Research Square

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

Patient overlap inflates ovarian cancer transcriptomic model performance

Jingyun Xu
Zhongda Hospital, Southeast University
Ying Li
152051734920163.com

Zhongda Hospital, Southeast University

Research Article

Keywords: ovarian cancer, transcriptomics, repeated biopsies, identity confounding, cross validation, negative control, reproducibility

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11432475/v1

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

Additional Declarations: No competing interests reported.

Abstract

Repeated biopsies from the same patient may enter both the training and test sets of an ovarian cancer transcriptomic model, leading to overly optimistic performance estimates. We compared specimen-stratified with patient-grouped threefold cross-validation in three public cohorts (107 within-cohort patients; 246 specimens or biopsies). We used three fixed classifiers and 20 repeated partitions, and aggregated out-of-fold scores at the patient level. The mean differences in area under the receiver operating characteristic curve (AUC) were 0.126, 0.339 and 0.137, with conditional patient-bootstrap intervals of 0.084–0.171, 0.241–0.450 and 0.087–0.197. After labels were randomized consistently within patients, specimen-split support vector machines still achieved mean AUCs of 0.652–0.830. Two additional sources (36 patients; 117 measurements) showed similar predictability of artificial labels. Across all five sources, patient-grouped within-fold null AUCs were near 0.5. Independently permuting specimen labels obscured this effect. In contrast, a spatial subset restricted to pretreatment biopsies retained an AUC difference of 0.131. These findings show how patient identity can inflate apparent performance in repeated-biopsy transcriptomic analyses. We provide an auditable ovarian cancer benchmark using established validation controls, not a new algorithm or a clinically validated predictor.

Introduction

High-grade serous ovarian carcinoma (HGSOC) shows marked heterogeneity across tumor sites and within individual lesions. Multi-site single-cell profiling has demonstrated distinct cellular ecosystems in primary tumors, metastases and ascites [1]. Spatial transcriptomic studies have also identified different tumor subclones and associated microenvironments within the same tumor section [2]. These observations highlight the value of collecting more than one specimen per patient. They also raise an important question for predictive modeling: how should correlated specimens be handled when the intended prediction concerns a patient who was not included in model development?

Public ovarian cancer datasets now provide opportunities to examine this question directly. Paired transcriptomic studies have documented changes after neoadjuvant chemotherapy and their associations with recurrence-related outcomes [3, 4]. Other studies include biopsies from different anatomical sites or treatment time points, allowing investigation of spatial and temporal variation in the tumor microenvironment [5]. In these settings, several biopsies can share a single patient-level clinical label. Treating them as independent observations during model validation may therefore give an incorrect impression of the amount of information available for prediction in a new patient.

The underlying methodological concern is established. Validation data should resemble the setting in which a model will be used, including whether future patients were represented during training [6]. When records from the same person appear in both training and test sets, a classifier may learn patient-associated patterns rather than a transferable clinical association. This has been termed identity confounding, and patient-consistent label permutation was proposed as a diagnostic approach [7]. Similarity between gene-expression samples can also distort apparent machine-learning performance

[8]. More broadly, data leakage has been identified as a recurring source of overly optimistic results in computational research [9]. Current reporting guidance emphasizes transparent description of the prediction task and model evaluation [10].

For ovarian cancer transcriptomic studies, the practical size and persistence of this validation artifact depend on the available specimens and their patient relationships. Paired treatment samples and spatially distinct biopsies need not produce identical patterns of overlap. A patient-level area under the receiver operating characteristic curve (AUC) also does not guarantee that model training was patient-disjoint: averaging predictions after validation cannot remove information already shared across the split. Negative-control results may likewise depend on whether labels are permuted among patients or independently among specimens. These distinctions warrant direct evaluation using recoverable patient and biopsy identifiers, rather than inference from model performance alone.

We therefore conducted a patient-resolved benchmark using five public ovarian cancer transcriptomic sources. Three cohorts, comprising 107 within-cohort patients and 246 specimens or distinct biopsies, supported analyses with recurrence-related labels. We compared specimen-level and patient-grouped cross-validation using three fixed classifiers. We also evaluated patient-consistent label randomization, specimen-level permutation, within-fold AUC calculations and a pretreatment spatial subset. Two additional sources were used solely for artificial-label controls. Our objective was to quantify and characterize validation optimism in these specific data structures and provide an auditable analysis workflow. We did not aim to introduce a new validation algorithm or develop a clinically deployable prognostic biomarker.

Methods

Study design and analysis specification

We conducted a secondary computational analysis of public processed expression matrices and published metadata accessed on 19 October 2026. Patients, biological biopsies or regions, and technical library measurements were treated as distinct units. In the three cohorts with patient-level recurrence-related outcomes, we compared discrimination under two cross-validation splitting strategies. Each cohort was analyzed separately; no datasets were pooled for model training. We did not assume that the original studies shared identical assays, endpoints or clinical aims.

Before fitting any audit classifiers, we fixed the three models, two splitting strategies, within-specimen rank transformation, 20 repeated threefold partitions and primary mean AUC contrast. The protocol also specified 100 patient-level label permutations and a pretreatment spatial sensitivity analysis. These datasets had been examined previously for separate biomarker feasibility questions. Thus, the specification represents an analysis lock rather than prospective preregistration or outcome-blind dataset selection. We added two further cohorts and the analyses of permutation units and fold-score

comparability later; these additions are identified separately. Source URLs, filenames, checksums and analysis scripts are provided with the supplementary materials.

Patient and biopsy reconstruction; clinical endpoints

We used the patient identifier as the independent grouping unit. Biopsies from different sites or treatment time points were kept as separate biological observations but linked to their patient. When identifiable, technical libraries from a single biopsy were averaged before validation. We checked that all retained specimens from a patient had the same binary clinical label. Patients were counted within each data source because reliable linkage of donors across GEO studies was not possible.

GSE201600 contained 62 paired NanoString specimens obtained before and after neoadjuvant chemotherapy from 31 patients, with 750 processed expression features [3]. We extracted patient identifiers, specimen timing, platinum-free interval (PFI) and progression status from the series metadata. Short PFI required a documented progression event before six months. Patients without progression and with less than six months of follow-up were not assigned to the noncase group. All 31 patients were eligible, including eight short-PFI cases. Two patients with progression at exactly six months contributed four specimens. In a boundary sensitivity analysis, we classified them as short PFI using a threshold of at most six months. Although frontline carboplatin and paclitaxel were reported, treatment histories were not uniform.

