·论著·

基于机器学习算法构建慢性萎缩性胃炎风险预测模型的研究

韩经略[1]　颜财旺[2]　胡非凡[1]　沈耀[1]　高晓娟[1]　赵嘉敏[1]　章华臻[1]　殷洁[1]　占强[1]　安方梅[1]

[1]南京医科大学附属无锡人民医院消化内科 南京医科大学无锡医学中心 国家消化系统疾病临床医学研究中心(西安)江苏省分中心,无锡 214023;[2]南京医科大学公共卫生学院,南京 211166

通信作者:安方梅,Email:fangmeian@njmu.edu.cn

【摘要】 目的　建立并验证一种基于机器学习的模型,用于预测发生慢性萎缩性胃炎(chronic atrophic gastritis,CAG)的可能性。**方法**　回顾性纳入2022年10月至2023年12月在无锡地区1 268例参与胃癌筛查队列的患者,通过问卷调查、血清学检测、上消化道内镜检查和病理活检获取研究数据。首先通过单因素logistic回归、Boruta算法和LASSO回归方法进行CAG特征筛选。其次采用8种机器学习算法,即logistic回归、SVM、GBM、神经网络、XGBoost、AdaBoost、LightGBM和CatBoost,使用5折交叉验证法训练并开发CAG预测机器学习模型。使用受试者工作特征曲线、校准曲线、决策曲线、特异度、灵敏度等指标评估这些模型的性能。最后利用SHAP分析对模型的每个特征及其决策依据进行解释,以提高该模型预测CAG的可读性。**结果**　通过单因素logistic回归、Boruta算法和LASSO回归3种方法共同识别出幽门螺杆菌(*Helicobacter pylori*,*HP*)感染、胃蛋白酶原比值(pepsinogen ratio,PGR)、年龄、吸烟史、性别及胃癌家族史为CAG的重要预测因素。进一步分析对比发现,logistic回归模型在测试集中受试者工作特征曲线下面积为0.805(95%*CI*:0.762~0.849),灵敏度为79%,特异度为71%,性能不弱于其他7种机器学习模型。最后利用SHAP方法识别出*HP*感染、PGR和年龄是影响机器学习模型预测CAG的主要核心因素。**结论**　基于人口学和临床因素的机器学习算法能准确预测CAG的发生率,*HP*感染、PGR和年龄是预测CAG的主要核心因素,而吸烟史、性别及胃癌家族史则可提升CAG的发病风险。这将有助于临床实践中早期发现和诊断CAG。

【关键词】 胃炎,萎缩性;　癌前疾病;　筛查;　机器学习算法;　预测模型

基金项目:无锡市重大项目(Z202208);南京医科大学无锡医学中心队列项目(WMCC202502);南京医科大学无锡医学中心重大项目(WMCM202501)

临床试验注册:中国临床试验注册中心(ChiCTR2400085856)

Development and validation of a machine learning-based risk prediction model for chronic atrophic gastritis

Han Jinglue[1], Yan Caiwang[2], Hu Feifan[1], Shen Yao[1], Gao Xiaojuan[1], Zhao Jiamin[1], Zhang Huazhen[1], Yin Jie[1], Zhan Qiang[1], An Fangmei[1]

[1]Department of Gastroenterology, Wuxi People's Hospital Affiliated to Nanjing Medical University, Wuxi Medical Center of Nanjing Medical University, Jiangsu Branch of National Clinical Research Center for Digestive Diseases (Xi'an), Wuxi 214023, China; [2]School of Public Health, Nanjing Medical University, Nanjing 211166, China

Corresponding author: An Fangmei, Email: fangmeian@njmu.edu.cn

DOI:10.3760/cma.j.cn321463-20250428-00031

收稿日期　2025-04-28　　**本文编辑**　朱悦

引用本文:韩经略, 颜财旺, 胡非凡, 等. 基于机器学习算法构建慢性萎缩性胃炎风险预测模型的研究[J]. 中华消化内镜杂志, 2026, 43(2): 108-115. DOI: 10.3760/cma.j.cn321463-20250428-00031.

【Abstract】 **Objective** To establish and validate a machine learning-based model for predicting the risk of chronic atrophic gastritis (CAG). **Methods** This retrospective study enrolled 1 268 participants from a gastric cancer screening cohort in Wuxi from October 2022 to December 2023. Data were collected through questionnaires, serological tests, upper gastrointestinal endoscopy, and pathological biopsies. Feature selection was performed by using univariate logistic regression, Boruta algorithm, and LASSO regression. Eight machine learning algorithms—logistic regression, SVM, GBM, neural network, XGBoost, AdaBoost, LightGBM, and CatBoost—were trained and developed using 5-fold cross-validation. Model performance was evaluated using multiple metrics, including receiver operating characteristic curve, calibration curves, decision curves, specificity, and sensitivity. SHAP analysis was applied to interprete feature contributions and decision basis, enhancing the model′s interpretability. **Results** Three methods (univariate logistic regression, Boruta algorithm, and LASSO regression) identified *Helicobacter pylori* (*HP*) infection, pepsinogen ratio (PGR), age, smoking history, gender, and family history of gastric cancer as key predictors for CAG. The logistic regression model achieved an area under the curve of 0.805 (0.762-0.849) in the test set, with a sensitivity of 79% and specificity of 71%, performing comparably with seven other machine learning models. SHAP analysis further highlighted *HP* infection, PGR, and age as the most influential features in predicting CAG. **Conclusion** Machine learning algorithms based on demographic and clinical factors can accurately predict CAG risk. *HP* infection, PGR, and age are identified as the core predictive factors, while smoking history, gender, and family history of gastric cancer may increase the risk of CAG. This approach could facilitate early detection and diagnosis of CAG in clinical practice.

