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.

Incremental value of pretreatment MRI features for predicting Miller–Payne grade 5 after neoadjuvant therapy in breast cancer: model development and exploratory temporal evaluation

Xiang Zhuang
Baotou Cancer Hospital
Kaiqi Tan
Inner Mongolia University of Science and Technology
Hui Zhang
Baotou Cancer Hospital
Jia Hao
Baotou Cancer Hospital
Chenwei Hou
Inner Mongolia University of Science and Technology
Zhihui Zhang
Inner Mongolia University of Science and Technology
Xinying Zhang
Chengdu University of Traditional Chinese Medicine
Renyi Zhu
Inner Mongolia University of Science and Technology
Qing Yang
Inner Mongolia University of Science and Technology
Yongqiang Xing
Inner Mongolia University of Science and Technology
Hu Meng
mh_imust@foxmail.com

Inner Mongolia University of Science and Technology

Research Article

Keywords: Miller–Payne, breast cancer, neoadjuvant therapy, machine learning, temporal evaluation, magnetic resonance imaging

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11266553/v1

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

Additional Declarations: No competing interests reported.

Incremental value of pretreatment MRI features for predicting Miller–Payne grade 5 after neoadjuvant therapy in breast cancer: model development and exploratory temporal evaluation

Xiang Zhuang[†a], Kaiqi Tan[†b], Hui Zhang[a], Jia Hao[a], Chenwei Hou[b], Zhihui Zhang[b], Xinying Zhang[c], Renyi Zhu[b], Qing Yang[b], Yongqiang Xing[b], Hu Meng[b*]

[a] Baotou Cancer Hospital, 18 Tuanjie Street, Qingshan District, Baotou 014030, Inner Mongolia, China

[b] Inner Mongolia Key Laboratory of Life Health and Bioinformatics, College of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, Inner Mongolia, China

[c] Chengdu University of Traditional Chinese Medicine, 1166 Liutai Avenue, Wenjiang District, Chengdu 611137, Sichuan, China

†These authors contributed equally to this work and share first authorship.

*Correspondence: Hu Meng

Inner Mongolia Key Laboratory of Life Health and Bioinformatics, College of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, Inner Mongolia, China. Email: mh_imust@foxmail.com

Abstract

Background: To quantify the incremental value of pretreatment magnetic resonance imaging (MRI) measurements beyond clinicopathological variables for predicting Miller‐Payne

grade 5 (MP5) after neoadjuvant therapy (NAT), and to describe performance in a same-centre temporal cohort.

Methods: Patients receiving NAT were divided chronologically into development (n = 371)$n = 371$ and same-centre temporal (n = 93)$n = 93$ cohorts. Core clinicopathological, MRI-integrated, and Expanded predictor sets were evaluated with six algorithms using nested five-fold cross-validation repeated five times; hyperparameters were tuned within each outer training fold. After the unified development comparison, the MRI-integrated eXtreme Gradient Boosting (XGBoost) model was selected post hoc as a candidate for further study.

Results: Development MRI-integrated receiver operating characteristic areas under the curve (ROC-AUCs) ranged from 0.8447 to 0.8644 across the six algorithms. XGBoost had the highest five-repeat mean (0.8644), only 0.0002 above Random Forest (0.8642). Its Core-to-MRI Δ AUC was +0.0124 (Bonferroni simultaneous bootstrap 95% CI for the six Core-to-MRI comparisons, -0.0036 to +0.0289; separately, Holm-adjusted P = 0.221); all six adjusted intervals crossed zero. In the temporal cohort (43/93 MP5), XGBoost achieved ROC-AUC 0.8507 (95% CI, 0.7618 – 0.9222), compared with 0.8535 (0.7665 – 0.9239) for Random Forest. The post hoc paired AUC difference was -0.0028 (XGBoost minus Random Forest; DeLong 95% CI, -0.0194 to 0.0139; two-sided P = 0.743). XGBoost had a Brier score of 0.1632 (95% CI, 0.1314 – 0.1994), joint-fit calibration intercept of 0.093 (95% CI, -0.427 to 0.634), and slope of 1.373 (95% CI, 0.888 – 2.395).

Conclusions: Clinicopathological variables provided good MP5 discrimination. Routine MRI measurements yielded small, uncertain incremental gains. XGBoost was selected after the unified development comparison. Its temporal discrimination and calibration estimates are

exploratory results from already available data; independent validation of discrimination, calibration, and clinical utility is needed before clinical use.

Keywords: Miller–Payne; breast cancer; neoadjuvant therapy; machine learning; temporal evaluation; magnetic resonance imaging

Background

Breast cancer remains a major global health burden [1]. Neoadjuvant therapy (NAT) permits assessment of tumour response before surgery [2,3]. The Miller–Payne (MP) system grades postoperative response by comparing tumour cellularity between pretreatment biopsy and post-treatment surgical specimens, from no reduction (MP1) to disappearance of invasive tumour cells from the primary tumour bed (MP5) [4,5]. MP5 measures disappearance of invasive tumour cells from the primary breast lesion; residual ductal carcinoma in situ may remain. This allowance also applies to the commonly used ypT0/is ypN0 definition of pathological complete response (pCR), although stricter pCR definitions exclude in situ disease. Thus, the breast-tumour criterion of MP5 corresponds to ypT0/is; when the sampled nodes are also ypN0, this meets the commonly used ypT0/is ypN0 pCR definition. MP grading alone does not assess regional lymph-node status [3–5].

Machine-learning studies have increasingly examined response to NAT, including conventional pCR prediction [6,7] and a systematic review of multimodal deep learning for pCR prediction [8]. Hou et al. analysed 1,053 patients with MP5 as the outcome and compared clinicopathological prediction models using five-fold cross-validation [9]. Bi et al. used pretreatment magnetic resonance imaging (MRI) radiomics to predict MP grades 1–2 versus 3–5, a different outcome from MP5 versus MP1–MP4, and reported calibration and decision curves [10]. Song et al. used 26 clinical-record variables from 166 patients to predict the five Miller–

Payne grades with four classifiers, a different outcome from MP5 versus MP1–MP4 [11]. MRI accuracy for assessing residual disease after neoadjuvant chemotherapy (NAC) varies across breast cancer subtypes [12]. The incremental value of routine MRI measurements beyond clinicopathological predictors for the specific MP5 endpoint remains uncertain. Tumour size, apparent diffusion coefficient (ADC), and time–intensity curve (TIC) pattern provide complementary morphological, cellular, and kinetic information [13,14].

Transportability and calibration are essential for clinical prediction. Internal cross-validation estimates development performance but cannot establish performance in later patients or other settings; calibration should therefore be assessed alongside discrimination [15–18]. Many prediction studies emphasise discrimination without fully examining calibration, temporal stability, or uncertainty around incremental performance. Evaluation in a later same-centre cohort provides a pragmatic assessment of temporal stability while remaining distinct from external validation.

We quantified the incremental value of routine pretreatment MRI measurements beyond clinicopathological predictors for MP5 after NAT. We compared Core, MRI-integrated, and Expanded predictor sets across six algorithms. The intended use at this stage was research-level pretreatment MP5 risk stratification; no treatment action or probability threshold was prespecified. Following the unified development comparison, MRI-integrated eXtreme Gradient Boosting (XGBoost) was chosen as a candidate for further study, and its performance was described in the existing same-centre temporal cohort.

Methods

Study population

This retrospective single-centre study included 464 eligible female patients with invasive breast cancer who received NAT followed by surgery and had pretreatment examinations dated from 15 July 2022 to 10 June 2026 (screened from 528 consecutive institutional records; Fig. 1). Patients were ordered by pretreatment examination date and partitioned chronologically in an approximately 80:20 ratio: 371 patients examined from 15 July 2022 to 1 August 2025 formed the development cohort, and 93 examined from 6 August 2025 to 10 June 2026 formed the same-centre temporal cohort. The resulting date boundary reflected the chronological partition rather than a clinically selected calendar cutoff; no eligible patient in the final cohort had a pretreatment examination dated 2–5 August 2025. The split was independent of MP5 status and model performance. All predictors were obtained before NAT, whereas the outcome was determined from postoperative surgical specimens. Exclusions comprised metastatic breast cancer (n = 8), a history of other malignancies (n = 18), severely incomplete or non-evaluable pretreatment clinical, pathological, or MRI data (n = 20), prior procedures or neoadjuvant treatment at another institution with unavailable required pretreatment data (n = 9), and unavailable or non-evaluable postoperative MP grading (n = 9). Treatment-related variables, including neoadjuvant regimen composition and targeted or immunotherapy exposure, were not collected for this analysis.

The primary outcome was MP5 after NAT, determined from the surgical specimen of the primary breast lesion. Pretreatment core-needle biopsy and post-NAT surgical specimens were evaluated by three senior pathologists according to established criteria, with disagreements resolved by consensus. MP5 was defined as complete absence of residual invasive tumour cells in the primary tumour bed; residual ductal carcinoma in situ was permitted. MP1–MP4

constituted non-MP5. A post hoc sample-size assessment used the pmsampsize R package (version 1.1.3; R version 4.4.3) of the Riley framework [19]. With the observed mean out-of-fold development C-statistic (equivalent to the area under the receiver operating characteristic curve [ROC-AUC]; 0.864), MP5 prevalence (39.6%), and 9, 14, and 22 candidate predictor parameters for the Core, MRI-integrated, and Expanded sets, respectively, the estimated minimum sample size was 368 cases (146 events) for the Core and MRI-integrated sets and 467 cases (185 events) for the Expanded set. For Core and MRI-integrated, the precision criterion for estimating the overall outcome risk was binding (368 cases for each); for Expanded, the criterion limiting optimism in model fit was binding (467 cases). Under the assumptions of this post hoc calculation, the development cohort (371 cases, 147 events) exceeded the numerical minimum for the Core and MRI-integrated sets but not the Expanded set; Expanded-set analyses were therefore exploratory. Because the observed, model-selected development performance informed these calculations, the required sample size may be underestimated; the assessment was descriptive rather than a prospective sample-size justification. The development cohort exceeded the calculated Core/MRI-integrated numerical minimum by only three cases and one event. Parameter counts excluded the intercept: Core comprised four continuous terms, two human epidermal growth factor receptor 2 (HER2) indicators, and three grade indicators (9); MRI-integrated added two continuous MRI terms and three TIC indicators (14); Expanded additionally included two continuous biomarker terms, four qualitative p53 indicators, and one numeric term each for cytokeratin 5/6 (CK5/6) and epidermal growth factor receptor (EGFR) (22). This calculation does not establish that the sample size was sufficient for each of the six algorithms.