GSE227100 included 48 paired pretreatment and post-treatment samples from 24 patients. Its processed, log-transformed TPM matrix contained 18,950 features [4]. We used the original Early and Late recurrence categories as binary outcomes (11 and 13 patients). We did not derive individual PFI values from these categories or treat them as equivalent to the PFI thresholds in other cohorts.

For GSE173420, we matched RNA filenames with the published clinical workbook [5]. We included high-grade serous ovarian carcinoma samples with an interpretable PFI and a documented pretreatment or post-treatment time point. Normal tissues, relapse samples, unmatched controls and records with unresolved endpoints were excluded. The source notes use a terminal 'b' to indicate ongoing follow-up. Patients without an event were assigned to the nonshort group only if follow-up reached at least 180 days. Short PFI was defined as less than 180 days. Duplicate clinical annotations were combined only when patient identity, sampling time and PFI agreed. We averaged technical libraries from the same biopsy but retained separate biological regions. This yielded 52 patients, including 22 short-PFI cases, and 136 biopsies. After removing three completely empty annotation rows, the matrisome matrix contained 1,026 features. This is a restricted expression panel, not whole-transcriptome data.

The three clinical-label cohorts contained 107 within-study patients and 246 specimens or biological biopsies. We analyzed them separately rather than treating them as one harmonized 107-patient cohort. Data availability, recoverable patient identifiers and suitable processed expression files determined

eligibility, making this a convenience benchmark. Dataset screening and patient-level mappings are documented in Table S1 and the accompanying files.

Expression representation and fold-isolated preprocessing

Expression values were converted to percentile ranks across available genes separately for each specimen. This transformation used no information from other specimens or outcome labels. It provided a consistent representation within each cohort, but did not make the different assays biologically equivalent. For every training fold, we selected up to 500 features with the highest variance in training observations, using a stable rule to resolve ties. We fitted StandardScaler on those features in the training data and applied the fitted scaler to the test data. Feature selection and scaling were repeated for every fold and permutation. We did not screen features by outcome, correct batches across cohorts or choose features using test-fold performance.

Fixed classifiers and validation partitions

We used three classifiers specified in advance: a radial-basis-function support vector machine (SVM; C = 1, gamma='scale', class_weight='balanced'), L2-penalized logistic regression (C = 1, class_weight='balanced', max_iter = 1,000), and one-nearest-neighbor classification. For patient-level aggregation, we applied a fixed logistic transformation to the SVM decision margin. This transformation was not fitted to calibrate probabilities. Logistic regression and the nearest-neighbor classifier provided positive-class probabilities. We did not tune hyperparameters, select a preferred model or choose algorithms according to the observed outcomes [11].

We compared specimen-level StratifiedKFold with patient-level StratifiedGroupKFold. Each used three folds and 20 randomized repeats with seeds 20261019–20261038. Patient grouping prevented a patient from appearing in both training and test sets within any fold. In specimen-level splitting, biopsies from the same patient could appear in both sets. For each repeat, we averaged out-of-fold specimen scores within patient, then calculated area under the receiver operating characteristic curve (AUC) using one score per patient. This gave patients equal outcome weight regardless of biopsy number. Averaging predictions at the patient level does not remove identity information already available during training.

For each clinical cohort, the primary estimand was specimen-split AUC minus patient-grouped AUC, averaged equally across three classifiers and 20 repeated partitions. We averaged AUC values rather than combining model predictions; the result is not an ensemble classifier. All nine cohort-by-model comparisons and their repeat-level values are reported in Tables S2 and S3. Because grouping also changes the number and composition of patients used for training, this descriptive difference cannot be interpreted as the isolated causal effect of identity leakage.

Conditional uncertainty and partition stability

We obtained percentile intervals from 2,000 stratified bootstrap resamples of patients, preserving the numbers of short- and nonshort-PFI patients. Each paired contrast used the same resampled patients across splitting strategies, models and repeat-level prediction arrays. We calculated AUC using a weighted Mann–Whitney U statistic. In the examples checked, this matched scikit-learn AUC recomputed on duplicated patient indices to numerical precision. We also calculated Bonferroni-adjusted conditional percentile bounds for the three primary cohort comparisons and examined variation across the 20 partitions.

These intervals were calculated from saved out-of-fold predictions. They did not include new patient selection, fold generation, feature selection or model fitting. They therefore do not represent confidence intervals for the full modeling pipeline. We did not treat the 20 repeated partitions or their folds as independent patients in significance tests. Instead, we report the size of the differences and their stability across partitions without treating 20 correlated splits as independent replications.

Patient-consistent negative controls and identity diagnostics

In each clinically labeled cohort, we permuted the binary labels among patients 100 times and assigned each resulting label to all specimens from that patient. We refitted the complete fold-specific SVM workflow, including variance-based feature selection and scaling, for every permutation and splitting strategy. Randomization removes the intended association between recurrence and expression while preserving label dependence within patient. An above-chance AUC after specimen splitting is therefore a diagnostic of identity-related predictability, not clinical discrimination [7]. We did not subtract the null AUC means from observed AUCs to construct an 'adjusted' biomarker result.

For actual-label specimen splits, we measured the proportion of held-out biopsies whose patient also contributed training samples. We also assessed whether the nearest training specimen, based on Euclidean distance after fold-specific selection and scaling, came from the same patient. We first averaged each proportion within patient and then across patients. Under independently assigned K folds, the reference probability that a held-out specimen from a patient with m specimens shares that patient with training is 1 − K^(1 − m). Because stratification changes the exact expectation, we used measured overlap proportions for the analyses (Figure S1; Table S4).

Two additional sources with artificial patient labels

GSE109934 provided 40 public NanoString RCC records from 20 identifiable pairs sampled before and after treatment [12]. We retained 730 Endogenous probes for modeling and excluded housekeeping and control probes. We also report quality-control summaries for each record (Table S5). The source study analyzed 19 pairs after excluding one for RNA quality, but the available sample metadata did not identify the excluded pair. We therefore retained all 20 public pairs for artificial-label experiments. We do not claim that this reproduces the source study's clinical population.

GSE162714 provided 77 measurements from 16 patients, including laser-captured tumor regions and tissue sections [13]. We calculated TPM from the reported counts and feature lengths. Four rows were removed because date-formatted feature symbols were ambiguous across genomic locations; 23,706 features remained for rank transformation. We did not reinterpret the reported time from diagnosis to progression as PFI. In each additional cohort, we generated approximately balanced artificial labels at the patient level and assessed 100 assignments with the fixed SVM under both splitting strategies. In total, these sources contributed 36 additional within-study patient identifiers and 117 measurements. They were used only as negative controls, not as independent clinical validation cohorts.

