Research Square

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

Proliferation Pathway Scores Predict Neoadjuvant Chemotherapy Response in Breast Cancer

Sujoy Banik
Rajshahi University of Engineering and Technology
Ushashi Bhattacharjee
Iowa State University
Boishakhi Das Toma
American International University-Bangladesh
Tirtho Roy
tirtho@iastate.edu

Iowa State University

Method Article

Keywords: Breast neoplasms, Neoadjuvant therapy, Pathologic complete response, Gene expression profiling, Hallmark gene sets, Treatment response prediction, Melanoma, Immunotherapy, Data leakage

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11579839/v1

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

Additional Declarations: No competing interests reported.

Proliferation Pathway Scores Predict Neoadjuvant Chemotherapy Response in Breast Cancer

Authors and affiliations removed for double-anonymous review

Abstract

Background: Curated pathway programs are widely used to predict therapy response from pretreatment tumor expression. Whether they add to clinical variables has rarely been tested with locked rules, untouched cohorts and patient deduplication.

Methods: A protocol frozen before any model fit defined the cohorts after a sample-level audit. HER2-negative breast cancer treated with neoadjuvant chemotherapy was the primary setting: GSE25055 (303 patients) for development, GSE25065 (180) for untouched validation and three cohorts for transportability; three melanoma immunotherapy cohorts were exploratory. Fifty Hallmark pathway scores, six clinical variables and 2,000 genes entered penalized logistic models. Controls included matched random gene sets and permuted labels. A hypothesis locked after these analyses was tested in five unseen breast cohorts, and a locked audit compared every array across 46 public breast series.

Results: Pretreatment pathway scores predicted pathologic complete response (pCR) in the untouched cohort (AUROC 0.706, AP 0.411, prevalence 0.23). The signal was proliferation: three proliferation sets matched the full 50-set model, and proliferation pointed toward pCR in all five initial breast cohorts after adjustment for receptor status and grade. In five unseen breast cohorts (1,121 patients, 285 pCR events), the frozen clinical + proliferation model improved AP over the frozen clinical model (pooled difference 0.042, 95% CI 0.011 to 0.073), although the recalibrated Brier score did not improve. Immune programs carried the signal in MAGE-A3 melanoma immunotherapy. Across 4,481 audited arrays, 401 were identical copies of arrays deposited earlier in another series, in 11 undeclared series pairs named together in at least 102 articles; among them were 53 copies of external-validation patients, and retaining duplicates inflated validation AUROC by up to 0.25. The full pathway library did not improve on clinical variables, and expression-matched random gene sets performed as well.

Conclusions: Proliferation is the transportable pretreatment signal of chemotherapy response in HER2-negative breast cancer. Three proliferation programs capture what a 50-program library offers and improve the ranking of patients beyond clinical variables in unseen cohorts. Public breast expression series contain hundreds of undeclared repeat arrays,

so array-level deduplication is essential before multi-cohort validation.

Keywords: Breast neoplasms; Neoadjuvant therapy; Pathologic complete response; Gene expression profiling; Hallmark gene sets; Treatment response prediction; Melanoma; Immunotherapy; Data leakage

1 Background

Tumors that share a histology and stage respond differently to the same therapy. In HER2-negative breast cancer treated before surgery with taxane and anthracycline chemotherapy, fewer than one in four patients reach pathologic complete response (pCR) [1], and pCR is associated with long-term survival at the patient level [2]. Clinical variables such as estrogen receptor (ER) status, grade and stage carry part of this information, but they summarize tumor biology through a few coarse categories.

Gene expression measured in a pretreatment biopsy records the state of the molecular programs that therapy acts on. Cytotoxic chemotherapy kills dividing cells, so proliferation programs should matter; immune-mediated tumor killing depends on a pre-existing immune response, so interferon and inflammatory programs should matter [3, 4]. Hallmark gene sets summarize such programs as 50 coherent biological states [5, 6], and single-sample scores turn each set into one number per patient [7]. These scores measure coordinated transcript abundance, not biochemical pathway activity, and this article calls them pathway scores.

Earlier work has answered parts of this question. A meta-analysis of 1,419 pretreatment breast cancer biopsies from 17 studies tested 125 expression signatures against pCR and found that most of their signal tracked proliferation, with immune signatures adding information in some subtypes [8]. Molecular, clinical and combined models of breast cancer pCR have been validated across cohorts [9]. In melanoma treated with anti-PD-1 antibodies, elastic-net models built on pathway signatures predicted response, more strongly from on-treatment than from pretreatment biopsies [10], and pathway and interaction-network features have been combined for checkpoint-inhibitor response [11]. Interpretable pathway prediction of response is therefore not new.

Three questions remain open. First, does a curated pathway representation add information to clinical variables in an untouched cohort, judged on probability-based metrics rather than discrimination alone [12, 13]? Second, is that information specific to curated programs? Random gene signatures can predict breast cancer outcome about as well as published ones, because a

few dominant programs such as proliferation correlate with thousands of genes [14, 15], so a pathway model that beats a null model has not shown that its biology is specific. Third, does the answer survive patient deduplication? Public cohorts are reused across GEO series, so the same patient can appear in development and validation data, and such hidden duplicates inflate apparent accuracy [16]. How often this happens across public breast cancer series has not been measured.

This study tests these three questions in HER2-negative breast cancer treated with neoadjuvant taxane-anthracycline chemotherapy, the primary setting, with one development cohort, one untouched external cohort [1] and three further breast cohorts for transportability [17–19]. Three melanoma immunotherapy cohorts [20–22] serve as exploratory context: the same feature representation is refitted within each, so they test whether the representation carries response information in a second disease and therapy class, not whether a breast model transfers. A protocol frozen before any model fit (Table 1) fixed every primary choice (Fig. 1).

Central question. Which pretreatment tumor programs carry transportable therapy-response information, and how much of a curated pathway library is needed to capture it once clinical variables, matched random sets and duplicate patients are accounted for? H1 tests incremental value over clinical variables in an untouched cohort. H2 compares pathway-level and gene-level models. H3 asks whether frozen models need local calibration. H4 identifies which programs carry the response signal in each disease context. H5, locked after these analyses, tests the resulting proliferation model in five breast cohorts unseen by the study.

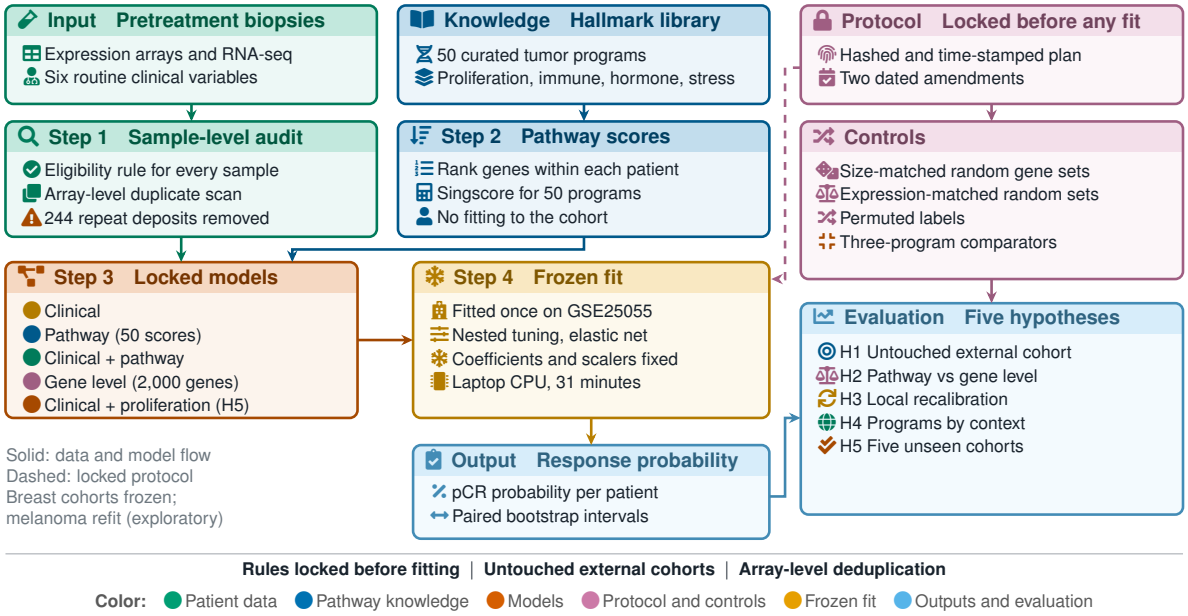


Fig. 1: Overview of the study. Pretreatment expression becomes 50 Hallmark pathway scores, four locked models are fitted once on GSE25055, and the frozen fits are evaluated on breast cohorts; melanoma cohorts refit the same representation as exploratory context. Circles mark where hypotheses 1 to 4 are tested; H5 (amendment 2) tests the proliferation model in five unseen breast cohorts (Fig. 4d). AP, average precision.

