

列线图预测前列腺癌患者的预后： 基于蛋白组学的研究

杨磊, 张志强*

安徽医科大学第二附属医院泌尿外科, 安徽 合肥

收稿日期: 2024年11月27日; 录用日期: 2024年12月21日; 发布日期: 2024年12月31日

摘要

前列腺癌是男性最常见且致命的恶性肿瘤之一, 具有较高的发病率和死亡率。本研究利用TCGA数据库蛋白质组学数据, 采用生物信息学方法利用蛋白组学构建前列腺癌预后模型。通过COX回归分析确定了关键的蛋白质, 并利用它们构建了风险评分模型。通过Kaplan-Mier曲线和ROC曲线等多种方法验证了该模型的预测性能。结果表明, 模型能够有效地将前列腺癌患者分为高风险组和低风险组, 具有很强的预后准确性。本研究为前列腺癌的精准医疗提供了理论依据, 并为临床个性化治疗提供了新的方向。

关键词

前列腺癌, 蛋白组学, TCGA数据库, 预后模型, 风险评分

A Nomogram for Predicting the Prognosis of Patients in Prostate Cancer: Research Based on Proteomics

Lei Yang, Zhiqiang Zhang*

Department of Urology, The Second Affiliated Hospital of Anhui Medical University, Hefei Anhui

Received: Nov. 27th, 2024; accepted: Dec. 21st, 2024; published: Dec. 31st, 2024

Abstract

Prostate cancer is one of the most common and lethal malignant tumors in men, which has a high incidence and mortality rate. This study leverages proteomics data from the TCGA database and

*通讯作者。

文章引用: 杨磊, 张志强. 列线图预测前列腺癌患者的预后: 基于蛋白组学的研究[J]. 临床个性化医学, 2024, 3(4): 2445-2454. DOI: 10.12677/jcpm.2024.34349

applies bioinformatics methods to build a protein expression prognostic model for prostate cancer. Key protein markers were identified COX regression analysis, and a risk scoring model was constructed based on these key proteins. The model's predictive performance was verified by various methods, including Kaplan-Mier Curve and ROC Curve. The results demonstrate that this model can effectively distinguish between high-risk and low-risk groups of patients about prostate cancer with strong prognostic accuracy. This study provides a theoretical basis for precision medicine in prostate cancer and offers a new direction for clinical personalized treatment.

Keywords

Prostate Cancer, Proteomics, TCGA Database, Prognostic Model, Risk Score

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 背景

前列腺癌是泌尿系统最常见的恶性肿瘤之一, 在世界范围内, 其发病率高居男性肿瘤第 2 位, 死亡率高居男性肿瘤第 5 位[1]。相比于西方和欧美国家, 在中国, 前列腺癌的发病率与死亡率稍低, 均位于男性肿瘤第 7 [2]。但随着筛查技术的进步与普及, 前列腺癌的发病率呈现逐年上升的态势, 且相较于西方欧美人群, 趋势更快更高[3]。前列腺癌是一种异质性很高的癌种, 在不同的人群中呈现出完全不同的进展情况, 治疗方法也不尽相同。因此, 前列腺癌的诊断与治疗亟需进一步的规范化和个体化。

蛋白质作为细胞组成的重要部分, 是生物体内最重要的生物大分子之一, 蛋白质在细胞中发挥许多关键的作用, 包括催化反应(如蛋白酶)、细胞信号传递、免疫反应以及细胞的结构支持等[4]。此外, 蛋白质的功能不局限于单个细胞, 在生物体内调节注入激素和生长因子等生物过程, 协助维持机体的稳定。因此, 蛋白质是维持生命活动的核心分子, 在前列腺癌的发生发展中起到了至关重要的作用。

TCGA 数据库(The Cancer Genome Atlas): 是由美国国家癌症和肿瘤研究所(NCI)和国家人类基因组研究所(NHGRI)于 2006 年联合启动, 目前已经有 20 多种组织类型的 30 多种癌症 11,000 多个病人的临床与基因表达信息[5]。其建立之初目的是理解癌症的分子机制, 绘制癌症基因图谱, 进而提高诊断、治疗和预防癌症的能力。目前这些信息全部开放供研究者下载使用, 本文所需数据均来自 TCGA 数据库。

2. 研究方法

2.1. 材料收集与整理

以“TCGA-PRAD”关键词搜索, 从 GDC 官网(<https://portal.GDC.cancer.gov/>)下载前列腺癌蛋白表达谱、临床数据和转录组数据, 使用 perl 软件整理文件, 得到 352 位前列腺癌患者的蛋白表达信息和临床信息。整理临床数据时, 将进展期前列腺癌患者(包括淋巴转移或骨转移)与死亡患者统归为不良预后结果。