【Key words】 Gastritis, atrophic; Precancerous condition; Screening; Machine learning algorithms; Predictive model

Fund program: Major Project of Wuxi City (Z202208); Cohort Project of Wuxi Medical Center of Nanjing Medical University (WMCC202502); Major Project of Wuxi Medical Center of Nanjing Medical University (WMCM202501)

Trial registration: Chinese Clinical Trial Registry (ChiCTR2400085856)

慢性萎缩性胃炎(chronic atrophic gastritis, CAG)是一种以胃黏膜固有腺体数量减少或消失为特征,常伴有肠上皮化生的慢性胃病,被认为是胃黏膜癌变过程中关键的前驱步骤[1]。研究发现,CAG 患者发生胃癌的风险是正常人的 4~6 倍,且 CAG 患者 20 年内患胃癌风险为 1/50[2]。另一项 10 年以上的随访研究发现,从初始诊断到确诊胃癌的中位间隔时间在 CAG 患者中为 1.6 年[3]。在东亚地区,据估计每年约有 1.8% 的 CAG 会进展为胃癌[4]。包含 107 项研究的 meta 分析表明,CAG 的全球患病率为 33%,其中,胃癌高发国家的 CAG 患病率显著高于低发国家(42% 比 23%)[5]。国内一项大规模调查显示,在 8 892 例慢性胃炎患者中,CAG 的比例为 17.7%,内镜诊断的准确率低于组织学诊断,且内镜对其诊断准确率仅为 50.3%[6]。在我国,由于大多数患者无明显症状而未做内镜检查,或即便有症状但惧怕做内镜检查,实际 CAG 患病率可能更高,估计我国 CAG 的患病率超过 20%[7]。因此,早期诊断 CAG 对于降低胃癌的发病率至关重要。

传统诊断 CAG 的方法依赖于内镜联合活检病理检查,但这作为筛查手段人力成本高、检查过程有创伤、患者依从性差,且诊断准确性依赖于专业人员的临床专业知识和主观判断。研究发现,内镜医师对于内镜下 CAG 的诊断存在异质性,且高年资内镜医师的诊断准确率优于低年资内镜医师[8]。鉴于以上现状,目前急需一种无创、能客观预测 CAG 的方法用于临床实践中。机器学习旨在让计算机系统从大量的数据中自动学习模式和规律,进而对未知的数据进行准确预测或分类[9]。通过机器学习算法可分析大量患者数据,并识别潜在的关联和模式,建立疾病预测模型,使医师能够更方便地预测发生疾病的可能性,从而精准筛选出高危人群。目前机器学习已用于消化内镜辅助诊断、胃肠道肿瘤的风险预测及预后估计等领域[10-11],但尚缺乏基于机器学习的 CAG 预测模型。

在本研究中,我们回顾性分析了本地区胃癌筛查队列患者的人口学信息和临床检验数据,探究了发生 CAG 的相关危险因素,基于机器学习算法构建了 CAG 预测模型,为 CAG 的内镜检查前高危人群筛选提供了可靠的临床证据。

材料与方法

一、受试者资料

本研究在实施前已获得南京医科大学附属无

锡人民医院伦理委员会批准，并遵循所有适用的法律法规和伦理标准，研究方案已提交并获得了无锡市人民医院科研伦理委员会的备案批准（KY23001）。所有参与者在参与研究前均已签署知情同意书，并充分了解研究内容、目的、可能的风险及其个人权益。

这项回顾性研究分析了2022年10月至2023年12月期间本地区胃癌筛查队列患者的数据。自2022年10月起，按社区分批邀请40岁以上永久居民参与该项目。通过面对面问卷调查收集患者的一般信息，包括性别、年龄、体重指数及受教育程度；患者的生活习惯，包括吸烟史、饮茶史、饮食喜好温度及口味、睡眠时间及睡眠质量；患者的病史，包括高血压及糖尿病病史，肿瘤及胃癌家族史。在面对面问卷调查后，每位参与者接受血清幽门螺杆菌（*Helicobacter pylori*，*HP*）抗体、胃蛋白酶原（pepsinogen，PG）检测及上消化道内镜和活检病理检查。排除缺乏病理活检报告及缺失血清学指标的参与者，以及病理报告提示为胃癌的患者。

二、血清*HP*抗体及PG检测

血清*HP*抗体检测采用胶体金免疫法分析（杭州安旭生物科技股份有限公司），PG检测采用双标记时间分辨荧光免疫分析法（无锡市江原实业技贸总公司），以上试验按照试剂盒说明书流程操作，根据PGⅠ和PGⅡ的比值计算得到胃蛋白酶原比值（pepsinogen ratio，PGR）（PGⅠ/PGⅡ）。

三、CAG的诊断标准

在悉尼系统OLGIM 5点活检诊断标准的基础上，结合当地具体情况，采用3点活检法，即分别在胃窦小弯距幽门2~3 cm处、胃体小弯近胃角4 cm处及胃体大弯近贲门8 cm处取胃组织活检。病理诊断存在萎缩合并肠上皮化生即可确诊。所有病理结果由2名拥有10年以上消化道病理诊断经验的病理医师独立诊断，存在分歧时再由更高一级专家复核作出最终诊断[12]。

四、研究设计

开发一种基于机器学习的模型，用于预测患者发生CAG的风险。将筛查队列患者的临床信息和检查结果数据集按7∶3划分为训练集和测试集，使用单因素逻辑（logistic）回归，Boruta算法[13]和最小绝对收缩和选择算子（LASSO）回归[14]进行特征筛选，以识别与CAG风险最相关的特征。随后，应用5折交叉验证方法，开发包括logistic回归、SVM、GBM、神经网络（Neural Network）、XGBoost、AdaBoost、LightGBM、CatBoost在内的8种机器学习模型，并进行优化。使用受试者工作特征曲线[15]、校准曲线、决策曲线、特异度、灵敏度等指标评估模型的性能。为了提高模型的透明度和可解释性，采用SHAP[16-17]方法解释预测结果并阐明每个特征对预测的影响。

