Sriwijaya Journal of Otorhinolaryngology e-ISSN 2987-131X
https://phlox.or.id/index.php/sjorl Vol. 3 No. 2 (2025)

Original Research Article

Cross-Cohort Transcriptomic Analysis Identifies an ECM–CAF Stromal Program but Does Not Validate a Seven-Gene Prognostic Score in Head and Neck Squamous Cell Carcinoma

Reisha Notonegoro[1], Bjorka Alma[2], Patricia Wulandari[3], Muhammad Yoshandi[4,*]

[1]*Department of Radiology, Bintan Family Hospital, Bintan, Indonesia*
[2]*Department of Molecular Medicine, Ordubad State Hospital, Ordubad, Azerbaijan*
[3]*Department of Psychiatry, CMHC Research Center, Palembang, Indonesia*
[4]*Department of Health Sciences, Tembilahan Community Health Center, Tembilahan, Indonesia*

ARTICLE INFO

Article history:
Received: *September 2, 2025*
Accepted: *October 1, 2025*
Published: *October 30, 2025*

Keywords:
Cancer-associated fibroblasts
Gene expression profiling
Head and neck squamous cell carcinoma
Prognosis
Tumor microenvironment

***Corresponding author:**
Muhammad Yoshandi
E-mail:
muh.yoshandi@phlox.or.id
ORCID: 0009-0005-1272-3331

DOI: 10.59345/sjorl.v3i2.307

ABSTRACT

Background: Head and neck squamous cell carcinoma (HNSCC) is biologically heterogeneous, and many public-data gene signatures lack cross-platform replication and unbiased evaluation.

Objective: We sought reproducible tumor–normal expression programs and tested the transportability of a derived seven-gene score.

Methods: TCGA-HNSC RNA-sequencing defined differentially expressed genes (DEGs) between 520 tumors and 44 normal tissues using TMM normalization and voom–limma, with paired sensitivity analysis. Concordant genes were replicated in 22 GSE6631 matched pairs and analyzed for GO and KEGG enrichment. A seven-gene overall-survival score was developed in event-stratified TCGA training data (n=361), evaluated in held-out TCGA (n=156), repeated nested resampling, and examined in GSE65858 (n=253). ESTIMATE and marker scores characterized tumor-microenvironment features.

Results: Among 5,657 TCGA and 168 GSE6631 DEGs, 151 replicated. Upregulated genes were enriched for extracellular-matrix organization, ECM–receptor interaction, integrin signaling, focal adhesion, and PI3K–Akt signaling. The score was associated with survival in training (HR per SD=1.68; C-index=0.642) but not held-out TCGA (HR=1.02; C-index=0.546) or unadjusted GSE65858 (HR=1.15; C-index=0.548). Median nested C-index was 0.564. HPV adjustment attenuated the GSE65858 association. The score correlated with CAF/fibroblast (ρ=0.269) and stromal scores (ρ=0.228).

Conclusion: Replicated expression changes support an ECM–CAF program, but the seven-gene score did not generalize and is not a validated prognostic model.

All authors have reviewed and approved the final version of the manuscript.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) encompasses malignancies arising from the mucosal epithelium of the oral cavity, oropharynx, hypopharynx, and larynx and contributes substantially to the contemporary global cancer burden.[1] HNSCC is not a single biological entity: tumors from different anatomic subsites differ in etiology, cellular composition, molecular drivers, lymphatic behavior, treatment response, and survival. Recent single-cell profiling has further demonstrated that human papillomavirus (HPV) status is associated with distinct epithelial, stromal, and immune states.[2]

Treatment requires close integration of otorhinolaryngology–head and neck surgery, radiation oncology, medical oncology, pathology, radiology, dental care, rehabilitation, nutrition, and supportive care. Surgery, radiotherapy, and systemic therapy can achieve durable control, but treatment may compromise speech, swallowing, airway protection, appearance, and quality of life. Contemporary neoadjuvant immunotherapy studies demonstrate biological activity while also underscoring that response is heterogeneous and requires careful clinical validation.[3] Conventional prognostic factors—including tumor site, tumor–node–metastasis stage, HPV status, performance status, comorbidity, smoking exposure, surgical margins, extranodal extension, and treatment modality—remain clinically essential. Nevertheless, patients with apparently similar clinicopathological characteristics may experience divergent outcomes. Molecular biomarkers could therefore complement, but not replace, established clinical assessment.

Bulk RNA sequencing offers a composite view of HNSCC biology: the measured transcriptome reflects both malignant cells and non-malignant components such as fibroblasts, endothelial cells, macrophages, lymphocytes, and extracellular matrix (ECM)-producing stromal populations. Consequently,

expression-based prognosis may arise from tumor-intrinsic biology, cellular composition, or both. Integrated single-cell and bulk analyses have specifically linked HNSCC fibroblast states with clinically relevant stromal programs.[4]

The tumor microenvironment (TME) is therefore especially relevant in HNSCC. Recent single-cell studies have mapped diverse malignant, stromal, myeloid, and lymphoid populations and their intercellular communication.[2] Cancer-associated fibroblasts (CAFs) can deposit and remodel ECM, support invasion, alter metabolism, and shape immune exclusion or dysfunction.[4] In bulk expression data, a stromal-associated risk signal may therefore be biologically informative even when it does not provide sufficiently accurate patient-level prognostic predictions.

Numerous recent HNSCC gene signatures have been proposed using microenvironmental, stemness, metastatic, HPV-associated immune, programmed-cell-death, early-stage immune, splicing-factor, senescence, disulfidptosis, single-biomarker, and cancer-stem-cell features.[5-15] Additional models have combined infiltrating immune-cell or chemokine-receptor signals with survival outcomes.[16,17] Many studies use TCGA for model development and GSE65858 for evaluation, creating a crowded literature in which different signatures may appear statistically significant despite limited agreement in selected genes. Common vulnerabilities include feature selection conducted before data splitting, repeated testing without adequate multiplicity control, dichotomization optimized separately within each cohort, incomplete reporting of calibration or discrimination, and reliance on the same public dataset across multiple studies. A statistically significant Kaplan–Meier separation in the development cohort does not establish transportability or clinical usefulness.

Rigorous prognostic research requires an explicit separation between biomarker discovery, prognostic-factor association, model development, and model-performance evaluation. Contemporary reporting and risk-of-bias frameworks emphasize transparent cohort definition, complete specification of preprocessing and predictors, unbiased evaluation, open science, and clear distinction between model development and performance assessment. In high-dimensional genomic data, regularization can reduce variance, but it does not eliminate optimism, batch effects, endpoint heterogeneity, population shift, or algorithmic unfairness. Negative or weak evaluation is therefore scientifically meaningful and should be reported rather than reframed as success.

This study was designed as a reproducible, multi-cohort computational investigation with three linked but distinct objectives. The primary biological objective was to identify genes differentially expressed between HNSCC and normal mucosa in TCGA and require replication in the independent, patient-matched GSE6631 cohort. The secondary modeling objective was to derive a parsimonious overall-survival score in TCGA and test, without outcome-based refitting, whether it generalized to held-out TCGA patients; GSE65858 provided a cross-platform association analysis rather than fully locked model validation. The third objective was to determine whether the score tracked broad immune or stromal expression features. Tumor–normal replication was not treated as evidence of prognostic validity, and failure in unseen survival data was interpreted as failure of the candidate score irrespective of its development-set significance.

2. Methods

2.1 Study design and reporting framework

This was a retrospective secondary analysis of publicly accessible, de-identified transcriptomic and clinical data. The analytical workflow separated differential-expression discovery and replication, functional enrichment, exploratory time-to-event model development, performance evaluation, and TME characterization. TCGA-HNSC served as the RNA-sequencing discovery and model-development cohort. GSE65858 provided cross-platform evaluation of association and ranking because it includes expression, overall survival, stage, tumor site, treatment category, and transcriptionally active HPV metadata.[18] GSE6631 replicated tumor–normal differential expression using patient-matched tumor and histologically uninvolved mucosa.[19] GSE65858 was not considered a strict external validation of a deployable model because its microarray data required cohort-specific standardization.

The report was structured with reference to the TRIPOD statement, and methodological quality and risk-of-bias interpretation were informed by PROBAST. We distinguish an association score from an individual prediction model throughout. No absolute survival probabilities, clinical decision thresholds, treatment recommendations, or claims of clinical utility were generated. A completed reporting checklist with manuscript locations should accompany journal submission.

All analyses were performed from a scripted and checksummed project directory. A fixed random seed of 20260721 was used for the TCGA split and resampling. Data-derived thresholds, coefficients, transformations, selection frequencies, and out-of-fold predictions were preserved in machine-readable outputs. The study was exploratory and was not preregistered. The public cohorts fixed the available sample size; uncertainty, event counts, and precision requirements for external evaluation were therefore emphasized rather than applying a binary post hoc power label.[20]

2.2 Data sources and cohort assembly

The TCGA-HNSC expression matrix was downloaded on July 21, 2026 from the UCSC Xena/GDC mirror as a project-level GENCODE v36 STAR-count matrix, with

corresponding Xena clinical annotations. The GDC STAR pipeline quantifies gene-level unstranded read counts after GRCh38 alignment. The mirrored matrix documentation specified $\log_2$(count+1); values were consequently back-transformed before count-based normalization. This reconstruction yields count-like, potentially non-integer values rather than original aliquot-level GDC files and is explicitly treated as a technical limitation. Gene identifiers were mapped using the retained GENCODE v36 probe map. When multiple identifiers mapped to one symbol, the row with the greatest variance was retained under a rule applied without survival information.