Sensitivity analyses and metric diagnostics

We repeated the three-classifier comparison under both splitting strategies using 88 pretreatment GSE173420 biopsies from 46 patients, including 19 with short PFI. In a separate GSE201600 sensitivity analysis, we changed the six-month event threshold. This increased the short-PFI group from eight to ten patients without changing the total of 31 patients (Tables S6 and S7). Both analyses were secondary to the prespecified primary definitions.

Scores pooled across test folds come from different fitted models and may not have directly comparable margins. We therefore calculated AUC within each test fold after aggregating scores by patient, and then averaged these fold-level values. For the negative controls, we tested 100 further patient-level label assignments using both pooled and within-fold AUC. All included test folds contained both outcome classes (Tables S8 and S9). We retained the original random-label results and the later diagnostic experiments as separate analyses.

We also compared whole-patient label permutation with independent permutation of specimen labels in the three clinical cohorts. Independent specimen permutation deliberately removes label consistency across a patient's biopsies. To compare these two null conditions without assigning an artificial patient-level outcome to inconsistent specimen labels, we used specimen-level ROC AUC weighted by the inverse of each patient's number of specimens. This ensured that each patient's specimens had a combined weight of one. We evaluated the fixed SVM under 100 assignments for each permutation unit and splitting strategy (Table S10). This weighted specimen-level AUC differs from the primary patient-aggregated AUC.

Reproducibility and interpretation limits

The supplementary package provides source links, SHA256 hashes, patient–specimen mappings, decisions on technical replicates and excluded features, fixed model settings, and fold assignments. The recorded environment used Python 3.12.10, NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1 and scikit-learn 1.9.1. Exact package requirements and commands are supplied with the code. Primary fold assignments (Table S11) and all patient-consistent null experiments (Table S12) can be inspected individually. Supplementary Information contains the full table index. These materials support independent auditing,

although we have not confirmed complete reproducibility by rerunning every analysis in a new computing environment. We did not recreate the original publications' prediction workflows or determine whether they used specimen-level splitting. TRIPOD + AI [10] informs relevant reporting considerations, but this work is a methods benchmark rather than a clinically deployable prediction model.

Results

Patient resolved benchmark populations

The three primary cohorts contained 31, 24 and 52 patients, contributing 62, 48 and 136 specimens or distinct biopsies, respectively (Fig. 1; Table 1). Each patient in the two paired cohorts contributed two specimens. In the matrisome cohort, 34 of 52 patients had multiple biopsies. Patient–specimen links are documented in the public-data audit records. The two additional cohorts contributed 20 and 16 patient identifiers and 40 and 77 measurements, respectively. These records were used only for artificial-label experiments. Overall, the five sources provided 363 measurements, of which 246 were included in the primary clinical-label analyses.

Table 1
Study populations

Source	Patients	Specimens or biopsies	Features	Analysis role
GSE201600	31	62	750	Clinical label
GSE227100	24	48	18950	Clinical label
GSE173420	52	136	1026	Clinical label
GSE109934	20	40	730	Artificial label only
GSE162714	16	77	23706	Artificial label only

Primary clinical sources use PFI < 6 months, source Early/Late recurrence categories, and PFI < 180 days, respectively. GSE173420 technical libraries of the same biological biopsy were collapsed first. The last two sources are mechanism controls with no imputed clinical endpoint. GSE109934 includes all 20 public pairs; its original paper analyzed 19 pairs after an RNA quality exclusion whose identifier was not recovered. Counts are within source and do not assert cross study donor linkage.

Specimen splitting increased the primary performance estimate

In GSE201600, mean AUC across the three classifiers was 0.621 with specimen splitting and 0.495 with patient grouping, a difference of 0.126 (conditional 95% bootstrap interval 0.084–0.171). The corresponding AUCs were 0.827 and 0.488 in GSE227100, giving a difference of 0.339 (0.241–0.450). In GSE173420, AUCs were 0.830 and 0.693, a difference of 0.137 (0.087–0.197; Fig. 2; Table 2). Bonferroni-adjusted conditional intervals remained above zero for all three primary comparisons.

Table 2
Primary validation contrast

Source	Patients	Sample split AUC	Patient split AUC	AUC gap and conditional 95% interval
GSE201600	31	0.621	0.495	0.126 (0.084 to 0.171)
GSE227100	24	0.827	0.488	0.339 (0.241 to 0.450)
GSE173420	52	0.830	0.693	0.137 (0.087 to 0.197)

Estimates average AUCs across three fixed classifiers and 20 repeated threefold partitions. Same patient scores were averaged before each AUC. Gaps equal sample split minus patient split. Intervals use 2,000 stratified patient bootstrap resamples conditional on fixed out of fold predictions; full pipeline retraining uncertainty is not included. Bonferroni adjusted conditional intervals and split ranges appear in the CSV. All model results, including the inconclusive GSE201600 nearest neighbor contrast, are retained in Table S1.

The size of the difference depended on the classifier. Across the nine cohort-by-model comparisons, mean AUC gaps ranged from 0.037 for one-nearest-neighbor classification in GSE201600 to 0.362 for logistic regression in GSE227100 (Table S2). The conditional interval for the smaller difference included zero. In GSE173420, patient-grouped SVM and logistic regression AUCs were still approximately 0.71. Thus, patient grouping did not remove all observed association in that cohort, although these results are internal rather than external clinical validation. The model-averaged gap was positive in 19 of 20 GSE201600 partitions and in all 20 partitions in each remaining primary cohort. We display partition-to-partition variation without treating repeats as independent observations.

Random patient labels remained predictable with specimen splitting

After permutation of clinical labels among patients, mean specimen-split SVM AUCs were 0.652, 0.830 and 0.696 in GSE201600, GSE227100 and GSE173420, respectively. The corresponding patient-grouped means were 0.504, 0.496 and 0.480 (Fig. 3; Table 3). Clinical labels had been randomized, but each patient's specimens still shared a label and their transcriptomic structure. The elevated specimen-split AUCs therefore cannot be attributed to recurrence discrimination alone.

Table 3
Patient label negative controls

Source	Sample split mean AUC	Patient split mean AUC	Patient label type
GSE201600	0.652	0.504	Shuffled clinical
GSE227100	0.830	0.496	Shuffled clinical
GSE173420	0.696	0.480	Shuffled clinical
GSE109934	0.674	0.505	Balanced artificial
GSE162714	0.913	0.445	Balanced artificial

One hundred patient consistent assignments per source, fixed RBF SVM, complete training fold preprocessing. These AUCs refer to labels with no clinical meaning. The primary metric pools out of fold patient scores; the small GSE162714 cohort has a downward grouped null bias. In the separate within fold null diagnostic, grouped means were 0.504–0.521 across all five sources (Table S8). Means should not be interpreted as clinical model accuracy or subtracted from observed AUC to create an adjusted biomarker performance.

