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.

Independent prognostic value of a 4-gene transcriptomic signature in lung adenocarcinoma across five external cohorts

Md. Abul Hossain
Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh
Sajid Nur Chowdhury
sajidnoorchowdhury@gmail.com

Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh
Samiul Islam
Department of Statistics, University of Rajshahi
Md. Abdul Latif
Department of Statistics, University of Rajshahi

Research Article

Keywords: lung adenocarcinoma, prognostic signature, survival modelling, external validation, data leakage, TRIPOD + AI, restricted mean survival time, transcriptomics

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10490641/v1

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

Additional Declarations: No competing interests reported.

Independent prognostic value of a 4-gene transcriptomic signature in lung adenocarcinoma across five external cohorts

Md. Abul Hossain[1], Sajid Nur Chowdhury[2*], Samiul Islam[3], Md. Abdul Latif[4]

[1,2*]Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh
[2,4] Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

[*]Corresponding author name and email address: Sajid Nur Chowdhury-sajidnoorchowdhury@gmail.com

Abstract

Background. Transcriptomic prognostic signatures for lung adenocarcinoma are frequently reported with in-sample discrimination and without independent external validation, and are prone to feature-selection leakage and undisclosed model selection. We asked whether a pre-specified, leakage-free signature adds prognostic value beyond standard-of-care staging, and whether it transports to independent cohorts.

Methods. In 504 TCGA lung adenocarcinoma patients (182 deaths), eight survival models were evaluated by repeated 5-fold nested cross-validation, with feature selection strictly inside folds and all metrics computed out-of-fold. Five classical and tree-based learners formed a confirmatory tier eligible for deployment; three neural models formed a pre-declared exploratory tier; a rank-average ensemble was the pre-specified primary model. A 4-gene signature (DKK1, MS4A1, CCL20, and one long non-coding RNA) was locked and combined with stage, age and sex in a cross-validated integrated model, then applied unchanged to five independent cohorts (1,254 patients, 475 events) and pooled by meta-analysis. Discrimination, calibration, decision-curve analysis, and reclassification were reported with 95% confidence intervals.

Results. Internal discrimination was modest (confirmatory-tier Harrell's concordance 0.607–0.623; deployed Cox 0.613, 0.596–0.627); two exploratory neural models did not exceed 0.60. The signature was independently prognostic after adjustment for stage, age and sex (hazard ratio 1.34 per standard deviation, 1.15–1.56; likelihood-ratio $p = 2 \times 10^{-4}$), robust to multiple imputation and to exclusion of squamous-confounded samples. Combining the molecular score with staging raised discrimination from 0.645 to 0.676, a leakage-free gain of 0.031. Proportional hazards were

violated for the risk-group comparison, where high-risk patients lost 293 days (169–418) of restricted mean survival over 5 years. External discrimination was homogeneous but modest (pooled Harrell's concordance 0.606, 0.578–0.634; $I^2 = 0\%$), competitive with a published 14-gene signature (0.616) at one-third the gene count. All 25 pre-specified reporting-guideline items were met.

Conclusions. A pre-specified, leakage-free 4-gene signature carries independent, modest prognostic value beyond stage, age and sex, and its combination with staging (Harrell's concordance 0.676) is the predictor we advance. Molecular-signature discrimination alone (approximately 0.60–0.62) is not clinically actionable, and external transport, though homogeneous, is modest and driven partly by one weak, event-dominant cohort. These are prognostic, not predictive, associations from a single retrospective cohort; prospective validation is required before clinical application.

Keywords: *lung adenocarcinoma; prognostic signature; survival modelling; external validation; data leakage; TRIPOD+AI; restricted mean survival time; transcriptomics*

Introduction

Lung adenocarcinoma is the most common histological subtype of lung cancer and a leading cause of cancer death worldwide[1], [2], [3] . Anatomic staging by the American Joint Committee on Cancer (AJCC) TNM system remains the principal determinant of prognosis and of adjuvant-therapy decisions after resection [4], [5]. Staging is, however, an incomplete predictor of individual outcome: patients with the same stage frequently follow markedly different survival trajectories, and a substantial fraction of early-stage patients relapse despite complete resection [6]. This residual, stage-independent variation in outcome has motivated a long search for molecular markers that could refine prognosis beyond what anatomic staging provides, and potentially help identify the early-stage patients most likely to benefit from adjuvant treatment.

Bulk tumour transcriptomic profiling is an attractive substrate for such markers: it is widely available, captures proliferative, metabolic and immune programmes relevant to tumour aggressiveness, and can be summarized into compact multi-gene risk scores [7]. Over the past two decades, dozens of transcriptomic prognostic signatures for lung adenocarcinoma and non-small-cell lung cancer have been proposed, many derived from public resources such as The Cancer

Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) cohorts [8], [9], [10], [11], [12], [13], [14], [15]. A small number, such as the Kratz et al. 14-gene assay for resected non-squamous non-small-cell lung cancer, have undergone independent external validation [16].

Despite this volume of work, the field has a well-documented reproducibility problem, and the practical clinical impact of transcriptomic prognostic signatures in lung adenocarcinoma has remained limited [17], [18]. Several recurring methodological weaknesses contribute. Feature selection is frequently performed on the full dataset before, rather than inside, cross-validation folds, which leaks outcome information into the selected gene set and inflates apparent performance [19], [20]. The model or signature reported is often chosen after inspecting held-out performance, without a pre-specified primary analysis, so that the published result is the best of many undisclosed comparisons. Concordance statistics are sometimes reported without checking the orientation of the risk score, which can silently invert the metric. Many signatures are validated in a single convenience cohort, or not externally at all, and are reported with discrimination alone, without calibration or an explicit comparison against the standard-of-care staging they are meant to improve upon. Each of these issues, individually, tends to bias reported performance in the optimistic direction, and each is difficult to detect from a finished manuscript.

We therefore approached transcriptomic prognostic modelling in lung adenocarcinoma primarily as a question of methodological rigor rather than of signature discovery. Using the TCGA lung adenocarcinoma cohort for development and five independent GEO cohorts for validation, we built a fully leakage-free pipeline in which feature selection is performed strictly inside cross-validation folds, the primary-endpoint model and the deployed gene signature are pre-specified before any out-of-fold result is inspected, discrimination is reported with two concordance estimators and an explicit orientation check, and the frozen signature is applied unchanged across five external cohorts spanning two platforms with formal between-cohort heterogeneity accounting. Throughout, we report calibration alongside discrimination, compare the signature against standard-of-care staging by a likelihood-ratio test and decision-curve analysis, benchmark it head-to-head against a published signature, and report negative and null findings — including the failure of deep survival models to add value, and the absence of incremental clinical utility in late-stage disease — rather than omitting them.

Our aim was not to claim a superior or novel signature, but to establish, under conditions that a skeptical reader can audit, whether a pre-specified transcriptomic signature adds prognostic information beyond stage, age and sex, and whether any such signal transports honestly to independent cohorts. We present the integrated clinical-plus-molecular model, not the transcriptomic signature in isolation, as the study's central result, and we are explicit that the absolute discrimination achieved does not, on its own, support clinical deployment. The contribution we intend is a reproducible, externally validated, and reporting-guideline-compliant template for transcriptomic prognostic modelling, together with an honest account of what such a signature can and cannot currently deliver in lung adenocarcinoma.

Methods

Study population and discovery cohort

The discovery cohort was the TCGA lung adenocarcinoma (TCGA-LUAD) dataset, obtained as STAR-aligned raw RNA-seq counts. We retained primary-tumour samples only (sample type code 01), one sample per patient (the sample with the highest sequencing depth where multiple were available), and patients with follow-up time greater than zero. Pseudoautosomal-region (PAR_Y) genes and low-count genes were removed using edgeR's [21] filterByExpr, an outcome-blind filter applied before any survival information was used. Applying these criteria to 592 input samples yielded 531 primary tumours, 527 with positive follow-up, and 504 unique patients after deduplication (182 deaths; 63.9% censored). Overall survival was the sole endpoint.

Normalization and transformation

For machine-learning and exploratory analyses, counts were transformed with the DESeq2 [22] variance-stabilizing transformation (VST, blind=TRUE). For descriptive differential-expression analysis, limma-voom [23], [24] was used; edgeR log-counts-per-million was retained for sensitivity checks. The VST and the low-count filter (below) were each fit once on the full discovery cohort, consistent with their outcome-blind design (blind=TRUE uses no design or group information, so no survival label enters either step); this is a global, unsupervised normalization choice, standard in RNA-seq pipelines, and is distinct from the strictly supervised steps described next. All outcome-dependent steps — the univariate Cox screen, elastic-net feature

selection, and every model fit — were performed strictly inside the cross-validation training folds, using only the VST-transformed values of the training-fold samples for that iteration.

Feature selection (performed strictly within folds)

All supervised feature selection was performed inside cross-validation folds, never on the full dataset, to prevent selection leakage. Within each outer training fold, we applied a variance pre-filter, a vectorized univariate Cox score test with Benjamini–Hochberg false-discovery-rate control, and an elastic-net Cox [25] model. A bootstrap stability-selection procedure (fixed regularization parameter, 1,000 resamples) identified a consensus gene set robust to resampling. Because gene presence on an external platform is metadata rather than an outcome, cross-platform measurability was additionally used as a leakage-free selection criterion, so that the deployed signature is measurable across the external validation cohorts rather than requiring cohort exclusion for missing genes.

Survival models and pre-specified model tiers

Eight survival models were evaluated: Cox proportional-hazards regression[26], elastic-net Cox, random survival forest[27], an extension of the Breiman random forest to right-censored outcomes[28] and XGBoost with a Cox objective[29], three pure-R neural Cox models (DeepSurv[30], CoxNNet[31], and a Cox partial-likelihood multilayer perceptron trained by analytic-gradient L-BFGS), and a rank-average Ensemble. The Ensemble is a fixed, zero-parameter combination of whichever base models return finite predictions for a given fold; it involves no additional fitting or tuning and therefore introduces no additional leakage.

Models were assigned to two pre-specified tiers before any out-of-fold result was inspected. A confirmatory tier (Cox, elastic-net Cox, random survival forest, XGBoost, and the Ensemble) comprised classical and tree-based survival learners with tight out-of-fold discrimination intervals, and was the only tier eligible for deployment, for the decision-curve and nomogram analyses, and for any primary claim. An exploratory tier (DeepSurv, CoxNNet, MLP) comprised neural Cox benchmarks included solely to test whether deep survival models add value at this sample and event count; these were pre-declared ineligible for deployment and are reported as a secondary, honestly bounded (wide-confidence-interval) negative result. The Ensemble was pre-specified as

the primary-endpoint model. Hyperparameters for the random survival forest and XGBoost were tuned per fold by a small grid search on training-fold performance only, never on the outer test fold.

Cross-validation and performance metrics

Models were evaluated by repeated (×5) 5-fold nested cross-validation, with all reported metrics computed on out-of-fold risk predictions. Discrimination was quantified by Harrell's concordance index[32], [33] and Uno's concordance index[34], [35], each with bootstrap 95% confidence intervals. Harrell's C was computed with survival::concordance using the documented orientation for risk scores (higher score corresponding to shorter survival) and was cross-checked against Uno's C for directional agreement before reporting, guarding against the sign-orientation error that can otherwise invert a concordance statistic. We additionally report time-dependent AUC at 1, 3 and 5 years, calibration by the integrated calibration index[36], [37] and predicted-versus-observed curves, and decision-curve analysis[38]. As a sensitivity analysis on the confidence-interval method, we also report a patient-level bootstrap (resampling patients with replacement).

Independent prognostic value over standard-of-care staging

To test whether the signature adds prognostic information beyond standard-of-care staging, a clinical-only Cox model (age, AJCC stage, sex) was compared with a clinical-plus-molecular model that added the leakage-free out-of-fold molecular risk score, by a likelihood-ratio test. Both C-indices in this comparison were bootstrap optimism-corrected (Harrell's method), so that the clinical-only and clinical-plus-signature models are compared on the same leakage-free footing rather than pitting two in-sample apparent C-indices against each other. Univariable and multivariable Cox regressions report the adjusted hazard ratio of the out-of-fold signature score (per standard deviation), with and without adjustment for stage, age and sex, together with a global Schoenfeld proportional-hazards check on the adjusted model. Decision-curve analysis and categorical net reclassification improvement[39] versus the clinical-only model are reported alongside discrimination. Decision-curve analysis was additionally stratified by stage group (early I–II versus late III–IV), since the adjuvant-therapy decision context differs by stage; a clinically relevant threshold range of approximately 10–30% was referenced for early-stage decisions.

