TCGA GEO

操利超,巴 颖,卢晓萍,张核子
(深圳市核子基因科技有限公司,广东 深圳 518071)

摘要:目的 利用 TCGA 和 GEO 数据库中的基因表达数据和临床信息,挖掘结肠癌预后相关的基因,并构建和评估结肠癌预后模型。**方法** 从 GEO 数据库中下载结肠癌相关的基因表达矩阵,包括 GSE44076、GSE28000 和 GSE39582,从 TCGA 数据库中下载结肠癌相关的 mRNA 表达数据矩阵和临床信息,通过 NCBI 数据库中在线分析软件 GEO2R 对三个 GEO 数据集进行差异基因分析,利用 R 包 limma 对 TCGA 数据集进行差异基因分析,获取共同的差异表达基因。通过单因子回归、LASSO 回归和多因子回归分析构建结肠癌相关的预后模型,进一步结合临床特征构建列线图模型,综合评估预后模型的性能。**结果** 成功构建结肠癌相关的预后模型,构建的预后模型 ROC 曲线下面积在 3 年时为 0.628,4 年时为 0.678,5 年时为 0.730;Wilcoxon 检验显示,较高的风险评分与较高的 T 分期(P=0.049)、N 分期(P=0.0015)、M 分期(P=0.003)和病理分期(P=0.0019)相关;结合预后风险评分模型、年龄、性别和病理分期等级构建了列线图,模型的 C-index 从 0.63 增加至 0.74。**结论** 本次构建的结肠癌预后模型在评估结肠癌患者复发风险分层、肿瘤分期等方面具有潜在意义。

关键词:结肠癌;预后模型;生物信息学;病理分期;复发
中图分类号:R714.24 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2021.24.006
文章编号:1006-1959(2021)24-0027-06

Prognostic Model of Colon Cancer Based on TCGA and GEO Database

CAO Li-chao,BA Ying,LU Xiao-ping,ZHANG He-zi
(Shenzhen Nucleus Gene Technology Co., Ltd.,Shenzhen 518071,Guangdong,China)

Abstract:Objective To explore the prognostic genes of colon cancer by using gene expression data and clinical information in TCGA and GEO databases, and to construct and evaluate the prognostic model of colon cancer.**Methods** The gene expression matrix related to colon cancer was downloaded from the GEO database, including GSE44076, GSE28000 and GSE39582. The mRNA expression data matrix and clinical information related to colon cancer were downloaded from the TCGA database. The differential gene analysis of the three GEO data sets was carried out through the online analysis software GEO2R in the NCBI database. The differential gene analysis of the TCGA data set was carried out R package limma to obtain the common differential expression genes. Prognostic models related to colon cancer were constructed through single factor regression, LASSO regression and multi-factor regression analysis, and the line chart model was further constructed combined with clinical characteristics to comprehensively evaluate the performance of the prognosis model.**Results** The colon cancer-related prognostic model was successfully constructed. The area under the ROC curve of the prognostic model was 0.628 at 3 years, 0.678 at 4 years and 0.730 at 5 years. Wilcoxon test showed that higher risk scores were correlated with higher T staging (P=0.049), N staging (P=0.0015), M staging (P=0.003) and pathological staging (P=0.0019). Combined with the prognostic risk score model, age, gender and pathological staging level, a line chart was constructed, and the C-index of the model increased from 0.63 to 0.74.**Conclusion** The constructed colon cancer prognosis model has potential significance in evaluating the recurrence risk stratification and tumor staging of colon cancer patients.

Key words: Colorectal cancer;Prognostic model;Bioinformatics;Pathological staging;Recurrence

结肠癌(colorectal cancer,CRC)是一种常见的恶性肿瘤,是世界上第二大致死原因[1]。尽管结肠癌的诊断和治疗已经取得了很大的进展,但结肠癌患者通常会出现复发和转移,导致 5 年生存率显著下降[2]。因此,迫切需要改善结肠癌患者的诊断、治疗和预后。近些年来,分子诊断技术已广泛应用于肿瘤的治疗、预后领域[3-5]。生物信息学和机器学习技术已广泛应用于肿瘤诊断或预后分子标志物的识别,这种分子标志物类型多种多样,如 microRNAs[6]、长链非编码 RNA[7]、差异表达基因[8]、DNA 甲基化[9]等。其中,差异表达基因作为潜在的肿瘤诊断或预后标志物应用最为广泛。为得到广泛验证的结肠癌相关的差异表达基因,本文利用生物信息学方法,从多个数据集、不同的数据库中寻找共同的结肠癌相关的差异表达基因,并进一步利用机器学习的方法,从这些差异基因中挑选出结肠癌预后相关的预测因子,并建立预后风险评估模型。