2 Methods

2.1 Study design and reporting

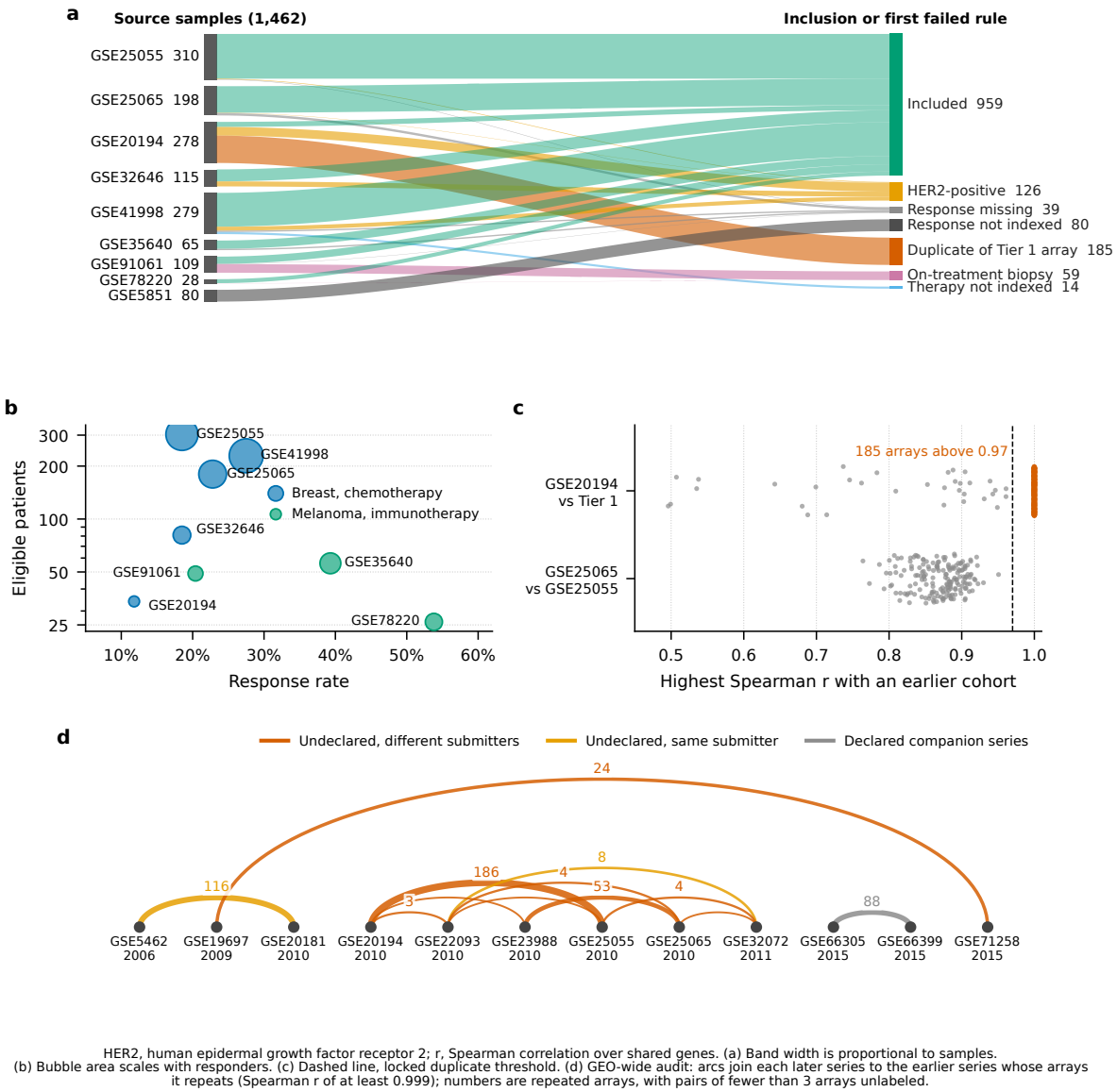
This study tests a predictive hypothesis with public, de-identified data. The written analysis protocol fixed the cohorts and their roles, eligibility rules, endpoints, predictors, pathway library, model families, tuning rule, validation design, metrics, robustness analyses and the decision rule for each hypothesis (Table 1). The protocol was frozen, and its SHA-256 hash and a UTC timestamp were recorded, before any model was fitted; the pipeline refuses to run if the file changes. Before freezing, the authors had inspected cohort metadata (response counts, covariate completeness and the cross-cohort duplicate scan) but no model output. The hash shows that the file did not change after freezing; it cannot by itself show what was known before, so the timeline is disclosed here and in the repository log. After an external pre-submission review, a first dated amendment added six analyses before they were run; all are labeled exploratory. A second dated amendment was written after the amendment 1 results were known and before any data from new cohorts were downloaded; it stated hypothesis H5, its decision rule and the candidate cohorts, and committed to reporting the result whatever its direction. Reporting follows TRIPOD and TRIPOD+AI [23, 24], and the design addresses the PROBAST risk-of-bias

domains [25].

2.2 Cohort discovery and eligibility

Candidate cohorts came from the Gene Expression Omnibus (GEO) [26] and their source publications. A cohort entered the study only if it met five conditions, each checked at the sample level: (1) human tumor tissue collected before the indexed therapy; (2) a single defined therapy or a defined regimen family; (3) a response label indexed in the repository record for that therapy, not inferred from survival or prognosis fields; (4) one record per patient, with patient linkage that can be verified; and (5) usable genome-wide expression data. In breast cancer cohorts, tumors that were HER2-positive by immunohistochemistry or fluorescence in situ hybridization were excluded, because the primary cohorts were designed for HER2-negative disease and HER2-targeted therapy changes response biology. Each source sample received an audit row that records inclusion or the first rule it failed (Fig. 2a). Replicate baseline samples from one patient were averaged after normalization, a rule fixed before outcomes were examined.

Eight cohorts passed the audit and one candidate failed (Table 2). The primary setting contains the two HER2-negative breast cancer cohorts of Hatzis et al. [1], treated with neoadjuvant taxane followed by anthracycline chemotherapy: GSE25055 for development (303 patients, 56 with pathologic complete response [pCR]) and GSE25065 as the untouched external validation cohort (180 patients, 41 with pCR). The source study reported 508 enrolled patients and 488 with evaluable response; the audit reproduced this difference as 20 samples whose repository response field is missing, and also removed six HER2-positive samples, one of which lacked a response record. Tier 2 contains three further HER2-negative breast cohorts with neoadjuvant chemotherapy and a pCR endpoint: GSE20194 (MAQC-II, paclitaxel then FAC) [17, 27], GSE32646 (paclitaxel then FEC) [18] and GSE41998 (doxorubicin and cyclophosphamide, then ixabepilone or paclitaxel) [19]. Tier 3 contains three exploratory melanoma cohorts: GSE91061 (nivolumab; pretreatment biopsies only) [21] and GSE78220 (anti-PD-1 therapy) [22], both checkpoint blockade, and GSE35640 (MAGE-A3 antigen-specific immunotherapy) [20], a different immunotherapy class that is analyzed separately. The colorectal cancer cetuximab cohort GSE5851 [28] failed the audit because its repository record does not index the response labels.



HER2, human epidermal growth factor receptor 2; r, Spearman correlation over shared genes. (a) Band width is proportional to samples. (b) Bubble area scales with responders. (c) Dashed line, locked duplicate threshold. (d) GEO-wide audit: arcs join each later series to the earlier series whose arrays it repeats (Spearman r of at least 0.999); numbers are repeated arrays, with pairs of fewer than 3 arrays unlabeled.

Fig. 2: Sample-level audit. (a) Flow of every source sample to inclusion or its first failed rule. (b) Eligible patients against response rate per cohort. (c) Highest cross-cohort Spearman correlation per array; the dashed line marks the locked duplicate threshold. (d) GEO-wide audit: identical arrays repeated between public breast series, with arc width proportional to the number of arrays.

2.3 Leakage checks

Two checks guarded against the same patient appearing in development and evaluation data. First, patient identifiers were compared between GSE25055 and GSE25065. Second, because identifiers differ across GEO series, every array of a later breast cohort was compared with every Tier 1 array by Spearman correlation over shared genes. This follows the correlation approach used to detect hidden duplicates in transcriptome databases [16]. Arrays above the locked threshold of $r = 0.97$ were treated as suspected duplicates and removed from the later cohort (Fig. 2c). A high correlation is a warning, not proof of identity, so the amendment checked each suspected pair against recorded age, ER status and pCR, compared counts at thresholds of 0.95, 0.97 and 0.99, and rescored the affected cohort with the suspected duplicates retained.

2.4 Endpoints

The breast cancer endpoint was pCR versus residual disease as defined by each source study (Tier 1 used no residual invasive cancer in breast and nodes [1, 29]); pCR is the accepted early endpoint of neoadjuvant trials and is associated with long-term outcome at the patient level [2]. For anti-PD-1 melanoma cohorts, complete or partial response by RECIST or immune-related RECIST was coded as response and stable or progressive disease as nonresponse [30]; samples with unknown response were excluded. For the MAGE-A3 cohort, the study-defined responder and nonresponder labels were used and not-evaluable samples were excluded [20]. No endpoint was derived from survival.

2.5 Expression processing and pathway scores

Processed expression values were taken from each series matrix. Probes were mapped to gene symbols with the GEO platform annotation; probes that map to several genes were dropped, and probes for one gene were collapsed with the median, an outcome-blind rule. Linear-scale data were transformed with $\log_2(x + 1)$. The two RNA-seq cohorts provide fragments per kilobase per million mapped reads (FPKM), transformed in the same way. GSE91061 labels genes by NCBI Entrez identifier; these were mapped to symbols with the Entrez-to-symbol columns of the GPL570 and GPL96 annotation files. Those columns are drawn from NCBI Gene, and using them keeps the RNA-seq symbols identical to the array symbols that define the gene universe; genes absent from both platforms are outside the universe and could not be scored in any cohort.

Pathway scores used the 50 MSigDB Hallmark gene sets (version 2024.1) [5], which represent

coherent biological states including proliferation, immune signaling, hormone response and stress programs [6, 31]. The gene universe was the 12,502 genes measured on GPL96, the oldest platform, so every cohort was scored over comparable genes. Sets with at least 10 measured genes and at least half of their genes measured were kept; all 50 sets passed. Each sample was scored with singscore [7], which ranks genes within one sample and returns the centered mean rank of the set's genes. The score uses only the sample itself, so it is unaffected by cohort composition and does not need fitting, unlike cohort-dependent methods such as GSVA [32]. A cohort-dependent alternative score, the mean of within-cohort gene z-scores, served as a sensitivity analysis. The scores measure coordinated transcript abundance and are not measurements of pathway activity.

Six clinical variables available before treatment were prespecified: age, estrogen receptor (ER) status, progesterone receptor (PR) status, histologic grade 3, clinical T stage and clinical nodal status. GSE41998 does not index grade or nodal status; the clinical models impute these from the development data, and its clinical comparisons are interpreted with that limitation. The melanoma cohorts index no comparable clinical covariates, so Tier 3 compares molecular models only.

2.6 Models

Five model families were fitted: a prevalence-only model (null); elastic-net logistic regression [33] on clinical variables (clinical); on the 50 Hallmark scores (pathway); on clinical variables plus Hallmark scores (combined); and on the 2,000 highest-variance genes of the training data (gene level). The gene-level model represents the high-dimensional alternative that pathway summaries are meant to replace. Within each training partition, missing clinical values were imputed with the median, features were standardized, and the 2,000 genes were selected by variance without labels, so no step used test data. Hyperparameters were chosen over $C \in \{0.001, 0.003, \ldots, 3\}$ and mixing parameter $\in \{0, 0.5, 1\}$ by stratified four-fold inner cross-validation on mean log loss, and the one-standard-error rule chose the most regularized candidate. Log loss was used for tuning because the primary endpoints are probabilistic. A constrained nonlinear model, ridge-penalized logistic regression on quadratic B-spline expansions of each clinical and pathway feature with main effects only, was prespecified as a sensitivity analysis and allowed only if the combined model beat the null model on both AUROC and Brier score in internal validation.

2.7 Validation

Internal validation. Repeated nested cross-validation on GSE25055 (5 outer folds, 5 repeats, tuning inside every outer training partition) estimated optimism [12].

External validation. The models fitted once on all of GSE25055 were frozen, with coefficients, scalers and hyperparameters fixed, and applied once to GSE25065, which shares the platform and processing of GSE25055. No step used any other GSE25065 patient to score a patient, and GSE25065 was not used for any choice.

