

鼻咽癌放疗后大出血风险机器学习预测模型构建^{*}

葛晓伟 李星丹 张伟祎 底瑞青 程 铭

(郑州大学第一附属医院 郑州 450008)

〔摘要〕 目的/意义 构建鼻咽癌放疗后大出血风险预测模型,并评价其预测效能。方法/过程 选取郑州大学第一附属医院 2016—2019 年鼻咽癌放疗后大出血的住院患者为研究对象,随机选取同等数量未出现大出血的患者为对照组,收集两组患者的病历指标数据,分别应用多种机器学习算法并选取最优算法构建模型。结果/结论 基于支持向量机算法的模型召回率为 0.94、 $F1$ 值为 0.93、精确度为 0.93,效果最好,可用于构建鼻咽癌放疗后大出血预测模型,为患者提供更精确的个体化预测,具有良好的临床应用前景。

〔关键词〕 鼻咽癌;大出血;疾病预测模型;机器学习

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.07.015

Construction of a Machine Learning Prediction Model for the Risk of Massive Hemorrhage After Radiotherapy for Nasopharyngeal Carcinoma

GE Xiaowei, LI Xingdan, ZHANG Weiyi, DI Ruiqing, CHENG Ming

The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450008, China

〔Abstract〕 **Purpose/Significance** To construct a risk prediction model for postoperative massive bleeding in nasopharyngeal carcinoma after radiotherapy, and to evaluate its predictive performance. **Method/Process** Inpatients with major bleeding after radiotherapy for nasopharyngeal cancer in the First Affiliated Hospital of Zhengzhou University from 2016 to 2019 are selected as the study objects, and the same number of patients without major bleeding are randomly selected as the control group. The medical record index data of the two groups of patients are collected, and various machine learning algorithms are applied respectively and the optimal algorithm is selected to build the model. **Result/Conclusion** The model based on support vector machine (SVM) algorithm has a recall rate of 0.94, an $F1$ value of 0.93, and a precision of 0.93, showing the best performance. It can be used to construct a prediction model for postoperative massive bleeding in nasopharyngeal carcinoma, and provide more accurate personalized prediction for patients, which has good clinical application prospects.

〔Keywords〕 nasopharyngeal cancer; massive bleeding; disease prediction model; machine learning

〔修回日期〕 2024-02-28

〔作者简介〕 葛晓伟,硕士,工程师,发表论文 9 篇;通信作者:程铭,博士,高级工程师,硕士生导师。

〔基金项目〕 河南省重点研发与推广专项(科技攻关)项目(项目编号:222102210112);河南省医学科技攻关计划软科学项目(项目编号:RKX202202021)。

1 引言

鼻咽癌放疗后出现大出血症状最为凶险，死亡率达 40%。即使抢救成功，大多数患者也会面临如肝肾衰竭、颅内感染、脑疝、偏瘫等并发症^[1-3]，严重影响患者疗效、预后及生存质量。因此，如何有效预测鼻咽癌放疗后大出血，并进行早期及时防护是临床亟待解决的重要问题。近年来计算机断层扫描（computed tomography, CT）模拟定位和适形放疗等新技术的应用减少了肿瘤周围正常组织的照射剂量，提高了放射治疗靶区和鼻咽癌肿瘤组织的适形度^[4]，但鼻咽癌放疗后患者大出血仍是不可忽略和最凶险的并发症。探索鼻咽癌放疗后大出血的风险因素可以明确高风险人群，及早给予干预，降低鼻咽癌放疗后出现大出血的可能。然而，目前只有相对较少的回顾性小规模研究讨论这些问题，且研究者列举的风险因素较多，究竟哪些因素可作为鼻咽癌放疗后大出血风险敏感预测指标，并用于建立预测模型成为亟待解决的问题之一。

