

Development and Validation of a Nomogram for Predicting the Electrophysiological Response to Postpartum Pelvic Floor Rehabilitation: A Machine Learning–Enhanced Study

JiaPing Lu

Jinhua People's Hospital

Heng Xu

xheng0248@sina.com

Jinhua People's Hospital

Qiulian Xu

Jinhua People's Hospital

Research Article

Keywords: postpartum, pelvic floor dysfunction, surface electromyography, electrophysiological response, prediction model, nomogram, machine learning, TRIPOD

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11199738/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 Validation of a Nomogram for Predicting the Electrophysiological Response to Postpartum Pelvic Floor Rehabilitation: A Machine Learning–Enhanced Study

Authors:JiaPing Lu¹, Heng Xu^{1*}, Qiulian Xu¹

1Department of Obstetric, Jinhua People's Hospital, Jinhua, China

Corresponding Author:

Heng Xu^{1*}

Department of Obstetric, Jinhua People's Hospital, Jinhua, China

Tel:8618757829616

Email: xheng0248@sina.com

Development and Validation of a Nomogram for Predicting the Electrophysiological Response to Postpartum Pelvic Floor Rehabilitation: A Machine Learning–Enhanced Study

Abstract

Background: Pelvic floor dysfunction (PFD) is highly prevalent after childbirth, and pelvic floor rehabilitation is first-line conservative management; however, individual treatment response varies substantially and no validated pretreatment tool exists to predict the electrophysiological response.

Objective: To develop and validate a nomogram predicting the electrophysiological response to rehabilitation from pretreatment characteristics, and compared with machine-learning (ML) models.

Methods: In this single-center retrospective cohort, 1,007 postpartum women with paired pre- and post-treatment pelvic floor surface electromyography (sEMG; Glazer protocol) were analyzed. The primary outcome was a $\geq 20\%$ relative improvement in fast-twitch maximum amplitude. Twenty-five candidate predictors spanning demographics, obstetric factors, baseline sEMG, and clinical features were screened by univariable analysis and LASSO, then entered into multivariable logistic regression. Validation comprised bootstrap optimism correction, a random 30% hold-out set, and temporal validation (2023–2024 $\rightarrow$ 2025–2026). Random forest, XGBoost, SVM, and a full-variable logistic model served as benchmarks. Mathematical coupling between the relative-change outcome and its baseline was examined using correlation analysis, the Olds test, and a parallel model for absolute improvement ($\geq 5 \mu\text{V}$).

Results: The final nomogram retained age (OR 1.038 per year, 95% CI 0.996–1.082) and baseline fast-twitch maximum (OR 0.931 per μV , 95% CI 0.918–0.944, $p < 0.001$). Bootstrap-corrected C-index was 0.754; AUC was 0.717 (random hold-out) and 0.742 (temporal validation), with adequate calibration (Hosmer–Lemeshow $p = 0.194$ and 0.318) and net benefit across threshold probabilities of 0.2–0.8. No benchmark outperformed the nomogram (random forest 0.720; XGBoost 0.713; SVM 0.693; full-variable logistic 0.697). The baseline–response association persisted for the absolute-improvement outcome (OR 0.954 per μV , $p < 0.001$), indicating the signal is not purely an artifact of the relative-change definition.

Conclusions: A two-variable nomogram predicts the electrophysiological response to postpartum pelvic floor rehabilitation with moderate, temporally robust discrimination, performing on par with or better than complex models. It may support pretreatment counseling and individualized planning as an adjunct to clinical judgment.

Keywords: postpartum; pelvic floor dysfunction; surface electromyography; electrophysiological response; prediction model; nomogram; machine learning; TRIPOD

1. Introduction

Pelvic floor dysfunction (PFD)—encompassing urinary incontinence, pelvic organ prolapse, fecal incontinence, and sexual dysfunction—affects a substantial proportion of adult women worldwide. In the United States, the prevalence of at least one symptomatic pelvic floor disorder approaches 24% and rises steeply with age and parity¹; comparable figures have been reported in European and Asian populations^{1,2}. Although rarely life-threatening, PFD markedly impairs quality of life, social participation, and psychological well-being, and imposes a growing surgical and economic burden^{3,4}.

Pregnancy and vaginal childbirth are the most consistently identified modifiable risk factors for PFD. Delivery-related injury to the levator ani complex and pudendal innervation leads to measurable declines in pelvic floor muscle strength within weeks of childbirth⁵, and vaginal delivery confers a long-term excess risk of urinary incontinence compared with cesarean section⁶. In China, cross-sectional data indicate that pelvic floor muscle strength is already significantly reduced one month postpartum, with obstetric factors such as forceps delivery, prolonged second stage, and macrosomia further aggravating the deficit⁶.

Pelvic floor muscle training (PFMT) and instrument-assisted rehabilitation (electrical stimulation, magnetic stimulation, and radiofrequency therapy) are established first-line conservative treatments: women with stress urinary incontinence receiving PFMT are up to eight times more likely to report cure than untreated controls^{7,8}. Yet response is far from uniform—in an assessor-blinded randomized trial of postpartum PFMT, more than one-third of trained women still reported urinary incontinence at six months⁹. This heterogeneity raises a pragmatic clinical question that remains unanswered: *at the first postpartum assessment, can we predict which women will show a measurable electrophysiological response to a standard rehabilitation course?*

Surface electromyography (sEMG) of the pelvic floor, performed using the Glazer protocol, provides an objective, reproducible quantification of fast-twitch contraction, tonic contraction, and endurance, with demonstrated test-retest reliability and clinical predictive validity^{10,11}. Baseline sEMG indices carry prognostic information in stress urinary incontinence¹², and recent artificial-intelligence work has confirmed that sEMG features can support high-precision PFD phenotyping¹³. Nevertheless, existing prediction studies in this field are small, rarely validated, and—to our knowledge—none has produced a clinically applicable tool that integrates baseline electrophysiology with routinely collected obstetric factors.

We therefore developed and validated a nomogram predicting $\geq 20\%$ improvement in fast-twitch maximum sEMG amplitude after a standard postpartum rehabilitation course, using pre-treatment characteristics from 1,007 women, in accordance with the TRIPOD reporting guideline¹⁴. Because complex algorithms are increasingly proposed for clinical prediction without clear evidence of superiority¹⁵, we explicitly compared the nomogram with random forest, XGBoost, support vector machine, and full-variable logistic regression models, subjected the nomogram to bootstrap, hold-out, and temporal validation, and formally interrogated the mathematical coupling inherent in a relative-change outcome.

The contributions of this study are threefold: (1) the first large-sample ($n > 1,000$) nomogram for the electrophysiological response to postpartum pelvic floor rehabilitation, requiring only two routinely available pretreatment variables; (2) a three-tier validation strategy including temporal validation that simulates prospective deployment; and (3) a machine-learning benchmark with SHAP interpretability plus dedicated analyses of mathematical coupling, demonstrating that the parsimonious model achieves performance equivalent to complex algorithms and that its core signal is not an artifact of the outcome definition.

2. Methods

2.1. Study Design and Participants

This single-center retrospective cohort study analyzed women who attended the pelvic floor rehabilitation clinic of Jinhua People's Hospital between January 2023 and January 2026. The study was conducted and reported following the TRIPOD statement¹⁵; methodological risk of bias was considered with reference to PROBAST¹⁶. The protocol was approved by the institutional ethics committee of Jinhua People's Hospital, with a waiver of individual informed consent for de-identified retrospective data.

Of 1,810 women with an initial pelvic floor assessment, we included those with paired pre- and post-treatment fast-twitch maximum sEMG measurements, yielding an analytical cohort of 1,007 women (Figure 1). Women without a post-treatment sEMG evaluation ($n = 803$) were excluded; they did not differ materially from included women in age or baseline sEMG profile (Supplementary Table S1).

