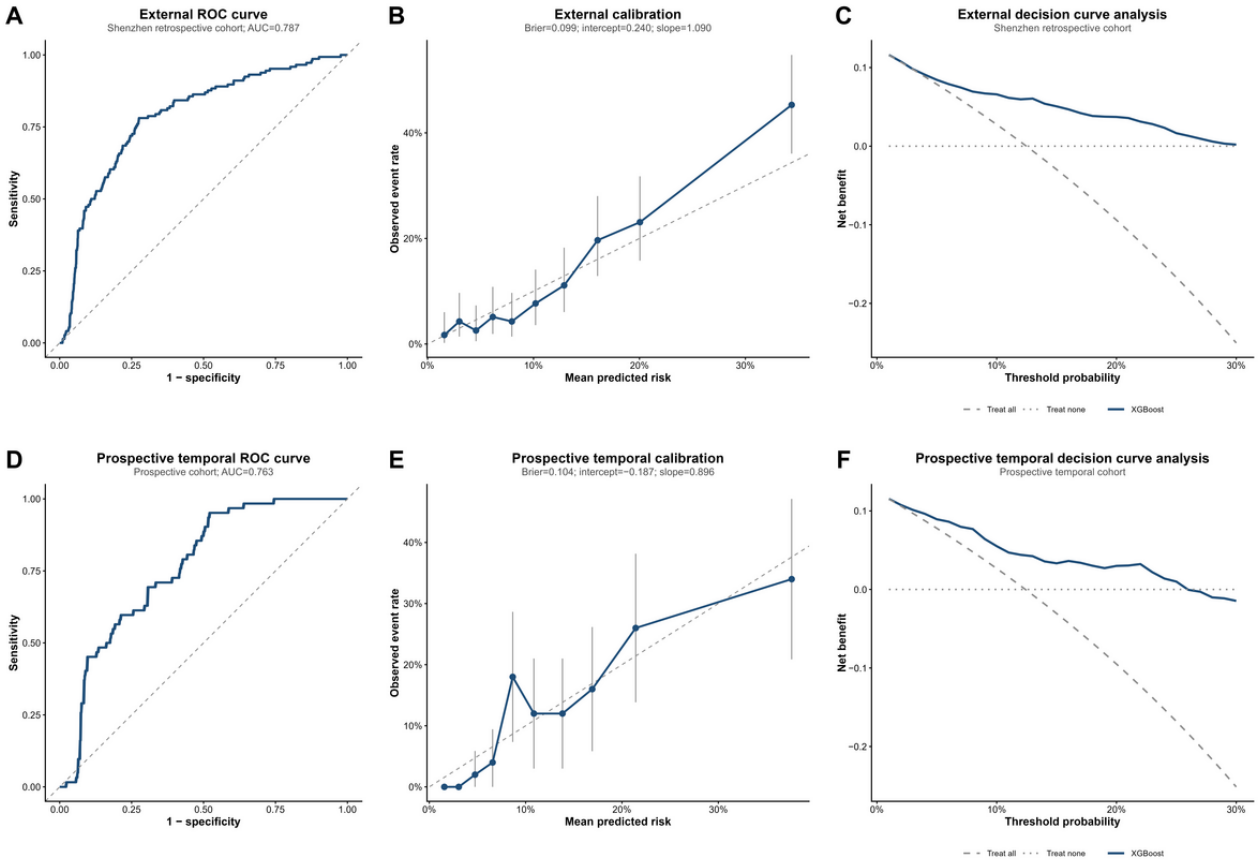
External Validation in the Shenzhen Cohort

In the geographic external validation cohort (n=1170; miscarriage events: n=146), the locked XGBoost model achieved an AUC of 0.787 (95% CI 0.748-0.827; Figure 4A). Calibration showed cohort-level drift (Brier score=0.099; calibration intercept=0.240; calibration slope=1.090; CITL=0.083; Figure 4B). DCA showed positive net benefit across thresholds 0.05 to 0.30 (Figure 4C). At the descriptive Youden threshold of 0.132, sensitivity was 0.781, specificity was 0.725, PPV was

XSL•FO
RenderX

0.288, and NPV was 0.959. The modest PPV indicates a substantial false-positive burden and supports use for risk ranking rather than diagnosis.

