Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and temporal validation of a model for identifying depressive symptoms in postpartum women

Haoying Wang
Xinjiang Medical University
Xiang Li
Hongxing Hospital, 13th Division of the Xinjiang Production and Construction Corps
Ying Zhang
Hongxing Hospital, 13th Division of the Xinjiang Production and Construction Corps
Wenjing Yong
Hongxing Hospital, 13th Division of the Xinjiang Production and Construction Corps
Shengfeng Chang
Xinjiang Medical University
Xin He
Xinjiang Medical University
Xiaofan Ma
Xinjiang Medical University
Shuqi Zhang
Xinjiang Medical University
Ting Jiang
jt5583@126.com

Xinjiang Medical University

Research Article

Keywords: Postpartum depression, EPDS, Interpretable machine learning, Temporal validation, Model calibration

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11318045/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Development and temporal validation of a model for identifying depressive symptoms in postpartum women

Haoying Wang[1]†, Xiang Li[2]†, Ying Zhang[3], Wenjing Yong[2], Shengfeng Chang[1], Xin He[1], Xiaofan Ma[1], Shuqi Zhang[1], Ting Jiang[1]

[1] School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang 830000, China

[2] Department of Clinical Laboratory, Hongxing Hospital, 13th Division of the Xinjiang Production and Construction Corps, Hami, Xinjiang 839000, China

[3] Department of Obstetrics, Hongxing Hospital, 13th Division of the Xinjiang Production and Construction Corps, Hami, Xinjiang 839000, China

Corresponding author: Ting Jiang, School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang 830000, China. Email: jt5583@126.com

† Haoying Wang and Xiang Li contributed equally to this work.

Author email addresses

Haoying Wang: 3377644901@qq.com

Xiang Li: 115130517@qq.com

Ying Zhang: zhangying.xj@163.com

Wenjing Yong: 878982917@qq.com

Shengfeng Chang: shengfeng1021@qq.com

Xin He: hx51870@163.com

Xiaofan Ma: 839584213@qq.com

Shuqi Zhang: zhangshuqi202203@163.com

Ting Jiang: jt5583@126.com

Abstract

Background: Identifying early postpartum depressive symptoms is important for perinatal mental healthcare. The incremental value of psychosocial scales beyond routine clinical information remains uncertain. We developed and temporally validated models identifying Edinburgh Postnatal Depression Scale (EPDS) screen positivity at a tertiary hospital in Xinjiang, China.

Methods: This cross-sectional study included 3595 women recruited between February 2023 and May 2026. Clinical information and scores from EPDS, the Connor-Davidson Resilience Scale, Perceived Social Support Scale, and Satisfaction With Life Scale were collected within six days postpartum. EPDS ≥10 defined screen positivity. Analyses used Mann–Whitney U and χ^2 tests. Admission dates determined development (2023–2024; n=2089) and temporal validation (2025–2026; n=1402) sets; 104 women with missing or invalid dates were excluded from modelling. Elastic net logistic regression, Explainable Boosting Machine (EBM), CatBoost, and XGBoost were compared using strict nested cross-validation, Platt calibration of out-of-fold predictions, and prespecified high-sensitivity threshold selection. Discrimination, calibration, decision curves, and review workload were evaluated.

Results: Overall, 892 women (24.8%) screened positive. Temporal validation areas under the receiver operating characteristic curve were 0.837, 0.832, 0.832, and 0.821, respectively. Corresponding precision–recall areas were 0.623, 0.619, 0.612, and 0.599. EBM had a Brier score of 0.129, calibration slope of 1.162, and grouped expected calibration error of 0.031. Its locked threshold of 0.146 yielded 91.4% sensitivity, 50.9% specificity, 36.0% positive predictive value, and 95.1% negative predictive value, requiring approximately 589 reviews per 1000 women. The psychosocial model improved discrimination over the basic clinical model by 0.250 (95% confidence interval 0.210–0.290). Adding clinical predictors to psychosocial scores offered no additional gain (difference −0.006; 95% confidence interval −0.016 to 0.004). Resilience, life satisfaction, and perceived social support ranked highest in permutation importance.

Conclusions: Psychosocial scales provided most information for identifying concurrent EPDS screen positivity; routine clinical variables offered no additional discrimination. The model cannot replace EPDS screening or psychiatric diagnosis. Multicentre prospective evaluation is needed.

Keywords: Postpartum depression; EPDS; Interpretable machine learning; Temporal validation; Model calibration

Background

Postpartum depression (PPD) is a common perinatal mental disorder characterised by persistent low mood, diminished interest, and reduced energy. Anxiety, sleep disturbances, self-blame, and thoughts of self-harm often accompany these symptoms [1, 2]. International and Chinese meta-analyses have reported pooled proportions of postpartum depression or depressive symptoms of approximately 17.7% and 21.0%,

respectively [3, 4]. Assessment instruments, thresholds, and measurement times varied across studies. PPD may increase maternal self-harm or suicide risk, impair mother–infant interactions, and affect children's emotional, cognitive, and behavioural development [1, 5]. Experts in China and internationally therefore recommend standardised screening during pregnancy and postpartum, linked to assessment, referral, and follow-up for positive results [6, 7]. Timely identification of women experiencing postpartum depressive symptoms remains an important clinical challenge.

The Edinburgh Postnatal Depression Scale (EPDS) is widely used to screen for perinatal depressive symptoms [8, 9]. Its Chinese versions have demonstrated good reliability and validity [10–12]. However, EPDS is a symptom screening instrument rather than a diagnostic tool. Its positive screening rate depends on the threshold and assessment timing [13]. We selected three scales to capture psychosocial information that routine medical records rarely document systematically. The Connor-Davidson Resilience Scale (CD-RISC) measures adaptive resources for coping with childbirth and changes in maternal roles. The Satisfaction With Life Scale (SWLS) measures cognitive evaluations of overall life circumstances. The Perceived Social Support Scale (PSSS) measures subjective perceptions of support from family, friends, and others. These scales cover three complementary dimensions: individual adaptive resources, subjective life appraisal, and social support resources. Their development and use in Chinese populations are described elsewhere [14–19]. We used them to assess whether psychosocial information improves identification of concurrent screen positivity beyond clinical information.

Combining psychosocial scales with clinical data enables assessment of the identification value of both information sources. Machine learning can handle multidimensional variables and potential nonlinear relationships and has been applied in PPD research. However, previous studies have identified different algorithms as optimal [20, 21]. Comparing multiple algorithms alone does not establish reproducibility or clinical value. Model development must also prevent information leakage between tuning and performance evaluation and report discrimination, calibration, and clinical consequences. We used Elastic Net Logistic as a transparent linear benchmark and the Explainable Boosting Machine (EBM) as the primary interpretable nonlinear model. CatBoost and XGBoost served as tree boosting benchmarks. Beyond comparing algorithms, we assessed probability estimates, threshold performance, and review workload in a subsequent period.

Participants were divided into two non-overlapping periods according to admission dates. Models were developed and locked using the earlier period and evaluated using new participants from the later period. This approach is termed temporal validation [22–24]. As both periods involved the same hospital, it assesses performance in women admitted subsequently but cannot replace external validation at independent institutions.

We aimed to develop a model identifying postpartum EPDS screen positivity and compare basic clinical, psychosocial, and full predictor sets. We quantified the incremental identification value of psychosocial scales and evaluated calibration, threshold-related workload, and subgroup stability. Screen positivity was explicitly distinguished from psychiatric diagnosis. Concurrent identification was not interpreted as prediction of future PPD onset.

Methods

Study participants

Women who delivered at the obstetric department of a tertiary hospital in Xinjiang between February 2023 and May 2026 completed postpartum questionnaires. All participants were enrolled after providing informed consent.

Inclusion criteria were: (1) age ≥18 years; (2) registration and inpatient delivery at the study hospital; (3) ability to understand and complete questionnaires independently or with an investigator's assistance; (4) voluntary participation with signed informed consent; and (5) ability to complete electronic questionnaires and scales through WeChat on a mobile phone. Exclusion criteria were: (1) a previous confirmed diagnosis of a mental disorder, such as schizophrenia or bipolar disorder; (2) severe cognitive, language, or consciousness impairment preventing survey completion; (3) missing information for an entire questionnaire or scale; and (4) explicit refusal after the study had been explained. The study followed the Declaration of Helsinki and received approval from the Ethics Review Committee of Xinjiang Medical University (approval number: XJYKDXR20230303023).

Data collection

This was a cross-sectional study. Women completed electronic questionnaires during their postpartum hospital stay. Researchers concurrently reviewed electronic medical records to verify pregnancy, delivery, and neonatal information. Questionnaires standardised the assessment of depressive symptoms, resilience,

life satisfaction, and perceived social support, which routine records often capture incompletely. They also enabled evaluation of the incremental value of psychosocial information.