2.2. 建立预测模型

通过 R-studio 建立预后预测模型, 将 352 位前列腺癌患者随机分为实验组和测试组, 使用单因素 COX 对实验组所有蛋白进行 COX 回归分析, 符合条件的蛋白使用多因素 COX 回归分析, 同种方法作用于测试组。当两组回归曲线的 AUC 均大于 0.7 时, 输出多因素结果。

2.3. 评估差异风险基因

根据模型筛选出的蛋白表达情况, 为每一位患者赋予 Risk 评分, 评分的计算方法为 $\text{risk score} = \sum_{i=1}^n (\exp i * \beta_i)$, 其中 $\exp$ 表示蛋白质的表达量, β 表示 LASSO 算法中患者的系数。以中位风险评分为临界值, 将患者分为高危组和低危组进行随访分析选取 Risk 评分的中位数为截断值, 将所有患者分为高低风险组。选取高低风险组为研究对象, 对单因素筛选出的所有预后相关蛋白进行差异风险分析。使用 “ggplot” R 包对结果进行可视化, 纵坐标为 FDR, 横坐标为差异倍数。

2.4. 主成分分析

为了评估高低风险组分类器效果, 使用 PCA (principal component analysis) 分析[6]方法进行处理。根据模型结果将 352 位前列腺癌患者分为高低风险两组, 对前列腺癌患者的所有蛋白以及风险蛋白进行分析, 使用 “scatterplot3d” R 包将结果可视化, 得到两组结果的 3D PCA 图。

2.5. 生存分析

使用 “survival [7]” 包中的 fit 函数, 针对实验组、测试组以及所有患者的高低风险组, 进行生存分析, 结果使用 Kaplan-Miere 曲线[8]表示结果。其中, 纵坐标为总体生存率, 横坐标为生存时间。接下来, 我们使用同种方法, 二分类法将所有患者分为单个模型蛋白的高低表达组, 进行生存分析, 同样对结果可视化处理。

2.6. 独立预后分析

为了进一步评估模型效能, 针对 Risk 评分、年龄、T 分期、N 分期, 使用单因素 COX 回归分析评估他们的风险系数(HR) [9], 同种方法作用于多因素 COX 回归分析。所得结果以森林图样式显示, 选取 95% 置信区间, 并显示他们的 P 值。

2.7. ROC 曲线

受试者操作曲线所围成的曲线下面积(AUC) [10]作为一种良好的模型预测指标, 分别评估模型在横向跨度(1、3、5 年生存期)以及纵向指标(Risk 评分、年龄、T 分期、N 分期)上的预测效能。其中, 横坐标为特异程度, 纵坐标为敏感程度, 将结果输出为图表。

2.8. Nomogram 图

Nomo 图[11]是根据模型现有参数, 为不同临床性状赋予评分, 根据评分之和可预测相对应的 1、3、5 年生存期, 进而指导前列腺癌患者的个体化治疗, 主要用来评估预测模型。使用 “rms” 包进行分析, 使用 “regplot” 包分别做出校准曲线和 Nomo 图。

3. 研究结论

3.1. 建立前列腺癌的预后模型

使用上述建模方法, 单因素 COX 回归筛选预后相关蛋白质, 所得结果如图 1(A)所示, 共得到 21 个生存相关蛋白质。其中, 共 17 种蛋白质 $HR > 1$, 代表它们与不良预后有关。将单因素结果进行 LASSO 回归, 避免数据过拟合, 进一步筛选变量, 不同危险因素 LASSO 系数路径图 1(B)和交叉验证曲线图 1(C)如下所示, LASSO 所得的风险系数以及多因素 COX 回归筛选后共得到 b-Actin、JNK_pT183Y185、MEK1、IRF-3_pS396 四种预后相关蛋白质, 使用它们来计算 Risk 评分, 根据 Risk 评分高低将所有患者分为高低风险组, 建立预后模型进而继续后续分析。