五、统计学分析

数据分析基于R语言（4.4.0），正态分布的计量资料以$\bar{x}\pm s$表示，组间比较使用t检验；非正态分布的计量资料采用$M(Q_1, Q_3)$表示，组间比较使用Mann-Whitney U检验；分类变量使用例（%）表示，采用卡方检验或Fisher精确概率法。LASSO特征筛选通过glmnet包可视化完成，Boruta特征选择通过Boruta包执行，SHAP分析结合shapviz包进行可视化。所有统计分析显著性阈值设为$P<0.05$。

结　果

一、患者特征

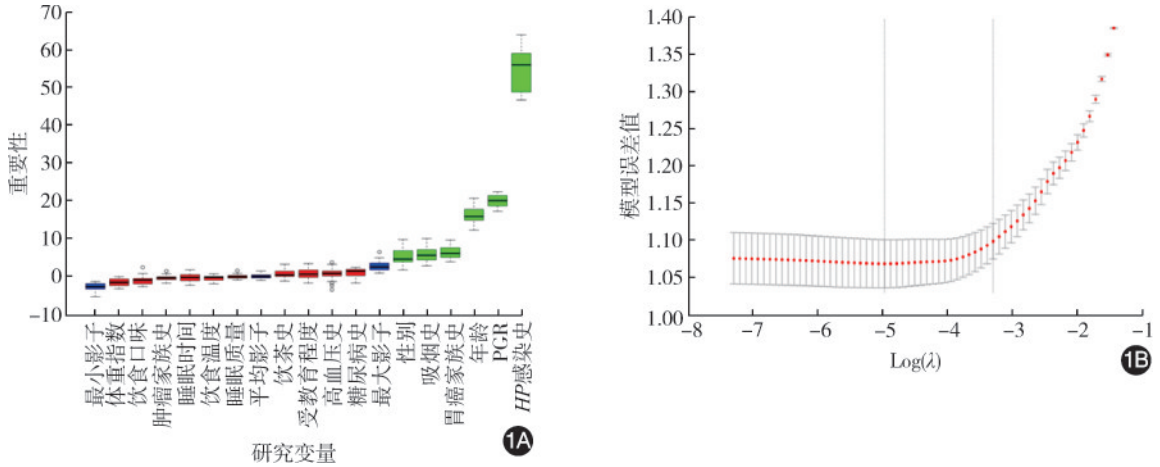
截至2023年12月，共招募了1 375例参与者，最后共1 268例参与者有完整的基线资料，被纳入到研究中。受试者中位年龄为57岁（40~77岁），其中男532例（42.0%）、女736例（58.0%）。根据胃镜及病理诊断结果，619例（48.8%）被确诊为非慢性萎缩性胃炎（Non-CAG），649例（51.2%）为CAG。在CAG患者中，341例（52.5%）伴发肠上皮化生，14例（2.2%）伴异型增生，中度异型增生12例、高度异型增生2例。

二、预测因子筛选

1 268例入组患者分为训练集887例和测试集381例，两组人数比例为7∶3。Boruta算法（图1A）、LASSO回归（λ.1se=0.034）（图1B）及logistic回归（表1）共同识别出6个CAG关键特征：*HP*感染史、PGR、年龄、吸烟史、性别和胃癌家族史。LASSO通过L1正则化压缩估计解决共线性问题，Boruta基于特征重要性筛选，单因素回归验证变量关联性，三者结果高度一致，最终纳入模型构建。

三、模型性能检测

采用5次重复训练构建模型，比较logistic回归、SVM、GBM、神经网络、XGBoost、AdaBoost、LightGBM、CatBoost的预测性能，基于5次独立实验的稳定性验证，通过受试者工作特征曲线下面积（area under the curve，AUC）指标在训练集（图2A）和测试集（图2B）进行模型评估。训练集结果显



注：PGR指胃蛋白酶原比值；*HP*指幽门螺杆菌

图1　慢性萎缩性胃炎（CAG）相关预测因子筛选　1A：Boruta算法CAG重要和非重要特征区分，重要特征用绿色表示；1B：LASSO回归CAG关键特征识别，图中左侧虚线（λ.min）表示所有λ值中最小的最优解，右侧虚线（λ.lse）代表在λ.min一个方差范围内可得到的最简模型对应的λ值