Proportional hazards and restricted mean survival time

Two distinct Schoenfeld proportional-hazards tests are reported and are not conflated: a global test on the multivariable clinical-plus-signature Cox model, and a separate test on the isolated high-versus-low risk-group Kaplan–Meier[40] comparison. Where the proportional-hazards assumption is violated, the hazard ratio is interpreted as a time-averaged effect, and an assumption-free effect size — the restricted mean survival time (RMST)[41] difference — is reported alongside the risk-group hazard ratio. The RMST truncation time was pre-specified as the maximum evaluation horizon (5 years, 1,825 days), fixed before any survival curve was inspected and capped at the last observed event time only where follow-up was shorter; it was not selected from the data. A permutation-based log-rank[42] test (2,000 relabelings) corroborated the asymptotic log-rank result for the risk-group comparison, which is appropriate because the risk-group label is itself derived from a fitted model.

Integrated clinical + molecular model

Clinical-only, molecular-only, and integrated (clinical-plus-molecular) Cox models were compared under repeated patient-level cross-validation on identical folds. The molecular term was the out-of-fold risk score and the mixing coefficients were themselves cross-validated, so the integrated C-index is leakage-free. The integrated model is presented as the deployable clinical predictor. A nomogram[43] was constructed on the integrated model conditional on its clearing a pre-specified minimum discrimination gate (Uno's $C \geq 0.60$); the gate decision and the model used are recorded in the corresponding results table.

Deployed signature and sensitivity analyses

The deployed gene set was selected to be robust under both bootstrap stability selection and nested-CV fold recurrence, and to be measurable across the external validation platforms. For final submission the deployed signature was locked to four pre-specified genes. A pre-specified signature-sensitivity analysis compared candidate gene sets under identical repeated cross-validation and recommended the smallest whose discrimination was statistically non-inferior (pre-specified non-inferiority margin 0.02 Harrell's C). Robustness to histological confounding was assessed by recomputing discrimination after excluding the top decile of an inferred squamous

score, and by correlating each deployed gene with that score. Robustness to missing clinical data was assessed by multiple imputation by chained equations[44] (m=10, survival-aware with a Nelson–Aalen cumulative-hazard term) over all 504 patients, compared with the complete-case multivariable estimate.

Model interpretability and network prioritization

Feature importance for the deployed portable Cox model was quantified by model-faithful permutation importance (the drop in Harrell's C when each gene is permuted). XGBoost SHAP values[45], with bootstrap stability (importance 95% confidence intervals and rank stability), were retained as an exploratory tree-model view rather than as the explanation of the deployed model. A co-expression network among the deployed genes was built from absolute Spearman correlations of VST expression, retaining edges above the 60th percentile of the off-diagonal absolute-correlation distribution (a data-driven density threshold rather than a fixed p-value); leading-eigenvector centrality was computed on the resulting weighted adjacency and fused with SHAP importance and stability frequency into an integrated per-gene priority score. A Hallmark pathway-crosstalk network (pathway co-variation with absolute correlation above the 80th percentile) was used to surface coordinated programmes, and is reported as descriptive context.

Systems-biology characterization

Functional characterization comprised over-representation analysis (GO[46], KEGG[47], Reactome, and Hallmark[48] gene sets), gene-set enrichment analysis (fgsea)[49], GSVA[50] pathway scoring compared between out-of-fold risk groups, and immune cell-type deconvolution[51], [52]. Unsupervised exploration (principal-component analysis, UMAP[53], t-SNE[54], consensus clustering, and non-negative matrix factorization[55] with rank selection by cophenetic and dispersion coefficients) is reported as exploratory context and underlies no primary claim.

External validation

A portable, cohort-standardized linear Cox signature with frozen training-derived coefficients and a training-median risk cutoff[56] was applied, unchanged, to five independent cohorts (GSE68465, GSE31210, GSE30219, GSE81089, and GSE72094), spanning four microarray cohorts and one

RNA-seq cohort. The frozen signature model was ridge-penalized ($glmnet, \alpha = 0, \lambda.1se$), rather than an unpenalized maximum-likelihood fit, to favour out-of-cohort generalization over in-sample fit; an unpenalized Cox fit was used only as a fallback if the ridge fit errored. Each external cohort was standardized (z-scored) on its own mean and standard deviation to remove platform offset, a per-cohort approach chosen over a joint batch-correction method such as ComBat[57],which would require pooling discovery and validation expression data and risks re-introducing the leakage the restof the pipeline is designed to avoid then scored, Kaplan–Meier-stratified, and calibrated (predicted versus observed absolute risk at 1, 3 and 5 years; calibration slope from the refit linear predictor). Where the deployed antisense lncRNA (BCAN-AS1) was not measurable on a microarray platform, a reduced overlap-genes-only portable Cox model was refit on the discovery cohort and applied, rather than imputing the missing measurement.

Per-cohort Harrell's C was pooled by both fixed-effect and DerSimonian–Laird random-effects inverse-variance meta-analysis[58], with between-cohort heterogeneity quantified by Cochran's Q, I^2[59], and τ^2. Transportability — the comparison of internal versus pooled-external discrimination, and whether the pooled external confidence interval excludes 0.5 — was pre-specified as the primary external-validation result. A leave-one-cohort-out analysis assessed the influence of individual cohorts on the pooled estimate. As an external benchmark, the deployed signature was compared against the Kratz et al. 14-gene resected-NSCLC prognostic gene list (12 of 14 genes mappable): coefficients for both signatures were re-fit on the discovery cohort using the identical VST + z-score + ridge-Cox pipeline and scored through the identical five-cohort external-validation machinery. This benchmark holds gene selection versus pipeline methodology constant and evaluates the Kratz gene list under our normalization; it does not replicate Kratz et al.'s original qPCR ΔCt assay, which is normalized against three specific housekeeping genes (ESD, TBP, YAP1) rather than cohort-level z-scoring, and the two should not be read as measuring identical things.

Reporting-guideline compliance and reproducibility

Compliance against TRIPOD+AI (BMJ 2024;385:e078378), PROBAST+AI (BMJ 2025;388:e082505), and REMARK was machine-audited: each checklist item's status was derived programmatically from whether the corresponding output artifact existed on disk for the run, and

any unmet item would be flagged. All 25 pre-specified items on the tracked checklist were met. This tracked checklist does not include a separate item for stratified subgroup performance (e.g., by sex or age), which the 2024 TRIPOD+AI statement added as part of its fairness domain; subgroup discrimination was not separately assessed in this study, and we report this as an explicit, named gap rather than as part of the 25/25 figure. A single global seed plus deterministic per-repeat and per-fold seeds make every stochastic step reproducible; all out-of-fold predictions were produced by the pipeline itself rather than reconstructed from externally saved folds; sessionInfo and a full package manifest were captured at run time; and a cache-cleared re-run from raw counts reproduces the reported numbers exactly. Analyses were conducted in R 4.5.3; key packages included survival (3.8-6), glmnet (5.0), and xgboost (3.2.1.1), with the complete versioned manifest provided alongside the code.

Results

We developed and validated a leakage-free transcriptomic prognostic model for lung adenocarcinoma (LUAD) using repeated nested cross-validation, benchmarked it against a published signature, and tested transportability across five independent external cohorts. Model selection, the primary endpoint, and the deployed gene signature were pre-specified before any out-of-fold result was inspected; deviations from pre-specification are stated explicitly where they occur, consistent with TRIPOD+AI and PROBAST+AI reporting principles. We present the integrated clinical + molecular model, not the transcriptomic signature in isolation, as the study's central result, and we report negative and null findings alongside positive ones throughout.

Cohort and study flow

Of 592 TCGA-LUAD RNA-seq samples considered, 531 were primary tumours (sample type 01), 527 had positive follow-up time, and 504 remained after deduplication to one sample per patient (highest sequencing depth) (Supplementary Table S1). The analysis cohort comprised 504 patients with 182 deaths recorded (63.9% censored); median age was 66 years (range 38–88, 487 patients with age recorded), 266 (52.8%) were female, and AJCC stage distribution was I: 266 (52.8%), II: 119 (23.6%), III: 81 (16.1%), IV: 24 (4.8%), with stage missing for 14 patients (Supplementary Table S2). Unsupervised exploration (principal-component analysis, UMAP, t-SNE and consensus clustering) did not reveal a dominant substructure independent of the survival model; NMF

consensus clustering supported an optimal rank of 2 (cophenetic coefficient 0.996), and the resulting clusters are reported as exploratory context rather than as a basis for any primary claim (Fig. 1).

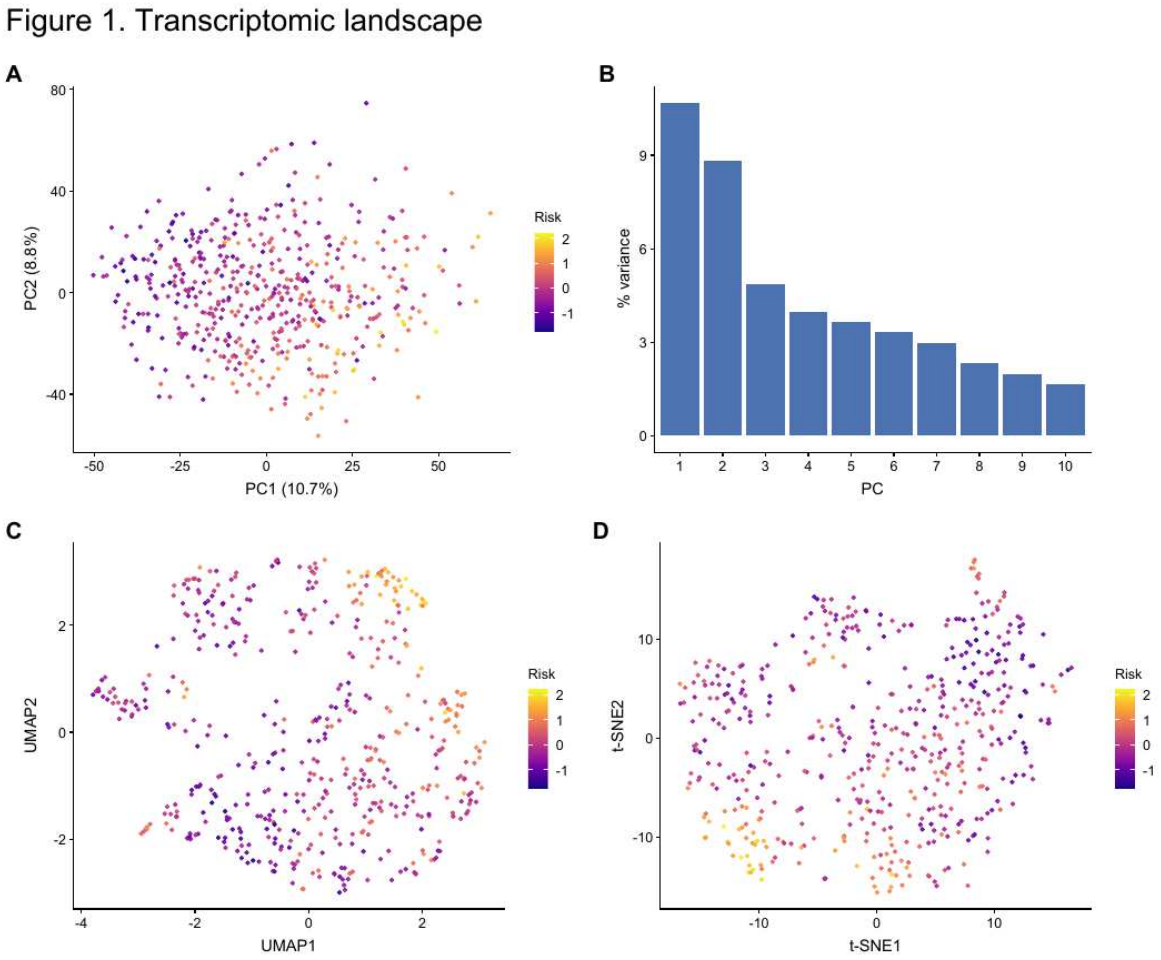


Figure 1. *Exploratory unsupervised landscape of the discovery cohort. (A) Principal-component analysis, coloured by out-of-fold risk score. (B) Scree plot of variance explained per principal component. (C) UMAP projection. (D) t-SNE projection. Provided as exploratory context; no primary claim depends on this figure.*

Internal discrimination: repeated nested cross-validation

Eight survival models were evaluated under repeated (×5) 5-fold nested cross-validation with feature selection strictly inside folds. Five classical/tree learners (Cox, elastic-net Cox, random survival forest [RSF], XGBoost, and a rank-average Ensemble) formed a pre-declared

confirmatory tier eligible for deployment; three neural Cox models (DeepSurv, CoxNNet, MLP) formed a pre-declared exploratory tier included to test whether deep survival models add value at this sample size, and were excluded a priori from deployment and from any primary claim. The Ensemble was pre-specified as the primary-endpoint model before any out-of-fold result was inspected (Table 1).

