

Development and external evaluation of MRI and clinical models for pathological complete response in breast cancer

Tianyi Yu

tianyi.1.yu@kcl.ac.uk

King's College London

Research Article

Keywords: Breast cancer, Neoadjuvant therapy, Magnetic resonance imaging, Multimodal prediction, Pathological complete response, External validation, Calibration

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11038113/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Whether fixed pretreatment MRI signal summaries improve pathological complete response probability prediction beyond age and receptor status was examined across two trial cohorts. External probability accuracy and the stability of individual predictions were assessed to inform the requirements for treatment-response communication.

Methods

Models were developed retrospectively using public, deidentified I-SPY2 data and evaluated externally using I-SPY1 data. Twelve baseline dynamic contrast-enhanced MRI summaries were combined with age, hormone receptor status and HER2 status. Clinical, MRI and fusion models were fitted by penalized logistic regression with nested five-fold cross-validation. All pipelines were frozen before external prediction. The paired external Brier score difference, clinical minus fusion, was estimated with a 95% interval from 2000 patient bootstrap resamples conditional on the fitted models. Discrimination, calibration and conventional MRI comparisons were assessed. Development and external patients were independently resampled in a further 1000 replicates, with preprocessing, tuning and fitting repeated.

Results

Models were developed in 982 patients with 316 complete responses. External evaluation was performed in 166 patients with 48 responses after five missing original outcomes and one prespecified phase-annotation discrepancy were excluded. Internal Brier scores of 0.20143 and 0.19975 were obtained for clinical and fusion models, respectively. Corresponding external scores of 0.17543 and 0.18454 were obtained, with a difference of -0.00911 (95% interval -0.02087 to 0.00218). External areas under the receiver operating characteristic curve were 0.773 for clinical, 0.528 for MRI and 0.707 for fusion models. Mean predicted response probabilities exceeded the observed 28.9% frequency. With repeated development, the median external Brier difference was -0.01598 (95% interval -0.07367 to 0.00646). Mean absolute prediction changes were 3.1, 8.4 and 7.3 percentage points for clinical, MRI and fusion models.

Conclusions

Improved external response probability prediction was not established after these fixed pretreatment MRI summaries were added. The MRI increment should be assessed against clinical and conventional imaging comparators, with calibration and uncertainty reported before treatment-response communication is considered. Neither equivalence nor clinical utility was established.

Background

Treatment response can be assessed before surgery during neoadjuvant therapy, and the extent of local treatment may be reduced. Its use is informed by tumour stage, grade and receptor status [1]. Pathological complete response (pCR) is associated with favourable subsequent outcomes, particularly in triple-negative and HER2-positive disease [2, 3, 4]. A pretreatment estimate of pCR could therefore inform discussions of expected response, provided that the probability is sufficiently accurate for the population and treatment setting in which it is communicated.

The primary research question was whether pCR probabilities could be predicted more accurately in a separate trial cohort when 12 fixed pretreatment dynamic contrast-enhanced (DCE) MRI signal summaries were added to age and receptor status. The clinical comparator comprised age, hormone receptor (HR) status and human epidermal growth factor receptor 2 (HER2) status. The question was evaluated through the paired external Brier score difference between clinical and fusion models. Development was undertaken in I-SPY2, followed by evaluation of frozen predictions in I-SPY1. The intended application was pretreatment response communication; effects of choosing one regimen over another were not estimated.

Tumour extent and enhancement can be characterized by breast MRI to complement clinical assessment. The predictive relevance of MRI measurements and their changes during therapy has been investigated in the I-SPY studies [5, 6]. For pretreatment prediction, intratumoral and peritumoral texture, extensive image-feature sets and multiparametric MRI have been evaluated using different validation designs [7, 8, 9]. These designs matter when an imaging increment is interpreted: performance in a held-out sample from one institution does not establish performance in a different clinical population.

In recent multimodal studies, MRI and clinical representations have been integrated in pooled cohorts [10], clinical information has been combined with digital pathology and molecular profiles [11], and MRI has been combined with biopsy whole-slide images and clinical variables [12]. Predictions have also been updated as information becomes available during therapy [13]. These studies support further investigation of complementary data, but their findings depend on the modalities, prediction time and patients included. A compact baseline MRI representation should be assessed on its own merits, with direct comparison against the same clinical information in the same patients.

For response communication, discrimination alone is insufficient. The area under the receiver operating characteristic curve (AUROC) measures ranking, whereas calibration concerns agreement between predicted probabilities and observed frequencies; both are relevant to overall probability error [14, 15]. External performance may change with the distribution of patient characteristics, treatment practice and measurement conditions [16, 17, 18]. Internal and external validation consequently address different questions [19]. In addition, a probability assigned to an individual may vary when the model is developed again in another sample from the same population [20].

These issues can be examined across established breast cancer cohorts using public imaging resources [21, 22, 23]. Alongside the primary external comparison, calibration, recorded treatment, receptor subgroups and conventional MRI measurements were assessed. Uncertainty from repeated model development and the stability of individual predictions were examined in a subsequent analysis. The study was designed to

estimate the added value of a defined MRI representation, without assuming that multimodal fusion would improve prediction.

Methods

Study design and prediction setting

A retrospective secondary analysis of public, deidentified data was performed without new recruitment or participant contact. The prediction time was defined as after the treatment plan was known and before the first dose; the outcome was pCR at surgery. The potential application was discussion of expected response between clinicians and patients. The models were evaluated as research tools, without a counselling intervention or clinical deployment.

The statistical analysis plan was recorded locally on 13 September 2026 before the first model fit. The conventional MRI comparison was added after the initial clinical benchmark and before any MRI outcome model was fitted. These dated project specifications were not publicly preregistered. Two technical corrections were documented before the primary MRI fits, and model objects and checksums were frozen before external prediction. Reporting was guided by TRIPOD + AI, with limitations assessed against PROBAST + AI [24, 25]. Detailed methods and results are provided in Additional file 1. The original protocol, source audit, executable code and analysis inputs are provided in Additional file 2; records and code for the repeated-development analysis are provided in Additional file 3.

Data sources and populations

Original clinical and outcome workbooks were obtained from The Cancer Imaging Archive (TCIA), and baseline image derivatives were obtained from BreastDCEDL, Zenodo release 18114231, with versioned GitHub metadata [26, 27, 28, 29, 30]. These resources were selected for public access, linked outcomes and separate trial populations. The 982 I-SPY2 participants in the derivative image release were considered for development. ACRIN 6698 records were drawn from the I-SPY2 population and were not counted as a separate external cohort. Images from 172 I-SPY1 participants were available in the release. I-SPY1 was excluded from supervised fitting, feature selection, tuning and recalibration.

Women with high-risk or locally advanced breast cancer were enrolled in the parent studies in specialist, multicentre settings. I-SPY1 recruitment was undertaken between May 2002 and March 2006; recruitment between May 2010 and November 2016 was described in the published I-SPY2 imaging analysis [5, 6]. Several investigational treatment combinations are evaluated within the adaptive I-SPY2 platform [31]. Exact accrual dates and site membership for the derivative subsets were not verified. Individual treatment and surgery dates were not used to define a fixed calendar prediction horizon. Because participating institutions may overlap between trials, the evaluation was interpreted as separate-trial transport rather than geographical independence. Sample sizes were determined by eligible public records; no prospective precision-based sample-size calculation was performed. Limited external precision was retained as a study limitation [32, 33].