1 材料与方法

1.1 数据下载和获取 通过 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo/)下载基因芯片表达数据集 GSE44076、GSE28000 和 GSE39582,每个参考数据集的正常和肿瘤样本情况见表 1。TCGA 中 mRNA 表达数据集和对应的临床信息从 UCSC Xena 平台(https://xenabrowser.net/datapages/)下载,选择队列为 GDC TCGA Colon Cancer(COAD),样本信息见表 2。

表 1 3 个 GEO 数据集的样本量情况

样本量	GSE44076	GSE28000	GSE39582
正常样本	98	34	19
肿瘤样本	98	81	566
样本总数	196	115	585

基金项目:深圳市可持续发展专项(编号:深科技创新〔2020〕180 号,专 2019N002)
作者简介:操利超(1984.4-),男,湖北黄冈人,硕士,工程师,主要从事基于多组学测序数据和临床指标的结直肠癌预测诊断和预后模型的研究
通讯作者:张核子(1972.4-),男,湖南永州人,硕士,工程师,主要从事肿瘤早筛技术和 ctDNA 精准医疗应用的研究

表 2 TCGA 数据集的样本信息[*n*(%)]

项目	肿瘤样本（*n*=432）	正常样本（*n*=39）	所有样本（*n*=471）	项目	肿瘤样本（*n*=432）	正常样本（*n*=39）	所有样本（*n*=471）
T 分期				性别			
T_1	11(2.5)	0	11(2.3)	女	200(46.3)	20(51.3)	220(46.7)
T_2	75(17.4)	5(12.8)	80(17.0)	男	232(53.7)	19(48.7)	251(53.3)
T_3	296(68.5)	28(71.8)	324(68.8)	年龄			
T_4	49(11.3)	6(15.4)	55(11.7)	≤60	136(31.5)	9(23.1)	145(30.8)
N/A	1(0.2)	0	1(0.2)	>60	296(68.5)	30(76.9)	326(69.2)
M 分期				肿瘤分期			
M_0	318(73.6)	25(64.1)	343(72.8)	Ⅰ	73(16.9)	4(10.3)	77(16.3)
M_1	60(13.9)	7(17.9)	67(14.2)	Ⅱ	166(38.4)	21(53.8)	187(39.7)
MX	47(10.9)	6(15.4)	53(11.3)	Ⅲ	122(28.2)	6(15.4)	128(27.2)
N/A	7(1.6)	1(2.6)	8(1.7)	Ⅳ	60(13.9)	7(17.9)	67(14.2)
N 分期				N/A	11(2.5)	1(2.6)	12(2.5)
N_0	254(58.8)	27(69.2)	281(59.7)				
N_1	100(23.1)	7(17.9)	107(22.7)				
N_2	78(18.1)	5(12.8)	83(17.6)				

1.2 差异基因分析和统计分析 利用 R 包分别对 3 个 GEO 数据集和 TCGA 数据集进行差异基因分析，过滤标准为 adjusted *P*-value<0.05 和差异倍数 1.5 倍（$|\log_2 FC|>0.585$），然后取交集，得到共同的上调差异基因和下调差异基因。

1.3 构建和评估预后风险评分模型 为了确定与生存相关的差异表达基因，使用 R 包 Survival 进行单变量 Cox 比例风险回归模型（*P*<0.05）。接着，使用 LASSO 回归分析进一步缩减预后因子数量，通过多因子回归分析确定每个预后因子的回归系数，建立预后风险评估模型，预测患者生存率。公式为：

$$\text{风险分数}=\sum \text{差异基因的回归系数 } x_i \times \text{归一化处理后的基因表达量 } \beta_i$$

1.4 绘制生存曲线和 ROC 曲线 根据风险评分预后模型，计算每个肿瘤样本的预后因子风险评分。利用 R 包 survivalROC 绘制 ROC 曲线，用以展示构建的风险评估模型的敏感性和特异性。在 ROC 曲线的转折点选择最佳风险评分临界值，转折点处真阳性和假阳性之间的差异最大。高于临界值的患者属于高危组，低于临界值的患者属于低危组。使用未配对 *t* 检验估计两组正态分布变量的统计显著性，并使用 R 包 Survminer 绘制两组的生存曲线。