随着人工智能技术的快速发展，机器学习在医疗领域，尤其是疾病风险预测方面得到广泛应用^[5-6]。传统统计方法擅长分析线性关系，或得出有无差异、相关性等结论，且常对变量有一定要求，如要求量的数据呈正态分布或变量间相互独立；机器学习算法对于复杂的非线性关系数据更有优势。基于机器学习的算法包括支持向量机（support vector machine, SVM）、人工神经网络、决策树、随机森林、朴素贝叶斯算法等^[7]，不仅可以很好地解决多重共线性问题，而且对数据分布限制较小。已有研究表明，机器学习算法用于预测麻醉诱导后低血压^[8]、脑卒中复发^[9]、住院患者术后死亡率^[10]、住院患者跌倒^[11]等方面可获得良好效果，与传统统计学方法相比，可以为患者提供更准确的个体预测，具有更好的预测性能和效果^[12]。何霞霞等^[13]收集呼吸内科 117 例患者资料，共提取 21 项 CT 影像学特征和 5 项临床参数，应用人工神经网络及逻辑回归模型比较肺癌辅助诊断模型准确度，结果显示人工神经网络模型判断肺部良恶性病变准确率更高，对 44 例

样本预测的准确度为 90.9%，逻辑回归模型为 68.2%。尽管机器学习算法在医学领域的应用和评估研究迅速增加，但构建鼻咽癌放疗后大出血预测模型进行个体化预测仍需进一步研究。

因此，选择最恰当的风险指标和分类算法，构建鼻咽癌放疗后大出血风险预测模型，并评价预测效能，有助于及时准确识别鼻咽癌放疗后大出血风险患者，及早给予针对性防护措施，降低患者大出血风险，提高患者生存质量，节约医疗资源。

2 对象与方法

2.1 病例来源与入选标准

本研究数据来源于郑州大学第一附属医院，选取其 2016—2019 年符合纳入、排除标准的鼻咽癌放疗后大出血患者 67 例，纳入标准：病理确诊为鼻咽癌并接受放疗者；放疗后出血，一次出血总量 > 300 毫升或出血量 > 100 毫升且反复发生者。排除标准：鼻内镜等检查为普通鼻腔出血者；病历资料不全者。

以年龄、性别为匹配因素随机抽取相应例数的鼻咽癌放疗后未出现大出血患者 67 例作为对照组，基于文献回顾性风险指标研究总结，收集两组患者的性别、年龄、抗凝药、造影、肿瘤分期、放疗剂量、放疗次数、血凝等 33 个临床指标共 9 975 条数据，见表 1。

表 1 临床指标

指标类别	详细指标
个人因素	性别、年龄
主观因素	头晕、烦躁、鼻塞、鼻咽出血量、鼻出血、鼻涕、脓涕、头疼、鼻疼、耳闷、口渴
药物因素	口服抗凝药
合并慢性病	高血压、糖尿病、其他慢性病
个人史	吸烟、饮酒
肿瘤因素	部位、分期、颅底侵犯、病理类型
检验因素	血红蛋白、红细胞、红细胞比容、血小板、血小板分布宽度、血管造影是否存在血管损伤、凝血酶原时间
治疗因素	放疗剂量、再程放疗、放疗次数

2.2 纳入指标数据预处理

从医院电子病历系统、实验室信息管理系统、影像存储与传输系统中抽取患者各项指标数据，建立原始特征库，抽取范围仅限前文表1中的临床指标信息，并根据抽取的原始数据初步筛选，对初步筛选后的数据进行缺失值与噪声处理，并量化部分非数值型指标。缺失值的产生原因包括人为手工数据、机器故障等，缺失值是随机的，采用平均值修正方法处理；删除波动太大的数据，删除与特征均值比大于60和小于0.01的数据。经过数据预处理，实际有效患者样本数据共6123条，以放疗后大出血患者数量为准，放疗后未出现大出血患者随机取相同数据作为样本集，通过设计预测评估框架，选择机器学习中最优的分类预测方法。

2.3 数据特征指标分析

特征指标的选定有助于降低数据的拟合度、提高算法的精度，减少训练时间，避免对预测新数据造成干扰，采用卡方检验方法对鼻咽癌放疗后是否出现大出血症状的患者各项指标进行差异性分析，见表2。选取前20个在鼻咽癌放疗后是否出现大出血区分上差异均较大的指标，拟纳入机器学习方法试验。

表2 鼻咽癌放疗后大出血分类各项指标卡方值

序号	指标	卡方值
1	性别	2.410×10^{-3}
2	年龄	6.935×10^{-2}
3	头晕	0
4	烦躁	0
5	鼻塞	1.009×10^{-1}
6	鼻咽出血量*	25.86
7	鼻出血*	1.096
8	鼻疼*	1.096
9	脓涕*	8.116
10	头疼*	7.482
11	血涕*	12.63
12	耳闷*	2.354
13	口渴*	2.126
14	口服抗凝药	1.820×10^{-2}
15	高血压*	2.153×10^{-1}
16	糖尿病	9.739×10^{-3}
17	其他慢性病	7.218×10^{-2}
18	吸烟	0

