基于 Lasso-Cox 回归模型的肺腺癌基因学预后风险分析*

山东第二医科大学公共卫生学院(261053)
卜伟晓　穆华夏　高梦瑶　苏维强　韩　梅　陶子琨　杨　希　徐雅琪　石福艳　王清华　王素珍△　孔雨佳△

【提　要】 **目的**　通过构建 Lasso-Cox 模型筛选肺腺癌差异表达基因,计算患者风险评分,构建肺腺癌预测模型,为肺腺癌的研究提供潜在的基因靶点,并为临床诊疗及预后提供新方向。**方法**　下载癌症基因组图谱(TCGA)和肿瘤基因表达数据库(GEO)的肺腺癌基因表达和临床数据,用 TCGA 数据库训练模型,并合并两数据库用以模型验证,筛选的肺腺癌差异表达基因(DEGs)通过多因素 Lasso-Cox 回归构建风险评分预后模型,结合临床资料以确定肺腺癌最终的独立预后预测因素。利用 GO 富集分析、KEGG 通路分析和 CIBERSORTx 免疫分析对风险模型差异表达基因进行生物学解释。**结果**　通过单变量 Cox 和 Lasso-Cox 回归分析,获得了与肺腺癌预后相关的 9 个差异表达基因。结合临床数据的多因素 Cox 回归模型显示,恶性肿瘤病史、N 分期、T 分期和风险评分是预后的独立影响因素。**结论**　本研究构建的肺腺癌预后模型可以有效预测患者的预后风险,为临床决策和个性化治疗提供理论基础。

【关键词】 Lasso-Cox 模型　预后预测　基因表达　肺腺癌
【中图分类号】 R181.23　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.03.006

Genetic Prognostic Risk Analysis of Lung Adenocarcinoma with Lasso-Cox Regression Model

Bu Weixiao, Mu Huaxia, Gao Mengyao, et al (*School of Public Health*, *Shandong Second Medical University* (261053), *Weifang*)

【**Abstract**】 **Objective**　To screen differentially expressed genes in lung adenocarcinoma by constructing Lasso-Cox model to provide potential gene targets for the research of lung adenocarcinoma and new directions for clinical diagnosis, treatment and prognosis by calculating patient risk score and constructing prediction model of lung adenocarcinoma. **Methods**　The gene expression and clinical data of lung adenocarcinoma were downloaded from the Cancer Genome Atlas (TCGA) and Gene Expression Omnibus database (GEO). The TCGA database was used to train model, and the two databases were combined for model validation. The screened differentially expressed genes (DEGs) of lung adenocarcinoma were analyzed by univariate Cox and multivariate Lasso-Cox to construct a risk score prognosis model. Risk score from the final Cox prediction model and clinical data were combined to determine independent prognostic factors. GO enrichment analysis, KEGG pathway analysis and CIBERSORTx immunoassay were used to evaluate the biological interpretation of differentially expressed genes in the risk model. **Results**　The analysis using univarate Cox and Lasso-Cox regreesion identified 9 differentially expressed genes associated with the prognosis of lung adenocarcinoma. Multivariate Cox regression analysis, incorporating clinical data, revealed that a history of malignant tumors, N stage, T stage, and the risk score were independent prognostic factors. **Conclusion**　The prognositic model of lung adenocarcinoma can effectively predict the prognosis risk and provide a theoretical basis for clinical decision-making and personalized treatment.

【**Key words**】 Lasso-Cox model; Prognositic prediction; Gene expression; Lung adenocarcinoma

肺癌作为目前世界公认的发病率和死亡率最高的恶性肿瘤[1-3],非小细胞肺癌占比超过 80%,其中最常见的组织型是肺腺癌(lung adenocarcinoma, LUAD)。多数患者首诊为晚期,因而预后较差。而高通量测序技术的快速进展为从基因表达层面进行 LUAD 的风险评估、早期诊断及肿瘤监测提供了新思路[4]。在基因表达层面的研究中,高维度基因资料的分析过程中往往存在多重共线性问题,使传统 Cox 比例风险模型[5-6]无法准确估计。通过在回归模型中添加正则化约束项[7-8]能够有效处理自变量之间的多重共线性,同时避免过度拟合[9]。因此,本研究基于公开数据库,筛选 LUAD 差异表达的基因,利用 Lasso-Cox 回归[10]构建风险模型获取风险评分,并结合临床数据构建预后预测模型,旨在为 LUAD 的预后评估和个性化预后方案的制定提供科学依据。

