DOI: 10.12235/E20230422
文章编号：1007-1989（2024）05-0036-12

论 著

基于自动化机器学习建立结肠镜肠道准备失败风险预测模型及评价*

王甘红[1]，陈健[2]，沈支佳[1]，奚美娟[1]，周燕婷[1]
[1.常熟市中医院（新区医院） 消化内科，江苏 常熟 215500；2. 苏州大学附属常熟医院（常熟市第一人民医院） 消化内科，江苏 常熟 215500]

摘要：目的 鉴于机器学习（ML）在医学模型中的广泛应用，以及其出色的学习和泛化特性，该研究采用自动化机器学习（AutoML）结合患者一般资料和临床状况，早期评估结肠镜前肠道准备的失败风险。**方法** 回顾性分析2022年1月—2023年1月在该院接受结肠镜检查的患者的临床资料。波士顿肠道准备评分（BBPS）≤5分被定义为肠道准备失败，＞5分为合格。将患者按8∶2的比例随机划分了训练集（n = 303）和验证集（n = 76）。采用最小绝对收缩和选择算子（LASSO）逻辑回归（LR）模型进行特征选择，构建列线图评分系统，并使用基于5种算法的AutoML建立模型。模型性能通过受试者操作特征曲线（ROC curve）、校准曲线、基于LR（Lasso回归）的决策曲线分析（DCA）、SHAP图和力图进行评估。**结果** 在379例患者中，105例（27.7%）肠道准备失败（BBPS≤5分）。21个研究变量在经LASSO 5折交叉验证后，获得10个变量，并构建了一款列线图评分系统，通过校正曲线表明了LASSO模型的可靠性。使用H2O平台和5种算法［梯度提升机（GBM）、深度学习（DL）、广义线性模型（GLM）、堆叠集成（Stacked Ensemble）和分布式随机森林（DRF）］开发了67个模型。经比较，Stacked Ensemble表现最佳，其曲线下面积（AUC）为0.871，对数损失值（LogLoss）为0.403，均方根误差（RMSE）为0.354，优于其他模型和传统的LR模型。变量重要性贡献图显示，服完泻药至检查间隔时间、便秘、是否完整服完泻药、年龄和家属陪同等因素对肠道准备失败的预测有重要影响。最后，SHAP图和力图揭示了变量在二分类预测结果中的分布特征，以及各变量对预测结果的影响。**结论** 基于Stacked Ensemble算法的AutoML模型，对肠道准备失败风险的早期预测有明显的临床实用性。同时，该研究构建了一款可供临床使用的列线图评分工具。

关键词： 波士顿肠道准备评分（BBPS）；结肠镜；自动化机器学习（AutoML）；预测模型；列线图
中图分类号： R574

Establishing and evaluating a risk prediction model for colonoscopy bowel preparation failure based on automated machine learning*

Wang Ganhong[1], Chen Jian[2], Shen Zhijia[1], Xi Meijuan[1], Zhou Yanting[1]
[1.Department of Gastroenterology, Changshu Hospital of Traditional Chinese Medicine (New District Hospital), Changshu, Jiangsu 215500, China; 2.Department of Gastroenterology, Changshu No.1 People's Hospital, Changshu, Jiangsu 215500, China]

Abstract: Objective Given the extensive application of machine learning (ML) in medical models and its remarkable learning and generalization capabilities, this study employed automated ML (AutoML) combined with

收稿日期：2023-09-11
* 基金项目：常熟市卫生健康委员会科技计划项目（No：CSWS202316）
[通信作者] 周燕婷，E-mail：szcs10132718@aliyun.com；Tel：13862250114

patient demographics and clinical conditions to early assess the risk of failure in bowel preparation prior to colonoscopy. **Methods** A retrospective analysis was conducted on patients who underwent colonoscopy examinations in Hospital 1 and Hospital 2 from January 2022 to January 2023, and their general and clinical information was collected. According to the Boston bowel preparation scale (BBPS), a BBPS of ≤ 5 was defined as a failure in bowel preparation, > 5 was deemed satisfactory. From the data of the two hospitals, we randomly divided the dataset into a training set (n = 303) and a validation set (n = 76) at an 8∶2 ratio. Least absolute shrinkage and selection operator (LASSO) logistic regression (LR) model was used for feature selection, a nomogram scoring system was constructed, and models were established using AutoML based on five algorithms. Model performance was evaluated through receiver operator characteristic curve (ROC curve), calibration curves, LR-based decision curve analysis (DCA), SHAP plots, and force plots. **Results** Among the 379 patients, 105 cases (27.7%) experienced bowel preparation failure (BBPS ≤ 5). 21 study variables were narrowed down to 10 through LASSO with 5-fold cross-validation, resulting in the development of a Nomogram chart with demonstrated reliability via calibration curves. Using the H2O platform and five algorithms [gradient boosting machine (GBM), deep learning (DL), generalized linear model (GLM), Stacked Ensemble and distributed random forest (DRF)], 67 models were developed. Stacked Ensemble outperformed the others with an area under the curve (AUC) of 0.871, LogLoss of 0.403, and RMSE of 0.354, surpassing traditional LR model and other models. Variable importance contribution plots indicated significant predictive influences from factors such as the interval between laxative ingestion and examination, history of constipation, completion of laxative regimen, age, and presence of a companion during the procedure. Finally, SHAP plots and force plots revealed variable distribution patterns in binary classification predictions and the impact of variables on predictive outcomes. **Conclusion** The AutoML model based on the Stacked Ensemble algorithm exhibits clear clinical utility in early prediction of bowel preparation failure risk. Moreover, a clinically applicable column chart scoring tool is constructed.

Keywords: Boston bowel preparation scale (BBPS); colonoscopy; automated machine learning (AutoML); predictive model; nomogram

据文献[1]报道，我国结直肠癌（colorectal carcinoma，CRC）年发病为40.8万人，占我国癌症发病率的第二位，中国CRC发病率呈持续上升趋势，发病率和死亡率呈现“双高”态势，结肠镜检查是筛查CRC的金标准。然而，结肠镜检查的安全性、息肉、早期癌症发现率和操作时间等，都与肠道准备质量密切相关。检查前的肠道准备质量，是决定结肠镜检查质量的核心影响因素。国内外指南[2–3]均强调了肠道准备良好的重要性。有研究[4–6]报道，门诊结肠镜检查前肠道准备不合格率为15%～30%；有文献[7]显示，住院患者肠道准备不合格率高达51%。提高结肠镜检查的肠道准备质量，减少肠道准备失败率，已成为当务之急。因此，在结肠镜检查前，筛查出存在高风险肠道准备失败的患者，显得尤为重要。目前，临床多采用Logistic回归分析建立预测模型，未能充分排除缺失变量及变量间多重共线性关系对研究结果的影响，当数据存在异常值或噪声时，会降低模型的准确性。以往有研究[8–9]已经证实机器学习（machine learning，ML）在建立疾病诊断、预后预测和生存分析模型等方面，有巨大潜力。近年来，一种名为自动化机器学习（automatic ML，AutoML）的先进技术得以发展[10–12]，其能够智能地从众多算法和超参数中选择，并构建适用于特定数据集的模型。AutoML借助智能早期停止、交叉验证、正则化和超参数优化等技术，能够在更短的时间内开发出更为精确的模型。本研究旨在利用H2O AutoML平台，在多个中心训练和验证一系列ML模型中，筛选出性能最佳模型，用于预测结肠镜检查患者肠道准备失败风险，以供临床参考。