2.2. Pelvic Floor Assessment and Rehabilitation

Pelvic floor sEMG was recorded with an intravaginal probe following the Glazer protocol¹¹, yielding pre-contraction resting tone (pre-rest), fast-twitch maximum amplitude (phasic, type-II-dominant), mean tonic contraction amplitude (type-I-dominant), mean endurance contraction amplitude, post-contraction resting tone (post-rest), and the composite Glazer score. Manual muscle testing of type-I and type-II fibers (Modified Oxford Scale, grades 0–5) was performed by trained therapists; digital assessment correlates significantly with sEMG¹². Rehabilitation comprised individualized courses of electrical stimulation, magnetic stimulation, combined magneto-electric therapy, or radiofrequency therapy per institutional routine, with post-treatment sEMG re-evaluation at course completion. Treatment modalities were pooled because the study objective was to predict response to the rehabilitation pathway as delivered in routine practice rather than to any single technique; the resulting heterogeneity is addressed in Section 4.6.

2.3. Outcome Definitions

The primary outcome was a binary indicator of electrophysiological response, defined as a relative gain of at least 20% in fast-twitch maximum amplitude:

$$Y_i = \mathbb{1} \left[\frac{\text{FastMax}_{i,\text{post}} - \text{FastMax}_{i,\text{pre}}}{\text{FastMax}_{i,\text{pre}}} \times 100\% \geq 20\% \right],$$

where $\mathbb{1}[\cdot]$ is the indicator function. A 20% threshold was chosen a priori as a pragmatic marker of response clearly exceeding the test–retest variability reported for Glazer-protocol sEMG¹¹. Secondary outcomes applied the same definition to tonic ($\geq 20\%$) and endurance ($\geq 20\%$) contraction means.

Because a relative-change outcome is mathematically coupled to its baseline, we additionally defined an **absolute-improvement outcome** as $\text{FastMax}_{\text{post}} - \text{FastMax}_{\text{pre}} \geq 5 \mu\text{V}$ (a threshold likewise exceeding the protocol’s typical test–retest variability¹¹) and analyzed it with the identical modeling pipeline as a prespecified sensitivity analysis (Section 2.7).

2.4. Candidate Predictors and Data Preprocessing

Twenty-five candidate predictors were assembled a priori from four domains (full list in Table 2): demographics (age, body mass index [BMI], gravidity, parity), obstetric factors (mode of delivery, birth weight, forceps delivery, labor analgesia), baseline electrophysiology (Glazer score, pre-rest tone, fast-twitch maximum, tonic and endurance means, post-rest tone) and clinical features (manual type-I/II strength, pain, pelvic organ prolapse [POP], stress urinary incontinence [SUI], dyspareunia, days from delivery to assessment, gestational diabetes [GDM], hypertensive disorders of pregnancy [HDP], amniotic fluid abnormality, and premature rupture of membranes [PROM]). Diagnoses followed IUGA/ICS terminology^{4,5}.

Data cleaning included: extraction of numeric age from mixed strings; setting implausible BMI values ($>60 \text{ kg/m}^2$, $n = 3$) and birth weights $>6,000 \text{ g}$ to missing, these being subsequently handled by imputation; conversion of birth weight to grams; and harmonization of categorical coding. Per the prespecified missing-data rules, variables with $>20\%$ missingness were excluded from the primary modeling pool, a threshold commonly adopted in prediction-model development¹⁷ (perineal injury, 24.3%; labor-stage durations, 28–30%, largely structural in cesarean deliveries; scarred uterus, 62.1%), and variables with $<20\%$ missingness were imputed by chained-equation iterative imputation with posterior draws (Bayesian ridge conditional models), with binary indicators re-rounded to 0/1¹⁸. The most incomplete retained predictor was forceps delivery (19.1% missing); all others were $\leq 3.2\%$ incomplete. Predictors with event prevalence $<2\%$ (precipitate labor, constipation, urgency urinary incontinence) were omitted to avoid separation. To assess whether excluding the high-missingness obstetric variables discarded predictive information, we performed an **extended-imputation sensitivity analysis** in which perineal injury, total labor duration, and second-stage duration were imputed and added to the candidate pool (Section 2.7).

2.5. Model Development

The cohort was randomly split 7:3 into training ($n = 704$) and validation ($n = 303$) sets, stratified by outcome. On the training set, candidate predictors were first screened by

univariable logistic regression (retention at $p < 0.10$; age, BMI, and mode of delivery were additionally designated for retention on clinical grounds regardless of their screening p -value), then subjected to least absolute shrinkage and selection operator (LASSO) logistic regression with 10-fold cross-validation at the $\lambda_{\min}$ penalty¹⁹. LASSO-retained variables entered a multivariable logistic regression with backward elimination (likelihood-ratio $p > 0.10$ for removal) to further simplify the LASSO-selected set while monitoring the Akaike information criterion, requiring variance inflation factors < 5 . The final model has the form

$$\hat{p}_i = \frac{1}{1 + \exp[-(\beta_0 + \sum_j \beta_j x_{ij})]},$$

and was translated into a points-based nomogram on a 0–100 scale²⁰.

2.6. Model Validation and Machine-Learning Benchmarking

Three validation tiers were prespecified¹⁸: (i) internal—200 bootstrap resamples of the training set to estimate the optimism-corrected C-index; (ii) random hold-out—direct application of the locked model to the 30% validation set; and (iii) temporal—the model refitted on 2023–2024 data ($n = 709$) and applied to 2025–2026 data ($n = 298$), simulating prospective use. Discrimination was quantified by C-index/AUC; calibration by grouped and smoothed calibration curves and the Hosmer–Lemeshow test²¹; and clinical utility by decision curve analysis (DCA)^{22,23}, with net benefit at threshold probability p_t defined as

$$\text{NB}(p_t) = \frac{\text{TP}}{n} - \frac{\text{FP}}{n} \cdot \frac{p_t}{1 - p_t},$$

where TP and FP are true- and false-positive counts and n the sample size.

Benchmark models—random forest (500 trees; max_features tuned over 2–8)²⁴, XGBoost (max_depth 3–6; learning rate 0.01–0.3; subsample 0.8)²⁵, and radial-basis SVM (cost 0.1–10)²⁶—were trained on all 25 predictors with 5-fold cross-validated grid search. To address the fairness of comparing a two-variable logistic model against ML models trained on 25 predictors, we additionally fitted an **unpenalized logistic regression using all 25 predictors** (full-variable logistic) and an L2-regularized counterpart as further benchmarks. All models were compared on the validation set by AUC, Brier score, sensitivity, specificity, and predictive values. XGBoost predictions were explained with SHAP values²⁷.

2.7. Subgroup, Sensitivity, and Coupling Analyses

Subgroup analyses refit the final model within strata of delivery mode, age (<35 vs ≥ 35 years), parity (primiparous vs multiparous), and baseline Glazer score (<60 vs ≥ 60), with formal interaction tests (predictor $\times$ stratum). Sensitivity analyses comprised: complete-case analysis; outcome thresholds of 10% and 30%; exclusion of assessments >180 days after delivery; exclusion of forceps deliveries; restriction to women with paired Glazer scores; the extended-imputation analysis (Section 2.4); and the absolute-improvement outcome (Section 2.3).

Mathematical coupling between baseline and change was quantified by (i) Spearman correlations of baseline fast-twitch maximum with relative and with absolute change; and (ii) the Olds (change-on-baseline) regression of absolute change on baseline²⁸, in which a slope significantly greater than -1 indicates partial rather than complete regression to the mean.

Analyses used Python 3.12 (statsmodels 0.14, scikit-learn 1.7, XGBoost 3.4, SHAP 0.52). Two-sided $p < 0.05$ denoted statistical significance unless otherwise specified; standardized mean differences (SMD) < 0.1 indicated baseline balance.

