论著·临床研究

随机森林模型和Logistic回归模型预测非计划再手术发生风险的效能比较▲

豆 娟[1] 王 旭[2] 吴嘉越[3] 赵英英[1]
（上海市第一人民医院1医务处，2神经外科，3信息处，上海市 200080）

【摘要】 **目的** 比较随机森林模型和Logistic回归模型预测非计划再手术发生风险的效能。**方法** 在手术麻醉系统中筛选一次住院期间申请2次手术的患者信息。提取所有非计划再次手术患者（n=219）作为研究组，对应科室的计划再次手术患者（n=14 311）作为对照组。运用随机森林模型和Logistic回归模型建立非计划再手术预测模型。采用受试者工作特征曲线下面积评价两种模型的预测效能。**结果** （1）Logistic回归分析结果显示，前次术中输血、罹患恶性肿瘤、合并疾病数量、前次手术切口愈合等级、前次手术级别、前次手术时长、前次手术切口类别是非计划再手术发生的影响因素（P<0.05）。Logistic回归预测模型的曲线下面积为0.922，灵敏度、特异度、准确率分别为92.59%、79.11%、79.28%。（2）随机森林模型特征变量的重要性排序结果显示，前次手术切口类别、前次术中输血、前次手术级别、前次手术切口愈合等级、合并疾病数量、罹患恶性肿瘤等变量的重要性更靠前。随机森林预测模型的曲线下面积为0.866，灵敏度、特异度、准确率分别为80.00%、93.33%、86.66%。Logistic回归预测模型曲线下面积大于随机森林预测模型，但差异无统计学意义（P>0.05）。**结论** 综合使用Logistic回归模型和随机森林模型，并将二者分析结果互为补充，可从各个方面预测非计划再次手术的风险因素，能获得更好的预测效能。

【关键词】 非计划再手术；随机森林模型；Logistic回归模型；风险因素；预测模型
【中图分类号】 R 197.32 【文献标识码】 A 【文章编号】 0253-4304（2024）04-0501-05
DOI：10.11675/j.issn.0253-4304.2024.04.07

Comparison of efficiency between random forest model and Logistic regression model for predicting the occurrence risk of unscheduled resurgery

DOU Juan[1], WANG Xu[2], WU Jiayue[3], ZHAO Yingying[1]
(1 Section of Medical Affairs, 2 Department of Neurosurgery, 3 Section of Information, Shanghai General Hospital, Shanghai 200080, China)

【**Abstract**】 **Objective** To compare the efficiency between random forest model and Logistic regression model for predicting the occurrence risk of unscheduled resurgery. **Methods** Patients′ information who requested two-times surgery during one hospital stay was screened from the surgical anesthesia system. All patients undergoing unscheduled resurgery (n=219) were extracted as study group, whereas patients in the corresponding departments who underwent scheduled resurgery (n=14,311) were as control group. The unscheduled resurgery prediction model was established by the application of random forest model and Logistic regression model. The prediction efficiency of the two models was evaluated by employing area under the receiver operating characteristic curve. **Results** (1) The results of Logistic regression analysis revealed that previous intraoperative blood transfusion, suffering from malignant tumor, number of comorbidity, previous healing classification of surgical incision, previous surgical level, previous surgical duration, previous category of surgical incision were the influencing factors for the occurrence of unscheduled resurgery (P<0.05). Area under the curve of Logistic regression prediction model was 0.922, and the sensitivity, specificity, and accuracy rate were 92.59%, 79.11%, and 79.28%, respectively. (2) The results of importance ranking of characteristic variables in random forest model indicated that the importance of variables such as previous category of surgical incision, previous intraoperative blood transfusion, previous surgical level, previous healing classification of surgical incision, number of comorbidity, and suffering from malignant tumor came top in. Area under the curve of random forest prediction model

▲基金项目：上海申康医院发展中心临床管理优化项目（SHDC12022622）
第一作者简介：豆娟，硕士，助理研究员，研究方向为医疗质量管理。
通信作者简介：赵英英，硕士，副主任医师，研究方向为急诊医学和医疗质量管理。

was 0.866, and the sensitivity, specificity, and accuracy rate were 80.00%, 93.33%, and 86.66%, respectively. Area under the curve of Logistic regression prediction model was larger than that of random forest prediction model, but no statistically significant difference was found (P>0.05). **Conclusion** Comprehensive use of Logistic regression model and random forest model, and complementing the results of the two analysis to each other, can predict the risk factors for unscheduled resurgery from various aspects, and obtain better prediction efficiency.