Outcome and eligibility

The binary endpoint was source-recorded pCR after neoadjuvant treatment, defined in the I-SPY reports as no residual invasive disease in the breast and regional nodes [3, 6]. Outcomes and receptor labels were obtained from the original clinical workbooks. Missing pCR was not inferred from survival, recurrence, residual size or derivative coding. Outcome assessment was performed in the original studies; pathology slides, assessor characteristics and patient-level assessment blinding were unavailable for this reanalysis.

Image sets were considered eligible if unique clinical linkage, an observed binary endpoint, three readable contrast-phase arrays, a compatible nonempty tumour mask, finite values and usable positive precontrast tumour signal were available. All three phases were obtained from one baseline examination. One I-SPY1 record with a documented phase-index discrepancy was excluded from the primary analysis and included under the specified alternative mapping in a sensitivity analysis. Clinical and fusion models were compared in exactly the same patients.

Technical checks were performed without outcome columns. Stored I-SPY2 arrays were reconciled with metadata dimension fields, without changes to phase indices, patient identifiers or voxels. Both binary mask encodings, 0/1 and 0/255, were accepted with identical nonzero foreground interpretation. Cohort inclusion was determined by a fixed identifier whitelist rather than filename cohort tokens. Segmentation, motion and physiological contrast timing were not independently assessed by these numerical checks. The supplied masks were not reannotated or reviewed by a breast radiologist for this study. Image quality assessment and measurement reproducibility are emphasized in imaging biomarker guidance, and these unassessed components were retained as limitations [34]. Full corrections and source checksums are provided in Additional files 1 and 2.

Predictors and preprocessing

Age, hormone receptor (HR) status, HER2 status and the HR-by-HER2 interaction were included in the clinical model. Age was represented by a cubic B-spline with four training-sample quantile knots and linear extrapolation. Missing age and receptor values were imputed with training-fold medians and modes. Age spline outputs were standardized within the training fold. Stage and grade were not included because harmonized inputs had not been verified; the comparator was therefore restricted to limited common information rather than a full clinical assessment.

Source receptor categories were retained without imposing new staining thresholds. Differences between the original I-SPY1 and I-SPY2 HR definitions are described in the derivative publication [23]. In the I-SPY1 dictionary, pretreatment receptor fields include an updated HER2 field containing central results when community results were absent. The resulting historical labels may differ from the information available at the intended consultation.

MRI features were fixed before outcome modelling. Within each supplied tumour mask, precontrast signal was normalized to the median of its strictly positive values. Early enhancement, late enhancement and delayed signal change were calculated using the larger of absolute precontrast signal and one tenth of that

median as the voxelwise denominator. Nine features were obtained from the 10th, 50th and 90th percentiles of these three arrays. Two features were obtained from the 10th and 90th percentiles of normalized precontrast signal; the fraction of voxels with late signal below early signal was included as the remaining feature. An inverse hyperbolic sine transformation was applied to the first 11 features; all 12 were standardized within training folds. Formulas are given in Additional file 1.

Signal distributions were summarized without spatial texture, physical tumour volume or pharmacokinetic fitting. Placeholder unit affines in the derivative headers were not treated as physical geometry. The representation was not claimed to be an Image Biomarker Standardization Initiative radiomics implementation [35]. No outcome-guided feature screening or response-supervised pretrained model was used. Class weighting and synthetic oversampling were not applied. This approach retained the observed outcome distribution for probability estimation; oversampling and related imbalance corrections can distort the calibration of logistic risk models [36].

Model development and evaluation

Clinical, MRI-only and concatenated fusion models were fitted by L2-penalized logistic regression with the same tuning budget. Use of a common estimator allowed the contribution of the additional inputs to be compared without changing the learning method. More complex algorithms have not consistently outperformed logistic regression in clinical prediction studies [37]. Candidate inverse penalties of 0.01, 0.1, 1, 10 and 100 were compared by mean validation Brier score. The value with the lowest score was selected, with stronger regularization preferred in exact ties.

Five outer stratified patient folds were used for internal evaluation, with five inner folds within each outer training sample. Preprocessing was fitted within the corresponding training partition. Common folds were used across models, and one outer-fold prediction was obtained for each participant from a model fitted without that participant. Tuning was separated from performance assessment by nested validation [38]. This separation is particularly relevant to imaging studies, in which feature selection outside validation has been shown to produce optimistic estimates [39]. The split seed was 20260913, with inner seeds increased by outer fold number. A final penalty was selected using development-only five-fold cross-validation, after which each model was fitted to all eligible development data and frozen. External probabilities were generated directly from the saved pipelines.

The primary estimand was defined as $\text{Brier}(\text{clinical}) - \text{Brier}(\text{fusion})$ in the external cohort; positive values favour fusion. A two-sided 95% percentile interval was estimated from 2000 paired patient bootstrap resamples, with seed 20260914. The pairing of outcomes and both predictions was preserved during resampling. The interval was conditioned on the fitted models; development-sample uncertainty and unidentified site clustering were not incorporated. No clinically important Brier difference was prespecified, and an interval spanning zero was not interpreted as equivalence.

AUROC, average precision, log loss, calibration-in-the-large and calibration slope were assessed as secondary metrics [14, 15]. Average precision summarized the precision–recall relationship and was interpreted alongside the event frequency, on which precision depends [40]. It was not calculated as

trapezoidal precision–recall area. Calibration-in-the-large was estimated with prediction log odds as a logistic offset; calibration slope was estimated by logistic regression with a freely estimated intercept. Probabilities were clipped to [0.000001, 0.999999] only for logarithmic calculations. Predictions were not updated by diagnostic calibration regressions. Descriptive external intervals were estimated from 2000 patient resamples of fixed predictions; failed diagnostic fits were recorded. Internal metrics and secondary intervals were reported descriptively, without multiplicity-adjusted inference.

For calibration plots, predictions were divided into five approximately equal-sized groups, with ties retained. Mean predicted probability, observed frequency, Wilson intervals and prediction distributions were displayed. The intervals described uncertainty in grouped frequencies, not in individual true probabilities. No classification threshold was selected. Decision-curve analysis requires a clinical action and threshold probabilities that reflect the relative consequences of false predictions [41]. These were not specified for the communication setting, so decision curves were not estimated.

Development and prediction stability

An additional analysis was specified after the original external results had been inspected to assess uncertainty from the final model-development procedure [20]. In each of 1000 replicates, 982 development patients and 166 external patients were sampled independently with replacement. Identical development and external draws were used for all three models. Distinct sampled development patients were stratified by outcome into five tuning folds; every duplicate copy was assigned to the same fold. Bootstrap multiplicity was retained in training and validation. Imputation, age spline knots, scaling, five-candidate penalty selection and final fitting were repeated using development observations only. External labels were used solely for evaluation. The resample count and procedure were fixed before the complete run.

The Brier contrast was computed within each paired external resample and summarized by its median and central 95% percentile interval. The original frozen-model comparison was retained. For prediction stability, all 166 original external patients were held fixed while probabilities from the 1000 refitted models were compared. The mean absolute change from each original prediction was averaged across replicates and patients. For each patient, the 2.5th–97.5th percentile prediction range was calculated; range widths and penalty-selection frequencies were summarized. These ranges were interpreted as stability measures, not confidence intervals for an individual's true response probability [20]. The final development procedure was repeated, but the outer internal assessment and fixed image extraction were not rerun. Site clustering and imaging measurement error were not incorporated.