1　资料与方法

1.1　一般资料

回顾性分析2022年1月—2023年1月在常熟两家大型综合医院接受结肠镜检查的379例患者的临床资料。其中，常熟市中医院（医院1）279例，常熟市第一人民医院（医院2）100例。按照波士顿肠道准备评分（Boston bowel preparation scale，BBPS）标准，

将结肠分 3 段（盲肠和升结肠，肝曲、横结肠和脾曲，以及降结肠、乙状结肠和直肠）进行评分，每段结肠根据清洁程度评分，从最差（0 分）到最佳（3 分），总分为每段结肠评分的累加，总分范围为 0 到 9 分。按照 BBPS，将患者分为两组：肠道准备合格组（BBPS > 5 分）和肠道准备失败组（BBPS ≤ 5 分）。在训练和验证 AutoML 模型时，按照 8∶2 的比例，将患者随机分为训练集（n = 303）和验证集（n = 76）。对拟行结肠镜检查的患者进行问诊，并收集相关数据，一般资料包括：年龄、性别、文化程度（小学以下、小学至中学和大学及以上）、家属陪同、体重指数（body mass index，BMI）、服完泻药至检查间隔时间（h）等；临床资料包括：患者来源（住院部或门诊）、结肠镜检查史、饮食情况（流质饮食、低纤维低渣饮食或其他）、既往病史（冠心病、肝硬化、炎症性肠病、便秘、高血压和糖尿病）、药物服用情况（抗抑郁药物、阿片类药物和钙通道阻滞剂）、吸烟、完整服完泻药、饮酒、腹部或盆腔手术史等。研究流程图见图 1。

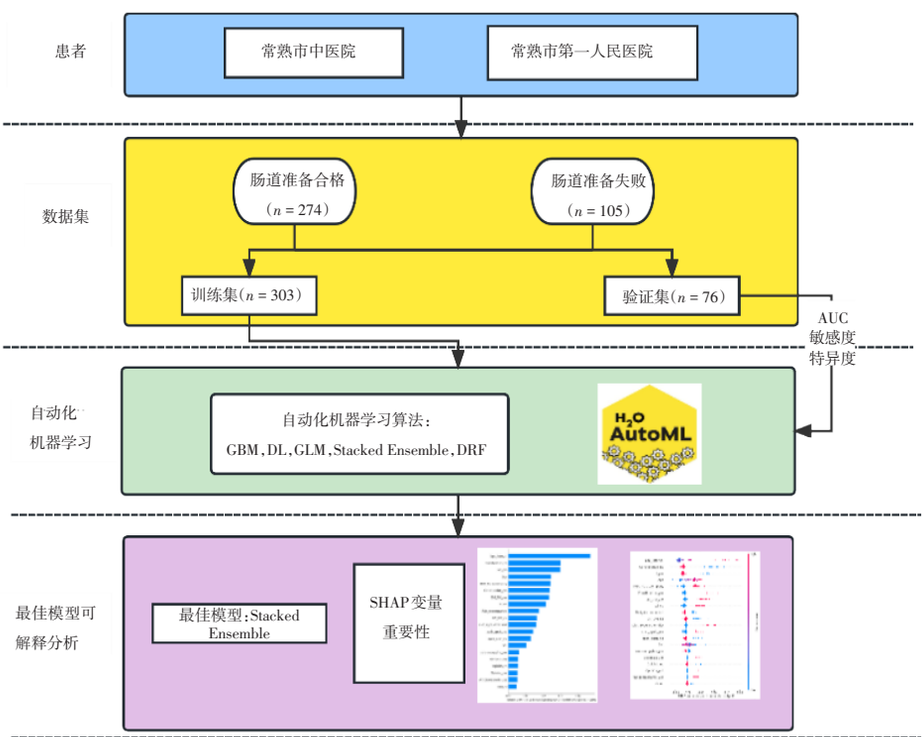


图 1　研究流程图

Fig.1　Flowchart of the study

纳入标准：采用指南[3]推荐的聚乙二醇电解质散分次剂量方案的患者；由经过统一培训的责任护士进行一对一肠道准备宣教的患者；若患者因肠道准备不充分而需进行多次结肠镜检查，仅将首次结肠镜检查纳入研究。排除标准：未采用指南[3]推荐的肠道准备方案；未完成全结肠检查，无法完成 BBPS 的患者。在进行结肠镜检查之前，所有患者都详细了解了方案的优点和潜在风险，并签署书面同意书。本研究符合 STROBE 指南[13]。

1.2　方法

由门诊医生或住院部医生开具结肠镜检查申请单，患者凭申请单至内镜中心预约服务台进行预约登记，现场由经过统一培训的责任护士进行一对一肠道准备宣教，介绍肠道准备注意事项、饮食准备和清洁剂服用方法后，协助有智能手机的患者，关注内镜中心公众号，并告知肠道准备微信宣教视频的打开方法，现场获得患者书面的知情同意书，解释本研究目的，并收集患者的一般资料和临床资料。在本中心使用 3 L 聚乙二醇电解质散分次服用方案，即：肠道检查前 1 天晚上 8 点服用 1 L，检查当天检查前 4 ~ 6 h 服用 2 L，服药期间可适当走动，并轻揉腹部加快排泄。清洁剂服用时间分为两种情况，结肠镜上午检查者：

第 1 次服用时间为前 1 天晚上 6 点，第 2 次服用时间为当天早上 4 点；结肠镜下午检查者：第 1 次服用时间为前 1 天晚上 8 点，第 2 次服用时间为当天早上 9 点。为确保服用最后一剂肠道准备药物后，至操作开始的时间在合适范围内（最佳推荐时间为 2 ~ 5 h），患者可根据实际预约时间稍作调整。由指定的 9 名主治及以上级别的内镜医师完成结肠镜检查，检查完成后，操作医师使用 BBPS 评估肠道准备质量，并将评估结果写入内镜报告单。

1.3　相关定义

1.3.1　肠道准备饮食方案　患者检查前采取指南[3]推荐的低渣饮食、低纤维饮食或流质饮食，当天结肠镜检查前，询问并记录患者实际饮食情况。“其他饮食”指除了“流质饮食”和“低纤维低渣饮食”之外的饮食方案，具体包括：高纤维饮食、高蛋白饮食和高脂饮食等。

