

Original Research

Machine Learning Prediction of Persistent Postpartum Hypertension Following Hypertensive Disorders of Pregnancy: A Secondary Analysis of a Prospective Cohort

Ky Dinh Nguyen¹, Thi Huyen Anh Nguyen^{2,3}, Quynh Anh Nguyen¹, The Lam Pham¹,
Thang Nguyen Manh^{2,3,*}

¹General Medicine, Hanoi Medical University, 100000 Hanoi, Vietnam

²Department of Obstetrics and Gynecology, Hanoi Medical University, 100000 Hanoi, Vietnam

³Department of Required Services, National Hospital of Obstetrics and Gynecology, 100000 Hanoi, Vietnam

*Correspondence: manhthang@hmu.edu.vn (Thang Nguyen Manh)

Academic Editor: Michael H. Dahan

Submitted: 17 April 2026 Revised: 16 June 2026 Accepted: 30 June 2026 Published: 13 October 2026

Abstract

Background: Persistent postpartum hypertension (PPPH) is a clinically significant complication of hypertensive disorders of pregnancy (HDP) that is frequently underrecognized due to high rates of loss to follow-up after hospital discharge. Early pre-discharge risk stratification represents an important unmet clinical need. This study aimed to develop and internally validate machine learning (ML) models using exclusively pre-discharge data to predict the risk of PPPH in women with HDP before hospital discharge. **Methods:** A secondary analysis of a prospective cohort from the National Hospital of Obstetrics and Gynecology in Hanoi, Vietnam, was conducted. The primary outcome was PPPH, defined as a systolic blood pressure (SBP) ≥ 140 mmHg and/or a diastolic blood pressure (DBP) ≥ 90 mmHg, or ongoing antihypertensive therapy at 6 weeks postpartum. Feature extraction was strictly limited to the period from hospital admission through 72 hours postpartum. Three ML algorithms were developed: Ridge logistic regression (LR), Random Forest, and extreme Gradient Boosting (XGBoost). Internal validation was performed using a 500-repeat bootstrap framework with Synthetic Minority Over-sampling Technique (SMOTE) applied within each resampling loop. Model calibration was assessed using the Brier score and calibration plots, while clinical utility was evaluated via decision curve analysis (DCA). Model interpretability was achieved using Shapley Additive Explanations (SHAP). **Results:** PPPH was observed in 39 of 301 women (12.96%). The Random Forest model, incorporating 25 pre-discharge clinical and biochemical features, was selected as the final model based on its superior calibration (Brier score: 0.130; calibration intercept: -0.794 ; calibration slope: 0.759; observed-to-expected [O/E] ratio: 0.544) and area under the receiver operating characteristic curve (AUROC) of 0.638 (95% confidence interval [CI]: 0.632–0.643). To prioritize the identification of women at risk of PPPH, the decision threshold was set at 0.157, achieving a sensitivity of 80.2% and a negative predictive value (NPV) of 92.5%. DCA demonstrated that the Ridge LR model provided the greatest net clinical benefit across the 15%–25% threshold probability range. SHAP analysis identified alanine aminotransferase (ALT), blood urea nitrogen (BUN), and maternal age as the top influential predictors. **Conclusions:** An ML-based pre-discharge risk-stratification tool for PPPH was developed and internally validated. Although the current model demonstrated modest discrimination and requires external validation in larger multicenter cohorts, the Random Forest model produced well-calibrated risk estimates and a high NPV (92.5%), offering a proof-of-concept for individualized pre-discharge risk stratification and the targeted allocation of postpartum monitoring resources.

Keywords: machine learning; postpartum hypertension; pre-discharge triage; predictive model; hypertensive disorders of pregnancy

1. Introduction

Hypertensive disorders of pregnancy (HDP) complicate approximately 5%–10% of pregnancies worldwide and remain a leading cause of maternal morbidity and mortality [1,2]. The pathophysiological alterations underlying HDP frequently persist beyond delivery, elevating the risk of acute maternal complications and long-term cardiovascular diseases [3,4]. Recent evidence suggests that the onset of hypertension in women with a history of HDP may occur much earlier than previously recognized. In a recent meta-analysis, Giorgione et al. [5] reported that postpartum hypertension was diagnosed within the first 2 years fol-

lowing pregnancy in 468 of 1646 women (28.4%). Moreover, the pooled analysis demonstrated a substantially increased risk of developing postpartum hypertension after HDP compared with normotensive pregnancies (odds ratio [OR]: 5.75, 95% confidence interval [CI]: 3.92–8.44; $I^2 = 49\%$) [5].

Current clinical guidelines recommend routine blood pressure reassessment at 6 weeks postpartum [6]. However, this strategy is severely limited by high rates of loss to follow-up, delaying the identification of women at high risk. This critical clinical gap underscores the need for robust pre-discharge risk-stratification tools [7].



Machine learning (ML) algorithms have increasingly demonstrated superior predictive performance over conventional statistical models by capturing complex, high-order nonlinear interactions in clinical data [8]. Despite these advances, existing predictive models for postpartum hypertension remain methodologically limited. The widespread incorporation of post-discharge variables in existing prediction models introduces target leakage, thereby artificially inflating estimates of predictive performance [9]. Furthermore, an overreliance on discriminatory metrics, coupled with the inherent “black-box” nature of advanced algorithms, compromises model calibration and clinical interpretability, posing a significant barrier to real-world translation.

Therefore, this study aimed to develop and internally validate ML models based exclusively on pre-discharge data for the early prediction and risk stratification of persistent postpartum hypertension (PPPH). In accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Artificial Intelligence (TRIPOD-AI) guidelines, the clinical utility of the models was rigorously evaluated using decision curve analysis (DCA). Furthermore, model interpretability was achieved via Shapley Additive Explanations (SHAP) to enhance clinical applicability [10,11]. To the best of our knowledge, this is the first study to develop a ML-based pre-discharge triage tool for PPPH using exclusively pre-discharge clinical data in a Vietnamese cohort, addressing a critical gap in the postpartum surveillance literature for low- and middle-income settings.

