Journal of Clinical Personalized Medicine 临床个性化医学, 2026, 5(2), 1-10
Published Online April 2026 in Hans. https://www.hanspub.org/journal/jcpm
https://doi.org/10.12677/jcpm.2026.52094

Hans汉斯

基于TCGA-HNSC放疗队列的lncRNA预后模型构建与内部验证：一项生物信息学研究

刘惟佳

华北理工大学生命科学学院，河北 唐山

收稿日期：2026年2月18日；录用日期：2026年3月11日；发布日期：2026年3月20日

摘　要

目的：头颈部鳞状细胞癌(HNSCC)放疗患者的总体生存(OS)存在显著异质性，急需特异性的分子风险分层工具。本研究旨在TCGA-HNSC队列的放疗(RT = Yes)人群中，筛选预后相关lncRNA并构建验证预后模型，为放疗后的个体化管理提供依据。方法：整合TCGA-HNSC转录组与临床数据，提取放疗患者(RT = Yes)样本。对表达数据进行过滤及标准化预处理后，依次采用单因素Cox和LASSO-Cox回归筛选特征基因，通过多因素Cox构建风险评分(RiskScore)模型。利用Kaplan-Meier曲线、timeROC、C-index、列线图(nomogram)及校准曲线等多维度指标，对模型的区分度与校准性能进行内部验证。结果：最终纳入25例放疗样本(含17例死亡事件)，筛选出2个关键lncRNA进入模型。多因素分析显示，HCFC1-AS1为显著保护因素(HR = 0.33, P = 0.008)，Lnc-ENSG257226呈保护趋势(HR = 0.47)；Bootstrap内部验证证实了模型系数的稳健性。模型C-index高达0.84，且RiskScore被证实为独立于年龄和分期的预后因子。模型能显著区分高、低风险组的OS差异(P = 0.0038)，列线图与校准曲线显示其在预测1年、3年生存率方面具有良好的一致性。结论：本研究在TCGA-HNSC放疗人群中构建了基于两个lncRNA的预后模型并完成了内部验证。该模型可有效用于放疗患者的OS风险分层及个体化预测，具备潜在临床价值，但未来仍需更大样本量及外部队列的进一步验证。

关键词

头颈部鳞状细胞癌，放射治疗，长链非编码RNA，Cox回归，LASSO，预后模型，TCGA

Construction and Internal Validation of a lncRNA Prognostic Model Based on a TCGA-HNSC Radiotherapy Cohort: A Bioinformatics Study

Weijia Liu

文章引用: 刘惟佳. 基于 TCGA-HNSC 放疗队列的 lncRNA 预后模型构建与内部验证: 一项生物信息学研究[J]. 临床个性化医学, 2026, 5(2): 1-10. DOI: 10.12677/jcpm.2026.52094

College of Life Science, North China University of Science and Technology, Tangshan Hebei

Received: February 18, 2026; accepted: March 11, 2026; published: March 20, 2026

Abstract

Objective: There exists significant heterogeneity in overall survival (OS) among patients with head and neck squamous cell carcinoma (HNSCC) undergoing radiotherapy, highlighting an urgent need for specific molecular risk-stratification tools. This study aimed to identify prognosis-associated long non-coding RNAs (lncRNAs) within the radiotherapy-treated (RT = Yes) subcohort of The Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma (TCGA-HNSC) project, and to construct and internally validate a corresponding prognostic model, thereby providing a basis for individualized post-radiotherapy management. Methods: Transcriptomic and clinical data from the TCGA-HNSC cohort were integrated to extract samples from patients who received radiotherapy (RT = Yes). Following filtering and normalization of the expression data, univariate Cox regression and LASSO-Cox regression were sequentially applied to screen feature genes. A multivariate Cox proportional hazards model was subsequently employed to construct a risk score (RiskScore) model. The model's discriminative ability and calibration performance were assessed through multi-dimensional internal validation metrics, including Kaplan-Meier curves, time-dependent receiver operating characteristic (timeROC) analysis, concordance index (C-index), nomogram construction, and calibration curves. Results: A total of 25 radiotherapy-treated samples (including 17 death events) were ultimately included in the analysis, leading to the identification of two key lncRNAs for model construction. Multivariate analysis identified HCFC1-AS1 as a significant protective factor (Hazard Ratio [HR] = 0.33, P = 0.008), while Lnc-ENSG257226 exhibited a protective trend (HR = 0.47). Bootstrap internal validation confirmed the robustness of the model coefficients. The model achieved a high C-index of 0.84, and the RiskScore was validated as a prognostic factor independent of age and disease stage. The model effectively stratified patients into high- and low-risk groups with significantly different OS outcomes (P = 0.0038). The nomogram and accompanying calibration curves demonstrated good agreement between predicted and observed 1-year and 3-year survival probabilities. Conclusion: In the radiotherapy-treated subpopulation of the TCGA-HNSC cohort, this study successfully developed and internally validated a two-lncRNA-based prognostic model. This model can effectively stratify OS risk and facilitate individualized outcome prediction for HNSCC patients receiving radiotherapy, showing potential clinical utility. However, future validation in larger, independent external cohorts is warranted.

