DOI:10.16016/j.2097-0927.202410117

基于机器学习的多模态数据建立Ⅰ期非小细胞肺癌患者术后复发预测模型

张 地[1],吴 毅[2],徐 瑜[1],王 帅[1],胡 玥[1],陈滑维[1],胡那娜[3,4],何 容[1],童雪玲[1],李梦侠[1] [1]陆军特色医学中心(第三军医大学大坪医院)肿瘤科,重庆;[2]陆军军医大学(第三军医大学)生物医学工程与影像医学系数字医学教研室,重庆;[3]渝粤病理科学研究中心,重庆;[4]金凤实验室,重庆

[**摘要**] **目的** 基于术前胸部CT影像组学特征联合临床信息构建机器学习模型,用于预测手术切除的Ⅰ期非小细胞肺癌患者术后5年复发风险。**方法** 回顾性纳入2014年1月至2019年12月陆军特色医学中心经手术病理确诊的Ⅰ期非小细胞肺癌患者217例(入组标准筛选自778例初筛病例),其中术后5年内复发53例,未复发164例。将病例按8∶2比例随机分为训练集(173例)和验证集(44例)。以肺癌原发灶主体作为感兴趣区域(ROI)提取影像组学特征,建立影像模型;同时基于患者人口学资料及术前检验等指标建立临床模型。进一步融合两类特征构建融合模型,并通过模型性能比较筛选最优预测模型。**结果** 本研究通过平扫CT影像特征筛选,最终纳入7个影像组学特征用于影像模型构建。在6种机器学习模型中,自适应增强算法(adaptive boosting, Adaboost)的训练集曲线下面积(AUC)为0.866(95%*CI* 0.808~0.923),准确度和特异度分别为0.832、0.884;验证集AUC为0.806(95%*CI* 0.630~0.983),准确度和特异度分别为0.795、0.971,综合比较预测性能最好。基于单因素及多因素Logistic回归分析筛选了4个临床特征参与构建临床模型,在训练集中AUC为0.874(95%*CI* 0.821~0.928),准确度和特异度分别为0.827、0.891;验证集中AUC为0.813(95%*CI* 0.677~0.948),准确度和特异度分别为0.636、0.600。进一步通过特征级融合策略整合影像组学与临床特征后,融合模型在训练集AUC达到0.953(95%*CI* 0.924~0.983),准确度和特异度分别为0.884、0.860;在验证集AUC达到0.852(95%*CI* 0.729~0.976),准确度和特异度分别为0.682、0.629,预测效能得到进一步提升。**结论** 融合术前CT影像组学特征与临床风险因素构建的联合模型,可为评估Ⅰ期非小细胞肺癌患者术后5年内的复发风险提供依据。

[**关键词**] 非小细胞肺癌;机器学习;计算机断层扫描;Ⅰ期;5年复发风险
[**中图法分类号**] R319;R734.2;R814.42 [**文献标志码**] A

Development of a postoperative recurrence prediction model for stage Ⅰ non-small cell lung cancer patients using multimodal data based on machine learning

ZHANG Di[1], WU Yi[2], XU Yu[1], WANG Shuai[1], HU Yue[1], CHEN Huawei[1], HU Nana[3,4], HE Rong[1], TONG Xueling[1], LI Mengxia[1] [1]Department of Oncology, Army Medical Center of PLA/Daping Hospital of Third Military Medical University, Chongqing; [2]Department of Digital Medicine, Faculty of Biomedical Engineering and Medical Imaging, Army Medical University (Third Military Medical University), Chongqing; [3]Yu-Yue Pathology Scientific Research Center, Chongqing; [4]Jinfeng Laboratory, Chongqing, China

[**Abstract**] **Objective** To develop a machine learning model integrating preoperative chest CT radiomic features with clinical data for predicting 5-year postoperative recurrence risk in stage Ⅰ non-small

[**通信作者**] 李梦侠,E-mail:limengxia@tmmu.edu.cn

2097-0927/Copyright © 2025,陆军军医大学(第三军医大学)。这篇开放获取文章遵循CC BY许可协议(https://creativecommons.org/licenses/by/4.0/)

引用格式:张地,吴毅,徐瑜,等.基于机器学习的多模态数据建立Ⅰ期非小细胞肺癌患者术后复发预测模型[J].陆军军医大学学报,2025,47(14):1602-1611.DOI:10.16016/j.2097-0927.202410117.

Citation:ZHANG D,WU Y,XU Y,*et al*. Development of a postoperative recurrence prediction model for stage Ⅰ non-small cell lung cancer patients using multimodal data based on machine learning[J]. J Army Med Univ,2025,47(14):1602-1611.DOI:10.16016/j.2097-0927.202410117.

cell lung cancer (NSCLC) patients undergoing surgical resection. **Methods** A total of 217 patients with pathologically confirmed stage Ⅰ NSCLC (selected from 778 initially screened cases based on our inclusion and exclusion criteria) treated in Army Medical Center of PLA between January 2014 and December 2019 were retrospectively enrolled, including 53 recurrence cases and 164 non-recurrence cases within 5-year follow-up. They were randomly divided into a training set (n=173) and a validation set (n=44) in a ratio of 8:2. Radiomic models were established based on extracted features from tumor-dominant regions of interest (ROI) on CT images, while clinical models were developed using demographic characteristics and preoperative laboratory examinations. A combined model was further constructed by integrating both feature sets, and model performance was compared to identify the optimal predictive model. **Results** This study screened the features from non-contrast CT images and ultimately selected 7 radiomic features for constructing radiomic model. Among 6 machine learning algorithms, the adaptive boosting (Adaboost) model demonstrated the best overall predictive performance, with an area under the curve (AUC) of 0.866 (95% *CI*: 0.808~0.923; accuracy: 0.832, specificity: 0.884) in the training set and of 0.806 (95% *CI*: 0.630~0.983; accuracy: 0.795, specificity: 0.971) in the validation set. Univariate and multivariate logistic regression analyses identified 4 clinical features for clinical model construction. The clinical model achieved an AUC value of 0.874 (95% *CI*: 0.821~0.928; accuracy: 0.827, specificity: 0.891) in the training set and 0.813 (95% *CI*: 0.677~0.948; accuracy: 0.636, specificity: 0.600) in the validation set. By integrating the 7 radiomic features and 4 clinical features using a feature-level fusion strategy, the combined model exhibited further improved predictive performance, with an AUC value of 0.953 (95% *CI*: 0.924~0.983; accuracy: 0.884, specificity: 0.860) and 0.852 (95% *CI*: 0.729~0.976; accuracy: 0.682, specificity: 0.629), respectively in the training set and the validation set. **Conclusion** The combined model integrating preoperative CT radiomic features with clinical risk factors may provide an evidence-based framework for evaluating 5-year postoperative recurrence risk in stage Ⅰ NSCLC patients.