Across the five confirmatory-tier learners, Harrell's C ranged from 0.607 (XGBoost) to 0.623 (RSF), and Uno's C from 0.574 (elastic-net Cox) to 0.586 (Cox); the pre-specified Ensemble reached Harrell's C 0.618 (0.612–0.625) and Uno's C 0.585 (0.576–0.594) (Table 1, Fig. 3A). Cox was selected as the deployed model because it had the highest Uno's C (0.586) among confirmatory-tier learners; we report this substitution of the deployed model for the pre-specified primary model explicitly rather than silently. Time-dependent discrimination for the deployed Cox model declined with horizon (AUC 0.670 at 1 year, 0.634 at 3 years, 0.581 at 5 years), and the integrated calibration index rose correspondingly (ICI 0.040, 0.086, 0.102 at the same horizons) (Table 2, Table 3, Fig. 3B–D). Harrell's C for Cox was independently cross-checked for direction against Uno's C (raw orientation 0.387 versus reversed 0.613, matching Uno's 0.587) (Table 1B), ruling out a sign-flip error present in earlier drafts of this pipeline. Hyperparameters for RSF (mtry, nodesize) and XGBoost were selected per fold by a small pre-specified grid search on training-fold performance only, never on the outer test fold (Supplementary Table S3). Each base learner's contribution to the pre-specified Ensemble varied by repeat and fold, with a higher median weight for confirmatory-tier learners (Cox, elastic-net, RSF, XGBoost: median 0.17–0.18 per fold) than for exploratory-tier learners (DeepSurv, CoxNNet, MLP: median 0.10–0.13) (Supplementary Table S4), broadly tracking each learner's individual discrimination.

The three exploratory neural Cox models did not exceed, and in two cases had 95% CI upper bounds below, 0.60: MLP Harrell's C 0.577 (0.556–0.600), DeepSurv 0.570 (0.556–0.596), CoxNNet 0.536 (0.514–0.563) (Table 1). We report this as a documented negative result — deep survival models did not outperform classical/tree learners at this sample size (n=504, 182 events) — rather than omitting it; neither DeepSurv nor CoxNNet contributes to any primary claim or to deployment.

Table 1. Internal discrimination, repeated (×5) 5-fold nested cross-validation, out-of-fold.

Model	Tier	Pre-specified primary	Harrell's C (95% CI)	Uno's C (95% CI)
Cox (deployed)	Confirmatory	No	0.613 (0.596–0.627)	0.586 (0.573–0.599)
Ensemble	Confirmatory	Yes	0.618 (0.612–0.625)	0.585 (0.576–0.594)
Random survival forest	Confirmatory	No	0.623 (0.614–0.629)	0.585 (0.576–0.596)
XGBoost	Confirmatory	No	0.607 (0.600–0.616)	0.579 (0.573–0.585)
Elastic-net Cox	Confirmatory	No	0.608 (0.594–0.623)	0.574 (0.565–0.590)
MLP	Exploratory	No	0.577 (0.556–0.600)	0.547 (0.525–0.570)
DeepSurv	Exploratory	No	0.570 (0.556–0.596)*	0.542 (0.528–0.556)
CoxNNet	Exploratory	No	0.536 (0.514–0.563)*	0.533 (0.499–0.564)

** 95% CI upper bound < 0.60.*

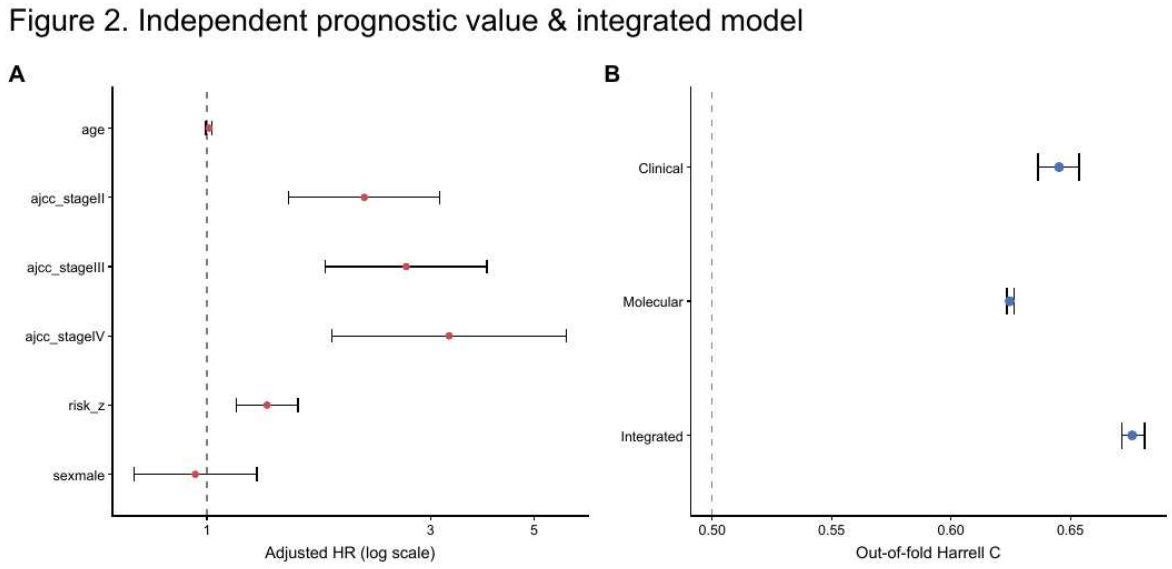


Figure 2. *Independent prognostic value and the integrated model. (A) Multivariable Cox adjusted hazard-ratio forest for the molecular signature (per SD) versus AJCC stage, age, and sex. (B) Out-of-fold Harrell's C for clinical-only, molecular-only, and integrated (clinical + molecular) models.*

As a sensitivity check on the confidence-interval method, a patient-level bootstrap (B=300) gave a wider but overlapping interval for the deployed Cox model — Harrell's C 0.610 (0.567–0.650), Uno's C 0.577 (0.519–0.595) (Supplementary Table S5) — consistent with the repeated-CV estimate. A non-negative-coefficient ridge-Cox stacking meta-learner fit on the base learners' out-of-fold predictions — which, unlike the rank-average Ensemble, has a valid baseline hazard — reached Harrell's C 0.620 (0.612–0.625), Uno's C 0.583 (Supplementary Table S6), closely matching the rank-average Ensemble and confirming the Ensemble result is not an artifact of its zero-parameter combination rule.

Independent prognostic value over standard-of-care staging

In patients with complete clinical covariates (n=480, 176 events), a clinical-only Cox model (age, AJCC stage, sex) had optimism-corrected C-index 0.655; adding the out-of-fold molecular signature score raised this to 0.677 (apparent C-indices 0.667 and 0.688; likelihood-ratio $\chi^2 = 13.69, df = 1, p = 2.15 \times 10^{-4}$) (Table 1C). Both C-indices are bootstrap optimism-corrected on the same leakage-free footing, and we headline the optimism-corrected gain (Δ0.022) rather than

the numerically larger apparent gain (Δ0.021 is, in fact, marginally smaller after correction than before, indicating minimal added optimism from the signature term itself).

In a multivariable Cox model, the signature score (per SD) remained an independent predictor after adjustment for age, stage and sex: hazard ratio 1.343 $(1.153 - 1.565, p = 1.53 \times 10^{-4})$, versus an unadjusted hazard ratio of 1.451 $(1.254 - 1.680, p = 6.05 \times 10^{-7})$ (Table 1D, Fig. 2A). A global Schoenfeld test on this multivariable model indicated the proportional-hazards assumption was not fully satisfied ($p = 9.16 \times 10^{-3}$); this is a distinct test from the risk-group Schoenfeld test reported below, and the two should not be conflated. Age (multivariable HR 1.010, p=0.214) and sex (HR 0.944, p=0.708) were not independently significant in this cohort; AJCC stage remained strongly prognostic at every level (stage II HR 2.167, stage III HR 2.662, stage IV HR 3.289; all $p < 10^{-4}$), as expected.

Integrated clinical + molecular model: the deployable predictor

Clinical-only, molecular-only, and integrated (clinical + molecular) Cox models were compared under repeated patient-level cross-validation on identical folds. Discrimination rose from 0.645 (0.636–0.654) clinical-only, and 0.625 (0.623–0.626) molecular-only, to 0.676 (0.672–0.681) integrated (Uno's C 0.644, 0.637–0.650) (Table 1G, Fig. 2B) — a leakage-free gain of +0.031 Harrell's C over staging alone, since the molecular term is itself out-of-fold and the mixing coefficients are cross-validated on the same folds. We treat this integrated model, not the molecular signature in isolation, as the study's deployable predictor and headline performance figure.

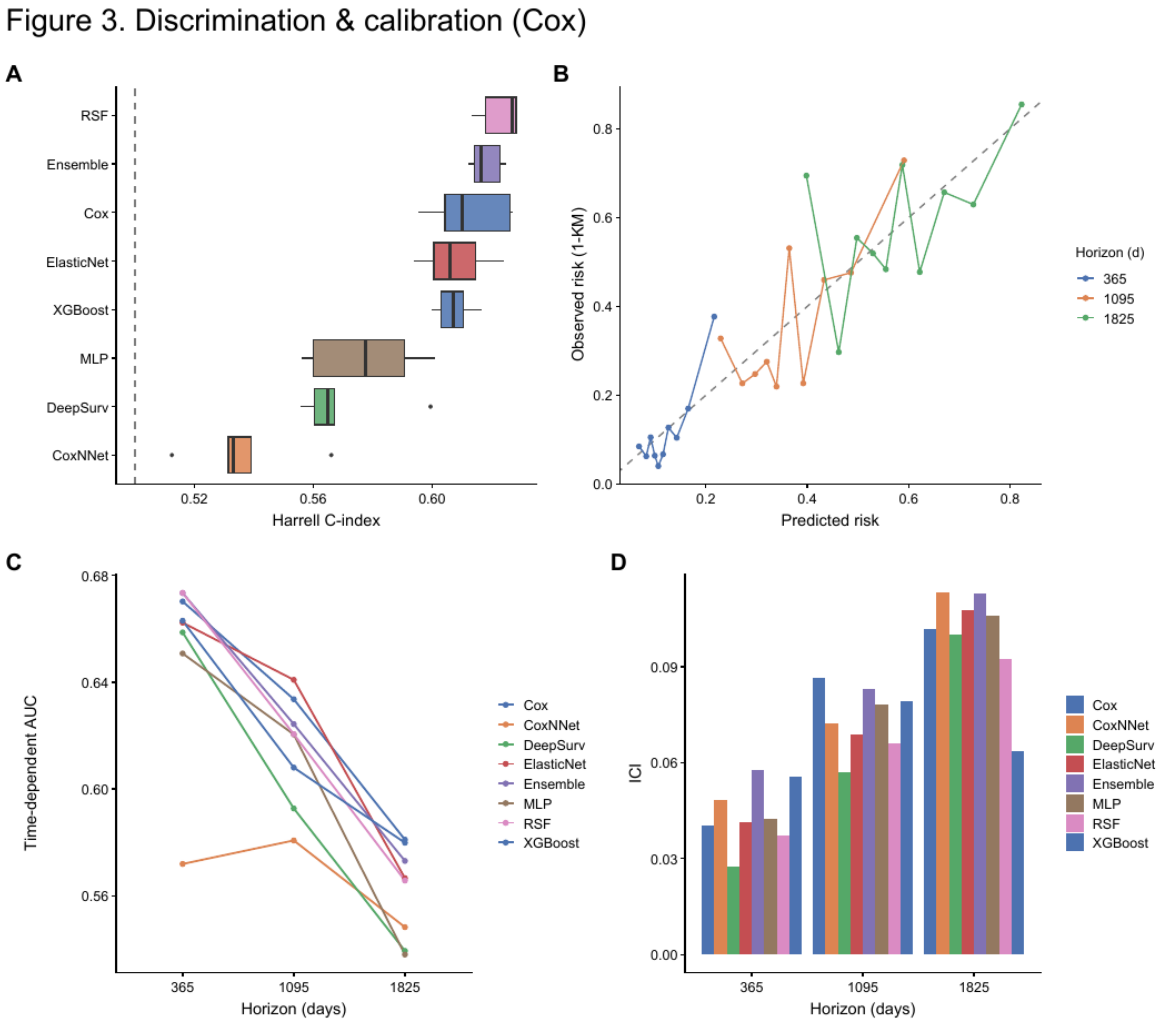


Figure 3. *Internal discrimination and calibration of the deployed model. (A) Harrell's C-index distribution by model across repeated (×5) 5-fold nested cross-validation; dashed line marks the pre-specified minimum-discrimination reference. (B) Predicted-versus-observed calibration for the deployed Cox model, by horizon. (C) Time-dependent AUC by model across horizons. (D) Integrated calibration index (ICI) by model and horizon.*

Clinical utility: decision-curve analysis and reclassification

Across the pooled cohort (n=480, 176 events), decision-curve analysis compared four strategies — treat all, treat none, clinical-only, and clinical + signature — across a threshold-probability range (Table 1E, Fig. 4). Categorical net reclassification improvement of adding the signature to clinical-only at 3 years was 0.278 (event-NRI 0.034, non-event-NRI 0.243; n=480, 176 events)

(Table 1F), attributable predominantly to improved classification of patients who did not experience the event.

