基于多肿瘤标志物和特征筛选的逻辑回归方法诊断良恶性胸腔积液的研究*

宋俊儒[1]　梁宝生[2△]　王思洋[3△]　陈阳育[4]

【提　要】 **目的**　基于特征筛选算法探索利用胸腔积液和血清中 CEA、CA125、CA153 和 CA199 四种肿瘤标志物的组合与筛选对鉴别良、恶性胸腔积液的诊断价值。**方法**　收集北京朝阳医院和武汉某医院收治的胸腔积液患者共 452 例，其中恶性胸腔积液患者 143 例、良性胸腔积液患者 309 例；取胸腔积液及配对血清标本，用化学发光法检测 CEA、CA125、CA153 和 CA199 浓度，辅以患者性别、年龄和医院所在城市三项人口学变量，首先应用独立性检验进行变量初筛，而后应用带惩罚项的逻辑回归和基于逻辑回归的模拟退火算法和遗传算法进行标志物筛选，根据受试者工作特征曲线下面积（area under the curve，AUC）和 DeLong 检验进行模型诊断效果的评估和比较。**结果**　特征筛选结果以及回归系数和 SHAP（shapley additive explanations）值一致表明胸腔积液 CA199、CA153 联合血清 CEA 为最优肿瘤标志物组合；在测试数据集上，该指标组合达到最高诊断精度（AUC = 0.923），显著高于最优单标志物模型（AUC = 0.877，P<0.001）和全标志物模型（AUC = 0.906，P = 0.044），灵敏度和特异度分别达到 0.811 和 0.939。**结论**　多项肿瘤标志物的联合应用相较单一标志物能够显著提升模型诊断精度，且合理的标志物筛选策略对提升诊断精度和简化模型有进一步帮助；本文推荐联合胸腔积液中 CA199 和 CA153 以及血清中 CEA 来建立诊断模型，并提供了该模型的列线图和实用化的网页计算器，为辅助临床诊断提供便利。

【关键词】 胸腔积液　肿瘤标志物　联合诊断　特征筛选　逻辑回归

【中图分类号】 R730.4　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.02.029

胸腔积液（pleural effusion，PE）是以胸膜腔内病理性液体积聚为特征的一种常见临床症候。恶性胸腔积液（malignant pleural effusion，MPE）是指原发于胸膜或由其他部位转移至胸膜的恶性肿瘤所引起的胸腔积液[1]；良性胸腔积液是肺外结核的主要症状[2]。临床上恶性肿瘤患者一旦出现 MPE 即意味病变已达晚期，若不及时治疗，平均生存期仅 3.3 个月[3-4]。因此，尽早准确鉴别良恶性胸腔积液对于诊断原发病、疾病预后和选择治疗方案都十分重要。

随着肿瘤免疫学的发展，肿瘤标志物凭借创伤小、取材方便、可重复性强等优点逐渐成为临床中肿瘤早期筛查的重要手段[5]。但单一标志物对 MPE 的诊断价值有限，其灵敏度常常不能令人满意、存在较大漏诊的可能性[6]；诸多研究表明多种肿瘤标志物联合检测是提高诊断效果的有效手段[7-8]，但现有研究多基于简单的并联试验，诊断精度有限。随着机器学习的广泛应用，已有研究表明使用机器学习构建标志物联合诊断模型能在不同程度上提高诊断效果，但相关研究有很多局限性[9-10]。特别地，多种肿瘤标志物的联合检测会增加医疗资源消耗和患者经济负担。如何在确保诊断效果的前提下，选取尽可能少且重要的标志物组合进行诊断，有重要的临床意义，但也是一个难题。基于此，本文使用血清和胸腔积液中四种肿瘤标志物浓度，辅以人口学变量，使用特征筛选方法探讨肿瘤标志物的联合与筛选对鉴别良、恶性胸腔积液的诊断价值。

数据与分析方法

1.数据来源

选取 2015 年 1 月至 2017 年 6 月收入首都医科大学附属北京朝阳医院呼吸与危重症医学科的所有伴有胸腔积液的成人患者 307 例，以及武汉某医院收治的胸腔积液患者 145 例，共计 452 例样本数据，其中恶性胸腔积液患者 143 例（年龄 21～86 岁，男性 76 例，女性 67 例），良性胸腔积液患者 309 例（年龄 16～88 岁，男性 219 例，女性 90 例）。恶性胸腔积液诊断的金标准为病理证实在胸腔积液和（或）胸膜活检标本中存在恶性肿瘤细胞。