续表 2

序号	指标	卡方值
19	饮酒	1.463×10^{-2}
20	部位*	5.468×10^{-1}
21	分期*	21.78
22	颅底侵犯*	1.747×10^{-1}
23	病理类型	9.543×10^{-2}
24	血红蛋白*	1.448×10^{-1}
25	红细胞*	9.946×10^{-2}
26	红细胞比容*	1.249×10^{-1}
27	血小板	9.293×10^{-2}
28	血小板分布宽度*	7.625×10^{-1}
29	血管造影存在血管损伤*	36.14
30	血凝酶原时间	1.372×10^{-3}
31	放疗剂量*	4.610
32	再程放疗*	32.49
33	放疗次数*	3.695

注：卡方检验统计样本的实际观测值与理论观测值之间的偏离程度，偏离程度决定卡方值的大小，卡方值越大，越不符合，越能用于区分；卡方值越小，偏差越小，越趋于符合；若卡方值为0，表明理论值完全符合。*为纳入指标。

纳入机器学习算法前，采用数据特征直方图、关系矩阵图分析选定的指标数据，见图1。大部分数据存在一定程度的偏态分布，需要对数据正态化处理，通过Box-Cox转换提高模型的准确度；另由关系矩阵可知，数据具有一定的负相关性，能够对数据分类预测。

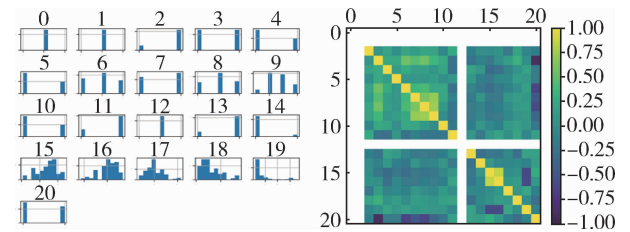


图1 数据特征直方图与关系矩阵

注：数据直方图表示数据变化情况，可以解析出数据的规则性，能够比较直观地展示数据质量特性的分布状态，便于判断其总体质量分布情况；数据关系矩阵图是分析数据分布情况的直观可视化方法，可以用来进行数据的质量控制和差异数据的具象化展示。

2.4 构建分类预测模型

通过设计基于机器学习的分类预测评估框架选择合适的分类预测算法。选择常用的逻辑回归算法

(logistic regression, LR)、分类与回归树算法 (classification and regression trees, CART)、SVM、朴素贝叶斯 (naive Bayes, NB) 和 K 近邻 (K - nearest neighbor, KNN) 5 种算法^[14], 采用 10 折交叉验证分离数据, 并通过准确度比较算法, 找到最优算法, 见图 2。LR 与 NB 结果分布比较紧凑, 但存在异常点, 考虑数据量的问题, 选取 KNN、CART 与 SVM 3 种算法, 但需进一步优化。对 SVM 算法调参, 对放疗后大出血患者与未出血患者分类预测, 使用参数 { ‘C’: 0.1, ‘kernel’: ‘linear’ } 时, 分类预测模型的性能最好为 0.94; 针对 KNN 算法, 使用参数 { ‘n_neighbors’: 3 } 时, 分类预测模型的性能最好为 0.88; 针对 CART 算法, 参数 { ‘n_estimators’: 160 } 时, 分类预测模型的性能最好, 具有最优的精确率为 0.93。

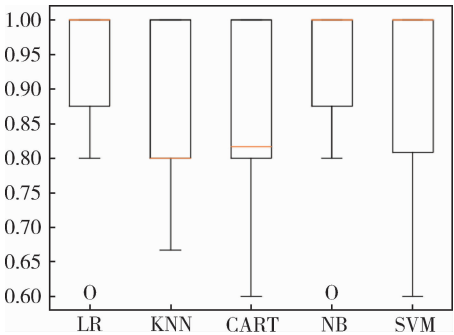


图 2 分类预测算法评估结果

注: 箱线图能够有效展示数据结果的特征, 直观地识别数据集异常值 (查看离群点), 判断数据集的数据离散程度和偏向 (观察盒子的长度, 上下隔间的形状, 以及延伸线的长度)。

