人工智能技术应用于医学数据分析

DOI：10.13406/j.cnki.cyxb.003380

基于机器学习的病理组学特征可预测乳腺癌患者对新辅助化疗的反应

张杰秋[1]，伍　棋[2]，王舰梅[2]，要小鹏[3]

（1. 西南医科大学公共卫生学院流行病与卫生统计学教研室，泸州　646000；2. 西南医科大学附属医院病理科，泸州　646000；3. 西南医科大学医学信息与工程学院医学工程技术教研室，泸州　646000）

【摘　要】目的：利用病理组学的方法开发用于预测乳腺癌（breast cancer，BC）患者新辅助化疗（neoadjuvant chemotherapy，NAC）反应的新型标志物。**方法：**回顾性纳入211例西南医科大学附属医院的非特殊浸润性BC患者（训练组：155例，验证组：56例），使用CellProfiler软件提取患者数字病理切片中的高维病理组学特征，利用Mann-Whitney *U*检验、Spearman相关系数和最小绝对值收敛和选择算子（least absolute shrinkage and selection operator，LASSO）算法进行特征逐层筛选。筛选后的最优特征通过支持向量机（support vector machine，SVM）方法在训练集中开发了病理组学特征（pathomics signature，PS）并在独立验证集中进行验证。PS与单因素有意义的临床病理因素（P<0.05）纳入多因素逻辑回归进行进一步验证。**结果：**PS的曲线下面积（area under the curve，AUC）为0.749（95%CI=0.672~0.827），验证集中AUC为0.737（95%CI=0.604~0.870）。多因素逻辑回归的结果显示，PS（*OR*=2.317）与人表皮生长因子受体2（human epidermal growth factor receptor 2，HER2）（*OR*=4.018）是BC患者NAC反应的独立预测因素。**结论：**PS可以帮助临床医生在治疗前准确预测NAC的反应，促进BC患者的个性化治疗。

【关键词】病理组学；新辅助化疗；乳腺癌；预测模型

【中图分类号】R737.9　　**【文献标志码】**A　　**【收稿日期】**2023-05-05

Pathomics signature based on machine learning can predict the response to neoadjuvant chemotherapy in breast cancer patients

Zhang Jieqiu[1]，*Wu Qi*[2]，*Wang Jianmei*[2]，*Yao Xiaopeng*[3]

（*1. Department of Epidemiology and Health Statistics，School of Public Health，Southwest Medical University；2. Department of Pathology，The Affiliated Hospital of Southwest Medical University；3. Department of Medical Engineering Technology，School of Medical Information and Engineering，Southwest Medical University*）

【Abstract】*Objective*：To develop a novel marker for predicting the response to neoadjuvant chemotherapy（NAC）in patients with breast cancer（BC）using the pathomics approach. ***Methods*：**A retrospective analysis was performed for 211 patients with non-specific invasive BC in The Affiliated Hospital of Southwest Medical University，among whom 155 were enrolled as training group and 56 were enrolled as validation group. CellProfiler software was used to extract high-dimensional pathomics signature from the digital pathological sections of patients，and then the Mann-Whitney *U* test，the Spearman correlation coefficient，and the least absolute shrinkage and selection operator（LASSO）algorithm were used for the stepwise screening of features. The optimal features after screening were used to develop pathomics signature（PS）by the support vector machine（SVM）method in the training set and validate in the independent validation set. PS and significant clinicopathological factors（P<0.05）identified in the univariate analysis were included in the multivariate logistic regression analysis for further validation. ***Results*：**PS had an area under the ROC curve of 0.749（95%CI=0.672-0.827）in the training set and 0.737（95%CI=0.604-0.870）in the validation set. The multivariate logistic regression analysis showed that PS［odds ratio（*OR*）=2.317］and human epidermal growth factor receptor 2（*OR*=4.018）were independent predictive factors for response to NAC in BC patients. ***Conclusion*：**PS can help clinicians accurately predict the response to NAC before treatment and improve the personalized treatment for BC patients.

