·技术与方法· doi:10.3969/j.issn.1671-8348.2023.13.020
网络首发 https://kns.cnki.net/kcms/detail/50.1097.R.20230403.0959.002.html(2023-04-03)

基于 KNN 算法与 logistic 回归的代谢综合征风险预测模型构建与对比研究*

张 慧[1],陈丹丹[2],邵 静[3],汤磊雯[2],吴静洁[2],薛二旭[2],叶志弘[2△]
(1.贵州省人民医院心内科,贵阳 550002;2.浙江大学医学院邵逸夫医院护理部,杭州 310016;
3.浙江大学医学院护理系,杭州 310012)

[摘要] **目的** 构建基于 K 最近邻(KNN)算法和 logistic 回归的代谢综合征预测模型并比较两种模型对代谢综合征的预测效能。**方法** 纳入 6 793 例研究对象进行数据分析,构建基于 KNN 算法和 logistic 回归的预测模型,对模型进行内部验证及外部验证,采用多维度指标对预测性能进行评估,对比两种预测模型的预测效能。**结果** 基于 KNN 算法预测模型的内部验证曲线下面积(AUC)为 0.776(95%*CI*:0.764~0.788)、校准截距为 0.028(95%*CI*:-0.031~0.089)、校准斜率为 1.181(95%*CI*:1.106~1.257)、布里尔分数为 0.157;外部验证 AUC 为 0.780(95%*CI*:0.768~0.791)、校准截距为 0.262(95%*CI*:0.207~0.317)、校准斜率为 1.053(95%*CI*:0.990~1.117)、布里尔分数为 0.167。基于 logistic 回归预测模型内部验证 AUC 为 0.783(95%*CI*:0.772~0.795)、校准截距为-0.008(95%*CI*:-0.088~0.073)、校准斜率为 0.995(95%*CI*:0.934~1.058)、布里尔分数为 0.156;外部验证 AUC 为 0.782(95%*CI*:0.771~0.793)、校准截距为-0.045(95%*CI*:-0.113~0.022)、校准斜率为 1.006(95%*CI*:-0.011~1.063)、布里尔分数为 0.164。**结论** 在代谢综合征的风险预测上,logistic 回归预测模型表现优于基于 KNN 算法预测模型。

[关键词] 代谢综合征;预测模型;K 最近邻算法;机器学习;logistic 回归
[中图法分类号] R589 [文献标识码] A [文章编号] 1671-8348(2023)13-2019-05

Construction and comparative study of metabolic syndrome risk prediction models based on KNN algorithm and logistic regression*

ZHANG Hui[1],*CHEN Dandan*[2],*SHAO Jing*[3],*TANG Leiwen*[2],*WU Jingjie*[2],*XUE Erxu*[2],*YE Zhihong*[2△]
(1.*Department of Cardiology,Guizhou Provincial People's Hospital,Guiyang,Guizhou* 550002,*China*;2.*Department of Nursing,Run Run Shaw Hospital,School of Medicine,Zhejiang University,Hangzhou,Zhejiang* 310016,*China*;3.*Department of Nursing,School of Medicine,Zhejiang University,Hangzhou,Zhejiang* 310012,*China*)

[**Abstract**] **Objective** To construct the metabolic syndrome prediction models based on the K nearest neighbor(KNN) algorithm and logistic regression,and to compare the predictive efficiencies between the two methods. **Methods** The included 6 793 study subjects conducted the data analysis. The prediction models based on KNN algorithm and logistic regression were constructed. The models conducted the internal validation and external validation. The multiple dimensions indicators were adopted to evaluate their predictive performances and the predictive efficiencies were compared between the two predictive models. **Results** The area under internal validation curve (AUC) of the prediction model based on the KNN algorithm was 0.776 (95%*CI*:0.764-0.788),the calibration intercept was 0.028 (95%*CI*:-0.031-0.089),the calibration slop was 1.181 (95%*CI*:1.106-1.257) and the Brier score was 0.157. In the external validation,AUC was 0.780 (95%*CI*:0.768-0.791),the calibration intercept was 0.262 (95%*CI*:0.207-0.317),the calibration slop was 1.053 (95%*CI*:0.990-1.117) and the Brier score was 0.167. AUC of internal validation in the prediction model based on the logistic regression was 0.783 (95%*CI*:0.772-0.795),the calibration intercept was -0.008 (95%*CI*:-0.088-0.073),the calibration slop was 0.995 (95%*CI*:0.934-1.058),the Brier score was 0.156,the external validation AUC was 0.782 (95%*CI*:0.771-0.793),the calibration intercept was -0.045 (95%*CI*:-0.113-0.022),the calibration slope was 1.006 (95%*CI*:-0.011-1.063) and the Brier

