

Cross-Cohort Reliability of Machine Learning for Breast Cancer Survival Prediction

NADIR Younes^{1*}, RACHDI Mohamed², AMHAIMAR Lahcen³,
BAKHOUYI Abdellah⁴, KHALIDI Abderrahim⁵, AZOUAZI Mohamed⁶

M2S2I Laboratory-National Higher School of Art and Design (ENSAD),
Hassan II University of Casablanca, Casablanca, Morocco^{1, 2, 3, 4, 5}

LTIM / Laboratory of Information Technology and Modeling-Faculty of Sciences Ben M'Sik (FSBM),
Hassan II University of Casablanca, Casablanca, Morocco⁶

Abstract—Transportability limitations can affect discrimination, probability calibration, and uncertainty estimates even when machine-learning pipelines use the same input variables. This study evaluated 5-year breast cancer survival models across METABRIC and TCGA-BRCA using age and ER, PR, and HER2 status. Logistic regression, radial-basis-function support vector machine, random forest, XGBoost, and soft voting underwent bidirectional external validation. The METABRIC soft-voting model was further assessed with source-only Platt scaling, split conformal prediction, selective prediction by singleton sets, cohort classification, and exploratory few-shot target adaptation. The principal contribution is an integrated reliability framework. METABRIC-to-TCGA ROC-AUC decreased numerically from 0.732 to 0.643, an observed difference of approximately 0.088 with a 95 percent confidence interval from minus 0.015 to 0.196. TCGA-to-METABRIC ROC-AUC decreased numerically from 0.674 to 0.586, also by approximately 0.088, with a 95 percent confidence interval from minus 0.127 to 0.282. Neither interval excluded zero. Source Platt scaling improved the observed internal Brier score and expected calibration error but produced worse observed values for both metrics externally. TCGA empirical conformal coverage was 81.8 percent versus nominal 90 percent. Exploratory target recalibration improved mean locked-test Brier score and expected calibration error at larger target sample sizes, whereas updating only the conformal threshold did not restore nominal mean coverage. These findings support separate evaluation of ranking, probability accuracy, and uncertainty under cohort shift and do not establish model superiority, clinical usefulness, or deployment readiness.

Keywords—Machine learning; transportability; external validation; probability calibration; conformal prediction; selective prediction; cohort shift; few-shot recalibration; survival prediction

I. INTRODUCTION

Five-year overall survival is a clinically interpretable endpoint, and breast cancer contributes substantially to global cancer incidence [1]. Prognosis and the data used to estimate it vary across patient populations and data-generating settings. Machine-learning models can combine clinical predictors through linear, kernel-based, tree-based, and ensemble relationships, yet apparent performance within a development cohort does not establish performance in a distinct population [2]. Differences in case mix, variable definitions, missingness, care pathways, and follow-up can alter discrimination and

probability calibration even when the same predictor names are available.

External validation is therefore central to evaluating model transportability [3,4]. Discrimination measures whether a model ranks patients with different outcomes; calibration measures agreement between predicted probabilities and observed outcome frequencies [4,5]. A transported model may retain ranking but yield inaccurate probabilities. A calibration map fitted in the source cohort may not improve calibration in the target cohort.

Uncertainty methods introduce a second transportability question. Split conformal classification can produce patient-specific prediction sets with finite-sample marginal coverage under exchangeability [6]. That guarantee does not automatically extend to a target cohort under distribution shift [7]. Empirical coverage and abstention behavior must consequently be evaluated in the target cohort rather than assumed. Limited labeled target data may permit probability recalibration or re-estimation of a conformal threshold, but small adaptation samples can introduce instability.

These issues were investigated in two clinical breast cancer cohorts, METABRIC and the TCGA-BRCA publication cohort. The study had four objectives: quantify observed internal-to-external performance changes in both transport directions; assess whether source-only Platt scaling transported from METABRIC to TCGA; evaluate source split-conformal coverage and conformal abstention internally and externally; and explore whether few-shot target adaptation improved probability calibration or restored nominal conformal coverage on a locked TCGA test set. The study design and protected information boundaries are summarized in Fig. 1 in Section III.

The principal contribution of this work is an integrated cross-cohort machine-learning reliability framework. It examines discrimination transportability, probability calibration, conformal coverage, selective prediction, cohort shift, and limited-label target adaptation under protected external evaluation.

Specifically, the study contributes: 1) bidirectional METABRIC↔TCGA external validation of five machine-learning models; 2) joint assessment of discrimination, probability calibration, and conformal uncertainty rather than reliance on accuracy or ROC-AUC alone; 3) evaluation of

*Corresponding author.

whether source-derived Platt calibration and empirical conformal coverage remain reliable after cohort transport; and 4) exploratory few-shot target adaptation with a permanently locked target test, separating probability recalibration from conformal-threshold adaptation.

The remainder of this study is organized as follows. Section II reviews related work on external validation, calibration, dataset shift, and conformal prediction. Section III describes the cohorts, endpoint, protected modeling workflow, calibration and conformal procedures, and statistical analysis. Section IV presents cross-cohort performance, calibration, uncertainty, domain-shift, and exploratory few-shot results. Section V discusses the findings and outlines future work; Section VI summarizes limitations, and Section VII concludes the study.

II. RELATED WORK

METABRIC and TCGA-BRCA have both supported research on breast cancer biology and outcomes, but they differ in study design, composition, and available clinical fields [8,9]. Similar variable names cannot be assumed to have identical meanings across the two studies. A systematic review of 31 machine-learning studies of 5-year breast-cancer survival found that all used internal validation, but only one used external validation and only one assessed calibration [2]. Later work externally evaluated 87 breast-cancer prediction models [10] and compared statistical and machine-learning approaches through internal-external validation [11]. These studies support assessing discrimination and calibration in new populations. Related image-based work combined CLAHE-enhanced mammograms with EfficientNetV2S for breast-cancer classification [12]. A systematic review also examined deep-learning methods for breast-cancer classification [13]. These image-based diagnostic tasks differ from transporting clinical survival probabilities across cohorts.

Clinical prediction guidance recommends protected separation of development, tuning, calibration, and final evaluation, together with uncertainty-aware reporting [3,4,14]. Dataset shift can produce performance changes after deployment. Recent clinical work studied stable prediction under temporal electronic-health-record shift [15]. Cohort-membership classifiers provide a complementary label-free measure of multivariable shift [16]. Covariate differences alone do not establish that transported probabilities or uncertainty sets remain reliable.

Calibration can change even when ranking is retained [5]. Platt scaling uses a parsimonious logistic score transformation [17]. Model-updating research includes slope-and-intercept recalibration with attention to sample size [18], broader guidance on recalibration and revision [19], target-time correction of calibration shift [20], and dynamic updating under changing clinical environments [21]. This literature supports evaluating limited-label updating. It does not justify using the target test set to choose an updating method or sample size.

Conformal prediction provides set-valued uncertainty under exchangeability, and weighted approaches address specified

covariate-shift settings [6,7]. More recent work extends conformal inference beyond exchangeability [22] and develops doubly robust calibration of prediction sets under covariate shift [23]. Selective-prediction methods separately formalize abstention or rejection [24,25]. The present study evaluates conventional source split-conformal sets and singleton-based selective prediction empirically after transport; it does not claim the guarantees of these shift-robust extensions.

Existing studies address different parts of this reliability problem. Breast-cancer external-validation studies commonly assess discrimination and calibration, whereas methodological work examines shift, conformal prediction, selective prediction, or model updating [2,10,11,19,22,23,25]. This study evaluates those components in one protected patient-level cross-cohort workflow.