Sample-type codes identified primary tumors (01), solid tissue normals (11), and metastatic samples. The differential-expression discovery analysis included 520 primary tumors and 44 normal samples; two metastatic samples were excluded. Patient identifiers were derived from TCGA barcodes, and the first available primary-tumor aliquot represented each patient in survival analysis. Forty-three patients with both a primary tumor and a solid normal specimen were retained for a paired sensitivity analysis. Overall-survival time was days to death for deceased patients and days to last follow-up for living patients. Exclusion of non-positive or missing time or event status left 517 individuals and 220 deaths.

GSE6631 and GSE65858 were downloaded from NCBI GEO on July 21, 2026. GSE6631 was profiled on the Affymetrix Human Genome U95 Version 2 Array and includes 44 profiles from 22 patients, each contributing tumor and clinically uninvolved matched mucosa.[19] The public metadata did not establish that these samples represented exposure-free healthy mucosa; field cancerization and tissue composition remain possible. GSE65858 used the Illumina HumanHT-12 v4 platform. After the source study's quality control, 270 arrays were available;[18] restriction to primary tumors with valid positive overall-survival time and event status yielded 253 patients and 85 deaths. Platform annotations mapped probes to symbols. The maximum-variance probe was retained when multiple probes mapped to one gene; this rule was outcome independent but may favor noisy probes and is considered a limitation.

2.3 Expression preprocessing and quality control

For TCGA, the reconstructed count-like matrix was converted to a DGEList object. Low-information genes were excluded using expression-aware filtering conditional on the design. TMM normalization addressed library-composition differences. Voom estimated the mean–variance relationship and observation-level precision weights before limma modeling. The paired TCGA analysis provided an important sensitivity check against imbalance in the primary unpaired tumor–normal comparison; it did not resolve possible batch, tissue-source, or reconstructed-count effects.

For GEO matrices, numeric values were inspected and $\log_2$ transformed if their upper distribution indicated an untransformed intensity scale. GSE65858 was already $\log_2$ transformed and normalized in the source study. Sample annotations were parsed directly from series-matrix metadata. Expression matrices were aligned to metadata by accession identifier. Genes with missing or ambiguous symbols were excluded, and duplicated symbols were collapsed as described above.

Data-integrity checks included expected file size where available, gzip testing, sample-expression alignment, and SHA-256 checksums for every retained input. Cohort counts, events, pairs, clinical missingness, data exclusions, package versions, and input provenance were written to audit outputs. No sample was removed on the basis of risk score or survival outcome beyond the stated requirement for a positive observed survival time and known event indicator. Outcome data from the held-out TCGA subset and GSE65858 were not used to select the seven genes or estimate their coefficients.

2.4 Differential gene-expression analysis

For the primary TCGA comparison, a design matrix contrasted primary tumors with solid tissue normals. Moderated linear models were fitted to voom-transformed expression values, and empirical-Bayes variance shrinkage was applied with robust estimation. A TCGA DEG was defined by an absolute $\log_2$ fold change of at least 1 and a Benjamini–Hochberg FDR below 0.01. The paired TCGA sensitivity analysis incorporated patient as a blocking factor in the design matrix and applied the same fold-change and FDR criteria.

For GSE6631, a paired limma model included patient identity and tumor–normal status. A GEO DEG was defined by an absolute $\log_2$ fold change of at least 1 and FDR below 0.05. A replicated DEG had to be significant in both the primary TCGA comparison and GSE6631, have the same gene symbol, and show the same direction of fold change. This rule was deliberately more stringent than using a discovery P-value alone and was intended to prioritize cross-platform disease signals over platform-specific findings.

2.5 Functional enrichment analysis

Replicated upregulated and downregulated genes were analyzed separately using the official g:Profiler application-programming interface. The statistical background was not the entire genome; it was the intersection of genes that passed expression processing and were testable in both TCGA and GSE6631. This custom background reduces bias caused by platform content and filtering. Over-representation analysis included GO Biological Process, Molecular Function, and Cellular Component categories as well as KEGG pathways. Multiple testing was controlled using the g:Profiler correction, and terms with adjusted P below 0.05 were considered significant. Direction-specific interpretation was retained because mixing increased and decreased genes can obscure biological meaning.

2.6 Candidate-gene restriction and survival endpoint

Only replicated DEGs measurable in the TCGA tumor expression matrix and GSE65858 were eligible for prognostic modeling. This requirement ensured that every candidate had evidence of differential expression in two tumor–normal comparisons and could be evaluated on the external platform. Overall survival was the sole time-to-event endpoint. Time was measured in days from the source clinical record. The event was death from any cause. No disease-specific survival inference was made.

Clinical covariates in the common adjusted model were age, source-record sex, and stage. Stage I–II was classified as early and III–IV as advanced; source cohorts predominantly used AJCC/UICC seventh-edition staging, which does not implement the current HPV-mediated oropharyngeal framework. Stage was missing for one TCGA patient, so complete-case analysis removed one individual and multiple imputation was not warranted. HPV status was available for only 120 of 517 TCGA patients and was not used for primary adjustment because this would select a small non-representative subset. Transcriptionally active HPV was available for 233 of 253 GSE65858 patients and was included in an additional sensitivity model. Tumor site and broad treatment category were summarized descriptively; treatment could not be harmonized sufficiently for causal or treatment-predictive inference.

2.7 Training and held-out internal testing

The 517 survival-eligible TCGA patients were divided into 70% training and 30% held-out test subsets, stratified by event status. This produced 361 training patients with 154 deaths and 156 test patients with 66 deaths. All gene screening, scaling parameters, stability selection, coefficient estimation, and the training risk threshold were derived without using the held-out test outcomes.

Candidate gene expression in the training set was centered and scaled. Each candidate was first assessed with a univariable Cox proportional-hazards model. Genes with P below 0.01 entered regularized selection; if fewer than ten passed, the script advanced the ten strongest associations as an algorithmic fallback. This outcome-guided screening was confined to training data but remains an exploratory dimension-reduction procedure that can induce selection optimism.

2.8 Stability selection and elastic-net Cox modeling

Feature stability was assessed using 50 repetitions of five-fold cross-validated LASSO Cox regression. Within each repetition, folds were randomly assigned under a repetition-specific seed. A gene's selection frequency was the proportion of repetitions in which its coefficient was non-zero at the cross-validated minimum-deviance penalty. The minimum-deviance penalty was used for stability counting because the more conservative one-standard-error penalty frequently selected the null model in smaller repeated folds. Genes selected in at least 30% of repetitions were considered stable, with a maximum of ten features advanced to the final model.

The final candidate signature was fitted in the TCGA training set using an elastic-net Cox model with $\alpha=0.5$ and ten-fold cross-validation. Coefficients were extracted at the one-standard-error penalty when this retained at least two genes; otherwise, the minimum-deviance penalty was used. Elastic net was selected because it combines L1 shrinkage with L2 stabilization when predictors are correlated. The linear score was the sum of standardized gene-expression values multiplied by fitted log-hazard coefficients. TCGA test values used training-set centers and scales, and the training median defined TCGA groups. For GSE65858, genes were standardized within that microarray cohort to reduce assay-scale incompatibility and TCGA-derived coefficients were retained. Because one patient's standardized score depends on the external cohort distribution, this is not a locked single-patient algorithm. The GSE65858 median was used only for descriptive Kaplan–Meier display. These results are termed external association or transport analysis, not definitive external model validation.

2.9 Prognostic performance and model comparison

The continuous risk score was evaluated per one-standard-deviation increase using Cox proportional-hazards regression. We report HRs with 95% confidence intervals, Wald P values, Harrell C-indices with standard errors and approximate 95% confidence intervals, and Schoenfeld-residual tests for every covariate and the global model. Adjusted models included age, sex, and stage. Incremental association was explored by comparing age/sex/stage with age/sex/stage plus score using change in C-index and a likelihood-ratio test. These comparisons were not interpreted as clinical utility because no absolute-risk model or decision context was specified.

Kaplan–Meier curves included group sizes, censoring, and numbers at risk; log-rank tests were descriptive. Because median dichotomization discards information and varies across cohorts, continuous-score estimates and discrimination were primary. Reverse Kaplan–Meier methods estimated median follow-up. No coefficient was refitted in the held-out TCGA set or GSE65858.

