

DOI:10.3872/j.issn.1007-385X.2026.08.008

· 转化医学 ·

基于多区域甲基化特征的年轻乳腺癌预后模型构建与验证

孙礼峰, 韩雷, 陈桂冬, 程亚楠, 于津浦(天津医科大学肿瘤医院肿瘤分子诊断中心, 恶性肿瘤国家临床医学研究中心, 天津市恶性肿瘤临床医学研究中心, 药物成药性评价与系统转化全国重点实验室, 乳腺癌防治教育部重点实验室, 天津市肿瘤免疫与生物治疗重点实验室, 天津 300060)

[摘要] **目的:** 探讨年轻乳腺癌独特的DNA甲基化特征, 筛选与预后相关的甲基化位点, 构建并验证预后预测模型, 并通过多组学整合及关键蛋白验证评估其临床价值。**方法:** 收集癌症基因组图谱(TCGA)数据库中的乳腺癌样本DNA甲基化及临床数据, 筛选年轻乳腺癌特有的甲基化位点。采用随机森林、LASSO回归及单因素和多因素Cox比例风险回归分析构建预后模型。使用受试者工作特征(ROC)曲线、校准曲线和一致性指数(C-index)评估模型性能, 并比较高、低风险组间的多组学特征差异, 在独立队列中通过免疫组化法检测模型关键基因LAMA5的蛋白表达, 分析其与预后的关系。**结果:** 年轻乳腺癌相较于老年乳腺癌, 在启动子区域(尤其是CpG岛)呈现显著低甲基化特征。通过多步骤差异分析获得3 489个年轻乳腺癌特有的甲基化位点, 经Cox比例风险回归筛选出12个与总生存期(OS)独立相关的位点, 其中启动子区、基因体区和基因间区位点各4个。基于筛选出的12个甲基化位点构建的预测模型, 在TCGA训练集中显示出良好的区分度, 高风险组总生存期显著更差($P = 0.0025$), 5、8和10年OS预测曲线下面积(AUC)达0.88~0.94。多组学分析显示, 高风险组丝裂原活化蛋白激酶(MAPK)信号通路激活, TP53基因突变频率更高, 雌激素受体 α 蛋白表达较低。将TP53突变状态及雌激素受体1(ESR1)mRNA表达水平等关键多组学指标纳入模型, 构建综合列线图, 其预测效能进一步提升。组织芯片验证显示, 模型关键基因LAMA5蛋白表达阴性患者的OS更差($P = 0.046$), 且富集于三阴性乳腺癌, 与模型高风险组的临床病理特征一致。**结论:** 整合启动子和非启动子区域信息的基于12个甲基化位点的模型可用于年轻乳腺癌患者的预后分层。联合年龄、TP53突变及ESR1 mRNA表达水平构建的综合列线图显示出较单一甲基化模型更高的预测效能。

[关键词] DNA甲基化; 年轻乳腺癌; 表观遗传学; 预后标志物; 预测模型

[中图分类号] R737.9; R730.7 **[文献标识码]** A **[文章编号]** 1007-385X(2026)08-0876-10

Development and validation of a multi-region DNA methylation signature-based prognostic model for breast cancer in young women

Sun Lifeng, Han Lei, Chen Guidong, Cheng Yanan, Yu Jinpu (Cancer Molecular Diagnostics Core, Tianjin Medical University Cancer Institute & Hospital, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, State Key Laboratory of Druggability Evaluation and Systematic Translational Medicine, Key Laboratory of Breast Cancer Prevention and Therapy, Tianjin Medical University, Ministry of Education, Tianjin Key Laboratory of Cancer Immunology and Biotherapy, Tianjin 300060, China)

[Abstract] **Objective:** This study aims to systematically characterize the unique DNA methylation landscape of breast cancer in young women (BCYW), identify prognosis-associated methylation loci, develop and externally validate a robust prognostic prediction model, and further elucidate its clinical utility *via* multi-omics integration analysis and protein-level validation of a pivotal candidate gene. **Methods:** DNA methylation and clinical data were obtained from The Cancer Genome Atlas (TCGA). BCYW-specific methylation loci were identified. A prognostic prediction model was developed using random forest, LASSO regression, and univariate and multivariate Cox regression analyses. The predictive performance of the constructed model was comprehensively evaluated using receiver operating characteristic (ROC) curves, calibration plots, and the concordance index (C-index). Multi-omics characteristics were compared between high-risk and low-risk subgroups. In an independent cohort, immunohistochemistry was performed to assess the association between LAMA5 protein expression and clinical outcomes. **Results:** Compared with their older counterparts with breast cancer, young patients exhibited prominent hypomethylation in promoter regions, which was particularly pronounced within CpG islands. Through a multi-step differential methylation analysis, a total of 3 489 BCYW-specific methylation loci were identified. Subsequent univariate and

[基金项目] 天津市医学重点学科建设项目(TJYXZDXK-3-003A)

[作者简介] 孙礼峰(1998—), 男, 硕士生

[通信作者] 于津浦(扫码获取作者通信方式)



multivariate Cox regression analyses further screened out 12 loci that were independently correlated with overall survival (OS), among which 4 were located in promoter regions, 4 in gene body regions, and the remaining 4 in intergenic regions. The prognostic prediction model constructed based on these 12 methylation loci exhibited satisfactory discriminative performance in the TCGA training cohort, where patients assigned to the high-risk subgroup presented significantly worse overall survival outcomes ($P = 0.0025$). The areas under the curve (AUCs) for 5-, 8-, and 10-year OS prediction ranged from 0.88 to 0.94. Multi-omics analysis revealed that the high-risk subgroup exhibited the activation of the mitogen-activated protein kinase (MAPK) signaling pathway, higher frequency of TP53 mutations, and lower expression of estrogen receptor alpha protein. Incorporating key multi-omics indicators, including TP53 mutation status and estrogen receptor 1 (ESR1) mRNA expression level, into the comprehensive nomogram further significantly improved the model's predictive performance. Tissue microarray-based immunohistochemical validation demonstrated that young breast cancer patients with negative LAMA5 protein expression exhibited significantly shorter OS ($P = 0.046$), and this phenotype was enriched in triple-negative breast cancer, consistent with the clinicopathological features of the high-risk group defined by the methylation model.

Conclusion: The 12-methylation-site model integrating promoter and non-promoter region information shows potential for prognostic stratification in young patients with breast cancer. The multi-omics nomogram constructed by combining age, TP53 mutations and ESR1 mRNA expression level demonstrated superior predictive performance over the methylation-only model.