2.统计分析方法

选取是否患有恶性胸腔积液作为响应变量，选取三项人口学变量（年龄、性别、医院所在城市）和八项肿瘤标志物（血清和胸腔积液中 CEA、CA125、CA153 和 CA199）浓度作为预测变量。年龄和各项肿瘤标志物浓度属于连续变量，进行标准化处理以消除量纲的影响；性别和医院所在城市属于定性变量，将其转变为 0～1 变量。首先利用卡方检验和均方差独立性检验（mean variance，MV）[11] 分别对定性和定量变量做初步筛选，再使用特征筛选算法做进一步细筛，最后纳入逻辑回归构建诊断模型。通过五折交叉验证进行模型诊断效果的评估和比较。使用 AUC 值作为模型评价

* 基金项目：国家自然科学基金（12031016，11971324，11901013）

1. 中国人民大学统计与大数据研究院（100872）
2. 北京大学公共卫生学院生物统计系
3. 中央财经大学统计与数学学院
4. 北京朝阳医院呼吸与危重症医学科
△通信作者：梁宝生，E-mail：liangbs@ hsc.pku.edu.cn；王思洋，E-mail：siyangw@ 163.com

指标，依据约登指数选取最优 cut-off 值，计算对应的灵敏度和特异度。使用 DeLong 检验[12]对不同模型诊断效果的差异进行显著性检验。

使用带惩罚项的逻辑回归以及基于逻辑回归的遗传算法和模拟退火算法共三类方法进行特征筛选。其中惩罚项函数使用 adaptive-lasso[13]、MCP（minimax concove penalty）[14]和 SCAD（smoothly clipped absolute deviation）[15]三种。模拟退火（simulated annealing，SA）算法和遗传算法（genetic algorithm，GA）是两种常用的随机性搜索方法，具有原理简单、通用性好等优点[16]；采用逻辑回归作为 GA 和 SA 的基分类器。利用筛选所得标志物组合建立模型，并与单标志物和全标志物模型进行比较，来探究标志物筛选对诊断效果的影响以及联合诊断是否有助于提升诊断效果。此外，使用 SHAP 值[17]衡量各标志物对预测结果的贡献度，从而比较其对鉴别良、恶性胸腔积液的诊断价值。采用 Python 和 R 软件完成数据分析和绘图，检验 P 值小于 0.05 认为具有统计学意义。

结　果

1.描述性统计和变量初筛

首先对各数值型变量进行描述性统计，发现患者年龄和八项肿瘤标志物浓度的取值在良、恶性胸腔积液患者间均有明显差异。卡方独立性检验结果显示性别与响应变量高度显著相关（$P<0.001$），而医院所在城市这一变量不显著（$P=0.685$），由此判断性别可能对良、恶性胸腔积液的鉴别有较高诊断价值，医院所在城市则予以剔除；MV 独立性检验结果显示年龄、胸腔积液和血清中四类标志物浓度均与响应变量高度显著相关（$P<0.001$），故全部保留。

表 1　诊断模型及对应的肿瘤标志物组合列表

模型名称	肿瘤标志物组合	特征筛选的说明
S1	PE:CA199,CA153;serum:CEA	使用 GA、SA、带 MCP 和 adaptive-lasso 惩罚项的逻辑回归筛选所得
S2	PE:CA199,CA153,CEA;serum:CEA	使用带 SCAD 惩罚项的逻辑回归筛选所得
M3	PE:CA153	最优单标志物模型
M9	PE:CA199,CA125,CA153,CEA serum:CA199,CA125,CA153,CEA	全标志物模型

2.统计建模结果

（1）标志物组合筛选

筛选所得标志物组合及对应诊断模型见表 1。其中，使用特征筛选算法所得模型记作 S1 和 S2；单标志物模型按胸腔积液 CA199、CA125、CA153、CEA 和血清 CA199、CA125、CA153、CEA 的顺序依次记作 M1～M8（限于篇幅仅列出最优单标志物模型 M3），全标志物模型记为 M9，用作对比研究，serum 代表血清。上述模型中均纳入性别和年龄两个变量。

