

· 论 著 ·

基于归一化回归算法的多参数模型用于原发性肝癌 微小血管侵犯预测的研究^{*}

王孜怡¹,肖 潇²,黄晨军²,童 林³,高春芳^{2,3},曹宏伟^{1△}1. 海军军医大学附属长海医院信息科,上海 200438;2. 上海中医药大学附属岳阳中西医结合医院
检验实验中心,上海 200437;3. 上海东方肝胆外科医院检验科,上海 200438

摘要:目的 基于归一化的甲胎蛋白(AFP)、甲胎异质体(AFP-L3)、异常凝血酶原Ⅱ(PIVKAⅡ)检测数据建立逻辑回归(LR)模型用于预测原发性肝癌微小血管侵犯(MVI)。方法 纳入手术后证实为原发性肝细胞癌(HCC)的患者 1 314 例,使用 Python 进行数据集 7 : 3(建模组:验证组)的划分,收集入组患者术前 AFP、AFP-L3、PIVKAⅡ检测结果和术后 MVI 分级信息,检测数据在(-1,1)区间归一化,建立 LR 模型并在建模组和验证组评价模型预测 MVI 的效能。结果 建模组区分 HCC 患者是否发生 MVI 的曲线下面积为 0.647,虽然与单独采用 AFP 指标的相同,但在验证组中,采用 LR 模型诊断 MVI 的曲线下面积为 0.720,高于 AFP、AFP-L3、PIVKAⅡ单独使用的诊断效能。结论 基于临床常用的 AFP、AFP-L3、PIVKAⅡ 3 项肝癌标志物的数据归一化和 LR 建模,可辅助临床预测 HCC 是否发生 MVI,对于 HCC 患者的精准施治和临床预后判断有积极意义。

关键词:微小血管侵犯; 归一化; 逻辑回归; 模型; 原发性肝癌**DOI:**10.3969/j.issn.1673-4130.2022.24.007**文章编号:**1673-4130(2022)24-2973-04**中图法分类号:**R735.7**文献标志码:**A

Multi-parameters model based on normalized regression algorithm in prediction of microvascular invasion of primary hepatocellular carcinoma^{*}

WANG Ziyi¹, XIAO Xiao², HUANG Chenjun², TONG Lin³, GAO Chunfang^{2,3}, CAO Hongwei^{1△}

1. Department of Information, Changhai Hospital, Naval Military Medical University, Shanghai 200438, China; 2. Clinical Laboratory Medicine Center, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200437, China; 3. Department of Clinical Laboratory, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai 200438, China

Abstract: Objective To establish a logistic regression (LR) model based on the normalized detection results of AFP, AFP-L3 and PIVKAⅡ for the predicting of microvascular invasion (MVI) in primary hepatocellular carcinoma(HCC). **Methods** A total of 1 314 cases of HCC verified by operation after entering the group used the Python to conduct the data set 7 : 3(modeling group : verification group) division. The detection results of AFP, AFP-L3 and PIVKAⅡ before operation, and postoperative MVI classification information were collected from the patients entering the group. The detection data were normalized in the section of (-1,1). The LR model was constructed. The efficacy of the model for predicting MVI was evaluated in the modeling group and verification group. **Results** The area under curve (AUC) of the model in distinguishing whether MVI occurring was 0.647, although which was similar to that of independently adopting AFP indicator alone; but in the verification group, AUC of adopting LR model for diagnosing MVI was 0.720, which was higher than the diagnostic efficacy of AFP, AFP-L3 and PIVKAⅡ used alone. **Conclusion** The normalization of data based on clinical common 3 liver cancer markers (AFP, AFP-L3 and PIVKAⅡ) and LR modeling could assist in clinical prediction whether MVI occurring in HCC, which has an important significance for the precision treatment and clinical prognosis in the patients with HCC.

