·方法学研究·

PROBAST+AI：基于传统或人工智能方法的预测模型研究质量、偏倚风险及适用性评价工具解读

邹汶廷[1,2#]，郑起程[1,2#]，黄　桥[1]，王永博[1]，任相颖[1]，肖时烽[1,2]，靳英辉[1]，阎思宇[1]

1. 武汉大学中南医院循证与转化医学中心（武汉 430071）
2. 武汉大学第二临床医学院（武汉 430071）

【摘要】近年来，人工智能（AI）或机器学习（ML）方法越来越多地被应用于开发和评估临床预测模型，其算法不同于传统的回归建模方法，导致已有的预测模型偏倚风险评估工具（PROBAST）在评价该类模型时表现出较多的局限性。为此，该工具于2025年更新为PROBAST+AI，在原有框架基础上进行了扩展，以应对基于AI/ML算法开发或评估临床预测模型时特有的方法学挑战。PROBAST+AI从研究对象与数据源、预测因子、结局、分析4个领域对模型开发进行质量评价、对模型评估进行偏倚风险评价，分别涵盖16、18个信号问题；从研究对象与数据源、预测因子、结局3个领域对模型的适用性进行评价。本文旨在对比工具前后版本的变化，对PROBAST+AI的重点内容与条目进行解读，并使用更新版工具对示例临床预测模型文献进行评估，以期帮助国内系统评价作者、临床医生和政策制定者对开发或评估基于传统或AI/ML方法的预测模型研究进行批判性评价。

【关键词】PROBAST+AI；预测模型；系统评价；偏倚风险；人工智能；机器学习
【中图分类号】R-05；R319　　【文献标识码】A

PROBAST+AI: an interpretation of the tool for assessing the quality, risk of bias and applicability of prediction models based on traditional or artificial intelligence methods

ZOU Wenting[1,2#], ZHENG Qicheng[1,2#], HUANG Qiao[1], WANG Yongbo[1], REN Xiangying[1], XIAO Shifeng[1,2], JIN Yinghui[1], YAN Siyu[1]

1. Center for Evidence-Based and Translational Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China
2. The Second Clinical Medical College, Wuhan University, Wuhan 430071, China
#Co-first authors: ZOU Wenting and ZHENG Qicheng
Corresponding authors: JIN Yinghui, Email: jinyinghui0301@163.com; YAN Siyu, Email: yansiyu@whu.edu.cn

【Abstract】In recent years, artificial intelligence (AI) or machine learning (ML) methods have been increasingly used in the development and evaluation of clinical prediction models. Their algorithms differ from traditional regression modeling methods, resulting in significant limitations of the existing Prediction Model Risk of Bias Assessment Tool (PROBAST) for their evaluation. To address these limitations, the tool is updated to PROBAST+AI in 2025. PROBAST+AI extends the

DOI: 10.12173/j.issn.1004-5511.202601172
基金项目：国家自然科学基金面上项目（82174230）；国家自然科学基金青年科学基金项目（82505373）
#共同第一作者
通信作者：靳英辉，博士，教授，博士研究生导师，Email：jinyinghui0301@163.com
阎思宇，助理研究员，Email：yansiyu@whu.edu.cn

original framework to specifically address methodological challenges unique to developing or evaluating clinical prediction models based on AI/ML algorithms. PROBAST+AI assesses the quality of model development and the risk of bias in model evaluation across 4 domains: participants and data sources, predictors, outcome, and analysis, encompassing 16 and 18 signaling questions, respectively. Furthermore, the tool evaluates the applicability of the model across 3 domains: participants and data sources, predictors, and outcome. This article aims to compare the changes between the original and updated versions of the PROBAST tool, interpret the key content and items of PROBAST+AI, and apply the updated tool to evaluate an example clinical prediction model publication, to help domestic systematic review authors, clinicians, and policymakers critically appraise studies that develop or evaluate prediction models based on traditional or AI/ML methods.

【Keywords】PROBAST+AI; Prediction model; Systematic review; Risk of bias; Artificial intelligence; Machine learning

临床预测模型作为医疗决策支持工具，广泛应用于各类健康场景的结局预测。针对同一临床问题，每年新开发的预测模型数量可达数千项[1-4]；然而模型质量参差不齐，许多模型存在质量低下、预测性能偏倚风险高以及算法偏倚等问题[5-6]。2019 年，Moons 教授牵头的团队开发了 PROBAST （Prediction Model Risk Of Bias Assessment Tool）工具[7]，以对预测模型的质量、偏倚风险以及适用性进行系统评估。

