

Outcome labels and patient reuse in clinical validation: reanalyses of the DIPK and TIL-30 breast-cancer evaluations

Yunhao Jiang

biojiang777@gmail.com

University of Edinburgh

Research Article

Keywords: clinical validation, outcome labels, patient reuse, drug response prediction, breast cancer, pathological complete response, reanalysis, data leakage

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11307948/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Public cancer-expression datasets may contain classifier predictions, mid-treatment responses and repeated patient records, which can be mistaken for intended outcomes or independent patients. We traced labels and patient identities in breast-cancer evaluations of DIPK, a deep-learning drug-response model trained on cell lines, and TIL-30, a 30-gene tumor-infiltrating lymphocyte classifier of pathological complete response (pCR).

Results

DIPK's 310 archived GSE25055 labels came from DLDA30, a gene-expression pCR classifier. In 306 patients, evaluating DIPK's saved scores against observed pathology reduced the area under the receiver operating characteristic curve (AUC, calculated here) from 0.852 to 0.748 (paired difference -0.103 ; bootstrap 95% interval, -0.181 to -0.030). The same label change lowered AUC from 0.854 to 0.758 for SLM-HRD, a correctly labeled signature, so the drop was not DIPK-specific; SLM-HRD's drop was uncertain in another cohort. TIL-30's 122 GSE41998 labels recorded complete response to initial doxorubicin/cyclophosphamide, not final pCR; AUC changed from 0.670 to 0.524 (paired difference -0.146 ; 95% interval, -0.300 to 0.008). DIPK's GSE20194 and GSE25055 cohorts shared 188 of GSE20194's 248 patients, reducing independent support. Patient identifiers linked 60 of TIL-30's 71 GSE20194 validation records (reported AUC 0.96) to 57 training patients; the effect on AUC was not estimated. DIPK's predicted response remained higher in observed responders, but exploratory within-cohort AUC gains from adding its scores to seven clinical variables had intervals including zero. Our score-based logistic models misestimated pCR rates when transferred between cohorts.

Conclusions

DIPK's scores remained associated with observed response, but their value beyond seven clinical variables was unresolved; TIL-30's examined AUCs rested on earlier-stage response or largely reused patients. Readers citing validation AUCs should check the scored outcome, its treatment stage and the number of distinct patients. These two purposively selected cases do not estimate how common such problems are.

Introduction

A response label in a public cancer-expression dataset may record a patient's observed outcome or a genomic classifier's prediction of it. We call the label against which predictions are scored the target. Scoring against a classifier's output measures agreement with that classifier, even when the intended claim concerns observed response. In GSE25055, for example, one field records each patient's observed

outcome, pathological complete response (pCR) or residual disease, and another records predictions made from the same patients' pretreatment tumor expression profiles by DLDA30, a 30-probe diagonal linear discriminant analysis classifier of pCR [1]. Another score computed from the same profiles can therefore agree with DLDA30 labels through shared expression patterns, not only through a common relation to observed response. Responses measured before treatment is complete may also differ from final pathology. Describing a target simply as "pCR" can obscure both distinctions.

Many drug-response models, including one of the two predictors examined here, are trained on cancer cell-line screens, and benchmarks test whether their performance survives changes in data and evaluation design. A community benchmark found substantial degradation across drug-screening datasets [2]; DrEval compares models with a naive baseline predicting only mean drug and cell-line effects and tests feature dependence and cross-dataset generalization [3]. Interpreting a clinical evaluation also requires identifying the outcome predicted, the distinct patients supporting it, and the information a score adds to clinical predictors.

Two of these, added information and patient independence, have precedents. A prospective study of 100 breast cancers found similar discrimination for a clinical nomogram and DLDA30, and both identified many of the same responders [4]. An earlier recombination proficiency score (RPS) study removed samples with matching raw signals and scan dates across five breast-cancer datasets and published their identifiers [5]. Hidden duplicate profiles occur in many public cancer transcriptome datasets and can inflate apparent prediction accuracy [6], as can similar samples split between training and validation sets [7]. What remains open for any released evaluation is whether its own outcome definitions and patient records justify its clinical claim.

DIPK (Deep neural network Integrating Prior Knowledge) is a deep-learning model predicting drug response as natural-log half-maximal inhibitory concentration (IC₅₀) from gene expression, prior gene-interaction networks and drug structure; lower values indicate greater sensitivity [8]. The clinically applied version was trained on cell-line data from Genomics of Drug Sensitivity in Cancer (GDSC). Its authors predicted paclitaxel IC₅₀ from pretreatment tumor expression profiles in GSE25055, GSE32646 and GSE20194, three breast-cancer cohorts given taxane-containing multi-drug neoadjuvant chemotherapy. Using Welch t-tests, they reported lower predicted IC₅₀ in patients labeled pCR than in other patients, but not how well the predictions separated these groups (discrimination) [8]. In DIPK's released evaluation, however, every archived GSE25055 label was a DLDA30 prediction rather than observed pathology, and the GSE20194 and GSE25055 cohorts shared patients. Our central reanalysis therefore holds patients and saved scores fixed and replaces DLDA30 labels with observed pathology in both the original group comparison and discrimination. To ask whether closer agreement with DLDA30 labels than with pathology is specific to DIPK, we evaluated released scores from SLM-HRD, a synthetic lethal metagene signature associated with homologous recombination deficiency; its published evaluation, in the SYLVER study, used observed pathology [9]. Exploratory analyses asked whether DIPK's agreement with DLDA30 labels is compatible with tumor subtype structure, recorded as estrogen-receptor status by immunohistochemistry (ER IHC) and PAM50 intrinsic molecular subtypes.

TIL-30 predicts pCR versus residual disease in triple-negative breast cancer from bulk expression of 30 genes selected from a tumor-infiltrating lymphocyte signature found by single-cell RNA sequencing [10]. Its released training input contains 170 records from GSE25055 and GSE25065, and GSE41998 and GSE20194 are two of its 11 external validation cohorts. Our later targeted examination of these inputs found that GSE41998 labels recorded complete response after the initial doxorubicin/cyclophosphamide stage rather than final pCR at surgery, and that most GSE20194 validation records linked to training patients. Unlike the overlap between DIPK's evaluation cohorts, this reuse crosses the training–validation split. The two studies, selected sequentially, not by a representative survey, show how a classifier-derived target, an earlier treatment-stage response and reused patients can alter the interpretation of clinical validation. The Results first establish, for both studies, what each target records and how many distinct patients support it. They then examine DIPK's agreement with DLDA30 labels, its association with observed response within ER IHC and PAM50 groups, what its scores add to clinical predictors within each cohort and, briefly, whether logistic models fitted here to clinical variables and the scores give accurate pCR probabilities when transferred unchanged between GSE25055 and GSE32646.

Materials and methods

Ethics and data use

This study used publicly available, deidentified Gene Expression Omnibus (GEO) data and published model outputs, without new participant recruitment or sample collection. No additional ethical approval was required.

Study design and provenance audit

This study reanalyzed released patient-level records and saved predictions from two published clinical evaluations, examined in sequence. The initial analysis used the three public breast-cancer datasets in DIPK's clinical evaluation; its primary comparison changes the outcome label on identical patients without changing their scores. Further analyses compared DIPK with released SLM-HRD scores, related its agreement with DLDA30 labels to recorded ER IHC and PAM50 groups, traced repeated patients and assessed prediction beyond clinical variables. A later targeted search of other published clinical evaluations identified two TIL-30 datasets, GSE41998 and GSE20194, whose released inputs suggested an outcome discrepancy and patient reuse, respectively. Other evaluations examined were not analyzed because their public resources lacked identifiable patient-level predictions or resolvable labels (S1 Text). The searched set was not defined in advance. Sample sizes were set by the released records and each analysis's eligibility criteria. Each comparison was specified before its effect was calculated but after earlier results were known; bootstrap and permutation draw counts were fixed in each specification and differ between analyses. Table 1 summarizes the evaluation sets and outcome definitions; S1 Text records the sequential exploratory design, analysis settings, draw counts and random seeds.

We recovered DIPK's clinical CSV files from the archive linked to its official repository and joined saved predictions, archived labels and frozen GEO metadata snapshots by exact GSM accession (S1 Text). Scores were DIPK's saved paclitaxel natural-log IC50 predictions with the sign reversed, as plotted in its clinical figure, so higher scores mean lower predicted IC50, greater predicted sensitivity and a higher predicted chance of response. The scores agreed with the repository's figure data, and our recalculated Welch t-tests agreed with the P values in that figure. DIPK's evaluation did not report the area under the receiver operating characteristic curve (AUC), which we calculated for this reanalysis. All response analyses use the saved scores unchanged, without retraining DIPK. Archive checks and a label-only check of the GSE25055 evaluation code of VAEN, a deep generative model for imputing drug response, are detailed in S1 Text.

Table 1
Datasets, evaluation units and outcome definitions.

Dataset and role	Evaluation records	Released target and observed outcome
GSE25055 DIPK and SLM-HRD reference	DIPK: 310 released records; 306 patients with known pathology. SLM-HRD: the same 306 patients.	DIPK labels are DLDA30 predictions: 121 positives among the 306 patients, versus 57 observed pCR and 249 RD. SLM-HRD source labels match pathology.
GSE32646 DIPK	115 patients	Released labels match observed pathology: 27 pCR and 88 non-pCR.
GSE20194 DIPK and TIL-30 validation	DIPK: 278 arrays; 248 source patients. TIL-30: 71 records; 68 source patients.	Both released label sets match pathology. DIPK includes 56 pCR arrays from 50 patients; TIL-30 includes 25 pCR records.
GSE25065 SLM-HRD reference	181 released scores	Source labels match pathology: 42 pCR and 139 RD. DLDA30 labels have 69 positive records.
GSE41998 TIL-30	122 released records	Released labels represent AC complete response (complete response versus any other response category after AC): 24 positives, versus 49 with final pCR at surgery. The two targets differ in 43 records.

Counts describe each evaluation separately; some patients appear in more than one dataset. GSE25055 and GSE25065 also contribute to TIL-30's 170-record training input (57 pCR, 113 RD), whose labels match observed pathology (S1 Text). GSE25055 clinical prediction analyses use the 301 patients with observed pathology and positive or negative ER IHC status, 56 with pCR. AC,

doxorubicin/cyclophosphamide; ER IHC, estrogen-receptor immunohistochemistry; pCR, pathological complete response; RD, residual disease.