材料与方法

1. 研究人群

从 TCGA 下载 LUAD 中基因表达量数据(HT-SeqCounts,$n=585$)、表型数据($n=877$)以及生存数据($n=738$)。删除 TCGA 基因表达量数据中非固体瘤/非癌旁组织病例,得到 524 例癌组织和 59 例癌旁组织基因的 mRNA 表达矩阵。验证数据集为 GEO 中的 GSE31210 数据集。由于测序平台不同等因素影响两数据库 mRNA 测序数据所包含的基因个数并不完全

*基金项目:国家自然科学基金(82003560);山东省自然科学基金(ZR2020MH340);山东省教育厅教改项目(M2021174; M2021327)
△通信作者:孔雨佳,E-mail: yujia_kyj80@163.com;王素珍,E-mail: wangsz@wfmc.edu.cn

相同，为后续模型的验证本研究只选取两数据库均包含的基因表达数据进行模型的构建。

将来自 TCGA 数据库的生存数据整理（剔除 524 位患者中未记录生存信息的个体）后得到 513 名患者生存信息，保留全部癌症样本的基因表达数据用以预后模型的建立。多因素 Cox 模型基于 TCGA 中筛选出的数据完整（即 513 位患者中临床信息均无缺失）的 489 名患者数据构建，来检验构建的基因表达数据预后模型的风险评分能否作为患者预后的独立风险因素。该队列患者基本情况如表 1 所示，中位生存时间为 21.73（13.90，35.70）月。

表 1 肺腺癌患者的临床和生存信息（n=489）[n（%）]

变量名	HT-SeqCounts
性别	
女	265（54.19）
男	224（45.81）
年龄（岁）	
≤50	39（7.98）
50~70	294（60.12）
>70	156（31.9）
双侧肿瘤病史	
无	401（82）
有	88（18）
T 分期	
T1	166（33.95）
T2	258（52.76）
T3	46（9.41）
T4	19（3.89）
M 分期	
M0	320（65.44）
M1	29（5.93）
Mx	140（28.63）
N 分期	
N0	324（66.26）
N1	85（17.38）
N2/N3	67（13.7）
Nx	13（2.66）
吸烟史	
当前吸烟	70（14.31）
从未吸烟	118（24.13）
戒烟>15 年	131（26.79）
戒烟<15 年	166（33.95）
戒烟，戒烟时间不详	4（0.82）
肿瘤类型	
腺泡细胞肿瘤	21（4.29）
腺瘤和腺癌	454（92.84）
囊性、粘液性和浆液性肿瘤	14（2.86）
生存状态	
死亡	172（35.17）
生存	317（64.83）

2. 研究方法

（1）筛选差异表达基因

对数据集的基因表达数据进行预处理标准化并构建比较矩阵，使用 R 语言中的“limma”包来识别 LUAD 癌组织与癌旁组织的差异表达基因（differently expressed genes，DEGs）。选取|log2FC|>2 且调整后的 P<0.05 的基因作为 DEGs。

（2）构建风险预后模型

对 TCGA 队列中 DEGs 经过单因素 Cox 回归分析后有统计学意义的基因进行 Adaptive-Lasso 惩罚的 Cox 回归分析，通过交叉验证选择最优 λ 值以得到基于风险评分的预后模型。

（3）评估预测模型

利用 StepMiner 方法[11]将 LUAD 患者根据风险评分中位数划分为高风险组和低风险组。利用主成分分析（Rtsne 包）、Kaplan-Meier 生存曲线分析（survival 包）和时间依赖的 ROC 曲线分析（survival、survminer 和 timeROC 包）评估风险评分的生存预测能力。TCGA 与 GEO 的合并数据集进行该风险评分模型的预测能力验证。

（4）基于风险模型的富集分析

应用 clusterProfiler 包进行基因本体论（Gene Ontology，GO）和京都基因组基因百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）分析来分析 DEGs 参与免疫球蛋白和免疫应答激活信号转达等的生物学过程。

（5）基于风险模型的免疫功能分析

使用 CIBERSORTx（https://cibersortx.stanford.edu）中的 LM22 免疫细胞基因特征以绝对模式运行，计算样本的 B 细胞幼稚、CD8+T 细胞、巨噬细胞 M0、巨噬细胞 M1、巨噬细胞 M2、活化 NK 细胞、静止 NK 细胞和中性粒细胞等免疫细胞的含量。