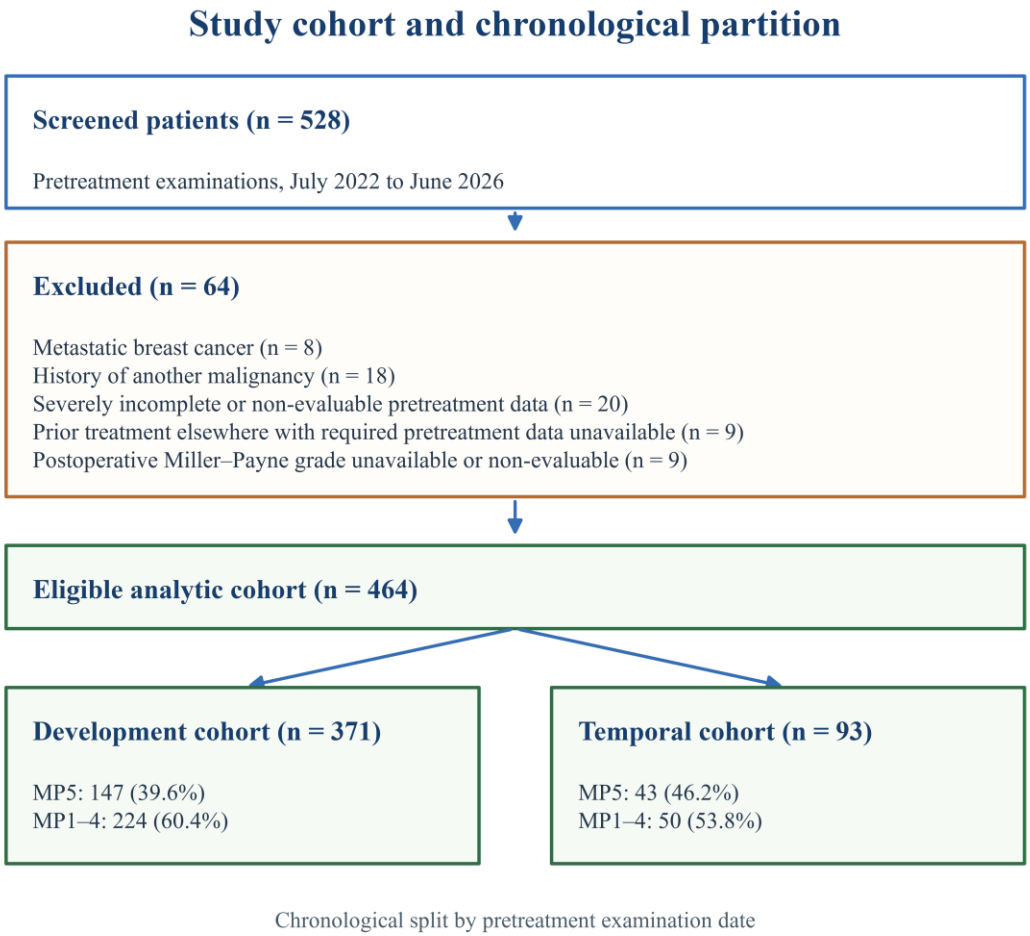


Fig. 1. Study flowchart.

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Review Committee of Baotou Cancer Hospital (approval no. 2025 Ethics Review No. 018). This was a retrospective study based on de-identified medical records and pre-treatment MRI data; the requirement for written informed consent from individual participants was therefore waived by the Ethics Review Committee.

Predictors and data preprocessing

Candidate predictors were obtained from routine pretreatment examinations. Tumour size, ADC ($\times 10^{-3}$ mm²/s), and TIC pattern were extracted from routine pretreatment MRI reports before NAT; images were not reread for this analysis. Clinicopathological predictors were age, oestrogen receptor (ER), progesterone receptor (PR), HER2, Ki-67, histological grade, androgen receptor (AR), p53 percentage expression, qualitative p53 staining pattern, CK5/6, and EGFR. When multiple tumour diameters were recorded, the maximum diameter was retained. ER, PR, and Ki-67 entered the models as continuous percentages; values recorded on a 0–1 scale were converted to percentages, interval ADC values were represented by midpoints, and HER2 status was parsed from the recorded immunohistochemistry (IHC) score and accompanying positive/negative annotation. IHC 3+ was classified as positive. IHC 0 or 1+ was classified as negative unless an accompanying annotation explicitly stated positive; IHC 2+ without a definitive accompanying positive/negative result, blank entries, and uninterpretable scores were classified as equivocal/unknown. Explicit positive/negative annotations were treated as the recorded clinical interpretation of an IHC 2+ entry; the underlying confirmatory test results were unavailable. No IHC 0/1+ entry carried a positive annotation and no IHC 3+ entry carried a negative annotation. Five IHC 2+ entries carried a positive annotation and were classified positive. Of 83 equivocal/unknown records, 75 had IHC 2+ without a definitive annotation (57 development, 18 temporal) and eight lacked an interpretable IHC score (two development, six temporal). This combined category could reflect unperformed confirmatory testing, inconclusive testing, or incomplete recording, which could not be separated. Numeric p53 expression and qualitative p53 status were parsed separately from the recorded report. Percentage-only entries contributed to the numeric field but were coded Unknown in the qualitative field; explicit

mutant/aberrant or wild-type entries contributed to the qualitative field without an inferred percentage. No staining-pattern category was inferred from a numeric expression threshold. Detailed MRI acquisition metadata, including scanner field strength, pulse sequences, and b-values, were not systematically available from routine reports.

Missing or uncertain HER2, histological grade, TIC pattern, and qualitative p53 values were encoded as explicit categories. For elastic-net logistic regression, radial basis function support vector machine (RBF-SVM), Random Forest, and XGBoost, continuous variables were median-imputed and categorical variables one-hot encoded using training-data parameters; elastic-net and RBF-SVM models were additionally standardised. CatBoost and TabPFN (version 2.1.3) handled missingness natively. Three predictor sets were prespecified: Core (age, ER, PR, HER2, Ki-67, and histological grade); MRI-integrated (Core plus tumour size, ADC, and TIC pattern), the primary set; and Expanded (MRI-integrated plus AR, p53 expression and staining pattern, CK5/6, and EGFR). No class-balancing procedure was applied.

Model development, performance assessment, and selection

Reporting was guided by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD+AI) statement [17] (Additional file 6: Table S6). Model selection after review of the unified development results and the already available temporal results is explicitly identified as post hoc; the temporal assessment should not be represented as a prospective validation of a newly prespecified XGBoost choice.

Six algorithms were compared in the development cohort: elastic-net logistic regression, RBF-SVM, Random Forest, XGBoost, CatBoost, and TabPFN [20]. The unified analysis repeated stratified nested five-fold cross-validation five times: within each repetition, each of five outer training folds used stratified three-fold inner cross-validation for hyperparameter

tuning [16,21]. TabPFN used a fixed configuration; software versions and result provenance are detailed in Additional file 4: Table S4. Each repetition yielded complete out-of-fold (OOF) predictions. Development ROC-AUC was the mean of the five repeat-level OOF AUCs (Fig. 2).

After the unified development comparison, MRI-integrated XGBoost was retained as an exploratory candidate because it had the highest unrounded mean OOF ROC-AUC; the small differences did not establish algorithmic superiority. The selected XGBoost was fitted using development-cohort data only; the 93 temporal outcomes were not used for its preprocessing, tuning, or fitting. Because the temporal results were already available when the final XGBoost choice was made, they are reported as exploratory assessment rather than confirmation of a prospectively locked choice. Fig. 3 presents Shapley additive explanations (SHAP) TreeExplainer values for the fitted MRI-integrated XGBoost on the raw log-odds scale. One-hot encoded columns were combined into original predictor concepts. Importance and summary panels use the 371 development patients and describe the fitted model in sample; the waterfall example is a temporal patient selected by proximity to the median predicted probability without reference to outcome. SHAP values are model attributions, not causal effects or held-out performance estimates.

Same-centre temporal evaluation and supplementary analyses

The development-fitted MRI-integrated models were applied to the 93-patient same-centre temporal cohort. Frozen predictions were used to assess ROC-AUC, precision–recall performance, Brier score, and calibration. Exploratory decision curves [22] are presented in Additional file 8: Fig. S1; no clinical action linked to a probability threshold was prespecified. AUCs across the six algorithms were reported descriptively. In one additional post hoc comparison, the XGBoost-minus-Random Forest AUC difference in the same 93 patients was

tested with a two-sided paired DeLong test [23], with a covariance-based Wald 95% CI. A 10,000-resample patient-level paired percentile-bootstrap interval was calculated as a sensitivity analysis; each resample retained the outcome and both model predictions for each selected patient. The nominal paired test and confidence intervals condition on the fitted models and do not account for uncertainty from model refitting or the post hoc selection of XGBoost and this comparison. Separately, temporal patients were grouped into tertiles of XGBoost-predicted MP5 probability. To assess the MRI increment in the temporal cohort, a Core-only XGBoost model was tuned and refitted using the 371 development patients only. Its frozen predictions were compared with those of the MRI-integrated XGBoost in the same 93 temporal patients. The paired MRI-minus-Core difference in ROC-AUC and its 95% percentile confidence interval were calculated using 5,000 patient-level bootstrap resamples. This post hoc analysis conditions on the fitted models and does not account for model-refitting or selection uncertainty. A separate sensitivity analysis excluded the eight patients without an interpretable HER2 IHC score and recomputed XGBoost Core and MRI AUCs from frozen predictions without refitting; a 5,000-resample paired patient bootstrap supplied a temporal ΔAUC interval (Additional file 10: Table S9).