* **基金项目**:贵州省人民医院人才项目(2022-18);浙江省医药卫生重大科技计划(WKJ-ZJ-1925)。 **作者简介**:张慧(1984-),主管护师,博士,主要从事慢性病管理方面的研究。 △ **通信作者**,E-mail:3192005@zju.edu.cn。

score was 0.164. **Conclusion** The logistic regression prediction model performance is better than the prediction model based on the KNN algorithm.

[**Key words**] metabolic syndrome; prediction model; KNN; machine learning; logistic regression

代谢综合征是一种由于多种代谢性疾病聚集出现的代谢紊乱症候群，其主要表现包括腹部肥胖、空腹血糖水平过高、高血压、甘油三酯水平过高、高密度脂蛋白胆固醇水平过低等[1]。研究发现，代谢综合征会增加 2 型糖尿病及心血管不良事件的发生风险，威胁个体的健康，加重医疗经济负担[1-2]。基于“健康中国 2030”提出的疾病防控“关口前移”思想，代谢综合征的防控措施应着重于建立精准的临床决策系统，从而开展风险预警、筛查与诊断，实现早期干预及治疗。

本课题组前期依据个体预后或诊断的多变量预测模型透明报告（transparent reporting of a multivariable prediction model for individual prognosis or diagnosis，TRIPOD），进行了代谢综合征风险预测模型的系统评价，筛选出一系列代谢综合征潜在风险预测因子[3-4]。K 最近邻（K-nearest neighbors，KNN）属于机器学习中理论较为成熟的一种数据挖掘分类及模式识别技术，其特点为懒惰学习（即不进行模型数据训练），支持特征高维度计算[5]。logistic 回归作为一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域，通用性较高。因此，本研究拟基于体检队列人群数据，采用 KNN 算法和 logistic 回归分别构建代谢综合征预测模型，并比较两者的预测性能，以期为代谢综合征风险早期预警与干预体系的深入推进与纵深发展提供重要依据。

1 资料与方法

1.1 一般资料

选取浙江大学医学院邵逸夫医院 2011—2014 年患者资料作为预测模型构建与内验证数据源；选取 2015—2018 年患者资料作为预测模型时段外验证数据源。纳入标准：(1)研究对象年龄≥18 岁；(2)随访初始阶段未被诊断为代谢综合征；(3)连续 4 年均参加健康检查。排除标准：孕产妇。本研究已通过医院伦理委员会批准（批准号：20181220-3）。

1.2 方法

本研究为回顾性队列研究。分别对构建的代谢综合征风险预测模型进行内部验证及外部验证，从而对预测模型的可复制性及可泛化性进行评价。内部验证方法采用 Bootstrap 方法（100 次重抽样），以期对预测模型的表现进行高估校正，避免过度拟合。外部验证采用时段验证法（temporal validation），使用独立数据集进行模型预测性能评估。

1.2.1 代谢综合征诊断标准

本研究将代谢综合征定为临床结局。代谢综合征诊断标准基于国际糖尿病联盟流行病预防工作组联合过渡声明 2009[6]，以下情况中满足 3 项者可诊断为代谢综合征：(1)男性腰围≥85 cm，女性腰围≥80 cm；(2)甘油三酯≥1.7 mmol/L 或者已经接受治疗；(3)男性高密度脂蛋白＜1.0 mmol/L、女性高密度脂蛋白＜1.3 mmol/L 或者已经接受治疗；(4)收缩压≥130 mmHg（1 mmHg＝0.133 kPa）、舒张压≥85 mmHg 或者已经接受治疗；(5)空腹血糖≥5.6 mmol/L 或者已经接受治疗。当研究对象在随访的 4 年内发生代谢综合征，即判断出现临床结局事件。

1.2.2 KNN 算法

KNN 算法属于一种基本分类和回归方法。给定一个训练数据并进行集中，将新的测试实例数据特征与训练数据集中对应的特征进行比较，找到训练数据集中与新的测试实例数据最邻近的 K 个实例数据，则该测试实例数据对应的类别就是 K 个数据中出现次数最多的类别[5]。在 KNN 中，通过欧式距离函数计算各个数据之间距离来作为非相似性指标，避免对象之间的匹配问题。

$$\text{欧式距离函数：} d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \quad (1)$$

