·临床研究· doi:10.3969/j.issn.1671-8348.2024.14.010
网络首发 https://link.cnki.net/urlid/50.1097.R.20240508.1632.024(2024-05-09)

列线图与机器学习算法预测中老年龋齿的比较研究*

赖丽冲[1],韦发烨[2],黄冬妹[3],曹晓莹[1],彭 捷[1],冯小玲[1],黄惠桥[4△]
(1.广西医科大学第二附属医院护理部,南宁 530007;2.广西医科大学第一附属医院泌尿外科,南宁 530021;3.广西医科大学第二附属医院康复医学科,南宁 530007;4.广西医科大学第二附属医院党委办公室,南宁 530007)

[摘要] **目的** 对比列线图与不同机器学习算法构建中老年人龋齿预测模型的效能。**方法** 采用多阶段分层随机抽样法,选取南宁市、贵港市、崇左市 510 名中老年人为研究对象,进行问卷调查及口腔检查。采用单因素分析和 Lasso 回归筛选相关变量,使用多因素 logistic 回归分析确定最终独立影响因素。基于显著特征,建立列线图预测模型,并运用线性判别分析(LDA)、偏最小二乘算法(PLS)、距离多普勒算法(RDA)、广义线性模型(GLM)、随机森林(RF)、支持向量机(SVM)核函数(SVM-Radial)及 SVM 线性核函数(SVM-Linear)7 种机器学习算法构建 7 种龋齿风险预测模型。采用受试者工作特征(ROC)曲线下面积(AUC)中位数评价各模型预测性能,以及不同变量筛选方法所构建模型的预测性能。**结果** 中老年人龋齿检出率为 71.18%。经过特征筛选后最终保留 5 个预测因子,分别是年龄($OR=0.945$,95%CI:0.917~0.973)、刷牙频率($OR=0.688$,95%CI:0.475~0.997)、过去 1 年是否洗牙($OR=0.303$,95%CI:0.103~0.890)、牙存留数($OR=1.062$,95%CI:1.038~1.087)和口腔健康评估量表(OHAT)得分($OR=1.363$,95%CI:1.234~1.505)。各模型对比结果显示,RF 算法所构建的预测模型表现最佳,AUC 中位数为 0.747,其次为列线图,AUC 中位数为 0.733。单因素+Lasso+多因素 logistic(简称 Lasso+logistic)筛选自变量构建预测模型的 AUC 中位数均高于 RF 算法筛选自变量构建的预测模型。**结论** 基于 Lasso+logistic 筛选变量,RF 较列线图及其他机器学习算法在中老年龋齿预测中提供了更可靠的预测性能。

[关键词] 中老年人;龋齿;预测;机器学习;列线图
[中图法分类号] R781.1 [文献标识码] A [文章编号] 1671-8348(2024)14-2130-08

Comparative study on nomogram and machine learning algorithms for predicting dental caries in middle-aged and elderly people*

LAI Lichong[1],*WEI Faye*[2],*HUANG Dongmei*[3],*CAO Xiaoying*[1],*PENG Jie*[1],*FENG Xiaoling*[1],*HUANG Huiqiao*[4△]
(1.*Department of Nursing,the Second Affiliated Hospital of Guangxi Medical University,Nanning,Guangxi* 530007,*China*;2.*Department of Urology Surgery,the First Affiliated Hospital of Guangxi Medical University,Nanning,Guangxi* 530021,*China*;3.*Department of Rehabilitation Medicine,the Second Affiliated Hospital of Guangxi Medical University,Nanning,Guangxi* 530007,*China*;4.*Party Committee Office,the Second Affiliated Hospital of Guangxi Medical University,Nanning,Guangxi* 530007,*China*)

[**Abstract**] **Objective** To compare the efficiency of nomogram and different machine learning algorithms for constructing the dental caries predictive models for middle-aged and elderly people. **Methods** The multi-stage stratified random sampling method was used to select 510 middle-aged and elderly people from Nanning City,Guigang City and Chongzuo City as the research subjects for conducting the questionnaire survey and oral cavity examination. The univariate analysis and Lasso regression were used to screen the related variables,and the multivariate logistic regression analysis was used to determine the final independent influencing factors. Based on the salient features,the nomogram predictive model was established,and the seven machine learning algorithms,including linear discriminant analysis (LDA),partial least squares (PLS),range Doppler algorithm (RDA),generalized linear models (GLM),random forest (RF),support vector machine

* **基金项目**:广西壮族自治区卫生健康委员会合作项目(2022039);广西壮族自治区卫生健康委员会自筹经费科研课题(Z-A20230629)。
△ **通信作者**,E-mail:hhq@sr.gxmu.edu.cn。