Keywords

Head and Neck Squamous Cell Carcinoma, Radiotherapy, Long Non-Coding RNA, Cox Regression, LASSO, Prognostic Model, TCGA

1. 引言

头颈部鳞状细胞癌是一组发生于口腔、咽部与喉部等上消化呼吸道黏膜的常见恶性肿瘤，其治疗策略强调多学科协作与器官功能保留。放射治疗在 HNSCC 的根治性治疗、术后辅助治疗以及局部晚期同步放化疗中均占据关键地位，但临床观察显示，即便接受相似放疗策略，患者的生存结局仍呈现显著差

异：部分患者可长期控制，而部分患者在放疗后仍发生复发、转移或早期死亡。这种“同治异效”的现象提示，仅依赖 TNM 分期、病理分级等传统指标难以充分解释放疗人群的预后异质性，建立更精细的分子风险分层体系具有现实意义[1]。

随着高通量测序技术成熟与 TCGA 等公共数据库开放，研究者得以在大样本真实世界数据中系统筛选与预后相关的分子特征。尤其是 lncRNA 在染色质调控、转录调控、miRNA 海绵效应、DNA 损伤修复与免疫微环境塑形等方面发挥作用，而上述环节与放疗敏感性及放疗后肿瘤控制密切相关[2]。因此，在放疗人群中构建基于 lncRNA 表达的预后模型，既可为临床提供风险分层工具，也可为进一步探索放疗相关生物学机制提供候选靶点[3]。本研究基于 TCGA-HNSC 队列，聚焦 RT = Yes 患者亚队列，采用“单因素 Cox 筛选–LASSO-Cox 降维–多因素 Cox 建模”的统计学习流程构建最简预后模型，并结合 KM、timeROC、列线图与校准曲线进行内部验证，以期形成可复现、可解释且具有潜在应用价值的放疗人群预后预测框架[4]。

2. 资料与方法

2.1. 数据来源与下载获取

本研究数据来自 TCGA-HNSC 队列，并通过 UCSC Xena 数据门户获取，数据类型包括 RNA-seq 表达矩阵、生存随访信息及临床/表型信息。表达矩阵采用 R 包 UCSCXenaTools 在 XenaData 元数据中定位“Head and Neck”相关 cohort，并进一步匹配包含“hnsc”与“htseq/counts”特征的数据集，使用 XenaGenerate - XenaQuery - XenaDownload - XenaPrepare 流程自动下载并读取，保证表达数据来源透明且可追溯。考虑到服务器连接波动可能影响自动化流程，本研究的生存与临床表型数据采用“从 UCSC Xena 门户手动下载至本地文件夹后再读取”的方式，脚本中通过匹配文件名关键词(如 survival、phenotype 或 clinical)自动识别本地文件并以 data.table::fread 读取。上述“双通道”策略既保持数据来源一致(均来自 UCSC Xena/TCGA)，又提升了流程稳定性，符合研究生课题与中文核心期刊对可重复性的要求[5]。

2.2. 数据预处理与整合

表达矩阵读取后首先统一样本条形码格式，将“.”替换为“-”，并截取前 12 位作为患者层面的 patient_id 以与临床随访对齐。随后依据 TCGA 条形码第 14~15 位筛选原发肿瘤样本(01)，以减少正常组织或非目标样本混入带来的偏倚。对于同一患者存在多个肿瘤样本的情况，将计数矩阵按 patient_id 聚合并取均值，获得病人层面表达矩阵。生存数据方面，脚本对不同命名规范进行兼容：优先使用“OS”作为结局事件列，若不存在则尝试“vital_status”或“fustat/event”等字段，并将死亡事件统一编码为 status = 1、存活为 status = 0；同时从时间相关列中提取 OS 时间并统一为数值型 time。放疗变量从临床/表型表中通过关键词检索(radiat/radioth/radiation/rt)定位候选列后标准化为 RT = Yes/No。最终以 patient_id 为关键将表达矩阵、生存信息与放疗变量进行内连接合并，剔除 time 缺失或≤0、status 缺失及 RT 缺失样本，形成最终分析集 df_all。

建模队列与统计分析流程

本研究采用在 RT = Yes 人群中构建 OS 预后模型。首先从 df_all 筛选 RT = Yes 样本得到 rt_yes 队列，并提取其基因表达矩阵。为适配 Cox 与 LASSO 建模并提升稳健性，对表达矩阵执行低表达过滤(保留在至少 50%样本中表达 > 0 的基因)，随后进行 log2 (x + 1)转换以减弱计数分布偏态，并对每个基因进行 Z-score 标准化以消除量纲差异。接着对每个基因进行单因素 Cox 回归，筛选 $P < 0.05$ 的候选基因进入 LASSO-Cox。LASSO 部分采用交叉验证选择最优惩罚参数 λ，在 lambda.min 对应的非零系数集合中确定特征 lncRNA；当特征数量过多时使用更保守的 lambda.1se。最终以筛选得到的特征 lncRNA 构建多因素