表1　慢性萎缩性胃炎风险因素单因素及多因素logistic回归分析

研究变量	单因素					多因素				
	*β*值	标准误	*Z*值	*P*值	*OR*值（95%*CI*）	*β*值	标准误	*Z*值	*P*值	*OR*值（95%*CI*）
性别（女/男）	−0.52	0.14	−3.79	<0.001	0.59（0.45~0.78）	−0.37	0.24	−1.53	0.125	0.69（0.43~1.11）
受教育程度（初中以上/初中及以下）	−0.23	0.15	−1.50	0.133	0.79（0.59~1.07）					
吸烟史（有/无）	0.69	0.15	4.57	<0.001	2.00（1.49~2.69）	0.38	0.26	1.44	0.151	1.46（0.87~2.46）
饮茶史（多于每周一次/每周一次或更少）	0.23	0.15	1.56	0.119	1.26（0.94~1.68）					
饮食喜好温度（过热或过凉/温度适中）	0.14	0.19	0.77	0.444	1.15（0.80~1.66）					
饮食喜好口味（偏咸/正常或偏淡）	0.50	0.30	1.64	0.101	1.64（0.91~2.98）					
睡眠质量（一般或较好/较差）	−0.08	0.15	−0.55	0.582	0.92（0.69~1.23）					
高血压史（有/无）	0.25	0.14	1.78	0.075	1.29（0.97~1.71）					
糖尿病史（有/无）	0.10	0.22	0.46	0.646	1.11（0.72~1.71）					
肿瘤家族史（有/无）	−0.07	0.15	−0.47	0.641	0.93（0.70~1.24）					
胃癌家族史（有/无）	0.54	0.22	2.47	0.014	1.71（1.12~2.62）	0.89	0.26	3.43	<0.001	2.44（1.47~4.06）
幽门螺杆菌感染史（有/无）	2.10	0.15	13.66	<0.001	8.18（6.05~11.05）	1.99	0.17	11.68	<0.001	7.29（5.22~10.17）
胃蛋白酶原比值	−0.17	0.02	−8.11	<0.001	0.85（0.81~0.88）	−0.12	0.02	−5.04	<0.001	0.89（0.85~0.93）
年龄（岁）	0.05	0.01	5.71	<0.001	1.05（1.03~1.07）	0.06	0.01	5.78	<0.001	1.07（1.04~1.09）
体重指数（kg/m^2）	0.01	0.02	0.54	0.588	1.01（0.97~1.06）					
睡眠时间（h）	0.05	0.05	0.90	0.370	1.05（0.95~1.16）					

示，CatBoost模型表现最优（AUC=0.842，95%*CI*：0.816~0.868），XGBoost与LightGBM次之（AUC=0.836），传统logistic回归达到中等水平（AUC=0.819）。测试集评估中，XGBoost保持最佳性能（AUC=0.815），logistic回归（AUC=0.805）与CatBoost（AUC=0.810）呈现良好泛化能力，而AdaBoost模型出现显著性能衰减（AUC=0.763）。通过决策曲线和校准曲线评估模型临床效用（图2C~2F），结合综合性能指标（准确率、灵敏度、特异度等，详见表2）的系统分析，发现各模型在训练及测试集间保持稳定性能。值得注意的是，尽管集成模型在预测精度上具有优势，但logistic回归展现出最佳的效能平衡特性。基于其模型简洁性、计算高效性及临床可解释性优势，本研究选择该模型进行深入分析。

多因素logistic回归预测模型提示*HP*感染史、胃癌家族史、年龄及PGR降低是CAG的独立危险因素（表1），其回归方程为y=−3.39−0.37×性别+0.38×吸烟史+0.89×胃癌家族史+1.99×*HP*感染史−0.12×PGR+0.06×年龄。模型验证显示良好预测效能：Hosmer-Lemeshow检验证实预测概率与观测概率无显著偏离（*P*>0.05），决策曲线分析表明在0.1~0.9风险阈值范围内具有临床适用性。

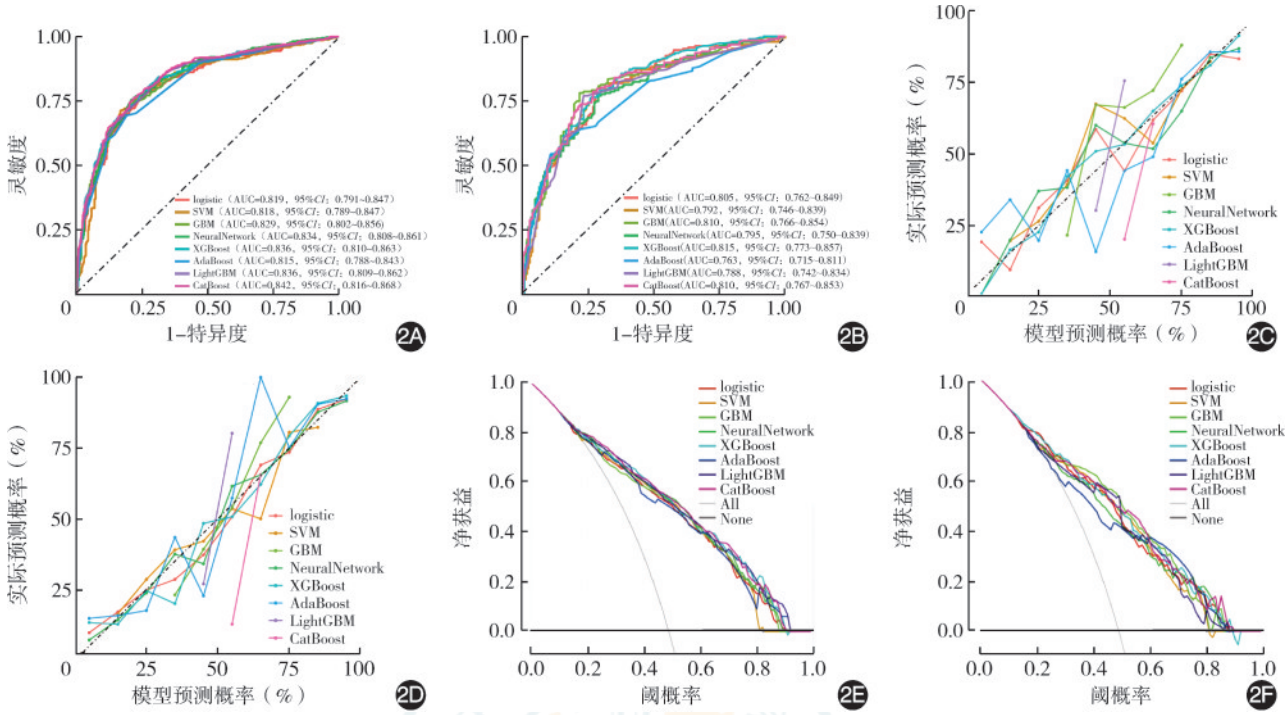


图 2　8 种不同预测模型的性能比较　2A：训练集受试者工作特征曲线；2B：测试集受试者工作特征曲线；2C：训练集校准曲线；2D：测试集校准曲线；2E：训练集决策曲线；2F：测试集决策曲线