【Key words】pathomics；neoadjuvant chemotherapy；breast cancer；predictive model

作者介绍：张杰秋，Email：ygcandidate@163.com，研究方向：卫生统计与临床预测模型。

通信作者：要小鹏，Email：xp_yao@swmu.edu.cn。

基金项目：四川省科技计划资助项目（编号：2022YFS0616）；泸州市科技计划资助（编号：2023SYF112）。

优先出版：https://link.cnki.net/urlid/50.1046.R.20231227.1610.002（2023-12-29）

乳腺癌（breast cancer，BC）是世界上女性中最常见的癌症，其相关发病率不断上升且发病人群逐渐年轻化[1-2]。新辅助化疗（neoadjuvant chemotherapy，NAC）已成为早期高危和局部晚期BC的标准治疗方案，可以帮助患者降期缩瘤以接受更保守的治疗[3]。治疗反应患者其预后通常也表现良好[4]。然而，由于肿瘤的异质性和复杂性，并非所有患者对NAC都具有敏感性。因为长期治疗过程仍会产生一些毒副作用[5]，也可能错过改变治疗计划的最佳时间，所以对BC患者NAC前预测其疗效至关重要。

目前，肉眼观察仍然是病理切片信息获取的主要方式。随着医学图像高通量处理技术的发展，以及对得到的高维数据的广泛探索与挖掘，“病理组学”引起了越来越多的关注。病理组学包含从数字病理学图像中捕获的各种数据生成定量特征。病理组学特征可提供有关肿瘤微环境的相关信息，目前研究已经在癌症风险分层、预后预测和辅助化疗疗效预测等方面进行[6-8]。本研究构建了一个基于机器学习的病理组学特征（pathomics signature，PS），通过体内微观角度预测BC患者对NAC的反应。这可能为临床医生提供辅助决策意见，以促进个体化治疗的过程。

1 资料与方法

1.1 研究对象的纳入和排除

这项回顾性研究获得了西南医科大学医院机构审查委员会（编号KY2022216）的批准，并放弃了书面知情同意的要求。研究对象入选标准如下：病理活检证实无远处转移的非特异浸润性BC；患者进行了6~8个周期的NAC；NAC完成后进行手术，并在术后病理切片上评估疗效；临床数据可用。排除标准：无组织病理学评估结果；多病灶同侧或双侧乳腺癌。

1.2 图像资料

病理医生在NAC之前使用粗针穿刺收集了BC患者的活检样本，然后对其进行病理制片。首先，将活检组织浸泡在浓度为10%的福尔马林中4 h，之后包埋在免疫组织化学石蜡中。随后，以4 μm的间隔对蜡块进行切片，并用苏木精和伊红进行染色准备用于病理评估。病理医生使用数字载玻片扫描仪（KFBio KF-PRO-020）以40倍扫描倍率扫描所有治疗前的组织病理切片，以获得患者的数字病理切片。在数字切片管理器中，将样本放大10倍，病理医生选择具有代表性样本区域，并获得了一张512×512像素的截图，然后由另一位病理医生进行确认，他们分别有3年和8年的BC病理诊断经验。如果2位病理学家有不同意见，他们将与第3位病理学家进行商讨以做出决定。

1.3 临床病理资料

通过查阅电子病历，获得BC患者的临床病理信息，包括年龄、T分期、N分期、Ki-67、雌激素受体（estrogen receptor，ER）、孕激素受体（progesterone receptor，PR）和人表皮生长因子受体2（human epidermal growth factor receptor 2，HER2）的表达。ER和PR阳性定义为≥1%的肿瘤细胞具有核染色阳性；免疫组化法（immunohistochemistry，IHC）染色为3的肿瘤被认为是HER2阳性，荧光原位杂交（fluorescence insitu hybridization，FISH）进一步确认，FISH扩增的结果被确定为HER2阳性；Ki-67的临界值为30%。使用Miller-Payne（MP）分级标准对术后样本进行病理反应的评价以获得研究结局变量[9]，MP系统分为G1~G5等级，无反应患者定义为G1~G3，反应患者定义为G4~G5。