General information questionnaire

Demographic variables included age, educational attainment, place of residence, living environment, total monthly household income, and employment status. Reproductive variables included gravidity, mode of delivery, pregnancy complications, and postpartum haemorrhage. Neonatal variables included sex, birth weight, gestational age category, neonatal asphyxia, birth defects, and neonatal disease. Women reported demographic information. Pregnancy, delivery, and neonatal outcomes were verified against electronic medical records.

Psychosocial questionnaires

Edinburgh Postnatal Depression Scale

The EPDS was used to assess postpartum depressive symptoms [8]. It comprises 10 items scored from 0 to 3, giving a total score of 0–30. Higher scores indicate more severe depressive symptoms. Validation in a pregnant mainland Chinese sample reported 80.0% sensitivity, 83.0% specificity, and a ROC-AUC of 0.856 at a cutoff of 9.5, equivalent to ≥10 [10]. A validation study among postpartum women in Chengdu reported a ROC-AUC of 0.896, Cronbach's α of 0.78, and test–retest reliability of 0.90 [11]. Based on Chinese validation evidence and the screening priority of sensitivity, we prespecified EPDS ≥10 as screen positive and <10 as screen negative [10, 11, 13]. In our sample, Cronbach's α was 0.882, and corrected item–total correlations ranged from 0.436 to 0.715. The Kaiser–Meyer–Olkin (KMO) value was 0.899. Bartlett's test of sphericity yielded χ^2=16119.11 (df=45, P<0.001), supporting internal consistency and suitability for factor analysis.

Connor-Davidson Resilience Scale

Resilience was assessed using the CD-RISC [14, 15]. Its 25 items cover tenacity, strength, and optimism and use a five-point Likert scale. Original responses of 1–5 were recoded to 0–4, yielding a total score of 0–100. Higher scores indicate greater resilience. The original scale reported Cronbach's α of 0.89 and test–retest reliability of 0.87 [14]. The Chinese version reported Cronbach's α of 0.91 and supported the three-dimensional structure of tenacity, strength, and optimism [15]. Total scores entered the models as continuous variables without within-sample dichotomisation. After excluding one out-of-range item value, Cronbach's α among 3594 respondents with complete valid responses was 0.963. Corrected item–total

correlations ranged from 0.345 to 0.834. The KMO value was 0.979, and Bartlett's test yielded χ^2=70144.63 (df=300, P<0.001).

Satisfaction With Life Scale

The SWLS assessed women's cognitive judgements of their overall life circumstances [16, 17]. This unidimensional scale comprises five positively scored items rated on a seven-point Likert scale (1–7), yielding total scores of 5–35. Higher scores indicate greater life satisfaction. The original scale reported Cronbach's α of 0.87 and two-month test–retest reliability of 0.82. A single factor explained 66% of the variance [16]. A Chinese general-population study reported Cronbach's α of 0.78 and split-half reliability of 0.70. Confirmatory factor analysis yielded a comparative fit index (CFI) of 0.96 and root mean square error of approximation (RMSEA) of 0.071 [17]. Total scores entered the models as continuous variables. In our sample, Cronbach's α was 0.907, and corrected item–total correlations ranged from 0.636 to 0.849. The KMO value was 0.881, and Bartlett's test yielded χ^2=13444.24 (df=10, P<0.001).

Perceived Social Support Scale

The PSSS assessed subjectively perceived social support. The scale was developed by Zimet and colleagues; its introduction and adaptation in Chinese were referenced through Jiang Qianjin's scale description [18, 19]. Its 12 items cover support from family, friends, and others. All items retain their original scoring direction and use a seven-point Likert scale (1–7), yielding total scores of 12–84. Higher scores indicate greater perceived social support. Total scores entered the models as continuous variables. After excluding one out-of-range item value, Cronbach's α among 3594 respondents with complete valid responses was 0.967. Corrected item–total correlations ranged from 0.797 to 0.861. The KMO value was 0.946, and Bartlett's test yielded χ^2=51029.81 (df=66, P<0.001).

Outcome definition

The primary outcome was EPDS screen positivity within six days postpartum. All predictors were assessed and collected within six days after delivery. Based on Chinese validation studies and the screening priority of sensitivity, EPDS ≥10 defined screen positivity and <10 defined screen negativity [10, 11, 13]. This outcome indicates potentially clinically significant depressive symptoms and is not equivalent to psychiatric diagnosis. Any reported thoughts of self-harm should immediately trigger a safety assessment and referral, regardless of model results.

Statistical analysis

Analyses were performed using Python. Before analysis, the data dictionary was fixed, and variable labels, coding directions, theoretical ranges, and date formats were checked. EPDS items were scored 0–3; original SWLS item scores were retained and summed. CD-RISC items were recoded to 0–4, and PSSS items retained their 1–7 coding. One out-of-range value was identified in each of the CD-RISC and PSSS. These values were treated as missing, and total scores were prorated using each participant's remaining valid items from the same scale. Continuous variables were summarised as medians and interquartile ranges and compared using Mann–Whitney U tests, with Z statistics reported. Categorical variables were reported as n (%) and compared using Pearson's χ^2 tests; Yates' continuity correction was applied to 2×2 tables. All tests were two-sided, with $P<0.05$ indicating statistical significance. As living environment contained sparse categories, its asymptotic P value requires cautious interpretation. Predictors were not selected based on univariable statistical significance.

The clinical predictor set included age and 15 categorical demographic, obstetric, and neonatal variables. The psychosocial predictor set comprised SWLS, CD-RISC, and PSSS total scores. The full set included all 19 variables (Table 1). Continuous variables were standardised and categorical variables were one-hot encoded within each training pipeline. Categories absent during training were ignored during encoding. All four algorithms used the same full predictor set; EBM was used to compare the three predictor sets. Elastic net logistic regression served as the linear benchmark [25]. EBM used additive boosting without pairwise interaction terms [26]. CatBoost and XGBoost served as tree boosting comparators [27, 28]. Predictors were not selected retrospectively using permutation importance.

Admission year defined two non-overlapping groups: development in 2023–2024 (n=2089; 537 screen-positive women) and temporal validation in 2025–2026 (n=1402; 325 screen-positive women). Dates for another 104 participants could not be assigned under these rules, so they were excluded from modelling. Temporal validation within the same hospital assesses applicability to subsequent admissions. It does not establish direct transferability to independent institutions [22–24, 29].

Model development used stratified nested cross-validation with five outer folds and three inner folds. Hyperparameters were selected by mean ROC-AUC [30]. Predictions for each outer fold came from a model whose tuning did not use those observations. Pooled out-of-fold probabilities were transformed to logits and used to fit a logistic recalibration model, implementing Platt calibration [31]. Among thresholds

achieving ≥90% sensitivity in calibrated development predictions, we selected the threshold with the highest specificity. Calibration and threshold selection used the same pooled development outcomes; no additional calibration cross-fitting was performed. Final hyperparameters were subsequently selected using five-fold cross-validation on all development data, and each model was refitted. The locked calibrator and threshold were then applied to the temporal validation group. Validation participants contributed to neither model fitting nor tuning, calibration, or threshold selection.

The final elastic net model used C=0.03 and an L1 ratio of 0.1. EBM used a learning rate of 0.02, 128 maximum bins, five minimum samples per leaf, 800 maximum rounds, and four outer bags, without interactions. CatBoost used 400 iterations, depth 4, a learning rate of 0.03, and L2 leaf regularisation of 3. XGBoost used 400 trees, maximum depth 3, a learning rate of 0.03, and row and column subsampling ratios of 0.8. The random seed was 20260806. Analyses were implemented in Python.

Discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC). PR-AUC was calculated using average precision. Overall predictive accuracy was assessed using the Brier score. Calibration intercepts and slopes were jointly estimated by logistic regression of outcomes on the logits of predicted probabilities [32]. Grouped expected calibration error (ECE) was also calculated. Participants were divided into 10 approximately equal-frequency groups by predicted probability. The absolute difference between observed positivity and mean predicted probability in each group was averaged using group-size weights. This grouped measure is not the integrated calibration index derived from a smoothed calibration curve. Calibration plots used Wilson 95% confidence intervals (CIs) for each group's observed positive proportion.

Decision curves assessed net benefit relative to reviewing all or none of the women [33]. Net benefit was calculated as $TP/n-(FP/n)\times t/(1-t)$, where t denotes the threshold probability and EPDS screening status is the reference outcome. CIs for performance metrics were estimated using 1000 bootstrap resamples. Differences in ROC-AUC, PR-AUC, and Brier scores were estimated using 2000 paired bootstrap resamples. These intervals reflect sampling variation in the validation population conditional on the fitted models. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and review workload were calculated from confusion matrices. Counts per 1000 women were proportionally extrapolated from the validation sample and rounded to integers.