Treatment and sensitivity analyses

All recorded neoadjuvant treatment groups were retained in the main development cohort. Their pCR frequencies were described without estimating comparative treatment effects. The recorded Arm field was added to clinical and fusion models in I-SPY2-only analyses; an equivalent participant-level field was not verified in I-SPY1. The external predictions were consequently evaluated across the observed treatment mix, and regimen-specific accuracy was not established. Baseline MRI was acquired before I-SPY2 allocation; analyses including Arm were framed at the subsequent pretreatment time when allocation was known.

Restriction to recorded paclitaxel with or without trastuzumab, external complete-receptor cases and the alternative phase mapping was assessed in prespecified analyses. HR/HER2 subgroup AUROC was reported only with at least 30 patients and at least five events and five nonevents. Adequate subgroup precision was not established by this reporting rule. Sociodemographic fairness and site-level performance were not evaluated because suitable harmonized fields were not verified. No patient or public partners were involved in this secondary research.

Conventional MRI comparison

Four conventional baseline MRI measurements were obtained from a TCIA workbook for 384 linked I-SPY2 patients: longest diameter, functional tumour volume, sphericity and background parenchymal enhancement [6, 42]. Clinical, clinical-plus-diameter, clinical-plus-four-measurement, MRI-signal and signal-fusion models were evaluated on the same subset and folds. Volume was transformed as $\log(1 + \text{volume})$, and measurements were standardized within training folds. Postbaseline measurements and changes were excluded as predictors.

The source subset had been selected for complete longitudinal measurements. Response-related selection could therefore remain even though only baseline predictors were used. Diameter units and measurement extent were not verified as equivalent across trials, and the comparator was not evaluated externally. The longitudinal combined model in the source study addressed a later prediction time [6].

Software and research assistance

Analyses were performed using Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0 and scikit-learn 1.8.0. Executable pipelines, fitted parameters, inputs, predictions and verification code are provided in Additional files 2 and 3. OpenAI ChatGPT/Codex was used to assist with study planning, public-source retrieval, code development and execution, results checking, and manuscript drafting and editing during September 2026. Patient observations and fitted model results were obtained from the cited data and executed statistical code; no synthetic research findings were generated. Responsibility for the scientific content and use of these tools is retained by Tianyi Yu.

Results

Participant flow and source checks

All 982 development image records were linked uniquely to the original clinical table, and 316 pCR events were identified (32.2%). Among 172 external image records, original pCR values were missing for five despite negative derivative labels, and the prespecified phase discrepancy was identified in one additional record. The primary external cohort was therefore restricted to 166 patients with 48 events (28.9%; Fig. 1 and Table 1). The five missing outcomes were excluded from every analysis. Missing age was recorded for three development patients, and at least one receptor field was missing for five external patients.

Table 1
Characteristics of development and external populations

Characteristic	I-SPY2 development	I-SPY1 external
Participants	982	166
pCR, n (%)	316 (32.2)	48 (28.9)
Age, median (IQR), years	49.0 (41.0–56.0)	48.7 (41.8–54.1)
Missing age, n	3	0
HR- / HER2-, n	359	42
HR- / HER2+, n	87	29
HR+ / HER2-, n	381	66
HR+ / HER2+, n	155	24
Any missing receptor, n	0	5

Five external patients with incomplete receptor information are excluded from receptor-combination counts. IQR, interquartile range. Values are reported before imputation.

Across eligible image sets, 12 finite features were obtained per patient. In 980 of 982 development records, two to four margin slices beyond the positive-mask extent described in metadata were identified in the stored arrays. Two masks were encoded as 0/255. A further 515 image files bearing I-SPY1 identifiers were identified in the development archive; all were excluded from model development by the fixed whitelist. These findings were resolved through source and technical checks before primary MRI fitting. Corrections were not selected according to predictive performance.

Internal and external performance

In internal evaluation, Brier scores were 0.20143 for the clinical model and 0.19975 for fusion, with corresponding AUROCs of 0.647 and 0.680 (Table 2). For MRI alone, an AUROC of 0.566 was estimated. Internal performance was numerically better after recorded Arm was added. Final inverse penalties of 0.1, 1 and 0.1 were selected for clinical, MRI and fusion models, respectively.

Table 2
Nested internal performance in the full development cohort

Model	AUROC	Brier score	Average precision
Clinical	0.647	0.20143	0.444
MRI signal	0.566	0.21615	0.378
Clinical + MRI signal	0.680	0.19975	0.479
Clinical + Arm	0.687	0.19755	0.500
Clinical + MRI signal + Arm	0.699	0.19468	0.519

I-SPY2: 982 patients, 316 pCR events. Five outer and five inner folds were used with common partitions. Lower probability error is indicated by lower Brier values. The recorded treatment-combination field is denoted by Arm. Estimates are reported descriptively; treatment effects were not identified.

In external evaluation, Brier scores were 0.17543 for the clinical model and 0.18454 for fusion. The primary paired difference was -0.00911 (95% interval -0.02087 to 0.00218), without an established improvement from fusion. External AUROCs were 0.773, 0.528 and 0.707 for clinical, MRI and fusion models, respectively (Table 3).

Table 3
External performance of frozen models

Metric	Clinical	MRI signal	Fusion
Brier score	0.17543 (0.15264 to 0.20120)	0.21918 (0.19085 to 0.25001)	0.18454 (0.15896 to 0.21284)
AUROC	0.773 (0.696 to 0.845)	0.528 (0.426 to 0.629)	0.707 (0.616 to 0.792)
Average precision	0.546 (0.425 to 0.710)	0.348 (0.252 to 0.497)	0.537 (0.408 to 0.673)
Calibration-in-the-large	-0.236 (-0.585 to 0.086)	-0.330 (-0.723 to 0.022)	-0.297 (-0.665 to 0.040)
Calibration slope	1.652 (1.076 to 2.456)	0.138 (-0.397 to 0.820)	1.079 (0.614 to 1.629)

I-SPY1: 166 patients, 48 pCR events. Descriptive 95% bootstrap intervals conditional on fixed predictions are shown in parentheses. The MRI calibration-slope interval was estimated from 1999 valid replicates after one diagnostic fit failed; 2000 were used for other intervals. Development-sample uncertainty and site clustering were not incorporated. Calibration targets are zero for calibration-in-the-large and one for slope.

Mean probabilities of 33.6%, 35.6% and 34.6% were predicted by clinical, MRI and fusion models, respectively, exceeding the observed 28.9% frequency. Calibration-in-the-large was estimated as -0.236 , -0.330 and -0.297 , with slopes of 1.652, 0.138 and 1.079. Grouped agreement and prediction distributions are shown in Fig. 2. Overall overprediction was observed for fusion despite a slope near one. Calibration intervals were wide. After one MRI slope fit failed among 2000 bootstrap samples, its interval was estimated from 1999 valid fits. Complete external metric intervals are provided in Additional file 1.

Sensitivity and conventional MRI analyses

After external patients with incomplete receptor fields were excluded, 161 patients with 46 events were retained; the paired difference was estimated as -0.01077 (95% interval -0.02233 to 0.00050). After the phase-flagged patient was included under the specified mapping, 167 patients with 49 events were evaluated, and a difference of -0.00828 (-0.02047 to 0.00303) was obtained. Similar paired differences were observed in both sensitivity analyses and the primary analysis (Fig. 3).

The conventional MRI analysis was performed in 384 patients with 113 source-recorded events. Internal Brier scores of 0.19379, 0.18857, 0.19252 and 0.19657 were obtained for clinical, clinical-plus-diameter, clinical-

plus-four-conventional-measurement and signal-fusion models, respectively. Corresponding AUROCs were 0.642, 0.675, 0.666 and 0.666. An external imaging benefit was not established by these selected-subset findings. In the corresponding published cohort, 114 events were reported [6]; an unresolved one-event difference was identified within the triple-negative group. The acquired original labels were retained without alteration.