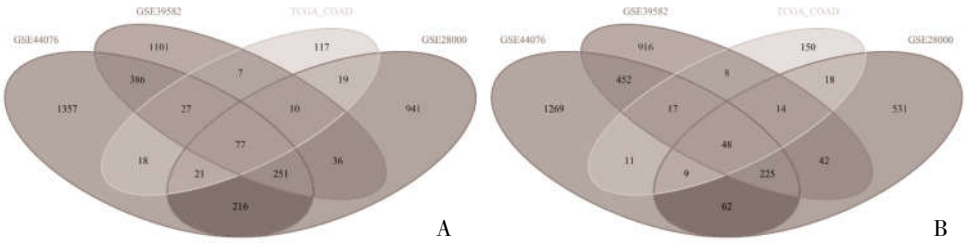
1.5 构建和验证列线图 为了提高预后模型的性能，通过整合风险评分模型和临床信息，包括年龄、性别和肿瘤分期，可视化不同患者特征的预后价值，构建列线图。该分析使用 R 软件包 rms 绘制校准曲线，以评估预测概率，并与理想预测线进行比较。此外，基于单因子回归分析的森林图说明了临床信息与 OS 之间的关系。其中，一致性指数（C-index）表明了列线图的预测准确性。

2 结果

2.1 差异基因分析 通过比较测试肿瘤样本组和正常样本组，TCGA 数据集和三个 GEO 数据集的差异基因数量分布见表 3 和图 1，可以看到共同的上调差异基因为 48 个，共同的下调差异基因数为 77 个。

表 3 差异表达基因的统计信息

项目	TCGA	GSE44076	GSE28000	GSE39582	基因交集
上调	275	2093	949	1722	48
下调	296	2353	1571	1895	77
总数	571	4446	2520	3617	125



注：A：下调基因；B：上调基因

图 1 三个 GEO 数据集和 TCGA 数据集的差异基因数量情况

2.2 结肠癌预后模型的建立 通过单因子回归分析表明,有 14 个 DEGs 与总生存期(OS)有关(P<0.05),见表 4。进一步 LASSO 回归将基因数量缩减为 10 个,见图 2。根据逐步回归模型,Akaike 信息标准(AIC)为 995.94,C 指数为 0.63。

2.3 结肠癌预后模型的性能评估 使用预后模型公式计算每个结肠癌患者的风险评分,然后根据 R 软件包 survminer 中预后因子相关风险评分的最佳临界值(cut-off 为 0.16),将结肠癌患者分为高评分组和低评分组。结果显示,随着风险得分的增加,生存时间呈现缩短的趋势,并且高危组的死亡比例比低危组高,高风险评分患者的 OS 比低评分患者预后更差,其中基因 CILP 和 C7 在低风险组表达量低,在高风险组表达量低,而其余 8 个基因趋势相反,见图 3。

2.4 预后模型的统计分析 为了进一步评估预后风险评分模型的性能,绘制 ROC 曲线和肿瘤分层分析。通过将风险评分的预后准确性作为一个连续变量进行研究,OS 预后模型的 ROC 曲线下面积(AUC)在 3 年时为 0.628,4 年时为 0.678,5 年时为 0.730,见图 4。Wilcoxon 检验表明,较高的风险评分与较高的病理分期(P=0.0019)、T 分期(P=0.049)、M 分期(P=0.003)、N 分期(P=0.0015)相关。

2.5 列线图模型的构建与验证 在列线图中,每个变量的得分可以在分数表上找到,然后通过计算总分来估计 3 年、4 年和 5 年的生存概率,见图 5A。森林图显示患者特征,包括年龄(>60)、肿瘤分期(Ⅲ和Ⅳ)和风险评分与 OS 相关(P<0.05),见图 5B。为了验证列线图的性能,绘制校准曲线,可观察到预测曲线接近理想曲线,性能良好,见图 5C~图 5E。此外,该列线图(C-index:0.74)的预测准确性高于风险评分模型(C-index:0.63)。