3. Results

3.1. Cohort Characteristics

The analytical cohort comprised 1,007 women (median age 31 years [IQR 28–33]; median interval from delivery to assessment 50 days [IQR 44–62]). SUI was the most common diagnosis (53.2%), followed by POP (17.8%); 28.4% had undergone cesarean delivery and 2.9% forceps-assisted delivery. The primary outcome occurred in 445 women (44.2%). Training and validation sets were well balanced: only birth weight (SMD = 0.142), age (0.101), and baseline post-rest tone (0.125) marginally exceeded the 0.1 threshold (Table 1). The study flow is summarized in Figure 1.

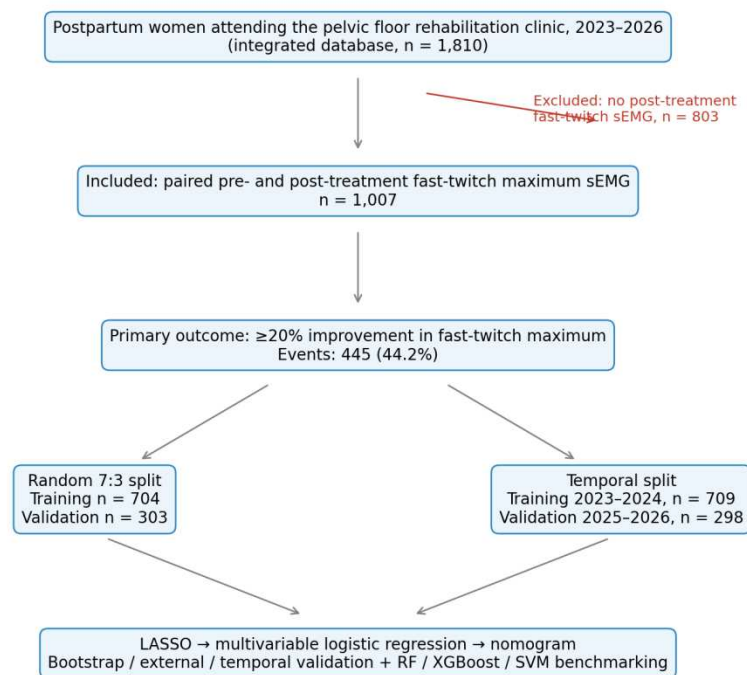


Figure 1. Study flow diagram. Of 1,810 women with an initial postpartum pelvic floor assessment, 1,007 with paired pre- and post-treatment fast-twitch maximum sEMG were analyzed. The cohort was split randomly (7:3) and temporally (2023–2024 vs 2025–2026) for model development and validation.

Table 1. Baseline characteristics of the training and validation sets. Values are median (IQR) or n (%).

Characteristic	Training set (n = 704)	Validation set (n = 303)	SMD
Age, years	31.0 (28.0, 33.0)	30.0 (28.0, 33.0)	0.101
BMI, kg/m ²	22.8 (20.8, 24.9)	22.7 (20.8, 25.0)	0.014
Parity	1.0 (1.0, 2.0)	1.0 (1.0, 2.0)	0.051
Gravidity	1.0 (1.0, 2.0)	1.0 (1.0, 2.0)	0.040
Cesarean delivery	206 (29.3)	80 (26.4)	0.064
Birth weight, g	3,350 (3,100, 3,600)	3,300 (3,000, 3,585)	0.142
Forceps delivery	21 (3.0)	8 (2.6)	0.021
Labor analgesia	482 (68.5)	218 (71.9)	0.076
Baseline Glazer score	63.4 (50.8, 72.9)	61.9 (50.6, 72.3)	0.077
Baseline pre-rest tone, μ V	4.0 (2.1, 6.4)	4.0 (2.3, 6.5)	0.087
Baseline fast-twitch maximum, μ V	31.1 (21.1, 41.5)	30.3 (20.3, 41.5)	0.045
Baseline tonic mean, μ V	20.0 (12.8, 28.6)	19.7 (12.4, 28.7)	0.031
Baseline endurance mean, μ V	16.6 (10.8, 24.5)	15.7 (10.5, 24.0)	0.048
Baseline post-rest tone, μ V	3.3 (1.6, 5.6)	3.7 (2.1, 5.9)	0.125
Manual type-I strength (grade)	2.0 (1.0, 2.0)	2.0 (1.0, 2.0)	0.008
Manual type-II strength (grade)	2.0 (2.0, 3.0)	2.0 (2.0, 3.0)	0.013
Pelvic pain	44 (6.2)	20 (6.6)	0.014
POP	132 (18.8)	47 (15.5)	0.086
SUI	370 (52.6)	164 (54.1)	0.031
Dyspareunia	42 (6.0)	19 (6.3)	0.013
Days from delivery to assessment	49.5 (44.0, 63.0)	50.0 (43.5, 61.5)	0.015
GDM	100 (14.2)	37 (12.2)	0.059
HDP	21 (3.0)	9 (3.0)	0.001
Amniotic fluid abnormality	70 (9.9)	32 (10.6)	0.020
PROM	107 (15.2)	50 (16.5)	0.036

SMD, standardized mean difference; POP, pelvic organ prolapse; SUI, stress urinary incontinence; GDM, gestational diabetes mellitus; HDP, hypertensive disorders of pregnancy; PROM, premature rupture of membranes.

3.2. Predictor Screening and LASSO Selection

Univariable screening retained 12 predictors at $p < 0.10$, including age and mode of delivery (Table 2); with BMI forced in on clinical grounds despite $p > 0.10$, 13 variables

entered LASSO. At $\lambda_{\min}$ (10-fold cross-validated AUC 0.755), LASSO retained six variables—age, BMI, forceps delivery, baseline fast-twitch maximum, baseline post-rest tone, and manual type-II strength (Figure 2). As anticipated, LASSO substantially shrank the mutually correlated baseline sEMG indices (Glazer score, tonic and endurance means), retaining fast-twitch maximum as their representative.

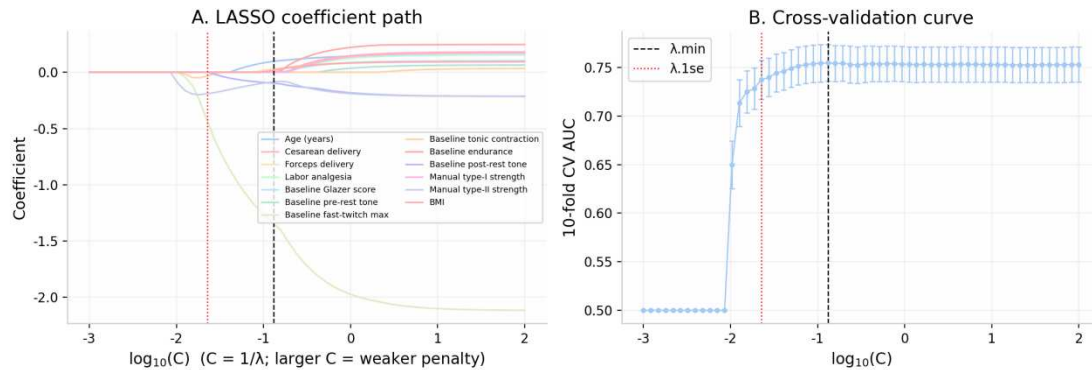


Figure 2. LASSO predictor selection on the training set. (A) Coefficient paths against $\log_{10}(C)$, where $C = 1/\lambda$ and larger C denotes weaker penalization; the dashed and dotted vertical lines mark $\lambda_{\min}$ and λ_{1se} . (B) Ten-fold cross-validated AUC with standard-error bars; $\lambda_{\min}$ achieved a cross-validated AUC of 0.755.