Cox 模型，并以线性预测值(lp)计算 RiskScore，按中位数将患者划分为高风险组与低风险组[6]。模型性能通过 KM 曲线与 log-rank 检验评估风险分层效果，通过 timeROC 评估 1/3/5 年判别能力，并报告 C-index 作为整体一致性指标；同时绘制风险评分分布与生存状态关联图、模型基因表达热图展示分层可解释性，并构建列线图用于个体化预测 1 年与 3 年生存概率，利用 Bootstrap 校准曲线评估预测一致性；此外，采用 Bootstrap 重抽样(1000 次)评估模型系数稳定性，并通过多因素 Cox 回归分析评估 RiskScore 的独立预后价值。统计检验均为双侧检验，以 $P < 0.05$ 为差异有统计学意义[7]。

3. 研究方法

3.1. RT = Yes 队列特征与建模概况

在完成表达、生存与放疗信息整合后，本研究在 RT = Yes 亚队列中开展建模分析。根据最终模型评估图示，RT = Yes 队列纳入 25 例，死亡事件数为 17 例。在该队列中，通过单因素 Cox 初筛与 LASSO 降维确定特征 lncRNA，并建立多因素 Cox 模型计算 RiskScore，实现放疗人群的预后风险分层与内部验证[8]。

3.2. LASSO-Cox 筛选特征 lncRNA

图 1 展示了候选基因进入 LASSO-Cox 后的交叉验证与系数收缩过程。随着惩罚参数 λ 增大，回归系数逐步被压缩，部分特征系数收缩至 0 从而被剔除，体现了 LASSO 在高维数据中控制复杂度与减少过拟合的优势。交叉验证用于在拟合误差与模型稀疏性之间取得平衡，最终在最优 λ 下保留非零系数的特征集合。本研究最终筛得 2 条 lncRNA 进入后续多因素 Cox 模型，形成最简结构的放疗人群预后模型，为临床解释与后续外部验证提供便利[9]。

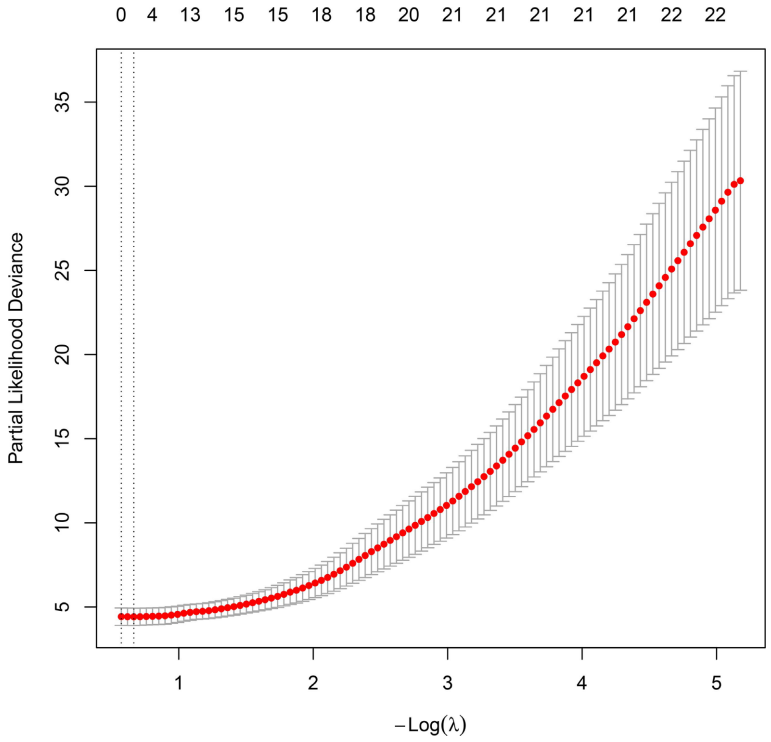


Figure 1. Selection of feature lncRNAs using LASSO-Cox regression
图 1. LASSO-Cox 筛选特征 lncRNA

3.3. 多因素 Cox 模型效应量与模型一致性

图 2 给出了最终多因素 Cox 模型的森林图及核心统计量。模型包含 HCFC1-AS1 与 Lnc-ENSG257226 两条 lncRNA，其中 HCFC1-AS1 表现为显著保护因素(HR = 0.33, 95% CI 0.15~0.75, P = 0.008)，提示其表达升高与死亡风险降低相关；Lnc-ENSG257226 同样呈保护方向(HR = 0.47)，但 P 值为 0.055 接近显著阈值且置信区间上界略超过 1，提示该效应可能受样本量限制影响，需要更大样本进一步检验。图 04 同时报告模型 C-index 为 0.84，表明在 RT = Yes 队列中模型具有较强的一致性与区分能力；AIC 等指标亦提示模型复杂度较低且拟合合理。针对 Lnc-ENSG257226 在原始模型中 P 值(0.055)边缘显著的问题，为进一步验证其在小样本中的可靠性，本研究采用了 Bootstrap 方法进行 1000 次重抽样验证。图 3 显示，HCFC1-AS1 与 Lnc-ENSG257226 的回归系数在重抽样分布中均呈现稳健的保护效应(Robust, Stable < 0)，且其 95%置信区间均未跨越 0 点。这表明尽管受样本量限制 P 值略高，但该 lncRNA 作为保护因素的稳定性良好，故保留在模型中。