【Key words】 Unscheduled resurgery, Random forest model, Logistic regression model, Risk factors, Prediction model

非计划再次手术是指同一患者在同一住院期间因多种原因而需要进行计划外的再次手术(含介入治疗)[1],导致非计划再次手术的因素包括医源性因素及非医源性因素,前者指经手术治疗后疗效不佳必须再次施行手术,后者指由于患者病情发展或出现严重术后并发症,需要再次进行计划外的手术。此类患者的病死率、住院费用、住院时间和医患纠纷明显增加[2-3]。目前国内外学者对非计划再次手术的研究主要集中于危险因素分析、并发症、发病率和死亡率、戴明环、品管圈和数据管理等方面[4],这说明很多国家已将其作为医疗质量与安全评价的负性评价指标进行监管。2023年国家卫生健康委员会印发的《全面提升医疗质量行动计划(2023—2025年)》[5],要求非计划重返手术室再次手术率不高于1.8‰。因此,预防和控制非计划再次手术的发生具有重要意义。

近年来,机器学习技术在医疗领域应用广泛。随机森林是经典的机器学习算法之一,对纳入模型的数据结构没有要求,不存在共线性与过拟合的情况,而Logistic回归模型简单易用,两者用途广泛。本研究将随机森林算法引入非计划再手术的风险预测研究中,基于多因素Logistic回归分析和随机森林算法构建非计划再手术风险预测模型,为临床及时采取预防措施与改善围术期医疗质量提供参考。

1 资料与方法

1.1 资料来源 资料来源于上海市某三级医院2021—2023年的电子病历系统及手术麻醉系统。在手术麻醉系统中筛选一次住院期间申请2次手术的患者信息。纳入标准:(1)同一次住院期间进行2次手术的患者;(2)再次手术间隔时间<31 d。排除标准:(1)第1次手术出院后再入院的患者;(2)手术类型为活检、穿刺等检查或治疗类操作;(3)临床资料不完整的患者。将非计划再次手术患者(n=219)作为研究组,根据对应的科室筛选出计划再次手术患者(n=14 311)作为对照组。

1.2 非计划再次手术判定流程 经系统接口改造后,在住院患者同一次住院期间申请第二次手术时,系统即自动跳出判断窗,操作医生需要确认是否为非计划二次手术并填写原因。同时,医务部门会接收到信息窗口内容,管理专员进行初筛和判断,有争议时申请专家审核,经过科室和管理部门的两轮审核后判定是否为非计划再次手术。

1.3 临床资料收集 在中国知网数据库中,以北大核心和中文社会科学引文索引认定的学术期刊为主要文献来源,以“非计划二次手术”“非计划再次手术”“质量评价”“危险因素”为检索词,检索相关文献,搜集可能导致非计划再次手术的主要因素。最终从患者、手术、术者3个方面,选取年龄、性别、支付方式、合并疾病数量、是否罹患恶性肿瘤、前次术中输血情况、前次手术时长、前次手术级别、前次主刀医师职称、前次手术切口类别、前次手术切口愈合等级11个指标进行分析。由医院信息处导出两组患者的病案数据,数据收集完成后随机抽取20%的数据进行核对。

1.4 统计学分析 采用SPSS 22.0软件和R语言3.6.0软件进行统计分析。

1.4.1 数据处理:采用K-means法对缺失值填补以改进数据的质量,提高数据分析的可行性和准确性。使用R语言的caret程序包随机将患者按7∶3分成训练集10 171例(研究组165例,对照组10 006例)和测试集4 359例(研究组54例,对照组4 305例),分别用于模型的构建与效能评价。

1.4.2 单因素分析:将11个指标纳入单因素分析。服从正态分布的计量资料以($\bar{x}\pm s$)表示,组间比较采用两独立样本t检验,不服从正态分布的资料以[$M(Q)$]表示,组间比较采用秩和检验;分类变量以例数或构成比表示,组间比较采用χ^2检验。纳入单因素分析中P<0.05的指标,采用训练集数据分别建立Logistic回归预测模型和随机森林预测模型。

1.4.3 Logistic回归预测模型的建立:在训练集数据上,使用R语言中的glm函数构建Logistic回归模型,并利用step函数对构建的初始Logistic模型进行基于赤池信息准则的逐步回归变量筛选,从而构建最佳的Logistic回归预测模型。