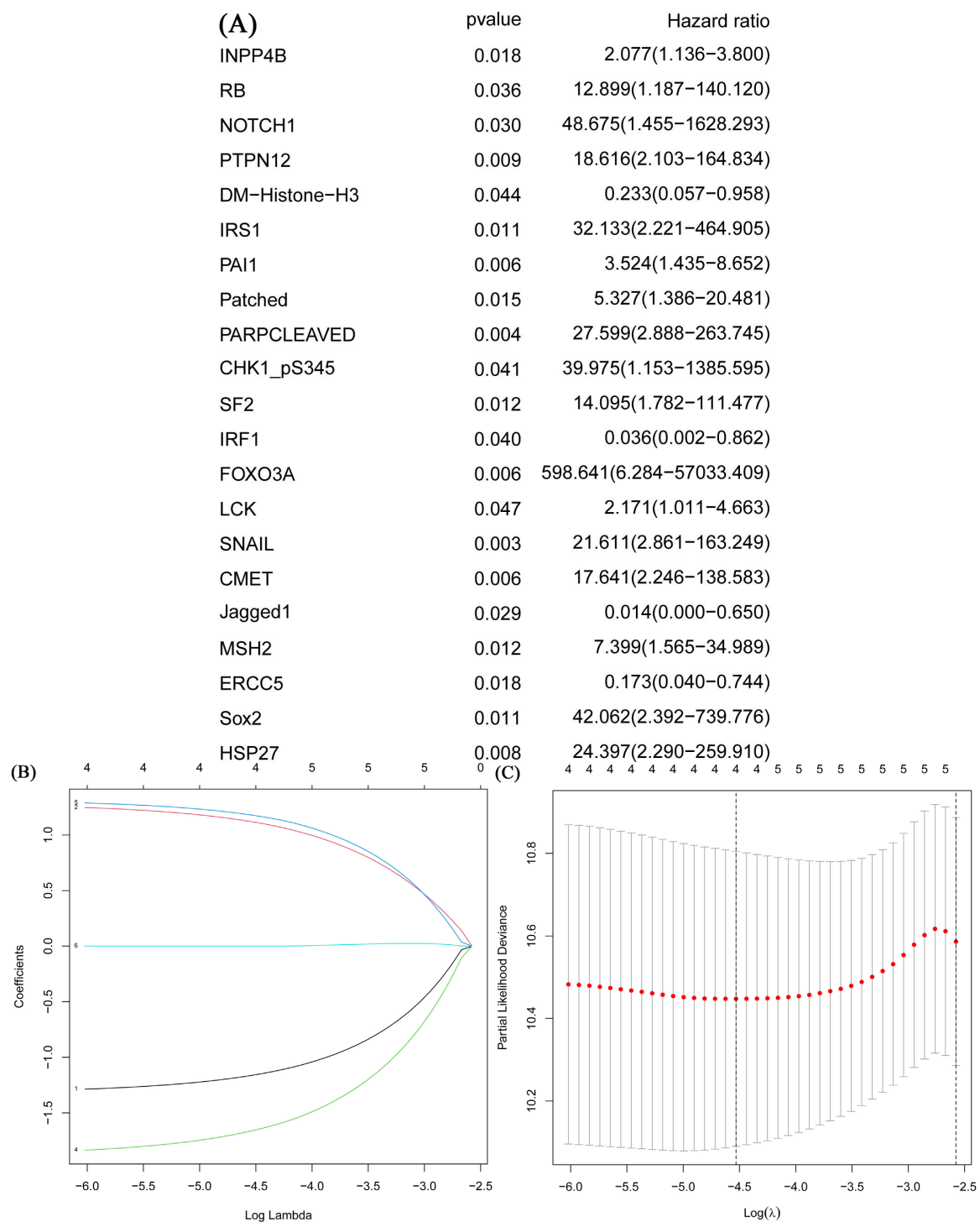
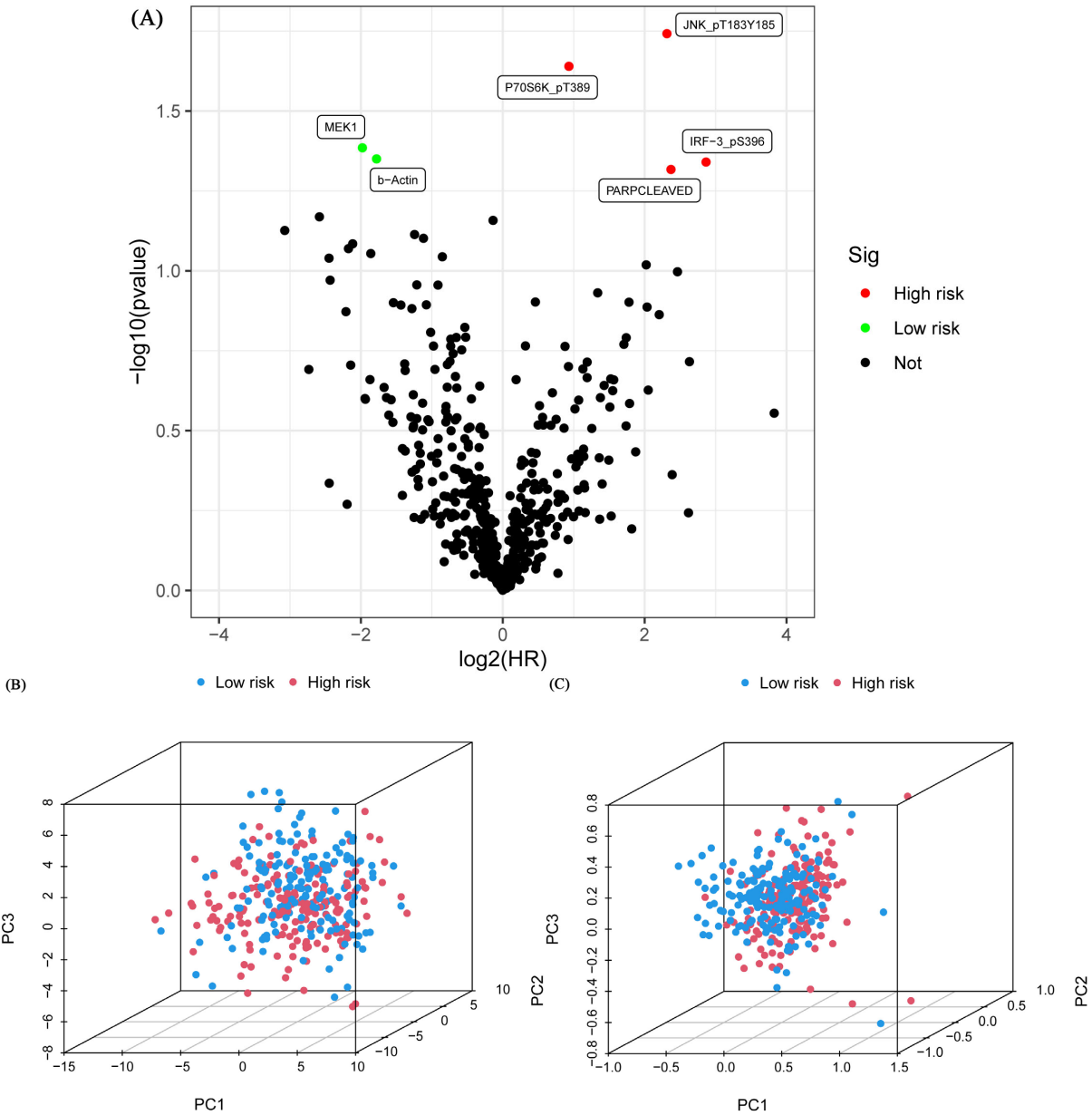


