·论 著·

基于CT影像特征的非小细胞肺癌复发相关性因素研究

鲁晓腾，许 青

复旦大学附属肿瘤医院放疗科，复旦大学上海医学院肿瘤学系，上海 200032

［摘要］ **背景与目的：**基于CT图像特征对非小细胞肺癌（non-small cell lung cancer，NSCLC）患者的复发相关性因素进行探究。**方法：**选用NSCLC-Radiogenomics数据库中的157组数据。首先，对肺部肿瘤及其图像特征进行提取；然后，使用独立样本*t*检验对特征数据进行单因素分析，使用logistic回归模型进行进一步分析，得到NSCLC复发情况的显著性相关因素；其次，使用Z-score标准化方法对数据进行标准化处理，采用合成少数过采样技术（synthetic minority over-sampling technique，SMOTE）算法对标准化后的数据进行平衡化操作；最后，利用随机森林、K最邻近算法（K-nearest neighbor，KNN）、支持向量机（support vector machine，SVM）、决策树算法以及留一交叉验证方法训练分类器并检验相关性因素对患者复发情况的预测能力。**结果：**独立样本*t*检验分析结果显示，Variance、Energy、Relative message、和熵以及Coarseness与NSCLC复发情况相关（$P<0.05$）。Logistic回归分析显示，Energy及和熵与NSCLC复发情况显著相关（$P<0.05$），分类器分类结果显示最高分类准确率为82.7%，最大曲线下面积（area under curve，AUC）为0.891，即这两种特征可以对患者复发情况作出较为准确的预测。**结论：**Energy以及和熵是非小细胞肺癌复发的显著性相关因素。

［关键词］ 非小细胞肺癌；图像特征；复发；分类器

DOI: 10.19401/j.cnki.1007-3639.2020.08.012

中图分类号：R734.2 **文献标志码：**A **文章编号：**1007-3639(2020)08-0636-05

A study on factors associated with recurrence of non-small cell lung cancer based on CT image features LU Xiaoteng, XU Qing (Department of Radiation Oncology, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China)

Correspondence to: XU Qing E-mail: qingxu68@hotmail.com

［**Abstract**］ **Background and purpose:** The purpose of this paper was to explore the factors associated with non-small cell lung cancer (NSCLC) patient's recurrence situation based on CT image features. **Methods:** A hundred and fifty-seven sets of data collected in NSCLC radiogenomics database were used in the experiment. The lung tumors were segmented, and image features were extracted. Independent samples *t* test was used to perform a univariate analysis. And logistic regression model was used to obtain the significant factors associated with NSCLC recurrence. *Z*-score normalization and synthetic minority over-sampling technique (SMOTE) methods were used to analyze data. Finally, random forest, K-nearest neighbor (KNN), support vector machine (SVM), decision-tree and leave-one-out cross validation were used to train classifier and test the validity of results. **Results:** The independent samples *t* test showed that Variance, Energy, Relative message, Add-entropy and Coarseness were related to NSCLC recurrence ($P<0.05$). And the logistic regression analysis showed that Energy and Add-entropy were significantly correlated with NSCLC recurrence ($P<0.05$). Furthermore, the classification results revealed that the best accuracy was 82.7% and the maximum area under curve (AUC) was 0.891. These two features could make a well prediction for NSCLC patient's recurrence. **Conclusion:** Energy and Add-entropy were the factors significantly associated with NSCLC recurrence.

［**Key words**］ Non-small cell lung cancer; Image features; Recurrence; Classifiers

在过去50年中，肺癌是全球发病率和死亡率增长最快的恶性肿瘤，并稳居中国恶性肿瘤之首。其中，非小细胞肺癌（non-small cell lung cancer，NSCLC）占肺癌患病总人数的80%~