近年随着人工智能（artificial intelligence，AI）和机器学习（machine learning，ML）模型的不断发展，其在临床预测模型领域的应用已相当广泛，表现出巨大的实用潜力，例如有文献指出 AI 模型在改善脓毒症的早期识别以尽早对患者进行抗菌药物治疗方面具有潜在价值[8]。然而，原 PROBAST 工具的部分条目在评估此类预测模型时逐渐表现出明显的局限性，如“是否避免了基于单变量分析的预测变量选择？”“最终模型中的预测变量及其权重是否与多变量分析结果一致？”[9]，AI 模型不存在传统意义上的“单变量筛选”或“多变量分析”过程，因此需针对 AI 预测模型进行适当修改[10]。考虑到大量用户的反馈、预测模型结合 AI/ML 方法的进步及普遍应用，以及近期多项综述研究指出基于 AI/ML 的预测模型研究质量普遍欠佳的现状，对原 PROBAST 版本进行更新有其必要性[11]，在此背景下，PROBAST+AI 应运而生[12]。本文旨在对 PROBAST+AI 工具进行系统解读，对比分析工具的更新要点，介绍新工具的使用方法，为国内相关领域的学者及目标用户群体（如临床研究者、方法学家）提供参考。

1 PROBAST+AI 的改进与补充

在临床预测模型领域，传统的数据建模方法均以回归模型为主。但近年来，随着 AI 的不断兴起，ML 也被大量运用到预测模型的开发中。而 PROBAST+AI 工具在评估预测模型的质量、偏倚风险以及适用性时，能更好地兼容传统的回归模型以及新兴的 AI/ML 模型，具体表现如下。

1.1 信号问题的设置

PROBAST+AI 较原 PROBAST 工具最直观的区别在于信号问题的设置变化，包括以下方面。

1.1.1 信号问题的拆分

原信号问题 1.1“是否使用了适当的数据来源，例如队列研究、随机对照试验或巢式病例对照研究数据？”在 PROBAST+AI 中被拆分成 1.1“是否使用了适当的数据源？”和 1.2“是否使用了适当的研究设计？”2 个信号问题，将“数据来源”与“研究设计”问题拆分后，可更清晰地评估数据本身的适用性以及研究设计是否恰当。

1.1.2 信号问题措辞的修改

共有 9 处修改：①提高表述清晰度，将“研究对象的纳入和排除是否恰当”具体化为“研究对象的纳入和排除是否产生了具有代表性的数据集”；将“所有预测因子在模型使用时是否可用？”改为“模型中包含的预测因子在模型使用时是否可用？”；将“相关模型性能指标是否得到恰当评估？”具体化为“模型预测性能是否得到了恰当评估（如校准度、区分度和净获益）？”；将“是否对存在数据缺失的研究对象进行了恰当的处理？”改为“分析中是否对存在数据缺失或

删失的研究对象进行了恰当的处理?”，以明确涵盖生存数据中的删失情况。②进行了术语标准化，将结局领域中2个相关信号问题的“outcome determination”（结局确定）术语改为“outcome assessment”（结局评估）；将结局领域另1个相关信号问题的“outcome defined and determined”（结局的定义和确定）术语改为“outcomes defined and assessed”（结局的定义和评估）。③评估重点的变化，将“是否有合适数量的具有结果的研究对象?”改为“是否有证据表明样本量是合适的?”，强调需提供样本量计算的依据；将“是否考虑了模型的过拟合、欠拟合和性能乐观估计?”改为“是否使用了方法来处理潜在的模型过拟合?”，考虑在AI/ML中，过拟合是最主要的风险而非欠拟合，因此强调了需评估是否有具体技术手段应对过拟合。

1.1.3 信号问题的增加

共有5处改进，以应对基于AI/ML算法的特有风险。AI/ML模型常涉及复杂的图像或数据预处理，若处理不一致会导致严重的偏倚和不公平，因此在“预测因子”领域中新增信号问题2.2“所有研究对象预测因子的预处理方式是否相似?”。基于AI/ML模型中常用的过采样/欠采样、重采样方法及可能存在的仅基于训练数据评估模型及数据泄露问题导致模型性能的高估，因此，在“分析”领域新增4个信号问题以进行充分评估。

1.1.4 信号问题的合并

将原信号问题3.1、3.2、3.3合并为1个信号问题3.1“结局是否被恰当地定义和评估”。因原有的3个问题（结果定义的适当性、标准化以及是否包含预测变量），逻辑上紧密相关，容易导致混淆或重复评价，合并后简化了评估流程。

1.1.5 信号问题的删除

共有4处修改：①2个信号问题因特定于传统回归方法，不适用于AI/ML模型，故被删除：“是否避免了基于单变量分析的预测变量选择?”“最终模型中的预测变量及其权重是否与多变量分析结果一致?”。因现代AI/ML算法通常不使用传统的单变量分析，不提供传统的“权重”或多变量分析结果。②2个信号问题因核心思想已被其他问题覆盖而被删除：“所有登记的研究对象是否都被纳入分析?”，这一考量已融入对缺失数据处理的评估中；“是否正确处理了数据的复杂性，如删失、竞争风险?”已被融入缺失、删失数据处理、过拟合处理等问题中。

1.1.6 信号问题的适用性改变

原适用于模型开发和模型评估的信号问题4.2“对连续和分类预测因子的处理是否合适?”，在新工具中修改为仅适用于模型开发，因模型评估时不应再对预测因子进行任何新的处理或重新定义。