Key words: microvascular invasion; normalized; logistic regression; model; primary hepatocellular carcinoma

^{*} 基金项目:上海市卫健委肿瘤分子医学协同创新集群(2019CXJQ03);上海市科委重点项目(17JC1404500)。

作者简介:王孜怡,女,助理工程师,主要从事疾病数据模型研究。 △ 通信作者,E-mail:chw777@163.com。

肿瘤是严重危害人群健康的公共卫生问题。根据最新发布的美国 2022 流行病学数据,肿瘤是仅次于心脏病的第二大致死性疾病,是 60 岁以上人群病死的重要病因。在排名前 5 位的肿瘤相关死亡病因中,男性肝癌位列第 5 位,其中在 40~59 岁人群则位列第 4,在 60~79 岁位列第 5,提示肝癌是 40~79 岁这一中老年年龄段人群的重要肿瘤相关死亡病因^[1]。全世界新发肝癌的 50% 病例在中国^[2],在我国肝癌是仅次于肺癌的高发病率和高病死率肿瘤^[3]。我国年龄和性别分层的流行病学数据表明,肝癌在 15 岁以上男性肿瘤的发病率和病死率中都位列前茅^[3]。我国 80% 以上的原发性肝癌与乙型肝炎病毒(HBV)感染有关,虽然近年来我国 HBV 表面抗原(HBsAg)阳性率呈现下降趋势,但是由于 HBsAg 携带者存量巨大,所以 HBV 感染作为原发性肝癌的高危因素在我国肝癌发病中仍然长期占据高位^[4-5]。目前早期肝癌的 5 年生存率高达 70% 以上,而晚期肝癌 5 年生存率不足 10%^[6],因此早期发现、精准诊断和治疗是提高生存率、降低病死率的重要环节。微小血管侵犯(MVI)是肝癌具有侵袭性、转移性生物学行为的组织学标志,可用于预测肝癌的预后、复发、生存,发生肝内、肝外转移时,肝癌细胞沿着血管迁移,故 MVI 也是临幊上肝癌复发的重要病理指标^[7-9]。通常 MVI 信息的获取需通过手术标本的组织学分析,有严格的手术采集要求和标准,包括准确采样肿瘤与非肿瘤交界面的组织标本等。若能术前精准预测,或者对于未接受手术治疗者进行 MVI 预测,则对患者精准治疗方案的制订和预后判断有重要意义。本研究基于临床常用的 3 种肝癌标志物——甲胎蛋白(AFP)、甲胎蛋白异质体(AFP-L3)、异常凝血酶原Ⅱ(PIVKA Ⅱ),对检测结果数据采用归一化处理并建模后,用于辅助判断原发性肝细胞癌(HCC)患者是否发生 MVI。

1 资料与方法

1.1 一般资料 本研究纳入 2016 年 1 月至 2017 年 12 月在上海东方肝胆外科医院住院并接受 HCC 手术治疗的患者 1 314 例,这些患者均经术后病理检测证实为 HCC 且在未行手术切除之前完成了 AFP、

AFP-L3 和 PIVKA Ⅱ 的检测,并剔除了正在使用维生素 K 和华法林治疗的研究对象,以免对 PIVKA Ⅱ 检测结果产生干扰。本研究经上海东方肝胆外科医院伦理委员会批准,所有患者知情同意。

1.2 仪器与试剂 AFP 检测方法为电化学发光法(Roche),检测设备为罗氏 Cobas e601 全自动免疫分析仪,检测试剂为罗氏诊断公司原装配套试剂。AFP-L3 采用凝集素亲和富集加全自动化学发光检测,凝集素亲和富集试剂盒为北京热景生物技术有限公司产品,富集后得到的核心岩藻糖基化蛋白检测方法同上述 AFP,作为 AFP-L3 数值。计算 AFP-L3/AFP 的百分比^[10],参考试剂盒说明,>10% 定义为 AFP-L3 阳性。PIVKA Ⅱ 采用酶化学发光法检测,检测设备为富士瑞必欧的 LUMIPULSE G1200 全自动免疫分析仪,试剂为配套原装试剂。

1.3 方法 对数据进行缺失值处理,对缺失值的处理有 3 种:删除记录、数据插补和不处理。本研究中采用删除记录的方法。采用逻辑回归(LR)方法,对 3 项常用的肝癌标志物 AFP、AFP-L3、PIVKA Ⅱ 建模,使用 Python 进行数据集 7:3(建模组:验证组)的划分并获得数据的 MVI 预测效率。

1.4 统计学处理 所用数据采用 Python 语言进行统计。归一化处理则将完整的数据缩放在(-1,1)区间内。非正态分布的计量资料以 $M(P_{25}, P_{75})$ 表示,两组间比较采用非参数 Mann-Whitney 检验。计数资料以例数或百分率表示,比较采用 χ^2 检验。以 $P < 0.05$ 为差异有统计学意义。在模型参数方面进行了调整:(1)为防止过拟合,进行正则化处理;(2)使用梯度下降法优化损失函数并调整标本权重以提高模型预测能力。

