DOI:10.16016/j.2097-0927.202502013

基于5种机器学习算法构建的怀孕概率预测模型及其性能比较

任 超[1,2],杨 桓[1],周妮娅[3,4],陈 卿[1],周文正[5],王 童[1],凌 曦[1],孙 磊[1],邹 鹏[1],梁卓越[1],敖 琳[1],刘晋祎[1],曹 佳[1] [1]陆军军医大学(第三军医大学)军事预防医学系毒理学研究所,重庆;[2]西部战区总医院泌尿外科,四川成都,[3]重庆市妇幼保健院科研外事科,重庆医科大学附属妇女儿童医院;[4]临床与公共卫生研究中心质量控制科,重庆;[5]重庆市妇幼疾病预防控制与公共卫生研究中心,重庆

[**摘要**] **目的** 选择5种机器学习算法构建模型并比较其预测孕前社会-心理-行为因素与夫妻备孕结局关联的性能表现。**方法** 基于重庆市孕前生殖健康与出生结局队列,选取2019年1月至2022年3月年于重庆市妇幼保健院入组的志愿者资料,队列共入组5 447对夫妻,以一对一问卷方式收集夫妻双方人口学和社会-心理-行为资料,共计221个变量。依据纳入排除标准,最终纳入4 097对夫妇,按7∶3简单随机化法划分训练集(*n*=2 867对)和验证集(*n*=1 230对),采用特征分析和共线性筛除选择潜在暴露因素;考虑基层卫生机构不易开展精子质量分析,通过在训练集和验证集中同时纳入或排除精子质量,构建包含精液常规参数的特征集1和排除精液常规参数的特征集2。采用Logistic回归、朴素贝叶斯、随机森林、梯度提升机、支持向量机5种算法构建备孕结局预测模型,并采用随机搜索合并网格搜索法优化模型参数。利用精确率、召回率、F1分数、受试者工作特征曲线下面积(area under the curve,AUC)、校准曲线比较各模型预测能力,并选择最优模型,比较有或无精液常规参数情况下,问卷资料对生育结局预测能力的变化。**结果** 特征集1共筛选得到24个变量,特征集2共筛选得到16个变量。特征集1中,梯度提升机效果较好,AUC相对较高(0.651),且F1分数表现较好(0.61);Logistic回归模型表现稳定(AUC=0.647),适合作为参照模型;随机森林(AUC=0.641)、朴素贝叶斯(AUC=0.641)和支持向量机(AUC=0.634)表现次之。特征集1中,梯度提升机验证集AUC为0.651(95%*CI*:0.629~0.681),预测精度为0.63,召回率为0.65,平均精确度值F1为0.61;特征集2中,其验证集AUC为0.649(95%*CI*:0.624~0.663),与特征集1相比未见明显降低,2个特征集校准曲线均与理想曲线接近。预测结果提示,在特征集1中,与备孕结局呈较高负相关的特征为女性年龄、男性年龄、夫妻无避孕措施1年内未怀孕;呈较高正相关的特征为女性妊娠史、精子总活力、入组前1年使用避孕措施。**结论** 在本队列资料中对比5种机器学习算法,梯度提升机性能略优;夫妻双方共有24项因素与备孕结局关联,排除精液常规参数的简化模型性能未见明显下降。利用机器学习方法,通过社会-心理-行为问卷预测夫妻备孕结局具有可行性。

[**关键词**] 机器学习;怀孕概率;预测模型;社会-心理-行为因素

[**中图法分类号**] R195.1;R319;R714.1 [**文献标志码**] A

[**基金项目**] 国家重点研发计划项目(2022YFC2702902)

[**通信作者**] 曹 佳,E-mail:caojia1962@tmmu.edu.cn

2097-0927/Copyright © 2025,陆军军医大学(第三军医大学)。这篇开放获取文章遵循CC BY许可协议(https://creativecommons.org/licenses/by/4.0/)

引用格式:任超,杨桓,周妮娅,等.基于5种机器学习算法构建的怀孕概率预测模型及其比较[J].陆军军医大学学报,2025,47(12):1376-1387.DOI:10.16016/j.2097-0927.202502013.

Citation:REN C,YANG H,ZHOU N Y,*et al*. Pregnancy probability prediction models based on 5 machine learning algorithms and comparison of their performance[J]. J Army Med Univ,2025,47(12):1376-1387. DOI:10.16016/j.2097-0927.202502013.

Pregnancy probability prediction models based on 5 machine learning algorithms and comparison of their performance

REN Chao[1,2], YANG Huan[1], ZHOU Niya[3,4], CHEN Qing[1], ZHOU Wenzheng[5], WANG Tong[1], LING Xi[1], SUN Lei[1], ZOU Peng[1], LIANG Zhuoyue[1], AO Lin[1], LIU Jinyi[1], CAO Jia[1] [1]Institute of Toxicology, Faculty of Military Preventive Medicine, Army Medical University (Third Military Medical University), Chongqing; [2]Department of Urology, General Hospital of Western Theater Command, Chengdu, Sichuan; [3]Department of Scientific Research Management and Foreign Affairs, Chongqing Health Center for Women and Children, Women and Children's Hospital of Chongqing Medical University, Chongqing; [4]Chongqing Research Center for Prevention and Control of Maternal and Child Diseases and Public Health, Chongqing; [5]Department of Quality Control Center, Clinical and Public Health Research, Chongqing, China

[**Abstract**] Objective To construct 5 machine-learning models and compare their performance in predicting the associations between pre-pregnancy socio-psycho-behavioral exposures of both spouses and preconception outcomes. Methods Based on Chongqing Preconception Reproductive Health and Birth Outcome Cohort of volunteers recruited from Chongqing Health Center for Women and Children during January 2019 and March 2022, 5 447 couples were recruited and surveyed through interviewer-interview for the demographic and social-psychological-behavioral data of both spouses (221 variables). According to the inclusion and exclusion criteria, 4 097 couples were finally included, and randomly assigned into a training set (n=2 867 spouses) and a validation set (n=1 230 spouses) at a ratio of 7∶3. Feature analysis and collinear screening were applied to select the potential exposure factors. In consideration of difficulty to carry out semen parameters analysis in primary healthcare institutions, feature Set 1 including sperm parameters and feature Set 2 excluding semen parameters were constructed by including or excluding sperm quality simultaneously in the training set and the validation set. Five algorithms, that is, Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting Machine, and Support Vector Machine, were used to construct preconception outcome prediction models, and the parameters of each model were optimized using random search combined with grid search. The predictive performance of each model was compared using precision, recall, F1 score, area under the receiver operating characteristic curve (AUC), and calibration curve. The optimal model was then selected by comparing the changes in the predictive ability of the questionnaire data for fertility outcomes with or without semen parameters. Results There were 24 variables screened out in feature Set 1, and 16 variables in feature Set 2. In feature Set 1, the gradient boosting machine performed better, with a relatively higher AUC value (0.651) and better F1 score (0.61). The logistic regression model performed stably (AUC value =0.647) and was suitable as the reference model. The random forest (AUC value=0.641), Naive Bayes (AUC value=0.641), and support vector machine (AUC value=0.634) performed second-best. By utilizing the gradient boosting machine, comparable results were found between the predictions from feature sets with or without semen parameters, as in feature Set 1, the AUC value of its validation set was 0.651 (95%*CI*: 0.629~0.681), the prediction accuracy was 0.63, the recall rate was 0.65, and the average precision value F1 was 0.61; and in feature Set 2, the AUC value of its validation set was 0.649 (95%*CI*: 0.624~0.663), and both the calibration curves were close to the ideal curve. The prediction results indicated that in feature Set 1, the features highly negatively correlated with preconception outcomes were female age, male age, and no pregnancy within 1 year without contraception, while the features highly positively correlated with preconception outcomes were female pregnancy history, total sperm vitality, and use of contraceptive measures before enrollment. Conclusion Among the 5 machine-learning algorithms performed in this cohort data, the gradient boosting machine shows slightly better performance. There are