Artificial labels produced similar results in the two additional sources. Mean specimen-split AUC was 0.674 in the 20 public GSE109934 pairs and 0.913 in GSE162714; pooled patient-grouped means were 0.505 and 0.445. The below-chance mean in the smaller spatial cohort led us to examine the AUC calculation. In a separate within-fold null analysis, patient-grouped means ranged from 0.504 to 0.521 across all five sources. For GSE162714, pooled and within-fold patient-grouped null means were 0.445 and 0.513, respectively, whereas specimen-split within-fold AUC remained 0.930. This pattern is consistent with sensitivity of the pooled null baseline to score comparability across folds. It did not explain the high specimen-split AUC (Table S9).

The permutation unit changed the apparent null baseline

Using patient-weighted specimen AUC, mean values after whole-patient label permutation and specimen splitting were 0.645, 0.789 and 0.684 in the three clinical cohorts. After independently permuting specimen labels, the corresponding means were 0.519, 0.495 and 0.488 (Fig. 4; Table 4). Thus, shuffling specimen labels can produce an apparently chance-level null even when randomized patient labels remain predictable. This difference reflects whether the within-patient measurement structure is preserved. It does not imply a new biological mechanism.

Table 4
Effect of permutation unit

Source	Whole patient permutation	Independent specimen permutation
GSE201600	0.645	0.519
GSE227100	0.789	0.495
GSE173420	0.684	0.488

Mean patient weighted specimen AUC for the sample split SVM across 100 assignments per permutation unit. Each specimen receives weight 1 divided by the number of specimens from that patient. This metric is different from the primary patient averaged score AUC. Patient consistent randomization preserves an artificial label shared across sibling biopsies, whereas specimen independent randomization intentionally destroys it. Full results under both splitting schemes appear in Table S7.

The effect persisted in pretreatment biopsies and metric controls

In the pretreatment-only GSE173420 subset, mean AUC was 0.730 with specimen splitting and 0.599 with patient grouping. The difference was 0.131 (conditional interval 0.066–0.214) and was positive on average in all 20 partitions (Fig. 5; Table 5; Table S6). The observed optimism therefore did not require mixing pretreatment and post-treatment samples. Repeated spatial biopsies alone were sufficient in this cohort. However, this comparison cannot separate the effects of sampling site, cellular composition and patient-correlated technical variation.

Table 5
Sensitivity analyses

Analysis	Patients, n	Sample AUC	Patient AUC	Gap and conditional 95% interval
GSE173420 pretreatment only	46	0.730	0.599	0.131 (0.066 to 0.214)
GSE201600 PFI at most 6 months	31	0.645	0.495	0.150 (0.100 to 0.201)

The pretreatment subset contains 88 distinct biopsies from 46 patients, including 19 short interval cases. The alternative six month boundary changes the number of short interval patients from eight to ten. These analyses retained the same models, preprocessing and repeated splitting protocol. Primary definitions remain unchanged. Within fold actual label analyses are in Table S4; full sensitivity model results are in Tables S5 and S6.

Changing the GSE201600 short-PFI threshold to at most six months increased the number of short-PFI patients from eight to ten. The mean AUC gap was then 0.150 (0.100–0.201; Table S7); the original boundary remained the primary definition. When AUC was calculated within individual folds, mean model-averaged gaps were 0.118, 0.335 and 0.154 in the three primary cohorts, with the same direction of effect (Table S8). In the actual-label specimen splits, patient-averaged overlap proportions were

approximately 0.66, 0.73 and 0.54. The corresponding nearest-training-specimen identity rates were approximately 0.17, 0.46 and 0.30 (Figure S1).

Discussion

Summary and interpretation of the main findings

Specimen-stratified cross-validation produced higher mean AUCs than patient-grouped cross-validation in all three ovarian cancer transcriptomic cohorts. With fixed classifiers and the same fold-specific preprocessing approach, the mean differences were 0.126, 0.339 and 0.137. Their size varied by model and dataset. These findings show that the unit used to split repeated-biopsy data can meaningfully alter estimates of discrimination. The differences should not be taken as estimates of the typical performance inflation across all ovarian cancer biomarker studies.

Randomized patient labels provide stronger evidence that patient-associated information can generate artificial discrimination. When we reassigned recurrence labels among patients but kept them consistent across each patient's specimens, specimen-split SVM AUCs remained high. Two additional cohorts showed the same pattern with entirely artificial labels. These results cannot reflect genuine discrimination of recurrence based on the randomized outcomes. Instead, they show that classifiers can learn patterns associated with individual patients in this validation setting. The experiments do not establish whether those patterns are mainly biological or arise from sample handling or other patient-correlated technical factors.

Contribution relative to earlier research

Earlier studies showed why validation must reflect the intended clinical population and how subject-consistent label permutation can uncover identity confounding [6, 7]. Similar differences between specimen-level and participant-level evaluation have since been reported for repeated biomechanical measurements [14]. Highly similar gene-expression samples can also compromise the independence of validation data [8]. A recent ovarian cancer NanoString preprint reported limited clinical predictability with pooled matched-sample data [15]. Its focus was related but different from the patient-consistent permutation tests used here. Our work does not introduce a new cross-validation method or identify a previously unknown source of bias. Rather, we reconstruct patient, biopsy and technical-replicate identities in accessible ovarian cancer datasets. We then quantify three clinical-label comparisons, reproduce the identity signal in two artificial-label cohorts and provide fold-level records that can be audited.

The controls help explain the AUC differences beyond a simple comparison of the two splitting strategies. Patient-consistent permutation removes the intended clinical link while retaining labels shared by a patient's biopsies. By contrast, independent specimen permutation breaks this structure and produces a different null distribution. Such a shuffle may appear reassuring even when patient-consistent labels remain predictable. This does not mean that specimen-level permutation is

inappropriate in every study. It can address a valid question when outcomes truly belong to independent specimens. Here, the clinical outcome belonged to the patient, so independent specimen permutation did not preserve the dependence relevant to identity confounding.