2 结 果

2.1 入组患者的临床基本信息 依据入组的原发性肝癌组织病理学 MVI 结果,将所有入组的 1 314 例 HCC 患者分为以下两组:无 MVI 组(MVI-0 组)和存在不同程度 MVI 的 MVI-1+MVI-2 组即 MVI-1+2 组,两组的基线信息见表 1。在纳入的 1 314 例 HCC 患者中,存在 MVI 的患者共 616 例,占 46.9%。

表 1 入组的 HCC 患者基本临床特征

项目	MVI-0 组($n=698$)	MVI-1+2 组($n=616$)	合计($n=1 314$)	P
性别[$n(\%)$]				>0.05
男	628(90.0)	562(91.2)	1 190(90.6)	
女	70(10.0)	54(8.8)	124(9.4)	
年龄[$M(P_{25}, P_{75})$, 岁]	54(26,81)	52(22,78)	53(22,82)	<0.05
AFP[$M(P_{25}, P_{75})$, $\mu\text{g/L}$]	22.9(0.6,1 261 300)	217.7(0.6,718 000)	71.5(0.6,11 261 300)	>0.05
AFP-L3[$n(\%)$]				<0.05
阴性	467(68.5)	293(48.3)	760(59.0)	
阳性	215(31.5)	313(51.7)	528(41.0)	
PIVKA Ⅱ[$M(P_{25}, P_{75})$, mAU/mL]	204(12,336 821)	815(11,265 092)	387.5 (11,336 821)	>0.05

2.2 模型与诊断 使用 Pycharm 软件建模,自建 LR 模型公式,即 $1.08 \times \text{AFP} + 0.34 \times \text{AFP-L3} + 0.26 \times \text{PIVKA II}$ 。使用 Python 进行数据集 7 : 3(建模组 : 验证组)的划分并预测 MVI 结果。

2.3 模型对 MVI 的预测价值 建模组中 LR 模型诊断 MVI 的 AUC 为 0.647, 与单独采用 AFP 诊断 MVI 的相同, 见表 2。但在验证组中, 与 AFP、AFP-L3、PIVKA II 检测相比, LR 模型诊断 MVI 的 AUC 为 0.720, AUC 高于单独应用 AFP-L3 和 PIVKA II 诊断 MVI 的 AUC(0.651、0.601), 也略高于单独应用 AFP 诊断 MVI 的 AUC(0.700)。当 LR 模型最佳截断值为 -1.16 时, 诊断 MVI 的灵敏度为 77.1%, 准确度为 64.7%。见表 3。

表 2 建模组各指标对 MVI 的诊断性能

指标	最佳截断值	AUC	灵敏度 (%)	特异度 (%)	准确度 (%)
模型	-1.16	0.647	66.7	51.5	60.0
AFP	-1.37 ng/mL	0.647	97.5	6.9	56.8
AFP-L3	0.14%	0.591	66.7	51.5	60.0
PIVKA II	-0.18 mAU/mL	0.604	99.0	0.3	54.6

表 3 验证组各指标对 MVI 的诊断性能

指标	最佳截断值	AUC	灵敏度 (%)	特异度 (%)	准确度 (%)
模型	-1.16	0.720	77.1	53.0	64.7
AFP	-1.37 ng/mL	0.700	98.7	3.0	49.2
AFP-L3	0.14%	0.651	77.1	53.0	64.7
PIVKA II	-0.18 mAU/mL	0.601	99.3	1.0	48.3

3 讨 论

本研究主要基于目前临床常用的肝癌标志物 AFP、AFP-L3、PIVKA II 开展对于 HCC 患者是否存在 MVI 的非创性预测,之前已有类似的研究^[9]。本研究通过数据归一化处理和 LR,建立了多参数诊断模型,用于 HCC 患者 MVI 的预测,预测患者是否发生 MVI 的 AUC 达到 0.720,高于单独应用 AFP-L3 和 PIVKA II(0.651、0.601),也略高于单独应用 AFP(0.700)。术前 MVI 的非创性预测有助于辅助临床制订更为精准的手术治疗方案,而对于非手术治疗的 MVI 预测,则可辅助临床判断 HCC 复发、转移及预后。