1.4 病理组学分析方法

1.4.1 总体分析路径 包括图像分割、特征提取、特征选择与PS的构建评价4个步骤，如图1所示。

1.4.2 病理组学特征提取 使用图像分析软件CellProfiler（版本4.0.7）提取所选病理图像的定量特征[10]。基于“Unmix Colors”模块，分离原始染色图像并将其转换为苏木精染色和伊红染色的灰度图像，原始染色图像也使用“ColorToGray”模块转换为灰度图像。测量过程分为总体测量与对象测量。第1次对总体图像进行测量中，获得了136个病理组学特征。第2次测量对苏木精染色图像进行对象测量。首先确定主要和次要对象，然后对其进行测量。对每张图片中的大量对象原始值求平均值、中位数值和标准差值，将其作为研究特征，共获得1 054个病理组学特征。具体提取过程如图2所示。

1.4.3 病理组学特征选择 为了去除冗余特征，对每个特征进行Mann-Whitney U检验，P值确定为0.05。随后，考虑到特征之间的相关性，对特征进行Spearman相关性分析，如果2个特征之间的相关系数绝对值大于0.9，则排除其中1个特征。然后，使用最小绝对值收敛和选择算子（least absolute shrinkage and selection operator，LASSO）算法来选择提取的特征，并使用5倍交叉验证来选择Lambda值以确定最佳特征子集。

1.4.4 构建PS 基于以上筛选的最优特征，通过非线性支持向量机（support vector machine，SVM）使用选择后的最优特征来构建病理组学预测模型，通过5倍交叉验证和网格搜索来确定高斯径向基函数核的最佳正则化参数C和Gamma。将病理组学模型的输出预测值当作PS。

1.4.5 独立预测因素的验证 以NAC疗效为因变量（0=无反应，1=反应），将单因素分析中有统计学意义（P<0.05）的变量纳入多因素逻辑回归模型筛选出BC患者NAC反应的独立预测因子。

1.5 统计学方法

采用Jupyter notebook和R-studio进行统计分析。采用卡方检验比较训练集及测试集2组间T分期、N分期、ER、PR、HER2、Ki-67是否有显著性差异；采用Kolmogorov-

