RESEARCH PAPER

KI-67, CD4, CD8 AND PD-L1 AS FOCUSED BIOMARKERS OF PATHOLOGICAL COMPLETE RESPONSE IN ER/PR/HER2-CLASSIFIED BREAST CANCER: A RETROSPECTIVE ANALYSIS OF GSE20194

Shiv kumar, prof Sunil kumar Saini, Prof Sanjiv Kumar Verma, Dr. Nadia Shirazi
[1]Research Scholar, Department of Oncology medicine, CCRC, SRHU Dehradun, [2]Director, CCRC, SRHU Dehradun,
[3]HOD, Oncology medicine, SRHU Dehradun

ABSTRACT
Pathological complete response (pCR) can be predicted by pretreatment, which can help with treatment planning for patients with breast cancer who are treated with neoadjuvant chemotherapy. The GSE20194 dataset was used for this retrospective secondary analysis, and the molecular assessments were limited to the assessment of Ki-67, CD4, CD8, and PD-L1, with the status of estrogen receptor (ER), progesterone receptor (PR) and HER2 retained for classification of the receptor classes. A total of 29 replicate and 19 quality-excluded arrays were removed, leaving 230 patients: 130 in the training group and 100 in the locked validation group. Forty-eight patients (20.9%) achieved pCR. The expression measure for MKI67, CD4 and CD8A was provided by GPL96 and no CD274 probe set was present, so that PD-L1 was not estimated or used as a proxy. A repeated stratified ten-fold cross-validated penalised logistic model included the three available marker expressions and ER, PR and HER2 status. The results were obtained by training out-of-fold AUC of .808. The locked validation AUC was .775 with .800 sensitivity and .729 specificity and a Brier score of .186 at the threshold of .499. The AUC for the receptor-only validation set was .811, while the available-marker-only validation set was .645. The findings are in favor of an 'integrated' marker approach but not for a full 4 marker panel as PD-L1 was not represented on the platform. 1) an external validation in a cohort of all four markers is necessary.

KEYWORDS
How to cite this article: Shiv kumar prof Sunil kumar Saini Prof Sanjiv Kumar Verma Dr. Nadia Shirazi. KI-67, CD4, CD8 AND PD-L1 AS FOCUSED BIOMARKERS OF PATHOLOGICAL COMPLETE RESPONSE IN ER/PR/HER2-CLASSIFIED BREAST CANCER: A RETROSPECTIVE ANALYSIS OF GSE20194. Int J Drug Deliv Technol. 2026;16(78s):1308-1321. DOI: 10.25258/ijddt.16.78s.142.

1. INTRODUCTION

1.1 Background and clinical context

Breast cancer is still a health issue worldwide and the most common cancer in women. About 2.4 million women were diagnosed with the disease and 694,000 women died of it in 2024, and inequalities in access to timely diagnosis and treatment (World Health Organization, 2026). In the clinic, breast cancer is not a single disease, as the expression of estrogen receptor, progesterone receptor, and HER2 results in biologically distinct subtypes with different prognosis and treatment sensitivity (Zhao et al., 2024). Neoadjuvant chemotherapy is used prior to definitive surgery to reduce the size of the tumour, to make it easier to operate on, to increase the chance of breast conserving surgery, and to assess in vivo response to treatment. It has a well-defined place in high-risk HER2+ and triple-negative disease, with pathological response providing direction for further escalation or de-escalation of systemic therapy (Dowling et al., 2023). But the answer is not uniform, some tumours are cured, others are resistant and patients are made toxic without gaining anything. Predicting response prior to surgery may thus guide treatment selection, allow for earlier adjustment of ineffective treatment, guide surgical planning, and identify patients who need more aggressive post-neoadjuvant therapy. An accurate pretreatment model is therefore crucial for individual decision making and effective utilization of oncology resources.

1.2 Pathological Complete Response

Pathological complete response is achieved when there is no residual invasive carcinoma in the resected breast or regional lymph nodes after neoadjuvant treatment (ypT0/Tis ypN0) (Dowling et al., 2023). Patients who do not fall into this category have residual disease, which may be microscopic foci or involve tumour and/or nodal involvement. Event-free and overall survival are better for those with pCR, which reflects chemosensitivity, and for those with residual disease, which reflects recurrence risk. In a study of stage, I triple-negative breast cancer, 97% of patients with pCR survived for 5 years compared with 90% of those without pCR (de Graaf et al., 2025). However, the frequency and prognosis of pCR is not the same for all subtypes. Rates are typically higher in triple-negative and HER2-positive cancers treated with targeted therapy, and lower in hormone-receptor-positive/HER2-negative disease, highlighting the importance of subtype-aware prediction (Zhao et al., 2024).

1.3 Focused Tumour Markers and Receptor Classification

Clinically relevant receptor phenotypes are determined by ER, PR and HER2 status, showing varying treatment sensitivity and pCR rate (Zhao et al., 2024). For these receptor-defined groups, this study involves just four pretreatments markers: Ki-67, a proliferation marker, CD4 and CD8, both markers of helper and cytotoxic T-cell infiltration, and PD-L1, an immune-checkpoint marker. These markers, in combination, reflect tumour proliferation and specific elements of antitumour immunity, and are repeatedly associated with the response to neoadjuvant treatment (Derouane et al., 2022; Fernandez-Martinez et al., 2023). The interpretation is dependent on the assay; gene-expression measurements are proxies of the level of expression and cannot be converted to an immunohistochemistry protein score. A focused model should then present clearly specified measurements, maintain the ER/PR/HER2 classification and make explicit where a marker is not available on the source platform.

These markers, in combination, reflect tumour proliferation and specific elements of antitumour immunity, and are repeatedly associated with the response to neoadjuvant treatment (Derouane et al., 2022; Fernandez-Martinez et al., 2023; Kumar, n.d.).

1.4 Research Gap

The majority of pCR models are based on large clinical and/or molecular panels, with the challenge of distinguishing the role of a small clinically interpretable marker set. CD4, CD8 and PD-L1 were also not considered as candidate molecular variables in the earlier analysis. The present focused re-analysis aimed to solve this mismatch by restricting the planned molecular panel to the expression of Ki-67, CD4, CD8 and PD-L1, preserving the receptor classification based on ER and PR and HER2, documenting platform availability prior to modelling, and not changing the validation cohort, which was defined by the source. This is also required to avoid silent imputation or substitution of biologically distinct immune genes.

1.5 Study Aim and Objectives

Objectives: To evaluate the focused marker panel of Ki-67, CD4, CD8 and PD-L1 as a potential predictor of pathological complete response (pCR) following neoadjuvant chemotherapy (NAC) in ER/PR/HER2 breast cancer. To (1) assign each requested marker to the GPL96 platform and report its availability, (2) explore univariable associations between the measurable markers and pCR, (3) create a narrowed model with available marker expression and ER, PR and HER2 status, (4) assess discrimination and calibration in the locked validation cohort, and (5) describe the performance of the model within ER/PR/HER2 derived receptor phenotypes. There were no markers that were not measured to be imputed or represented by a substitute gene.