2.10 Repeated nested resampling of the complete survival-selection pipeline

To quantify instability beyond a single split, we performed five-times repeated five-fold nested resampling in all 517 TCGA survival-eligible tumors. Each outer training fold independently repeated gene centering and scaling, univariable Cox screening, ten repetitions of five-fold LASSO stability selection, the 30% stability threshold, and cross-validated elastic-net Cox fitting. The reduced ten-repetition stability loop made full nesting computationally feasible and was applied identically in all 25 outer models. Training-derived centers, scales, selected genes, and coefficients were applied to the untouched outer fold.

Harrell C-index and HR per standard deviation were calculated from out-of-fold predictions. We report the C-index for each complete repeat, the distribution across 25 outer folds, model size, and gene-selection frequency. This analysis estimates ranking performance of the data-driven procedure; it does not create a deployable absolute-risk model.

2.11 Tumor microenvironment analysis

The ESTIMATE algorithm was applied to TCGA primary-tumor expression profiles to derive stromal, immune, and combined estimate scores. These scores summarize expression patterns associated with non-malignant admixture and should not be interpreted as absolute cell counts. To provide transparent lineage-oriented summaries, predefined marker-gene sets were used to compute z-score averages for CD8/cytotoxic T cells, regulatory T cells, natural killer cells, macrophages, CAFs/fibroblasts, and endothelial cells. Each gene was standardized across tumors before marker averaging. A lineage score required at least two measurable genes.

Associations between the continuous score and each TME score were assessed using Spearman correlation. Wilcoxon tests compared descriptive score groups. Benjamini–Hochberg correction was applied separately to correlation and group-comparison families. We documented overlap between the seven genes and marker sets and calculated rank-based partial correlations between the score and lineage markers while controlling for ESTIMATE stromal score. These quantities represent expression-inferred features, not cell fractions or spatial infiltration. Orthogonal deconvolution, pathology, spatial profiling, immunohistochemistry, or single-cell data would be required for cellular confirmation.[2]

2.12 Clinical heterogeneity and sensitivity analyses

TCGA anatomic-neoplasm subdivisions were grouped as oral cavity, oropharynx, larynx, hypopharynx, or other/unspecified using outcome-independent text rules. GSE65858 site categories were retained from the source record. Kruskal–Wallis tests evaluated descriptive score differences across sites. In GSE65858, a sensitivity Cox model included continuous score, age, sex, stage, and transcriptionally active HPV. Risk-score distributions were compared between HPV-positive and HPV-negative tumors using Wilcoxon testing. These analyses were exploratory and were not powered for interactions, site-specific prognosis, or treatment-effect prediction.

2.13 Statistical software and reproducibility

Analyses were performed in R 4.5.3 using edgeR 4.8.2, limma 3.66.0, survival 3.8.6, glmnet 4.1-10, tidyestimate 1.1.1, data.table, ggplot2, and patchwork. Functional enrichment used the g:Profiler API, database version e114_eg62_p19_27110d83, accessed July 21, 2026. All tests were two sided. Unless a more stringent analysis-specific threshold was defined, FDR below 0.05 indicated statistical significance. Raw-input checksums, processed tables, scripts, out-of-fold predictions, figures, and session metadata were retained. A permanent public archive with DOI and completed TRIPOD and REMARK checklists are required before submission.

3. Results

3.1 Cohort characteristics and analytical flow

The expression discovery dataset comprised 566 TCGA-HNSC profiles: 520 primary tumors, 44 solid normal samples, and two excluded metastatic samples. Forty-three patients contributed matched tumor and normal samples for sensitivity analysis. Overall survival was evaluable for 517 primary-tumor patients with 220 deaths. Median age was 61 years (IQR, 53–69); 382 patients were recorded as male and 135 as female. Stage was available for 516 patients: 110 stage I–II and 406 stage III–IV. Grouped primary sites were oral cavity (n=339), larynx (n=116), oropharynx (n=52), and hypopharynx (n=10). HPV annotation was available for only 120 patients (41 positive, 79 negative). Curative-treatment method was missing for 331 patients; available categories included surgery (n=100), concurrent chemotherapy (n=43), radiotherapy (n=36), and non-concurrent chemotherapy (n=7). Reverse Kaplan–Meier median follow-up was 2.89 years (95% CI 2.66–3.23).

GSE6631 contributed 44 microarray profiles representing 22 matched pairs. GSE65858 contributed 253 eligible primary tumors with 85 deaths. Median age was 58.3 years (IQR, 52.3–68.3); 210 patients were recorded as male and 43 as female. Forty-nine had stage I–II and 204 stage III–IV disease. Sites were oropharynx (n=101), oral cavity/cavum oris (n=78), larynx (n=43), hypopharynx (n=28), and unclassified (n=3). Transcriptionally active HPV was available for 233 patients: 35 positive and 198 negative. Broad treatment categories were multimodal (n=175), monotherapy (n=75), and palliative (n=3). Median follow-up was 2.75 years (95% CI 2.58–2.90). GSE65858 therefore differed from TCGA in platform, site distribution, HPV completeness, treatment annotation, event proportion, and case mix. Cohort composition, data platforms, analytical roles, and follow-up are detailed in Table 1.

Table 1. Cohort characteristics and analytical roles

Characteristic	TCGA-HNSC	GSE6631	GSE65858
Platform and role	STAR-count RNA sequencing; DEG discovery, model development/evaluation, and TME analysis	Affymetrix U95Av2; paired DEG replication	Illumina HumanHT-12 v4; cross-platform survival association
Included expression samples	520 primary tumors and 44 solid normals	22 tumors and 22 matched uninvolved mucosae	253 primary tumors
Survival sample/events	517/220	Not used	253/85
Age, median (IQR), years	61 (53–69)	Not consistently available	58.3 (52.3–68.3)
Recorded sex	382 male; 135 female	Not consistently available	210 male; 43 female
Stage I–II / III–IV	110 / 406; 1 missing	Not used	49 / 204
Main anatomic sites	Oral cavity 339; larynx 116; oropharynx 52; hypopharynx 10	Mucosal HNSCC; paired normal source	Oropharynx 101; oral cavity 78; larynx 43; hypopharynx 28; 3 unclassified
HPV availability	120/517: 41 positive, 79 negative	Not used	233/253: 35 transcriptionally active positive, 198 negative
Treatment availability	Curative method missing in 331/517	Not used	Multimodal 175; monotherapy 75; palliative 3
Median follow-up, years (95% CI)	2.89 (2.66–3.23)	Not applicable	2.75 (2.58–2.90)

3.2 Differential expression in TCGA and paired replication

Using an absolute $\log_2$ fold-change threshold of 1 and FDR below 0.01, the TCGA comparison identified 5,657 DEGs: 2,764 upregulated and 2,893 downregulated in tumor. The paired TCGA sensitivity analysis identified 2,835 upregulated and 3,135 downregulated genes, supporting widespread transcriptional remodeling while controlling for patient-specific background. In GSE6631, using the paired design and FDR below 0.05, 61 genes were upregulated and 107 were downregulated.

Intersection by symbol, statistical significance, and direction produced 151 replicated DEGs, consisting of 54 upregulated and 97 downregulated genes. The smaller replicated set reflected the older microarray's limited probe content, stricter paired comparison, and cross-platform requirement. Nevertheless, fold-change direction among replicated genes was highly coherent. Importantly, all seven genes later included in the candidate prognostic score were also significant with the same direction in the TCGA paired sensitivity analysis, reducing concern that their differential expression was driven solely by unmatched case composition.

The replicated set included pronounced loss of tissue-differentiation or contractile genes and elevation of genes associated with matrix interaction, proteolysis, and motility. *COBL* and *MYL1* were downregulated in tumor in both datasets, whereas *CTSC*, *FSCN1*, *MYO1B*, *PLAU*, and *SEMA3C* were upregulated. These findings indicate that the prognostic candidates emerged from a tumor–normal disease signal reproduced across RNA sequencing and microarray technology rather than from survival screening alone. Cross-platform fold-change concordance and the location of the 151 replicated DEGs are shown in Figure 1.

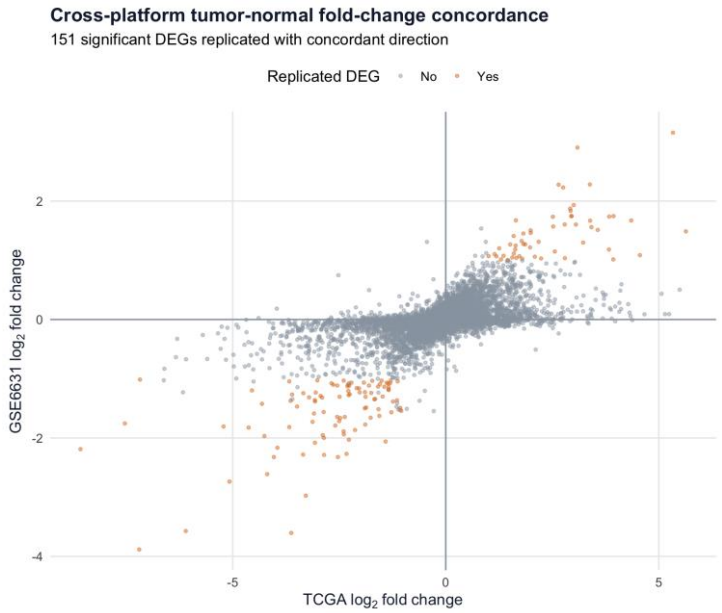

Figure 1. Cross-platform concordance of tumor–normal differential expression. Fold changes from TCGA-HNSC and the paired GSE6631 cohort are shown, with the 151 replicated DEGs highlighted.