Transportability. The frozen models were applied to the Tier 2 breast cohorts with two preprocessing modes. In the protocol mode (target adaptation), each molecular feature was standardized within the target cohort without labels and mapped onto the development scale before the frozen model was applied; this mode uses other target patients to score a patient. In the amendment mode (frozen preprocessing), features entered the frozen model unchanged, as they would for a single new patient. Clinical variables were not rescaled in either mode. Differences between the two modes reflect the adaptation step, not the pathway representation.

Melanoma context (exploratory). A model fitted on breast cancer treated with chemotherapy cannot be applied to melanoma treated with immunotherapy. Each melanoma cohort was therefore analyzed with repeated nested cross-validation inside the cohort (5 outer folds, 10 repeats, 3 inner folds), comparing pathway-level, gene-level and null models. These analyses ask whether the same feature representation carries response information in another disease and therapy class; they do not test transfer of a fitted model.

2.8 Metrics and statistical analysis

The co-primary metrics were average precision (AP), which summarizes the precision-recall curve and reflects performance on the minority responder class [34], and the Brier score, a proper scoring rule for predicted probabilities. Secondary metrics were the area under the receiver operating characteristic curve (AUROC), the scaled Brier score (one minus the ratio to the Brier score of the cohort prevalence), calibration intercept and calibration slope [35, 36], and prevalence. Uncertainty was estimated with 2,000 stratified patient-level bootstrap resamples, using the same resample for every model so that differences are paired; intervals are percentile 95% intervals. Cohort-level differences were combined with DerSimonian-Laird random-effects models [37], with heterogeneity reported as I^2. Frozen external evaluations (breast) and within-cohort refits (melanoma) estimate different quantities, so they were summarized separately; a

summary across all seven cohorts is shown only for orientation.

Sample size. The sample size was set by the public cohorts available; no formal calculation was made. The development cohort has 56 events for up to 56 candidate predictors in the combined model, too few for unpenalized regression, which is why all models used penalized regression with nested tuning. The external cohort has 41 events, so its intervals are wide, as reported.

Decision rules. H1 was declared supported only if the 95% interval for the paired AP difference (combined minus clinical) on GSE25065 lay above zero and the interval for the paired Brier difference lay below zero. Requiring both conditions is an intersection-union test, which does not need a split of the 5% error rate. H2 estimates the AUROC difference between pathway-level and gene-level models; no equivalence margin was prespecified, so the analysis reports the difference and its interval and makes no claim of equal or non-inferior performance. For H3, intercept-only recalibration was estimated on four fifths of a target cohort and scored on the remaining fifth (stratified five-fold cross-fitting, 20 repeats), so that no patient's outcome was used to recalibrate that patient's prediction; the protocol's in-sample version is also reported and is optimistic by construction. H4 describes context-associated pathway signals: within each cohort, response was regressed on each standardized pathway score, P values were adjusted with the Benjamini-Hochberg procedure [38], and log-odds ratios were pooled within context by random effects. In the amendment, breast associations were adjusted for ER, PR and grade 3, and anti-PD-1 and MAGE-A3 melanoma cohorts were pooled separately. Cancer type and therapy class vary together in this design, so differences between contexts cannot be attributed to the therapy. Secondary analyses are descriptive; Holm adjustment was prespecified for the secondary family.

2.9 Robustness, negative controls and stability

Five analyses probed the primary comparison. (1) Size-matched random gene sets: 100 draws replaced the Hallmark sets inside the combined model, refitted and evaluated as above. (2) Expression- and variance-matched random gene sets (amendment): each Hallmark gene was replaced by a non-member gene from the same development-cohort bin of mean expression and standard deviation (10 by 10 quantile bins), so that random sets match the abundance and variability of curated genes; random signatures can match curated ones in breast cancer because proliferation genes dominate expression variance [14, 15]. (3) Parsimonious comparators

Table 1: Hypotheses, findings and locked verdicts. H1 is confirmatory; H2 to H4 are secondary; H5 was locked in amendment 2 and tested on unseen cohorts.

Hypothesis	Locked test and rule	Result (95% CI)	Main finding	Locked verdict
H1 (primary)	Clinical + pathway vs clinical on GSE25065; AP CI > 0 and Brier CI < 0	AP 0.057 (-0.095, 0.178); Brier -0.004 (-0.016, 0.008)	Pathway scores predict pCR in an untouched cohort (AUROC 0.706; AP 0.411 vs prevalence 0.23)	Increment over clinical not shown
H2 (estimation)	Pathway vs gene level, AUROC difference; estimate only	Breast -0.061 (-0.097, -0.026); melanoma -0.003	Gene-level detail adds discrimination in breast; no difference in melanoma	Estimated
H3 (descriptive)	Cross-fitted intercept recalibration in four breast cohorts	Lower Brier in 11 of 16 model-cohort pairs	Local recalibration improves most frozen models	Mostly met
H4 (descriptive)	Adjusted pathway log-odds by context	Breast 8, MAGE-A3 14, anti-PD-1 0; shared 1	Proliferation predicts breast pCR; immune programs predict MAGE-A3 response	Context-specific
H5 (amendment 2)	Clinical + proliferation vs clinical in five unseen cohorts; pooled AP CI > 0 and recalibrated Brier CI < 0	AP 0.042 (0.011, 0.073); Brier -0.0014 (-0.0044, 0.0016)	Three proliferation programs improve ranking over clinical variables in 1,121 unseen patients	AP condition met; Brier condition not met

(amendment): clinical variables plus three proliferation sets (E2F targets, G2M checkpoint, MYC targets V1), and clinical variables plus three immune sets (interferon gamma response, allograft rejection, inflammatory response), fitted with the same tuning. (4) Label permutation: 50 refits on permuted development labels. (5) The alternative pathway score, balanced class weights, and ER-positive and ER-negative subgroups of GSE25065. Pathway stability was measured with 200 bootstrap refits of the combined model on GSE25055 at the selected hyperparameters, recording each feature's selection frequency and sign consistency; features selected in at least 70% of refits with at least 90% sign agreement were called stable.

2.10 Confirmatory test in unseen cohorts (amendment 2)

The amendment 1 analyses suggested that three proliferation sets carried the breast signal. Because that observation came from data the study had already used, a second amendment locked it as a new hypothesis before new data were downloaded. H5 states that the clinical + proliferation model, fitted once on GSE25055 with the amendment 1 tuning and then frozen, predicts pCR better than the frozen clinical model in HER2-negative breast cancer cohorts unseen by this study. Seven candidate GEO series were named in advance: GSE16446, GSE20271 [39], GSE22093 and GSE23988 [40], GSE42822 [41], GSE22226 (I-SPY 1) [42] and GSE194040 (I-SPY 2) [43]. A cohort was eligible if it had pretreatment expression with gene-symbol annotation, a binary pCR record, HER2-negative status and at least 10 pCR events after exclusions. HER2 status came from the recorded test, or from the series design for GSE22093 and GSE23988, whose source study analyzed HER2-negative tumors only [40]. Arrays with Spearman $r > 0.97$

to any array already in the study, or to an array in an earlier new series, were removed. Features entered the frozen models with development-only preprocessing, and clinical variables missing from a cohort were imputed with development medians. H5 was declared supported if the random-effects pooled AP difference had a 95% interval above zero and the pooled difference in Brier score after cross-fitted intercept recalibration had an interval below zero. Intercept recalibration does not change the ranking of patients; it was applied because H3 showed that frozen intercepts do not transport. Per-cohort differences used 2,000 paired stratified bootstraps. I-SPY 1 used two-color arrays with log ratios against a common reference, and I-SPY 2 released only hormone-receptor status, which entered the clinical model as ER status.

2.11 GEO-wide duplicate audit

To measure how far repeat deposits extend beyond the study cohorts, a third protocol, locked and hashed before any additional series was downloaded, defined an audit of public breast cancer treatment-response series. A GEO DataSets search (breast; neoadjuvant, preoperative, complete response or chemotherapy; *Homo sapiens*; expression profiling by array; Affymetrix HG-U133 family or X3P platforms), followed by a keyword filter on title and summary, at least 20 samples and tumor tissue only, gave 47 series after two superseries and three non-tumor series were removed (Additional file 3). Arrays on platforms that share HG-U133A probe identifiers were compared on their 22,215 shared probes, and X3P arrays were compared among themselves. For each array, the highest Spearman correlation with any array of another series, and with any other array of its own series, was recorded. Pairs above the locked threshold of 0.97 were then classified from the GEO records: identical arrays ($r \geq 0.999$); declared companion series, linked by a SuperSeries or SubSeries relation; and undeclared pairs, split by whether the submitting contact was the same. Age, ER status and pCR were compared wherever both records indexed them, and pairs that exceeded 0.97 without identical arrays or metadata agreement were not counted. Within a pair, the earlier-deposited series was taken as the original. Europe PMC full-text search counted articles that name both series of a confirmed pair; only open-access full texts are indexed, so this count is a lower bound.

2.12 Computation

All analyses ran on a laptop CPU (Apple A18 Pro, 6 cores, 8 GB memory, recorded by the operating system at run time) with Python 3.9.6, NumPy 2.0.2, pandas 2.3.3 and scikit-learn 1.6.1

(recorded with the results); no GPU was used. Parallel work was limited to four processes during tuning, and all random seeds were fixed. The pipeline records wall time and peak resident memory for every stage (Table 4). Two commands download the inputs, verify their SHA-256 checksums and rebuild every result, table, figure and number in this article.

3 Results

3.1 Cohort audit and patient deduplication

Of 1,462 source samples in nine candidate series, 959 passed the sample-level audit, giving eight cohorts and 958 patients with one row per patient (Table 2, Fig. 2). The Tier 1 count reproduced the source study: of 508 enrolled patients, 20 lack a response record and 6 are HER2-positive, one patient meets both conditions, and the remaining 483 form the development (303) and external validation (180) cohorts. No patient identifier appeared in both Tier 1 cohorts, and no GSE25065 array exceeded the duplicate threshold against GSE25055 (highest $r = 0.95$).

The duplicate scan flagged 185 of 219 HER2-negative GSE20194 arrays as copies of Tier 1 arrays. All 185 flagged arrays had a Pearson correlation of at least 0.993 with their match, and the count did not change between thresholds of 0.97 and 0.99 (185 at both). Recorded metadata agreed within the flagged pairs far more often than within the best-matching unflagged pairs: pCR agreed in 100% versus 88%, ER status in 99% versus 76%, and age within one year in 97% versus 38%. The flagged arrays are therefore treated as repeat deposits of Tier 1 patients. Retaining them would have inflated the frozen gene-level AUROC on GSE20194 from 0.850 to 0.939 and the clinical AUROC from 0.567 to 0.820, because these patients had trained the models. After deduplication GSE20194 kept 34 patients with 4 responders, too few events for a stable estimate, and its intervals are wide. The same scan in amendment 2 found that 52 of 61 eligible arrays in GSE23988 and 7 of 97 in GSE22093 repeat patients deposited elsewhere; 51 of the GSE23988 copies repeat patients of GSE25065, the external validation cohort. Within the study's own cohorts and the amendment 2 candidates, 244 arrays were repeat deposits of patients in another series; every matched pair had a Spearman correlation of at least 0.999 and the same recorded pCR, and the full array-to-array map is given in Additional file 2. The GSE20194 overlap agrees with a recent report that linked 188 of its 248 patients to GSE25055 through patient identifiers [44]; we found no earlier report of the GSE23988 and GSE22093 overlaps, neither of which is flagged in the GEO records.