本研究中选用 AFP、AFP-L3、PIVKA II 的依据是基于目前国际和临幊上已获得一定认可用于 HCC 诊断的 GALAD 模型。GALAD 模型最初于 2014 年由 JOHNSON 等^[11]建立,2016 年 BERHANE 等^[12]基于国际多中心队列进行了充分验证,结果提示其可用于 HCC 诊断并具有较单个指标更好的诊断性能。

GALAD 模型包括性别(G)、年龄(A)以及 3 种血清学标志物 AFP(A)、AFP-L3(L)、PIVKA II(D)共 5 项。本团队在前期国内多中心队列研究中证明其对于 HCC 的早期诊断具有很好价值^[12-13],但 GALAD 模型用于 MVI 的预测效果尚有限^[13]。本研究在预实验中发现,加入性别、年龄并没有显著提升该模型的效率,因此本研究仅针对临床常用的指标展开预测 MVI 的建模性和验证性研究。

本研究采用的建模方法是 LR 方法,LR 是一种线性分类器,可以定量地描述某一现象和某些因素之间的函数关系,将各变量的已知值代入回归方程可以求出应变量的估计值,从而可以进行预测等相关研究。LR 方法是目前临床多参数建模性研究中的最常用方法^[14]。目前已经获得 NMPA 注册证的微小 RNA panel 就是采用了 LR 模型^[15],其作为 HCC 标志物已被纳入 2022 版的中国原发性肝癌防治指南^[5]。

数据缺失和数据标准化处理是建模性研究中的关键问题。数据缺失主要包括记录缺失和字段信息缺失等情况,对数据分析有较大影响,使结果不确定性增加。通常对缺失值的处理有 3 种:删除记录、数据插补和不处理。本项目采用删除记录的方法,最大程度减少了因数据缺失对模型稳定性带来的影响。在多指标评价体系中,由于各评价指标的性质不同,通常具有不同的量纲和数量级。当各指标间的水平相差很大时,如果直接用原始指标值进行分析,就会突出数值较高的指标在综合分析中的作用,相对削弱数值水平较低指标的作用。因此,为了保证结果的可靠性,需要对原始指标数据进行标准化处理。目前数据的标准化方法有多种,归结起来可以分为直线型方法(如极值法、标准差法)、折线型方法(如三折线法)、曲线型方法(如半正态性分布)。不同的标准化方法,对系统的评价结果会产生不同的影响,但在数据标准化方法的选择上,还没有通用的法则可以遵循。其中最典型的方法就是数据的归一化处理,即将数据统一映射到(-1,1)区间上。归一化是一种简化计算的方式,即将有量纲的表达式,经过变换,化为无量纲的表达式,成为纯量。归一化是让不同维度之间的特征在数值上有一定比较性,可以大大提高分类器的准确性。各个特征对结果做出的贡献相同时,可以比较出参数之间分类效果的差异。归一化后有提升模型的收敛速度和精度的优势。采用归一化处理数据后,本研究中所建立的模型较本团队前期建立的 MVI 模型的预测效率有了明确提升^[16]。在本研究中,其应用优势还在于能缩小由于不同临幊实验室采用不同检测系统可能对模型稳定性带来的影响,这也是本研究的亮点之一。

综上所述,本研究基于临床常用的 3 项肝癌标志物(AFP、AFP-L3、PIVKA II),通过数据归一化和 LR 建模,可辅助临床进行 HCC 患者发生 MVI 的预测,对于 HCC 患者的精准施治和临床预后判断提供了全新手段。未来需要进一步多中心研究验证该模型的稳定性和有效性,并在实践中不断优化。