1.4.4 随机森林预测模型的建立:利用训练集数据,调用Random Forest程序包进行随机森林模型的训练,设置种子数为1 234,通过importance()和

imp$MeanDecreaseAccuracy()函数输出结果。根据基尼指数平均值排序筛选出对非计划再次手术影响较大的变量，依据计算袋外错误率选择最优的随机森林模型特征个数，建立随机森林分类预测模型。经测试集检验，当ntree=500、mtry=3时，随机森林预测模型达到最优。

1.4.5 预测模型的效能评价：基于测试集数据应用两种预测模型对非计划再手术风险进行预测，绘制受试者工作特征（receiver operating characteristic，ROC）曲线，根据灵敏度、特异度、准确率、曲线下面积（area under the curve，AUC）评价模型的预测效能，采用Z检验比较AUC的差异。

2 结 果

2.1 非计划再次手术发生率及单因素分析结果 2021—2023年该院总手术病例数为137 787例，非计划再次手术219例，总体发生率为0.15%，低于国家基准值（0.44%）[6]。单因素分析结果显示，两组患者性别、罹患恶性肿瘤情况、合并疾病数量、前次手术时长、前次术中输血、前次主刀医师职称、前次手术切口类别、前次手术切口愈合等级、前次手术级别比较，差异有统计学意义（P<0.05），见表1。

表1 单因素分析结果[n(%)]

指标	研究组（n=219）	对照组（n=14 311）	χ^2值	P值
年龄				
0~<65岁	114(52.1)	8 305(58.0)	3.163	0.075
≥65岁	105(47.9)	6 006(42.0)		
性别				
男性	140(63.9)	8 188(57.2)	3.972	0.046
女性	79(36.1)	6 123(42.8)		
支付方式				
自费	32(14.6)	1 700(11.9)	1.534	0.215
医保	187(85.4)	12 611(88.1)		
罹患恶性肿瘤				
否	33(15.1)	7 326(51.2)	112.602	<0.001
是	186(84.9)	6 985(48.8)		
合并疾病数量				
0~2种	108(49.3)	11 159(78.0)		
3~5种	65(29.7)	2 622(18.3)	200.767	<0.001
6~10种	46(21.0)	530(3.7)		
前次手术时长				
<3 h	51(23.3)	8 297(58.0)	106.180	<0.001
≥3 h	168(76.7)	6 014(42.0)		
前次术中输血				
否	54(24.7)	12 419(86.8)	684.957	<0.001
是	165(75.3)	1 892(13.2)		
前次主刀医师职称				
住院/主治医师	12(5.5)	1 565(10.9)	6.637	0.010
副高/正高医师	207(94.5)	12 746(89.1)		
前次手术切口类别				
0类/Ⅰ类	113(51.6)	9 460(66.1)	20.190	<0.001
Ⅱ类/Ⅲ类	106(48.4)	4 851(33.9)		
前次手术切口愈合等级				
甲级愈合	168(76.7)	11 806(82.5)	4.977	0.026
其他	51(23.3)	2 505(17.5)		
前次手术级别				
一/二级	18(8.2)	3 529(24.7)		
三级	80(36.5)	4 631(32.3)	32.562	<0.001
四级	121(55.3)	6 151(43.0)		

2.2 Logistic 回归模型分析结果 以是否发生非计划再手术为因变量，以性别、罹患恶性肿瘤、合并疾病数量、前次术中输血情况、前次手术时长、前次手术级别、前次主刀医师职称、前次手术切口类别、前次切口愈合等级为自变量（变量赋值方法见表2）。进行多因素 Logistic 回归分析，direction 参数设置为“backword”，通过 summary 函数输出最终模型的结果。结果显示，前次术中输血、罹患恶性肿瘤、合并疾病数量、前次手术切口类别、前次手术切口愈合等级、前次手术时长、前次手术级别是非计划再手术的影响因素（P<0.05），见表3。据模型系数得出模型表达式：Logit$P$=−9.224+2.627×前次术中输血+1.764×前次手术切口愈合等级+1.752四级手术+1.684×罹患恶性肿瘤+1.485×三级手术+ 0.945×前次手术切口类别+ 0.724×前次手术时长+ 0.507×合并疾病数量。

表2 变量赋值方法

变量	赋值方法	变量	赋值方法
非计划再次手术	否=0，是=1	前次手术时长	<3 h=0，≥3 h=1
性别	男性=0，女性=1	前次手术级别	以一级/二级为参照设置哑变量
罹患恶性肿瘤	否=0，是=1	前次主刀医师职称	住院/主治医师=0，副高/正高医师=1
合并疾病数量	0~2种=0，3~5种=1，6~10种=2	前次手术切口类别	0类/Ⅰ类=0，Ⅱ类/Ⅲ类=1
前次术中输血	否=0，是=1	前次切口愈合等级	甲级愈合=0，其他愈合=1

