

基于 CT 影像的早期肺癌风险评估

何怡雯, 陈 阳, 吴文浩, 侯学文, 李浩东, 聂生东

(上海理工大学 健康科学与工程学院, 上海 200093)

摘要: 为了克服横断面型数据的不完整性, 提出了一种基于合成分析的早期肺癌风险预测模型。收集 247 组来自医院的患者数据进行实验, 使用合成分析法结合肺结节良恶性判断结果、吸烟史、家族病史等肺癌危险因素建立风险评估模型。实验表明, 未纳入影像学良恶性分类结果的合成分析模型的准确率为 83.20%, 而纳入影像学良恶性分类结果的合成分析模型的准确率达到 87.40%。结合基于 CT 影像的肺结节良恶性判断结果能有效提高早期肺癌风险评估模型的准确性。

关键词: 早期肺癌; 危险因素; 合成分析; 风险评估

中图分类号: TP 391.41 文献标志码: A

Risk assessment methods of early lung cancer based on CT images

HE Yiwen, CHEN Yang, WU Wenhao, HOU Xuewen, LI Haodong, NIE Shengdong

(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The incidence and mortality of lung cancer have always been high. The establishment of an effective method to predict the risk of early lung cancer is of great significance for improving the survival rate of lung cancer patients. To overcome the incompleteness of cross-sectional data, an early lung cancer risk prediction model was proposed based on synthetic analysis. 247 groups of cases from hospitals were collected for experiments, and a risk assessment model was established using synthetic analysis method combined with lung nodule benign and malignant judgment results, smoking history, family history and other lung cancer risk factors. The accuracy rate of the synthetic analysis model that did not include the imaging benign and malignant classification results was 83.20 %, and the accuracy of the synthetic analysis model that included the imaging benign and malignant classification results reached 87.40 %. Combining the results of benign and malignant lung nodules based on CT images can effectively improve the accuracy of early lung cancer risk assessment models.

Keywords: early lung cancer; risk factors; synthetic analysis; risk assessment

收稿日期: 2021-06-21

基金项目: 国家自然科学基金重点项目 (81830052); 上海市科技创新行动计划 (18441900500); 上海市自然科学基金资助项目 (20ZR1438300)

第一作者: 何怡雯 (1995-), 女, 硕士研究生。研究方向: 医学图像处理与分析。E-mail: 940244473@qq.com

通信作者: 聂生东 (1962-), 男, 教授。研究方向: 医学图像处理与分析。E-mail: nsd4647@163.com

肺癌是世界上最常见的恶性肿瘤之一，发病率和死亡率常年居高不下。由于肺癌的并发症状不易被识别、易转移等特性^[1-3]，绝大部分的肺癌患者直到中晚期才被确诊，导致错过了最佳的治疗时机，因而肺癌患者的 5 年生存率低下。研究证明，在肺癌早期及时确诊并得到有效的治疗，可以有效地提高患者的 5 年生存率^[4]。因此，构建一种可靠、准确的早期肺癌风险评估模型具有重要的应用价值。多项研究表明，肺癌是环境和自身各种因素综合作用所引起的，常见的危险因素有年龄、性别、吸烟、肺部疾病史及家族恶性肿瘤史等。年龄作为肺癌的危险因素之一也已得到确认^[5]，相关研究显示^[6]40 岁以下人群肺结节的检出率小于 3%，60 岁以上人群的肺结节检出率超过 50%。此外，研究表明，女性的肺癌患病风险比男性要高，同时长时间处于压抑状态的人群也更容易产生肺结节。吸烟作为环境中最主要的危险因素已经得到了大家的公认^[7]。而自身的内在主要危险因素有患病本人的肺部疾病史和家族恶性肿瘤史。