（6）构建预后预测模型

将 LUAD 患者的风险评分模型的风险评分与临床信息进行单因素和多因素 Cox 回归，分析 LUAD 患者的预后因素和独立预后因素，用森林图展示，同时绘制预后模型的列线图。

结　果

1. 差异表达基因的筛选结果

从 TCGA 得到的训练集（524 个癌组织和 59 个癌旁组织基因）的 mRNA 表达矩阵分析中，共筛选出 1294 个 mRNA 差异表达基因，其中上调基因 572 个，下调基因 722 个。

2. Lasso-Cox 回归构建风险预后模型

对 1294 个 DEGs 进行单因素 Cox 回归分析，得到 575 个相关基因（P<0.05）。采用 Lasso-Cox 回归模型及交叉优化（图 1），构建包含 9 个基因特征的风险预后模型（λ=0.068）。

风险评分=（0.0157×$Exp_{C1QTNF6}$）+（-0.0852×$Exp_{CYP17A1}$）+（-0.0164×$Exp_{FAM189A2}$）+（-0.0336×Exp_{FCRL2}）

$+(0.0154\times Exp_{IGF2BP1})+(0.0120\times Exp_{KRT6A})+(0.0379\times Exp_{RGS20})+(0.0396\times Exp_{RHOV})+(0.0007\times Exp_{SPRR1B})$。

PCA 分析显示该模型涉及的变量可以将患者很好地区分为两组(图 2)。

3. 风险预后模型的评估

如图 3 所示,根据风险评分将 TCGA 中的 LUAD 患者分为高风险组和低风险组,绘制的风险曲线与 Kaplan-Meier 生存曲线可见随着患者风险系数升高,死亡风险增加,生存时间缩短($P<0.0001$)。1、3 和 5 年 ROC 曲线下面积(AUC)分别为 0.726、0.711 和 0.724。