3 分类预测结果分析

选择召回率 (recall, R)、精确率 (precision, P) 和 F1 值评价鼻咽癌放疗后患者出现大出血与否分类预测效果。R 中 TP 是将两组相似的患者特征指标归为一类, FP 是将两段不相似的患者特征指标归为一类。同理, 根据不相似的患者特征指标归为不同类及相似的归为不同类计算 P。

$$R = \frac{TP}{TP + FP}$$

(1)

F_{β} 值是 R、P 的调和平均值。其中, 取 $\beta = 1$,

使 F1 值中的 R、P 权重相同, 根据相似患者特征指标的归类与总文本数量计算 P。

$$F_{\beta} = \frac{(\beta^2 + 1) PR}{(\beta^2 P + R)} \quad (\beta^2 \in [0, +\infty])$$

(2)

算法评估结果, 见表 3。SVM 具有更好的预测效果, 鼻咽癌放疗后患者大出血预测 P 达 0.93, 这可能与 SVM 泛化性能较强, 可解决小样本情况下机器学习问题的特性有关; 而 KNN 对异常值不敏感, 对不相关功能和数据规模敏感; CART 能够处理大规模数据集和高维数据, 但其贪心性质的搜索策略, 可能会找到次优解而不是全局最优解, 且对噪声和异常值比较敏感, 可能会过度训练, 导致泛化能力下降。

表 3 鉴别分类预测结果

算法	P	R	F1
KNN	0.92	0.91	0.91
CART	0.87	0.88	0.87
SVM	0.93	0.94	0.93

4 讨论

本研究基于对文献报道中鼻咽癌放疗后大出血影响因素的整理, 收集两组患者相关数据, 分析指标差异性, 将差异较大的指标纳入机器学习算法。血管造影存在血管损伤、鼻咽出血量、再程放疗、分期、血涕 5 个指标的重要值排在前 5 位, 是影响鼻咽癌放疗后大出血发生的重要因素。

研究^[15]证实, 放疗造成的血管损伤是鼻咽癌放疗后大出血的重要影响因素。放疗后局部组织低供血、低供氧和微循环障碍, 会影响胶原蛋白和细胞再生, 动脉过早粥样硬化、外膜纤维化、动脉壁坏死阻塞, 极易诱发血管破溃、假性动脉瘤形成或破裂出血。再程放疗也是其影响因素之一, 放疗会造成鼻腔黏膜纤毛排列紊乱、黏连、上皮化生, 细胞内空泡形成, 细胞间隙扩大, 通透性增强, 鼻腔充血水肿, 分泌物增多, 当患者再程放疗或接受较高放射剂量时, 鼻咽部黏膜上皮损伤进一步加重, 腺体分泌减少, 黏液纤毛功能障碍, 黏膜炎症及溃疡加重, 诱发血管破裂出血。肿瘤分期也是不容忽视的影响因素, 肿瘤分期越晚, 肿瘤组织浸润越广

泛,特别是颅底破坏明显、侵及血管者,更易发生大出血。在大出血先兆征象方面,吴润坤等^[16]分析 85 例鼻咽癌患者临床资料,45 例放疗后大出血的患者出血前均出现反复头痛、口鼻腔反复小出血、鼻腔分泌物恶臭增多和程度不等的持续性贫血等先兆征象,这与本研究结果一致,评估患者是否反复出现鼻出血、血涕等,应当作为医护人员观察评估的重要因素,警惕患者大出血。

本研究评估 5 种机器学习模型预测鼻咽癌放疗后大出血风险,首先采用分类预测算法评估结果的箱线图,排除存在异常点的 LR 与 NB 算法,选择 KNN、CART 与 SVM 调参优化模型,对其试验结果采用召回率、精确率和 $F1$ 值评价指标进行比较,研究结果表明,SVM 具有最好的精确性,预测精确率达 0.93,基于机器学习算法构建风险预测模型对于预测鼻咽癌放疗后大出血具有重要意义。