1.2.3 logistic 回归

采用多因素 logistic 回归构建代谢综合征风险预测模型，根据既往数据构建的 logistic 回归方程如下[4]：$P=1/\{1+\exp-[-12.486+0.048\times$年龄$+0.354\times$总胆固醇（TC）$+0.003\times$尿酸（UA）$+0.937\times$(log)谷丙转氨酶（ALT）$+0.243\times$体重指数（BMI）$]\}$ (2)

1.2.4 样本量

根据 TRIPOD，每个自变量的事件数（events per variable，EPV）是衡量样本量的重要指标，采用机器学习算法构建风险预测模型，EPV 应该大于 200[7]。本研究中，KNN 算法纳入 5 个预测因子构建预测模型，因此发生代谢综合征的研究对象应不低于 1 000 例。

1.2.5 潜在预测因子

根据 TRIPOD，预测模型的潜在预测因子的选择应该基于相关预测模型系统评价的筛选结果[7]。本课题组前期开展的代谢综合征风险预测模型系统评价共整理出 25 个潜在风险预测因子[3]，最终通过惩罚回归系数法筛选出 5 个预测因子（年龄、TC、UA、ALT、BMI）构建代谢综合征风险预测模型[4]。本研究将基于上述 5 个预测因子构建与验证两种代谢综合征风险预测模型。

1.3 统计学处理

针对模型的预测性能表现，采用多维度指标进行评估，主要分为区分度、校准度、综合表现[8-9]。区分

度采用受试者工作特征(receiver operation characteristic,ROC)曲线下面积(area under the curve,AUC)进行预测模型的评估。AUC 为 0.7～<0.8,提示模型区分度可接受;AUC 为 0.8～<0.9,提示模型区分度好;AUC≥0.9,提示模型区分度理想。校准度采用校准图、校准截距、校准斜率进行评估,该类指标可以评估模型预测的绝对风险值是否正确。校准截距越接近 0、校准斜率越接近 1,提示模型预测效能越高。采用布里尔分数对风险预测模型的区分度和校准度的整体表现进行评价。布里尔分数评分临界值为 0.25,该数值越小表示模型表现越好。针对数据集中出现的缺失值,采用了多重插补的方法进行处理[7]。当变量缺失超过 50%时,该变量会被舍弃且不进行插补,避免该类数据经多重插补后出现偏误。采用 RV3.6.2 语言软件进行构建和计算。在对不同预测模型进行对比时,基于既往研究者在不同预测模型对比的 meta 分析和实证分析中所采用的策略,对不同预测模型的不同表现指标进行直接比较。

2 结　果

2.1 研究对象基线特征

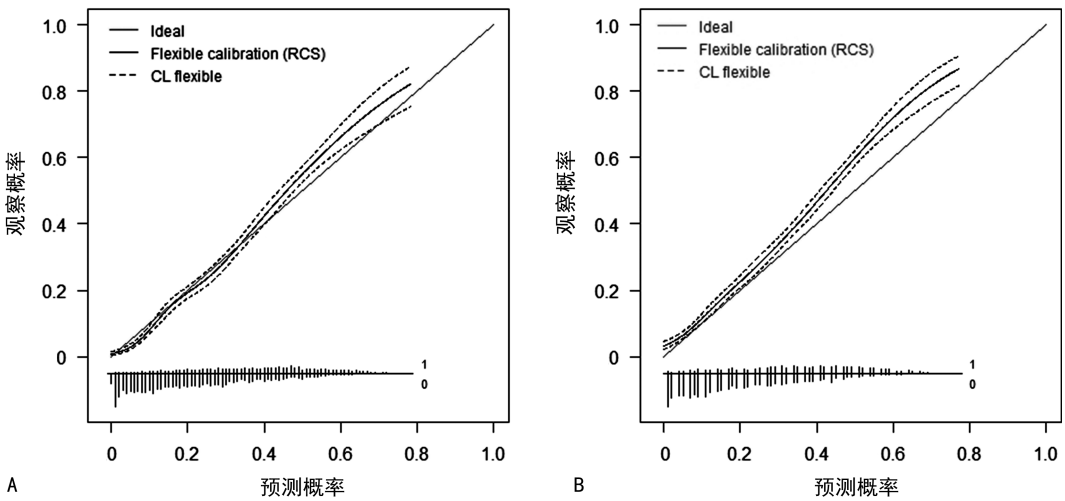
本研究构建预测模型及内验证数据集中,共纳入研究对象 230 170 例,排除了孕产妇及未成年人 260 例、没有连续参加 4 年体检的研究对象 220 606 例、随访初始阶段已被诊断为代谢综合征的研究对象 2 511 例,最终有 6 793 例研究对象符合标准,纳入数据分析。预测模型开发及内验证数据集中,共有 1 750 例研究对象发生代谢综合征,发病率为 25.76%。在模型外验证数据集中,共纳入研究对象 290 463 例,排除孕产妇及未成年人 342 例、没有连续参加 4 年体检的研究对象 278 770 例、随访初始阶段已被诊断为代谢综合征的研究对象 3 670 例,最终有 7 681 例研究对象符合标准,纳入数据分析。模型外验证数据集中,共有 2 222 例研究对象发生代谢综合征,发病率为 28.93%。模型开发队列与外验证队列预测因子及临床结局数据,见表 1。