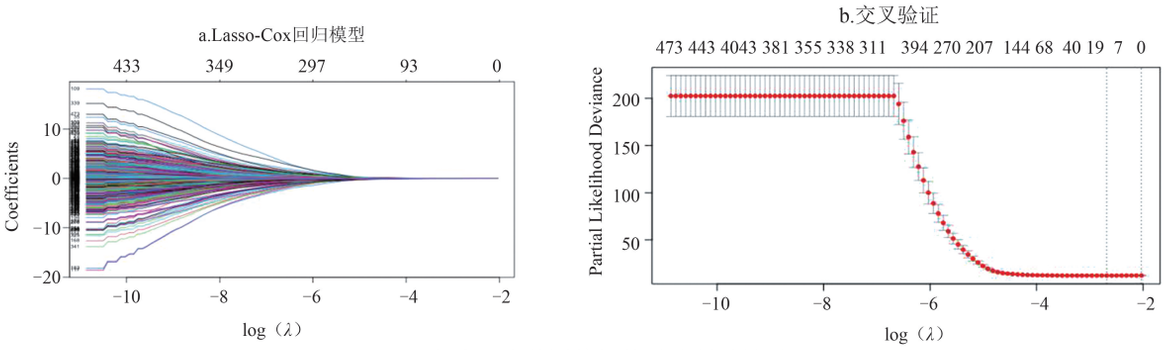

图 1　包含 9 个基因特征的风险预后模型

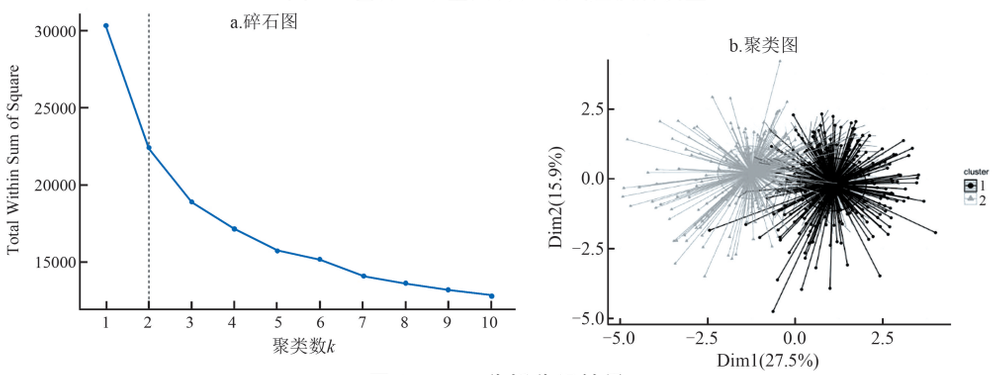

图 2　PCA 分析分组结果

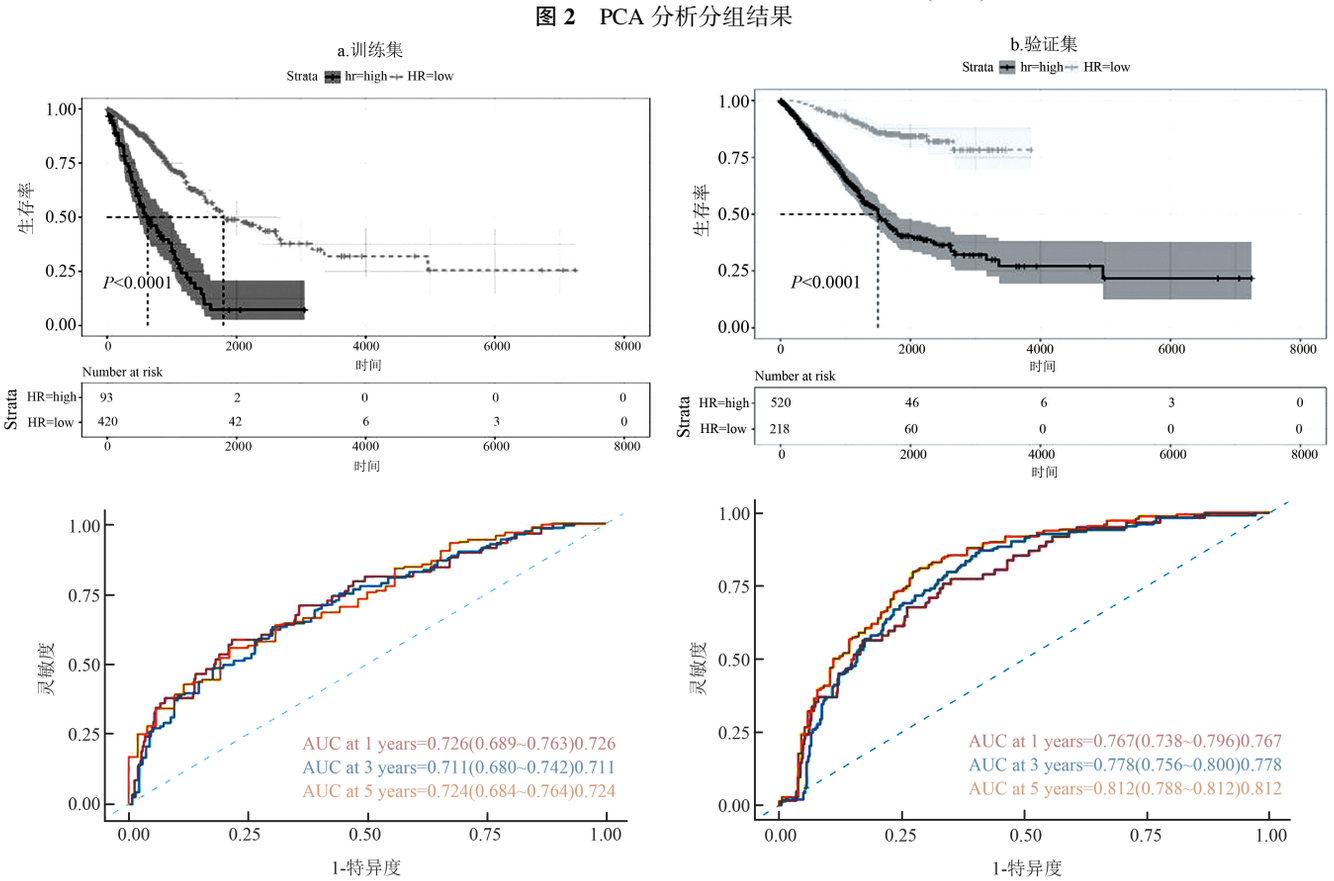

图 3　风险评分模型 Kaplan-Meier 曲线和 ROC 曲线

利用 GSE31210 芯片中的 226 例样本与 TCGA 队列 513 例样本合并构成的数据集（$n=738$，删掉一个离群值）对模型进行验证，预测患者风险评分，采用 StepMiner 法根据风险评分将患者分为高风险组和低风险组。1、3、5 年 AUC 分别为 0.767、0.778 和 0.812。