[Key words] DNA methylation; breast cancer in young women (BCYW); epigenetics; prognostic biomarker; prediction model

[Chin J Cancer Biother, 2026, 33(8): 876-885. DOI:10.3872/j.issn.1007-385X.2026.08.008]

乳腺癌是全球女性最常见的恶性肿瘤,近年来其发病呈年轻化趋势^[1-3]。在亚洲国家,年轻乳腺癌(发病年龄 ≤ 40 岁)占比高达10%~20%^[4-5]。与老年患者相比,年轻乳腺癌患者预后更差、复发风险更高,其分子亚型多以人类表皮生长因子受体2(human epidermal growth factor receptor 2, HER2)阳性和三阴性乳腺癌(triple-negative breast cancer, TNBC)为主^[5-6]。现有研究从基因组^[7-8]、转录组^[9]及蛋白质组^[10]层面揭示了年轻乳腺癌的部分分子特征,如GATA结合蛋白3(GATA binding protein 3, GATA3)和TP53基因突变频率较高,磷脂酰肌醇-3-激酶催化亚基 α (phosphoinositide-3-kinase catalytic subunit alpha, PIK3CA)基因突变频率较低,以及雌激素受体1(estrogen receptor 1, ESR1)表达下调等。DNA甲基化作为稳定的表观遗传修饰,与乳腺癌的分子异质性及预后密切相关^[11-12]。现有研究多聚焦于启动子区甲基化,但近年来发现基因体区甲基化同样具有重要生物学意义,可通过影响转录延伸、可变剪接及染色质结构参与肿瘤发生发展^[13-15]。然而,针对年轻乳腺癌这一特殊人群,同时关注启动子区与非启动子区甲基化特征及其预后价值的系统性研究仍有限。此外,肿瘤发生发展是多层面分子事件协同作用的结果,单一组学指标难以全面反映疾病本质,近期研究提示多组学整合可能是提升预后预测效能的有效策略^[11]。因此,本研究同时聚焦启动子区与非启动子区域,系统分析年轻乳腺癌特有的DNA甲基化模式,筛选关键的预后相关甲基化位点来构建预后预测模型,并结合多组学数据解析不同风险组的分子特征,探究整合TP53突变状态和ESR1 mRNA表达水平等多组学指标对上述预测模型的增效作用。最后,通过免疫组织化学方法对模型关键基因层粘

连蛋白 $\alpha 5$ (laminin subunit alpha 5, LAMA5)进行蛋白水平的独立验证,为12个甲基化位点模型的生物学合理性提供实验支持,并且为年轻乳腺癌的精准预后评估提供新的表观遗传学依据。

1 材料与方法

1.1 数据来源、分组与预处理

从癌症基因组图谱(The Cancer Genome Atlas, TCGA)数据库下载784例乳腺癌Illumina HumanMethylation450 BeadChip (450K) DNA甲基化数据,并从cBioPortal癌症基因组学数据库(cBioPortal for Cancer Genomics)平台获取对应的临床信息,样本匹配后共纳入773例乳腺癌样本。根据诊断年龄,筛选出年轻组(≤ 40 岁, $n = 73$)和老年组(≥ 60 岁, $n = 352$)用于后续差异甲基化分析,其余样本不纳入年轻组与老年组的差异甲基化比较。使用芯片分析甲基化流程(Chip Analysis Methylation Pipeline, ChAMP)包对甲基化数据进行质量控制、缺失值填补及标准化处理。差异甲基化位点的筛选标准^[12]如下:(1)癌组织与正常组织比较,校正后 $P < 0.05$ 且 $|\Delta\beta| > 0.20$ ($\Delta\beta$ 为差异 β 值);(2)年轻乳腺癌与老年乳腺癌组织比较,校正后 $P < 0.05$ 且 $|\Delta\beta| > 0.10$ 。为扩大年轻乳腺癌队列的样本量,本研究进一步从基因表达综合数据库(Gene Expression Omnibus, GEO)下载GSE72251、GSE72308、GSE75067、GSE78754及GSE37754共5个数据集的甲基化及临床数据,经去重及批次效应校正后合并,筛选其中年龄 ≤ 40 岁的年轻乳腺癌样本($n = 96$),与TCGA年轻乳腺癌队列($n = 73$)合并,形成扩展年轻乳腺癌队列($n = 169$),并基于此队列进行亚组生存分析。

1.2 生物信息学分析

使用limma包对TCGA年轻乳腺癌队列高风险组与低风险组进行转录组差异表达分析,筛选阈值为 $P < 0.05$,且倍数变化(fold change, FC)满足 $|\log_2FC| > 0.50$ 的差异表达基因并进行基因本体(Gene Ontology, GO)和京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)富集分析。同时,基于所有差异基因的表达量排序,利用基因集富集分析(gene set enrichment analysis, GSEA)进行富集分析。使用survival包和survminer包进行生存分析并绘制Kaplan-Meier生存曲线。基因拷贝数变异(copy number variation, CNV)分析通过GenePattern平台的基因组鉴定显著靶标(Genomic Identification of Significant Targets in Cancer, GISTIC)2.0模块完成。

1.3 甲基化预后模型的构建与评价

针对年轻乳腺癌特有的差异甲基化位点,首先采用单因素Cox比例风险回归分析进行初步筛选($P < 0.05$),随后应用LASSO回归(10折交叉验证)进一步降维筛选候选位点。将筛选后的位点纳入多因素Cox回归分析($P < 0.10$),确定独立预后相关甲基化位点。根据多因素Cox回归模型中各位点的回归系数构建风险评分公式,并以风险评分的中位数为界,将患者分为高风险组和低风险组。采用时间依赖的受试者工作特征(receiver operating characteristic, ROC)曲线、校准曲线(Bootstrap自助法,重复1 000次)及一致性指数(concordance index, C-index)评估模型的预测性能。将年龄、甲基化风险评分、TP53突变状态及ESR1 mRNA表达水平纳入列线图(nomogram),并采用校准曲线、时间依赖性ROC曲线和C-index进行评价。

1.4 高、低风险组多组学特征分析

TCGA年轻乳腺癌队列包括73例样本,其中5例样本因突变信息缺失而未纳入分析,共68例具有体细胞突变谱数据的样本纳入多组学分析。使用突变注释格式工具(Mutation Annotation Format Tools, maftools)包比较高、低风险组之间的基因突变谱差异。转录组及蛋白质组的差异表达分析采用limma包完成,筛选阈值为 $P < 0.05$ 且 $|\log_2FC| > 0.50$ 。

1.5 LAMA5蛋白表达的免疫组织化学验证

临床标本来源:本研究所用乳腺癌组织芯片购自上海芯超生物科技有限公司(产品编号:HBreD129Su01)。该芯片共包含129例乳腺癌组织样本,样本收集经伦理委员会审查批准(伦理批件号:SHYJS-CP-1804031),并取得患者知情同意。该组织芯片样本中63例未计入统计(49例因组织脱片未完成染色步骤,14例样本因缺乏足量肿瘤组织而

无法判读),最终共66例样本纳入临床病理特征及生存分析。根据组织芯片提供的临床病理信息,66例样本有明确的分子分型信息,Luminal型34例,HER2过表达型18例,三阴性乳腺癌14例。

免疫组织化学染色:石蜡切片脱蜡至水,经柠檬酸盐缓冲液(pH 6.0)微波抗原修复。3% H_2O_2 灭活内源性过氧化物酶,5% BSA封闭。滴加兔抗人LAMA5多克隆抗体(OriGene,货号TA322581S,工作浓度1:10),4℃处理过夜。次日滴加HRP标记的山羊抗兔IgG聚合物,室温孵育。DAB显色,显微镜下控制显色时间,苏木素复染、分化、返蓝后,脱水、透明、封片。