2.2 模型表现

基于 KNN 算法构建的代谢综合征预测模型通过 Bootstrap 方法的内部验证 AUC 为 0.776(95%*CI*:0.764～0.788),校准截距为 0.028(95%*CI*:−0.031～0.089)、校准斜率为 1.181(95%*CI*:1.106～1.257)、布里尔分数为 0.157;外部验证 AUC 为 0.780(95%*CI*:0.768～0.791)、校准截距为 0.262(95%*CI*:0.207～0.317)、校准斜率为 1.053(95%*CI*:0.990～1.117)、布里尔分数为 0.167。基于 logistic 回归预测模型内部验证 AUC 为 0.783(95%*CI*:0.772～0.795)、校准截距为−0.008(95%*CI*:−0.088～0.073)、校准斜率为 0.995(95%*CI*:0.934～1.058)、布里尔分数为 0.156;外部验证 AUC 为 0.782(95%*CI*:0.771～0.793)、校准截距为−0.045(95%*CI*:−0.113～0.022)、校准斜率为 1.006(95%*CI*:−0.011～1.063)、布里尔分数为 0.164。与基于 KNN 算法构建的预测模型比较,基于 logistic 回归的代谢综合征风险预测模型 AUC 更高,校准斜率接近 1,校准截距接近 0,布里尔分数更低。见表 2。通过 KNN 算法和 logistic 回归预测模型的校准图(图 1、2)可见,KNN 算法内外部验证校准曲线欠佳。

表 1　研究对象基线表($\bar{x}\pm s$)

预测因子	模型开发队列(n=6 793)	外验证队列(n=7 681)
年龄(岁)	41.488±10.411	41.988±11.246
TC(mmol/L)	4.419±0.846	4.834±0.893
UA(μmol/L)	303.305±86.971	331.758±83.226
BMI(kg/m^2)	22.802±2.726	22.697±2.681
ALT(U/L)	21.453±20.271	20.984±15.619