(SVM) kernel function (SVM-Radial), and SVM linear kernel function (SVM-Linear), were used to construct the seven kinds of dental caries risk predictive models. The area under the receiver operating characteristic (ROC) curve (AUC) was adopted to evaluate the predictive performance of various models and the predictive performance of models constructed using different variable screening methods. **Results** The detection rate of dental caries in middle-aged and elderly people was 71.18%. After feature screening, the five predictive factors were ultimately retained, which were the age (OR=0.945, 95%CI: 0.917－0.973), brushing frequency (OR=0.688, 95%CI: 0.475－0.997), whether having teeth cleaning in the past one year (OR=0.303, 95%CI: 0.103－0.890), number of remaining teeth (OR=1.062, 95%CI: 1.038－1.087) and oral health assessment tool (OHAT) score (OR=1.363, 95%CI: 1.234－1.505). The results of comparison of various models showed that the predictive model constructed by the RF algorithm performed the best, the median of AUC was 0.747, followed by the nomogram, and the median of AUC was 0.733. The median of AUCs in the prediction model constructed by single factor＋Lasso＋multivariate logistic (Lasso＋logistic) screening independent variables were higher than those constructed by RF algorithm screening independent variables. **Conclusion** Based on Lasso＋logistic screening variables, RF provide more reliable predictive efficiency in predicting dental caries in middle-aged and elderly people than nomogram and the other machine learning algorithms.

[**Key words**] middle-aged and elderly people; dental caries; prediction; machine learning; column diagram;

中老年人龋齿患病在全球范围内普遍存在，不同国家、年龄组和社会经济地位人群之间差异较大[1]，在我国第 4 次口腔健康流行病学调查中 55～74 岁中老年人患龋率高达 96.8%[2]。有研究表明，老年人龋齿与认知障碍[3]、抑郁焦虑[4]、营养不良[5]等疾病密切相关，严重危害老年人的身体健康和生存质量，有效预防并早期识别龋齿，以及进行口腔卫生干预对健康老龄化的发展意义重大[6]。

由于老年人龋齿影响因素多、结构复杂，有关于老年龋齿预防与筛查的研究大多使用传统的统计学方法，拟合程度、准确度有待系统证实。列线图是一种建立在多因素回归分析基础上，用以表达预测模型中各个变量之间相互关系的可视化图形，它能将复杂的回归方程可视化、具象化，提高预测模型结果的可读性[7]。近年来，机器学习因具有快速、高精度、高效和安全等特点，在人工智能辅助诊断和预测模型中的应用被广泛关注[8]。有研究探索了部分机器学习老年人龋齿预测模型的性能[9]，但仍缺乏列线图与多种机器学习的应用比较，优势及效能未能明确，并且不同变量筛选方法所构建的模型比较研究也较少。本研究建立可视化的中老年人龋齿列线图预测模型，同时应用较为成熟的多种机器学习模型筛选变量并建立重复性及泛性化较好的预测模型，进而探讨出最佳的龋齿预测模型。

1 资料与方法

1.1 一般资料

采用多阶段分层随机抽样法。第 1、2 阶段行按规模大小成比例概率抽样(probability proportionate to size, PPS)。在南宁市、贵港市、崇左市随机抽取 3 个城区(西乡塘区、港北区、江州区)，3 个县或县级市(横州市、桂平市、扶绥县)，从每个区(县/县级市)抽取 1 个街道或乡镇，再从选取的街道或乡镇抽取 1 个居委会。第 3 阶段采用简单随机抽样，从每个居委会抽取 55 岁及以上常住中老年人，共计 510 名，于 2022 年 10－11 月进行中老年人口腔现状调查。纳入标准：(1)年龄≥55 岁；(2)意识清醒，具备一定的沟通表达能力；(3)在广西居住≥6 个月；(4)知情同意且自愿参加。排除标准：(1)有严重精神、认知障碍疾病者；(2)未完成所有调查，中途要求退出研究者。本研究经广西医科大学第二附属医院研究伦理委员会批准(审批号：2022-KY0776)。

1.2 方法

1.2.1 问卷调查与口腔检查

采用研究团队自制问卷进行调查，内容包括一般资料、口腔卫生行为及口腔卫生服务利用现状；采用汉化的口腔健康评估量表(oral health assessment tool, OHAT)[10]进行整体口腔评分；按照国家卫生标准《口腔健康调查：检查方法》[11]进行口腔检查，评估有无患龋齿。