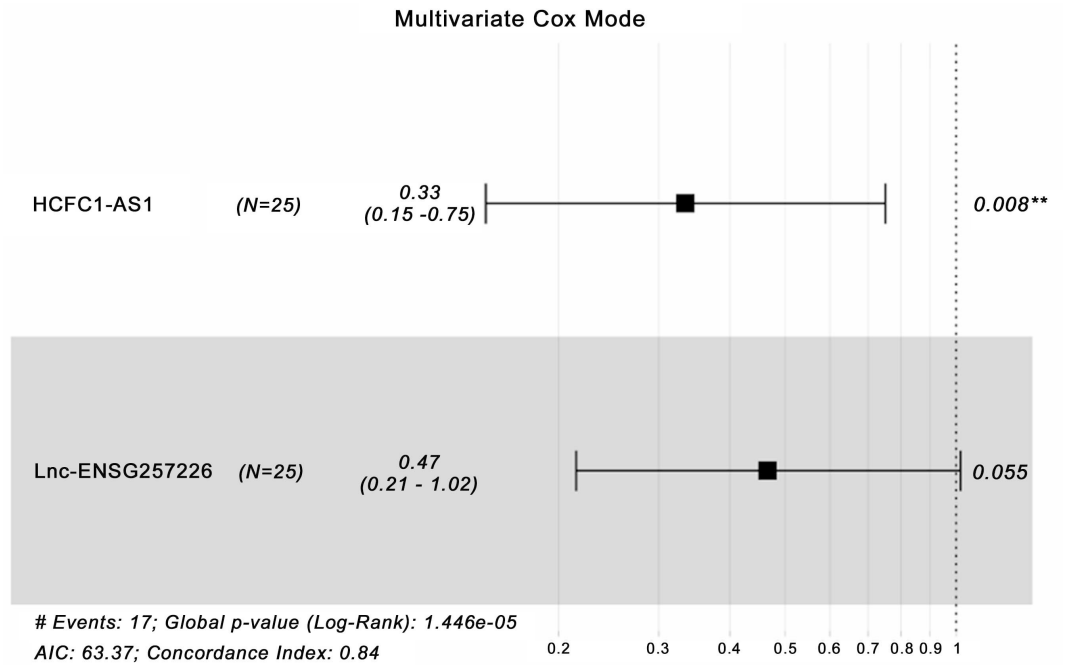


Figure 2. Forest plot and key statistics of the multivariate Cox model
图 2. 多因素 Cox 模型的森林图及核心统计量

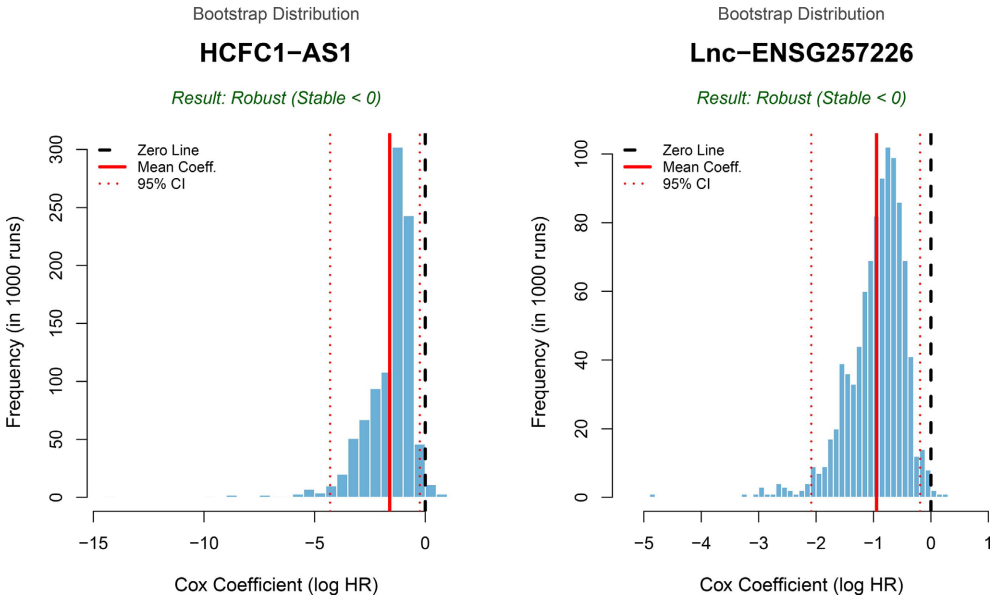
3.4. RiskScore 风险分层与 Kaplan-Meier 生存差异

以 RiskScore 中位数为阈值将 RT = Yes 患者分为高风险组与低风险组后，KM 曲线显示两组 OS 曲线明显分离，高风险组生存概率下降更快，而低风险组总体生存更优；log-rank 检验 P = 0.0038，提示 RiskScore 可在放疗人群中有效区分预后显著不同的亚群。该结果从生存结局角度证明了模型的风险分层价值，并为后续个体化随访与治疗强化策略提供统计学依据[10] (图 4)。