2. Materials and Methods

2.1 Study Design and Setting

This study was conducted as a secondary analysis of a previously published prospective cohort [12]. The primary cohort enrolled 336 women with HDP who delivered at the National Hospital of Obstetrics and Gynecology, Hanoi, Vietnam, between March and November 2023. Detailed descriptions of the study design, data collection procedures, and primary results have been reported elsewhere [12]. All study procedures were conducted in accordance with the Declaration of Helsinki. This study was reported in accordance with the TRIPOD-AI and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.

2.2 Study Population

Of the 336 women enrolled in the primary prospective cohort, the present analysis was restricted to 301 women with new-onset HDP (gestational hypertension, $n = 67$; preeclampsia-eclampsia, $n = 234$). Women with preexisting chronic hypertension ($n = 16$) or superimposed preeclampsia on chronic hypertension ($n = 19$) were excluded, as their postpartum blood pressure trajectories reflect persistent prepregnancy pathophysiology rather than pregnancy-

induced cardiovascular dysregulation (Fig. 1, Ref. [12]). Eligibility criteria for the primary cohort included women aged ≥ 18 years with a confirmed pregnancy at ≥ 20 weeks of gestation who were diagnosed with HDP according to American College of Obstetricians and Gynecologists criteria [6].

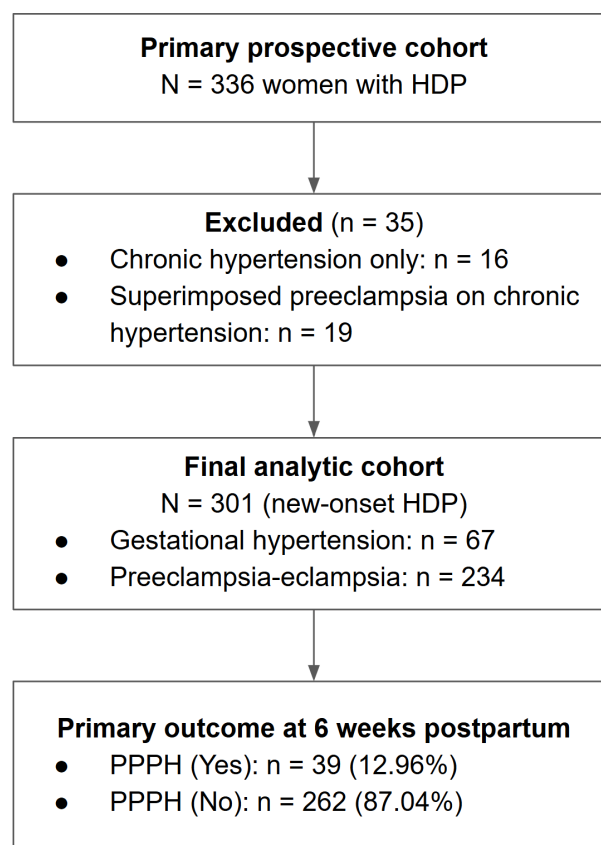


Fig. 1. Participant flowchart. Data were derived from the primary prospective cohort [12]. HDP, hypertensive disorders of pregnancy; PPPH, persistent postpartum hypertension.

2.3 Study Outcome

The primary outcome was PPPH, defined as a systolic blood pressure (SBP) ≥ 140 mmHg and/or diastolic blood pressure (DBP) ≥ 90 mmHg, or the ongoing use of antihypertensive therapy, at 6 weeks postpartum. This time point aligns with current postpartum follow-up guidelines [6] and captures a critical clinical window for therapeutic intervention. In the primary cohort, the prevalence of PPPH was 12.96% at 6 weeks postpartum and 7.6% at 3 months postpartum [12].

Prior to hospital discharge, all participants received a validated home blood pressure monitor (Omron M2 Basic HEM7120-E; Omron Healthcare, Kyoto, Japan). Standardized training on home blood pressure measurement techniques was provided to each participant and a designated

family member by trained research staff in accordance with the Ministry of Health of Vietnam and American Heart Association (AHA) guidelines [12]. Blood pressure assessment at 6 weeks postpartum was conducted by trained research staff via standardized telephone interview, during which participants were instructed to report two consecutive morning blood pressure measurements obtained after 5 minutes of seated rest. Women who reported an SBP ≥ 140 mmHg or DBP ≥ 90 mmHg, or who reported ongoing use of antihypertensive medication, were classified as having PPPH.

2.4 Feature Selection and Timeframe

The feature extraction window was strictly restricted to the period from hospital admission through 72 hours postpartum to prevent target leakage and ensure the model's applicability as a pre-discharge risk stratification tool. Candidate predictors included maternal demographic and obstetric characteristics, admission clinical parameters, biochemical markers of end-organ involvement, and early postpartum blood pressure trajectories on postpartum days 1 and 3.

2.5 Data Preprocessing and ML Model Development

Data preprocessing and all ML pipelines were executed using the tidymodels framework in R (version 4.5.0, R Foundation for Statistical Computing, Vienna, Austria). The normality of continuous variables was assessed using the Shapiro-Wilk test, complemented by visual inspection of normal Q-Q plots and histograms. Given evidence of departure from normality (Shapiro-Wilk test, $p < 0.05$), continuous variables were summarized as medians with interquartile ranges (Q1–Q3) and compared between groups using the Mann-Whitney U (Wilcoxon rank-sum) test. Categorical variables were summarized as frequencies and percentages [n (%)]. Pearson's chi-square test was used when its assumptions were satisfied, whereas Yates' continuity-corrected chi-square test was applied to 2×2 tables when the minimum expected cell frequency was between 1 and <5 . A two-sided $p < 0.05$ was considered statistically significant. Among the 25 candidate predictors, 24 had complete data, while missing values occurred exclusively for lactate dehydrogenase (LDH; 7.3%, 22/301 cases). As no predictor exceeded the predefined 20% missingness threshold, all 25 predictors were retained for model development. Continuous variables were standardized by Z-score normalization, and categorical variables were converted to dummy variables. Predictors exhibiting zero variance were removed before model development. All data-dependent preprocessing procedures were incorporated into the tidymodels recipe and estimated exclusively within the training portion of each bootstrap resampling iteration, with the resulting preprocessing parameters subsequently applied to the corresponding assessment data to prevent information leakage.