Relevance to ovarian cancer sampling and clinical use

Treatment timing and repeated spatial sampling create distinct, although related, validation problems. Samples collected before and after chemotherapy can reveal treatment-associated biology. However, expression measured after treatment cannot serve as a predictor for decisions made before treatment. Restricting GSE173420 to pretreatment samples removed that temporal issue but retained repeated tumor regions. The mean AUC gap remained 0.131. Thus, in this cohort, mixing treatment time points was not necessary to produce the observed difference. This analysis does not distinguish effects of tumor site, stromal composition, clonal relationships or specimen processing. Nor can its magnitude be assumed to apply to other cohorts.

Repeated tumor regions are not inherently a problem and should not be excluded without a scientific reason. They are important for describing heterogeneity within a patient. What matters is the intended prediction setting. For a model designed for patients not previously encountered, no biopsy from a test patient should enter training, feature selection or model development. Predicting a new lesion in a patient already represented during training is a different task, with a different interpretation of performance [6]. The presence of a common patient-level recurrence label does not make these prediction settings equivalent.

Why metric definition also matters

Averaging out-of-fold biopsy scores within patient prevented heavily sampled patients from contributing disproportionate weight to the AUC. It did not correct the underlying train–test overlap. The splitting unit and the unit used to calculate performance therefore need to be reported separately. We also found that pooling predictions from different fitted folds can affect the null AUC if their scores are not directly comparable. In the smallest spatial artificial-label cohort, pooled patient-grouped null AUC averaged 0.445, compared with approximately 0.513 when calculated within folds. The latter was near chance, whereas artificial labels remained highly predictable under specimen splitting. This explains an important feature of the numerical results without changing their main interpretation.

Patient-grouped AUC did not fall to chance in every analysis. In GSE173420, the SVM and logistic regression still yielded AUCs of approximately 0.71. These associations should not be dismissed solely because specimen splitting gave higher estimates. They should not, however, be regarded as evidence of a clinically ready recurrence predictor. Validation remained internal to a small, selected cohort, the measurements covered only a restricted matrisome panel, and the main analysis combined treatment time points. We did not assess calibration, decision benefit or transfer to another center or assay.

Strengths, limitations and implications

Several features strengthen the interpretation of this benchmark. We reconstructed patient and biopsy units, used the same prespecified classifiers in both validation schemes, performed separate label-randomization experiments and supplied detailed audit records. These steps make the origin of each prediction and its validation partition easier to examine. Important limitations remain. The clinical analyses covered only 107 within-study patients across three cohorts, and the cohorts differed in outcome definitions, treatment settings and gene coverage. They were selected for public accessibility rather than through a systematic search. Decisions about technical replicates and censored follow-up also depend on the accuracy of the original annotations. For GSE109934, the unidentified RNA-quality exclusion means our public-sample set is not identical to the cohort analyzed in the source paper.

The splitting strategies differ in more than patient overlap. They also change which independent patients contribute to training and may alter class distributions across groups and folds. We therefore cannot attribute the entire observed AUC gap solely to patient identity. The randomized-label controls show that patient-associated structure can generate optimistic performance, but they do not quantify its exact share of the clinical-label difference. Our patient-bootstrap intervals are conditional on saved predictions and do not include new fold assignment, feature selection or model fitting. They consequently capture less uncertainty than a bootstrap that repeats the full analysis. The correlated partitions describe stability and do not increase the number of independent patients. These limitations preclude causal claims about the complete AUC gap or conclusions about the prevalence of leakage and clinical benefit.

Future studies using repeated ovarian cancer biopsies should preserve recoverable patient and biopsy identifiers. Reports should specify when specimens were collected in relation to the intended clinical decision, describe fold-specific preprocessing and state whether patients are shared across training and test sets. For patient-level outcomes, label-randomization controls should preserve the patient as the unit of permutation. External validation in a separate patient population, with comparable outcomes and measurements, is necessary to assess transportability. The sample should also be large enough for precise estimates of discrimination and calibration [16]. TRIPOD + AI and PROBAST + AI offer useful frameworks for reporting and bias assessment [10, 17]. Our benchmark supports these practices but cannot replace prospective clinical validation of a prediction model.

Conclusions

Across three public ovarian cancer transcriptomic cohorts, comprising 107 within-cohort patients and 246 specimens or biological biopsies, specimen-based cross-validation produced mean AUCs 0.126, 0.339 and 0.137 higher than patient-grouped cross-validation. The same direction was found in the pretreatment spatial subset and in sensitivity analyses using other score calculations and outcome boundaries.

Artificial labels assigned consistently within patient remained predictable under specimen splitting in all five datasets. In contrast, patient-grouped within-fold null AUCs were near chance. Permuting specimen labels independently largely concealed this dependence. These results show why splitting strategy, label-

permutation unit and AUC calculation all matter when multiple biopsies share one patient's clinical outcome.

This study offers an auditable ovarian cancer example of established methods for detecting identity confounding across different repeated-biopsy settings. It does not show that the source studies used inappropriate validation, establish a clinically validated prediction model or quantify treatment benefit. For models intended to generalize to previously unseen patients, patient-disjoint evaluation and negative controls that preserve within-patient dependence remain necessary safeguards.

Declarations

Data availability and ethics

We analyzed publicly available GEO datasets GSE201600, GSE227100, GSE173420, GSE109934 and GSE162714, together with published clinical supplements where applicable. Supplementary Code and Inputs provides source URLs, SHA256 checksums, patient-to-specimen mappings, executable analysis scripts and the processed public files used here. Full numerical results are included in Supplementary Data. No controlled-access sequencing data were used.

Ethics statement

This study reanalyzed previously released public datasets and published clinical annotations. No participants were recruited or contacted, and no new clinical interventions or laboratory experiments were conducted.

Patient overlap benchmark tables

Main numerical results for the computational validation study. Complete model, partition and negative control outputs are supplied as CSV tables in Supplementary Data.

Author Contribution

J.X. performed data curation, formal analysis, methodology, and visualization, and wrote the original draft of the manuscript. Y.L. supervised the study and critically reviewed and edited the manuscript. Both authors read and approved the final version of the manuscript.

Data Availability

The transcriptomic datasets analyzed in this study are publicly available from the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE201600, GSE227100, GSE173420, GSE109934, and GSE162714 (https://www.ncbi.nlm.nih.gov/geo/). Additional clinical information was obtained from the supplementary materials of the original publications. The derived numerical results, patient-to-specimen mappings, cross-validation fold assignments, and analysis code are provided in the Supplementary Materials accompanying this article. No new patient samples or sequencing data were generated.