1.3.2　便秘　诊断主要取决于症状，有排便困难费力、排便次数减少（每周 < 3 次）、粪便干结或量少，可诊断为便秘。

1.3.3　预约等待时间　指患者从开具结肠镜检查申请单，到实际进行结肠镜检查的等待时间。

1.3.4　服完泻药至检查间隔时间　指从肠道准备结束，到结肠镜检查之间的时间间隔（单位为 h）。

1.4　结肠镜肠道准备失败预测模型构建

1.4.1　特征选择和 Logistic 回归模型的构建　采用最小绝对收缩和选择算子（least absolute shrinkage and selection operator，LASSO）逻辑回归（LASSO regression，LR）来优化特征选择，并解决变量之间可能存在的多重共线性问题。使用“λ.1–s.e.”标准进行参数调整。

1.4.2　模型性能评估和列线图的构建　利用受试者操作特征曲线（receiver operator characteristic curve，ROC curve）、校准曲线和决策曲线分析（decision curve analysis，DCA）来评估模型的预测性能。通过 LASSO 回归模型的特征收缩和选择，识别出独立的风险因素。基于这些独立风险因素，构建列线图评分系统（Nomogram），以更直观地展示风险评估结果。

1.4.3　AutoML　H2O.ai 平台（www.h2o.ai）中的 H2O 软件包已应用于实现 AutoML 分析，其实现了很多先进的 ML 算法，并通过自动化的流程帮助研究者找到最佳模型。使用的算法包括：梯度提升机（gradient boosting machine，GBM）、深度学习（deep learning，DL）、广义线性模型（generalized linear model，GLM）、堆叠集成（Stacked Ensemble）和分布式随机森林（distributed random forest，DRF）。比较多个模型在验证集上的预测性能，并选择表现最优的模型。为了增强模型的可解释性，进一步突破其“黑盒”特性，利用变量的重要性，用排序图和 SHAP 值（SHapley Additive exPlanations）对模型结果做可视化展示。通过这种方式，可以更直观地理解模型中的各个特征，以及预测结果的贡献大小，提高了模型的透明度和可理解性。

1.5　统计学方法

使用 SPSS 27.0 软件进行统计学处理。符合正态分布的计量资料以均数 ± 标准差（$\bar{x} \pm s$）表示，组间比较采用两独立样本 t 检验；计数资料以例（%）表示，组间比较采用 χ^2 检验。$P < 0.05$ 为差异有统计学意义。ML 部分的统计分析使用 R 4.2.3 软件，涉及的 R 包有：版本为 3.10.3.5 的 h2o 包，版本为 0.12.0 的 tableone 包，版本为 1.3.0 的 tidyverse 包，以及版本为 1.0.2 的 tidyquant 包。

2　结果

2.1　两组患者基线特征比较

本研究中，共有 379 例患者入选。其中，105 例（27.7%）患者肠道准备失败（BBPS ≤ 5 分）。两组患者家属陪同、便秘、糖尿病、冠心病、肝硬化、腹盆腔手术史、服完泻药至检查时间间隔、结肠镜检查史、抗抑郁药物、阿片类药物、钙通道阻滞剂、饮食类型、完全服完泻药和年龄等比较，差异均有统计学意义（$P < 0.05$）；高血压、炎症性肠病、性别、BMI、文化程度、吸烟和饮酒比较，差异均无统计学意义（$P > 0.05$）。见表 1。小提琴图清晰地呈现了患者的年龄、服完泻药至检查间隔时间和 BMI 的分布。见图 2。

表 1 两组患者基线资料比较

Table 1 Comparison of baseline data between the two groups

组别	年龄/岁	性别 例(%)		BMI/(kg/m²)	文化程度 例(%)			家属陪同 例(%)	
		女	男		小学以下	小学至中学	大学及以上	无	有
肠道准备合格组(n = 274)	51.10±17.29	132(48.2)	142(51.8)	24.45±3.98	49(17.9)	193(70.4)	32(11.7)	85(31.0)	189(69.0)
肠道准备失败组(n = 105)	57.70±16.00	47(44.8)	58(55.2)	24.49±3.61	19(18.1)	73(69.5)	13(12.4)	65(61.9)	40(38.1)
t/χ^2值	−3.39†	0.23		−0.08†	0.04			29.00	
P值	0.000	0.631		0.929	0.979			0.000	

组别	吸烟 例(%)		饮酒 例(%)		便秘史 例(%)		糖尿病 例(%)	
	无	有	无	有	无	有	无	有
肠道准备合格组(n = 274)	240(87.6)	34(12.4)	237(86.5)	37(13.5)	252(92.0)	22(8.0)	236(86.1)	38(13.9)
肠道准备失败组(n = 105)	92(87.6)	13(12.4)	87(82.9)	18(17.1)	64(61.0)	41(39.0)	66(62.9)	39(37.1)
t/χ^2值	0.01		0.54		50.48		23.98	
P值	0.982		0.461		0.000		0.000	

组别	高血压 例(%)		冠心病 例(%)		肝硬化 例(%)		炎症性肠病 例(%)	
	无	有	无	有	无	有	无	有
肠道准备合格组(n = 274)	183(66.8)	91(33.2)	247(90.1)	27(9.9)	264(96.4)	10(3.6)	259(94.5)	15(5.5)
肠道准备失败组(n = 105)	63(60.0)	42(40.0)	86(81.9)	19(18.1)	90(85.7)	15(14.3)	96(91.4)	9(8.6)
t/χ^2值	1.25		4.09		12.27		0.76	
P值	0.263		0.043		0.000		0.383	

组别	腹盆腔手术史 例(%)		结肠镜检查史 例(%)		抗抑郁药服用史 例(%)		阿片类药物服用史 例(%)	
	无	有	无	有	无	有	无	有
肠道准备合格组(n = 274)	246(89.8)	28(10.2)	186(67.9)	88(32.1)	265(96.7)	9(3.3)	265(96.7)	9(3.3)
肠道准备失败组(n = 105)	71(67.6)	34(32.4)	84(80.0)	21(20.0)	90(85.7)	15(14.3)	89(84.8)	16(15.2)
t/χ^2值	25.65		4.86		13.69		15.72	
P值	0.000		0.027		0.000		0.000	