3.5. RiskScore 的独立预后价值评估

为评估 RiskScore 是否独立于临床特征预测患者预后，我们将 RiskScore 与患者年龄(Age)、肿瘤分期(Stage)一同纳入多因素 Cox 比例风险回归分析。如图 5 所示，在校正了年龄与分期后，RiskScore 依然显

示出极显著的预后关联(HR = 9.0, 95% CI 2.42~33.3, P = 0.001)，证明其能够提供超越传统临床指标的额外预后信息。而年龄(P = 0.051)与分期在该小样本队列中未显示出独立统计学显著性。



注：红色实线代表平均系数，红色虚线代表 95%置信区间。两基因的 95% CI 均位于 0 线左侧，提示结果稳健。

Figure 3. Bootstrap validation of the stability of model gene coefficients (1000 resamples)
图 3. 模型基因的 Bootstrap 系数稳定性验证(1000 次重抽样)

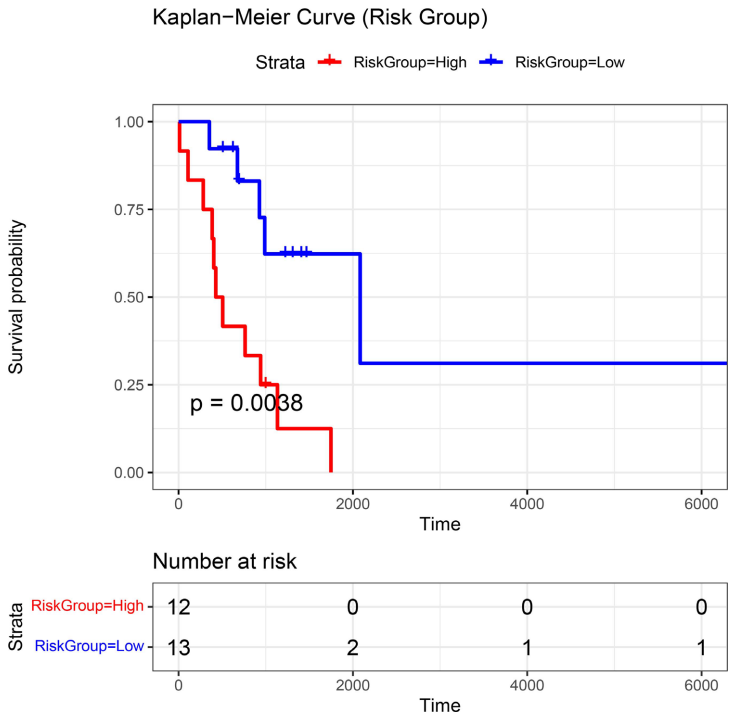
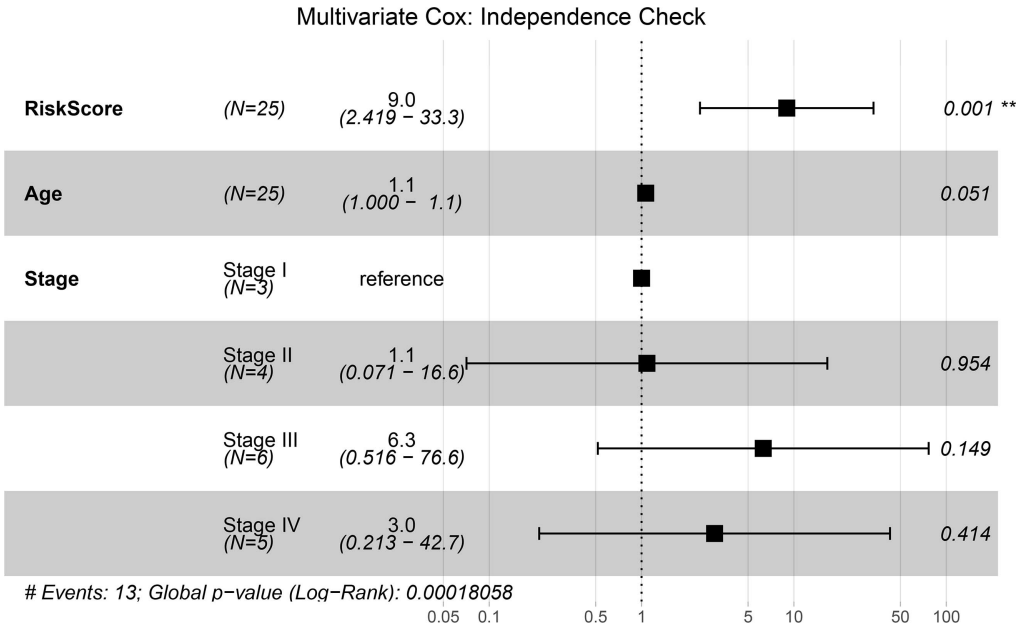


Figure 4. Kaplan-Meier survival difference plot
图 4. Kaplan-Meier 生存差异图



注：RiskScore 在校正年龄和分期后 $P < 0.05$，证实其为独立预后因子。

Figure 5. Forest plot of multivariate Cox regression analysis incorporating clinical features
图 5. 结合临床特征的多因素 Cox 回归分析森林图