3.2 Repeat deposits extend across public breast series

The GEO-wide audit read 46 of 47 series (one had no expression table) with 4,481 expression arrays. 401 arrays in 7 series were identical ($r \geq 0.999$) to arrays deposited earlier in another series, across 11 series pairs, and no GEO record declared these relations (Fig. 2d; Additional file 2). Wherever both records indexed them, pCR, ER status and age agreed within the matched pairs. The largest pairs were GSE25055 repeating 186 GSE20194 arrays (185 of them among the eligible HER2-negative arrays above), GSE20181 repeating 116 arrays of GSE5462 from the same submitter, and GSE25065 repeating 53 GSE23988 arrays. Outside the MD Anderson series, GSE71258 repeated 24 arrays of GSE19697, and GSE32072 repeated 13 arrays from three earlier series. A further 88 identical arrays belonged to a declared companion pair and were not counted, and 5 series pairs that exceeded 0.97 without identical arrays or metadata agreement were not counted either. No series contained identical arrays within itself. At least 102 open-access articles published from 2010 to 2026 name both series of a confirmed pair. Naming both series does not show that an article combined them for training and validation, but each of these analyses was exposed to the problem.

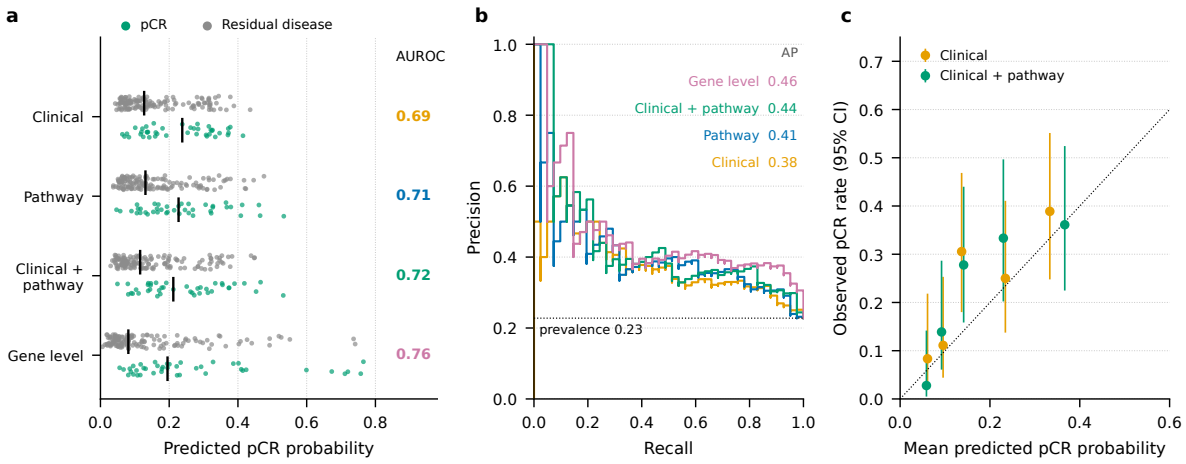
3.3 Pathway scores add no measurable information to clinical variables (H1)

On the untouched GSE25065 cohort, every frozen model separated responders from nonresponders better than prevalence alone (Fig. 3, Table 3). The pathway model reached an AUROC of 0.706 (95% CI 0.619 to 0.791) and an AP of 0.411 against a prevalence of 0.23, and it lowered the Brier score relative to prevalence by −0.015 (−0.028 to −0.002). Pretreatment pathway scores therefore carry response information.

They did not add measurable information to clinical variables. The combined model changed AP by 0.057 (−0.095 to 0.178) and the Brier score by −0.004 (−0.016 to 0.008) relative to the clinical model (Fig. 4a). Both intervals include zero, so H1 is not supported under the locked rule. The AP interval is wide: it excludes losses larger than 0.10 but cannot exclude a gain as large as 0.18, so the data neither establish nor rule out a clinically useful increment. Internal cross-validation gave the same ordering (AP 0.379 clinical, 0.414 combined; Fig. 4b), and the external values fell within the spread of the internal repeats. Across the four breast cohorts, the random-effects AP difference between the combined and clinical models was −0.002 (−0.133 to 0.130).

Table 2: Cohorts after the sample-level audit. Patients are those eligible after all exclusions, with responders in parentheses. Tier 4 cohorts were added in amendment 2 to test H5.

Cohort	Tier	Cancer	Indexed therapy	Endpoint	Platform	Samples	Patients	Prev.	Exclusions (first failed rule)
GSE25055	1	Breast (HER2-negative)	Taxane then anthracycline (T/FAC or T/FEC), neoadjuvant	pCR vs residual disease	GPL96	310	303 (56)	0.18	HER2-positive (IHC or FISH): 4; Response not recorded: 3
GSE25065	1	Breast (HER2-negative)	Taxane then anthracycline (T/FAC or T/FEC), neoadjuvant	pCR vs residual disease	GPL96	198	180 (41)	0.23	Response not recorded: 16; HER2-positive (IHC or FISH): 2
GSE20194	2	Breast (HER2-negative)	Paclitaxel then FAC, neoadjuvant	pCR vs residual disease	GPL96	278	34 (4)	0.12	Duplicate of Tier 1 array: 185; HER2-positive (IHC or FISH): 59
GSE32646	2	Breast (HER2-negative)	Paclitaxel then FEC, neoadjuvant	pCR vs non-pCR	GPL570	115	81 (15)	0.18	HER2-positive (IHC or FISH): 34
GSE41998	2	Breast (HER2-negative)	AC then ixabepilone or paclitaxel, neoadjuvant	pCR vs no pCR	GPL571	279	229 (63)	0.28	HER2-positive (IHC or FISH): 27; Second-phase therapy not indexed: 14; Response not recorded: 9
GSE35640	3	Melanoma (metastatic)	MAGE-A3 antigen-specific immunotherapy	Clinical responder vs non-responder	GPL570	65	56 (22)	0.39	Response not evaluable: 9
GSE91061	3	Melanoma (advanced)	Nivolumab (anti-PD-1)	RECIST CR/PR vs SD/PD	RNA-seq (GPL9052)	109	49 (10)	0.20	On-treatment biopsy: 58; Response unknown: 2
GSE78220	3	Melanoma (metastatic)	Pembrolizumab or nivolumab (anti-PD-1)	irRECIST CR/PR vs PD	RNA-seq (GPL11154)	28	26 (14)	0.54	On-treatment biopsy: 1
GSE5851	3	Colorectal (metastatic)	Cetuximab monotherapy	Not indexed	GPL571	80	Excluded	–	Response not indexed in repository: 80
GSE16446	4	Breast (HER2-negative)	Epirubicin monotherapy (ER-negative)	pCR vs residual disease	GPL570	120	Excluded (6 events)	–	HER2 not negative by FISH: 58; Response not recorded: 3
GSE20271	4	Breast (HER2-negative)	FAC or paclitaxel then FAC	pCR vs residual disease	GPL96	178	152 (19)	0.12	HER2 not negative: 26
GSE22093	4	Breast (HER2-negative)	Anthracycline-based	pCR vs residual disease	GPL96	103	90 (27)	0.30	Duplicate of earlier array: 7; Response not recorded: 6
GSE22226	4	Breast (HER2-negative)	AC then taxane (I-SPY 1)	pCR vs residual disease	Agilent two-color	150	84 (13)	0.15	HER2 not negative: 62; Response not recorded: 4
GSE23988	4	Breast (HER2-negative)	FAC then docetaxel-capecitabine	pCR vs residual disease	GPL96	61	Excluded (2 events)	–	Duplicate of earlier array: 52
GSE42822	4	Breast (HER2-negative)	FEC then docetaxel (TX)	pCR vs residual disease	GPL96	91	54 (20)	0.37	HER2 not negative: 37
GSE194040	4	Breast (HER2-negative)	Paclitaxel (+/- investigational agent) then AC (I-SPY 2)	pCR vs residual disease	Agilent	988	741 (206)	0.28	HER2 not negative: 245; Pretreatment expression missing: 2



AP, average precision; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; pCR, pathologic complete response. Frozen models on GSE25065. (a) Each dot is one patient; bars, medians. (c) Quintiles of predicted probability; points above the diagonal mean underprediction.

Fig. 3: Frozen models on the untouched GSE25065 cohort. (a) Predicted pCR probability for each patient by observed outcome, with AUROC. (b) Precision-recall curves with average precision. (c) Calibration by quintile of predicted probability with 95% confidence intervals.

Table 3: Primary external validation on GSE25065. Internal values are means over five repeats of nested cross-validation on GSE25055; external values have 95% bootstrap CIs.

Model or paired difference	Internal CV		External, GSE25065				
	AUROC	AP	AUROC [95% CI]	AP [95% CI]	Brier [95% CI]	Cal. int.	Cal. slope
Prevalence only	0.490	0.182	0.500 [0.50, 0.50]	0.228 [0.23, 0.23]	0.178 [0.178, 0.178]	0.26	0.88
Clinical	0.773	0.379	0.688 [0.60, 0.77]	0.380 [0.29, 0.53]	0.166 [0.155, 0.179]	0.35	0.92
Pathway (Hallmark)	0.770	0.397	0.706 [0.62, 0.79]	0.411 [0.32, 0.55]	0.163 [0.150, 0.176]	0.34	0.98
Clinical + pathway	0.780	0.414	0.718 [0.64, 0.80]	0.437 [0.33, 0.57]	0.162 [0.148, 0.176]	0.38	0.98
Gene level	0.782	0.463	0.764 [0.69, 0.84]	0.460 [0.36, 0.60]	0.160 [0.138, 0.180]	0.46	0.79
Additive spline (sensitivity)	–	–	0.713 [0.63, 0.79]	0.431 [0.33, 0.57]	0.163 [0.147, 0.178]	0.40	0.85
Clinical + pathway − clinical	–	–	0.031 [−0.05, 0.10]	0.057 [−0.09, 0.18]	−0.004 [−0.016, 0.008]	–	–
Pathway − clinical	–	–	0.019 [−0.09, 0.11]	0.031 [−0.13, 0.18]	−0.003 [−0.018, 0.011]	–	–
Pathway − gene level	–	–	−0.058 [−0.12, −0.00]	−0.049 [−0.14, 0.05]	0.003 [−0.011, 0.017]	–	–
Clinical − prevalence	–	–	0.188 [0.10, 0.27]	0.152 [0.07, 0.30]	−0.011 [−0.023, 0.001]	–	–