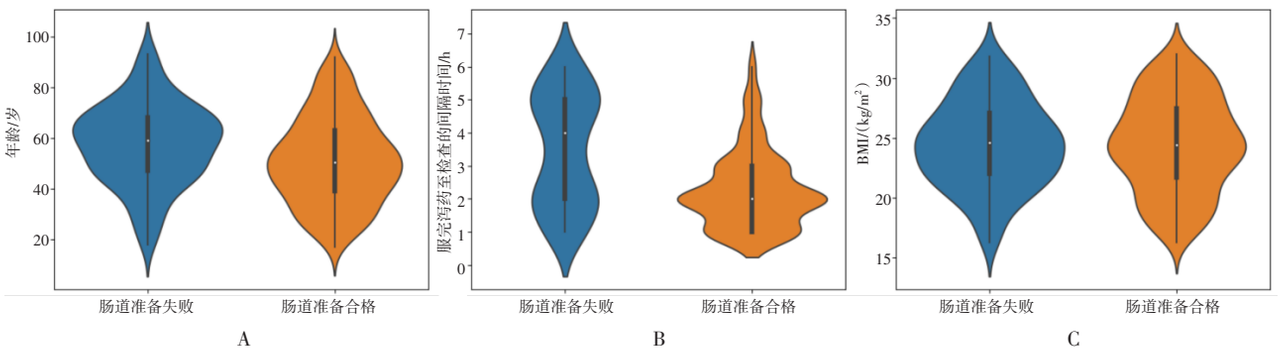
组别	饮食类型 例(%)			钙通道阻滞剂 例(%)		完整服完泻药 例(%)		服完泻药至检查间隔时间/h
	流质饮食	低纤维低渣	其他	无	有	否	是	
肠道准备合格组(n = 274)	66(24.1)	196(71.5)	12(4.4)	248(90.5)	26(9.5)	31(11.3)	243(88.7)	2.29±1.17
肠道准备失败组(n = 105)	18(17.1)	65(61.9)	22(21.0)	83(79.0)	22(21.0)	45(42.9)	60(57.1)	3.49±1.70
t/χ^2值	25.91			8.01		53.14		−7.76†
P值	0.000			0.005		0.000		0.000

注：†为t值。

2.2 单变量和多变量Logistic回归分析

2.2.1 LASSO回归模型 采用LASSO回归模型处理21个预测变量间的多重共线性问题，并以“λ_1se（0.062）”为标准，通过5折交叉验证来训练模型。见图3。经过筛选，有10个变量（年龄、家属陪同、便秘史、糖尿病史、肝硬化史、腹盆腔手术史、抗抑郁药服用史、饮食类型、完整服完泻药和服完泻药至检查间隔时间）被选入最终的LASSO回归模型。

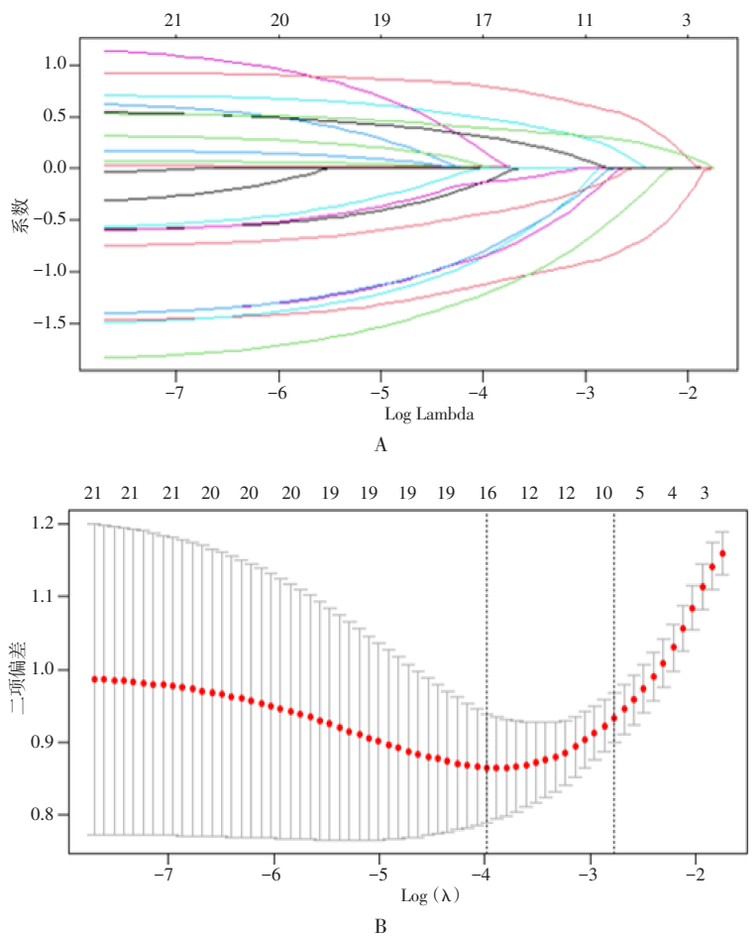
2.2.2 列线图评分系统 基于以上选定的变量，开发了一种列线图评分系统，以便在临床实践中预测肠道准备失败的风险。见图4。



A：年龄；B：服完泻药至检查间隔时间；C：BMI。

图2　变量分布的小提琴图

Fig.2　Violin plot of variable distribution



A：回归系数，随着λ值的增加，系数的绝对值减小；B：通过5折交叉验证，确定LR中的最优λ值。

图3　基于LASSO回归模型的肠道准备失败风险预测因子的惩罚图

Fig.3　Penalty plot of predictive factors for bowel preparation failure risk based on LASSO regression model analysis

2.2.3　LASSO 回归模型的校准曲线　LASSO回归模型在训练集和验证集中，校准曲线的平均绝对误差分别为0.026和0.036。见图5。

2.2.4　LASSO 回归模型的 ROC curve　在训练集中，LASSO回归模型的曲线下面积（area under the curve，AUC）为0.881，敏感度为0.85，特异度为0.81；在验证集中，LASSO回归模型的AUC为0.734，敏感度为0.80，特异度为0.72。见图6。