The paclitaxel-with-or-without-trastuzumab analysis was performed in 209 patients with 39 events. Internal AUROCs of 0.381, 0.497 and 0.429 were obtained for clinical, MRI and fusion models, respectively, with Brier scores of 0.15360, 0.15269 and 0.15408. The strongest candidate penalty was selected in all 15 outer-fold fits. Interpretation of this subset was limited by weak ranking and unstable calibration. Heterogeneous performance was observed across external receptor subgroups; among 42 HR-negative/HER2-negative patients, an AUROC of 0.199 was obtained for fusion, compared with 0.681 for the clinical model. These small subgroup estimates were reported descriptively. Treatment frequencies, subgroup results and complete sensitivity metrics are provided in Additional file 1.

Development and prediction stability

All 1000 planned resampling replicates were completed. In total, 3000 final pipelines were refitted after 75000 tuning fits; no failed replicates or model-fitting warnings were recorded. Patient separation was verified in all 5000 tuning partitions. A median paired external Brier difference of -0.01598 was obtained, with a 95% two-sample bootstrap interval from -0.07367 to 0.00646 (Fig. 4A). The interval was wider than the original interval conditional on frozen models. Positive contrasts were observed in 10.1% of replicates; this fraction was reported descriptively and was not interpreted as a posterior probability or confirmatory p value.

Mean absolute changes from the original external probabilities were 3.1, 8.4 and 7.3 percentage points for clinical, MRI and fusion models, respectively. Median patient-level 95% stability-range widths were 12.9, 28.8 and 29.3 percentage points (Table 4; Fig. 4C–E). Variation in selected penalties was also observed (Fig. 4B). These estimates were obtained without using external outcomes for recalibration or model selection.

Table 4
Prediction stability across repeated model development

Model	Brier median (95% range)	Mean absolute change, pp	Median range width (IQR), pp
Clinical	0.17587 (0.17195–0.18147)	3.1	12.9 (10.6–16.9)
MRI signal	0.22000 (0.20320–0.27971)	8.4	28.8 (16.3–49.6)
Fusion	0.19037 (0.17615–0.24993)	7.3	29.3 (19.5–48.1)

All 166 original external patients were held fixed while 1000 development samples were drawn and all models were retuned and refitted. The Brier range summarizes variation across development resamples with this external sample fixed. Mean absolute change was measured from the original frozen probability and averaged across replicates and patients. Each patient’s 95% prediction stability range was calculated, and the median and IQR of its width are reported. These ranges are not confidence intervals for true individual risk.

pp, percentage points. The independently resampled external Brier contrast is reported in Results and Fig. 4A.

Discussion

Improved external pCR probability prediction was not established when the 12 fixed MRI signal summaries were added to age and receptor status. The small internal improvement was not retained in I-SPY1, where a numerically lower Brier score was obtained for the clinical model. The paired interval included zero, so neither a reliable imaging increment nor equivalence was demonstrated. The interpretation was unchanged when development and external sampling uncertainty were considered together. This finding concerns the representation and comparator studied here; it does not establish that breast MRI lacks predictive information.

Added value of pretreatment MRI

Signal distributions within supplied tumour masks were summarized by the present representation. Spatial texture, peritumoral tissue, diffusion measurements and physical tumour extent were not represented. Intratumoral and peritumoral texture was evaluated by Braman and colleagues, while multiparametric MRI was combined with clinicopathological information and tested in external hospital cohorts by Liu and colleagues [7, 9]. An independent split from one institution was used by Cain and colleagues [8]. Such differences in both inputs and validation design prevent a direct ranking of these studies by their reported AUROC.

The same distinction applies to more extensive multimodal models. Clinical, digital pathology and molecular information was combined by Sammut and colleagues, with development in 168 patients and evaluation in 75 independent patients [11]. MRI, biopsy whole-slide images and clinical information were integrated by Mao and colleagues, with assessment in 417 external and 109 prospectively recruited patients [12]. A stratified hold-out set from pooled cohorts was used by Goyal and colleagues [10], whereas information accumulated during therapy was incorporated in the longitudinal framework of Gao and colleagues [13]. The current analysis contributes a narrower comparison: whether a reproducible baseline signal representation improves probability estimates beyond common clinical fields when an entire trial cohort is withheld from development.

Conventional tumour measurements also deserve direct comparison. On the same 384-patient subset, a numerically lower Brier score was obtained with baseline diameter plus clinical information than with signal fusion or the four-measurement comparator. These findings were obtained internally in a subset selected for longitudinal completeness and cannot establish an external imaging benefit. Earlier I-SPY analyses incorporated changes during treatment [5, 6]. Functional tumour volume has also been evaluated against recurrence-free survival [43], and changes in apparent diffusion coefficient after 12 weeks of therapy were evaluated as pCR predictors in ACRIN 6698 [44]. Their findings concern different information or endpoints from strictly pretreatment pCR prediction.

Reliability of probabilities across settings

A higher clinical-model AUROC was obtained in I-SPY1 than in internal validation, although the mean predicted probability was higher than the observed response frequency. Better external ranking is not necessarily evidence of better transportability: AUROC can be affected by the distribution of patient characteristics even when model coefficients are unchanged [45]. The contributions of case mix, treatment and acquisition differences were not separated in this study. The external result should therefore be interpreted through the paired comparison and calibration findings, rather than through AUROC alone.

For fusion, a calibration slope close to one was accompanied by negative calibration-in-the-large and greater probability error than for the clinical comparator. Different properties of agreement are assessed by the mean, slope and shape of the calibration relationship [46]. The five-group plots provided a descriptive summary but could not establish accurate probabilities throughout the prediction range. Flexible calibration curves can reveal local departures that grouped summaries obscure, although their interpretation is also limited by sample size [47]. With 48 external events and broad calibration intervals, reliable probabilities for individual consultations were not demonstrated.

An additional source of uncertainty was identified through repeated development. Mean absolute changes from the original probabilities were larger for MRI and fusion than for the clinical model, and wider patient-level stability ranges were observed. Average model performance can conceal this variation in the probability assigned to a particular patient [20]. Regularization does not remove development uncertainty, because the strength of penalization must itself be estimated [48]. This variation was incorporated in the repeated analysis through penalty reselection and refitting of all training-dependent preprocessing. It did not establish the reliability of the underlying image measurements, and the percentile ranges should not be interpreted as confidence intervals for a patient's true probability of response.

Treatment context and clinical interpretation

Treatment must be considered when a prediction is interpreted. A probability averaged over an observed treatment mix answers a different question from the outcome expected if a particular regimen were assigned [49]. Internal prediction was improved when recorded Arm was included, but an equivalent external participant-level field was not verified. The external models were therefore evaluated across the observed treatment mix, without establishing accuracy for individual regimens. Historical receptor thresholds and the incorporation of central testing further limit correspondence with information available at a contemporary consultation. Evidence from KEYNOTE-522 includes neoadjuvant pembrolizumab plus chemotherapy followed by adjuvant pembrolizumab [50]; performance in the earlier I-SPY1 cohort cannot be assumed to apply to that treatment context.