[**Key words**] non-small cell lung cancer; machine learning; computed-tomography; stage Ⅰ; 5-year recurrence risk

Corresponding author: LI Mengxia, E-mail: limengxia@tmmu.edu.cn

肺癌作为全球范围内发病率和死亡率极高的恶性肿瘤[1],在我国同样占据恶性肿瘤新发及死亡人数首位,其中非小细胞肺癌(non-small cell lung cancer,NSCLC)占比高达80.0%~85.0%[2]。近年来,随着高分辨率CT技术的普及,Ⅰ期NSCLC的诊断比例已从2010年的23.5%上升至29.1%[3]。尽管处于疾病早期,Ⅰ期NSCLC 5年复发率高达20.0%~48.7%[4-5]。目前Ⅰ期NSCLC术后辅助治疗仍存在争议,ⅠA期患者不推荐术后化疗,而ⅠB期也仅推荐存在低分化癌、血管浸润、楔形切除、脏层胸膜受累和淋巴结状态不明的患者进行术后辅助治疗,其原因在于目前尚未精准寻找到Ⅰ期NSCLC复发高危患者以给予合适的术后辅助计划。因此,进一步识别5年复发高危患者,并优化术后辅助治疗及随访策略可能是提高Ⅰ期NSCLC预后的重要手段。

以往对肿瘤预后的研究主要利用临床风险因素来构建模型,随着医学大数据时代的到来,影像组学作为一种基于医学图像高维定量分析的方法,已被证明能够提取出比传统目测更为细致的特征,为治疗策略提供更丰富的信息[6]。通过开发具有精细特征的适当模型,影像组学在NSCLC的不同应用中展现出了良好的预测和预后效果,包括病变诊断[7]、淋巴结转移预测[8]以及治疗反应评估[9]等。

鉴于疾病成因的复杂性和多变性,以及各种检查手段的局限性,单一类型的检查难以全面评估各种医学情况。为了提高诊断的准确性、减少风险并提升治疗效果,有必要将多种检查手段和医学评价指标纳入医学诊疗过程,并利用计算机辅助技术帮助医师更全面、准确地把握描述患者不同体征和检查的信息。综合使用多种不同类型数据(多模态数据)的方法已成为智能医疗系统的重要驱动力,其应用研究范围已涵盖从筛查发现、疾病诊断到治疗预后评估的各个方面。

研究结果显示,多模态影像组学模型在预测Ⅰ~Ⅳ期NSCLC患者总生存期(overall survival,OS)方面优于单模态和临床模型[10]。

本研究回顾性收集2014年1月至2019年12月在陆军特色医学中心接受手术切除的Ⅰ期NSCLC患者数据,整合临床基本信息、实验室检查及术前胸部CT影像等多模态数据,构建Ⅰ期肺癌术后5年复发的预后评估模型,旨在筛选术后5年复发高危患者,为个体化辅助治疗及随访策略的制定提供依据。

1 资料与方法

1.1 研究对象

回顾性收集2014年1月至2019年12月陆军特色医学中心收治的778例Ⅰ期NSCLC患者。纳入标准:①患者具有明确的Ⅰ期NSCLC病理学诊断;②具有可获得的术前3个月内的CT图像;③患者具有临床和放射学随访。排除标准:①非R0切除。根据2005年国际肺癌研究协会分期委员会基于既往的定义和认识,提出了肺癌外科手术"完全性切除",即为R0切除;②手术前接受过全身抗肿瘤治疗;③CT图像中有运动伪影(包括呼吸)或金属物体造成的明显条纹伪影;④拒绝手术、缺乏病理结果,或其他原因导致临床资料不全,无法判断具体复发时间;⑤有其他恶性肿瘤病史。

本研究为回顾性观察性研究,根据相关伦理规定,免除了受试者知情同意。在数据处理过程中,已从研究数据中彻底移除了所有可能识别患者身份的信息。本研究严格遵循《赫尔辛基宣言》的相关规定,已通过陆军特色医学中心伦理委员会审批(医研伦审2024第76-02号)。

1.2 临床资料收集

本研究纳入样本的临床资料包括:①基线临床信息:患者人口学特征(年龄、性别),生活方式史(吸烟史、饮酒史),基础疾病状况,以及术后辅助治疗方案等信息;②术前1周内的实验室检查数据:血常规、肿瘤标志物及炎性标志物等信息;③术前3个月内的CT影像资料。

1.3 CT影像学检查

采用GE 64排CT扫描仪对NSCLC患者进行扫描。管电压100 kV,矩阵512×512,图像层厚5 mm。患者取仰卧位,屏气深吸气后扫描,扫描范围为肺尖至双侧肋膈角水平(包括全肺)。通过图像归档及影像系统将CT扫描图像以DICOM格式导出。先由一位初级医师对图像的质量进行评价,如遇到问题,由1名高年资医师再次审核。若2位医师的意见不统一则进行协商,达到一致意见。

1.4 图像预处理和靶区勾画

为了获取局部结构信息,将图像重新采样到1 mm×1 mm×1 mm体素大小。对于各向同性和各向异性图像,采用了最近邻法来抑制重采样产生的伪影。使用z-score标准化方法将图像调整为肺窗(窗宽:1 300 HU,窗位:-300 HU),并将标准化后的强度值裁剪到(-1,1)范围内,以实现模型的快速收敛。将预处理后的图像以NIfTI格式导入到ITK-SNAP(3.8.0)软件中,由一名具有5年临床经验的医师采用平扫图像沿肿瘤边缘逐层手动勾画靶区直至覆盖整个肿瘤。为评估靶区的可用性,由另一位具有15年临床经验的医师检查靶区勾画情况。如果感兴趣区域(ROI)范围受到质疑,则在2人达成一致后重新勾画。