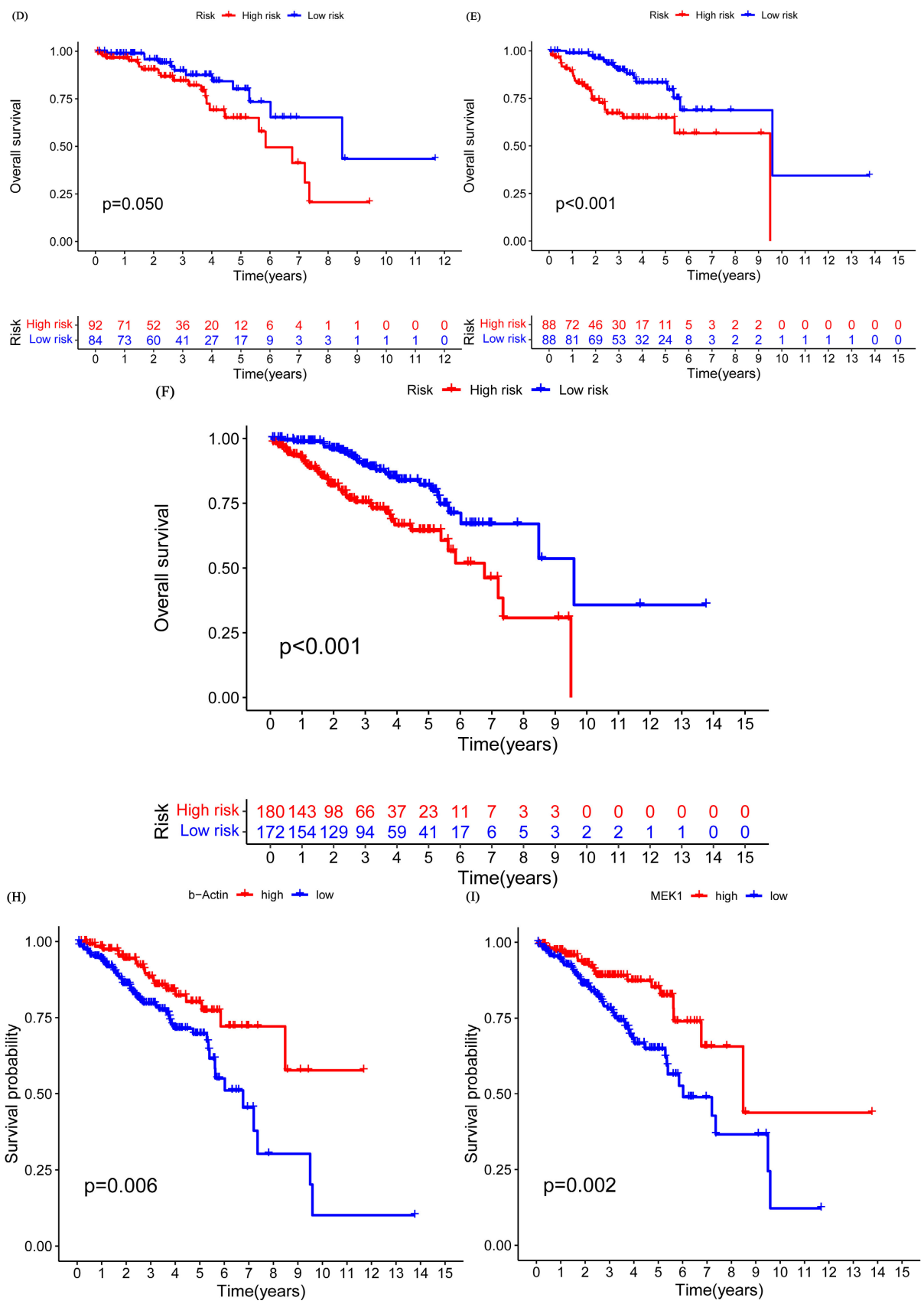
Figure 1. (A) Univariate COX regression analysis; (B) LASSO coefficient path plot; (C) LASSO regularization plot
图 1. (A) 单因素 COX 回归分析; (B) LASSO 系数路径图; (C) LASSO 正则化路径图