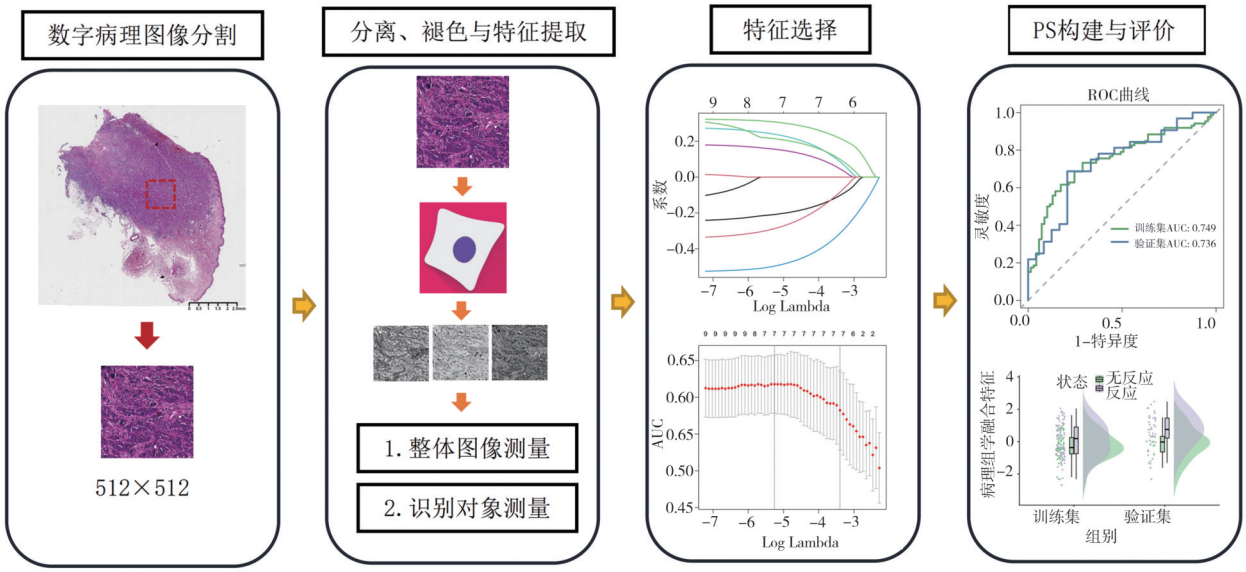


图1　病理组学工作流程图

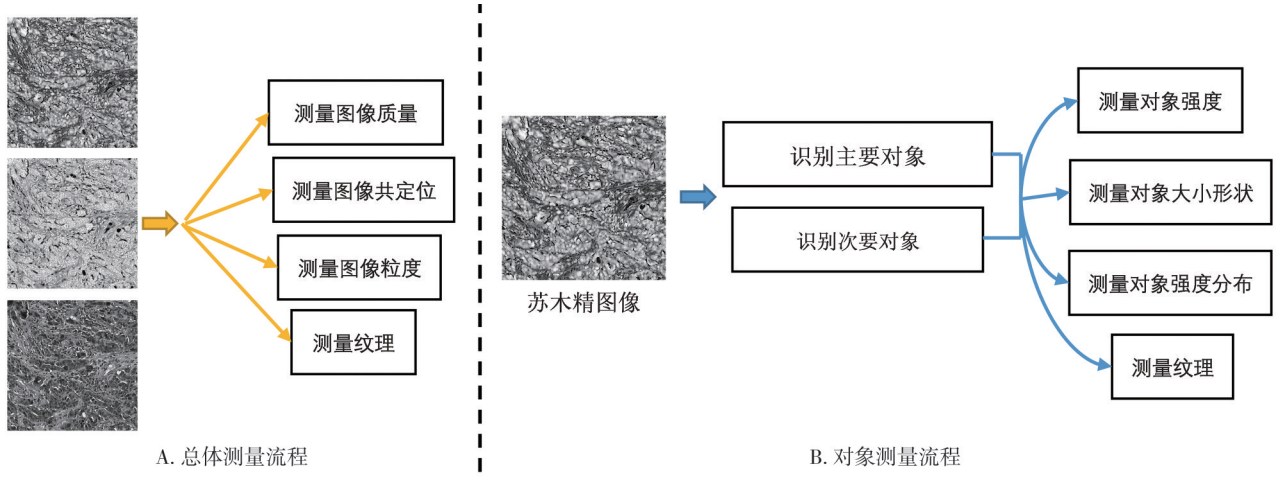


图2　病理组学特征提取

Smirnov 检验对训练集及验证集 2 组患者的年龄进行正态性检验，满足正态分布的采用独立样本 t 检验，用均数 ± 标准差（$\bar{x} \pm s$）形式表达；否则采用 Mann-Whitney U 检验比较 2 组之间的差异有无统计学意义，用 $M_d(P_{25}, P_{75})$ 表示。受试者工作特征（receiver operating characteristic，ROC）曲线下面积（area under the curve，AUC）来评估病理组学模型的辨别性能；多因素逻辑回归适用于独立预测因素的进一步验证。检验水准 α=0.05。