结果判读:所有切片由2位病理科医师独立阅片并评分。染色强度分级记录为:0分(无着色)、1分(弱阳性)、2分(中等阳性)、3分(强阳性);阳性细胞比例以阳性肿瘤细胞占有肿瘤细胞的百分比表示。阳性细胞比例 $> 0\%$ 且染色强度 ≥ 1 分的样本为LAMA5阳性组;阳性细胞比例为0%且染色强度为0分的样本为LAMA5阴性组。若2位初判医师意见不一致,由第3位病理科医师复核。采用Cohen's kappa一致性检验评估2位初判医师间的一致性,2者均判为阳性者21例,均判为阴性者39例,观察一致率为90.91%(60/66),kappa值为0.804(95% CI: 0.66~0.95),提示判读结果具有高度一致性。

1.6 统计学处理

所有统计分析均在R软件(版本4.3.3)中完成。对连续变量进行正态性检验,符合正态分布的连续变量以 $\bar{x} \pm s$ 表示,2组间比较采用独立样本 t 检验;不符合正态分布的连续变量2组间比较采用Wilcoxon秩和检验,多组间比较采用Kruskal-Wallis检验。分类变量组间比较采用 χ^2 检验。生存曲线采用Kaplan-Meier法绘制并以log-rank检验比较。多重比较采用Benjamini-Hochberg法进行假发现率(false discovery rate, FDR)校正。以 $P < 0.05$ 或 $P < 0.01$ 表示差异具有统计学意义。

2 结 果

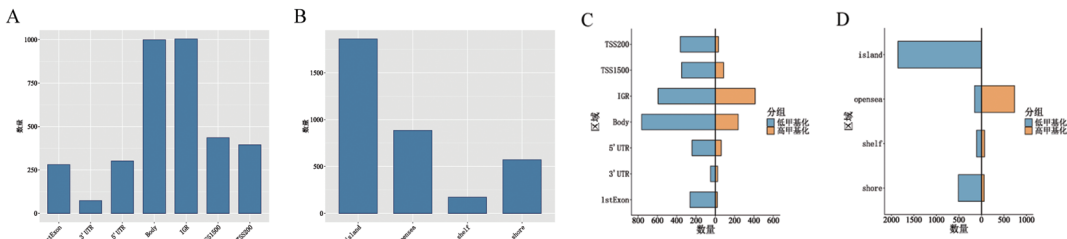
2.1 年轻乳腺癌特异性DNA甲基化图谱及其与基因组变异的关联

为探究年轻乳腺癌特有的DNA甲基化改变,本研究首先比较年轻与老年乳腺癌患者的全基因组甲基化水平,结果显示2组间总体差异无统计学意义($P = 0.46$),但年轻乳腺癌在启动子区域呈现低甲基化趋势($P = 0.076$),尤其在CpG岛区域更为显著($P = 1.3 \times 10^{-5}$),而在基因体和基因间区(intergenic region, IGR)未见显著差异($P = 0.58$)。

为系统鉴定年轻乳腺癌特异性DNA甲基化位点,通过多步骤差异分析策略:先将年轻及老年乳腺

癌分别与正常组织比较,剔除2组变化趋势一致的位点,再直接比较年轻与老年乳腺癌样本,最终获得3 489个特异性差异甲基化位点。从基因组定位来看,这些位点主要分布于启动子区(40.5%)、基因体区(28.6%)和基因间区(28.8%),3者合计占比近

98.0%(图1A),是年轻乳腺癌特异性甲基化改变的核心构成。进一步分析发现,启动子区(低甲基化占85.1%)和基因体区(低甲基化占72.0%)均以低甲基化占主导,而基因间区低甲基化位点亦多于高甲基化位点(低甲基化占59.0%)(图1B~D)。

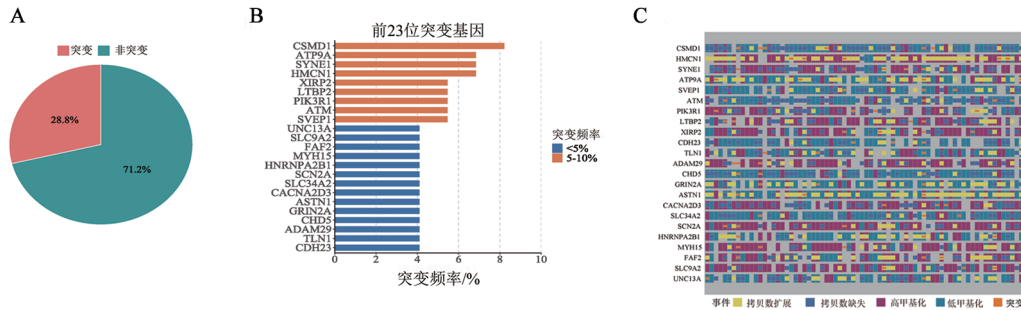


A:不同基因区域位点总数;B:CpG岛相关区域位点总数;C:不同基因区域高/低甲基化位点分布;D:CpG岛相关区域高/低甲基化位点分布。

图1 年轻乳腺癌特异性差异甲基化位点的基因组分布及甲基化方向

在启动子区或基因体区存在差异甲基化位点的1 242个基因中,358个基因发生体细胞突变,以CUB和Sushi多结构域蛋白1(CUB and Sushi multiple domains 1, CSMD1)突变频率最高(图2A、B)。关键预后相关基因(CSMD1、PIK3R1、ATM)均检测到反复出现的启动子或基因体差异甲基化位点,其中CSMD1和共济失调毛细血管扩张突变(ataxia telangiectasia

mutated, ATM)主要表现为启动子区低甲基化,常伴随拷贝数缺失及突变;而磷脂酰肌醇-3-激酶调节亚基1(phosphoinositide-3-kinase regulatory subunit 1, PIK3R1)则以高甲基化为主(图2C)。上述结果提示,启动子区和基因体区的DNA甲基化改变与基因组变异在特定基因上存在协同关系,共同参与年轻乳腺癌的发生发展。



A:差异甲基化基因的突变频率;B:Top 23基因突变频率/%;C:Top 23基因的甲基化、拷贝数与突变关联。

图2 年轻乳腺癌差异甲基化相关基因的基因组变异特征

2.2 预后相关甲基化位点筛选及12位点风险评分模型的构建与评价

基于上述3 489个年轻乳腺癌特异性差异甲基化位点,本研究进一步筛选与患者预后相关的关键位点。通过单因素Cox回归分析($P < 0.05$)初步筛选出954个与总生存期显著相关的候选位点,随后采用LASSO回归进行10折交叉验证,以 $\lambda_{\min}$ 为阈值进一步降维,筛选出46个最具代表性的预后相关位点(图3A、B)。最后将上述46个位点纳入多因素Cox回归分析($P < 0.10$),通过逐步回归最终确定12个独立的预后相关甲基化位点(图3C)。在最终获得的12