3.3. Final Model and Nomogram

Backward elimination of the LASSO-retained set converged on a two-predictor model (both variance inflation factors = 1.0; McFadden pseudo- $R^2 = 0.151$): each additional year of age increased the odds of $\geq 20\%$ improvement by 3.8% (OR 1.038, 95% CI 0.996–1.082, $p = 0.076$), and each 1- μV higher baseline fast-twitch maximum decreased the odds by 6.9% (OR 0.931, 95% CI 0.918–0.944, $p < 0.001$) (Table 2). Age was retained as prespecified on clinical grounds—having been forced into the candidate pool a priori—because its removal did not improve model fit (Akaike information criterion 826.7 with age versus 827.9 without). The complete fitted equation is $\text{logit}(p) = 0.8243 + 0.0375 \times \text{age} - 0.0715 \times \text{FastMax}_0$, with age in years and FastMax_0 (baseline fast-twitch maximum) in μV . The corresponding nomogram is shown in Figure 3: for example, a 30-year-old woman with a baseline fast-twitch maximum of 15 μV has a predicted response probability of approximately 0.70, whereas the same woman with a baseline of 60 μV has a predicted probability below 0.35.

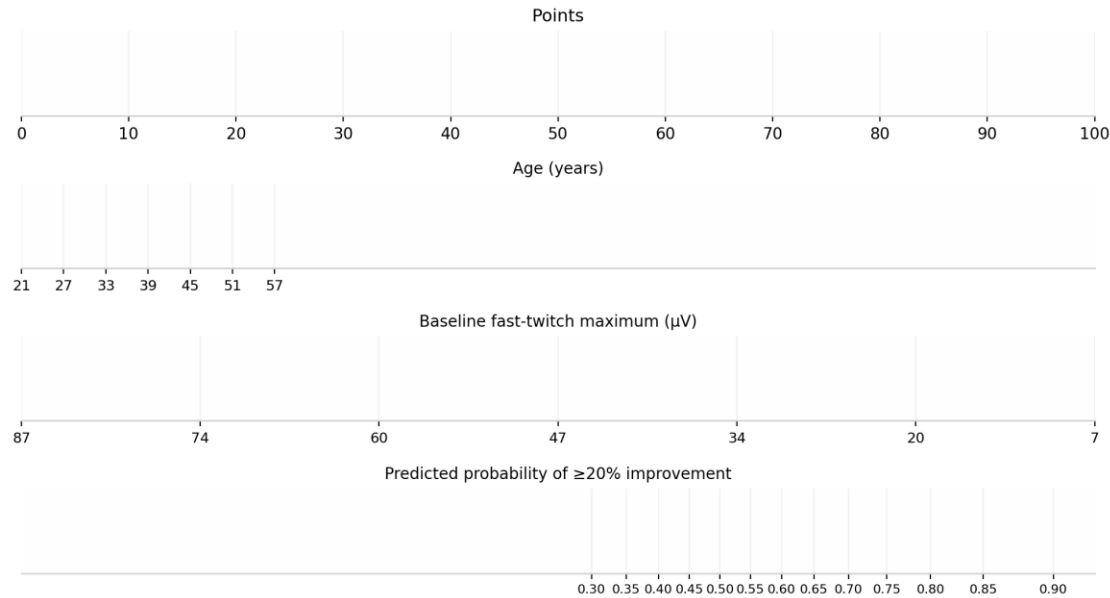


Figure 3. Nomogram for predicting $\geq 20\%$ improvement in fast-twitch maximum amplitude after postpartum pelvic floor rehabilitation. Each predictor value is projected vertically to the Points axis; the sum of points is projected to the bottom axis to read the predicted response probability. The fast-twitch axis is scaled to the 1st–99th percentile of the training distribution.

Table 2. Univariable and multivariable logistic regression for $\geq 20\%$ improvement in fast-twitch maximum (training set, $n = 704$).

Variable	Univariable OR (95% CI)	p	Multivariable OR (95% CI)	p
Age, per year	1.046 (1.007–1.085)	0.019	1.038 (0.996–1.082)	0.076
BMI, per kg/m^2	0.995 (0.952–1.039)	0.806	—	—
Parity, per birth	0.831 (0.649–1.064)	0.142	—	—
Gravidity, per pregnancy	1.054 (0.910–1.221)	0.482	—	—
Cesarean delivery	0.654 (0.469–0.913)	0.013	—	—
Birth weight, per g	1.000 (1.000–1.000)	0.728	—	—
Forceps delivery	2.599 (1.036–6.521)	0.042	—	—
Labor analgesia	1.503 (1.086–2.081)	0.014	—	—

Variable	Univariable OR (95% CI)	p	Multivariable OR (95% CI)	p
Baseline Glazer score, per point	0.961 (0.950–0.972)	<0.001	—	—
Baseline pre-rest tone, per μ V	0.931 (0.892–0.972)	0.001	—	—
Baseline fast-twitch maximum, per μ V	0.931 (0.918–0.944)	<0.001	0.931 (0.918–0.944)	<0.001
Baseline tonic mean, per μ V	0.930 (0.914–0.946)	<0.001	—	—
Baseline endurance mean, per μ V	0.936 (0.920–0.953)	<0.001	—	—
Baseline post-rest tone, per μ V	0.879 (0.833–0.927)	<0.001	—	—
Manual type-I strength, per grade	0.430 (0.337–0.549)	<0.001	—	—
Manual type-II strength, per grade	0.417 (0.338–0.514)	<0.001	—	—
Pelvic pain	0.784 (0.419–1.466)	0.446	—	—
POP	1.286 (0.880–1.879)	0.194	—	—
SUI	1.013 (0.752–1.364)	0.934	—	—
Dyspareunia	0.766 (0.403–1.454)	0.414	—	—
Days from delivery, per day	0.999 (0.993–1.006)	0.815	—	—
GDM	1.197 (0.783–1.829)	0.406	—	—
HDP	2.099 (0.859–5.131)	0.104	—	—
Amniotic fluid abnormality	0.827 (0.500–1.368)	0.459	—	—
PROM	1.290 (0.854–1.947)	0.226	—	—

OR, odds ratio; CI, confidence interval. Multivariable model retained by LASSO followed by backward elimination ($p > 0.10$ for removal); final-model estimates in bold.

3.4. Model Performance and Validation

The apparent C-index on the training set was 0.756; bootstrap validation estimated an optimism of only 0.003, yielding a corrected C-index of 0.754 (Table 3). In the random hold-out validation set, the locked model achieved an AUC of 0.717 with good calibration (Hosmer–Lemeshow $p = 0.194$; Figure 4B). Under temporal validation—refitting on 2023–2024 and testing on 2025–2026—the model achieved an AUC of 0.742 (Brier score 0.205; Hosmer–Lemeshow $p = 0.318$), indicating robust transportability across calendar periods. Decision curve analysis demonstrated net benefit over the treat-all and treat-none strategies across threshold probabilities of approximately 0.2–0.8 in both the training and validation sets (Figure 4C, D).

Table 3. Discrimination and calibration of the nomogram across validation tiers.