表 2　8 种不同预测模型指标比较

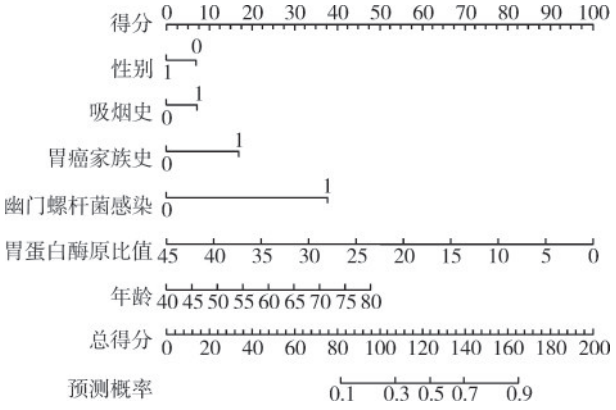
模型	训练集					测试集				
	准确率	灵敏度	特异度	精确率	F1 评分	准确率	灵敏度	特异度	精确率	F1 评分
logistic	0.76	0.75	0.77	0.76	0.76	0.75	0.79	0.71	0.72	0.75
SVM	0.78	0.71	0.83	0.8	0.76	0.76	0.80	0.71	0.73	0.76
GBM	0.77	0.75	0.78	0.77	0.76	0.78	0.78	0.79	0.78	0.78
神经网络	0.78	0.75	0.80	0.78	0.77	0.74	0.77	0.71	0.72	0.74
XGBoost	0.77	0.78	0.76	0.75	0.77	0.75	0.76	0.74	0.74	0.75
AdaBoost	0.75	0.69	0.81	0.78	0.73	0.72	0.54	0.89	0.83	0.66
LightGBM	0.76	0.82	0.70	0.72	0.77	0.77	0.77	0.77	0.76	0.77
CatBoost	0.77	0.76	0.78	0.77	0.77	0.76	0.79	0.73	0.74	0.76

四、列线图

基于关键预测因子开发列线图模型以实现风险量化评估，通过各变量对应分值的线性叠加，可直观计算个体患者的 CAG 发病率。该可视化工具在保持模型预测精度的同时，显著提升了临床决策支持系统的可解释性（图 3）。

五、基于 SHAP 的 CAG 模型可解释性分析

基于 SHAP 可解释性框架，系统解析 logistic 回归模型预测 CAG 风险的决策机制，按照特征重要性进行排序显示，*HP* 感染史、PGR 及年龄构成了 CAG 核心预测因子（图 4A）。其中，*HP* 阳性、PGR 降低、高龄、存在胃癌家族史及吸烟史对预测结果有正向贡献，均可显著提升 CAG 风险，而女性性别呈现保护效应（图 4B）。



注：第 1 行代表得分尺度，每 1 个变量的得分基于第 1 行的得分尺度；第 8 行代表 6 个预测变量的总得分；第 9 行代表慢性萎缩性胃炎预测概率

图 3　慢性萎缩性胃炎的列线图模型

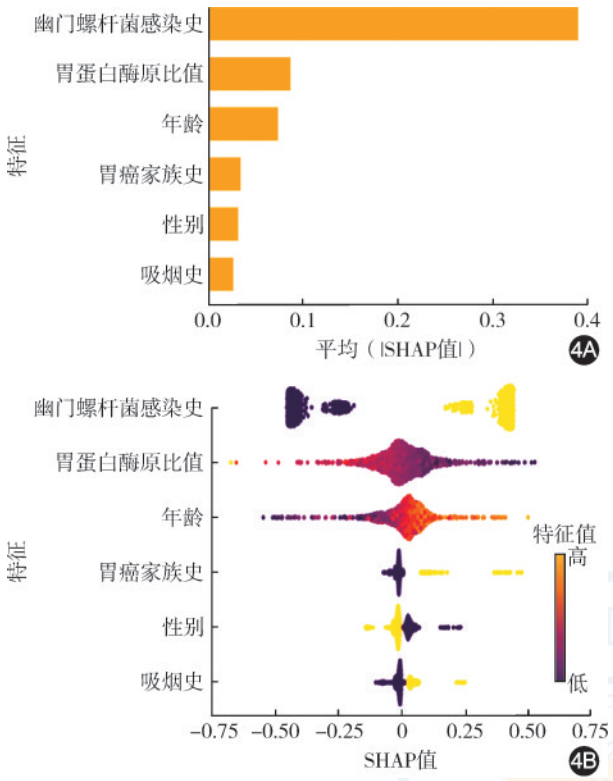


图 4　慢性萎缩性胃炎 logistic 回归模型的可解释性分析 4A：SHAP 特征重要性条形图，条形图描述了在最终预测模型开发中的重要性排序；4B：SHAP 特征蜂窝图，图中每个点代表一个特征，黄点表示较高的特征值，蓝点表示较低的特征值

讨　论

近年来，基于代谢组学、影像组学及内镜数据的胃癌预测模型在临床应用中取得突破性进展。有研究通过整合血浆代谢物特征与机器学习算法，构建了高精度胃癌诊断模型和预后模型[18]，其性能显著优于传统肿瘤标志物和临床经验判断。这些模型不仅能准确区分早期胃癌，还可通过动态监测代谢物变化实现风险分层，为个体化治疗提供依据。尽管已知 CAG 的严重程度与胃癌风险呈正相关[19]，但目前研究尚缺乏 CAG 向胃癌演进过程的动态预测模型。通过整合多模态数据构建风险评估体系可为 CAG 患者的个体化随访间隔时间和干预时机提供量化依据，从而提升胃癌防控体系的整体效能。