3.4 The breast signal is proliferation, and curated sets are not more specific than random sets

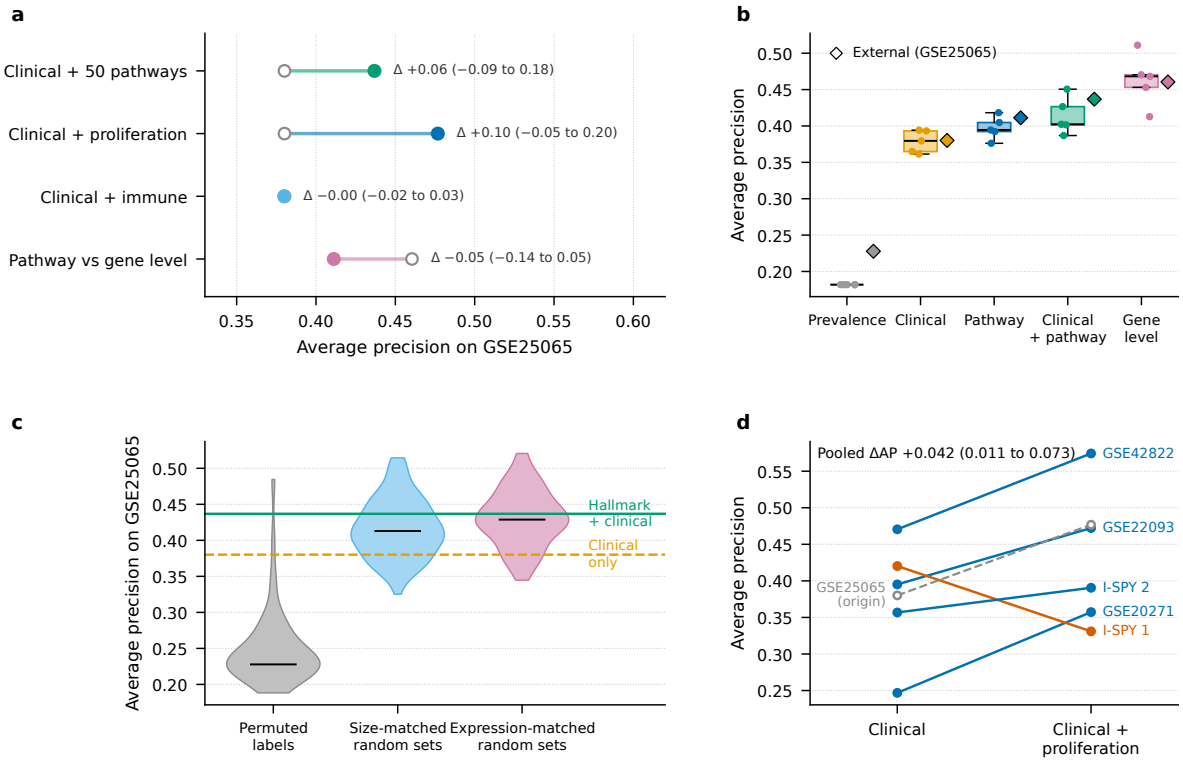
The response information in the pathway scores was not specific to the curated Hallmark sets (Fig. 4c). Combined models built on 100 size-matched random gene sets reached a median AP of 0.413 on GSE25065 ($P = 0.31$ for an AP at least as high as the Hallmark model), and models built on 100 random sets matched to the Hallmark genes in mean expression and variability reached a median AP of 0.429 ($P = 0.42$). Permuting the development labels removed the signal (median AP 0.228, $P = 0.04$), so the models learned real response information, but that information is carried by expression variation in general rather than by the curated membership.

Three proliferation sets explained the breast signal. A clinical model with the E2F target, G2M checkpoint and MYC target scores added 0.097 AP to the clinical model (−0.047 to 0.205) and reached an AUROC of 0.743, and the full 50-set model performed no better than this three-set model (AP difference −0.040, −0.102 to 0.023). Three immune sets added nothing (AP difference −0.000, −0.025 to 0.028). These comparators were added in the amendment and are exploratory.

3.5 The proliferation gain in ranking replicates in five unseen cohorts (H5)

Amendment 2 tested the clinical + proliferation model, frozen on GSE25055, against the frozen clinical model in breast cohorts the study had not used (Table 2, Tier 4). Of seven prespecified candidates, five met the eligibility rules, with 1,121 HER2-negative patients and 285 pCR events: GSE20271 (152 patients) [39], GSE22093 (90) [40], GSE22226 (I-SPY 1, 84) [42], GSE42822 (54) [41] and GSE194040 (I-SPY 2, 741) [43]. GSE16446 had 6 pCR events among HER2-negative patients, and GSE23988 kept 9 patients after deduplication, so the locked rules excluded both.

Adding the three proliferation scores raised AP in 4 of 5 cohorts, with differences from −0.089 in I-SPY 1 to 0.110 in GSE20271. The pooled AP difference was 0.042 (Fig. 4d; 95% CI 0.011 to 0.073; $I^2 = 0\%$), and the pooled AUROC difference was 0.035 (−0.002 to 0.072). The Brier score after cross-fitted intercept recalibration did not improve (difference −0.0014, −0.0044 to 0.0016). The locked rule required both conditions, so H5 is not supported as stated. The AP gain held without I-SPY 2, whose only released clinical variable is hormone-receptor status (0.074, 0.004 to 0.144), and in the three single-channel cohorts with full clinical covariates (0.092, 0.018 to 0.166); these two restrictions were not prespecified. Proliferation scores therefore improved the



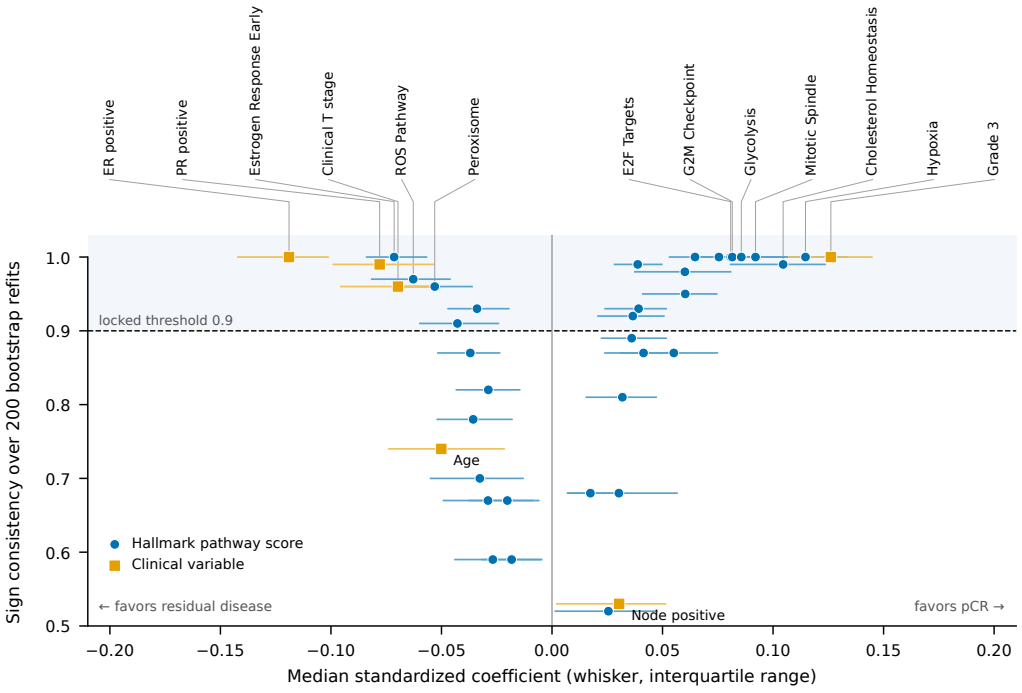
AP, average precision; Δ, difference with 95% bootstrap CI. (a) Open circle, reference model; filled circle, compared model. (b) Boxes summarize five repeats of nested cross-validation on GSE25055. (c) 50 permutations and 100 random-set draws. (d) Five breast cohorts unseen by the study (amendment 2); blue lines rise, red lines fall; the pooled estimate is DerSimonian-Laird random effects.

Fig. 4: Incremental value and specificity on GSE25065. (a) Average precision of the reference and compared models with the paired difference. (b) Average precision over five repeats of internal nested cross-validation (boxes) and on GSE25065 (diamonds). (c) The Hallmark + clinical model against distributions of permuted-label and random gene-set controls. (d) Amendment 2 (H5): average precision of the frozen clinical and clinical + proliferation models in five unseen cohorts and the origin cohort, with the pooled random-effects gain.

ranking of patients by pCR probability beyond clinical variables in unseen cohorts, but the gain was too small to lower overall prediction error.

3.6 Coefficients are stable and follow known breast cancer biology

Bootstrap refits of the combined model kept the sign of 22 features in at least 90% of 200 refits (Fig. 5). The penalty selected by nested tuning was ridge (mixing parameter 0), so every feature stayed in every refit and stability is reported as sign consistency and coefficient spread. Higher grade and negative ER and PR status were associated with pCR, as were higher proliferation (G2M checkpoint, E2F targets, mitotic spindle, MYC targets), hypoxia and glycolysis scores; higher early estrogen response scores were associated with residual disease. These coefficients describe associations within a correlated set of features, and the magnitudes are small (standardized coefficients below 0.15 under strong shrinkage).



ER, estrogen receptor; PR, progesterone receptor; pCR, pathologic complete response. All 56 features of the combined model fitted on GSE25055; the selected penalty was ridge, so every feature is kept in every refit. Shaded band, features above the locked sign-consistency threshold.

Fig. 5: Stability of the combined model over 200 bootstrap refits on GSE25055. Each point is one of 56 features: median coefficient against the share of refits with the same sign; whiskers show the interquartile range.