表 4 与结肠癌预后相关的基因信息

Gene Name	*HR*	*HR*.95L	*HR*.95H	*P*	Gene Name	*HR*	*HR*.95L	*HR*.95H	*P*
SLC4A4	0.904572	0.834725	0.980264	0.01440	DNASE1L3	0.906838	0.826916	0.994485	0.037761
ZG16	0.931458	0.880508	0.985356	0.013363	HEPACAM2	0.913524	0.851271	0.980329	0.012016
CD177	0.916153	0.84319	0.995428	0.038623	CLCA1	0.943625	0.901078	0.988181	0.013699
CLCA4	0.940267	0.885705	0.99819	0.043451	ITLN1	0.929528	0.880765	0.98099	0.007859
C7	1.091254	1.002827	1.187479	0.042823	MMP3	0.918399	0.846147	0.996821	0.041738
UGT2A3	0.931236	0.870192	0.996561	0.039442	MMP10	0.878862	0.797419	0.968622	0.009255
CILP	1.080853	1.000544	1.167609	0.048409	COMP	1.085740	1.006479	1.171243	0.033425

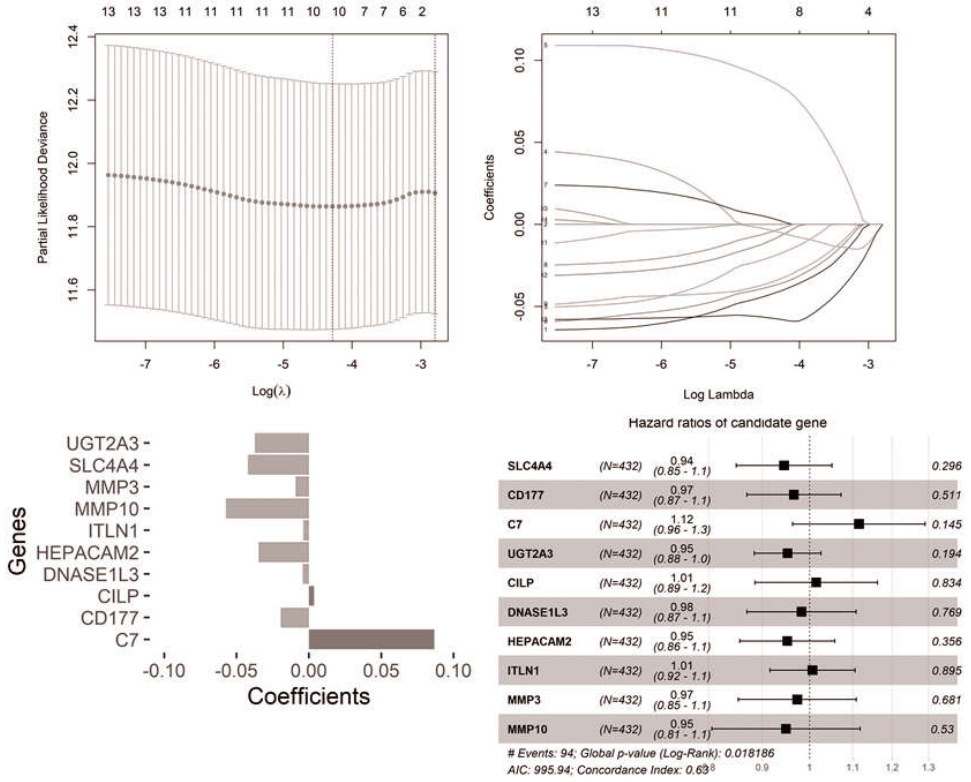


图 2 LASSO 回归分析结果

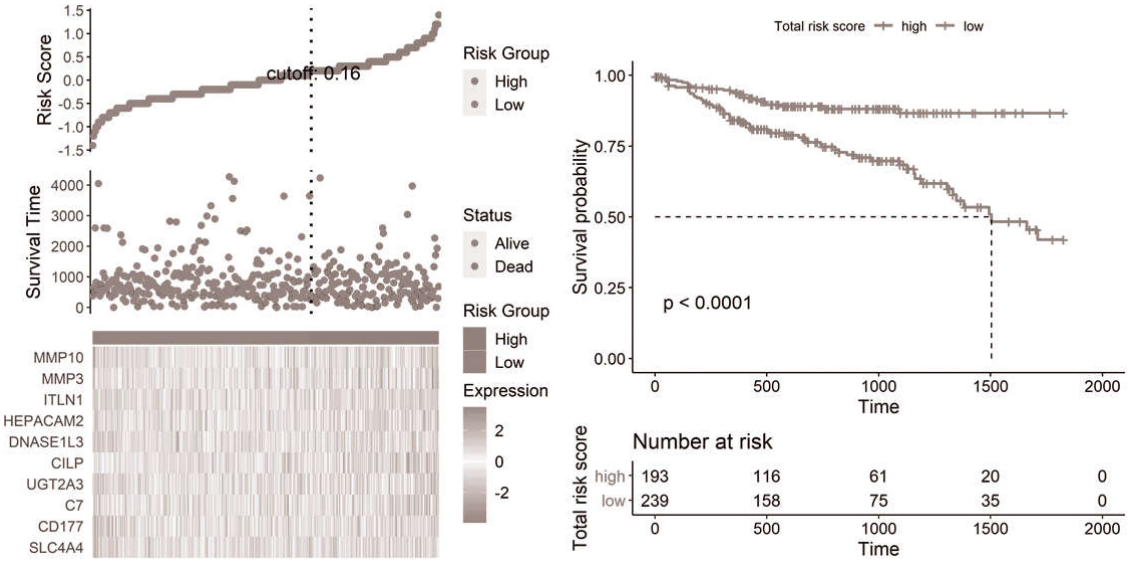


图 3 预后风险评分分组和评估

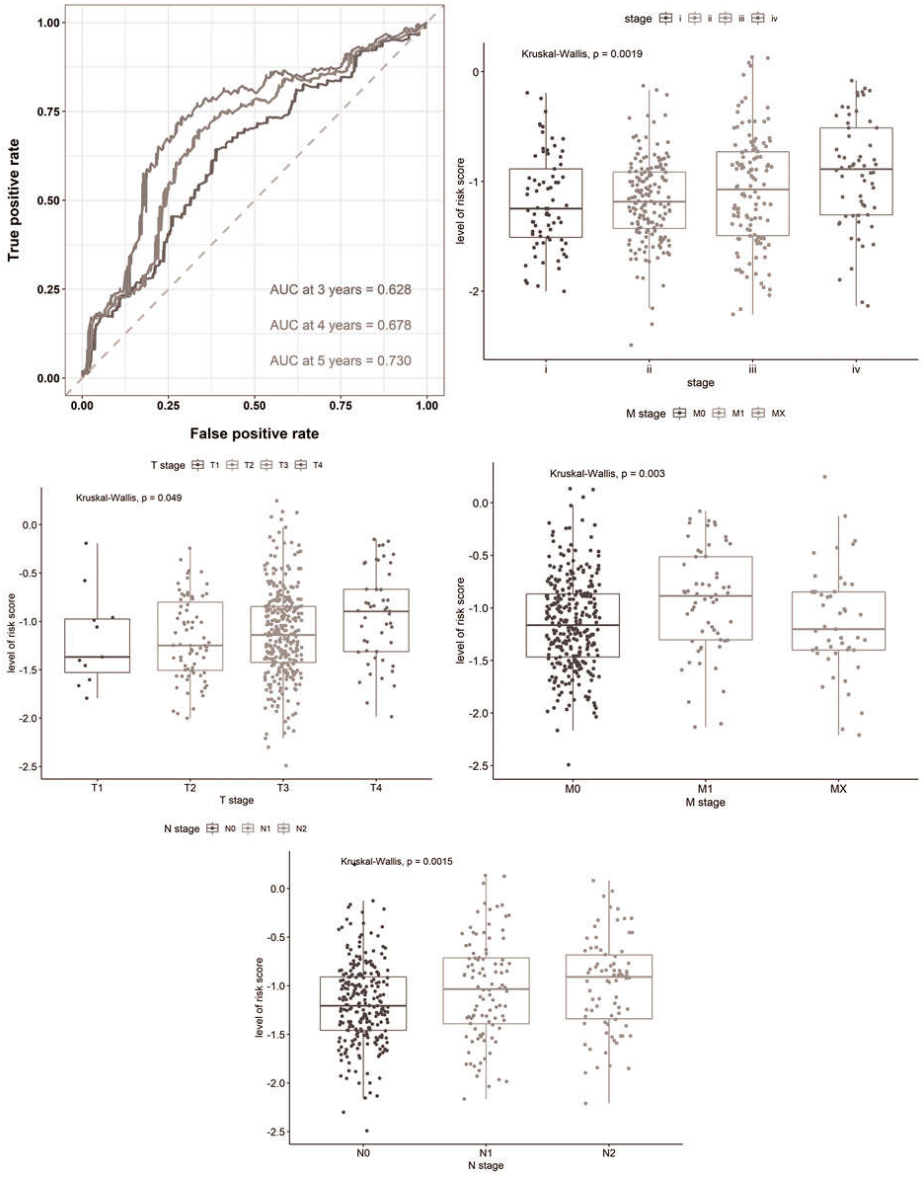


图 4 预后风险评估模型性能评价