本研究基于机器学习技术构建多种机器学习预测模型，系统探索 CAG 的危险因素及预测效能。研究创新性地将 Boruta 算法与 LASSO 回归联合单因素 logistic 回归分析方法整合，通过特征重要性排序与变量筛选验证，精准筛选出与 CAG 密切相关的核心预测因子。结果发现，所选特征中 *HP* 感染、PGR 和年龄与 CAG 的关系最为密切。

作为我们模型得到的 CAG 核心预测因子之一，*HP* 感染是慢性胃炎最主要的病因[7]。*HP* 的黏附与毒力作用，破坏黏膜屏障，同时 *HP* 调节炎症因子释放，引起慢性炎症与免疫失调，通过多靶点、多层次的致病机制，*HP* 持续感染最终导致胃黏膜腺体破坏、功能丧失，成为 CAG 发生发展的核心驱动因素[20]。一项前瞻性 CAG 研究中发现，2/3 的患者有 *HP* 感染证据[21]。一项 meta 分析发现，*HP* 阳性患者的 CAG 发病率是 *HP* 阴性患者的 5 倍，这表明 CAG 发病与 *HP* 感染之间存在密切关系[22]。研究中我们也发现 CAG 患者 *HP* 阳性的比率是非患者的 7.29 倍，这与既往的结果相一致。

PGR 作为我们模型得到的另一个 CAG 核心预测因子，是评估胃黏膜功能的敏感指标，表达值下降与 CAG 的发病密切相关[23]。PGR 的血清浓度能够客观体现胃黏膜腺体的分泌能力状况，其检测结果可作为评估胃黏膜是否存在广泛性腺体萎缩的重要生物学指标[24]。我们的分析结果也表明随着 PGR 值的降低，CAG 的风险逐步增加。

年龄作为我们模型得到的第 3 个 CAG 核心预测因子，与 CAG 的发生风险已有多项研究证实。英国胃肠病学会发布的胃癌管理指南明确指出，CAG 的发生风险与年龄之间呈现出正相关态势[25]。美国一项研究在对 6 年时间内接受胃镜检查的多达 48 万例受试者展开了深入分析，结果显示慢性活动性胃炎的患病率有着明显的年龄变化趋势，从 20 岁时的 5% 逐步攀升至 40 岁时的 12%；而 CAG 患病率同样随年龄增长而升高，在 60 岁时约为 5%，到了 80 岁则升高至 10%，此后每增长 10 岁，其患病率大致会增高 5%[26]。值得注意的是，CAG 发病呈现出的这种年龄依赖特征和 *HP* 感染有着极为密切的关联。由于 *HP* 持续感染的时间不断累积，再加上炎症反应对胃黏膜造成的损伤日益加剧，这一系列情况致使 CAG 的发生风险持续升高[7]。

本研究中，虽然通过 Boruta 算法、LASSO 回归和单因素 logistic 回归共同识别筛选出 *HP* 感染史、PGR、年龄、吸烟史、性别和胃癌家族史为影响 CAG 发生的 6 个关键特征，但进一步通过 SHAP 方法识别得到 *HP* 感染、PGR 和年龄是影响机器学习模型预测 CAG 的主要核心因素。胃癌家族史作为胃癌的重要遗传易感因素，可能通过表观遗传修饰或肿瘤微环境重塑等机制，增加胃癌及癌前病变的发生发展[27]。然而现有流行病学证据对此存在争议：在

东亚地区，由于*HP*感染率长期居高不下（60%~90%）[28]，胃癌家族史可能更多反映的是*HP*感染的家族聚集性特征，而非独立的遗传易感性标志。本研究构建的预测模型中，虽然胃癌家族史被认为是CAG的独立危险因素，但胃癌家族史对整体风险预测的贡献度较小。一方面，家族成员共享的饮食习惯和环境暴露可能构成混杂因素，另一方面，特定的遗传易感基因突变可能通过调控诸如宿主对*HP*的免疫应答等通路，间接影响胃癌发生进程[29]。性别与胃癌发病存在潜在关联。基于胃癌流行病学的多维度证据显示，雌激素可能通过特定生物学机制发挥胃黏膜保护作用。英国的“Million Women”研究项目构建了131万女性的前瞻性队列，发现绝经后女性发生胃癌的风险是绝经前女性的1.59倍，已绝经女性中，绝经时间每提早5年，胃癌发生风险会增加1.18倍[30]。雌激素替代治疗会减少胃癌风险（*RR*=0.77），而使用抗雌激素药物他莫昔芬会增加胃癌发生风险（*RR*=1.87）[31]。本研究中发现性别是CAG的重要特征，但进一步的多因素logistic回归分析得到性别不是CAG的独立危险因素。这可能因为性别并非影响CAG的直接风险因素，其影响机制更可能通过性别差异相关的变量介导实现，包括但不限于生活方式差异、激素水平波动等多因素的协同作用。另外本研究的结果尚需更大样本量的多中心数据去验证。