Due to the relatively small sample size ($n = 301$) and class imbalance (outcome prevalence, 12.96%), a 500-repeat standard bootstrap resampling framework with replacement was employed for internal validation. Optimism was estimated as the mean difference between the apparent and bootstrap-validated performance metrics across all 500 resamples. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) [13]. Three supervised ML algorithms were developed: Ridge logistic regression (LR) (incorporating L2 regularization and splines), Random Forest, and Extreme Gradient Boosting (XGBoost) [11]. Importantly, both KNN imputation and SMOTE were implemented exclusively within each bootstrap resampling iteration to prevent data leakage. To assess the incremental benefit of oversampling, each algorithm was additionally trained without SMOTE using identical preprocessing and validation procedures. These results are reported in **Supplementary 3** for reference. Hyperparameters were optimized via grid search within each bootstrap resampling loops, with the area under the precision-recall curve (AUPRC) used as the optimization metric. The candidate search ranges and final optimal hyperparameter values for all machine learning models following grid search optimization within the 500-repeat bootstrap resampling framework are detailed in **Supplementary 4**. Although the nominal events-per-variable (EPV) ratio was 1.56 (39 events for 25 candidate predictors), we deliberately avoided univariable screening prior to modeling to preserve complex pathophysiological interactions and prevent selection bias. Instead, the risk of overfitting was strictly mitigated through algorithm-specific embedded regularization and hyperparameter constraints. Specifically, Ridge LR utilized an L2 penalty, Random Forest applied strict node size and feature split limitations, and XGBoost was restricted to shallow decision stumps with a conservative learning rate.

2.6 Model Evaluation, Threshold Optimization, and Interpretability

Discriminatory performance was assessed using the area under the receiver operating characteristic curve (AUROC) and the AUPRC, with both metrics reported as 95% CIs derived from the bootstrap distribution. Model calibration was evaluated using the Brier score, calibration intercept (the log-odds intercept from a logistic recalibration model fitted to the bootstrap-validated predictions), calibration slope (the corresponding regression coefficient), and the observed-to-expected (O/E) event ratio. Calibration was further assessed visually using calibration plots with Locally Weighted Scatterplot Smoothing (LOESS) smoothing and 95% CIs. Because SMOTE overrepresents the minority class during training, it can inflate predicted probabilities. Therefore, out-of-bag predictions were recalibrated using Platt scaling prior to calibration and clinical utility assessments. The Youden index, which assumes equal clinical costs for false-negative and false-positive classification

Table 1. Baseline characteristics and clinical trajectories of the study population.

Patient Characteristics	PPPH at 6 Weeks			Test Statistic	p-value
	Overall (n = 301)	Persistent HTN at 6W (n = 39)	Resolved HTN at 6W (n = 262)		
Maternal Age (years), Median (Q1–Q3)	31.0 (27.0–35.0)	33.0 (28.0–38.0)	31.0 (27.0–35.0)	W = 6003	0.078
Prepregnancy body mass index (BMI) (kg/m ²), Median (Q1–Q3)	21.6 (19.9–23.9)	22.5 (19.7–25.5)	21.6 (19.9–23.8)	W = 5489	0.455
Gestational Weight Gain (kg), Median (Q1–Q3)	13.0 (10.0–17.0)	12.0 (8.0–15.0)	14.0 (10.0–17.0)	W = 4146	0.057
Multiple Pregnancy, n (%)	70 (23.3%)	5 (12.8%)	65 (24.8%)	$\chi^2 = 2.73$	0.098
Gestational Diabetes Mellitus, n (%)	53 (17.6%)	5 (12.8%)	48 (18.3%)	$\chi^2 = 0.71$	0.400
History of Preeclampsia, n (%)	27 (9.0%)	6 (15.4%)	21 (8.0%)	$\chi^2 = 1.45$	0.229
History of Gestational Hypertension, n (%)	31 (10.3%)	6 (15.4%)	25 (9.5%)	$\chi^2 = 0.70$	0.402
Gestational Age at Admission (weeks), Median (Q1–Q3)	36.1 (33.1–37.4)	33.3 (31.4–36.4)	36.2 (33.3–37.6)	W = 3478	0.001
Admission SBP (mmHg), Median (Q1–Q3)	150.0 (150.0–160.0)	160.0 (150.0–170.0)	150.0 (150.0–160.0)	W = 6044	0.059
Admission DBP (mmHg), Median (Q1–Q3)	100.0 (90.0–100.0)	100.0 (100.0–110.0)	100.0 (90.0–100.0)	W = 6354	0.007
HDP Classification, n (%)				$\chi^2 = 0.48$	0.488
Gestational Hypertension	67 (22.3%)	7 (17.9%)	60 (22.9%)		
Preeclampsia-Eclampsia	234 (77.7%)	32 (82.1%)	202 (77.1%)		
Severe Clinical Symptoms, n (%)	42 (14.0%)	11 (28.2%)	31 (11.8%)	$\chi^2 = 7.58$	0.006
Proteinuria (g/24 h), Median (Q1–Q3)	1.3 (0.3–4.5)	4.0 (0.7–7.4)	1.3 (0.3–4.2)	W = 6153	0.039
Creatinine (μmol/L), Median (Q1–Q3)	67.5 (59.1–76.4)	70.2 (63.1–83.9)	67.5 (58.9–76.3)	W = 5741	0.213
Blood Urea Nitrogen (BUN) (mmol/L), Median (Q1–Q3)	4.8 (3.7–6.1)	5.5 (4.1–7.6)	4.8 (3.7–6.0)	W = 6424	0.010
Aspartate Aminotransferase (AST) (U/L), Median (Q1–Q3)	23.2 (19.3–31.0)	24.9 (20.4–34.1)	23.0 (19.3–30.8)	W = 5650	0.286
Alanine Aminotransferase (ALT) (U/L), Median (Q1–Q3)	14.7 (10.4–21.8)	17.8 (14.5–24.7)	14.0 (9.8–21.2)	W = 6747	0.001
Lactate Dehydrogenase (LDH) (U/L), Median (Q1–Q3)	205.6 (176.9–256.4)	217.9 (174.2–280.5)	204.4 (178.2–251.0)	W = 4815	0.460
Missing data for LDH	22	2	20		
Platelet Count ($\times 10^9/L$), Median (Q1–Q3)	197.0 (156.0–232.0)	185.0 (156.0–233.0)	198.0 (155.0–232.0)	W = 4981	0.801
Postpartum Day 1 SBP (mmHg), Median (Q1–Q3)	140.0 (130.0–150.0)	140.0 (130.0–153.0)	138.5 (130.0–147.0)	W = 6222	0.028
Postpartum Day 1 DBP (mmHg), Median (Q1–Q3)	90.0 (80.0–94.0)	90.0 (80.0–100.0)	90.0 (80.0–91.0)	W = 6052	0.060
Postpartum Day 3 SBP (mmHg), Median (Q1–Q3)	126.0 (120.0–136.0)	130.0 (120.0–147.0)	124.0 (120.0–134.0)	W = 6689	0.002
Postpartum Day 3 DBP (mmHg), Median (Q1–Q3)	80.0 (78.0–90.0)	85.0 (80.0–94.0)	80.0 (74.0–89.0)	W = 6387	0.010
Hypertension on Postpartum Day 3, n (%)	86 (28.6%)	19 (48.7%)	67 (25.6%)	$\chi^2 = 8.91$	0.003
Mode of Delivery (Cesarean vs. Vaginal), n (%)				$\chi^2 = 0.00$	1.000
Vaginal Delivery	21 (7.0%)	3 (7.7%)	18 (6.9%)		
Cesarean Section	280 (93.0%)	36 (92.3%)	244 (93.1%)		