（2）诊断模型比较

表 2 展示了不同诊断模型在测试数据集上的表现。不同模型诊断效果比较的 DeLong 检验结果见表 3。可见，单标志物模型中基于胸腔积液 CA153 的 M3 模型表现最优，AUC = 0.877，灵敏度和特异度分别为 0.713 和 0.919；基于血清 CEA（M8）和胸腔积液 CA199（M1）的单标志物模型次之，AUC 分别为 0.854 和 0.853。全标志物模型 M9 相比于最优单标志物模型 M3，整体诊断效果有所提升，AUC = 0.906，相比提升3.3%，灵敏度和特异度为0.748和0.987，相对M3分别提升了 4.9%和 7.4%，但是全标志物模型相比最优单标志物模型的诊断优势无统计学意义（$P=0.052$）。相比于全标志物模型，特征筛选所得模型 S1 和 S2 的诊断效果均有显著提升（$P<0.05$），相对 M9 灵敏度提升了 8.4%。基于最优诊断模型 S1，即联合胸腔积液 CA153、CA199 与血清 CEA 的逻辑回归模型，绘制列线图（图 1）并开发部署了网页计算器（链接 https://woodysjr.shinyapps.io/0319/），以便辅助临床诊断应用。

表 2　不同诊断模型的诊断效果

诊断模型	AUC	灵敏度	特异度	cut-off 值
S1	0.923	0.811	0.939	0.282
S2	0.923	0.811	0.939	0.281
M1	0.853	0.671	0.880	0.341
M2	0.742	0.790	0.608	0.277
M3	0.877	0.713	0.919	0.312
M4	0.839	0.608	0.916	0.410
M5	0.769	0.580	0.828	0.387
M6	0.712	0.888	0.440	0.236
M7	0.810	0.727	0.744	0.283
M8	0.854	0.734	0.825	0.293
M9	0.906	0.748	0.987	0.465

表 3　DeLong 检验结果

模型	M3 vs M9	M3 vs S1	M9 vs S1	M3 vs S2	M9 vs S2	S1 vs S2
AUC	0.877/0.906	0.877/0.923	0.906/0.923	0.877/0.923	0.906/0.923	0.923/0.923
P 值	0.052	<0.001	0.044	<0.001	0.048	0.280