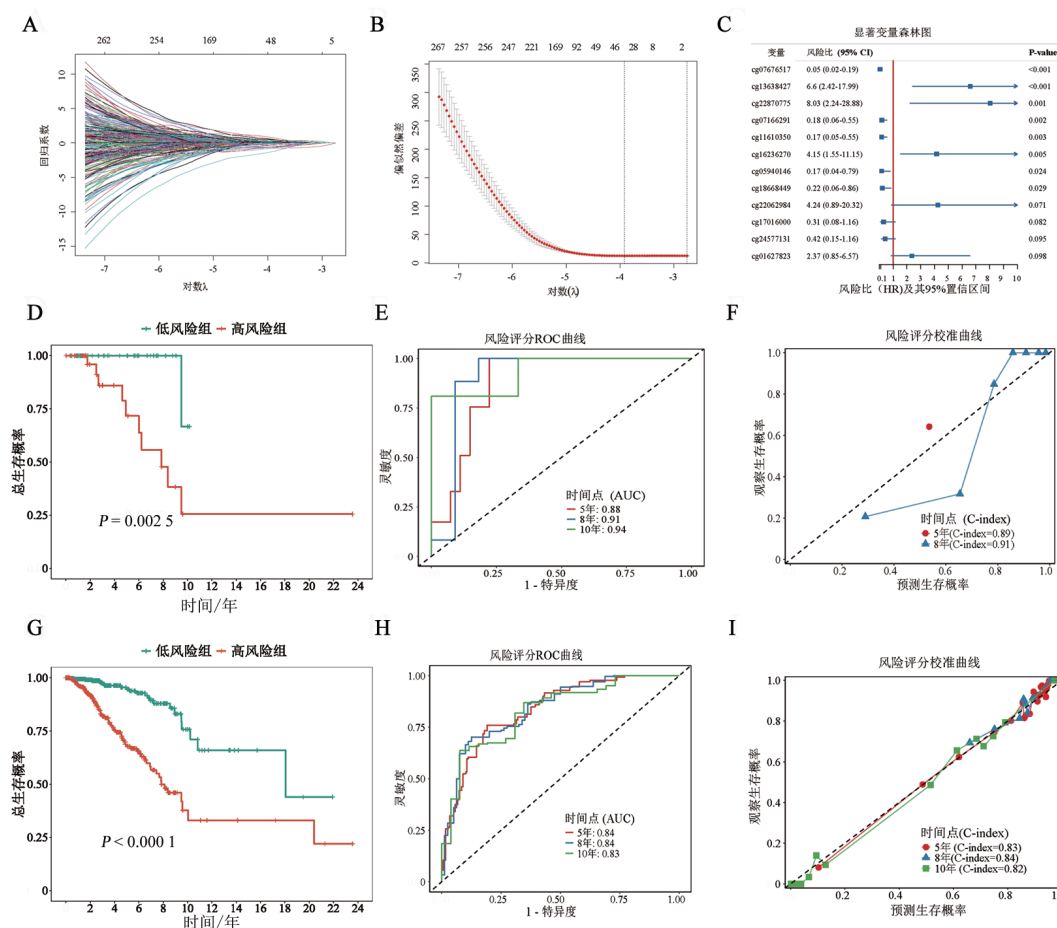
个位点中,启动子区位点4个、基因体区位点4个、基因间区位点4个。随机森林变量重要性分析显示,上述12个位点多数位列重要性排名前20,提示非启动子区甲基化在年轻乳腺癌的预后预测中亦具有重要价值。

首先仅利用启动子区域的4个甲基化位点构建初步模型。在TCGA年轻乳腺癌队列中,高风险组总生存期(overall survival, OS)显著低于低风险组($P < 0.0001$);在TCGA全体乳腺癌患者中,高风险组OS仍显著较差($P < 0.0001$),但模型对生存时间的预测效能相对有限,5年、8年和10年OS的曲线下面积

(area under the curve, AUC)介于0.69~0.74之间。

为进一步提升预测性能,整合启动子区和非启动子区域的12个甲基化位点,基于多因素Cox回归系数构建最终风险评分: $\text{Risk score} = \sum_{i=1}^{12} \beta_i \times \text{Coef}_i$, 其中 β_i 为第*i*个位点的甲基化水平(β 值), Coef_i 为第*i*个位点的多因素Cox回归系数。在TCGA年轻乳腺癌队列($n=73$)中,按风险评分中位数划分的高风险组OS显著低于低风险组($P=0.0025$,

图3D),模型预测AUC达0.88~0.94,校准曲线显示预测与实际观察值具有良好一致性(图3E~F)。将该模型扩展至TCGA全体乳腺癌患者($n=773$),高风险组的OS显著低于低风险组(均 $P<0.0001$,图3G),模型预测AUC维持在0.83~0.84的高水平,校准曲线及一致性指数均提示良好的预测效能(图3H~I)。



A: LASSO回归参数路径图; B: 10折交叉验证误差曲线; C: 12个关键位点多因素Cox回归森林图; D: 年轻乳腺癌高/低风险组OS曲线; E: 年轻乳腺癌时间依赖性ROC曲线; F: 年轻乳腺癌校准曲线; G: 全体乳腺癌高/低风险组OS曲线; H: 全体乳腺癌时间依赖性ROC曲线; I: 全体乳腺癌校准曲线。

图3 12位点甲基化预后模型的构建及预测效能评价

2.3 12位点甲基化风险评分与临床病理特征及亚组预后的关系

为评估整合非启动子区域甲基化位点的临床价值,本研究比较了仅启动子模型与完整12位点模型在TCGA年轻乳腺癌队列中的风险分层与临床病理特征的关联。仅启动子4个甲基化位点模型划分的高、低风险组在雌激素受体(estrogen receptor, ER)状态、孕激素受体(progesterone receptor, PR)状态、基于50个基因表达的预测分析(Prediction Analysis of Microarray 50, PAM50)分子分型及肿瘤-淋巴结-转