4. GO 及 KEGG 通路富集分析

对单因素 Cox 分析中，筛选出的预后相关的 DEGs 进行 GO 生物富集分析、KEGG 通路分析。结果表明，这些预后相关基因主要与核分裂过程、细胞器分裂过程、微管蛋白结合和 G 蛋白偶联肽受体活性等过程以及细胞周期、神经活性配体受体相互作用通路相关（图 4）。KEGG 通路分析显示，基因表达在细胞周期相关通路中下调，在神经活性配体受体相互作用通路中上调。

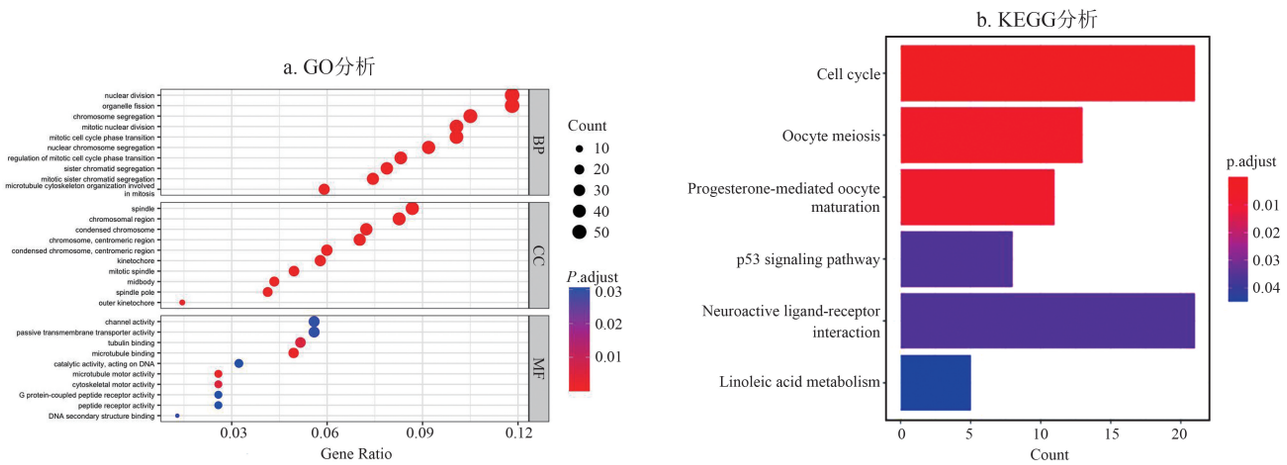

图 4 TCGA 队列中两风险组间差异表达基因的 GO 分析与 KEGG 分析

5. CIBERSORTx 免疫浸润分析

在 TCGA 队列中比较不同风险组间免疫细胞流行率，发现高风险组均具有较低水平的免疫细胞浸润，尤其在树突状细胞静止及肥大细胞静止中，在高风险组中的免疫细胞含量均低于低风险组（图 5）。

6. 构建 LUAD 患者预后预测模型

将 TCGA 队列中 LUAD 患者的临床信息以及风险模型计算的风险评分作为因素进行单因素和多因素 Cox 回归分析，由于 TCGA 数据库中临床数据缺失等问题，选取了年龄、性别、恶性肿瘤疾病史、肿瘤组织类型以及癌症 TMN 分期纳入分析。多因素 Cox 分析发现风险评分可作为独立预后因素（图 6）。

同时多因素 Cox 回归得到恶性肿瘤病史、N 分期、T 分期为患者预后的独立影响因素。列线图可以看出风险评分、T 分期、有肿瘤疾病史、肿瘤组织类型、N 分期及吸烟史对模型预测贡献较大（图 7）。