EBM was selected for detailed reporting after considering discrimination, calibration, and interpretability in temporal validation. This selection was influenced by validation results and was not an independent confirmation of a prespecified optimal algorithm. Permutation importance was calculated using 50 repetitions and expressed as the reduction in ROC-AUC. Error bars represent the standard deviation (SD) across repetitions. Exploratory subgroup results were reported only for groups containing at least 30 participants, 10 screen-positive women, and 10 screen-negative women.

The study design and model development workflow are shown in Fig. 1.

Use of artificial intelligence

DeepSeek was used to assist with language editing, formatting, writing code for statistical checks, and checking consistency across analytical outputs.

Results

Demographic information, pregnancy outcomes, and scale characteristics

Of 3595 women, 892 (24.8%) screened positive on the EPDS within six days postpartum and 2703 (75.2%) screened negative. SWLS, CD-RISC, and PSSS scores were lower in the screen-positive group (all $P<0.001$; Table 2). The primary modelling analysis included 3491 women, of whom 862 screened positive. Among the 104 excluded women, 30 (28.8%) screened positive. Their numerically higher positive rate and exclusion due to date quality warrant caution when generalising temporal validation results.

Predictor sets and incremental value

Performance of the three predictor sets is presented in Table 3. In temporal validation, ROC-AUCs for the basic clinical, psychosocial, and full sets were 0.588, 0.838, and 0.832, respectively. Corresponding PR-AUCs were 0.320, 0.628, and 0.619, and Brier scores were 0.175, 0.128, and 0.129. Paired comparisons showed that the psychosocial model improved on the basic clinical model: ΔAUC=0.250 (95% CI 0.210–0.290), ΔPR-AUC=0.308 (95% CI 0.248–0.365), and ΔBrier=−0.047 (95% CI −0.056 to −0.039). The full model versus the psychosocial model yielded ΔAUC=−0.006 (95% CI −0.016 to 0.004). Psychosocial scales therefore provided the main incremental identification information, while routine clinical variables showed no additional discrimination gain.

The primary modelling analysis included 3491 women: 2089 in the development set (537 screen positive) and 1402 in the temporal validation set (325 screen positive). Annual counts were 408 in 2023, 1681 in

2024, 1368 in 2025, and 34 in 2026. The small 2026 sample precluded separate conclusions about stability. Another 104 women with missing or invalid dates were excluded from primary modelling and included only in descriptive sensitivity assessments.

Model performance in temporal validation

Temporal validation results for the four full models are presented in Table 4 and Fig. 2. ROC-AUCs for Elastic Net Logistic, EBM, CatBoost, and XGBoost were 0.837 (95% CI 0.811–0.860), 0.832 (0.807–0.857), 0.832 (0.808–0.856), and 0.821 (0.796–0.846), respectively. Their PR-AUCs were 0.623, 0.619, 0.612, and 0.599; Brier scores were 0.130, 0.129, 0.131, and 0.134. Calibration slopes were 1.417, 1.162, 1.283, and 1.210, and grouped ECE values were 0.054, 0.031, 0.039, and 0.035. The paired ΔAUC for EBM versus Elastic Net Logistic was −0.004 (95% CI −0.013 to 0.004), with the interval including zero. Given similar discrimination, better calibration, and direct representation of nonlinear contributions, EBM was selected for primary reporting. This does not establish its absolute superiority over other algorithms. The high-sensitivity EBM threshold locked using calibrated development out-of-fold predictions was 0.146. Applied to temporal validation, it yielded 91.4% sensitivity (95% CI 88.0–94.3%), 50.9% specificity (95% CI 47.9–53.8%), 36.0% PPV, and 95.1% NPV. Approximately 589 women per 1000 would be flagged for review, comprising 212 true positives and 377 false positives. Approximately 20 screen-positive women would be missed, while 391 would be true negatives. This threshold prioritises reducing missed positive screens at the cost of increased review workload.

Full model performance and EBM classification performance at different thresholds are presented in Tables 4 and 5.

In temporal validation, EBM had a Brier score of 0.129 (95% CI 0.119–0.139), calibration intercept of −0.049 (95% CI −0.224 to 0.126), and calibration slope of 1.162 (95% CI 1.019–1.319). Its grouped ECE was 0.031. ROC and precision–recall curves, calibration curves with Wilson 95% CIs, decision curves, and expected workload per 1000 women are shown in Figs. 3–6.

Model interpretability

After 50 permutation repetitions in the temporal validation set, resilience, life satisfaction, and perceived social support ranked highest in importance. Mean ROC-AUC reductions were 0.093, 0.083, and 0.015, respectively. Error bars indicate the SD across repetitions. As the three scales were correlated, permutation

could redistribute importance among correlated predictors. This ranking reflects model reliance, not independent effects, a ranking of risk factors, or causal relationships. Results are shown in Fig. 7.

In exploratory subgroups, EBM ROC-AUCs were 0.863, 0.816, and 0.837 for women aged <30, 30–34, and ≥35 years, respectively. ROC-AUCs were 0.840 for vaginal delivery, 0.830 for caesarean delivery, and 0.833 for preterm delivery. Only 34 women were included in 2026; estimates for that year were not interpreted definitively. Excluding sensitive attributes from modelling does not establish fairness. Subgroup results served only to describe stability.

Discussion

This study examined identification of EPDS screen positivity during the early postpartum hospital stay. Strict nested cross-validation separated tuning, out-of-fold calibration, and threshold locking from temporal validation, reducing information leakage. Psychosocial predictors provided substantial incremental information beyond basic clinical variables. Adding routine clinical variables did not further improve discrimination. EBM and Elastic Net Logistic showed similar discrimination. EBM's lower Brier score and grouped ECE supported its detailed presentation but did not establish superior overall performance over the linear benchmark. Selection of EBM for detailed reporting was influenced by temporal validation results and requires further assessment in new independent samples. This reporting approach accords with TRIPOD+AI principles: specifying model use, separating development from validation, and reporting discrimination, calibration, and clinical consequences [34].

Resilience had the highest permutation importance in the final EBM. The CD-RISC characterises adaptive resources under stress and adversity [14, 15]. During early postpartum hospitalisation, childbirth, pain, sleep changes, and role transitions occur together. The concurrence of lower resilience and greater EPDS symptom burden is therefore clinically interpretable. However, CD-RISC and EPDS were measured concurrently by self-report in a cross-sectional design. Importance indicates contribution to classification and cannot establish that low resilience preceded or caused depressive symptoms.

Life satisfaction also carried substantial model information. The SWLS assesses cognitive judgements about overall life circumstances [16]. Its relationship with EPDS screening status may reflect both underlying psychological associations and construct overlap or common method bias from concurrent self-report. High permutation importance should therefore be interpreted as identification information within a broader psychological assessment. It cannot establish an independent protective factor or

intervention target. Longitudinal studies should measure SWLS before EPDS and assess its predictive value against later clinical outcomes.

Perceived social support contributed relatively less identification information. Previous studies have linked partner support and conflict to postpartum depressive symptoms and included social support in broader assessments of PPD risk factors [35, 36]. The PSSS covers support from family, friends, and others and may inform resource assessment following a positive screen. However, its scores are also influenced by family structure, caregiving expectations, and the hospital setting. Measurement performance and model calibration require reassessment before use across regions or institutions.

The basic clinical model achieved a ROC-AUC of only 0.588. The full model yielded $\Delta AUC = -0.006$ relative to the psychosocial model, indicating no additional discrimination from the available demographic, delivery, and neonatal information. This does not imply that clinical factors are unrelated to PPD. It may reflect the limited candidate variables, insufficient information detail, and the use of concurrent symptom screening as the outcome. Large electronic health record studies can incorporate psychiatric history, medication use, and longitudinal care information to predict outcomes within one year postpartum [37]. Their research question differs from concurrent identification during hospitalisation. Clinical value therefore cannot be compared directly using AUC alone.

At the locked threshold of 0.146, EBM achieved high sensitivity and NPV but limited specificity and PPV. Approximately 589 women per 1000 required review, including 377 false positives. The model is therefore better suited to research on stratification and review workload than to replacing EPDS. Before clinical piloting, departments should specify who performs reviews, actions following positive flags, and acceptable workload. Multicentre external validation, input checks, cost–consequence analysis, and patient safety monitoring are also needed. As the three psychosocial scales contain 42 items, the model cannot currently be claimed to reduce questionnaire burden.