References

1. Zheng X, Wang X, Cheng X, et al. Single-cell analyses implicate ascites in remodeling the ecosystems of primary and metastatic tumors in ovarian cancer. Nat Cancer. 2023;4:1138–56. https://doi.org/10.1038/s43018-023-00599-8.
2. Denisenko E, de Kock L, Tan A, et al. Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones. Nat Commun. 2024;15:2860. https://doi.org/10.1038/s41467-024-47271-y.
3. James NE, Woodman M, De La Cruz P, et al. Adaptive transcriptomic and immune infiltrate responses in the tumor immune microenvironment following neoadjuvant chemotherapy in high grade serous ovarian cancer reveal novel prognostic associations and activation of pro-tumorigenic pathways. Front Immunol. 2022;13:965331. https://doi.org/10.3389/fimmu.2022.965331.
4. Adzibolosu N, Alvero AB, Ali-Fehmi R, et al. Immunological modifications following chemotherapy are associated with delayed recurrence of ovarian cancer. Front Immunol. 2023;14:1204148. https://doi.org/10.3389/fimmu.2023.1204148.
5. Pietilä EA, Gonzalez-Molina J, Moyano-Galceran L, et al. Co-evolution of matrisome and adaptive adhesion dynamics drives ovarian cancer chemoresistance. Nat Commun. 2021;12:3904. https://doi.org/10.1038/s41467-021-24009-8.
6. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6:1–9. https://doi.org/10.1093/gigascience/gix019.
7. Chaibub Neto E, Pratap A, Perumal TM, et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digit Med. 2019;2:99. https://doi.org/10.1038/s41746-019-0178-x.
8. Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression data. iScience. 2022;25:104788. https://doi.org/10.1016/j.isci.2022.104788.
9. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4:100804. https://doi.org/10.1016/j.patter.2023.100804.
10. 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.
11. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.

12. Arend RC, Londoño AI, Montgomery AM, et al. Molecular response to neoadjuvant chemotherapy in high-grade serous ovarian carcinoma. Mol Cancer Res. 2018;16:813–24. https://doi.org/10.1158/1541-7786.MCR-17-0594.
13. Artibani M, Masuda K, Hu Z, et al. Adipocyte-like signature in ovarian cancer minimal residual disease identifies metabolic vulnerabilities of tumor-initiating cells. JCI Insight. 2021;6:e147929. https://doi.org/10.1172/jci.insight.147929.
14. Karbalaie A, Abtahi F, Häger CK. Participant-aware model validation for repeated-measures data: comparative cross-validation study. JMIR AI. 2026;5:e87728. https://doi.org/10.2196/87728.
15. Van Kleunen LB, Bowman G, Stockman SE et al. On the predictability of progression-free survival in ovarian cancer from NanoString gene expression data. bioRxiv [preprint]. 2026 Apr 24:2026.04.22.719856. https://doi.org/10.64898/2026.04.22.719856
16. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40:4230–51. https://doi.org/10.1002/sim.9025.
17. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505.

Figures

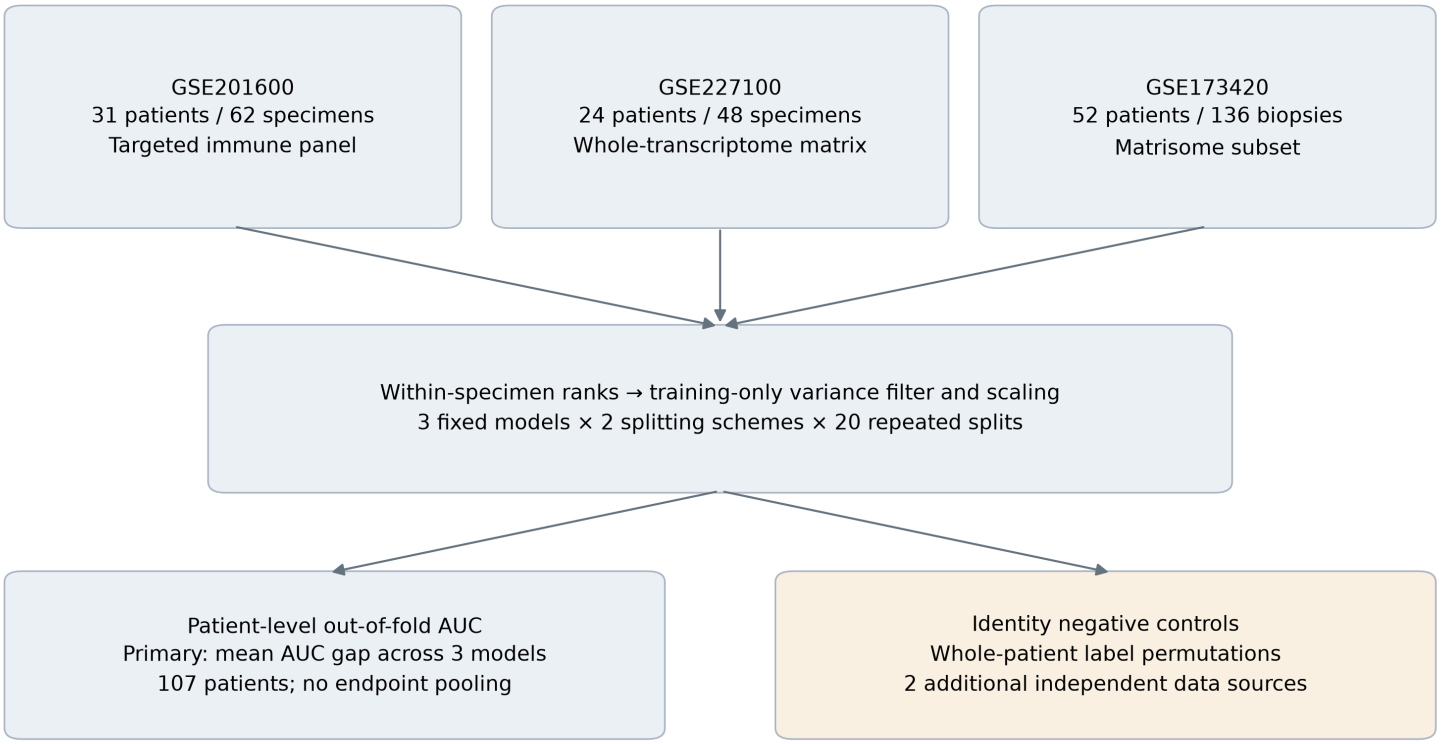
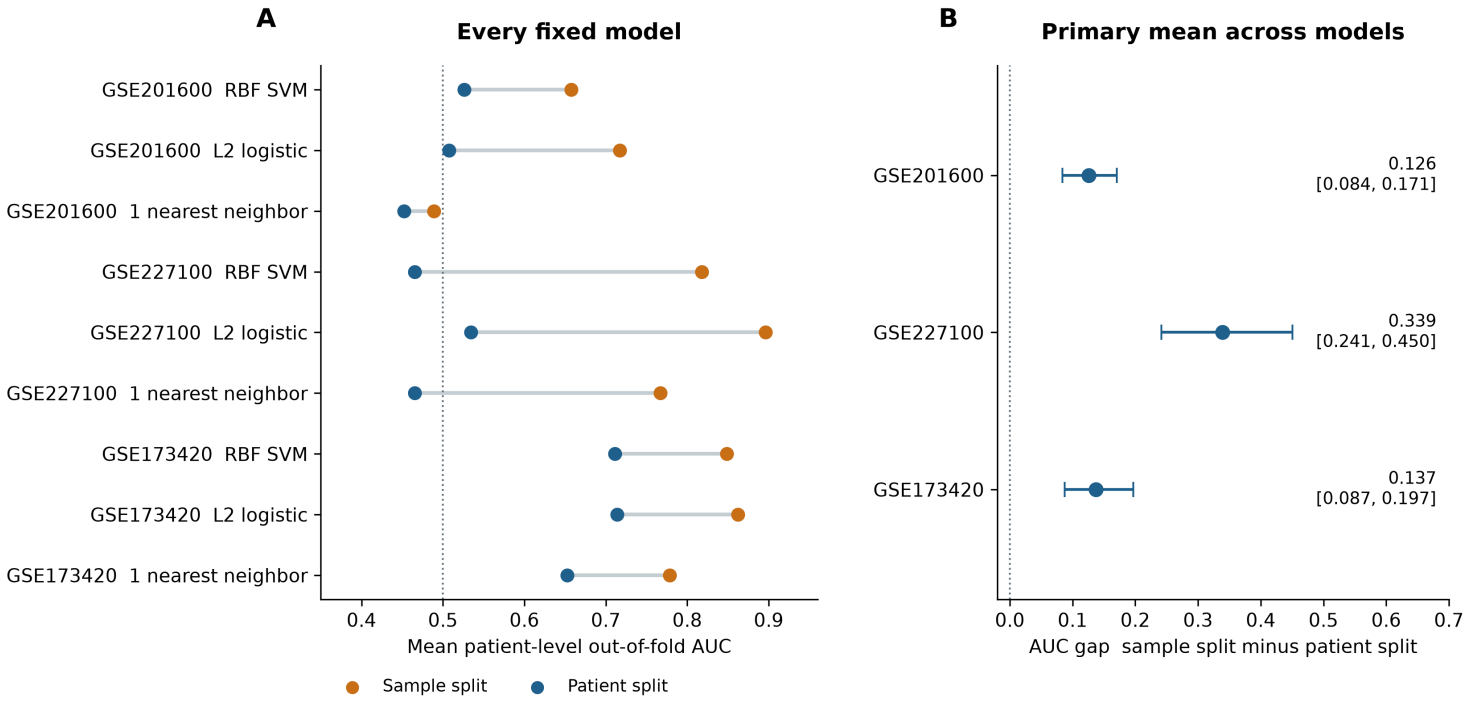