移(tumor-node-metastasis, TNM)分期方面均无显著差异,提示该模型虽能区分生存结局,但未能富集侵袭性临床表型。

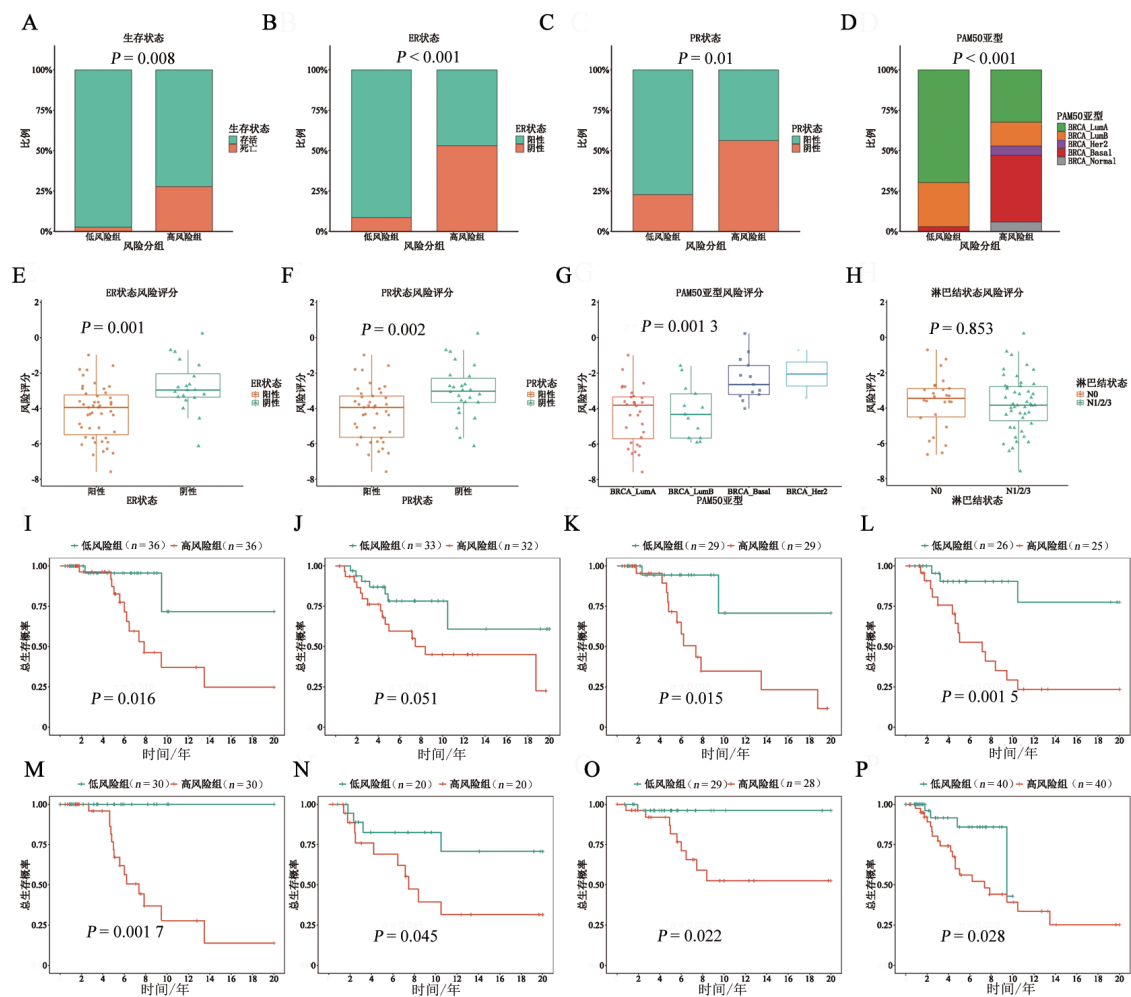
相比之下,完整12位点模型划分的高风险组表现出显著的不良临床病理特征:高风险组死亡比例更高($P=0.008$),ER阴性($P<0.001$)及PR阴性($P=0.01$)比例升高,基底样型及HER2过表达亚型占比更高($P<0.001$)(图4A~D)。风险评分与ER、PR及PAM50分型显著相关,而与TNM分期无关。进一步分析显示,ER阳性、PR阳性及Luminal型患者的风险

评分显著降低(图4E~H)。在ER阳性、PR阳性、PR阴性、Luminal型、非Luminal型、N0及非N0亚组中, 高风险组的总生存期显著低于低风险组; 其中ER阴性亚组差异无统计学意义($P=0.051$)(图4I~P)。高低风险组详细样本量及风险比(hazard ratio, HR)详见表1。这些结果表明, 该甲基化风险评分在不同临床亚组中均具有一定的预后分层能力。

2.4 年轻乳腺癌12位点甲基化模型高、低风险组的多组学特征

对TCGA年轻乳腺癌队列高低风险组进行多组学分析, 共鉴定出639个差异表达基因, 其中高风险组上调364个, 下调275个。GO富集分析显示, 上述差异表达基因显著富集于钾离子跨膜转运与膜电位调节、激素代谢/分泌/转运及信号释放等生物学过程(图5A)。KEGG通路富集分析显示, 差异基因主要涉及神经活性配体-受体相互作用、雌激素信号通路、

环磷酸腺苷(cyclic adenosine monophosphate, cAMP)信号通路及IL-17信号通路等(图5B)。GSEA结果显示高风险组中MAPK相关信号通路呈激活趋势, 其中GOBP_ERK1_AND_ERK2_CASCADE通路的归一化富集评分(normalized enrichment score, NES)为1.52(图5C)。MAPK通路是乳腺癌发生发展中的关键信号转导通路, 其异常激活与肿瘤侵袭、转移及化疗耐药密切相关, 且在年轻乳腺癌中尤为突出。基因突变谱分析显示, 高风险组中TP53基因的突变频率显著高于低风险组, 而PIK3CA和GATA3基因的突变频率显著降低(图5D)。此外, 雌激素受体相关基因ESR1 mRNA表达水平在高风险组中显著下调。蛋白质组学分析结果与转录组结果一致, 显示高风险组中雌激素受体 α 蛋白表达显著下调(图5E)。免疫细胞浸润分析进一步表明, 高风险组中浆细胞和CD8⁺T细胞的浸润比例显著降低(图5F)。

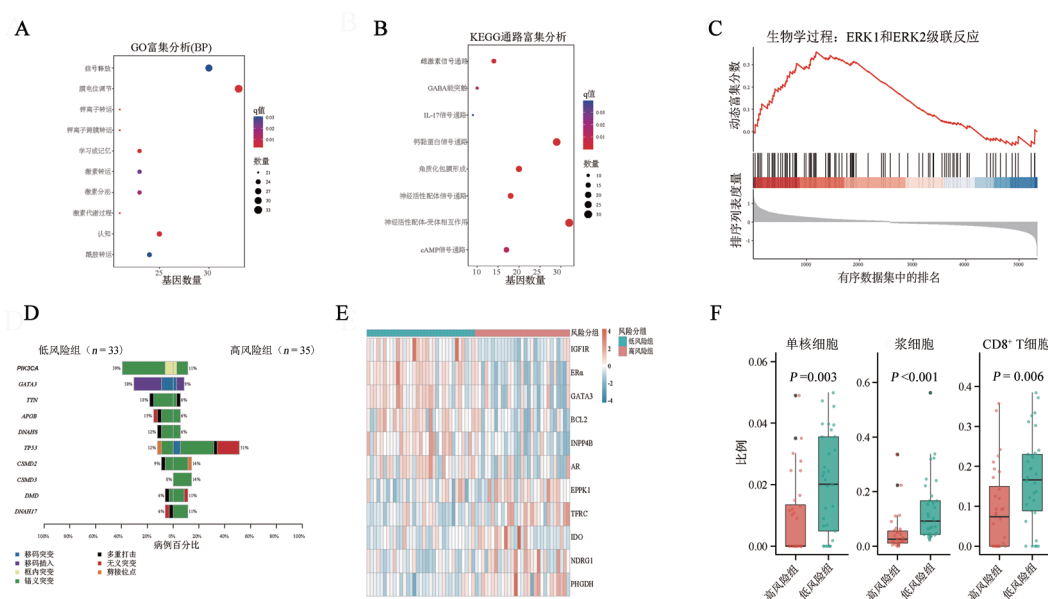


A~D: 高/低风险组间临床特征分布(TCGA年轻队列, $n=73$); E~H: 不同临床特征组的风险评分比较(TCGA年轻队列, $n=73$); I~P: 各临床亚组(ER⁺/ER⁻、PR⁺/PR⁻、Luminal型/非Luminal型、N0/非N0)内高/低风险组的OS曲线(基于TCGA+GEO扩展年轻乳腺癌队列, $n=169$)。