如上，PROBAST+AI更新了多个信号问题，并明确将模型开发与模型评价区分成2个独立的评价板块。

1.2 模型性能的评估

更新后的PROBAST工具阐明并具体区分了3个有关模型性能评估的概念：表观性能（apparent performance）、内部验证性能（internal validation performance）以及外部验证性能（external validation performance）。其主要区别在于：表观性能是指直接使用与模型开发相同的数据评估模型性能，容易高估模型的真实性能；内部验证性能是指在开发数据集上基于数据划分、重采样（如交叉验证）等方法评估模型性能，可以部分校正乐观偏倚（optimism bias），即减少对模型性能的高估；外部验证性能则是指使用未用于模型开发（包括内部验证）的研究对象数据估计模型性能，如基于不同人群、不同时间、不同地点或不同数据来源的新数据测试模型性能。理论上，外部验证性能最能真实反映模型的泛化能力，而表观性能反映模型泛化能力的参考价值最低，三者概念的梳理详见图1。

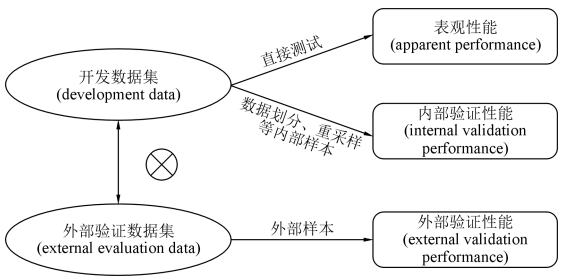

图1 表观性能、内部验证性能与外部验证性能的关系

Figure 1. Relationship among apparent performance, internal validation performance and external validation performance

注：图标“⊗”表示互斥；开发数据集指用于构建、拟合（开发或训练）预测模型的数据集。

1.3 模型的公平性

相较原PROBAST工具，PROBAST+AI特别关注模型的公平性（即不基于年龄、性别、种族或民族、社会经济地位等特征对特定人群造成不利影响）和算法偏倚（指由于训练数据中的偏倚、模型设计缺陷或评估过程不充分，导致模型在特定人群或情境下产生系统性误差或不公平结果的现象），并贯穿模型开发和评价的4个领域。例如领域1“研究对象与数据源”明确强调数据纳排过程中不能有意排除亚组人群，以确保纳入数据的公平性，避免算法偏倚。

1.4 术语体系的统一

PROBAST+AI协调统一了不同专业（如统计学、AI、数据科学、流行病学等）间的术语体系，例如“AI”这一术语在健康预测领域通常用于指代不属于广义线性模型（如Logistic回归）或生存分析模型（如Cox回归）范畴的统计学习方法。基于支持向量机、基于树的学习模型（如随机森林）和神经网络（如深度学习）的分析模型则被归类为AI。同时在该领域，“ML”常被作为AI的同义词使用。然而，工作组表示未来传统统计方法与AI/ML二者可能不再进行严格区分[13]。

1.5 使用范围的扩大

PROBAST+AI扩充了原工具的使用范围，大幅提高了实践价值。例如，PROBAST+AI不仅可在科学研究领域用于预测模型研究的系统评价，还可在多个现实场景中（如在制订医疗保健指南、政策，或决定是否将预测模型应用于日常实践中）用于评价1个或多个特定预测模型的适用性、质量以及偏倚风险等。

2 PROBAST+AI的评价内容、目标用户及适用范围

2.1 目标用户

PROBAST+AI工具开发团队强调，该工具主要使用者为预测模型（开发或评估）研究的研究人员、作者和审稿人，同时潜在参考对象为任何需要评估预测模型适用性、质量及偏倚风险的各行各业人士，如医护人员、医疗政策制定者，甚至是患者、普通民众、模型开发的研究对象等。

2.2 适用范围

PROBAST+AI既适用于AI/ML预测模型研究，也适用于传统回归预测模型研究。此外，与原工具相同，PROBAST+AI既适用于开发新预测模型的研究，也适用于对已有模型进行验证的研究。且同时适用于诊断模型和预后模型，PROBAST+AI对这两类模型在信号问题设计上无差异，但在同一信号问题的判断上可能存在差异（见下文“信号问题1.2：是否使用了适当的研究设计？”）。

2.3 评价内容

PROBAST+AI用于评估预测模型、算法及其研究的质量、偏倚风险及适用性。在使用时，PROBAST+AI分为模型开发与模型评估两个独立模块，分别对模型进行质量及偏倚风险评估。

3 PROBAST+AI 工具使用步骤

PROBAST+AI工具的应用通常遵循以下四步：步骤1，明确预测模型评估或预测模型系统评价的预期用途；步骤2，对预测模型进行分类；步骤3，分领域评价质量、偏倚风险和适用性；步骤4，整体评价质量、偏倚风险和适用性（图2）。其中步骤1每次评估或每个系统评价进行一次，步骤2需对每个预测模型每个相关结果进行评价，步骤3和4需对文献中每个不同的模型分为建立与评估两部分进行评价，各步骤具体内容如下。