2. MATERIALS AND METHODS

2.1 The study design and reporting framework

This study was still a retrospective secondary analysis of a de-identified cohort and observation prediction-model development and validation study. No treatment was given. Ki-67, CD4, CD8 and PD-L1 were the only proposed molecular markers for the focused reanalysis, and ER, PR and HER2 were the remaining markers for receptor-based classification. Model estimation was performed after determining platform feasibility. Only the source-designated validation cohort was locked, all outcome informed model development was restricted to training cohort. Reporting continued to be done according to TRIPOD+AI/STROBE guidelines (Collins et al., 2024; EQUATOR Network, 2026).

2.2 The study's data source and setting

The data were downloaded from the National Centre for Biotechnology Information Gene Expression Omnibus (GEO) series GSE20194. These are the MAQC-II breast-cancer series provided by the University of Texas MD Anderson Cancer Centre, Houston. The GPL96 Affymetrix Human Genome U133A platform was used to measure gene expression. The specimens were fine-needle aspirates from newly diagnosed stage I–III tumours, with a high proportion of neoplastic cells. The source participants were recruited sequentially from 2000 to 2008 as part of a pharmacogenomic marker discovery study and then treated with neoadjuvant chemotherapy and surgery. GEO lists 230 cancers that can be analysed, 130 of which are used for training and 100 for independent validation (National Centre for Biotechnology Information, 2022).

2.3 Study Population and Eligibility Criteria

Documented stage I–III breast cancer, available pretreatment tumour specimen, received neoadjuvant chemotherapy, completed surgery with pathological response assessment, known pCR or residual-disease outcome, and source training or validation designation were required for eligibility. Eligibility was defined based on clinical annotation, not on the basis of expression patterns. Technical replicates were excluded from the analysis to ensure that only one independent observation per patient was included. Records that were flagged by the source quality control process, records with no response data, and records that were not able to be matched across clinical, expression, and platform files were also excluded. No patients were excluded due to missing predictor values, but missingness was dealt with analytically as described below.

2.4 Secondary Data Extraction and Focused Marker Mapping

The GSE20194 clinical annotation workbook, complete processed series matrix and GPL96

annotation were obtained from GEO. GSM accessions were standardised and used to link clinical and expression records. A total of 29 replicate arrays were rejected and 19 arrays with poor quality data were excluded in the sources, leaving 230 independent arrays. The entire matrix was tested for exact gene-symbol mappings for MKI67, CD4, CD8A and CD274, which were not restricted to the 19-gene worksheet used in the previous analysis. Both clinical values and GEO-processed expression intensities have been kept as supplied without the addition of any marker value that was not present, back-calculated or substituted.

Table 1. Derivation of the analysis sample and source-designated cohort partition

Stage or cohort	Total, n	pCR, n	Residual disease, n	pCR rate
GEO array records screened	278	—	—	—
Replicate arrays excluded	29	—	—	—
Quality-excluded arrays removed	19	—	—	—
Training cohort	130	33	97	25.4 %
Validation cohort	100	15	85	15.0 %
Final analysis sample	230	48	182	20.9 %

Note. pCR = pathological complete response. Exclusion categories are mutually exclusive. The validation cohort was preserved exactly as designated by the source.

2.5 Outcome Definition

This resulted in pathological complete response following neoadjuvant chemotherapy. As defined in the source, at surgery there was no residual invasive cancer in either the breast or lymph nodes, according to pCR. Residual disease meant that there was still some invasive cancer in the breast, the lymph nodes or both. For modelling, pCR was coded as a binary variable (1 if pCR was achieved, 0 if residual disease was present). If the responses were unclear or not provided, these were not included in the analysis.

2.6 Focused Markers and ER/PR/HER2 Classification

The planned molecular panel only included Ki-67 (MKI67), CD4, CD8 (CD8A) and PD-L1 (CD274). If a gene represented multiple GPL96 probe sets that were each mapped exactly, the probe with the largest interquartile range over the 230 analysis cases was chosen (without employing response labels). This outcome-blind rule selected 212022_s_at for MKI67, 203547_at for CD4 and 205758_at for CD8A. No probe set in the GPL96 annotation was mapped to CD274 and it was not a patient-level missing value, so it was not substituted in numerical modelling (National Center for Biotechnology Information, 2021). There were no changes in ER, PR or HER2 status and the four descriptive receptor subphenotypes were defined as HR+/HER2-, HR+/HER2+, HR-/HER2+ and triple-negative (defined as HR negative). The focused combined model included the three marker expressions and binary ER, PR and HER2 status.

Table 2. Focused markers, ER/PR/HER2 classification, operational coding, and model handling

Domain	Variables	Operational coding	Model handling
Receptor classification	ER, PR and HER2 status	Binary status; HR defined as ER or PR positive	Retained as model terms and used to define receptor phenotypes
Proliferation marker	Ki-67: MKI67 / 212022_s_at	Continuous GEO-processed expression	Training-derived centring and scaling
Helper T-cell marker	CD4 / 203547_at	Continuous GEO-processed expression	Training-derived centring and scaling
Cytotoxic T-cell marker	CD8: CD8A / 205758_at	Continuous GEO-processed expression	Training-derived centring and scaling

Immune-checkpoint marker	PD-L1: CD274	No CD274 probe set on GPL96	Unavailable; excluded without imputation or substitution	
Outcome	pCR versus residual disease	pCR = 1; residual disease = 0	Binary logistic target	

Note. ER = oestrogen receptor; PR = progesterone receptor; HER2 = human epidermal growth factor receptor 2; HR = hormone receptor; MKI67 = Ki-67 gene; CD8A = CD8 alpha-chain gene; CD274 = PD-L1 gene; IQR = interquartile range. Expression values are transcript-level proxies, not IHC protein scores. PD-L1 was unavailable because GPL96 has no CD274 probe set.

2.7 Data Preprocessing and Missing-Data Management

Complete binary ER, PR and HER2 status, as well as complete MKI67, CD4 and CD8A expression measurement, were used in the focused analysis. PD-L1 was not imputed as patient level missing data but was classified at the platform level. There were no imputations needed for an outcome or focused predictor. The three expression variables were centred and scaled in each split of cross validation by subtracting the analysis-fold mean and dividing by the analysis-fold standard deviation; ER, PR and HER2 indicators were kept binary. Without further tuning the full training scaling parameters were then mapped to the locked validation cohort.