3.3 GO and KEGG enrichment identifies an ECM--adhesion program

Functional analysis yielded 366 significant terms: 254 among upregulated replicated genes and 112 among downregulated genes. The upregulated set was dominated by extracellular and stromal biology. The most significant cellular-component term was extracellular matrix (29 of 54 submitted genes; adjusted $P=7.39\times10^{-31}$). Extracellular matrix structural constituent was the leading molecular-function signal (21 genes; adjusted $P=1.71\times10^{-23}$). Extracellular matrix organization and extracellular structure organization were the leading biological processes (23 genes each; adjusted $P=4.23\times10^{-21}$).

KEGG results converged with GO. ECM–receptor interaction included 16 of the 54 upregulated genes and had an adjusted P of 5.58×10^{-19}. Integrin signaling (16 genes; adjusted $P=4.38\times10^{-15}$), focal adhesion (16 genes; adjusted $P=4.44\times10^{-13}$), human papillomavirus infection (17 genes; adjusted $P=9.11\times10^{-12}$), and PI3K–Akt signaling (16 genes; adjusted $P=1.64\times10^{-10}$) were also enriched. The HPV pathway label should not be interpreted as evidence that all contributing tumors were HPV-driven; KEGG pathways share adhesion, cell-cycle, and survival components with broader oncogenic programs.

Downregulated terms included cornified-envelope formation, epidermal development, keratinocyte differentiation, and cytoskeletal or muscle-associated programs. This combination suggests simultaneous activation of matrix-remodeling and motility-related biology with attenuation of mature epithelial differentiation. Because enrichment was based on the cross-cohort replicated set and a custom measurable-gene background, the results are less likely to reflect a single platform's annotation bias. The leading GO and KEGG signals, overlaps, and adjusted P values are detailed in Table 2.

Table 2. Selected enriched terms among replicated upregulated genes

Source	Term	Overlap	Adjusted P
GO:CC	Extracellular matrix	29/202	7.39×10^{-31}
GO:MF	Extracellular matrix structural constituent	21/114	1.71×10^{-23}
GO:BP	Extracellular matrix organization	23/190	4.23×10^{-21}
KEGG	ECM–receptor interaction	16/73	5.58×10^{-19}
KEGG	Integrin signaling	16/129	4.38×10^{-15}
KEGG	Focal adhesion	16/176	4.44×10^{-13}
KEGG	Human papillomavirus infection	17/258	9.11×10^{-12}
KEGG	PI3K–Akt signaling pathway	16/269	1.64×10^{-10}

3.4 Derivation of a seven-gene candidate score

The event-stratified TCGA split assigned 361 patients with 154 deaths to training and 156 patients with 66 deaths to the held-out test subset. Replicated DEGs measurable in both TCGA and GSE65858 entered training-only survival screening. Repeated LASSO stability selection was performed across 50 five-fold cross-validation runs. Seven genes were ultimately retained by the elastic-net Cox model: *COBL*, *CTSC*, *FSCN1*, *MYO1B*, *PLAU*, *MYL1*, and *SEMA3C*. Each was selected in all 50 runs conditional on the single training-set screen. These 100% frequencies must not be interpreted as total-pipeline stability because univariable screening was not repeated within those original runs.

All final coefficients were positive, meaning that a higher standardized value contributed to a higher fitted risk score. This does not mean that every gene was overexpressed in tumor. *COBL* and *MYL1* were substantially downregulated relative to normal tissue, yet higher expression among tumors was associated with higher hazard within the training cohort. Differential expression and within-tumor prognostic association answer different questions and can legitimately have apparently discordant directions.

The largest penalized coefficient belonged to *COBL* (0.198), followed by *CTSC* (0.150), *FSCN1* (0.115), *MYO1B* (0.093), *PLAU* (0.075), *MYL1* (0.053), and *SEMA3C* (0.023). The shrinkage coefficients should be interpreted as components of a composite standardized score, not as unbiased biological effect estimates. The training univariable HRs ranged from 1.24 to 1.34 per standardized expression unit. Gene-level coefficients, tumor–normal directions, fold changes, and conditional selection frequencies are detailed in Table 3; the corresponding conditional stability profile and coefficient distribution are shown in Figure 2.

Table 3. Composition of the candidate seven-gene signature

Gene	Elastic-net coefficient	Tumor direction	TCGA $\log_2$FC	GSE6631 $\log_2$FC	Selection frequency
COBL	0.198	Down	−3.673	−1.045	100%
CTSC	0.150	Up	1.147	1.103	100%
FSCN1	0.115	Up	1.824	1.276	100%
MYO1B	0.093	Up	1.593	1.410	100%
PLAU	0.075	Up	2.515	1.735	100%
MYL1	0.053	Down	−3.952	−2.166	100%
SEMA3C	0.023	Up	1.004	1.070	100%

The fitted development score was 0.197812×z(*COBL*) + 0.150431×z(*CTSC*) + 0.115036×z(*FSCN1*) + 0.093403×z(*MYO1B*) + 0.075179×z(*PLAU*) + 0.052834×z(*MYL1*) + 0.023258×z(*SEMA3C*), where each TCGA z value used the training-set gene mean and standard deviation retained in the machine-readable model output. The formula produces a relative linear predictor only; no baseline survival function or absolute-risk estimate was specified.

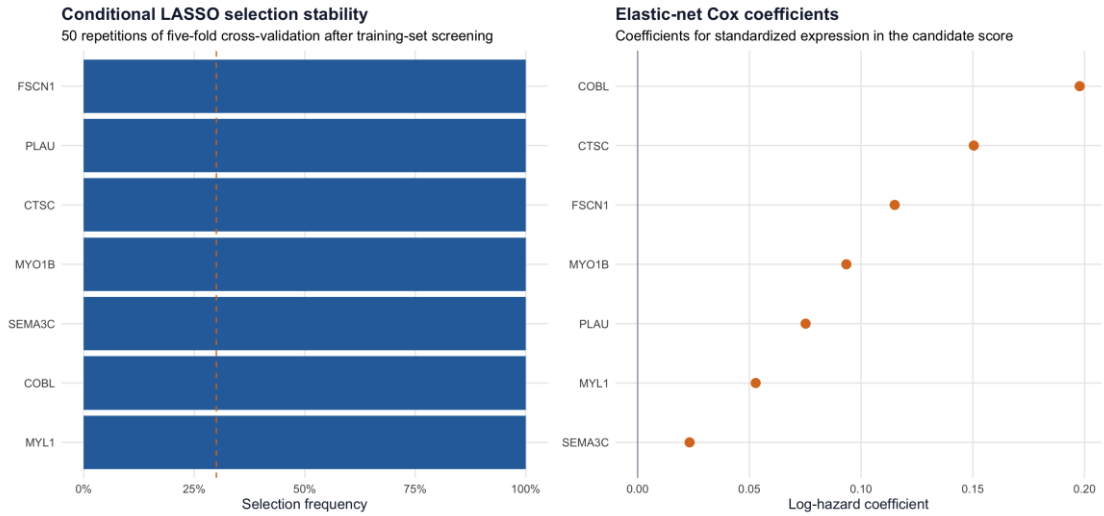

Figure 2. Conditional stability selection and penalized coefficients. LASSO selection frequencies after the training-set screen and elastic-net log-hazard coefficients for the retained genes. Total-pipeline stability is reported separately using nested resampling.

3.5 Training performance does not reproduce in the held-out TCGA test set

In the TCGA training set, a one-standard-deviation increase in the score was associated with a 68% increase in the hazard of death (HR=1.676, 95% CI 1.379–2.037; P=2.18×10^{-7}). The C-index was 0.642 (95% CI 0.595–0.689). After adjustment for age, sex, and stage, the association remained strong (HR=1.704, 95% CI 1.396–2.080; P=1.62×10^{-7}), and the full-model C-index was 0.663 (95% CI 0.616–0.710). Neither the score-specific nor global proportional-hazards test indicated a violation (P=0.136 and P=0.647, respectively).

The held-out test set provided a substantially different result. The unadjusted HR was 1.019 (95% CI 0.798–1.301; P=0.880) and the C-index was 0.546 (95% CI 0.460–0.631), compatible with chance ranking. Adjustment did not recover the molecular signal: the HR was 1.051 (95% CI 0.804–1.374; P=0.716), and the full-model C-index was 0.627 (95% CI 0.551–0.703). Score-specific and global proportional-hazards tests were non-significant (P=0.703 and P=0.447), indicating that non-proportionality did not explain the absence of association.