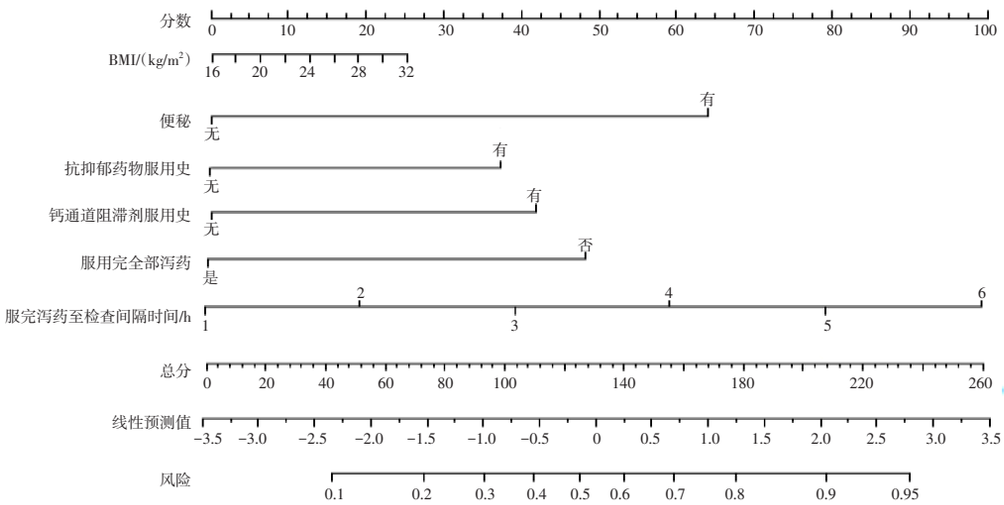
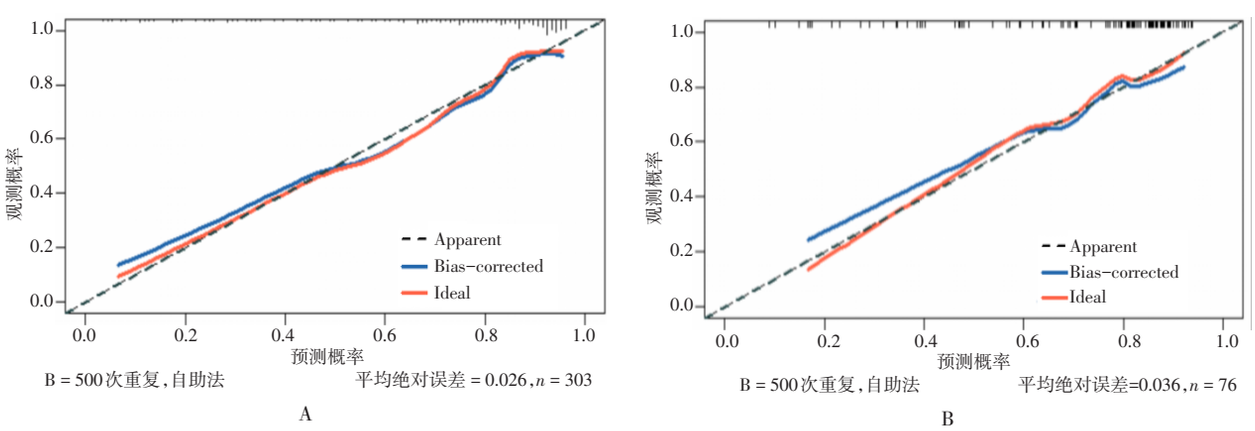


图 4　LASSO 回归模型预测肠道准备失败风险的列线图

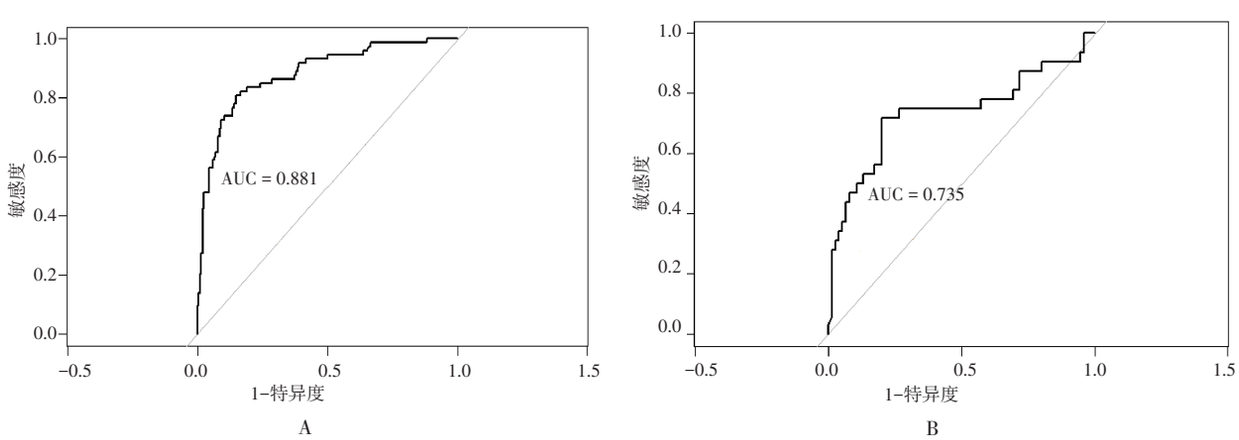
Fig.4　Nomogram of LASSO regression model for predicting the risk of bowel preparation failure



A：训练集；B：验证集。

图 5　LASSO 回归模型在训练集和验证集中的校准曲线

Fig.5　Calibration curves of the LASSO regression model in the training and validation sets



A：训练集；B：验证集。

图 6　LASSO 回归模型在训练集和验证集中的 ROC curve

Fig.6　ROC curve of the LASSO regression model in the training and validation sets

2.3　AutoML

所有临床数据经预处理后，载入 H2O 平台 AutoML 框架中，按 8∶2 的比例随机分为训练集和验证集。经过自动化变量选择和建模调参，总共开发了 67 个基于 5 种 ML 算法（GBM、DL、GLM、Stacked Ensemble 和 DRF）的模型。Stacked Ensemble 模型的 AUC 最高，表现最为出色。随着训练样本的增加，其学习曲线逐渐稳定，并在交叉验证和验证曲线之间显示出良好的拟合，没有出现过拟合现象。见图 7。

在经过 AutoML 筛选后，4 种性能最优的模型分别为 Stacked Ensemble_BestOfFamily、DeepLearning、GBM 和 Stacked Ensemble_AllModels。Stacked Ensemble_BestOfFamily 模型在整体性能上表现最为优秀，AUC 为 0.871，对数损失值（LogLoss）为 0.403，均方根误差（root mean square error，RMSE）为 0.354，超越了其他模型。DeepLearning 模型的表现也十分出色，尽管在 LogLoss、RMSE 和均方误差（mean square error，MSE）上稍逊于 Stacked Ensemble_BestOfFamily，但其 AUC 达到了 0.868，精确率–召回率 AUC 为 0.713。见表 2。

2.4　基于最佳模型的可解释性分析

2.4.1　变量重要性贡献　经过 AutoML 筛选后，获取性能最佳模型 Stacked Ensemble_BestOfFamily。服完泻药至检查间隔时间是最重要的特征，其次是便秘史、是否完整服完泻药、年龄和家属陪同。此外，Stacked Ensemble 模型和 LASSO 模型共同的重要变量是服完泻药至检查间隔时间、是否完整服完泻药、是否家属陪同和便秘病史。Stacked Ensemble_BestOfFamily 在验证集中的变量重要性贡献见图 8。