近年来，血清肿瘤标志物在肺癌诊断方面的价值也逐渐被重视起来^[8]。相关研究显示，随着肺癌患者的病情不断恶化，对应患者体内的癌胚抗原(carcinoembryonic antigen, CEA)、鳞状上皮细胞癌抗原(squamous cell carcinoma antigen, SCCA)和细胞角质蛋白 19 片段(cytokeratin-19-fragment, CYFRA21-1)在血清中的表达也随之上升。因此，通过联合这 3 种血清肿瘤标志物在血清中的含量情况进行分析，可以作为早期肺癌诊断的一种有效手段。CEA 与 SCCA 的正常值为 5.0 mg/mL，CYFRA21-1 的正常值为 3.3 mg/mL。当检测值大于正常值时，可以认为该患者的肺癌患病风险较高。

计算机断层扫描(computed tomography, CT)是目前筛查肺癌最主要的技术手段。相对于支气管镜与 X 线胸片检查等其他检查方式，CT 可以显示横断面的三维结构图像，能够很好地帮助医生区分病灶区域和正常组织，对于早期肺癌的筛查有着更好的检查效果。低剂量 CT 技术不仅能降低患者的放射暴露时间，而且能获取足够清晰的胸部影像，从而成为早期肺癌筛查的主要方式。

由于数据的规范性、普适性等原因，国内外大部分对于肺癌的风险评估研究还属于前瞻性研究。目前国外已经有一些个人或机构建立了一些较为有效且稳定的早期肺癌风险评估模型。例

如，美国德克萨斯大学安德森癌症研究中心研究的肺癌风险评估模型^[9]；哈佛大学公共卫生学院开发的哈佛癌症风险评估工具可以对多种恶性肿瘤进行风险评估^[10]；Bach 等^[11]建立的肺癌预警模型主要适用于目前或既往存在吸烟史的人群；Liverpool Lung Project (LLP) 模型^[12]是通过研究肺炎史、肺癌家族史、石棉暴露史、年龄、性别和吸烟史来建立的肺癌风险预测模型；Spitz 等^[9]建立的早期肺癌风险评估模型针对不吸烟、既往吸烟和目前正在吸烟的人群其交叉验证统计值分别为 0.59、0.63 和 0.65；Cassidy 等^[13]建立的肺癌风险预测模型最终验证的 AUC 值为 0.70；王旭等^[14]将与肺癌相关的有效分子生物标志物与环境因素及机体内因素结合在一起进行联合分析，研发出了相对客观的适用于东北地区的肺癌预警评估模型；同样地，张亚琛等^[15]使用肺结节分布特征与其他危险因素建立了适合河北地区的肺癌早期预警模型；浙江大学团队研制出一种可以检测并分析患者呼出气标志物的仪器^[16]，可以对患者呼出的气体中的挥发性有机物进行定性分析，再结合其他肺癌危险因素建立肺癌的早期诊断模型，为肺癌的早期预防和诊断提供了比较成熟的技术支持。刘龙飞^[17]的研究内容主要集中在对于已存在的各种模型的优化。

综上所述，现有的方法都是将重点放在对不同特征数据的处理上面，将各组学的特征数据进行融合，再选择合适的统计学方法，最后建立肺癌风险评估模型。

本文则针对于横断面数据，提出一种基于合成分析的早期肺癌风险预测模型，以克服横断面数据的不完整性。横断面数据是呈现出“倒金字塔”型的一类数据，不同病例有不同的数据缺失，从而导致具有不同因素的样本数存在区别。因此，传统的多因素回归方法无法应用到所有数据上面，而合成分析法是一种能够很好地适应于横断面型数据的统计学分析方法。

1 材料与方法

1.1 数据筛选

本文使用的数据来自上海市肺科医院，一共 247 组病例。其中，200 组作为训练数据，剩余作为测试数据。每组数据包括完整的患者术前最后一套 CT 影像和其他信息，并且临床医生根据每例

患者的病理诊断报告给每组病例划分了危险等级，共分为5个等级，第5级代表危险程度最高，第1级表示无患癌风险，以此作为建模中的结果变量。具体信息如表1所示。从表1可以看出，本文所使用的数据属于典型的横断面型数据，合成分析法可以在缺乏纵向队列数据时，基于横断面数据构建有效的疾病预测模型^[18]。将单个因素回归模型和单个因素之间的相关性整合到一起形成一个多因素回归模型，从而克服横断面数据的不完整性。因此，本文使用合成分析法来建立最终的早期肺癌风险评估模型。