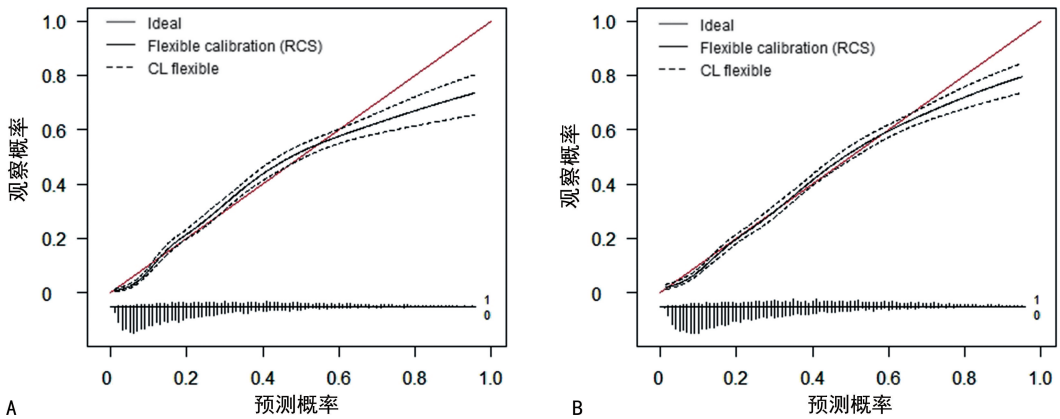


A:内部验证校准图;B:外部验证校准图。

图 1　KNN 预测模型校准曲线

表 2 KNN 算法模型与 logistic 回归预测模型内、外验证预测性能比较

项目	内部验证		外部验证	
	logistic 回归模型	KNN 算法模型	logistic 回归模型	KNN 算法模型
AUC[$\bar{x}$(95%CI)]	0.783(0.772～0.795)	0.776(0.764～0.788)	0.782(0.771～0.793)	0.780(0.768～0.791)
校准截距[$\bar{x}$(95%CI)]	−0.008(−0.088～0.073)	0.028(−0.031～0.089)	−0.045(−0.113～0.022)	0.262(0.207～0.317)
校准斜率[$\bar{x}$(95%CI)]	0.995(0.934～1.058)	1.181(1.106～1.257)	1.006(−0.011～1.063)	1.053(0.990～1.117)
布里尔分数	0.156	0.157	0.164	0.167



A：内部验证校准图；B：外部验证校准图。

图 2 logistic 回归预测模型校准曲线

3 讨 论

疾病风险预测模型有助于临床医务人员早期识别、筛选、诊断疾病并协助其做出临床决策，对疾病的防控及医疗资源的优化具有重要意义。机器学习具有自学习、高容量和不受限于数据分布的特征，能模拟或者无限地接近于人类的大脑学习行为，精确地给出机器学习计算效果，在疾病的早期筛查与诊断等方面存在独到的优势。随着机器学习的兴起与蓬勃发展，其已在医学研究与临床照护领域被广泛运用。

本研究构建基于 KNN 算法和 logistic 回归两种代谢综合征预测模型并进行了预测效能对比，结果提示，基于 KNN 算法构建的代谢综合征预测模型内、外部验证 AUC 低于基于 logistic 回归构建的预测模型。基于 logistic 回归构建的预测模型内、外部验证的校准截距、校准斜率优于基于 KNN 算法构建的代谢综合征预测模型。在综合表现方面，基于 logistic 回归构建的预测模型内、外验证的布里尔分数优于基于 KNN 算法构建的代谢综合征预测模型。GRAVESTEIJN 等[10]的研究结果提示，机器学习构建的中重度创伤性脑损伤预测模型表现并不优于基于 logistic 回归构建的预测模型，这与本研究结果相似。相较于机器学习算法，logistic 回归预测模型更具有优势。logistic 回归与机器学习基于不同的技术原理构建预测模型，算法上各有千秋。机器学习模型对大量数据的处理及运算方面有明显优势，且具有较高的自适应、自学习能力，能对复杂数据进行非线性拟合，从而提高预测结果的准确性。但是，机器学习具有“黑匣子”特性，基于机器学习构建的模型无法对其决策过程和结果进行解释，在临床运用推广时面临可解释性较差的问题[11]；此外，机器学习方法还具有“数据饥饿”特性，在数据量不够的情况，会对估计结果产生偏倚[12]。相对而言，logistic 回归构建的预测模型通过最小二乘法的线性回归来拟合最佳曲线，可通过公式对每一个预测变量的回归系数进行详细展示，协助医务人员判别不同预测因子对预测结果影响的强度和方向，从而进行临床决策[13]。有学者推荐，当基于 logistic 回归构建的预测模型表现不弱于基于机器学习构建的预测模型时，考虑到机器学习的低解释性及“黑匣子”特性，应该选择前者进行推广使用[13-14]。本研究结果显示，基于 KNN 算法构建的代谢综合征预测模型弱于 logistic 回归，且综合前期课题组构建的神经网络、决策树、支持向量机预测模型，推荐使用基于 logistic 回归的代谢综合征风险预测模型。

本研究依据 TRIPOD 报告声明，对风险预测模型研究的样本量进行了计算，研究所用数据集的样本量达到机器学习的 EPV 要求。为验证模型的可复制性和可泛化性，对两种预测模型进行了内部验证（Bootstrap 法）及外部验证（时段验证），其预测性能的评估指标分别从 ROC、校准曲线、校准截距、校准斜率、布里尔分数等多个维度进行了全面评估。而当前不少研究都存在缺乏内部验证或者外部验证、模型预测效能指标报告不全面、样本量不达标等问题，导致有些预测模型具有整体高偏倚风险[3]，其预测值的准确度较低，无法在临床推广运用辅助临床决策。DHI-