表3 多因素 Logistic 回归分析结果

变量	β值	SE值	Wald χ^2值	P值	OR值（95% CI）
性别	0.002	0.152	0.000	0.988	1.002（0.744，1.351）
前次主刀医师职称	0.424	0.323	1.722	0.189	1.528（0.811，2.879）
前次术中输血	2.627	0.175	225.650	<0.001	13.833（9.819，19.489）
罹患恶性肿瘤	1.684	0.205	67.655	<0.001	5.387（3.606，8.046）
合并疾病数量	0.507	0.102	24.754	<0.001	1.661（1.360，2.028）
前次手术切口类别	0.945	0.166	32.471	<0.001	2.573（1.859，3.561）
前次手术切口愈合等级	1.764	0.207	72.979	<0.001	5.838（3.895，8.751）
前次手术时长	0.724	0.186	15.185	<0.001	2.062（1.433，2.968）
前次手术级别					
三级手术	1.485	0.274	29.387	<0.001	4.414（0.2581，7.552）
四级手术	1.752	0.270	42.205	<0.001	5.765（3.398，9.780）
常量	−9.224	0.375	608.442	<0.001	0.000（—）

2.3 随机森林预测模型分析结果 采用随机森林算法对单因素分析中有统计学意义的指标进行重要性评分。训练样本为10 171例，ntree=500，mtry=3，基于此参数设置对测试集数据进行分类。各变量重要性排序见表4。

表4 随机森林预测模型变量的重要性排序

变量	平均基尼降低值	排序
前次手术切口类别	9.579	1
前次术中输血	8.797	2
前次手术级别	8.456	3
前次手术切口愈合等级	5.083	4
合并疾病数量	4.487	5
罹患恶性肿瘤	3.568	6
前次手术时长	2.5000	7
性别	1.974	8
前次主刀医师职称	0.984	9

2.4 随机森林模型和 Logistic 回归模型预测性能比较 基于测试集数据比较两种模型的预测效能。Logistic 回归预测模型的AUC为0.922（95% CI：0.895，0.950），随机森林预测模型的 AUC 为0.866（95% CI：0.797，0.936）。Logistic 回归预测模型的AUC大于随机森林预测模型的 AUC，但差异无统计学意义（Z=1.469，P=0.144）。选择约登指数最高时的阈值作为最佳临界点，此时 Logistic 回归预测模型的灵敏度、特异度、准确率分别为92.59%、79.11%、79.28%，而随机森林预测模型的灵敏度、特异度、准确率分别为80.00%、93.33%、86.66% ，见图 1。

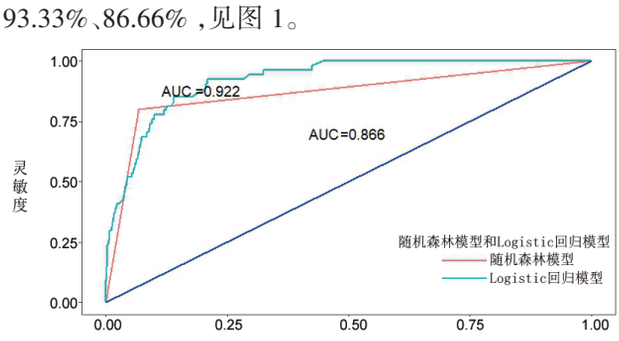

图1 两种模型预测发生非计划再手术的ROC曲线图

3　讨　论

非计划再次手术发生率已作为医疗质量与安全评价的负性评价指标，降低非计划再次手术率可以强化医疗安全，减少医疗资源的浪费；增加医生诊疗的自信心，减少额外的工作量，提高患者就医满意度，减轻患者心理和经济负担。预测非计划再手术风险有助于提高医院围术期的管理水平，减少手术意外伤害，降低并发症发生率。

Logistic回归模型有较多的限制条件：各观察对象之间相互独立，不存在交互作用，Logit P与自变量的关系为线性，样本量须大于自变量个数的10倍。与Logistic回归模型相比，随机森林模型需要的假设条件更少，对多元共线性不敏感[7]，不易出现过度拟合的情况，且具有可视、可读和操作简单的特点，可获知各个特征的重要性[8]。