表 1 肺科医院数据信息

Tab.1 Data information of pulmonary hospitals

因 素	组数
年龄	247
性别	247
吸烟史	207
肺部疾病史	55
家族病史	205
肿瘤标记物	201

将基于 CT 影像的判断结果结合其他肺癌危险因素来进行建模。首先，使用基于三维条件随机场优化的 3D U-Net 分割算法^[19]对所使用数据进行肺结节分割，提取完整的肺结节区域；然后，使用基于聚类辅助的随机森林分类算法^[20]进行良恶性分类；最后，再使用合成分析法结合良恶性分类结果和吸烟史、家族病史等肺癌危险因素建立最终的早期肺癌风险评估模型。模型以第 1~5 级的危险等级为结果变量，等级越高表示危险程度越高，模型中其他变量的类型和赋值情况如表 2 所示。

1.2 模型建立

回归是一种从连续型统计数据中得到数学模型的方法，其中，线性回归模型的实现较为简单且快速，特别适合线性数据及较小的数据量，但对于高度复杂的非线性数据不能很好地进行拟合。另外，回归模型具有很好的可解释性，可以直接看出每个特征对结果的影响程度，有利于建模时决策分析。多因素线性回归体现了多个独立输入变量与输出因变量之间的关系，合成分析法通常有两种形式去构建多因素回归模型：一是逐步地纳入多个单因素变量，先构建第一个危险因

表 2 肺癌风险因素及赋值

Tab.2 Risk factors and assignment for lung cancer

危险因素	属性	赋值
年龄	连续值	k
性别	二分类	男-1, 女-0
吸烟史	二分类	吸烟-1, 不吸烟-0
肺部疾病史	二分类	有-1, 无-0
家族病史	二分类	有-1, 无-0
CEA	二分类	正常值-1, 非正常值-0
SCCA	二分类	正常值-1, 非正常值-0
CYFRA21-1	二分类	正常值-1, 非正常值-0

素与结果变量直接的基础方程，再逐次将其他危险因素添加到方程里面，形成最终的多因素回归方程；二是基于现有的基础方程，直接构建多因素回归模型。本文主要研究以第一种形式来建立早期肺癌的风险评估模型。在实际肺癌风险评估模型建立过程中，单个因素(如性别、年龄、吸烟史、肺部疾病史和家族病史等)与结果变量之间的回归分析等数据主要是通过其他资料文献获得，影像学判断结果和 3 种肿瘤标志物与结果变量之间的回归系数通过单因素分析获得。

现介绍使用合成分析法构建早期肺癌风险评估模型的具体步骤。

步骤 1 基础方程构建。

设第一个危险因素为自变量 F_1 ，肺癌危险等级为因变量 I_1 ，构建单因素回归模型 $I_1 = a + b_1F_1$ ， b_1 为 F_1 的回归系数。

步骤 2 纳入第二个危险因素。

纳入第二个危险因素 F_2 ，并根据基础方程计算所对应的回归系数 f_2 ， b_2 为先前统计分析得到的回归系数，将 b_2 与 f_2 进行比较，如果差值较小，则令 $b_2 = f_2$ ；否则，令 $b_2 = (b_2 - f_2)$ 。

步骤 3 构建包含前 2 个危险因素的回归方程。

构建新方程 $I_2 = a + b_1F_1 + b_2F_2$ ，此时，包含 2 个危险因素的方程 I_2 则反映了危险因素 F_1 和危险因素 F_2 与疾病之间的关系。

步骤 4 重复步骤 2~3，逐次纳入所有危险因素，得到最终的回归方程 $I_n = a + b_1F_1 + \cdots + b_nF_n$ ， n 为危险因素数目。