Outcome correction and patient identity

The primary analysis asked how DIPK's discrimination in GSE25055 changes when the archived DLDA30 labels are replaced by observed pathology, with patients and scores held fixed. Throughout, observed pathology means pCR versus non-pCR after chemotherapy as recorded by each dataset's source study; in GSE25055, this was the observed pCR/residual disease (RD) field, not the pCR-or-minimal-residual-disease composite used for some source-study evaluations [1]. The analysis included the same 306 patients under both label definitions, excluding four with missing observed outcomes. Rank-based AUCs and their paired difference (observed pathology minus DLDA30 labels) were estimated with 2,000 paired patient bootstrap draws and percentile 95% intervals. We call an estimate resolved when its 95% interval excludes the null value (zero for a difference, 0.5 for an AUC, 1 for an odds ratio), and uncertain or unresolved otherwise. Because GSE25055 combines two recruitment sources, MD Anderson Cancer Center (MDACC) and the I-SPY trial, a sensitivity analysis repeated the comparison within each recruitment source and averaged the two within-source AUCs with the same fixed patient-count weights under both label definitions, using 2,000 paired draws within each recruitment source; unlike the pooled AUC, this average excludes pairs of patients from different recruitment sources. A separate 2,000-draw paired bootstrap supplied two secondary estimates: average precision under both label definitions, reported with the positive fraction because changing the labels changes that fraction; and the AUC difference between DIPK scores and the binary DLDA30 call against observed pathology. Continuous DLDA30 scores were unavailable.

For GSE20194, we counted distinct patients and searched for patients also deposited in GSE25055 under different GEO accessions. We matched the official sample-information workbook and CodeBook [11] to GEO by specimen title and CEL filename, and cross-checked the patient identifiers recorded by the source study (source patient identifiers) against specimen aliases. GSE20194 identifiers were linked only to GSE25055 patients from its MDACC recruitment source. We compared raw CEL filenames and, for three pairs chosen by a rule fixed in advance, the intensity, standard-deviation and pixel-count values read directly from both raw files (S1 Text). We also compared all linked GSE25055 accessions with the earlier RPS study's duplicate-exclusion list [5], which flags duplicate samples across five datasets but does not name each one's partner.

Because some GSE20194 patients contributed repeated arrays and some also appear in GSE25055, we evaluated DIPK's AUC there by array and by patient. For the array-weighted AUC across all 278 arrays, naive array resampling was compared with patient-cluster resampling, which keeps every array of each sampled patient; both estimate the same array-weighted quantity. Three patient-level analyses used one saved score per source patient identifier: the mean score of each of all 248 patients; the score of the single original MicroArray Quality Control (MAQC) training or validation array of each of 230 patients, a subset excluding the replicate and outlier arrays flagged in the source study's workbook; and the mean

score of each of the 60 patients without an identifier linked to GSE25055. Each analysis used 2,000 bootstrap draws. MAQC flags describe source-study eligibility, not DIPK training membership.

Targeted reanalysis of TIL-30

For TIL-30, we analyzed the predicted probabilities saved in the official repository [10] for GSE41998 and GSE20194 unchanged, without retraining, thresholding or changing the score direction. Higher probabilities indicate pCR, the positive class in the source training data. In the archived code, the 30 features and their standardization came from the 170-record training input alone, whereas a class-imbalance resampling method and classifier were chosen separately for each external cohort. The article calls the comparison based on these cohort-specific choices its upper-bound comparison; apart from the GSE20194 reporting difference described in Results, AUCs recalculated from the saved outputs match those it reports. The article states that these choices were based on training-set cross-validation, but the code that made them is not in the repository, so we could not verify how they were selected; this concerns the reported performance level, not the label and identity findings examined here (S1 Text). The article's separate benchmark, which applies one fixed pipeline to every signature and cohort, was not recalculated. Saved probabilities were linked to GEO samples by GSM accession in GSE41998 and by original row order in GSE20194.

Patients in GSE41998 received neoadjuvant doxorubicin/cyclophosphamide (AC), followed by ixabepilone or paclitaxel [12]. GEO records both the response category after AC and final pCR at surgery; a separate pCR-or-minimal-residual-disease field was not used. The AC categories grade tumor response after AC, not surgical pathology; GEO does not state how response was assessed. After confirming that all 122 released labels matched AC complete response (complete response versus any other AC response category), we specified the paired AUC difference when the same saved scores are evaluated against final pCR instead (final pCR minus AC complete response), with 5,000 paired record bootstrap draws and percentile 95% intervals. The same draws supplied secondary average-precision estimates, reported with both positive fractions. A later check found two sets of three records; each set shared a specimen-name prefix and matching clinical annotations and may come from one patient. A post hoc sensitivity analysis therefore resampled the 118 candidate specimen-prefix groups (116 single records and the two sets of three) instead of single records, with 5,000 draws; each draw kept every record of a sampled group.

To check whether TIL-30's GSE20194 validation patients also appear in its training input, we traced all 71 validation records to source patient identifiers and checked their links to the 170-record training input using the identity rules and raw-filename comparisons above, reporting records and distinct source patients separately. S1 Text gives source-output agreement, differences from article tables and numerical checks.

Paired comparison with a released biological signature

To examine whether a higher AUC against DLDA30 labels than against observed pathology was specific to DIPK, we used released scores from SLM-HRD as a reference [9]. Sheet g of the SYLVER Fig. 4 source data supplied the 73-gene signature's scores and outcome labels for cohort Hatzis 1, identified in the article as GSE25055. Removing the common H. prefix matched every SYLVER sample identifier to a unique GEO sample ID and then GSM accession; no row-order, score or outcome matching was used. We then compared the published labels with the GEO observed-pathology and DLDA30 fields. Following its source study, higher SLM-HRD scores, like higher DIPK scores, indicate a higher predicted chance of response.

Before calculating these comparisons, we specified AUC and average precision for both unchanged scores under both targets in the same 306 patients. The primary contrast asked whether replacing observed pathology with DLDA30 labels raised DIPK's AUC more than SLM-HRD's: the DIPK-minus-SLM-HRD AUC difference against DLDA30 labels minus the same difference against observed pathology, so a positive value means that DIPK's AUC rose more. We used 5,000 paired patient bootstrap draws shared across scores, labels and contrasts; the same draws supplied AUCs within each recruitment source. DLDA30 labels are a substitute target, applied to SLM-HRD scores only in these comparisons.

To examine whether SLM-HRD's AUC also differs between the two targets in a separate cohort, a second comparison used its 181 released scores for GSE25065, the validation cohort of the GSE25055 source study [1, 9]. After checking eligibility and patient identifiers, we specified the paired AUC and average-precision differences (observed pathology minus DLDA30 labels), using 5,000 patient bootstrap draws. A sensitivity analysis within the same draws excluded the 28 accessions on the RPS duplicate-exclusion list [5]. One known-RD patient lacked a released score; instead of imputing it, we placed that score at every possible rank among the 181 known scores to bound the AUCs and their paired difference.

Permutation of DLDA30 labels within recorded groups

This exploratory diagnostic asked how high the saved scores' AUC against DLDA30 labels would be expected to be if, within recorded patient groups, those labels were unrelated to the scores. Keeping all 306 GSE25055 patients with their observed outcomes and saved scores, we permuted DLDA30 labels within observed-outcome groups and, separately, within groups defined by observed outcome and recruitment source. Each permutation preserves the number of DLDA30-positive labels in each group, and therefore all 90 disagreements with pathology, while randomizing which patients within a group carry them. We derived the exact expected AUC from rank sums and used 10,000 permutations per scheme for central 95% reference ranges. The outcome-only expectation follows the established AUC relationship for class-conditional label mixtures [13]. The closer a reference expectation lies to the actual AUC, the more of the agreement between scores and DLDA30 labels is compatible with the recorded groups alone.

After inspecting those results, we specified three further schemes that also grouped patients, within observed outcome and recruitment source, by ER IHC status, PAM50 subtype, or both; PAM50 was the

principal molecular-background comparison. Four indeterminate and one missing ER values were kept as separate categories. Permutation cannot move labels in a stratum whose patients all share one DLDA30 label, so we report how many patients were in strata containing both labels. Four conditional AUCs measured how well the saved scores separated DLDA30-pCR from DLDA30-RD labels among observed-RD patients and, separately, among observed-pCR patients, and how well they separated observed pCR from RD within the DLDA30-RD group and, separately, within the DLDA30-pCR group. A separate 2,000-draw patient bootstrap estimated sampling intervals for the exact expectations, the actual-minus-expected differences and the conditional AUCs. Permutation ranges show the AUCs expected if DLDA30 labels were exchangeable within recorded categories, whereas bootstrap intervals describe patient sampling with saved scores fixed. S1 Text gives formulas, seeds, complete estimates and undefined-draw counts.

Response association within ER and PAM50 groups

These analyses asked whether the saved scores remain associated with observed response among patients who share an ER IHC status or PAM50 subtype, and after adjustment for these variables. Within-group AUC used only pCR/non-pCR pairs of patients in the same group, so each group was weighted by its number of eligible pairs; intervals came from 2,000 patient bootstrap draws, and draws lacking an outcome class were not redrawn. Logistic models related observed pCR to the standardized saved score, age and either ER IHC or PAM50. The GSE25055 ER analysis included 301 patients, excluding four indeterminate and one missing IHC status without substituting expression-revised labels; the PAM50 analysis retained 306 patients and the GSE32646 ER analysis 115. Odds ratios (ORs) are per within-analysis score standard deviation, with Wald intervals (S1 Text gives settings and full estimates).

Exploratory predictive-value analyses

These analyses asked whether adding the saved score improves held-out prediction of observed pCR beyond clinical predictors, a different question from association. They used the same 301 GSE25055 patients and all 115 GSE32646 patients. The compared models were training prevalence alone, age plus ER, the saved score alone, and age plus ER plus score; in GSE25055, PAM50 could also be added to the clinical predictors. Apart from training prevalence, every model used ridge (L2-penalized) logistic regression with an unpenalized intercept and a fixed, untuned penalty (scikit-learn $C = 1$, the inverse of penalty strength). We call these models, all fitted in this study, secondary models. They convert clinical variables and saved DIPK scores into pCR probabilities, which DIPK itself does not provide because it outputs a predicted log IC50.

A broader clinical comparison asked whether the score still adds discrimination when more clinical information is available. It added progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status as recorded in each source study, clinical T stage, nodal positivity and histological grade to age and ER. These seven predictors were compared without and with the saved

score in each cohort and, in GSE25055, also with PAM50 added to both models. Indeterminate PR, HER2 and grade values were treated as missing; missing binary values were imputed with the training mode and missing grades with the training median. Numeric predictors were standardized within training; T stage and grade entered as ordinal scores, and PAM50 indicators were not standardized. Each source study's receptor definitions were retained, and no treatment-response or follow-up variable entered a predictor set. S1 Text gives field definitions, missingness and the limited cross-cohort comparability of HER2 status.