24 factors being associated with preconception outcomes in both spouses, and the performance of the simplified model excluding semen parameters is not significantly declined. It is feasible to use machine-learning methods to predict human preconception outcomes through social-psychological-behavioral questionnaires.

[**Key words**] machine learning; preconception outcomes; prediction model; social-psychological-behavioral factors

Supported by the National Key Research and Development Program of China (2022YFC2702902). Corresponding author: CAO Jia, E-mail: caojia@tmmu.edu.cn

全球生育率持续下降已成为21世纪人口可持续发展的重大挑战。联合国数据显示[1-3],超半数国家和地区总和生育率跌破人口更替水平(2.1),中国2023年人口自然增长率降至-1.48‰,女性终身不育率较2010年增长近4倍。更加严峻的是,全球约15%育龄夫妇面临不孕症困扰,我国辅助生殖技术年周期数突破百万。然而当前生育力评估面临因素考察不全[4]、医疗资源不均[5]、性别关注失衡[6]等挑战,且多数手段因依赖侵入性检查而出现了高成本、患者低依从性等瓶颈[7]。

现有研究发现,人类生育力主要受遗传、社会-心理-行为等因素影响,例如睡眠[8]、压力[9]和生活方式(吸烟、饮酒等)[10-11]。出生队列以非回顾方式采集暴露因素并随访生殖结局,明确了暴露与结局间的时间先后关系,是生育力评估方法研究中的重要工具之一[12]。但既往出生队列多自怀孕建档时启动构建,缺失了孕前暴露、尤其是男性暴露的信息[13]。本课题组前期构建的重庆市孕前生殖健康与出生结局队列(pre-pregnancy reproductive health and birth outcomes cohort, PREBIC)从备孕期开始建立基线,同步纳入夫妻双方,并持续随访至子代出生或2年仍未孕,是评价人类生育力的理想工具之一[14]。但队列研究中面临的暴露因素众多、测量指标多样、变量异质性大、变量间关系复杂等问题,使传统统计学分析仅能从局部展开关联性分析,缺失了对队列数据的宏观尺度把握[15]。

随着大数据和机器学习技术的快速发展,基于人工智能的健康管理平台正对人类健康管理产生深刻影响。在处理和分析复杂的医疗数据时,机器学习在准确度和规律提取能力上优于传统统计学工具。近年来,机器学习在生殖相关领域应用也逐渐增多,主要应用于妊娠并发症[16]、辅助生殖概率等预测[17]。国外有学者利用女性疾病史、实验室检查(包括精液分析和性交后检查)和妇科体格检查的数据对怀孕概率进行预测,虽未取得理想的预测模型[18-20],但启发了使用复杂信息预测怀孕概率的方法;后有学者使用队列研究中的女性社会人口因素、行为因素、病史和生殖史和少部分男性信息构建了女性妊娠预测模型[13]。而目前尚未见研究报道平等纳入夫妻双方孕前社会-心理-行为暴露以预测备孕成功概率的模型。

本研究基于PREBIC队列,从夫妻双方人口学资料、生活方式、心理状态等多维度资料中筛选变量,比较逻辑回归(logistic regression, LR)、朴素贝叶斯(naive Bayes, NB)、随机森林(random forest, RF)、梯度提升机(gradient boosting machine, GBM)和支持向量机(support vector machine, SVM)5种机器学习算法在本队列数据集的预测能力,并采用较优模型比较是否纳入精液常规参数对备孕结局的预测能力的影响,以评估仅通过社会-心理-行为问卷预测生育力的可行性,为开发基层适用、家庭自检的筛查工具提供理论依据。

1 资料与方法

1.1 研究对象

基于PREBIC队列[14],本研究收集2019年1月至2022年3月于重庆市妇幼保健院入组的夫妻共5 447对,依据纳排标准,最终纳入4 097对夫妻。纳入标准:①年龄20~49岁;②在重庆当地居住满2年,加入本研究前3个月未离开重庆市,且每年离开重庆时间累计不超过3个月;③1年内有生育意愿且接受本研究调查。排除标准:①确诊的泌尿生殖系统疾病史,如明确诊断器质性病变的不孕不育、睾丸/卵巢发育不良综合征、生殖系统外伤,或列于国家免费孕前优生健康检查项目调查表内的泌尿生殖系统炎症或性传播疾病;②青春期后腮腺炎或医学诊断的遗传病史;③参加研究之前3个月内男性有发热史;④进行过绝育手术(包括复通手术);⑤男性禁欲时间<2 d或>7 d者。本研究获得重庆市妇幼保健院[2018伦审(科)

020号]和陆军军医大学(2019第003-02号)伦理委员会审核批准,同时通过中国人类遗传资源管理办公室批(国科遗办审字〔2021〕CJ2417号)。每位参与者在充分了解研究目的、实施方案及潜在风险后,签署了知情同意书。

1.2 数据收集

PREBIC采用标准化电子问卷系统和经过统一培训的调查员,以一对一问卷方式调查人口学信息、疾病史、日常环境暴露因素、社会-心理-行为因素等共221个变量资料(表1),问卷回收率100%。为确保结果的可信性、与国内外研究的可比性,本调查中尽量采用国际通用或已经验证的问卷或量表。通过问卷、重庆医科大学附属妇女儿童医院产科登记信息和重庆市妊娠检查登记系统查询3种方式随访志愿者妊娠信息,确定备孕结局,其中任一一项出现医学诊断的妊娠即确定为该夫妻达成自然怀孕;其中自招募之日起2年内自然怀孕的判定为备孕成功,其余判定为备孕失败,通过辅助生殖达成怀孕的夫妇纳入备孕失败组。

1.3 数据预处理

1.3.1 缺失值处理 删除样本缺失值>20%的变量,对缺失值≤20%的变量,使用众数(针对分类变量)或平均数(针对连续变量)填补缺失值。

1.3.2 变量编码 二分类变量按0、1编码;有序分类变量从0开始进行标签编码;以One-Hot编码方式(哑变量)处理名义分类变量;为消除连续变量之间的量纲差异,按公式$\frac{Xi-\bar{X}}{SD}$标准化处理所有连续变量。