1.5 特征提取和特征筛查

1.5.1 影像学特征及筛选 所有影像组学特征均通过PyRadiomics(http://pyradiomics.readthedocs.io)中内部特征分析程序提取。本研究对所有提取的影像组学特征运用以下3种方式进行降维和筛选:①仅保留统计检验中$P<0.05$的特征。②对于重复性高的特征(图1A),采用皮尔逊相关系数来计算特征之间的相关性,并且在任意2个特征之间保留1个相关系数>0.9的特征。为了最大程度地保留描述特征的能力,使用贪婪递归删除策略进行特征筛选,即每次删除当前集合中冗余度最高的特征。③采用最小绝对收缩和选择算子(Lasso)逻辑回归方法。根据正则化权重λ,Lasso将所有回归系数向零收缩,并将许多不相关特征的系数精确地设置为0。为了找到最优λ,采用10折交叉验证方法,并使用最小标准,其中最终的λ值产生了最小的交叉验证误差。保留的特征(非零系数)被用于回归模型拟合并组合成一个影像组学特征。随后,通过保留特征与其模型系数的线性组合为每位患者获得一个影像组学分数。

1.5.2 临床特征筛选 临床因素基线分析采用统计检验,并进行单因素及多因素Logistic回归分析比较临床特征,选取$P<0.05$的临床参数用于构建临床模型。

1.6 模型构建及评估

经过Lasso回归进行特征筛选后,基于筛选所得的特征构建影像模型,该模型整合了支持向量机(support voctor machines,SVM)、随机森林(RandomForest,RF)、K-最近邻(K-nearest neighbors,KNN)、自适应增强算法(Adaboost)、梯度提升决策树(light gradient boosting

machine, LightGBM)以及多层感知器(multilayer perceptron, MLP)6种机器学习算法。为保持研究设计的可比性,临床预测模型的构建遵循以下原则:首先基于影像模型的性能评估结果选择最优建模策略;其次采用与影像组学研究完全相同的随机队列划分方案进行临床模型的训练与验证。最终通过特征级融合方法,将优化后的影像组学特征与临床特征相结合,建立多维度融合预测模型。

1.7 统计学方法

采用SPSS Statistics(v 26.0)和Python(v 3.7.12)开发定制化分析程序,并在Onekey多模态分析平台(v 2.2.3)中调用其集成的Python工具包:基于scikit-learn(v 1.0.2)构建Lasso特征选择模型;通过PyRadiomics(v 3.0)提取影像组学特征;运用statsmodels(v 0.13.2)完成统计分析。连续变量采用$\bar{x}\pm s$表示,计数资料以例数(%)呈现。统计推断过程包含3个阶段:首先采用独立样本t检验(正态分布数据)、Mann-Whitney U检验(非正态分布数据)和卡方检验(分类变量)进行组间差异分析;通过单因素及多因素Logistic回归分析筛选NSCLC术后5年复发的独立预测因子;通过绘制受试者工作特征(ROC)曲线,计算曲线下面积(AUC)值、灵敏度、特异度等指标,对模型效能进行评估。运用Delong检验定量比较临床模型、影像模型与融合模型之间的预测效能差异。所有统计检验均采用双侧检验,且显著性阈值设定为$P<0.05$(95% *CI*)。

2 结果

2.1 一般临床资料

本研究严格遵循纳入排除标准,最终确立217例Ⅰ期NSCLC研究队列,其中男性123例(56.7%)、女性94例(43.3%),男女比例1.31∶1。观察到5年内有53例(24.4%)患者出现复发转移(原位及同侧肺部转移24例,远处转移29例),未复发患者164例(75.6%)。按8∶2的比例将队列随机划分为训练集(n=173)和验证集(n=44),2组基线特征经统计检验证实具有均衡性(表1)。预后相关因素筛查显示:在训练集内,复发组患者吸烟史比例(59.09% *vs* 40.91%)、碱性磷酸酶(ALP)水平[(92.93±32.32)U/L]及D-二聚体浓度[(147.06±249.32)μg/L]显著高于未复发组($P<0.05$)。单因素

表1 Ⅰ期NSCLC患者手术前基线特征[n(%),$\bar{x}\pm s$]

特征	训练集				验证集				训练集与验证集对比
	全部(n=173)	未复发(n=129)	复发(n=44)	P值	全部(n=44)	未复发(n=35)	复发(n=9)	P值	P值
性别				0.278				1.000	0.058
女	81(46.82)	64(49.61)	17(38.64)		13(29.55)	10(28.57)	3(33.33)		
男	92(53.18)	65(50.39)	27(61.36)		31(70.45)	25(71.43)	6(66.67)		
年龄				0.306				0.702	0.685
<50	23(13.29)	15(11.63)	8(18.18)		3(6.82)	2(5.71)	1(11.11)		
50~59	55(31.79)	38(29.46)	17(38.64)		14(31.82)	10(28.57)	4(44.44)		
60~69	74(42.77)	60(46.51)	14(31.82)		21(47.73)	18(51.43)	3(33.33)		
≥70	21(12.14)	16(12.40)	5(11.36)		6(13.64)	5(14.29)	1(11.11)		
吸烟史				0.047				0.657	0.340
从未	95(54.91)	77(59.69)	18(40.91)		20(45.45)	17(48.57)	3(33.33)		
曾有	78(45.09)	52(40.31)	26(59.09)		24(54.55)	18(51.43)	6(66.67)		
白细胞(10^9/L)	6.14±1.88	6.13±1.97	6.18±1.58	0.492	6.51±2.59	6.28±1.80	7.44±4.58	0.873	0.459
血红蛋白(g/L)	136.59±16.01	136.01±16.30	138.27±15.18	0.530	140.58±15.78	141.71±14.28	136.16±21.08	0.352	0.095
血小板(10^9/L)	202.45±63.45	204.67±63.03	195.93±64.96	0.680	186.37±67.28	189.14±70.96	175.58±52.56	0.600	0.095
白蛋白(g/L)	42.13±3.26	42.02±3.12	42.44±3.67	0.466	41.05±4.37	41.46±4.43	39.46±3.98	0.142	0.246
碱性磷酸酶(U/L)	85.43±28.34	82.87±26.51	92.93±32.32	0.035	82.93±21.21	81.55±21.00	88.30±22.43	0.375	0.977
D-二聚体(μg/L)	120.66±171.54	111.66±135.42	147.06±249.32	0.047	90.37±57.48	82.53±57.43	120.88±49.26	0.069	0.381
癌胚抗原(ng/mL)	3.04±4.92	3.14±5.48	2.75±2.76	0.844	3.68±5.80	3.84±6.41	3.05±2.34	0.694	0.563
术后化疗				0.007				0.771	0.514
无	110(63.58)	90(69.77)	20(45.45)		25(56.82)	19(54.29)	6(66.67)		
有	63(36.42)	39(30.23)	24(54.55)		19(43.18)	16(45.71)	3(33.33)		