Every comparison used the same ten stratified five-fold partitions within each cohort. AUC, average precision, Brier score and logarithmic loss were calculated on each complete held-out prediction vector and averaged over the ten repeats; the range of paired differences across repeats shows only their dependence on how patients were split into folds. For the expanded clinical comparisons, 1,000 patient bootstrap draws per cohort repeated imputation, standardization and fitting. Each draw used the same resampled patients, with the same copy numbers, for every model and repeat, and all copies of a patient stayed in that patient's original fold, so no patient was in both training and validation within a fit. The 2.5th and 97.5th percentiles of the paired increments give exploratory intervals reflecting patient resampling and secondary-model fitting, conditional on the original partitions and saved DIPK scores. Because cross-validation errors are dependent, folds and repeats were not treated as independent observations [14].

Transfer of secondary models between cohorts

This analysis asked whether secondary models fitted in one cohort still rank patients and estimate pCR probabilities accurately in another. We call the cohort in which a model is fitted the source cohort and the cohort to which it is applied unchanged the target cohort. The prevalence-only model and the age/ER, score-only and age/ER/score secondary models were fitted to all eligible source-cohort patients (301 in GSE25055 or 115 in GSE32646) and applied with the source cohort's standardization and coefficients to the target cohort, in both directions; target-cohort data were not used for centering or recalibration. Mean predicted probability minus the observed pCR fraction in the target cohort measured mean-risk error (calibration-in-the-large). Uncertainty was estimated in two ways, each with 2,000 draws per direction shared across models: resampling target-cohort patients with the source-cohort fits held fixed, and resampling source- and target-cohort patients independently while refitting standardization and coefficients in each source-cohort sample. The expanded clinical models were evaluated within cohorts only.

Numerical verification

Key estimates agreed across separate R and Python implementations (including SciPy and scikit-learn) for the primary AUCs, patient aggregation, the adjusted logistic models, the penalized secondary-model coefficients and the case-weighted bootstrap refits; direct pair enumeration, metric formulas and fold checks confirmed discrimination, probability losses, patient coverage and training–test separation. Reported differences, including paired and cross-validated increments, were calculated from unrounded

estimates and may differ by up to 0.001 from differences between the rounded values shown. S1 Text reports all checks and complete estimates.

A separate execution check reran the released five-model DIPK ensemble on its archived GSE25055 features to assess whether applying DIPK to a new cohort could be justified. On CPU and under two GPU precision settings, the maximum absolute differences from the saved predictions across all 310 records ranged from 0.0005 to 0.0013 log-IC50 units, all above the maximum-error tolerance of 10^{-4} log-IC50 units, which was recorded before execution and not relaxed. Because this discrepancy remains unresolved, no new-cohort inference was undertaken (S1 Text).

Results

Replacing DLDA30 labels with observed pathology lowers AUC in GSE25055

All 310 archived labels in DIPK's GSE25055 evaluation matched the GEO field `dllda30_prediction`, which holds DLDA30's binary pCR/RD call. Among the 306 patients with known pathology, DLDA30 labeled 121 as pCR, whereas 57 had observed pCR. The two labels disagreed for 90 patients: 77 labeled pCR by DLDA30 had observed RD, and 13 labeled RD had observed pCR (Fig. 1A; Fig. 1D shows saved DIPK scores for each label combination). Replacing DLDA30 labels with observed pathology therefore lowers the pCR fraction from 39.5% to 18.6%. By contrast, all archived labels in DIPK's GSE32646 cohort matched observed pathology (27 pCR and 88 non-pCR; saved-score AUC 0.761).

In GSE25055, with patients and saved scores held fixed, AUC fell from 0.852 (bootstrap 95% interval, 0.806–0.894) against DLDA30 labels to 0.748 (0.675–0.816) against observed pathology; the paired difference was -0.103 (-0.181 to -0.030 ; Fig. 1B and 1E). DIPK's original clinical figure tested a different quantity, the difference in mean saved score between response groups. This difference was 0.265 negative-log-IC50 units against DLDA30 labels (Welch $P = 1.18 \times 10^{-28}$) and remained positive at 0.202 against observed pathology (Welch $P = 7.96 \times 10^{-9}$). Replacing the labels thus reduced both this difference and the AUC, but the association that DIPK originally tested remained.

AUC point estimates also fell within each recruitment source (Fig. 1E): from 0.847 to 0.771 among 227 MDACC patients (paired difference -0.076 ; -0.153 to 0.006) and from 0.873 to 0.668 among 79 I-SPY patients (-0.204 ; -0.394 to -0.032). Only the I-SPY interval excludes zero; no between-source difference was tested (Table S11). Averaging the two within-source AUCs with fixed patient-count weights (227/306 and 79/306) gave a difference of -0.109 (-0.186 to -0.035), so a reduction remains when only patients from the same source are compared.

Average precision fell from 0.798 against DLDA30 labels to 0.468 against observed pathology, a paired difference of -0.330 (-0.460 to -0.180 ; second 2,000-draw bootstrap). This change reflects both the different labels and their lower pCR fraction.

The binary DLDA30 call itself provides a direct comparator. Against observed pathology, it had AUC 0.731 versus 0.748 for DIPK's saved scores; the paired DIPK-minus-DLDA30 difference was only 0.017 (– 0.059 to 0.090; same bootstrap as average precision). DIPK's strong agreement with DLDA30 labels therefore does not show that it predicts observed pathology better than DLDA30 does. The call's AUC reflects a single operating point (sensitivity 44/57, specificity 172/249), whereas a continuous score can also rank patients within each call, so this comparison may favor DIPK. Whether such agreement with DLDA30 labels is specific to DIPK is examined after the TIL-30 and patient-identity results (Fig. 1C and 1F).

TIL-30's released GSE41998 labels record complete response after AC, not final pathology

The second outcome discrepancy concerns treatment stage. The TIL-30 article's Table 1 describes the GSE41998 outcome as “pCR status”, with 51 pCR and 74 non-pCR among 125 records [10]; these counts are consistent with final pathology in the 122 released records (49 pCR), not with AC complete response (24). However, all 122 saved TIL-30 labels matched AC complete response in the GEO “ac response” field: the 24 positives had complete response after AC, and the 98 negatives had partial response, stable disease or progressive disease. These labels differed from final pathology in 43 records: nine had AC complete response without final pCR, and 34 had final pCR without AC complete response (Fig. 2A). With the same records and scores, AUC changed from 0.670 (95% interval, 0.537–0.795) against AC complete response to 0.524 (0.420–0.633) against final pathology (Fig. 2B), a paired difference of – 0.146 whose interval included zero (– 0.300 to 0.008; Fig. 2C; Table S20). Resampling the 118 candidate specimen groups instead of records gave almost the same interval (– 0.303 to 0.008).

Average precision rose from 0.385 to 0.427, an uncertain paired difference (0.042; –0.163 to 0.250), while the positive fraction rose from 19.7% to 40.2% (Fig. 2E). Because a random ranking has average precision approximately equal to the positive fraction, this rise does not indicate better ranking: average precision was well above that chance level against AC complete response but close to it against final pathology. The article's three additional records cannot be identified in the released outputs, so inference is restricted to the 122 released records, whose AUC against the released labels (0.670) agrees with the archived TIL-30 notebook and the article's reported 0.67.

Shared patients reduce independent support across DIPK's evaluation cohorts and cross TIL-30's training–validation split

In both studies, source patient identifiers link GSE20194 records to GSE25055 records, although the two series share no GSM accession; the consequences differ. DIPK used both series as clinical evaluation cohorts, and 188 of its GSE20194 patients link to GSE25055 (Fig. 3A). This overlap limits how many distinct patients support DIPK's clinical results; because DIPK was trained on cell-line data, it does not by

itself show training leakage. TIL-30's training input included GSE25055, and GSE20194 was one of its validation cohorts; 60 of its 71 GSE20194 validation records link to 57 training patients (Fig. 2D), so this reuse crosses the published training–validation split. All 57 are among DIPK's 188 linked patients.

The 188 linked patients correspond to 199 GSE20194 arrays; for 187 of these patients, at least one array had exactly the same raw CEL filename in both series. Of three raw-array pairs chosen by a fixed rule (Fig. 3B), the two with matching filenames, M157 and M806, were identical in all 506,944 values of each stored field (intensity, standard deviation and pixel count) after conversion to float32 precision. The third, M367, the only linked patient without a filename match, had different intensities and was kept as a link on the source patient identifier alone. Observed outcomes agreed on all linked records. Recorded age differed for 26 of the 188 linked patients (28 of 199 arrays; most by one year, at most by ten years), and ER IHC status differed for one patient; these covariates were not harmonized. No GSM or raw-filename link to GSE32646 was identified.

The earlier RPS duplicate-exclusion list [5] contained 188 GSE25055 accessions, 186 of which matched the 188 accessions linked here to GSE20194 (Fig. 3C; Table S13); each list contained two accessions absent from the other. This agreement corroborates most links, while the differences show that the two curation rules are not interchangeable.

Within GSE20194, some patients also contributed more than one array: DIPK's 278 arrays represented 248 source patient identifiers, 218 appearing once and 30 twice. Of these, 188 link to GSE25055 and 60 have no known link (Fig. 3A). Labels matched observed pathology in all 278 arrays, including 56 pCR arrays, and outcomes agreed within every repeated-patient group. The source study's workbook assigned 130 arrays to its MAQC training set, 100 to its validation set, 29 to a replicate category and 19 to an outlier category (one also a replicate); the 230 training and validation arrays came from 230 distinct patients.

GSE20194 retained discrimination when each source patient was counted once: AUC was 0.757 for patient-mean scores across all 248 patients and 0.746 in the 230-patient original MAQC training and validation subset (Table 2; Fig. 3D). The 60 patients without a known GSE25055 link gave AUC 0.712, with a wider interval (0.571–0.834). Because these summaries differ in their scores or patients, differences between them do not estimate an effect of patient reuse.

Table 2
GSE20194 results by evaluation unit and eligibility.

Evaluation set	Units	pCR	AUC	Patient bootstrap 95% interval
All released arrays	278 arrays	56	0.736	0.664–0.807 (patient clusters)
Mean saved score per source patient	248 patients	50	0.757	0.684–0.827
Original MAQC training plus validation subset	230 patients	48	0.746	0.664–0.818
Patient means without a GSE25055 identifier link	60 patients	18	0.712	0.571–0.834

Intervals use 2,000 patient bootstrap draws; in the array row, each sampled patient contributes all of their arrays. Patient means average negative saved predicted log IC50. MAQC labels refer to the source study.