Data are presented as median (Q1–Q3) or n (%). Continuous variables were compared using the Wilcoxon rank-sum test. Categorical variables were compared using Pearson's χ^2 test (minimum expected frequency ≥ 5) or Yates-corrected χ^2 test (1 to < 5). W and χ^2 denote the corresponding test statistics. Bold values indicate $p < 0.05$. LDH had 22 missing observations (2 PPPH; 20 resolved); all other variables were complete. Abbreviations: PPPH, persistent postpartum hypertension; HTN, hypertension; BMI, body mass index; HDP, hypertensive disorders of pregnancy; SBP, systolic blood pressure; DBP, diastolic blood pressure; BUN, blood urea nitrogen; AST, aspartate aminotransferase; ALT, alanine aminotransferase; LDH, lactate dehydrogenase.

ns, was considered inappropriate for this high-sensitivity risk stratification application, in which missed PPPH cases carry substantially higher clinical consequences than false-positive predictions. Accordingly, the decision threshold for the champion model was set to achieve a prespecified sensitivity of $\geq 80\%$. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were calculated at the selected threshold. A confusion matrix was constructed to illustrate the clinical trade-offs associated with the selected operating point. DCA was performed to quantify the net clinical benefit across a range of threshold probabilities, comparing each model against treat-all and treat-none strategies [14]. Model interpretability was achieved using SHAP [10].

3. Results

3.1 Baseline Characteristics and Clinical Trajectories

Of the 336 women enrolled in the primary cohort, 301 met the eligibility criteria for the present analysis (Fig. 1). PPPH at 6 weeks postpartum was observed in 39 patients (12.96%). Baseline demographic characteristics, obstetric history, and clinical parameters stratified by PPPH status are detailed in Table 1.

Maternal age, prepregnancy body mass index (BMI), and the distribution of HDP subtypes did not differ significantly between the 2 groups. Conversely, women who subsequently developed PPPH exhibited a more severe clinical profile upon admission. This clinical severity was characterized by a significantly earlier gestational age at admission ($p = 0.001$), a substantially higher prevalence of severe symptoms ($p = 0.006$), and an elevated admission DBP ($p = 0.007$).

Furthermore, biochemical assessments indicated more pronounced end-organ involvement in the PPPH group, as evidenced by significantly elevated levels of 24-hour proteinuria ($p = 0.039$), blood urea nitrogen (BUN) ($p = 0.010$), and alanine aminotransferase (ALT) ($p = 0.001$). Disparities in hemodynamic trajectories remained evident during the early postpartum period. Postpartum day 1 SBP was significantly higher in the PPPH group ($p = 0.028$). Specifically, both SBP and DBP on postpartum day 3 were significantly higher in women progressing to persistent hypertension ($p = 0.002$ and $p = 0.010$, respectively). Hypertension on postpartum day 3 was also more prevalent in the PPPH group ($p = 0.003$).

3.2 Model Performance and Calibration

All models underwent a 500-repeat bootstrap internal validation to mitigate optimism bias. Crucially, synthetic minority oversampling was strictly confined within each bootstrap resampling loop to prevent data leakage.

Optimism-corrected performance metrics are detailed in Table 2. Ridge LR exhibited the highest discrimination (AUROC: 0.648; 95% CI: 0.642–0.654), followed by the Random Forest model (AUROC: 0.638; 95% CI: 0.632–

0.643) and XGBoost (AUROC: 0.631; 95% CI: 0.625–0.636). Random Forest was ultimately selected as the final model due to its superior overall calibration profile, with the lowest Brier score (0.130), a calibration intercept closest to zero (-0.794), a calibration slope closest to unity (0.759), and an O/E ratio of 0.544. These findings reflect the most accurate agreement between predicted probabilities and observed risks, as supported by the calibration plots (Fig. 2).

To maximize clinical utility as a potential pre-discharge risk-stratification tool, the decision threshold for the Random Forest model was set at 0.157, prioritizing sensitivity over specificity. At this threshold, the model achieved a sensitivity of 80.2%, a specificity of 35.9%, a PPV of 15.5%, and an NPV of 92.5% (Table 2). For comparison, a strategy that classified all patients as low risk would yield a baseline NPV of approximately 87.0%, given the event prevalence of 12.96%. As such, the model provides an absolute NPV improvement of 5.5 percentage points through individualized risk stratification.

The confusion matrix applied to the primary cohort ($n = 301$) is presented in Table 3. Of the 39 PPPH cases, 31 were correctly classified as high-risk (true positives), whereas 8 were missed (false negatives). The trade-off inherent to this high-sensitivity operating point was an elevated false-positive rate, with 168 of 262 women who did not develop PPPH classified as high-risk, reflecting the low baseline prevalence of the outcome.

The confusion matrix applied to the primary cohort ($n = 301$) is presented in Table 3, illustrating the clinical trade-offs of the selected threshold on the original unresampled dataset. Of the 39 PPPH cases, 31 were correctly classified as high-risk (true positives), whereas 8 were missed (false negatives). The trade-off inherent to this high-sensitivity operating point was an elevated false-positive rate, with 168 of 262 women who did not develop PPPH classified as high-risk, reflecting the low baseline prevalence of the outcome.