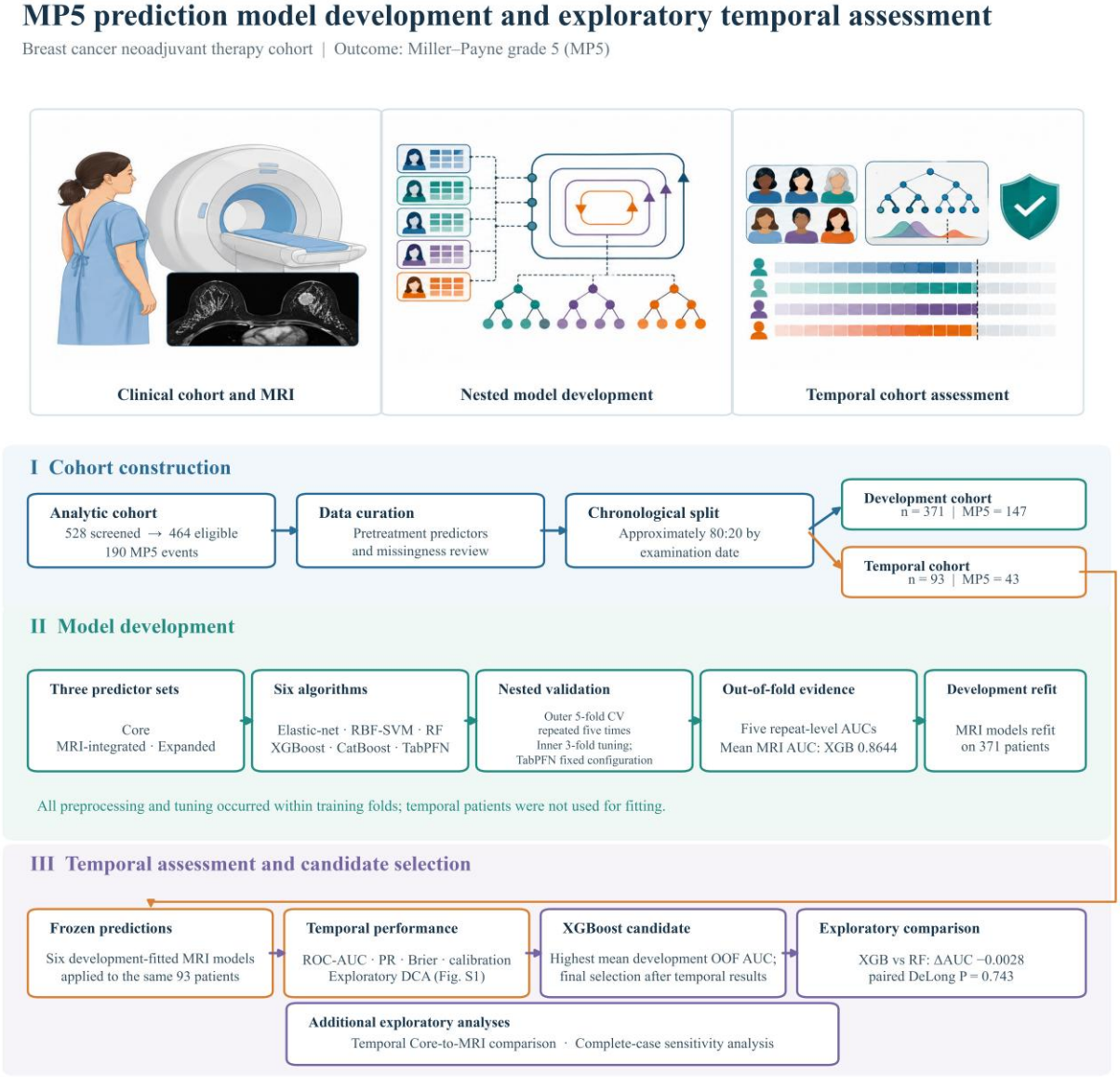


Fig. 2. Analysis workflow. The 371-patient development cohort underwent a unified six-algorithm, three-predictor-set comparison with repeated nested cross-validation. MRI-integrated XGBoost was selected post hoc as the candidate model based on the highest development mean AUC. Its application to the already available 93-patient same-centre temporal cohort is an exploratory assessment, not a prospectively locked validation.

Statistical analysis

Continuous variables are presented as mean ± standard deviation or median (interquartile range), and categorical variables as number (percentage). Differences between the development and temporal cohorts were described using absolute standardised mean differences (SMDs): the

pooled standard deviation was used for continuous variables and the pooled Bernoulli variance for each categorical level. Missing values were excluded variable-wise. Associations with MP5 were examined using univariable and multivariable Firth penalised logistic regression, reporting odds ratios, profile penalised-likelihood 95% CIs, and penalised likelihood-ratio P values. For regression analyses, ADC was expressed per 0.1×10^{-3} mm²/s increment. MP5 had no missing values. Machine-learning analyses included all 371 development patients with training-fold preprocessing. Firth regression excluded truly missing continuous values; the MRI-integrated Firth model included 355 patients. Variable-specific missingness appears in Additional file 1: Table S1. The multiplicity family comprised six paired Core-to-MRI AUC comparisons, one for each algorithm. These six comparisons used simultaneous bootstrap 95% intervals and Holm-adjusted P values. MRI-integrated-to-Expanded comparisons were outside this six-test family and are descriptive.

Development ROC-AUC for each information set was the arithmetic mean of five complete repeat-level OOF AUCs; Core-to-MRI ΔAUC was the difference between the two unrounded means for each algorithm. For 5,000 patient-level bootstrap resamples, the same patient indices were applied synchronously to Core and MRI predictions across all five repeats. Each resample yielded the difference between mean repeat-level AUCs. The six Core-to-MRI multiplicity-adjusted intervals used Bonferroni percentile limits at $0.05/(2 \times 6)$ and $1 - 0.05/(2 \times 6)$, respectively. For each algorithm, the two-sided bootstrap P value for a null ΔAUC of zero was calculated as [1 + number of resamples with $|\Delta AUC^* - \text{observed } \Delta AUC| \geq |\text{observed } \Delta AUC|$]/(5,000 + 1). Thus, the bootstrap differences were centred at the observed difference for testing, whereas the percentile confidence limits used the uncentred bootstrap distribution. All 5,000 resamples contained both outcome classes. The six resulting P values were adjusted by the

Holm procedure. MRI-to-Expanded comparisons were not included in this family. These patient-resampling intervals and tests condition on the fixed OOF predictions. They do not propagate new cross-validation splits, preprocessing, tuning, model refitting, or post hoc selection uncertainty. Individual temporal ROC-AUC intervals used 5,000 patient-level percentile-bootstrap resamples of frozen predictions. The additional XGBoost-versus-Random Forest paired DeLong test used the covariance of AUCs measured in the same patients. Calibration was assessed using a jointly estimated logistic intercept and slope from MP5 status regressed on the logit of each frozen predicted probability (clipped to 10^{-6} – $[1 - 10^{-6}]$), the Brier score (mean squared prediction error), and grouped calibration curves. Temporal intercept, slope, and Brier 95% CIs used 5,000 patient-level percentile-bootstrap resamples; calibration fits required both outcome classes (Additional file 3: Supplementary methods, temporal calibration intervals). Temporal-cohort calibration and decision-curve analyses were conducted post hoc and interpreted descriptively; decision curves covered threshold probabilities of 0.10 – 0.70. A paired 10,000-resample patient bootstrap supplied a percentile sensitivity interval. A post hoc complete-case XGBoost sensitivity analysis excluded patients with missing age, ER, PR, Ki-67, tumour diameter, or ADC, while retaining recorded HER2 equivocal/unknown, grade Unknown, and TIC unknown categories. Only the originally selected hyperparameters and outer splits were reused. For each restricted outer training fold, preprocessing (including imputation and encoding) was fitted afresh on that fold's complete cases before XGBoost was refitted; for temporal predictions, preprocessing and the model were refitted on all complete development cases using the original final hyperparameters. Paired 5,000-resample bootstrap intervals were nominal, outside the six-test family, and conditional on these splits and tuning choices; they did not account for post hoc model selection (Additional file 9: Table S8).

Results

Study population and baseline characteristics

The analysis included 371 patients in the development cohort and 93 in the same-centre temporal cohort. MP5 occurred in 147/371 patients (39.6%) and 43/93 patients (46.2%), respectively. Baseline characteristics are summarised in Table 1.

Table 1. Baseline characteristics and absolute standardised differences between the development and temporal cohorts.