2.8 Partitioning of the sample size and data set

Essentially, there were 230 patients in the secondary data set: 130 source-designated training and 100 source-designated validation. Table 1 indicates that 48 patients had pCR and 182 residual disease; 33 and 15 in training and 15 and 85 in validation, respectively. In this way, the prevalence of pCR was 20.9% in general and varied depending on partitions. The small number of events limited the complexity of the models and class imbalance could lead to erroneous raw accuracy. Stratified resampling, class-weighted estimates and balanced performance measures were thus prespecified. No random repartition or expansion of sample-size was done post hoc.

2.9 Focused Model Development

The focused candidate set contained six estimable terms: MKI67, CD4 and CD8A expression plus ER, PR and HER2 status. Although a penalised logistic-regression framework was chosen, only 33 pCR events were available in training and the three receptor indicators and immune/proliferation measurements were correlable. Full retention of clinical question (four markers): CD274 remained matched instead of STAT1, CXCL9, PD-L2 or any other immune transcript.

Repeated stratified ten-fold cross-validation was performed on the 130 training cases to evaluate the elastic-net mixing parameter and the regularisation strength that yielded the inverse. Mixing values were varied between 0 and 1 and the C values were varied between .01 and 10. Main tuning criterion was mean out-of-fold log loss; secondary tuning criterion was ROC AUC. Outcome imbalance was countered by inverse-frequency class weights. Tuning chose a mixing value of 0 and C = .30; as a result, the final "fitted" model was the ridge boundary of this elastic-net grid. The classification cut-off (.499) was determined by out-of-fold model training predictions by maximising the Youden index.

Then, for each train case (all 130 cases), the model coefficients were fitted once after tuning and preprocessing. For the focused model, all 6 estimable terms were in the model. Individual pCR probabilities were obtained with the inverse-logit transformation of the model intercept and weighted predictors. For validation, no refitting, recalibration, feature replacement or threshold revision was done.

2.10 Model Performance and validation

ROC AUC used a 95% confidence interval to quantify discrimination. Sensitivity, specificity, accuracy, positive predictive value and negative predictive value were included at the locked threshold with CIs. In case of imbalance, supportive metrics were balanced accuracy and precision-recall AUC.

The examination of calibration was performed with the help of the a smoothed observed-versus-predicted calibration plot. Calibration intercept measured systematic over- or under-prediction, calibration slope was an evaluation of unwarranted or inadequate prediction variability, and Brier value was a measurement of the overall probabilistic error. Repeated stratified ten-fold cross-validation was used to estimate the optimism-corrected training performance. Validation metrics were created with 2,000 patient-level bootstrap resamples to construct ninety-five percent confidence intervals.

The model locked only applied to 100 of the cases in the validation group. The selection of a refitting, recalibration, feature replacement, preprocessing revision or threshold selection were not permitted in validation (Riley et al., 2024). If both pCR and residual-disease outcomes were available, descriptive subgroup AUCs were computed for HR+/HER2-, HR+/HER2+, HR-/HER2+ and triple-

negative phenotypes. Subgroup analysis results were not used to change the model.

2.11 Process of conducting statistical analysis and using software

Gaussian (continuous) variables were summarised as mean (SD) or median (IQR) and categorical variables as counts and percentages. The Welch's t test was used to compare marker distributions between cohorts. Univariable logistic regression was used to calculate unadjusted odds ratios with 95% CIs. The performance intervals for model discrimination and classification and calibration have been assessed as indicated above, using 2000 patient-level bootstrap resamples. Analyses were conducted in Python 3.12 with pandas, NumPy, SciPy and scikit-learn, all with fixed random seeds to ensure reproducibility (Python Software Foundation, 2026).

2.12 Data Governance and Ethical considerations

The GSE20194 consists of deidentified, publicly available secondary data; no direct interaction with the patients and no identifiers and confidential data were gathered (National Centre for Biotechnology Information, 2022). The GEO repository conditions were applied to files.

3. RESULTS

3.1 Patient Selection and Cohort Formation.

There were 278 records of GEO arrays within the source workbook. Upon elimination of 29 replicate arrays, 249 unique records were left; another 19 arrays with source quality-exclusion designation were eliminated. The confirmed analytic sample thus consisted of 230 patients, which represented 82.7 per cent of the records that were first screened. Among them, 130 (56.5) were the prespecified training cohort and 100 (43.5) were the locked validation cohort. Table 3 reports the arithmetic and cohort denominators, and Figure 1 reports the respective patient-selection pathway. The exclusion of the sources did not take away any patient based on outcome or predictor value.

Table 2. Patient selection and cohort formation

Selection stage	n	Percentage
Initial GEO array records	278	100.0% of initial
Replicate records excluded	29	10.4% of initial
Unique records after replicate removal	249	89.6% of initial
Quality-excluded records	19	7.6% of 249
Final analytic sample	230	82.7% of initial
Training cohort	130	56.5% of final
Validation cohort	100	43.5% of final

Note. Percentages for training and validation cohorts use the final analytic sample as the denominator.

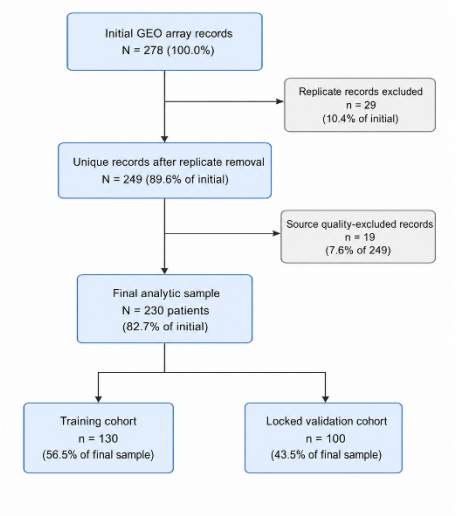

Figure 1. Patient-selection flow diagram

3.2 Baseline Patient, Tumour and Treatment Characteristics

Table 4 displays the baseline characteristics and Figure 2 presents a visual comparison. Mean age was 52.02 years (SD = 10.75), and 153 patients (66.5%) were White. The most common tumour category was T2 disease (57.4%), 164 patients (71.3%) were node positive, 123 (53.5%) had grade 3 tumours and 181 (78.7%) had pure invasive ductal carcinoma. ER, PR and HER2 positivity occurred in 61.3%, 45.2% and 17.4%, respectively. TFAC was administered to most patients (78.3%). There were no differences between training and validation cohorts in terms of age, race, T stage, N stage, ER or PR status (all Sig. > .05). The largest non-significant imbalances involved grade 3 disease (58.5% versus 47.0%, Sig. = .084) and pure IDC (74.6% versus 84.0%, Sig. = .085). HER2 positivity was higher in the training set compared to the validation set (Sig. < .001). There was also a difference in the regimen distributions (Sig. < .001): TFAC received 96.2% of training treatment but 55.0% of validation treatment, whereas the seven validation regimens were unknown.