通信作者：许 青 E-mail: qingxu68@hotmail.com

85%[1-2]。随着科技的进步和医疗的发展，NSCLC患者的存活状况得到了一定程度的改善，但绝大多数患者初诊时已属Ⅲ~Ⅳ期，所以其5年生存率只有15%[3]。NSCLC的一个显著的生物学特征是复发。约50%的NSCLC患者在术后5年内发生肿瘤的复发，其中大部分在2年内发生。由此可知，肿瘤复发是影响NSCLC患者预后的重要因素之一。因此，对影响肿瘤复发的因素进行探究在临床上具有重要的指导意义。

在现有的研究中，少有文献表明在统计学分析得到相关性因素后会补充分类实验对结果进行验证。故本文在提取图像特征基础上，采用严谨的统计学分析方法得到NSCLC复发的相关性因素，进而训练分类器对统计学分析实验结果进行验证。

1 材料和方法

1.1 实验材料

本研究所用数据均来自癌症影像档案（The Cancer Imaging Archive，TCIA）公共访问中的NSCLC Radiogenomics数据库[4]。本实验使用数据库中157例NSCLC患者的临床数据、治疗前的CT图像数据以及治疗后患者的复发情况等。其中，男性95例，女性62例；肺腺癌136例，肺鳞癌21例；复发患者30例，非复发患者127例。图像数据包括157组CT序列图像，每幅图像的大小为512像素×512像素。实验平台为64位Window 10操作系统，i7-4770-3.4 GHz CPU，16 GB内存；使用的专业软件是Matlab 2015a、SPSS 22.0以及Weka 3.8。

1.2 方法

首先，使用区域生长法、数学形态学方法以及数据库中提供的图像数据提取患者图像的感兴趣区（region of interest，ROI）。然后，提取ROI的直方图统计特征、形态学特征和纹理特征。接着，特征数据进行统计学分析得到NSCLC复发的显著性相关因素。最后，使用合成少数过采样技术（Synthetic Minority Over-sampling Technique，SMOTE）算法对数据进行平衡化处理，进而训练分类器对实验结果进行验证。实验方法流程图见图1。

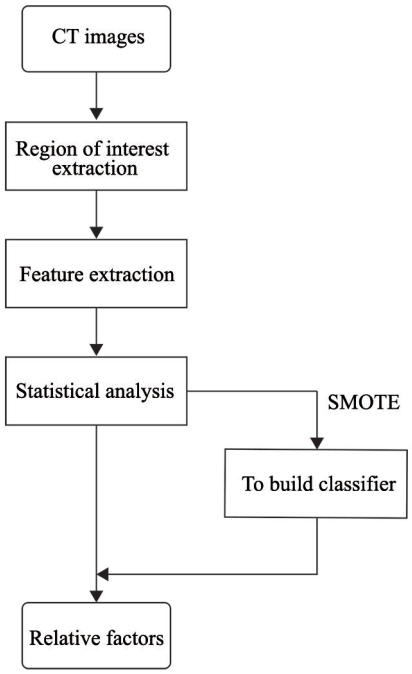

图 1 方法流程图

Fig. 1 Method flow chart

1.2.1 特征提取

根据图像特征定义及计算公式提取图像特征，其中包括直方图统计特征、形态学特征和纹理特征。本文提取上述3类特征共计65种，具体特征如下所示：

直方图统计特征：Mean、Variance、Skewness、Kurtosis、Energy、Entropy。

形态学特征：面积、离心率、紧凑度、矩形度、径向均值，径向方差等。

纹理特征：GLCM纹理特征、NGTDM纹理特征、差分统计纹理特征、小波纹理特征等。

1.2.2 统计学处理

本文采用SPSS 22.0中的独立样本t检验和logistic回归模型对特征数据进行分析。在独立样本t检验中，将所有特征数据作为输入变量进行单因素分析，得到每种特征和患者复发情况的相关性。在得到单因素分析结果后，将与复发情况相关（P<0.05）的特征数据作为logistic回归分析的输入协变量输入到分析模型中，进而得到与患者复发显著相关（P<0.05）的特征变量。