饮食和生活方式是影响慢性胃炎特别是CAG进展与演变的重要因素[7]。一项社区调查发现患有慢性疾病的居民中有12.5%每天吸烟，且总体患病率中慢性胃炎处于首位[32]。香烟中有害物质会刺激胃黏膜下的血管收缩，导致胃部血液循环不畅，同时干扰前列腺素的合成，损害胃黏膜屏障，进而增加CAG发生的可能性[33]。我们的研究中将吸烟史、饮茶史、饮食喜好温度及口味、睡眠时间及睡眠质量这些和研究者饮食、生活习惯密切相关的变量纳入到研究中，结果发现吸烟史与CAG可能存在关联，但对模型的结果影响很小。本研究同时发现，饮茶史、饮食喜好温度及口味、睡眠时间及睡眠质量不会明显增加CAG的发病风险。分析其原因可能是由于采用简化的问卷量表丢失关键暴露信息，且部分暴露因素存在剂量效应阈值，而研究中受试者可能未达到该阈值。此外，可能还存在一些混杂因素未被排除，或是保护性因素的补偿机制，但在统计模型中未被有效捕捉。这些非客观因素在CAG发病风险中的作用均需要后续进行纵向队列研究，从动态观察生活习惯改变与胃黏膜病理进展的关系中得到验证。

CatBoost等模型在训练集及测试集中的表现优越，但其可能存在的过拟合倾向及复杂的决策过程使得logistic回归模型成为最佳选择。logistic回归模型因其预测准确性和可解释性而受到青睐，这些特性对于实际临床应用至关重要。构建疾病预测模型的重要性在于能够识别高风险患者，并减少可能落入高风险类别的个体风险，从而整体上惠及患者。因此，机器学习模型的临床可解释性在医学实践中具有重要价值。本研究构建了logistic回归模型及列线图用于预测患者CAG发病风险。临床应用中，医护人员既可通过logistic回归公式计算CAG发病风险，也可基于列线图进行标准化评估：首先对患者各项临床指标进行单项赋分，将各分值累加后定位至总分轴对应刻度，随后垂直向下投影至风险预测轴，即可直观获取患者的CAG发病风险预测值，临床应用方便，便于推广。

本研究尚存在局限性。首先，作为回顾性研究，存在信息收集偏差影响结果的可能性。其次，本研究样本量较小，且样本仅来自单一中心，这可能会限制研究结果的普适性。未来研究预期将纳入前瞻性设计和多中心数据，同时整合更多患者数据并利用先进的机器学习技术，以提升模型的稳定性和普适性，最终开发出更精准的CAG早期预测工具。

本研究通过logistic回归模型，成功开发出一种预测工具，用于预测CAG的发生情况。该模型预测准确率较高，通过识别和评估关键预测因素，能有效预测CAG。本研究方法建立的模型通过进一步优化后推广应用，可将胃癌防治战线前移，科学管理高危人群，从而降低胃癌的发病率。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　韩经略：数据统计分析，论文撰写与修订；颜财旺：研究实施，数据管理；胡非凡、沈耀、高晓娟、赵嘉敏、章华臻、殷洁：数据采集与录入，研究实施；占强：研究项目总体统筹，监督及资金获取；安方梅：资金获取，论文审阅、关键性修改及最终校对；所有作者已阅读并同意最终稿件的提交

参　考　文　献

[1] 中国中西医结合学会消化系统疾病专业委员会，吕宾，王彦刚，等. 慢性萎缩性胃炎中西医结合诊疗专家共识（2025年）[J]. 中国中西医结合消化杂志，2025，33(3)：230-241. DOI: 10.3969/j.issn.1671-038X.2025.03.04.

[2] Song H, Ekheden IG, Zheng Z, et al. Incidence of gastric cancer among patients with gastric precancerous lesions:

observational cohort study in a low risk Western population[J]. BMJ, 2015,351:h3867. DOI: 10.1136/bmj.h3867.

[3] de Vries AC, van Grieken NC, Looman CW, et al. Gastric cancer risk in patients with premalignant gastric lesions: a nationwide cohort study in the Netherlands[J]. Gastroenterology, 2008, 134(4): 945-952. DOI: 10.1053/j.gastro.2008.01.071.

[4] Huang RJ, Choi AY, Truong CD, et al. Diagnosis and management of gastric intestinal metaplasia: current status and future directions[J]. Gut Liver, 2019,13(6):596-603. DOI: 10.5009/gnl19181.

[5] Marques-Silva L, Areia M, Elvas L, et al. Prevalence of gastric precancerous conditions: a systematic review and meta-analysis[J]. Eur J Gastroenterol Hepatol, 2014, 26(4): 378-387. DOI: 10.1097/MEG.0000000000000065.

[6] Du Y, Bai Y, Xie P, et al. Chronic gastritis in China: a national multi-center survey[J]. BMC Gastroenterol, 2014, 14: 21. DOI: 10.1186/1471-230X-14-21.

[7] 中华医学会消化病学分会, 中华医学会消化病学分会消化系统肿瘤协作组. 中国慢性胃炎诊治指南(2022年,上海)[J]. 中华消化杂志, 2023, 43(3): 145-175. DOI: 10.3760/cma.j.cn311367-20230117-00023.

[8] 王珩宇, 陈稳, 陈明锴, 等. 慢性萎缩性胃炎内镜下木村-竹本分型诊断异质性研究[J]. 中华消化内镜杂志,2025,42(4): 307-313. DOI: 10.3760/cma.j.cn321463-20250114-00499.

[9] Forte GC, Altmayer S, Silva RF, et al. Deep learning algorithms for diagnosis of lung cancer: a systematic review and meta-analysis[J]. Cancers (Basel), 2022, 14(16): 3856. DOI: 10.3390/cancers14163856.

[10] Taninaga J, Nishiyama Y, Fujibayashi K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: a case-control study[J]. Sci Rep, 2019, 9(1): 12384. DOI: 10.1038/s41598-019-48769-y.

[11] 王林俊, 沈义凯, 杨昆, 等. 基于术前血清肿瘤标记物的胃癌预后预测模型:一项多中心回顾性研究[J]. 中国实用外科杂志, 2025, 45(1): 95-108. DOI: 10.19538/j. cjps.issn1005-2208.2025.01.16.