参考文献

- [1] SIEGEL R L, MILLER K D, FUCHS H F, et al. Cancer statistics, 2022[J]. CA Cancer J Clin, 2022, 72(1): 7-33.
- [2] XIE L, XIA R, ZHUANG G. Cost-effectiveness of antiviral treatment after resection in hepatitis B virus-related hepatocellular carcinoma patients with compensated cirrhosis[J]. Hepatology, 2018, 68(4): 1476-1486.
- [3] ZHENG R S, ZHANG S W, ZENG H M, et al. Cancer incidence and mortality in China, 2016[J]. JNCC, 2022, 2(1): 1-9.
- [4] 中华医学会感染病学分会,中华医学会肝病学分会.慢性乙型肝炎防治指南:2019 年版[J].中华肝脏病杂志,2019,27(12):938-961.
- [5] 中华人民共和国国家卫生健康委员会医政医管局.原发性肝癌诊疗指南:2022 年版[J].中华肝脏病杂志,2022, 30(4):367-388.
- [6] ZHAO J, GREENE C M, GRAY S G, et al. Long noncoding RNAs in liver cancer: what we know in 2014[J]. Expert Opin Ther Targets, 2014, 18(10): 1207-1218.
- [7] ROAYAIE S, BLUME I N, THUNG S N, et al. A system of classifying microvascular invasion to predict outcome after resection in patients with hepatocellular carcinoma [J]. Gastroenterology, 2009, 137(3): 850-855.
- [8] 丛文铭,步宏,陈杰,等.原发性肝癌规范化病理诊断指
- [9] 黄晨军,孙小娟,童林,等.基于肝癌三联检的原发性肝细胞癌微血管侵犯的术前预测研究[J].现代免疫学,2020, 40(2): 115-120.
- [10] 季君,顾星,高致远,等.凝集素捕获微量离心柱法检测甲胎蛋白异质体在原发性肝癌中的应用评价[J].检验医学,2011, 26(4): 256-259.
- [11] JOHNSON P J, PIRRIE S J, COX T F, et al. The detection of hepatocellular carcinoma using a prospectively developed and validated model based on serological biomarkers[J]. Cancer Epidemiol Biomarkers Prev, 2014, 23(1): 144-153.
- [12] BERHANE S, TOYODA H, TADA T, et al. Role of the GALAD and BALAD-2 serologic models in diagnosis of hepatocellular carcinoma and prediction of survival in patients[J]. Clin Gastroenterol Hepatol, 2016, 14: 875-886.
- [13] HUANG C, FANG M, XIAO X, et al. Validation of the GALAD model for early diagnosis and monitoring of hepatocellular carcinoma in Chinese multicenter study[J]. Liver Int, 2022, 42(1): 210-223.
- [14] 黄晨军,高春芳.注重肝病标志物及其模型算法应用:实践与展望[J].中华检验医学杂志,2021, 44(6): 457-461.
- [15] ZHOU J, YU L, GAO X, et al. Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma. [J]. J Clin Oncol, 2011, 29(36): 4781-4788.
- [16] 童林,高致远,黄晨军,等.GALAD 模型在原发性肝细胞癌诊断及微血管侵犯预测中的作用[J].中华检验杂志,2019, 42(12): 1037-1041.

(收稿日期:2022-05-11 修回日期:2022-10-17)

(上接第 2972 页)

- [11] MORGAN CHRISTOPHER E, ZHANG Z M, BONO-MO ROBERT A, et al. An analysis of the novel fluorocycline TP-6076 bound to both the ribosome and multidrug efflux pump AdeJ from *Acinetobacter baumannii* [J]. mBio, 2022, 13(1): e0373221.
- [12] BRATU S, LANDMAN D, MARTIN D A, et al. Correlation of antimicrobial resistance with beta-lactamases, the OmpA-like porin, and efflux pumps in clinical isolates of *Acinetobacter baumannii* endemic to New York City[J]. Antimicrob Agents Chemother, 2008, 52(9): 2999-3005.
- [13] ZHANG L, LIANG W, XU S G, et al. CarO promotes adhesion and colonization of *Acinetobacter baumannii* through inhibiting NF- κ B pathways[J]. Int J Clin Exp Med, 2019, 12: 2518-2524.
- [14] ZHU L J, CHEN X Y, HOU P F. Mutation of CarO participates in drug resistance in imipenem-resistant *Acine-*

tobacter baumannii [J]. J Clin Lab Anal, 2019, 33(8): e22976.

- [15] 高春艳,侯盼飞.膜孔蛋白改变在鲍曼不动杆菌对亚胺培南耐药中的作用[J].中国微生态学杂志,2018, 30(10): 1135-1138.
- [16] ABBASI E, GOUDARZI H, HASHEMI A, et al. Decreased CarO gene expression and OXA-type carbapenemases among extensively drug-resistant *Acinetobacter baumannii* strains isolated from burn patients in Tehran, Iran[J/OL]. Acta Microbiol Immunol Hung, [2022-04-17]. <https://pubmed.ncbi.nlm.nih.gov/31706000/>.
- [17] 潘红超,商安全,张康见,等.亚胺培南耐药的鲍氏不动杆菌 OXA、AdeABC、CarO 基因及生物膜的研究[J].中华医院感染学杂志,2016, 26(20): 4572-4575.

(收稿日期:2022-04-20 修回日期:2022-10-15)