图4 12位点甲基化风险评分与临床特征及亚组预后的关系

表1 年轻乳腺癌不同临床亚组中高、低风险组的样本量及HR、95% CI

亚组	高风险组	低风险组	HR	95% CI	P
ER 阳性	36	36	5.21	1.16~23.4	0.016
ER 阴性	32	33	2.29	0.96~5.42	0.051
PR 阳性	29	29	5.30	1.18~23.86	0.015
PR 阴性	25	26	5.91	1.7~20.57	0.0015
Luminal 亚型	30	30	1.2×10^9	0~Inf	0.0017
非 Luminal 亚型	20	20	3.08	0.96~9.82	0.045
N0	28	29	7.75	0.98~61.36	0.022
非 N0	40	40	3.18	1.07~9.45	0.028



A:GO 富集分析;B:KEGG 通路富集分析;C:GSEA 富集通路;D:基因突变频谱;E:蛋白质组差异表达热图;F:免疫细胞浸润差异。

图5 年轻乳腺癌 12 位点甲基化模型高、低风险组的多组学差异

2.5 整合甲基化风险评分与临床、多组学指标的预后列线图构建

基于前述多组学分析结果,将年龄、TP53 基因突变状态、ESR1 mRNA 表达水平及 12 个 DNA 甲基化位点构建的风险评分共同纳入模型,构建了整合年龄、甲基化风险评分、TP53 突变状态及 ESR1 mRNA 表达水平的综合列线图(图 6A)。该列线图将上述信息整合为可视化的预后评估工具,为预后评估提供参考。时间依赖性 ROC 曲线提示该模型具有良好的区分能力(图 6B),校准曲线显示预测概率与实际生存结局具有良好一致性(图 6C)。

2.6 模型关键基因 LAMA5 的表达及预后价值验证

上述 12 个甲基化位点预后模型中包含一个位于 LAMA5 基因上的位点(cg18668449)。为验证该模型相关基因的临床意义,本研究选择 LAMA5 基因进行后续验证,分别在 mRNA 和蛋白水平评估其表达与预后的关联。LAMA5 编码层粘连蛋白 $\alpha 5$ 亚基,是基底膜的

核心组分,在维持乳腺上皮组织完整性和抑制肿瘤侵袭中发挥重要作用;文献^[16]报道 LAMA5 在乳腺癌中表达下调,且其低表达与患者不良预后相关。

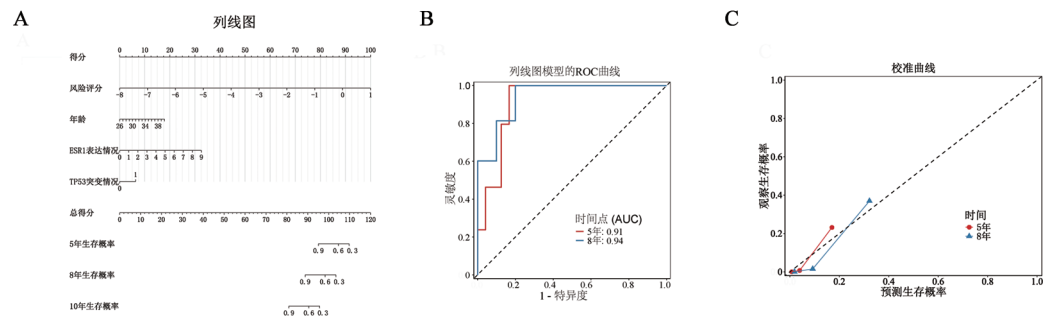
为验证该基因的预后价值,本研究首先在 TCGA 全体乳腺癌队列($n = 958$)中评估 LAMA5 mRNA 表达与临床预后的关联。以最佳截断值分组,LAMA5 mRNA 低表达组($n = 106$)的总生存期显著低于高表达组($n = 852$)($P < 0.05$)。临床病理特征分析显示,LAMA5 mRNA 低表达组中基底样型占比显著更高($P < 0.001$),ER 阴性及 PR 阴性比例亦显著升高(均 $P < 0.001$);此外,LAMA5 mRNA 低表达组的 TP53 突变频率显著高于高表达组(57% vs 32%)。

为排除分子分型对 LAMA5 预后价值的混杂影响,本研究进一步在 TCGA 乳腺癌队列的不同分子亚型中评估 LAMA5 mRNA 表达与预后的关联。将 Luminal A 型与 Luminal B 型合并为 Luminal 型,分别在 Luminal 型、HER2 过表达型及基底样型中按

LAMA5 mRNA 表达水平分组进行生存分析。结果显示,在 Luminal 型和基底样型中,LAMA5 mRNA 低表达组的总生存期均显著劣于 LAMA5 mRNA 高表达组(均 $P < 0.05$),而在 HER2 过表达型中 2 组间 OS 差异无统计学意义($P > 0.05$),提示 LAMA5 的预后价值在 Luminal 型及基底样型中更为显著。

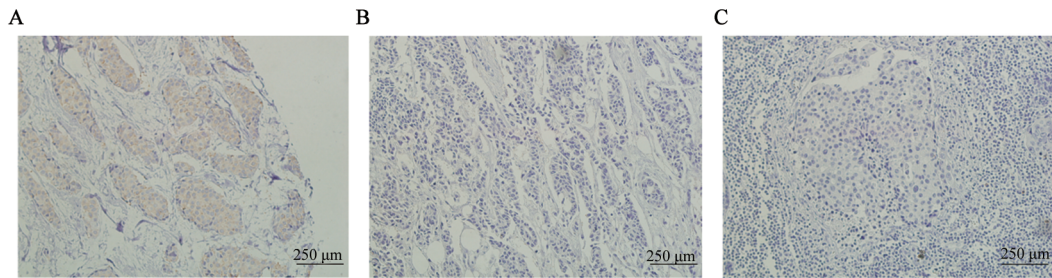
利用免疫组化方法比较该蛋白在不同分子分型

中的表达特征,结果显示,LAMA5 蛋白在 Luminal 型乳腺癌中呈高表达(图 7A),而在 HER2 过表达型与三阴性乳腺癌中表达显著减弱或缺失(图 7B~C)。上述结果与 TCGA 数据库中 LAMA5 mRNA 在 Luminal 型中高表达、在基底样型及 HER2 过表达型中低表达的生物信息学分析结果一致。