III. MATERIALS AND METHODS

A. Study Design and Cohorts

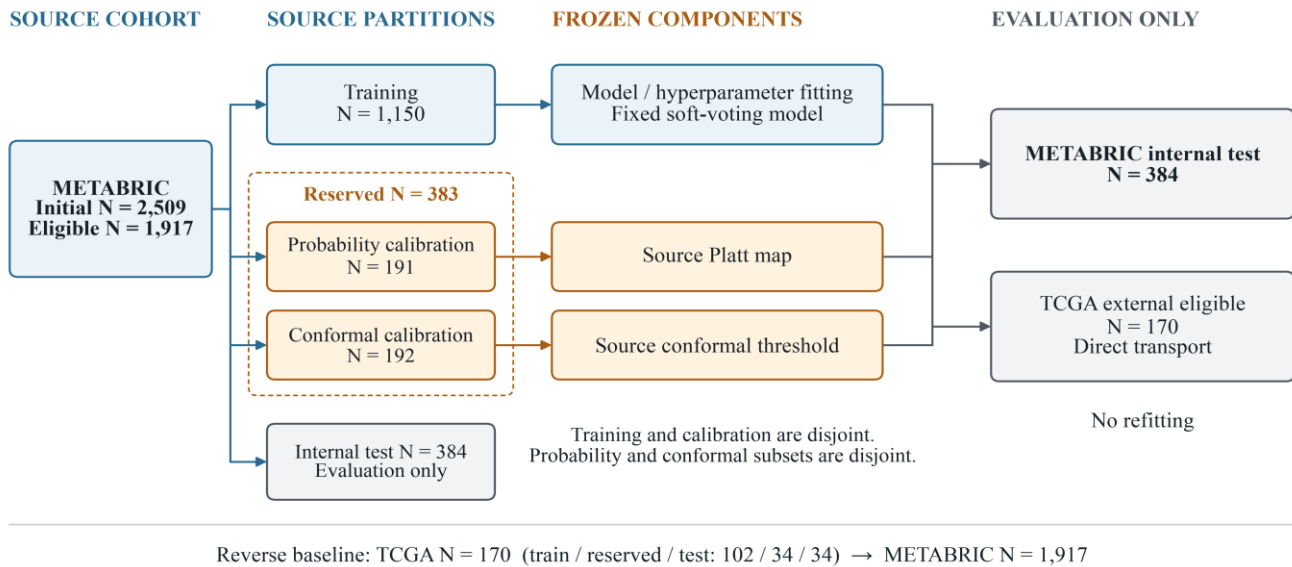
This was a retrospective secondary analysis of clinical data from METABRIC (*brca_metabric*) and TCGA-BRCA (*brca_tcga_pub*). Only patient-level and sample-level clinical data obtained through the official cBioPortal REST API were used; no mutation, expression, copy-number, imaging, or complete study-archive data were included. The unit of analysis was the patient. Stable patient identifiers were required and checked for duplication before modeling.

Each cohort served once as the source and once as the external cohort. The bidirectional baseline analyses transported METABRIC models to all eligible TCGA patients and TCGA models to all eligible METABRIC patients. Source calibration, split-conformal, and domain-shift analyses used the fixed METABRIC soft-voting model with TCGA as the external cohort. The few-shot analysis was exploratory and post hoc. Fig. 1 shows the protected workflows. Panel A separates source training, disjoint probability and conformal calibration, and evaluation without refitting; it also shows the complementary TCGA-to-METABRIC baseline direction. Panel B shows separate target probability and conformal-threshold branches, each evaluated only on the locked TCGA test.

B. Five-Year Overall-Survival Outcome

Overall survival status was harmonized from *OS_STATUS*, with 0: LIVING mapped to living and 1: DECEASED mapped to deceased. Overall survival time was parsed from nonnegative *OS_MONTHS*. The binary endpoint used a 60-month horizon. Death on or before 60 months was assigned *y_5yr*=0; documented survival beyond 60 months was assigned *y_5yr*=1, irrespective of a later death or censoring event. Patients recorded as living or censored on or before 60 months were excluded because their 5-year outcome was not observed. Patients with missing, invalid, conflicting, or otherwise unsupported required outcome fields were also excluded. The resulting class proportions describe the outcome-eligible analytic samples and should not be interpreted as population survival estimates.

A Cross-cohort modeling and source reliability evaluation



B Exploratory target adaptation with a protected locked test

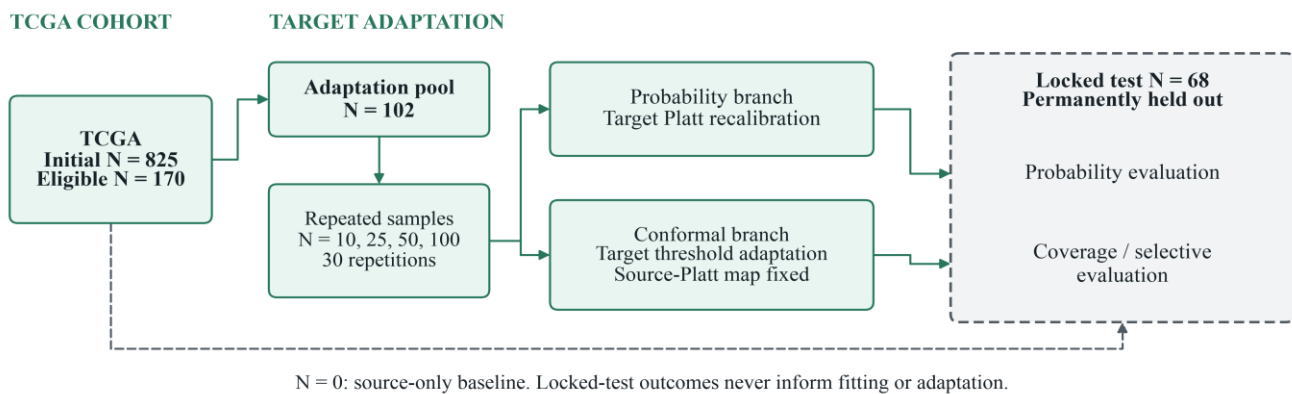


Fig. 1. Study design and protected information boundaries. (A) Bidirectional cross-cohort validation and detailed METABRIC-to-TCGA source calibration and conformal evaluation. (B) Exploratory few-shot target adaptation used a 102-patient TCGA adaptation pool; the 68-patient locked test was reserved for evaluation only.

C. Predictor Harmonization

Candidate shared clinical variables were audited before mapping. Age was parsed in years from AGE_AT_DIAGNOSIS in METABRIC and AGE in TCGA; nonnumeric values or values outside 0–120 years were treated as missing. ER_STATUS, PR_STATUS, and HER2_STATUS were mapped only when the observed value was exactly Positive or Negative. Unknown, equivocal, indeterminate, unavailable, unsupported, conflicting, or absent values were treated as missing.

Tumor stage was not included in primary modeling. METABRIC numeric stages 1–4 could be mapped to I–IV, but TCGA TUMOR_STAGE values represented T categories rather than supported overall stages. Consequently, harmonized stage was missing for all 170 outcome-eligible TCGA patients. The final predictor set was age at diagnosis, estrogen-receptor status, progesterone-receptor status, and HER2 status.

D. Source Partitions and Preprocessing

Random seed 42 was used for source modeling. Within each eligible source cohort, a stratified split allocated approximately 60% to training, 20% to a reserved calibration partition, and 20% to an untouched internal test. METABRIC contributed 1,150 training, 383 reserved calibration, and 384 internal-test patients. TCGA contributed 102, 34, and 34 patients, respectively. Hyperparameter selection used only the training split; the internal test and all external patients were not used for fitting or selection.

Preprocessing was encapsulated in the model pipeline. Age underwent median imputation and standardization. Receptor variables underwent most-frequent imputation followed by one-hot encoding with unknown categories ignored at transformation. Every preprocessing operation was fitted within the applicable training data, including separately within each cross-validation fold, and then applied without refitting to calibration and evaluation patients.