Variable	Development	Temporal	|SMD|
Age (years)	55.60 ± 10.01	57.29 ± 11.82	0.154
Tumour maximum diameter (cm)	3.24 ± 1.69	3.17 ± 1.78	0.044
ADC ($\times 10^{-3}$ mm²/s)	0.91 ± 0.17	0.92 ± 0.23	0.071
ER (%)	47.59 ± 41.28	35.11 ± 41.08	0.303
PR (%)	24.01 ± 34.30	18.06 ± 32.88	0.177
AR (%)	59.95 ± 33.83	55.04 ± 37.21	0.138
p53 (%)	32.70 ± 35.01	44.00 ± 38.61	0.307
Ki-67 (%)	42.13 ± 26.26	46.18 ± 26.31	0.154
HER2 status			
Equivocal/unknown	59 (15.9%)	24 (25.8%)	0.246
Negative	221 (59.6%)	53 (57.0%)	0.052
Positive	91 (24.5%)	16 (17.2%)	0.181
TIC pattern			
Plateau	112 (30.2%)	42 (45.2%)	0.313
Unknown	4 (1.1%)	0 (0.0%)	0.148
Wash-in	15 (4.0%)	3 (3.2%)	0.044
Wash-out	240 (64.7%)	48 (51.6%)	0.267
ER status			
Negative (<1%)	142 (38.4%)	44 (47.3%)	0.181
Positive (≥1%)	228 (61.6%)	49 (52.7%)	0.181
PR status			
Negative (<1%)	198 (53.8%)	63 (67.7%)	0.288
Positive (≥1%)	170 (46.2%)	30 (32.3%)	0.288
CK5/6			
Negative	273 (83.5%)	69 (81.2%)	0.061
Positive	54 (16.5%)	16 (18.8%)	0.061

Variable	Development	Temporal	|SMD|
EGFR			
Negative	142 (71.4%)	26 (57.8%)	0.287
Positive	57 (28.6%)	19 (42.2%)	0.287
Ki-67 category			
≤20%	100 (27.3%)	20 (21.7%)	0.130
>20%	266 (72.7%)	72 (78.3%)	0.130
Histologic grade			
I	15 (4.0%)	0 (0.0%)	0.290
II	210 (56.6%)	47 (50.5%)	0.122
III	77 (20.8%)	29 (31.2%)	0.240
Unknown	69 (18.6%)	17 (18.3%)	0.008
p53 immunostaining pattern			
mutant/aberrant	12 (3.2%)	3 (3.2%)	0.000
negative-expression	39 (10.5%)	6 (6.5%)	0.146
qualitative-positive-unspecified	1 (0.3%)	0 (0.0%)	0.074
Unknown	309 (83.3%)	83 (89.2%)	0.174
wild-type	10 (2.7%)	1 (1.1%)	0.119
Miller–Payne grade			
MP1–4	224 (60.4%)	50 (53.8%)	0.134
MP5	147 (39.6%)	43 (46.2%)	0.134

Note: Continuous variables are presented as mean ± standard deviation and categorical variables as number (percentage). Percentages were calculated using non-missing observations as the denominator. Absolute SMDs describe cohort differences; categorical SMDs compare each level with all other non-missing levels of the same variable. Variable-specific missingness is detailed in Additional file 1: Table S1. SD, standard deviation; SMD, standardised mean difference; ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; Ki-67, proliferation index; AR, androgen receptor; p53, tumour protein p53; CK5/6, cytokeratin 5/6; EGFR, epidermal growth factor receptor; ADC, apparent diffusion coefficient; TIC, time–intensity curve; MP, Miller–Payne.

Table 2. Univariable Firth regression of candidate predictors in the development cohort.

Variable	*n*	OR	95% CI	*P*
Age (per 1 year)	369	0.989	0.968–1.010	0.294
ER (per 1%)	370	0.973	0.967–0.979	<0.001
PR (per 1%)	368	0.974	0.966–0.982	<0.001
Ki-67 (per 1%)	366	1.031	1.022–1.041	<0.001

Variable	*n*	OR	95% CI	*P*
Tumour maximum diameter (per 1 cm)	369	0.992	0.876–1.121	0.900
ADC (per 0.1 × 10^{-3} mm^2/s)	364	1.156	1.020–1.325	0.023
AR (per 1%)	296	0.991	0.985–0.998	0.012
p53 (per 1%)	190	1.014	1.005–1.023	0.001
CK5/6 positive	327	2.290	1.277–4.147	0.005
EGFR positive	199	4.134	2.192–7.960	<0.001
HER2 status	371	—	—	<0.001
Equivocal/unknown vs negative	371	2.186	1.185–3.997	0.013
Positive vs negative	371	14.457	8.036–27.239	<0.001
Histologic grade	371	—	—	<0.001
I vs II	371	0.048	<0.001–0.367	<0.001
III vs II	371	1.536	0.911–2.595	0.107
Unknown vs II	371	0.806	0.454–1.406	0.451
TIC pattern	371	—	—	0.249
Plateau vs wash-out	371	1.296	0.820–2.042	0.266
Unknown vs wash-out	371	1.723	0.262–11.332	0.550
Wash-in vs wash-out	371	2.518	0.908–7.423	0.076
p53 immunostaining pattern	371	—	—	0.431
Mutant/aberrant vs unknown	371	2.204	0.719–7.181	0.165
Negative-expression vs unknown	371	1.537	0.790–2.986	0.204
Qualitative-positive-unspecified vs unknown	371	0.539	0.004–10.182	0.692
Wild-type vs unknown	371	0.754	0.180–2.591	0.663

Note: OR, odds ratio; CI, confidence interval; ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; Ki-67, proliferation index; ADC, apparent diffusion coefficient; TIC, time–intensity curve; AR, androgen receptor; p53, tumour protein p53; CK5/6, cytokeratin 5/6; EGFR, epidermal growth factor receptor. Global P values are reported for multilevel categorical variables. The grade I versus II lower confidence limit is displayed as <0.001 at the reported precision; this interval was obtained by profile penalised likelihood.

Table 3. Multivariable Firth regression for the MRI-integrated predictor set (*n* = 355).

Variable	OR	95% CI	*P*
Age (years)	1.001	0.972–1.032	0.943
ER (%)	0.982	0.972–0.992	<0.001
PR (%)	0.999	0.986–1.012	0.871
HER2 status	—	—	<0.001
Equivocal/unknown vs negative	3.276	1.536–7.125	0.002
Positive vs negative	14.750	7.062–32.934	<0.001
Ki-67 (%)	1.027	1.012–1.042	<0.001
Histologic grade	—	—	0.460
I vs II	0.165	0.001–1.692	0.152

Variable	OR	95% CI	*P*
III vs II	0.756	0.345–1.643	0.480
Unknown vs II	0.962	0.418–2.166	0.926
Tumour maximum diameter (cm)	0.770	0.641–0.915	0.003
ADC (per 0.1 × 10^{-3} mm^2/s)	1.120	0.924–1.377	0.260
TIC	—	—	0.196
Plateau vs wash-out	1.818	0.951–3.531	0.071
Wash-in vs wash-out	3.218	0.516–20.556	0.213
Unknown vs wash-out	4.112	0.027–84.713	0.455

Note: ORs were estimated by Firth penalised logistic regression; 95% CIs by profile penalised likelihood; and P values by penalised likelihood-ratio tests. Global P values are reported for multilevel categorical variables. Reference categories were HER2-negative, histological grade II, and TIC wash-out. OR, odds ratio; CI, confidence interval; ER, oestrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; ADC, apparent diffusion coefficient; TIC, time–intensity curve; Ki-67, proliferation index.

Predictors and incremental performance across predictor sets

In univariable Firth analysis, ER, PR, HER2, Ki-67, ADC, AR, p53 percentage expression, CK5/6, EGFR, and histological grade were associated with MP5 (Table 2). In the multivariable MRI-integrated model, ER, HER2, Ki-67, and tumour size were associated with MP5 (Table 3). Across the six algorithms, development Core AUCs ranged from 0.8397 to 0.8576, and MRI-integrated AUCs from 0.8447 to 0.8644 (Table 4). Core-to-MRI ΔAUC point estimates ranged from +0.0050 to +0.0124, but all six simultaneous 95% intervals crossed zero and all Holm-adjusted P values exceeded 0.05. Expanded-set AUCs ranged from 0.8379 to 0.8608 and did not establish additional value. Expanded-set findings remained exploratory given marker missingness and the sample-size assessment.

Table 4. Development ROC-AUCs and multiplicity-adjusted Core-to-MRI comparisons across six algorithms.

Algorithm	Core AUC	MRI-integrated AUC	Expanded AUC	Core→MRI ΔAUC (simultaneous 95% CI)	Holm-adjusted P
Elastic-net	0.8517	0.8622	0.8557	+0.0105 (-0.0070 to +0.0287)	0.416
RBF-SVM	0.8397	0.8447	0.8379	+0.0050 (-0.0160 to +0.0258)	0.522
RandomForest	0.8576	0.8642	0.8594	+0.0065 (-0.0058 to +0.0198)	0.416
XGBoost	0.8520	0.8644	0.8592	+0.0124 (-0.0036 to +0.0289)	0.221
CatBoost	0.8546	0.8640	0.8608	+0.0094 (-0.0058 to +0.0247)	0.416
TabPFN	0.8547	0.8628	0.8572	+0.0081 (-0.0036 to +0.0210)	0.416

The AUC for each predictor set is the arithmetic mean of five complete repeat-level OOF ROC-AUCs. Across the five repeats, MRI-integrated AUC SDs were 0.0031 (elastic-net), 0.0070 (RBF-SVM), 0.0029 (Random Forest), 0.0019 (XGBoost), 0.0032 (CatBoost), and 0.0023 (TabPFN); these SDs describe split-to-split variation, not sampling uncertainty. Core-to-MRI ΔAUC is the difference between the corresponding unrounded mean AUCs, not an AUC calculated from probabilities averaged across repetitions. For the six algorithm-specific Core-to-MRI comparisons, Bonferroni simultaneous 95% percentile-bootstrap CIs use tail probabilities $0.05/(2 \times 6)$, and the reported P values are Holm-adjusted. MRI-to-Expanded comparisons are outside this six-test family. Intervals condition on frozen OOF predictions and do not include model-selection or refitting uncertainty.