1.2.2 列线图预测模型的构建及验证

基于显著特征建立列线图预测模型，运用 Rstudio4.1.1 软件的 rms 包，使用受试者工作特征(receiver operating characteristic, ROC)曲线下面积(area under the curve, AUC)评价模型的区分能力，AUC 为 0.60～0.75，表示模型具有一定的区分能力，＞0.75 表示模型区分能力较好；使用 Bootstrap 抽样(n=1 000)对模型进行验证，利用一致性指数(concordance index, C-index)评价模型预测的准确性，0.71～0.90 表示模型准确度中等，＞0.90 表示模型准确度高；利用 Hosmer-Lemeshow 检验评估模型的校准度，P＞0.05 表示模型有较好的校准能力[12]。

1.2.3 机器学习算法预测模型的构建及验证

运用 Rstudio4.1.1 软件的 caret 包，选取常用的 7 种机器学习算法：线性判别分析（linear discriminant analysis，LDA）、偏最小二乘算法（partial least squares，PLS）、距离多普勒算法（range Doppler algorithm，RDA）、广义线性模型（generalized linear models，GLM）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）核函数（SVM-Radial）及 SVM 线性核函数（SVM-Linear），采用重复 10 次的 10 折交叉验证构建中老年人龋齿预测模型，在 10 次交叉验证中，对象集被随机分为 10 个大小大致相等的部分；在每次迭代中，以 90％的训练和 10％的验证百分比分割集，将数据集分成训练集（n＝459）和验证集（n＝51）两组，所有预测模型的训练和测试折叠完全相同。最后，使用 AUC 中位数比较列线图及基于 7 种算法的预测模型的效能。

1.2.4 筛选变量方法

首先，利用 Lasso 回归再次筛选特征变量，选择自适应法最优 λ 值，调整可能的混杂因素，使用 logistic 回归进行多因素分析。同时本研究中，观察到 RF 构建模型的效能最佳，而 RF 也可用于变量筛选，方法为使用 sbf 函数特征过滤，给每一维度的特征赋予权重，然后依据权重排序，显示各特征的重要性。采用 AUC 中位数和 7 种算法比较单因素＋Lasso＋多因素 logistic（以下简称 Lasso＋logistic）与 RF 筛选变量所构建模型的预测效能。

1.3 统计学处理

应用 SPSS23.0 和 Rstudio4.1.1 软件进行统计学分析。符合正态分布的计量资料以 $\bar{x}\pm s$ 表示，采用 t 检验进行单因素分析，非正态分布的计量资料以 $M(Q_1,Q_3)$ 表示，采用秩和检验进行单因素分析；计数资料以例数或百分比表示，采用 χ^2 检验进行单因素分析。基于单因素分析结果及传统的龋齿患病危险因素[13]，使用 logistic 回归进行多因素分析，以 $P<0.05$ 为差异有统计学意义。

2 结　果

2.1 中老年人患龋齿影响因素的单因素分析

本研究共纳入中老年人 510 名，口腔检查结果显示，363 名中老年人患龋齿（龋齿组），147 名中老年人未患龋齿（正常组），龋齿检出率为 71.18％；75 岁及以上老年人的无颌牙率达 11.93％（21/176）。单因素分析结果显示，年龄、共同生活人口数、使用牙签情况、看牙频率、牙存留数、OHAT 得分与中老年人患龋齿有关（$P<0.05$），见表 1。