Validation tier	n	C-index / AUC	Brier score	Hosmer– Lemeshow
Training set (apparent)	704	0.756	—	—
Training set (bootstrap-corrected, 200 resamples)	704	0.754	—	—
Random hold-out validation set	303	0.717	0.214	$\chi^2 = 11.14$, $p = 0.194$
Temporal validation set (2025–2026)	298	0.742	0.205	$\chi^2 = 9.30$, $p = 0.318$

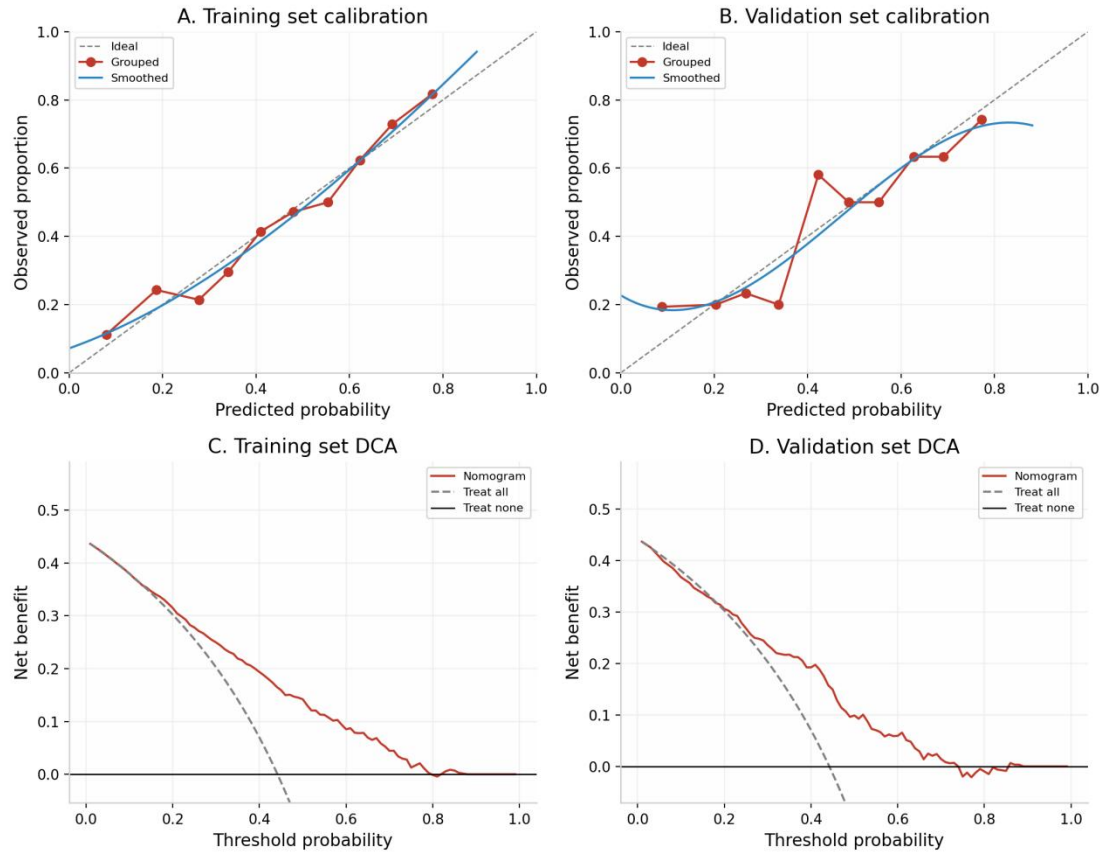


Figure 4. Calibration and clinical utility of the nomogram. (A, B) Calibration curves in the training and validation sets: grouped deciles (red), smoothed calibration (blue), and the ideal diagonal (dashed). (C, D) Decision curve analysis in the training and validation sets: the nomogram (red) provides higher net benefit than treat-all (dashed gray) and treat-none (black) strategies across threshold probabilities of approximately 0.2–0.8.

3.5. Machine-Learning and Full-Variable Benchmarks

On the validation set, random forest (AUC 0.720), XGBoost (0.713), and SVM (0.693) did not materially outperform the two-variable nomogram (0.717), with AUC differences ranging from 0.003 (random forest) to 0.024 (SVM) (Table 4, Figure 5). Critically, the unpenalized full-variable logistic regression using all 25 predictors achieved an apparent training AUC of 0.777 but only 0.697 on the validation set—evidence of overfitting and a performance *below* the parsimonious nomogram; an L2-regularized full-variable logistic model likewise reached only 0.697. Thus, the equivalence of the nomogram to ML models is not an artifact of comparing a reduced logistic model against full-information algorithms: given the same 25 predictors, logistic regression performs worse, not better. SHAP analysis of the XGBoost model confirmed the dominance of baseline fast-twitch maximum (mean $|\text{SHAP}| = 0.736$, roughly sevenfold the second-ranked variable, age [mean $|\text{SHAP}| = 0.103$]), with low baseline values driving predictions toward response (Figure 6).

Table 4. Performance of the nomogram versus benchmark models on the validation set (n = 303).

Model	Predictors	AUC	Brier score	Sensitivity	Specificity	PPV	NPV
Logistic nomogram	2 (selected)	0.717	0.214	0.597	0.704	0.615	0.688
Full-variable logistic regression	25 (all)	0.697	—	—	—	—	—
L2-regularized full logistic	25 (all)	0.697	—	—	—	—	—
Random forest	25 (all)	0.720	0.212	0.560	0.734	0.625	0.678
XGBoost	25 (all)	0.713	0.217	0.530	0.746	0.623	0.667
SVM (RBF kernel)	25 (all)	0.693	0.221	0.515	0.722	0.595	0.652

Best value per metric in bold. PPV, positive predictive value; NPV, negative predictive value. Threshold 0.5. The full-variable logistic model achieved an apparent training-set AUC of 0.777, indicating overfitting.

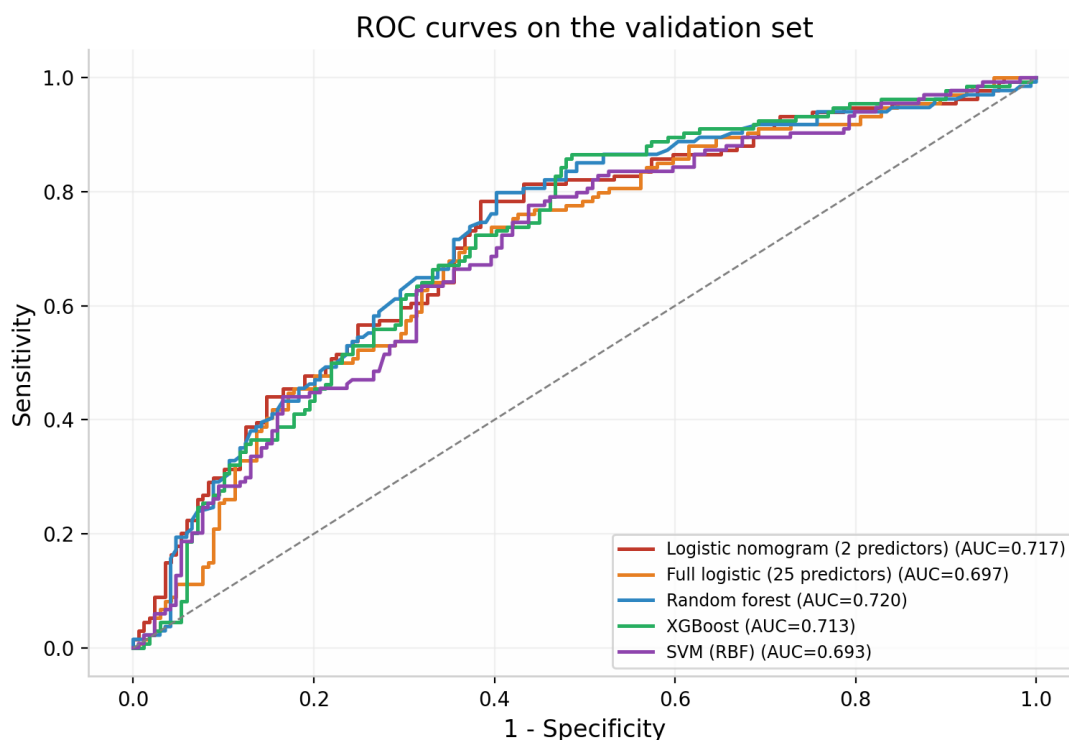


Figure 5. Receiver operating characteristic curves on the validation set for the two-variable logistic nomogram, full-variable logistic regression, random forest, XGBoost, and SVM. The nomogram, random forest, and XGBoost curves are nearly superimposed, while the full-variable logistic model underperforms despite using all 25 predictors.