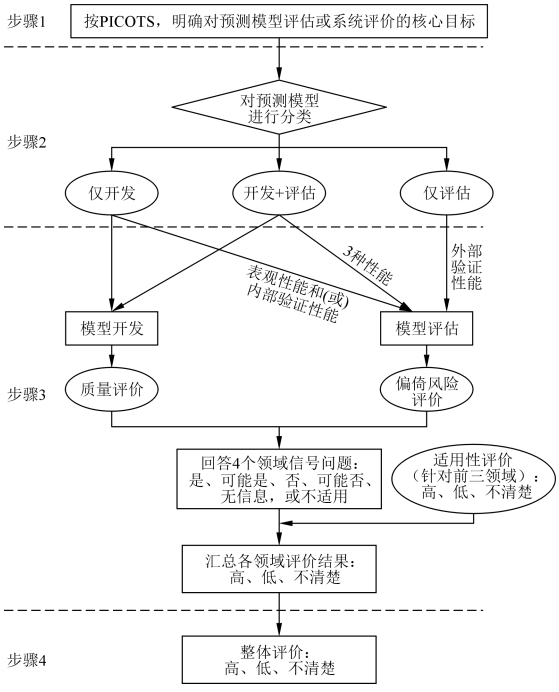

图2 PROBST+AI工具使用步骤

Figure 2. Steps for using the PROBST+AI tool

注：PICOTS. Population，Index model，Comparator model，Outcome，Timing，Setting and intended use of the prediction model[14]；3种性能. 表观性能、内部验证性能及外部验证性能。

3.1 步骤1

明确预测模型评估或预测模型系统评价的预期用途：推荐采用由Cochrane预后方法组指南提出并在CHARMS（Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies）中描述的PICOTS（Population，Index model，Comparator model，Outcome，Timing，Setting and intended use of the prediction model）标准框架（示例见表1）[14]，从模型的目标人群、待评价的预测模型、对照预测模型、预测结局、时间范围、应用场景及预期用途出发，明确对预测模型进行评估或系统评价的核心目标。

3.2 步骤2

对预测模型进行分类：PROBAST+AI工具针对不同类型的预测模型研究设计了不同的信号问题，因此评价者须明确预测模型类型为“仅开发”“仅评估”（此处指外部验证评估）或“开发加评估”。在完成预测模型类型的分类时，需明确区分“模型更新”与“模型验证”2种场景，前者涉及对模型结构的实质性改变（如引入新的预测因子），应使用“模型开发”模块评估；后者仅考察既有模型在新数据中的泛化能力，不涉及模型本身的改动，仅使用“模型评估”模块。

3.3 步骤3

分领域评价质量、偏倚风险和适用性：针对模型开发部分开展质量评价，针对模型评价部分开展偏倚风险评价。此外，模型评估部分需区分模型表观性能评估、内部验证性能评估与外部验证性能评估。对于“仅开发”预测模型研究，通常需要同时进行模型开发和模型评估两部分的评估，后者包括表观性能评估和（或）内部验证性能评估；对于“仅评估”预测模型研究，只需进行模型评估部分的工具评估，并进行外部验证性能评估；对于“开发+评估”预测模型研究，需同时进行模型开发和模型评估两部分的评估。若研究同时报告多种性能，记录模板可参见原文附件Supplementary Table 3[12]。

PROBAST+AI信号问题分为4个领域（表2），这些问题可选择“是”“可能是”“否”“可能否”“无信息”或酌情选择“不适用”进行回答。“是”或“可能是”代表高质量或低偏倚风险（以下简称为“低风险”）；“否”或“可能否”代表低质量或高偏倚风险（以下简称为“高风险”）；“无信息”仅在报告信息不足时使用；“不适用”可用于特定类型预测模型或场景不适用的信号问题。

在对信号问题评价基础上，可将预测模型各领域和整体质量与偏倚风险分为“低”“高”“不清楚”。汇总每个领域内结果时，借鉴“短板理论”，即仅当该领域涵盖的所有信号问题均回答“是”或“可能是”时，方可判定为“低风险”；只要出现任何一个“否”或“可能否”的信号问题，该领域即为“高风险”。当存在特殊原因时也可判断“低风险”，如，仅信号问题4.1“是否避免了仅基于表观性能进行模型评估”为“否”，但模型开发的样本量远超模型复杂度，则过拟合风险极低，此时表观性能可能接近真实性能，因此仍可认为该领域偏倚风险较低。当某个信号问题被判断为“无信息”而同时其他信号问题均为“低风险”时，则该领域归为“不清楚”。

适用性评价针对前3个领域开展，但没有信号问题，各领域适用性可评估为“低”“高”或“不清楚”。

表1 PICOTS内容及应用示例

Table 1. The content of PICOTS and its application example

PICOTS	应用示例
Population（目标人群）	有潜在发生重度脓毒症或脓毒性休克高危风险的患者
Index model（待评价的预测模型）	基于ML算法（如XGBoost、逻辑回归等）开发的重度脓毒症早期预测模型
Comparator model（对照预测模型）	MEWS（改良的早期预警评分）、SOFA（序贯器官衰竭评估量表）、SIRS（全身炎症反应综合征标准）
Outcome（预测结局）	重度脓毒症的发生（基于Sepsis-3标准或其他临床诊断标准）
Timing（时间范围）	T0（预测起始点）：患者就诊或入院时；预测时间窗：提前48 h预测重度脓毒症发生风险
Setting and intended use of prediction model（预测模型的应用场景及预期用途）	美国急诊科和住院病房（含重症监护病房及普通病房），预期用途为早期重度脓毒症实时预警，从而降低不良结局风险