3.7 Gene-level models discriminate better than pathway models in breast cancer (H2)

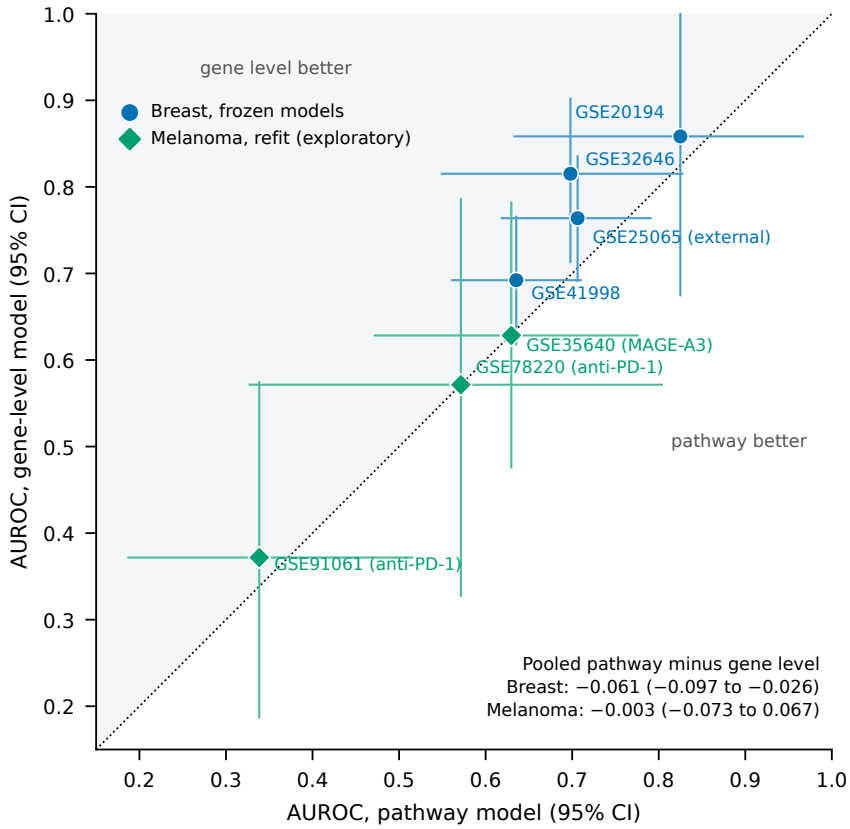
The frozen gene-level model had a higher AUROC than the pathway model on GSE25065 (difference −0.058, −0.120 to −0.001) and in every Tier 2 cohort (Fig. 6; Table 4). The random-effects difference over the four breast cohorts was −0.061 (−0.097 to −0.026), which favors the gene-level model. In the melanoma cohorts the two representations did not differ (−0.003, −0.073 to 0.067). No equivalence margin was prespecified, so these estimates are reported without a non-inferiority claim. The pathway model used 50 features and the gene-level model 2,000, and both were tuned and fitted in under one minute on the same CPU.

3.8 Frozen models need cohort-specific calibration (H3)

The frozen models underestimated pCR on GSE25065, with calibration intercepts of 0.35 to 0.46 and slopes of 0.79 to 0.98 (Table 3, Fig. 3c); the development cohort had a lower pCR rate (0.18) than GSE25065 (0.23). Intercepts in the Tier 2 cohorts ranged from −0.65 to 0.56, with signs that differed between cohorts. Intercept recalibration that never used a patient's own outcome lowered the Brier score in 11 of 16 model-cohort pairs, including all four models on GSE25065 (combined model 0.162 to 0.159; Table 1). The gains were small, and in the remaining pairs recalibration raised the Brier score slightly. Standardizing molecular features within each target cohort, the protocol's transport mode, changed AUROC by at most 0.06 relative to frozen development-only preprocessing (Table 4), so the transport results do not depend on that step.

3.9 Pathway associations are context-associated (H4)

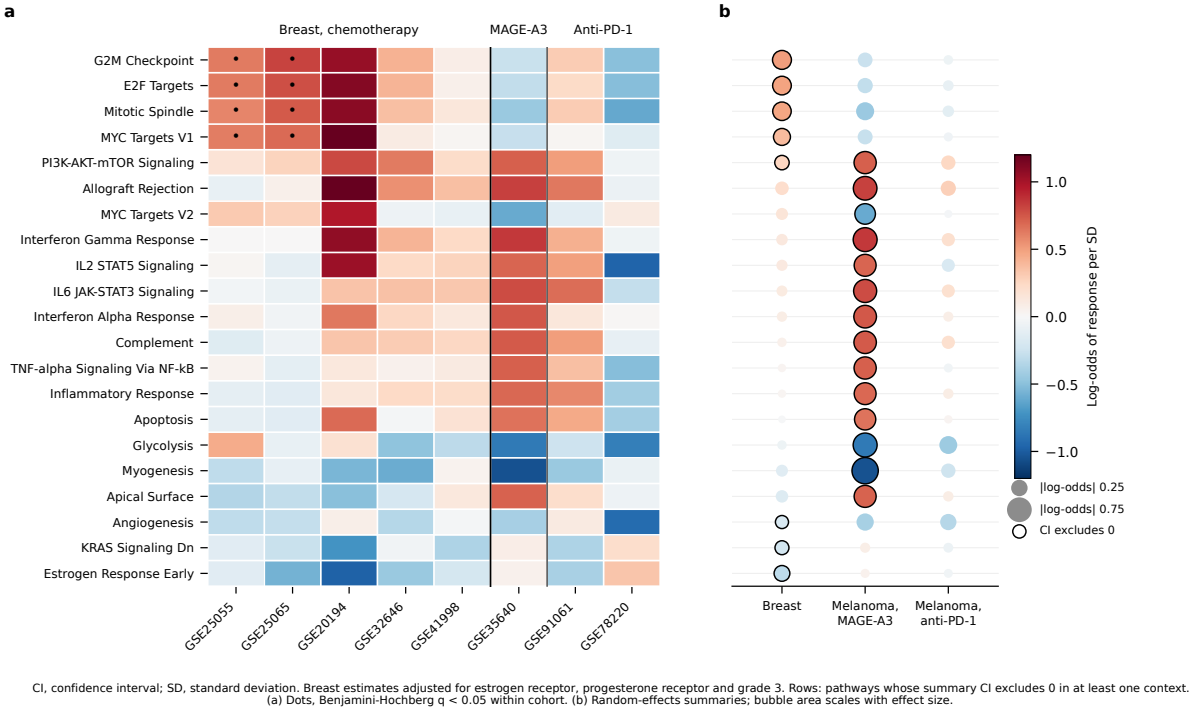
After adjustment for ER, PR and grade 3, 8 of 50 pathway scores were associated with pCR across the five breast cohorts (Fig. 7). Proliferation scores carried the strongest positive associations (G2M checkpoint 0.50 log-odds per standard deviation, 95% CI 0.16 to 0.83; E2F targets 0.48), and early estrogen response the strongest negative association (−0.32, −0.59 to −0.05). In the MAGE-A3 melanoma cohort, 14 scores were associated with clinical response, led by interferon gamma response (0.85), allograft rejection (0.82) and IL6-JAK-STAT3 signaling (0.78). No score reached significance in the pooled anti-PD-1 cohorts (49 and 26 patients). Only 1 score, PI3K-AKT-mTOR Signaling, was associated with response in the same direction in breast cancer and in the MAGE-A3 cohort. Proliferation, which marked chemotherapy response, showed no association with immunotherapy response, and the immune programs that



AUROC, area under the receiver operating characteristic curve; CI, confidence interval. Points above the diagonal favor the gene-level model; whiskers are 95% bootstrap CIs; pooled values are DerSimonian-Laird estimates.

Fig. 6: Pathway versus gene-level discrimination across cohorts. AUROC of the pathway model against the gene-level model per cohort with 95% bootstrap intervals; points above the diagonal favor the gene-level model.

marked MAGE-A3 response showed none with chemotherapy response after clinical adjustment. Cancer type and therapy class change together in this design, so these differences are associated with the treatment context and cannot be attributed to the therapy itself.



CI, confidence interval; SD, standard deviation. Breast estimates adjusted for estrogen receptor, progesterone receptor and grade 3. Rows: pathways whose summary CI excludes 0 in at least one context. (a) Dots, Benjamini-Hochberg q < 0.05 within cohort. (b) Random-effects summaries; bubble area scales with effect size.

Fig. 7: Context-associated pathway signals. (a) Per-cohort log-odds of response per standard deviation of each score. (b) Random-effects summaries by context; bubble area scales with effect size and outlined bubbles have intervals that exclude zero; breast estimates adjusted for ER, PR and grade 3.

3.10 Exploratory melanoma refits

Within-cohort nested cross-validation gave a pathway-model AUROC of 0.630 (0.472 to 0.775) in the MAGE-A3 cohort and 0.571 (0.327 to 0.804) in GSE78220. In GSE91061, with 10 responders among 49 patients, the AUROC was 0.338 (0.187 to 0.515), below chance, which indicates that the refit procedure overfits in a cohort this small. Gene-level refits behaved the same way. These cohorts show that the representation can carry response information in a second disease and therapy class, and they do not test transfer of a fitted model.

3.11 Robustness and computation

The results did not depend on the scoring method, class weighting or the constrained nonlinear model (Table 4). The mean z-score alternative gave a combined AP of 0.475, balanced class weights gave 0.425, and the additive spline model, run because the combined model beat the null

model in internal validation, gave an AUROC of 0.713 and an AP of 0.431. In the ER-positive subgroup of GSE25065 (19 events), the combined model reached an AUROC of 0.749 against 0.681 for the clinical model; in the ER-negative subgroup (22 events) the values were 0.631 and 0.605. These subgroups are small and are reported descriptively. The full confirmatory pipeline ran in 30.9 minutes with a peak memory of 843 MB on a laptop CPU, without a GPU.

Table 4: Transportability, negative controls, sensitivity analyses and resources. Amendment analyses are exploratory.

Analysis	Result	Comparison
Transport, GSE20194	AUROC: clinical 0.567, combined 0.800, pathway 0.825, gene 0.858	AP difference (combined − clinical) 0.320 (0.024, 0.645)
Transport, GSE32646	AUROC: clinical 0.745, combined 0.723, pathway 0.698, gene 0.815	AP difference (combined − clinical) −0.155 (−0.332, 0.022)
Transport, GSE41998	AUROC: clinical 0.746, combined 0.655, pathway 0.635, gene 0.692	AP difference (combined − clinical) −0.054 (−0.152, 0.047)
Cross-cancer CV, GSE35640	AUROC: pathway 0.630 (0.472, 0.775), gene 0.628 (0.476, 0.782)	AP: pathway 0.535, prevalence 0.355
Cross-cancer CV, GSE91061	AUROC: pathway 0.338 (0.187, 0.515), gene 0.372 (0.187, 0.574)	AP: pathway 0.163, prevalence 0.212
Cross-cancer CV, GSE78220	AUROC: pathway 0.571 (0.327, 0.804), gene 0.571 (0.327, 0.786)	AP: pathway 0.657, prevalence 0.421
Negative control: random gene sets	Median AP 0.413, AUROC 0.721 (100 draws)	$P(\text{AP} \geq \text{observed}) = 0.31$
Negative control: permuted labels	Median AP 0.228, AUROC 0.500 (50 draws)	$P(\text{AP} \geq \text{observed}) = 0.04$
Sensitivity: mean z-score pathway scorer (events = 41)	AUROC: pathway 0.714, clinical + pathway 0.723	AP: pathway 0.437, clinical + pathway 0.475
Sensitivity: balanced class weights (events = 41)	AUROC: clinical 0.687, pathway 0.695, clinical + pathway 0.717, gene level 0.739	AP: clinical 0.373, pathway 0.424, clinical + pathway 0.425, gene level 0.425
Sensitivity: er-positive subgroup (events = 19)	AUROC: clinical 0.681, pathway 0.743, clinical + pathway 0.749	AP: clinical 0.305, pathway 0.329, clinical + pathway 0.324
Sensitivity: er-negative subgroup (events = 22)	AUROC: clinical 0.605, pathway 0.612, clinical + pathway 0.631	AP: clinical 0.440, pathway 0.520, clinical + pathway 0.531
Negative control: expression-matched random sets	Median AP 0.429, AUROC 0.732 (100 draws)	$P(\text{AP} \geq \text{observed}) = 0.42$
Comparator: clinical + proliferation sets	AUROC 0.743 (0.659, 0.820); AP 0.477 (0.365, 0.606)	AP difference vs clinical 0.097 (−0.047, 0.205)
Comparator: clinical + immune sets	AUROC 0.687 (0.597, 0.771); AP 0.380 (0.296, 0.525)	AP difference vs clinical −0.000 (−0.025, 0.028)
Leakage: GSE20194 with suspected duplicates retained	AUROC gene 0.939 vs 0.850 deduplicated	Clinical 0.820 vs 0.567
Preprocessing: frozen vs target adaptation	Clinical + pathway AUROC, frozen vs adapted: GSE20194 0.742 vs 0.800; GSE32646 0.740 vs 0.723; GSE41998 0.660 vs 0.655	Largest absolute AUROC change over molecular models 0.058
Resources (full pipeline)	Wall time 30.9 min on a laptop CPU	Peak memory 843 MB, no GPU