A sensitivity analysis comparing SMOTE versus no-SMOTE pipelines is presented in **Supplementary 3**. Models trained without SMOTE exhibited improved probabilistic calibration relative to the primary SMOTE models, with calibration intercepts approaching zero (Ridge LR: 0.002; Random Forest: -0.118) and O/E ratios near 1.0 (1.002 and 0.909, respectively). However, the omission of class imbalance correction substantially reduced sensitivity for detecting the minority class. When evaluated at the Youden index threshold, the unadjusted Ridge LR, Random Forest, and XGBoost models achieved sensitivities of 46.6%, 53.2%, and 62.2%, respectively, all failing to meet the prespecified threshold of $\geq 80\%$ for the primary risk stratification models. These findings confirm that SMOTE was necessary to achieve adequate sensitivity in this imbalanced cohort, albeit at the expected cost of probability inflation requiring subsequent Platt scaling recalibration.

Table 2. Optimism-corrected performance of ML models.

Model	AUROC (95% CI), <i>p</i> -value	AUPRC (95% CI)	Brier Score (95% CI)	O/E Ratio	Calibration intercept	Calibration slope	Threshold	Sens	Spec	PPV	NPV	F1 score
Ridge LR (SMOTE)	0.648 (0.517–0.776), <i>p</i> -value < 0.001	0.248 (0.137–0.414)	0.157 (0.143–0.172)	0.374	–1.305	1.434	0.294	0.803	0.331	0.150	0.920	0.252
Random Forest (SMOTE)*	0.638 (0.518–0.745), <i>p</i> -value < 0.001	0.211 (0.119–0.355)	0.130 (0.112–0.152)	0.544	–0.794	0.759	0.157	0.802	0.359	0.155	0.925	0.260
XGBoost (SMOTE)	0.631 (0.513–0.754), <i>p</i> -value < 0.001	0.218 (0.114–0.378)	0.155 (0.137–0.176)	0.398	–1.220	1.191	0.252	0.809	0.344	0.153	0.925	0.257

*Random Forest was selected as the primary model based on its overall predictive performance, including its lower Brier score and calibration characteristics.

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; CI, confidence interval; LR, logistic regression; O/E, observed-to-expected; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; SMOTE, Synthetic Minority Over-sampling Technique; NPV, negative predictive value; ML, machine learning; XGBoost, eXtreme Gradient Boosting.

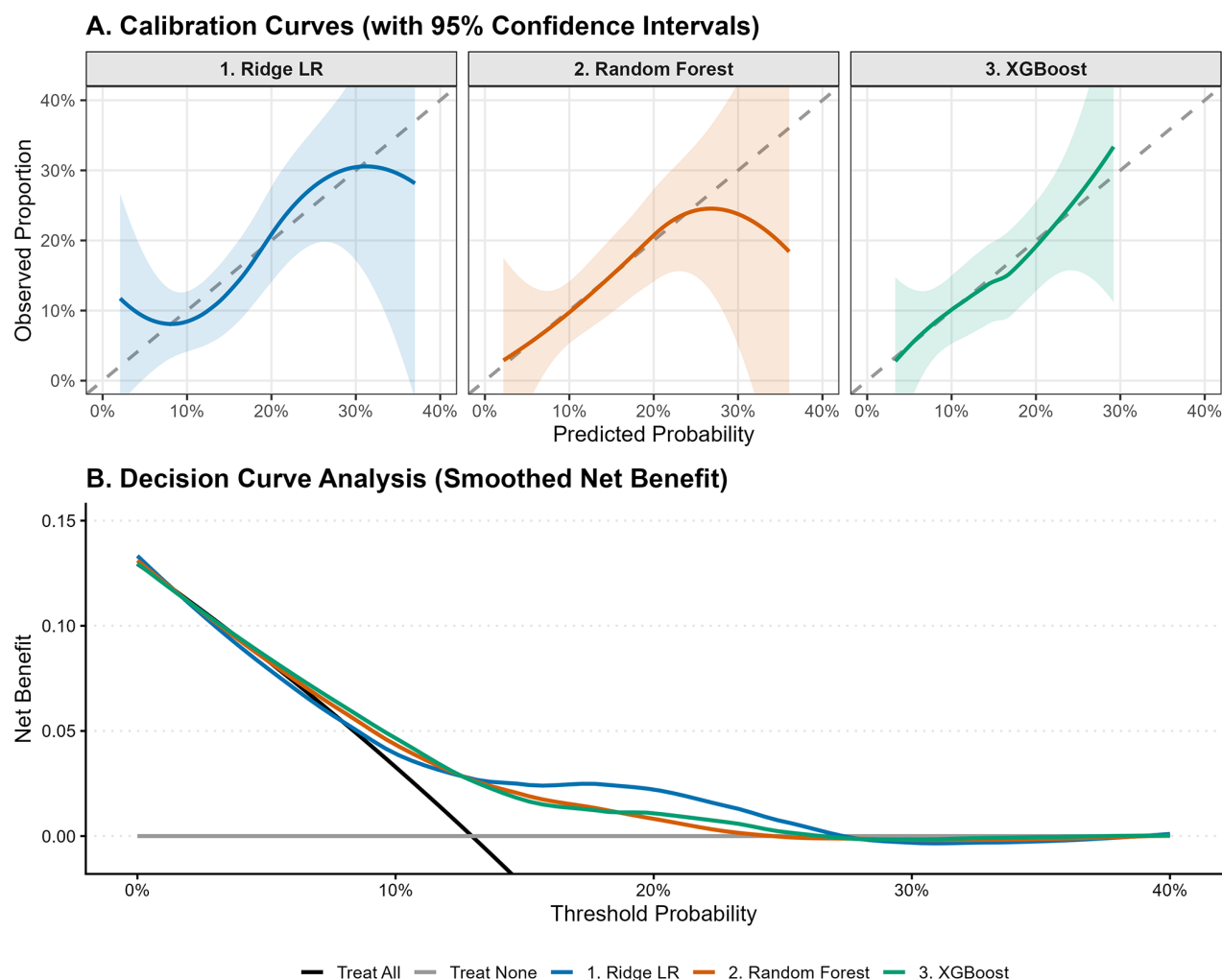


Fig. 2. Calibration and clinical utility dashboard of the predictive models. (A) Calibration curves for the Ridge Logistic Regression (LR), Random Forest, and XGBoost models. The curves display the agreement between the predicted probabilities (x-axis) and the observed proportions (y-axis), with the shaded areas representing the 95% confidence intervals (CI). The dashed diagonal line represents perfect calibration. (B) DCA illustrating the smoothed net benefit of the models across a range of threshold probabilities. The models are compared against the default clinical strategies of treating all patients (solid black line) and treating no patients (solid grey line). LR, logistic regression; DCA, decision curve analysis; XGBoost, eXtreme Gradient Boosting.