Figure 4. Retrospective geographic external and prospective temporal validation of the prelocked Extreme Gradient Boosting (XGBoost) miscarriage risk model. (A-C) Receiver operating characteristic (ROC) curve, calibration plot, and decision curve analysis in the Shenzhen retrospective geographic external validation cohort. (D-F) Corresponding analyses in the prospective temporal validation cohort. Validation predictions were generated using the locked model applied without modification. See Multimedia Appendix 2 for a higher-resolution version of this figure.



Prospective Temporal Validation Cohort

In the prospective temporal validation cohort (n=500; miscarriage events: n=62), the locked XGBoost model achieved an AUC of 0.763 (95% CI 0.709-0.816; Figure 4D). Calibration remained imperfect and differed in direction from the geographic external cohort (Brier score=0.104; calibration intercept=–0.187; calibration slope=0.896; CITL=–0.013; Figure 4E). DCA showed positive net benefit primarily across thresholds 0.05 to 0.25 (Figure 4F). At the descriptive Youden threshold of 0.082, sensitivity was 0.952, specificity was 0.479, PPV was 0.206, and NPV was 0.986. These findings support counseling and low-risk identification but not diagnostic classification or treatment selection.

Risk Ranking Across Cohorts and Exploratory Subgroup Analyses

Fixed risk groups were defined once from the prelocked internal reference predictions and applied without re-optimization to both validation cohorts (Table S7 in Multimedia Appendix 1 [11,12,14]). Observed miscarriage rates across the 4 predefined fixed risk groups increased from 0.6% (2/362) to 37.2% (135/363) in the prelocked internal reference set, from 2.9% (9/314) to 34.5% (77/223) in the geographic external validation cohort, and from 0.8% (1/125) to 29.4% (32/109) in the prospective temporal validation cohort. These fixed cut points were used solely to evaluate transportable risk ranking and were not optimized treatment thresholds. The development-specific recalibrated risk groups shown in Figure 3B constitute a separate descriptive analysis and must not be conflated with these fixed cut points applied for validation.

Exploratory subgroup discrimination across female age, BMI, AMH, basal LH, testosterone, EMT, and center ranged from 0.862 to 0.905, with generally small mean calibration errors (Figure S9 in Multimedia Appendix 3). These analyses lacked sufficient power to support subgroup-specific inferences and require independent external confirmation.