注：示例取自文献[15]，同本文第5部分示例。

表2 PROBAST+AI模型开发及评估部分的领域及信号问题

Table 2. Assessment domains and signaling questions in PROBAST+AI for model development and model evaluation

模型开发	模型评估
一、研究对象与数据源	
1.1 是否使用了适当的数据源？	1.1 是否使用了适当的数据源？
1.2 是否使用了适当的研究设计？	1.2 是否使用了适当的研究设计？
1.3 研究对象的纳入和排除是否产生了具有代表性的数据集？	1.3 研究对象的纳入和排除是否产生了具有代表性的数据集？
适用性评价：纳入的研究对象与模型研究的问题或评估者对模型的预期用途是否相符？	
二、预测因子	
2.1 所有研究对象的预测因子是否以相同方式定义和评估？	2.1 所有研究对象的预测因子是否以相同方式定义和评估？
2.2 所有研究对象预测因子的预处理方式是否相同？	2.2 所有研究对象预测因子的预处理方式是否相同？
2.3 预测因子的评估是否在不知晓结局数据的情况下进行？	2.3 预测因子的评估是否在不知晓结局数据的情况下进行？
2.4 模型中包含的预测因子在模型使用时是否可用？	2.4 模型中包含的预测因子在模型使用时是否可用？
适用性评价：模型中预测因子的定义、预处理、评估和评测时机与模型研究的问题或评估者的预期用途是否相符？	
三、结局	
3.1 结局是否被恰当地定义和评估？	3.1 结局是否被恰当地定义和评估？
3.2 所有研究对象的结局是否以相同的方式定义和评估？	3.2 所有研究对象的结局是否以相同的方式定义和评估？
3.3 是否在未使用或不知晓预测因子信息的情况下进行结局评估？	3.3 是否在未使用或不知晓预测因子信息的情况下进行结局评估？
3.4 预测因子评估与结局评估之间的时间间隔是否设置合理？	3.4 预测因子评估与结局评估之间的时间间隔是否设置合理？
适用性评价：结局的定义、评估以及评定时机与模型研究的问题或评估者的预期用途是否相符？	
四、分析	
4.1 是否有证据表明样本量是合适的？	4.1 是否避免了仅基于表观性能进行模型评估？
4.2 对连续和分类预测因子的处理是否合适？	4.2 是否有证据表明样本量是合适的？
4.3 分析中是否对存在数据缺失或删失的研究对象进行了恰当的处理？	4.3 分析中是否对存在数据缺失或删失的研究对象进行了恰当的处理？
4.4 如果使用了处理类别不平衡的方法，模型及其预测是否进行了重新校准？[a]	4.4 如果使用了处理类别不平衡的方法，模型评估是否在未经不平衡校正的数据集中进行？[a]
4.5 是否使用了方法来处理潜在的模型过拟合？	4.5 如果通过数据拆分创建训练集和测试集，是否有证据表明避免了数据泄露？[abc]
	4.6 如果使用重采样方法评估模型性能，重采样过程是否复现了模型开发的所有步骤？[ab]
	4.7 模型预测性能是否得到了恰当评估（如校准度、区分度和净获益）？

注：[a]若模型开发或评估过程未使用“处理类别不平衡的方法”“数据拆分”“重采样方法”，则判定相应信号问题为“不适用”；[b]模型评估部分领域4中各信号问题需分别对模型的表观性能、内部验证性能及外部验证性能进行评价，其中4.5及4.6仅适用于内部验证性能的评估；[c]数据泄露指模型开发阶段的数据以某种方式被无意间包含在模型评估阶段，通常会导致对模型性能的过度乐观估计或不准确的预测或分类[12]。

3.4 步骤4

整体评价质量、偏倚风险和适用性：在步骤3评价基础上，可将整体质量、偏倚风险及适用性评为“低”“高”“不清楚”。与步骤3类似，基于“短板理论”，4个领域均判断为“低风险”则整体为“低风险”；只要1个领域被评为“高风险”，则整体为“高风险”；若存在某个或某些领域为“不清楚”，且其他领域均为“低风险”，则整体为“不清楚”。适用性的总体评价也是如此。需要注意的是，有时总体判断会因为某些特殊情况发生变化。如一项模型评估研究因未进行外部验证，领域4被认为是“高风险”，但若该模型是基于足够大的数据集开发并且进行了一些内部验证，在其他3个领域均为“低风险”时，该研究的总体偏倚风险可能会由“高风险”调整为“低风险”。

可以列表展示各研究或模型在4个领域及总体的质量评价、偏倚风险评价及适用性的评估结果，条形图展示各领域及总体上高、低或不清楚的研究或模型比例。针对多模型评估的处理为当一项研究报告了多个不同预测模型的开发或评估时，应对每个模型分别进行完整的PROBAST+AI评估。