Limitations

This study has several limitations. First, its single-centre cross-sectional design identifies concurrent screening status within six days postpartum and cannot predict future onset. Second, the outcome was EPDS screen positivity rather than a structured psychiatric diagnosis. Symptoms within six days postpartum cannot establish subsequent persistent depression. Third, concurrent self-report of the three predictor scales and EPDS introduces construct overlap, common method bias, and additional assessment

burden. These factors may inflate discrimination and limit clinical efficiency. Fourth, the calibrator and threshold were estimated using the same development out-of-fold outcomes without additional cross-fitting. Selection of EBM for detailed reporting was also influenced by validation results. Strict nested validation reduces but does not eliminate model development bias [34, 38]. EBM interpretations remain affected by correlated variables, sample density, and sparse categories. Fifth, 104 women with missing or invalid dates were excluded from primary modelling. Their numerically higher positive rate may introduce selection bias. Sixth, temporal validation used the same hospital and data source and cannot replace external validation at independent institutions [24]. Only 34 women were included in 2026, limiting evidence of annual stability. Seventh, subgroup analyses were exploratory, with few events in some groups, and cannot establish fairness. Future multicentre prospective studies should externally validate the model against independent clinical interviews or later outcomes. They should compare model-assisted and direct EPDS pathways for workload, costs, missed positive screens, and referral completion.

Conclusions

Among 3595 hospitalised postpartum women, 892 (24.8%) screened positive on EPDS within six days postpartum. Strict nested validation and temporal validation showed similar discrimination across four models, with EBM offering good calibration and interpretability. Paired incremental analyses indicated that psychosocial scales provided most information for concurrent identification; routine clinical variables added no discrimination gain. The model cannot predict future onset, and the 42 psychosocial scale items increase assessment burden. It should not replace EPDS or support automated decisions before multicentre external validation and clinical impact studies are completed.

List of abbreviations

PPD, Postpartum depression; EPDS, Edinburgh Postnatal Depression Scale; CD-RISC, Connor-Davidson Resilience Scale; SWLS, Satisfaction With Life Scale; PSSS, Perceived Social Support Scale; EBM, Explainable Boosting Machine; ROC-AUC, Area under the receiver operating characteristic curve; PR-AUC, Area under the precision–recall curve (calculated using average precision in this study); ECE, Expected calibration error; CI, Confidence interval; PPV, Positive predictive value; NPV, Negative predictive value; SD, Standard deviation; KMO, Kaiser–Meyer–Olkin measure of sampling adequacy;

CFI, Comparative fit index; RMSEA, Root mean square error of approximation; TP, True positive; FP, False positive.

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Review Committee of Xinjiang Medical University (approval number: XJYKDXR20230303023). All participants provided informed consent.

Availability of data and materials

The datasets analysed during the current study are not publicly available and are not shared externally because of participant privacy and ethical restrictions.

Competing interests

The authors declare that they have no competing interests.

Funding

Supported by the National Natural Science Foundation of China (82360669), the Key Laboratory of Population Health and Eugenics of Anhui Province (JKYS20231), and the Xinjiang Uygur Autonomous Region “14th Five-Year Plan” Higher Education Characteristic Disciplines–Public Health and Preventive Medicine.

Authors' contributions

HW, SC, and TJ designed the study and drafted the manuscript. HW, XL, YZ, WY, XH, XM, and SZ collected the data. HW and SC analysed the data and interpreted the results. HW, XL, YZ, WY, SC, XH, XM, SZ, and TJ revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

We thank all women who participated in this study and the clinical staff who assisted with data collection.

References

[1] O'Hara MW, McCabe JE. Postpartum depression: current status and future directions. Annu Rev Clin Psychol. 2013;9:379-407. doi:10.1146/annurev-clinpsy-050212-185612.

[2] Wisner KL, Moses-Kolko EL, Sit DK. Postpartum depression: a disorder in search of a definition. Arch Womens Ment Health. 2010;13(1):37-40. doi:10.1007/s00737-009-0119-9.

[3] Hahn-Holbrook J, Cornwell-Hinrichs T, Anaya I. Economic and health predictors of national postpartum depression prevalence: a systematic review, meta-analysis, and meta-regression of 291 studies from 56 countries. Front Psychiatry. 2018;8:248. doi:10.3389/fpsyt.2017.00248.

[4] Mu TY, Li YH, Pan HF, Zhang L, Zha DH, Zhang CL, et al. Postpartum depressive mood (PDM) among Chinese women: a meta-analysis. Arch Womens Ment Health. 2019;22(2):279-287. doi:10.1007/s00737-018-0885-3.

[5] Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry. 2020;19(3):313-327. doi:10.1002/wps.20769.

[6] World Health Organization. WHO recommendations on maternal and newborn care for a positive postnatal experience. Geneva: World Health Organization; 2022. https://www.who.int/publications/i/item/9789240045989. Accessed 9 Oct 2026.

[7] Chen J, Zou T, Zhao DQ, Xiao ZW, Wu XQ, Perinatal Mental Disorders Collaborative Group of the Chinese Society of Psychosomatic Medicine, Chinese Medical Association. Expert consensus on screening, diagnosis and treatment of perinatal mental disorders. Chinese General Practice. 2023;26(28):3463-3470. Chinese. doi:10.12114/j.issn.1007-9572.2023.0114.

[8] Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1987;150:782-786. doi:10.1192/bjp.150.6.782.

[9] Hewitt CE, Gilbody SM, Mann R, Brealey S. Instruments to identify post-natal depression: which methods have been the most extensively validated, in what setting and in which language? Int J Psychiatry Clin Pract. 2010;14(1):72-76. doi:10.3109/13651500903198020.

[10] Wang Y, Guo X, Lau Y, Chan KS, Yin L, Chen J. Psychometric evaluation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale. Int J Nurs Stud. 2009;46(6):813-823. doi:10.1016/j.ijnurstu.2009.01.010.

[11] Lau Y, Wang Y, Yin L, Chan KS, Guo X. Validation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale in Chengdu mothers. Int J Nurs Stud. 2010;47(9):1139-1151. doi:10.1016/j.ijnurstu.2010.02.005.

[12] Lee DTS, Yip SK, Chiu HFK, Leung TYS, Chan KPM, Chau IOL, et al. Detecting postnatal depression in Chinese women. Validation of the Chinese version of the Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1998;172:433-437. doi:10.1192/bjp.172.5.433.

[13] Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ. 2020;371:m4022. doi:10.1136/bmj.m4022.

[14] Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety. 2003;18(2):76-82. doi:10.1002/da.10113.

[15] Yu X, Zhang J. Factor analysis and psychometric evaluation of the Connor-Davidson Resilience Scale (CD-RISC) with Chinese people. Soc Behav Pers. 2007;35(1):19-30. doi:10.2224/sbp.2007.35.1.19.

[16] Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. J Pers Assess. 1985;49(1):71-75. doi:10.1207/s15327752jpa4901_13.

[17] Xiong CQ, Xu YL. Reliability and validity of the Chinese version of the Satisfaction With Life Scale in the general population. Chinese Journal of Health Psychology. 2009;17(8):948-949. Chinese. doi:10.13342/j.cnki.cjhp.2009.08.026.

[18] Zimet GD, Dahlem NW, Zimet SG, Farley GK. The Multidimensional Scale of Perceived Social Support. J Pers Assess. 1988;52(1):30-41. doi:10.1207/s15327752jpa5201_2.

[19] Jiang QJ. Perceived Social Support Scale. Chinese Journal of Behavioral Medical Science. 2001;10(Special issue):41-42. Chinese.

[20] Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health. 2021;8(11):e29838. doi:10.2196/29838.

[21] Cellini P, Pigoni A, Delvecchio G, Moltrasio C, Brambilla P. Machine learning in the prediction of postpartum depression: a review. J Affect Disord. 2022;309:350-357. doi:10.1016/j.jad.2022.04.093.

[22] Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. doi:10.1136/bmj-2023-074819.

[23] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49-58. doi:10.1093/ckj/sfaa188.

[24] Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022;22(1):316. doi:10.1186/s12874-022-01801-8.

[25] Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22. doi:10.18637/jss.v033.i01.

[26] Nori H, Jenkins S, Koch P, Caruana R. InterpretML: a unified framework for machine learning interpretability. arXiv [Preprint]. 2019;1909.09223. https://arxiv.org/abs/1909.09223. Accessed 9 Oct 2026.

[27] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31:6638-6648.

[28] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785-794. doi:10.1145/2939672.2939785.

[29] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207.

[30] Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-781. doi:10.1016/s0895-4356(01)00341-9.