Stratifying by stage — the clinically relevant split for adjuvant-therapy decisions, with an NCCN-referenced threshold range of approximately 10–30% — showed the incremental benefit of adding the signature is concentrated in early-stage (I–II) disease (n=376, 117 events): at a 20% threshold, net benefit was 0.165 for clinical + signature versus 0.155 for clinical-only, a modest gain. In late-stage (III–IV) disease (n=104, 59 events), net benefit for clinical + signature was numerically identical to clinical-only at the same thresholds (Table 1E), i.e., no measurable incremental clinical utility was observed in this smaller subgroup. We report this negative subgroup finding rather than omitting it; it is also clinically plausible, since staging is a stronger and less discretionary driver of treatment decisions in late-stage disease, and this subgroup analysis is comparatively underpowered (59 events).

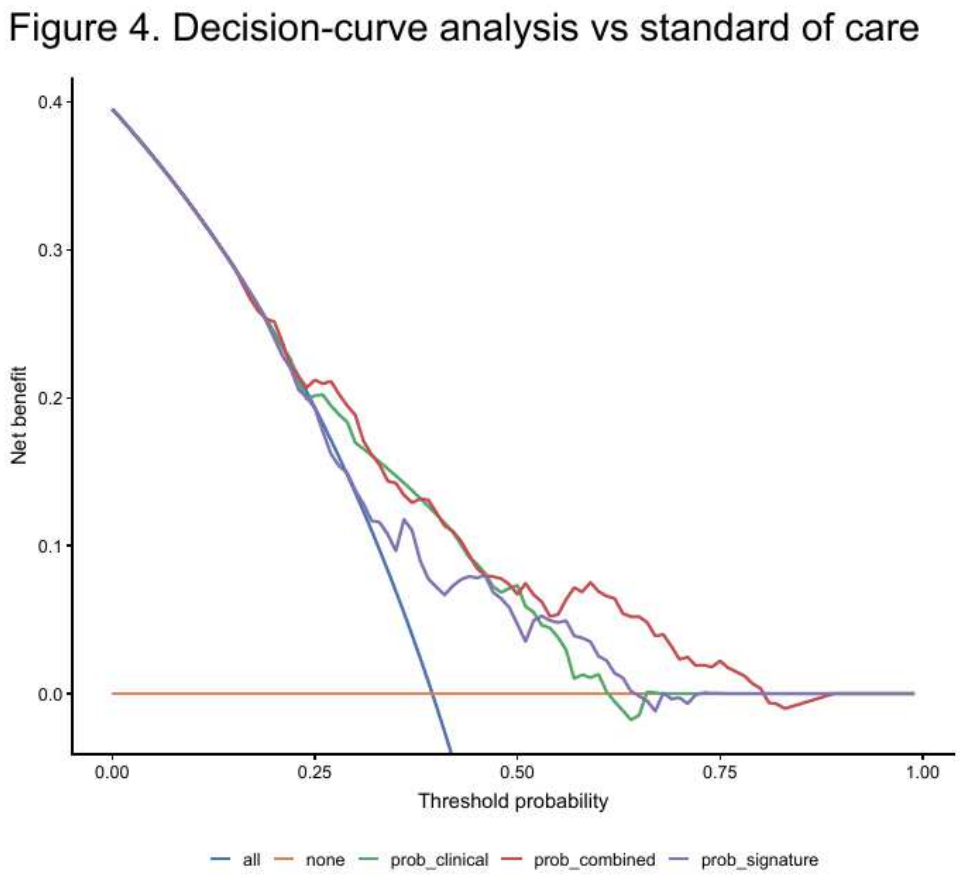


Figure 4. *Decision-curve analysis. Net benefit versus threshold probability for treat-all, treat-none, clinical-only, and clinical + signature strategies (pooled cohort, n=480, 176 events).*

Risk stratification, proportional-hazards violation, and restricted mean survival time

Patients above the out-of-fold median risk score had a hazard ratio of 1.799 (1.337–2.421) for death versus those below it (log-rank $p = 8.44 \times 10^{-5}$; permutation-based log-rank $p = 5.00 \times 10^{-4}$, 2,000 relabelings) (Table 10, Fig. 7A). A Schoenfeld test on this risk-group comparison indicated the proportional-hazards assumption was violated (p=0.0181), so this hazard ratio should be read as a time-averaged effect rather than a constant risk multiplier. Over the pre-specified 5-year (1,825-day) horizon, the assumption-free restricted mean survival time (RMST) was 293 days shorter ($169 - 418\ days, p = 3.86 \times 10^{-6}$) in the high-risk group (Table 10), and we present this RMST difference as the primary effect-size estimate for this comparison.

A clinical nomogram was built on the integrated clinical + molecular model after it cleared a pre-specified minimum discrimination gate (Uno's C ≥ 0.60); the integrated model's Uno's C of 0.644 cleared this threshold (Table 11, Fig. 7B–C). Nomogram variables were molecular risk (per SD), age, AJCC stage, and sex; a nomogram built directly on the molecular signature alone would not have cleared this gate at the pre-specified threshold and was not produced.

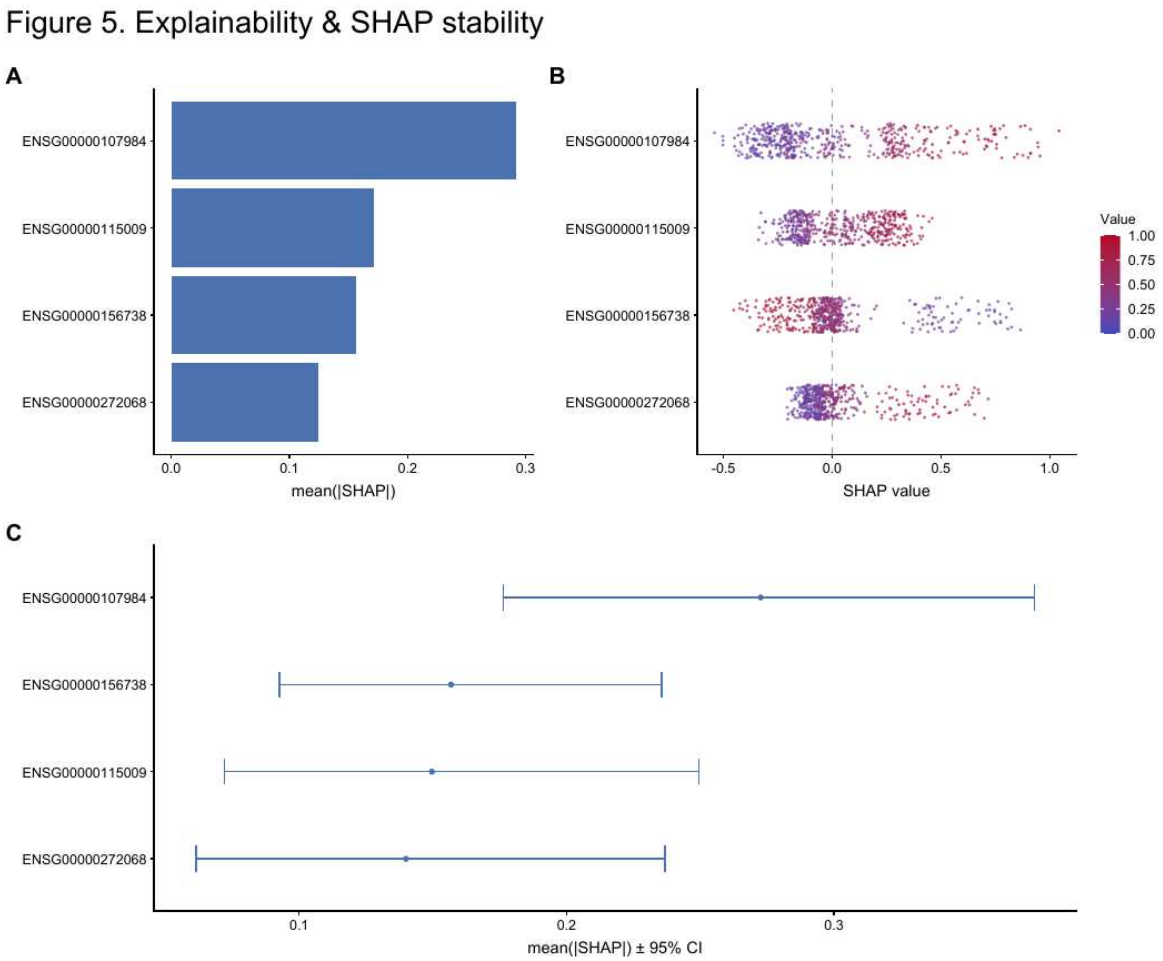


Figure 5. *SHAP-based interpretability of the deployed 4-gene signature. (A) Global feature importance. (B) Beeswarm summary plot. (C) Dependence plots. (D) Individual-prediction waterfall example. (E) Feature-interaction plot. (F) Bootstrap stability of importance ranking (95% CI, rank stability).*

Deployed signature: derivation, parsimony, and interpretability

The deployed signature was locked, prior to final analysis, to a 4-gene hard bootstrap consensus set: DKK1 (ENSG00000107984), MS4A1/CD20 (ENSG00000156738), CCL20 (ENSG00000115009), and BCAN-AS1 (ENSG00000272068, an antisense lncRNA) (Supplementary Table S7). In a pre-specified signature-sensitivity analysis comparing candidate gene sets by identical repeated-CV Cox scoring, this 4-gene set was statistically non-inferior to a larger 20-gene top-frequency set (Δ Harrell's C = −0.006 against a pre-specified 0.02 non-inferiority margin) and was the recommended, smallest non-inferior signature (Supplementary Table S8); absolute discrimination in that sensitivity analysis (Harrell's C 0.696) is optimistic relative to the full nested-CV estimate above and is reported there only for relative comparison across candidate signatures, not as a competing headline figure.

Two independent interpretability procedures agreed on gene ranking. Bootstrap-stability SHAP on the exploratory XGBoost model ranked DKK1 highest (mean |SHAP| 0.292, rank 1 of 4) (Table 4, Table 5, Fig. 5), and model-faithful permutation importance on the deployed portable Cox model — the more appropriate explanation for the deployed model itself — also ranked DKK1 highest (ΔHarrell's C = 0.087, versus 0.046, 0.025 and 0.016 for MS4A1, CCL20 and BCAN-AS1, respectively) (Supplementary Table S9). A co-expression network analysis independently identified DKK1 as the top-priority hub gene (eigenvector centrality 0.590, integrated priority score 0.841) (Supplementary Table S10), concordant with both importance metrics.

Figure 6. Pathway biology

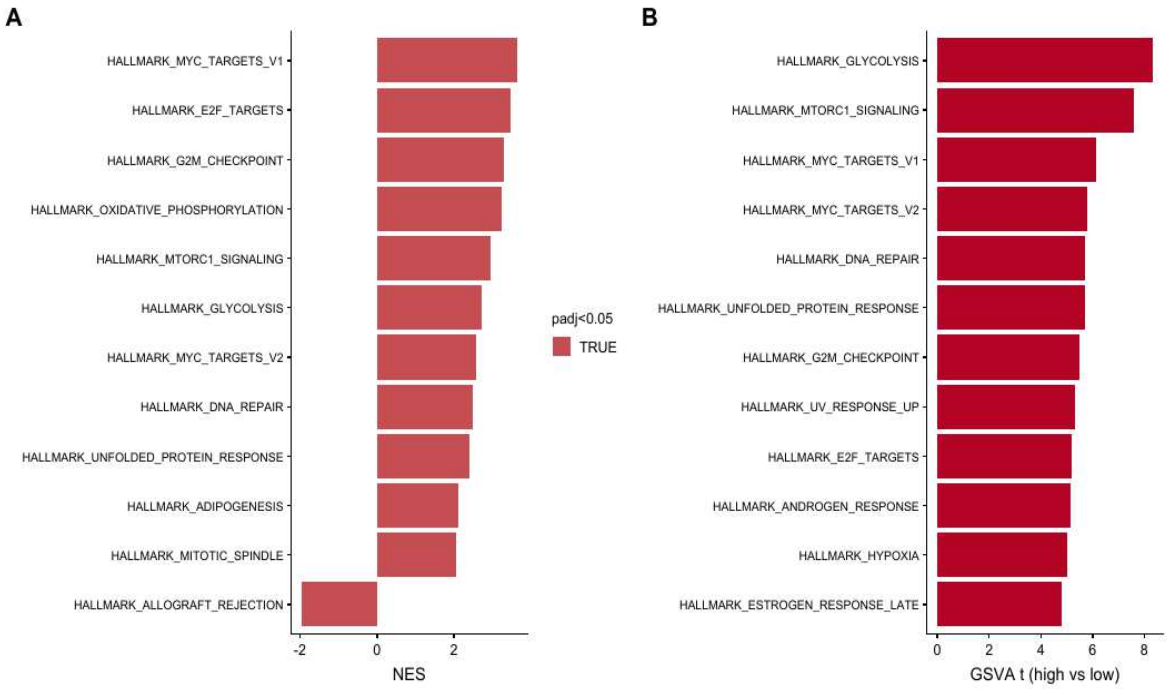


Figure 6. *Systems-biology context. Over-representation dotplots (GO, KEGG, Reactome, Hallmark), gene-set enrichment analysis (GSEA), GSVA pathway scores by out-of-fold risk group, and immune cell-type deconvolution. Exploratory context; not used to support any primary claim.*

Sensitivity analyses

Squamous confounding. Because histological misclassification is a known confounder for LUAD signatures, discrimination was re-evaluated after excluding the top decile of an inferred squamous score. Discrimination was materially unchanged: 0.696 (0.694–0.698) in all samples (n=504) versus 0.699 (0.697–0.703) in the LUAD-pure subset (n=453) (Supplementary Table S11), and only 1 of the 4 deployed genes (DKK1, $\rho = 0.308$) correlated with the squamous score at $|\rho| \geq 0.3$. The deployed signature does not appear to be primarily a histology detector within this LUAD cohort.