Repeated arrays raise a separate question: whether counting arrays rather than patients understates uncertainty. The array-weighted AUC of 0.736 had 95% intervals of 0.665–0.802 under naive array resampling and 0.664–0.807 under patient-cluster resampling; the width increased only from 0.137 to 0.143 (Fig. 3E; Supplementary Figure S1), so grouping the arrays of the 30 patients with two arrays changed precision little. This analysis concerns repeated arrays within GSE20194, not the 188 patients linked to GSE25055; only 60 GSE20194 patients have no known link to that cohort (Table 2).

TIL-30's GSE20194 validation input contains 71 records from 68 source patients. Its 60 linked records correspond to 57 patients in the GSE25055 part of the training input, which supplies 112 of the 170 training records; 56 of these 60 links also have matching raw CEL filenames (Fig. 2D; Table S21). The other 11 records represent 11 patients without a known training link. All 71 validation labels agree with recorded pathology. The TIL-30 article reports AUC 0.96 for this input, whereas the saved outputs give 0.9548 (0.95 at two decimals); this reporting difference remains unresolved (S1 Text).

A correctly labeled signature also has higher AUC against DLDA30 labels than against pathology in GSE25055

Returning to GSE25055, we asked whether a higher AUC against DLDA30 labels than against observed pathology is specific to DIPK, using released SLM-HRD scores (Fig. 1C and 1F). All 306 released SLM-HRD records matched unique GSE25055 patients, and all published labels agreed with observed pathology (57 pCR, 249 RD); the four GEO records absent from the release were exactly those with unknown response. The published labels therefore differed from DLDA30 labels in the same 90 patients as in the DIPK evaluation.

SLM-HRD had AUC 0.758 against observed pathology (paired patient-bootstrap 95% interval, 0.683–0.826) and 0.854 (0.811–0.894) against DLDA30 labels, a counterfactual target not used by its source

study (Fig. 1C). The pathology-minus-DLDA30 difference was -0.097 (-0.177 to -0.021 ; Fig. 1F). DIPK-minus-SLM-HRD AUC was -0.009 against pathology (-0.071 to 0.055) and -0.003 against DLDA30 labels (-0.046 to 0.040). The change between these two contrasts, which equals DIPK's AUC loss minus SLM-HRD's when the target changes from DLDA30 labels to pathology, was 0.007 (-0.060 to 0.068); its interval includes zero. Thus, a separately derived signature with correct source labels reached almost the same AUC against DLDA30 labels as DIPK (0.854 vs 0.852) and a similar pathology-minus-DLDA30 difference (-0.097 vs -0.103): DIPK's high AUC against DLDA30 labels is not specific to DIPK. The data do not resolve which score better predicts observed pathology. Tables S17 and S18 give average-precision contrasts and both recruitment-source sensitivity analyses.

In GSE25065, all 181 released labels matched pathology (42 pCR, 139 RD), whereas 65 differed from DLDA30 labels. SLM-HRD AUC was 0.748 against pathology and 0.786 against DLDA30 labels, a paired difference of -0.039 (95% interval, -0.145 to 0.067). Excluding the 28 RPS-flagged samples left 153 patients and a difference of -0.064 (-0.175 to 0.039 ; Fig. 1F; Table S19). Placing the one known-RD patient without a released score at every possible rank bounded the difference at -0.039 to -0.037 , so this missing score cannot explain the uncertainty. Average precision decreased under pathology in both analysis sets, alongside lower positive-label prevalence. Both AUC-difference intervals include zero and the GSE25055 value (-0.097), so GSE25065 neither confirmed nor excluded a decrease of that size.

Recorded subtype structure is compatible with much of DIPK's agreement with DLDA30 labels

Because a high AUC against DLDA30 labels was not specific to DIPK in GSE25055, we examined which patient pairs and recorded patient groups underlie DIPK's agreement with these labels. Of the 22,385 pairs of one DLDA30-positive and one DLDA30-negative patient, 13,816 (61.7%) compare patients with the same observed pathology (Fig. 4C). Among the 249 patients with observed RD, the saved scores separated DLDA30-pCR from DLDA30-RD patients with AUC 0.856 (patient-bootstrap 95% interval, 0.804 – 0.904), about as strongly as in the full cohort (0.852). Among the 57 with observed pCR, this AUC had a lower point estimate and was less precise: 0.678 (0.494 – 0.846). Conversely, AUC against observed pathology was 0.703 (0.518 – 0.881) within the DLDA30-RD group and 0.581 (0.466 – 0.694) within the DLDA30-pCR group (Fig. 4D; Table S15).

We next asked how high this AUC would be expected to be if DLDA30 labels were linked to the saved scores only through recorded patient groups. Permuting the labels among patients who shared observed pathology and recruitment source (Fig. 4A) gave an expected AUC of 0.575 (central 95% permutation range, 0.517 – 0.633 ; Fig. 4B), compared with the actual 0.852 . Holding ER fixed as well raised the expectation to 0.748 (range, 0.714 – 0.784), still clearly below the actual AUC (patient-bootstrap 95% interval for the actual-minus-expected difference, 0.056 – 0.145). Holding PAM50 instead of ER gave 0.826 (range, 0.802 – 0.851). The actual AUC lay just above this range, with 240 of 10,000 permutations at or above it, whereas the actual-minus-expected difference of 0.025 had a patient-bootstrap interval that just included zero (-0.0001 to 0.053). Holding both ER and PAM50 gave a similar expectation of 0.829 ,

with a range containing the actual AUC and a difference of 0.023 whose interval included zero (Table S14).

Much of the high AUC against DLDA30 labels is therefore compatible with recorded subtype structure. Finer grouping also fixes more labels: with pathology, recruitment source and PAM50 held fixed, only 188 patients were in groups containing both DLDA30 labels (166 with ER also fixed), so the expectation can rise because fewer labels can move, not only through shared subtype information. The comparison therefore does not measure how much of the agreement is due to subtype. The expectation with ER fixed is a permutation reference, not a corrected AUC; its value of 0.748 matches the AUC against observed pathology only by coincidence.

Response association persists, but added prediction remains uncertain

Leaving DLDA30 labels aside, we asked whether the saved scores remain associated with observed response within ER IHC or PAM50 groups and after adjustment for them. Discrimination was lower within background groups, most clearly within PAM50 subtypes: AUC 0.578 (0.471–0.681), compared with 0.748 across the same 306 patients. In the largest subtype, Basal (120 patients, 44 with pCR), it was 0.582 (0.463–0.691). Both intervals include 0.5, so discrimination within subtypes was not resolved above chance; each remaining subtype had only three or four pCR cases (Table S1), too few to establish differences between subtypes. Within ER groups, AUC was 0.657 (0.561–0.746) in 301 GSE25055 patients and 0.702 (0.571–0.813) in GSE32646, compared with 0.761 across all 115 GSE32646 patients. In logistic models, the saved score nevertheless remained positively associated with pCR: age/ER-adjusted OR 2.14 (Wald 95% interval, 1.45–3.16) and age/PAM50-adjusted OR 1.81 (1.21–2.71) in GSE25055, and age/ER-adjusted OR 2.10 (1.20–3.70) in GSE32646. The odds ratio is one adjusted slope fitted to all patients, whereas within-group AUC uses only same-group pairs, so a resolved association can coexist with within-subtype discrimination not resolved above chance (Fig. 5A–B; Table S2).

Held-out prediction asks a different question: whether adding the score improves prediction for patients not used in fitting. Against age plus ER, mean held-out AUC increased from 0.733 to 0.788 in GSE25055 and from 0.689 to 0.757 in GSE32646. Adding PAM50 to the GSE25055 baseline reduced the mean score increment from 0.056 to 0.035 (AUC 0.748 to 0.783). These increments are summarized by partition ranges (Tables S5 and S7); within cohorts, patient-resampling intervals were calculated only for the expanded comparisons below. Adding the score lowered Brier error in every GSE25055 partition, but in GSE32646 it slightly increased Brier error in one of ten partitions (maximum + 0.0004; Table S5), and the score-only model had a marginally lower mean Brier error than the age/ER/score model (0.157 vs 0.158; Table S4).

Against seven clinical variables, the score increments were smaller: mean AUC rose from 0.746 to 0.772 in GSE25055 (difference 0.025), from 0.703 to 0.756 in GSE32646 (0.053) and, with PAM50 also included in GSE25055, from 0.744 to 0.765 (0.021; Fig. 5C–D). Mean Brier error decreased from 0.136 to 0.130,

0.166 to 0.159 and 0.136 to 0.131, respectively. The expanded predictor sets did not yield uniformly better models. In GSE25055, the age/ER/score model had a higher mean AUC (0.788) than either expanded model with the score (0.772 and 0.765), so the smaller increments there partly reflect lower AUC of the larger models with the score, not only stronger clinical baselines. In GSE32646, the increment was smaller mainly because the baseline AUC was higher (0.703 with seven clinical variables vs 0.689 with age and ER). The age/ER/score models also had lower mean Brier errors than the expanded models with the score (Tables S4 and S9).

All three expanded comparisons improved AUC and Brier error in each of the ten partitions, but patient resampling with refitting gave AUC increment intervals of -0.006 to 0.075 in GSE25055, -0.019 to 0.152 in GSE32646 and -0.009 to 0.069 with PAM50 added in GSE25055. All three Brier increment intervals also included zero, as did the average-precision and log-loss contrasts (Fig. 5E–F; Table S10). Because the partitions reuse the same patients, this consistency does not show that the gain would hold in new patients.

Secondary models transferred between cohorts misestimated mean response

Better ranking does not guarantee accurate probabilities in a new cohort. We therefore fitted the age/ER and age/ER/score secondary models in each source cohort and applied them unchanged to the target cohort, in both directions. With GSE25055 as the source cohort, adding the saved score improved ranking in GSE32646: AUC increased from 0.677 to 0.787, a paired difference of 0.110 (95% interval, 0.002–0.195). Here and below, intervals resample both cohorts and refit the models in each draw (Fig. 6C–D and Table S8 also give target-only resampling). Its interval is the only added-AUC interval with model refitting in the study that excludes zero, and only narrowly; its comparator is age and ER alone, because the seven-variable models were not transferred. Despite this gain, the age/ER/score model assigned a mean pCR probability of 55.8% to a target cohort with 23.5% observed pCR, whereas the age/ER model assigned 16.6% (Fig. 6B). Brier error increased from 0.165 to 0.261, although the paired difference was uncertain (0.096; -0.016 to 0.259).

With GSE32646 as the source cohort, adding the score raised AUC in GSE25055 from 0.705 to 0.769, but this gain was uncertain (0.064; -0.036 to 0.113). The age/ER/score model assigned a mean pCR probability of 6.5% against 18.6% observed pCR, whereas the age/ER model assigned 23.9%; the age/ER/score model's Brier error was numerically higher, by 0.015 (-0.024 to 0.045). The age/ER/score model's mean-risk error was + 32.3 percentage points (8.7 to 50.5) in the forward direction (GSE25055 to GSE32646) and -12.1 percentage points (-18.5 to -1.6) in the reverse direction (Fig. 6B; Table S8).