When all 517 eligible TCGA cases were analyzed together, the score was associated with survival (unadjusted HR=1.391, 95% CI 1.201–1.611; P=1.02×10^{-5}; C-index=0.610, 95% CI 0.568–0.651). This estimate includes development observations and is not an unbiased evaluation result.

3.6 Nested resampling confirms weak and unstable generalization

The complete survival-selection procedure was refitted in 25 outer training sets. Median out-of-fold C-index across the five complete repeats was 0.564 (range, 0.560–0.570). Across individual outer folds, median C-index was 0.574 and the 2.5th–97.5th percentile range was 0.492–0.622. The median selected model contained seven genes (2.5th–97.5th percentile, 4–10 after rounding). Only *PLAU* (23/25 outer models; 92%), *EXPH5* (76%), *MYO1B* (72%), *BLNK* (68%), and *SPINK5* (60%) appeared in at least 60% of outer models. Among the original seven-gene score, *SEMA3C* appeared in 36%, *COBL* in 32%, *MYL1* in 20%, and *CTSC* and *FSCN1* in fewer than 20%. Thus, the original 100% conditional frequencies did not reproduce when screening and fitting were repeated inside outer resamples. These findings support model instability rather than a locked seven-gene predictor.

3.7 Cross-platform association in GSE65858 is weak and HPV-dependent

In GSE65858, the continuous score was not significant in unadjusted analysis (HR per standard deviation=1.152, 95% CI 0.933–1.423; P=0.189). The C-index was 0.548 (95% CI 0.485–0.612), indicating weak ranking. Score-specific and global proportional-hazards tests were non-significant (P=0.596 and P=0.318). Descriptive cohort-median Kaplan–Meier groups did not differ significantly (log-rank P=0.30).

After adjustment for age, sex, and stage, the score was statistically significant (HR=1.277, 95% CI 1.029–1.584; P=0.026), and the full-model C-index was 0.667 (95% CI 0.608–0.727). The clinical model alone had a C-index of 0.634; adding the score increased it by 0.033 (likelihood-ratio P=0.025). However, HPV-positive tumors had substantially lower scores than HPV-negative tumors (median −0.410 vs 0.080; Wilcoxon P=6.69×10^{-11}). Among 233 patients with transcriptionally active HPV data, adding HPV to the

model attenuated the score association to HR=1.090 (95% CI 0.855–1.390; P=0.486), whereas HPV positivity was associated with lower hazard (HR=0.374, 95% CI 0.154–0.910; P=0.030). The earlier age/sex/stage-adjusted association therefore appears materially confounded by viral etiology and should not be interpreted as independent external support.

Incremental model results reinforced this conclusion. In TCGA training, adding the score to age, sex, and stage increased C-index from 0.583 to 0.663 (ΔC=0.080; likelihood-ratio $P=3.44\times10^{-8}$). In the held-out test set, the increase was 0.003, from 0.624 to 0.627 (P=0.716). Across all TCGA cases the increase was 0.046, but this estimate reused development observations. The score also varied substantially by TCGA primary site (Kruskal–Wallis $P=3.76\times10^{-18}$), with median values highest in oral-cavity tumors and lowest in oropharyngeal tumors. Site differences in GSE65858 were weaker (P=0.061). These patterns further limit pan-HNSCC interpretation. Continuous-score effect estimates and discrimination are detailed in Table 4, while descriptive survival curves with numbers at risk are shown in Figure 3.

Table 4. Prognostic performance of the continuous risk score

Cohort/model	n/events	HR per SD (95% CI)	P	C-index
TCGA training, unadjusted	361/154	1.676 (1.379–2.037)	2.18×10^{-7}	0.642 (0.595–0.689)
TCGA training, adjusted	360/154	1.704 (1.396–2.080)	1.62×10^{-7}	0.663 (0.616–0.710)*
TCGA held-out test, unadjusted	156/66	1.019 (0.798–1.301)	0.880	0.546 (0.460–0.631)
TCGA held-out test, adjusted	156/66	1.051 (0.804–1.374)	0.716	0.627 (0.551–0.703)*
GSE65858, unadjusted association	253/85	1.152 (0.933–1.423)	0.189	0.548 (0.485–0.612)
GSE65858, adjusted association	253/85	1.277 (1.029–1.584)	0.026	0.667 (0.608–0.727)*
GSE65858, HPV-adjusted sensitivity	233/73	1.090 (0.855–1.390)	0.486	Not estimated

*C-index for the full model including age, sex, stage, and risk score.

3.8 The score is associated with expression-inferred stromal and CAF features

Among 517 TCGA tumors, the strongest TME association was observed for the CAF/fibroblast marker score (Spearman ρ=0.269; nominal $P=5.39\times10^{-10}$; $FDR=4.85\times10^{-9}$). The ESTIMATE stromal score was also positively correlated with risk (ρ=0.228; $FDR=7.51\times10^{-7}$), as was the combined estimate score (ρ=0.138; FDR=0.0048). Smaller positive correlations were observed for macrophage (ρ=0.126; FDR=0.0089) and endothelial scores (ρ=0.123; FDR=0.0089).

In contrast, the global immune score was not significantly correlated with the score (ρ=0.048; FDR=0.418). CD8/cytotoxic T-cell, regulatory-T-cell, and natural-killer-cell scores were weakly negative and non-significant in unadjusted correlations. None of the seven candidate genes overlapped any lineage marker set, excluding direct gene duplication as the source of these correlations.

After controlling for the ESTIMATE stromal score using rank-based partial correlation, the CAF/fibroblast association persisted but was attenuated (partial ρ=0.147; FDR=0.0025). Partial associations with CD8/cytotoxic, regulatory-T, and NK markers became modestly negative, suggesting that the unadjusted bulk correlations were influenced by general stromal admixture. These exploratory results connect the score with ECM/CAF biology but cannot distinguish cell abundance, cell state, tumor purity, or spatial exclusion. They support biological coherence, not cellular causality or infiltration quantification. The complete correlation estimates are detailed in Table 5, and their direction and FDR status are visualized in Figure 4.

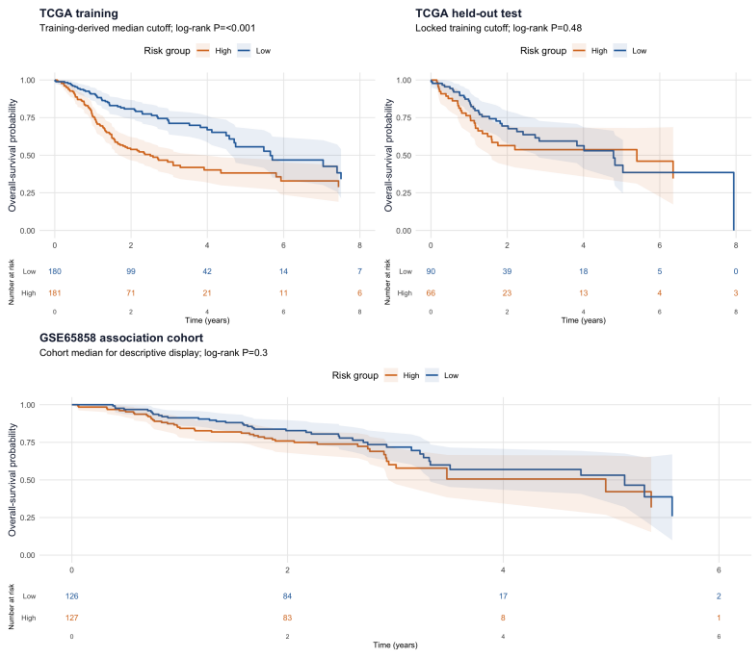

Figure 3. Descriptive Kaplan–Meier analysis with numbers at risk. TCGA groups used the training-derived cutoff; GSE65858 used its cohort median solely for visualization. Continuous-score modeling was primary, and significant separation was confined to training.