The endpoint also limits the clinical interpretation. A favourable prognosis is associated with pCR, but an individual's survival or comparative treatment benefit is not identified by a predicted pCR probability [2, 3, 4]. Prognostic information is retained in the amount of residual disease among patients who do not achieve pCR, as shown in the pooled analysis of residual cancer burden [51]. In KATHERINE, patients with observed residual invasive HER2-positive disease after neoadjuvant therapy were randomized between postoperative treatments [52]. A pretreatment probability cannot replace that pathological assessment. The present findings support neither treatment omission nor model-guided regimen selection. Evaluation of such

decisions would require a specified clinical action and its consequences, beyond the probability-accuracy question addressed here [41].

Strengths and limitations

Source auditing was central to the validity of the comparison. Five patients would have been assigned negative outcomes if derivative labels had been used without the original workbook. Margin slices, mask encodings and inconsistent cohort tokens also required reconciliation before fitting. Original labels, fixed identifier lists and correction records were retained, and all models were compared in the same eligible patients. These records allow the analysis to be reconstructed and its exclusions to be examined. They do not resolve the one-event discrepancy with the published conventional MRI subset or establish anatomical segmentation accuracy.

Measurement reliability remains a separate limitation. Variation in breast tumour segmentation has been shown to affect the robustness of MRI radiomic features [53]. In ACRIN 6698, repeatable apparent diffusion coefficient measurements were obtained under standardized acquisition and quality assurance [54]; that evidence does not establish repeatability of the DCE signal summaries used here. Image perturbations have been evaluated as a way to study feature robustness in CT radiomics [55], but no such validation was performed for this representation. Supplied masks were used without independent breast-radiologist review, and segmentation, motion and acquisition uncertainty were not quantified by patient resampling.

The clinical comparator was restricted because harmonized stage and grade had not been verified. An increment over age and receptor status would not necessarily be retained against a fuller clinical assessment. Selection for image availability had already occurred in the derivative release, with further selection for longitudinal completeness in the conventional subset. Historical specialist-trial populations may not represent contemporary practice, patients treated outside specialist centres or men with breast cancer. Sociodemographic fairness was not assessed. Development sample adequacy depends on the number of candidate parameters, anticipated model performance and the precision required; it cannot be inferred from cohort size alone [56]. External calibration and subgroup precision were limited by 166 patients and 48 events [32, 33].

The additional stability analysis was specified after the original external findings were known and was interpreted alongside the frozen primary analysis. Resampling did not increase the number of independent external patients, and unidentified site clustering remained unaccounted for. Neither an independent replication by another analyst nor a clinical-impact study was performed. Before response communication is considered, further evaluation would require a broader clinical comparator, independently assessed image quality, verified treatment and acquisition metadata, and prospectively defined targets for probability precision [18, 33]. Any model revised after inspection of I-SPY1 should be assessed in a new, unused cohort.

Conclusions

Improved external pCR probability prediction was not established after 12 fixed pretreatment MRI signal summaries were added to age and receptor status in this study of 982 development and 166 external patients. The apparent internal gain was not retained across trials, and greater variability in individual

predictions was observed for fusion than for the clinical model. The value of MRI for response communication should therefore be judged through external probability accuracy, development stability and comparison with established clinical and imaging information. Clinical utility and reliable regimen-specific prediction remain unestablished.

Abbreviations

AUROC, area under the receiver operating characteristic curve; DCE, dynamic contrast-enhanced; HER2, human epidermal growth factor receptor 2; HR, hormone receptor; IQR, interquartile range; MRI, magnetic resonance imaging; pCR, pathological complete response; TCIA, The Cancer Imaging Archive.

Declarations

Ethics approval and consent to participate

Ethics approval and additional participant consent were not required for this secondary analysis of publicly accessible, deidentified datasets. No new participants were recruited or contacted, and no identifiable information was collected. Institutional ethics approval and informed consent were reported for the original I-SPY studies [5, 6]. No new institutional approval or formal waiver was issued for this reanalysis.

Consent for publication

Not applicable. No identifiable participant information or individual clinical images are presented.

Availability of data and materials

The original clinical data can be accessed through TCIA, and the image derivatives through the cited BreastDCEDL release [26, 27, 28, 29, 30, 42]. The datasets supporting the conclusions are included in Additional file 2 as analysis tables and predictions, together with code, fitted pipelines, dated specifications, source checksums and reproduction instructions. Repeated-development inputs, resample indices, predictions, tuning records and reproduction code are included in Additional file 3. The large image archives can be obtained from their original repository. Source-specific licences and attributions are retained; the BreastDCEDL derivative release is licensed under Creative Commons Attribution-NonCommercial 4.0 International. The original data are not relicensed by this study. No public registration or separate code-repository identifier is asserted.

Competing interests

No competing interests are declared by TY.

Funding

No funding was received for this research.

Authors' contributions

The study was conceived by TY. Public data were curated, the analysis workflow was developed and executed, results were checked and interpreted, figures were prepared, and the manuscript was drafted and revised by TY using the assistance described in Methods. Responsibility for the work and the decision to submit is retained by TY as the sole author.

Acknowledgements

The contributions of the participants and original study teams to the publicly available imaging and clinical resources are gratefully acknowledged.

References

1. Korde LA, Somerfield MR, Carey LA, Crews JR, Denduluri N, Hwang ES, et al. Neoadjuvant Chemotherapy, Endocrine Therapy, and Targeted Therapy for Breast Cancer: ASCO Guideline. *J Clin Oncol*. 2021;39:1485-1505. <https://doi.org/10.1200/jco.20.03399>
2. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384:164-172. [https://doi.org/10.1016/s0140-6736\(13\)62422-8](https://doi.org/10.1016/s0140-6736(13)62422-8)
3. Esserman LJ, Berry DA, DeMichele A, Carey L, Davis SE, Buxton M, et al. Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL--CALGB 150007/150012, ACRIN 6657. *J Clin Oncol*. 2012;30:3242-3249. <https://doi.org/10.1200/jco.2011.39.2779>
4. Spring LM, Fell G, Arfe A, Sharma C, Greenup R, Reynolds KL, et al. Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis. *Clin Cancer Res*. 2020;26:2838-2848. <https://doi.org/10.1158/1078-0432.ccr-19-3492>
5. Hylton NM, Blume JD, Bernreuter WK, Pisano ED, Rosen MA, Morris EA, et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. *Radiology*. 2012;263:663-672. <https://doi.org/10.1148/radiol.12110748>
6. Li W, Newitt DC, Gibbs J, Wilmes LJ, Jones EF, Arasu VA, et al. Predicting breast cancer response to neoadjuvant treatment using multi-feature MRI: results from the I-SPY 2 TRIAL. *NPJ Breast Cancer*. 2020;6:63. <https://doi.org/10.1038/s41523-020-00203-7>
7. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res*. 2017;19:57. <https://doi.org/10.1186/s13058-017-0846-1>

8. Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast Cancer Res Treat.* 2019;173:455-463. <https://doi.org/10.1007/s10549-018-4990-9>
9. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study. *Clin Cancer Res.* 2019;25:3538-3547. <https://doi.org/10.1158/1078-0432.ccr-18-3190>
10. Goyal M, Shukla T, Hassanpour S. Multimodal AI Methods for Predicting Pathologic Complete Response in Breast Cancer. *J Imaging Inform Med.* 2026. Published online 26 August 2026. <https://doi.org/10.1007/s10278-026-02188-4>
11. Sammut SJ, Crispin-Ortuzar M, Chin SF, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature.* 2022;601:623-629. <https://doi.org/10.1038/s41586-021-04278-5>
12. Mao N, Dai Y, Zhou H, Lin F, Zheng T, Li Z, et al. A multimodal and fully automated system for prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer. *Sci Adv.* 2025;11:eadr1576. <https://doi.org/10.1126/sciadv.adr1576>
13. Gao Y, Ventura-Diaz S, Wang X, He M, Xu Z, Weir A, et al. An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. *Nat Commun.* 2024;15:9613. <https://doi.org/10.1038/s41467-024-53450-8>
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>
15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21:128-138. <https://doi.org/10.1097/ede.0b013e3181c30fb2>
16. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68:279-289. <https://doi.org/10.1016/j.jclinepi.2014.06.018>
17. Collins GS, Dhiman P, Ma J, Schlusser MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ.* 2024;384:e074819. <https://doi.org/10.1136/bmj-2023-074819>
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. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245-247. <https://doi.org/10.1016/j.jclinepi.2015.04.005>
20. 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>
21. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* 2013;26:1045-1057. <https://doi.org/10.1007/s10278-013-9622-7>

22. Garrucho L, Kushibar K, Reidel CA, Joshi S, Osuala R, Tsirikoglou A, et al. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Sci Data*. 2025;12:453. <https://doi.org/10.1038/s41597-025-04707-4>
23. Fridman N, Solway B, Fridman T, Barnea I, Goldstein A. BreastDCEDL: A standardized deep learning-ready breast DCE-MRI dataset of 2,070 patients. *Sci Data*. 2026;13:264. <https://doi.org/10.1038/s41597-026-06589-6>
24. 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>
25. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. <https://doi.org/10.1136/bmj-2024-082505>
26. The Cancer Imaging Archive. I-SPY 2 TRIAL imaging cohort [dataset]. <https://doi.org/10.7937/TCIA.D8Z0-9T85>. Accessed 13 September 2026.
27. The Cancer Imaging Archive. ACRIN 6698 [dataset]. <https://doi.org/10.7937/TCIA.KK02-6D95>. Accessed 13 September 2026.
28. The Cancer Imaging Archive. I-SPY 1 [dataset]. <https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK>. Accessed 13 September 2026.
29. Fridman N, Solway B, Fridman T, Barnea I. BreastDCEDL: A Deep Learning–Ready Breast DCE-MRI Dataset. Version 4 [dataset]. Zenodo; 2025. <https://doi.org/10.5281/zenodo.18114231>. Accessed 13 September 2026.
30. Fridman N, et al. BreastDCEDL source metadata. GitHub commit 961c458a393b3173ce923e7bc78409790153a15e. <https://github.com/naomifridman/BreastDCEDL/tree/961c458a393b3173ce923e7bc78409790153a15e>. Accessed 13 September 2026.
31. Nanda R, Liu MC, Yau C, Shatsky R, Pusztai L, Wallace A, et al. Effect of Pembrolizumab Plus Neoadjuvant Chemotherapy on Pathologic Complete Response in Women With Early-Stage Breast Cancer: An Analysis of the Ongoing Phase 2 Adaptively Randomized I-SPY2 Trial. *JAMA Oncol*. 2020;6:676-684. <https://doi.org/10.1001/jamaoncol.2019.6650>
32. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021;40:4230-4251. <https://doi.org/10.1002/sim.9025>
33. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*. 2024;384:e074821. <https://doi.org/10.1136/bmj-2023-074821>
34. Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14:75. <https://doi.org/10.1186/s13244-023-01415-8>
35. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based

- Phenotyping. *Radiology*. 2020;295:328-338. <https://doi.org/10.1148/radiol.2020191145>
36. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. 2022;29:1525-1534. <https://doi.org/10.1093/jamia/ocac093>
 37. 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. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
 38. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>
 39. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights Imaging*. 2021;12:172. <https://doi.org/10.1186/s13244-021-01115-1>
 40. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432. <https://doi.org/10.1371/journal.pone.0118432>
 41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565-574. <https://doi.org/10.1177/0272989x06295361>
 42. The Cancer Imaging Archive. Multi-feature MRI NACT data workbook. <https://wiki.cancerimagingarchive.net/download/attachments/70230072/Multi-feature%20MRI%20NACT%20Data.xlsx?api=v2>. Accessed 13 September 2026.
 43. Hylton NM, Gatsonis CA, Rosen MA, Lehman CD, Newitt DC, Partridge SC, et al. Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology*. 2016;279:44-55. <https://doi.org/10.1148/radiol.2015150013>
 44. Partridge SC, Zhang Z, Newitt DC, Gibbs JE, Chenevert TL, Rosen MA, et al. Diffusion-weighted MRI Findings Predict Pathologic Response in Neoadjuvant Treatment of Breast Cancer: The ACRIN 6698 Multicenter Trial. *Radiology*. 2018;289:618-627. <https://doi.org/10.1148/radiol.2018180273>
 45. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172:971-980. <https://doi.org/10.1093/aje/kwq223>
 46. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. <https://doi.org/10.1016/j.jclinepi.2015.12.005>
 47. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33:517-535. <https://doi.org/10.1002/sim.5941>
 48. Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol*. 2021;132:88-96. <https://doi.org/10.1016/j.jclinepi.2020.12.005>

49. van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol.* 2020;35:619-630. <https://doi.org/10.1007/s10654-020-00636-1>
50. Schmid P, Cortes J, Dent R, McArthur H, Pusztai L, Kümmel S, et al. Overall Survival with Pembrolizumab in Early-Stage Triple-Negative Breast Cancer. *N Engl J Med.* 2024;391:1981-1991. <https://doi.org/10.1056/nejmoa2409932>
51. Yau C, Osdoit M, van der Noordaa M, Shad S, Wei J, de Croze D, et al. Residual cancer burden after neoadjuvant chemotherapy and long-term survival outcomes in breast cancer: a multicentre pooled analysis of 5161 patients. *Lancet Oncol.* 2022;23:149-160. [https://doi.org/10.1016/s1470-2045\(21\)00589-1](https://doi.org/10.1016/s1470-2045(21)00589-1)
52. von Minckwitz G, Huang CS, Mano MS, Loibl S, Mamounas EP, Untch M, et al. Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. *N Engl J Med.* 2019;380:617-628. <https://doi.org/10.1056/nejmoa1814017>
53. Granzier RWY, Verbakel NMH, Ibrahim A, van Timmeren JE, van Nijnatten TJA, Leijenaar RTH, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep.* 2020;10:14163. <https://doi.org/10.1038/s41598-020-70940-z>
54. Newitt DC, Zhang Z, Gibbs JE, Partridge SC, Chenevert TL, Rosen MA, et al. Test-retest repeatability and reproducibility of ADC measures by breast DWI: Results from the ACRIN 6698 trial. *J Magn Reson Imaging.* 2019;49:1617-1628. <https://doi.org/10.1002/jmri.26539>
55. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. *Sci Rep.* 2019;9:614. <https://doi.org/10.1038/s41598-018-36938-4>
56. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, 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>