Missing clinical data. The complete-case multivariable analysis above used 480 of 504 patients (95%). Multiple imputation by chained equations (m=10, survival-aware, 24 imputed patients, full n=504) gave a signature hazard ratio of 1.312 ($1.130 - 1.525, p = 4.87 \times 10^{-4}$) (Supplementary

Table S12), closely matching the complete-case estimate (1.343, 1.153–1.565); the independent-prognostic-value claim is not an artifact of case exclusion.

Cross-platform gene coverage. DKK1, MS4A1 and CCL20 were measurable on all five external platforms; BCAN-AS1 was measurable only on Cohort 4 (RNA-seq), consistent with sparser long non-coding RNA annotation on the older microarray platforms used by Cohorts 1, 2, 3 and 5 (Table 2). A reduced 3-of-4-gene model was therefore used for four of the five external cohorts; this is a reportable limitation of cross-platform measurability rather than an imputation choice.

Biological context

Over-representation analysis of the event-associated differential expression signature against the Hallmark gene sets identified E2F targets as the top enriched term (104 of 199 pathway genes represented, fold enrichment 2.67, $p.adjust = 1.63 \times 10^{-24}$), consistent with a proliferation-associated transcriptional programme distinguishing patients who experienced the event (Table 6). Separately, differential expression between patients who experienced the event and those censored was enriched for proliferation and metabolic Hallmark pathways by gene-set enrichment analysis (MYC targets v1, NES=3.63, $padj = 5 \times 10^{-49}$; E2F targets, NES=3.47, $padj = 5.4 \times 10^{-42}$; G2M checkpoint, NES=3.29, $padj = 1.2 \times 10^{-34}$; oxidative phosphorylation, NES=3.25, $padj = 2.4 \times 10^{-33}$) (Table 7, Fig. 6). Within the cohort, GSVA pathway scores were higher in the out-of-fold high-risk group for glycolysis (t=8.31, $padj = 4.2 \times 10^{-14}$), mTORC1 signaling (t=7.61, $padj = 3.3 \times 10^{-12}$) and MYC targets (t=6.15, $padj = 2.6 \times 10^{-8}$) (Table 8), an association consistent with a more proliferative, metabolically active tumour state in high-risk patients; this is a transcriptomic association, not evidence of a causal mechanism. Immune cell-type deconvolution and a Hallmark pathway crosstalk network (87 pathway-pathway edges above the 80th percentile of absolute correlation) are reported as exploratory systems-biology context (Table 9, Supplementary Table S13, Fig. 6); a weighted gene co-expression module correlated only weakly with vital status (best module $\rho = -0.169$) and is not used to support any primary claim.

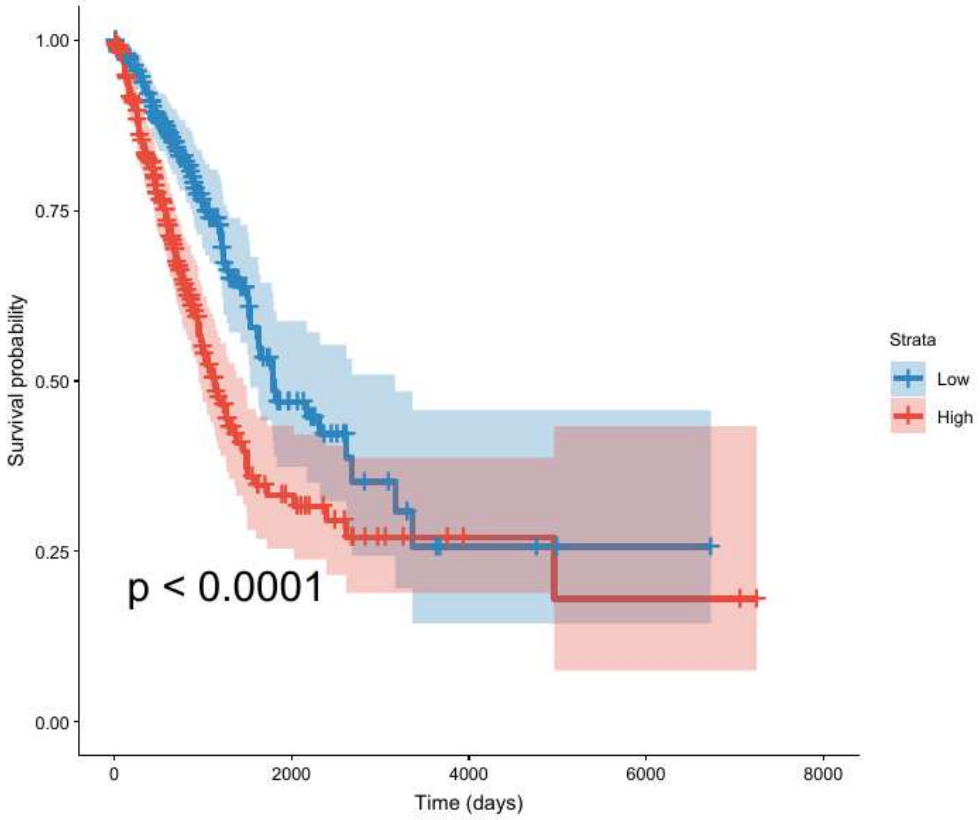


Figure 7. *Risk stratification and nomogram. (A) Kaplan–Meier survival by out-of-fold risk group (high vs. low), with log-rank p-value. (B) Nomogram built on the integrated clinical + molecular model (molecular risk, age, AJCC stage, sex). (C) Nomogram calibration.*

External validation across five independent cohorts

The frozen 4-gene Cox model — ridge-penalized (glmnet, λ.1se) rather than an unpenalized maximum-likelihood fit, to favor out-of-cohort generalization — was applied unchanged, with training-derived coefficients and risk cutoff, to five independent cohorts spanning two platforms and 1,254 patients / 475 events (Table 2, Fig. 8) (Supplementary Table S14). Per-cohort Harrell's C ranged from 0.591 (Cohort 1, GSE68465) to 0.642 (Cohort 3, GSE30219). Cohort 1 remains the weakest-performing cohort and, despite the addition of Cohort 5, still contributes the largest single share of pooled events (49.7% of 475); leave-one-cohort-out analysis confirms it is the most influential cohort, though its influence is now modest (removing Cohort 1 shifts pooled Harrell's C by +0.013; removing any other cohort shifts it by ≤ 0.004) (Supplementary Table S15).

Table 2. External validation summary, five independent cohorts.

Cohort	Platform	n	Events	Sig. genes	Harrell's C (95% CI)	Calib. slope
Cohort 1 (GSE68465)	Microarray	442	236	3/4	0.591 (0.550–0.631)	0.346
Cohort 2 (GSE31210)	Microarray	226	35	3/4	0.633 (0.539–0.726)	0.696
Cohort 3 (GSE30219)	Microarray	83	43	3/4	0.642 (0.550–0.734)	0.813
Cohort 4 (GSE81089)	RNA-seq	105	48	4/4	0.595 (0.505–0.685)	0.473
Cohort 5 (GSE72094)	Microarray	398	113	3/4	0.615 (0.560–0.670)	0.574
Pooled (fixed-effect)	2 platforms	1254	475	—	0.606 (0.578–0.634)	—

Pooled estimate by fixed-effect inverse-variance meta-analysis; the random-effects estimate is numerically identical $I^2 = 0\%, \tau^2 = 0$*).*

Figure 8. External validation (transportability)

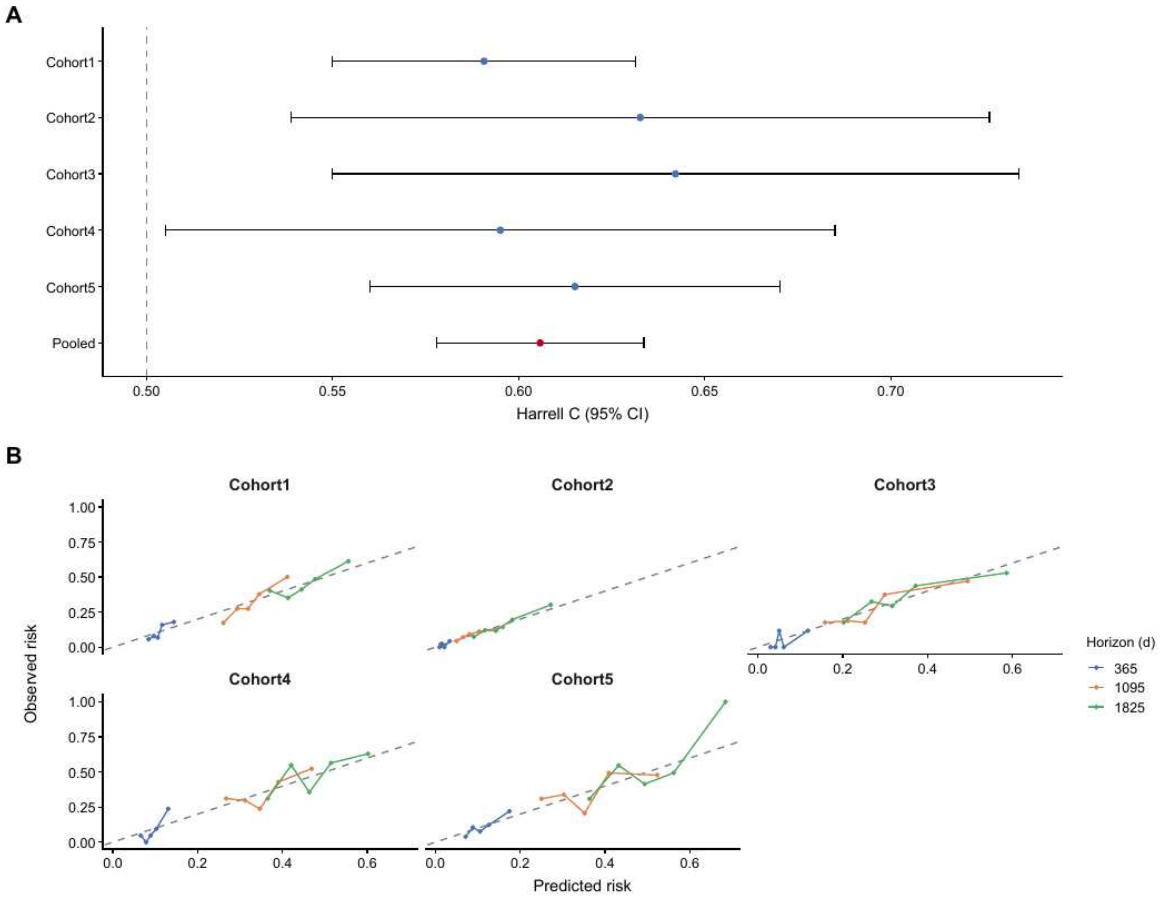


Figure 8. *External validation across five independent cohorts. (A) Per-cohort and pooled Harrell's C forest plot (fixed-effect meta-analysis; dashed line = 0.5). (B) Per-cohort calibration: predicted versus observed risk, by horizon.*

Pooled discrimination was 0.606 (0.578–0.634) by fixed-effect meta-analysis; the random-effects estimate is identical, with no detected between-cohort heterogeneity ($I^2 = 0\%, \tau^2 = 0$, Cochran's Q=1.60, k=5) (Supplementary Table S16). With k=5 cohorts the 95% prediction interval (0.561–0.651) remains imprecisely estimated and we headline the random-effects confidence interval instead. This pooled external estimate is close to the internal deployed-model estimate (Cox Harrell's C 0.613), an internal-to-external gap of 0.007. Calibration slopes were below 1 in every cohort (range 0.346–0.813) (Table 2), indicating cross-platform over-dispersion of the frozen linear predictor rather than a preprocessing artifact, since both training and each external cohort are independently platform-standardized; discrimination, being rank-based, is not affected by this.

The single RNA-seq cohort (Cohort 4, Harrell's C 0.595) did not perform systematically better or worse than the mean of the four microarray cohorts (0.620), so pooled transportability does not appear to depend on a single platform.