1.3.3 方差过滤 设定0.01、0.10、0.50共3个不同水平的方差阈值[21],采用方差过滤法依次筛选所有自变量,移除方差低于各阈值的特征,比较不同筛选方法找到最佳筛选阈值。对比分析不同阈值下的筛选结果,确定最佳的方差筛选阈值。

1.3.4 特征集构建与特征选择 志愿者按7∶3简单随机分入训练集(n=2 867对)和验证集(n=1 230对),在训练集中进行特征筛选。为研究仅通过问卷预测备孕成功概率的可行性,本研究构建了2个特征集,在特征集1中纳入精液常规参数、特征集2中去除精液常规参数。采用单因素Logistic回归分析对2个特征集中的变量进行初步筛选,剔除与备孕预测结果无显著关联的变量(连续变量和分类变量分别采用Logistic回归模型和χ^2检验计算关联程度,分别以$P\geqslant0.05$和$P\geqslant0.10$作为排除变量的标准);运用特征递归消除法进一步筛选关键特征;为避免引入共线性较强的自变量,使用Lasso回归模型检验上一步骤中得到的自变量,对自变量相关性>0.9的变量组,综合考虑变量的实际意义、统计指标,选择1个更具代表性和重要性的变量替代整组强共线性变量。

1.4 模型构建

采用scikit-learn(v 1.6.1,https://scikit-learn.org/stable/index.html)包,在Python(v 3.1.1)中完成变量筛选和模型构建。在训练集中采用LR、NB、RF、GBM

表1 基线调查中收集的志愿者信息

类型	变量名称
日常环境暴露因素	放射线、高温(>37 ℃)、低温(<0 ℃)、噪音、震动、高原、粉尘、农药、有机物、重金属、新房装修情况、油烟接触情况
人口学信息、社会因素	年龄、青春期启动时间[a]、脱发情况[a]、家庭年收入、文化程度、民族、居住地区[b]、疫苗接种情况(风疹、乙肝、人乳头瘤病毒)、生育史[c]、月经史[c]、妊娠史[c]、过去1年未采取避孕措施未怀孕、近1年使用相关避孕措施(避孕套、避孕药、避孕环、体外射精、皮下埋植剂、月经周期避孕)
行为因素	体质量指数(BMI)、腰臀比、吸烟、被动吸烟、饮酒、饮茶[d]、喝咖啡[d]、辣食、营养和膳食补充剂使用[d](多重维生素、叶酸、钙补充剂)、手机使用、睡眠情况(匹兹堡睡眠质量指数)、运动习惯[d](运动等级量表)、出行交通方式、工作倒班情况、腌制食品、熏制食品、过夜剩菜、大豆及其制品、烧烤、碳酸饮料、含糖软饮料、牛奶、近半年接触放射线(X线、CT、PET-CT)、近半年接触核磁共振
心理因素	工作压力(工作压力量表)、应激感受(知觉压力量表)
疾病史	贫血、高血压、高血脂、心脏病、糖尿病、癫痫、甲状腺疾病、慢性肾炎、结核、淋病、梅毒、衣原体感染、乙型肝炎、精神心理疾患、肿瘤、腹部、盆腔手术
医学测量	食指无名指指长比、男性精液检测[a]:精子量、精子浓度、精子总数、精子总活力、前向运动精子、非前向运动精子、不活动精子

a:仅收集男性相关数据;b:居住地区按照重庆市区划代码后两位赋值;c:仅收集女性相关数据;d:数据预处理时删除

和SVM 5种算法构建预测模型。调整超参数优化模型预测能力,同时引入正则化项微调超参数优化模型预测能力。利用随机搜索合并网格搜索方法确定各模型的最优超参数,并反复调整直至获得最佳结果。

1.5 模型检验和性能指标

使用精确度、召回率、F1分数、受试者工作特征(receiver operating characteristic, ROC)曲线下面积(area under the curve, AUC)评价各模型预测效果。采用5次交叉验证进行模型拟合,并在验证集中进行验证,以评估预测结果。以校准曲线评估模型预测概率与实际发生概率的一致性,采用ROC、AUC评估模型预测区分度。

1.6 统计学分析

使用STATA(Version 17.0, Texas,USA)完成所有变量描述性分析。采用秩和检验和χ^2检验比较训练集和验证集之间的基线特征分布。对于分类变量采用χ^2检验,以例(%)的方式呈现;对符合正态分布的连续变量,采用t检验比较其在备孕成功与失败组之间的分布差异,并使用$\bar{x}\pm s$表示;对不符合正态分布的连续变量,采用Mann-Whitney U检验比较其组间分布差异,并以$M(P_{25},P_{75})$展示,以评估暴露因素与生育结局的关联。

2 结果

2.1 人口学资料

纳入研究的志愿者年龄为女性(28.8±3.8)岁,男性(30.6±4.3)岁。纳入的夫妻中,2 576对(62.9%)在2年内成功怀孕;妊娠组与非妊娠组年龄存在统计学差异(P<0.01),妊娠组女性年龄为(28.2±3.5)岁,男性为(30.1±4.0)岁;非妊娠组女性年龄为(29.8±4.1)岁,男性为(31.6±4.7)岁。妊娠组中,具有大专及以上学历的女性比例更高(84.0% *vs* 77.3%,P<0.01);同时,在入组前避孕措施使用率方面,妊娠组显著高于非妊娠组(41.5% *vs* 33.5%,P<0.01)。

2.2 变量筛选

当定义方差阈值为0.1时,精确度达到最大值,筛得190个变量。以全部190个变量构建特征集1,其中包含3项精液常规参数(精子总数、精子浓度、精子总活力),去除该部分后构建特征集2。经过单因素分析、特征递归消除、Lasso回归特征筛选(特征集1中最优Alpha值为9.3×10^{-4},特征集2中最优Alpha值为5×10^{-5},图1)等步骤后,最终得到夫妻双方入组前1年使用避孕措施情况和无避孕措施1年未生育;女性特征包含年龄、月经是否规律、妊娠史、自然流产史、腹部手术史、工作中常久坐、文化程度、近半年吸烟情况、近半年饮酒情况、每周被动吸烟天数、环境相关的包括是否经常接触0 ℃以下环境和近半年有机物暴露情况;男性特征包含年龄、精液常规参数、倒班情况、压力情况、近半年吸烟及被动吸烟情况、出行方式、环境相关的是放射线检查共24个变量,构建特征集1。特征集2得到16个变量,去除3项精液常规参数后,筛选结果不再包含男性出行方式、压力情况和被动吸烟3项特征,女性特征减少了近半年饮酒和有机物暴露2项特征,所有变量在训练集和验证集组间无统计学差异(表2)。