表 1　中老年人患龋齿影响因素的单因素分析

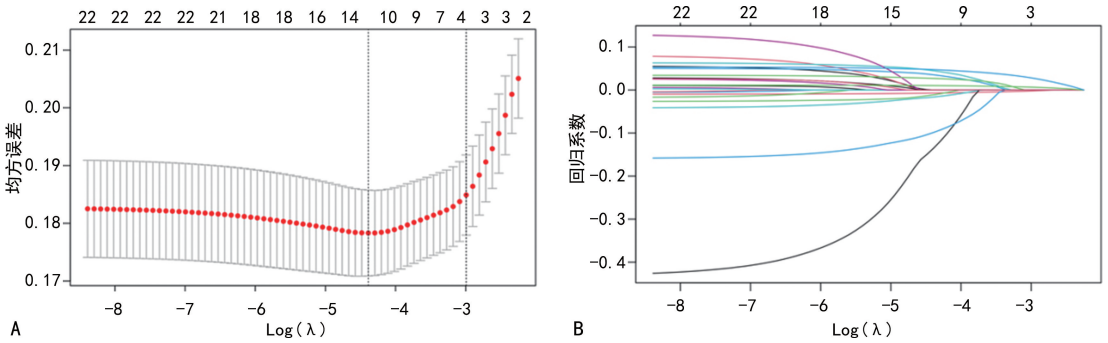
项目	龋齿组（n＝363）	正常组（n＝147）	$\chi^2/t/Z$	P
性别[n（％）]			0.051	0.821
男	105(28.93)	44(29.93)		
女	258(71.07)	103(70.07)		
年龄（$\bar{x}\pm s$，岁）	70.95±7.87	73.87±8.37	−3.724	＜0.001
户口[n（％）]			0.019	0.890
城市	96(26.45)	38(25.85)		
农村	267(73.55)	109(74.15)		
民族[n（％）]			1.982	0.371
汉族	156(42.98)	54(36.73)		
壮族	203(55.92)	92(62.59)		
其他	4(1.10)	1(0.68)		
文化程度[n（％）]			2.070	0.723
没上过学	90(24.79)	35(23.81)		
小学	165(45.45)	61(41.50)		
初中	67(18.46)	35(23.81)		
高中/中专	35(9.64)	13(8.84)		
大专及以上	6(1.65)	3(2.04)		
有无伴侣[n（％）]			0.462	0.497
有	229(63.09)	88(59.86)		
无	134(36.91)	59(40.14)		
共同生活人口数［$M(Q_1,Q_3)$］	2(2,5)	3(2,5)	−2.269	0.023
月收入[n（％）]			0.798	0.850

续表 1　中老年人患龋齿影响因素的单因素分析

项目	龋齿组(n=363)	正常组(n=147)	$\chi^2/t/Z$	P
无固定收入	139(38.29)	58(39.46)		
＜1 000 元	117(32.23)	48(32.65)		
1 000～3 000 元	92(25.34)	33(22.45)		
＞3 000 元	15(4.13)	8(5.44)		
吸烟史[n(%)]			1.695	0.193
有	75(20.66)	23(15.65)		
无	288(79.34)	124(84.35)		
饮酒史[n(%)]			0.030	0.863
有	49(13.50)	19(12.93)		
无	314(86.50)	128(87.07)		
食用甜食的习惯[n(%)]			0.131	0.717
有	69(19.01)	30(20.41)		
无	294(80.99)	117(79.59)		
喝甜饮料的习惯[n(%)]			0.008	0.929
有	63(17.36)	26(17.69)		
无	300(82.64)	121(82.31)		
刷牙频率[n(%)]			3.955	0.138
≥2 次/d	190(52.34)	91(61.90)		
1 次/d	154(42.42)	49(33.33)		
很少/从不	19(5.23)	7(4.76)		
使用牙签[n(%)]			7.109	0.008
是	193(53.17)	59(40.14)		
否	170(46.83)	88(59.86)		
使用冲牙器[n(%)]			0.881	0.327
是	2(0.55)	2(1.36)		
否	361(99.45)	145(98.64)		
看牙频率[n(%)]			10.218	0.017
半年内	15(4.13)	8(5.44)		
半年至 1 年	13(3.58)	6(4.08)		
1 年以上	184(50.69)	94(63.95)		
从没看过牙	151(41.60)	39(26.53)		
过去 1 年是否洗牙[n(%)]			1.267	0.260
是	12(3.31)	8(5.44)		
否	351(96.69)	139(94.56)		
高血压病史[n(%)]			2.893	0.089
有	117(32.23)	59(40.14)		
无	246(67.77)	88(59.86)		
糖尿病病史[n(%)]			1.209	0.272
有	30(8.26)	8(5.44)		
无	333(91.74)	139(94.56)		
牙存留数($\bar{x}\pm s$,颗)	21.35±8.06	16.38±12.16	4.565	＜0.001
OHAT 得分($\bar{x}\pm s$,分)	5.14±2.35	3.86±2.58	5.357	＜0.001

2.2 Lasso 回归筛选特征变量

结合单因素分析差异有统计学意义的变量及传统的龋齿患病危险因素进行 Lasso 回归分析。选择自适应法最优 λ 值(λ=0.012 41),此时纳入的预测变量包括 10 个:年龄、民族、文化程度、刷牙频率、使用牙签情况、使用冲牙器情况、看牙频率、过去 1 年是否洗牙、牙存留数及 OHAT 得分,见图 1。变量赋值见表 2。



A:筛选特征变量的 Lasso 系数分布;B:Lasso 模型自适应法最优 λ 值选择;λ:正则化参数。

图 1 Lasso 回归模型筛选预测变量

表 2 影响因素自变量赋值

项目	赋值方法
龋齿	否=0,是=1
年龄	连续型变量
民族	汉族=1,壮族=2,其他=3
文化程度	没上过学=1,小学=2,初中=3,高中/中专=4,大专及以上=5
刷牙频率	很少/从不=1,1 次/d=2,≥2 次/d=3
使用牙签	否=0,是=1
使用冲牙器	否=0,是=1
看牙频率	半年内=1,半年至 1 年=2,1 年以上=3,从没看过牙=4
过去 1 年是否洗牙	否=0,是=1
牙存留数	连续型变量
DHAT 得分	连续型变量