Logistic回归分析显示,除冠状动脉粥样硬化和术后化疗外,其余12项候选变量与预后存在潜在关联($P<0.05$,表2);结合多因素分析结果,最终确立年龄、吸烟史、ALP水平和D-二聚体浓度参与临床模型的构建(表3)。

2.2 影像组学特征提取和筛选

2.2.1 影像特征统计 本研究基于影像组学标准化流程,从肿瘤原发灶区域提取7大类共1 834个定量影像组学特征向量。特征架构包含:360个一阶统计特征(描述病灶密度分布)、14个形态特征(量化空间几何特性),以及1 460个高阶纹理特征,其中灰度共生矩阵(gray level cooccurence matrix, GLCM)440个、灰度运行长度矩阵(gray level run length matrix,GLRLM)320个、灰度区域大小矩阵(gray level size zone matrix,GLSZM)320个、邻域灰度差异矩阵(neighbouring gray tone difference matrix,NGTDM)100个和灰度依赖矩阵(gray level dependence matrix,GLDM)280个。特征筛选过程可视化展示如图1A所示:初步获得969个具有统计学意义的特征($P<0.05$);进一步通过相关性分析进行特征去冗余(相关系数阈值>0.9时保留方差较大者),最终纳入179个低共线性、高鉴别力的影像组学特征用于后续建模分析。

2.2.2 Lasso特征选择 通过最小绝对收缩与选择算子(Lasso)逻辑回归模型,选择非零系数以建立影像特征评分(图1B)。通过10折交叉验证,最终得到如图1C所示的均方误差曲线,其中最低点($\lambda=0.059\ 6$)

表2 Ⅰ期NSCLC患者临床特征单因素分析

特征名称	Log(*OR*)	95%*CI*下限	95%*CI*上限	*OR*	95%*CI*下限	95%*CI*上限	*P*值
饮酒史	−0.916	−1.375	−0.458	0.400	0.253	0.633	0.001
性别	−0.879	−1.255	−0.502	0.415	0.285	0.605	0.000
CHD	−0.762	−1.515	−0.009	0.467	0.220	0.991	0.096
吸烟史	−0.693	−1.088	−0.298	0.500	0.337	0.742	0.004
年龄	−0.423	−0.536	−0.309	0.655	0.585	0.734	0.000
癌胚抗原	−0.261	−0.366	−0.157	0.770	0.694	0.855	0.000
白细胞	−0.164	−0.210	−0.117	0.849	0.811	0.890	0.000
NSE	−0.082	−0.105	−0.059	0.922	0.900	0.943	0.000
白蛋白	−0.025	−0.032	−0.018	0.975	0.969	0.982	0.000
碱性磷酸酶	−0.010	−0.014	−0.007	0.990	0.986	0.993	0.000
血红蛋白	−0.008	−0.010	−0.006	0.992	0.990	0.994	0.000
血小板	−0.005	−0.007	−0.004	0.995	0.993	0.996	0.000
D-二聚体	−0.004	−0.006	−0.002	0.996	0.994	0.998	0.005
术后化疗	−0.486	−0.912	−0.059	0.615	0.402	0.943	0.061

CHD:冠状动脉粥样硬化性心脏病;NSE:神经元特异性烯醇化酶

表3 Ⅰ期NSCLC患者临床特征多因素分析

特征名称	Log(*OR*)	95%*CI*下限	95%*CI*上限	*OR*	95%*CI*下限	95%*CI*上限	*P*值
年龄	−0.497	−0.853	−0.141	0.608	0.426	0.868	0.022
饮酒史	−0.383	−1.241	0.475	0.682	0.289	1.608	0.463
NSE	−0.030	−0.085	0.025	0.970	0.919	1.025	0.372
白细胞	−0.057	−0.246	0.131	0.944	0.782	1.140	0.617
血红蛋白	−0.004	−0.025	0.017	0.996	0.975	1.017	0.744
血小板	−0.003	−0.008	0.002	0.997	0.992	1.002	0.380
D-二聚体	0.001	−0.001	0.003	1.001	0.999	1.003	0.420
癌胚抗原	−0.003	−0.075	0.068	0.997	0.928	1.070	0.938
白蛋白	0.008	−0.065	0.080	1.008	0.937	1.083	0.865
碱性磷酸酶	0.013	0.002	0.023	1.013	1.002	1.023	0.042
性别	0.197	−0.764	1.158	1.218	0.466	3.184	0.736
吸烟史	1.036	−0.011	2.083	2.818	0.989	8.029	0.104

