·XXXX·

基于水平测试成绩的执业医师理论考试早期预警模型：双队列验证研究

叶熊[1]　肖辉[2]

[1]上海健康医学院临床医学院，上海 201318；[2]上海交通大学医学院附属第一人民医院呼吸和危重症医学科，上海 201620

通信作者：肖辉，Email：xiaohui771210@163.com

【摘要】 目的　基于临床医学本科水平测试理论考试成绩，构建跨院校可泛化的预测模型，提前2年识别执业医师理论考试不合格的高风险学生。**方法**　采用回顾性设计，以373名锦州医科大学2017级五年制临床医学专业本科学生的水平测试和执业医师理论考试成绩为训练集，以上海健康医学院同年级126名本科学生的相应成绩为独立外部验证集。收集学生本科四年级第二学期末（2021年5月）的水平测试理论考试转换分（converted score，CS）与百分等级（percentile rank，PR），以及毕业后第一年（2023年8月）的执业医师理论考试成绩。采用Spearman相关分析评价水平测试两指标与执业医师理论考试成绩的相关性；构建Logistic回归模型预测执业医师理论考试不合格风险；通过曲线下面积（area under the curve，AUC）评估模型区分度，通过校准曲线评估校准度，通过决策曲线分析（decision curve analysis，DCA）评估净获益；基于约登指数确定水平测试理论成绩最佳截点。**结果**　CS、PR与执业医师理论考试总分均呈中等正相关（r=0.61、0.58，P<0.001）。CS单变量模型表现最佳：训练集AUC为0.797（95%CI=0.746~0.848），独立外部验证集AUC为0.895（95%CI=0.819~0.971），显示模型具有良好的跨院校泛化能力。以CS≤175分为最佳截点，预测执业医师理论考试不合格的敏感度为79.55%，特异度为84.21%，阳性预测值为82.80%，阴性预测值为77.80%。模型校准良好（Brier=0.13），DCA显示，在10%~70%阈值范围内，基于CS模型的净获益优于“全部干预”策略。**结论**　水平测试转换分单项指标预测模型在独立外部院校验证中表现稳健，可提前2年有效预测执业医师理论考试不合格风险，为医学生学业风险分层与教育资源精准投放提供简便易得、可推广的指标。

【关键词】 临床医学本科；　水平测试；　执业医师资格考试；　预测效度；　教育预警；　单变量模型

【中图分类号】 R-0；R-4；R33

基金项目：上海高校教师专业发展工程项目（A3-2601-26-311007）

DOI：10.3760/cma.j.cn116021-20260105-02329

An early warning model for National Medical Licensing Examination theoretical test based on Clinical Medicine Undergraduate Level Test: a dual-cohort validation study

Ye Xiong[1], Xiao Hui[2]

[1]School of Clinical Medicine, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China; [2]Department of Respiratory and Critical Care Medicine, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 201620, China

Corresponding author: Xiao Hui, Email: xiaohui771210@163.com

【Abstract】 Objective　To develop a cross-institutionally generalizable predictive model based on the theoretical test scores of Clinical Medicine Undergraduate Level Test (CMULT), and to identify students at high risk of failing the theoretical test of National Medical Licensing Examination (NMLE) two years in

advance. **Methods** In this retrospective study, the theoretical test scores of CMULT and NMLE of 373 undergraduate clinical medicine students (five-year program, enrolled in 2017) from JinZhou medical university served as the training set, and the corresponding scores of 126 undergraduate students enrolled in the same year from Shanghai University of Medicine & Health Sciences served as the independent external validation set. For each student, the converted score (CS) and percentile rank (PR) on the CMULT theoretical test administered at the end of the second semester of the fourth undergraduate year (May 2021), as well as the NMLE theoretical test score obtained in the first year after graduation (August 2023), were collected. A Spearman correlation analysis was used to assess the relationship between CS/PR and NMLE theoretical test scores. A logistic regression model was constructed to predict the risk of failure in the NMLE theoretical test. Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), calibration was assessed using calibration plots, and net benefit was evaluated by decision curve analysis (DCA). The optimal cutoff value for CMULT theoretical test scores was determined using the Youden index. **Results** Both CS and PR showed moderate positive correlations with the total NMLE theoretical test score (r=0.61 and 0.58, respectively, P<0.001). The univariable model based on CS demonstrated the best predictive performance, with an AUC of 0.797 (95%CI: 0.746-0.848) in the training set and 0.895 (95%CI: 0.819-0.971) in the independent external validation set, which indicated good cross-institutional generalizability. CS cutoff of ≤175 yielded a sensitivity of 79.55%, specificity of 84.21%, positive predictive value of 82.80%, and negative predictive value of 77.80% in predicting failure in the NMLE theoretical test. The model showed good calibration (Brier score=0.13), and DCA demonstrated that the CS-based model provided greater net benefit than the "treat-all" strategy across a threshold probability range of 10%-70%. **Conclusions** The predictive model based solely on CS of the CMULT demonstrated robust predictive performance in independent external validation. It can effectively identify students at high risk of failing the NMLE theoretical test two years in advance, providing a simple, accessible, and scalable metric for risk stratification and targeted resource allocation in medical education.