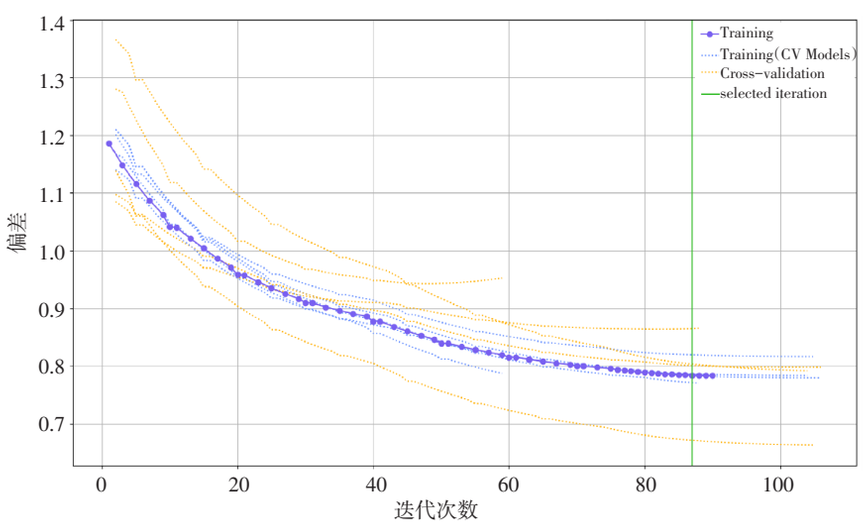


图 7　最佳 ML 模型 Stacked Ensemble 的学习曲线

Fig.7　Learning curve of the optimal ML model Stacked Ensemble

表 2　ML 模型在验证集上的预测性能比较

Table 2　Comparison of predictive performance of ML models on the validation set

模型 ID	AUC	准确率	LogLoss	精确率–召回率 AUC	平均类别错误率	RMSE	MSE
Stacked Ensemble_BestOfFamily	0.871	0.832	0.403	0.704	0.187	0.354	0.125
Deep Learning	0.868	0.823	0.489	0.713	0.207	0.365	0.133
GBM	0.865	0.808	0.414	0.721	0.213	0.365	0.133
Stacked Ensemble_AllModels	0.860	0.816	0.406	0.728	0.186	0.353	0.125

2.4.2　最佳模型的 SHAP 图　在验证集中，性能表现最佳模型的 SHAP 图中，每一行表示一个特征，红色表示该特征的值较高的数据点，蓝色值表示该特征的值较低的数据点，越靠右的点表示这个特征对预测

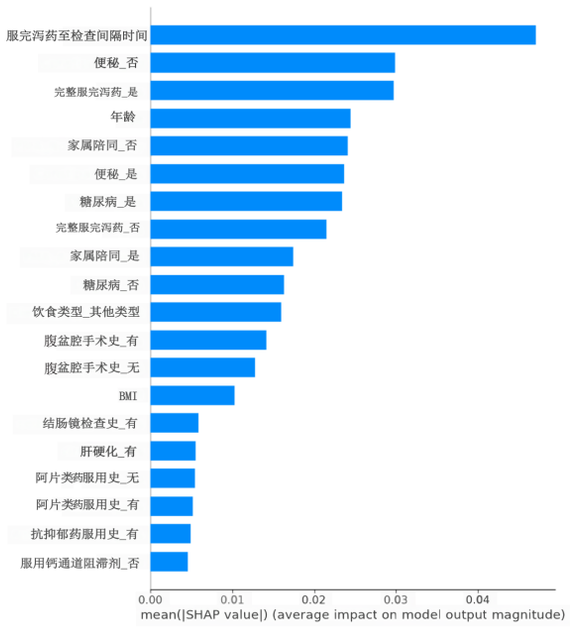


图 8　变量重要性贡献图

Fig.8　Variable importance contribution plot

结局（肠道准备失败）的正向影响越高。变量重要性排名前 7 的变量依次为：服完泻药至检查间隔时间、便秘史、是否完整服完泻药、年龄、家属陪同、糖尿病史和饮食类型。见图 9。变量值越接近 1，患者发生肠道准备失败的可能性就越大。如：图中便秘史的红色部分集中在轴 = 0 的右侧，表示有便秘史的患者，发生肠道准备失败的风险更高，进行肠道准备宣教时，需要获得更多的关注。

2.4.3　每个特征影响模型对单个数据实例的预测力图　对于第 12 号患者（图 10A），其实际诊断为阳性。模型为其估计的阳性概率为 70.0%，与实际诊断相符，进一步证实了模型的准确性。模型预测 87 号患者（图 10B）肠道准备失败的概率仅为 17.0%。鉴于此预测值较低，与其真实标签相一致，再次验证了模型的预测准确性。

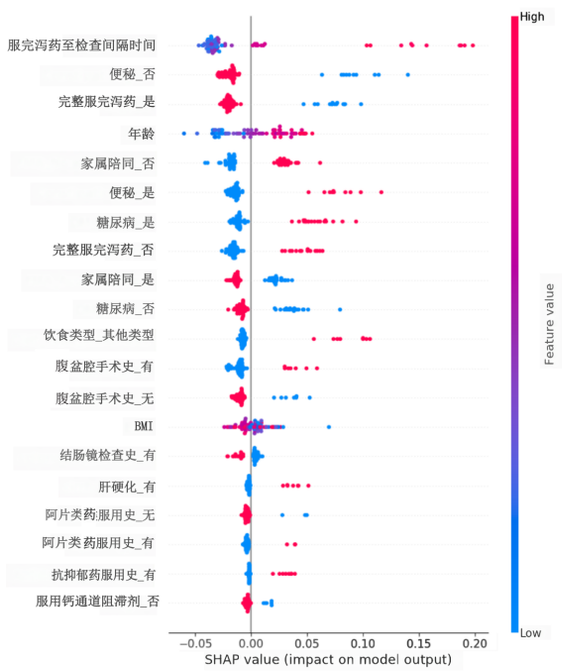
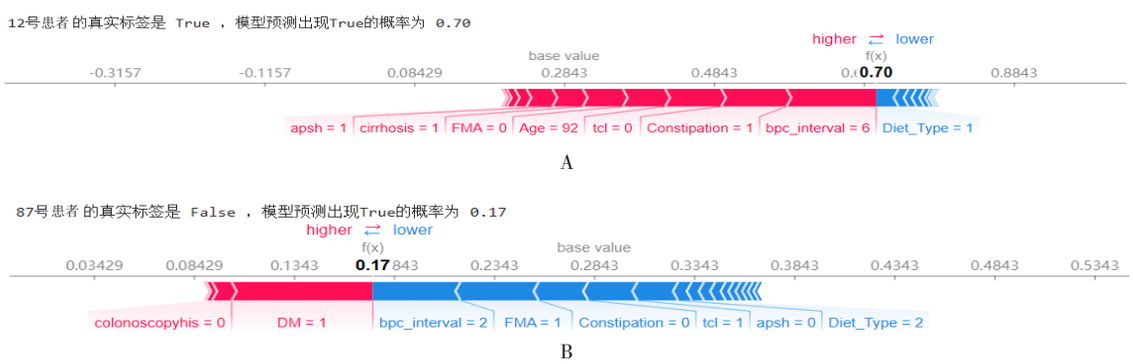