Table 3. Confusion matrix for the final Random Forest model at the high-sensitivity threshold (0.157) applied to the study cohort (n = 301).

Prediction	Actual PPPH (Positive)	Actual Resolved (Negative)	Total Predicted
Predicted High-Risk (Positive)	31 (True Positive)	168 (False Positive)	199
Predicted Low-Risk (Negative)	8 (False Negative)	94 (True Negative)	102
Total Actual	39	262	301

Note: Cell counts represent the apparent classification performance (raw integer patient counts) on the original unresampled dataset (n = 301) to illustrate the concrete clinical patient allocation at the selected operating point. These raw counts correspond to an apparent sensitivity of 79.5% (31/39) and an apparent NPV of 92.2% (94/102). In contrast, the performance metrics reported in Table 2 represent the optimism-corrected estimates (sensitivity: 80.2%, NPV: 92.5%) derived from the 500-repeat bootstrap resampling framework to adjust for overfitting bias. Due to the nature of statistical resampling, these corrected expectations cannot be mapped back to integer patient counts without generating fractional values.

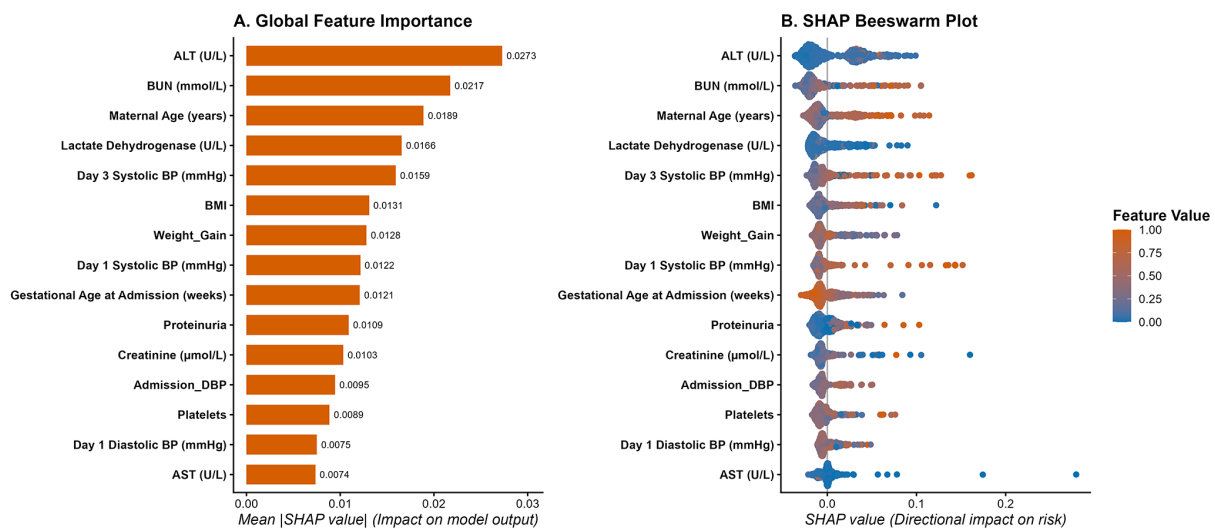


Fig. 3. Explainable Artificial Intelligence (XAI) mapping of the final Random Forest model using Shapley Additive Explanations (SHAP). (A) Global feature importance bar plot ranking the top 15 predictors by their mean absolute SHAP values to illustrate the overall magnitude of their impact on the model's output. (B) SHAP beeswarm plot showcasing the directional effect of the top 15 predictors across all individual patients ($n = 301$). Each dot represents a single patient's data point, color-coded by the actual clinical feature value from low (blue) to high (red). Positive SHAP values ($x\text{-axis} > 0$) indicate an increased predicted risk of persistent postpartum hypertension at 6 weeks. The remaining 10 low-impact variables are omitted from this visualization for optimal visual clarity (their precise contributions are detailed in Supplementary 5). Abbreviations: ALT, Alanine Aminotransferase; BUN, Blood Urea Nitrogen; BP, Blood Pressure; DBP, Diastolic Blood Pressure; SBP, Systolic Blood Pressure; LDH, Lactate Dehydrogenase; SHAP, Shapley Additive Explanations.

3.3 Net Clinical Benefit and Clinical Utility

DCA was used to quantify the net clinical benefit of the models (Fig. 2B). Across the clinically relevant threshold probability range of approximately 15% to 25%, the Ridge LR model consistently achieved the highest net benefit, outperforming both XGBoost and Random Forest models. The clinical utility of the Random Forest model declined more rapidly, intersecting the treat-none strategy at a threshold of approximately 23%. Beyond a 28% threshold, corresponding to the region of greater statistical uncertainty reflected in the calibration curves, all models provided minimal or no additional net benefit compared with the default clinical strategies.

3.4 Model Interpretation and Clinical Relevance

SHAP analysis was performed to elucidate the model's decision-making process and ensure clinical transparency. The final Random Forest model incorporated the following pre-discharge variables: ALT, gestational age at admission, admission SBP, admission DBP, postpartum day 3 SBP, postpartum day 3 DBP, postpartum day 1 SBP, postpartum day 1 DBP, hypertension on postpartum day 3, BUN, proteinuria (g/24 h), severe clinical symptoms, history of preeclampsia, history of gestational hypertension, gestational weight gain, pre-pregnancy BMI, maternal age, platelet count, creatinine, LDH, AST, mode of delivery, multiple pregnancy, gestational diabetes mellitus, and HDP subtype classification.