【**Key words**】 Clinical medicine undergraduate; Clinical Medicine Undergraduate Level Test; National medical licensing examination; Predictive validity; Educational early-warning; Single-variable model

Fund program: Shanghai University Faculty Professional Development Program (A3-2601-26-311007)

DOI: 10.3760/cma.j.cn116021-20260105-02329

临床医学专业本科水平测试(简称“水平测试”)由国家医学考试中心和全国医学教育发展中心共同发起,是在医学院校全日制临床医学本科学生进入临床实习前开展的一项全国统一考试。水平测试与执业医师资格考试(the National Medical Licensing Examination,NMLE)是衡量医学人才培养质量的两道重要关口,分别承担院校教育评价与行业准入的核心职能[1]。水平测试综合考查医学生对医学核心知识与临床技能的系统掌握程度,为教育质量提供标准化、可追溯的反馈[2];NMLE是医师合法执业的安全底线与专业准入标准,是评价医师资格考试申请者是否具备从事医疗工作所必需的专业知识与实践技能的准入考试,其成绩是检验高等医学院校医学教育水平的重要参数[3-4]。2场考试虽间隔2年、分属不同阶段,却在知识图谱、能力架构与认知层次上高度衔接,命题逻辑呈递进关系,难度参数存在显著重叠。随着《中国教育现代化2035》提出“构建精准化、智能化人才培养体系”,教育数据挖掘与学习分析技术被赋予新使命,如何将“结果评价”前置为“过程预警”已成为医学教育研究的前沿热点。目前,水平测试与NMLE成绩关联性的实证证据仍极为稀缺,前者在医学教育中的潜在价值尚未被系统量化,导致“教—考—用”闭环缺失关键一环。本研究基于2所医学院校2017级临床医学本科学生在水平测试(本科阶段第4学年末)和毕业后1年参加NMLE(即本科毕业并完成1年住院

医师规范化培训后，距水平测试约2年）的理论考试成绩数据，构建仅含水平测试理论考试转换分（converted score，CS）与百分等级（percentile rank，PR）2项核心指标的极简Logistic回归预测模型，系统评估其对首次NMLE理论考试结果的预测效度。研究从区分度[受试者工作特征（receiver operating characteristic），ROC曲线下面积]、校准度（Hosmer-Lemeshow检验、校准曲线）、净获益（决策曲线分析）与政策可行性（外部验证、可解释性、部署成本）4个维度进行立体验证，旨在：①量化水平测试理论成绩的早期预警阈值，提前2年预警执业医师理论考试不合格风险；②为医学院校提供低成本、可推广的学业预警工具，精准识别需干预人群；③为教育管理部门动态调整人才培养方案、优化资源配置提供循证依据。本文将填补国内“院校教育评价—行业准入资格”纵向预测研究的空白，为推动医学教育评价从“经验驱动”走向“数据驱动”提供可复制、可扩展的实践路径。

1　对象与方法

1.1　研究对象与数据来源

训练集：373名锦州医科大学2017级全日制临床医学专业本科学生的水平测试和NMLE理论考试成绩。外部验证集：126名上海健康医学院同年级、同专业学生的相应考试成绩，考试节点与训练集完全一致。纳入标准：水平测试和NMLE理论考试成绩完整，且为首次参加NMLE。数据采集时间节点：2021年5月，本科四年级第二学期末，完成水平测试；2023年8月，本科毕业后第一年，首次参加执业医师考试。