图 9　验证集中 Stacked Ensemble 模型的 SHAP 图

Fig.9　SHAP plot of the Stacked Ensemble model in the validation set



A：12 号患者；B：87 号患者；红色表示对预测为阳性结局产生正向影响的变量，蓝色表示对预测为阳性结局产生负向影响的变量。

图 10　验证集中 Stacked Ensemble 模型的力图

Fig.10　Force plot of the Stacked Ensemble model in the validation set

3　讨论

3.1　肠道准备质量对结肠镜检查的影响

合格的肠道准备是提高结肠镜检查有效性和安全性的关键。确保肠道准备质量，不仅能够提高肠道的清洁度，还能提高结肠镜检查在诊断和治疗过程中的精确性，使患者在接受结肠镜检查时，能够获得最佳的医疗效果和体验。有研究[14-16]显示，不充分的肠道准备与腺瘤和高级腺瘤的检出率降低有关，肠道准备失败的患者，可明显降低腺瘤检出率[17]。一项前瞻性研究[18]发现，当肠道准备不充分时，≥ 5 mm 的腺瘤漏诊率增加了 3 倍。不充分的肠道准备，是盲肠插管失败的最不利预测因素之一[19]，且会导致患者的不满意体验[20]增加，结肠镜检查的监测间隔缩短[21]，住院时间延长，以及医疗成本的增加[22]。一项前瞻性多中心随机临床试验[23]中，对于 BBPS 在任何肠段得分 < 2 分的患者，在短期内进行了结肠镜的复查，腺瘤检出率为 45.3%，高级别腺瘤检出率为 10.9%，锯齿状息肉检出率为 14.3%。在普通人群中行结肠镜的筛查，可以发现 CRC，具有较高的成本效益比，但是，不合格的肠道准备会降低成本效益比[24]。由此可见，多种可改变和不可改变的因素会影响肠道准备的质量[25-26]。因此，临床应将高质量的肠道准备视为结肠镜检查成功的重要组成部分。本研究旨在利用多种 ML 算法，预测结肠镜检查肠道准备失败风险，以期为临床医护人员在肠镜检查前宣教过程中提供参考。

3.2　预测发生肠道准备失败风险的列线图

本研究共纳入 379 例结肠镜检查患者，选用 21 个预测变量，以“λ_1se（0.062）”为标准，通过 5 折交叉验证筛选变量，建立了 LASSO 回归模型，该模型在训练集和验证集中，校准曲线的平均绝对误差分别为 0.026 和 0.036，模型预测的风险和实际观察到的风险非常接近，从而证明了模型的可靠性较高。使用该模型建立了一个列线图评分系统，在临床工作中，医护人员可以根据患者的具体情况，参照此列线图，将每个因素的得分相加后，得到一个总分，对照列线图中最下方的“风险”，即可获得该患者发生肠道准备失败风险的预测概率，利于辅助临床进行更高质量的肠道准备宣教工作，提高结肠镜检查效果。

3.3　预测发生肠道准备失败风险的 AutoML

本研究使用基于 H2O 平台的 AutoML 技术，开发了 67 个 ML 模型。其中，Stacked Ensemble_BestOfFamily 模型展现出最佳的综合预测性能，其在验证队列中 AUC 达到了 0.871，表明：该模型在区分肠道准备失败和成功的患者上具有较高的准确性。同时，其 LogLoss 值为 0.403，表明：预测错误的概率较低。且该模型的平均类别错误率为 0.187，RMSE 为 0.354，MSE 为 0.125，这些数据都进一步证明了该模型的高度精确性和可靠性。笔者通过 AutoML 构建的肠道准备失败风险预测模型，具有较高的预测性能（AUC = 0.871）和预测准确率（0.832）。与徐苗苗等[27]和郭盛丽等[28]建立的 Logistic 回归模型（AUC 分别为 0.720 和 0.824）相比，本研究建立的模型在预测肠道准备失败风险上，具有更高的可靠性。Stacked Ensemble 是一种集成学习方法，可以利用不同模型的优势，弥补各个模型的弱点，从而提高整体预测性能。本研究结果表明，集成学习方法在肠道准备失败风险预测领域，临床实用性高。

3.4　对 ML 的可视化处理

近年来，ML 飞速发展，具有对数据的挖掘与处理的强大能力，在临床实践相关预测模型中，发挥了重要作用。为克服 ML 中不可解释的“黑盒”效应，笔者对预测性能最佳的模型，进行了可视化处理，以便更好地理解模型的工作原理。笔者先绘制了变量重要性贡献图，该图展示了各变量对模型预测能力的贡献程度，预测肠道准备失败风险，贡献度从高到低排名，依次为：服完泻药至检查间隔时间、便秘史、完整服完泻药、年龄、家属陪同、糖尿病史和饮食类型。然后，利用 SHAP 图揭示了变量在二分类结局（肠道准备失败与否）中的分布特征，从而反映了各变量对预测结果的影响，如：从图中可以清晰地看到，便秘史红点和蓝点分布于中线两侧，表明：当便秘史这个变量为不同取值时（有便秘和无便秘），会对预测结果产生影响。从 SHAP 图中，可以看到 Stacked Ensemble 模型预测的最关键特征是服完泻药至检查间隔时间。从肠道准备结束到结肠镜检查之间的时间间隔过长，可预测肠道清洁不足。该结果与本研究中的 LASSO 回归模型的结果一致。欧洲胃肠道内镜学会肠道准备指南[2]推荐：在结肠镜检查前 5 h 内开始服用最后一份泻药，并且至少在结肠镜检查开始前 2 h 完成，也表明：肠道准备结束至结肠镜检查之间的时间间隔不宜过长。这种对模型可视化解释的方法，为临床提供了深入了解模型机制的途径，有助

于提高模型在医学研究中的透明度和可信度。

3.5 本研究的局限性

1）由于结肠镜预约等待时间普遍较长，未能将当天肠道准备方案与分次给药方案进行对比研究，有待在优化结肠镜预约流程后，纳入更多采用单次肠道准备方案的患者；2）肠道准备不足可能对结肠镜检查，尤其是CRC的发现，造成负面影响，有待在下一步的研究中，确定一个肠道准备的阈值，对低于该阈值需要复查的病例进行进一步研究，以确认阈值的准确性和有效性。

综上所述，基于AutoML技术的ML模型，特别是表现优异的Stacked Ensemble_BestOfFamily模型，是预测肠道准备失败风险的一种高度可靠的模型。通过模型可视化，揭示了关键变量对预测结果的影响，进一步增强了模型的透明度和可信度，为提高结肠镜检查质量，提供了有力的支持。

参 考 文 献 ：

[1] ZHENG R, ZHANG S, ZENG H, et al. Cancer incidence and mortality in China, 2016[J]. Journal of the National Cancer Center, 2022, 2(1): 1-9.

[2] HASSAN C, EAST J, RADAELLI F, et al. Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) guideline-update 2019[J]. Endoscopy, 2019, 51(8): 775-794.

[3] 中国医师协会内镜医师分会消化内镜专业委员会, 中国抗癌协会肿瘤内镜学专业委员会. 中国消化内镜诊疗相关肠道准备指南 (2019, 上海)[J]. 中华内科杂志, 2019, 58(7): 485-495.