Table 3. Baseline patient, tumour, and treatment characteristics

Characteristic	Overall (N = 230)	Training (n = 130)	Validation (n = 100)	Sig.
Age, years, mean (SD)	52.02 (10.75)	52.43 (11.44)	51.48 (9.81)	.499
Race: White	153 (66.5)	85 (65.4)	68 (68.0)	.849
↳ Black	25 (10.9)	13 (10.0)	12 (12.0)	—
↳ Hispanic	34 (14.8)	21 (16.2)	13 (13.0)	—
↳ Asian/other	18 (7.8)	11 (8.5)	7 (7.0)	—
T stage: T0–1	23 (10.0)	13 (10.0)	10 (10.0)	.609
↳ T2	132 (57.4)	70 (53.8)	62 (62.0)	—
↳ T3	34 (14.8)	21 (16.2)	13 (13.0)	—
↳ T4	41 (17.8)	26 (20.0)	15 (15.0)	—
N stage: N0	66 (28.7)	39 (30.0)	27 (27.0)	.935
↳ N1	107 (46.5)	60 (46.2)	47 (47.0)	—
↳ N2	27 (11.7)	14 (10.8)	13 (13.0)	—
↳ N3	30 (13.0)	17 (13.1)	13 (13.0)	—
Grade: Grade 1–2	107 (46.5)	54 (41.5)	53 (53.0)	.084
↳ Grade 3	123 (53.5)	76 (58.5)	47 (47.0)	—
Histology: Pure IDC	181 (78.7)	97 (74.6)	84 (84.0)	.085
↳ Other/combined	49 (21.3)	33 (25.4)	16 (16.0)	—
ER status: Negative	89 (38.7)	50 (38.5)	39 (39.0)	.934
↳ Positive	141 (61.3)	80 (61.5)	61 (61.0)	—
PR status: Negative	126 (54.8)	77 (59.2)	49 (49.0)	.122
↳ Positive	104 (45.2)	53 (40.8)	51 (51.0)	—
HER2 status: Negative	190 (82.6)	97 (74.6)	93 (93.0)	< .001
↳ Positive	40 (17.4)	33 (25.4)	7 (7.0)	—
Regimen: TFAC	180 (78.3)	125 (96.2)	55 (55.0)	< .001
↳ Non-TFAC	43 (18.7)	5 (3.8)	38 (38.0)	—
↳ Unknown	7 (3.0)	0 (0.0)	7 (7.0)	—

Note. Values are n (%) unless otherwise stated. Sig. values are from Welch's independent-samples t-test for age and Pearson chi-square tests for categorical variables. IDC = invasive ductal carcinoma; TFAC = paclitaxel followed by fluorouracil, doxorubicin, and cyclophosphamide.

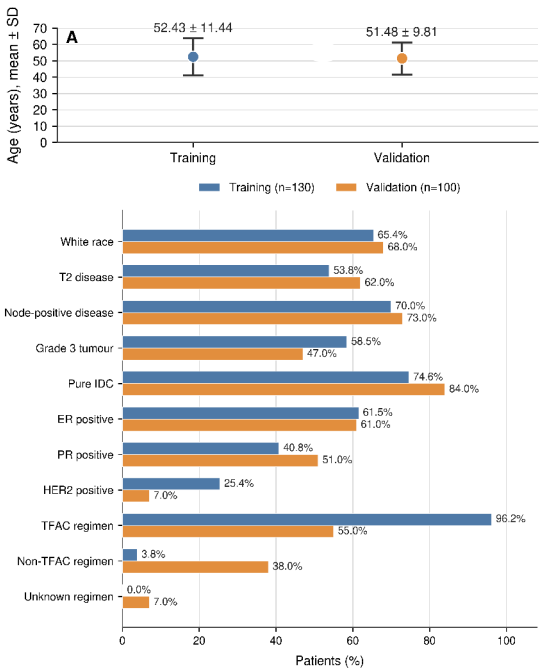

Figure 2. Comparison of baseline characteristics between cohorts

3.3 Distribution of Pathological Response

In total, 48/230 patients achieved pCR (20.9%) while 182 (79.1%) had residual disease as reported in Table 5. There were 33 pCR events (25.4%) in the training cohort and 15 pCR events (15.0%) in the validation cohort, while there were 97 and 85 residual-disease outcomes in the training and validation cohorts, respectively. There was a difference of 10.4 percentage points in absolute pCR-rate. The score difference between cohorts was close to being significant at the .05 level (Pearson Sig. = .055). These proportions are shown in Figure 3 and the imbalance between classes is made obvious. For this reason, accuracy was taken into account as well as discrimination, class-specific performance and calibration.

Table 4. Distribution of pathological response

Response	Overall (N = 230)	Training (n = 130)	Validation (n = 100)	Pearson Sig.
pCR	48 (20.9)	33 (25.4)	15 (15.0)	.055
Residual disease	182 (79.1)	97 (74.6)	85 (85.0)	—

Note. Values are n (%). pCR = pathological complete response.

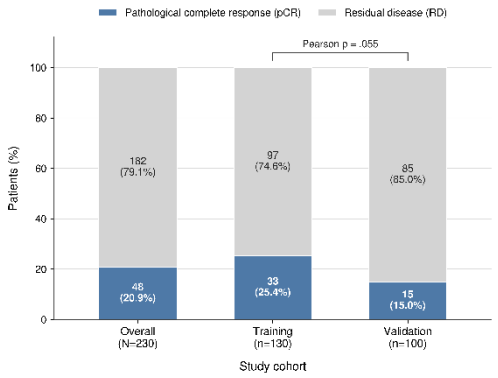

Figure 3. Pathological-response distribution by cohort

3.4 Focused Marker Availability, Distribution and Missingness

The four intended markers are summarised in Table 6 and Figure 4. MKI67 expression was similar between training and validation (8.06 [SD 1.30] versus 7.90 [SD 1.31], Sig. = .349). CD4 expression was higher in validation (8.53 [SD .83]) than training (8.21 [SD 1.14], Sig. = .014), and CD8A expression was also higher in validation (9.50 [SD .96]) than training (8.50 [SD 1.27], Sig. < .001). All three of these were expression variables that were complete. PD-L1 could not be summarised because CD274 is not represented on GPL96. In 230 cases it was not available, no value was imputed or inferred for it.

Table 6. Availability, distributions and missingness of the focused markers

Marker	Gene / selected probe	Training mean (SD)	Validation mean (SD)	Missing or unavailable, n	Sig.
Ki-67	MKI67 / 212022_s_at	8.06 (1.30)	7.90 (1.31)	0	.349
CD4	CD4 / 203547_at	8.21 (1.14)	8.53 (.83)	0	.014
CD8	CD8A / 205758_at	8.50 (1.27)	9.50 (.96)	0	< .001
PD-L1	CD274 / no GPL96 probe set	Not available	Not available	230	—

Note. Values are GEO-processed expression units. Sig. values are from Welch's t tests comparing training and validation. PD-L1 is unavailable because no CD274 probe set is present on GPL96; this is a platform-level limitation rather than ordinary patient-level missingness.