1.2　研究方法

1.2.1　变量定义

1.2.1.1　预测变量

水平测试理论成绩CS：该指标为国家医学考试中心基于原始分数转换后提供的考生成绩，反映考生在本次测试中的能力水平。水平测试理论成绩PR：该指标为考生个体成绩在所有参测考生中的相对位置，由国家医学考试中心反馈。学生性别和不同实习医院：将性别（以男性为参照）和实习医院（以第1个实习医院为参照）作为分类变量纳入模型。

1.2.1.2　结局变量

NMLE理论考试总分≥360分者记为“通过（1）”，反之记为“未通过（0）”。

1.2.2　统计策略（Prism 10.0）

线性关联：Pearson相关评价水平测试CS、PR与NMLE理论考试总分的相关性。模型构建：多变量Logistic回归，$\mathrm{Logit}(P)=\beta_0+\beta_1\times X_1+\beta_2\times X_2+\beta_3\times X_3$。

1.3　模型性能

区分度：ROC曲线下面积（area under the curve，AUC）、灵敏度、特异度。②校准度：Hosmer-Lemeshow（H-L）检验、校准曲线。③阈值优选：以最大Youden指数确定预警切点Cut-off值。④净收益：决策曲线分析（decision curve analysis，DCA）。

2　结果

2.1　基线差异

训练集与验证集基线差异见表1。验证集CS均值（181.40±33.66）显著高于训练集（169.20±29.65），均值差12.21分（95%*CI*为5.57~18.85，Welch's $t=-3.63$，$df=194.60$，$P=0.0004$）；效应量$\mathrm{eta}^2=0.063$，为小到中等程度差异。F检验显示2组方差齐性成立（$F=1.29$，$P=0.073$），基础分布见图1。

2.2　Pearson相关性

水平测试理论考试中，CS（$r=0.61$，$P<0.001$）和PR（$r=0.58$，$P<0.001$）与NMLE理论考试总分呈中等偏强的正相关。

表1　训练集与验证集基线情况比较

变量	训练集（n=373）	验证集（n=126）	t/χ^2值	P值
年龄（参加NMLE时）	24.40±2.30	23.60±1.90	0.47	0.637
女性（%）	51.47	58.73	3.06	0.080
水平测试PR	32.91±23.81	43.80±27.15	−4.01	<0.001
水平测试CS	169.20±29.65	181.40±33.66	−3.63	0.0004
NMLE理论总分	372.70±54.37	377.60±55.51	−0.86	0.389
NMLE通过率（%）	66.22	69.84	0.41	0.512

2.3　模型选择

以CS和PR单变量、两者合并为双变量及合并后再增加性别与不同实习医院变量分别做Logistic回归建模。4种候选模型的AUC均为0.797，提示预测区分能力一致；PR单变量模型的H-L检验P=0.024，拒绝“拟合良好”假设，校准度差；多变量模型AUC仅提升0.009；仅含CS的模型H-L检验χ^2=9.52，P=0.296，校准度良好，且变量最少，最终选择CS单变量模型进行后续验证研究。模型参数β_0=−6.715，β_1=0.045。

2.4　外部验证集表现

在独立外部队列（n=126）中，采用训练集获得的CS单变量Logistic回归模型参数对外部个体进行概率预测。基于模型概率预测（默认阈值0.5）：AUC为0.895（95%CI=0.819~0.971）（图2），H-L χ^2=8.546（P=0.382）。校准曲线显示，8等分预测概率与实际观察率高度吻合（图3），校准斜率为1.24（95%CI=0.95~1.53），截距为−0.22（95%CI=−0.45~0.00），R^2=0.95，最大绝对误差<0.05，无系统偏移。Brier分数为0.13，准确率为81.75%，阳性预测值为83.51%，阴性预测值为75.86%。DCA显示，在阈值概率10%~70%范围内，基于模型预测的净获益优于“全干预”策略（图4）。基于最佳截点的二分类：根据最大Youden指数（0.638）确定最佳预警截点为CS≤175。该截点下，预测执业医师理论考试不合格的敏感性为79.55%，特异性为84.21%，阳性预测值为82.80%，阴性预测值为77.80%。