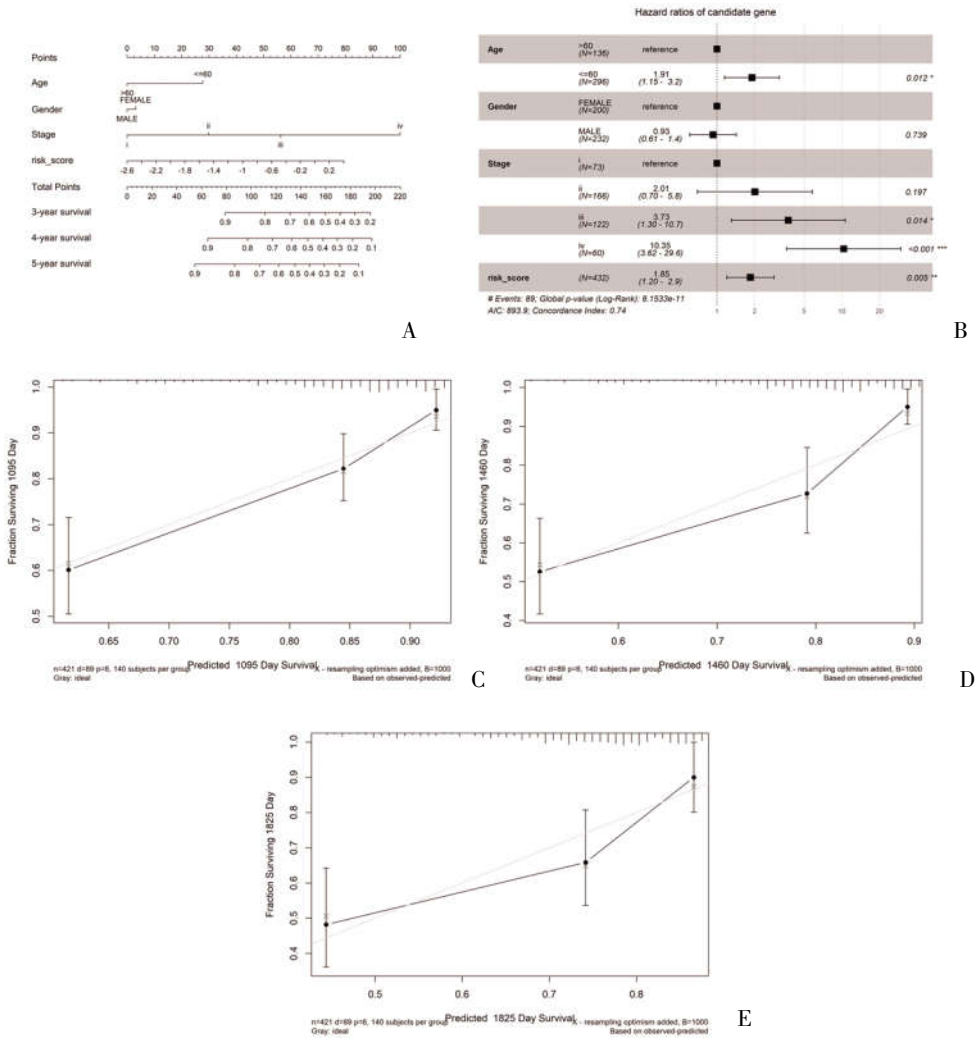


图 5 列线图模型的构建与验证

3 讨论

本研究中利用 3 个 GEO 数据集和 TCGA 数据集来挖掘共同的差异表达基因。其中，GEO 数据集是基于基因芯片平台，TCGA 数据集是基于二代测序平台，不同数据集交叉验证，使得得到的差异表达基因具有相对广泛适用性。回归分析研究表明，有 10 个差异表达基因与结肠癌预后显著相关。其中，SLC4A4 全称 Solute Carrier Family 4 Member 4，已有研究报道该基因在结肠癌患者中低表达，与结肠癌癌较差的预后相关，也与淋巴结浸润和远处转移有关[10,11]。CD177 被认为是一种干细胞因子受体，有实验证明 CD177 可作为结直肠癌患者对含贝伐单抗的抗癌治疗反应的潜在预测生物标记物[12]。另有研究[13]提出 CD177 调节胃癌中的肿瘤细胞粘附和迁移，是生存预后的因素。有报道表明[14]，CD177 的异常表达与结肠癌的发生发展相关。C7 作为一种潜在的肿瘤抑制因子，被报道与前列腺癌的免疫相关预后生物标志物[15]。此外，细胞膜上表达的 C7 是过度促炎反应的调节因子[16]，而非小细胞肺癌（NSCLC）中 C7 的低表达可能是肿瘤抑制剂，与肿瘤进展和预后相关[16]。UGT2A3 是葡萄糖醛酸转移酶（UGT）家族成员之一，UDP-葡萄糖醛酸转移酶负责外源和内源性化合物的葡萄糖醛酸化，包括药物、环境麻醉剂、类固醇、神经递质、胆汁酸和其他激素。据报道，UGT2A3 主要在与药物清除有关的组织中表达水平最高，其中肝脏是表达最多的器官，其次是胃肠道和肾脏。研究表明[17]，原发性结肠癌肝转移患者的 UGT2A3 水平明显高于无肝转移患者。DNASE1L3 与自身免疫性疾病相关，它可以消化凋亡细胞释放的微粒中的染色质，这是一种潜在的自身抗原，它的过度积累将导致身体产生自身免疫反应[18,19]。