Post hoc XGBoost candidate selection and model interpretation

MRI-integrated XGBoost had the highest unrounded mean development OOF AUC (0.8644), followed closely by Random Forest (0.8642) and CatBoost (0.8640); these small differences did not establish algorithmic superiority. In the fitted XGBoost, mean absolute SHAP importance was highest for HER2, ER, and Ki-67. The importance, summary, and interaction

panels describe the fitted model on the raw log-odds scale. The waterfall illustrates one temporal prediction (probability 0.483; model baseline probability 0.396)(Fig. 3).

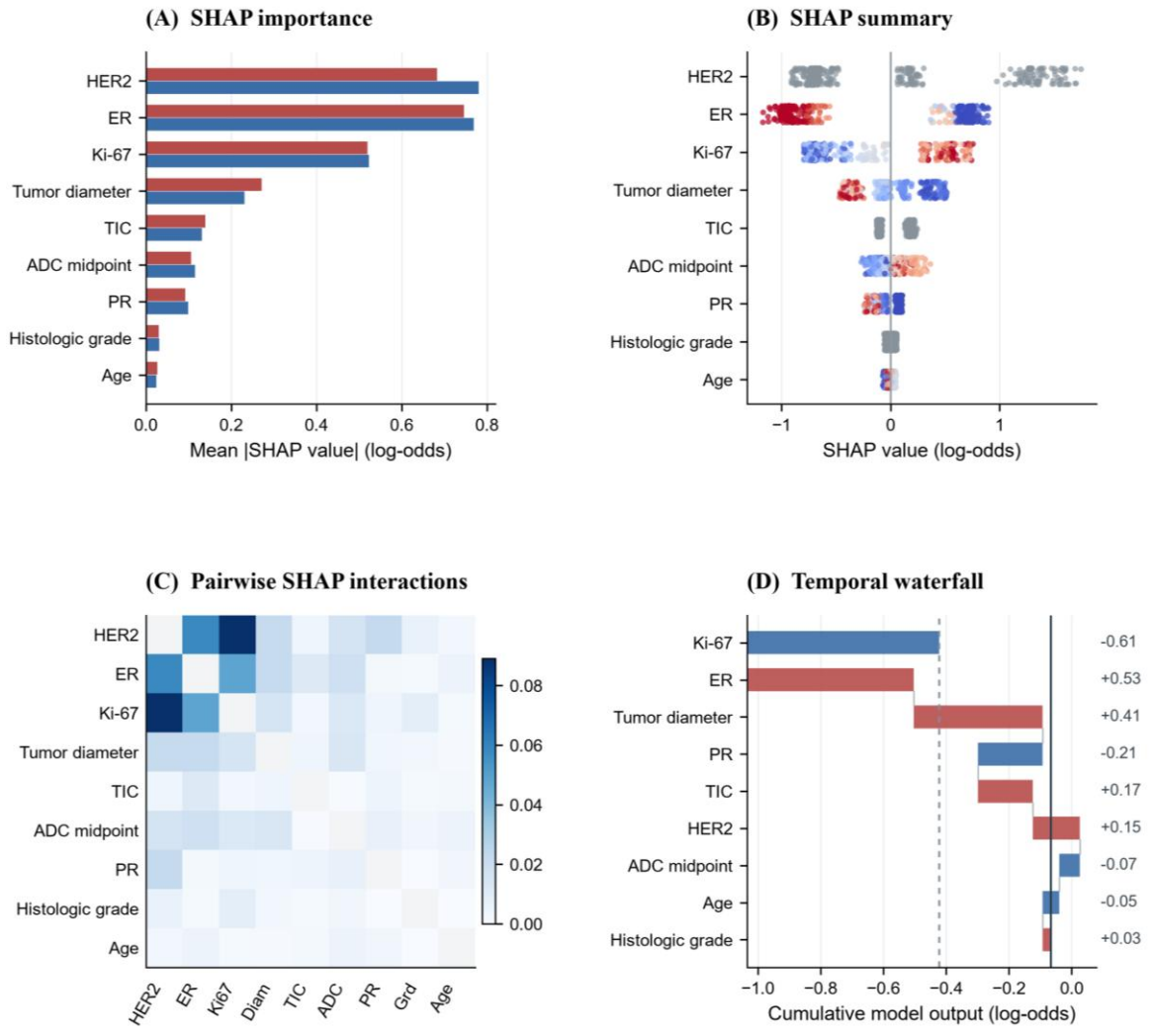


Fig. 3. SHAP interpretation of the development-fitted MRI-integrated XGBoost. (A) Mean absolute predictor-level SHAP importance in 371 development patients. (B) Signed SHAP summary in the same patients. (C) Predictor-level SHAP interaction summary; diagonal terms represent main effects. (D) Waterfall explanation for one temporal patient chosen as closest to the median XGBoost prediction without using outcome (predicted MP5 probability 0.483; baseline probability 0.396). TreeExplainer values are on the raw log-odds scale; encoded features were grouped into clinical predictor concepts. Development panels are in sample and are not estimates of held-out performance.

Same-centre temporal evaluation, calibration, and exploratory decision analysis

In the 93-patient temporal cohort, MRI-integrated XGBoost yielded ROC-AUC 0.8507 (95% CI, 0.7618–0.9222), average precision 0.8128, and Brier score 0.1632 (95% CI, 0.1314–0.1994). Its joint-fit calibration intercept was 0.093 (95% CI, −0.427 to 0.634) and slope was

1.373 (95% CI, 0.888 to 2.395); the wide intervals limit firm conclusions about calibration (Fig. 4). The other five MRI-integrated models had temporal AUCs of 0.8084 (elastic-net), 0.8186 (RBF-SVM), 0.8535 (Random Forest), 0.8186 (CatBoost), and 0.8516 (TabPFN); complete MRI-integrated performance metrics are provided in Additional file 5: Table S5. Exploratory receptor-defined subgroup estimates are provided in Additional file 2: Table S2; small subgroup sizes limit interpretation. In the additional post hoc paired comparison, the XGBoost-minus-Random Forest ΔAUC was −0.0028 (DeLong standard error, 0.0085; 95% CI, −0.0194 to 0.0139; z = −0.328; two-sided P = 0.743). The 10,000-resample paired bootstrap percentile interval was −0.0205 to 0.0155. Neither interval excluded zero; this is not evidence of equivalence. Development and temporal precision–recall, ROC, and calibration curves are compared in Fig. 4. Exploratory decision curves are shown in Additional file 8: Fig. S1; their interpretation is limited by the small temporal cohort, calibration uncertainty, and the absence of a specified clinical decision. XGBoost temporal predicted-probability tertiles contained 3/31 (9.7%), 14/31 (45.2%), and 26/31 (83.9%) MP5 cases, respectively (Cochran–Armitage χ^2 = 34.32, df = 1, P < 0.001; Fig. 5). These cohort-derived cut-offs are not proposed as clinical risk thresholds; the trend test does not establish calibration or clinical utility. In the exploratory temporal Core-to-MRI comparison for XGBoost, the Core-only model had a ROC-AUC of 0.8314 (95% CI, 0.7458–0.9094), compared with 0.8507 for the MRI-integrated model. The paired MRI-minus-Core ΔAUC was +0.0193 (95% percentile-bootstrap CI, −0.0238 to +0.0664). The interval included zero, so these temporal data did not establish an incremental discrimination benefit from MRI (Additional file 7: Table S7). In the complete-case sensitivity analysis, 355/371 development patients (141 MP5) and 92/93 temporal patients (42 MP5) were retained. After complete-case refitting, XGBoost Core-to-MRI ΔAUC was +0.0130 in

development (Core 0.8521, MRI-integrated 0.8651; nominal 95% CI, +0.0006 to +0.0257) and +0.0202 in the temporal cohort (Core 0.8217, MRI-integrated 0.8419; nominal 95% CI, −0.0271 to +0.0724). The point estimates were similar to the full-cohort estimates, but the development sensitivity CI was unadjusted and does not overturn the six-comparison-adjusted primary finding (Additional file 9: Table S8). Excluding the eight patients without an interpretable HER2 IHC score from the frozen-prediction evaluation left 369 development patients (146 MP5) and 87 temporal patients (41 MP5). XGBoost MRI-minus-Core ΔAUC was +0.0123 in development and +0.0231 in the temporal cohort (95% paired bootstrap CI, −0.0222 to +0.0718). This post hoc subset analysis did not refit the models or reassess the HER2 odds ratio (Additional file 10: Table S9).