Time-dependent AUC at 1/3/5 years was 0.637/0.632/0.565 (Cohort 1), 0.698/0.617/0.664 (Cohort 2), 0.750/0.667/0.643 (Cohort 3), 0.736/0.583/0.477 (Cohort 4), and 0.651/0.622/0.648 (Cohort 5) (Supplementary Table S17). Discrimination was consistently highest at 1 year and generally declined at longer horizons; the decline was most pronounced in Cohort 4, where the 5-year point estimate (0.477) fell below chance. We report this plainly: Cohort 4 has the smallest number of patients still at risk by 5 years (n=105, 48 events overall), so this estimate carries wide uncertainty, but it is a negative result at that specific horizon in that specific cohort and is not smoothed over by the pooled 3-year summary above.

Benchmark against a published prognostic signature

The deployed 4-gene signature was compared against the Kratz et al. 14-gene resected-NSCLC prognostic gene list (12 of 14 original genes mappable), with coefficients for both signatures re-fit on the discovery cohort under the identical pipeline and scored on the current five-cohort external set through the identical validation procedure — holding methodology constant rather than replicating Kratz et al.'s original qPCR/housekeeping-normalized assay (Supplementary Table S18). Pooled external discrimination was 0.616 ($I^2 = 0\%$) for the 12-gene Kratz set versus 0.605 ($I^2 = 0\%$) for the 4-gene deployed signature — a small ($\Delta \approx 0.011$) advantage for the larger published signature. We report this as a competitive, not superior, result for a signature roughly one-third the size.

Reporting-guideline compliance

All 25 pre-specified reporting items — 14 TRIPOD+AI, 6 PROBAST+AI, and 5 REMARK items — were met, machine-audited against the artifacts actually produced by this run rather than asserted independently of them. This tracked checklist does not include a separate item for stratified subgroup performance (e.g., by sex or age); such subgroup discrimination was not separately assessed and is reported as an explicit gap, not folded into the 25/25 figure (Supplementary Table S19). PROBAST+AI domains covered participant selection, predictor

definition, outcome definition, and analysis (feature selection inside folds, appropriate handling of censoring, and reporting of calibration alongside discrimination); each domain-level judgment is reported individually in the compliance table rather than as a single aggregate score.

Summary of principal results

Taken together, these results support three claims at different levels of confidence. We have high confidence that the deployed molecular signature carries statistically independent prognostic information beyond stage, age and sex, and that this holds under multiple imputation and after excluding squamous-confounded samples. We have moderate confidence that the resulting integrated clinical + molecular model (Harrell's C 0.676 internally) is the more clinically relevant predictor than the molecular signature alone, and that it transports to five independent cohorts with homogeneous, modest discrimination (pooled C 0.606, $I^2 = 0\%$). We have low confidence — and state this explicitly — that this signature alone, at C≈0.60–0.62, is ready for any clinical use without prospective validation; the Discussion addresses this distinction in full.

Discussion

Principal findings, in order of evidentiary weight

This study's contribution is best read as three claims of decreasing certainty, and we present them in that order rather than leading with the most favourable-sounding number. First, a pre-specified, leakage-free 4-gene transcriptomic signature carries statistically independent prognostic information beyond AJCC stage, age and sex in this cohort (multivariable hazard ratio 1.343 per SD, $p = 1.53 \times 10^{-4}$), a result that survives multiple imputation of missing covariates and exclusion of squamous-confounded samples. Second, an integrated clinical + molecular model — itself cross-validated so that its reported discrimination carries no additional optimism beyond what is already present in the out-of-fold molecular score — reaches Harrell's C 0.676, a leakage-free +0.031 gain over staging alone; this integrated model, not the transcriptomic signature in isolation, is the predictor we advance as this study's central result. Third, the signature's discrimination transports to five independent cohorts spanning two platforms with homogeneous behaviour ($I^2 = 0\%, \tau^2 = 0$), a materially stronger claim than internal cross-validation alone,

though the absolute transported discrimination — pooled Harrell's C 0.606 — remains modest by any clinical standard.

We deliberately do not lead with molecular-signature discrimination on its own. Internally it sits at $C \approx 0.61 - 0.62$ across the confirmatory-tier models; externally, pooled C is 0.606. Both estimates clear chance — their 95% confidence intervals exclude 0.5 — but neither reaches a magnitude we would independently call clinically actionable, and we say so in the Results section itself rather than reserving that admission for the Discussion.

By "leakage-free" throughout this Discussion we mean specifically: feature selection performed inside, not before, cross-validation folds; every optimism-corrected comparison corrected symmetrically on both sides; and out-of-fold predictions used for every reported discrimination, calibration, and clinical-utility statistic rather than in-sample fitted values. We note this explicitly because these three properties are each individually easy to violate without an obviously wrong-looking result, and violating any one of them tends to inflate discrimination in the direction that makes a signature look more clinically promising than it is — which is precisely the direction a hopeful analyst is least likely to interrogate. One exception to strict fold-internal fitting is explicit and, we think, defensible: the DESeq2 variance-stabilizing transformation and the low-count gene filter are each fit once on the full discovery cohort rather than per fold, because both are outcome-blind by construction (no survival label or group variable enters either step). This is standard practice for RNA-seq normalization and is a materially different, much smaller-risk category of leakage than supervised feature selection using outcome labels — but it is not literally fold-internal, and we would rather name that precisely than let "leakage-free" imply a stronger guarantee than the pipeline actually provides.

A methodology contribution, not a gene-discovery claim

All three genes in the deployed signature with characterized biology — DKK1, CCL20, and MS4A1/CD20 — are independently described as LUAD/NSCLC-relevant in prior literature; we are not reporting novel biology, and we do not claim to be. What we believe is less common in this specific literature is the methodology behind how the signature was derived and validated: feature selection performed strictly inside cross-validation folds rather than before splitting; a pre-specified and tiered model hierarchy, with confirmatory learners eligible for deployment decided

before any out-of-fold result was inspected and exploratory neural models excluded from deployment a priori; dual Harrell's-and-Uno's-C reporting with an explicit sign-orientation check; leakage-free optimism correction applied symmetrically to both sides of every baseline-versus-signature comparison; and external validation across five independent cohorts with formal random-effects heterogeneity accounting, rather than a single convenience held-out set.

Transcriptomic prognostic-signature papers in oncology have a well-documented history of feature-selection leakage, undisclosed model selection after seeing test-set results, and orientation ambiguity in concordance statistics — problems that are each individually easy to introduce and, once introduced, easy to miss on read-through. Given that history, we regard reproducible, leakage-free methodology — not discrimination magnitude — as this study's primary contribution. We have tried to frame the Results section accordingly: confidence intervals are stated before point estimates, model substitutions from the pre-specified plan are disclosed rather than silently applied, and the exploratory neural models' negative result (two of three failing to clear a 0.60 discrimination lower bound) is reported rather than omitted. We do not read this negative result as evidence that deep survival modelling is unsuited to this problem in general: n=504 with 182 events is a small dataset by the standards deep architectures are usually trained at, and a fair test of whether deep learning adds value here would more plausibly require transfer learning or pretraining on a larger pan-cancer cohort rather than training from scratch on a single, modestly sized dataset, which is what we did. We report the comparison as it was pre-specified rather than as the strongest possible test of deep learning specifically. A reader who takes away only the C-index numbers from this paper will, we think, have missed its main point.

Position within a crowded literature

Multi-gene prognostic signatures for resected LUAD are one of the more heavily populated corners of oncology bioinformatics; dozens of candidate signatures, mostly derived from TCGA or overlapping public cohorts, have been proposed over the past decade, and most are validated in at most one or two external cohorts, if any. A new 4-gene signature reporting $C \approx 0.61$ internally is not, by itself, a distinguishing result — comparable or higher point estimates appear elsewhere in this literature, often without the leakage controls described above. The more useful question this study can answer is not "is this signature better than the others" but "does a fully leakage-free, pre-specified, multi-platform-validated pipeline still find a modest, homogeneous, competitive signal"

— and the honestly reported answer is yes to homogeneous and competitive, but also yes to modest.

Comparison with a published signature

Run through an identical five-cohort external-validation procedure, the deployed 4-gene signature was numerically outperformed by the Kratz et al. 14-gene resected-NSCLC gene list (pooled Harrell's C $0.605\ vs.\ 0.616, \Delta \approx 0.011$), with both signatures homogeneous across cohorts ($I^2 = 0\%$ for each). We are explicit that this benchmark re-fits Kratz's published gene list under our own pipeline (VST-normalized expression, cohort z-scoring, ridge-Cox), not Kratz et al.'s original qPCR ΔCt assay normalized against three housekeeping genes; it isolates the contribution of gene selection while holding modelling methodology constant, and should not be read as a reproduction of the original clinical assay's real-world performance. Within that scope, we report the result as a competitive, not superior, comparison: a signature roughly one-third the size reaches within about one percentage point of a substantially larger, well-cited gene list.

This comparison also understates how far ahead the Kratz assay is on a dimension we cannot benchmark computationally: clinical validation. The original 14-gene assay, in its native qPCR form, has since been tested in AIM-HIGH, an international randomized phase 3 trial (45 centres, France/Germany/USA) that used the assay to select stage IA–IIA non-squamous NSCLC patients for adjuvant chemotherapy and reported improved disease-free survival in the molecular-high-risk arm [60]. No signature discussed in this manuscript — the Kratz gene list as we re-scored it, or our own deployed signature — has undergone anything resembling this. We see AIM-HIGH not as a competing result to argue against, but as a concrete illustration of the evidentiary distance between a discrimination statistic in a retrospective cohort and a validated clinical decision tool; it is the kind of trial our own signature would need before any claim of clinical utility, and we reference it here to make that distance explicit rather than abstract.

We see two defensible readings of this comparison and do not privilege either. A generous reading is that parsimony carries practical value independent of raw discrimination: fewer genes mean fewer platform-dependent measurement failures, illustrated directly within this study by its own long non-coding RNA, which was measurable on only one of five external platforms and forced a reduced 3-of-4-gene model on the other four. A more skeptical reading is that both signatures —

one derived in 2012, one derived here with considerably more methodological machinery — are converging on a similar, modest ceiling for transcriptome-only prognostic discrimination in resected LUAD, and that neither represents a qualitative advance over the other. If that skeptical reading is closer to correct, it is itself a useful, if unglamorous, finding for the field: it suggests the ceiling may lie in the information content of bulk transcriptomic profiling for this outcome, rather than in any particular gene-selection or modelling strategy.

We also note what this comparison does not show: that our signature is preferable to Kratz's on any dimension other than gene count. A reader whose priority is maximum discrimination would reasonably prefer the published alternative on the evidence presented here; we report the comparison because an absent benchmark against existing signatures is itself a common, avoidable gap in this literature, not because the result favours us.

Biological plausibility, held at arm's length from the data

DKK1 — the top-ranked gene by three independent methods within this study (SHAP importance, permutation importance on the deployed model, and co-expression network centrality) — is reported elsewhere in NSCLC as pro-tumorigenic through non-canonical, β-catenin-independent signalling, despite its canonical textbook role as a Wnt pathway antagonist [61], [62], [63]. CCL20 is reported as associated with shorter survival in NSCLC, consistent with a myeloid/Th17 chemokine-recruitment axis [64], [65]. MS4A1/CD20 marks tumour-infiltrating B cells, a recognized favourable-immunity axis in several solid tumours including NSCLC [66]. These associations are drawn from literature external to both our discovery cohort and our five validation cohorts, and we read them as offering independent corroboration that the signature is tracking established biology rather than discovery-cohort-specific noise.

We stop short of any causal interpretation. Our own data are purely observational and transcriptomic; GSVA-level pathway associations with out-of-fold risk group — elevated glycolysis, mTORC1 signalling, and MYC-target expression in the high-risk group — are consistent with, but do not establish, a more proliferative high-risk phenotype, since gene expression differences between risk groups could equally reflect downstream consequences of tumour burden rather than upstream drivers of it.

BCAN-AS1 (HGNC:55233; also annotated as RP11-284F21.9), the antisense long non-coding RNA in the deployed signature (ENSG00000272068), has been reported to stabilize c-Myc in pancreatic ductal adenocarcinoma: in its m6A-modified form it acts as a scaffold for a SNIP1–c-Myc complex that blocks SKP2-mediated c-Myc ubiquitination and degradation [67]. Under its earlier alias RP11-284F21.9, it has also been reported to promote lung carcinoma cell proliferation and invasion via a miR-627-3p/CCAR1 axis [68]. Neither study establishes the c-Myc-stabilizing mechanism in lung adenocarcinoma specifically, and we do not claim it is demonstrated in our cohort; we present it as biologically plausible candidate context rather than a proven mechanism here. It is nonetheless a notable convergence with our own data: gene-set enrichment analysis independently identified MYC-target enrichment as the top pathway associated with the event group, and BCAN-AS1 entered the deployed signature through a purely statistical, outcome-blind selection procedure with no biological annotation used in gene choice — the two findings were derived independently of one another and connect only post hoc, not by design.