As shown in Fig. 3A, the global feature importance ranking identified ALT, BUN, maternal age, LDH, and postpartum day 3 SBP as the five most influential predictors. The SHAP beeswarm plot (Fig. 3B) further delineated the direction and magnitude of these features on individual risk estimates. Higher values of markers of hepatic and renal injury (ALT, BUN), elevated LDH, advanced maternal age, and elevated postpartum blood pressure parameters consistently yielded positive SHAP values, directly escalating the predicted risk of PPPH. Conversely, earlier gestational age at admission was strongly associated with a greater predicted risk of PPPH.

4. Discussion

In this study, we developed and internally validated a pre-discharge triage tool to predict PPPH in women with new-onset HDP. By restricting all predictors to the 72-hour pre-discharge window, we eliminated target leakage, which frequently compromises models that incorporate post-discharge variables [9]. Although overall discrimination was modest (AUROC range 0.631–0.648), we strategically prioritized superior model calibration, as evidenced by the lowest Brier score (0.130), the calibration intercept closest to 0 (−0.794), and the calibration slope closest to 1.0 (0.759) achieved by the Random Forest model [15]. Accurate calibration ensures that predicted probabilities reflect real-world risks, an essential prerequisite for personalized clinical decision-making.

The predictive features identified by SHAP are strongly consistent with the established pathophysiology of HDP. The ML framework autonomously prioritized elevated hepatic enzyme levels (ALT), markers of renal end-organ involvement (BUN), advanced maternal age, and elevated postpartum blood pressure trajectories (day 3 SBP). Biologically, elevated ALT reflects hepatic endothelial dysfunction [16], elevated BUN indicates renal involvement and reduced glomerular filtration characteristic of severe HDP, elevated LDH reflects systemic hemolysis and endothelial injury, and persistently elevated postpartum blood pressure on day 3 directly predicts failure of cardiovascular normalization following delivery [17,18]. Advanced maternal age further increases cardiovascular risk through age-related reductions in vascular compliance and endothelial reserve. The algorithm's capacity to independently capture these complex, nonlinear biological patterns supports its clinical plausibility.

The practical utility of these models was confirmed by DCA. Within the clinically relevant threshold probability range of 15%–25%, the Ridge LR model achieved the highest net clinical benefit, outperforming both XGBoost and Random Forest, as well as the default treat-all and treat-none strategies [14]. At the selected threshold of 0.157, the Random Forest model achieved an NPV of 92.5%, representing a 5.5 percentage point improvement over the no-model baseline of 87.0%. This performance enables individualized risk stratification rather than a uniform monitoring approach for all patients. Consequently, this mechanism supports the strategic reallocation of limited clinical resources toward intensive monitoring and early pharmacological intervention for high-risk individuals.

The AUPRC values (0.211–0.248) represent a meaningful improvement over the no-skill baseline of approximately 0.13, consistent with the inherent challenge of predicting a heterogeneous outcome in a small, imbalanced cohort with 39 outcome events. Future studies incorporating larger sample sizes and additional candidate predictors, such as angiogenic biomarkers, which have recently demonstrated significant peripartum kinetic alterations in gestational hypertension [19], or abbreviated urine protein quantification [20], may further improve discriminatory performance.

Limitations

This study has several limitations. First, the single-center design and modest sample size ($n = 301$, including 39 outcome events) limit the generalizability of findings to broader and more diverse clinical settings. While this small number of events yielded a low nominal EPV ratio, modern ML frameworks rely on algorithmic regularization rather than strict EPV thresholds to prevent overfitting. Furthermore, our SHAP analysis confirmed that the model's predictive engine is highly parsimonious; clinical predictions are predominantly driven by a core subset of

5 to 10 features. This results in a robust effective clinical EPV (e.g., 39 events for the top 5 features = 7.8) that exceeds traditional safety thresholds and supports practical bedside feasibility. Second, ascertainment of the 6-week outcome through telephone interviews rather than in-person measurements may have introduced self-reporting or misclassification bias. To optimize data quality, all participants received a validated home blood pressure monitor at no cost, supplemented by structured follow-up conducted by trained research staff. Although these measures likely improved participant retention and data completeness, future protocols would benefit from incorporating scheduled, in-person clinical assessments. Third, the bootstrap validation framework provides only internal validation; therefore, overfitting cannot be fully excluded despite optimism correction. Fourth, the predictive model was developed using a limited set of pre-discharge variables, whereas PPPH is a complex endocrine and hemodynamic process that is likely influenced by additional factors not captured in this dataset. External validation across larger, multicenter cohorts is required before this risk stratification tool can be considered for integration into routine clinical practice. Fifth, our predictive modeling was restricted to the 6-week postpartum period. Although 3-month follow-up data were available (event rate, 7.6%), modeling this later endpoint was not statistically feasible due to the extreme class imbalance and the failure to meet acceptable EPV thresholds, which would inevitably induce severe model overfitting.

5. Conclusions

This study developed and internally validated an ML-based pre-discharge risk stratification framework for PPPH in women with new-onset HDP. Although the current model demonstrated modest discrimination (AUROC: 0.638) and required external validation in larger, multicenter cohorts before clinical implementation, the Random Forest model exhibited a well-calibrated risk profile, with SHAP analysis identifying ALT, BUN, and maternal age as the primary predictors. The NPV of 92.5% at the selected high-sensitivity threshold, representing an absolute improvement over the no-model baseline, supports the potential utility of this proof-of-concept framework for individualized pre-discharge risk stratification and the targeted allocation of postpartum monitoring resources toward high-risk individuals.

Abbreviations

ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; AUPRC, Area Under the Precision-Recall Curve; BMI, Body Mass Index; BUN, Blood Urea Nitrogen; CI, Confidence Interval; DBP, Diastolic Blood Pressure; DCA, Decision Curve Analysis; HDP, Hypertensive Disorders of Pregnancy; KNN, K-Nearest Neighbors; LDH, Lactate Dehydrogenase; LOESS, Locally Weighted Scatterplot Smoothing; LR, Logistic Regression; ML, Ma-

chine Learning; NPV, Negative Predictive Value; O/E, Observed-to-expected; OR, Odds Ratio; PPPH, Persistent Postpartum Hypertension; PPV, Positive Predictive Value; SBP, Systolic Blood Pressure; SHAP, Shapley Additive Explanations; SMOTE, Synthetic Minority Over-sampling Technique; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; TRIPOD-AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Artificial Intelligence; XGBoost, Extreme Gradient Boosting.