1.2.3 建立分类器

实验中包括复发组患者30例，非复发组患者127例。若用此不平衡数据来训练分类器，得到

的分类器健壮性不高，泛化性不强，分类结果也不准确。所以需要对特征数据进行平衡化处理。经过分析后，本文选用SMOTE算法对数据进行平衡处理。SMOTE算法是由Chawla等[5]提出的一种利用原始样本合成少数类样本的过采样方法。具体算法流程如下：

① 求出少数类样本中每个样本的K个最邻近样本；

② 根据不同类样本数量，确定采样倍率N；

③ 随机从少数类样本中选择N个样本作为父代样本（可重复选取）；

④ 在每个父代样本和某一最邻近样本中进行插值，生成新的子代样本。

数据平衡化处理后，采用Z-Score标准化方法对数据进行标准化处理，该方法根据原始数据的标准差和平均值进行标准化，经过标准化后的数据符合标准正态分布，具体公式如式（1.1）所示（μ为数据组的平均值，σ为数据组的标准差）：

$$X^{*}=\frac{X-\mu}{\sigma} \qquad (1.1)$$

本实验使用随机森林法、K最邻近算法、支持向量机以及决策树方法训练分类器对数据进行分类实验。除此之外，为了充分利用有限的数据，本实验使用留一交叉验证的方式进行分类实验。虽然该方法计算量较大，但是数据利用率高，更适合小样本数据的分类实验。

2　结　　果

2.1　独立样本*t*检验

根据患者在随访期间是否发生复发将其分为两组（复发组30例，非复发组127例）。将65种特征数据和患者复发情况分别作为检验变量和分组变量输入到独立样本*t*检验模型中，经分析后发现，Variance、Energy、Relative message、和熵以及Coarseness与患者的复发情况相关（$P<0.05$，表1）。

表 1　独立样本*t*检验分析结果

Tab. 1　The results of independent sample *t*-test

Variable	*P* value	*SE*
Variance	0.041	0.158
Energy	0.042	0.204
Relative message	0.014	0.147
Add-entropy	0.046	0.203
Coarseness	0.037	0.162

2.2　Logistic回归分析

将表2.1中的5种特征数据作为输入协变量输入到logistic回归分析模型中，同时将患者复发情况数据作为分组变量输入。分析后发现，和熵以及Energy两个特征与NSCLC患者复发情况显著相关（$P<0.05$，表2）。

表 2　Logistic回归分析结果

Tab. 2　The results of logistic regression analysis

Variable	*B*	*SE*	*Wald*	*P* value	Exp (B)	95% Exp (B)	CI
Energy	0.389	0.195	3.979	0.046	1.475	1.007	2.162
Add-entropy	-1.994	0.952	4.390	0.036	0.136	0.021	0.879

2.3　分类器检测

将SMOTE中的采样倍率*N*分别设置为0、2、3和4对数据进行平衡化操作，以便分析不同采样倍率对分类结果的影响。采用四种不同分类算法和留一交叉验证的方法训练分类器进行分类实验，我们将根据准确率、灵敏度、特异度和曲线下面积（area under curve，AUC）这4个指标来评判分类结果的质量，具体结果见表3~6。

表 3　随机森林分类结果

Tab. 3　The results of random forest classification

Item	*N*=0	*N*=2	*N*=3	*N*=4
Accuracy	0.758	0.779	0.810	0.816
Sensitivity	0.913	0.787	0.780	0.740
Specificity	0.100	0.767	0.842	0.880
AUC	0.581	0.849	0.886	0.891

AUC: Area under curve

表 4 KNN分类结果

Tab. 4 The results of KNN classification

Item	*N*=0	*N*=2	*N*=3	*N*=4
Accuracy	0.694	0.779	0.818	0.827
Sensitivity	0.827	0.748	0.748	0.732
Specificity	0.133	0.822	0.892	0.907
AUC	0.480	0.768	0.815	0.834

KNN: K-nearest neighbor

表 5 SVM分类结果

Tab. 5 The results of SVM classification

Item	*N*=0	*N*=2	*N*=3	*N*=4
Accuracy	0.809	0.594	0.628	0.592
Sensitivity	1.000	0.945	0.638	0.378
Specificity	0.000	0.100	0.617	0.773
AUC	0.500	0.522	0.627	0.576