3.6. 时间依赖 ROC 评估模型在不同时间点的判别能力

图 6 为 timeROC 曲线，分别评估 RiskScore 对 1 年、3 年与 5 年结局的判别表现。图中不同时间点的 ROC 曲线整体显示出一定的区分能力，提示模型不仅可用于整体风险分层，也具有在特定时间窗口内预测生存结局的潜力[11]。

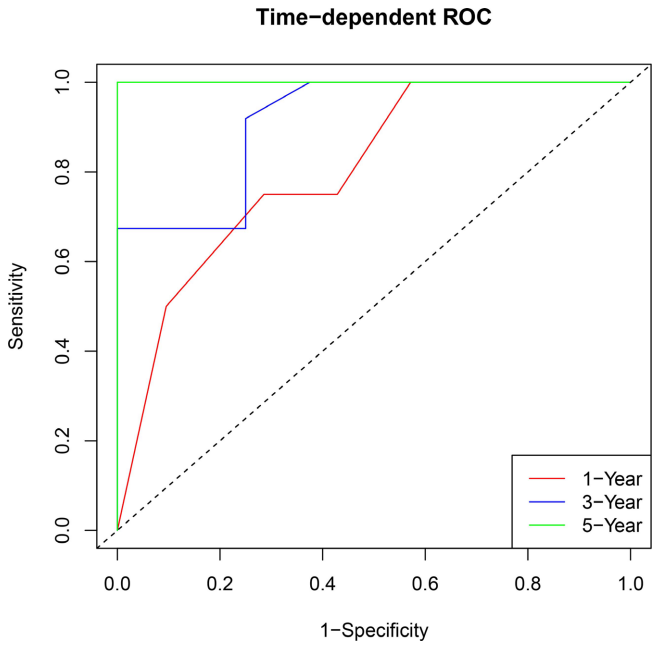


Figure 6. ROC curve
图 6. ROC 曲线图

3.7. 风险三联图与表达模式支持模型可解释性

图 7 从 RiskScore 分布、生存状态与基因表达模式三个层面提供了模型分层的直观证据。图上方显示样本按 RiskScore 从低到高排列后，高风险组集中于高分区间而低风险组集中于低分区间，表明分组阈值能够形成清晰分层；中间散点图以生存时间为纵轴、样本顺序为横轴，死亡事件在高 RiskScore 区域更为集中，且部分死亡样本生存时间较短，提示高风险组确实对应较差结局；下方热图展示模型基因在样本排序下的表达变化，风险组注释条件与表达模式呈一定一致性，使得 RiskScore 不仅是统计拟合的产物，也具有表达层面的模式支撑[12]。

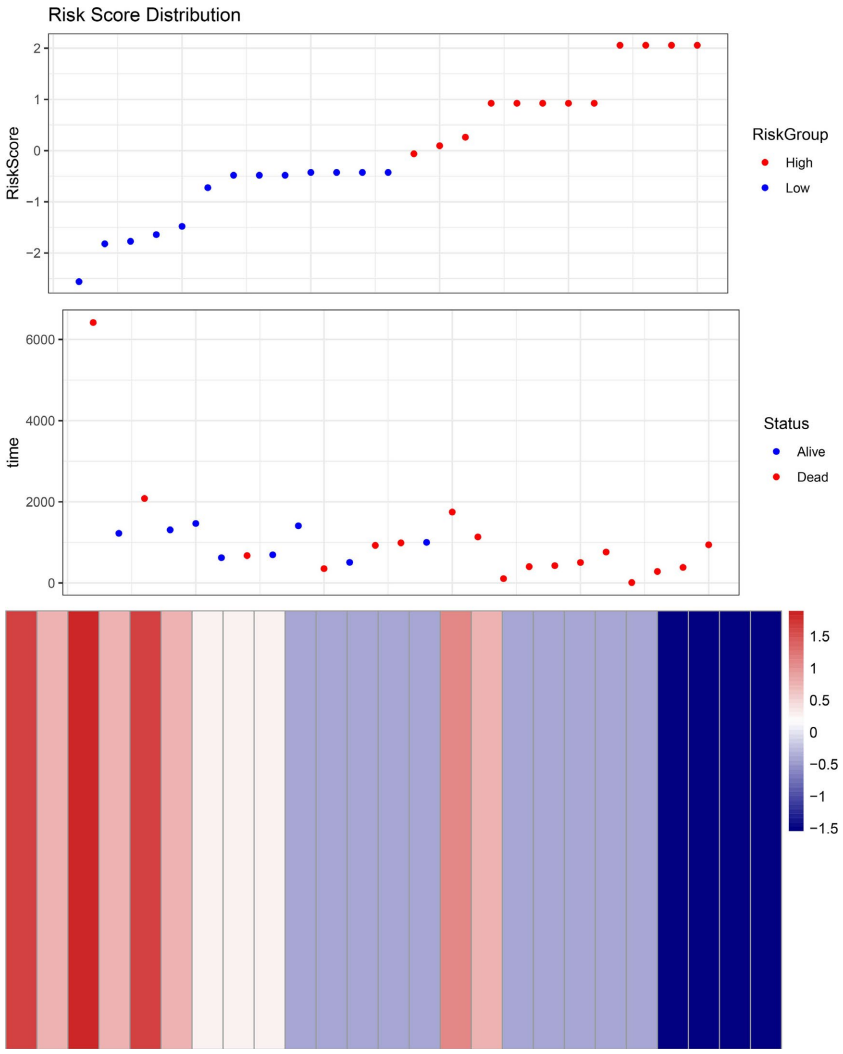


Figure 7. Risk triplet plot
图 7. 风险三联图

3.8. 列线图与校准曲线体现模型的临床工具化潜力

图 8 为基于 Cox 模型构建的列线图，将 RiskScore 映射为个体化的 1 年与 3 年生存概率预测。使用时可在 RiskScore 轴定位患者分值，投影至积分轴累积总分后读取对应时间点的生存概率，从而将模型从“统计意义”转化为“可操作的临床预测工具”[13]。