A Hallmark pathway crosstalk network built from GSVA scores (87 pathway-pathway edges above the 80th percentile of absolute correlation) places proliferation and metabolic programmes in close co-variation, which is broadly consistent with the enrichment results above but should be read as descriptive systems-biology context rather than as an independently informative finding: the same underlying risk-group split drives both the crosstalk network and the enrichment analysis, so the two are not independent lines of evidence for a shared mechanism, merely two views of the same association.

Transportability, and where it remains fragile

Homogeneous transport across five cohorts and two platforms ($I^2 = 0\%, \tau^2 = 0$) is, in our assessment, the strongest single piece of evidence in this study. We added a fifth cohort specifically to test whether the original four-cohort homogeneity might be an artifact of a small number of cohorts, and the pooled point estimate barely moved (0.602 with four cohorts, 0.606 with five) while precision improved modestly. Leave-one-cohort-out analysis on the current five-cohort set shows no single cohort dominates the pooled estimate: removing the most influential cohort — Cohort 1, still the largest single contributor of pooled external events despite the addition of Cohort 5 — shifts pooled Harrell's C by only 0.013, and removing any of the other four cohorts shifts it by 0.004 or less.

That said, transportability here means "modest discrimination replicates," not "discrimination is high," and calibration slopes below 1 in every external cohort (range 0.346–0.813) indicate the frozen linear predictor is systematically over-dispersed when transported across platforms, even though discrimination itself is preserved. We would also want to see this signature tested in a sixth, larger, and ideally more demographically diverse cohort before treating five-cohort homogeneity as a fully settled question: all five validation cohorts to date, like the TCGA discovery cohort itself, are drawn from broadly similar (predominantly US/European) sampling frames, and transportability outside that frame is untested. The single RNA-seq cohort performed within the range of the four microarray cohorts, which is reassuring for platform-generalizability but is a sample of one and should not be over-interpreted as proof that platform no longer matters.

Clinical-utility interpretation

The stage-stratified decision-curve finding — a modest incremental net benefit from adding the signature in early-stage (I–II) disease, and no measurable incremental benefit in late-stage (III–IV) disease — strikes us as clinically coherent rather than a simple power artifact. Late-stage treatment decisions in LUAD are, on the whole, less threshold-sensitive and less discretionary, whereas the stage I/II adjuvant-therapy decision is precisely the kind of threshold-dependent, discretionary decision that decision-curve analysis is designed to inform. We nonetheless flag that the late-stage subgroup (n=104, 59 events) is underpowered to detect a moderate difference, so an absence of measured incremental benefit there is not strong evidence of a true absence.

Neither decision-curve result should be read as evidence that this signature is ready to inform an actual adjuvant-therapy decision today. Decision-curve analysis quantifies utility conditional on a model's discrimination and calibration being trustworthy in the population where it will be used; establishing that trustworthiness in a genuinely new, prospectively enrolled population is exactly what retrospective decision-curve analysis cannot do on its own, however favourable the curve looks in cross-validation or external validation.

The same caution applies to the nomogram built on the integrated model. Its underlying molecular score was externally validated, but the nomogram itself — the mapping from inputs to an absolute predicted risk — was calibrated and evaluated only internally, never against outcomes in an external cohort. We present it as a summary of the integrated model's internal behaviour, not a

validated absolute-risk tool; external calibration of the nomogram specifically is a prerequisite before it informs any individual patient's risk estimate.

Limitations, ranked by how much weight a critical reader should give them

We rank these in the order we expect a skeptical reviewer would raise them, not the order most flattering to the study.

1. Absolute discrimination is modest, and this is the central limitation, not a footnote. Internal Harrell's C is approximately 0.61–0.62 across confirmatory-tier models; pooled external C is 0.606. Two of eight internal models — the exploratory-tier DeepSurv and CoxNNet — failed to clear a 95% confidence-interval lower bound of 0.60 at all. This ceiling, not any single design choice made along the way, is what keeps us from framing this work as describing a clinical-grade biomarker.

2. Single discovery cohort. All feature selection and internal cross-validation derive from one source (TCGA-LUAD, n=504, 182 events). Five-cohort external validation mitigates, but cannot fully substitute for, an independent second discovery effort with its own feature-selection process.

3. A 4-gene signature, one member of which (BCAN-AS1, an antisense lncRNA) is largely unmeasurable outside RNA-seq. Four of five external cohorts — 80% of the external validation set by cohort count, and 427 of 475 pooled external events (90%) — were necessarily scored on a reduced 3-of-4-gene model rather than the full deployed signature. This is a genuine cross-platform portability limitation of the specific genes selected here, not something resolved by better imputation of the missing measurement, and it means the pooled external estimate we headline is, strictly, an estimate for the 3-gene reduced model with a single 4-gene cohort's contribution, not a five-cohort validation of the 4-gene signature as deployed.

4. No treatment or adjuvant-therapy covariate is available in any cohort we analysed. We cannot rule out that treatment allocation, rather than tumour biology alone, partly explains the observed risk-group survival difference. In our own assessment this is the most consequential unaddressed confound in the study, more so than any purely statistical limitation above.

5. No protein-level or orthogonal-platform validation. All evidence presented here, including the external validation, is transcriptomic. Whether DKK1, CCL20, or MS4A1 protein levels track the RNA signal in this specific cohort is untested in this study.

6. Prognostic, not predictive. No treatment-randomized or treatment-interaction data exist in any cohort analysed here. We make no claim, implicit or explicit, that this signature can inform treatment selection between two or more options; it estimates prognosis under whatever treatment patients in these cohorts actually received.

7. The proportional-hazards assumption is violated, in two separate analyses. This holds both for the multivariable clinical + signature Cox model and, separately, for the isolated risk-group Kaplan–Meier comparison. We report hazard ratios alongside their proportional-hazards test results throughout, and for the risk-group comparison specifically we report an assumption-free restricted-mean-survival-time difference; a reader extracting a single hazard ratio from this work without that surrounding context would be treating a time-averaged effect as if it were a constant one.

8. All cohorts are demographically similar. The discovery cohort and all five validation cohorts are drawn from predominantly US/European sampling frames; transportability to other populations and healthcare settings is untested.

9. Retrospective throughout. Nothing in this study constitutes, or approaches, evidence of clinical readiness. This is a statement about the scope of what retrospective analysis can establish, not a hedge or an expression of false modesty.

10. Global, rather than fold-internal, fitting of the VST normalization. The variance-stabilizing transformation and the low-count gene filter were each fit once on the full discovery cohort rather than per cross-validation fold. Both are outcome-blind by construction, which distinguishes this from supervised feature-selection leakage and is standard practice for RNA-seq normalization, but it is not literally fold-internal, and a fully strict pipeline would refit both per fold. We regard the resulting leakage risk as small — no survival label enters either step — but we would rather name this precisely than let our leakage-free framing imply a stronger guarantee than the pipeline provides.

What would change our conclusions, and what we would want to see next

A sixth external cohort with better-balanced events than Cohort 1 — still the largest single contributor of pooled events — would be the most informative next experiment for transportability. A treatment-annotated cohort, even retrospective, would let us test whether the risk-group survival difference survives adjustment for adjuvant therapy, the most important open confound. Protein-level validation of DKK1, CCL20 and MS4A1 would upgrade the biological-plausibility argument from literature-corroborated to directly tested.

None of these additions would, on their own, move this signature from "prognostic, homogeneously transportable, and modest" to "clinically actionable." That transition requires prospective, ideally trial-embedded, validation in a population where a treatment decision genuinely turns on the signature's output — a study design outside the scope of any retrospective analysis, however rigorously conducted, including this one.

What reporting-guideline compliance does, and does not, buy

All 25 pre-specified TRIPOD+AI [69](which updates the original TRIPOD statement[70], PROBAST+AI [71](which extends the original PROPBAST tool[72]), and REMARK [73] items on our tracked checklist were met in this study, machine-audited against the artifacts this run actually produced rather than asserted by the authors independently of them. We are explicit about what this does and does not establish. It establishes that the analysis was conducted and reported in a way that lets a reader check our work — every number in the Results traces to a named table or figure, every discrimination estimate is reported with an appropriate interval, and calibration is reported alongside discrimination rather than in its place. It does not establish that the underlying signal is strong, and reporting-guideline compliance should never be read as a proxy for clinical validity. A perfectly reported study of a weak biomarker is still a study of a weak biomarker; we simply think readers are better served knowing that with confidence than not knowing it at all.

We also flag a specific, named gap in this compliance claim rather than let the headline 25/25 figure stand unqualified: the 2024 TRIPOD+AI statement added a fairness domain that modifies the model-performance item to recommend reporting discrimination separately across key subgroups (for example, by sex or age). Our tracked checklist, built before we had reconciled it against every subitem of the current statement, does not include this as a separate tracked item,

and we did not compute stratified subgroup discrimination in this analysis. We would rather name this omission explicitly than have a reader discover it independently and reasonably wonder what else the compliance claim might be silently missing.

Reproducibility as part of the contribution

Every stochastic step in this pipeline is controlled by a single global seed plus deterministic per-repeat, per-fold seeds; out-of-fold predictions are produced internally rather than reconstructed from externally saved folds; and a cache-cleared re-run from raw counts reproduces every number reported here exactly. We regard this as part of the contribution, not incidental infrastructure: a signature that cannot be exactly regenerated by an independent user is not, in a meaningful sense, fully reported. We recommend code and a versioned container be deposited with a citable identifier (e.g., Zenodo) alongside publication, consistent with Genome Medicine's data- and code-availability requirements.

Conclusion

A pre-specified, leakage-free 4-gene signature adds independent, modest prognostic information beyond stage, age and sex in this cohort, and its combination with staging — not the signature in isolation — is the predictor we advance, at Harrell's C 0.676 internally with homogeneous, modest transport across five independent cohorts (pooled C 0.606, $I^2 = 0\%$). Absolute discrimination does not, on its own, support any claim of clinical readiness. We think the strongest and most defensible contribution of this work is methodological: a fully leakage-free, pre-specified, multi-cohort externally validated, and TRIPOD+AI/PROBAST+AI/REMARK-compliant template for transcriptomic prognostic modelling, offered in a subfield where this level of methodological rigor remains the exception rather than the rule. We would rather this study be judged on whether its numbers can be trusted and reproduced than on whether its C-index is the highest reported for LUAD — and we have tried to write both the Results and this Discussion so that a careful reader can make that judgement without needing to take our word for it.

References

[1] H. Sung *et al.*, “Global Cancer Statistics 2020: {GLOBOCAN} Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] W. D. Travis *et al.*, “The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification,” *Journal of Thoracic Oncology*, vol. 10, no. 9, pp. 1243–1260, 2015, doi: 10.1097/JTO.0000000000000630.

[3] R. S. Herbst, D. Morgensztern, and C. Boshoff, “The Biology and Management of Non-Small Cell Lung Cancer,” *Nature*, vol. 553, no. 7689, pp. 446–454, 2018, doi: 10.1038/nature25183.

[4] P. Goldstraw *et al.*, “The {IASLC} Lung Cancer Staging Project: Proposals for Revision of the {TNM} Stage Groupings in the Forthcoming (Eighth) Edition of the {TNM} Classification for Lung Cancer,” *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39–51, 2016, doi: 10.1016/j.jtho.2015.09.009.

[5] F. C. Detterbeck, D. J. Boffa, A. W. Kim, and L. T. Tanoue, “The Eighth Edition Lung Cancer Stage Classification,” *Chest*, vol. 151, no. 1, pp. 193–203, 2017, doi: 10.1016/j.chest.2016.10.010.

[6] H. Uramoto and F. Tanaka, “Recurrence After Surgery in Patients with {NSCLC},” *Transl. Lung Cancer Res.*, vol. 3, no. 4, pp. 242–249, 2014, doi: 10.3978/j.issn.2218-6751.2013.12.05.

[7] L. J. van ’t Veer and R. Bernards, “Enabling Personalized Cancer Medicine Through Analysis of Gene-Expression Patterns,” *Nature*, vol. 452, no. 7187, pp. 564–570, 2008, doi: 10.1038/nature06915.

[8] Cancer Genome Atlas Research Network, “Comprehensive Molecular Profiling of Lung Adenocarcinoma,” *Nature*, vol. 511, no. 7511, pp. 543–550, 2014, doi: 10.1038/nature13385.

[9] T. Barrett *et al.*, “{NCBI} {GEO}: Archive for Functional Genomics Data Sets---Update,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991--D995, 2013, doi: 10.1093/nar/gks1193.

[10] A. Colaprico *et al.*, “{TCGAbiolinks}: An {R/Bioconductor} Package for Integrative Analysis of {TCGA} Data,” *Nucleic Acids Res.*, vol. 44, no. 8, p. e71, 2016, doi: 10.1093/nar/gkv1507.

[11] K. Shedden *et al.*, “Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study,” *Nat. Med.*, vol. 14, no. 8, pp. 822–827, 2008, doi: 10.1038/nm.1790.