2.3 中老年人患龋齿影响因素的多因素 logistic 回归分析

将是否患龋齿作为因变量,Lasso 回归分析筛选的 10 个预测变量作为自变量,进行多因素 logistic 回归分析,结果显示:年龄、刷牙频率、过去 1 年是否洗牙、牙存留数、OHAT 得分为中老年人患龋齿的独立影响因素($P<0.05$),见表 3。

2.4 中老年人龋齿列线图预测模型的构建与验证

中老年人龋齿列线图预测模型,见图 2。ROC 曲线结果显示,该模型的 AUC 为 0.742(95%*CI*:0.693~0.791),见图 3A。内部验证结果显示,该模型预测的准确性中等(C-index=0.742,95%*CI*:0.683~0.785),并具有较好的校准能力($\chi^2=10.920$,$P=0.142$),见图 3B。通过重复 10 次的 10 倍交叉验证,列线图构建模型的 AUC 中位数为 0.733(95%*CI*:0.708~0.740),见图 4。

2.5 不同机器学习算法预测模型与列线图预测模型的比较

基于筛选出的 5 个显著特征建立 IDA、PLS、RDA、GLM、RF、SVM-Radial、SVM-Linear 共 7 个预测模型,应用 AUC 中位数进行预测效能评估,经过 10 倍交叉验证,列线图与 7 种机器学习算法构建模型的 AUC 见图 4,按中位数由小到大排列。模型对比结果显示,7 种机器学习算法构建模型均有一定的区分能力,其中 RF 表现出最佳性能,AUC 中位数为 0.747,其次为列线图,AUC 中位数为 0.733。

表 3 中老年人患龋齿影响因素的 logistic 回归分析

变量	*β*	*SE*	*Wald*	*P*	*OR*	95%*CI*
年龄	−0.057	0.015	14.163	<0.001	0.945	0.917~0.973
刷牙频率(≥2 次/d *vs*. 1 次/d=2、很少/从不)	−0.374	0.189	3.915	0.048	0.688	0.475~0.997
牙签使用情况(是 *vs*. 否)	0.397	0.230	2.987	0.084	1.487	0.948~2.332
过去 1 年是否洗牙(是 *vs*. 否)	−1.193	0.549	4.715	0.030	0.303	0.103~0.890
牙存留数	0.060	0.012	25.460	<0.001	1.062	1.038~1.087
OHAT 得分	0.310	0.051	37.406	<0.001	1.363	1.234~1.505
常量	3.217	1.456	4.882	0.027	24.957	

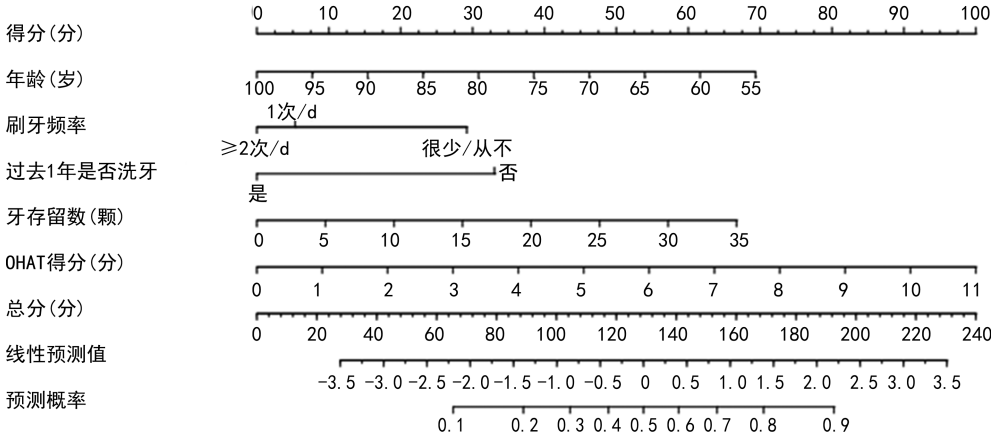
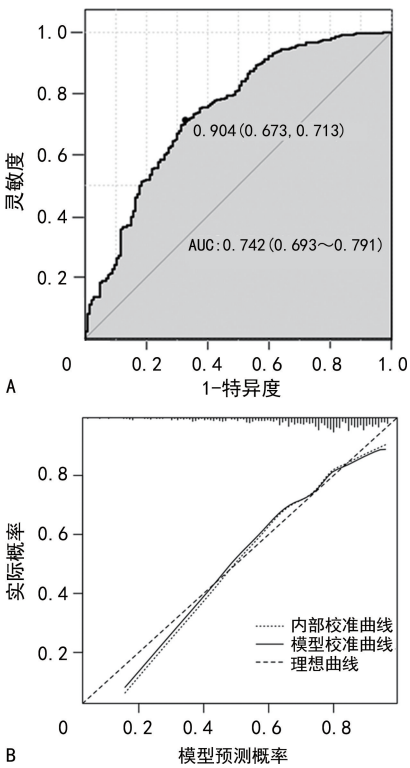
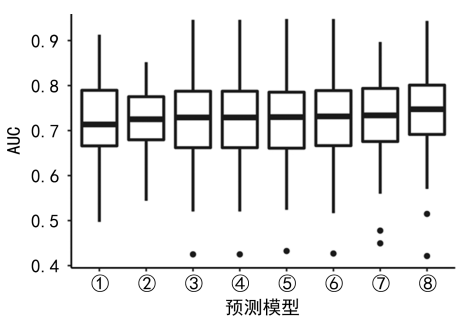