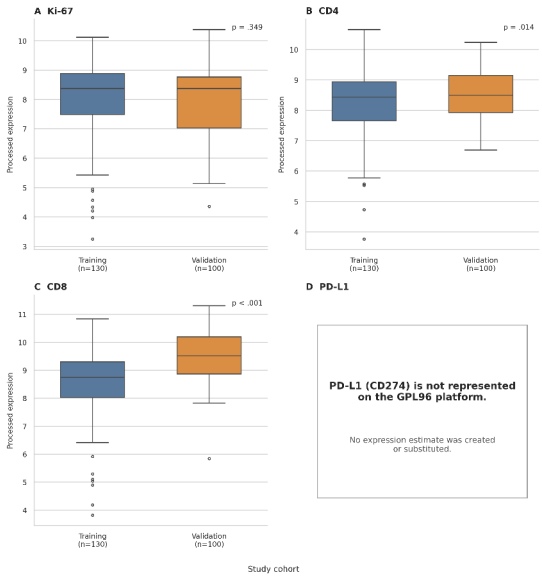

Figure 4. Distributions of Ki-67, CD4 and CD8 expression and documented PD-L1 platform availability

3.5 Focused Univariable Associations With pCR

The results of the focused univariable analysis are presented in Table 7 and Figure 5. ER positivity was associated with lower odds of pCR (OR = .069, 95% CI [.025, .188]), as was PR positivity (OR = .182, 95% CI [.065, .511]). HER2 positivity was associated with higher odds (OR = 2.502, 95% CI [1.065, 5.879]). Per training-cohort standard deviation, higher MKI67 expression had OR = 3.220 (95% CI [1.635, 6.340]), higher CD4 expression had OR = 1.884 (95% CI [1.135, 3.129]), and higher CD8A expression had OR = 1.745 (95% CI [1.050, 2.900]). The three marker associations were all significant. As there was no measurement of CD274, there was no estimate of PD-L1. The results presented are unadjusted and not effects of the final model.

Table 7. Focused univariable logistic-regression associations with pCR in the training cohort

Predictor	n	B	S.E.	Wald	df	Sig.	Exp(B)	95% CI for Exp(B)
ER positive	130	-2.673	.511	27.400	1	< .001	.069	.025-.188
PR positive	130	-1.702	.526	10.461	1	.001	.182	.065-.511
HER2 positive	130	.917	.4336	4.431	1	.035	2.502	1.065-5.879
Ki-67 (MKI67), per SD	130	1.169	.3446	11.446	1	< .001	3.220	1.635-6.340
CD4, per SD	130	.634	.2595	5.995	1	.014	1.884	1.135-3.129
CD8 (CD8A), per SD	130	.557	.2595	4.622	1	.032	1.745	1.050-2.900
PD-L1 (CD274)	0	—	—	—	—	—	—	Not estimable

Note. Molecular-expression effects are reported per one training-cohort SD. Binary receptor effects compare the positive category with the negative reference category. PD-L1 was not estimable because CD274 was absent from GPL96. Results are unadjusted. Exp(B) = odds ratio.

3.6 Focused Final Model and Coefficients

Repeated training-cohort tuning chose a ridge model based on elastic-net grid (mixing value = 0; C = .30). All six honorable words were retained (Table 8). ER positivity had the largest absolute coefficient (B = -1.196), followed by MKI67 expression (B = .654), PR positivity (B = -.569), HER2 positivity (B = .335), CD4 expression (B = .229) and CD8A expression (B = .036). The fitted intercept was .434. Using training-standardised expression values, the linear predictor was: logit(pCR) = .434 + .654z(MKI67) + .229z(CD4) + .036z(CD8A) - 1.196(ER+) - .569(PR+) + .335(HER2+). The predicted probability was predicted as 1/[1 + exp(-logit)]. A pCR was defined as having a probability of .499 or greater. The term PD-L1 was not a zero coefficient term, it was not included in the source platform and hence was not fitted.

Table 8. Coefficients in the focused receptor-plus-marker penalised logistic model

Rank	Retained term	B	Exp(B)	Direction for pCR
1	ER positive versus negative	-1.196	.302	Lower probability
2	Ki-67 (MKI67) expression	.654	1.923	Higher probability
3	PR positive versus negative	-.569	.566	Lower probability

4	HER2 positive versus negative	.335	1.398	Higher probability
5	CD4 expression	.229	1.257	Higher probability
6	CD8 (CD8A) expression	.036	1.037	Higher probability

Note. Terms are ordered by absolute coefficient. Expression predictors are reported per one training-derived standardisation unit; receptor indicators compare positive with negative status. The model intercept was .434. Exp(B) values are descriptive penalised-model estimates, and conventional inferential confidence intervals are not reported.

3.7 Training-Cohort Performance

The AUC was .808 (95% CI [.718, .890]) for the repeated out-of-fold training predictions (Table 9, Figure 6). At the .499 threshold, sensitivity was .848, specificity .773, accuracy .792, PPV .560 and NPV .938. The 28 true positives, 75 true negatives, 22 false positives and five false negatives resulted in a level of confusion. The Brier score was .173 (95% CI [.140, .206]). Calibration intercept was -1.024 (95% CI [-1.490, -.557]) and slope was 1.081 (95% CI [.630, 1.532]), indicating systematic probability overestimation despite useful ranking.

3.8 Validation-Cohort Performance

Locked evaluation of the 100-case validation cohort yielded an AUC of .775 (95% CI [.630, .894]) (Table

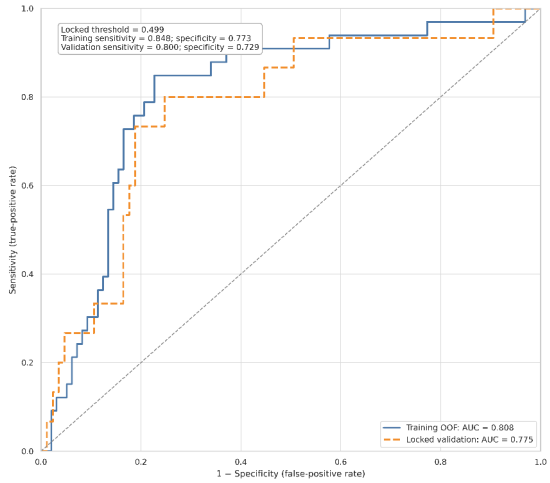

Figure 6. ROC curves for the focused model in training and locked validation

3.9 Sensitivity and Receptor-Phenotype Analyses

The focused sensitivity analyses are presented in Table 9. The ER/PR/HER2 receptor-only model achieved a validation AUC of .811, whereas the available-marker-only model achieved a validation