4 PROBAST+AI 工具各领域信号问题解读

4.1 模型开发

4.1.1 研究对象与数据源

信号问题1.1 是否使用了适当的数据源？

数据来源可追溯时判断为“是/可能是”，不可追溯（包括未说明数据来源）或与研究问题适配度低时判断为“否/可能否”。对于数据收集和测量程序信息不足的公共数据库数据源应引起关注。此外，缺乏研究对象抽样和测量细节的数据源可能隐藏有公平性问题。

信号问题1.2　是否使用了适当的研究设计？

预后模型采用前瞻性纵向队列设计或随机对照试验的数据判断为“是/可能是”，但随机对照试验需将随机分配的干预措施作为独立的预测因子纳入模型分析以权衡治疗效应；诊断模型采用横断面研究设计判断为“是/可能是”，对于需随访以获得结局信息的模型也可以使用队列研究。针对病例队列或巢式病例对照等选择性抽样设计，若采用逆采样分数法（inverse sampling fraction）重新加权校正后可判断为“是/可能是”；若未进行校正则判断为“否/可能否”。现有数据集或常规医疗登记数据因并未专门设计用于开发预测模型，往往存在更多的数据质量问题，通常被判断为“否/可能否”。

信号问题1.3　研究对象的纳入和排除是否产生了具有代表性的数据集？

选择性纳入、排除或招募研究对象，使之与预测模型预期使用人群的代表性不符，可判断为“否/可能否”，如预后模型中纳入已知结局的对象，或诊断模型中排除了难以诊断的对象（属于诊断模型中常见的选择偏倚，从而引入疾病谱偏倚[16]）。此外，该问题也涉及公平性问题，当未包含、明确排除或过采样边缘化亚组（例如，按年龄、性别、种族/民族、地理位置或特定健康状况/合并症定义的亚组个体）时也应判断为“否/可能否”。当使用现有数据源或公共数据库且纳入排除标准未知时，需评估该数据集是否有代表性，或判断为“无信息”。所有研究对象采用相同的纳排标准且无上述选择性问题或对亚组的不当处理时，判断为“是/可能是”。

适用性评价：主要考察纳入的研究对象及其数据是否与步骤1中定义的系统评价的问题或预测模型的预期用途相符。若研究对象构成符合系统评价的问题或预测模型的预期用途，则适用性可判为“高适用性”，否则判为“低适用性”。对于随机对照试验，由于其通常涉及更严格的纳入排除标准、更少的预测因子、更短的随访时间，需仔细考虑基于此的预测模型研究的适用性问题。

4.1.2　预测因子

信号问题2.1　所有研究对象的预测因子是否以相同方式定义和评估？

当所有研究对象同一预测因子的定义、分类阈值和评估方法一致时判断为“是/可能是”，不一致时判断为“否/可能否”。尤其要关注需主观判断或依赖测定者测量技能的预测因子，是否明确说明了测定者信息。该问题也涉及公平性问题，需关注是否对所有研究对象，包括边缘化亚组使用了不同的预测因子定义或评估方法。

信号问题2.2　所有研究对象预测因子的预处理方式是否相同？

当所有研究对象（特别注意边缘化亚组）的预测因子预处理方式相同时，判断为“是/可能是”，不一致时判断为“否/可能否”。当使用现有数据源或公共数据库且预处理步骤未知时，需评估是否具有一致性，或判断为“无信息”。

信号问题2.3　预测因子的评估是否在不知晓结局数据的情况下进行？

预测因子的评估在不知结局数据的情况下完成，或虽未提供盲法信息但预测因子的评估时间明显早于结局评估时，判断为“是/可能是”，其中前瞻性研究可直接判断为“是/可能是”；已知结局数据进行的预测因子评估判断为“否/可能否”；未提供盲法信息且预测因子评估时间与结局评估时间间隔未知时，判断为“无信息”。

信号问题2.4　模型中包含的预测因子在模型使用时是否可用？

进行预测时，所有预测因子的数据应可获取，判断为“是/可能是”，反之则为“否/可能否”；不清楚是否可获取时为“无信息”。

适用性评价：考察模型中预测因子的定义、预处理、评估和评估时机与系统评价问题或评估者的预期用途是否相符，若相符，则判为“高适用性”，否则判为“低适用性”。

4.1.3　结局

信号问题3.1　结局是否被恰当地定义和评估？

采用预先指定或标准化的结局定义和评估方法时，如基于指南或已发表研究，判断为“是/可能是”；采用非标准化的结局定义或较差的结局

评估方法时，判断为“否/可能否”，尤其对于需主观判断或特定测量技能的结局时，该问题发生的可能性更高。对于连续量表结局基于自定义或数据驱动设定阈值，或对复合结局有意排除或纳入特定成分结局，均可判断为“否/可能否”。当使用现有数据源或公共数据库时也应着重考虑结局的定义和测量方法是否与预先设计的预测模型研究不同。

信号问题3.2　所有研究对象的结局是否以相同的方式定义和评估？

所有研究对象的结局定义、阈值、测量方法均相同，且对不同中心、边缘化亚组无差异实施时，判断为“是/可能是”，否则判断为“否/可能否”，如诊断模型研究中仅对预测因子阳性者实施参考标准，发生部分结局验证（如未对首检指标阴性者进一步确证诊断）或差异结局验证（如对首检指标阴性者使用准确性较低的替代标准评估），这属于诊断模型中常见的部分证实偏倚[16]。对于需要主观判断或特定测量技能的结局，或基于现有数据源或公共数据库时，该问题发生的可能性更大。