Table 5. Correlations between the candidate score and expression-inferred TME features in TCGA

TME feature	Unadjusted ρ	FDR	Stromal-adjusted partial ρ	Partial FDR
CAF/fibroblast	0.269	4.85×10^{-9}	0.147	0.0025
Stromal	0.228	7.51×10^{-7}	Not applicable	Not applicable
Combined estimate	0.138	0.0048	Not estimated	Not estimated
Macrophage	0.126	0.0089	−0.062	0.158
Endothelial	0.123	0.0089	−0.077	0.095
Immune	0.048	0.418	Not estimated	Not estimated
CD8/cytotoxic T	−0.030	0.490	−0.108	0.021
Regulatory T	−0.043	0.428	−0.185	0.00013
Natural killer	−0.032	0.490	−0.115	0.018

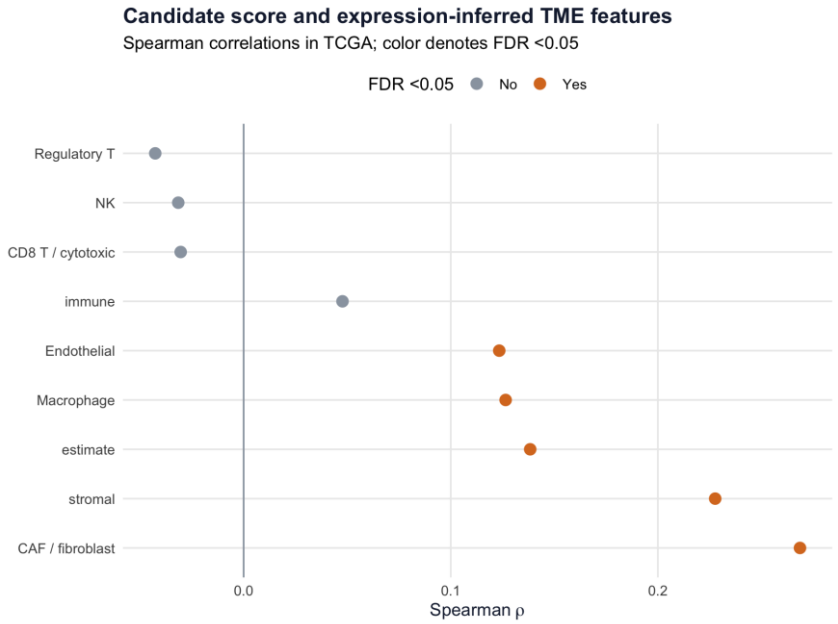

Figure 4. Correlation of the candidate score with expression-inferred tumor-microenvironment features. Colored points denote FDR below 0.05; estimates are bulk-expression correlations and not cell fractions.

4. Discussion

4.1 Principal findings

This cross-cohort study produced two distinct conclusions. The biological conclusion is comparatively robust: HNSCC tumor–normal changes replicated between TCGA RNA sequencing and paired GEO microarrays converge on ECM organization, integrin/focal-adhesion signaling, proteolysis, and stromal remodeling. The modeling conclusion is negative: a seven-gene score showed development-set association but failed in held-out TCGA, yielded a nested-resampling C-index near 0.56, and showed weak cross-platform association in GSE65858. The age/sex/stage-adjusted GSE65858 association disappeared after accounting for transcriptionally active HPV. The score is therefore an exploratory correlate of heterogeneous stromal biology, not a validated prognostic model.

This distinction matters because genomic prognostic literature is vulnerable to selective emphasis. A model can achieve a very small development P value while providing little discrimination for new patients. Here, the training HR of 1.68 and ΔC of 0.080 contrasted with a held-out HR near 1.0, ΔC of 0.003, and instability of the selected genes across outer resamples. Reporting only the complete TCGA cohort or the incompletely adjusted GSE65858 model would have produced a misleading success narrative.

4.2 Cross-cohort differential expression reveals a stromal-remodeling phenotype

The replicated DEG strategy was intended to reduce false discovery and platform dependence. TCGA and GSE6631 differ in technology, era, sample processing, and probe coverage. Requiring significance and direction concordance therefore sacrifices sensitivity but increases confidence that retained signals represent reproducible disease biology. The 151-gene intersection was far smaller than the 5,657 TCGA DEGs, which is expected given the older array and stringent pairing. Yet its functional enrichment was exceptionally coherent.

ECM organization and ECM–receptor interaction are plausible features of HNSCC invasion. Tumor cells encounter and remodel basement membrane and interstitial matrix as they invade locally, track along nerves or lymphatics, and access regional lymph nodes. Integrin and focal-adhesion pathways translate extracellular cues into cytoskeletal organization, survival signaling, migration, and therapy response. PI3K–Akt enrichment is also compatible with growth-factor signaling and stromal–epithelial interaction. The downregulation of epidermal differentiation and cornified-envelope programs complements this picture by suggesting loss of mature squamous identity.

Single-cell evidence provides an important cellular context. Recent HNSCC profiling has described multiple epithelial, immune, and fibroblast states, while integrated single-cell and bulk analyses have linked selected CAF programs with worse overall survival.[2,4] Our bulk expression findings cannot distinguish whether the ECM signal originates from malignant cells, fibroblasts, or their interaction. However, the positive correlation between risk, CAF markers, and ESTIMATE stromal score supports a mixed tumor–stroma interpretation rather than a purely tumor-cell-autonomous program.

The absence of a strong adaptive immune association is also informative. HNSCC immune ecosystems vary with HPV etiology and contain heterogeneous T-cell, B-cell, myeloid, and regulatory states.[2] A global immune score can mask opposing cell states; similarly, a small marker panel cannot resolve exhaustion, spatial exclusion, clonality, or tertiary lymphoid structures. Our results do not imply that immunity is unimportant. They indicate that this specific candidate score aligns more consistently with stroma than with broad CD8, regulatory-T, NK, or total immune signals.

4.3 Biological interpretation of the seven genes

Several signature genes have mechanistic links to invasion or matrix interaction. *MYO1B* encodes an unconventional myosin involved in membrane dynamics and cytoskeletal movement, but its role in the present bulk-transcriptomic score remains inferential. *FSCN1* encodes an actin-bundling protein that promotes cellular protrusions and migration; recent HNSCC work has associated *FSCN1* with radiation resistance in the context of *PIK3CA* alteration.[21] *PLAU* encodes urokinase-type plasminogen activator, a protease that can facilitate pericellular proteolysis and matrix degradation; *PLAU* has also appeared in programmed-cell-death-related HNSCC models.[8] *SEMA3C*, a semaphorin involved in guidance and signaling, has been evaluated through computational and experimental work in tongue squamous cell carcinoma.[22]

CTSC encodes cathepsin C, a lysosomal cysteine protease that activates several granule-associated serine proteases and may influence inflammation, proteolysis, and tumor–myeloid interactions. *COBL* is involved in actin nucleation and neuronal morphogenesis, whereas *MYL1* encodes a skeletal-muscle myosin light chain. The latter two genes were strongly downregulated in tumors compared with normal tissue, possibly reflecting loss of differentiated tissue components or altered cellular composition. Their positive within-tumor coefficients caution against interpreting a prognostic score as a simple tumor-upregulation panel. Gene-specific biological context, likely bulk-expression compartment, and interpretive cautions are detailed in Table 6.

Table 6. Biological context and interpretive cautions for the seven candidate genes

Gene	Principal biological context	Tumor–normal direction	Likely bulk-expression compartment	Main interpretive caution
COBL	Actin nucleation and cellular morphogenesis	Down	Tumor and normal tissue-composition signal uncertain	Positive within-tumor coefficient conflicts with marked tumor downregulation and was selected in only 32% of nested outer models
CTSC	Lysosomal protease activation and inflammatory/myeloid biology	Up	Malignant, myeloid, or mixed	Cell of origin is unresolved; low nested selection frequency
FSCN1	Actin bundling, protrusion formation, migration, and reported radioresistance	Up	Malignant cells and possibly stromal/immune subsets	Functional plausibility does not validate the composite score; low nested selection frequency
MYO1B	Membrane–cytoskeleton dynamics and migration-related biology	Up	Predominantly malignant-cell hypothesis	Selected in 72% of nested outer models, but effect remains assay and site dependent
PLAU	Pericellular proteolysis and ECM degradation	Up	Malignant and stromal compartments	Most stable original gene (92%), but not specific to one cell population
MYL1	Skeletal-muscle contractile program	Down	Likely normal tissue or muscle admixture	Positive within-tumor coefficient may reflect sampling composition; 20% nested frequency
SEMA3C	Semaphorin guidance, migration, and tumor–stroma signaling	Up	Malignant or stromal	Limited HNSCC-specific validation; 36% nested frequency

The exact seven-gene combination was not identified in our focused literature search, but absence of an exact match does not prove novelty. Several individual genes and biological themes have already been reported in HNSCC. Furthermore, the public datasets used here have been repeatedly mined. A defensible novelty claim would require a systematic review of PubMed, Embase, Web of Science, Scopus, preprints, and relevant patents, with explicit search dates and gene-name synonyms. Until then, "candidate seven-gene signature" is more accurate than "novel prognostic signature."

4.4 Comparison with previous HNSCC signatures

The field contains many public-data-derived signatures. Stemness-related and metastatic models have reported associations with survival and immune or stromal features.[6,7,15] Immune-related, HPV-focused, programmed-cell-death, splicing-factor, senescence, and chemokine-receptor-centered models have also been developed.[5,8-14,16,17] Some studies report apparently robust Kaplan–Meier or time-dependent receiver-operating-characteristic performance in GSE65858.

Our study differs in three respects. First, candidate genes were required to show concordant tumor–normal differential expression across RNA sequencing and a matched-pair microarray cohort before survival modeling. Second, a held-out TCGA subset was preserved throughout feature selection and fitting. Third, failure of the primary continuous-score evaluation was retained as the central conclusion. These design choices reduce the chance of presenting a cohort-specific association as a generalizable predictor.