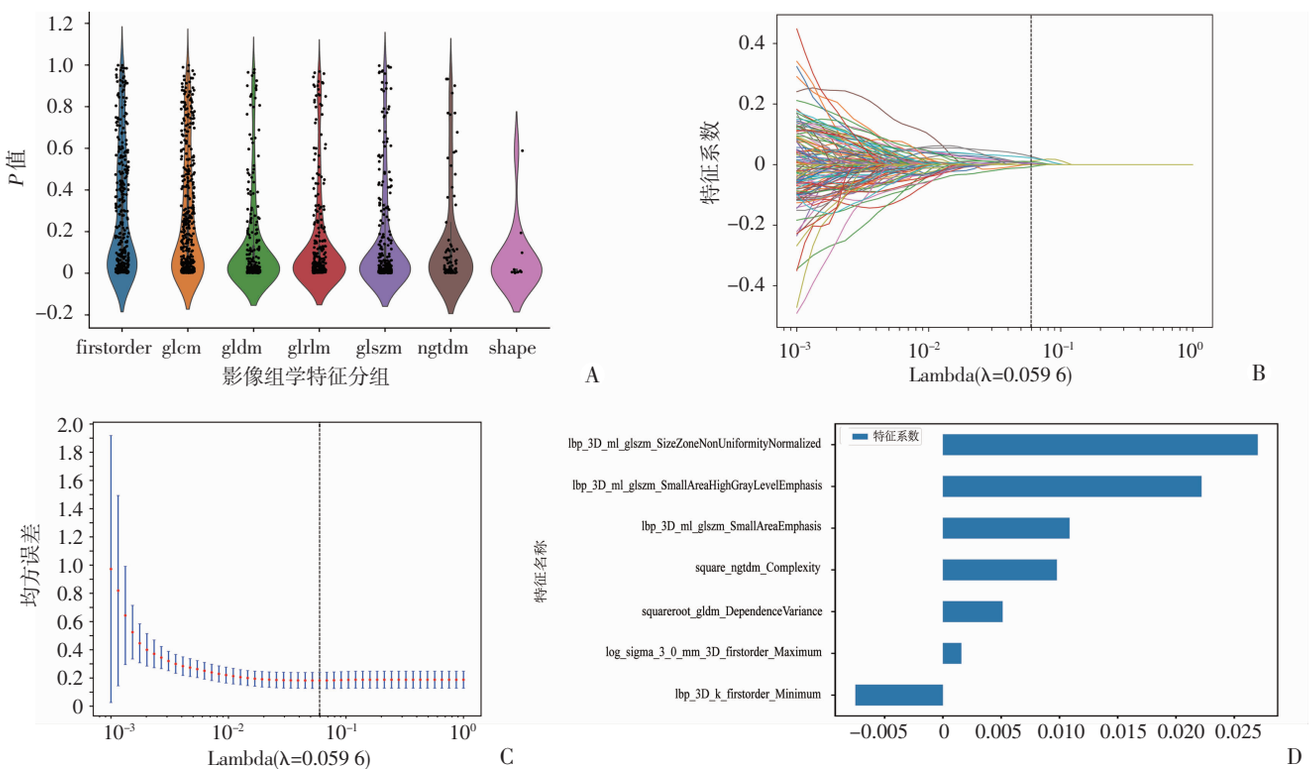
即为最优参数。此时模型在保留关键预测特征(非零系数)的同时,实现最小的预测误差。最终选取7个非零特征进行线性拟合,对应的标准化系数值如图1D所示,图中显示不同特征对Ⅰ期NSCLC术后复发风险预测的贡献方向和强度。特征系数的绝对值越大,提示与预后的相关性越大。

2.3　影像模型的构建与效能

选取6种经典机器学习算法来构建影像模型,应用AUC值、准确度、敏感度、特异度及95%*CI*等指标评估模型的效能。结果显示Adaboost模型在训练集和验证集上均实现了最佳性能,AUC值分别为0.866(95% *CI* 0.808~0.923)和0.806(95% *CI* 0.630~0.983),准确度为0.795,特异度为0.971(图2,表4)。

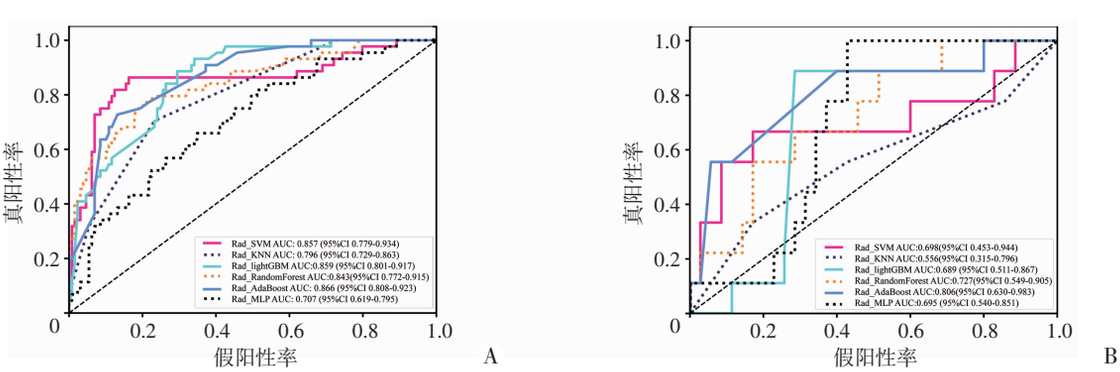
2.4　临床模型构建与效能

共采集14项临床预测变量,涵盖:①人口统计学特征(年龄、性别等);②术前实验室检验指标(D-二聚体、ALP、血常规参数等);③术后辅助治疗信息。经单因素及多因素Logistic回归分析,最终确定年龄、吸烟史、ALP水平和D-二聚体浓度作为4项临床预测模型的构建变量。基于前期影像模型的性能评估结果,采



A:影像学特征统计;B:最小绝对收缩与选择算子特征筛查(10倍交叉验证的系数);不同颜色代表不同特征;C:10折交叉验证的均方误差(MSE);D:基于所选特征的影像特征评分直方图

图1　Ⅰ期NSCLC患者CT影像组学特征的提取及筛查



A:训练集;B:验证集;Adaboost在训练集和测试集中均取得了最佳表现

图2　Ⅰ期NSCLC患者CT影像组学特征在不同模型上的ROC曲线

用Adaboost集成学习算法构建临床预测模型,结果显示:训练集AUC达0.874(95%*CI*:0.821~0.928),验证集AUC=0.813(95%*CI*:0.677~0.948)。模型鉴别性能指标显示,验证集灵敏度与特异度分别为0.778和0.600,提示模型具有较好的预测能力(图3A,表5)。