MAN 等[15]开展了一项关于机器学习构建预测模型的系统评价，发现大量有关运用机器学习构建预测模型的研究也存在上述缺陷与短板。TRIPOD 报告声明的初衷是为了统一及规范全球有关构建临床预测模型的方法与报告步骤，从而改进临床预测模型质量，降低临床预测模型的风险偏倚，而忽视 TRIPOD 报告声明会使研究者构建的临床预测模型可能存在方法学缺陷、预测结果可解释度低、模型过度拟合、可移植性较差等问题。依据 TRIPOD 报告声明，研究者完整、透明、准确地对预测模型进行报告，将会增强模型的可解释性、可重复性，促进其在临床实践的推广和运用。

本研究也存在某些不足之处。(1)本研究数据为单中心数据，外验证方法仅采用时段验证方法，未来研究应该采用多中心数据，并进行时空验证法，进一步对模型的可泛化性和效度进行验证。(2)相较于 logistic 回归，KNN 算法属于机器学习的方法范畴，建模时能够灵活应对大量高维度数据。而本研究中仅使用 5 个预测因子进行建模，可能会限制 KNN 算法的计算能力和灵活度。

综上所述，基于 KNN 算法构建的代谢综合征风险预测模型表现不如基于 logistic 回归构建的预测模型。基于 logistic 回归构建的预测模型简单，易于理解及使用，可作为代谢综合征的早期风险预警及监测的有力工具，也能为制定代谢综合征风险人群的分层干预措施提供重要的参考依据。研究应基于多中心研究数据对该模型进行空间验证，以此进一步验证该模型的效度。此外，后续研究可考虑从新兴的基因组学、蛋白质组学和影像学技术指标中，筛选出有预测增量值的预测因子纳入模型中进行更新，从而增加模型的预测性能。

参考文献

[1] ALBERTI K G, ZIMMET P, SHAW J. Metabolic syndrome: a new world-wide definition. A consensus statement from the international diabetes federation[J]. Diabet Med, 2006, 23(5): 469-480.

[2] BAHAR A, KASHI Z, KHERADMAND M, et al. Prevalence of metabolic syndrome using international diabetes federation, national cholesterol education panel-adult treatment panel Ⅲ and iranian criteria: results of tabari cohort study[J]. J Diabetes Metab Dis, 2020, 19(1): 205-211.

[3] ZHANG H, SHAO J, CHEN D, et al. Reporting and methods in developing prognostic prediction models for metabolic syndrome: a systematic review and critical appraisal[J]. Diabetes Metab Syndr Obes, 2020, 13: 4981-4992.

[4] ZHANG H, CHEN D, SHAO J, et al. Machine learning-based prediction for 4-year risk of metabolic syndrome in adults: a retrospective cohort study[J]. Risk Manag Healthc Policy, 2021, 14: 4361-4368.

[5] 张凡，高仲合，牛琨. 基于 KNN 的网络流量异常检测研究[J]. 通信技术，2021，54(5)：1235-1239.

[6] ALBERTI K G, ECKEL R H, GRUNDY S M, et al. Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity[J]. Circulation, 2009, 120(16): 1640-1645.

[7] MOONS K G M, WOLFF R F, RILEY R D, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration[J]. Ann Intern Med, 2019, 170(1): W1-33.

[8] ZHANG H, CHEN D, SHAO J, et al. Development and internal validation of a prognostic model for 4 year-risk of metabolic syndrome in adults: a retrospective cohort study[J]. Diabet Metab Synd Ob, 2021, (14): 2229-2237.

[9] ZHANG H, CHEN D, SHAO J, et al. External validation of the prognostic prediction model for 4-year risk of metabolic syndrome in adults: a retrospective cohort study[J]. Diabetes Metab Syndr Obes, 2021, 14: 3027-3034.

[10] GRAVESTEIJN B Y, NIEBOER D, ERCOLE A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury[J]. J Clin Epidemiol, 2020, 122: 95-107.

[11] GABEL J, DESAPHY J, ROGNAN D. Beware of machine learning-based scoring functions-on the danger of developing black boxes[J]. J Chem Inf Model, 2014, 54(10): 2807-2815.

[12] PLOEG T, AUSTIN P C, STEYERBERG E W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints[J]. BMC Med Res Methodol, 2014, 14: 137.

(下转第 2029 页)