[12] C.-Q. Zhu *et al.*, “Prognostic and Predictive Gene Signature for Adjuvant Chemotherapy in Resected Non-Small-Cell Lung Cancer,” *Journal of Clinical Oncology*, vol. 28, no. 29, pp. 4417–4424, 2010, doi: 10.1200/JCO.2009.26.4325.

[13] S. D. Der *et al.*, “Validation of a Histology-Independent Prognostic Gene Signature for Early-Stage, Non-Small-Cell Lung Cancer Including Stage {IA} Patients,” *Journal of Thoracic Oncology*, vol. 9, no. 1, pp. 59–64, 2014, doi: 10.1097/JTO.0000000000000042.

[14] H.-Y. Chen *et al.*, “A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer,” *New England Journal of Medicine*, vol. 356, no. 1, pp. 11–20, 2007, doi: 10.1056/NEJMoa060096.

[15] H. Tang *et al.*, “A 12-Gene Set Predicts Survival of Patients with Lung Adenocarcinoma,” *Clinical Cancer Research*, vol. 19, no. 6, pp. 1577–1586, 2013, doi: 10.1158/1078-0432.CCR-12-2321.

[16] J. R. Kratz *et al.*, “A Practical Molecular Assay to Predict Survival in Resected Non-Squamous, Non-Small-Cell Lung Cancer: Development and International Validation Studies,” *Lancet*, vol. 379, no. 9818, pp. 823–832, 2012, doi: 10.1016/S0140-6736(11)61941-7.

[17] J. P. A. Ioannidis, “Why Most Published Research Findings Are False,” *PLoS Med.*, vol. 2, no. 8, p. e124, 2005, doi: 10.1371/journal.pmed.0020124.

[18] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard, “Navigating the Pitfalls of Applying Machine Learning in Genomics,” *Nat. Rev. Genet.*, vol. 23, no. 3, pp. 169–181, 2022, doi: 10.1038/s41576-021-00434-9.

[19] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, “Pitfalls in the Use of {DNA} Microarray Data for Diagnostic and Prognostic Classification,” *J. Natl. Cancer Inst.*, vol. 95, no. 1, pp. 14–18, 2003, doi: 10.1093/jnci/95.1.14.

[20] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine Learning Algorithm Validation with a Limited Sample Size,” *PLoS One*, vol. 14, no. 11, p. e0224365, 2019, doi: 10.1371/journal.pone.0224365.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "{edgeR}: A {Bioconductor} Package for Differential Expression Analysis of Digital Gene Expression Data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010, doi: 10.1093/bioinformatics/btp616.

[22] M. I. Love, W. Huber, and S. Anders, "Moderated Estimation of Fold Change and Dispersion for {RNA-seq} Data with {DESeq2}," *Genome Biol.*, vol. 15, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[23] M. E. Ritchie *et al.*, "{limma} Powers Differential Expression Analyses for {RNA}-Sequencing and Microarray Studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, 2015, doi: 10.1093/nar/gkv007.

[24] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "{voom}: Precision Weights Unlock Linear Model Analysis Tools for {RNA-seq} Read Counts," *Genome Biol.*, vol. 15, p. R29, 2014, doi: 10.1186/gb-2014-15-2-r29.

[25] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for {Cox}'s Proportional Hazards Model via Coordinate Descent," *J. Stat. Softw.*, vol. 39, no. 5, pp. 1–13, 2011, doi: 10.18637/jss.v039.i05.

[26] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[27] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random Survival Forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008, doi: 10.1214/08-AOAS169.

[28] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[29] T. Chen and C. Guestrin, "{XGBoost}: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, New York, NY, USA: ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[30] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "{DeepSurv}: Personalized Treatment Recommender System Using a {Cox} Proportional Hazards Deep Neural Network," *BMC Med. Res. Methodol.*, vol. 18, p. 24, 2018, doi: 10.1186/s12874-018-0482-1.

[31] T. Ching, X. Zhu, and L. X. Garmire, "{Cox-nnet}: An Artificial Neural Network Method for Prognosis Prediction of High-Throughput Omics Data," *PLoS Comput. Biol.*, vol. 14, no. 4, p. e1006076, 2018, doi: 10.1371/journal.pcbi.1006076.

[32] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the Yield of Medical Tests," *JAMA*, vol. 247, no. 18, pp. 2543–2546, 1982, doi: 10.1001/jama.1982.03320430047030.

[33] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Stat. Med.*, vol. 15, no. 4, pp. 361–387, 1996, doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[34] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the {C}-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data," *Stat. Med.*, vol. 30, no. 10, pp. 1105–1117, 2011, doi: 10.1002/sim.4154.

[35] H. Uno *et al.*, "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis," *Journal of Clinical Oncology*, vol. 32, no. 22, pp. 2380–2385, 2014, doi: 10.1200/JCO.2014.55.2208.

[36] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, "A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data," *J. Clin. Epidemiol.*, vol. 74, pp. 167–176, 2016, doi: 10.1016/j.jclinepi.2015.12.005.

[37] P. C. Austin, F. E. Harrell, and D. van Klaveren, "Graphical Calibration Curves and the Integrated Calibration Index ({ICI}) for Survival Models," *Stat. Med.*, vol. 39, no. 21, pp. 2714–2742, 2020, doi: 10.1002/sim.8570.

[38] A. J. Vickers and E. B. Elkin, "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.

[39] M. J. Pencina, R. B. D'Agostino, R. B. D'Agostino, and R. S. Vasan, "Evaluating the Added Predictive Ability of a New Marker: From Area Under the {ROC} Curve to Reclassification and Beyond," *Stat. Med.*, vol. 27, no. 2, pp. 157–172, 2008, doi: 10.1002/sim.2929.

[40] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958, doi: 10.1080/01621459.1958.10501452.

[41] P. Royston and M. K. B. Parmar, “Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome,” *BMC Med. Res. Methodol.*, vol. 13, p. 152, 2013, doi: 10.1186/1471-2288-13-152.

[42] N. Mantel, “Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration,” *Cancer Chemother. Rep.*, vol. 50, no. 3, pp. 163–170, 1966.

[43] A. Iasonos, D. Schrag, G. V Raj, and K. S. Panageas, “How to Build and Interpret a Nomogram for Cancer Prognosis,” *Journal of Clinical Oncology*, vol. 26, no. 8, pp. 1364–1370, 2008, doi: 10.1200/JCO.2007.12.9791.

[44] S. van Buuren and K. Groothuis-Oudshoorn, “{mice}: Multivariate Imputation by Chained Equations in {R},” *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.

[45] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Curran Associates, Inc., 2017, pp. 4765–4774.

[46] M. Ashburner *et al.*, “Gene Ontology: Tool for the Unification of Biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000, doi: 10.1038/75556.

[47] M. Kanehisa and S. Goto, “{KEGG}: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000, doi: 10.1093/nar/28.1.27.

[48] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database ({MSigDB}) Hallmark Gene Set Collection,” *Cell Syst.*, vol. 1, no. 6, pp. 417–425, 2015, doi: 10.1016/j.cels.2015.12.004.

[49] A. Subramanian *et al.*, “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005, doi: 10.1073/pnas.0506580102.

[50] S. Hänzelmann, R. Castelo, and J. Guinney, “{GSVA}: Gene Set Variation Analysis for Microarray and {RNA-seq} Data,” *BMC Bioinformatics*, vol. 14, p. 7, 2013, doi: 10.1186/1471-2105-14-7.

[51] A. M. Newman *et al.*, “Robust Enumeration of Cell Subsets from Tissue Expression Profiles,” *Nat. Methods*, vol. 12, no. 5, pp. 453–457, 2015, doi: 10.1038/nmeth.3337.

[52] B. Chen, M. S. Khodadoust, C. L. Liu, A. M. Newman, and A. A. Alizadeh, “Profiling Tumor Infiltrating Immune Cells with {CIBERSORT},” in *Cancer Systems Biology: Methods and*

Protocols, vol. 1711, in Methods in Molecular Biology, vol. 1711. , New York, NY: Humana Press, 2018, pp. 243–259. doi: 10.1007/978-1-4939-7493-1_12.

[53] L. McInnes, J. Healy, and J. Melville, "{UMAP}: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018. doi: 10.48550/arXiv.1802.03426.

[54] L. van der Maaten and G. Hinton, "Visualizing Data Using {t-SNE}," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[55] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and Molecular Pattern Discovery Using Matrix Factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004, doi: 10.1073/pnas.0308531101.

[56] C. Contal and J. O'Quigley, "An Application of Changepoint Methods in Studying the Effect of Age on Survival in Breast Cancer," *Comput. Stat. Data Anal.*, vol. 30, no. 3, pp. 253–270, 1999, doi: 10.1016/S0167-9473(98)00096-6.

[57] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting Batch Effects in Microarray Expression Data Using Empirical {Bayes} Methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007, doi: 10.1093/biostatistics/kxj037.

[58] R. DerSimonian and N. Laird, "Meta-analysis in Clinical Trials," *Control. Clin. Trials*, vol. 7, no. 3, pp. 177–188, 1986, doi: 10.1016/0197-2456(86)90046-2.

[59] J. P. T. Higgins and S. G. Thompson, "Quantifying Heterogeneity in a Meta-Analysis," *Stat. Med.*, vol. 21, no. 11, pp. 1539–1558, 2002, doi: 10.1002/sim.1186.

[60] D. R. Spigel *et al.*, "Adjuvant Chemotherapy for Stage {IA--IIA} Non-Squamous, Non-Small-Cell Lung Cancer Identified as Molecular High-Risk by a 14-Gene Expression Profile ({AIM-HIGH}): An International, Randomised, Phase 3 Trial," *Lancet Respiratory Medicine*, vol. 13, no. 10, pp. 887–896, 2025, doi: 10.1016/S2213-2600(25)00213-9.

[61] J. Yang, Y. Liu, X. Mai, S. Lu, L. Jin, and X. Tai, "{STAT1}-Induced Upregulation of {LINC00467} Promotes the Proliferation Migration of Lung Adenocarcinoma Cells by Epigenetically Silencing {DKK1} to Activate {Wnt}/$β$-Catenin Signaling Pathway," *Biochemistry and Biophysical Research Communications*, vol. 514, no. 1, pp. 118–126, 2019, doi: 10.1016/j.bbrc.2019.04.107.

[62] Q. Zhang, M. Zhao, S. Lin, and others, "Prediction of Prognosis and Immunotherapy Response in Lung Adenocarcinoma Based on {CD79A}, {DKK1} and {VEGFC}," *Heliyon*, vol. 9, no. 8, p. e18503, 2023, doi: 10.1016/j.heliyon.2023.e18503.

[63] A. Zeybek and others, "{DKK1} Expression and Prognosis in Non-Small-Cell Lung Cancer," *Mol. Biol. Rep.*, 2021.

[64] J. Wang and Y. Chen, "{CCL20} in Non-Small-Cell Lung Cancer," *Int. Immunopharmacol.*, 2020.

[65] S. Liu *et al.*, "Development and Validation of a Basement Membrane Inflammatory Response Gene Signature in Lung Adenocarcinoma," *Sci. Rep.*, vol. 15, p. 21700, 2025, doi: 10.1038/s41598-025-05582-0.

[66] J. Chen *et al.*, "Single-Cell Transcriptome and Antigen-Immunoglobin Analysis Reveals the Diversity of {B} Cells in Non-Small Cell Lung Cancer," *Genome Biol.*, vol. 21, no. 1, p. 152, 2020, doi: 10.1186/s13059-020-02064-6.

[67] G. Wu *et al.*, "{LncRNA} {BCAN-AS1} Stabilizes c-Myc via {N6}-Methyladenosine-Mediated Binding with {SNIP1} to Promote Pancreatic Cancer," *Cell Death Differ.*, vol. 30, pp. 2213–2230, 2023, doi: 10.1038/s41418-023-01213-2.

[68] D. Li and others, "{RP11-284F21.9} ({BCAN-AS1}) Promotes Lung Carcinoma Proliferation and Invasion via the Regulation of {miR-627-3p}/{CCAR1}," *Oncol. Rep.*, 2020.

[69] G. S. Collins *et al.*, "{TRIPOD+AI} Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods," *BMJ*, vol. 385, p. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[70] K. G. M. Moons *et al.*, "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis ({TRIPOD}): Explanation and Elaboration," *Ann. Intern. Med.*, vol. 162, no. 1, pp. W1--W73, 2015, doi: 10.7326/M14-0698.

[71] K. G. M. Moons *et al.*, "{PROBAST+AI}: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods," *BMJ*, vol. 388, p. e082505, 2025, doi: 10.1136/bmj-2024-082505.

[72] R. F. Wolff *et al.*, "{PROBAST}: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.

[73] L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, and G. M. Clark, "Reporting Recommendations for Tumor Marker Prognostic Studies ({REMARK})," *J. Natl. Cancer Inst.*, vol. 97, no. 16, pp. 1180–1184, 2005, doi: 10.1093/jnci/dji237.