2.5 融合模型构建与效能

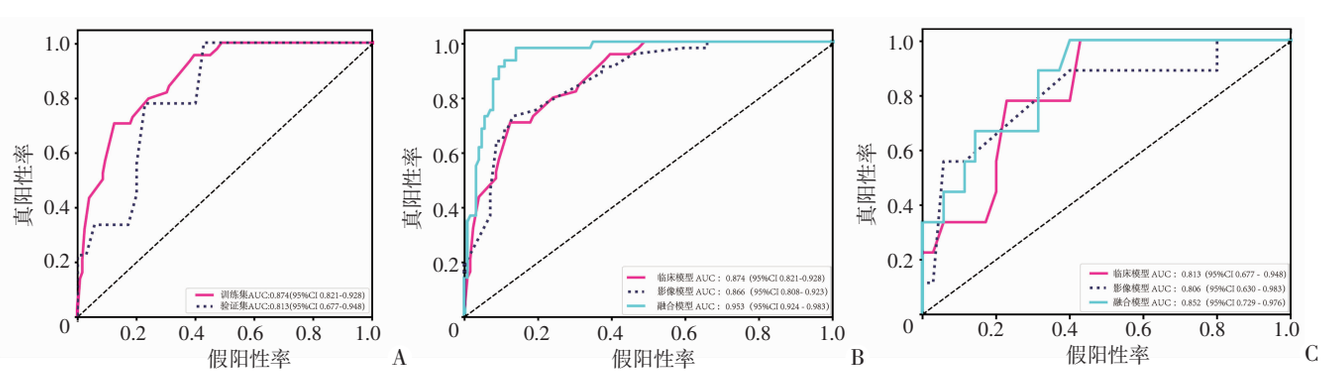
采用基于Adaboost算法的特征融合策略,将优化后的4项临床预测因子与7个影像组学特征进行多维度整合。模型验证结果显示:在训练集和验证集中AUC值分别为0.953(95% *CI* 0.924~0.983)和0.852(95% *CI* 0.729~0.976)(图3B、C,表5)。通过DeLong检验进行模型间效能比较发现:在训练集阶段,融合模型较临床模型(P=0.000)及影像组学模型(P=0.002)均有统计学差异;而在验证集阶段,融合模型同临床模型与影像模型之间的差异无统计学意义。

3 讨论

近年来,影像组学技术逐渐应用于各类肿瘤的诊断和预后的研究中,不仅提高了肿瘤诊断的准确性和效率,还为肿瘤的良恶性鉴别、制定合理的治疗方案提供了重要依据[11-12]。陆亮等[13]通过回顾性研究151例经病理证实为周围型肺癌的患者,分别进行胸部CT扫描,提取其中的影像学特征来构建Logistic回归模型,并以患者性别、年龄、病灶直径、分叶征、瘤体

表4 基于影像组学特征的6种机器学习模型预测性能

模型名称	准确度	AUC	95% *CI*	灵敏度	特异度	阳性预测值	阴性预测值	精密度	召回率	F1分数	置信度阈值
支持向量机											
训练集	0.838	0.857	0.779~0.934	0.841	0.837	0.638	0.939	0.638	0.841	0.725	0.249
验证集	0.773	0.698	0.453~0.944	0.556	0.829	0.455	0.879	0.455	0.556	0.500	0.255
K-最近邻											
训练集	0.786	0.796	0.729~0.863	0.318	0.946	0.667	0.803	0.667	0.318	0.431	0.400
验证集	0.795	0.556	0.315~0.796	0.000	1.000	0.000	0.795	0.000	0.000	NaN	0.600
梯度提升决策树											
训练集	0.746	0.859	0.801~0.917	0.864	0.705	0.500	0.938	0.500	0.864	0.633	0.275
验证集	0.636	0.689	0.511~0.867	0.222	0.743	0.182	0.788	0.182	0.222	0.200	0.312
随机森林											
训练集	0.786	0.843	0.772~0.915	0.750	0.798	0.559	0.904	0.559	0.750	0.641	0.264
验证集	0.750	0.727	0.549~0.905	0.444	0.829	0.400	0.853	0.400	0.444	0.421	0.335
自适应增强算法											
训练集	0.832	0.866	0.808~0.923	0.682	0.884	0.667	0.891	0.667	0.682	0.674	0.477
验证集	0.795	0.806	0.630~0.983	0.111	0.971	0.500	0.810	0.500	0.111	0.182	0.498
多层感知器											
训练集	0.647	0.707	0.619~0.795	0.636	0.651	0.384	0.840	0.384	0.636	0.479	0.305
验证集	0.636	0.695	0.540~0.851	0.889	0.571	0.348	0.952	0.348	0.889	0.500	0.308



A:临床模型训练集和验证集AUC值;B:临床模型、影像模型和融合模型训练集ROC曲线对比;C:临床模型、影像模型和融合模型验证集ROC曲线对比

图3 临床模型和融合模型的ROC曲线

表5 Ⅰ期NSCLC患者临床、影像和融合3种模型预测性能

模型名称	准确度	AUC	95%*CI*	灵敏度	特异度	阳性预测值	阴性预测值	精密度	召回率	F1分数	置信度阈值
临床模型											
训练集	0.827	0.874	0.821~0.928	0.636	0.891	0.667	0.878	0.667	0.636	0.651	0.483
验证集	0.636	0.813	0.677~0.948	0.778	0.600	0.333	0.913	0.333	0.778	0.467	0.452
影像模型											
训练集	0.832	0.866	0.808~0.923	0.682	0.884	0.667	0.891	0.667	0.682	0.674	0.477
验证集	0.795	0.806	0.630~0.983	0.111	0.971	0.500	0.810	0.500	0.111	0.182	0.498
融合模型											
训练集	0.884	0.953	0.924~0.983	0.955	0.860	0.700	0.982	0.700	0.955	0.808	0.286
验证集	0.682	0.852	0.729~0.976	0.889	0.629	0.381	0.957	0.381	0.889	0.533	0.220

密度及胸膜牵拉征6个变量来构建CT临床特征模型;将单独采用CT临床特征模型与2种特征相结合的多模态模型相比较,结果发现影像组学联合CT特征模型预测精度较好,这提示多模态模型具有更加准确的预测价值。本研究分别基于平扫CT图像和临床资料构建影像模型和临床模型,并将影像特征与临床风险因素相融合,形成了一个用于预测Ⅰ期NSCLC 5年复发风险的融合模型。通过多种模型间的比较,选出最优模型对Ⅰ期NSCLC术后5年复发风险进行预测。