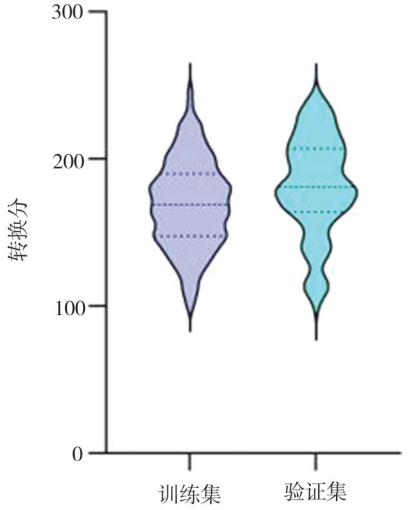


图1　训练集与验证集CS基础分布

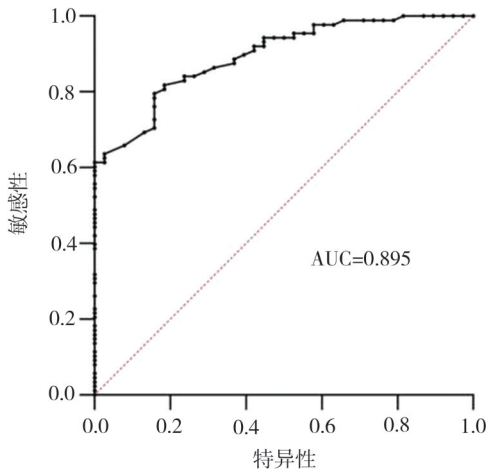


图2　验证集ROC曲线

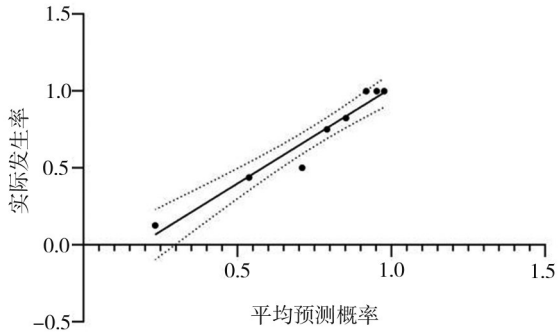


图3　验证集校准曲线

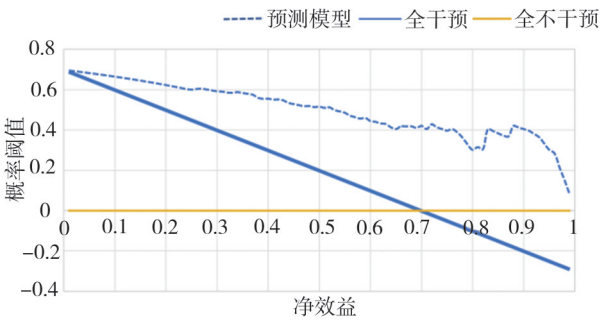


图4　决策曲线

2.5　预测公式

基于水平测试理论考试的CS预测NMLE理论考试通过概率的Logistic回归模型为：$P_{raw}=\frac{1}{1+\exp[-(-6.751+0.045\times CS)]}$，其中$P_{raw}$为原始预测概率。采用Platt缩放在Logit空间对原始预测概率进行后处理校准：Logit(Pcal) = 1.24 × Logit(Praw) − 0.22，其中Logit(p) = ln[p/(1 − p)]。校准参数通过在验证集上最小化Brier评分估计得到。经校准后，模型在外部验证集中的Brier评分为

0.13，校准斜率为 1.24（95%*CI*=0.95~1.53），截距为 −0.22（95%*CI*=−0.45~0.00），显示校准良好。

3　讨论

本研究首次证实，临床医学本科水平测试理论成绩中的CS和PR，均与毕业后NMLE理论考试分数呈中度正相关（r分别为0.61和0.58，均P<0.001）。在提前18~24个月的时间窗内，CS单变量模型已具备足够的预测效力，可前瞻性地筛查出未来NMLE理论考试不合格的高风险人群。以下围绕模型效能、指标优势、落地优势及研究局限与未来方向展开讨论。