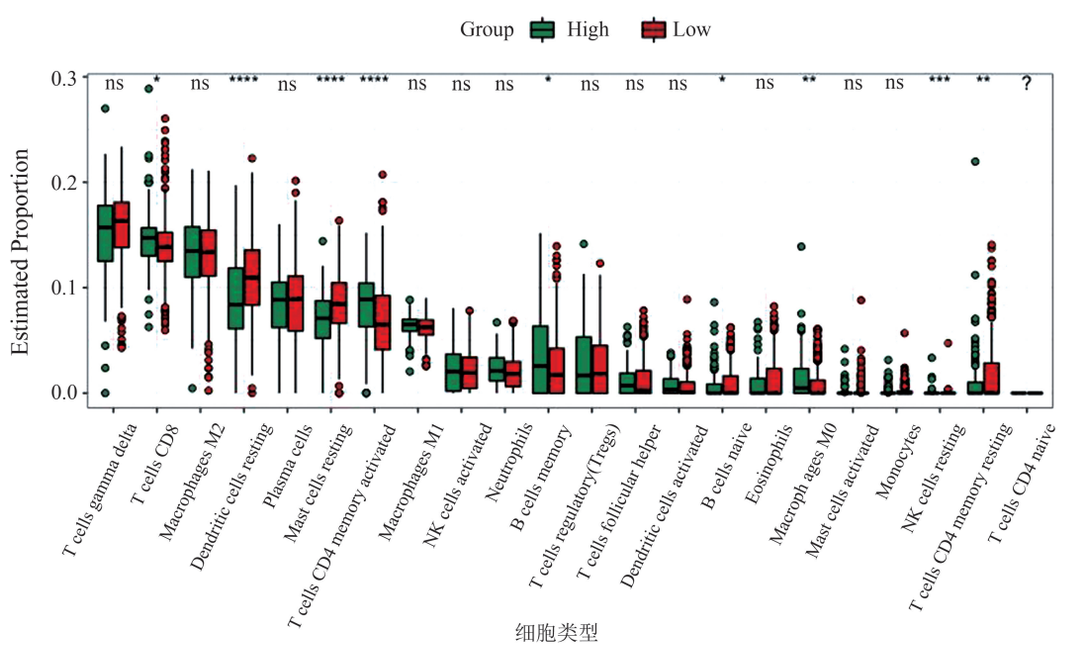

图 5 CIBERSORTx 免疫浸润分析

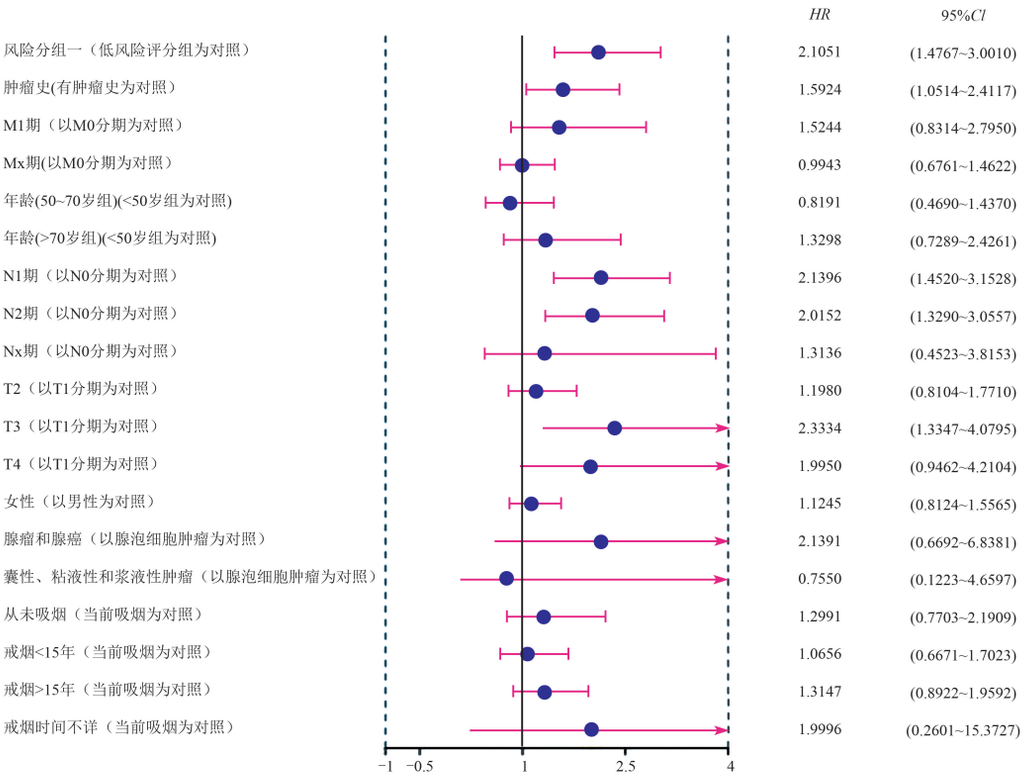

图 6　结合临床信息和风险评分的多因素 Cox 分析

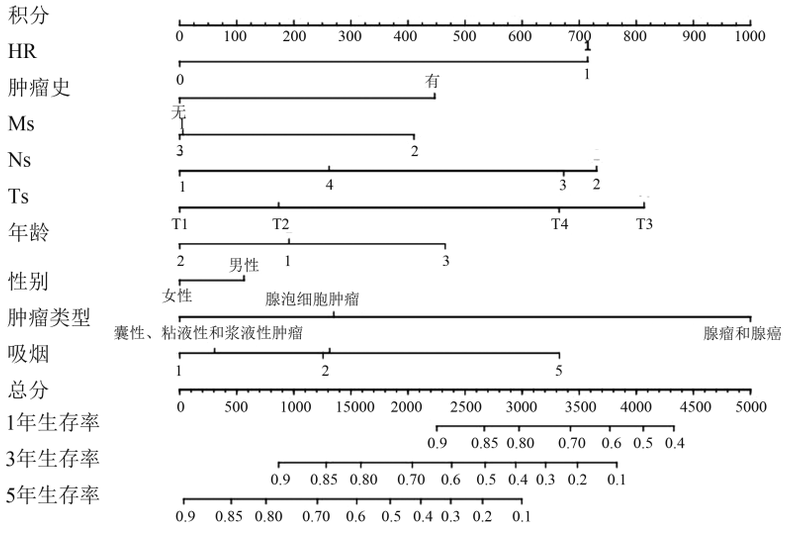

图 7　预后相关列线图

讨　论

本研究基于 TCGA 队列中 LUAD 预后相关 DGEs，采用自适应 Lasso-Cox 回归[12-13]构建了由肿瘤病史、N 分期、T 分期和由 9 个基因组成的风险评分构建的预测模型。GO 功能分析、KEGG 通路富集分析和 CIBERSORTx 免疫浸润分析，显示通过关键基因获得的风险评分可作为 LUAD 的独立预后因素进行风险评估。

补体 C1q 肿瘤坏死因子相关蛋白 6（complement C1q tumor necrosis factor-related protein 6，C1QTNF6）被发现可参与多种炎症过程，在肿瘤发病发展过程中扮演着重要角色，该基因相关蛋白与肿瘤血管生成、细胞增殖及迁移能力等生物功能相关[14]。*C*1QTNF6 低表达可显著抑制细胞增殖、迁移和侵袭能力[15-16]。*IGF2BP*1 被证实可加速细胞周期进程，促进肺癌细胞的持续增殖[17]，因而与肺癌患者的肿瘤发生和进展有关[18]，这与我们的结果一致。*KRT6A* 的高表达可以促进非小细胞的增殖和侵袭[19-20]，对肺癌生存和免疫治疗反应有较好的预测[21]。*SPRR1B* 基因有成为肿瘤