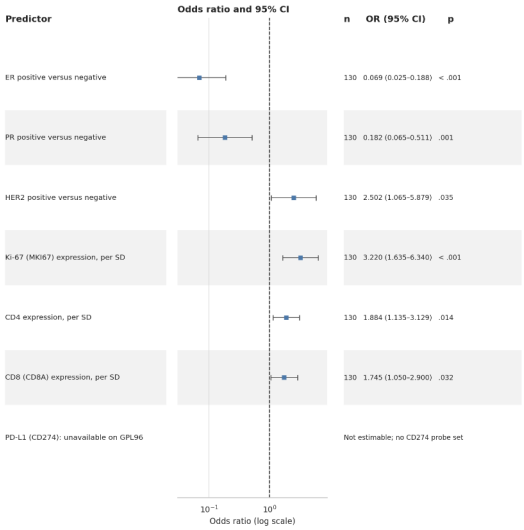

Figure 5. Focused univariable associations with pathological complete response

9 and Figure 6). At the unchanged .499 threshold, sensitivity was .800 (95% CI [.579, 1.000]), specificity .729 (95% CI [.629, .820]), accuracy .740 (95% CI [.640, .820]), PPV .343 (95% CI [.186, .500]) and NPV .954 (95% CI [.897, 1.000]). The confusion matrix contained 12 true positives, 62 true negatives, 23 false positives and three false negatives. Brier score was .186 (95% CI [.145, .235]). Calibration intercept was -1.689 (95% CI [-2.321, -1.058]) and slope was .917 (95% CI [.365, 1.469]) (Figure 7). AUC was .033 lower and Brier score was .013 higher compared to training. There are 15 pCR events for the large interval (validation).

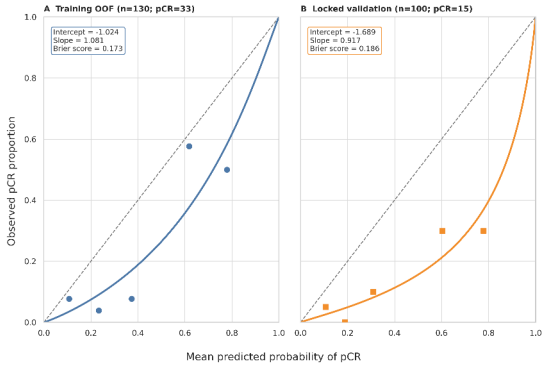

Figure 7. Calibration of focused-model pCR probabilities in training and locked validation

AUC of .645 and the focused receptor-plus-marker model a validation AUC of .775. Therefore, MKI67, CD4 and CD8A did not add to the validation discrimination that receptor status alone. Due to an unavailability of PD-L1, a complete four marker

model was not possible. Validation AUCs were .443 in HR+/HER2- disease (63 patients, two pCR events), .375 in HR-/HER2+ disease (six patients, four events) and .460 in triple-negative disease (30 patients, nine events). HR+/HER2+ disease: AUC was not estimable due to 1 validation patient not achieving pCR. These subgroup data are descriptive and not very precise in any way. The focused combined model achieved validation sensitivity .800, specificity .729, accuracy .740, PPV .343 and NPV .954 at the prespecified locked .499 classification threshold.

Table 9. Focused-model performance, locked-validation confusion matrix, and sensitivity analyses

Panel A. Primary focused-model performance estimates

Cohort	AUC (95% CI)	Sensitivity	Specificity	Accuracy	PPV	NPV	Brier score	Calibration intercept; slope
Training OOF	.808 (.718-.890)	.848 (.710-.967)	.773 (.686-.856)	.792 (.723-.862)	.560 (.417-.692)	.938 (.878-.987)	.173 (.140-.206)	-1.024 [-1.490, -.557]; 1.081 [.630, 1.532]
Locked validation	.775 (.630-.894)	.800 (.579-1.000)	.729 (.629-.820)	.740 (.640-.820)	.343 (.186-.500)	.954 (.895-1.000)	.186 (.145-.235)	-1.689 [-2.321, -1.058]; .917 [.365, 1.469]

Panel B. Locked-validation confusion matrix at the .499 threshold

Actual validation outcome	Predicted residual disease	Predicted pCR	Total
Residual disease	62	23	85
pCR	3	12	15
Total	65	35	100

Panel C. Focused marker and ER/PR/HER2 phenotype analyses

Validation analysis	n	pCR events	AUC	Sensitivity	Specificity	Accuracy	PPV	NPV
ER/PR/HER2 receptor-only model	100	15	.811	—	—	—	—	—
Available-marker-only model	100	15	.645	—	—	—	—	—
Focused receptor-plus-marker model	100	15	.775	—	—	—	—	—
Complete	0	—	—	—	—	—	—	—

four-marker panel								
HR+/HER2+ subgroup	1	0	—	—	—	—	—	—
HR+/HER2- subgroup	63	2	.443	—	—	—	—	—
HR-/HER2+ subgroup	6	4	.375	—	—	—	—	—
Triple-negative	30	9	.460	—	—	—	—	—
subgroup								
Focused combined model, threshold = .499	100	15	.775	.800	.729	.740	.343	.954

Note. The locked primary threshold was .499. Performance intervals used 2,000 patient-level bootstrap resamples. Training estimates used repeated out-of-fold predictions. HR positivity means ER or PR positivity. OOF = out of fold; PPV = positive predictive value; NPV = negative predictive value. PD-L1 was not estimable because GPL96 contains no CD274 probe set. Subgroup AUCs with very small event counts must not be interpreted as stable inferential conclusions.

4. DISCUSSION

4.1 Key Findings

This targeted retrospective study assessed the intended Ki-67, CD4, CD8 and PD-L1 panel with the ER/PR/HER2 classification left intact. Of 230 patients, 48 (20.9%) achieved pCR. Three desired measurements on the transcripts were available: MKI67, CD4, and CD8A. No measurable PD-L1 was found as GPL96 does not have a probe set for this gene. The focused model included the three marker expressions and ER, PR and HER2 status and performed with training AUC .808 and locked-validation AUC .775, .800 sensitivity and .729 specificity. The available-marker-only AUC was .645, however the receptor status alone yielded a higher AUC for validation (.811). However, the overall discrimination of the model was acceptable, but the three expression markers did not provide any validation benefit and without PD-L1 data the model could not solve the complete four-marker question.

4.2 Clinical and Biological Interpretation

The negative correlation between ER and PR positivity with pCR and the positive correlation with HER2 positivity confirmed the advantage of keeping the ER/PR/HER2 classification with them (Zhao et al., 2024; Prat et al., 2022). The best marker association was for MKI67, which is indicative of higher sensitivity to chemotherapy of tumor cells with fast proliferation rates. CD4 and CD8A also were positively associated with pCR in univariable analysis, indicating an immune-infiltration signal. If the other terms in focus were adjusted for out, CD4 still had a small positive correlation, while the CD8A had a near zero correlation. These transcript measurements are not the same as Ki-67, CD4 or CD8 IHC scores.