Figure 1

Patient-resolved benchmark design. We analyzed the three clinical cohorts separately; together, they included 107 within-cohort patient identifiers. Three fixed classifiers were evaluated under two splitting strategies, each repeated 20 times. The two additional sources were used for artificial-label controls, not clinical validation.



Intervals: 95% patient bootstrap, conditional on fixed out-of-fold predictions; not full retraining intervals.

Figure 2

Model performance with specimen-level and patient-grouped splitting. Panel A shows the nine cohort-by-classifier comparisons, each averaged over 20 partitions. Orange indicates specimen splitting and blue indicates patient grouping. Panel B shows the primary mean AUC difference averaged across the three classifiers. Error bars represent conditional 95% percentile intervals from patient bootstrap resampling of fixed out-of-fold predictions. They do not capture uncertainty from retraining the full pipeline.

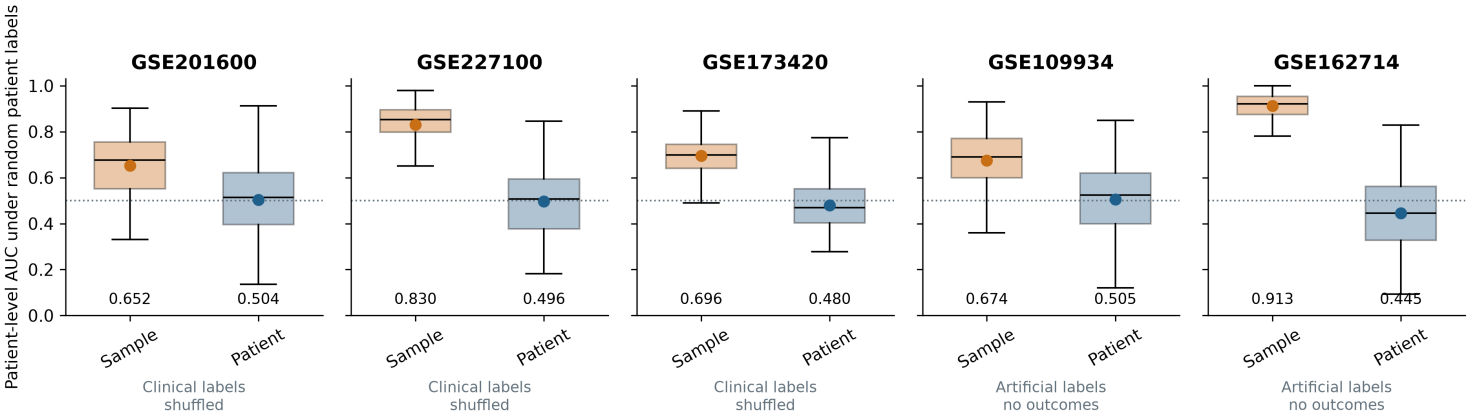
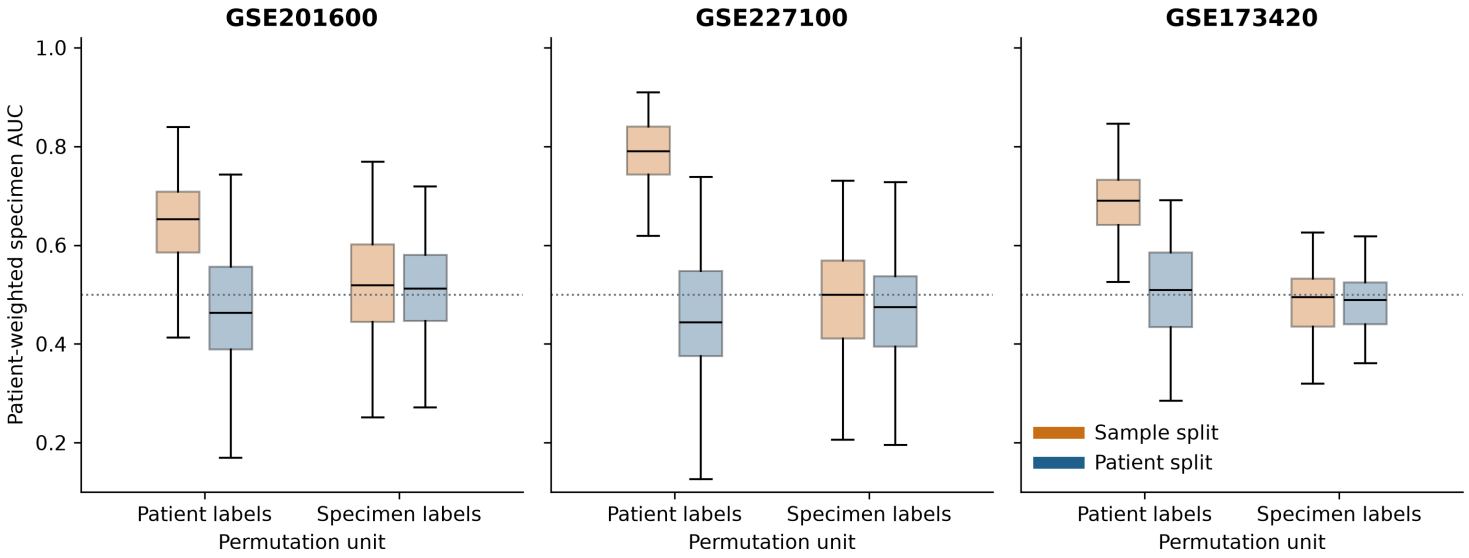