本研究使用 Windows 平台的 SPSS24.0 软件实现。

2 实验结果与分析

2.1 因素分析

本文将基于 CT 影像的良恶性判断结果和其他危险因素相结合, 建立了一个多因素回归模型, 对肺癌进行风险评估。年龄、性别、吸烟等肺癌危险因素在肺癌诊断模型中的效果已经被确认, 但良恶性的判断结果对于模型的影响尚未明确。因此, 在建模之前, 首先要分析基于 CT 影像的肺结节良恶性判断结果是否适合作为肺癌危险因素之一来进行建模。由于数据类型的限制, 使得本文无法使用常用的变量筛选方法来进行分析, 所以, 通过单因素方差分析良恶性判断结果与结果变量之间的相关性来进行确认, 分析结果如表 3 所示。

表 3 不同分类器良恶性分类结果的单因素分析结果
Tab.3 Univariate analysis results of benign and malignant classification results of different classifiers

分类器	分类准确率/%	P值
SVM	89.12	0.023
NB	90.15	0.014
LDA	91.20	0.020
RF	94.28	0.002

整个分析在 SPSS24.0 软件中进行, 分别将 4 组不同分类器的良恶性分类结果与风险评估的结果变量作了单因素方差分析, 分别计算了各组的 P 值。由 4 组的 $P < 0.05$ 可以看出, 良恶性分类结果因素对于肺癌风险评估结果变量的影响是显著的, 因此, 可以初步证明肺结节良恶性判断结果可以作为危险因素之一建立肺癌风险评估模型。在之后的实际建模结果评估的时候, 可以进一步对此结论进行验证。

2.2 合成分析模型构建

首先统计分析了具体的肺癌单个因素和对应的回归系数, 如表 4 所示。将良恶性判断结果作为第一个变量来进行建模, 3 种肿瘤标志物的回归系数则通过单次的单因素 Logistic 回归分析得到。

a. 基础方程。
以肺癌危险等级为因变量, 基于 CT 影像的肺结节良恶性判断结果 A 为第 1 个危险因素作自变量, 构建基础方程为

$$I_1 = 2.124 + 2.012A$$

b. 加入第 2 个变量: 年龄 B 。

表 4 肺癌危险因素和对应回归系数
Tab.4 Risk factors and corresponding regression coefficients of lung cancer

危险因素	回归系数
年龄	-0.042
性别	-1.36
吸烟	1.334
肺部疾病史	1.269
家族病史	1.134
CEA	1.568
SCCA	2.087
CYFRA21-1	1.652

$$I_2 = I_1 - 0.042B$$

c. 加入第 3 个变量: 性别 C 。
基于方程 I_2 , 根据肺科医院整理的数据, 计算每个具有年龄因素的样本的 I_2 值, 以 I_2 为因变量、性别为自变量进行单因素线性回归分析, 得到回归系数为 0.005, 因此, $b_3 = (-1.36 - 0.005) = -1.365$, 构建新的方程为

$$I_3 = I_2 - 1.365C$$

d. 加入第 4 个变量: 吸烟 D 。
基于第 3 个方程 I_3 , 根据同一数据集计算每个具有吸烟因素的样本的 I_3 值, 再以 I_3 为因变量、吸烟因素为自变量进行单因素线性回归分析, 得到回归系数 f_3 为 1.213, 与 b_3 相差较小, 因此, 包含 3 个变量的方程为

$$I_4 = I_3 + 1.213D$$

e. 加入第 5 个变量: 肺部疾病史 E 。
根据数据集基于第 4 个方程 I_4 计算每一个具有肺部疾病史因素的样本的 I_4 值, 并以 I_4 为因变量、肺部疾病史为自变量进行分析, 得到回归系数 $f_5 = 0.034$, 则 $b_5 = (1.269 - 0.034) = 1.235$, 新的方程为

$$I_5 = I_4 + 1.235E$$

f. 加入第 6 个变量: 家族病史 F 。
根据方程 I_5 和数据集计算每个包含家族病史的样本的 I_5 值, 并以 I_5 作为因变量、以家族病史因素为自变量作回归分析, 得到回归系数 $f_6 = 0.107$, 因此 $b_6 = (1.134 - 0.107) = 1.027$, 得到新的方程

$$I_6 = I_5 + 1.027F$$

g. 随后, 依次分别加入第 7, 8, 9 个变量 G , H , J : 肿瘤标志物 CEA, SCCA 和 CYFRA21-1。
重复前面的步骤加入肿瘤标志物因素, 计算得出 $b_7 = 1.387$, $b_8 = 1.890$, $b_9 = 1.542$, 构建新方