Figures

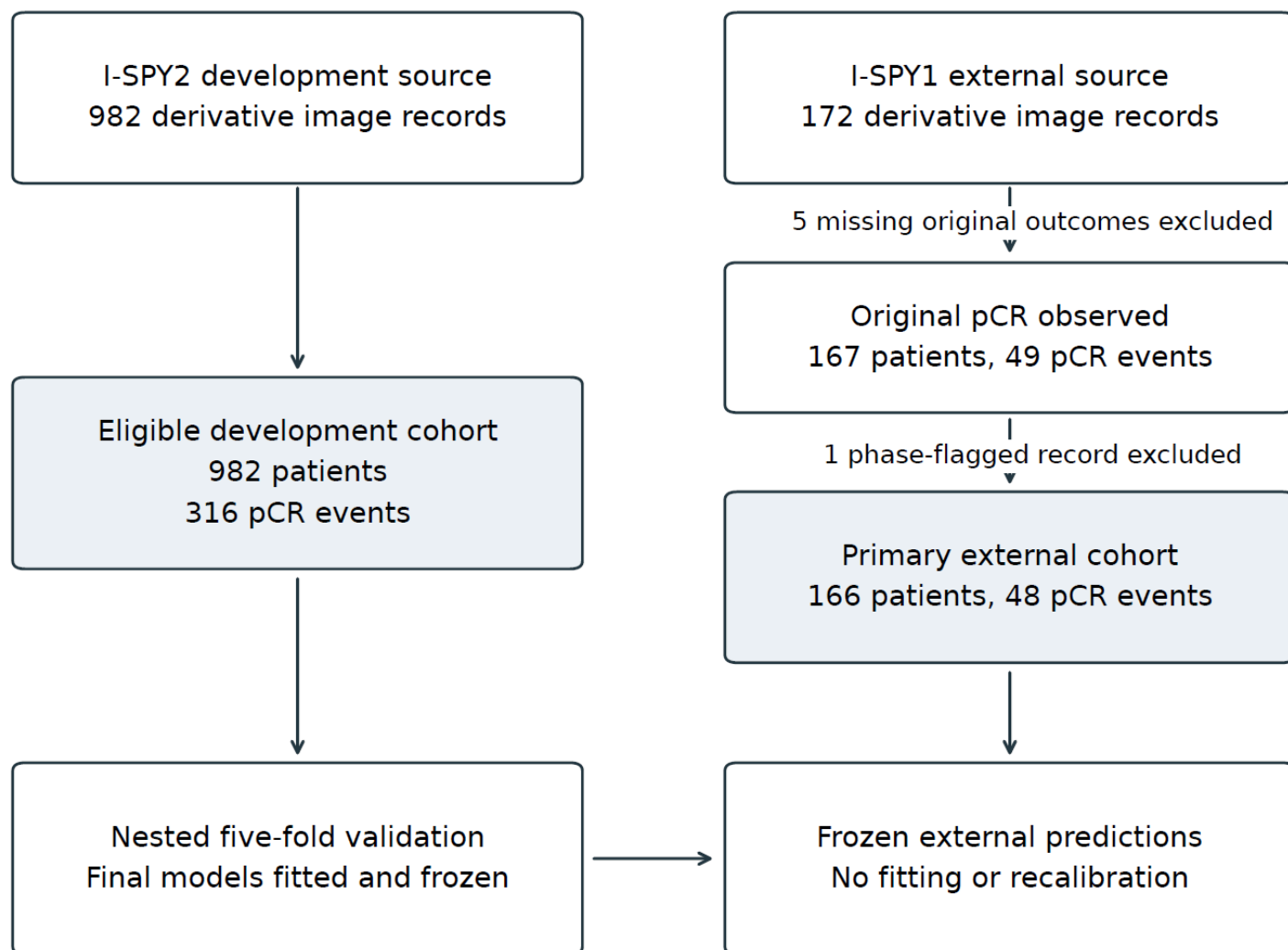


Figure 1

Participant selection for development and external evaluation

Counts are reported from the selected derivative image release rather than the full parent trial enrolment. Development was performed in 982 patients with 316 pCR events. Five external records with missing original pathological outcomes and one additional record with a prespecified phase-annotation discrepancy were excluded. The primary external population was restricted to 166 patients with 48 events. In the alternative phase-mapping sensitivity, 167 patients with 49 events were included; missing original outcomes were excluded throughout.

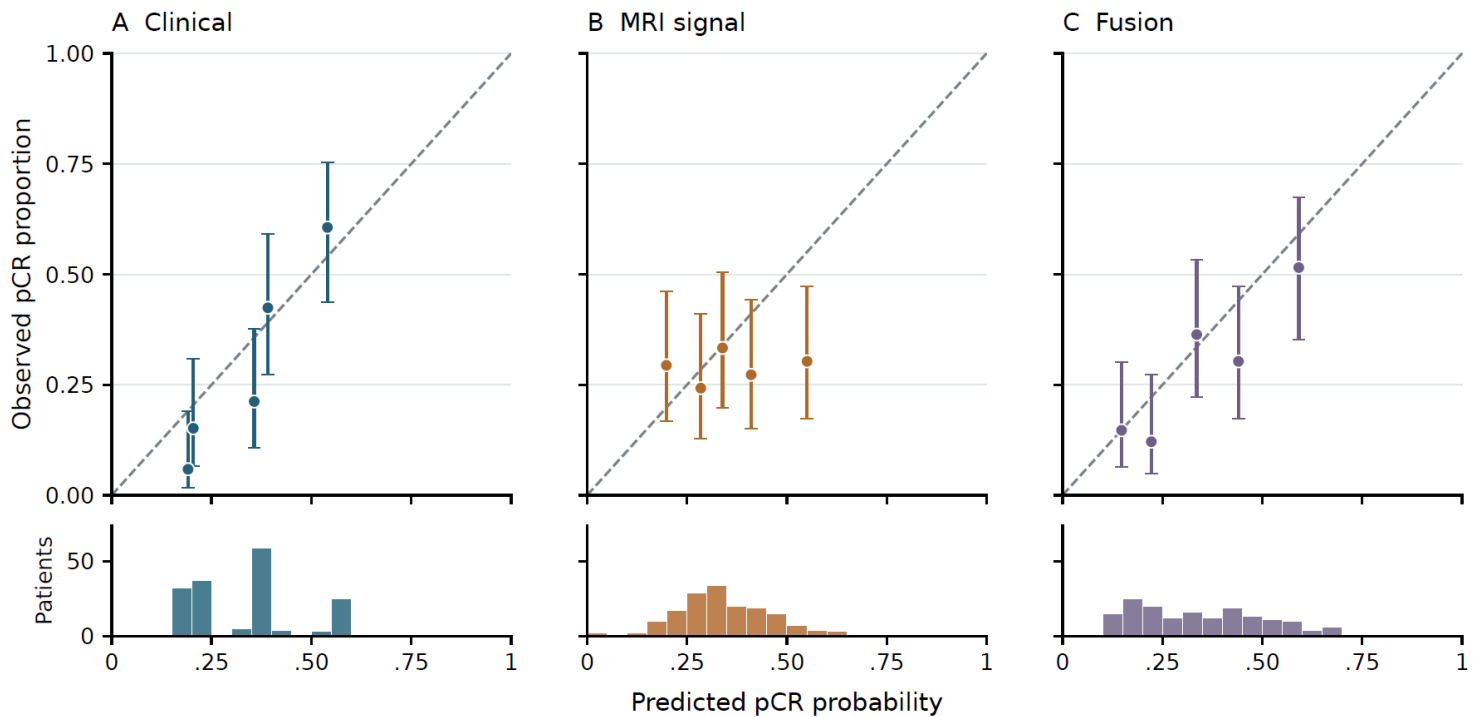


Figure 2

External probability calibration and prediction distributions

Frozen I-SPY1 probabilities are shown for clinical, MRI-signal and fusion models in 166 patients with 48 pCR events. Mean predicted probabilities and observed frequencies are represented by points in five approximately equal-sized prediction groups, with ties retained. Pointwise 95% Wilson intervals for grouped observed frequencies are shown by error bars; individual probability uncertainty, model-development uncertainty and site clustering are not represented. Perfect agreement is indicated by the dashed line, and prediction distributions are displayed in the lower panels. Predictions were not recalibrated.