[31] Kull M, Silva Filho T, Flach P. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. Proc Mach Learn Res. 2017;54:623-631. Available from: https://proceedings.mlr.press/v54/kull17a.html. Accessed 9 Oct 2026.

[32] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7.

[33] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. doi:10.1177/0272989x06295361.

[34] 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 machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

[35] Dennis CL, Ross L. Women's perceptions of partner support and conflict in the development of postpartum depressive symptoms. J Adv Nurs. 2006;56(6):588-599. doi:10.1111/j.1365-2648.2006.04059.x.

[36] Hutchens BF, Kearney J. Risk factors for postpartum depression: an umbrella review. J Midwifery Womens Health. 2020;65(1):96-108. doi:10.1111/jmwh.13067.

[37] Hochman E, Feldman B, Weizman A, Krivoy A, Gur S, Barzilay E, et al. Development and validation of a machine learning-based postpartum depression prediction model: a nationwide cohort study. Depress Anxiety. 2021;38(4):400-411. doi:10.1002/da.23123.

[38] Moons KGM, Damen JAA, 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. 2025;388:e082505. doi:10.1136/bmj-2024-082505.

Tables

Table 1 Variable definitions and coding

Variable	Coding and processing
Age	Continuous; entered in years and standardised
Educational attainment	1=Junior secondary or below; 2=Senior secondary/vocational secondary; 3=Junior college/bachelor’s degree; 4=Postgraduate or above
Place of residence	1=Rural; 2=Urban
Monthly household income	1=≤3000 CNY; 2=3001–5000 CNY; 3=5001–8000 CNY; 4=>8000 CNY
Employment status	1=Employed; 2=Not employed
Living environment	1=Good; 2=Average; 3=Poor
Gravidity	1=1 pregnancy; 2=2 pregnancies; 3=3 pregnancies; 4=≥4 pregnancies
Mode of delivery	1=Vaginal; 2=Caesarean

Variable	Coding and processing
Postpartum haemorrhage	1=Yes; 2=No
Pregnancy complications	1=No; 2=Yes
Neonatal sex	1=Male; 2=Female
Birth weight	1=Low birth weight; 2=Normal birth weight; 3=Macrosomia
Gestational age category	1=Term; 2=Preterm; 3=Post-term
Neonatal asphyxia	1=Yes; 2=No
Birth defects	1=No; 2=Yes
Neonatal disease	1=No; 2=Yes
Life satisfaction (SWLS)	Sum of five original item scores; range 5–35; continuous
Resilience (CD-RISC)	Sum of 25 items recoded to 0–4; range 0–100; continuous
Perceived social support (PSSS)	Sum of 12 items scored 1–7; range 12–84; continuous
Outcome	Ten EPDS items scored 0–3; total ≥10 defines screen positivity

Note: Categorical variables were one-hot encoded and continuous variables standardised within training pipelines. One out-of-range item value occurred in each of the CD-RISC and PSSS. Total scores were prorated using each participant's remaining valid items.

Table 2 Participant characteristics by EPDS screening status

Variable	Total (N=3595)	Screen negative (n=2703)	Screen positive (n=892)	Test statistic	*P value*
Age	32.0 (29.0–34.0)	32.0 (29.0–34.0)	31.0 (29.0–34.0)	Z=0.644	0.520

Variable	Total (N=3595)	Screen negative (n=2703)	Screen positive (n=892)	Test statistic	*P value*
Educational attainment	—	—	—	χ^2=13.167	0.004
Junior secondary or below	170 (4.7)	119 (4.4)	51 (5.7)		
Senior secondary/vocational secondary	880 (24.5)	629 (23.3)	251 (28.1)		
Junior college/bachelor's degree	2056 (57.2)	1586 (58.7)	470 (52.7)		
Postgraduate or above	489 (13.6)	369 (13.7)	120 (13.5)		
Place of residence	—	—	—	χ^2=0.371	0.542
Rural	166 (4.6)	121 (4.5)	45 (5.0)		
Urban	3429 (95.4)	2582 (95.5)	847 (95.0)		
Monthly household income	—	—	—	χ^2=14.182	0.003
≤3000 CNY	163 (4.5)	120 (4.4)	43 (4.8)		
3001–5000 CNY	592 (16.5)	425 (15.7)	167 (18.7)		
5001–8000 CNY	1078 (30.0)	785 (29.0)	293 (32.8)		
>8000 CNY	1762 (49.0)	1373 (50.8)	389 (43.6)		
Employment status	—	—	—	χ^2=8.821	0.003
Employed	2852 (79.3)	2176 (80.5)	676 (75.8)		
Not employed	743 (20.7)	527 (19.5)	216 (24.2)		
Living environment	—	—	—	χ^2=28.030	<0.001
Good	2822 (78.5)	2178 (80.6)	644 (72.2)		
Average	765 (21.3)	520 (19.2)	245 (27.5)		
Poor	8 (0.2)	5 (0.2)	3 (0.3)		

Variable	Total (N=3595)	Screen negative (n=2703)	Screen positive (n=892)	Test statistic	*P value*
Gravidity	—	—	—	χ^2=6.913	0.075
1 pregnancy	316 (8.8)	250 (9.2)	66 (7.4)		
2 pregnancies	2024 (56.3)	1490 (55.1)	534 (59.9)		
3 pregnancies	834 (23.2)	639 (23.6)	195 (21.9)		
≥4 pregnancies	421 (11.7)	324 (12.0)	97 (10.9)		
Mode of delivery	—	—	—	χ^2=5.271	0.022
Vaginal	950 (26.4)	741 (27.4)	209 (23.4)		
Caesarean	2645 (73.6)	1962 (72.6)	683 (76.6)		
Postpartum haemorrhage	—	—	—	χ^2=3.970	0.046
Yes	1052 (29.3)	767 (28.4)	285 (32.0)		
No	2543 (70.7)	1936 (71.6)	607 (68.0)		
Pregnancy complications	—	—	—	χ^2=5.822	0.016
No	2613 (72.7)	1993 (73.7)	620 (69.5)		
Yes	982 (27.3)	710 (26.3)	272 (30.5)		
Neonatal sex	—	—	—	χ^2=0.535	0.465
Male	1709 (47.5)	1275 (47.2)	434 (48.7)		
Female	1886 (52.5)	1428 (52.8)	458 (51.3)		
Birth weight	—	—	—	χ^2=26.140	<0.001
Low birth weight	259 (7.2)	163 (6.0)	96 (10.8)		
Normal birth weight	3106 (86.4)	2377 (87.9)	729 (81.7)		
Macrosomia	230 (6.4)	163 (6.0)	67 (7.5)		

Variable	Total (N=3595)	Screen negative (n=2703)	Screen positive (n=892)	Test statistic	*P value*
Gestational age category	—	—	—	χ^2=19.614	<0.001
Term	3185 (88.6)	2430 (89.9)	755 (84.6)		
Preterm	357 (9.9)	241 (8.9)	116 (13.0)		
Post-term	53 (1.5)	32 (1.2)	21 (2.4)		
Neonatal asphyxia	—	—	—	χ^2=28.355	<0.001
Yes	117 (3.3)	63 (2.3)	54 (6.1)		
No	3478 (96.7)	2640 (97.7)	838 (93.9)		
Birth defects	—	—	—	χ^2=14.961	<0.001
No	3544 (98.6)	2677 (99.0)	867 (97.2)		
Yes	51 (1.4)	26 (1.0)	25 (2.8)		
Neonatal disease	—	—	—	χ^2=11.061	<0.001
No	3487 (97.0)	2637 (97.6)	850 (95.3)		
Yes	108 (3.0)	66 (2.4)	42 (4.7)		
Life satisfaction	30.0 (26.0–31.0)	30.0 (28.0–32.0)	25.0 (20.0–29.0)	Z=24.748	<0.001
Resilience	73.0 (61.0–84.0)	75.0 (67.0–88.0)	59.0 (50.0–70.0)	Z=24.822	<0.001
Perceived social support	72.0 (61.0–78.0)	72.0 (66.0–81.0)	61.0 (50.0–72.0)	Z=20.210	<0.001

Note: Continuous variables are presented as median (P25–P75) and compared using Mann–Whitney U tests, with Z statistics reported. Categorical variables are presented as n (%) and compared using Pearson's χ^2 tests, with Yates' continuity correction for 2×2 tables. Percentages use the number of participants in each column as the denominator. P values reflect overall variable-level tests. Living environment contains sparse categories; its asymptotic P value requires caution.