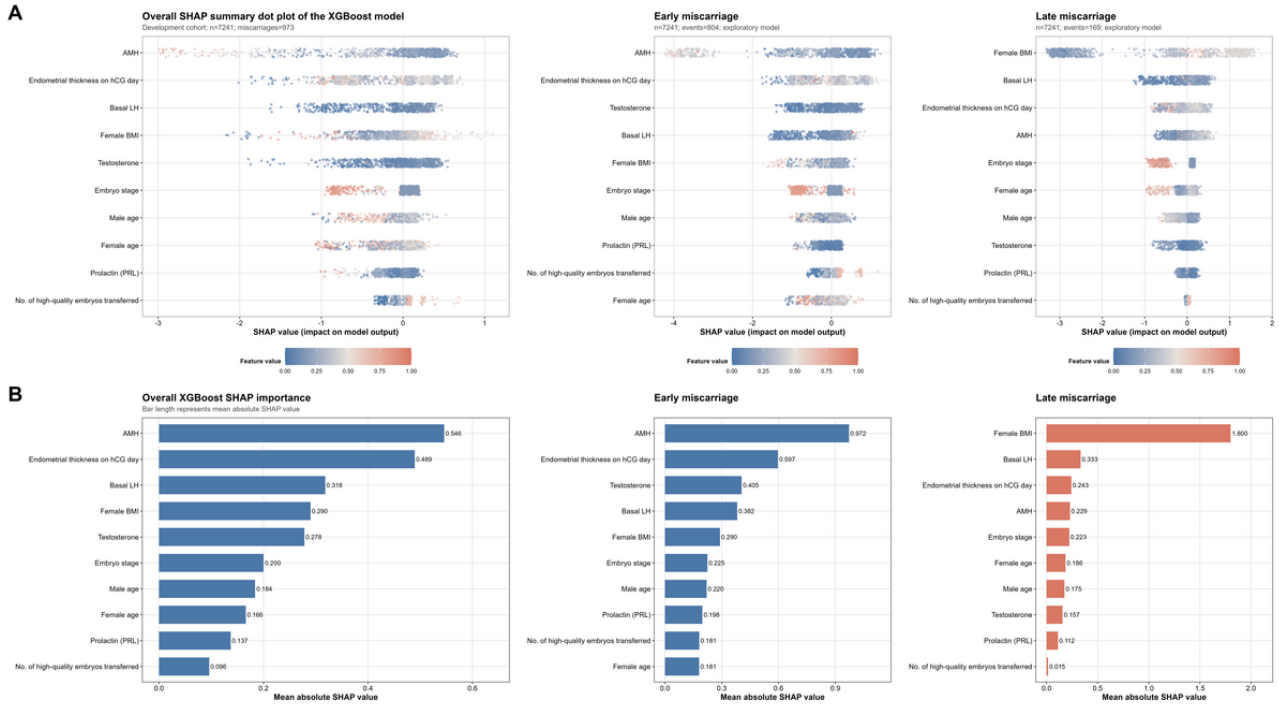
SHAP-Based Model Interpretation

SHAP analyses quantified predictive contributions of individual features rather than causal relationships. In the overall model, AMH, EMT on hCG day, and basal LH were the leading contributors, followed by female BMI, testosterone, embryo stage, male age, female age, PRL, and number of high-quality embryos (Figures 5A and 5B). AMH, EMT, and testosterone ranked highly in the exploratory early-miscarriage model, whereas female BMI and basal LH ranked highly in the exploratory late-miscarriage model (Figures 5A and 5B).

XSL•FO
RenderX

Subtype sample sizes, training strategy, calibration, and performance are reported in Table S8 in Multimedia Appendix 1 [11,12,14].

Figure 5. Shapley additive explanations (SHAP)–based interpretation of the overall locked Extreme Gradient Boosting (XGBoost) model and exploratory miscarriage-subtype models. (A) SHAP summary dot plots for the overall model, exploratory early-miscarriage model, and exploratory late-miscarriage model; point color represents feature value. (B) Corresponding bar charts of mean absolute SHAP values for each model. SHAP values quantify contributions to model-predicted log odds. SHAP describes fitted-model behavior and does not establish causality. Early- and late-miscarriage subtype analyses were exploratory and not predefined primary outcomes. AMH: anti-Müllerian hormone; bLH: basal luteinizing hormone; Emb. stage: embryo stage (cleavage or blastocyst); F age: female age; F-BMI: female BMI; hCG day EMT: endometrial thickness on the day of human chorionic gonadotropin administration; HQE T(n): number of high-quality embryos transferred; M age: male age; PRL: prolactin; T: testosterone. See Multimedia Appendix 2 for a higher-resolution version of this figure.



SHAP dependence plots showed nonlinear fitted-model associations for AMH, EMT, basal LH, and female BMI in the overall model (Figure S10 in Multimedia Appendix 3), with corresponding exploratory subtype patterns in Figure S11 in Multimedia Appendix 3. These plots should not be interpreted as causal relationships, biologically optimal ranges, or treatment targets.

Discussion

Principal Findings

In this multicenter study of 8911 women with PCOS with clinical pregnancy following fresh embryo transfer, we developed and validated a locked XGBoost miscarriage risk stratification model using routine clinical, endocrine, ultrasound, and embryo features. XGBoost, LightGBM, and GBM had comparable discrimination in repeated cross-validation, with divergent calibration and clinical utility. The locked XGBoost model retained moderate discrimination in external and temporal validation. Fixed risk groups showed stepwise increases in miscarriage rates across risk strata, supporting risk ranking rather than definitive diagnosis. Exploratory SHAP analyses identified key endocrine, endometrial, metabolic, and embryo-related predictors.

Comparison With Previous Studies

Previous studies have documented increased miscarriage and adverse pregnancy risks in women with PCOS after assisted reproduction [16,17], but population-level associations do not provide individualized postclinical pregnancy risk estimates. Existing miscarriage prediction studies differ substantially in population, prediction time, outcome definition, predictors, and validation design; a structured comparison with studies by Yuan et al [11], Liu et al [12], and Ma and Zhou [14] is provided in Table S9 in Multimedia Appendix 1 [11,12,14]. Our prelocked XGBoost model was fixed before validation, while repeated cross-validation comparisons showed that LightGBM and GBM achieved comparable discrimination, with LightGBM demonstrating better probability accuracy on several raw out-of-fold metrics. The drop from an internal repeated cross-validation AUC of 0.885 to a geographic external AUC of 0.787 and a prospective temporal AUC of 0.763, together with cohort-specific calibration drift, demonstrates that predictive performance, updating requirements, and applicability must be evaluated separately in each target clinical setting.