Availability of Data and Materials

The datasets generated and analyzed during the current study are not publicly available due to institutional and patient confidentiality policies but are available from the corresponding author upon reasonable request.

Author Contributions

KDN contributed to the conceptualization, data curation, formal analysis, and drafting of the manuscript. TLP and QAN contributed to data collection and critical revision. THAN and TNM contributed to the conceptualization and design of the study, supervised the research, and critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

The study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board for Ethics in Biomedical Research of Hanoi Medical University (Approval Number: 820/GCN-HĐĐNCYSH-ĐHYHN, dated 08 March 2023). Written informed consent was obtained from all participants of the primary prospective cohort prior to their inclusion in the study, which included permission for secondary analysis.

Acknowledgment

The authors would like to thank all patients who participated in the primary prospective cohort at the National Hospital of Obstetrics and Gynecology, Hanoi, Vietnam, and the clinical and laboratory staff for their invaluable support during data collection. The authors also acknowledge the Master's and Doctoral Scholarship Programme of VinGroup Innovation Foundation (VINIF).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Gemini 3.1 Pro in order to check spelling, grammar, and refine the academic language. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at <https://doi.org/10.31083/CEOG53049>.

References

- [1] Duley L. The global impact of pre-eclampsia and eclampsia. *Seminars in Perinatology*. 2009; 33: 130–137. <https://doi.org/10.1053/j.semperi.2009.02.010>
- [2] Folk DM. Hypertensive Disorders of Pregnancy: Overview and Current Recommendations. *Journal of Midwifery & Women's Health*. 2018; 63: 289–300. <https://doi.org/10.1111/jmwh.12725>
- [3] Mosca L, Benjamin EJ, Berra K, Bezanson JL, Dolor RJ, Lloyd-Jones DM, et al. Effectiveness-based guidelines for the prevention of cardiovascular disease in women—2011 update: a guideline from the American heart association. *Circulation*. 2011; 123: 1243–1262. <https://doi.org/10.1161/CIR.0b013e31820faaf8>
- [4] Bellamy L, Casas JP, Hingorani AD, Williams DJ. Pre-eclampsia and risk of cardiovascular disease and cancer in later life: systematic review and meta-analysis. *BMJ (Clinical Research Ed.)*. 2007; 335: 974. <https://doi.org/10.1136/bmj.39335.385301.BE>
- [5] Giorgione V, Ridder A, Kalafat E, Khalil A, Thilaganathan B. Incidence of postpartum hypertension within 2 years of a pregnancy complicated by pre-eclampsia: a systematic review and meta-analysis. *BJOG : an International Journal of Obstetrics and Gynaecology*. 2021; 128: 495–503. <https://doi.org/10.1111/1471-0528.16545>
- [6] Hypertension in pregnancy. Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. *Obstetrics and Gynecology*. 2013; 122: 1122–1131. <https://doi.org/10.1097/01.AOG.0000437382.03963.88>
- [7] Garovic VD, Dechend R, Easterling T, Karumanchi SA, McMurtry Baird S, Magee LA, et al. Hypertension in Pregnancy: Diagnosis, Blood Pressure Goals, and Pharmacotherapy: A Scientific Statement From the American Heart Association. *Hypertension (Dallas, Tex. : 1979)*. 2022; 79: e21–e41. <https://doi.org/10.1161/HYP.0000000000000208>
- [8] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. 2021; 2: 160. <https://doi.org/10.1007/s42979-021-00592-x>
- [9] Kaufman S, Rosset S, Perlich C. Leakage in Data Mining: Formulation, Detection, and Avoidance. *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 556–563). Association for Computing Machinery: NY, United States. 2011. <https://doi.org/10.1145/2020408.2020496>
- [10] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *arXiv*. 2017. <https://doi.org/10.48550/arXiv.1705.07874> (preprint)
- [11] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or ma-

- chine learning methods. *BMJ (Clinical Research Ed.)*. 2024; 385: e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [12] Nguyen MT, Nguyen DK, Truong TH, Le SD, Nguyen THA. Postpartum blood pressure patterns in women with hypertensive disorders of pregnancy. *Ho Chi Minh City Journal of Medicine*. 2025; 28: 68–75. <https://doi.org/10.32895/hcjm.m.2025.07.09> (In Vietnamese)
- [13] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002; 16: 321–357. <https://doi.org/10.1613/jair.953>
- [14] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ (Clinical Research Ed.)*. 2016; 352: i6. <https://doi.org/10.1136/bmj.i6>
- [15] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019; 17: 230. <https://doi.org/10.1186/s12916-019-1466-7>
- [16] Rana S, Lemoine E, Granger JP, Karumanchi SA. Preeclampsia: Pathophysiology, Challenges, and Perspectives. *Circulation Research*. 2019; 124: 1094–1112. <https://doi.org/10.1161/CIRCRESAHA.118.313276>
- [17] Raymond D, Peterson E. A critical review of early-onset and late-onset preeclampsia. *Obstetrical & Gynecological Survey*. 2011; 66: 497–506. <https://doi.org/10.1097/OGX.0b013e3182331028>
- [18] Melchiorre K, Sharma R, Thilaganathan B. Cardiovascular implications in preeclampsia: an overview. *Circulation*. 2014; 130: 703–714. <https://doi.org/10.1161/CIRCULATIONAHA.113.003664>
- [19] Bokulić A, Butorac D, Djaković I, Čukelj M, Stasenکو S, Grdić Rajković M. Peripartum Changes in Neuroendocrine, Angiogenic, and Inflammatory Biomarkers in Gestational Hypertension. *Clinical and Experimental Obstetrics & Gynecology*. 2026; 53: 50841. <https://doi.org/10.31083/CEOG50841>
- [20] Okuyan II, Güngör ES, Kızıldemir YZ, Karslı SE, Zebitay AG. Diagnostic Value of an 8-Hour Urine Protein Measurement in Preeclampsia. *Clinical and Experimental Obstetrics & Gynecology*. 2026; 53: 45273. <https://doi.org/10.31083/CEOG45273>