Figure 6. SHAP summary plot of the XGBoost model (validation set). Each point is a patient; color encodes the feature value (red = high, blue = low). Baseline fast-twitch maximum dominates model output, with low values (blue) pushing predictions toward treatment response.

3.6. Subgroup, Sensitivity, and Extended-Imputation Analyses

The inverse association between baseline fast-twitch maximum and response was remarkably consistent across all eight subgroups (ORs 0.919–0.943 per μV), and no interaction reached significance (all $p > 0.10$) (Table 5, Figure 7). Sensitivity analyses—complete-case analysis ($n = 760$), 10% and 30% thresholds, exclusion of assessments >180 days postpartum, exclusion of forceps deliveries, and restriction to women with paired Glazer scores—yielded AUCs of 0.729–0.764 with the same two-predictor structure (Table 6). In the extended-imputation analysis, perineal injury entered the LASSO model (univariable $p = 0.082$) but labor durations did not ($p = 0.905$ and 0.221 for total and second-stage duration), and the extended model’s validation AUC (0.716) was unchanged from the primary model (0.717), indicating that excluding these high-missingness variables did not discard predictive information. Applied to secondary outcomes, the same predictors achieved AUCs of 0.707 (tonic mean $\geq 20\%$; event rate 49.7%) and 0.670 (endurance mean $\geq 20\%$; event rate 49.1%).

Table 5. Subgroup analyses: odds ratio of response per 1- μV increase in baseline fast-twitch maximum, with interaction tests.

Subgroup	n	Events	OR (95% CI)	AUC	Interaction p
----------	---	--------	-------------	-----	---------------

Subgroup	n	Events	OR (95% CI)	AUC	Interaction p
Vaginal delivery	721	335	0.932 (0.918–0.945)	0.741	0.480
Cesarean delivery	286	110	0.940 (0.920–0.960)	0.737	0.480
Age <35 years	839	361	0.939 (0.928–0.951)	0.730	0.185
Age ≥35 years	168	84	0.919 (0.892–0.947)	0.800	0.185
Primiparous	637	290	0.940 (0.927–0.953)	0.735	0.319
Multiparous (parity ≥2)	369	155	0.928 (0.910–0.947)	0.764	0.319
Baseline Glazer <60	431	270	0.929 (0.902–0.957)	0.665	0.431
Baseline Glazer ≥60	576	175	0.943 (0.926–0.961)	0.684	0.431

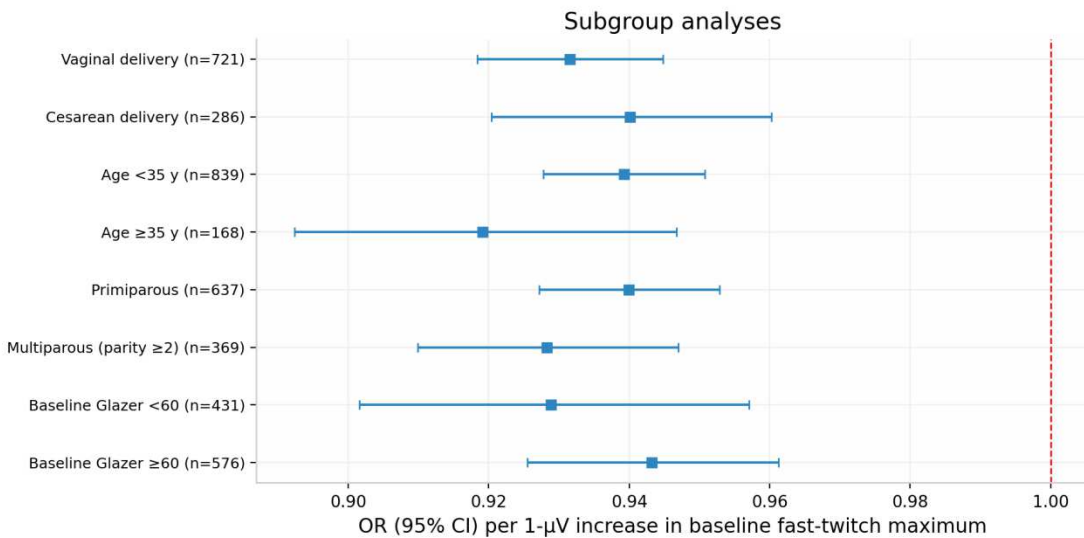


Figure 7. Forest plot of the association between baseline fast-twitch maximum and treatment response across subgroups. Point estimates (squares) and 95% confidence intervals (whiskers) are stable across all strata; no significant interaction was detected.

3.7. Mathematical Coupling and the Absolute-Improvement Outcome

Baseline fast-twitch maximum correlated more strongly with relative change (Spearman $\rho = -0.479$, $p < 0.001$) than with absolute change ($\rho = -0.348$, $p < 0.001$) (Figure 8). The Olds regression of absolute change on baseline gave a slope of -0.646 ($p < 0.001$), significantly greater than -1 , consistent with partial—but not complete—regression to the mean: women with lower baselines genuinely gained less in absolute terms than the ceiling effect alone would predict.

Modeling the absolute-improvement outcome ($\Delta \geq 5 \mu\text{V}$; event rate 45.5%) with the identical pipeline produced a four-predictor model retaining **age (OR 1.052 per year, 95% CI 1.011–1.095, $p = 0.012$)** and **baseline fast-twitch maximum (OR 0.954 per μV , 95% CI 0.939–0.969, $p < 0.001$)**, together with manual type-I strength (OR 1.367, $p = 0.051$) and HDP (OR 4.335, 95% CI 1.516–12.396, $p = 0.006$). The direction and significance of both core predictors therefore survive the transition from a relative to an absolute

outcome, although their effect sizes attenuate (as expected once definitional coupling is removed) and overall discrimination is more modest (validation AUC 0.609 vs 0.717). We interpret this pattern as evidence that the nomogram’s core signal reflects a genuine, physiologically grounded association that is *amplified*—not created—by the relative-change definition.

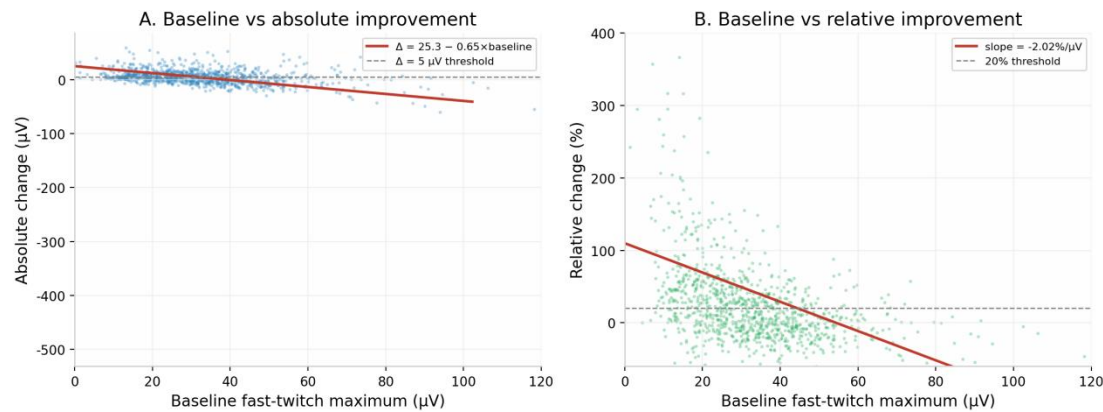


Figure 8. Baseline fast-twitch maximum versus treatment-induced change. (A) Absolute change against baseline with the Olds regression line (slope -0.646) and the $5\text{-}\mu\text{V}$ response threshold (dashed). (B) Relative change against baseline with the fitted slope and the 20% response threshold (dashed). The baseline–change association persists for absolute improvement, albeit attenuated, indicating partial rather than purely definitional coupling.