A: 列线图; B: 时间依赖性 ROC 曲线; C: 校准曲线。

图6 整合甲基化风险评分、年龄、TP53 突变和 ESR1 mRNA 表达的年轻乳腺癌预后列线图



A: Luminal 型; B: HER2 过表达型; C: 三阴性乳腺癌。

图7 LAMA5 蛋白在乳腺癌不同分子亚型中的表达(免疫组织化学染色, ×10)

进一步对 66 例具有完整随访资料的患者进行生存分析,分为 LAMA5 蛋白阴性组($n = 40$)和 LAMA5 蛋白阳性组($n = 26$)。结果显示,2 组间分子分型构成差异具有统计学意义($P = 0.037$)。相比 LAMA5 蛋白阳性组,LAMA5 蛋白阴性组的三阴性乳腺癌占比更高(30% vs 7.7%);肿瘤直径更大[(3.88 ± 1.59) cm vs (2.76 ± 0.93) cm, $P = 0.003$],而 2 组在年龄、淋巴结状态、病理分级及美国癌症联合委员会(American Joint Committee on Cancer, AJCC)分期方面差异无统计学意义(表 2)。生存分析显示,LAMA5 蛋白阴性组的总生存期显著低于阳性组($P = 0.046$)。上述结果从 mRNA 及蛋白层面均提示,LAMA5 转录水平下调及蛋白表达缺失与乳腺癌的侵袭性表型及不良预后密切相关,为模型关键基因的临床意义提供了实验支持。

3 讨论

本研究描绘了年轻乳腺癌独特的 DNA 甲基化图

谱。与老年乳腺癌相比,年轻乳腺癌在启动子区域,尤其是 CpG 岛内,呈现普遍的低甲基化特征,而在基因体区域未见显著差异。这一发现提示,启动子区低甲基化可能是年轻乳腺癌区别于老年患者的重要表观遗传标志。然而,尽管基因体区域整体甲基化水平在 2 组间无显著差异,后续分析表明该区域的特定甲基化位点仍具有重要的预后预测价值。

通过整合启动子区与非启动子区(包括基因体区和基因间区)甲基化信息,我们构建了一个由 12 个甲基化位点组成的预后预测模型。值得注意的是,在该模型中,位于基因体区域的 4 个甲基化位点与启动子区位点共同构成了核心预测因子,且随机森林变量重要性分析显示基因体区域位点的贡献较为突出。为明确整合基因体区域甲基化位点的增量价值,本研究比较了仅含启动子位点的模型与完整 12 位点模型的性能。结果显示,仅启动子模型虽能区分生存结局,但其预测效能有限,且无法有效富集 ER 阴性、PR 阴性及基底样型等侵袭性临床表型;而完整

12 位点模型不仅预测效能显著提升,还表现出与不良临床病理特征的强相关性。这一对比提示基因体区甲基化位点可能具有增量预后价值,支持了既往关于“基因体甲基化在肿瘤发生发展中具有与启动子甲基化同等重要生物学意义”的观点^[17-18]。近期独立队列的复制研究亦表明,多 CpG 甲基化特征与乳腺癌生存结局相关,进一步支持 DNA 甲基化标志物用于乳腺癌预后分层的潜力^[19]。尽管该模型基于年轻与老年乳腺癌的差异甲基化位点构建,但其在老年乳腺癌队列中同样具有一定的预测效能,提示该甲基化信号具有一定程度的泛化性。然而,该模型在年轻队列中的 AUC 各时间点均高于老年队列,且在年轻乳腺癌多个临床亚组中仍表现出一定的预后分层能力(图 4I~P),表明其核心预后信号在年轻患者中具有更好的适用性。该模型在不同年龄人群中的预测效能,尚需在更大样本队列中进一步验证。

表2 LAMA5蛋白阴性组与阳性组临床病理特征比较 $[\bar{x} \pm s, n(\%)]$

临床病理特征	阴性组 (n = 40)	阳性组 (n = 26)	P
年龄/岁	53.88 ± 14.20	51.92 ± 12.91	0.574
肿瘤大小/cm	3.88 ± 1.59	2.76 ± 0.93	0.003
阳性淋巴结数/个	2.97 ± 3.79	2.00 ± 1.98	0.241
病理分级			0.605
I 级	5(12.5%)	2(7.7%)	
II 级	33(82.5%)	24(92.3%)	
III 级	2(5.0%)	0(0.0%)	
T 分期			0.007
T1	4(10.0%)	9(34.6%)	
T2	29(72.5%)	17(65.4%)	
T3	7(17.5%)	0(0.0%)	
N 分期			0.790
N0	15(38.5%)	8(32.0%)	
N1~N3	24(61.5%)	17(68.0%)	
AJCC 分期			0.261
I 期	2(5.1%)	1(4.0%)	
II 期	22(56.4%)	19(76.0%)	
III 期	15(38.5%)	5(20.0%)	
分子分型			0.037
Luminal 型	16(40.0%)	18(69.2%)	
HER2 过表达型	12(30.0%)	6(23.1%)	
三阴性型	12(30.0%)	2(7.7%)	

病理分级、T 分期、分子分型数据完整(阴性组 $n = 40$, 阳性组 $n = 26$); N 分期、AJCC 分期各有 2 例数据缺失(阴性组 1 例, 阳性组 1 例)。各分类变量均以有效例数作为分母计算百分比。

进一步分析表明,该甲基化风险评分在年轻乳

腺癌的多个临床亚组中仍表现出一定的预后分层能力。基于风险评分划分的高风险组患者,呈现出与不良预后相关的多组学分子特征。在基因组层面,该组患者 TP53 突变频率显著升高,而 PIK3CA 和 GATA3 等管腔型相关基因突变频率降低;在转录组与蛋白质组层面,雌激素受体信号通路显著减弱,同时 MAPK 相关信号通路异常激活。这些特征与既往关于年轻乳腺癌在基因组^[7-8]、转录组^[9]及蛋白质组^[10]层面的研究结果高度一致,进一步验证了本模型风险分组的生物学合理性。基于上述发现,本研究尝试构建了整合年龄、表观遗传、基因组及转录组信息的综合列线图。该列线图可作为多组学信息整合的可视化预后评估工具,为临床决策提供参考。这一延伸性分析提示,多组学指标的整合可能具有进一步优化预后评估的潜能,但其临床效能仍需独立队列验证。

为从蛋白水平验证模型关键基因的临床意义,本研究针对模型中所包含的 LAMA5 基因进行了独立验证。LAMA5 编码层粘连蛋白 $\alpha 5$ 亚基,是细胞外基质的关键组分,维持基底膜完整性^[20]。本研究中,TCGA 数据库生物信息学分析提示 LAMA5 mRNA 低表达与乳腺癌不良预后及侵袭性分子表型相关。更重要的是,在包含多种乳腺癌亚型的组织芯片中通过免疫组化检测发现,相比 LAMA5 蛋白阳性组, LAMA5 蛋白阴性组的肿瘤直径更大、三阴性乳腺癌占比更高,且总生存期显著缩短。这一结果支持 LAMA5 蛋白表达与乳腺癌临床表型及预后相关,但尚不能据此确定 cg18668449 甲基化与 LAMA5 表达改变之间的因果关系。LAMA5 mRNA 低表达组中同时观察到 TP53 高频突变,与本研究高风险组的基因组特征一致,提示 2 者可能共同参与侵袭性表型的形成,但具体机制尚待阐明。将乳腺癌分子亚型分为 Luminal 型、HER2 过表达型及基底样型后的亚组分析进一步显示, LAMA5 mRNA 和蛋白层面的预后价值在 Luminal 型保持一致,提示 LAMA5 对预后的预测作用与 PAM50 分子亚型相关。然而,鉴于组织芯片验证队列样本量有限($n = 66$)且结局事件数不足,当前 Kaplan-Meier 分析仅提示 LAMA5 表达缺失与不良预后相关,尚无法通过多因素 Cox 回归评估其独立预后价值,仍需在更大样本的独立队列中验证。