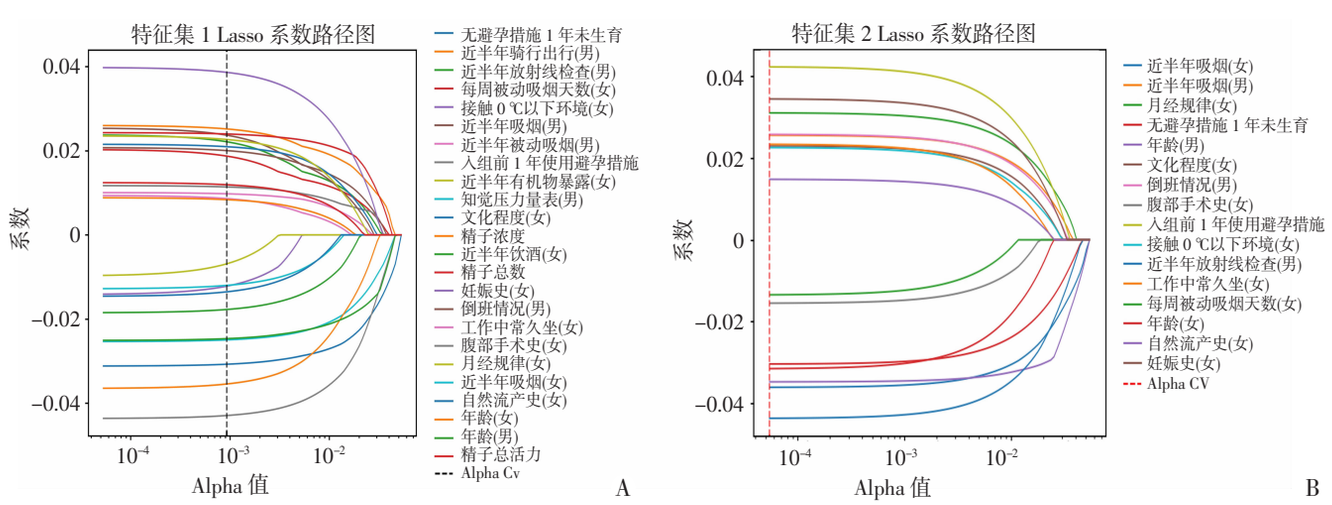


图1 特征集1(A)和特征集1(B)的Lasso系数路径曲线图

表2 特征集1训练集和验证集的特征分布对比[$M(P_{25},P_{75})$,例(%)]

变量	总数据集[a](n=4 097)	训练集[a](n=2 867)	验证集[a](n=1 230)	P
年龄(女)/岁	28(26,31)	28(26,31)	28(26,31)	0.24
年龄(男)/岁	30(28,33)	30(28,33)	30(28,33)	0.37
精子总活力/%[b]	54(51,57)	54(51,57)	55(51,57)	0.81
精子浓度/(10^6/mL)[b]	66(45,86)	66(45,87)	67(45,86)	0.55
精子总数/10^6个[b]	221(144,314)	219(143,311)	224(145,319)	0.33
无避孕措施1年未生育				0.64
否	2 975(72.6)	2 088(72.8)	887(72.1)	
是	1 122(27.4)	779(27.2)	343(27.9)	
入组前1年使用避孕措施				0.36
否	2 518(61.5)	1 749(61.0)	769(62.5)	
是	1 579(38.5)	1 118(39.0)	461(37.5)	
月经规律(女)				0.56
否	998(24.4)	691(24.1)	307(25.0)	
是	3 099(75.6)	2 176(75.9)	923(75.0)	
妊娠史(女)				0.26
否	1 820(44.4)	1 257(43.8)	563(45.8)	
是	2 277(55.6)	1 610(56.2)	667(54.2)	
自然流产史(女)				0.87
否	3 108(75.9)	2 177(75.9)	931(75.7)	
是	989(24.1)	690(24.1)	299(24.3)	
腹部手术史(女)				0.52
否	3 407(83.2)	2 377(82.9)	1 030(83.7)	
是	690(16.8)	490(17.1)	200(16.3)	
经常接触0 ℃以下环境(女)				0.30
否	4 039(98.6)	2 830(98.7)	1 209(98.3)	
是	58(1.4)	37(1.3)	21(1.7)	
工作中常久坐(女)				0.44
否	2 937(71.7)	2 045(71.3)	892(72.5)	
是	1160(28.3)	822(28.7)	338(27.5)	
文化程度(女)				0.42
小学	20(0.5)	13(0.5)	7(0.5)	
初中	244(5.9)	173(6.0)	71(5.8)	
高中	495(12.1)	356(12.4)	139(11.3)	
大专或本科	3 003(73.3)	2 092(73.0)	911(74.1)	
研究生及以上	335(8.2)	233(8.1)	102(8.3)	
近半年吸烟情况(女)				0.35
不吸烟	3 828(93.4)	2 694(94.0)	1 134(92.2)	
吸烟	117(2.9)	74(2.6)	43(3.5)	
已戒烟	152(3.7)	99(3.5)	53(4.3)	
近半年饮酒情况(女)[b]				0.51
不喝	2 791(70.0)	1 951(69.8)	840(70.4)	
偶尔喝	1 125(28.2)	788(28.2)	337(28.3)	
已戒酒	44(1.1)	34(1.2)	10(0.8)	
经常喝	29(0.7)	29(0.8)	6(0.5)	

续表2 特征集1训练集和验证集的特征分布对比[$M(P_{25},P_{75})$,例(%)]

变量	总数据集[a](n=4 097)	训练集[a](n=2 867)	验证集[a](n=1 230)	P
每周被动吸烟天数(吸入烟雾>15 min)(女)				0.93
几乎每天	345(8.4)	233(8.1)	112(9.1)	
每周>3 d	162(4.0)	107(3.7)	55(4.5)	
每周1~3 d	304(7.4)	220(7.7)	84(6.8)	
平均不到1 d	544(13.3)	395(13.8)	149(12.1)	
没有	2 742(66.9)	1 912(66.7)	830(67.5)	
近半年有机物暴露(女)[b]				0.93
从未接触	284(6.9)	196(6.8)	88(7.1)	
偶尔接触(<1次/月)	3 652(89.1)	2 559(89.3)	1 093(88.9)	
接触(≥1次/月)	161(4.0)	112(3.9)	49(4.0)	
倒班情况(男)				0.26
固定白班	3 334(81.4)	2 346(81.8)	988(80.3)	
倒班	763(18.6)	521(18.2)	242(19.7)	
近半年骑行出行(男)[b]				0.96
有	731(17.8)	511(17.8)	220(17.9)	
无	3 366(82.2)	2 356(82.2)	1 010(82.1)	
知觉压力量表(男)[b]				0.71
压力感知低	1 278(31.2)	893(31.2)	385(31.3)	
压力感知适中	1 901(46.4)	1 326(46.4)	575(46.8)	
压力感知较高	883(21.6)	619(21.6)	264(21.5)	
压力感知很高	35(0.9)	29(0.9)	6(0.5)	
近半年接受放射检查(男)				0.65
有	338(8.3)	222(7.7)	116(9.4)	
无	3 759(91.8)	2 645(92.3)	1 114(90.6)	
近半年吸烟情况(男)				0.12
不吸烟	2 122(51.8)	1 508(52.6)	614(49.9)	
吸烟	535(13.1)	380(13.3)	155(12.6)	
已戒烟	1 440(35.2)	979(34.2)	461(37.5)	
近半年被动吸烟情况(男)[b]				0.39
几乎每天接触	268(6.5)	185(6.5)	83(6.8)	
每周>3 d	177(4.3)	126(4.4)	51(4.2)	
每周1~3 d	658(16.1)	474(16.5)	184(15.0)	
没有	2 994(73.1)	2 082(72.6)	912(74.2)	