4 Discussion

This study identifies proliferation as the transportable pretreatment signal of chemotherapy response in HER2-negative breast cancer. Pathway scores predicted pCR in an untouched external cohort, the signal came from E2F targets, G2M checkpoint, mitotic spindle and MYC targets, these programs pointed toward pCR in all five breast cohorts after adjustment for ER,

PR and grade, and three proliferation sets captured what the full 50-set library offered. The same representation found a different signal in antigen-specific melanoma immunotherapy, where immune programs carried the response information. The locked tests also set clear bounds on these results: the full library added no measurable information to clinical variables, random gene sets matched in expression and variability performed as well, a gene-level model discriminated better and frozen models needed local calibration. Proliferation, not breadth of curated biology, carries the transportable signal. The strongest evidence comes from amendment 2, in which the three-set model was locked before new data were downloaded and then raised AP over clinical variables in five unseen cohorts with 1,121 patients, with no heterogeneity across cohorts. That gain concerns ranking: it did not lower recalibrated Brier score, so the model identifies which patients are more likely to reach pCR without making the predicted probabilities more accurate overall.

These findings agree with the meta-analysis of Stover et al. [8], which found that most published pCR signatures tracked proliferation. The present study adds three elements to that result. First, it tests incremental value with a locked rule on probability-based metrics in an untouched cohort, rather than discrimination alone. Second, it compares curated sets with random sets matched to the curated genes in expression level and variability, which removes the explanation that random sets succeed only because they contain abundant or variable genes [14]; the matched sets still performed as well as the Hallmark sets. Third, it maps repeat-deposited patients across public breast cohorts at the array level. Hidden duplicates in transcriptome databases are a known hazard [16], and the overlap between GSE20194 and GSE25055 was recently reported from patient identifiers [44]; here, retaining those duplicates would inflate validation AUROC by up to 0.25. The amendment 2 scan found the same problem in two further series not previously reported, including 51 copies of external-validation patients in GSE23988, bringing the total within the study's own cohorts to 244 repeat-deposited arrays. The GEO-wide audit shows that the problem is not confined to these cohorts: 401 identical arrays recur across 11 undeclared series pairs, including pairs outside the MD Anderson series, and at least 102 articles name both members of such a pair. The array-to-array map (Additional file 2) lets other analysts remove these duplicates directly. Evaluations that combine these series without an array-level check can therefore report optimistic transportability.

The breast results identify a narrow mechanistic reading. Proliferation scores (G2M checkpoint, E2F targets, mitotic spindle and MYC targets) remained associated with pCR after ad-

justment for ER, PR and grade, and early estrogen response remained associated with residual disease. Taxanes and anthracyclines act on dividing cells, and highly proliferative, hormone-independent tumors reach pCR more often [1, 45]. A three-set proliferation model performed as well as the 50-set model, so in this setting the curated library adds no information beyond that axis. In the MAGE-A3 melanoma cohort, interferon gamma, allograft rejection and inflammatory programs were associated with response, consistent with the immune signature reported by the source study [20] and with interferon-related profiles in checkpoint blockade [3]. Immune programs showed no association with chemotherapy response in breast cancer after clinical adjustment, and proliferation showed none with immunotherapy response. The two disease contexts therefore drew on different programs. Cancer type and therapy class change together here, so the design cannot separate disease biology from treatment mechanism, and the associations are not evidence that a program modifies the effect of a therapy.

Pathway prediction of response has been reported before in both cancers [9–11]. The anti-PD-1 cohorts in this study were too small for refitted models to perform above chance in one of them, and they gave no pathway association after correction. The positive pathway-signature results in anti-PD-1 melanoma came mainly from on-treatment biopsies [10], whereas this study used pretreatment biopsies only; the null result here is consistent with that difference and with limited power, and it does not contradict the earlier work.

Three practical points follow. First, the frozen models underestimated pCR when the external cohort had a higher response rate than the development cohort, and calibration intercepts differed in sign across cohorts, so any deployment of such models needs local recalibration on patients other than those being scored. Second, the choice between development-only and target-cohort feature standardization changed discrimination by at most 0.06, so the transport results do not hinge on that step. Third, the full pipeline, including 2,000-resample bootstraps and 150 negative-control refits, ran in about half an hour on a laptop CPU with less than 1 GB of memory, so the analysis can be repeated and extended to new cohorts without specialized hardware.

4.1 Limitations

The study has eight limitations. (1) The external AP interval for H1 is wide and includes a gain of 0.18; the data cannot rule out a clinically useful increment, and a larger cohort would be needed to bound it. (2) The protocol was frozen before any model was fitted but after cohort

metadata had been inspected, and it was not registered with a third party. (3) The amendment 1 analyses were added after the confirmatory results were known and are exploratory; H5 was generated from them and locked before new data were seen, but it met only one of its two locked conditions, and the H5 restrictions without I-SPY 2 and to full-covariate cohorts were not prespecified. I-SPY 1 used two-color ratio arrays and I-SPY 2 released only hormone-receptor status, which weakens the clinical comparator in that cohort. (4) After deduplication GSE20194 kept 4 responders, and GSE41998 lacks grade and nodal status, so these transport estimates are imprecise. (5) The melanoma cohorts are small, mix two immunotherapy classes and lack comparable clinical covariates; they test the representation, not a transferred model. (6) Duplicate identity rests on expression and metadata concordance rather than on source identifiers, although for GSE20194 it agrees with identifier-based linkage [44]; the GEO-wide audit covered Affymetrix series found by one search, so repeats on other platforms or in series it missed are not counted, and the article count covers open-access full texts only. (7) Pathway scores measure coordinated transcript abundance, not protein or pathway activity, and the models predict response under the treatment received, not benefit relative to another treatment. (8) Performance was not assessed across race, ethnicity or age groups, because most cohorts do not record race and the numbers of events are small.

4.2 Future validation

The frozen coefficients, scalers and code are released so that the clinical and three-set proliferation models can be evaluated on a new, consecutively accrued HER2-negative neoadjuvant cohort with pretreatment expression and centrally reviewed pCR. That evaluation should fix the recalibration procedure in advance, report AP and Brier score with paired intervals, and include an array-level duplicate check against public series. In immunotherapy, a test of pretreatment immune programs would need cohorts large enough to separate checkpoint blockade from vaccine therapy.

5 Conclusions

In HER2-negative breast cancer treated with neoadjuvant taxane-anthracycline chemotherapy, pretreatment pathway scores predicted pathologic complete response in an untouched external cohort, and the signal was proliferation. Proliferation programs were associated with pCR in

the same direction across five breast cohorts after adjustment for receptor status and grade, and a three-program proliferation model captured what the full 50-program library offered. Locked before new data were seen, this three-program model improved the ranking of patients by pCR probability over clinical variables in five unseen cohorts with 1,121 patients, although it did not lower overall prediction error. Immune programs carried the response signal in antigen-specific melanoma immunotherapy instead, so response programs follow the disease and therapy context. The full curated library did not improve on clinical variables, matched random gene sets performed as well and frozen models needed local calibration, which argues for compact, mechanism-anchored models over large pathway panels. Public breast cancer series in GEO contained 401 identical arrays repeated without declaration across 11 series pairs, and retaining such duplicates inflated validation AUROC by up to 0.25, so array-level deduplication should precede any multi-cohort validation of expression-based response models.

List of abbreviations

AC, doxorubicin and cyclophosphamide; AP, average precision; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CV, cross-validation; ER, estrogen receptor; FAC, fluorouracil, doxorubicin and cyclophosphamide; FEC, fluorouracil, epirubicin and cyclophosphamide; FPKM, fragments per kilobase per million mapped reads; GEO, Gene Expression Omnibus; HER2, human epidermal growth factor receptor 2; I-SPY, Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis; MAGE-A3, melanoma-associated antigen A3; pCR, pathologic complete response; PR, progesterone receptor; RECIST, Response Evaluation Criteria in Solid Tumors; RE, random effects; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Supplementary information

Additional file 1. TRIPOD+AI reporting checklist with the location of each item in the manuscript (PDF).

Additional file 2. GEO-wide duplicate map: every array pair above Spearman $r = 0.97$ between series, with the later and earlier series and arrays, the correlation, whether the arrays are identical ($r \geq 0.999$), the pair class and agreement of age, ER status and pCR where indexed

(CSV).

Additional file 3. Series in the GEO-wide audit with platform, number of arrays, submission date, submitting institute and number of confirmed repeat arrays (CSV).

References

[1] Christos Hatzis, Lajos Pusztai, Vicente Valero, Daniel J. Booser, Laura Esserman, Ana Lluch, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*, 305(18):1873–1881, 2011. doi: 10.1001/jama.2011.593.

[2] Patricia Cortazar, Lijun Zhang, Michael Untch, Keyur Mehta, Joseph P Costantino, Norman Wolmark, Hervé Bonnefoi, David Cameron, Luca Gianni, Pinuccia Valagussa, Sandra M Swain, Tatiana Prowell, Sibylle Loibl, D Lawrence Wickerham, Jan Bogaerts, Jose Baselga, Charles Perou, Gideon Blumenthal, Jens Blohmer, Eleftherios P Mamounas, Jonas Bergh, Vladimir Semiglazov, Robert Justice, Holger Eidtmann, Soonmyung Paik, Martine Piccart, Rajeshwari Sridhara, Peter A Fasching, Leen Slaets, Shenghui Tang, Bernd Gerber, Charles E Geyer, Richard Pazdur, Nina Ditsch, Priya Rastogi, Wolfgang Eiermann, and Gunter von Minckwitz. Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. *The Lancet*, 384(9938):164–172, 2014. doi: 10.1016/s0140-6736(13)62422-8.