This model is not designed for miscarriage diagnosis or treatment selection decision-making. Its intended prediction time is after ultrasound-confirmed clinical pregnancy and before miscarriage outcome ascertainment, with intended end users being reproductive medicine clinicians. Its potential clinical use is restricted to structured risk communication and planning

XSL•FO
RenderX

proportionate follow-up within existing clinical care standards. Owing to modest PPV and a substantial false-positive burden, a high-risk model output should only prompt enhanced clinical review and patient counseling and must not trigger automatic therapeutic intervention. For patients at low predicted risk, the comparatively high NPV supports clinical reassurance and routine follow-up. DCA provides exploratory estimates of potential net benefit under hypothetical risk-threshold assumptions and cannot substitute for evidence from a prospective clinical impact trial.

SHAP analysis identified that AMH, EMT, basal LH, and female BMI were the most predictive features for miscarriage in this population with PCOS. AMH and EMT showed the strongest predictive impact on overall and early miscarriage, while testosterone played an important role in predicting early miscarriage, consistent with prior evidence linking abnormal ovarian reserve, endometrial receptivity, and hyperandrogenism to early miscarriage in women with PCOS [18,19]. For late miscarriage, female BMI and basal LH emerged as the top predictive variables, in line with clinical observations that metabolic dysregulation and hypothalamic-pituitary-ovarian axis disturbance contribute to poorer ongoing pregnancy maintenance [20-22]. Large-scale meta-analyses reported that the high BMI of patients with PCOS was associated with a significantly higher spontaneous abortion rate following IVF [23,24]. These findings underscore that miscarriage risk in PCOS is not a single, homogeneous outcome but rather a composite of distinct endocrine, metabolic, and endometrial pathways that operate at different gestational stages. These findings may help generate hypotheses regarding biological pathways and future prospective clinical studies. Importantly, SHAP values reflect predictive importance within the model and do not imply causal inference; our results support the clinical utility of routinely available variables for personalized risk assessment.

Strengths of this study include a large multicenter cohort design, rigorous separation of datasets for model development, geographic external validation, and prospective temporal validation, and exclusive use of preoutcome predictors to minimize data leakage. Model evaluation covered discrimination, precision-recall performance, calibration, clinical utility, risk stratification, and formal statistical model comparisons, with transparent documentation of variable processing and hyperparameter specifications.

Limitations

Several limitations warrant consideration. First, the development data were retrospective, and residual confounding and center-specific protocols may have influenced model behavior. Second, the unit of analysis was the IVF/ICSI clinical pregnancy cycle; some women contributed more than 1 cycle, creating possible nonindependence. Third, exclusion of biochemical pregnancies produced a selected postconception cohort, so the model does not estimate failure to establish clinical pregnancy or biochemical pregnancy loss. Fourth, prospective temporal validation included only 500 cycles and 62 miscarriages, limiting precision. Fifth, modest PPV implies a substantial false-positive burden. Sixth, important variables, including genetic data, lifestyle factors, medication exposure, previous miscarriage history, AFC, infertility duration, and stimulation characteristics, were not consistently harmonized for inclusion in the locked predictor set. Seventh, although the benchmark was expanded to LightGBM, CatBoost, and TabNet, it was not an exhaustive comparison of all modern tabular architectures, such as FT-Transformer, TabTransformer, Self-Attention and Intersample Attention Transformer, or TabPFN. Eighth, calibration differed across algorithms and cohorts; development-sample recalibration was secondary and was neither applied nor validated externally. Ninth, early- and late-miscarriage models and subgroup analyses were exploratory and were not independently validated. Tenth, sample size adequacy was assessed post hoc using the Riley et al [15] framework for binary prediction models, rather than being calculated a priori at study inception [25]. Risk-of-bias appraisal according to the Prediction Model Risk of Bias Assessment Tool for AI underscores a key limitation: models with good internal performance may show performance attenuation and calibration drift in new cohorts [26]. Despite adherence to TRIPOD+AI reporting standards [13], local monitoring and potential recalibration are mandatory before clinical deployment. Finally, no prospective clinical intervention or impact study was conducted; the model cannot replace clinical judgment, and local evaluation and, when appropriate, recalibration is required before implementation.