E. Models and Hyperparameter Selection

Five model families were evaluated: logistic regression, radial-basis-function support vector machine with probability output, random forest, XGBoost, and an equal-weight soft-voting ensemble of the four component models. Component hyperparameters were selected by five-fold stratified cross-validation within the source training split, using ROC-AUC as the optimization metric. The small grids varied logistic-regression regularization strength and class weighting; support-vector regularization, kernel scale, and class weighting; random-forest tree count, maximum depth, and minimum leaf size; and XGBoost tree count, depth, and learning rate. The soft-voting model averaged the four fitted component probabilities with weights 1:1:1:1. All stochastic estimators used seed 42.

F. Reproducibility Details

Exact five-fold source-training grids were as follows: logistic regression $C \in \{0.1, 1.0, 10.0\}$ and $\text{class_weight} \in \{\text{None}, \text{balanced}\}$; RBF-SVM $C \in \{0.5, 1.0, 2.0\}$, $\text{gamma} \in \{\text{scale}, 0.1\}$, and $\text{class_weight} \in \{\text{None}, \text{balanced}\}$; random forest $n_estimators \in \{100, 200\}$, $\text{max_depth} \in \{\text{None}, 5\}$, and $\text{min_samples_leaf} \in \{1, 5\}$; and XGBoost $n_estimators \in \{100, 200\}$, $\text{max_depth} \in \{2, 3\}$, and $\text{learning_rate} \in \{0.05, 0.1\}$. XGBoost subsample and colsample_bytree were fixed at 0.8. Selection used source-training ROC-AUC only.

Selected METABRIC-source settings were logistic regression $C=1.0$ with balanced weights; RBF-SVM $C=1.0$, $\text{gamma}=0.1$, balanced; random forest 100 trees, depth 5, minimum leaf 5; and XGBoost 100 trees, depth 2, learning rate 0.05. Selected TCGA-source settings were logistic regression $C=10.0$, balanced; RBF-SVM $C=2.0$, $\text{gamma}=\text{scale}$, balanced; random forest 200 trees, depth 5, minimum leaf 5; and XGBoost 100 trees, depth 2, learning rate 0.1. Each soft-voting model used the selected components with weights 1:1:1:1.

The verified environment was Python 3.11.9 with numpy 2.4.6, pandas 3.0.5, scipy 1.17.1, scikit-learn 1.9.0, xgboost 3.2.0, matplotlib 3.11.1, statsmodels 0.14.6, joblib 1.5.3, and requests 2.34.2. Age parsing, receptor mappings, training-only imputation, one-hot encoding, and scaling are specified above and implemented within fitted pipelines. The source Platt map recorded slope 0.907921 and intercept 0.470539; these are transformation coefficients, not external calibration-slope or calibration-intercept performance measures.

Clinical records for `brca_metabric` and `brca_tcg_a_pub` were retrieved on 9 August 2026 (UTC) from <https://www.cbiportal.org/api/> using study-specific clinical-attributes and clinical-data endpoints; no separate server API-version string was recorded. The reproducibility archive includes deterministic extraction and harmonization logic, aggregate results, and SHA-256 patient-manifest hashes rather than patient identifiers in the paper. The local repository had no recorded commit and no verified paper-specific public URL or DOI at camera-ready preparation.

G. Internal and External Validation

Each fitted source model was evaluated on its source internal test and directly on all outcome-eligible patients in the other cohort. No target outcome was used for preprocessing, hyperparameter selection, calibration, threshold choice, or

model selection. Metrics were ROC-AUC, precision-recall AUC, accuracy, balanced accuracy, sensitivity, specificity, F1 score, Matthews correlation coefficient, and Brier score. The positive class was $y_{5yr}=1$. The fixed 0.5 probability threshold was used as a descriptive operating point and was not selected or optimized for a clinical decision context. After inspection of internal descriptive performance, the METABRIC soft-voting model—whose observed internal-test ROC-AUC was highest among the five METABRIC-source models—was carried forward for the detailed calibration, conformal, domain-shift, and few-shot analyses. This was a descriptive, post hoc model-specific choice; these secondary analyses are not used to support a comparative superiority claim.

H. Probability Calibration

Calibration analysis used the reproduced METABRIC soft-voting model and the identical source partitions and component hyperparameters. The 383-patient METABRIC reserved calibration partition was divided, stratified with seed 42, into 191 patients for probability calibration and 192 for conformal calibration. A logistic Platt map was fitted on the logit of the uncalibrated soft-voting probability using only the 191-patient probability-calibration subset. Probabilities were clipped to 10^{-6} and $1-10^{-6}$ only for logit transformation. Uncalibrated and Platt-scaled predictions were evaluated on the METABRIC internal test and all 170 eligible TCGA patients using ROC-AUC, Brier score, ten-bin expected calibration error (ECE), and log loss. Reliability plots display observed outcome frequency against mean predicted probability with bin counts (Fig. 3).

I. Split Conformal Prediction and Abstention

The Platt-scaled model was fixed before conformal calibration. For each of the 192 conformal-calibration patients, the nonconformity score was one minus the probability assigned to the observed class. At $\alpha=0.10$, the finite-sample corrected empirical quantile used $\text{rank_ceil}((n+1)(1-\alpha))$. A candidate class was included when its nonconformity score did not exceed the resulting threshold. The same source-derived threshold was applied to the METABRIC internal test and all eligible TCGA patients.

Outcomes were empirical coverage, mean prediction-set size, singleton rate, defer rate, and accuracy among singleton sets. Selective prediction was operationalized through conformal singleton status: singleton sets yielded a class prediction, whereas empty or two-class sets were deferred. Because target exchangeability was not assumed, TCGA coverage was interpreted as empirical external coverage rather than a guaranteed quantity [6,7].

J. Domain-Shift Analysis

A separate cohort-membership classifier was used to distinguish eligible METABRIC from eligible TCGA patients using only age and ER, PR, and HER2 status. The classifier was a class-balanced logistic-regression pipeline with the same training-fitted preprocessing rules. Five-fold stratified cross-validation with seed 42 produced ROC-AUC and balanced accuracy for cohort discrimination. No survival outcome was used. Descriptive comparisons included age, receptor positivity among nonmissing observations, and predictor missingness.

K. Exploratory Few-Shot Target Adaptation

The 170 eligible TCGA patients were split with stratification and seed 2026 into an adaptation pool of 102 patients and a permanently locked test of 68 patients. The locked test was not used for fitting, calibration, threshold selection, model selection, or choice of target sample size. It contained 22 $y_{5yr}=0$ and 46 $y_{5yr}=1$ patients.

Target sample sizes were $N=0, 10, 25, 50$, and 100. $N=0$ used the source Platt map and source conformal threshold. For each N greater than zero, a stratified sample was drawn from the adaptation pool in 30 repetitions using fixed seeds. Probability adaptation fitted a target Platt map, including slope and intercept, on the sampled patients' fixed METABRIC model scores and evaluated the resulting probabilities on the unchanged locked test. The cancer classifier and its preprocessing were never retrained.

Target conformal adaptation was evaluated as a parallel branch. The source model and source Platt probability function remained fixed, and only the conformal threshold was re-estimated from the N labeled target patients. This target-derived threshold was then evaluated on the same locked test using the fixed source-Platt probabilities. Thus, target labels influenced the probability-recalibration parameters in one branch or the conformal threshold in the other, but never the cancer predictor or locked-test evaluation choices.

L. Protocol Deviations and Exploratory Analyses

The original protocol intent was retained as the confirmatory plan: target sample sizes $N = 0, 25, 50, 100, 200$ and intercept-only target recalibration. The executed few-shot analysis differed by using $N = 0, 10, 25, 50, 100$ and fitting both a slope and an intercept through Platt recalibration. The $N = 10$ condition and fitted slope were not in the initial plan, and $N = 200$ was not feasible because the realized adaptation pool contained 102 patients. The executed few-shot analysis is therefore designated exploratory and post hoc. Its results are hypothesis-generating and are not treated as confirmatory evidence for the original specification.