本研究亦存在局限性。首先,本研究为回顾性分析,且 TCGA 训练集中年轻乳腺癌样本量较小,可能导致选择偏倚,这凸显了在大规模、多中心前瞻性队列中进一步验证和优化本模型的必要性。其次,本研究将年龄、TP53 突变、ESR1 表达及甲基化风险评分纳入列线图模型,旨在探索多组学指标整合对

预后评估的潜在增量价值。然而,考虑到 TP53 和 ESR1 在高/低风险组间存在显著差异,其与风险评分的相关性可能导致模型参数共线性问题,存在一定的过拟合风险。因此,该列线图的临床推广需谨慎,其效能有待在独立前瞻性队列中进一步验证。

[参考文献]

- [1] Wang X, Xia C F, Wang Y, et al. Landscape of young breast cancer under 35 years in China over the past decades: a multicentre retrospective cohort study (YBCC-Catts study)[J]. *eClinicalMedicine*, 2023, 64: 102243. DOI:10.1016/j.eclinm.2023.102243.
- [2] Yap Y S, Lu Y S, Tamura K, et al. Insights into breast cancer in the east vs the west: a review[J]. *JAMA Oncol*, 2019, 5(10): 1489-1496. DOI:10.1001/jamaoncol.2019.0620.
- [3] Sung H, Filho A M, Laversanne M, et al. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries[J]. *CA Cancer J Clin*, 2026, 76(4): e70090. DOI:10.3322/caac.70090.
- [4] Paluch-Shimon S, Cardoso F, Partridge A H, et al. ESO-ESMO fifth international consensus guidelines for breast cancer in young women (BCY5) [J]. *Ann Oncol*, 2022, 33(11): 1097-1118. DOI: 10.1016/j.annonc.2022.07.007.
- [5] Jenkins E O, Deal A M, Anders C K, et al. Age-specific changes in intrinsic breast cancer subtypes: a focus on older women[J]. *Oncologist*, 2014, 19(10): 1076-1083. DOI:10.1634/theoncologist.2014-0184.
- [6] Kumar R, Abreu C, Toi M, et al. Oncobiology and treatment of breast cancer in young women[J]. *Cancer Metastasis Rev*, 2022, 41 (3): 749-770. DOI:10.1007/s10555-022-10034-6.
- [7] Waks A G, Kim D, Jain E, et al. Somatic and germline genomic alterations in very young women with breast cancer[J]. *Clin Cancer Res*, 2022, 28(11): 2339-2348. DOI: 10.1158/1078-0432.ccr-21-2572.
- [8] Luen S J, Viale G, Nik-Zainal S, et al. Genomic characterisation of hormone receptor-positive breast cancer arising in very young women[J]. *Ann Oncol*, 2023, 34(4): 397-409. DOI: 10.1016/j.annonc.2023.01.009.
- [9] Parsons A, Sauras Colón E, Manjunath M, et al. Cell populations in human breast cancers are molecularly and biologically distinct with age[J]. *Nat Aging*, 2025, 5(12): 2546-2563. DOI: 10.1038/s43587-025-00984-1.
- [10] Raj-Kumar P K, Liu J F, Soltis A R, et al. Proteogenomic characterization of invasive breast tumors in young women[J]. *npj Breast Cancer*, 2025, 11: 94. DOI:10.1038/s41523-025-00793-0.
- [11] Sun S H, Chen S, Wang N, et al. DNA methylation profiling deciphers three EMT subtypes with distinct prognoses and therapeutic vulnerabilities in breast cancer[J]. *J Cancer*, 2024, 15 (15): 4922-4938. DOI:10.7150/jca.96096.
- [12] Gao L J, Zhang L, Liu J Y, et al. Discovery and validation of cell-free DNA methylation markers for specific diagnosis, differentiation from benign tumors, and prognosis of breast cancer [J]. *Breast Cancer Res*, 2025, 27(1): 108. DOI:10.1186/s13058-025-02066-x.
- [13] McGuire M H, Dasari S K, Yao H, et al. Gene body methylation of the lymphocyte-specific Gene CARD11 results in its overexpression and regulates cancer mTOR signaling[J]. *Mol Cancer Res*, 2021, 19(11): 1917-1928. DOI:10.1158/1541-7786.mcr-20-0753.
- [14] Shen H T, Hung C S, Davis C, et al. Hypermethylation of the gene body in SRCIN1 is involved in breast cancer cell proliferation and is a potential blood-based biomarker for early detection and a poor prognosis[J]. *Biomolecules*, 2024, 14(5): 571. DOI: 10.3390/biom14050571.
- [15] Zhang H K, Dong P L, Fan H L, et al. Gene body hypomethylation of pyroptosis-related genes NLRP7, NLRP2, and NLRP3 facilitate non-invasive surveillance of hepatocellular carcinoma[J]. *Funct Integr Genomics*, 2023, 23(2): 198. DOI: 10.1007/s10142-023-01114-z.
- [16] Zhang X L, Yan H F, Bie L H, et al. The role of LAMA5 in breast cancer progression and its potential in immunotherapy[J]. *Discov Oncol*, 2025, 16(1): 1387. DOI:10.1007/s12672-025-03113-x.
- [17] Jones P A. Functions of DNA methylation: islands, start sites, gene bodies and beyond[J]. *Nat Rev Genet*, 2012, 13(7): 484-492. DOI: 10.1038/nrg3230.
- [18] Xi Y Y, Lin Y, Guo W J, et al. Multi-omic characterization of genome-wide abnormal DNA methylation reveals diagnostic and prognostic markers for esophageal squamous-cell carcinoma[J]. *Signal Transduct Target Ther*, 2022, 7: 53. DOI: 10.1038/s41392-022-00873-8.
- [19] Zarean E, Li S, Wong E M, et al. Tumour DNA methylation markers associated with breast cancer survival: a replication study [J]. *Breast Cancer Res*, 2025, 27(1): 9. DOI: 10.1186/s13058-024-01955-x.
- [20] Englund J I, Ritchie A, Blaas L, et al. Laminin alpha 5 regulates mammary gland remodeling through luminal cell differentiation and Wnt4-mediated epithelial crosstalk[J]. *Development*, 2021, 148 (12): dev199281. DOI:10.1242/dev.199281.

[收稿日期] 2026-07-02

[修回日期] 2026-08-18

[本文编辑] 黄静怡