5 结语

使用恰当的机器学习算法构建预测模型,可以辅助临床决策。对于鼻咽癌放疗患者而言,了解放疗后大出血并发症发生的风险程度非常重要。本研究数据来自郑州大学第一附属医院收集的患者住院病历,无须进行额外的临床评估,得到的预测模型可用于识别鼻咽癌放疗后大出血风险,帮助医护人员识别高风险患者,及早给予干预,降低其发生率,减少医疗成本。但是,对纳入研究病例的病历完整性有较高要求,因而两组纳入样本量有限,对于重要指标的纳入,以及最优模型的筛选,可能存在偏倚;下一步将考虑多中心研究,进一步扩大样本量,以构建预测精确率更高的模型,为应用于临床鼻咽癌大出血患者的预测提供支持。

作者贡献: 葛晓伟负责实验设计与实施、论文撰写;李星丹负责数据收集与整理;张伟祎负责图表制作;底瑞青负责提供指导、论文审核与修订;程铭负责概念构思与设计。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

1 张文静,底瑞青,叶琳,等.鼻咽癌放疗后鼻出血干预

- 研究现状 [J]. 护理研究, 2021, 35 (1): 99 - 104.
- 2 倪松,徐震纲,王晓雷,等.头颈部肿瘤术后致命性大出血分析 [J]. 中华肿瘤杂志, 2010, 32 (1): 60 - 63.
- 3 JIA G D, SUN X S, LI X Y, et al. Clinical characteristics and prognostic factors affecting survival after radical radiotherapy for early and late post - treatment metastatic nasopharyngeal carcinoma [J]. BMC cancer, 2023, 23 (1): 7.
- 4 吴烽芳,黄艺峰,刘雅雪,等.鼻咽癌放疗后鼻咽部出血原因分析及治疗对策探讨 [J]. 中国耳鼻咽喉颅底外科杂志, 2019, 25 (5): 542 - 545.
- 5 陈瑜. 人工智能在护理领域的应用及思考 [J]. 护士进修杂志, 2020, 35 (3): 247 - 251.
- 6 乔红艳,郭邦俊,张龙江. 机器学习在心血管影像中的研究进展 [J]. 中华医学杂志, 2019, 99 (17): 1353 - 1357.
- 7 董国辉,高波,潘云. 人工智能在鼻咽癌诊断与预后领域的研究进展 [J]. 现代肿瘤医学, 2023, 31 (9): 1765 - 1768.
- 8 KENDALE S, KULKARNI P, ROSENBERG A D, et al. Supervised machine - learning predictive analytics for prediction of postinduction hypotension [J]. Anesthesiology, 2018, 129 (4): 675 - 688.
- 9 郑雯丽,邓仁丽,邱业银,等. 缺血性脑卒中复发风险预测模型的研究进展 [J]. 中华神经医学杂志, 2022, 19 (4): 408 - 412.
- 10 LEE C K, HOFER I, GABEL E, et al. Development and validation of a deep neural network model for prediction of postoperative in - hospital mortality [J]. Anesthesiology, 2018, 129 (4): 649 - 662.
- 11 任媛渊,丁福,付荣娟. 国外基于机器学习的住院患者跌倒风险预测模型构建研究及其启示 [J]. 医学信息学杂志, 2023, 44 (10): 63 - 67, 80.
- 12 冯利,岳小飞. 9 种结合式机器学习算法在肿瘤早期诊断中的准确性比较研究 [J]. 医学信息学杂志, 2022, 43 (5): 25 - 29.
- 13 何霞霞,张红升,李迪,等. 基于 CT 影像评分的人工神经网络模型对肺部良恶性病变的判别价值 [J]. 郑州大学学报 (医学版), 2018, 53 (6): 723 - 726.
- 14 XIANG L. A prototype selection technique based on relative density and density peaks clustering for k nearest neighbor classification [J]. Intelligent data analysis, 2023 (3): 27.
- 15 YU Y H, XIA W X, SHI J L, et al. A model to predict the risk of lethal nasopharyngeal necrosis after re - irradiation with intensity - modulated radiotherapy in nasopharyngeal carcinoma patients [J]. Chinese journal of cancer, 2016, 35 (1): 59.
- 16 吴润坤,陈振清,陈玉兰,等. 鼻咽癌患者放疗后大出血的危险因素、征象及防治 [J]. 内科, 2017, 12 (3): 392 - 394.