标志物的潜力[22]。*CYP17A1* 基因可以通过调节内质网和微粒体中的两个胆固醇合成睾酮过程中的关键酶的活性,来影响 LUAD 的发生发展过程[23-25]。FCRL 家族的某些成员(*FCRL3*、*FCRL2*、*FCRL5*)在肝癌的侵袭和转移中也有作用[26]。本研究发现这些基因是 LUAD 的重要基因标志物并与预后显著相关,可为 LUAD 的分子治疗提供较有价值的靶点。

本研究也存在一些局限,需在后续的研究中改进。本研究未对 LUAD 进行包含这 9 个关键基因在内的风险模型进行外部验证,在未来的研究中,有待开展相关数据收集和分析以验证模型的泛化性。

综上所述,本研究基于 Lasso-Cox 结合临床数据和组学数据构建了预后预测模型,具有良好的性能。相较于单纯以临床数据为基础的研究,揭示了影响 LUAD 预后的潜在生物标记物,为患者进行预后评估和个性化精准治疗方案的选择奠定了理论基础。

参 考 文 献

[1] 张雨苇,梁民勇. miRNA 在呼吸系统疾病的研究进展.中国医学创新,2021,18(18):180-184.

[2] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin,2021,71(3):209-249.

[3] 杨燕珍,李玉付,何伟,等.细胞焦亡在非小细胞肺癌发生发展和治疗预后中作用及机制的研究进展.现代肿瘤医学,2022,30(23):4396-4401.

[4] Yim J, Zhu LC, Chiriboga L, et al. Histologic features are important prognostic indicators in early stages lung adenocarcinomas. Mod Pathol,2007,20(2):233-241.