The information boundary was nevertheless preserved. The 68-patient locked test remained disjoint from the 102-patient adaptation pool and untouched by model fitting, probability recalibration, conformal-threshold estimation, threshold or sample-size selection, and stopping decisions. The fixed locked test was used only for evaluation after those choices had been set, and the cancer classifier was never retrained.

Probability recalibration and conformal adaptation were separate exploratory branches. Target Platt recalibration evaluated probability calibration by fitting slope and intercept

on the sampled target patients. Conformal target adaptation did not use the target-Platt probabilities: it updated only the nonconformity threshold using the sampled target labels while keeping the cancer model and source-Platt probabilities fixed.

M. Statistical Analysis

Point estimates and 95% confidence intervals for ROC-AUC, precision-recall AUC, Brier score, sensitivity, and specificity were computed using 2,000 patient-level nonparametric bootstrap replicates and percentile limits [3,4]. The same resampled patient indices were used across models evaluated on a common cohort, permitting paired comparisons. Internal-minus-external ROC-AUC differences and external-minus-internal Brier differences were calculated from bootstrap distributions. Pairwise model comparisons were performed for ROC-AUC and Brier score within each evaluation setting; two-sided bootstrap P values were adjusted within each family using Holm's method [26]. Conformal coverage received Wilson 95% confidence intervals.

For few-shot adaptation, each N greater than zero was summarized across 30 target-sampling repetitions by the mean, standard deviation, and 2.5th and 97.5th percentiles. These percentiles quantify variation due to the sampled adaptation patients while holding the 68-patient locked test fixed; they are not patient-level confidence intervals for a new target population. Analyses used Python 3.11 and the fixed software environment recorded in the project.

IV. RESULTS

A. Cohorts and Outcome Eligibility

METABRIC contained 2,509 patients, of whom 1,917 had an observable binary 5-year outcome; 427 (22.27%) were classified as death by 60 months and 1,490 (77.73%) as known survival beyond 60 months. Of 592 excluded patients, 64 were alive or censored by 60 months and 528 lacked valid overall-survival months. TCGA contained 825 patients, of whom 170 were outcome eligible; 55 (32.35%) were classified as death by 60 months and 115 (67.65%) as known survival beyond 60 months. Of 655 exclusions, 648 were alive or censored by 60 months, and seven lacked valid overall-survival months.

Among eligible patients, mean age was 61.19 years in METABRIC and 57.22 years in TCGA. ER-positive proportions among nonmissing observations were 76.21% and 70.55%, PR-positive proportions were 52.97% and 60.74%, and HER2-positive proportions were 12.32% and 13.10%, respectively. TCGA predictor missingness was 4.12% for ER, 4.12% for PR, and 14.71% for HER2, compared with 0%, 0.05%, and 0.05% in METABRIC. Cohort characteristics are presented in Table I.

TABLE I. COHORT CHARACTERISTICS FOR THE OUTCOME-ELIGIBLE ANALYTIC SAMPLES. RECEPTOR PERCENTAGES USE NONMISSING DENOMINATORS

Cohort	Initial N	Eligible N	Death by 60 months	Survival beyond 60 months	Age, mean (SD)	ER positive	PR positive	HER2 positive	Missing ER / PR / HER2
METABRIC	2,509	1,917	427 (22.27%)	1,490 (77.73%)	61.19 (12.94)	76.21%	52.97%	12.32%	0.00% / 0.05% / 0.05%
TCGA	825	170	55 (32.35%)	115 (67.65%)	57.22 (14.42)	70.55%	60.74%	13.10%	4.12% / 4.12% / 14.71%

B. Bidirectional Internal and External Performance

For METABRIC-source models, the soft-voting ensemble had the highest observed internal ROC-AUC: 0.732 (95% CI, 0.672–0.786) on 384 internal-test patients. Its ROC-AUC was 0.643 (95% CI, 0.557–0.728) on all 170 eligible TCGA patients (Table II; Fig. 2). The observed internal-to-external numerical decrease was approximately 0.088 (95% CI, -0.015 to 0.196). Because this interval included zero, the decrease was not statistically significant. Brier score increased from 0.160 (95% CI, 0.142–0.177) internally to 0.214 (95% CI, 0.182–0.247) externally, an external-minus-internal difference of 0.054 (95% CI, 0.020–0.090). Precision-recall AUC was 0.903 internally and 0.768 externally.

For TCGA-source models, the RBF support vector machine had the highest observed internal ROC-AUC: 0.674 (95% CI, 0.460–0.861) on 34 patients. Applied to all 1,917 eligible METABRIC patients, its ROC-AUC was 0.586 (95% CI, 0.554–0.615). The observed internal-to-external numerical decrease was also approximately 0.088 (95% CI, -0.127 to 0.282), and this interval likewise included zero. Its Brier score

was 0.206 (95% CI, 0.155–0.263) internally and 0.183 (95% CI, 0.177–0.190) externally, an external-minus-internal difference of -0.023 (95% CI, -0.080 to 0.028). The TCGA-source internal estimate is substantially less precise because the source internal test contains only 34 patients; the two transport directions therefore should not be interpreted as providing symmetric statistical evidence.

Threshold-based performance was uneven. The METABRIC-source soft-voting model had sensitivity 0.956 and specificity 0.174 internally, and sensitivity 0.930 and specificity 0.109 in TCGA. Several RBF support-vector evaluations assigned every patient to the positive class at the fixed descriptive 0.5 operating point, yielding sensitivity 1.000 and specificity 0.000; the zero specificity reflects this fixed operating point and is not a recommended clinical threshold. Complete metrics and bootstrap intervals for all five models are provided in Table II. Holm-adjusted pairwise model-comparison results are retained in the verified statistical output, but the approximately 0.088 cross-cohort numerical decreases above are interpreted from their confidence intervals and are not labeled significant.

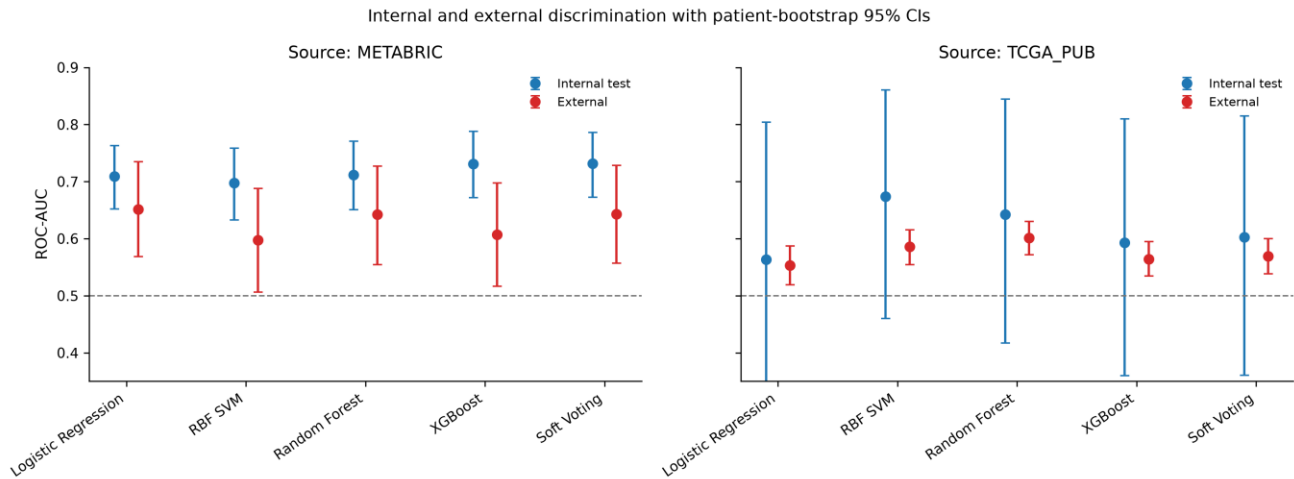


Fig. 2. Internal and external ROC-AUC for both transport directions with 95 percent bootstrap confidence intervals.