2　结　果

2.1　训练集与测试集临床病理特征比较

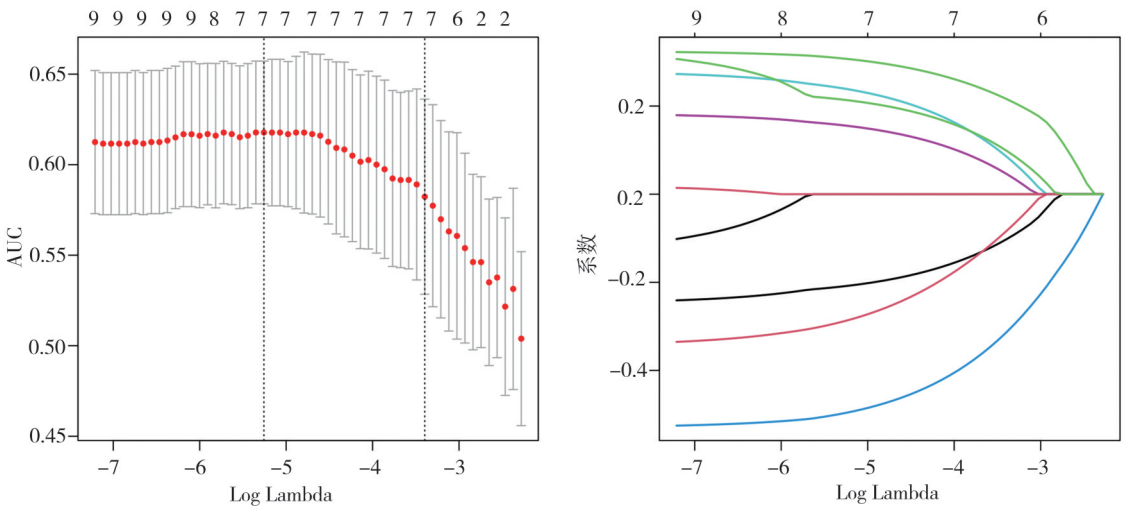
从 2020 年 2 月到 2022 年 3 月，共有 211 例非特异浸润性 BC 患者入选该研究，按照时间顺序将患者分为训练集和验证集。首次治疗时间在 2021 年 10 月之前的患者被纳入训练集，其余患者被纳入独立验证集，训练集与验证集的比率约为 7∶3。训练集 155 例患者中，有反应患者 86 例，无反应患者 69 例；验证集 56 例患者中，有反应患者 32 例，无反应患者 24 例。结果显示 HER2 的发现与训练和验证集中的 NAC 反应显著相关（P<0.05），见表 1。

2.2　病理组学特征提取与选择

每位患者提取 1 190 个病理特征，使用 z-score 对所有特征进行标准化。在训练集中，首先对所有特征进行 Mann-Whitney U 检验，筛选出 P 值小于 0.05 的 27 个特征；然后对这些特征进行 Spearman 相关性分析，得到 9 个最优特征。使用五折交叉验证的 LASSO 对 9 个筛选后的特征进行进一步筛选。五折交叉验证以模型效能评价指标 AUC 为衡量标准，AUC 最高时的 lambda 值（min=0.008 296，1se=0.030 515）为最佳值。筛选出 7 个最佳特征用于后续的模型建立。图 3 显示出 LASSO 的特征选择过程，最佳特征的特征名与系数值如图 4 所示。

表1 临床病理特征单因素分析[$(\bar{x}\pm s)$;n,%]

特征	训练集（n=155）		P值	验证集（n=56）		P值
	无反应（n=69）	反应（n=86）		无反应（n=24）	反应（n=32）	
年龄	49.00±8.54	49.50±7.69	0.932	50.30±9.39	49.70±10.40	0.809
T分期			0.298			0.270
T_1	2(2.9)	5(5.8)		2(8.3)	1(3.1)	
T_2	48(69.6)	63(73.3)		15(62.5)	26(81.2)	
T_3	7(10.1)	11(12.8)		3(12.5)	4(12.5)	
T_4	12(17.4)	7(8.1)		4(16.7)	1(3.1)	
N分期			0.028			1.000
N_0	20(29.0)	21(24.4)		6(25.0)	8(25.0)	
N_1	35(50.7)	59(68.6)		13(54.2)	16(50.0)	
N_2	11(15.9)	6(7.0)		3(12.5)	5(15.6)	
N_3	3(4.4)	0(0.0)		2(8.3)	3(9.4)	
ER			0.161			0.152
阴性	29(42.0)	47(54.7)		8(33.3)	18(56.2)	
阳性	40(58.0)	39(45.3)		16(66.7)	14(43.8)	
PR			0.287			0.263
阴性	23(33.3)	37(43.0)		9(37.5)	18(56.2)	
阳性	46(66.7)	49(57.0)		15(62.5)	14(43.8)	
HER2			<0.001			<0.001
阴性	53(76.8)	38(44.2)		20(83.3)	10(31.2)	
阳性	16(23.2)	48(55.8)		4(16.7)	22(68.8)	
Ki-67(%)			0.251			0.553
<30	20(29.0)	17(19.8)		7(29.2)	6(18.8)	
≥30	49(71.0)	69(80.2)		17(70.8)	26(81.2)	