信号问题3.3　是否在未使用或不知晓预测因子信息的情况下进行结局评估？

结果评估在未使用或不知晓预测因子信息，且专家裁决复杂结果时仍保持独立性，判断为“是/可能是”；结果评估者知晓预测因子信息，或使用预测因子定义或评估预测结局时，判断为“否/可能否”；未明确说明结果评估是否独立于预测因子则为“无信息”。

信号问题3.4　预测因子评估与结局评估之间的时间间隔是否设置合理？

时间间隔符合疾病自然史且边缘化群体无评估延迟时判断为“是/可能是”；时间间隔过长或过短导致关联性失真，急性病未实现同步评估或组织样本采集时点分离时判断为“否/可能否”；未明确说明时间间隔设计或未验证边缘化群体时效公平性判断为“无信息”。

适用性评价：考察结局的定义、评估以及评定时机与系统评价问题或评估者的预期用途是否相符。结局应当采用适用于目标系统评价问题或评估者预期使用模型相关背景的方法在适当的时间间隔进行定义和评估，若相符，则判为“高适用性”，否则判为“低适用性”。

4.1.4　分析

信号问题4.1　是否有证据表明样本量是合适的？

总体而言样本量越大越好，在判断时应谨慎使用经验法则，可根据已有针对连续、二分类、多分类或生存结局的预测模型的最小样本量标准进行判断[17-19]。模型的复杂性、预测因子-结局关联的预期强度等均会影响所需样本量。一些基于AI的预测模型，其模型复杂性和估计的独立参数量往往难以评估，可能需要巨大的样本量来产生稳定的预测并减小过拟合风险，特别是当超参数未仔细调整时。若原始开发者未提供有关样本量充足性的信息，判断为“无信息”。

信号问题4.2　对连续和分类预测因子的处理是否合适？

将连续型预测因子转换为分类变量，或随意删除、合并分类预测因子的类别，或基于数据驱动以最大化模型表观预测性能的变量分类，均可判断为“否/可能否”；回归模型中，对连续预测因子使用样条函数或分数多项式探索非线性关联，或采用基于树的学习方法（如随机森林），可判断为“是/可能是”。

信号问题4.3　分析中是否对存在数据缺失或删失的研究对象进行了恰当的处理？

未因预测因子或结局的缺失而排除研究对象，或采用了合理的缺失数据处理方法（如多重插补方法、竞争风险模型、缺失指标方法、使用代理分裂的基于树的学习方法等），判断为“是/可能是”，否则判断为“否/可能否”。当无法获知是否从分析中排除了具有任何缺失或删失数据的研究对象时，虽可判断为“无信息”，但建议判断为“否/可能否”，因大多建模技术会自动排除在任何预测因子或结局上具有缺失值的记录。注意该问题也涉及公平性问题，当对缺失或删失数据的处理在边缘化亚组中存在差异时，也应判断为“否/可能否”。

信号问题4.4　如果使用了处理类别不平衡的方法，模型及其预测是否进行了重新校准？

使用了改变数据分布的校正方法（如欠采样、过采样、SMOTE算法）但通过重新加权或重新估计截距等方法重新校准模型预测，可判断为“是/可能是”，未进行重新校准则判断为“否/可能否”；若未使用处理类别不平衡的方法，则此

问题不适用。

信号问题4.5　是否使用了方法来处理潜在的模型过拟合?

样本量相对于模型复杂性足够大，或应用正则化/收缩方法，或避免了未经惩罚的数据驱动的预测因子选择（如单变量筛选、向前逐步选择）时可判断为“是/可能是”，否则判断为“否/可能否”；若样本量过小，即使使用了上述方法也难以避免过拟合的问题，仍可判为“否/可能否”。

4.2　模型评估

模型评估 4.2.1 至 4.2.3 领域（研究对象与数据源、预测因子、结局）的评估与模型开发模块基本一致。若模型开发与模型评估均基于相同数据集，并采用相同的预测因子和结局指标定义及测量方法，则模型开发与模型评估在这 3 个领域的信号问题回答将完全一致；若模型评估使用了与模型开发不同的外部数据源，则模型开发与模型评估在前 3 个领域的回答可能不同。下文仅就 4.2.4 领域进行解读。

4.2.4　分析

信号问题4.1　是否避免了仅基于表观性能进行模型评估?

若模型性能评估使用了与模型开发完全不同的数据，而避免了只有表观性能的偏倚风险，则判为“是/可能是”；反之若使用了与模型开发相同或部分相同（数据泄露）的数据，则判为“否/可能否”；即使此信号问题被评为“否/可能否”（即未避免仅使用表观性能进行评估），但若模型开发的样本量相对于模型复杂度而言足够大，如极大样本应用于简单模型，考虑到过拟合风险极低，表观性能接近真实性能，也可接受由此带来的性能估计偏倚为低风险。

信号问题4.2　是否有证据表明样本量是合适的?