Saved scores were higher on average in GSE32646 (mean 2.965, SD 0.328) than in the age/ER-eligible GSE25055 patients (mean 2.244, SD 0.217; Fig. 6A). Because each score-based model kept its source-cohort score standardization and assigned higher pCR probabilities to higher scores, the generally higher GSE32646 scores raised predicted probabilities in the forward direction and the lower GSE25055 scores

lowered them in the reverse direction. Score-only models showed the same pattern (mean predicted pCR probabilities 75.2% and 3.9%; Table S6), whereas the age/ER models erred less, and in the opposite direction, in each case. The score shift is thus consistent with the direction of both errors.

Discussion

The reanalyses show two ways in which a clinical validation can answer a question other than the one attached to it. DIPK's archived GSE25055 target is the output of the DLDA30 classifier, and TIL-30's released GSE41998 target is AC complete response, recorded after the first treatment stage rather than as final pathology. Matching the released labels to GEO fields establishes these differences directly, whether or not a performance difference is statistically resolved: against pathology, both AUC point estimates were lower, but only DIPK's reduction was resolved. These findings complement drug-screening benchmarks [2, 3] by tracing the clinical meaning of the outcomes used in released evaluations.

These discrepancies concern particular released evaluations, not whole methods. DIPK's GSE32646 and GSE20194 labels match observed pathology, and the TIL-30 article explicitly discloses survival-based surrogate targets in other cohorts, which were not evaluated here. The same GSE25055 label choice appears outside DIPK: VAEN's public GSE25055 evaluation code [15] uses the DLDA30 prediction field to define its response groups, but VAEN's patient-level predictions were not recovered, so its performance against observed pathology remains unknown.

Patient tracing narrows the evidence in a different way. Duplicate profiles are a known hazard in public expression databases [6, 7], and the earlier RPS and BA100 chemotherapy-response predictor studies removed duplicate samples from these breast-cancer data [5, 16]; the published RPS curation corroborates most of the present links. Here, the links bear directly on released clinical evaluations. In DIPK, 188 of the 248 GSE20194 patients link to GSE25055, so its AUCs against observed pathology in the three cohorts (0.748, 0.761 and 0.757) are not three independent confirmations: they rest on two cohorts without identified overlap plus at most 60 further GSE20194 patient identifiers. In TIL-30, 60 of 71 GSE20194 validation records link to 57 of these 188 patients, whose GSE25055 records are in its training input. The same duplication thus has different consequences: it reduces distinct support when the two series serve as separate evaluation cohorts, and compromises training-validation separation when one supplies training data and the other validation data. Neither overlap counts nor AUC comparisons between patient subsets show how much, if at all, reuse changed performance.

The correctly labeled SLM-HRD reference shows that DIPK's high agreement with DLDA30 labels is not specific to DIPK: in GSE25055, this separately derived signature reached almost the same AUC against DLDA30 labels and lost a similar amount when the target changed to observed pathology. Against observed pathology, DIPK's AUC was not resolved as different from SLM-HRD's or from that of the binary DLDA30 call, and in GSE25065 the SLM-HRD difference was uncertain, so the pattern is established only in GSE25055. Holding PAM50 subtype fixed brought a permutation reference close to DIPK's actual AUC

against DLDA30 labels, although finer strata also leave fewer labels free to move. These results are compatible with shared subtype-associated expression but do not identify what the network learned.

For DIPK, the association between saved scores and observed response persists after outcome correction and after adjustment for age and ER or PAM50 subtype, although in GSE25055 discrimination within PAM50 subtypes was not resolved above chance. Association does not establish added prediction. Earlier studies compared genomic predictors with clinical information, such as a clinical nomogram with DLDA30 [4] and, in ER-negative patients, scores from the integral genomic signature (iGenSig) drug-response model with tumor stage, PR status, menopausal status and age [17]. Here, adding the score to age and ER raised mean held-out AUC in both cohorts, but these comparisons lack patient-resampling intervals. Against seven clinical variables, the mean gains were smaller and every patient-resampling interval included zero, so added prediction beyond these variables remains uncertain.

Calibration is a separate question from discrimination [18]. When secondary models fitted in one cohort were applied unchanged to the other, adding the score narrowly improved discrimination over age and ER from GSE25055 to GSE32646, yet the model with the score assigned a mean pCR probability of 55.8% to a cohort with 23.5% observed pCR; in the reverse direction it assigned 6.5% against 18.6%. Both mean-risk errors were resolved, whereas the Brier-score differences were not, so a higher AUC did not ensure accurate absolute probabilities. The errors matched the direction of the shift in saved scores between cohorts, but population, platform and processing differences prevent attributing them to one cause.

For clinical reuse, an evaluation therefore needs a traceable outcome field and treatment stage, a defensible count of distinct patients, and a comparison that addresses the intended predictive claim. The first two determine what was tested and on whom; the third distinguishes response association from added prediction. Judged against these requirements, the released records support narrower conclusions than their reported validation results suggest.

Limitations

The two studies were selected sequentially through targeted examination of public inputs, so they cannot estimate how often such errors occur across clinical prediction research. All comparisons between targets condition on saved scores and exclude the original training and selection uncertainty, and no new-cohort DIPK test was possible because rerunning DIPK did not reproduce its saved predictions within tolerance. The TIL-30 analysis covers two released evaluation sets, not its complete pipeline or other analyses.

Identity and outcome claims rest on the deposited records and metadata. Identifier matching and three raw-array spot checks do not exhaustively screen for duplicate profiles, so the 60 unlinked DIPK patient identifiers and the 11 unlinked TIL-30 records may still include repeated patients. The TIL-30 links establish reuse but not which training records each saved model was fitted on.

The released-score comparisons describe specific scores, not learning algorithms. SLM-HRD differs from DIPK in biological derivation and preprocessing, and its GSE25065 analysis is evidence for that signature, not new DIPK validation. Because published duplicate-exclusion flags lack partner identifiers, both GSE25065 sets may still contain patients shared with other studies.

The background and conditional analyses are exploratory. PAM50 may carry information relevant to both observed pathology and DLDA30 labels, so we do not treat it as a nuisance variable to remove.

Conditional AUCs use different patient subsets, so their differences do not identify a causal contribution of subtype. Several groups have few pCR events: within recruitment sources, I-SPY contributes only 14 to the label comparison and 13 to the clinical comparison. MDACC and I-SPY samples also share archived processing, and I-SPY holds all missing PR, HER2 and grade values and all three records that the source study recorded as HER2-positive (S1 Text), so these within-source analyses are sensitivity checks, not validation in independently processed data. Because patients received combination regimens, no response analysis can isolate a paclitaxel-specific effect.

The secondary models use specified, unoptimized predictor forms, imputation rules and penalties; they are not established clinical prediction tools. Patient-resampling intervals condition on the original cross-validation partitions and saved DIPK scores, and their 95% coverage in these small cohorts is unverified. Transfer resampling includes uncertainty from fitting the secondary models but not from developing DIPK, and because receptor definitions, clinical practice and score scales differ across cohorts, the transfer analysis tests reuse of unchanged secondary models, not prospective clinical validation.

Conclusions

Tracing outcomes to GEO fields showed that DIPK's GSE25055 evaluation used another classifier's predictions as its target and that TIL-30's GSE41998 evaluation used AC complete response, recorded after the first treatment stage. With records and scores fixed, AUC point estimates were lower against pathology in both evaluations, but only DIPK's reduction was resolved. Patients shared between GSE20194 and GSE25055 reduced independent support across DIPK's evaluation cohorts and crossed TIL-30's training–validation split. DIPK's saved scores remained associated with observed response, but their added prediction beyond seven clinical variables was unresolved within cohorts. Score-based logistic models fitted here narrowly improved ranking over age and ER in one transfer direction but misestimated mean pCR rates in both. In these cases, the released validation results did not by themselves establish the outcome's source and treatment stage, patient independence or, for DIPK, added prediction over clinical variables; each needs to be checked directly.

Declarations

Data availability

The source clinical data are publicly available in GEO under accession codes GSE25055, GSE32646, GSE20194, GSE25065, and GSE41998. Saved DIPK predictions are available through the original repository and its linked archive. Released SLM-HRD scores are provided in Source Data Fig. 4 of the SYLVER study [9]. TIL-30 inputs and saved outputs are available in the official repository at the reviewed commit. Processed data, numerical results and analysis code are available on GitHub and archived as version 1.0.1 in Zenodo (doi:10.5281/zenodo.23243329). Third-party scores, labels and lists are not redistributed; the archive includes a script that downloads them from their sources and verifies each file.

Code availability

Analysis code is available on GitHub and archived in Zenodo (doi:10.5281/zenodo.23243329).

Supporting information

S1 Text. Supplementary methods and complete results. Includes Tables S1–S21, Supplementary Figure S1, the claim-to-evidence table and source references.

Author contributions

Yunhao Jiang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

During the preparation of this work, the author used Claude (Anthropic) for spelling, grammar and language editing; explaining code and error messages; checking numbers and code logic; and searching the literature and public data.