A. 系数分布图　　B. 5折交叉验证图

图3 LASSO选择特征

2.3 PS的建立与验证

使用SVM的机器学习算法进行模型的建立。首先在jupyter notebook中建立基础的模型，然后在训练集中使用五倍交叉验证与网格搜索的方法进行模型最佳参数C与gamma的寻找，最优参数结果为C=10，gamma=0.01。建立好的模型在训练集与验证集中进行预测。训练集中，PS的AUC值为0.749（95%CI=0.672~0.827）；验证集中AUC值为0.737（95%CI=0.604~0.870）。多因素逻辑回归的结果显示，PS（*OR*=2.317）与HER2（*OR*=4.018）是BC患者NAC反应的独立预测因素，见图5、表2。

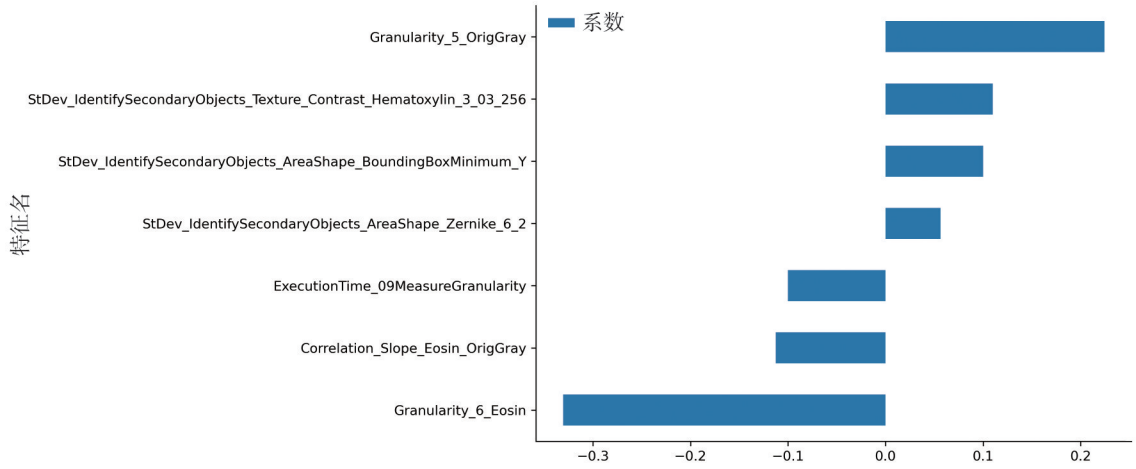


图4　特征系数图

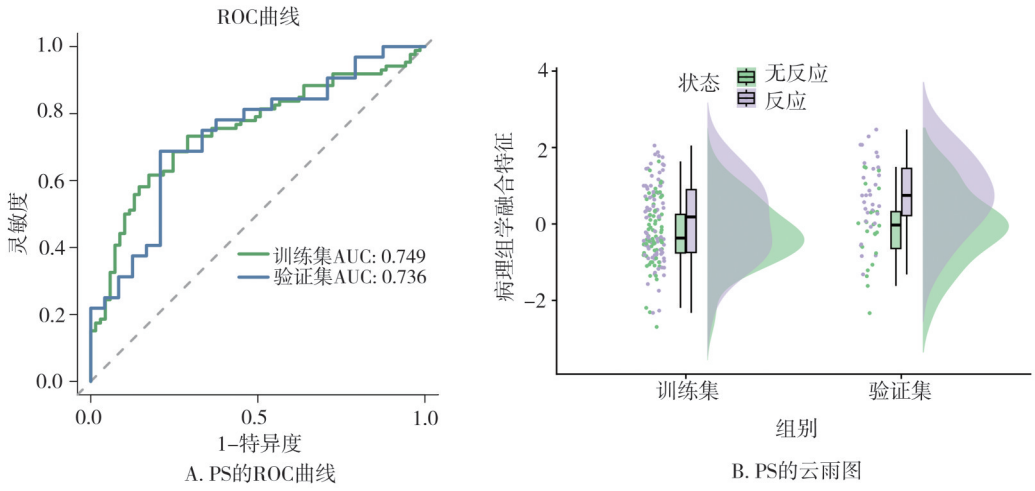


图5　PS的模型预测结果

表2　BC患者NAC反应预测因素的多因素逻辑回归分析

特征	*OR*	95%CI	*P*值
PS	2.317	1.630~3.398	<0.001
HER2	4.011	2.113~7.820	<0.001

3　讨　论

准确预测NAC的益处是临床中BC患者治疗方案选择中不可或缺的一部分。本研究使用LASSO算法从高维病理组学特征中筛选出7个最优特征，建立PS用于治疗前预测BC患者NAC的疗效。PS在训练集与独立验证集中表现稳定，预测效能较好。另外，本研究也证明了PS与HER2是BC患者NAC疗效预测中的重要预测因素。

BC是一种高度异质性肿瘤，当肿瘤对NAC产生反应时，肿瘤微环境将发生变化，但这种变化不易被肉眼察觉。得益于机器学习算法和图像分析方法的快速发展，影像组学已逐渐用于BC患者NAC的反应预测[11-13]。然而，由于癌症组织和放射图像之间的间接性，影像组学只能从体外摄影角度提供有关肿瘤的微量信息，可能会错过肿瘤细胞和细胞外基质中包含的重要信息。病理组学通过体内角度捕捉肿瘤的微观结构，并提供肿瘤病变中细胞与微环境的特征。一些学者和研究人员从数字病理切片中提取图像特征进行定量分析，病理组学特征被用于预测胃癌辅助化疗的疗效和预后、结直肠癌微卫星不稳定性等[8,14]，并取得了良好的效果。本研究中，病理组学特征在预测BC患者NAC疗效方面也显示出稳定的预测能力，它兼顾整体图像与局部图像，特征来源具有广泛性。据悉，这是第一项利用数字病理切片中提取的全定量成像特征预测BC患者NAC反应的研究。