Conclusions

We developed and evaluated a prelocked XGBoost model for miscarriage risk ranking after confirmed clinical pregnancy in women with PCOS following fresh IVF/ICSI. The model retained moderate discrimination in geographic external and prospective temporal validation but showed performance attenuation and cohort-specific calibration drift. It may support counseling and proportionate follow-up only after local evaluation; it should not be used for diagnosis, treatment selection, or automated decision-making. Exploratory subtype and SHAP findings require independent validation.

Acknowledgments

The authors thank all participants, clinicians, and laboratory staff for their contributions to this study.

Funding

This work was supported by the National Natural Science Foundation of China (grant 82401879), Shenzhen Medical Research Fund (grant A2503091), Shenzhen Science and Technology Program (grants JCYJ20240813114903006 and ZDCY20250901112903004), Guangdong Basic and Applied Basic Research Foundation Project (grant 2025A1515140005), and Hubei Provincial Natural Science Foundation (grant 2025AFB660).

XSL•FO
RenderX

Data Availability

Deidentified patient-level data are not publicly posted because of institutional and privacy restrictions but may be available from the corresponding authors upon reasonable request and approval by the participating institutions. Analysis code and the locked model object for repeated cross-validation, model comparison, cross-fitted recalibration, and figure generation are available from the corresponding authors for reproducibility review.

Generative AI Use

During manuscript preparation, the authors used ChatGPT (GPT-5.6 Sol; OpenAI) to assist with language editing and grammar correction. All authors reviewed, revised, and take full responsibility for the final content.

Authors' Contributions

Conceptualization: MQ
Methodology: JT
Software: JT
Data curation: YL, YZ, LL, XS, FW, HG, SZ
Resources: YL, YZ, KZ, AT
Investigation: YL, YZ
Formal analysis: LL, XS, FW, HG, SZ
Validation: LL, XS, FW, HG, SZ, KZ, AT
Supervision: KZ, AT, MQ
Funding acquisition: MQ
Writing: JT, MQ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Tables showing reproducibility and predictor handling, internal model benchmarking and focused statistical comparison, secondary development-sample recalibration and risk stratification, exploratory miscarriage-subtype analyses, and published model comparison.
[DOCX File , 4091 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Full resolution versions of Figures 1–5.
[ZIP File (Zip Archive), 8786 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Internal model comparisons, bootstrap stability analyses, head-to-head evaluations, threshold utilities, subgroup performance, and Shapley additive explanations feature dependence profiles supporting the evaluation and interpretability of the Extreme Gradient Boosting miscarriage prediction model.
[DOCX File , 5603 KB-Multimedia Appendix 3]

References

1. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod. Dec 2016;31(12):2841-2855. [FREE Full text] [doi: 10.1093/humrep/dew218] [Medline: 27664216]
2. Neven AC, Forslund M, Ranasinha S, Sethi P, Dhungana RR, Mousa A, et al. Prevalence of polycystic ovary syndrome: a global and regional systematic review and meta-analysis. Hum Reprod Update. May 01, 2026;32(3):277-312. [FREE Full text] [doi: 10.1093/humupd/dmaf030] [Medline: 41528735]
3. Teede HJ, Khomami MB, Morman R, Laven JS, Joham AE, Costello MF, et al. Polyendocrine metabolic ovarian syndrome, the new name for polycystic ovary syndrome: a multistep global consensus process. Lancet. Jun 06, 2026;407(10545):2329-2339. [FREE Full text] [doi: 10.1016/S0140-6736(26)00717-8] [Medline: 42119588]
4. Joham AE, Norman RJ, Stener-Victorin E, Legro RS, Franks S, Moran LJ, et al. Polycystic ovary syndrome. Lancet Diabetes Endocrinol. Sep 2022;10(9):668-680. [doi: 10.1016/S2213-8587(22)00163-2] [Medline: 35934017]