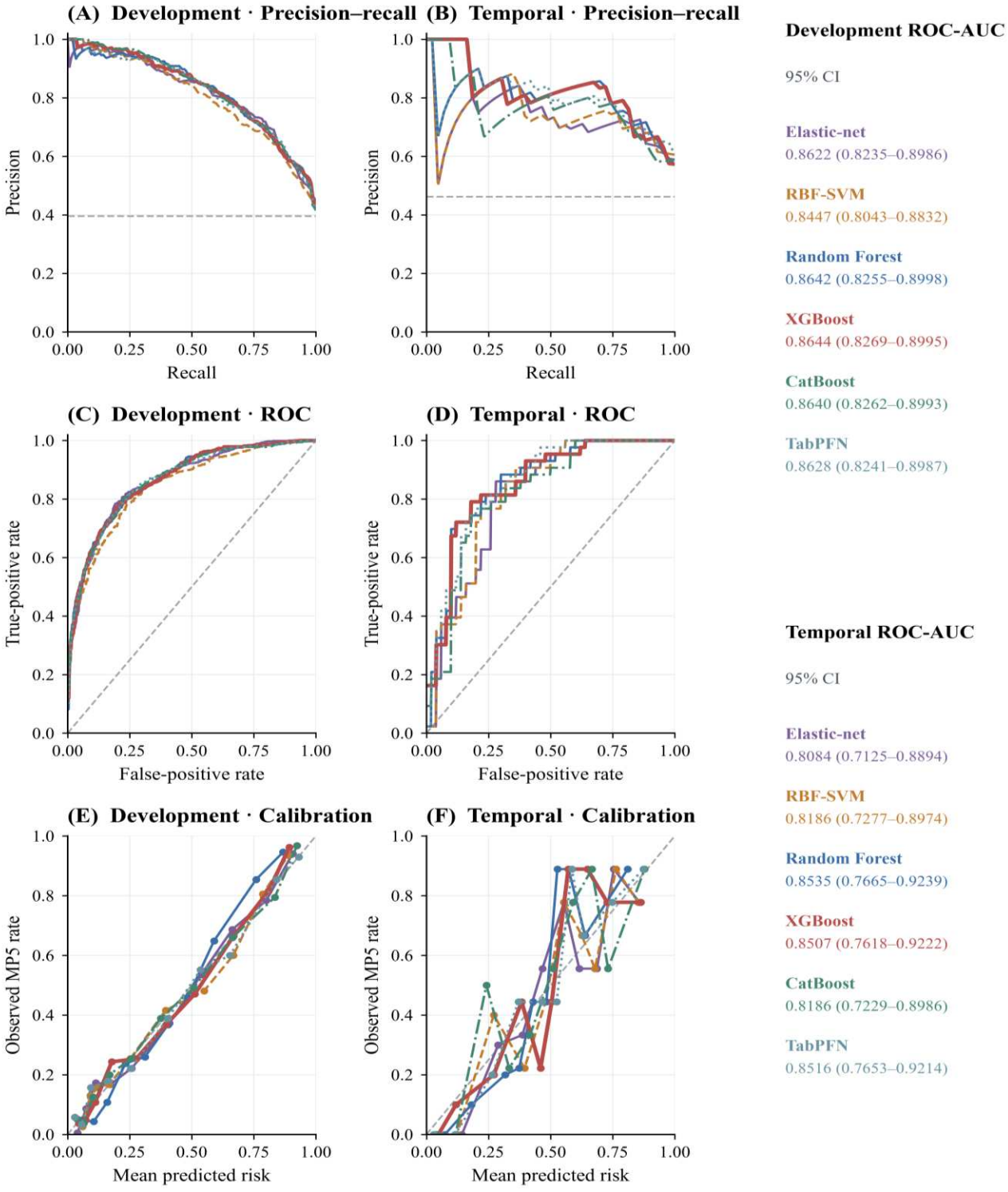

(A) Development · Precision–recall
(B) Temporal · Precision–recall
Precision
Recall
(C) Development · ROC
(D) Temporal · ROC
True-positive rate
False-positive rate
(E) Development · Calibration
(F) Temporal · Calibration
Observed MP5 rate
Mean predicted risk
Development ROC-AUC
95% CI
Elastic-net
0.8622 (0.8235–0.8986)
RBF-SVM
0.8447 (0.8043–0.8832)
Random Forest
0.8642 (0.8255–0.8998)
XGBoost
0.8644 (0.8269–0.8995)
CatBoost
0.8640 (0.8262–0.8993)
TabPFN
0.8628 (0.8241–0.8987)
Temporal ROC-AUC
95% CI
Elastic-net
0.8084 (0.7125–0.8894)
RBF-SVM
0.8186 (0.7277–0.8974)
Random Forest
0.8535 (0.7665–0.9239)
XGBoost
0.8507 (0.7618–0.9222)
CatBoost
0.8186 (0.7229–0.8986)
TabPFN
0.8516 (0.7653–0.9214)

Fig. 4. Six-algorithm comparison for MRI-integrated MP5 prediction. Development results use five complete out-of-fold prediction sets for the same 371 patients: ROC and precision–recall curves were interpolated to common grids and averaged pointwise; ten equal-count calibration-bin coordinates were averaged across repeats. The displayed development AUCs and their patient-level bootstrap CIs were calculated from the five repeat-level AUCs, using shared patient indices across repeats, rather than from the averaged curves. Temporal curves use one frozen prediction per each of 93 patients and ten equal-count calibration bins. Panels compare precision–recall, ROC, and calibration curves in both cohorts. AUC values and 95% confidence intervals are identified in the figure. Temporal model comparisons are descriptive. Exploratory decision curves are provided separately in Additional file 8: Fig. S1.

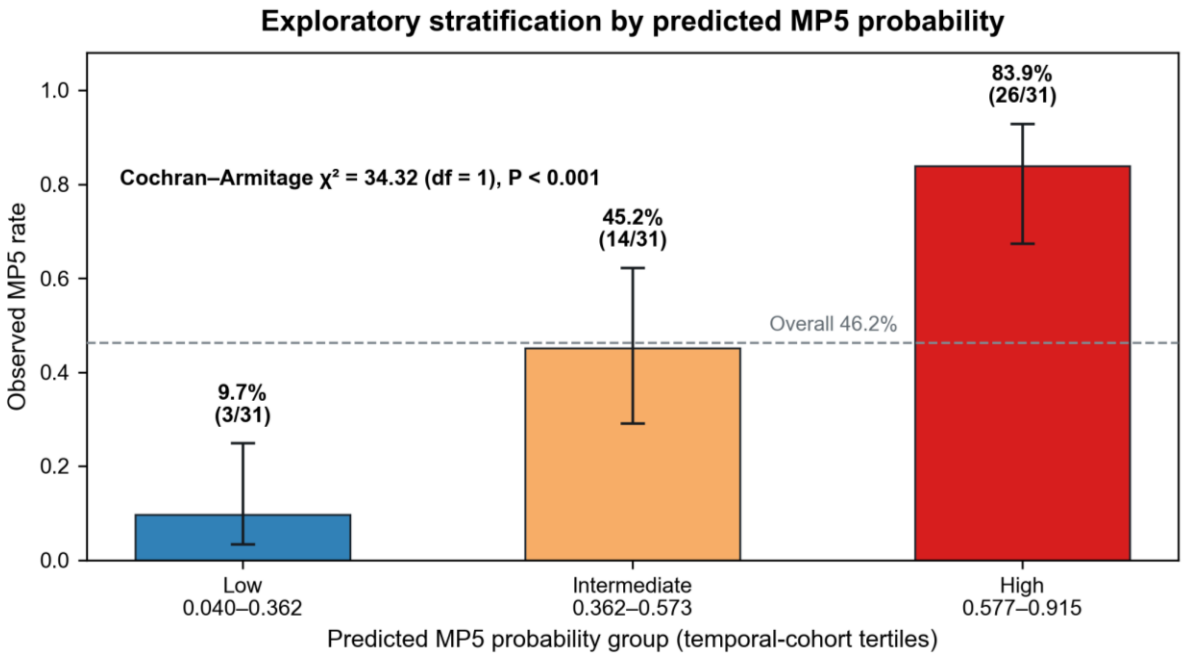


Fig. 5. Exploratory stratification by predicted MP5 probability in the temporal cohort using the MRI-integrated XGBoost model. Tertiles each contain 31 patients; observed MP5 frequencies were 3/31, 14/31, and 26/31. Error bars show Wilson 95% binomial confidence intervals, and the dashed line marks the overall MP5 rate (43/93; 46.2%). The Cochran–Armitage test assessed ordered trend. Tertile cut-offs were derived from these 93 patients, so this display is descriptive and requires independent validation.

Discussion

Clinicopathological predictors provided good MP5 discrimination, whereas adding routine pretreatment MRI measurements produced small, uncertain gains. MRI-integrated XGBoost had the highest mean development AUC by a narrow margin; its temporal performance requires independent evaluation.

The Miller–Payne grading system evaluates response within the primary breast tumour by comparing tumour cellularity in the pretreatment biopsy and post-treatment surgical specimen. MP5 indicates complete disappearance of invasive tumour cells from the primary

tumour bed, although residual ductal carcinoma in situ may remain [4,5]. This endpoint captures the invasive response of the breast lesion but does not establish whether regional lymph nodes are free of residual disease. It is therefore distinct from commonly used definitions of pCR that require the absence of residual invasive disease in both the breast and sampled lymph nodes [3,5]. Accordingly, the present model estimates the probability of an MP5 in the primary breast tumour, rather than pCR across the breast and sampled nodes. The distinction also limits direct numerical comparisons with studies predicting pCR or using other MP-grade groupings. Non-pCR after NAT has been associated with poorer prognosis [24]. Its predictions cannot replace pathological examination or independently support modification of systemic treatment or surgery.

HER2, ER, and Ki-67 were prominent clinicopathological contributors in the fitted XGBoost attribution analysis, consistent with the Firth associations and prior NAT response studies [9,25–28]. Experimental work also supports a functional role for progesterone receptor signalling in ER-positive breast cancer [29], but this does not establish incremental predictive value for MP5. SHAP magnitude depends on model structure and predictor dependence and should not be read as a causal ranking. The interaction panel and single-patient waterfall explain the fitted model and one prediction, respectively; they do not independently validate clinical usefulness [30,31].

The MRI component of this study was deliberately pragmatic. Tumour size, ADC, and TIC pattern were derived from routine pretreatment MRI measurements without radiomic segmentation, automated feature extraction, or specialised post-processing. This reduced feature-extraction demands within this study while restricting the imaging information available for modelling to three measurements. Report-derived measurements, absent image rereading, and

unavailable acquisition details may have introduced measurement noise; the small observed MRI increment therefore cannot be interpreted as evidence that MRI itself contains no additional information. These measurements represent different aspects of tumour phenotype: size describes pretreatment disease burden, ADC provides an indirect measure of water diffusion and tissue cellularity, and TIC pattern characterises contrast-enhancement kinetics. Previous studies have reported associations between these MRI characteristics and pathological response after NAT, although performance varies across populations, molecular subtypes, acquisition protocols, and response definitions [13,14,32].

Adding the three MRI variables produced small positive development ΔAUC point estimates across algorithms, but all six Bonferroni-adjusted simultaneous 95% intervals included zero. Routine pretreatment MRI measurements capture selected aspects of tumour morphology, diffusion, and enhancement, some of which may overlap with tumour biology represented by receptor status, proliferation, and grade. Whether richer imaging features add reproducible information beyond the Core predictor set remains to be tested.