有研究表明，DNASE1L3 可能是结肠癌免疫浸润的重要生物标志物，并将为结肠癌免疫治疗靶点的选择提供理论依据[20]。HEPACAM2 是粘附基因免疫球蛋白家族的成员，已有研究表明该基因与结肠癌的发生发展有关[21]。ITLN1 可作为多种癌症的肿瘤抑制因子，如胃癌[22]、卵巢癌[23]、神经母细胞瘤[24]和结肠癌[25]。有研究通过免疫组织化学发现[25]，148 例大肠癌中 87 例（59%）ITLN1 蛋白低表达，ITLN1 表达低的结肠癌患者的 M 分级高于 ITLN1 表达高

的结肠癌患者(*P*=0.0017),且 ITLN1 表达高的患者比 ITLN1 表达低的患者预后更为良好。MMP3 和 MMP10 均属于基质金属蛋白酶(matrix metalloproteinase,MMP)家族,该家族成员参与正常生理过程中细胞外基质的分解,如胚胎发育、生殖和组织重构,以及疾病的发生发展。研究表明[26,27],MMP3 和 MMP10 均可作为结肠癌的预后相关。CILP 基因编码软骨中间层蛋白,在早期骨关节病软骨中增加,目前还未见该基因与肿瘤的相关性报道。

综上所述,本研究基于以上 10 个预后相关的基因,构建了结肠癌风险评分模型和列线图模型,结果表明,构建的模型在结肠癌预后和肿瘤分层中表现良好,具有一定的应用价值。

参考文献:

[1]Bray F,Ferlay J,Soerjomataram I,et al.Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J].CA Cancer J Clin,2018 (68):394-424.

[2]Siegel RL,Miller KD,Goding Sauer A,et al.Colorectal cancer statistics,2020[J].CA Cancer J Clin,2020(70):145-164.

[3]Wang JY,Wang CL,Wang XM,et al. Comprehensive analysis of microRNA/mRNA signature in colon adenocarcinoma[J].Eur Rev Med Pharmacol Sci,2017,21(9):2114-2129.

[4]Müller MF, Ibrahim AE, Arends MJ. Molecular pathological classification of colorectal cancer [J].Virchows Arch,2016,469(2): 125-134.

[5]Rojas V,Hirshfield KM,Ganesan S,et al. Molecular Characterization of Epithelial Ovarian Cancer: Implications for Diagnosis and Treatment[J].Int J Mol Sci,2016,17(12):2113.

[6]Conti I,Simioni C,Varano G,et al.MicroRNAs Patterns as Potential Tools for Diagnostic and Prognostic Follow-Up in Cancer Survivorship[J].Cells,2021,10(8):2069.

[7]Chandra Gupta S,Nandan Tripathi Y.Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets[J].Int J Cancer,2017,140(9):1955-1967.

[8]Chen K,Zhu P,Liao Y,et al.An Apoptotic Gene Signature for the Prognosis of Hepatocellular Carcinoma [J].Onco Targets Ther,2021(14):1589-1604.