XSL•FO
RenderX

5. Gao X, Zhao S, Du Y, Yang Z, Tian Y, Zhao J, China Women's Reproductive Metabolic Network, et al. Data-driven subtypes of polycystic ovary syndrome and their association with clinical outcomes. Nat Med. Dec 2025;31(12):4214-4224. [doi: 10.1038/s41591-025-03984-1] [Medline: 41162652]
6. Matorras R, Pijoan JI, Laínz L, Díaz-Nuñez M, Sainz H, Pérez-Fernandez S, et al. Polycystic ovarian syndrome and miscarriage in IVF: systematic revision of the literature and meta-analysis. Arch Gynecol Obstet. Aug 2023;308(2):363-377. [doi: 10.1007/s00404-022-06757-0] [Medline: 36058943]
7. Bahri Khomami M, Hashemi S, Shorakae S, Harrison CL, Piltonen TT, Romualdi D, et al. Systematic review and meta-analysis of birth outcomes in women with polycystic ovary syndrome. Nat Commun. Jul 04, 2024;15(1):5592. [FREE Full text] [doi: 10.1038/s41467-024-49752-6] [Medline: 38965241]
8. Yang T, Zhao J, Liu F, Li Y. Lipid metabolism and endometrial receptivity. Hum Reprod Update. Nov 02, 2022;28(6):858-889. [FREE Full text] [doi: 10.1093/humupd/dmac026] [Medline: 35639910]
9. Zhang F, Wu J, Ge L, Huang Y, Huang Y. A review of immunometabolic crosstalk in PCOS-related pregnancy loss: mechanisms and emerging therapeutic strategies. Gynecol Endocrinol. Dec 31, 2025;41(1):2600160. [FREE Full text] [doi: 10.1080/09513590.2025.2600160] [Medline: 41384809]
10. Matsuyama S, Whiteside S, Li SY. Implantation and decidualization in PCOS: unraveling the complexities of pregnancy. Int J Mol Sci. Jan 18, 2024;25(2):1203. [FREE Full text] [doi: 10.3390/ijms25021203] [Medline: 38256276]
11. Yuan G, Lv B, Du X, Zhang H, Zhao M, Liu Y, et al. Prediction model for missed abortion of patients treated with IVF-ET based on XGBoost: a retrospective study. PeerJ. Jan 30, 2023;11:e14762. [FREE Full text] [doi: 10.7717/peerj.14762] [Medline: 36743954]
12. Liu L, Liu B, Wu H, Gan Q, Huang Q, Li M. Optimizing predictive features using machine learning for early miscarriage risk following single vitrified-warmed blastocyst transfer. Front Endocrinol (Lausanne). Apr 16, 2025;16:1557667. [FREE Full text] [doi: 10.3389/fendo.2025.1557667] [Medline: 40309447]
13. Collins GS, Moons K, Dhiman P, Riley R, Beam A, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
14. Ma L, Zhou Q. Predictors of miscarriage in polycystic ovary syndrome patients with threatened abortion: development and validation of a nomogram model. Front Endocrinol (Lausanne). Jan 09, 2026;16:1689878. [FREE Full text] [doi: 10.3389/fendo.2025.1689878] [Medline: 41585786]
15. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FEJ, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. Stat Med. Mar 30, 2019;38(7):1276-1296. [FREE Full text] [doi: 10.1002/sim.7992] [Medline: 30357870]
16. Xie N, Zhao W. Adverse pregnancy and perinatal outcomes in women with polycystic ovary syndrome undergoing assisted reproductive technology: a systematic review and meta-analysis. Front Med (Lausanne). Oct 10, 2025;12:1656389. [FREE Full text] [doi: 10.3389/fmed.2025.1656389] [Medline: 41140676]
17. Ban M, Sun Y, Chen X, Zhou X, Zhang Y, Cui L. Association between maternal polycystic ovarian syndrome undergoing assisted reproductive technology and pregnancy complications and neonatal outcomes: a systematic review and meta-analysis. J Ovarian Res. Jan 06, 2024;17(1):6. [FREE Full text] [doi: 10.1186/s13048-023-01331-x] [Medline: 38184624]
18. Liu X, Han Y, Wang X, Zhang Y, Du A, Yao R, et al. Serum anti-Müllerian hormone levels are associated with early miscarriage in the IVF/ICSI fresh cycle. BMC Pregnancy Childbirth. Apr 02, 2022;22(1):279. [FREE Full text] [doi: 10.1186/s12884-022-04591-5] [Medline: 35366830]
19. Zhou H, Yang W, Liu N, Liu Q, Jin J, Zhao C, et al. Hyperandrogenemia induces trophoblast ferroptosis and early pregnancy loss in patients with PCOS via CMA-dependent FTH1 degradation. Adv Sci (Weinh). Feb 2026;13(12):e06091. [doi: 10.1002/advs.202506091] [Medline: 41400098]
20. Cai H, Mol BW, Gordts S, Wang H, Wang T, Li N, et al. Early and late pregnancy loss in women with polycystic ovary syndrome undergoing IVF/ICSI treatment: a retrospective cohort analysis of 21 820 pregnancies. BJOG. Jun 2021;128(7):1160-1169. [doi: 10.1111/1471-0528.16590] [Medline: 33142019]
21. Wang L, Yu X, Xiong D, Leng M, Liang M, Li R, et al. Hormonal and metabolic influences on outcomes in PCOS undergoing assisted reproduction: the role of BMI in fresh embryo transfers. BMC Pregnancy Childbirth. Mar 28, 2025;25(1):368. [FREE Full text] [doi: 10.1186/s12884-025-07480-9] [Medline: 40155948]
22. Alenezi SA, Khan R, Amer S. The impact of high BMI on pregnancy outcomes and complications in women with PCOS undergoing IVF-a systematic review and meta-analysis. J Clin Med. Mar 10, 2024;13(6):1578. [FREE Full text] [doi: 10.3390/jcm13061578] [Medline: 38541804]
23. Zhang L, Feng Y, Sun X, Yi S, Xiao X, Ma F. Impact of body mass index on assisted reproductive technology outcomes in patients with polycystic ovary syndrome: a meta-analysis. Reprod Biomed Online. Jun 2024;48(6):103849. [doi: 10.1016/j.rbmo.2024.103849] [Medline: 38574459]
24. Sun YF, Zhang J, Xu YM, Cao ZY, Wang YZ, Hao GM, et al. High BMI and insulin resistance are risk factors for spontaneous abortion in patients with polycystic ovary syndrome undergoing assisted reproductive treatment: a systematic review and meta-analysis. Front Endocrinol (Lausanne). Dec 2020;11:592495. [FREE Full text] [doi: 10.3389/fendo.2020.592495] [Medline: 33343510]