3.2. 评估模型的预后价值

使用差异分析筛选高低风险组的不同风险蛋白, 风险倍数 log2 化, 取 2 为截断值, 显示所有具有差异的蛋白质, FDR 均大于 1。将模型基因标注在图 2(A)中, 可以看到 b-Actin、MEK1 两种蛋白在高风险

组中显著下调, JNK_pT183Y185、IRF-3_pS396 两种蛋白在高风险组中显著上调, 这表明前者可能是机体的保护因素, 后者可能与不良预后有关, 是健康的危险因素。接下来, 我们进一步使用了主成分分析, 所得结果在图 2(B)~(C)中。比较所有蛋白与模型风险蛋白分组, 可以看出模型蛋白在高低风险组中具有内在的相似度, 其表达模式具有潜在的联系, 这表明使用 Risk 评分作为分类器具有良好的效能。最后, 生存分析进一步判断高低风险组的生存情况, 如图 2(D)~(F)所示, 在测试集、实验集以及所有患者中, 使用 Risk 作为分类器划分的高低风险组中, 高风险组的生存期以及生存率均显著小于低风险组, 这表明我们的模型可以用来预测患者的生存期。此外, 针对四种模型蛋白质, 我们分别针对每一种模型蛋白的高低表达组作为分类器, 进行生存分析, 可视化结果如图 2(H)~(K)。b-Actin、MEK1 高表达组的生存期显著高于低表达组, JNK_pT183Y185、IRF-3_pS396 高表达组的生存期显著低于低表达组, 这与前文的 HR 差异分析结果相一致, 进一步明确了模型蛋白的预后价值。





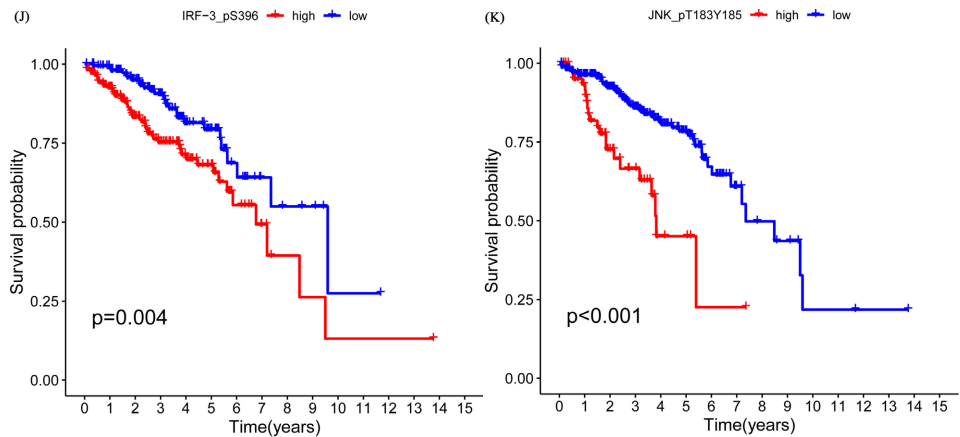
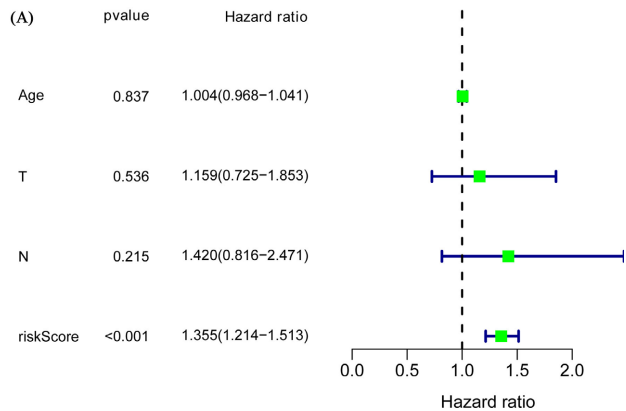


Figure 2. (A) Differential risk coefficient plot of high and low-risk groups; (B) Principal component analysis plot of high and-risk patients based on the distribution of all proteins; (C) Principal component analysis plot of high and low-risk patients based on the distribution of model proteins; (D) Kaplan-Meier survival curves of high and low-risk groups in the test set; (E) Kaplan-Meier survival curves of high and low-risk in the experimental set; (F) Kaplan-Meier survival curves of high and low-risk groups in all patients; (H) Kaplan-Meier survival of high and low expression groups of b-Actin protein; (I) Kaplan-Meier survival curves of high and low expression groups of MEK1; (J) Kaplan-Meier survival curves of high and low expression groups of IRF-3_Ps396 protein; (K) Kaplan-Mier survival curves of high and low expression groups of JNK_pT183Y185 protein

图 2. (A) 高低风险组的差异风险系数图; (B) 依据所有蛋白质的分布在高低风险组患者中的主成分分析图; (C) 依据模型蛋白的分布在高低风险组患者中的主成分分析图; (D) 测试集高低风险组患者的生存曲线; (E) 实验集高低风险组患者的生存曲线; (F) 所有患者高低风险组患者的生存曲线; (H) b-Actin 蛋白高表达组和低表达组患者的生存曲线; (I) MEK1 蛋白高表达组和低表达组患者的生存曲线; (J) IRF-3_Ps396 蛋白高表达组和低表达组患者的生存曲线; (K) JNK_pT183Y185 蛋白高表达组和低表达组患者的生存曲线

3.3. 验证模型与使用模型

COX 回归曲线与 ROC 曲线被用来验证模型的先进性。首先，我们对不同临床性状使用了单因素 COX，年龄、T 分期、N 分期以及 Risk 评分作为自变量评估风险系数，所得结果如图 3(A)所示，只有 Risk 评分的结果在统计学上有意义， $P < 0.001$ 。接下来，进一步使用多因素 COX 评估，所得结果与单因素相一致，如图 3(B)所示。这表明，我们所使用的 Risk 评分系统是前列腺癌的独立预后因素，具有评估前列腺癌预后的价值。最后，ROC 曲线被用来验证横向和纵向跨度上模型的优越性，AUC 均大于 0.65。其中，时间 ROC 曲线如图 3(C)，针对 1 年的评估效能最好，3 年次之，5 年预测准确率稍差。临床性状 ROC 如图 3(D)，Risk 评分系统明显优于其他临床性状。建立、评估与验证模型之后，针对模型的使用，我们分析了 Nomo 图，如图 3(E)所示，所有临床性状被赋予打分，根据临床性状打分之用来预测前列腺癌患者的 1、3、5 年生存率。Nomo 图的验证校准曲线如图 3(F)所示。



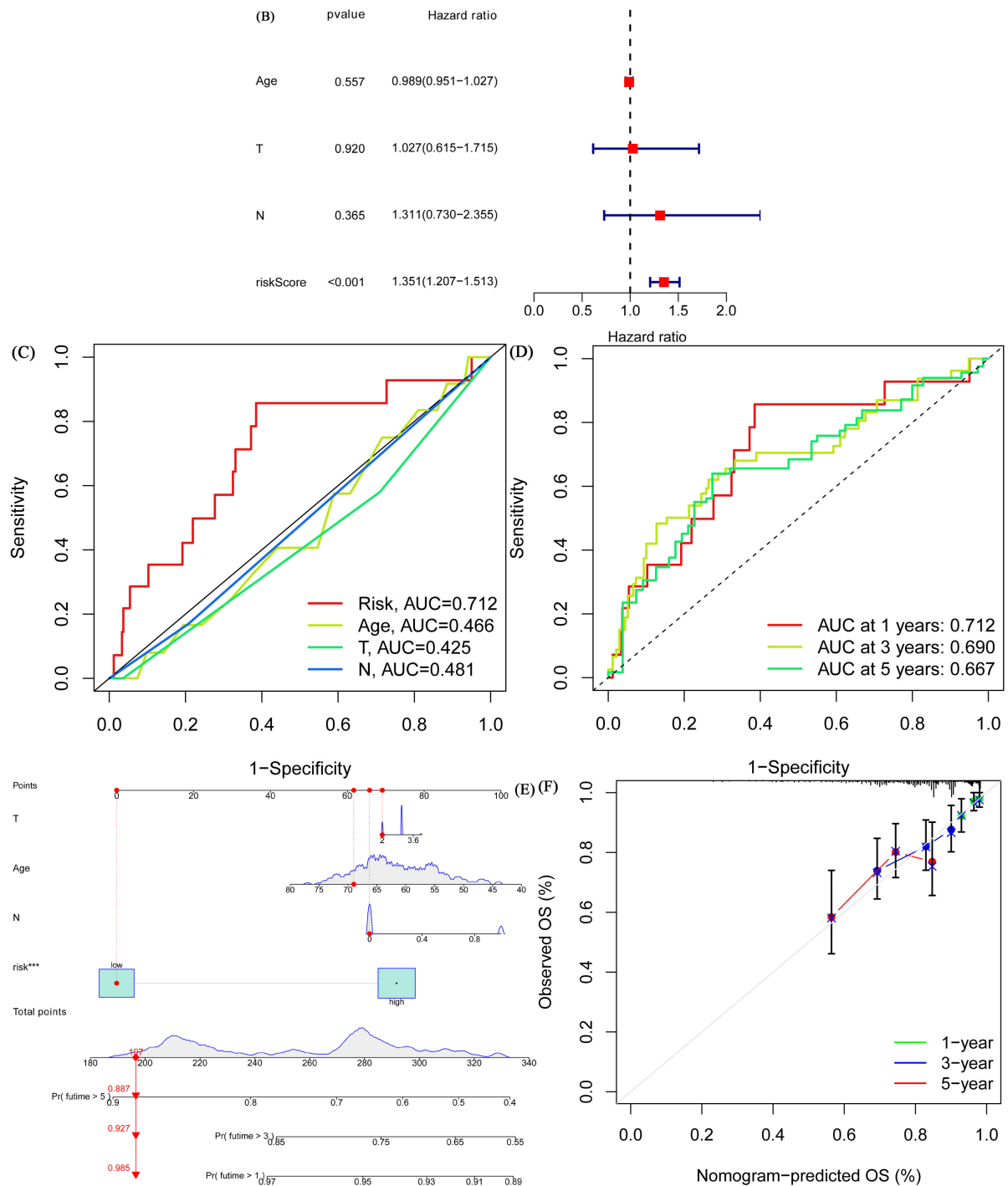


Figure 3. (A) Univariate COX regression for different clinical features; (B) Multivariate COX regression for different clinical features; (C) ROC curves for different clinical features; (D) ROC curves for 1, 3, and 5-year time points; (E) Nomo for prostate cancer patients; (F) Calibration curve for the nomogram

图 3. (A) 针对不同临床性状的单因素 COX 回归; (B) 针对不同临床性状的多因素 COX 回归; (C) 针对不同临床性状的 ROC 曲线; (D) 针对 1、3、5 年时间的 ROC 曲线; (E) 前列腺癌患者的 Nomo 图; (F) Nomo 图的校准曲线

4. 研究结果与讨论

通过本研究构建基于蛋白组学数据的前列腺癌预后模型, 不仅帮助我们识别了 b-Actin、MEK1、

JNK_pT183Y185、IRF-3_pS396 四种蛋白质作为新的生物标志物, 还能在临床实践中提供重要的预后信息。通过分析蛋白质表达模式与患者生存期、疾病进展等临床结局之间的关系, 我们能够为前列腺癌患者提供个体化的预后评估。JNK_pT183Y185、IRF-3_pS396 的高表达可能提示肿瘤侵袭性较强, 患者预后较差; 而 b-Actin、MEK1 的低表达可能与较好的预后相关。

此外, 前列腺癌的预测模型不仅限于生存预测, 还可以为临床治疗方案的选择提供依据。对于高风险患者, 可能需要更加积极的治疗干预, 如手术治疗、化疗或放射治疗; 而低风险患者则可以选择监测观察或轻度干预, 如药物去势治疗等。通过对预后模型的精准应用, 能够有效提高治疗的精确性, 避免不必要的过度治疗, 降低患者的治疗负担, 提高生活质量。

尽管基于 TCGA 数据库的蛋白组学预后模型在前列腺癌研究中取得了显著的成果, 但仍然存在一些局限性。首先, TCGA 数据库中包含的数据主要来自美国等特定地区, 可能存在一定的种族和地域偏差。其次, 蛋白组学数据的异质性较大, 不同实验平台和样本的差异可能会影响数据的可比性和一致性。因此, 在未来的研究中, 除了扩大样本量, 还需要进行多中心、跨区域的验证, 以提高模型的普适性和可靠性。

综上所述, 基于 TCGA 公共数据库的蛋白组学数据, 结合生物信息学分析工具, 我们成功构建了前列腺癌的预后模型。通过对蛋白质表达模式的深入分析, 我们不仅揭示了与前列腺癌预后相关的潜在标志物, 还为临床实践提供了有价值的参考依据。然而, 模型的建立仍面临着数据异质性、样本规模以及算法选择等方面的挑战, 未来的研究可以通过多中心、跨组学的数据整合和机器学习技术的应用, 进一步提升前列腺癌预后模型的准确性和临床实用性。

基金项目

2023 年度安徽省高校自然科学研究项目, 项目编号: 2023AH053161。

参考文献

- [1] Bergengren, O., Pekala, K.R., Matsoukas, K., Fainberg, J., Mungovan, S.F., Bratt, O., *et al.* (2023) 2022 Update on Prostate Cancer Epidemiology and Risk Factors—A Systematic Review. *European Urology*, **84**, 191-206. <https://doi.org/10.1016/j.eururo.2023.04.021>
- [2] Sun, D., Li, H., Cao, M., He, S., Lei, L., Peng, J., *et al.* (2020) Cancer Burden in China: Trends, Risk Factors and Prevention. *Cancer Biology and Medicine*, **17**, 879-895. <https://doi.org/10.20892/j.issn.2095-3941.2020.0387>
- [3] Wang, Y., Yan, Q., Fan, C., Mo, Y., Wang, Y., Li, X., *et al.* (2023) Overview and Countermeasures of Cancer Burden in China. *Science China Life Sciences*, **66**, 2515-2526. <https://doi.org/10.1007/s11427-022-2240-6>
- [4] Gurevich, V.V. (2019) Protein Multi-Functionality: Introduction. *Cellular and Molecular Life Sciences*, **76**, 4405-4406. <https://doi.org/10.1007/s00018-019-03271-6>
- [5] Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) Review the Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Współczesna Onkologia*, **1**, 68-77. <https://doi.org/10.5114/wo.2014.47136>
- [6] Jang, J.H. (2021) Principal Component Analysis of Hybrid Functional and Vector Data. *Statistics in Medicine*, **40**, 5152-5173. <https://doi.org/10.1002/sim.9117>
- [7] In, J. and Lee, D.K. (2019) Survival Analysis: Part II—Applied Clinical Data Analysis. *Korean Journal of Anesthesiology*, **72**, 441-457. <https://doi.org/10.4097/kja.19183>
- [8] Gomes, A.P., Costa, B., Marques, R., Nunes, V. and Coelho, C. (2024) Kaplan-Meier Survival Analysis: Practical Insights for Clinicians. *Acta Médica Portuguesa*, **37**, 280-285. <https://doi.org/10.20344/amp.21080>
- [9] Su, C., Xue, J. and Liu, N. (2023) Cox Regression Analysis of Prognostic Factors of Intensity-Modulated Radiotherapy in Patients with Bladder Cancer. *Archivos Españoles de Urología*, **76**, 411-417. <https://doi.org/10.56434/j.arch.esp.urol.20237606.50>
- [10] Hoo, Z.H., Candlish, J. and Teare, D. (2017) What Is an ROC Curve? *Emergency Medicine Journal*, **34**, 357-359. <https://doi.org/10.1136/emmermed-2017-206735>

- [11] Zhang, W., Ji, L., Wang, X., Zhu, S., Luo, J., Zhang, Y., *et al.* (2022) Nomogram Predicts Risk and Prognostic Factors for Bone Metastasis of Pancreatic Cancer: A Population-Based Analysis. *Frontiers in Endocrinology*, **12**, Article 752716. <https://doi.org/10.3389/fendo.2021.752176>