[5] 罗文婷. Cox 模型中协变量调整下参数带约束的统计推断.南宁:南宁师范大学,2021.

[6] Grant SW, Hickey GL, Head SJ. Statistical primer: multivariable regression considerations and pitfalls. Eur J Cardiothorac Surg,2019,55(2):179-185.

[7] Li R, Chang C, Justesen JM, et al. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. Biostatistics,2022,23(2):522-540.

[8] Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. Brief Bioinform,2021,22(1):77-87.

[9] 刘丹,郑少智. Cox 模型中的自适应 Lasso 变量选择.统计与决策,2016(10):7-10.

[10] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med,1997,16(4):385-395.

[11] Liu Y, Ji Y, Qiu P. Identification of thresholds for dichotomizing DNA methylation data. EURASIP J Bioinform Syst Biol,2013,2013(1):8.

[12] Ebrahimi V, Sharifi M, Mousavi-Roknabadi RS, et al. Predictive determinants of overall survival among re-infected COVID-19 patients using the elastic-net regularized Cox proportional hazards model: a machine-learning algorithm. BMC Public Health,2022,22(1):10.

[13] Li L, Liu ZP. Detecting prognostic biomarkers of breast cancer by regularized Cox proportional hazards models. J Transl Med,2021,19(1):514.

[14] Takeuchi T, Adachi Y, Nagayama T. Expression of a secretory protein C1qTNF6, a C1qTNF family member, in hepatocellular carcinoma. Anal Cell Pathol(Amst),2011,34(3):113-121.

[15] Rao X, Lu Y. C1QTNF6 Targeted by MiR-184 Regulates the Proliferation, Migration, and Invasion of Lung Adenocarcinoma Cells. Mol Biotechnol,2022,64(11):1279-1287.

[16] Lin G, Lin L, Lin H, et al. C1QTNF6 regulated by miR-29a-3p promotes proliferation and migration in stage I lung adenocarcinoma. BMC Pulm Med, 2022,22(1):285.

[17] Shi R, Yu X, Wang Y, et al. Expression profile, clinical significance, and biological function of insulin-like growth factor 2 messenger RNA-binding proteins in non-small cell lung cancer. Tumour Biol,2017,39(4):1010428317695928.

[18] Shen Q, Xu Z, Sun G, et al. TFAP4 Activates IGF2BP1 and Promotes Progression of Non-Small Cell Lung Cancer by Stabilizing TK1 Expression through m6A Modification. Mol Cancer Res,2022,20(12):1763-1775.

[19] Zhu Q, Zhang C, Qu T, et al. MNX1-AS1 Promotes Phase Separation of IGF2BP1 to Drive c-Myc-Mediated Cell-Cycle Progression and Proliferation in Lung Cancer. Cancer Res,2022,82(23):4340-4358.

[20] Ma C, Li F, He Z, et al. A more novel and powerful prognostic gene signature of lung adenocarcinoma determined from the immune cell infiltration landscape. Front Surg,2022,9:1015263.

[21] Che D, Wang M, Sun J, et al. KRT6A Promotes Lung Cancer Cell Growth and Invasion Through MYC-Regulated Pentose Phosphate Pathway. Front Cell Dev Biol,2021,9:694071.

[22] Martin LL, Kubeil C, Simonov AN, et al. Electrochemistry of cytochrome P450 17α-hydroxylase/17, 20-lyase (P450c17). Mol Cell Endocrinol,2017,441:62-67.

[23] 牛万祥,周晨旭,牛朝诗.CYP17A1 和 AR 在脑胶质瘤中的表达及意义.国际神经病学神经外科学杂志,2017,44(1):1-4.

[24] 林中民,王芳,张泉波,等.姜黄素衍生物 B06 对 2 型糖尿病大鼠睾酮合成的影响.中国病理生理杂志,2016,32(2):352-358.

[25] 李辉,张丽娜,宋瑞佳,等.微小 RNA-221 靶向组织金属蛋白酶抑制因子 2 介导 Akt/mTOR 信号通路对非小细胞肺癌移植瘤小鼠模型的影响.解剖学报,2022,53(6):754-761.

[26] Poonia B, Ayithan N, Nandi M, et al. HBV induces inhibitory FcRL receptor on B cells and dysregulates B cell-T follicular helper cell axis. Sci Rep,2018,8(1):15296.

(责任编辑:郭海强)