a:该数据集由成对的数据组成;b:未进入特征集2中的变量

2.3 不同模型预测能力比较

使用特征集1预测备孕成功的概率,结果显示:5种机器学习算法的AUC值为0.634~0.674;其中在训练集中,RF的AUC最高(0.674),SVM最低(0.639),5种模型在训练集上的区分能力相对接近。特征集1中,验证集AUC值为0.634~0.651,GBM的AUC最高(0.651),SVM最低(0.634),验证集的AUC普遍低于训练集,但下降幅度不大,说明模型经过参数调整后泛化能力较好;GBM在验证集上的准确率表现相对更优,在训练集和验证集上的F1分数表现较为稳定,且在精确率和召回率之间取得了较好的平衡,表明其对分布未知数据有较强的分类能力;SVM在训练集上的精确率和召回率较高,但在验证集上表现下降,可能存在过拟合(表3)。特征集2分析显示:验证集中GBM模型的ROC AUC为0.649(95%*CI*:0.624~0.663),AUC值相较特征集1无明显变化(图2E),预测精度为0.63、召回率为0.65,平均精确度值F1为0.62(表3)。图2以2个特征集各自的AUC曲线直观

地展示了预测效果,并使用校准曲线来判断模型的可靠程度,验证集ROC曲线显示,GBM的AUC值最高(0.651),LR的AUC(0.647)次之,SVM的AUC值相对较低(0.634),但所有模型的总体预测水平差距不大且都有待提高(图2B);通过对比校准曲线发现,随迭代次数增加,GBM预测结果与理想情况趋于一致(图2)。

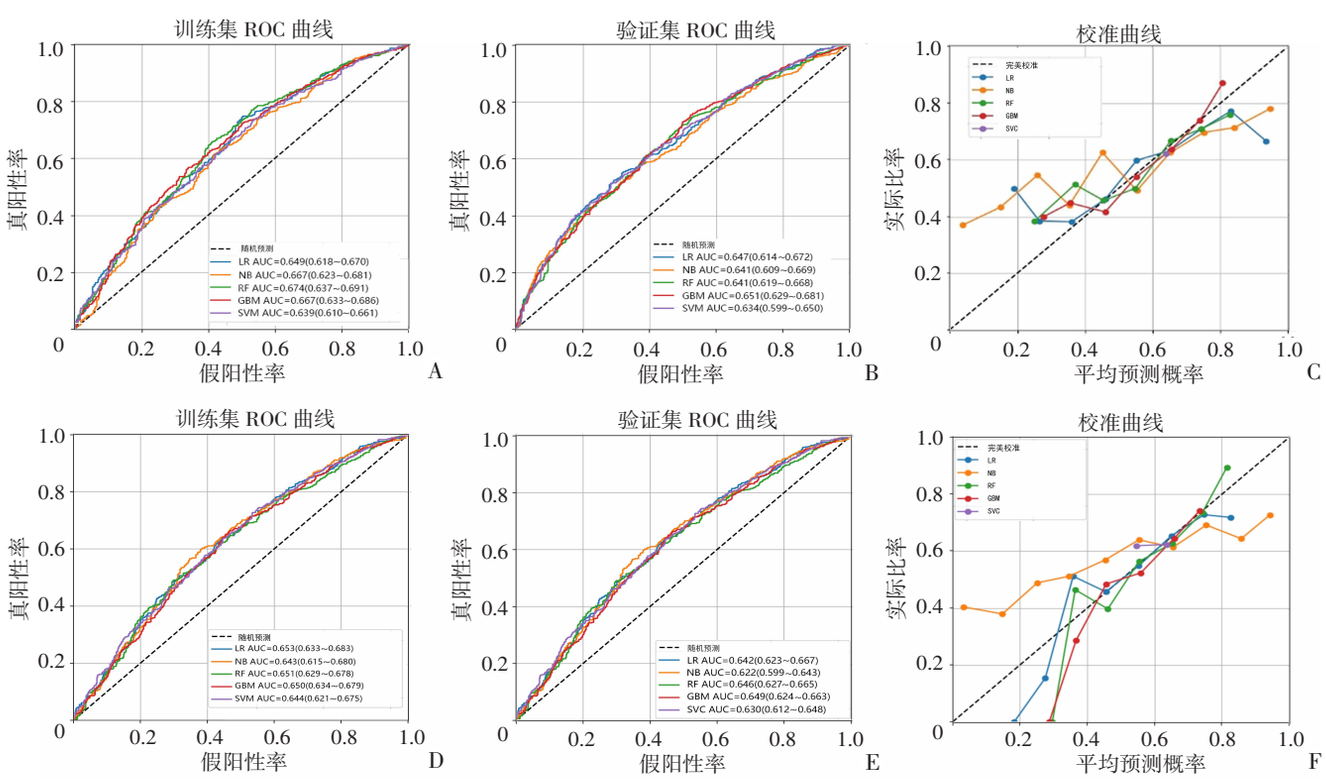
2.4　预测得到的影响因素

使用GBM进一步分析特征集1的暴露因素,发现筛得的24个变量对结局的贡献度有较大差异(图3A),

表3　模型预测性能比较

模型	AUC(95%*CI*)		精确度		召回率		F1分数[a]	
	训练集	验证集	训练集	验证集	训练集	验证集	训练集	验证集
特征集1								
LR	0.649(0.618~0.670)	0.647(0.614~0.672)	0.63	0.62	0.65	0.65	0.60	0.60
NB	0.667(0.623~0.681)	0.641(0.609~0.669)	0.65	0.62	0.67	0.64	0.65	0.62
RF	0.674(0.637~0.691)	0.641(0.619~0.668)	0.75	0.63	0.67	0.64	0.59	0.54
GBM	0.667(0.633~0.686)	0.651(0.629~0.681)	0.76	0.63	0.67	0.65	0.73	0.61
SVM	0.639(0.610~0.661)	0.634(0.599~0.650)	0.78	0.61	0.76	0.63	0.73	0.58
特征集2								
LR	0.653(0.633~0.683)	0.642(0.623~0.667)	0.64	0.64	0.66	0.66	0.61	0.62
NB	0.643(0.615~0.680)	0.622(0.599~0.643)	0.64	0.62	0.65	0.63	0.64	0.62
RF	0.651(0.629~0.678)	0.646(0.627~0.665)	0.71	0.63	0.68	0.65	0.62	0.58
GBM	0.650(0.634~0.679)	0.649(0.624~0.663)	0.73	0.63	0.72	0.65	0.69	0.62
SVM	0.644(0.621~0.675)	0.630(0.612~0.648)	0.66	0.61	0.65	0.63	0.65	0.61