3.1　模型效能：单变量CS预测模型在跨院校验证中表现稳健

本研究的关键发现是，仅保留CS单一变量，即可在独立外部验证集（n=126）中实现高区分度（AUC=0.895，95%*CI*=0.819~0.971）与高校准度（H-L检验 P=0.382；校准斜率为 1.24，95%*CI*=0.95~1.53；截距为−0.22，95%*CI*=−0.45~0.00；R^2=0.95；Brier评分为0.13；最大绝对误差<0.05）。DCA进一步显示，当干预阈值设定在10%~70%时，模型带来的净获益始终优于“全员刷题”或“不干预”策略，为教育治理提供了明确的决策区间。以CS≤175分作为预警切点，敏感度为79.55%、特异度为84.21%，Youden指数为0.638，在漏报风险与误报成本之间取得平衡，可直接用于教务系统中的“红黄绿”学业风险分层。尽管本研究验证集CS基础水平显著高于训练集（Cohen's d=0.50），但经线性重校准后，预测概率与实际发生率仍高度吻合，提示模型对整体分数偏移具有良好的稳健性。这种经外部验证且可重校准的预测模型，符合预测模型研究的方法学规范，确保了跨人群应用的可靠性[5]。

3.2　指标优势：CS作为标准化成绩的应用价值

本研究采用水平测试CS而非原始分进行预测，其核心优势在于CS为标准化成绩，由国家医学考试中心统一发布。院校无需获取本地原始分分布即可直接代入预测公式，实现跨省、跨校的“零代码”复现，便于推广应用。回归系数可直观呈现教育效应量，Logistic回归系数β=0.045，对应的比值比$OR = \exp(0.045) \approx 1.046$，表明CS每增加1分，NMLE理论考试通过的概率（odds，即通过概率与不通过概率之比）增加约4.60%。预警截点CS≤175分对应全国常模下的绝对水平，而非随试卷难易浮动的原始分，今后仅需参照官方发布的成绩报告即可长期沿用，无需重新寻找切点，真正实现“一次建模、多年通用”的早期预警目标。这种基于标准化成绩的预测方法，通过将不同年度、不同卷种的原始分数转换至统一量尺，保证了分数的互通可比和长期稳定性，为医学教育评价提供了标准化的测量工具[6-8]。

3.3　落地优势：从模型构建到教育治理的路径转化

与传统多变量预测模型相比，本研究的“单变量 + Platt 缩放校准”方案具备三大落地优势。在数据获取方面，水平测试成绩由国家医学考试中心统一公布，院校无需额外收集隐私敏感信息，显著压缩了数据获取的合规成本与时间成本。在模型应用方面，仅需在Excel表格或院校教务成绩录入和分析系统中内置公式即可完成预测与校准，无需依赖统计软件或算法工程师，有效降低了技术应用门槛。在干预时机方面，水平测试通常在第四学年结束前完成，距离学生毕业尚有12~18个月，为“精准补弱”争取了宝贵的干预时间。在教学管理场景下，模型可将有限经费从大班刷题转向真正的高危人群。教务系统可在成绩公布后1周内自动输出全校风险列表，向红区学生推送个性化学习包（如基于历年错题图谱与知识颗粒度分析），并触发一对一导师跟踪；对黄区学生开放线上强化模块；对绿区学生减少重复培训，释放教学资源。这种精准干预模式契合医学精准教育（precision medical education）的4P理念，即前瞻性（proactive）、个性化（personalized）、参与性（participatory）和预测性（predictive），通过数据驱动的个性化反馈实现教育资源的优化配置[9-10]。

3.4　研究局限与未来方向

尽管本研究构建的预测模型获得了较为稳健的初步结果，但仍存在以下局限性，需在后续工作中持续完善。本模型仅预测NMLE理论考试是否合格，未纳入实践技能考核情况，可能低估学生的整体考试风险。后续可对接实践技能考核数据，构建“理论—技能”联合预测模型。虽采用完全独立的院校样本进行外部验证，但仍局限于单一区域与单

一教考模式，且样本量仅126例。未来需开展多中心、跨年度（至少3届）外部验证，并探索模型在不同招生类别（如“5+3”一体化、农村订单定向等）中的可移植性。虽然本研究纳入了不同实习医院变量，但未发现医院类型对模型判别力有显著贡献。然而，由于未预设等效边界，尚不能认定实习医院效应为零。同时，缺乏学生毕业后规培医院等级、NMLE考前冲刺培训时长、心理资本、学习行为（如在线学习痕迹、题库使用频率）等关键协变量，可能存在未测量的混杂偏倚。下一步可联合学习行为分析与心理问卷测评，构建“学业数据+行为数据+心理数据”的多维预测框架，并采用轻量级梯度提升机或神经网络算法，比较其与传统Logistic回归模型的增量价值。