图 2　中老年人龋齿列线图预测模型



A:列线图预测模型的 ROC 曲线;B:列线图预测模型内部验证的校准曲线(Bootstrap 重复抽样 1 000 次,平均绝对误差=0.023,n=510)。

图 3　中老年人龋齿列线图预测模型的验证

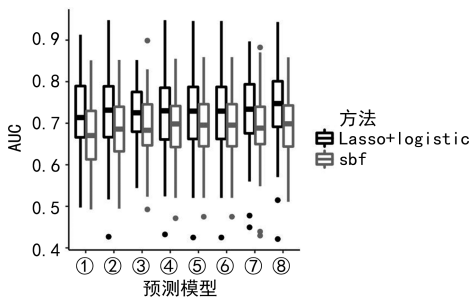


①:VM-Radial;②:PLS;③:IDA;④:RDA;⑤:GLM;⑥:SVM-Linear;⑦:Nomogram(列线图);⑧:RF。

图 4　7 种不同机器学习算法构建的预测模型与列线图预测模型的 AUC 比较

2.6　基于不同筛选变量方法构建的模型比较

比较 Lasso+logistic 和 caret 包自带的 sbf 函数筛选自变量所构建模型的预测性能。sbf 预测自变量采用 RF 模型,最后筛选出 6 个自变量,分别是:年龄、使用牙签情况、看牙频率、共同生活人口数、牙周炎、OHAT 得分。对比结果显示,在列线图预测模型和 7 种机器学习算法中,Lasso+logistic 筛选变量构建的模型 AUC 中位数均大于 RF 算法,见图 5。



①:VM-Radial;②:SVM-Linear;③:PLS;④:GLM;⑤:IDA;⑥:RDA;⑦:Nomogram(列线图);⑧:RF。

图 5　Lasso+logistic 与 sbf 函数筛选自变量构建模型的预测性能比较

3　讨　　论

3.1　中老年人龋齿的预测因素

本研究使用了列线图和 7 种机器学习算法来预测中老年人的龋齿患病率,并进行了比较。数据基于分层抽样,涵盖高中低发展水平城市,并平衡城乡患病差异,提高了研究对象的代表性。研究结果显示,年龄、刷牙频率、过去 1 年是否洗牙、牙存留数及 OHAT 得分是对模型预测中老年人龋齿具有重要意义的输入变量。由于随着年龄的增大,老年人的平均牙存留数减少,无颌牙率升高,龋齿检出人数也相对减少,本研究中 75 岁及以上老年人的无颌牙率达 11.93%。同时,刷牙频率≥2 次/d 与过去 1 年有洗牙经历是中老年人龋齿的保护性因素,这与 EDMAN 等[6]、王诗维等[14]研究结果一致。有 meta 分析显示[15],不刷牙者发生龋齿的风险为刷牙者的 1.967 倍,保持牙齿清洁是减少龋齿发生的重要途径。此

外，老年人的口腔保健意识与龋齿患病相关[16]，保持刷牙频率与定期到医院洗牙都是其良好的口腔健康知信行表现，提示医务工作者需要加强对老年人的口腔健康宣教，改善口腔卫生环境，鼓励老年人定期到牙科就诊，尽早发现龋齿问题，防止病情波及牙根造成牙齿被迫拔除。本研究中 OHAT 得分是中老年人龋齿的重要预测因子，OHAT 全面评估了口腔环境状况，可在一定程度上反映口腔健康的知识、态度、行为。荷兰的一项研究显示，该量表在老年人的非正式护理人员中可用性较强[17]，不用专业培训即可完成评分，可以考虑成为社区中老年人龋齿患病情况便捷评估的重要手段。