TABLE II. BIDIRECTIONAL INTERNAL AND EXTERNAL PERFORMANCE. THE POSITIVE CLASS WAS KNOWN SURVIVAL BEYOND 60 MONTHS; SENS. DENOTES SENSITIVITY AND SPEC. SPECIFICITY AT THE FIXED 0.5 THRESHOLD

Source	Evaluation	Model	N	ROC-AUC [95% CI]	PR-AUC [95% CI]	Brier [95% CI]	Sens.	Spec.
METABRIC	METABRIC Internal test	Logistic Regression	384	0.709 [0.652–0.763]	0.901 [0.868–0.930]	0.228 [0.215–0.241]	0.695	0.581
METABRIC	METABRIC Internal test	RBF SVM	384	0.698 [0.633–0.758]	0.878 [0.836–0.918]	0.161 [0.139–0.183]	1.000	0.000
METABRIC	METABRIC Internal test	Random Forest	384	0.712 [0.651–0.770]	0.888 [0.847–0.924]	0.156 [0.133–0.178]	0.963	0.151
METABRIC	METABRIC Internal test	XGBoost	384	0.731 [0.672–0.788]	0.895 [0.855–0.930]	0.152 [0.130–0.173]	0.993	0.105
METABRIC	METABRIC Internal test	Soft Voting	384	0.732 [0.672–0.786]	0.903 [0.867–0.933]	0.160 [0.142–0.177]	0.956	0.174
METABRIC	TCGA External eligible	Logistic Regression	170	0.651 [0.568–0.735]	0.819 [0.744–0.883]	0.235 [0.213–0.256]	0.713	0.455
METABRIC	TCGA External eligible	RBF SVM	170	0.597 [0.506–0.688]	0.751 [0.657–0.842]	0.223 [0.185–0.263]	1.000	0.000
METABRIC	TCGA External eligible	Random Forest	170	0.643 [0.554–0.727]	0.788 [0.699–0.864]	0.221 [0.183–0.265]	0.957	0.091
METABRIC	TCGA External eligible	XGBoost	170	0.607 [0.517–0.698]	0.746 [0.652–0.837]	0.226 [0.185–0.270]	0.957	0.127
METABRIC	TCGA External eligible	Soft Voting	170	0.643 [0.557–0.728]	0.768 [0.672–0.862]	0.214 [0.182–0.247]	0.930	0.109

TCGA	TCGA Internal test	Logistic Regression	34	0.563 [0.300–0.804]	0.698 [0.517–0.904]	0.231 [0.161–0.311]	0.739	0.455
TCGA	TCGA Internal test	RBF SVM	34	0.674 [0.460–0.861]	0.811 [0.639–0.954]	0.206 [0.155–0.263]	1.000	0.000
TCGA	TCGA Internal test	Random Forest	34	0.642 [0.417–0.844]	0.793 [0.626–0.938]	0.207 [0.140–0.283]	0.957	0.091
TCGA	TCGA Internal test	XGBoost	34	0.593 [0.360–0.810]	0.746 [0.568–0.922]	0.210 [0.124–0.306]	0.913	0.273
TCGA	TCGA Internal test	Soft Voting	34	0.603 [0.360–0.815]	0.768 [0.595–0.920]	0.205 [0.141–0.277]	0.957	0.364
TCGA	METABRIC External eligible	Logistic Regression	1,917	0.553 [0.519–0.587]	0.786 [0.761–0.813]	0.303 [0.292–0.314]	0.517	0.581
TCGA	METABRIC External eligible	RBF SVM	1,917	0.586 [0.554–0.615]	0.812 [0.787–0.838]	0.183 [0.177–0.190]	1.000	0.000
TCGA	METABRIC External eligible	Random Forest	1,917	0.601 [0.571–0.630]	0.838 [0.816–0.861]	0.191 [0.184–0.198]	0.799	0.218
TCGA	METABRIC External eligible	XGBoost	1,917	0.564 [0.534–0.595]	0.817 [0.794–0.840]	0.228 [0.218–0.239]	0.715	0.349
TCGA	METABRIC External eligible	Soft Voting	1,917	0.569 [0.538–0.600]	0.813 [0.790–0.837]	0.213 [0.205–0.220]	0.755	0.314

C. Calibration and Conformal Transport

On the METABRIC internal test, source Platt scaling preserved ROC-AUC at 0.732 while changing Brier score from 0.160 to 0.157, ECE from 0.066 to 0.044, and log loss from 0.494 to 0.485. In TCGA, ROC-AUC remained 0.643, but

source-only Platt scaling worsened the Brier score from 0.214 to 0.222, ECE from 0.078 to 0.104, and log loss from 0.616 to 0.641. Fig. 3 shows the corresponding reliability plots, and Table III reports the calibration and conformal metrics. Thus, the source calibration map showed a small favorable internal change but did not transport as an improvement in external calibration.

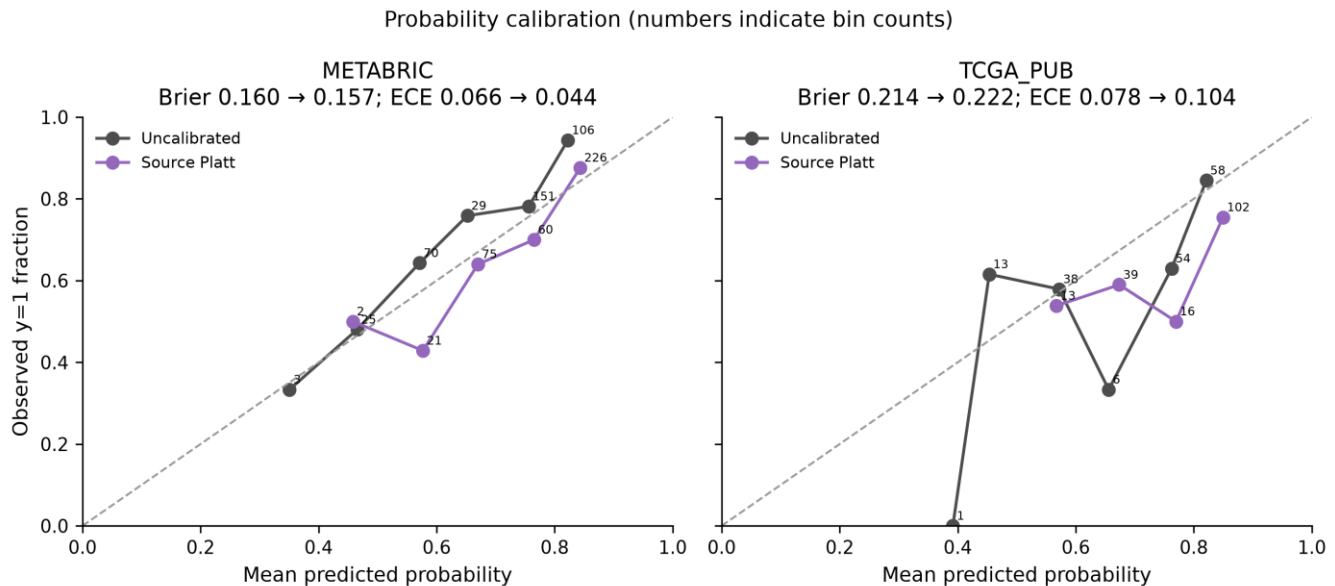


Fig. 3. Calibration behavior of the fixed METABRIC soft-voting model before and after source Platt scaling.

At nominal 90% coverage, empirical conformal coverage was 0.891 (Wilson 95% CI, 0.855–0.918) on the METABRIC internal test and 0.818 (Wilson 95% CI, 0.753–0.868) in TCGA (Fig. 4). The external interval did not include 0.900. Mean prediction-set size was 1.302 internally and 1.329 externally.

Singleton and defer rates were 0.698 and 0.302 internally, compared with 0.671 and 0.329 externally. Accuracy among singleton predictions was 0.843 internally and 0.728 externally. These results show external undercoverage of the source-derived conformal sets.