[9]Constancio V,Nunes SP,Henrique R,et al.DNA Methylation-Based Testing in Liquid Biopsies as Detection and Prognostic Biomarkers for the Four Major Cancer Types [J].Cells,2020,9(3): 624.

[10]Yang H,Liu H,Lin HC,et al.Association of a novel seven-gene expression signature with the disease prognosis in colon cancer patients[J].Aging (Albany NY),2019,11(19):8710-8727.

[11]Chen X,Chen J,Feng Y,et al.Prognostic Value of SLC4A4 and its Correlation with Immune Infiltration in Colon Adenocarcinoma[J].Med Sci Monit,2020(26):e925016.

[12]Schiffmann LM,Fritsch M,Gebauer F,et al.Tumour-infiltrating neutrophils counteract anti-VEGF therapy in metastatic colorectal cancer[J].Br J Cancer,2019,120(1):69-78.

[13]Toyoda T,Tsukamoto T,Yamamoto M,et al.Gene expression analysis of a Helicobacter pylori-infected and high-salt diet-treated mouse gastric tumor model: identification of CD177 as a novel prognostic factor in patients with gastric cancer[J].BMC Gastroenterol,2013(13):122.

[14]Shangkuan WC,Lin HC,Chang YT,et al.Risk analysis of colorectal cancer incidence by gene expression analysis[J].Peer J, 2017(5):e3003.

[15]Chen Z,Yan X,Du GW,et al.Complement C7 (C7),a Potential Tumor Suppressor, Is an Immune-Related Prognostic Biomarker in Prostate Cancer (PC)[J].Front Oncol,2020(10):1532.

[16]Ying L,Zhang F,Pan X,et al.Complement component 7 (C7), a potential tumor suppressor, is correlated with tumor progression and prognosis[J].Oncotarget,2016,7(52):86536-86546.

[17]Wang S,Zhang C,Zhang Z,et al.Transcriptome analysis in primary colorectal cancer tissues from patients with and without liver metastases using next-generation sequencing [J].Cancer Med,2017,6(8):1976-1987.

[18]Sisirak V,Sally B,D´Agati V,et al.Digestion of Chromatin in Apoptotic Cell Microparticles Prevents Autoimmunity [J].Cell, 2016,166(1):88-101.

[19]Weisenburger T,von Neubeck B,Schneider A,et al.Epistatic Interactions Between Mutations of Deoxyribonuclease 1-Like 3 and the Inhibitory Fc Gamma Receptor IIB Result in Very Early and Massive Autoantibodies Against Double-Stranded DNA [J].Front Immunol,2018(9):1551.

[20]Liu J,Yi J,Zhang Z,et al.Deoxyribonuclease 1-like 3 may be a potential prognostic biomarker associated with immune infiltration in colon cancer[J].Aging (Albany NY),2021,13(12):16513-16526.

[21]Wu Z,Liu Z,Ge W,et al.Analysis of potential genes and pathways associated with the colorectal normal mucosa-adenoma-carcinoma sequence[J].Cancer Med,2018,7(6):2555-2566.

[22]Li D,Zhao X,Xiao Y,et al.Intelectin 1 suppresses tumor progression and is associated with improved survival in gastric cancer[J].Oncotarget,2015,6(18):16168-16182.

[23]Au-Yeung CL,Yeung TL,Achreja A,et al.ITLN1 modulates invasive potential and metabolic reprogramming of ovarian cancer cells in omental microenvironment[J].Nat Commun,2020,11(1):3546.

[24]Li D,Mei H,Pu J,et al.Intelectin 1 suppresses the growth, invasion and metastasis of neuroblastoma cells through up-regulation of N-myc downstream regulated gene 2 [J].Mol Cancer, 2015(14):47.

[25]Kawashima K,Maeda K,Saigo C,et al.Adiponectin and Intelectin-1: Important Adipokine Players in Obesity-Related Colorectal Carcinogenesis[J].Int J Mol Sci,2017,18(4):866.

[26]Klupp F,Neumann L,Kahlert C,et al.Serum MMP7, MMP10 and MMP12 level as negative prognostic markers in colon cancer patients[J].BMC Cancer,2016(16):494.

[27]Zeng C,Chen Y.HTR1D, TIMP1, SERPINE1, MMP3 and CNR2 affect the survival of patients with colon adenocarcinoma [J].Oncol Lett,2019,18(3):2448-2454.

收稿日期:2021-09-23;修回日期:2021-10-03
编辑/成森