Figure 3

Patient-label negative controls. Each box summarizes 100 SVM runs in which labels were assigned consistently within patient. Clinical labels were permuted in the first three cohorts, whereas the last two

cohorts received balanced artificial labels. Boxes indicate the interquartile range; black lines mark medians, whiskers extend to observations within 1.5 interquartile ranges, and colored points show means. Outlier symbols are omitted from the plot, but all runs are reported in Table S12. These labels have no clinical interpretation. For GSE162714, mean patient-grouped null AUC was 0.445 when scores were pooled across folds and 0.513 when AUC was calculated within folds. This difference is consistent with limited comparability of scores from separately fitted models (Table S9).



Diagnostic metric differs from the primary patient-aggregated AUC. Each patient has total specimen weight 1.

Figure 4

Effect of permutation unit. Permuting labels among patients preserves the shared label structure of repeated measurements, whereas permuting specimens independently breaks it. AUC was calculated at the specimen level, weighting each specimen by the inverse of its patient's specimen count. This differs from the patient-aggregated metric used in the primary analysis. Each box represents 100 assignments. Colors denote the splitting strategy; box and whisker conventions are as in Figure 3.

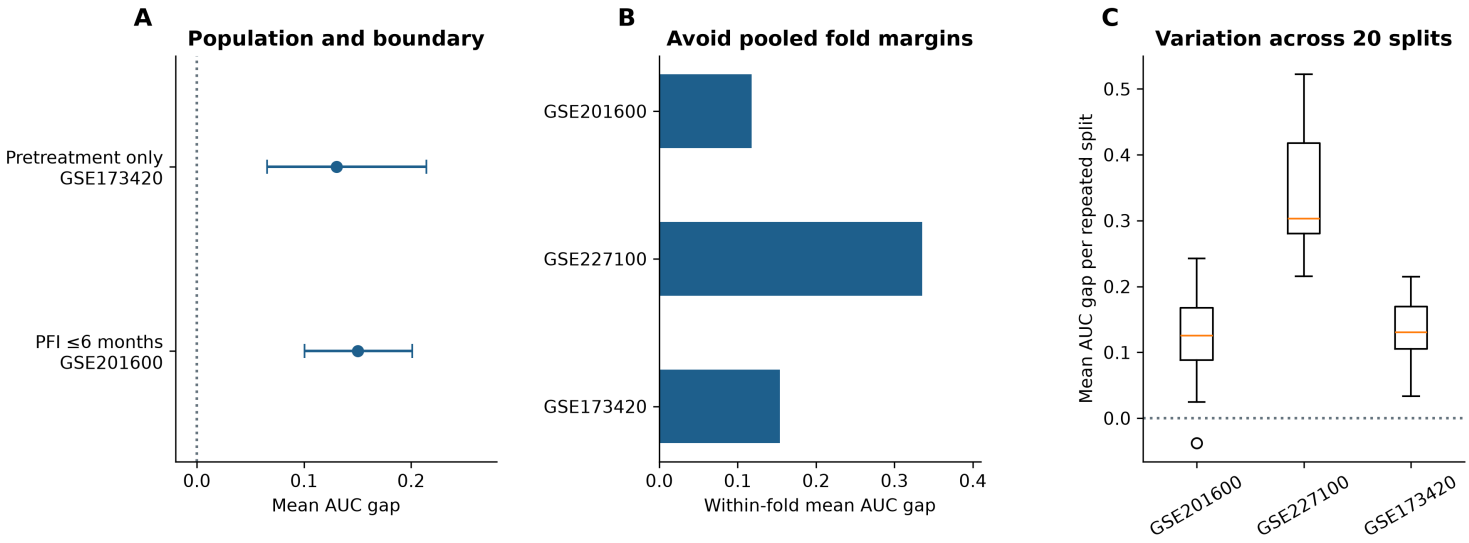


Figure 5

Sensitivity analyses. Panel A shows mean AUC differences in pretreatment GSE173420 biopsies and under the alternative six-month event boundary in GSE201600. Error bars indicate conditional 95% patient-bootstrap intervals. Panel B shows mean AUC differences calculated within test folds and then averaged over folds and classifiers. Panel C shows the distribution of each cohort's model-averaged AUC difference across 20 repeated partitions, including the one negative partition in GSE201600. Partitions do not represent independent patients.

Supplementary Files

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

- 06SupplementaryData.zip
- SupplementaryInformation.docx