在特征选择方面,与传统影像学存在的主观性和误差相比,影像组学通过自动化的方式快速提取大量强度、形状、纹理等多种定量特征,减少了人为因素的干扰,提高了分析的准确性[14]。本研究确定了一系列具有高稳定性和可重复性的特征,包括一阶统计特征、灰度区域大小矩阵、邻域灰度差矩阵和灰度依赖矩阵等,这些特征对于医师来说易于比较和使用。在平扫CT图像基础上,本研究从1 834个影像组学特征中运用Lasso等方法筛选出7个关键特征,其中灰度区域大小矩阵的特征权重最高,影像组学方法在提取肉眼无法观察到的高维定量图像特征方面具有重要作用。

ROC曲线分析显示,不同算法构建的影像模型预测效能各不相同。其中,Adaboost模型在训练集和验证集中的AUC值分别为0.866和0.806,表现最优,拟合性较好。Adaboost是一种集成学习算法,其核心思想是针对同一个训练集训练不同的分类器(弱分类器),然后把这些弱分类器集合起来,构成一个更强的最终分类器(强分类器),更侧重于降低分类偏差[15]。在多个影像组学研究中,Adaboost模型被广泛应用于预测肿瘤的分期和疗效评价等领域,并取得了较好的效果[16-18]。

本研究通过多因素Logistic回归分析,确立年龄和ALP水平为Ⅰ期NSCLC患者术后复发的独立预测因子。其生物学机制可解释为:年龄增长导致的生理功能衰退和免疫监视能力下降,已被多项研究证实与肿瘤复发风险呈正相关;ALP是一种磷酸单酯水解酶,在碱性条件下起催化水解和去除磷酸基团的作用[19]。血清ALP在某些病理条件下,尤其是骨转移时,呈现水平升高趋势,与生存期缩短相关[20]。临床基线分析显示,吸烟史、ALP和D-二聚体3个特征在复发与未复发的患者间存在统计学差异(P<0.05)。吸烟是导致肺癌发生的主要危险因素,虽然没有证据表明烟草的主要成分之一尼古丁本身会诱发癌症,但不同的研究已经确定尼古丁在体内会促进癌细胞的生长和内皮细胞的增殖[21]。张凯等[22]研究显示,术前D-二聚体水平可能是评估根治术后NSCLC患者预后的指标之一,随着D-二聚体水平增高,预后表现越来越差。本研究最终将4个临床特征用于建模。临床模型在训练集和验证集中的AUC值分别为0.874和0.813,灵敏度为0.778,特异度为0.600,说明纳入模型构建的4个临床特征能较好地对NSCLC预后进行预测。

受试者工作曲线结果显示,融合模型的AUC值在训练集和验证集中均高于影像模型或临床模型。然而,Delong检验结果提示融合模型在验证集中与影像模型和临床模型无明显差异,原因可能有以下几个方面:①特征冗余与冲突:在融合模型中,影像组学特征(如纹理、形状等)与临床特征(如年龄、ALP)可能表征不同的病理维度,若融合时未有效筛选互补特征,可能导致特征空间重叠或冲突,导致某些特征对预测结果的贡献度不高或产生负面效应。②融合策略局限性:融合过程中可能未考虑到特征的重要性和权重,导致信息整合不充分。影像组学特征与临床特征间的交互作用可能具有更复杂的非线性特性。③数据量不足:临床特征通常维度低且样本量有限,影像组

学模型训练数据量远大于临床模型,融合时可能因数据分布差异导致泛化性下降。同时需结合AUC、F1-score等指标综合评估模型效能。

本研究存在以下局限性:①影像特征的提取仅基于平扫图像,未纳入动脉期及静脉期影像学特征,可能遗漏与血流动力学相关的关键信息;②当前构建的融合模型尚未进行外部验证,需扩大样本量并进行多中心验证以提升模型的泛化能力;③研究设计未整合术中实时参数与术后组织病理学数据,可能影响预测系统的临床适配性;④在模型构建过程中,未探索其他融合策略以优化模型性能。

综上所述,本研究结果表明,基于CT影像组学和临床特征的多模态数据模型对于预测Ⅰ期NSCLC术后5年复发风险具有较高的价值和可重复性,这有助于临床医师进行有效的风险评估,并指导Ⅰ期肺癌患者的术后分层管理。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] SUNG H, FERLAY J, SIEGEL R L, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2021, 71(3): 209-249. DOI: 10.3322/caac.21660.

[2] 中华人民共和国国家卫生健康委员会. 原发性肺癌诊疗指南(2022年版)[J]. 中国合理用药探索, 2022, 19(9): 1-28. DOI: 10.3969/j.issn.2096-3327.2022.09.002.
National Health Commission of the People's Republic of China. Guidelines for the diagnosis and treatment of primary lung cancer (2022 edition)[J]. Chin J Ration Drug Use, 2022, 19(9): 1-28. DOI: 10.3969/j.issn.2096-3327.2022.09.002.

[3] SINGAREDDY A, FLANAGAN M E, SAMSON P P, *et al*. Trends in stage I lung cancer[J]. Clin Lung Cancer, 2023, 24(2): 114-119. DOI: 10.1016/j.cllc.2022.11.005.

[4] RUAN Z G, ZHUO X, XU C Y. Diagnosis, treatment, and prognosis of stage IB non-small cell lung cancer with visceral pleural invasion[J]. Front Oncol, 2024, 13: 1310471. DOI: 10.3389/fonc.2023.1310471.

[5] SIEGEL R L, MILLER K D, WAGLE N S, *et al*. Cancer statistics, 2023[J]. CA Cancer J Clini, 2023, 73(1): 17-48. DOI: 10.3322/caac.21763.

[6] AERTS H J W L. The potential of radiomic-based phenotyping in precision medicine: a review[J]. JAMA Oncol, 2016, 2(12): 1636-1642. DOI: 10.1001/jamaoncol.2016.2631.

[7] WU G Y, WOODRUFF H C, SHEN J, *et al*. Diagnosis of invasive lung adenocarcinoma based on chest CT radiomic features of part-solid pulmonary nodules: a multicenter study[J]. Radiology, 2020, 297(2): 451-458. DOI: 10.1148/radiol.2020192431.