Radiomic and multimodal studies address related but different prediction tasks [7,14,33]. Bi et al. used pretreatment MRI radiomics to distinguish MP grades 1–2 from grades 3–5 [10], whereas our endpoint was MP5 versus MP1–MP4 and our question was the incremental value of three routine pretreatment MRI measurements beyond the Core predictor set. A radiopathomic model combined clinicopathological data, venous-phase contrast-enhanced CT radiomics, and whole-slide pathology features to predict pCR in a chronologically separated validation cohort [34]. Zhou et al. assessed wavelet-transformed MRI radiomics [35], and Zhang et al. integrated early- and peak-phase DCE-MRI radiomics with deep-learning features in 204 development patients, with external validation in 30 patients (AUC 0.825, 95% CI 0.713–0.886) [36].

Separately, He et al. studied clinicopathological prediction of pCR [37]. Differences in outcome, reference model, imaging method, and validation design preclude direct AUC ranking against the present MP5 results.

Our results do not establish that pretreatment MRI is generally uninformative. The observed incremental discrimination from the three routine pretreatment MRI measurements was small, and a benefit beyond clinicopathological predictors was not established. More detailed imaging approaches should demonstrate reproducible improvement over the same clinical reference model in independent populations before their added complexity is considered clinically useful [10,33].

Although AR, p53, CK5/6, and EGFR were individually associated with MP5 in our univariable analyses, previous studies reported inconsistent findings for these markers and NAT response [38–40]. Expanded-set AUCs were lower than the corresponding MRI-integrated AUCs for all six algorithms. This may reflect redundancy with Core variables, but interpretation is limited by marker availability. In the development cohort, missingness varied across AR (20.2%), p53 percentage expression (48.8%), CK5/6 (11.9%), and EGFR (46.4%); qualitative p53 was coded Unknown in 83.3%. Missingness and uncertain coding may have attenuated the apparent contributions of these markers. These exploratory findings should therefore not be interpreted as evidence that the biomarkers are uninformative; they require reassessment in larger cohorts with more complete measurement. The two p53 variables represent different reporting channels: a percentage-only report could supply a numeric value while leaving the qualitative category Unknown. Thus, Unknown qualitative coding should not be interpreted as evidence that p53 testing was not performed.

The close MRI-integrated AUCs across algorithms emphasise the shared finding: any incremental discrimination from the three routine MRI measurements was small and uncertain. Selecting the largest of six correlated development AUCs can make the selected estimate optimistic [41]. The post hoc temporal XGBoost–Random Forest comparison did not establish superiority or equivalence, and the already available temporal outcomes cannot independently validate the algorithm choice. The XGBoost calibration slope was 1.373, with a 95% CI spanning 1, and its Brier score was 0.1632. Absolute-risk use and threshold-based decisions require evaluation in genuinely independent data [18,22,42].

This study compared three predictor sets and six algorithms using repeated nested cross-validation and described discrimination, calibration, prediction error, and exploratory decision curves in later same-centre patients. It provides a pragmatic benchmark for testing whether more detailed imaging adds value beyond the Core predictor set. Small AUC gains should not be assumed to improve clinical decisions; independent assessment of calibration and decision utility is needed before individual use.

This retrospective single-centre study included only female patients, and same-centre temporal assessment cannot establish transportability across institutions. The temporal cohort was small (93 patients; 43 MP5 events), leaving wide uncertainty around discrimination, calibration, and decision utility. XGBoost was selected after the unified development comparison when temporal results were already available; its temporal estimates are exploratory and potentially affected by selection and reporting bias. The paired bootstrap resampled patients conditional on fixed OOF predictions from models with extensively overlapping training sets. Its intervals and P values do not represent full modelling-pipeline uncertainty, including new splits, preprocessing, tuning, refitting, or post hoc selection; Bonferroni and Holm adjustment address

multiplicity only. MRI predictors came from routine reports without reader-level reproducibility assessment or complete acquisition metadata. Anti-HER2 therapy and other treatment details were unavailable; this may have affected estimated associations and temporal performance, but the direction of any effect cannot be determined from the available data. Expanded-set markers had substantial or uneven missingness, limiting interpretation of their incremental predictive value. More generally, data quality deserves attention when comparing modelling algorithms [43]. SHAP plots explain fitted model output rather than held-out performance, and temporal tertile cut-offs were derived from the same cohort. Complete-case refitting excluded only 16 development patients and one temporal patient; it could not resolve ambiguity in the recorded HER2, grade, or TIC categories. The separate eight-case HER2 IHC exclusion analysis assessed discrimination only; it left 75 unresolved IHC 2+ records in the cohort and could not establish whether the equivocal/unknown category's regression association reflected tumour biology, testing completeness, recording practices, or treatment differences. Because original tuning choices were reused, this was not a fully retuned complete-case pipeline. Larger independent multicentre studies should validate discrimination, calibration, and decision utility and test whether richer imaging features improve on the Core predictor set.

Conclusions

Clinicopathological variables provided good discrimination for MP5 after NAT. Routine MRI measurements produced small positive Core-to-MRI ΔAUC point estimates, but all six Bonferroni-adjusted simultaneous 95% intervals included zero. The post hoc XGBoost candidate had a temporal AUC of 0.8507 (95% CI, 0.7618–0.9222), but its calibration estimates were imprecise and the already available temporal results cannot independently validate its selection.

Independent validation of discrimination, calibration, and clinical utility is needed before patient-level use.

List of abbreviations

ADC: apparent diffusion coefficient

AR: androgen receptor

CI: confidence interval

CK5/6: cytokeratin 5/6

EGFR: epidermal growth factor receptor

ER: oestrogen receptor

HER2: human epidermal growth factor receptor 2

IHC: immunohistochemistry

MP: Miller–Payne

MP5: Miller–Payne grade 5 in the primary breast tumour

MRI: magnetic resonance imaging

NAT: neoadjuvant therapy

OOF: out-of-fold

OR: odds ratio

pCR: pathological complete response

PR: progesterone receptor

RBF-SVM: radial basis function support vector machine

ROC-AUC: area under the receiver operating characteristic curve

SHAP: SHapley Additive exPlanations

TIC: time–intensity curve

TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – Artificial Intelligence

XGBoost: eXtreme Gradient Boosting

ypT0/is ypN0: no residual invasive carcinoma in the breast or sampled regional lymph nodes; residual in situ carcinoma may remain

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Review Committee of Baotou Cancer Hospital (approval no. 2025 Ethics Review No. 018). This was a retrospective study based on de-identified medical records and pre-treatment MRI data; the requirement for written informed consent from individual participants was therefore waived by the Ethics Review Committee.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. Code for the unified six-algorithm modelling workflow is supplied as Additional file 11; the archive does not contain patient-level data.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Joint-Fund Project for Scientific Research of Public Hospitals, Inner Mongolia Academy of Medical Sciences (No. 2024GLLH0758); the National Natural Science Foundation of China (No. 62271132); the Natural Science Foundation of Inner Mongolia (No. 2025MS03116); and the 2025/2026 Inner Mongolia Key Laboratory of Life Health and Bioinformatics Project (2025KYPT0135/2026KYPT0062).

Authors' contributions

XZ and KT contributed equally to this work and share first authorship. XZ and KT performed the investigation, data curation, software development, formal analysis, and visualisation and drafted the manuscript. ZZ and CH contributed resources and software development. XYZ and RZ performed formal analysis and validation. HZ and JH contributed biological interpretation and reviewed and edited the manuscript. QY and YX contributed methodology and validation. HM supervised the study, acquired funding, contributed methodology, reviewed and edited the manuscript, and administered the project. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Additional files

Additional file 1: Table S1. Variable-specific missingness of candidate predictors in the development cohort.

Additional file 2: Table S2. Exploratory MRI-integrated XGBoost performance by receptor-defined group in the temporal cohort.

Additional file 3: Supplementary methods and Table S3. Six-algorithm hyperparameter grids, post hoc XGBoost candidate refit, SHAP procedures, temporal calibration intervals and valid bootstrap counts, and performance-curve construction.

Additional file 4: Table S4. Software versions

Additional file 5: Table S5. Development and temporal MRI-integrated performance for the six algorithms.

Additional file 6: Table S6. TRIPOD+AI reporting checklist.

Additional file 7: Table S7. Exploratory temporal performance and probability tertiles for the XGBoost candidate.

Additional file 8: Fig. S1. Exploratory decision curves for the MRI-integrated models in the development and temporal cohorts.

Additional file 9: Table S8. Complete-case XGBoost Core-to-MRI sensitivity analysis in the development and temporal cohorts.

Additional file 10: Table S9. HER2 IHC missing-score exclusion sensitivity analysis using frozen XGBoost predictions.

Additional file 11: Code archive for the unified six-algorithm modelling workflow (ZIP).