（3）不同肿瘤标志物诊断价值重要性评估

为了进一步衡量各种肿瘤标志物对鉴别良恶性胸腔积液的重要性，图 2 展示了全标志物模型中各肿瘤标志物的回归系数和 SHAP 值。可见，最优诊断模型

S1 中的三项标志物(胸腔积液 CA153、CA199 与血清 CEA)SHAP 值和回归系数明显高于其他标志物,表明三者对良、恶性胸腔积液具有较高诊断价值。

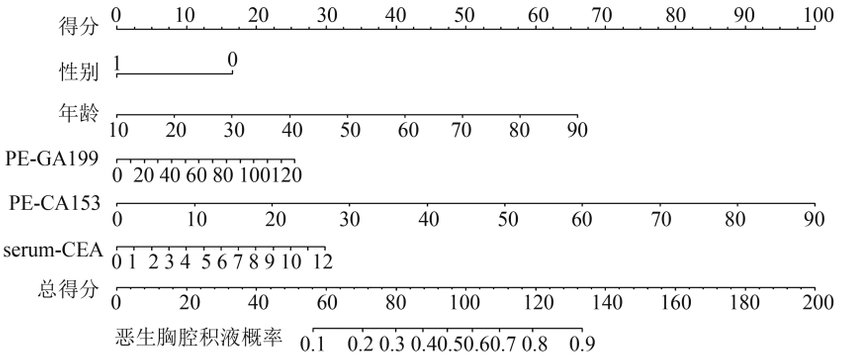

图 1　基于最优诊断模型 S1 的列线图

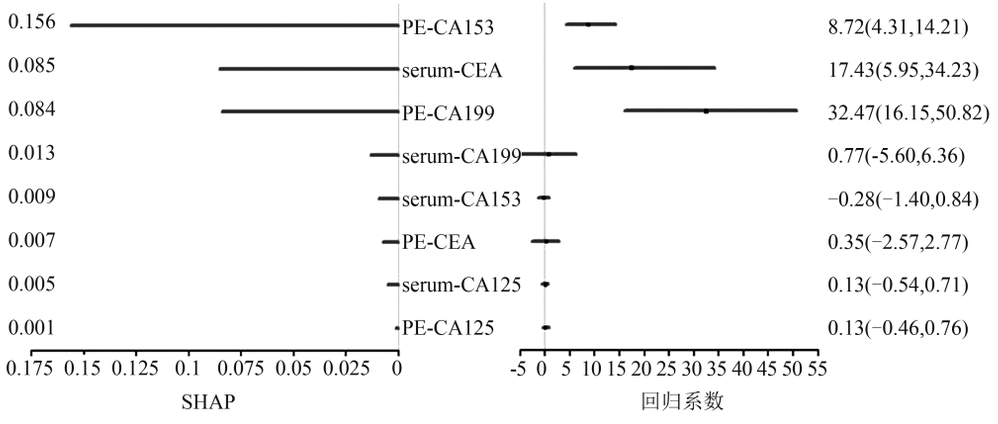

图 2　肿瘤标志物的 SHAP 值和回归系数森林图

讨　论

本文通过前瞻性试验研究,采用逻辑回归建立包含人口学变量和肿瘤标志物浓度的良、恶性胸腔积液诊断模型,验证了肿瘤标志物联合诊断具有较高的诊断价值,并探索了遗传算法、模拟退火算法以及带惩罚项的逻辑回归等特征筛选方法对进一步提高诊断精度的价值。主要研究结论体现在以下两方面:①CA125、CA199、CEA 和 CA153 的联合检测能够显著提升恶性胸腔积液的诊断精度、尤其是灵敏度,这与现有研究结论一致[7-10]。在此基础上本文进一步验证了特征筛选方法在优化诊断模型中的应用价值,发现合理筛选的肿瘤标志物组合能够提升诊断精度。例如,陈阳育等[10]利用胸腔积液和血清中肿瘤标志物分别建模,发现胸腔积液 CEA 联合 CA199 和 CA153 能够达到最高诊断精度(AUC = 0.9)[10],而本研究利用特征筛选方法,结合模型诊断效果、回归系数和 SHAP 值,发现胸腔积液 CA153、CA199 联合血清 CEA 为最优的肿瘤标志物组合方式,测试集 AUC 比陈阳育等的结果提高了 2.5%,灵敏度和特异度达到 0.811 和0.939。②血清及胸腔积液 CA125 仅在单标志物模型中的灵敏度最优(分别为 0.888 和 0.790),但特异度欠佳(分别为 0.440 和 0.608),而在多标志物联合建模过程中,血清和胸腔积液 CA125 所表现出的诊断价值均显著低于其他标志物。这与金宸等[7]和林永志等[18]的研究结论一致。另外,孙美琪等[19]也指出 CA125 在结核性胸腔积液和 MPE 中的阳性率均较高,难以鉴别。

本文仅涉及临床中常见的四种标志物,现有研究表明 CYFRA21-1、NSE、SCC 等标志物亦具有一定诊断价值[20-21],未来应引入更多标志物,并尝试其他特征筛选方法。此外,可以针对具体病因建立诊断模型,进一步分析标志物联合检测与筛选的诊断价值。综上所述,多种肿瘤标志物的联合能够显著改善诊断效果,而合理的肿瘤标志物筛选策略不仅有助于改善诊断精度,减小模型复杂度,降低计算成本,还有助于节约医疗成本,减轻患者的经济负担,改善就医体验。

参　考　文　献

[1] 中国恶性胸腔积液诊断与治疗专家共识组. 恶性胸腔积液诊断与治疗专家共识. 中华内科杂志, 2014, 53(3):252-256.

[2] 解春林, 黄韬, 卜俊晖, 等. 胸水标志物在诊断恶性与结核性胸水中的研究进展. 实用医学杂志, 2019, 35(6):1009-1011.

[3] Sahn SA. Management of malignant pleural effusions. Monaldi Arch Chest Dis, 2001, 56(5):394-399.

[4] 王立伟, 焦顺昌. 恶性胸腔积液的综合治疗新进展. 中国肿瘤临床, 2006(4):236-239.

(下转第 290 页)