TABLE III. SOURCE CALIBRATION AND SPLIT-CONFORMAL TRANSPORT AT ALPHA 0.10. ECE USED TEN EQUAL-WIDTH BINS. TCGA COVERAGE IS EMPIRICAL EXTERNAL COVERAGE

Evaluation	N	Uncalibrated AUC	Uncalibrated Brier / ECE / log loss	Platt Brier / ECE / log loss	Coverage [Wilson 95% CI]	Mean set size	Singleton	Defer	Singleton accuracy
METABRIC internal	384	0.732	0.160 / 0.066 / 0.494	0.157 / 0.044 / 0.485	0.891 [0.855–0.918]	1.302	0.698	0.302	0.843
TCGA external	170	0.643	0.214 / 0.078 / 0.616	0.222 / 0.104 / 0.641	0.818 [0.753–0.868]	1.329	0.671	0.329	0.728

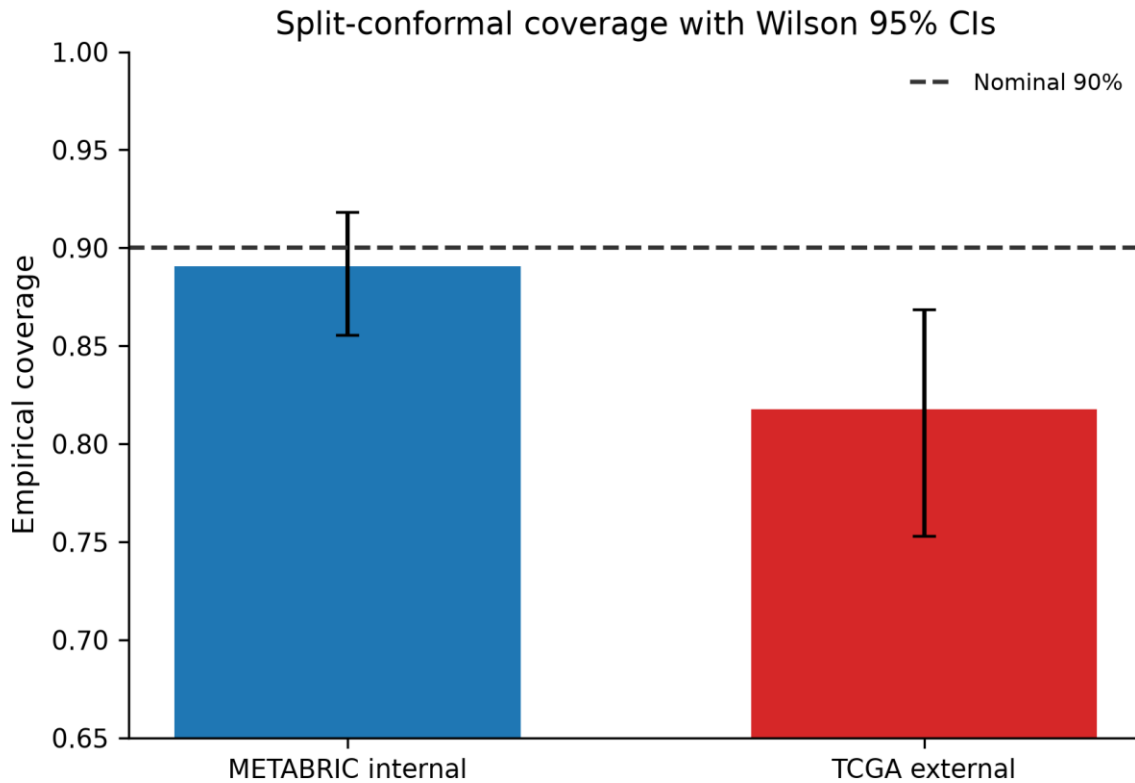


Fig. 4. Internal and external empirical split-conformal coverage relative to the nominal 90 percent level.

D. Domain Shift

Eligible TCGA patients were younger on average than METABRIC patients by 3.97 years, with a standardized mean difference of -0.290. Receptor prevalence and missingness also differed, most notably for HER2 missingness. The cohort-membership classifier achieved mean five-fold ROC-AUC 0.604 (SD, 0.049) and balanced accuracy 0.574 (SD, 0.047). These label-free results indicate measurable, although incomplete, separability using the four harmonized predictors and their missingness patterns.

E. Exploratory Few-Shot Target Adaptation

The TCGA adaptation pool contained 102 patients (33 y_5yr=0, 69 y_5yr=1), and the locked test contained 68 patients (22 and 46, respectively). Table IV and Fig. 5 summarize repeated target adaptation. With N=0, the locked-test Brier score was 0.226, and ECE was 0.128. At N=10, mean Brier score was

0.254 (SD, 0.059) and mean ECE was 0.162 (SD, 0.109), showing instability at the smallest target sample. Mean Brier score and ECE were 0.219 and 0.066 at N=25, 0.214 and 0.047 at N=50, and 0.213 and 0.036 at N=100. Relative to N=0, larger target samples therefore improved both locked-test Brier score and ECE without retraining the cancer classifier. Singleton accuracy at N=10 was not summarized because three of the 30 repetitions produced no singleton predictions, making per-repetition singleton accuracy undefined and causing the frozen all-repetition aggregation to return NA.

Target conformal adaptation did not restore mean nominal coverage. Mean coverage was 0.809 at N=0, 0.882 at N=10, 0.886 at N=25, 0.866 at N=50, and 0.868 at N=100. The highest mean was 0.886 at N=25, below 0.900. Mean defer rates were 0.294, 0.543, 0.554, 0.469, and 0.456 across the same N values. Fig. 5 displays Brier score, ECE, and conformal coverage with 2.5th–97.5th percentiles and the nominal 90% reference line.

TABLE IV. EXPLORATORY FEW-SHOT TARGET ADAPTATION ON THE FIXED 68-PATIENT TCGA LOCKED TEST. VARIATION IS ACROSS TARGET SAMPLES, NOT PATIENT-BOOTSTRAP CONFIDENCE INTERVALS

Target N	Repetitions	Brier mean (SD)	ECE mean (SD)	Log loss mean (SD)	Coverage mean (SD)	Mean set size	Defer mean (SD)	Singleton accuracy mean (SD)
0	1	0.226 (0.000)	0.128 (0.000)	0.653 (0.000)	0.809 (0.000)	1.294	0.294 (0.000)	0.729 (0.000)
10	30	0.254 (0.059)	0.162 (0.109)	1.426 (2.632)	0.882 (0.065)	1.543	0.543 (0.240)	NA
25	30	0.219 (0.009)	0.066 (0.047)	0.633 (0.025)	0.886 (0.044)	1.554	0.554 (0.172)	0.740 (0.029)
50	30	0.214 (0.001)	0.047 (0.020)	0.618 (0.003)	0.866 (0.013)	1.469	0.469 (0.048)	0.747 (0.012)
100	30	0.213 (0.000)	0.036 (0.009)	0.615 (0.000)	0.868 (0.000)	1.456	0.456 (0.000)	0.757 (0.000)

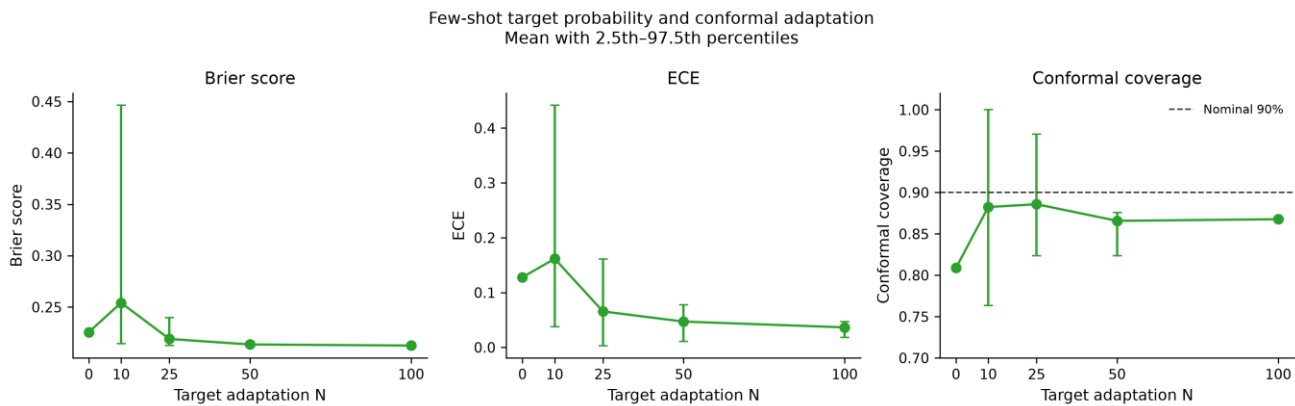


Fig. 5. Exploratory few-shot target recalibration and conformal threshold adaptation on the fixed TCGA locked test.