PS具有一定的预测优势。第一个优势是使用方便快捷的图像处理方法来提取定量病理组学特征。迄今为止，病理组学特征的提取方式尚未达成共识。主要的提取方式为CellProfiler、Qupath与深度学习模型。CellProfiler不仅是一款免费的开源软件，还是一个易于使用且可重复的平台，允许临床医生自动批量测量生物图像并已用于数字病理学分析。相对于复杂的深度学习分割与模型建立方式，其处理方式具有普适性。第二个优势体现在使用机器学习方法构建简单易泛化的模型。过滤法与嵌入法相结合，筛选出最佳的7个纹理特征，使用少量特征进行模型构建其不易出现过拟合现象；且使用的SVM算法可以在模型的复杂性和学习能力之间寻求最佳折中[15]，综合使得PS具有较强的泛化能力与临床实用性。

同时，本研究还发现HER2也是预测BC患者NAC疗效的重要指标。据研究报道，HER2阳性BC患者的NAC病理完全反应率为50%~80%，高于其他亚型[16]。其原因可能是此亚型乳腺癌较高的生物学异质性与靶向药物的使用[17]。本篇研究也得出同样的结论，多因素分析结果显示，HER2是BC患者NAC疗效的独立预测因素（*OR*=4.011，95%CI=2.113~7.820，*P*<0.05）。

本研究中仍存在一些局限性。首先，本研究是回顾性研究，基于机器学习的PS还需要进行多中心、前瞻性临床试验的进一步验证。其次，由于人工制作切片染色具有差异性，数字病理切片可能会显示出一些颜色异质性，这可能会影响分析。未来还需要在病理组学分析的各个阶段进行制定标准化指南以推进病理组学的发展。

总之，本研究基于高维病理组学特征，成功构建了一个用于预测BC患者NAC反应的新型标志物，显示出稳定的预测能力，对BC患者的个性化医疗发展具有重要临床意义，显示出较高的推广应用价值。

参 考 文 献

[1] Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023[J]. CA Cancer J Clin, 2023, 73(1): 17-48.