程为

$$I_9 = I_6 + 1.387G + 1.890H + 1.542J$$

h. 构建最终模型。
为了保证最终方程对于整体数据的平均肺癌危险等级保持不变，需要在方程 I_9 的基础上减去一个平均值 2.832，得到最终的风险评估方程 I_F 。

$$I_F = 2.124 + 2.012A - 0.042B - 1.365C + 1.213D + 1.235E + 1.027F + 1.387G + 1.890H + 1.542J - 2.832$$

2.3 模型评价

使用组内相关系数 (intraclass correlation coefficient, ICC)对医生划分的危险等级与合成分析模型得到的等级进行检验，得到的 ICC 值大于 0.75，表示这两组变量对同一个样本的一致性高，说明了本文得到的结果与医生给出的标准相比重复性高且可信度良好。

为进一步验证基于 CT 影像的肺结节良恶性判断结果对于肺癌风险评估模型的积极影响，按照同样的方法构建了不包含影像学良恶性分类结果的合成分析模型，得到了风险评估方程 I_C 。

$$I_C = 1.325 - 0.042B - 1.234C + 1.343D + 1.132E + 1.235F + 1.467G + 1.290H + 1.643J - 1.693$$

分别计算了合成分析模型 I_F 和 I_C 在相同测试数据上表现的性能，主要指标包括准确性、敏感性、特异性。对比结果如表 5 所示，图 1 显示了 2 个

表 5 模型性能对比结果

Tab.5 Comparison results of model performance

合成分析模型	准确率/%	敏感性/%	特异性/%
I_F	87.40	89.01	90.22
I_C	83.20	84.83	87.41

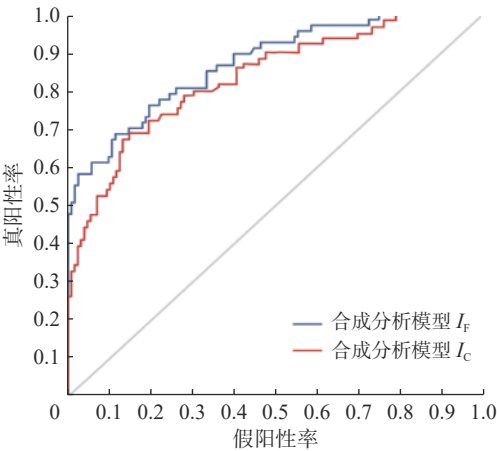


图 1 模型 ROC 曲线

Fig.1 ROC curve of the model

模型的 ROC 曲线(接收者操作特征曲线)对比结果。

由表 5 的结果可以看出，模型 I_F 相较于模型 I_C 具有更好的评价结果，本文提出的合成分析模型的准确率达到 87.40%。同时从图 1 中的 ROC 曲线对比也可以证明，加入良恶性判断结果之后的肺癌风险评估模型具有更好的评估性能。

3 结 论

研究了使用合成分析方法建立基于横断面数据的早期肺癌风险评估模型。基于病人 CT 影像的肺结节良恶性判断结果、年龄、性别、吸烟、家族病史、肺部疾病史和 3 种肿瘤标志物为研究变量，分别构建了包含和不包含肺结节良恶性分类结果因素的模型，并在相同数据集上对比 2 个模型。结果显示，包含肺结节良恶性判断结果的模型具有更好的预测性能，也进一步验证了通过单因素回归分析得出的结论，基于 CT 影像的肺结节良恶性判断结果变量对于肺癌风险评估模型的构建是有益的。但本方法也存在一些不足之处。首先，对于不同因素回归系数的确定和纳入模型顺序的确定具有主观性，在整体上对模型的性能评估造成一定的影响；此外，在模型对比的过程中，由于不同模型使用的数据，涉及到的自变因素和结果变量的设定具有一定的差别，很难客观地去分析不同模型之间的差异性。希望在未来的研究中，随着研究数据的增多，更多肺癌的危险因素被确定和加入到风险评估模型当中。能够建立一个稳定、准确的早期肺癌风险评估模型。

参考文献:

[1] BOWEN E F, RAYNER C F J. Patient and GP led delays in the recognition of symptoms suggestive of lung cancer[J]. *Lung Cancer*, 2002, 37(2): 227-228.