References

1. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A et al (2011) A genomic predictor of response and survival following taxane–anthracycline chemotherapy for invasive breast cancer. *JAMA* 305(18):1873–1881. 10.1001/jama.2011.593
2. Partin A, Vasanthakumari P, Narykov O, Wilke A, Koussa N, Jones SE et al (2026) Benchmarking community drug response prediction models: datasets, models, tools, and metrics for cross-dataset generalization analysis. *Brief Bioinform* 27(1):bbaf667. 10.1093/bib/bbaf667
3. Bennett J, Iversen P, Picciani M, Wilhelm M, Baum K, List M (2026) Critical evaluation of drug response prediction models with DrEval. *Nat Commun* 17:4238. 10.1038/s41467-026-72903-w
4. Lee JK, Coutant C, Kim YC, Qi Y, Theodorescu D, Symmans WF et al (2010) Prospective comparison of clinical and genomic multivariate predictors of response to neoadjuvant chemotherapy in breast

- cancer. Clin Cancer Res 16(2):711–718. 10.1158/1078-0432.CCR-09-2247
5. Pitroda SP, Bao R, Andrade J, Weichselbaum RR, Connell PP (2017) Low recombination proficiency score (RPS) predicts heightened sensitivity to DNA-damaging chemotherapy in breast cancer. Clin Cancer Res 23(15):4493–4500. 10.1158/1078-0432.CCR-16-2845
 6. Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M (2016) The doppelgänger effect: hidden duplicates in databases of transcriptome profiles. J Natl Cancer Inst 108(11):djw146. 10.1093/jnci/djw146
 7. Wang LR, Choy XY, Goh WWB (2022) Doppelgänger spotting in biomedical gene expression data. iScience 25(8):104788. 10.1016/j.isci.2022.104788
 8. Li P, Jiang Z, Liu T, Liu X, Qiao H, Yao X (2024) Improving drug response prediction via integrating gene relationships with deep learning. Brief Bioinform 25(3):bbae153. 10.1093/bib/bbae153
 9. Haider S, Brough R, Madera S, Iacovacci J, Gulati A, Wicks A et al (2025) The transcriptomic architecture of common cancers reflects synthetic lethal interactions. Nat Genet 57:522–529. 10.1038/s41588-025-02108-2
 10. Banerjee S, Tiwari VK, Raman K, Inayatullah M (2026) Tumor-infiltrating lymphocytes display prognostic signatures associated with chemotherapy response in TNBC patients. iScience 29(6):116295. 10.1016/j.isci.2026.116295
 11. GEO GSE20194 (2026) MAQC-II Project: human breast cancer data set. Series record and official sample-information workbook, Sample_Info and CodeBook sheets. Retrieved 26
 12. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM et al (2013) Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or paclitaxel in early-stage breast cancer. Clin Cancer Res 19(6):1587–1595. 10.1158/1078-0432.CCR-12-1359
 13. Menon A, van Rooyen B, Ong CS, Williamson RC (2015) Learning from corrupted binary labels via class-probability estimation. Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research. ;37:125–134. Article and full text
 14. Bates S, Hastie T, Tibshirani R (2024) Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc 119(546):1434–1445. 10.1080/01621459.2023.2197686
 15. Jia P, Hu R, Pei G, Dai Y, Wang YY, Zhao Z (2021) Deep generative neural network for accurate drug response imputation. Nat Commun 12:1740. 10.1038/s41467-021-21997-5
 16. Fournier MV, Goodwin EC, Chen J, Obenaus JC, Tannenbaum SH, Brufsky AM (2019) A Predictor of Pathological Complete Response to Neoadjuvant Chemotherapy Stratifies Triple Negative Breast Cancer Patients with High Risk of Recurrence. Sci Rep 9:14863. 10.1038/s41598-019-51335-1
 17. Wang XS, Lee S, Zhang H, Tang G, Wang Y (2022) An integral genomic signature approach for tailored cancer therapy using genome-wide sequencing data. Nat Commun 13:2936. 10.1038/s41467-022-30449-7
 18. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: the Achilles heel of predictive analytics. BMC Med 17:230. 10.1186/s12916-019-1466-7

Figures

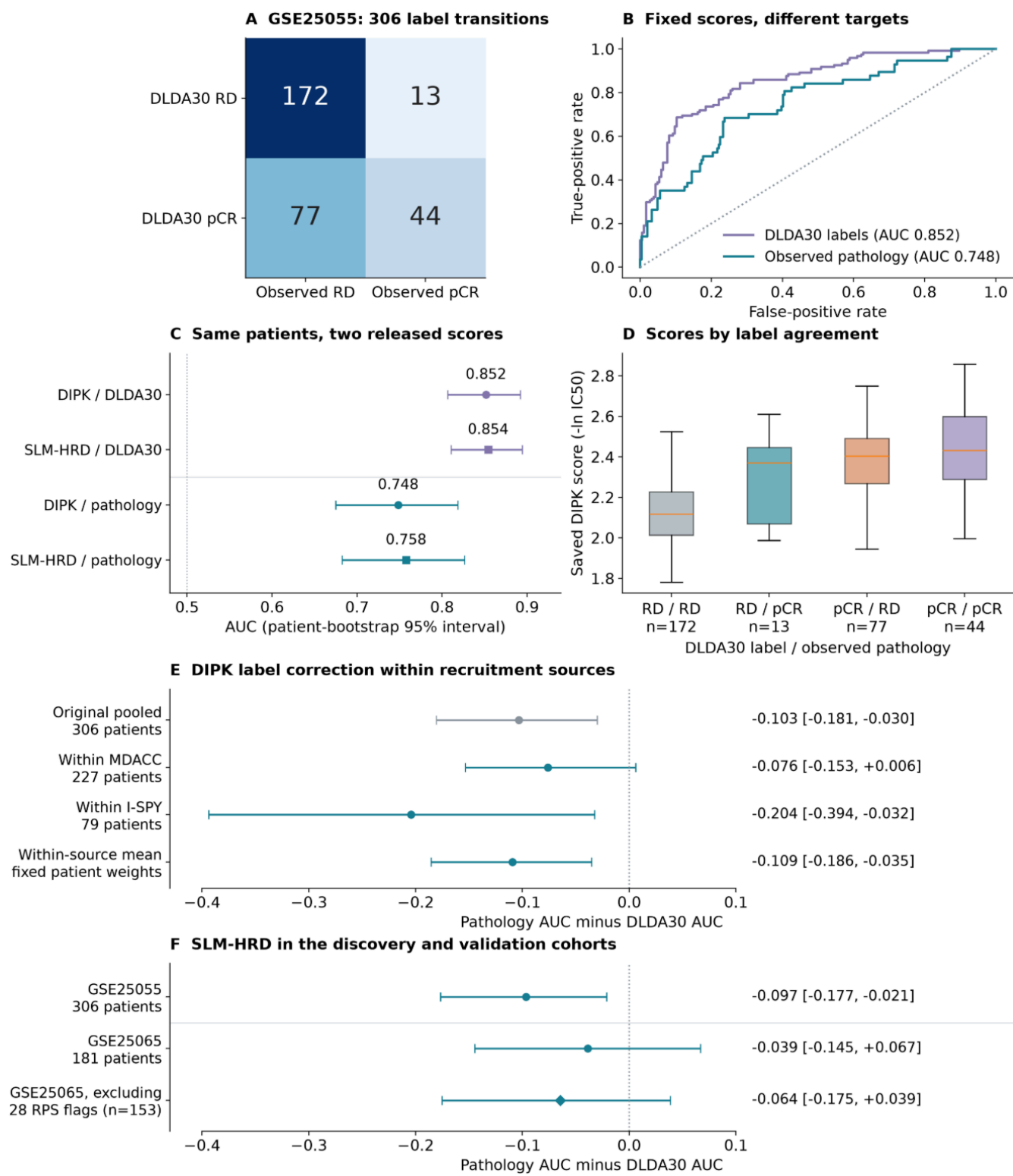


Figure 1

Unchanged scores evaluated against DLDA30 labels and observed pathology. A–E use the 306 GSE25055 patients with known pathology. A, archived DLDA30 labels versus observed pathology. B, receiver operating characteristic (ROC) curves of the unchanged DIPK scores against the two targets,

DLDA30 labels and observed pathology. C, AUCs of DIPK and released SLM-HRD scores against both targets, with percentile 95% intervals from the 5,000-draw patient bootstrap shared by the two scores, so DIPK's intervals differ slightly from the 2,000-draw intervals in the text. SLM-HRD's source-study labels match observed pathology. D, saved DIPK scores for each combination of archived DLDA30 label and observed outcome; boxes show quartiles and medians, whiskers extend to 1.5 interquartile ranges, and outliers are not shown. E, DIPK pathology-minus-DLDA30 AUC differences with the 2,000-draw intervals reported in the text; the last three rows resample patients within each recruitment source, and the last row averages the MDACC and I-SPY AUCs with fixed patient-count weights. F, the same difference for SLM-HRD in GSE25055 and GSE25065, the discovery and validation cohorts of the source study [1], with intervals from each cohort's 5,000-draw patient bootstrap. The diamond excludes the 28 GSE25065 accessions on the RPS duplicate-exclusion list [5], leaving 153 patients; its interval uses the same 181-patient draws. E–F share an axis; negative values indicate lower AUC against pathology.

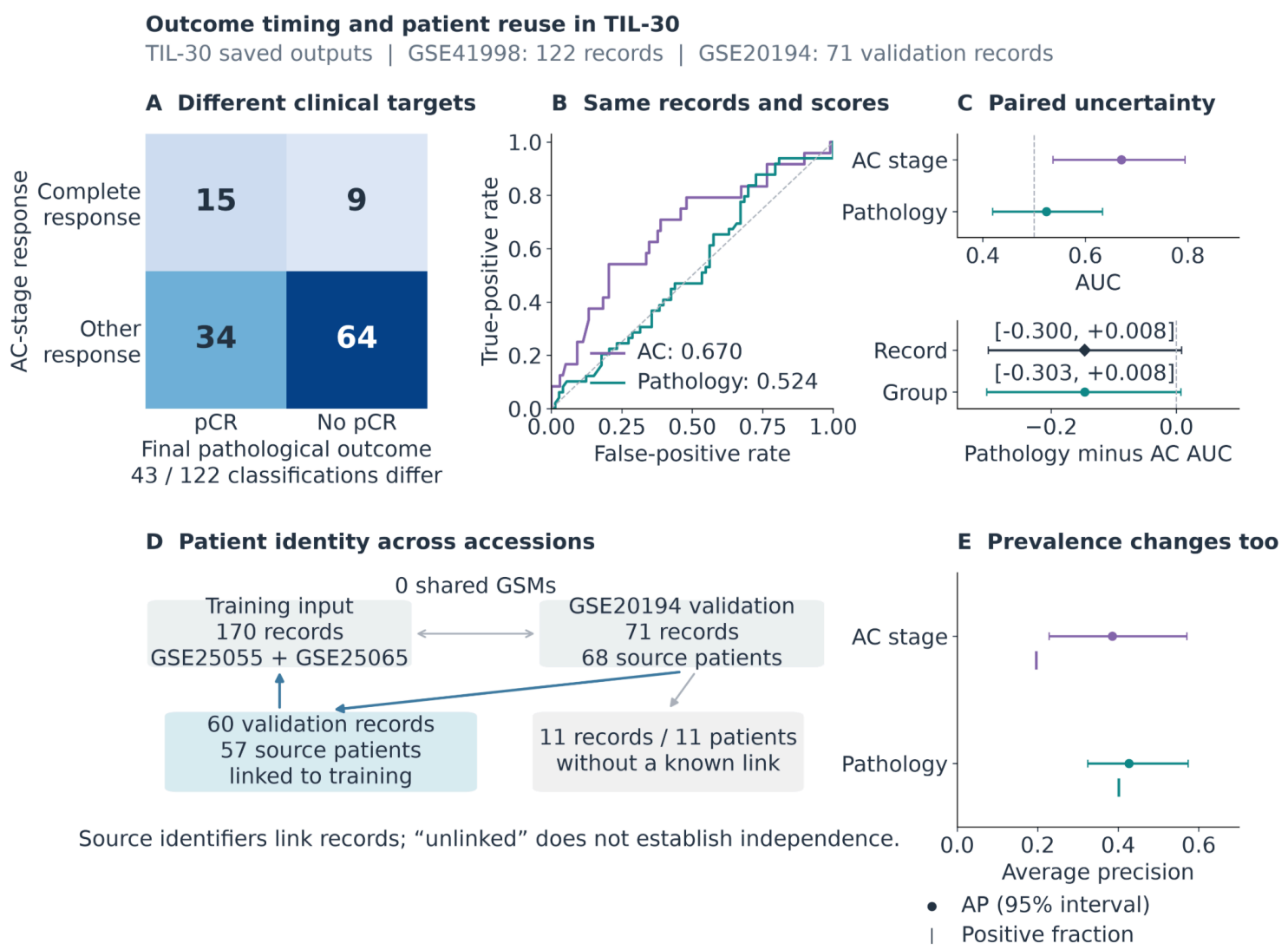


Figure 2

Outcome timing and patient reuse in the released TIL-30 evaluation. A–C and E use the same 122 GSE41998 records and unchanged scores. A, AC complete response (rows, “AC-stage response”) versus