本研究中，基于测试集进行验证，用于预测非计划再次手术风险时，两种预测模型的AUC均大于0.8，具有较好的预测效果；Logistic回归预测模型和随机森林预测模型的准确率分别为79.28%、86.66%，灵敏度分别为92.59%、80.00%，特异度分别为79.11%、93.33%。进一步分析后发现，Logistic回归预测模型的灵敏度高于随机森林预测模型，但特异度和准确率均低于随机森林预测模型，且两个模型的AUC差异并无统计学意义（P>0.05）。因此，将两种预测模型综合使用，二者分析结果互为补充，能够获得更好的预测效能。

随机森林预测模型变量重要性评分结果显示，前次手术切口类别、前次术中输血、前次手术级别、前次手术切口愈合等级、合并疾病数量、罹患恶性肿瘤等变量的重要性更靠前。有研究显示，手术时长、前次手术切口类别、前次手术切口愈合等级、前次手术级别等因素对于早期识别、预防非计划再手术有积极意义[9-13]。Logistic回归模型结果显示，前次术中输血、罹患恶性肿瘤、合并疾病数量、前次手术切口愈合等级、前次手术级别、前次手术时长、前次手术切口类别是非计划再手术发生的重要因素（P<0.05）。两种模型的分析结果相似，说明预测结果可靠。因此，前次术中输血、前次手术切口类别、罹患恶性肿瘤、前次手术级别、前次手术切口愈合等级、合并疾病数量或可作为研究非计划再手术风险因素的重点关注因素。

综上所述，综合使用Logistic回归模型和随机森林模型，并将二者分析结果互为补充，可从各个方面预测非计划再次手术的风险因素，能获得更好的预测效能，从而能够尽早采取相应的干预措施，降低非计划再次手术发生的风险。本研究的不足之处：收集的指标不够全面；纳入病例的疾病异质性较大，不同疾病或手术的治疗过程各异，混杂因素会对建立预测模型造成影响。因此，下一步将使用更全面的样本集，针对不同疾病或手术做进一步的研究。

参　考　文　献

[1] 高　阳.腰椎退变融合手术非计划再次手术的原因分析[J].中国微创外科杂志，2023，23(1)：45-49.

[2] 游　霞，韩善梅，林王椿.综合医院非计划二次手术情况分析[J].中国病案，2019，20(9)：23-25.

[3] 张梦玲，陈　浩，钱明平，等.利用医政APP进行非计划二次手术监管[J].中国卫生质量管理，2022，29(1)：37-38，56.

[4] 常　青，何紫棠，张国杰，等.基于CiteSpace知识图谱的国内外非计划再手术研究现状及热点可视化分析[J].中国医刊，2021，56(9)：990-995.

[5] 国家卫生健康委员会，国家中医药管理局.全面提升医疗质量行动计划（2023-2025年）[EB/OL].(2023-05-29)[2024-03-26].http://www.nhc.gov.cn/yzygj/s3585/202305/cfe6b26bce624b9f894cef021a363f3e.shtml.

[6] 陈　虹，王吉善.医院评审评价标准指南[M].北京：科学科技文献出版社，2015：301.

[7] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报，2013，50(4)：1190-1197.

[8] 傅华珍，丁小容，陈　丹，等.基于随机森林算法的糖尿病肾病患者血液透析中发生低血压预测模型的建立[J].中国中西医结合肾病杂志，2023，24(6)：493-496.

[9] 徐　芳，赵汝成，方　超，等.上海市某三甲专科医院非计划再次手术原因及影响因素分析[J].中国医院管理，2023，43(4)：62-65.

[10] Schairer WW，Sing DC，Vail TP，et al. Causes and frequency of unplanned hospital readmission after total hip arthroplasty[J]. Clin Orthop Relat Res，2014，472(2)：464-470.

[11] 张家齐，刘　磊，赵　珂，等.胸外科术后30d内非计划二次手术的临床特点[J].中国医学科学院学报，2022，44(5)：809-814.

[12] 李　平，黄昌明，郑朝辉，等.胃癌根治术后非计划再手术对临床疗效的影响及其发生的危险因素分析（附4124例报告）[J].中华消化外科杂志，2018，17(6)：564-570.

[13] 徐雨晨，曹云帆，吴开明，等.重庆市某三甲医院非计划再次手术影响因素研究[J].医学与社会，2020，33(10)：85-88，93.

（收稿日期：2024-01-10　修回日期：2024-03-12）