The discrepancy between our weak GSE65858 discrimination and positive results reported for other signatures could reflect genuine differences in gene choice. It could also arise from analytic flexibility, including cohort-specific risk cutoffs, preprocessing within all data before splitting, endpoint definitions, missing-data exclusions, or iterative reuse of GSE65858 during model development. Repeated public use can transform an ostensibly external cohort into a de facto benchmark that influences model choices. Truly independent temporal or geographic validation remains necessary.

Recent cancer-stem-cell and tumor-infiltrating immune-cell models have linked expression patterns to prognosis across public and clinical cohorts.[15,16] These studies underscore that HNSCC prognosis is likely governed by multiple partially overlapping processes rather than a single universal gene panel. A clinically useful model may need to integrate site, HPV biology, stage, treatment, comorbidity, tumor-intrinsic expression, and spatial TME features rather than rely on a small bulk-transcriptome score alone.

4.5 Why did the signature fail held-out validation?

Several explanations are plausible. First, univariable screening followed by repeated LASSO and penalty tuning explores many candidate models. The original 100% conditional selection frequencies did not repeat the upstream screen. Once the complete pipeline was nested, the selected set varied widely and out-of-fold C-index remained approximately 0.56. A single 70:30 split also creates sampling variability, and 66 held-out

events yield a broad interval; nevertheless, the point estimate near unity and nested results provide little support for useful prognostic ranking.

Second, HNSCC heterogeneity destabilized the pan-site score. TCGA scores differed strongly by anatomic site, and HPV-positive GSE65858 tumors had much lower scores than HPV-negative tumors. Adjustment for transcriptionally active HPV removed the GSE65858 score association. The candidate therefore partly encodes differences in viral etiology, anatomy, or associated tissue composition. Treatment pathways also differed, but incomplete and non-comparable annotations prevented treatment-adjusted or treatment-interaction analyses.

Third, cross-platform transfer is difficult. TCGA used RNA sequencing whereas GSE65858 used an Illumina microarray. Cohort-specific external standardization was necessary, so the analysis could not apply a locked single-patient rule. Platform-specific dynamic range, probe specificity, batch effects, RNA quality, and normalization can alter each gene's contribution. A deployable assay would require fixed probe/transcript definitions, training-derived transformations, coefficients, a baseline hazard, and prespecified prediction times.

Fourth, bulk expression confounds abundance and state. A high stromal score may reflect desmoplasia, lower tumor purity, sampling location, necrosis, or the proportion of invasive front captured in the specimen. If these features vary across pathology workflows, the score may not transport. Spatially resolved sampling of the tumor center and invasive margin could provide more reproducible biological measurement.

Finally, the score was optimized for association with overall survival, which is influenced by non-cancer mortality, salvage treatment, second primary tumors, and comorbidity. Disease-specific survival, progression-free survival, locoregional control, and distant metastasis might align differently with ECM-related biology. Endpoint choice should be matched to the intended clinical use.

4.6 Methodological strengths

The study has several strengths. It used three independent public cohorts with non-overlapping roles: TCGA for discovery and model development, GSE6631 for paired tumor–normal replication, and GSE65858 for cross-platform survival association. Inputs were checksummed and cohort assembly was auditable. Differential expression accounted for mean–variance structure and pairing where available. Functional enrichment used a custom tested-gene background and retained directionality.

The survival workflow used a fixed event-stratified split, training-only scaling, shrinkage, continuous-score evaluation, C-index intervals, covariate-specific and global proportional-hazards diagnostics, and comparison with clinical covariates. Additional nested resampling repeated the full selection procedure in 25 outer models, directly exposing instability. GSE65858 outcomes were not used to select genes or tune coefficients, and its cohort-specific preprocessing was disclosed rather than mislabeled as a deployable external validation.

The study also integrated prognosis with TME biology using both ESTIMATE and transparent marker sets. The convergence of replicated DEG enrichment with CAF and stromal correlations provides internal biological coherence. Reproducible scripts, versioned packages, input checksums, and machine-readable results support independent audit.

4.7 Limitations and risk of bias

This study remains exploratory and has important limitations. All cohorts were retrospective and public. Clinical annotation was incomplete, particularly for HPV, treatment, smoking, comorbidity, margins, extranodal extension, and recurrence. The common adjusted model included only age, sex, and broad stage. The GSE65858 age/sex/stage association was eliminated by HPV adjustment, demonstrating important residual confounding rather than stable independent prognostic value.

TCGA normal tissues are fewer than tumors and may not represent truly healthy mucosa. Field cancerization, inflammation, adjacency, and variable stromal or skeletal-muscle content can influence expression. GSE6631 used clinically uninvolved paired mucosa and an older array with limited coverage, but its anatomical sampling details were insufficient to exclude field effects. The TCGA Xena matrix was back-transformed from $\log_2(\text{count}+1)$, producing count-like values rather than original aliquot-level raw-count files. Paired TCGA concordance supports the signal but does not fully validate this technical reconstruction. Replication across datasets improves specificity but can omit relevant genes absent from the older platform.

The original seven-gene derivation still depends on one random training/test split. Repeated nested resampling was added, but only five complete repeats were computationally feasible and the inner stability loop was reduced from 50 to ten repetitions. It estimates procedure-level discrimination and selection variability but does not yield a final locked model or bootstrap-corrected absolute risk. The original 50-run selection frequencies remain conditional on a single upstream screen.

The external cohort contained 85 events. Contemporary methodological research for validating time-to-event prediction models recommends sample-size planning based on precision of calibration, discrimination, and clinical utility rather than simple rules of thumb.[20] The available event count limits precise assessment, particularly of calibration. We did not estimate calibration-in-the-large, calibration slope, time-specific Brier score, or decision-curve net benefit because the cross-platform score was relative and no locked absolute-risk model with baseline survival was specified. These omissions mean the analysis evaluates prognostic association and ranking, not full clinical prediction.

TME estimates were derived from bulk expression. ESTIMATE provides broad admixture signals, not absolute fractions. The candidate genes did not overlap the marker sets, and partial correlations controlled for general stromal score; nevertheless, all quantities came from the same expression matrix and remain mathematically and biologically correlated. Cytotoxic markers can be shared across T and NK cells. No orthogonal deconvolution, histopathological stromal quantification, spatial transcriptomics, or immunohistochemistry was available.

No laboratory validation was conducted. Gene-expression differences require confirmation by quantitative PCR, RNA in situ hybridization, or protein-level assays in independent tissues. Functional experiments are needed to test whether *CTSC*, *FSCN1*, *MYO1B*, *PLAU*, or *SEMA3C* causally mediate stromal interaction, invasion, or treatment response. *COBL* and *MYL1* require particular scrutiny because their tumor–normal and within-tumor prognostic directions differ.

Finally, the study did not establish clinical utility. A biomarker must improve decisions at relevant thresholds, demonstrate assay reproducibility, and add value beyond readily available clinical variables. The held-out result argues against immediate clinical translation.

Fairness and subgroup performance were not comprehensively evaluated. Public metadata encoded sex in binary administrative categories and provided limited race, ethnicity, socioeconomic, geographic, and access-to-care information. Site and HPV analyses revealed substantial case-mix dependence, but small subgroup event counts precluded reliable interaction testing. The score should not be assumed to perform equally across demographic or clinical groups.

4.8 Implications for future research and clinical translation

The next stage should prioritize biological validation of the ECM–CAF program rather than further optimization of this seven-gene score. Any renewed prediction effort should begin with an explicit target population, prediction time, clinical use, and assay. Site- and HPV-specific modeling requires prespecified interactions and adequate event counts; treatment context should distinguish surgery, definitive chemoradiation, and systemic-therapy cohorts.

Future model development should repeat every preprocessing and selection step within resampling and preserve an untouched, sufficiently sized external cohort. Multiple imputation should be considered when clinically important covariates have meaningful missingness and plausible assumptions. A locked model should report absolute-risk predictions, calibration slope, calibration-in-the-large, time-dependent Brier scores, and decision curves only for a credible clinical decision. External evaluation size should be planned from precision targets for calibration, discrimination, and net benefit.[20]

Biologically, the ECM–CAF signal deserves independent investigation even if the seven-gene prognostic score is abandoned. Digital pathology could quantify stromal area and invasive-front architecture. Multiplex immunohistochemistry or immunofluorescence could measure CAF subtypes, macrophages, endothelial cells, and spatial relationships with CD8 T cells. Single-cell or spatial transcriptomics could determine whether the risk-associated genes are expressed by malignant cells, fibroblasts, myeloid cells, or multiple compartments. Paired primary and nodal-metastasis samples would be particularly informative. Orthogonal circulating-biomarker work in HNSCC further illustrates the value of evaluating prognosis outside a single bulk-expression platform.[23] Recent studies of tertiary lymphoid structures and malignant epithelial-cell states also demonstrate how cell-resolved approaches can reveal prognostic biology that is obscured in a small bulk-expression score.[24,25]

For translational assay development, gene measurements must be standardized. A prospective protocol would specify tissue handling, tumor-content thresholds, RNA extraction, platform, normalization, quality controls, and a locked scoring algorithm. If a qPCR panel is pursued, primer specificity, reference genes, analytical precision, and inter-laboratory reproducibility must be established before clinical validation. A protein-based alternative might be more feasible if immunohistochemical expression maps consistently to the transcriptomic signal.