final pCR (columns); “Other response” combines partial response, stable disease and progressive disease. B, ROC curves against AC complete response (“AC”; “AC stage” in C and E) and final pathology. C, upper: AUCs with percentile 95% intervals from 5,000 record draws; lower: the paired pathology-minus-AC AUC difference when resampling records (“Record”) or the 118 candidate specimen groups defined by shared specimen-name prefixes (“Group”), with 5,000 draws each and the same record-weighted point estimate. D, the 170-record training input (GSE25055 and GSE25065) and the 71-record GSE20194 validation input (68 source patients) share no GSM accessions, but source patient identifiers link 60 validation records to 57 training patients. E, average precision with record-bootstrap 95% intervals; vertical ticks mark the positive fraction, the approximate average precision of a random ranking. Table S20 gives the complete estimates.

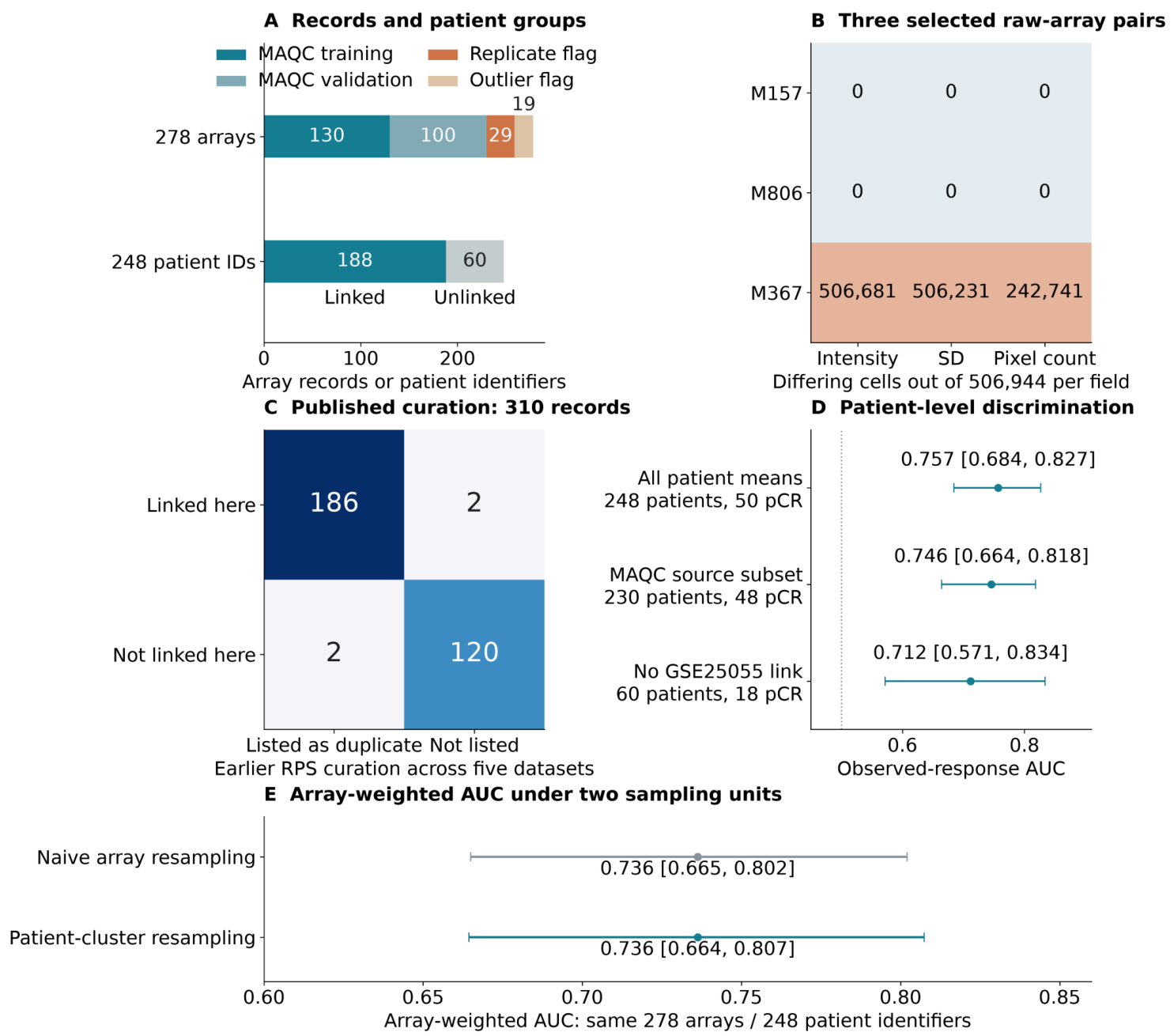


Figure 3

Shared patients, raw-array checks and repeated arrays in DIPK's GSE20194 cohort. A, the source study's MAQC categories for the 278 arrays (one of the 19 outlier arrays is also a replicate) and links from the 248 patient identifiers to GSE25055. B, number of differing values out of 506,944 per field in three raw-array pairs selected by a fixed rule, after conversion to float32 precision; SD, stored standard deviation. M367 remains a link on the patient identifier alone. C, GSE20194 links found here versus the earlier RPS duplicate-exclusion list [5] across all 310 GSE25055 accessions. D, AUCs of patient-level scores with percentile 95% intervals from 2,000 patient draws; rows use different eligibility rules. E, the same array-weighted AUC under naive array resampling and patient-cluster resampling, with percentile 95% intervals from 2,000 draws each; Supplementary Figure S1 in S1 Text gives the full comparison.

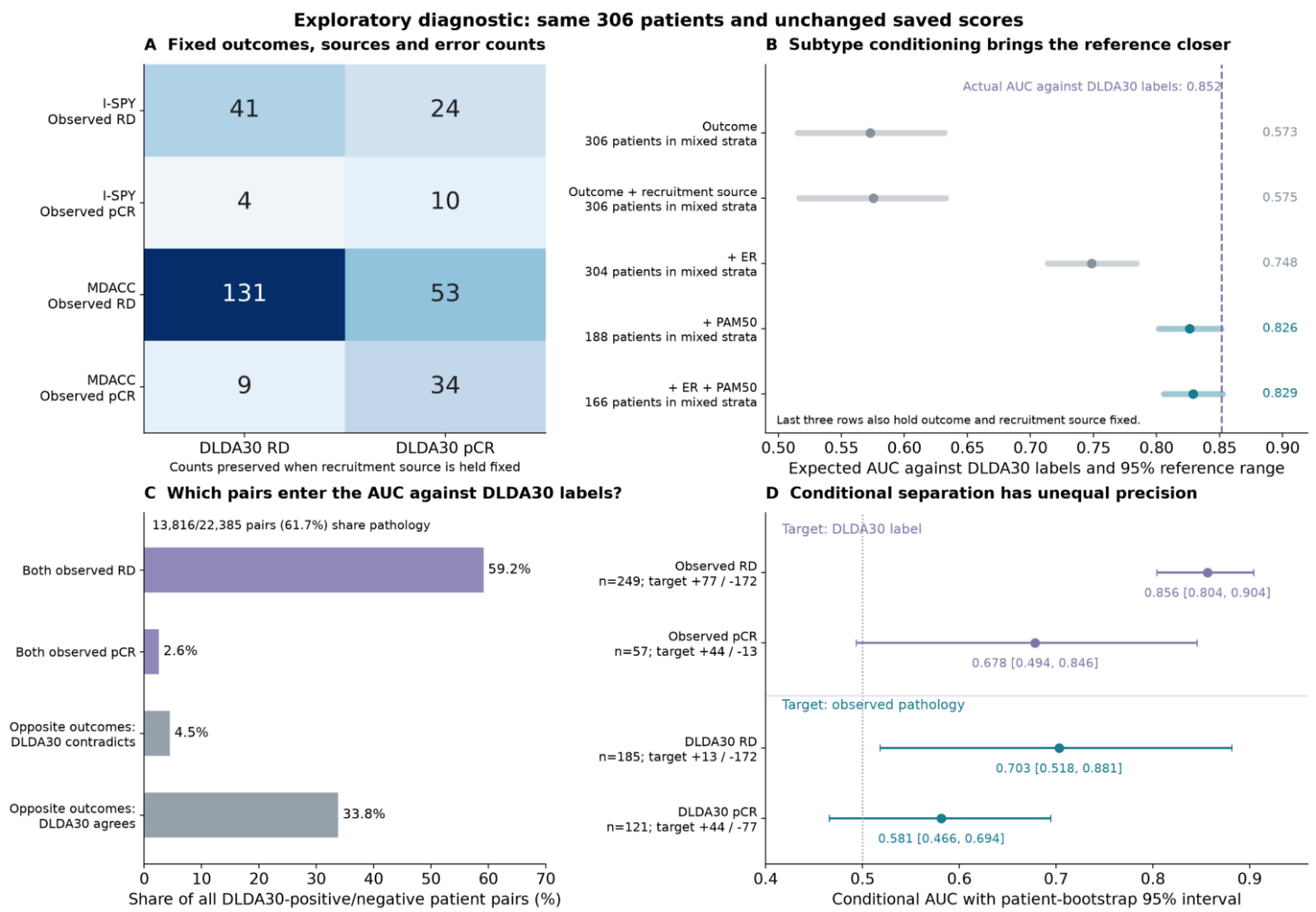


Figure 4

Agreement with DLDA30 labels, molecular background and conditional discrimination. All panels use the same 306 GSE25055 patients and unchanged saved scores. A, observed outcome by DLDA30 label within each recruitment source; permutations holding recruitment source fixed preserve these counts. B, exact expected AUCs against DLDA30 labels under each permutation scheme, with central 95% ranges from 10,000 permutations per scheme (reference ranges, not confidence intervals); the dashed line marks the actual value. Row labels name the categories held fixed and the number of patients in mixed

strata (groups containing both DLDA30 labels); other labels stay fixed. C, pairs of one DLDA30-positive and one DLDA30-negative patient, by the two patients' observed outcomes; "share pathology" means both patients had the same observed outcome. For pairs with opposite outcomes, "DLDA30 agrees" means that the DLDA30-positive patient had observed pCR, and "DLDA30 contradicts" that this patient had observed RD. D, the four conditional AUCs with percentile intervals from 2,000 patient draws; purple estimates use DLDA30 labels as the target and teal estimates observed pathology, and positive means pCR for the indicated target. Tables S14–S16 give reference sampling intervals, recruitment-source-specific conditional estimates and pair decompositions under both targets.