SVM: Support vector machine

表 6 决策树分类结果

Tab. 6 The results of decision-tree classification

Item	*N*=0	*N*=2	*N*=3	*N*=4
Accuracy	0.809	0.719	0.765	0.744
Sensitivity	1.000	0.843	0.724	0.606
Specificity	0.000	0.544	0.808	0.860
AUC	0.078	0.759	0.807	0.799

3 讨 论

目前针对NSCLC复发相关因素的研究，主要从以下四个方面开展，即临床病理学特征、肿瘤血液标志物、特殊基因及图像特征。Park等[6]通过对171例NSCLC患者的临床病理学特征和复发情况进行单、多因素分析后发现，T分期、N分期、病理学分期及淋巴管浸润是患者术后复发的独立危险因素（P<0.05）。McFarlane等[7]结合t检验、Kaplan-Meier生存分析及COX回归模型对影响NSCLC患者复发因素进行研究后发现，去泛素化酶-USP 17是患者复发的独立危险因素。Perumal等[8]通过log-rank检验和Kaplan-Meier检验对NSCLC患者的基因信息进行分析发现，染色体缩聚相关的基因与患者的不良预后及复发情况显著相关。Ko等[9]提取并分析145例患者的^{18}F-FDG PET及CT图像形态学特征后发现，这两类特征在早期NSCLC患者的复发情况预测方面可以起到关键作用。同时，部分结论也在Pyka等[10]的工作中得到验证。

直方图统计特征是根据ROI区域的灰度分布直方图提取的一系列特征，与非小细胞肺癌的预后有密切的联系[11]。形态特征是一类较为直观的特征。在临床方面，医师常根据肿瘤的分叶征、毛刺征、空洞征等形态学征象对患者的预后情况进行判断。同时，Ko等[9]的研究也证明了形态特征与NSCLC患者的复发预后具有较强的相关性。纹理特征包括物体表面的性质和结构，也在一定程度上反映物质与周围环境的关系[12]。有研究亦表明纹理特征与NSCLC患者的术后复发情况及存活时间等预后信息具有较强相关性[10]。

本文基于NSCLC患者的CT图像特征设置回顾性分析实验来研究患者复发的显著性相关因素，并用其训练分类器对患者的复发情况进行预测来验证实验结果的正确性。本研究有如下创新点。首先，本研究提取了可量化的CT图像特征作为研究对象，量化的图像特征可以更加精准有效地反映图像特性。将这些量化特征输入到统计学分析模型中，可以分析得到更加准确的相关性因素。其次，可量化的图像特征同样有利于分类器的训练和预测分类。通过独立样本t检验发现，Variance、Energy、Relative message、和熵及Coarseness与NSCLC患者的复发情况有关。同时，经logistic回归分析后发现，和熵及Energy与患者的复发情况显著相关。

其次，本实验设置了分类器检测实验对统计学分析结果的可靠性进行验证。从分类器的分类结果来看，我们可以根据4个参考量的值来判断分类结果的质量，分别是准确率，灵敏度，特异性和AUC，其中最高的准确率为82.7%，最大AUC为0.891，说明通过统计学分析筛选出的显著性相关因素可以较好地对患者复发情况进行预测。

除此之外，我们采用了SMOTE的数据处理

方法解决了数据不平衡的问题，训练了更为准确、稳定的分类器。当*N*=0时，分类结果具有较高准确率；但分类结果的特异性和AUC最大仅有0.133和0.581，这表明当*N*=0时，分类器的分类结果并无实际意义。另外，表中数据显示，当*N*=2、3或4时，分类器的分类结果会更加精准。其中，随机森林算法和K最邻近算法（K-nearest neighbor，KNN）算法分类结果显示，当*N*=4时，分类器具有更好的分类结果，分类准确率为81.6%和82.7%；支持向量机（support vector machine，SVM）算法和决策树算法分类结果显示，当*N*=3时，分类器具有更好的分类结果，分类准确率为62.8%和76.5%。将4组分类结果进行对比可以发现，随机森林算法、KNN算法、决策树算法建立的分类模型较SVM算法分类模型具有更好的分类效果，分类更为精准，结果更加稳定。综合来看，随机森林算法建立的分类模型具有最好的分类效果。