Table 6. Sensitivity and additional analyses using the two-predictor (age + baseline fast-twitch maximum) structure unless otherwise noted.

Analysis	n	Events	AUC
Primary analysis (imputed data)	1,007	445	0.744
Complete-case analysis	760	351	0.741
Outcome threshold $\geq 10\%$	1,007	527	0.729
Outcome threshold $\geq 30\%$	1,007	352	0.764
Excluding assessments >180 days postpartum	1,003	444	0.744
Excluding forceps deliveries	978	424	0.743
Restricted to paired Glazer scores	1,001	443	0.744
Extended imputation (perineal injury + labor durations added to pool)	1,007	445	0.716 (validation)
Absolute-improvement outcome ($\Delta \geq 5\text{ }\mu\text{V}$; four-predictor model)	1,007	458	0.609 (validation)

AUC estimated on the full analytical sample for comparability unless marked as validation-set performance.

4. Discussion

4.1. Principal Findings

In this cohort of 1,007 postpartum women, a parsimonious nomogram based on age and baseline fast-twitch maximum sEMG amplitude predicted the electrophysiological response to rehabilitation with moderate discrimination (bootstrap-corrected C-index 0.754) that was preserved under temporal validation (AUC 0.742) and adequate calibration. Neither three machine-learning algorithms nor a full-variable logistic regression using all 25 predictors exceeded the two-variable nomogram, and decision curve analysis indicated clinical usefulness across a wide, decision-relevant range of threshold probabilities. The model's central signal—lower baseline fast-twitch amplitude predicts a greater likelihood of response—was consistent across delivery mode, maternal age, parity, and baseline severity strata, across all sensitivity analyses, and—critically—persisted when the outcome was redefined in absolute rather than relative terms.

Discrimination was weaker for the secondary outcomes than for the primary fast-twitch outcome (AUC 0.707 for tonic and 0.670 for endurance improvement), a gradient we consider biologically plausible. The endurance contraction is sustained for ten seconds and therefore integrates motivation, instruction comprehension, and fatigue resistance to a greater degree than a phasic maximal contraction, increasing measurement variability and attenuating any baseline–response association. Moreover, endurance improvement may depend more heavily on training adherence and type-I fiber adaptation—mechanisms only partially reflected in pretreatment characteristics—so the lower predictability of this endpoint is expected rather than anomalous.

4.2. Comparison with Previous Work

Prior prognostic studies of pelvic floor rehabilitation have typically been small, single-arm, and focused on symptom outcomes in stress incontinence rather than on objective electrophysiological recovery¹³. Ptazkowski et al. reported prognostic value of sEMG in SUI, but without multivariable modeling or validation¹³. Recent AI-assisted analyses of pelvic floor sEMG have targeted diagnostic classification rather than treatment-response prediction¹⁴. To our knowledge, the present study is the first to (i) exceed 1,000 participants, (ii) provide a clinically applicable nomogram, and (iii) include temporal validation in this field. Our benchmark finding—no advantage of ML over logistic regression, and inferior performance of the full-variable logistic model—aligns with a systematic review of 71 clinical prediction studies showing equivalent discrimination when comparisons are methodologically fair¹⁶, and reinforces the principle that transparent, well-calibrated simple models should be preferred when performance is equivalent²⁹.

4.3. Mathematical Coupling, Regression to the Mean, and the Interpretation of the Core Signal

Because the primary outcome is defined as a *relative* improvement of $\geq 20\%$, a lower baseline mechanically requires a smaller absolute gain to qualify as a response (e.g., $+2\ \mu\text{V}$ from a $10\text{-}\mu\text{V}$ baseline versus $+10\ \mu\text{V}$ from a $50\text{-}\mu\text{V}$ baseline). The inverse association

between baseline fast-twitch maximum and response is therefore, in part, definitional. We addressed this threat explicitly rather than rhetorically. First, the baseline–change correlation was materially weaker for absolute than for relative change ($\rho = -0.348$ vs -0.479), and the Olds regression slope of -0.646 was significantly shallower than the -1 expected under pure regression to the mean—both indicating that part of the signal is genuine. Second, and decisively, when the outcome was redefined as an absolute gain of ≥ 5 μV , both core predictors retained their direction and significance (baseline fast-twitch maximum: OR 0.954, $p < 0.001$; age: OR 1.052, $p = 0.012$), with attenuated effect sizes as expected once definitional coupling is removed. Third, the association was stable across all subgroups and all sensitivity analyses, including the 10% and 30% relative thresholds that bracket the primary definition.

We therefore conclude that the signal is *amplified, not created*, by the outcome definition. Clinically, the nomogram answers the question “who is likely to show a $\geq 20\%$ relative electrophysiological gain?”—a legitimate and intuitive counseling endpoint—while the absolute-outcome analysis confirms that low-baseline women also experience genuinely larger recovery. However, this electrophysiological endpoint should not be conflated with a direct biological measure of tissue recovery: we caution against interpreting the nomogram as identifying who “biologically responds” to rehabilitation, and we recommend that future prospective studies anchor sEMG change to a minimal clinically important difference established against symptom-based outcomes.

4.4. Electrophysiological Response versus Symptomatic Benefit

A further boundary of interpretation deserves emphasis: the predicted endpoint is an *electrophysiological* improvement, not symptom relief or quality-of-life gain. While baseline sEMG indices carry prognostic information¹³ and digital muscle assessment correlates with sEMG¹², the strength of association between sEMG change and symptomatic change is variable across studies⁸. Whether the electrophysiological improvement predicted by this nomogram translates into symptom relief or quality-of-life gains remains to be established; the nomogram should be used as an adjunct to, not a replacement for, clinical judgment. We have deliberately framed the title, outcomes, and conclusions of this study in electrophysiological terms to avoid overgeneralization, and we regard prospective linkage of sEMG response to validated patient-reported outcome measures (PFDI-20 and PFIQ-7³⁰, PISQ-12³¹) as a key next step for this line of research.

4.5. Clinical Implications

The nomogram requires only two variables available at the first postpartum visit—age and the fast-twitch maximum from the routine baseline sEMG evaluation—and can be scored at the bedside in seconds without software. We envision two uses. First, pretreatment counseling: a woman with a predicted response probability of 0.70 can be reassured that a standard course is likely to yield measurable electrophysiological improvement, supporting adherence—a known determinant of PFMT success⁸. Second, treatment planning: for women with low predicted probability (e.g., <0.35), clinicians may consider intensified protocols (combined modalities, more sessions, or adjunctive biofeedback) and closer monitoring, although such triage should be prospectively tested before adoption. The wide

net-benefit range on DCA (thresholds 0.2–0.8) suggests the model is informative under diverse clinician preferences for false-positive versus false-negative trade-offs.

4.6. Strengths and Limitations

Strengths include the large sample, with an event-per-variable ratio above 20, and the prespecified preprocessing and missing-data rules. The three-tier validation design incorporated temporal validation, and discrimination, calibration, and clinical utility were reported comprehensively per TRIPOD¹⁵. Finally, the machine-learning benchmark was explicitly and fairly constructed, and the mathematical-coupling threat was interrogated quantitatively rather than rhetorically.