[8] 赵媛,王洪峰,赵林,等.能谱CT影像组学非小细胞肺癌淋巴结转移的预测模型构建[J]. 中国临床医学影像杂志, 2024, 35(9): 628-632. DOI: 10.12117/jccmi.2024.09.005.
ZHAO Y, WANG H F, ZHAO L, *et al*. Construction of a predictive model for lymph node metastasis in non-small cell lung cancer based on spectral CT radiomics[J]. J China Clin Med Imag, 2024, 35(9): 628-632. DOI: 10.12117/jccmi.2024.09.005.

[9] SAAD M B, SHOWKATIAN E, AL-TASHI Q, *et al*. I-SABR-SELECT: a radiomics-based model for personalized immunotherapy for early-stage non-small cell lung cancer[J]. Int J Radiat Oncol, 2024, 120(2): S42-S43. DOI: 10.1016/j.ijrobp.2024.07.064.

[10] AMINI M, NAZARI M, SHIRI I, *et al*. Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma[J]. Phys Med Biol, 2021, 66(20): 205017. DOI: 10.1088/1361-6560/ac287d.

[11] 王述平. 基于MRI影像组学的肝癌病型及分化等级预测研究[D]. 保定: 河北大学, 2023. DOI: 10.27103/d.cnki.ghebu.2023.000672.
WANG S P. Prediction of liver cancer type anddifferentiation grade based on MRI radiomics[D]. Baoding: Hebei University, 2023. DOI: 10.27103/d.cnki.ghebu.2023.000672.

[12] 李旭. CT影像组学在食管癌淋巴结转移及浸润深度预测中的研究[D]. 济南: 山东大学, 2023.
LI X. A computed tomography-based clinical-radiomics model for prediction of lymph node metastasis and Infiltration depth in esophageal carcinoma[D]. Jinan: Shandong University, 2023.

[13] 陆亮, 徐圆. 影像组学联合CT特征对周围型肺腺癌及鳞癌的鉴别价值研究[J]. 医疗卫生装备, 2021, 42(10): 48-52, 63. DOI: 10.19745/j.1003-8868.2021211.
LU L, XU Y. Diagnostic value of radiomics combined with CT features for differentiating peripheral lung adenocarcinoma from squamous cell carcinoma[J]. Chin Med Equip J, 2021, 42(10): 48-52, 63. DOI: 10.19745/j.1003-8868.2021211.

[14] 师庆科,李楠,叶枫. 机器学习在临床研究中的应用进展[J]. 中国数字医学, 2025, 20(3): 1-10. DOI: 10.3969/j.issn.1673-7571.2025.03.001.
SHI Q K, LI N, YE F. Advancements in the application of

machine learning for clinical research[J].China Digit Med, 2025, 20(3): 1-10. DOI: 10.3969/j.issn.1673-7571.2025.03.001.

[15] SONG F, SONG X, FENG Y D, *et al*. Radiomics feature analysis and model research for predicting histopathological subtypes of non-small cell lung cancer on CT images: a multi-dataset study[J]. Med Phys, 2023, 50(7): 4351-4365. DOI: 10.1002/mp.16233.

[16] 刘婧娟, 刘炜, 薛华丹, 等. 基于增强CT的影像组学定量特征模型在预测胃癌分期中的价值[J]. 放射学实践, 2022, 37(9): 1061-1067. DOI: 10.13609/j.cnki.1000-0313.2022.09.002.
LIU J J, LIU W, XUE H D, *et al*. The value of contrast enhanced CT-based radiomics analysis in predicting gastric cancer stage[J]. Radiol Pract, 2022, 37(9): 1061-1067. DOI: 10.13609/j.cnki.1000-0313.2022.09.002.

[17] 黄蓉, 卢晓旭, 孙学明, 等. ^{18}F-FDG PET/CT联合临床建立多模态影像组学模型预测局部晚期头颈部鳞癌对于新辅助免疫联合化学治疗的反应[J]. 中国医学影像技术, 2024, 40(10): 1493-1498. DOI: 10.13929/j.issn.1003-3289.2024.10.008.
HUANG R, LU X X, SUN X M, *et al*. Multimodal models established combined ^{18}F-FDG PET/CT radiomics with clinical data for evaluating response of locally advanced head and neck squamous cell carcinoma to neoadjuvant immuno-chemotherapy [J]. Chin J Med Imag Technol, 2024, 40(10): 1493-1498. DOI: 10.13929/j.issn.1003-3289.2024.10.008.

[18] 贺安晶. 基于CT影像组学预测晚期肺腺癌靶向治疗短期疗效[D]. 南昌: 南昌大学, 2024. DOI: 10.27232/d.cnki.gnchu.2024.003659.
HE A J. Prediction of short-term efficacy of targeted therapy for advanced lung adenocarcinoma based on CT radiomics [D]. Nanchang: Nanchang University, 2024. DOI: 10.27232/d.cnki.gnchu.2024.003659.

[19] HARRIS H. The human alkaline phosphatases: what we know and what we don't know[J]. Clin Chim Acta, 1990, 186(2): 133-150. DOI: 10.1016/0009-8981(90)90031-m.

[20] ZHANG L, GONG Z X. Clinical characteristics and prognostic factors in bone metastases from lung cancer[J]. Med Sci Monit, 2017, 23: 4087-4094. DOI: 10.12659/msm.902971.

[21] CATASSI A, SERVENT D, PALEARI L, *et al*. Multiple roles of nicotine on cell proliferation and inhibition of apoptosis: implications on lung carcinogenesis [J]. Mutat Res, 2008, 659(3): 221-231. DOI: 10.1016/j.mrrev.2008.04.002.

[22] 张凯, 李树建, 陈小乐. 血浆D-二聚体水平在非小细胞肺癌患者预后评估中的价值[J]. 医学综述, 2022, 28(5): 1032-1035. DOI: 10.3969/j.issn.1006-2084.2022.05.035.
ZHANG K, LI S J, CHEN X L. Value of plasma D-dimer level in prognostic assessment of patients with non-small cell lung cancer[J]. Med Recapitul, 2022, 28(5): 1032-1035. DOI: 10.3969/j.issn.1006-2084.2022.05.035.

(**收稿**:2024-10-29;**修回**:2025-03-19)

(编辑 张 维)