V. DISCUSSION

The main finding was an observed numerical decrease in discrimination in both transport directions. At the reported precision, the METABRIC soft-voting model and the highest-observed-AUC TCGA internal model each showed an internal-to-external ROC-AUC difference of approximately 0.088. This numerical similarity is descriptive: both 95% confidence intervals included zero, so neither decrease was statistically established. The TCGA-source internal estimate is substantially less precise because its internal test contained only 34 patients; the two directions therefore provide complementary evidence about directionality but not symmetric statistical evidence. The soft-voting model was carried forward for detailed uncertainty analysis only after inspection of internal descriptive performance. Accordingly, those model-specific analyses are secondary and do not support a claim of prospectively selected superiority.

Probability calibration exposed a different aspect of transportability. Source Platt scaling improved the observed METABRIC internal Brier score, ECE, and log loss, but the same fixed map produced worse observed values for all three metrics in TCGA. ROC-AUC was unchanged because the fitted transformation preserved score ordering. Discrimination and calibration therefore answered different questions. These metric changes show that the source map did not transport as an observed improvement in TCGA; the ten-bin ECE alone is not treated as proof of general calibration failure. Source-only recalibration should be evaluated externally as part of the transported prediction system. It cannot be assumed to be universally corrective.

The conformal analysis showed a parallel limitation of source-derived uncertainty after transport. Empirical coverage was 89.1% on the METABRIC internal test, close to the nominal 90%, but 81.8% in TCGA. The TCGA Wilson interval ended at 86.8%, and external evaluation also showed a higher defer rate and lower singleton accuracy. Split-conformal validity depends on an exchangeability relationship between calibration and evaluation observations; that relationship was not established across these cohorts. The result is therefore external empirical undercoverage, not a failure of conformal theory or a contradiction of the source guarantee.

The label-free domain analysis provided context without identifying a cause. A cohort-classifier ROC-AUC of 0.604 showed that METABRIC and TCGA were partly distinguishable using the four predictors and their missingness patterns. TCGA patients were younger on average, receptor distributions differed, and HER2 missingness was higher. These findings are compatible with covariate shift, but they do not attribute the performance changes to any single variable or exclude unmeasured differences in selection, follow-up, or care. Because the shift analysis was not used to select features or models, it informs interpretation rather than comparative model ranking.

The exploratory, post hoc few-shot analysis separated probability recalibration from conformal threshold adaptation. Target Platt recalibration improved mean locked-test Brier score and ECE at $N=25$, 50, and 100 relative to $N=0$, with the lowest means at $N=100$. $N=10$ was worse and highly variable. Estimating both a calibration slope and intercept from very limited labels was unstable. These probability results suggest that labeled target data can correct aspects of the score-to-risk relationship, but they remain conditional on one 68-patient locked test and do not establish a generally effective adaptation sample size.

The conformal branch answered a narrower and separate question. It kept the cancer model and source-Platt probabilities fixed and updated only the conformity threshold using target adaptation patients. This simple threshold update did not restore mean 90% coverage at any evaluated N ; the highest mean coverage was 88.6% at $N=25$, and coverage was 86.8% at $N=100$. The lack of monotonic recovery indicates that a target-derived threshold alone was insufficient in this small experiment. It does not compare alternative conformal methods and cannot support a superiority claim about any uncertainty procedure.

These results concern the transportability of the complete prediction workflow. Internal discrimination did not determine external discrimination, and internal calibration improvement did not imply external calibration improvement. Near-nominal internal conformal coverage also did not imply target coverage. Cross-cohort evaluation should keep target data protected and label post hoc adaptation explicitly. Ranking, probability accuracy, and set coverage should be reported separately. The

present findings do not establish clinical usefulness or deployment readiness.

A. Future Work

Future external-validation studies should use larger cohorts and clinically prespecified operating thresholds, with threshold uncertainty and decision-curve/net-benefit analysis. Calibration assessment should add calibration intercept and slope, flexible calibration curves with uncertainty bands, and a prevalence-only Brier reference. Richer harmonization should examine overall stage where semantically defensible, principled treatment of cohort-specific HER2 missingness, and molecular or other prognostic features.

Direct comparison with established clinical prognostic tools such as PREDICT and the Nottingham Prognostic Index should be undertaken when their required inputs can be reconstructed consistently across cohorts. A natural extension is a sequential target-adaptation pipeline in which target probability recalibration is followed by conformal recalibration using disjoint target data. Censoring-aware approaches, including Cox regression, random survival forests, and survival boosting, should be evaluated to use partially observed follow-up more efficiently while preserving a clinically interpretable horizon; such methods do not by themselves eliminate selection concerns. Temporal, geographic, prospective, and workflow-based validation also remain necessary.

VI. LIMITATIONS

First, the outcome-eligible TCGA cohort contained only 170 patients; its source internal test contained 34, and the few-shot locked test contained 68. These sample sizes yielded wide confidence intervals and limited the resolution of performance and adaptation estimates. In particular, the TCGA-source direction has lower precision and should not receive evidentiary weight symmetric to the METABRIC-source direction. Reusing one locked test across 30 target-sampling repetitions enabled controlled comparisons, but the reported percentiles are conditional on that fixed test and quantify adaptation-sample variation rather than test-population uncertainty.

Second, only 170 of 825 TCGA patients and 1,917 of 2,509 METABRIC patients met the binary 5-year outcome rule. In TCGA, 648 patients were excluded because they were alive or censored by 60 months, and seven had invalid or missing overall-survival months. The resulting complete-outcome analysis may represent a selected subset; no censoring-aware survival model was fitted.

Third, the shared predictor set contained only age and ER, PR, and HER2 status. This deliberate restriction enabled transparent cross-cohort semantic alignment but excluded important prognostic and molecular information. Tumor stage was not used because TCGA TUMOR_STAGE represented T categories, and no supported overall-stage mapping was available; cohort-specific receptor, particularly HER2, missingness also remained. Alternative stage or HER2 harmonization strategies were not evaluated.

Fourth, the detailed calibration, conformal, domain-shift, and few-shot analyses were performed only for METABRIC-to-TCGA transport using the soft-voting model; equivalent

reverse-direction uncertainty analyses were not performed. The reported ten-bin ECE is a descriptive, bin-dependent calibration summary and is particularly sensitive to finite-sample variation in the 170-patient external TCGA cohort. Calibration intercept and slope as external-validation measures, flexible calibration curves with uncertainty bands, and comparison with a prevalence-only Brier reference were not part of the accepted analysis. Brier-score and log-loss results remain complementary but do not remove these limitations.

Fifth, the fixed 0.5 threshold was only a descriptive operating point. Threshold-dependent sensitivity and specificity, including zero specificity in several RBF-SVM evaluations, must not be interpreted as recommended clinical operating points. Although frozen patient-bootstrap outputs contain confidence intervals for sensitivity and specificity at 0.5, no clinical threshold optimization or uncertainty analysis around a clinically chosen threshold was performed, and no decision-curve or net-benefit analysis was conducted.