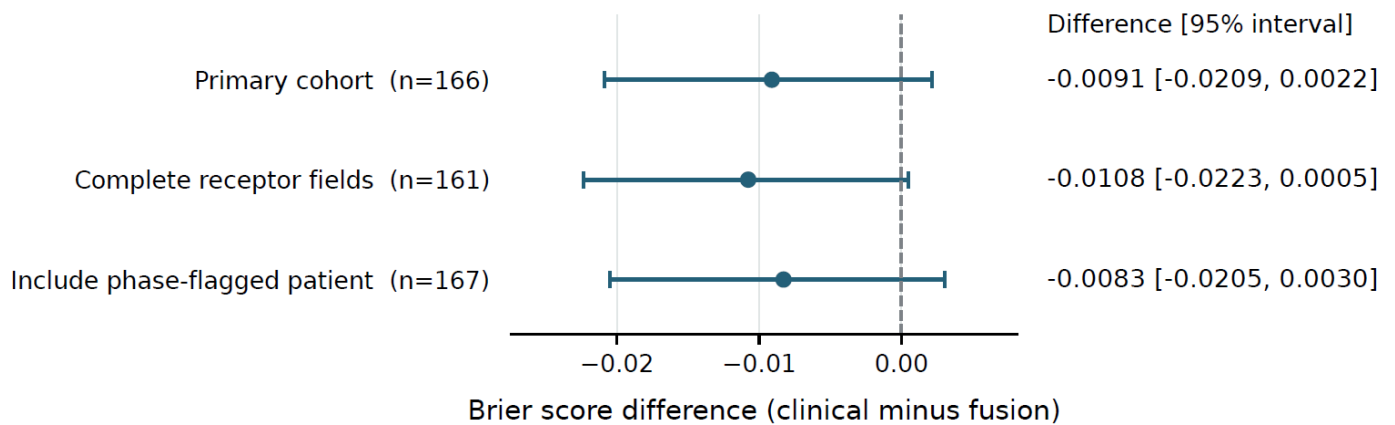


Figure 3

Paired external Brier differences and prespecified sensitivities

Differences are defined as Brier(clinical) minus Brier(fusion), with positive values favouring fusion. The 95% percentile intervals from 2000 paired external patient bootstrap resamples are shown by bars and were conditioned on the frozen models. The primary, complete-receptor and alternative-phase populations were formed from 166, 161 and 167 patients, respectively. Development-sample uncertainty and unidentified site clustering were not incorporated. No difference is indicated by the vertical dashed line; equivalence is not established by overlap with zero.

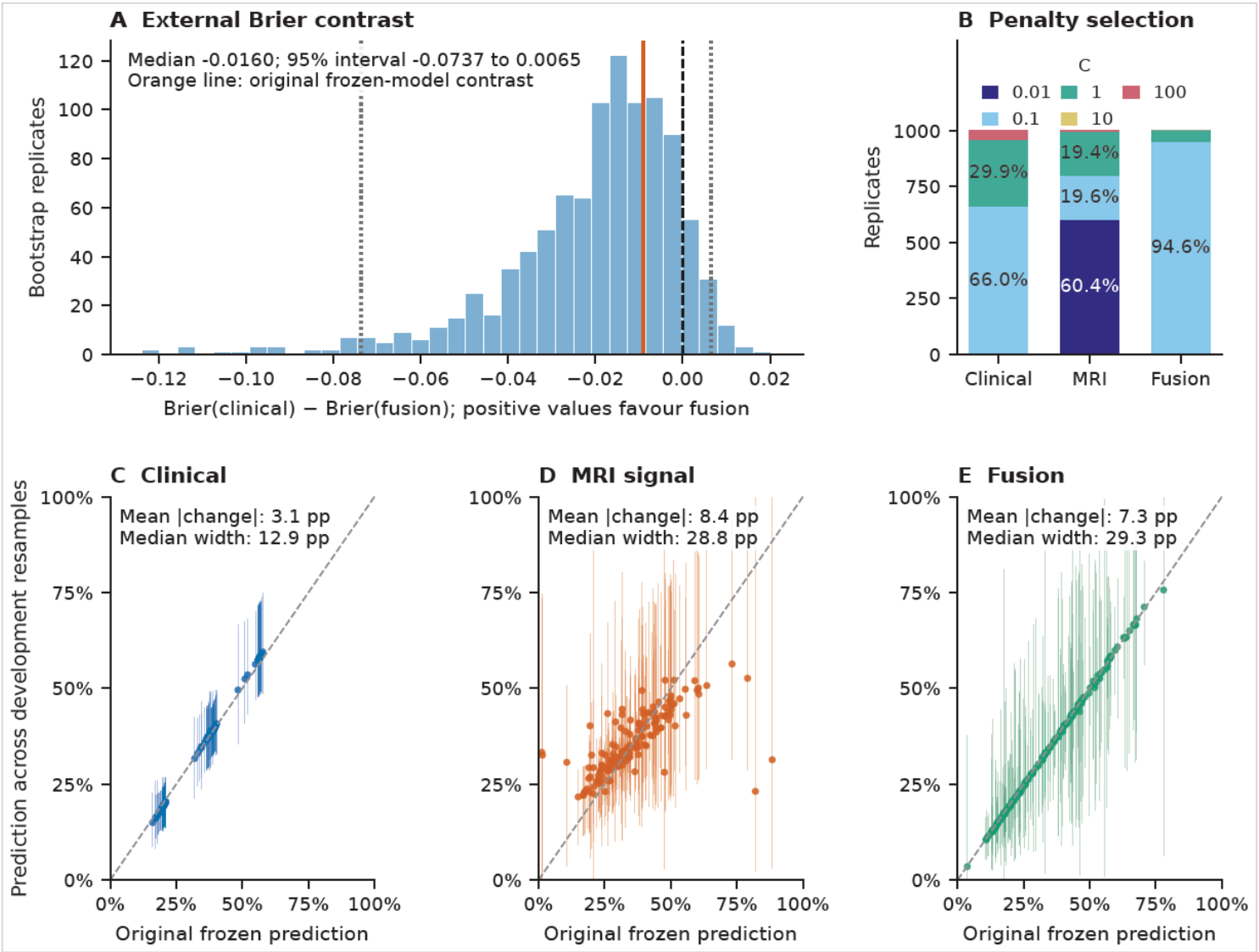


Figure 4

External model comparison and patient prediction stability after repeated development

(A) Paired Brier contrasts are shown from 1000 independent development and external patient resamples. No contrast is indicated by the black dashed line; the original frozen-model contrast is shown in orange, and the central 95% percentile limits are shown by dotted lines. (B) Frequencies of the selected inverse penalty C are shown for the 1000 final fits of each model. Stronger regularization is represented by smaller C. (C–E) All 166 original external patients were held fixed. Original predicted probabilities are shown on the horizontal

axes; medians across refitted models are shown by points and central 95% prediction stability ranges by vertical lines. Agreement with the original prediction is indicated by the diagonal. Probability differences are expressed in percentage points (pp). These ranges quantify development stability and are not confidence intervals for individual true response probabilities. Training-dependent preprocessing, tuning and final fitting were repeated; fixed image extraction and the outer internal assessment were not repeated.

Supplementary Files

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

- [Additionalfile1.docx](#)
- [TRIPODAIChecklist.docx](#)
- [Additionalfileslegends.docx](#)