[12] Zhao J, Tian W, Zhang X, et al. The diagnostic value of serum trefoil factor 3 and pepsinogen combination in chronic atrophic gastritis: a retrospective study based on a gastric cancer screening cohort in the community population[J]. Biomarkers, 2024, 29(6): 384-392. DOI: 10.1080/1354750X.2024.2400927.

[13] Wang X, Ren J, Ren H, et al. Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta[J]. Sci Rep, 2023,13(1):12718. DOI: 10.1038/s41598-023-40036-5.

[14] Frost HR, Amos CI. Gene set selection via LASSO penalized regression (SLPR)[J]. Nucleic Acids Res, 2017, 45(12): e114. DOI: 10.1093/nar/gkx291.

[15] Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine[J]. Phys Med Biol, 2018, 63(7): 07TR01. DOI: 10.1088/1361-6560/aab4b1.

[16] Jiang C, Xiu Y, Qiao K, et al. Prediction of lymph node metastasis in patients with breast invasive micropapillary carcinoma based on machine learning and SHapley Additive exPlanations framework[J]. Front Oncol, 2022, 12: 981059. DOI: 10.3389/fonc.2022.981059.

[17] Takefuji Y. Beyond XGBoost and SHAP: unveiling true feature importance[J]. J Hazard Mater, 2025, 488: 137382. DOI: 10.1016/j.jhazmat.2025.137382.

[18] Chen Y, Wang B, Zhao Y, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer [J]. Nat Commun, 2024, 15(1): 1657. DOI: 10.1038/s41467-024-46043-y.

[19] 中国抗癌协会胃癌专业委员会, 中国医师协会外科医师分会上消化道外科医师委员会, 中国人群健康风险管理协作组-胃癌专业组. 中国人群胃癌风险管理公众指南(2023版)[J]. 中华医学杂志, 2023, 103(36): 2837-2849. DOI: 10.3760/cma.j.cn112137-20230608-00968.

[20] Fischbach W, Malfertheiner P. Helicobacter pylori infection [J]. Dtsch Arztebl Int, 2018, 115(25): 429-436. DOI: 10.3238/arztebl.2018.0429.

[21] Annibale B, Negrini R, Caruana P, et al. Two-thirds of atrophic body gastritis patients have evidence of Helicobacter pylori infection[J]. Helicobacter, 2001, 6(3): 225-233. DOI: 10.1046/j.1083-4389.2001.00032.x.

[22] Adamu MA, Weck MN, Gao L, et al. Incidence of chronic atrophic gastritis: systematic review and meta-analysis of follow-up studies[J]. Eur J Epidemiol, 2010, 25(7): 439-448. DOI: 10.1007/s10654-010-9482-0.

[23] 杜囚鹏, 曹建彪, 郭汉斌, 等. 胃蛋白酶原亚群测定与萎缩性胃炎相关性研究及效价比分析[J]. 中国药物经济学, 2011(1):64-71. DOI: 10.3969/j.issn.1673-5846.2011.01.008.

[24] Malfertheiner P, Megraud F, O'Morain CA, et al. Management of Helicobacter pylori infection-the Maastricht V/Florence Consensus Report[J]. Gut, 2017, 66(1): 6-30. DOI: 10.1136/gutjnl-2016-312288.

[25] Banks M, Graham D, Jansen M, et al. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma[J]. Gut, 2019, 68(9):1545-1575. DOI: 10.1136/gutjnl-2018-318126.

[26] Genta RM, Turner KO, Sonnenberg A. Demographic and socioeconomic influences on Helicobacter pylori gastritis and its pre-neoplastic lesions amongst US residents[J]. Aliment Pharmacol Ther, 2017, 46(3): 322-330. DOI: 10.1111/apt.14162.

[27] Yaghoobi M, McNabb-Baltar J, Bijarchi R, et al. What is the quantitative risk of gastric cancer in the first-degree relatives of patients? A meta-analysis[J]. World J Gastroenterol, 2017, 23(13):2435-2442. DOI: 10.3748/wjg.v23.i13.2435.

[28] Zhou XZ, Lyu NH, Zhu HY, et al. Large-scale, national, family-based epidemiological study on Helicobacter pylori infection in China: the time to change practice for related disease prevention[J]. Gut, 2023,72(5):855-869. DOI: 10.1136/gutjnl-2022-328965.

[29] Hansford S, Kaurah P, Li-Chang H, et al. Hereditary diffuse gastric cancer syndrome: CDH1 mutations and beyond[J]. JAMA Oncol, 2015, 1(1): 23-32. DOI: 10.1001/jamaoncol.2014.168.

[30] Green J, Roddam A, Pirie K, et al. Reproductive factors and risk of oesophageal and gastric cancer in the Million Women Study cohort[J]. Br J Cancer, 2012, 106(1): 210-216. DOI: 10.1038/bjc.2011.525.

[31] Camargo MC, Goto Y, Zabaleta J, et al. Sex hormones, hormonal interventions, and gastric cancer risk: a meta-analysis[J]. Cancer Epidemiol Biomarkers Prev, 2012, 21(1):20-38. DOI: 10.1158/1055-9965.EPI-11-0834.

[32] 孙宏玉, 孙玉梅, 孙敬怡, 等. 基于智能健康监测系统的社区居民健康状况及影响因素分析[J]. 中华护理杂志,2020, 55(12): 1836-1843. DOI: 10.3761/j. issn. 0254-1769.2020.12.014.

[33] 梁国英,曲智慧,李庆伟. 慢性萎缩性胃炎致病因素的中西医研究进展[J]. 中国中西医结合消化杂志, 2022,30(5): 378-382.