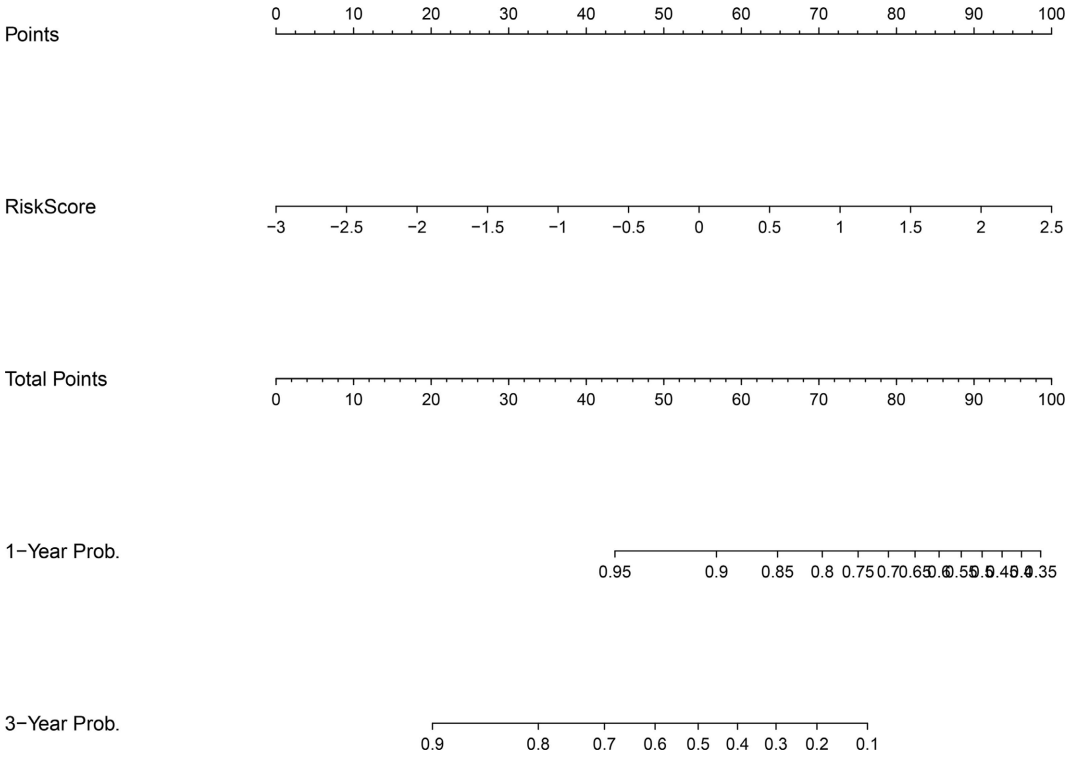


Figure 8. Nomogram predicting 1~3-year survival probability.
图 8. 列线图预测 1~3 年的生存概率

4. 讨论

本研究针对 TCGA-HNSC 放疗患者构建并内部验证了两 lncRNA 预后模型，在 RT = Yes 队列中获得较为一致的证据链：LASSO 实现特征降维并保留最简模型结构，多因素 Cox 提示 HCFC1-AS1 具有显著保护作用且模型 C-index 达到 0.84，RiskScore 可显著区分 OS 差异，并在 timeROC、风险三联图与热图中得到判别与可解释性支持。特别是，针对小样本研究常见的稳健性顾虑，本研究通过 Bootstrap 重抽样技术与多因素校正分析，进一步夯实了模型的统计学基础。从中文核心稿件的“完整性”角度看，本研究不仅给出了统计建模结果，还以列线图与校准曲线将模型工具化呈现，符合临床转化型生信研究[14]。

从生物学意义上看，lncRNA 可能通过多种机制影响放疗后肿瘤控制，例如调节 DNA 损伤修复通路、影响细胞周期检查点、改变肿瘤免疫微环境，进而影响复发与生存。HCFC1-AS1 在本研究中呈显著保护效应，提示其可能参与增强放疗敏感性或抑制放疗后肿瘤进展的过程。值得注意的是，Lnc-ENSG257226 在原始 Cox 模型中 P 值为 0.055 (边缘显著)，但这可能受限于样本量(N = 25)导致的统计效能不足。本研究进一步的 Bootstrap 内部验证显示，该基因在 1000 次重抽样中的回归系数 95%置信区间完全位于 0 点左侧，未出现跨越 0 点的情况。这一结果有力地证明了 Lnc-ENSG257226 并非随机噪声，而是在放疗人群中具有稳定的保护性预后价值。鉴于 TCGA 临床表型中放疗信息的记录形式可能存在差异，未来研究可结合功能富集、免疫浸润估计、DNA 修复评分等分析深化机制解释，以提升模型的生物学可解释性与临床信任度。