The immune-marker result is a specific one and should be interpreted with care. The abundance of immune cells may be represented in the CD4 and CD8 transcripts, but the fine-needle aspirates included in GSE20194 were not representative of the immune microenvironment (National Center for Biotechnology Information, 2022), because they had relatively high levels of neoplastic cells and relatively low levels of stromal cells. Immune gene-expression signatures may have different performance than direct proteins or tumour-infiltrating lymphocytes (Fernandez-Martinez et al., 2023). Most importantly, the biology of PD-L1 will not be deduced from CD4, CD8A or other checkpoint ligand; the data would have to be obtained directly from a suitable transcriptomic or IHC dataset.

The best marker association was for MKI67, which is indicative of higher sensitivity to chemotherapy of tumor cells with fast proliferation rates. CD4 and CD8A also were positively associated with pCR in univariable analysis, indicating an immune-infiltration signal, consistent with previous evidence linking Ki-67 and tumour-infiltrating lymphocyte

markers with response to neoadjuvant chemotherapy (Kumar, n.d.).

4.3 Results of the study with those of other research.

The pCR rate at 20.9% was lower than reported in the multi-centre modelling study of 36.3% (Jung et al., 2023), but was higher than reported in the National Cancer Database (NCDB) study (56,209 patients; Zhao et al., 2024). It was more similar to the pCR rate of the TransNEO multi-omic cohort (Sammut et al., 2022) at 26%. The disparities could be attributed to the use of mostly TFAC in the present cohort and receptor-directed HER2 therapy and chemoimmunotherapy (CIT) in contemporary populations for triple-negative disease (Schmid et al, 2022).

The AUC in the focused validation was 0.775, much like the AUC of 0.785 found by Zhao et al. (2024) in a similar study, although it was lower than some of the other multi-omic models (Jung et al., 2023; Sammut et al., 2022). The receptor-only model was superior to the receptor-plus-marker model in the present analysis, with limited marker-only discrimination. This does not imply that the biological value of Ki-67, CD4, CD8 or PD-L1 is not important. However, it suggests that the three microarray transcripts available were not transportable discriminators in this limited historical validation cohort, and the fourth requested transcript was not evaluable.

4.4 Results of the clinical relevance and potently useful item

A focused pretreatment marker model may eventually aid in response counselling and/or selection for response-adapted research, but this model is not ready for clinical use. Measurement of the requested panel should be obtained for the entire panel, consistently, ideally via prespecified IHC assays or newer transcriptomic assays, and then analysed within ER/PR/HER2-defined groups. The absence of a strong prediction of pCR should not be a reason to withhold standard treatment. There is still a need for calibration and assessment of net benefit and potential clinical impact prior to implementation (Collins et al., 2024; Zhao et al., 2024).

The validation NPV was high (0.954), while the PPV was low (0.343) and the intercept of the calibration was negative and systematic (1.689). A future calculator would have to include the predicted probability, uncertainty, ER/PR/HER2 phenotype and contribution of each measured marker. It would also have to be specific to the assay type with the ability to not return a PD-L1 result if CD274 or PD-L1 IHC was not measured.

4.5 Strengths

The analysis used public, de-identified and auditable data with transparent sample selection. Unrelated proliferation, DNA-repair, angiogenic and epithelial-mesenchymal genes were not retained as molecular scope was limited to four genes requested by the supervisor. Outcome blinding was done for exact probe mapping, the validation cohort was locked and ER/PR/HER2 were preserved. Another benefit was the reporting of unavailability of PD-L1 (rather than the creation, imputation or substitution of a CD274 measurement). All three of the following were computed: discrimination, calibration, and Brier score and confusion counts.

4.6 Limitations

This was a secondary analysis of a historical, single source cohort. The imprecision of the subgroups estimates were due to a small number of validations and pCR events (N = 15) were available (N = 48). There was a single HR+/HER2+ patient in the validation set and there were significant differences between training and validation sets for HER2 positivity and treatment modality. Ki-67, CD4 or CD8 IHC protein scores were not measured, but rather microarray transcripts of MKI67, CD4 and CD8A. The entire supervisor-specified four-marker panel has not been tested as PD-L1 is not present on GPL96. Also, FNA can misrepresent infiltrating immune cells if the purity of the tumour-cell population is high. Probability overestimation occurred due to class weighting and neither good ranking nor high NPV means clinical utility. Limited generality with contemporary HER2 targeted therapy and chemo-immunotherapy.

4.7 Future Research

Future studies should involve an independent contemporary cohort in which all of the following markers (Ki-67, CD4, CD8 and PD-L1) are measured by the same, prespecified assay. The four marker panel should be assessed in ER/PR (plus HER2) defined phenotypes, and sufficient pCR events in each. MicroRNA expression (direct IHC scores) should not be used interchangeably with modern RNA-sequencing measurements of MKI67, CD4, CD8A/CD8B and CD274. External validation, re-calibration, decision-curve analysis, fairness assessment and prospective clinical-impact testing is needed before use.

5. CONCLUSION

5.1 Overall Conclusion

This targeted reanalysis saved ER/PR/HER2-based breast cancer classification and limited the panel of molecules intended to be analyzed to Ki-67, CD4, CD8, and PD-L1. For GSE20194/GPL96, used expression measurements of MKI67, CD4 and CD8A, but not CD274; therefore, no estimation or replacement of PD-L1. The focused receptor-plus-marker model had a higher AUC .808 in the training set and .775 in the locked-validation set, with sensitivity of .800, specificity of .729, and NPV of .954. The model based on receptor status alone had a better performance (AUC .811) than the model with the two markers (AUC .645). The results will continue to guide research into the markers indicated, but will not confirm the full four-marker panel or readiness for clinical use.

5.2 Clinical and research implications of 5.2K

The short-term research lesson is that the study shouldn't use a PD-L1 finding from GSE20194. They need a new or additional dataset that contains all four markers to answer the question the supervisor is asking. There, the receptor phenotype (ER, PR, HER2) should remain the foundation and incremental value of Ki-67, CD4, CD8 and PD-L1 should be assessed against this receptor-only baseline. Any future decision support system should provide results for assay type, uncertainty, calibration and what would happen if markers were missing.