Table 3 Performance of predictor sets in the temporal validation set

Predictor set	ROC-AUC (95% CI)	PR-AUC	Brier score	ΔAUC vs basic clinical	ΔAUC vs psychosocial
Basic clinical	0.588 (0.551–0.625)	0.320	0.175	Reference	—
Psychosocial	0.838 (0.812–0.862)	0.628	0.128	0.250 (0.210–0.290)	Reference
Full	0.832 (0.807–0.857)	0.619	0.129	0.244 (0.209–0.279)	-0.006 (-0.016–0.004)

Note: All three predictor sets used EBM and were evaluated in the same temporal validation sample (n=1402; 325 screen-positive women). ΔAUC was calculated using unrounded predicted probabilities. Its 95% CI was estimated using 2000 paired bootstrap resamples.

Table 4 Performance of the four full models in the temporal validation set

Model	ROC-AUC (95% CI)	PR-AUC	Brier score (95% CI)	Intercept	Slope (95% CI)	Grouped ECE	Threshold
Elastic Net Logistic	0.837 (0.811–0.860)	0.623	0.130 (0.120–0.139)	0.220	1.417 (1.244–1.601)	0.054	0.157
EBM	0.832 (0.807–0.857)	0.619	0.129 (0.119–0.139)	-0.049	1.162 (1.019–1.319)	0.031	0.146
CatBoost	0.832 (0.808–0.856)	0.612	0.131 (0.121–0.140)	0.034	1.283 (1.132–1.454)	0.039	0.143
XGBoost	0.821 (0.796–0.846)	0.599	0.134 (0.124–0.143)	-0.013	1.210 (1.049–1.372)	0.035	0.157

Note: PR-AUC was calculated using average precision. Grouped ECE is the weighted absolute calibration error across 10 approximately equal-frequency groups defined by predicted probability. Calibration intercepts and slopes were jointly estimated. CIs were estimated using 1000 bootstrap resamples. Thresholds were locked in the development group, with full numerical precision used in calculations.

Table 5 EBM classification performance and review workload at different thresholds

Threshold strategy	Threshold probability	Sensitivity	Specificity	PPV/NPV	Reviews per 1000 women
Locked development threshold	0.146	91.4%	50.9%	36.0%/95.1%	589
Fixed threshold 5%	0.050	98.8%	17.5%	26.5%/97.9%	863
Fixed threshold 10%	0.100	95.1%	34.1%	30.3%/95.8%	727
Fixed threshold 20%	0.200	85.8%	64.0%	41.8%/93.7%	476

Note: The locked EBM threshold was 0.14586053281440056, displayed as 0.146. Fixed thresholds are included only for descriptive comparison. Reviews required per 1000 women=1000×(true positives+false positives)/1402, rounded to the nearest integer.

Figure legends

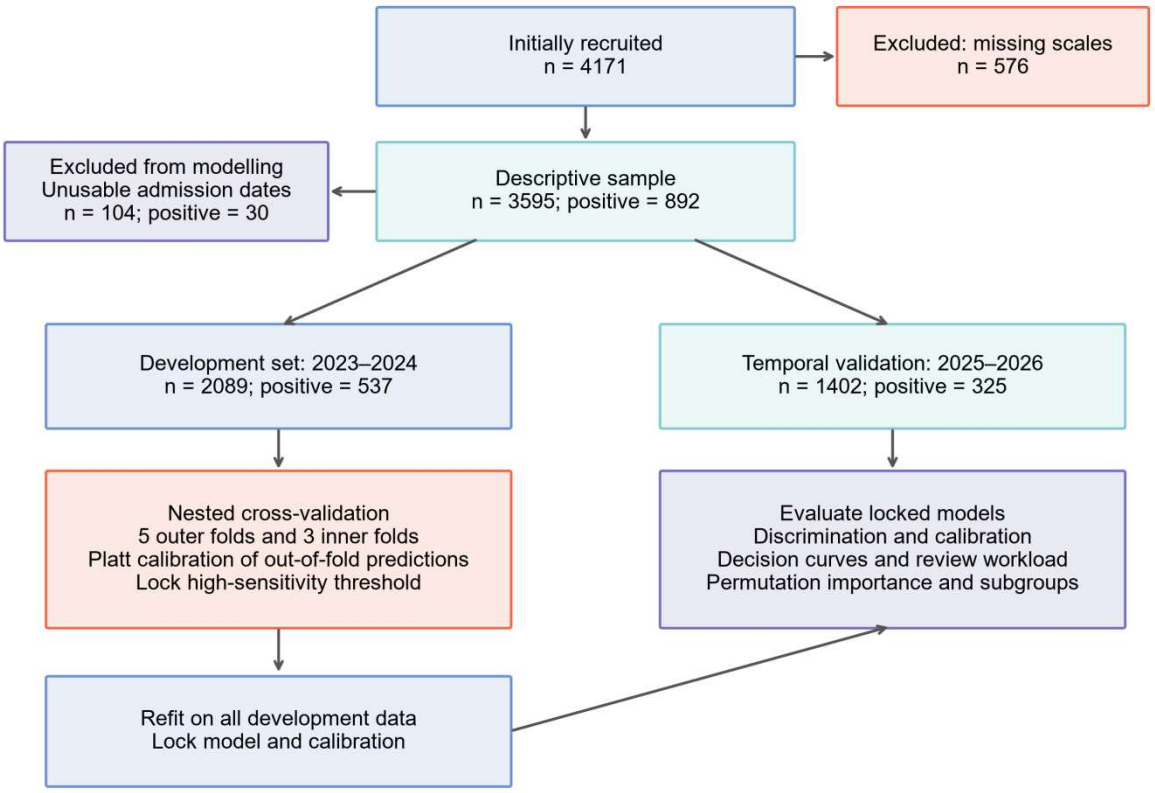


Figure 1 Participant flow and model development

Initially, 4171 women were recruited; 576 were excluded because of missing scales. The descriptive analysis included 3595 women. Of these, 104 were excluded from modelling because their admission dates could not be assigned to the modelling periods. Development and validation groups were separate subsets. Earlier participants were used for model fitting, calibration, and threshold selection; later participants were used for temporal validation.

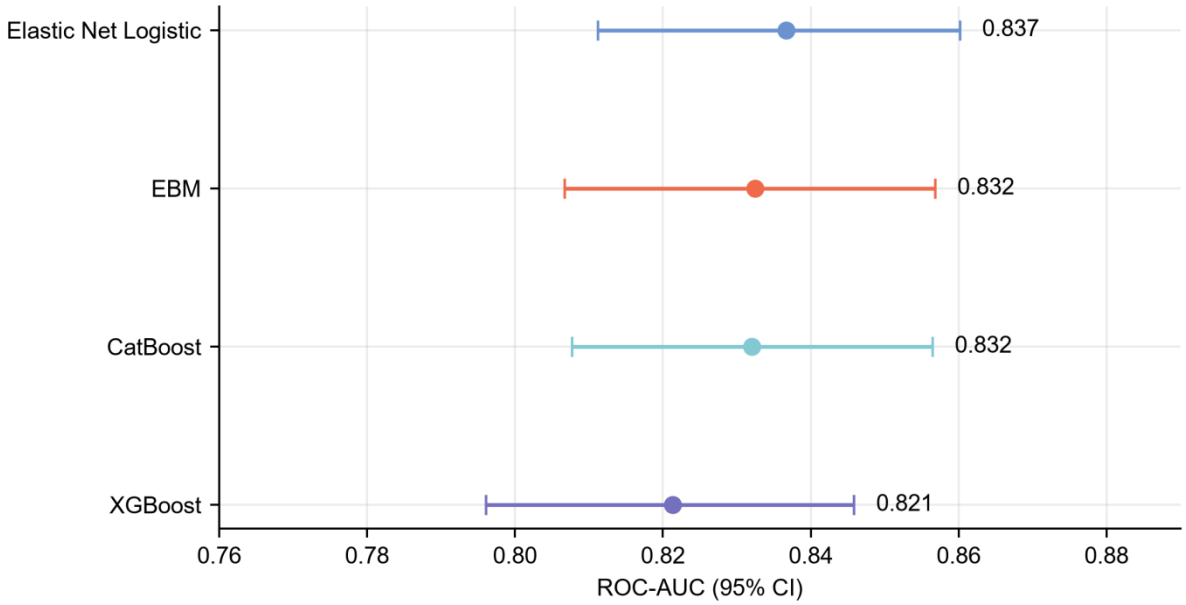


Figure 2 ROC-AUCs and 95% CIs of the four models in temporal validation

Points indicate ROC-AUC estimates; horizontal lines indicate percentile-based 95% CIs from 1000 bootstrap resamples. All models used the full set of 19 predictors and were evaluated in the same 1402 participants.