[2] Giaquinto AN, Sung H, Miller KD, et al. Breast cancer statistics, 2022[J]. CA Cancer J Clin, 2022, 72(6): 524-541.

[3] Loi S. The ESMO clinical practise guidelines for early breast cancer: diagnosis, treatment and follow-up: on the winding road to personalized medicine[J]. Ann Oncol, 2019, 30(8): 1183-1184.

[4] Cortazar P, Zhang LJ, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis[J]. Lancet, 2014, 384(9938): 164-172.

[5] Jang MK, Park S, Park C, et al. Body composition change during neoadjuvant chemotherapy for breast cancer[J]. Front Oncol, 2022, 12: 941496.

[6] Chen ST, Jiang LR, Zheng XY, et al. Clinical use of machine learning-based pathomics signature for diagnosis and survival prediction of bladder cancer[J]. Cancer Sci, 2021, 112(7): 2905-2914.

[7] Chen ST, Jiang LR, Gao F, et al. Machine learning-based pathomics signature could act as a novel prognostic marker for patients with clear cell renal cell carcinoma[J]. Br J Cancer, 2022, 126(5): 771-777.

[8] Chen DX, Fu MT, Chi LJ, et al. Prognostic and predictive value of a pathomics signature in gastric cancer[J]. Nat Commun, 2022, 13(1): 6903.

[9] Corben AD, Abi-Raad R, Popa I, et al. Pathologic response and long-term follow-up in breast cancer patients treated with neoadjuvant chemotherapy: a comparison between classifications and their practical application[J]. Arch Pathol Lab Med, 2013, 137(8): 1074-1082.

[10] Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes[J]. Genome Biol, 2006, 7(10): R100.

[11] 刘 晨，陈小波，黄晓媚，等. 基于MRI影像组学预测乳腺癌新辅助化疗后肿瘤退缩模式的研究[J]. 磁共振成像，2023，14(3)：28-35.

Liu C, Chen XB, Huang XM, et al. MRI-based radiomics for prediction of tumor regression pattern to neoadjuvant chemotherapy in breast cancer[J]. Chin J Magn Reson Imag, 2023, 14(3): 28-35.

[12] 李蔓英，李 彬，罗 佳，等. 基于灰阶超声的影像组学模型预测乳腺癌新辅助化疗效果[J]. 中国医学影像技术，2019，35(9)：1331-1335.

Li MY, Li B, Luo J, et al. Ultrasound-based radiomics model in predicting efficacy of neoadjuvant chemotherapy in breast cancer[J]. Chin J Med Imag Technol, 2019, 35(9): 1331-1335.

[13] Mao N, Shi YH, Lian C, et al. Intratumoral and peritumoral radiomics for preoperative prediction of neoadjuvant chemotherapy effect in breast cancer based on contrast-enhanced spectral mammography[J]. Eur Radiol, 2022, 32(5): 3207-3219.

[14] Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer[J]. Theranostics, 2020, 10(24): 11080-11091.

[15] Huang MW, Chen CW, Lin WC, et al. SVM and SVM ensembles in breast cancer prediction[J]. PLoS One, 2017, 12(1): e0161501.

[16] Houssami N, MacAskill P, von Minckwitz G, et al. Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy[J]. Eur J Cancer, 2012, 48(18): 3342-3354.

[17] Baselga J, Bradbury I, Eidtmann H, et al. Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial[J]. Lancet, 2012, 379(9816): 633-640.

（责任编辑：周一青）