The present results also support a broader lesson for computational reproducibility in biomedical research. Large-scale analytical pipelines can accelerate repository access, data harmonization, statistical modeling, visualization, and documentation. However, the ease of exploring many analytical paths can amplify researcher degrees of freedom. Robust analytical workflows should preserve provenance, record failed analyses, separate training from evaluation, and elevate negative validation. Scientific value lies not in producing a significant signature, but in producing a reliable account of what the data support.

5. Conclusions

Cross-platform replication identified 151 HNSCC tumor–normal DEGs converging on ECM organization, integrin signaling, focal adhesion, and related stromal biology. A seven-gene score was associated with survival only in development data and tracked CAF/fibroblast and stromal expression. It failed in held-out TCGA, showed C-index near 0.56 under nested resampling, and had no independent association in GSE65858 after HPV adjustment.

Accordingly, the seven-gene combination is neither a validated prognostic model nor a basis for clinical decisions. The reproducible contribution is the ECM–CAF stromal program and a transparent negative evaluation showing how site, HPV, platform, and

selection instability limit pan-HNSCC scores. Spatial, single-cell, pathology, and laboratory studies should localize and test this biology before any renewed clinical prediction program.

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified data from TCGA and GEO. No new human participants were recruited and no direct patient identifiers were accessed. Ethics statements and informed-consent procedures for the original cohorts are described by the respective data providers.

Consent for publication

Not applicable to the present secondary analysis.

Availability of data and materials

TCGA-HNSC data are available through the National Cancer Institute Genomic Data Commons and associated UCSC Xena/GDC hubs. GSE6631 and GSE65858 are publicly available from NCBI GEO.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

R.N. and M.Y. conceived and designed the study. R.N., B.A. and P.W. performed data curation, software implementation, formal analysis and investigation. R.N. drafted the original manuscript; B.A., P.W. and M.Y. critically reviewed and edited it. R.N. and B.A. prepared the visualizations. M.Y. supervised and administered the project and is the guarantor. All authors approved the final version.

Acknowledgments

The authors acknowledge the patients and investigators who contributed data to TCGA and GEO, as well as the developers and maintainers of the open-source software used in this study.

6. References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–263. doi:10.3322/caac.21834.
2. Kürten CHL, Kulkarni A, Cillo AR, Santos PM, Roble AK, et al. Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. Nat Commun. 2021;12:7338. doi:10.1038/s41467-021-27619-4.
3. Vos JL, Elbers JBW, Krijgsman O, Traets JJH, Qiao X, et al. Neoadjuvant immunotherapy with nivolumab and ipilimumab induces major pathological responses in patients with head and neck squamous cell carcinoma. Nat Commun. 2021;12:7348. doi:10.1038/s41467-021-26472-9.
4. Zhang Q, Wang Y, Xia C, Ding L, Pu Y, et al. Integrated analysis of single-cell RNA-seq and bulk RNA-seq reveals distinct cancer-associated fibroblasts in head and neck squamous cell carcinoma. Ann Transl Med. 2021;9:1017. doi:10.21037/atm-21-2767.
5. Ran QC, Long SR, Ye Y, Xie C, XuXiao ZL, et al. Mining TCGA database for prognostic genes in head and neck squamous cell carcinoma microenvironment. J Dent Sci. 2021;16:661–667. doi:10.1016/j.jds.2020.09.017.
6. Luo Y, Xu WB, Ma B, Wang Y. Novel stemness-related gene signature predicting prognosis and indicating a different immune microenvironment in HNSCC. Front Genet. 2022;13:822115. doi:10.3389/fgene.2022.822115.
7. Shen Y, Li L, Lu Y, Zhang M, Huang X, et al. Establishment and validation of a comprehensive prognostic model for patients with HNSCC metastasis. Front Genet. 2021;12:685104. doi:10.3389/fgene.2021.685104.
8. Chen Y, Nie J, Li X, Fan T, Deng X, et al. Identification of immune-related prognostic biomarkers associated with HPV-positive head and neck squamous cell carcinoma. J Immunol Res. 2021;2021:6661625. doi:10.1155/2021/6661625.
9. Xu C, Xu H, Liu B. Head and neck squamous cell carcinoma-specific prognostic signature and drug sensitive subtypes based on programmed cell death-related genes. PeerJ. 2023;11:e16364. doi:10.7717/peerj.16364.
10. Wang L, Zhang Y, Li H, Peng J, Gao C, et al. Identification of an immune-related signature as a prognostic classifier for patients with early-stage head and neck squamous cell carcinoma. Transl Cancer Res. 2024;13:1367–1381. doi:10.21037/tcr-23-1791.
11. Chen C, Huang F, Li X, Liu L, Zhang J, et al. Identification of splicing factors signature predicting prognosis risk and the mechanistic roles of novel oncogenes in HNSCC. Biochim Biophys Acta Mol Basis Dis. 2024;1870:167115. doi:10.1016/j.bbadis.2024.167115.
12. Wang A, Xiao N, Wang H, Yao Q, Li J, et al. Development of a novel senescence-related gene signature to predict clinical outcomes, immune landscape, and chemotherapeutic sensitivity in oral squamous cell carcinoma. Head Neck. 2024;46:1112–1125. doi:10.1002/hed.27698.
13. Huang X, Yang J, Wang Q, Fu R, Wen X, et al. Disulfidptosis in head and neck squamous carcinoma: integrative bioinformatic and in-vitro analysis. Oral Dis. 2024;30:4993–5006. doi:10.1111/odi.14977.
14. Selvan AJA, Kannan B, Pandi C, Jayaseelan VP, Arumugam P. EXT2: a novel prognostic and predictive biomarker for head and neck squamous cell carcinoma. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;137:282–289. doi:10.1016/j.oooo.2023.11.011.
15. Kim SI, Woo SR, Noh JK, Lee MK, Lee YC, et al. Association between cancer stem-cell gene-expression signatures and prognosis in head and neck squamous cell carcinoma. BMC Cancer. 2022;22:1077. doi:10.1186/s12885-022-10184-4.
16. Xu T, Xu M, Xu Y, Cai X, Brenner MJ, et al. Developing and validating the model of tumor-infiltrating immune cell

to predict survival in patients receiving radiation therapy for head and neck squamous cell carcinoma. Transl Cancer Res. 2024;13:394–412.
doi:10.21037/tcr-23-2345.

17. Meng L, He X, Hong Q, Qiao B, Zhang X, et al. CCR4, CCR8, and P2RY14 as prognostic factors in head and neck squamous cell carcinoma are involved in the remodeling of the tumor microenvironment. Front Oncol. 2021;11:618187.
doi:10.3389/fonc.2021.618187.
18. Wichmann G, Rosolowski M, Krohn K, Kreuz M, Boehm A, et al. The role of HPV RNA transcription, immune response-related gene expression and disruptive *TP53* mutations in diagnostic and prognostic profiling of head and neck cancer. Int J Cancer. 2015;137:2846–2857. doi:10.1002/ijc.29649.
19. Kuriakose MA, Chen WT, He ZM, Sikora AG, Zhang P, et al. Selection and validation of differentially expressed genes in head and neck cancer. Cell Mol Life Sci. 2004;61:1372–1383. doi:10.1007/s00018-004-4069-0.
20. Riley RD, Collins GS, Ensor J, Archer L, Booth S, et al. Minimum sample-size calculations for external validation of a clinical prediction model with a time-to-event outcome. Stat Med. 2022;41:1280–1295. doi:10.1002/sim.9275.
21. Li S, Huang XT, Wang MY, Chen DP, Li MY, et al. *FSCN1* promotes radiation resistance in patients with *PIK3CA* gene alteration. Front Oncol. 2021;11:653005.
doi:10.3389/fonc.2021.653005.
22. Dou H, Song C, Wang X, Feng Z, Su Y, et al. Integrated bioinformatics analysis of *SEMA3C* in tongue squamous cell carcinoma using machine-learning strategies. Cancer Cell Int. 2024;24:58. doi:10.1186/s12935-024-03247-y.
23. Zhang X, Weeramange CE, Hughes BGM, Vasani S, Liu ZY, et al. Circulating tumour cells predict recurrences and survival in head and neck squamous cell carcinoma patients. Cell Mol Life Sci. 2024;81:233.
doi:10.1007/s00018-024-05269-1.
24. Ruffin AT, Cillo AR, Tabib T, Liu A, Onkar S, et al. B cell signatures and tertiary lymphoid structures contribute to outcome in head and neck squamous cell carcinoma. Nat Commun. 2021;12:3349. doi:10.1038/s41467-021-23355-x.
25. Xiong M, Hu JJ, Yao ML, Song TT, Zhao L, et al. Single-cell sequencing of head and neck carcinoma: transcriptional landscape and prognostic model based on malignant epithelial cell features. FASEB J. 2024;38:e23354.
doi:10.1096/fj.202301287RR.