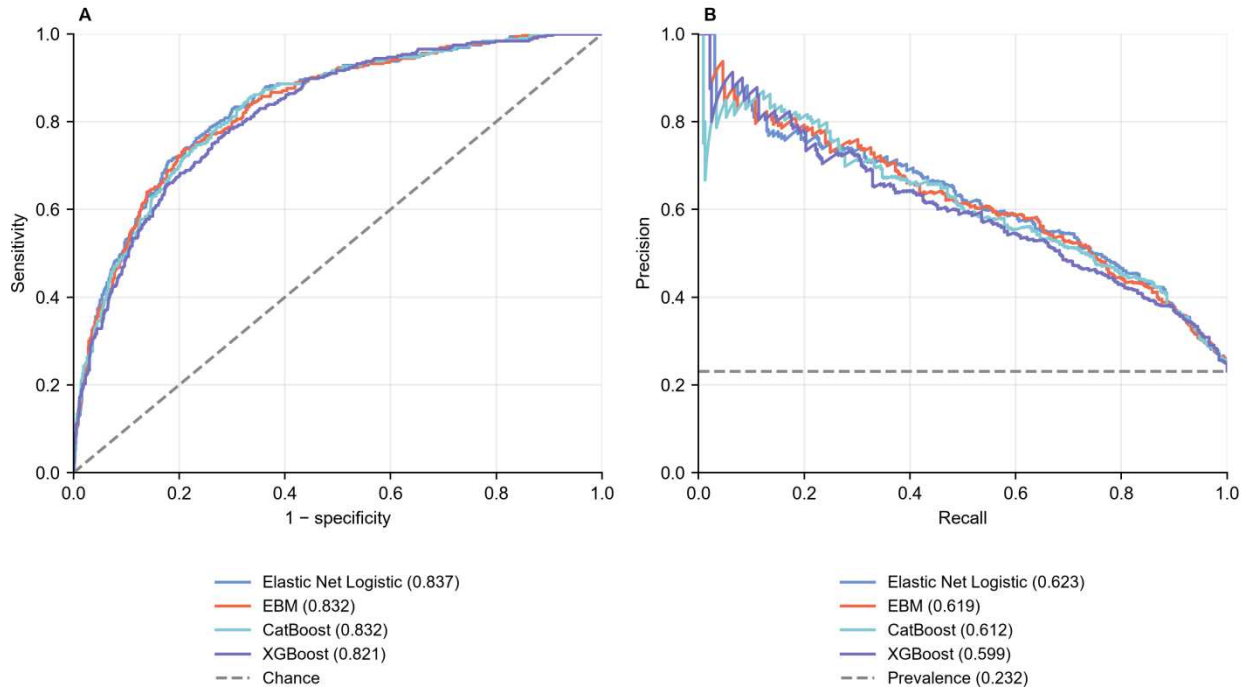


Figure 3 ROC and precision–recall curves of the four models in temporal validation

Panel A shows ROC curves and panel B shows precision–recall curves. Legends report ROC-AUC and average precision, respectively. The diagonal indicates chance-level discrimination. The horizontal reference line in panel B represents the validation group's positive screening proportion (325/1402=0.232).

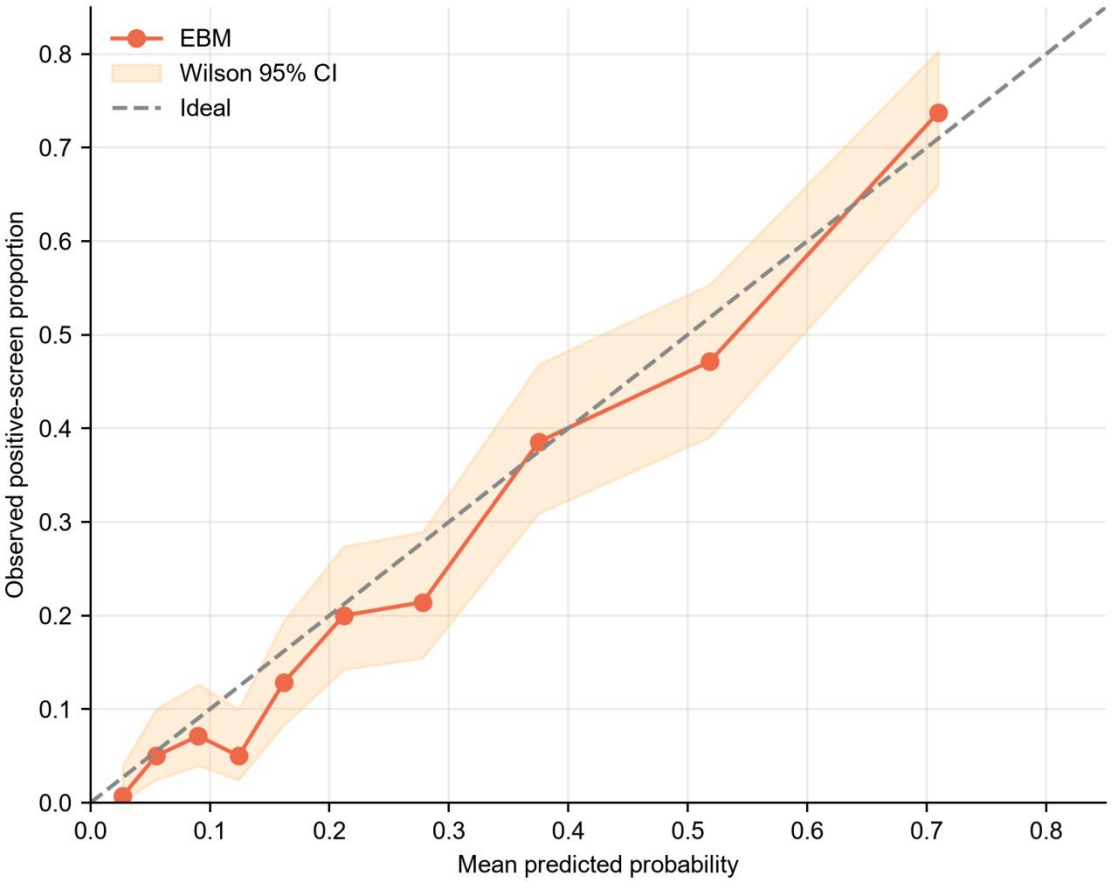


Figure 4 Calibration of the full EBM model in temporal validation

Participants were divided into 10 approximately equal-sized groups by predicted probability. Points represent group mean predicted probabilities and observed EPDS-positive proportions. Shading connects groupwise Wilson 95% CIs and is not a simultaneous confidence band for a smoothed calibration function. The dashed diagonal indicates ideal calibration.

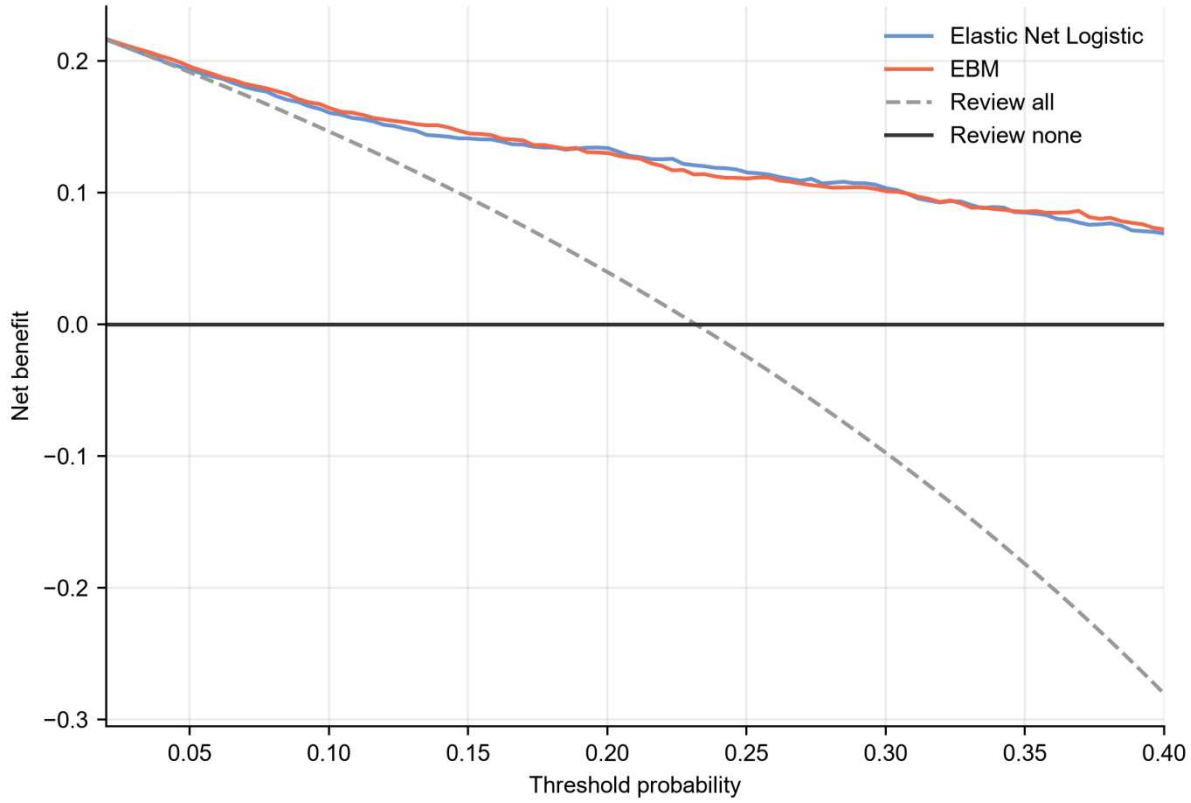


Figure 5 Decision curves for Elastic Net Logistic and EBM

Net benefit relative to reviewing all or none of the women was evaluated using EPDS screening status in temporal validation as the reference outcome. Threshold probabilities of 0.02–0.40 are displayed. These curves do not establish benefit for clinically diagnosed depression or superiority over direct EPDS screening.