3.2 列线图与机器学习算法构建的预测模型比较

交叉验证后 AUC 中位数的比较有助于减少构建模型的过拟合情况[18]，因此本研究选用 AUC 中位数来比较各模型的性能。本研究中，列线图和机器学习算法构建的中老年龋齿预测模型 AUC 中位数均显示较好的效能，模型的引入可为基层社区医务人员进行中老年人龋齿筛查提供一种辅助工具。机器学习方法具有较强的非线性学习能力和随机性等优势，可以通过训练多个特征量，寻找它们之间深层、多维的关联，从而给出更为准确的判断[19]。列线图提供了一种吸引人、可视化的方法来预测患病风险，且不需使用互联网或计算机，可以加强人群与医护人员的共同决策[20]。对二者进行比较并分析各自优势有助于临床应用。本研究中，RF 的预测性能优于列线图和其他 6 种机器学习算法，它是由多个决策树算法组成的集合，分类精度和模型的准确性高，处理多维数据能力强，尤其适用于分类、回归及聚类等机器学习任务[21]。此外，RF 的原理包含每次随机选取样本变量，异常值不会对结果造成影响，能更好地适应不同的数据集，因此所得的预测结果更稳健[22]。SADEGH-ZADEH 等[23]与 QU 等[24]也对比了决策树算法、K 临近算法、线性回归、多层感知器、RF、SVM 模型对早期儿童龋齿风险的预测性能，结果均表明 RF 的准确率等预测性能指标最好，与本研究结果相似。

3.3 基于不同变量筛选方法构建的预测模型比较

本研究对 Lasso＋logistic 与 sbf 函数筛选变量所构建的模型进行了比较，以增加预测模型的严谨性。RF 可以同时构建多棵决策树，故筛选变量时具有分类性能好、训练快等优点，并且能计算出各特征变量的重要程度进行排序，但在计算时偏向选择投票最多的特征，可能产生过度匹配的问题[25]。而运用 Lasso＋logistic 筛选变量，可以引入范数惩罚函数，使非零系数的特征减少，在选择变量的同时达到缩减模型的效果，很好地降低了过度拟合问题，并且还可以减少变量的多重共线性，具有良好的泛化能力与鲁棒性[26]。由于本研究参数较多，解释变量直接相关性较高以致单个解释变量不显著，而 Lasso 回归可削弱过拟合及多重共线性的影响，建立更具有概化能力的预测模型，因此该方法更适用于本研究筛选中老年龋齿预测变量，并显示出较好的效能。另外，RF 的特征选择过程与训练过程相互独立，特征选择使用 sbf 函数过滤法，虽然在速度上占优势，但可能会删除非常有实质意义的变量[27]，本研究中 RF 算法在构建模型中表现出较好的效能而在变量筛选中效能略低于 Lasso 回归可能与之有关。

综上所述，选择最合适的算法和变量筛选方法对构建模型至关重要，需应用多算法来构建预测模型并对比，防止模型出现“欠拟合”或“过拟合”现象；同时，应结合临床实践评定模型的效能及可行性，基于已有的行政数据、登记数据或电子健康档案等医疗大数据和机器学习算法，开发预测性能更高的预测模型。此外，不同变量筛选方法与多算法对比为模型构建提供了新的视角，可应用于其他领域的个体疾病预测及治疗决策[28]，提高预测模型的实用性与科学性。

参考文献

[1] BORG-BARTOLO R，ROCCUZZO A，MOLINER-OMOURELLE P，et al. Global prevalence of edentulism and dental caries in middle-aged and elderly persons：a systematic review and meta-analysis[J]. J Dent，2022，127：104335.

[2] 王兴，冯希平，李志新. 第四次全国口腔流行病学调查报告[M]. 北京：人民卫生出版社，2018：29-33.

[3] GU W，LI J，LI F，et al. Association between oral health and cognitive function among Chinese older adults：the Taizhou imaging study[J]. BMC Oral Health，2023，23(1)：640.

[4] GUO P，ZOU C，AN N，et al. Emotional symptoms，dietary patterns and dental caries：a cross-sectional study in adolescents[J]. Oral Dis，2024，30(4)：2653-2662.