a:F1为精确度和召回率的调和平均值



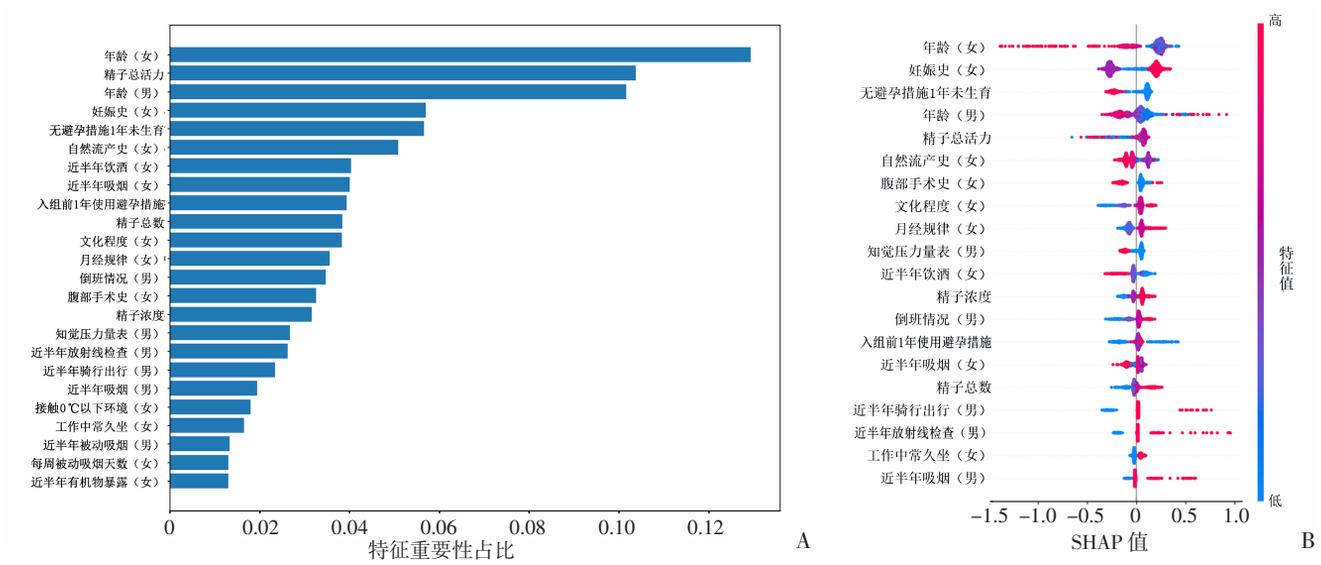
A、B:分别为特征集1的训练集和验证集ROC曲线;C:特征集1验证集的校准曲线;D、E:分别为特征集2训练集和验证集的ROC曲线;F:特征集2验证集的校准曲线

图2　5种机器学习相关特征曲线

在此基础上,使用沙普利加性解释(Shapley additive explanation, SHAP)图,综合考虑特征间的交互作用,对各特征的重要性排序发现,与备孕结局相关的前3个变量依次为女性年龄(负相关)、妊娠史(正相关)和无避孕措施1年未生育(负相关),其余重要特征贡献见图3B。在特征集2的16个变量中采用相同方法分析发现,与备孕结局相关的前3个变量为妊娠史(正相关)、入组前1年使用避孕措施(正相关)和女性年龄(负相关),其余重要特征贡献见图4B。在去除特征集1中的3个精液常规参数后,不但筛选得到的特征数量发生了改变,特征重要性的排序也出现了变化。

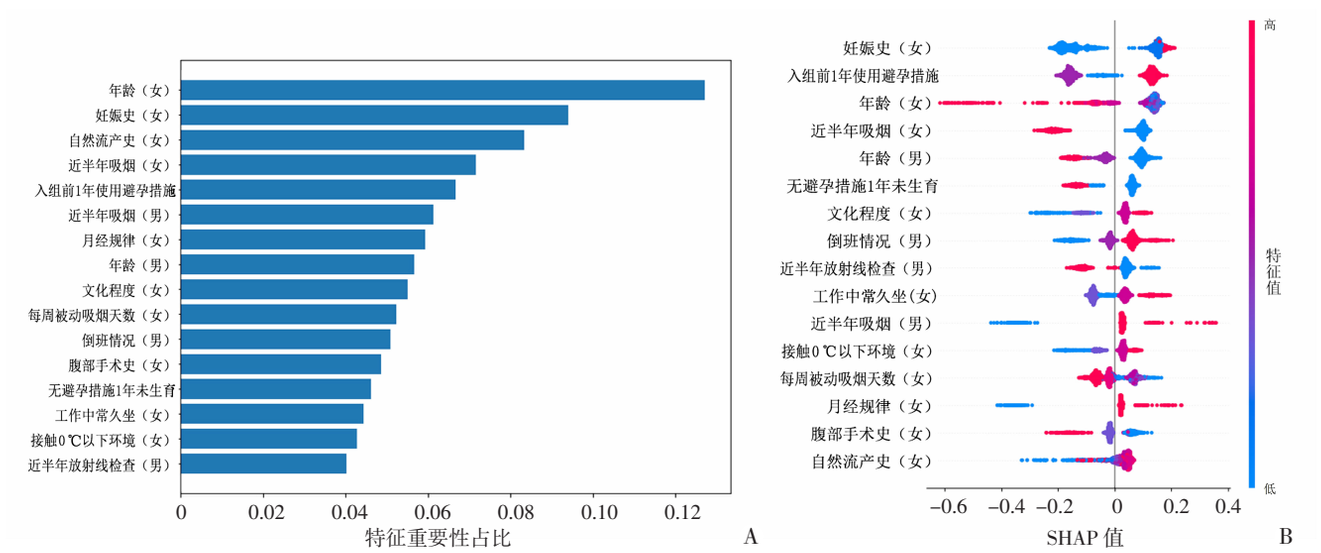
3 讨论

本研究基于PREBIC分析4 097对计划怀孕夫妇的数据,以5种监督学习方法,构建了针对育龄期且有生育意愿夫妇的备孕结局预测模型。通过比较5种方法发现,GBM在本研究数据集中具有最高预测准确度,且预测所得关联因素(例如女性年龄、女性妊娠



A:特征重要性直方图,反映特征在模型中的平均贡献;B:SHAP全局解释图,反映计算每个特征的局部贡献聚合后的全局重要性排序

图3 GBM预测模型的直方图与SHAP图(特征集1)



A:特征重要性直方图,反映特征在模型中的平均贡献;B:SHAP全局解释图,反映计算每个特征的局部贡献聚合后的全局重要性排序

图4 GBM预测模型的直方图与SHAP图(特征集2)