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229-263. https://doi.org/10.3322/caac.21834

[2] Hong R, Xu B. Breast cancer: an up-to-date review and future perspectives. Cancer Commun (Lond). 2022;42:913-936. https://doi.org/10.1002/cac2.12358

[3] Provenzano E. Neoadjuvant Chemotherapy for Breast Cancer: Moving Beyond Pathological Complete Response in the Molecular Age. Acta Med Acad. 2021;50:88-109. https://doi.org/10.5644/ama2006-124.328

[4] Ogston KN, Miller ID, Payne S, Hutcheon AW, Sarkar TK, Smith I, et al. A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. Breast. 2003;12:320-327. https://doi.org/10.1016/S0960-9776(03)00106-1

[5] Park CK, Jung WH, Koo JS. Pathologic Evaluation of Breast Cancer after Neoadjuvant Therapy. J Pathol Transl Med. 2016;50:173-180. https://doi.org/10.4132/jptm.2016.02.02

[6] Meti N, Saednia K, Lagree A, Tabbarah S, Mohebpour M, Kiss A, et al. Machine learning frameworks to predict neoadjuvant chemotherapy response in breast cancer using clinical and pathological features. JCO Clin Cancer Inform. 2021;5:66-80. https://doi.org/10.1200/CCI.20.00078

[7] Hajri R, Aboudaram C, Lassau N, Assi T, Antoun L, Ribeiro JM, et al. Prediction of Breast Cancer Response to Neoadjuvant Therapy with Machine Learning: A Clinical, MRI-

Qualitative, and Radiomics Approach. Life (Basel). 2025;15:1165. https://doi.org/10.3390/life15081165

[8] Krasniqi E, Filomeno L, Arcuri T, Ferretti G, Gasparro S, Fulvi A, et al. Multimodal deep learning for predicting neoadjuvant treatment outcomes in breast cancer: a systematic review. Biol Direct. 2025;20:72. https://doi.org/10.1186/s13062-025-00661-8

[9] Hou W, Yao Q, Niu DF, Xue WC. Clinicopathological characteristics related to Miller/Payne grading system of breast carcinoma after neoadjuvant therapy and establishment of novel prediction models. Zhonghua Bing Li Xue Za Zhi. 2022;51:743-748. [Article in Chinese]. https://doi.org/10.3760/cma.j.cn112151-20220413-00277

[10] Bi C, Chen A, Ran F, Hu Z, Sun S, Wang R, et al. Pretreatment MRI radiomics for predicting pathological Miller-Payne grading in breast cancer following neoadjuvant chemotherapy. Cancer Imaging. 2026;26:25. https://doi.org/10.1186/s40644-026-00990-5

[11] Song D, Man X, Jin M, Li Q, Wang H, Du Y. A decision-making supporting prediction method for breast cancer neoadjuvant chemotherapy. Front Oncol. 2021;10:592556. https://doi.org/10.3389/fonc.2020.592556

[12] Janssen LM, den Dekker BM, Gilhuijs KGA, van Diest PJ, van der Wall E, Elias SG. MRI to assess response after neoadjuvant chemotherapy in breast cancer subtypes: A systematic review and meta-analysis. NPJ Breast Cancer. 2022;8:107. https://doi.org/10.1038/s41523-022-00475-1

[13] Surov A, Pech M, Meyer HJ, Bitencourt AGV, Fujimoto H, Baxter GC, et al. Evaluation of pretreatment ADC values as predictors of treatment response to neoadjuvant chemotherapy in patients with breast cancer: a multicenter study. Cancer Imaging. 2022;22:68. https://doi.org/10.1186/s40644-022-00501-2

[14] He M, Su J, Ruan H, Song Y, Ma M, Xue F. Nomogram based on quantitative dynamic contrast-enhanced magnetic resonance imaging, apparent diffusion coefficient, and clinicopathological features for early prediction of pathologic complete response in breast

cancer patients receiving neoadjuvant chemotherapy. Quant Imaging Med Surg. 2023;13:4089-4102. https://doi.org/10.21037/qims-22-869

[15] Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science. 2024;383:164-167. https://doi.org/10.1126/science.adg8538

[16] Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A guide to cross-validation for artificial intelligence in medical imaging. Radiol Artif Intell. 2023;5:e220232. https://doi.org/10.1148/ryai.220232

[17] 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. https://doi.org/10.1136/bmj-2023-078378

[18] Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. BMJ. 2024;384:e074820. https://doi.org/10.1136/bmj-2023-074820

[19] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38:1276-1296. https://doi.org/10.1002/sim.7992

[20] Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, et al. Accurate predictions on small data with a tabular foundation model. Nature. 2025;637:319-326. https://doi.org/10.1038/s41586-024-08328-6

[21] Dunias ZS, Van Calster B, Timmerman D, Boulesteix AL, van Smeden M. A comparison of hyperparameter tuning procedures for clinical prediction models: A simulation study. Stat Med. 2024;43:1119-1134. https://doi.org/10.1002/sim.9932

[22] Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. https://doi.org/10.1186/s41512-019-0064-7

[23] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-845. https://doi.org/10.2307/2531595

[24] Resende U, Cabello C, Ramalho SOB, Zeferino LC. Prognostic assessment of breast carcinoma submitted to neoadjuvant chemotherapy with pathological non-complete response. BMC Cancer. 2019;19:601. https://doi.org/10.1186/s12885-019-5812-0

[25] Jain P, Doval DC, Batra U, Goyal P, Bothra S, Agarwal C, et al. Ki-67 labeling index as a predictor of response to neoadjuvant chemotherapy in breast cancer. Jpn J Clin Oncol. 2019;49:329-338. https://doi.org/10.1093/jjco/hyz012

[26] Zhang Y, Qiu S, Wang X. Factors influencing pathological complete response after neoadjuvant chemotherapy in breast cancer: A single-center retrospective study focusing on ER and HER-2 status. Int J Breast Cancer. 2026;2026:6694023. https://doi.org/10.1155/ijbc/6694023

[27] Gianni L, Pienkowski T, Im YH, Tseng LM, Liu MC, Lluch A, et al. 5-year analysis of neoadjuvant pertuzumab and trastuzumab in patients with locally advanced, inflammatory, or early-stage HER2-positive breast cancer (NeoSphere): A multicentre, open-label, phase 2 randomised trial. Lancet Oncol. 2016;17:791-800. https://doi.org/10.1016/S1470-2045(16)00163-7

[28] Chen H, Gui X, Zhou Z, Su F, Gong C, Li S, et al. Distinct ER and PR expression patterns significantly affect the clinical outcomes of early HER2-positive breast cancer: A real-world analysis of 871 patients treated with neoadjuvant therapy. Breast. 2024;75:103733. https://doi.org/10.1016/j.breast.2024.103733

[29] Scabia V, Ayyanan A, De Martino F, Agnoletto A, Battista L, Laszlo C, et al. Estrogen receptor positive breast cancers have patient specific hormone sensitivities and rely on progesterone receptor. Nat Commun. 2022;13:3127. https://doi.org/10.1038/s41467-022-30898-0

[30] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67. https://doi.org/10.1038/s42256-019-0138-9

[31] Brankovic A, Cook D, Rahman J, Delaforce A, Li J, Magrabi F, et al. Clinician-informed XAI evaluation checklist with metrics (CLIX-M) for AI-powered clinical decision support systems. NPJ Digit Med. 2025;8:364. https://doi.org/10.1038/s41746-025-01764-2

[32] Bozer A, Yılmaz C, Çetin Tunçez H, Kocatepe Çavdar D, Adıbelli ZH. MR imaging features predictive of pathologic complete response and survival outcomes for breast cancer patients undergoing neoadjuvant chemotherapy. Magn Reson Med Sci. 2025;24:mp.2024-0137. https://doi.org/10.2463/mrms.mp.2024-0137

[33] Li W, Partridge SC, Newitt DC, Steingrimsson J, Marques HS, Bolan PJ, et al. Breast multiparametric MRI for prediction of neoadjuvant chemotherapy response in breast cancer: The BMMR2 challenge. Radiol Imaging Cancer. 2024;6:e230033. https://doi.org/10.1148/rycan.230033

[34] Zhang J, Wu Q, Yin W, Yang L, Xiao B, Wang J, et al. Development and validation of a radiopathomic model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer patients. BMC Cancer. 2023;23:431. https://doi.org/10.1186/s12885-023-10817-2

[35] Zhou J, Lu J, Gao C, Zeng J, Zhou C, Lai X, et al. Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in MRI. BMC Cancer. 2020;20:100. https://doi.org/10.1186/s12885-020-6523-2

[36] Zhang Y, Cai J, Cui C, Qi S, Zhao D. Predicting breast cancer response to neoadjuvant therapy by integrating radiomic and deep-learning features from early-and-peak phases of DCE-MRI. BMC Cancer. 2025;25:1747. https://doi.org/10.1186/s12885-025-15095-8

[37] He C, Yu T, Yang L, He L, Zhu J, Chen J. Clinical prediction of pathological complete response in breast cancer: a machine learning study. BMC Cancer. 2025;25:933. https://doi.org/10.1186/s12885-025-14335-1

[38] Lee SM, Lee JE, Nam SJ, Kim SW, Yu J, Chae BJ, et al. A practical immunohistochemistry-based model for predicting pathologic complete response in estrogen receptor-strong positive and HER2-negative breast cancer. J Breast Cancer. 2026;29:128-140. https://doi.org/10.4048/jbc.2025.0242

[39] da Silva JL, Rodrigues FR, de Mesquita GG, Fernandes PV, Thuler LCS, de Melo AC. Triple-negative breast cancer: Assessing the role of immunohistochemical biomarkers on neoadjuvant treatment. Breast Cancer (Dove Med Press). 2021;13:31-44. https://doi.org/10.2147/BCTT.S287320

[40] Lee EG, Lee DE, Kim HH, Han JH, Lee S, Kang HS, et al. Androgen receptor as a predictive marker for pathologic complete response in hormone receptor-positive and HER-2-negative breast cancer with neoadjuvant chemotherapy. Cancer Res Treat. 2023;55:542-550. https://doi.org/10.4143/crt.2022.834

[41] Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65:e2200302. https://doi.org/10.1002/bimj.202200302

[42] Vickers AJ, Woo S. Decision curve analysis in the evaluation of radiology research. Eur Radiol. 2022;32:5787-5789. https://doi.org/10.1007/s00330-022-08685-8

[43] Hu Y, Zhang X, Slavin V, Belsti Y, Tiruneh SA, Callander E, et al. Beyond comparing machine learning and logistic regression in clinical prediction modelling: Shifting from model debate to data quality. J Med Internet Res. 2025;27:e77721. https://doi.org/10.2196/77721