[5] CHAN A K Y，TSANG Y C，JIANG C M，et al. Diet，nutrition，and oral health in older adults：a review of the literature[J]. Dent J (Basel)，2023，11(9)：222.

[6] EDMAN K，HOLMLUND A，NORDERYD O. Caries disease among an elderly population：a 10-year longitudinal study[J]. Int J Dent Hyg，2021，19(2)：166-175.

[7] SHAO Z，WANG Z，BI S，et al. Establishment and validation of a nomogram for progression to diabetic foot ulcers in elderly diabetic patients[J]. Front Endocrinol (Lausanne)，2023，14：1107830.

[8] SILVA G F S，FAGUNDES T P，TEIXEIRA B

C,et al. Machine learning for hypertension prediction:a systematic review[J]. Curr Hypertens Rep,2022,24(11):523-533.
[9] 刘璐. 基于人工神经网络技术的老年龋预测模型的构建及预测方法学比较的泛化能力验证研究[D]. 沈阳:中国医科大学,2021.
[10] 王勘琼,朱树贞,詹艳,等. 口腔健康评估量表的汉化及信效度检验[J]. 中华现代护理杂志,2019,25(28):3607-3610.
[11] 原国家卫生和计划生育委员会. 口腔健康调查:检查方法:WS/T 472-2015[S]. 北京:中国标准出版社:2016.
[12] CAI Z,LIN H,LI Z,et al. A prediction nomogram for postoperative gastroparesis syndrome in right colon cancer:a retrospective study[J]. Langenbecks Arch Surg,2023,408(1):148.
[13] CHOU R,PAPPAS M,DANA T,et al. Screening and interventions to prevent dental caries in children younger than 5 years:updated evidence report and systematic review for the US preventive services task force[J]. JAMA,2021,326(21):2179-2192.
[14] 王诗维,杨建军,张松梓. 老年龋病患者流行病学特征及 Carisolv 化学机械疗法的应用[J]. 中国老年学杂志,2023,43(17):4336-4339.
[15] TESHOME A,MUCHE A,GIRMA B. Prevalence of dental caries and associated factors in East Africa,2000—2020:systematic review and meta-analysis[J]. Front Public Health,2021,9:645091.
[16] ALLEN F,FAN S Y,LOKE W M,et al. The relationship between self-efficacy and oral health status of older adults[J]. J Dent,2022,122:104085.
[17] HO B V,VAN DE RIJT L J M,WEIJENBERG R A F,et al. Oral Health Assessment Tool (OHAT) deputized to informal caregivers:go or no go? [J]. Clin Exp Dent Res,2022,8(1):76-83.
[18] LI R,ZHU J,ZHONG W D,et al. Comprehensive evaluation of machine learning models and gene expression signatures for prostate cancer prognosis using large population cohorts[J]. Cancer Res,2022,82(9):1832-1843.
[19] 高泽鹏,李庆峰. 利用机器学习方法对几个核物理问题的深入研究[J]. 核技术,2023,46(8):92-99.
[20] ALABI R O,MÄKITIE A A,PIRINEN M,et al. Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer[J]. Int J Med Inform,2021,145:104313.
[21] HU J,SZYMCZAK S. A review on longitudinal data analysis with random forest[J]. Brief Bioinform,2023,24(2):bbad002.
[22] 李宁. 机器学习模型在子宫内膜癌生存预测中的应用研究[D]. 济南:山东财经大学,2023.
[23] SADEGH-ZADEH S A,RAHMANI QERANQAYEH A,BENKHALIFA E,et al. Dental caries risk assessment in children 5 years old and under via machine learning[J]. Dent J (Basel),2022,10(9):164.
[24] QU X,ZHANG C,HOUSER S H,et al. Prediction model for early childhood caries risk based on behavioral determinants using a machine learning algorithm[J]. Comput Methods Programs Biomed,2022,227:107221.
[25] 曹海涛,朱静,马云鹏,等. 机器学习在肠道菌群宿主表型预测中的应用[J]. 生物技术进展,2023,13(5):671-680.
[26] 张沥今,魏夏琰,陆嘉琦,等. Lasso 回归:从解释到预测[J]. 心理科学进展,2020,28(10):1777-1791.
[27] YAO N,PAN J,CHEN X,et al. Discovery of potential biomarkers for lung cancer classification based on human proteome microarrays using stochastic gradient boosting approach[J]. J Cancer Res Clin Oncol,2023,149(10):6803-6812.
[28] ELORANTA S,BOMAN M. Predictive models for clinical decision making:deep dives in practical machine learning[J]. J Intern Med,2022,292(2):278-295.

(收稿日期:2023-12-25 修回日期:2024-03-29)
(编辑:冯 甜)