[3] Digestive Endoscopy Special Committee of Endoscopic Physicians Branch of Chinese Medical Association, Cancer Endoscopy Committee of China Anti-Cancer Association. Chinese guideline for bowel preparation for colonoscopy (2019, Shanghai) [J]. Chinese Journal of Internal Medicine, 2019, 58(7): 485-495. Chinese

[4] MAHMOOD S, FAROOQUI S M, MADHOUN M F. Predictors of inadequate bowel preparation for colonoscopy: a systematic review and Meta-analysis[J]. Eur J Gastroenterol Hepatol, 2018, 30(8): 819-826.

[5] KUNNACKAL JOHN G, THULUVATH A J, CARRIER H, et al. Poor health literacy and medication burden are significant predictors for inadequate bowel preparation in an urban tertiary care setting[J]. J Clin Gastroenterol, 2019, 53(9): e382-e386.

[6] KUMAR A, SHENOY V, BUCKLEY M C, et al. Endoscopic disease activity and biologic therapy are independent predictors of suboptimal bowel preparation in patients with inflammatory bowel disease undergoing colonoscopy[J]. Dig Dis Sci, 2022, 67(10): 4851-4865.

[7] GARBER A, SARVEPALLI S, BURKE C A, et al. Modifiable factors associated with quality of bowel preparation among hospitalized patients undergoing colonoscopy[J]. J Hosp Med, 2019, 14(5): 278-283.

[8] HAN T Y, CHENG T, LIAO Y, et al. Development and validation of a novel prognostic score based on thrombotic and inflammatory biomarkers for predicting 28-day adverse outcomes in patients with acute pancreatitis[J]. J Inflamm Res, 2022, 15: 395-408.

[9] LIU J J, HU L, ZHOU B, et al. Development and validation of a novel model incorporating MRI-based radiomics signature with clinical biomarkers for distinguishing pancreatic carcinoma from mass-forming chronic pancreatitis[J]. Transl Oncol, 2022, 18: 101357.

[10] SIRIBORVORNRATANAKUL T. Human behavior in image-based Road Health Inspection Systems despite the emerging AutoML[J]. J Big Data, 2022, 9(1): 96.

[11] WEVER M, TORNEDE A, MOHR F, et al. AutoML for multi-label classification: overview and empirical evaluation[J]. IEEE Trans Pattern Anal Mach Intell, 2021, 43(9): 3037-3054.

[12] ALSHAREF A, AGGARWAL K, SONIA, et al. Review of ML and AutoML solutions to forecast time-series data[J]. Arch Comput Methods Eng, 2022, 29(7): 5297-5311.

[13] VON ELM E, ALTMAN D G, EGGER M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies[J]. Lancet, 2007, 370(9596): 1453-1457.

[14] TARIQ H, KAMAL M U, SAPKOTA B, et al. Evaluation of the combined effect of factors influencing bowel preparation and adenoma detection rates in patients undergoing colonoscopy[J]. BMJ Open Gastroenterol, 2019, 6(1): e000254.

[15] CALDERWOOD A H, THOMPSON K D, SCHROY P C, et al. Good is better than excellent: bowel preparation quality and adenoma detection rates[J]. Gastrointest Endosc, 2015, 81(3): 691-699.

[16] AFIFY S, TAG-ADEEN M, ABU-ELFATTH A, et al. Quality indicators for colonoscopy in Egypt: a prospective multicenter study[J]. Arab J Gastroenterol, 2022, 23(4): 253-258.

[17] GUO R, WANG Y J, LIU M, et al. The effect of quality of segmental bowel preparation on adenoma detection rate[J]. BMC Gastroenterol, 2019, 19(1): 119.

[18] CLARK B T, PROTIVA P, NAGAR A, et al. Quantification of adequate bowel preparation for screening or surveillance colonoscopy in men[J]. Gastroenterology, 2016, 150(2): 396-405.

[19] HSU C M, LIN W P, SU M Y, et al. Factors that influence cecal intubation rate during colonoscopy in deeply sedated patients[J]. J Gastroenterol Hepatol, 2012, 27(1): 76-80.

[20] BUGAJSKI M, WIESZCZY P, HOFF G, et al. Modifiable factors associated with patient-reported pain during and after screening colonoscopy[J]. Gut, 2018, 67(11): 1958-1964.

[21] ANDERSON J C, BARON J A, AHNEN D J, et al. Factors associated with shorter colonoscopy surveillance intervals for patients with low-risk colorectal adenomas and effects on outcome[J]. Gastroenterology, 2017, 152(8): 1933-1943.

[22] YADLAPATI R, JOHNSTON E R, GREGORY D L, et al. Predictors of inadequate inpatient colonoscopy preparation and its association with hospital length of stay and costs[J]. Dig Dis Sci, 2015, 60(11): 3482-3490.

[23] PANTALEÓN SÁNCHEZ M, GIMENO GARCIA A Z, BERNAD CABREDO B, et al. Prevalence of missed lesions in patients with inadequate bowel preparation through a very early repeat colonoscopy[J]. Dig Endosc, 2022, 34(6): 1176-1184.

[24] FRAZIER A L, COLDITZ G A, FUCHS C S, et al. Cost-effectiveness of screening for colorectal cancer in the general population[J]. JAMA, 2000, 284(15): 1954-1961.

[25] LEE J, KIM T O, SEO J W, et al. Shorter waiting times from education to colonoscopy can improve the quality of bowel preparation: a randomized controlled trial[J]. Turk J Gastroenterol, 2018, 29(1): 75-81.

[26] CHAN W K, SARAVANAN A, MANIKAM J, et al. Appointment waiting times and education level influence the quality of bowel preparation in adult patients undergoing colonoscopy[J]. BMC Gastroenterol, 2011, 11: 86.

[27] 徐苗苗, 付秀荣, 张娜, 等. 老年结肠镜检查患者肠道准备失败风险评分模型的构建及验证[J]. 中华护理杂志, 2022, 57(11): 1337-1344.

[27] XU M M, FU X R, ZHANG N, et al. Development and validation of a risk score model for inadequate bowel preparation for colonoscopy in elderly patients[J]. Chinese Journal of Nursing, 2022, 57(11): 1337-1344. Chinese

[28] 郭盛丽, 朱婷, 林威娜, 等. 老年人结肠镜检查肠道准备失败风险预测模型的建立与验证[J]. 护理研究, 2023, 37(3): 392-398.

[28] GUO S L, ZHU T, LIN W N, et al. Establishment and validation of risk prediction model for inadequate bowel preparation before colonoscopy in the elderly[J]. Chinese Nursing Research, 2023, 37(3): 392-398. Chinese

（曾文军　编辑）

本文引用格式：

王甘红, 陈健, 沈支佳, 等. 基于自动化机器学习建立结肠镜肠道准备失败风险预测模型及评价[J]. 中国内镜杂志, 2024, 30(5): 36-47.

WANG G H, CHEN J, SHEN Z J, et al. Establishing and evaluating a risk prediction model for colonoscopy bowel preparation failure based on automated machine learning[J]. China Journal of Endoscopy, 2024, 30(5): 36-47. Chinese