[2] CORNER J, HOPKINSON J, FITZSIMMONS D, et al. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis[J]. *Thorax*, 2005, 60(4): 314-319.

[3] CORNER J, HOPKINSON J, ROFFE L. Experience of health changes and reasons for delay in seeking care: a UK study of the months prior to the diagnosis of lung cancer[J]. *Social Science & Medicine*, 2006, 62(6): 1381-1391.

[4] CALLISTER M E J, BALDWIN D R. How should pulmonary nodules be optimally investigated and managed?[J]. *Lung Cancer*, 2016, 91: 48-55.

[5] JENKS S. Is lung cancer incidence increasing among never-smokers?[J]. [Journal of the National Cancer Institute](#), 2016, 108(1): djv418.

[6] BONNER M R, FREEMAN L E B, HOPPIN J A, et al. Occupational exposure to pesticides and the incidence of lung cancer in the agricultural health study[J]. [Environmental Health Perspectives](#), 2017, 125(4): 544–551.

[7] HOSSEINI M, NAGHAN P A, KARIMI S, et al. Environmental risk factors for lung cancer in Iran: a case-control study[J]. [International Journal of Epidemiology](#), 2009, 38(4): 989–996.

[8] 王晓成, 胡卫盟, 徐晓明, 等. SCCA、NSE、CEA 联合检测对肺癌的诊断价值 [J]. [河北医药](#), 2015, 37(22): 3393–3395.

[9] SPITZ M R, HONG W K, AMOS C I, et al. A risk model for prediction of lung cancer[J]. [Journal of the National Cancer Institute](#), 2007, 99(9): 715–726.

[10] COLDITZ G A, ATWOOD K A, EMMONS K, et al. Harvard report on cancer prevention volume 4: Harvard cancer risk index[J]. [Cancer Causes & Control](#), 2000, 11(6): 477–488.

[11] BACH P B, KATTAN M W, THORNQUIST M D, et al. Variations in lung cancer risk among smokers[J]. [Journal of the National Cancer Institute](#), 2003, 95(6): 470–478.

[12] CASSIDY A, MYLES J P, VAN TONGEREN M, et al. The LLP risk model: an individual risk prediction model for lung cancer[J]. [British Journal of Cancer](#), 2008, 98(2): 270–276.

[13] CASSIDY A, DUFFY S W, MYLES J P, et al. Lung cancer risk prediction: a tool for early detection[J]. [International Journal of Cancer](#), 2007, 120(1): 1–6.

[14] 王旭. 东北地区肺癌早期预警模型的构建及应用 [D]. 长春: 吉林大学, 2014.

[15] 张亚琛. 河北省肺癌和肺结节危险因素分析及风险评估模型的建立 [D]. 石家庄: 河北医科大学, 2018.

[16] 王福园. 基于呼出气中挥发性有机物的肺癌诊断模型优化和临床应用 [D]. 杭州: 浙江大学, 2018.

[17] 刘龙飞. 非匹配数据自动化批量上传及肺癌风险评估模型优化研究 [D]. 兰州: 兰州大学, 2014.

[18] 石福艳. 常见慢性病危险因素测量及风险等级评估方法研究 [D]. 西安: 第四军医大学, 2015.

[19] WU W H, GAO L, DUAN H H, et al. Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization[J]. [Medical Physics](#), 2020, 47(9): 4054–4063.

[20] WU W H, HU H H, GONG J, et al. Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis[J]. [Physics in Medicine & Biology](#), 2019, 64(3): 035017.

(编辑: 石 瑛)