REFERENCES

1. Chai, Y., Chen, Y., Zhang, D., Wei, Y., Li, Z., Li, Q., & Xu, B. (2022). Homologous recombination deficiency and BRCA1/2 gene mutation for predicting the effect of platinum-based neoadjuvant chemotherapy of early-stage triple-negative breast cancer: A systematic review and meta-analysis. *Journal of Personalized Medicine, 12*(2), 323. https://doi.org/10.3390/jpm12020323
2. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378
3. de Graaf, M., Gielen, R. C. A. M., Balduzzi, S., Siesling, S., Linn, S. C., & Kok, M. (2025). Pathological complete response and survival after neoadjuvant chemotherapy in patients with stage I TNBC: A registry-based study. *ESMO Open, 10*(12), 105923. https://doi.org/10.1016/j.esmoop.2025.105923
4. Derouane, F., van Marcke, C., Berlière, M., Gerday, A., Fellah, L., Leconte, I., Van Bockstal, M. R., Galant, C., Corbet, C., & Duhoux, F. P. (2022). Predictive biomarkers of response to neoadjuvant chemotherapy in breast cancer: Current and future perspectives for precision medicine. *Cancers, 14*(16), 3876. https://doi.org/10.3390/cancers14163876
5. Dowling, G. P., Keelan, S., Toomey, S., Daly, G. R., Hennessy, B. T., & Hill, A. D. K. (2023). Review of the status of neoadjuvant therapy in HER2-positive breast cancer. *Frontiers in Oncology, 13*, 1066007. https://doi.org/10.3389/fonc.2023.1066007
6. Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ, 386*, e078276. https://doi.org/10.1136/bmj-2023-078276
7. EQUATOR Network. (2026, June 8). *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies.* https://www.equator-network.org/reporting-guidelines/strobe/
8. Fang, S., Xia, W., Zhang, H., Ni, C., Wu, J., Mo, Q., Jiang, M., Guan, D., Yuan, H., & Chen, W. (2024). A real-world clinicopathological model for predicting pathological complete response to neoadjuvant chemotherapy in breast cancer. *Frontiers in Oncology, 14*, 1323226. https://doi.org/10.3389/fonc.2024.1323226
9. Fernandez-Martinez, A., Pascual, T., Singh, B., Nuciforo, P., Rashid, N. U., Ballman, K. V., Campbell, J. D., Hoadley, K. A., Spears, P. A., Paré, L., Brasó-Maristany, F., Chic, N., Krop, I., Partridge, A., Cortés, J., Llombart-Cussac, A., Prat, A., Perou, C. M., & Carey, L. A. (2023). Prognostic and predictive value of immune-related gene-expression signatures versus tumour-infiltrating lymphocytes in early-stage ERBB2/HER2-positive breast cancer. *JAMA Oncology, 9*(4), 490–499. https://doi.org/10.1001/jamaoncol.2022.6288
10. Jung, J.-J., Kim, E.-K., Kang, E., Kim, J. H., Kim, S. H., Suh, K. J., Kim, S. M., Jang, M., Yun, B. L., Park, S. Y., Lim, C., Han, W., & Shin, H.-C. (2023). Development and external validation of a machine-learning model to predict pathological complete response after neoadjuvant chemotherapy in breast cancer. *Journal of Breast Cancer, 26*(4), 353–362. https://doi.org/10.4048/jbc.2023.26.e14
11. Kumar, S., Saini, S. K., & Verma, S. K. (2026). Role of Ki-67, tumor-infiltrating lymphocytes (TILs-CD4 and CD8), and PD-L1 in predicting pCR after neoadjuvant chemotherapy for breast cancer. *International Journal of Science and Technology (IJST), 3*(2), 230–241. https://doi.org/10.70558/IJST.2026.v3.i2.241275
12. Liang, Y.-K., Deng, Z.-K., Chen, M.-T., Qiu, S.-Q., Xiao, Y.-S., Qi, Y.-Z., Xie, Q., Wang, Z., Jia, S., Zeng, D., & Lin, H. (2021). CXCL9 is a potential biomarker of immune infiltration associated with favourable prognosis in ER-negative breast cancer. *Frontiers in Oncology, 11*, 710286. https://doi.org/10.3389/fonc.2021.710286
13. National Center for Biotechnology Information. (2021, March 9). *GPL96: [HG-U133A] Affymetrix Human Genome U133A Array.* Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96

14. National Center for Biotechnology Information. (2022, June 6). *GSE20194: MAQC-II project: Human breast cancer (BR) data set.* Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20194
15. Prat, A., Guarneri, V., Pascual, T., Brasó-Maristany, F., Sanfeliu, E., Paré, L., Schettini, F., Martínez, D., Jares, P., Griguolo, G., Dieci, M. V., Cortés, J., Llombart-Cussac, A., Conte, B., Marín-Aguilera, M., Chic, N., Puig-Butillé, J. A., Martínez, A., Galván, P., ... Perou, C. M. (2022). Development and validation of the new HER2DX assay for predicting pathological response and survival outcome in early-stage HER2-positive breast cancer. *EBioMedicine, 75*, 103801. https://doi.org/10.1016/j.ebiom.2021.103801
16. Python Software Foundation. (2026). *Python* (Version 3.12) [Computer software]. https://www.python.org/
17. Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ, 384*, e074820. https://doi.org/10.1136/bmj-2023-074820
18. Sammut, S.-J., Crispin-Ortuzar, M., Chin, S.-F., Provenzano, E., Bardwell, H. A., Ma, W., Cope, W., Dariush, A., Dawson, S.-J., Abraham, J. E., Dunn, J., Hiller, L., Thomas, J., Cameron, D. A., Bartlett, J. M. S., Hayward, L., Pharoah, P. D., Markowetz, F., Rueda, O. M., ... Caldas, C. (2022). Multi-omic machine-learning predictor of breast cancer therapy response. *Nature, 601*, 623–629. https://doi.org/10.1038/s41586-021-04278-5
19. Schmid, P., Cortes, J., Dent, R., Pusztai, L., McArthur, H., Kümmel, S., Bergh, J., Denkert, C., Park, Y. H., Hui, R., Harbeck, N., Takahashi, M., Foukakis, T., Fasching, P. A., Cardoso, F., Zhao, J., Aktan, G., Karantza, V., O'Shaughnessy, J., & KEYNOTE-522 Investigators. (2022). Event-free survival with pembrolizumab in early triple-negative breast cancer. *The New England Journal of Medicine, 386*(6), 556–567. https://doi.org/10.1056/NEJMoa2112651
20. van den Ende, N. S., Nguyen, A. H., Jager, A., Kok, M., Debets, R., & van Deurzen, C. H. M. (2023). Triple-negative breast cancer and predictive markers of response to neoadjuvant chemotherapy: A systematic review. *International Journal of Molecular Sciences, 24*(3), 2969. https://doi.org/10.3390/ijms24032969
21. World Health Organization. (2026, July 3). *Breast cancer.* https://www.who.int/news-room/fact-sheets/detail/breast-cancer
22. Zhao, F., Polley, E., McClellan, J., Howard, F. M., Olopade, O. I., & Huo, D. (2024). Predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer using a machine learning approach. *Breast Cancer Research, 26*, 148. https://doi.org/10.1186/s13058-024-01905-7