XSL•FO
RenderX

25. Riley RD, Ensor J, Snell KI, Harrell FEJ, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. Mar 18, 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
26. Moons KG, Damen J, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]

Abbreviations

AFC: antral follicle count
AMH: anti-Müllerian hormone
AUC: area under the receiver operating characteristic curve
AUPRC: area under the precision-recall curve
CITL: calibration-in-the-large
DCA: decision curve analysis
ECE: expected calibration error
EMT: endometrial thickness
FSH: follicle-stimulating hormone
GBM: gradient boosting machine
hCG: human chorionic gonadotropin
ICSI: intracytoplasmic sperm injection
IVF: in vitro fertilization
LASSO: least absolute shrinkage and selection operator
LH: luteinizing hormone
LightGBM: light gradient boosting machine
NPV: negative predictive value
PCOS: polycystic ovary syndrome
PPV: positive predictive value
PRL: prolactin
SHAP: Shapley additive explanations
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for AI
XGBoost: Extreme Gradient Boosting

Edited by I Steenstra; submitted 20.May.2026; peer-reviewed by T-H Lee, T Panboonyuen; comments to author 28.Jul.2026; revised version received 11.Sep.2026; accepted 13.Oct.2026; published 05.Nov.2026

Please cite as:
Tong J, Zhang Y, Liu Y, Liu L, Shang X, Wu F, Guo H, Zhang S, Zhao K, Tang A, Qu M
Miscarriage Risk Prediction for Women With Polycystic Ovary Syndrome Who Achieve Clinical Pregnancy After Fresh Embryo Transfer: Multicenter Development, Geographic External Validation, and Prospective Temporal Validation Study
J Med Internet Res 2026;28:e101829
URL: https://www.jmir.org/2026/1/e101829
doi: 10.2196/101829
PMID:

©Juncheng Tong, Yingying Zhang, Yanhui Liu, Lifang Liu, Xiaoteng Shang, Fanggui Wu, Haoxuan Guo, Sibo Zhang, Kai Zhao, Aifa Tang, Mengyuan Qu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 05.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX