

构建结肠癌免疫相关长链非编码 RNA 风险评分模型

吴彬¹, 姚颐^{1,2}, 董熠¹, 杨思琦¹, 宋启斌^{1,2}

(1. 武汉大学人民医院, 湖北 武汉 430060; 2. 湖北省肿瘤精准医学研究中心, 湖北 武汉 430200)

摘要: [目的] 联合免疫相关长链非编码 RNA (long non-coding RNA, lncRNA), 探索用于评估结肠癌预后新模型。[方法] 我们从癌症基因组图谱 (the cancer genome atlas, TCGA) 数据库中下载结肠癌患者的临床数据和基因表达信息, 从分子标记数据库 v4.0 获得免疫相关的基因。Perl 软件和 R 软件用于数据处理和分析。利用 R 语言的相关性检验获得结肠癌免疫相关 lncRNAs。结合临床数据, 利用单因素 Cox 回归分析筛选出与结肠癌预后相关的免疫相关 lncRNAs, 随后进一步多因素分析, 筛选出构建风险评分模型的 lncRNAs。根据风险评分的中位数将患者分为高风险组和低风险组, 运用 Kaplan-Meier (K-M) 生存分析及独立预后因素评估对模型进行评价, 并将此模型联合其他临床因素构建列线图, 对个体进行生存率预测。[结果] 单因素 Cox 回归分析筛选出 33 个与结肠癌预后相关的免疫相关 lncRNA, 多因素 Cox 回归分析最终确定 12 个免疫相关 lncRNA 用来构建风险评分模型。以中位风险评分作为临界值, 患者可被分为高风险组和低风险组, 低风险和高风险组的 5 年生存率分别为 86.1% 和 42.7%。此外, 风险评分模型可作为结肠癌的独立预后因子, 联合结肠癌其他临床因素和风险评分, 建立了列线图以预测结肠癌个体生存率, 该列线图的 C 指数为 0.807 (95%CI: 0.762~0.854), 校准图显示预测值与实际观测值一致性较好。[结论] 由 12 个免疫相关 lncRNA 构成的风险评分模型可用于评估结肠癌预后, 并根据构建的列线图可预测结肠癌患者的生存率。

关键词: 结肠癌; 免疫; lncRNA; 预后; 生存分析

中图分类号: R735.3 **文献标识码:** A **文章编号:** 1671-170X (2020) 11-0966-06

doi: 10.11735/j.issn.1671-170X.2020.11.B008

Construction of Risk Score Model for Prognosis of Colon Cancer with Immune-related Long Non-coding RNA

WU Bin¹, YAO Yi^{1,2}, DONG Yi¹, YANG Si-qi¹, SONG Qi-bin^{1,2}

(1. Renmin Hospital of Wuhan University, Wuhan 430060, China; 2. Hubei Provincial Research Center for Precision Medicine of Cancer, Wuhan 430200, China)

Abstracts: [Objective] To construct a risk score model for prognosis of colon cancer with immune-related long non-coding RNA (lncRNA). [Methods] The clinical data and gene expression information of colon cancer were down loaded from the cancer genome atlas database and the immune-related genes were obtained from the molecular marker database v4.0. Perl software and R software were used for data processing and analysis. The immune-related lncRNAs of colon cancer were obtained through the correlation test of R language. Univariate Cox regression analysis was used to screen out immune-related lncRNAs related to the prognosis of colon cancer, and multivariate Cox regression was used to screen out lncRNAs for constructing a risk score model. According to the median risk score, patients were divided into high-risk groups and low-risk groups. Kaplan-Meier survival analysis, and independent prognostic factors were used to evaluate the model. Combining the constructed model with other clinical factors a nomogram was constructed to predict the survival of patients. [Results] Univariate Cox regression analysis screened out 33 immune-related lncRNAs associated with the prognosis of colon cancer, and multivariate Cox regression analysis finally determined 12 immune-related lncRNAs to construct a risk score model. With the median risk score as the cut-off value, patients were divided into high-risk and low-risk groups. The 5-year survival rates of the low-risk and high-risk groups were 86.1% and 42.7%, respectively. The C index of the developed nomogram was 0.807 (95%CI: 0.762~0.854). The calibration chart showed that the predicted value was in good agreement with the actual observed value. [Conclusion] The risk score model composed of 12 immune-related lncRNAs can be used to assess the prognosis of colon cancer, and the constructed nomogram can predict the survival probability of colon cancer patients.

Subject words: colon cancer; immune; lncRNA; prognosis; survival analysis

通信作者: 姚颐, 副主任医师, 博士; 武汉大学人民医院肿瘤中心, 湖北省武汉市武昌区张之洞路 99 号 (430060); E-mail: yaoyi2018@whu.edu.cn

宋启斌, 主任医师, 博士生导师, 博士; 武汉大学人民医院肿瘤中心, 湖北省武汉市武昌区张之洞路 99 号 (430060); E-mail: qibinson@whu.edu.cn

收稿日期: 2020-07-26; 修回日期: 2020-09-28

结肠癌是常见的消化道恶性肿瘤,其发病率和死亡率均居世界前列,而早期诊断和预后判断手段的缺乏是结肠癌患者死亡率高的主要原因^[1-2]。现已发现长链非编码 RNA (long non-coding RNA, lncRNA) 参与了肿瘤发展的多个环节^[3],亦有研究提示其在结肠癌预后评估中的重要作用^[4-5],但免疫相关的 lncRNAs 在结肠癌预后中的意义少见报道。本研究拟构建结肠癌免疫相关的 lncRNA 的风险评分模型,以期为结肠癌患者预后判断和免疫治疗疗效的预测提供新的参考依据。

1 资料与方法

1.1 结肠癌 lncRNA 和临床数据的收集和处理

从 TCGA 数据库 (<https://portal.gdc.cancer.gov/>) 中下载结肠癌相关的 RNA 序列 (RNA-seq) 和临床数据。下载的数据集包括 473 个肿瘤样本和 41 个正常样本。Perl 软件用于数据整合以及 lncRNA 表达数据和相应临床数据的提取,其中基因表达数据提取项目包括基因名称、样本编号和表达值,临床数据提取项目包括患者编号、生存时间、生存状态、年龄、性别、TNM 分期和 T、N、M 分期。

1.2 免疫相关 lncRNA 的挖掘及建立风险评分模型

从分子标记数据库 v4.0 (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>) 中提取了 332 个免疫相关基因。利用 R 软件将提取的 lncRNA 和免疫相关基因进行相关性检验以筛选出免疫相关的 lncRNAs。

利用免疫相关 lncRNAs 和相应的结肠癌临床数据建立预后模型。首先,利用单因素 Cox 回归分析评估免疫相关 lncRNAs 表达水平与总生存期 (overall survival, OS) 之间的关系,并以 $P < 0.01$ 为筛选标准,筛选出与结肠癌预后有关的免疫相关 lncRNAs。随后,将所筛出的免疫相关 lncRNAs 纳入多因素 Cox 回归分析,根据最优的赤池信息 (Akaike information criterion, AIC) 模拟准则,筛选出最终构成风险评分模型的免疫相关 lncRNAs。模型公式如下所示:

$$\text{风险评分} = \sum_{i=1}^N (E_i * C_i) \quad [6-7]$$

其中 N 表示构建风险评分模型的免疫相关 lncRNA 数目, C_i 表示免疫相关 lncRNA 系数, E_i 表示免疫相关 lncRNA 的表达水平。

1.3 模型评价和生存分析

依据风险评分模型,每位患者可获得相应的风险评分。以中位风险评分作为临界值,患者可被分为高风险组和低风险组。采用 K-M 法对高、低风险组进行生存分析。

为进一步评价此模型,将风险评分与其他临床特征进行单因素和多因素 Cox 回归分析,以考查其是否为独立预后因子。

1.4 列线图的构造

用于构建结肠癌列线图的变量包括年龄、性别、分期、TNM 分期和风险评分,绘制校准曲线以评估实际生存率和预测生存率之间的一致性。

列线图中,参照患者的变量在相应的坐标轴上找到相应点,过此点作垂线,与分数轴上交点的读数就是变量的分数,每个变量的分数之和即总分数。以同样方式,在生存率轴上读取总分数的数值,该分数便是患者在相应时间内存活的概率^[8-9]。

2 结果

2.1 免疫相关 lncRNA 预后模型的建立

本研究汇总了 TCGA 数据库中 452 例结肠癌患者的临床基线资料 (Table 1)。利用 R 语言的相关性检验筛选出 657 个免疫相关的 lncRNAs。针对这些 lncRNAs,同时结合结肠癌的临床信息,单因素 Cox 回归分析初步筛选出 33 个与结肠癌预后相关的免疫相关 lncRNAs (Figure 1)。将这些 lncRNA 纳入多因素 Cox 回归分析,最终确定 12 个免疫相关 lncRNAs 可作为结肠癌的预后指标 (Table 2),其中,11 个是高风险的 lncRNA (AC105219.1、PCAT6、AP001469.3、AL391422.4、AL162274.2、AC156455.1、AC139768.1、LINC01063、AC023509.4、AC068580.3、AC073283.1, 系数 > 0), 1 个是低风险的 lncRNA (LINC00174, 系数 < 0)。根据 Table 2 中免疫相关 lncRNA 的系数,获得预后模型公式以评估每位患者的风险评分,如下所示:

$$\begin{aligned} \text{风险评分} = & 0.305 \times E_{AC105219.1} + 0.364 \times E_{PCAT6} + 0.509 \times \\ & E_{AP001469.3} + 1.323 \times E_{AL391422.4} + 0.521 \times E_{AL162274.2} + 0.276 \times E_{AC156455.1} + \\ & 0.654 \times E_{AC139768.1} + 0.859 \times E_{LINC01063} + 0.566 \times E_{AC023509.4} + 1.036 \times \\ & E_{AC068580.3} + 0.510 \times E_{AC073283.1} - 0.570 \times E_{LINC00174}, \end{aligned}$$

其中 E 表示免疫相关 lncRNA 表达水平。

2.2 免疫相关 lncRNA 风险评分模型的评价

根据所得模型计算每位患者相应风险评分。将所有评分进行排序,以中位风险分数为临界值,将所有患者分为高风险组和低风险组(Figure 2A)。高风险组患者的生存状况较差(Figure 2B)。热图中显示 12 个 lncRNAs 在每个样品中的表达水平,其中粉色线条表示低风险组,蓝色线条表示高风险组(Figure 2C),且表达的差异均具有统计学意义($P < 0.05$)。生存分析显示,与高风险组相比,低风险组的 OS 显著性延长($P < 0.001$):低风险和高风险组的 5 年生存率分别为 86.1%(95%CI:76.8%~96.5%)和 42.7%(95%CI:30.87%~59.0%)(Figure 3)。

将风险评分和其他临床因素一起做

Table 1 Clinical baseline data for colon cancer patients of TCGA database

Clinical characteristics	Number
Gender	
Male	238
Female	214
Age(years-old)	
<65	169
≥65	283
Stage	
I	76
II	178
III	125
IV	62
Unknown	11
T stage	
Tis	1
T1	10
T2	77
T3	308
T4	56
N stage	
N0	269
N1	103
N2	80
M stage	
Mx	49
M0	334
M1	62
Unknown	7

单因素和多因素 Cox 回归分析以分析其是否为独立预后因子,从 Figure 4A 和 4B 中可以看出,无论单因素还是多因素分析,风险评分的 P 值均小于 0.001,说明此风险评分模型可作为结肠癌独立的预后因子。此外,风险评分的危险比大于 1,可见,相对于低风险评分组,高风险评分组是结肠癌的高危风险因素。

2.3 列线图的建立和验证

结合风险评分,我们对结肠癌样本进行 3 年和 5 年生存率分析。如 Figure 5A 所示,分配给每个变量的分数与其对生存的风险贡献成正比。经过测试,列线图的内部验证 C 指数为 0.807(95%CI:0.762~0.854),可信度较高。此外,模型的校准图显示,通过该列线图得出的生存率预测值与实际观测值一致性

Table 2 Coefficient and hazard ratio of 12 immune-related lncRNAs identified from multivariate Cox regression

lncRNA	Coef	HR	95%CI
AC105219.1	0.305	1.356	0.969~1.898
PCAT6	0.364	1.440	0.976~2.124
AP001469.3	0.509	1.664	0.856~3.234
AL391422.4	1.323	3.757	1.901~7.425
AL162274.2	0.521	1.685	0.939~3.024
AC156455.1	0.276	1.319	0.966~1.799
AC139768.1	0.654	1.925	0.892~4.151
LINC01063	0.859	2.361	1.313~4.247
AC023509.4	0.566	1.761	0.989~3.134
AC068580.3	1.036	2.819	1.464~5.428
AC073283.1	0.510	1.665	1.023~2.709
LINC00174	-0.570	0.565	0.304~1.049

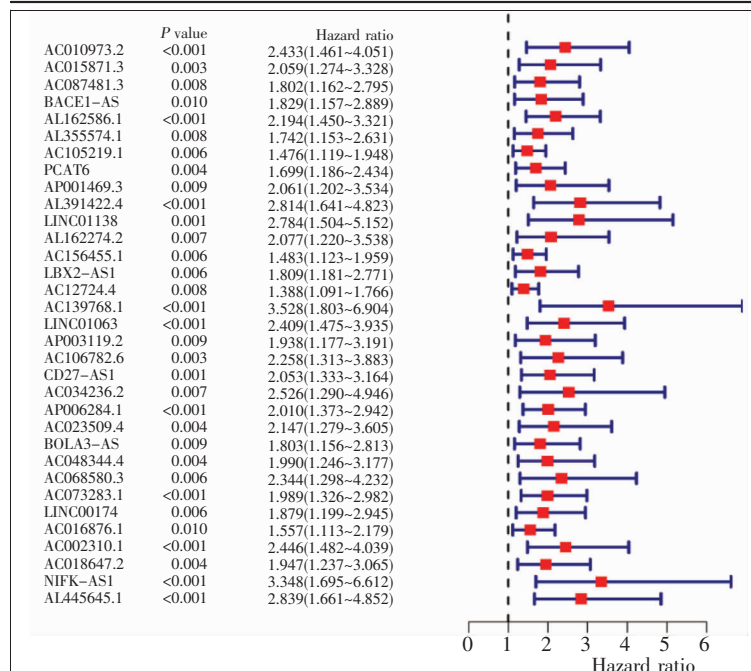


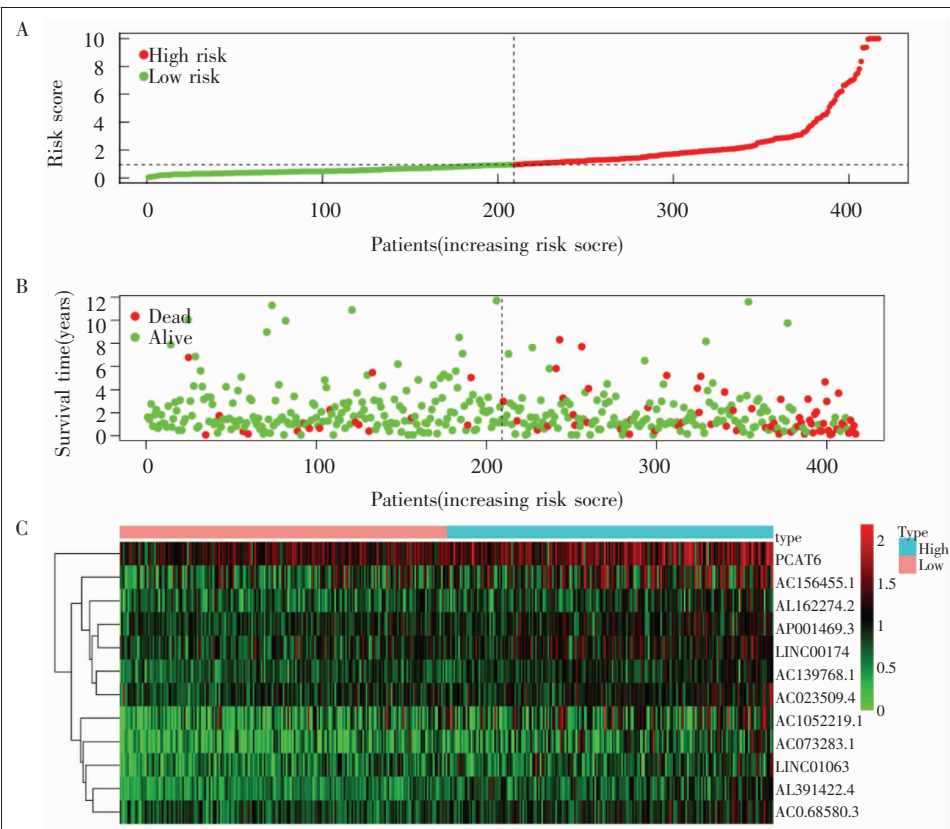
Figure 1 Hazard ratio of 33 immune-related lncRNAs associated with overall survival with univariate Cox regression

较好 (Figure 5B、5C)。

3 讨论

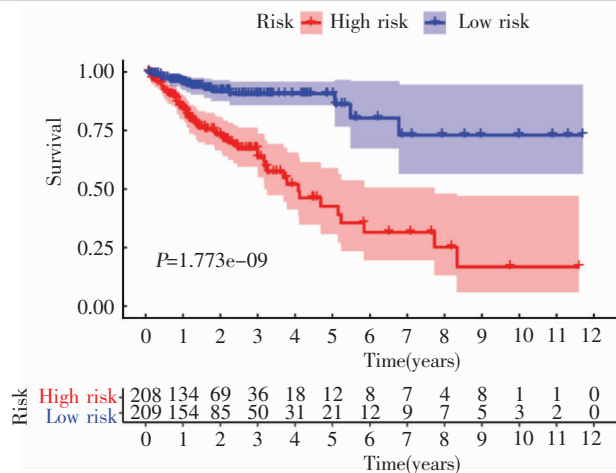
良好的疗效预测和预后判断方法是临床诊治结肠癌的关键。而现有的以临床特征或血清分子标志物 (CEA、CA199 等) 作为预后判断的手段存在一定局限性^[10]。lncRNA 是一组长度大于 200 个核苷酸的非编码 RNA 分子, 可通过影响邻近基因或其他染色体上的远距离基因来调节编码基因的表达^[11]。现已发现在结肠癌中有很多异常表达的 lncRNA (如 SNHG11^[12]、POU6F2 - AS2^[13]、SNHG7^[14] 等), 作为结肠癌的特异性的生物学标志物在预后、诊断、疗效预测和预后判断等方面发挥重要作用。然而, 肿瘤具有复杂的分子特征, 单一的 lncRNA 表达模式可能不足以准确判断患者的预后, 多种潜在的 lncRNA 生物标志物可以提高预后判断的准确性^[2,4]。

本研究构建了由 12 个免疫相关 lncRNA 组成的风险评分模型以期更好地评估结肠癌患者的预后。已有证据表明, lncRNAs 是免疫反应的关键调节因子, 参与免疫反应和基因激活, 并促进免疫表型^[15-16]。随着免疫疗法在癌症治疗中取得瞩目的成果, 免疫相关的 lncRNA 逐渐成为一个新的热点, 且已被证明是胰腺癌^[17]、恶性神经胶质瘤^[18]、



A: Immune-related lncRNAs risk score distribution of the colon cancer patients from TCGA RNA-Seq cohort; B: patients' survival status and time distributed by risk score; C: heat map of the 12 immune-related lncRNAs expression profiles in colon cancer patients. Rows denote immune-related lncRNAs and columns represent patients. Green to red indicates a trend from low to high expression. Dashed line indicates the immune-related lncRNAs signature cutoff arranging patients into low- and high-risk groups. RNA-seq, RNA sequencing; TCGA, The Cancer Genome Atlas.

Figure 2 Immune-related lncRNAs predictive risk score analysis



Significant correlation between the OS of patients in high-risk and low-risk groups ($P < 0.001$)

Figure 3 Kaplan-Meier for risk score of colon cancer patients

肾癌^[19]和乳腺癌^[20]等多种肿瘤的预后因子。与前人研究类似, 本研究也证实了上述 12 个免疫相关 lncRNA

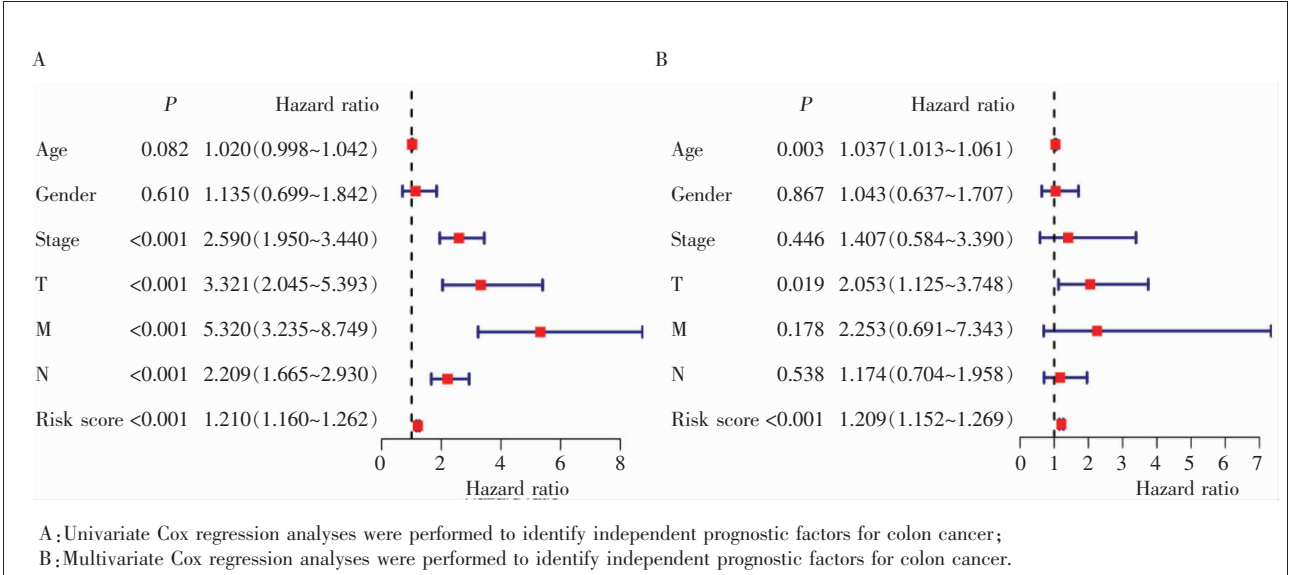
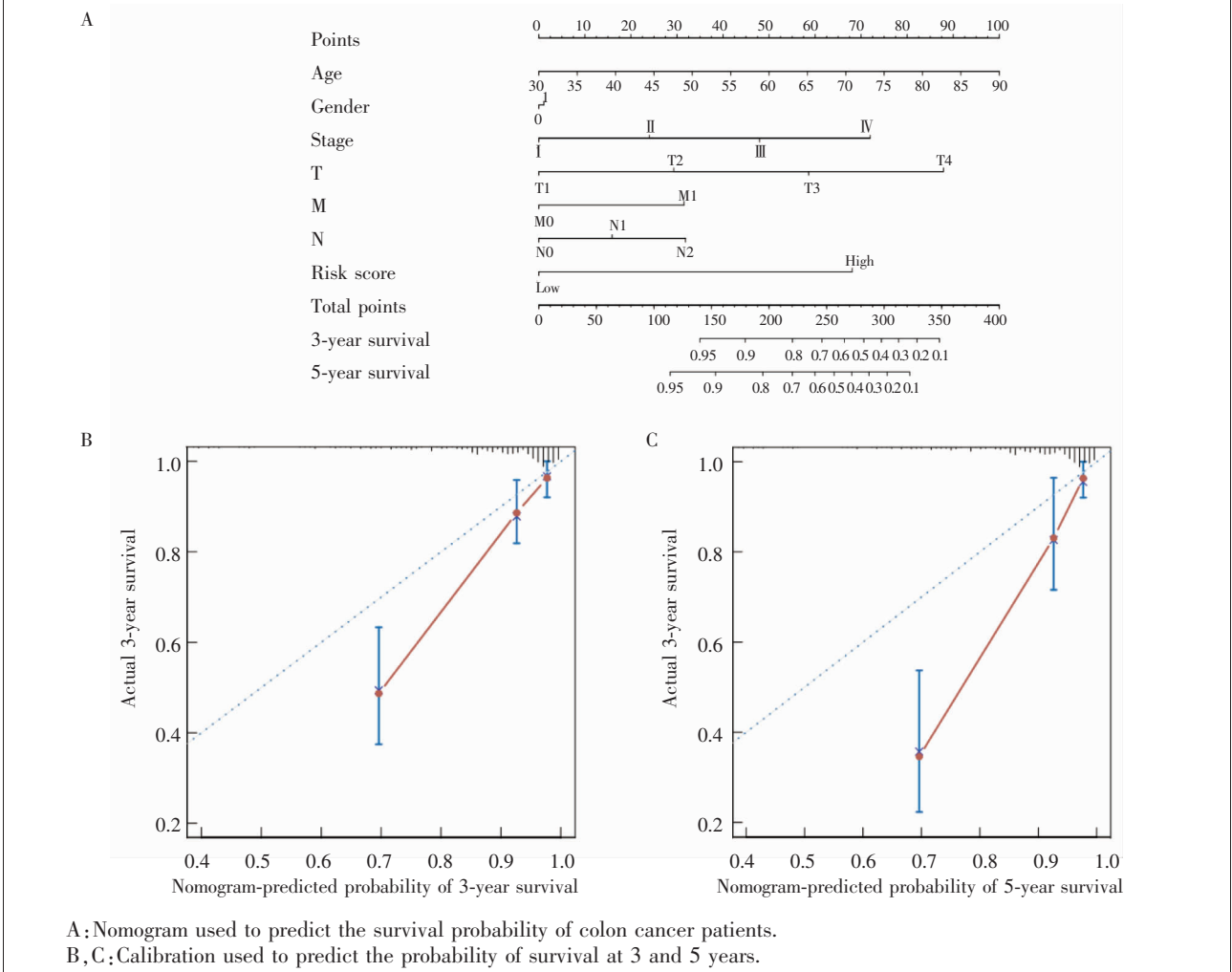


Figure 4 Prognostic factors for colon cancer



(AC105219.1、PCAT6、AP001469.3、AL391422.4、AL162274.2、AC156455.1、AC139768.1、LINC01063、AC023509.4、AC068580.3、AC073283.1 和 LINC00174) 构建的结肠癌风险评分模型在结肠癌预后中的价值。考虑到肿瘤的预后受多因素的影响,我们进一步联合其他临床特征,构建了列线图以预测结肠癌个体生存率,为结肠癌的临床决策和指导提供依据。

不可否认,TCGA 数据存在一定缺陷,主要缘于其数据主要来源于某几个国家、样本量有限、部分数据更新不及时、测序技术或质量控制存在差异,均会对数据的准确性造成影响;而且 RNA-seq 数据也不能完全代表肿瘤基因的情况;且本文用于构建模型和验证模型的数据集为同一数据集,检测本模型的预后判断效能及潜在机制,仍需通过真实世界临床资料的分析 and 基础研究以进一步验证。因此,为更好地临床应用,后续应进行更多的功能研究和体内外实验,以检验我们风险模型的准确性。

参考文献:

- [1] Yang L, Cui J, Wang Y, et al. FAM83H-AS1 is upregulated and predicts poor prognosis in colon cancer[J]. *Biomed Pharmacother*, 2019, 118: 109342.
- [2] Lv J, Guo Y, Yan L, et al. Development and validation of a five-lncRNA signature with prognostic value in colon cancer[J]. *J Cell Biochem*, 2020, 121(8-9): 3780-3793.
- [3] Aboudehen K. Regulation of mTOR signaling by long non-coding RNA[J]. *Biochim Biophys Acta Gene Regul Mech*, 2020, 1863(4): 194449.
- [4] Galamb O, Bartók BK, Kalmár A, et al. Diagnostic and prognostic potential of tissue and circulating long non-coding RNAs in colorectal tumors[J]. *World J Gastroenterol*, 2019, 25(34): 5026-5048.
- [5] Cheng Y, Geng L, Wang K, et al. Long noncoding RNA expression signatures of colon cancer based on the ceRNA network and their prognostic value[J]. *Dis Markers*, 2019, 2019: 7636757.
- [6] Li X, Li Y, Yu X, et al. Identification and validation of stemness-related lncRNA prognostic signature for breast cancer[J]. *J Transl Med*, 2020, 18(1): 331.
- [7] Cai C, Yang L, Tang Y, et al. Prediction of overall survival in gastric cancer using a nine-lncRNA[J]. *DNA Cell Biol*, 2019, 38(9): 1005-1012.
- [8] Yu C, Zhang Y. Development and validation of prognostic nomogram for young patients with gastric cancer[J]. *Ann Transl Med*, 2019, 7(22): 641.
- [9] Hou X, Wang D, Zuo J, et al. Development and validation of a prognostic nomogram for HIV/AIDS patients who underwent antiretroviral therapy: data from a china population-based cohort[J]. *EBioMedicine*, 2019, 48: 414-424.
- [10] Liu S, Tan Q, Song Y, et al. Anti-p53 autoantibody in blood as a diagnostic biomarker for colorectal cancer: a meta-analysis[J]. *Scand J Immunol*, 2020, 91(2): e12829.
- [11] Krackowska W, Jagodziński PP. The long non-coding RNA landscape of atherosclerotic plaques[J]. *Mol Diagn Ther*, 2019, 23(6): 735-749.
- [12] Xu W, Zhou G, Wang H, et al. Circulating lncRNA SNHG11 as a novel biomarker for early diagnosis and prognosis of colorectal cancer[J]. *Int J Cancer*, 2020, 146(10): 2901-2912.
- [13] Xu G, Zhu H, Xu J, et al. Long non-coding RNA POU6F2-AS2 promotes cell proliferation and drug resistance in colon cancer by regulating miR-377/BRD4 [J]. *J Cell Mol Med*, 2020, 24(7): 4136-4149.
- [14] Liu KL, Wu J, Li WK, et al. LncRNA SNHG7 is an oncogenic biomarker interacting with microRNA-193b in colon carcinogenesis[J]. *Clin Lab*, 2019, 65(11): 2199-2204.
- [15] Wei C, Liang Q, Li X, et al. Bioinformatics profiling utilized a nine immune-related long noncoding RNA signature as a prognostic target for pancreatic cancer [J]. *J Cell Biochem*, 2019, 120(9): 14916-14927.
- [16] Li X, Meng Y. Immune-related lncRNA risk signatures predict survival of IDH wild-type and MGMT promoter unmethylated glioblastoma[J]. *Biomed Res Int*, 2020, 2020: 1971284.
- [17] Yu Y, Zhang W, Li A, et al. Association of long noncoding RNA biomarkers with clinical immune subtype and prediction of immunotherapy response in patients with cancer [J]. *JAMA Netw Open*, 2020, 3(4): e202149.
- [18] Ahmad I, Valverde A, Ahmad F, et al. Long noncoding RNA in myeloid and lymphoid cell differentiation, polarization and function[J]. *Cells*, 2020, 9(2): 269.
- [19] Khadimaikar S, Kumar P, Pandi SN, et al. Immune associated lncRNAs identify novel prognostic subtypes of renal clear cell carcinoma[J]. *Mol Carcinog*, 2019, 58(4): 544-553.
- [20] Liu Z, Mi M, Li X, et al. lncRNA OSTN-AS1 may represent a novel immune-related prognostic marker for triple-negative breast cancer based on integrated analysis of a ceRNA network[J]. *Front Genet*, 2019, 10: 850.