史、男性精子总活力等)与领域内的当前认识相符。其中,入组前使用避孕措施是一项既往关注较少的行为学因素,现认为,此类夫妇通常在生育规划上更为主动和理性,因此减少了意外怀孕带来的心理压力;同时研究认为过去使用避孕措施并不会损害生育力[22]。同时,北美孕前队列(pregnancy study online,PRESTO)使用了Logistic回归模型的AUC值作为怀孕结局预测指标,其精度为0.70[13],本研究的预测水平与其相当(AUC=0.651)。综合来看,夫妻相对年轻、有孕产经验但无自然流产史、对备孕有合理的规划的夫妻备孕成功率较高。上述结果表明,在缺乏更详细医学信息的情况下,利用自我报告数据,开发具有较高判别能力的预测模型具有可行性。

3.1 机器学习方法的优势

在公共卫生领域,人工智能拥有广泛的应用前景,计算机分析通过自学习过程,利用特征和结局间的关联模式,建立疾病分布模式、因素关联的精准预测,机器学习分析未来可能会媲美甚至赶超人工分析。机器学习算法不仅能够处理和分析复杂医疗数据,还能提供个性化的健康管理建议。近年机器学习在生殖相关领域应用不断增多,例如应用于妊娠并发症[23]、辅助生殖概率预测[19]等,国外有部分学者针对女性人口学特征和疾病史等因素预测怀孕概率,但均未得到较为理想的预测准确度[17-19],故建议利用校准精度评判模型。国内外目前暂无通过夫妻双方社会-心理-行为因素对备孕结局的预测模型,而本研究通过孕前夫妇社会-心理-行为因素暴露情况构建的备孕结局GBM预测模型,为自然备孕人群预测提供了有价值的参考。作为家庭生育计划的首要步骤,评估生育潜力对我国生育政策管理具有重要的现实意义。

3.2 GBM的应用场景

GBM基于决策树逐步构建,可以捕捉复杂的非线性关系、处理分类数据和数值型混合数据,与队列资料中暴露因素混合多样的特征高度吻合。在本研究中,因暴露因素的确定来源于志愿者自报告问卷,数据异质性大、分布差异显著,GBM有较高的鲁棒性,在处理此类数据中表现出较高的稳定性[24];同时本研究通过引入正则化项,可有效防止过拟合,提高模型泛化能力。因此相较其他模型,GBM取得了更好的预测结果[25]。例如有研究表明,在慢性病风险预测[26]和ICU患者再入院预测[27]方面,GBM基于多项临床检查资料,呈现出最佳预测效能,ROC、AUC分别达到0.850 1和0.772,优于Logistic回归等传统方法。进一步证实了在特定的场景、尤其在复杂高维数据中,GBM相对具有更高的预测性能和稳定性[28]。

3.3 问卷资料预测生育力的可行性

通过在特征集2中人为排除3项男性精液常规参数,发现GBM在2个特征集中的预测准确度和危险因素发现能力方面均相当,表明仅使用社会-心理-行为因素预测生育力具有一定的可行性。值得注意的是,当排除精液常规参数后,女性妊娠史、入组前1年使用避孕措施及夫妻双方吸烟等特征的重要性出现明显提升,这一现象可能源于上述变量与精液常规检查存在潜在关联。既往研究发现[29],避孕措施可间接反映男性精子质量;而主动、被动吸烟已被证实会显著降低精液参数[30],这或许也支持了本研究通过生活方式评估生育力的科学假设。机器学习模型通过捕捉变量间的高阶交互作用,能够更全面地解析此类复杂关联,从而与传统统计学模型的单因素分析形成方法学差异[31]。

3.4 本研究的不足

本研究存在以下局限性:①生育力影响因素十分复杂,而现有问卷调查可能遗漏了部分影响生育结局的重要信息,降低了模型的预测准确度,环境暴露仅通过问卷也无法反映具体的暴露情况。例如PRESTO纳入了膳食问卷(母乳喂养史和多种维生素叶酸补充筛选为关键预测因子)和尝试怀孕时间[13],得到了更好的预测性能。但在本研究中,叶酸补充剂等一些暴露因素因缺失值较多而未纳入,母乳喂养史未在问卷中涉及。未来可通过优化问卷、增大样本量的方式进一步提高预测能力。②本研究虽然进行了内部验证和机器学习相关超参数的调整,但由于算力限制,可能尚未达到最佳参数设置,未来可进一步优化模型或使用深度学习等方法提高预测能力。③本研究为单中心研究,样本量有限且缺乏外部验证,研究结果需要在其他人群进行验证以提高模型泛化性。

综上所述,使用社会-心理-行为问卷方式,利用机器学习算法预测夫妻备孕结局,已经能够达到与国外相关报道基于医学检查数据相当的预测准确度。尤其是在大数据和人工智能高速发展的背景下,综合应用此类无医学检查的信息,是医学、尤其是公共卫生领域的新兴应用方向之一。随数据量增加、调查因素优化和算法发展,其预测性能还将进一步提升,必将更加准确地预测包括备孕结局在内的多种健康结局。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] MATSUURA H. Further acceleration in fertility decline in 2023: deviation of recently published provisional fertility estimates in selected OECD countries from those in the 2022 revision of the world population prospects[J]. Biodemography Soc Biol, 2024, 69(2): 55-56. DOI: 10.1080/19485565.2024.2367861.

[2] 郭瑛,陈卿,王童,等.育龄男性空气污染物暴露水平与不育风险关联:重庆市PREBIC队列研究[J].陆军军医大学学报,2023,45(9):893-901.
GUO Y, CHEN Q, WANG T, *et al*. Association between exposure to air pollutants and risk of infertility in men of childbearing age: a PREBIC cohort study in Chongqing[J].J Amry Med Univ,2023,45(9):893-901.

[3] 张翠玲,姜玉,庄亚儿,等.中国女性终身不育水平估计:基于第七次全国人口普查数据的分析[J].人口研究,2023,47(3):78-93.
ZHANG C L, JIANG Y, ZHUANG Y E, *et al*. An assessment of lifetime childlessness in China based on the 7th population census[J]. Popul Res, 2023, 47(3): 78-93.

[4] PEDERSEN E S, MIKKELSEN E M, SØRENSEN H T, *et al*. The association between the use of fertility indicators and fecundability in a Danish preconception cohort[J]. Paediatr Perinat Epidemiol, 2024, 38(8): 641-650. DOI: 10.1111/ppe.13108.

[5] MA Y B, HU C Y, CAI G Q, *et al*. Associations of exposure to ambient fine particulate matter constituents from different pollution sources with semen quality: Evidence from a prospective cohort[J]. Environ Pollut, 2024, 343: 123200. DOI: 10.1016/j.envpol.2023.123200.

[6] WIECZOREK K, SZCZĘSNA D, RADWAN M, *et al*. Exposure to air pollution and ovarian reserve parameters[J]. Sci Rep, 2024, 14(1): 461. DOI: 10.1038/s41598-023-50753-6.