Finally, target probability recalibration estimated both slope and intercept and was unstable at very small N. The accepted exploratory design evaluated probability recalibration and conformal-threshold adaptation as separate branches to isolate their respective effects while maintaining the locked evaluation boundary; a combined probability-plus-conformal target-adaptation pipeline was not evaluated. Direct comparison with PREDICT or the Nottingham Prognostic Index was also outside scope because the required inputs could not be reconstructed consistently from the deliberately restricted harmonized predictor set. This retrospective two-cohort study included no prospective, temporal, geographic, workflow-embedded, subgroup-fairness, or clinical-impact validation. Clinical deployment readiness cannot be inferred.

VII. CONCLUSION

Using age and receptor status, machine-learning models showed observed numerical decreases in ROC-AUC in both METABRIC-to-TCGA and TCGA-to-METABRIC transport. Both decreases were approximately 0.088, but their 95% confidence intervals included zero. Source-only Platt calibration improved the observed internal probability metrics but produced worse observed external Brier score and ECE. Source split-conformal prediction showed external empirical undercoverage in TCGA, reaching 81.8% rather than nominal 90% coverage. Exploratory few-shot target recalibration improved locked-test Brier score and ECE at larger N, but target conformal-threshold adaptation did not restore mean 90% coverage. These findings support separate evaluation of discrimination, calibration, and uncertainty under cohort shift and do not establish clinical usefulness or deployment readiness.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

NADIR Younes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing, Project administration, Supervision.

AMHAIMAR Lahcen: Methodology, Visualization, Software.

Mohamed Azouazi: Supervision, Validation, Writing – review and editing.

KHALIDI Abderrahim: Supervision, Validation, Writing – review and editing.

RACHDI Mohamed: Methodology, Formal analysis, Writing – review and editing.

BAKHOUYI Abdellah: Methodology, Investigation, Data curation.

DATA AVAILABILITY

This study used patient-level and sample-level clinical data for cBioPortal study identifiers `brca_metabric` and `brca_tcga_pub`, retrieved through the official REST API on 9 August 2026 (UTC). No mutation, expression, copy-number, imaging, or complete study-archive data were used. Source clinical data remain available through cBioPortal subject to the originating studies' data-use terms; no raw protected, private, or directly identifying data are redistributed with the reproducibility archive.

CODE AVAILABILITY

The executable analysis workflow comprises numbered Python scripts, fixed environment requirements, machine-readable aggregate result tables, split-manifest hashes, and figure-generation code. A publication-ready release candidate has been prepared for author deposit, but no paper-specific public repository URL, release tag, commit, or archival DOI was verified at camera-ready preparation; consequently, no public identifier is claimed here.

ETHICS STATEMENT

This study was a secondary analysis of publicly available, de-identified clinical data obtained through cBioPortal. No new participants were recruited or contacted, and no directly identifying information was accessed. Ethical approval and informed consent for the original data collection were governed by the respective METABRIC and TCGA study protocols.

FUNDING

This research received no external funding.

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language refinement and editorial preparation. All AI-assisted checks, numerical results, interpretations, and manuscript content were critically reviewed and validated by the authors. The authors take full responsibility for the content of the publication.

REFERENCES

- [1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263. <https://doi.org/10.3322/caac.21834>
- [2] Li J, Zhou Z, Dong J, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One.* 2021;16(4):e0250370. <https://doi.org/10.1371/journal.pone.0250370>
- [3] Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ.* 2024;384:e074819. <https://doi.org/10.1136/bmj-2023-074819>
- [4] Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ.* 2024;384:e074820. <https://doi.org/10.1136/bmj-2023-074820>
- [5] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>
- [6] Angelopoulos AN, Bates S. Conformal prediction: a gentle introduction. *Found Trends Mach Learn.* 2023;16(4):494–591. <https://doi.org/10.1561/22000000101>
- [7] Tibshirani RJ, Barber RF, Candès EJ, Ramdas A. Conformal prediction under covariate shift. In: *Advances in Neural Information Processing Systems* 32. 2019:2526–2536. https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html
- [8] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature.* 2012;486:346–352. <https://doi.org/10.1038/nature10983>
- [9] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490:61–70. <https://doi.org/10.1038/nature11412>
- [10] Huetting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. External validation of 87 clinical prediction models supporting clinical decisions for breast cancer patients. *Breast.* 2023;69:382–391. <https://doi.org/10.1016/j.breast.2023.04.003>
- [11] Clift AK, Dodwell D, Lord S, Petrou S, Brady M, Collins GS, Hippisley-Cox J. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. *BMJ.* 2023;381:e073800. <https://doi.org/10.1136/bmj-2022-073800>
- [12] Elmejjari C, El Khalfaoui Y, Nadir Y, Alibouch B, Qbadou M. A hybrid approach for breast cancer classification using CLAHE-enhanced mammograms and EfficientNetV2S. In: Mejdoub Y, Elamri A, Kardouchi M, editors. *Connected Objects, Artificial Intelligence, Telecommunications and Electronics Engineering. Lecture Notes in Networks and Systems*, vol. 1584. Cham: Springer; 2026. p. 536–542. https://doi.org/10.1007/978-3-032-01536-5_81
- [13] Elmejjari C, Nadir Y, Qbadou M. Deep learning-based techniques for breast cancer classification: a systematic review. In: 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). IEEE; 2024. p. 1–7. <https://doi.org/10.1109/IRASET60544.2024.10548784>
- [14] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [15] Lee S, Yin C, Zhang P. Stable clinical risk prediction against distribution shift in electronic health records. *Patterns.* 2023;4(9):100828. <https://doi.org/10.1016/j.patter.2023.100828>
- [16] Rabanser S, Günnemann S, Lipton ZC. Failing loudly: an empirical study of methods for detecting dataset shift. In: *Advances in Neural Information Processing Systems* 32. 2019:1394–1406. https://proceedings.neurips.cc/paper_files/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html
- [17] Lin HT, Lin CJ, Weng RC. A note on Platt's probabilistic outputs for support vector machines. *Mach Learn.* 2007;68(3):267–276. <https://doi.org/10.1007/s10994-007-5018-6>
- [18] Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* 2004;23(16):2567–2586. <https://doi.org/10.1002/sim.1844>
- [19] Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol.* 2022;22:316. <https://doi.org/10.1186/s12874-022-01801-8>

- [20] Shashikumar SP, Amrollahi F, Nemati S. Unsupervised detection and correction of model calibration shift at test-time. *Annu Int Conf IEEE Eng Med Biol Soc.* 2023;2023:1–4. <https://doi.org/10.1109/EMBC40787.2023.10341086>
- [21] Tanner KT, Keogh RH, Coupland CAC, Hippisley-Cox J, Diaz-Ordaz K. Dynamic updating of clinical survival prediction models in a changing environment. *Diagn Progn Res.* 2023;7:24. <https://doi.org/10.1186/s41512-023-00163-z>
- [22] Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. *Ann Stat.* 2023;51(2):816–845. <https://doi.org/10.1214/23-AOS2276>
- [23] Yang Y, Kuchibhotla AK, Tchetgen Tchetgen E. Doubly robust calibration of prediction sets under covariate shift. *J R Stat Soc Series B Stat Methodol.* 2024;86(4):943–965. <https://doi.org/10.1093/jrsssb/qkae009>
- [24] Geifman Y, El-Yaniv R. SelectiveNet: A deep neural network with an integrated reject option. In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR. 2019;97:2151–2159. <https://proceedings.mlr.press/v97/geifman19a.html>
- [25] Hendrickx K, Perini L, Van der Plas D, Meert W, Davis J. Machine learning with a reject option: a survey. *Mach Learn.* 2024;113:3073–3110. <https://doi.org/10.1007/s10994-024-06534-x>
- [26] Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65–70. <https://www.jstor.org/stable/4615733>