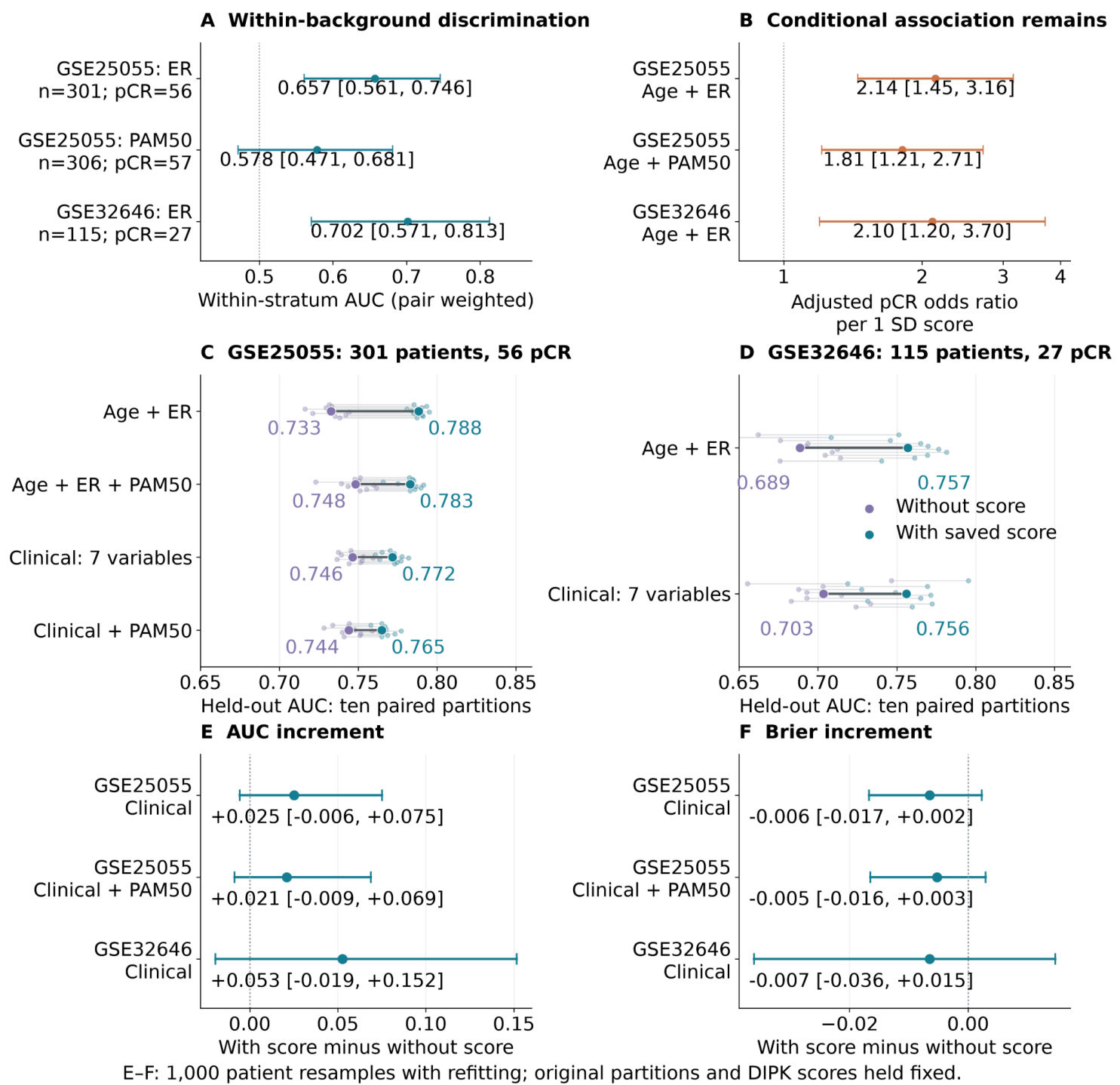


Figure 5

Clinical background, held-out prediction and uncertainty in incremental value. A, pair-weighted AUC within ER or PAM50 strata, with percentile 95% intervals from 2,000 patient draws. B, exploratory age-adjusted odds ratios per within-analysis score SD, additionally adjusted for the indicated background variable, with Wald 95% intervals. A–B use 301 GSE25055 patients (56 pCR) for ER and 306 (57 pCR) for PAM50; C–F use the same 301 GSE25055 patients; GSE32646 analyses use 115 patients (27 pCR) throughout. C–D, paired held-out AUCs without and with the saved score; small points show ten common partitions and large points their means. The seven clinical variables are age, ER, PR, HER2 status as recorded in each source study, clinical T stage, nodal positivity and histological grade. E–F, mean paired increments in AUC and Brier score with 2.5th–97.5th percentile intervals from 1,000 patient resamples with refitting. Positive AUC and negative Brier increments favor adding the score.

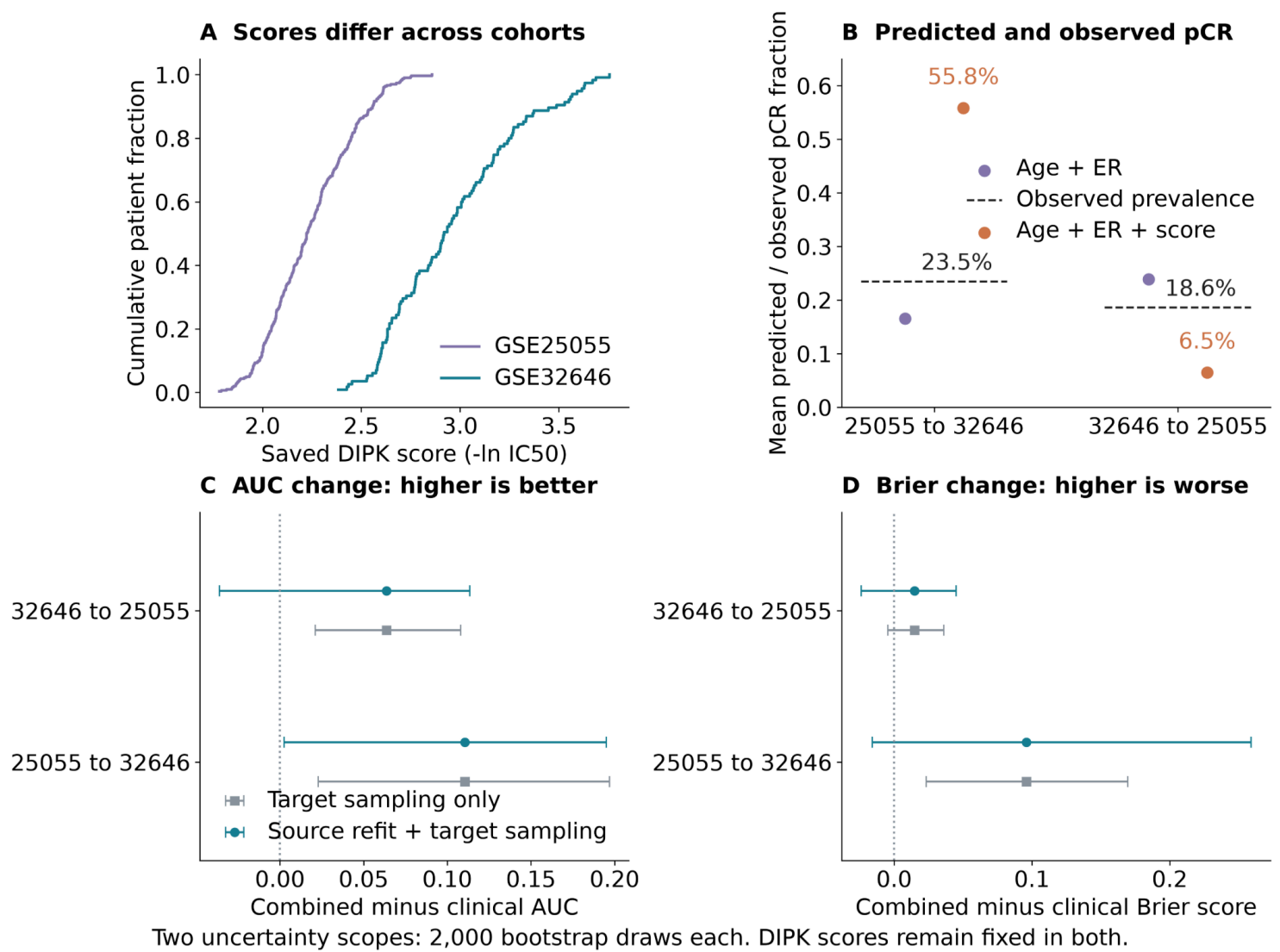


Figure 6

Cohort transfer separates ranking from probability accuracy. The clinical model uses age and ER; the combined model adds the saved DIPK score. Probabilities come from these secondary logistic models

fitted in this study, not from DIPK, which outputs predicted log IC50. A, empirical cumulative distributions of saved DIPK scores in the age/ER-eligible GSE25055 and GSE32646 patients (301 and 115). B, mean predicted pCR probabilities when models fitted in the source cohort are applied unchanged to the target cohort; dashed lines (“Observed prevalence”) mark the target cohort’s observed pCR fraction. Direction labels such as “25055 to 32646” name the source cohort first and the target cohort second. C–D, combined-minus-clinical differences in AUC and Brier score, with percentile 95% intervals from 2,000 draws per analysis. “Target sampling only” resamples target-cohort patients with the source-cohort fits held fixed; “Source refit + target sampling” also resamples source-cohort patients and refits the models in each draw. Positive AUC differences favor adding the score; positive Brier differences indicate worse probability accuracy.

Supplementary Files

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

- [DIPKsupplement.docx](#)