4 结论与建议

水平测试不仅是临床医学本科教育阶段的“课程结业考”，更可视为医学教育治理的“战略前置传感器”。本研究构建的这种“低成本、高可行、易推广”的早期学业预警工具，为医学教育从“结果评价”向“过程赋能”转变提供了可复制的实践路径。基于本研究结果，提出以下建议：构建闭环干预机制，建立“水平测试—NMLE不合格风险预测—精准干预—效果再评价”的数字化闭环，教务系统根据CS自动分级，触发导师—学生双向匹配，进行动态追踪，用校准偏差反向评估干预成效；成立“水平测试—执业医师考试预测联盟”，共享脱敏数据，每年滚动更新模型参数，推动模型持续迭代，发布《中国临床医学本科生执医理论风险白皮书》，为院校提供横向对标与政策循证依据；在“5+3”一体化、专硕并轨、农村订单定向等不同培养路径中分队列验证模型，探索预测模型对毕业后1年内首次参加执医考试与多次延考人群的适用边界；同时对接2个考试的实践技能考核数据，开发“理论—技能”联合预测模型，实现对NMLE综合通过率的精准预估。模型预警可能对被标记学生产生标签化效应，需要考虑配套心理疏导与正向激励措施，避免“自我实现预言”对学生的负面影响[11]。

利益冲突 所有作者声明不存在利益冲突

作者贡献声明 叶熊：研究设计、数据分析、论文撰写；肖辉：研究设计、论文修改

参考文献

[1] 魏丽娜，吴红斌. 医教协同对医学生学业表现的影响分析：基于全国临床医学本科生调查数据和水平测试数据的实证分析[J]. 高教探索，2024(5)：75-84. DOI: 10.3969/j.issn.1673-9760.2024.05.015.

[2] 胡吉富，李利华，李绍波，等. 以岗位胜任能力和执业医师资格考试要求为核心的医学人才培养模式研究与实践[J]. 中国高等医学教育，2021(1)：43-44. DOI: 10.3969/j.issn.1002-1701.2021.01.022.

[3] 刘芷宁，刘馨，葛丽丽. 基于执业医师考试成绩分析的教学改革研究[J]. 锦州医科大学学报（社会科学版），2024，22(3)：70-72. DOI: 10.13847/j.cnki.lnmu(sse).2024.03.014.

[4] 史宏灿，龚卫娟，郑英，等. 以国家执业医师资格考试为参照的临床医学教学改革思考[J]. 中华医学教育探索杂志，2016，15(5)：459-464. DOI: 10.3760/cma.j.issn.2095-1485.2016.05.007.

[5] 彭文军，杨惠莲，陈虹汝，等. 基于执业医师分阶段考试的医学生问诊SP教学与传统教学的效果分析[J]. 中华医学教育探索杂志，2023，22(9)：1351-1355. DOI: 10.3760/cma.j.cn116021-20210311-01417.

[6] 洪云，张冉，宋默涵，等. 国家医学学业水平测试在临床医学专业考核中的分析与思考[J]. 皖南医学院学报，2023，42(3)：290-292. DOI: 10.3969/j.issn.1002-0217.2023.03.023.

[7] Peng P, Liu L, Wu Q, et al. Establishment and validation of a nomogram for dropout intention in Chinese early year medical undergraduates [J]. BMC Med Educ, 2024, 24(1): 868. DOI: 10.1186/s12909-024-05835-y.

[8] 张泉慧，何惧，任杰，等. 临床医学专业（本科）水平测试的等值方法比较研究[J]. 中华医学教育杂志，2022，42(7)：577-580. DOI: 10.3760/cma.j.cn116021-20250910-02212.

[9] Santen SA, Lomis K, Richardson J, et al. Precision education in medicine: a necessary transformation to better prepare physicians to meet the needs of their patients [J]. AEM Educ Train, 2024, 8(6): e11041. DOI: 10.1002/aet2.11041.

[10] 孙宝志. 医学精准教育概念的提出及启示[J]. 医学与哲学，2024，45(19)：63-67. DOI: 10.12014/j.issn.1002-0772.2024.19.13.

[11] Marcotte K, Yang P, Millis MA, et al. Ethical considerations of using learning analytics in medical education: a critical review [J]. Adv Heal Sci Educ, 2025, 30(1): 87-101. DOI: 10.1007/s10459-024-10402-7.

（收稿日期：2026-01-05）

（本文编辑：邬小梅）