[7] 王超群.中国普惠性生育保障制度设计及其筹资需求[J].社会保障评论,2024,8(2):135-147.
WANG C Q. Designing China's inclusive maternity protection system and assessing its funding needs[J]. Soc Secur Rev,2024,8(2):135-147.

[8] YAO Q Y, YUAN X Q, LIU C, *et al*. Associations of sleep characteristics with outcomes of IVF/ICSI treatment: a prospective cohort study[J]. Hum Reprod, 2022, 37(6): 1297-1310. DOI: 10.1093/humrep/deac040.

[9] ZOU P, WANG X G, SUN L, *et al*. *Semen* quality in Chinese college students: associations with depression and physical activity in a cross-sectional study[J]. Psychosom Med, 2018, 80(6): 564-572. DOI: 10.1097/PSY.0000000000000595.

[10] 李蕴宁,刘瑗,强荣,等.1920例遗传咨询者的细胞遗传学研究和分析[J].空军军医大学学报,2024,45(6):666-669. DOI:10.13276/j.issn.2097-1656.2024.06.011.
LI Y N, LIU Y, QIANG R, *et al*. Cytogenetic study and analysis of 1920 genetic consultants[J]. J Air Force Med Univ, 2024, 45(6): 666-669. DOI: 10.13276/j.issn.2097-1656.2024.06.011.

[11] MIKKELSEN E M, RIIS A H, WISE L A, *et al*. Alcohol consumption and fecundability: prospective Danish cohort study[J]. BMJ, 2016, 354: i4262. DOI: 10.1136/bmj.i4262.

[12] MERRILL M L, CIRILLO P M, TERRY M B, *et al*. Prenatal exposure to the pesticide DDT and hypertension diagnosed in women before age 50: a longitudinal birth cohort study[J]. Environ Health Perspect, 2013, 121(5): 594-599. DOI: 10.1289/ehp.1205921.

[13] YLAND J J, WANG T Y, ZAD Z, *et al*. Predictive models of pregnancy based on data from a preconception cohort study[J]. Hum Reprod, 2022, 37(3): 565-576. DOI: 10.1093/humrep/deab280.

[14] 陈卿,周文正,周妮娅,等.重庆市孕前生殖健康与出生结局队列:设计简介与建设概况[J].中华流行病学杂志,2022,43(7):1134-1139. DOI:10.3760/cma,j,cn112338-20220219-00134.
CHEN Q, ZHOU W Z, ZHOU N Y, *et al*. Preconception reproductive health and birth outcomes cohort in Chongqing: the cohort profile[J]. Clin J Epidemiol, 2022, 43(7): 1134-1139. DOI: 10.3760/cma, j, cn112338-20220219-00134.

[15] FORTIER I, WEY T W, BERGERON J, *et al*. Life course of retrospective harmonization initiatives: key elements to consider[J]. J Dev Orig Health Dis, 2023, 14(2): 190-198. DOI: 10.1017/S2040174422000460.

[16] JHEE J H, LEE S, PARK Y, *et al*. Prediction model development of late-onset preeclampsia using machine learning-based methods[J]. PLoS One, 2019, 14(8): e0221202. DOI: 10.1371/journal.pone.0221202.

[17] CEDARS M I. Artificial intelligence in assisted reproductive technology: how best to optimize this tool of the future[J]. Fertil Steril, 2023, 120(1): 1-2. DOI: 10.1016/j.fertnstert.2023.05.150.

[18] VAN DER STEEG J W, STEURES P, EIJKEMANS M C, *et al*. Pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile couples[J]. Hum Reprod, 2007, 22(2): 536-542. DOI: 10.1093/humrep/del378.

[19] EIMERS J M, TE VELDE E R, GERRITSE R, *et al*. The

prediction of the chance to conceive in subfertile couples [J]. Fertil Steril, 1994, 61(1): 44-52. DOI: 10.1016/s0015-0282(16)56451-6.

[20] COPPUS S J, VAN DER VEEN F, OPMEER B C, *et al*. Evaluating prediction models in reproductive medicine[J]. Hum Reprod, 2009, 24(8): 1774-1778. DOI: 10.1093/humrep/dep109.

[21] LIU J J, FENG H C, TANG Y F, *et al*. A novel hybrid algorithm based on Harris Hawks for tumor feature gene selection[J]. Peer J Comput Sci, 2023, 9: e1229. DOI: 10.7717/peerj-cs.1229.

[22] DAMTIE Y, KEFALE B, AREFAYNIE M, *et al*. Fertility return after hormonal contraceptive discontinuation and associated factors among women attended Family Guidance Association of Ethiopia Dessie model clinic, Northeast Ethiopia: a cross-sectional study[J]. PLoS One, 2023, 18(7): e0287440. DOI: 10.1371/journal.pone.0287440.

[23] MCLERNON D J, LEE A J, MAHESHWARI A, *et al*. Predicting the chances of having a baby with or without treatment at different time points in couples with unexplained subfertility[J]. Hum Reprod, 2019, 34(6): 1126-1138. DOI: 10.1093/humrep/dez049.

[24] LI F, XIE H W, SU S B, *et al*. Investigation of critical factors influencing the underestimation of hearing loss predicted by the ISO 1999 predicting model[J]. BMC Public Health, 2023, 23(1): 2239. DOI: 10.1186/s12889-023-17138-w.

[25] TALISMAN S, GUEDALIA J, FARKASH R, *et al*. NICU admission for term neonates in a large single-center population: a comprehensive assessment of risk factors using a tandem analysis approach[J]. J Clin Med, 2022, 11(15): 4258. DOI: 10.3390/jcm11154258.

[26] YANG J M, JU X L, LIU F, *et al*. Prediction for the risk of multiple chronic conditions among working population in the United States with machine learning models[J]. IEEE Open J Eng Med Biol, 2021, 2: 291-298.

[27] HEGSELMANN S, ERTMER C, VOLKERT T, *et al*. Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines[J]. Front Med (Lausanne), 2022, 9: 960296. DOI: 10.3389/fmed.2022.960296.

[28] CHENG S Y, ZHANG Q H, MIN H, *et al*. Development of a predictive model for N-dealkylation of amine contaminants based on machine learning methods[J]. Toxics, 2024, 12(12): 931. DOI: 10.3390/toxics12120931.

[29] AMORY J K. Standardized outcome measures for reporting results from male contraceptive efficacy trials [J]. Andrology, 2025. DOI: 10.1111/andr.70042.

[30] ZHAO J Q, WANG X B, LENG X, *et al*. Dietary fat and fatty acid consumptions and the odds of asthenozoospermia: a case-control study in China[J]. Hum Reprod Open, 2023, 2023(3): hoad030. DOI: 10.1093/hropen/hoad030.

[31] ONTIVERO-ORTEGA M, FAES L, CORTES J M, *et al*. Assessing high-order effects in feature importance *via* predictability decomposition[J]. Phys Rev E, 2025, 111(3): L033301. DOI: 10.1103/PhysRevE.111.L033301.

(**收稿**:2025-02-06;**修回**:2025-05-17)

(编辑 张 维)