最后，为了高效地利用有限的数据，本实验采用留一交叉验证的方法完成分类实验。通过该方法，几乎所有的实验数据都用来训练分类器，同时在实验过程中不存在随机因素。因此，该方法能训练出更合适的分类器进而得到更为准确的实验结果。

虽然得到了较好的实验结果，但本研究依然存在一些不足。首先，本文为回顾性研究，仅使用了157例患者的临床及影像学数据，所以实验结果的健壮性及稳定性需要更多的回顾性及前瞻性研究来验证。其次，本研究使用的实验数据来自TCIA公共数据集，实验数据并没有包含当地医院的实验数据。在获取了医院的相关数据后，可以在下阶段实验中将患者的血液标志物和基因信息等因素和CT图像特征相结合进行更深层次的研究。除此之外，虽然我们提取了65种图像特征，囊括了图像的三类特征，但是特征的总量依然有待提高。最后，本研究只是一个技术发展性研究，其具体临床价值以及如何辅助临床医师作出准确的预后判断还需要进一步的实验证明。

通过统计学分析和分类实验的验证，Energy及和熵两种图像特征与NSCLC患者的复发情况显著相关。

［参　考　文　献］

［1］ ALBERG A J, BROCK M V, FORD J G, et al. Epidemiology of lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines［J］. Chest, 2013, 143(5): E1-E29.

［2］ CHEN W, ZHENG R, BAADE P D, et al. Cancer statistics in China, 2015［J］. CA Cancer J Clin, 2016, 66(2):115.

［3］ PASTOR M D, NOGAL A, MOLINA-PINELO S, et al. Identification of proteomic signatures associated with lung cancer and COPD［J］. J Proteomics, 2013, 89(16): 227.

［4］ PRIOR F W, CLARK K, COMMEAN P, et al. TCIA: an information resource to enable open science［J］. Conf Proc IEEE Eng Med Biol Soc, 2012, 2013(2013): 1282-1285.

［5］ CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique［J］. J Art Intelligence Res, 2002, 16(1): 321-357.

［6］ PARK C, LEE I J, JANG S H, et al. Factors affecting tumor recurrence after curative surgery for NSCLC: impacts of lymphovascular invasion on early tumor recurrence［J］. J Thorac Dis, 2014, 6(10):1420.

［7］ MCFARLANE C, MCFARLANE S, PAUL I, et al. The deubiquitinating enzyme USP17 is associated with nonsmall cell lung cancer (NSCLC) recurrence and metastasis［J］. Oncotarget, 2013, 4(10): 1836-1843.

［8］ PERUMAL D, SINGH S, YODER S J, et al. A novel five gene signature derived from stem-like side population cells predicts overall and recurrence-free survival in NSCLC［J］. PLoS One, 2012, 7(8): e43589.

［9］ KO K H, HSU H H, HUANG T W, et al. Predictive value of ^{18}F-FDG PET and CT morphologic features for recurrence in pathological stage ⅠA non-small cell lung cancer［J］. Medicine, 2015, 94(3): e434.

［10］ PYKA T, BUNDSCHUH R A, ANDRATSCHKE N, et al. Textural features in pre-treatment［F18］-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy［J］. Radiation Oncol, 2015, 10(1): 100.

［11］ KAMIYA A, MURAYAMA S, KAMIYA H, et al. Kurtosis and skewness assessments of solid lung nodule density histograms: differentiating malignant from benign nodules on CT［J］. Jpn J Radiol, 2014, 32(1): 14-21.

［12］ LEE G, LEE H Y, PARK H, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art［J］. Euro J Radiol, 2016, 86: 297-307.

（收稿日期：2019-11-18　修回日期：2020-02-25）