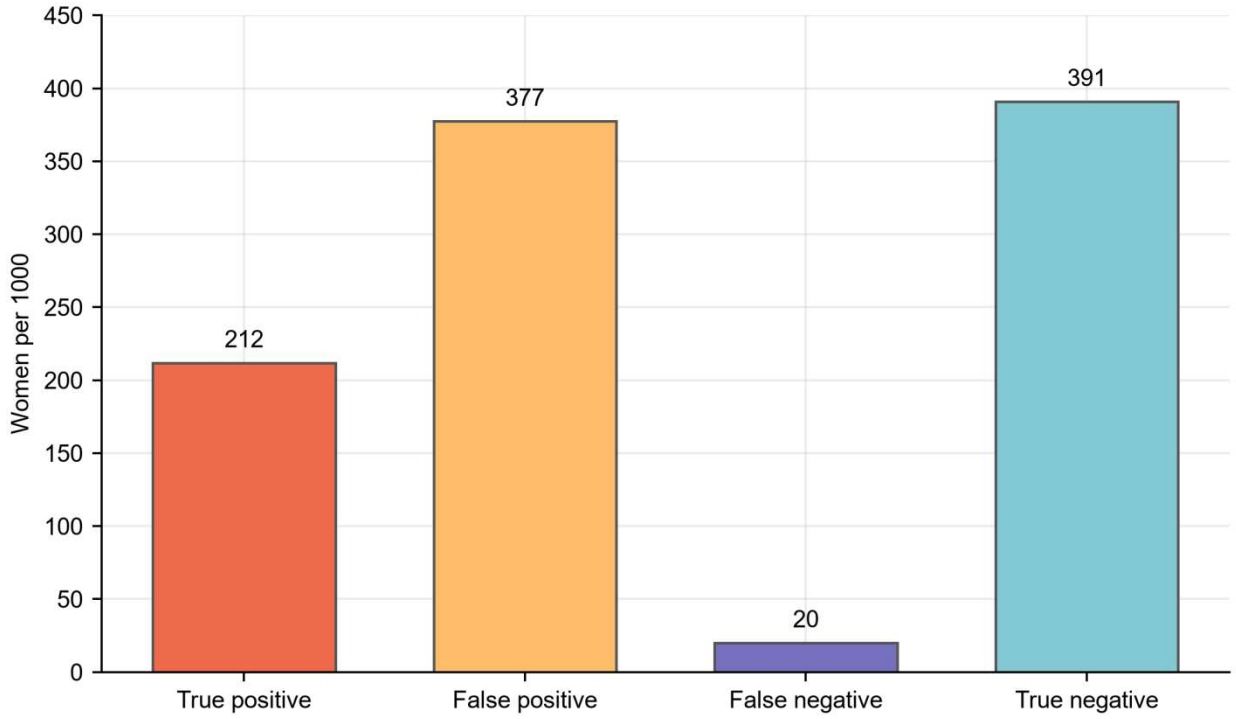


Figure 6 Classification and review workload per 1000 women at the locked threshold

Counts were proportionally extrapolated from the EBM validation confusion matrix at the locked threshold. Among 1402 women, there were 297 true positives, 529 false positives, 28 false negatives, and 548 true negatives. Rounded counts per 1000 women were 212, 377, 20, and 391, respectively; 589 women would be flagged for review.

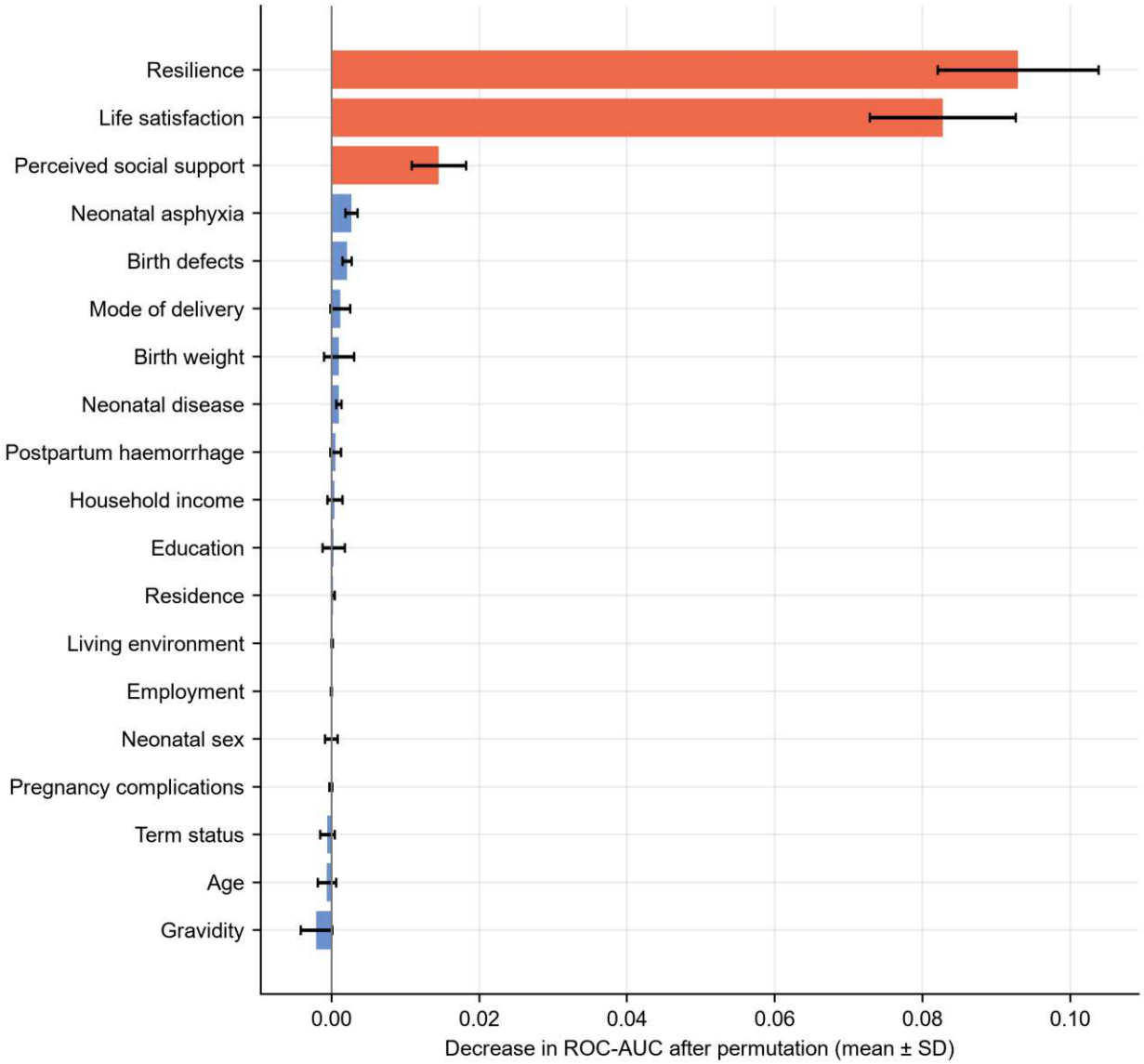


Figure 7 Global EBM permutation importance and variability across repetitions

Bars indicate mean ROC-AUC reductions across 50 permutation repetitions in the temporal validation sample; error bars indicate SDs. Negative values indicate a slight increase in AUC after permutation. Importance describes model reliance on variables and does not establish independent associations or causal effects.