Several limitations deserve emphasis. First, this was a single-center retrospective study without external multicenter validation; temporal validation mitigates but does not replace geographic transportability testing. Second, patient-reported outcomes were missing in >99% of records, so the endpoint is objective sEMG change rather than symptom relief (Section 4.4). Third, perineal injury grade and labor durations exceeded our 20% missingness threshold and were excluded from the primary pool; our extended-imputation analysis suggests no predictive information was lost (validation AUC 0.716 vs 0.717), but imputation of largely structural missingness (labor durations in cesarean deliveries) rests on strong assumptions, and we recommend that future prospective cohorts collect these obstetric variables systematically. Fourth, the relative-change primary outcome is intrinsically baseline-dependent; we have quantified this coupling and confirmed persistence of the signal under an absolute-change definition, but the two endpoints answer different clinical questions. Fifth, treatment modality was heterogeneous (electrical, magnetic, combined, radiofrequency) and not modeled; efficacy may differ by modality. Finally, 803 women lacked post-treatment sEMG and were excluded, which could introduce attrition bias; however, excluded women did not differ materially from those included in age or baseline sEMG profile (Supplementary Table S1), suggesting that any such bias is likely limited.

5. Conclusions

A two-variable nomogram—age and baseline fast-twitch maximum sEMG amplitude—predicts the electrophysiological response to postpartum pelvic floor rehabilitation with moderate, well-calibrated, and temporally robust discrimination, performing equivalently to machine-learning models and better than a full-variable logistic model built on 25 predictors. The core signal survives redefinition of the outcome in absolute terms, indicating it is amplified rather than created by mathematical coupling. Its simplicity supports immediate bedside use for pretreatment counseling as an adjunct to clinical judgment. The essential next step is prospective, multicenter validation that anchors electrophysiological response to patient-reported outcomes and systematically collects perineal injury and labor-stage durations.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional ethics committee of Jinhua People's Hospital. This study was conducted in accordance with the Declaration of Helsinki. This is a retrospective observational study using de-identified, anonymized electronic delivery records without any patient intervention, identifiable personal information, or potential physical or psychological harm to participants. In line with relevant national and institutional ethical regulations, ethical approval and informed consent were formally waived for this minimal-risk study. Consequently, no ethical approval number was assigned.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to institutional privacy policies but are available from the corresponding author on reasonable request and with appropriate institutional approval.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

Lu Jiaping was responsible for data collection, database sorting, preliminary statistical analysis and manuscript drafting. Xu Qiulian participated in clinical data screening, case inclusion and exclusion judgment, and revision of the results section. Xu Heng conceived and designed the research framework, guided the whole statistical analysis process, revised the full manuscript repeatedly, and acted as the corresponding author of this paper. All authors read and approved the final submitted manuscript.

Acknowledgments

The authors gratefully acknowledge the contributions of all obstetricians, midwives, and nursing staff whose clinical work made this study possible.

References

- ¹ Milsom I, Gyhagen M. The prevalence of urinary incontinence. *BJOG*. 2019;126(2):211–218.
- ² Hannestad YS, Rortveit G, Sandvik H, Hunskaar S. A community-based epidemiological survey of female urinary incontinence: the Norwegian EPINCONT study. *J Clin Epidemiol*. 2000;53(11):1150–1157.
- ³ Haylen BT, de Ridder D, Freeman RM, et al. An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. *Neurourol Urodyn*. 2010;29(1):4–20.
- ⁴ Bø K, Frawley HC, Haylen BT, et al. An IUGA/ICS joint report on the terminology for the conservative and nonpharmacological management of female pelvic floor dysfunction. *Neurourol Urodyn*. 2017;36(2):221–244.
- ⁵ Zhang Y, Li L, Wang X, Chen Q. Pelvic floor function and its influencing factors one month postpartum: a cross-sectional study in China. *PLoS One*. 2023;18(1):e0308563.
- ⁶ Rortveit G, Daltveit AK, Hannestad YS, Hunskaar S. Urinary incontinence after vaginal delivery or cesarean section. *N Engl J Med*. 2003;348(10):900–907.
- ⁷ Dumoulin C, Cacciari LP, Hay-Smith EJC. Pelvic floor muscle training versus no treatment, or inactive control treatments, for urinary incontinence in women. *Cochrane Database Syst Rev*. 2018;10:CD005654.
- ⁸ Mørkved S, Bø K. Effect of pelvic floor muscle training during pregnancy and after childbirth on prevention and treatment of urinary incontinence: a systematic review. *Br J Sports Med*. 2014;48(4):299–310.
- ⁹ Hilde G, Stær-Jensen J, Siafarikas F, et al. Postpartum pelvic floor muscle training and urinary incontinence: a randomized controlled trial. *Obstet Gynecol*. 2013;122(6):1231–1238.
- ¹⁰ Glazer HI, Romanzi L, Polaneczky M. Pelvic floor muscle surface electromyography: reliability and clinical predictive validity. *J Reprod Med*. 1999;44(9):779–782.
- ¹¹ Romanzi LJ, Polaneczky M, Glazer HI. Simple test of pelvic muscle contraction during pelvic examination: correlation to surface electromyography. *Neurourol Urodyn*. 1999;18(6):603–612.

-
- ¹² Ptazkowski K, Malkiewicz B, Zdrojowy R, et al. The prognostic value of the surface electromyographic assessment of pelvic floor muscles in women with stress urinary incontinence. *J Clin Med*. 2020;9(6):1967.
- ¹³ Chen J, Yao J, Chen W. AI-based pelvic floor surface electromyography reference ranges and high-precision pelvic floor dysfunction diagnosis. *EBioMedicine*. 2025;117:105755.
- ¹⁴ Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55–63.
- ¹⁵ Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
- ¹⁶ Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58.
- ¹⁷ Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Springer; 2019.
- ¹⁸ van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
- ¹⁹ Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996;58(1):267–288.
- ²⁰ Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol*. 2008;26(8):1364–1370.
- ²¹ Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
- ²² Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
- ²³ Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74(6):796–804.
- ²⁴ Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
- ²⁵ 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*. 2016:785–794.
- ²⁶ Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297.

-
- ²⁷ Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. Curran Associates; 2017:4765–4774.
- ²⁸ Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chronic Dis.* 1962;15:969–977.
- ²⁹ Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128–138.
- ³⁰ Barber MD, Walters MD, Bump RC. Short forms of two condition-specific quality-of-life questionnaires for women with pelvic floor disorders (PFDI-20 and PFIQ-7). *Am J Obstet Gynecol.* 2005;193(1):103–113.
- ³¹ Rogers RG, Coates KW, Kammerer-Doak D, Khalsa S, Qualls C. A short form of the Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire (PISQ-12). *Int Urogynecol J Pelvic Floor Dysfunct.* 2003;14(3):164–168.

Supplementary Material

Supplementary Table S1. Baseline characteristics of women excluded for missing post-treatment sEMG versus the analytical cohort. Values are median (IQR).

Characteristic	Excluded (n = 803)	Analytical cohort (n = 1,007)	SMD
Age, years	30.0 (28.0, 33.0)	31.0 (28.0, 33.0)	0.079
BMI, kg/m ²	22.1 (20.8, 24.3)	22.8 (20.8, 24.9)	0.089
Baseline Glazer score	61.4 (49.3, 70.8)	63.0 (50.8, 72.6)	0.054
Baseline pre-rest tone, μ V	4.3 (2.6, 7.1)	4.0 (2.2, 6.5)	0.014
Baseline fast-twitch maximum, μ V	29.5 (20.1, 40.2)	30.6 (20.9, 41.5)	0.098
Baseline tonic mean, μ V	18.6 (12.2, 26.7)	19.9 (12.5, 28.6)	0.089
Baseline endurance mean, μ V	15.9 (10.4, 21.9)	16.4 (10.6, 24.2)	0.094

SMD, standardized mean difference. All SMDs are below 0.1, indicating no material baseline differences between excluded and included women.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile2TRIPODchecklist.docx](#)