[3] Mark Ayers, Jared Lunceford, Michael Nebozhyn, Erin Murphy, Andrey Loboda, David R. Kaufman, Andrew Albright, Jonathan D. Cheng, S. Peter Kang, Veena Shankaran, Sarina A. Piha-Paul, Jennifer Yearley, Tanguy Y. Seiwert, Antoni Ribas, and Terrill K. McClanahan. Ifn-γ–related mrna profile predicts clinical response to pd-1 blockade. *Journal of Clinical Investigation*, 127(8):2930–2940, 2017. doi: 10.1172/jci91190.

[4] Carsten Denkert, Sibylle Loibl, Aurelia Noske, Marc Roller, Berit Maria Müller, Martina Komor, Jan Budczies, Silvia Darb-Esfahani, Ralf Kronenwett, Claus Hanusch, Christian von Törne, Wilko Weichert, Knut Engels, Christine Solbach, Iris Schrader, Manfred Dietel, and Gunter von Minckwitz. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *Journal of Clinical Oncology*, 28 (1):105–113, 2010. doi: 10.1200/jco.2009.23.7370.

[5] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

[6] Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discovery*, 12(1):31–46, 2022. doi: 10.1158/2159-8290.cd-21-1059.

[7] Momeneh Foroutan, Dharmesh D. Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19 (1), November 2018. doi: 10.1186/s12859-018-2435-4.

[8] Daniel G. Stover, Jonathan L. Coloff, William T. Barry, Joan S. Brugge, Eric P. Winer, and Laura M. Selfors. The role of proliferation in determining response to neoadjuvant chemotherapy in breast cancer: A gene expression–based meta-analysis. *Clinical Cancer Research*, 22(24):6039–6050, 2016. doi: 10.1158/1078-0432.ccr-16-0471.

[9] Jeffrey A. Thompson, Brock C. Christensen, and Carmen J. Marsit. Methylation-to-expression feature models of breast cancer accurately predict overall survival, distant-recurrence free survival, and pathologic complete response in multiple cohorts. *Scientific Reports*, 8(1), March 2018. doi: 10.1038/s41598-018-23494-0.

[10] Kuang Du, Shiyou Wei, Zhi Wei, Dennie T. Frederick, Benchun Miao, Tabea Moll, Tian Tian, Eric Sugarman, Dmitry I. Gabrilovich, Ryan J. Sullivan, Lunxu Liu, Keith T. Flaherty, Genevieve M. Boland, Meenhard Herlyn, and Gao Zhang. Pathway signatures derived from on-treatment tumor specimens predict response to anti-pd1 blockade in metastatic melanoma. *Nature Communications*, 12(1), October 2021. doi: 10.1038/s41467-021-26299 -4.

[11] Dohee Lee, Jaegyoon Ahn, and Jonghwan Choi. Pathnetdrp: a novel biomarker discovery framework using pathway and protein–protein interaction networks for immune checkpoint inhibitor response prediction. *BMC Bioinformatics*, 26(1), May 2025. doi: 10.1186/s12859 -025-06125-0.

[12] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), February 2006. doi: 10.1186/1471-2105-7-91.

[13] Richard D Riley, Joie Ensor, Kym I E Snell, Frank E Harrell, Glen P Martin, Johannes B Reitsma, Karel G M Moons, Gary Collins, and Maarten van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ*, page m441, 2020. doi: 10.1136/bmj.m441.

[14] David Venet, Jacques E. Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, 7(10):e1002240, 2011. doi: 10.1371/journal.pcbi.1002240.

[15] Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frédéric Schütz, Darlene R Goldstein, Martine Piccart, and Mauro Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4), 2008. doi: 10.1186/bcr2124.

[16] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, and Michael Birrer. The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *Journal of the National Cancer Institute*, 108(11):djw146, 2016. doi: 10.1093/jnci/djw146.

[17] Leming Shi, Gregory Campbell, Wendell D. Jones, Fabien Campagne, Zhining Wen, Stephen J. Walker, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–838, 2010. doi: 10.1038/nbt.1665.

[18] Tomohiro Miyake, Takahiro Nakayama, Yasuto Naoi, Noriaki Yamamoto, Yoko Otani, Seung J. Kim, Kenzo Shimazu, Atsushi Shimomura, Naomi Maruyama, Yasuhiro Tamaki, and Shinzaburo Noguchi. Gstp1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *Cancer Science*, 103(5):913–920, 2012. doi: 10.1111/j.1349-7006.2012.02231.x.

[19] Christine E. Horak, Lajos Pusztai, Guan Xing, Ovidiu C. Trifan, Cristina Saura, Ling-Ming Tseng, Stephen Chan, Rosanne Welcher, and David Liu. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or paclitaxel in early-stage breast cancer. *Clinical Cancer Research*, 19(6):1587–1595, 2013. doi: 10.1158/1078-0432.ccr-12-1359.

[20] Fernando Ulloa-Montoya, Jamila Louahed, Benjamin Dizier, Olivier Gruselle, Bart Spiessens, Frédéric F. Lehmann, Stefan Suciu, Wim H.J. Kruit, Alexander M.M. Eggermont, Johan Vansteenkiste, and Vincent G. Brichard. Predictive gene signature in mage-a3 antigen-specific cancer immunotherapy. *Journal of Clinical Oncology*, 31(19):2388–2395, 2013. doi: 10.1200/jco.2012.44.3762.

[21] Nadeem Riaz, Jonathan J. Havel, Vladimir Makarov, Alexis Desrichard, Walter J. Urba, Jennifer S. Sims, F. Stephen Hodi, Salvador Martín-Algarra, Rajarsi Mandal, William H. Sharfman, Shailender Bhatia, Wen-Jen Hwu, Thomas F. Gajewski, Craig L. Slingluff, Diego Chowell, Sviatoslav M. Kendall, Han Chang, Rachna Shah, Fengshen Kuo, Luc G.T. Morris, John-William Sidhom, Jonathan P. Schneck, Christine E. Horak, Nils Weinhold, and Timothy A. Chan. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949.e16, 2017. doi: 10.1016/j.cell.2017.09.028.

[22] Willy Hugo, Jesse M. Zaretsky, Lu Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, Elizabeth Seja, Shirley Lomeli, Xiangju Kong, Mark C. Kelley, Jeffrey A. Sosman, Douglas B. Johnson, Antoni Ribas, and Roger S. Lo. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, 2016. doi: 10.1016/j.cell.2016.02.065.

[23] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMJ*, 350(jan07 4):g7594–g7594, 2015. doi: 10.1136/bmj.g7594.

[24] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

[25] Robert F. Wolff, Karel G.M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. Probast: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019. doi: 10.7326/m18-1376.

[26] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, 2012. doi: 10.1093/nar/gks1193.

[27] Vlad Popovici, Weijie Chen, Brandon D Gallas, Christos Hatzis, Weiwei Shi, Frank W Samuelson, Yuri Nikolsky, Marina Tsyganova, Alex Ishkin, Tatiana Nikolskaya, Kenneth R Hess, Vicente Valero, Daniel Booser, Mauro Delorenzi, Gabriel N Hortobagyi, Leming Shi, W Fraser Symmans, and Lajos Pusztai. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Research*, 12(1), January 2010. doi: 10.1186/bcr2468.

[28] Shirin Khambata-Ford, Christopher R. Garrett, Neal J. Meropol, Mark Basik, Christopher T. Harbison, Shujian Wu, Tai W. Wong, Xin Huang, Chris H. Takimoto, Andrew K. Godwin, Benjamin R. Tan, Smitha S. Krishnamurthi, Howard A. Burris, Elizabeth A. Poplin, Manuel Hidalgo, Jose Baselga, Edwin A. Clark, and David J. Mauro. Expression of epiregulin and amphiregulin and *K-ras* mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of Clinical Oncology*, 25(22):3230–3237, 2007. doi: 10.1200/jco.2006.10.5437.

[29] W. Fraser Symmans, Florentia Peintinger, Christos Hatzis, Radhika Rajan, Henry Kuerer, Vicente Valero, Lina Assad, Anna Poniecka, Bryan Hennessy, Marjorie Green, Aman U. Buzdar, S. Eva Singletary, Gabriel N. Hortobagyi, and Lajos Pusztai. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 25(28):4414–4422, 2007. doi: 10.1200/jco.2007.10.6823.

[30] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised

recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. doi: 10.1016/j.ejca.2008.10.026.

[31] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. doi: 10.1016/j.cell.2011.02.013.

[32] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14(1), January 2013. doi: 10.1186/1471-2105-14-7.

[33] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[34] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[35] E. W. Steyerberg and Y. Vergouwe. Towards better clinical prediction models: seven steps for development and an abcd for validation. *European Heart Journal*, 35(29):1925–1931, 2014. doi: 10.1093/eurheartj/ehu207.

[36] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1), December 2019. doi: 10.1186/s12916-019-1466-7.

[37] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986. doi: 10.1016/0197-2456(86)90046-2.

[38] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[39] Adel Tabchy, Vicente Valero, Tatiana Vidaurre, et al. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clinical Cancer Research*, 16(21):5351–5361, 2010. doi: 10.1158/1078-0432.CCR-10-1265.

[40] Takayuki Iwamoto, Giampaolo Bianchini, Daniel Booser, et al. Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *Journal of the National Cancer Institute*, 103(3):264–272, 2011. doi: 10.1093/jnci/djq524.

[41] Kai Shen, Yuan Qi, Nan Song, et al. Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on US Oncology 02-103 clinical trial. *BMC Medical Genomics*, 5:51, 2012. doi: 10.1186/1755-8794-5-51.

[42] Laura J. Esserman, Donald A. Berry, Maggie C. U. Cheang, et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Research and Treatment*, 132(3):1049–1062, 2012. doi: 10.1007/s10549-011-1895-2.

[43] Denise M. Wolf, Christina Yau, Julia Wulfkuhle, et al. Redefining breast cancer subtypes to guide treatment prioritization and maximize response: predictive biomarkers across 10 cancer therapies. *Cancer Cell*, 40(6):609–623.e6, 2022. doi: 10.1016/j.ccell.2022.05.005.

[44] Yunhao Jiang. Outcome labels and patient reuse in clinical validation: reanalyses of the DIPK and TIL-30 breast-cancer evaluations. Research Square preprint, 2026.

[45] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *JNCI: Journal of the National Cancer Institute*, 98(4):262–272, 2006. doi: 10.1093/jnci/djj052.