本研究亦存在需在投稿中主动说明的局限，但我们在现有数据条件下进行了严格的补充验证以最大程度弥补这些不足。首先，RT = Yes 样本量为 25 例且事件 17 例，总体规模偏小。虽然小样本通常面临过拟合风险，但本研究通过 LASSO 正则化降维以及严格的 1000 次 Bootstrap 内部验证，证实了模型系数

的稳定性，表明模型并未因样本量少而出现严重的估计偏差。其次，关于外部验证，我们系统检索了 GEO 等公共数据库，但受限于现有 HNSCC 公开数据集中缺乏详细的放疗信息或完整的生存随访数据，目前无法找到适用的独立队列进行外部验证。为弥补这一缺憾，我们补充进行了结合临床特征的独立预后分析，结果显示 RiskScore 在校正了年龄与分期后，P 值仍达到 0.001，HR 高达 8.97。这一极显著的统计学差异表明，即便在缺乏外部验证的情况下，该模型在当前队列中对高危患者的识别能力也是强效且独立于传统分期的。再次，本研究未进一步区分同步放化疗、放疗剂量分割、HPV 状态等更精细的临床因素。基于上述原因，本模型目前定位为放疗人群预后分层的候选工具，在临床大规模推广前，仍期待未来开展基于大样本、多中心的各种族前瞻性研究进行验证。

5. 结论

基于 TCGA-HNSC 数据，本研究在放疗(RT = Yes)患者中构建了由 HCFC1-AS1 与 Lnc-ENSG257226 组成的两 lncRNA 预后模型。该模型可通过 RiskScore 显著区分高低风险患者 OS 差异，并表现出较好的内部判别能力与一致性；同时，列线图与校准曲线为个体化预测提供了可视化工具支持。尽管如此，受样本量与临床信息粒度限制，模型仍需在更大规模、具有明确放疗信息的外部队列中进一步验证，以评估其泛化能力与临床增益，并探索相关 lncRNA 在放疗敏感性中的潜在机制。

参考文献

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., *et al*. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[2] Johnson, D.E., Burtness, B., Leemans, C.R., Lui, V.W.Y., Bauman, J.E. and Grandis, J.R. (2020) Head and Neck Squamous Cell Carcinoma. *Nature Reviews Disease Primers*, **6**, Article No. 92. https://doi.org/10.1038/s41572-020-00224-3

[3] Alterio, D., Marvaso, G., Ferrari, A., Volpe, S., Orecchia, R. and Jereczek-Fossa, B.A. (2019) Modern Radiotherapy for Head and Neck Cancer. *Seminars in Oncology*, **46**, 233-245. https://doi.org/10.1053/j.seminoncol.2019.07.002

[4] Gregoire, V., Leftych, K. and White, J. (2020) Radiotherapy for Head and Neck Cancer: The Current State of the Art. *Journal of Clinical Oncology*, **38**, 54-63.

[5] Schmitt, A.M. and Chang, H.Y. (2016) Long Noncoding RNAs in Cancer Pathways. *Cancer Cell*, **29**, 452-463. https://doi.org/10.1016/j.ccell.2016.03.010

[6] Ang, H., Wu, Z., Zhang, J. and Su, B. (2013) Salivary LncRNAs as Potential Biomarkers for the Diagnosis and Prognosis of Head and Neck Squamous Cell Carcinoma. *Molecular Cancer*, **12**, 1-9.

[7] Gao, Y., Wang, P., Wang, D. and Wang, H. (2020) Long Non-Coding RNA-Based Signature for Prognosis Prediction in Head and Neck Squamous Cell Carcinoma. *Scientific Reports*, **10**, 1-10.

[8] Luo, H., Xu, C., Chen, X., Stubblefield, B., Chang, H.Y. and Wang, J. (2020) LncRNA Mechanistic Linkages to DNA Damage Response and Genomic Stability in Cancer. *Journal of Hematology & Oncology*, **13**, Article 159.

[9] Cancer Genome Atlas Network (2015) Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas. *Nature*, **517**, 576-582. https://doi.org/10.1038/nature14129

[10] Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., *et al*. (2020) Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nature Biotechnology*, **38**, 675-678. https://doi.org/10.1038/s41587-020-0546-8

[11] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[12] Blanche, P., Dartigues, J. and Jacqmin-Gadda, H. (2013) Estimating and Comparing Time-Dependent Areas under Receiver Operating Characteristic Curves for Censored Event Times with Competing Risks. *Statistics in Medicine*, **32**, 5381-5397. https://doi.org/10.1002/sim.5958

[13] Iasonos, A., Schrag, D., Raj, G.V. and Panageas, K.S. (2008) How to Build and Interpret a Nomogram for Cancer Prognosis. *Journal of Clinical Oncology*, **26**, 1364-1370. https://doi.org/10.1200/jco.2007.12.9791

[14] Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. CRC.