与模型开发中的信号问题 4.1 评估方式一致，参考最新样本量指南进行评估，谨慎使用经验法则，若缺乏用于评估样本量的充分信息时，可判断为“无信息”。注意当在特定亚组进行模型性能评估时，亚组的有效样本量可能过小，更容易判为“否/可能否”。

信号问题4.3　分析中是否对存在数据缺失或删失的研究对象进行了恰当的处理?

与模型开发中的信号问题 4.3 评估方式类似。在模型评估时，如该预测模型的预期用途是在预测因子可能缺失的环境中（如使用常规医疗数据），那么使用在实践中预期处理缺失数据的方法可能在模型评估时能更好地反映其实际性能。此外，当在外部数据中评估模型性能时，可能遇到某个预测因子的系统性缺失，如研究者只是简单地从模型中省略该预测因子时则该问题为高风险。

信号问题4.4　如果使用了处理类别不平衡的方法，模型评估是否在未经不平衡校正的数据集中进行?

若模型开发过程中采用了某种解决类别不平衡的采样方法（如欠采样、过采样），但模型评估在未进行不平衡校正的数据集上评估，则判为“是/可能是”；若在模型开发时采用了类别不平衡校正方法，但在校正后的数据集上进行模型性能评估，则判为“否/可能否”。

信号问题4.5　如果通过数据拆分创建训练集和测试集，是否有证据表明避免了数据泄露?

此项仅适用于进行内部验证性能评估，不适用于表观性能和外部验证性能评估。未出现数据泄露（训练集与测试集数据无重叠），则判为“是/可能是”；若预测模型的参数是基于评估模型的数据再次调优（而非仅基于开发数据进行调优），则可能出现数据泄露，判为“否/可能否”。鉴于数据泄露会高估性能，使用适当的内部或外部验证技术克服此问题，也可判为低偏倚风险。

信号问题4.6　如果使用重采样方法评估模型性能，重采样过程是否复现了模型开发的所有步骤?

此项仅适用于进行内部验证性能评估，不适用于表观性能和外部验证性能评估。若研究者采用了某种形式的内部验证，如基于自助法（bootstrapping）或交叉验证的重采样，当重采样程序复制了模型开发的全部步骤，包括插补、变量选择、超参数调优和模型构建，则判为“是/可能是”，否则为“否/可能否”。

信号问题4.7　模型预测性能是否得到了恰当评估（如校准度、区分度和净获益）?

模型性能评估包括评估整个预测结局值范围内观察值和估计值的一致性（即校准度，基于校准图而非仅基于拟合优度统计检验）；评估模型预测是否因观察到的结局值而异（即区分度，如 C-index）；理想情况下，还应包括基于模型的正

确决策的某种度量（如基于决策曲线的净获益度量）。因此，若同时提供了校准度和区分度，则判为“是/可能是”；若缺乏2组指标中的任意一种，或仅提供模型的表观性能评估，或仅基于模型概率阈值报告敏感性或特异性等分类指标（常见于诊断模型），则判为“否/可能否”。此外，也需注意边缘化亚组中模型的预测性能。

5 示例分析

本研究以“Validation of a machine learning algorithm for early severe sepsis prediction：a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals”[15]为例。此研究覆盖了PROBAST+AI评估的全部核心场景，包括模型的开发与验证、内部验证性能与外部验证性能的评估、AI算法与传统预测模型的应用，因此被选做示例分析。该研究旨在利用多样化患者数据集，针对可疑的脓毒症患者，开发并验证一种可提前48 h预测严重脓毒症发生的ML算法，实现脓毒症的早期准确诊断，以降低患者不良结局的发生风险。文章报告了一种基于ML算法的预测模型，同时包括此模型的开发与评估2个部分，测定了该模型的内部验证性能与外部验证性能。使用PROBAST+AI工具对其进行质量、偏倚风险和适用性评价，评价结果见附件表1和表2，具体原因分析如下。

5.1 模型开发部分

领域1，该研究开发的脓毒症预测模型是基于回顾性研究的已有数据集（Dascena Analysis Dataset），纳入的数据并非为此模型的开发所专门设计，但具有可追溯性且具有明确的纳排标准，因此问题1.1判断为“是”，问题1.2判断为“可能否”；同时研究对象的纳入和排除产生了具有代表性的数据集，问题1.3判断为“是”，因此领域1为“低质量”。因纳入的研究对象及其数据与预测模型的预期用途相符，适用性评价为“高适用性”。

领域2，此模型所有研究对象同一预测因子的定义、分类阈值和评估方法一致，问题2.1判断为“是”；未提及是否进行预处理及预处理方法，问题2.2判断为“无信息”；预测因子的评估是在不知晓结局数据的情况下完成且进行预测时所有预测因子的数据可获取，问题2.3与2.4均判断为“是”，因此领域2质量为“不清楚”。同时模型中预测因子的预处理与评估者的预期用途关系不明确，因此该领域适用性评价为“不清楚”。

领域3，此模型在开发过程中使用了合适的脓毒症定义，定义中并未包含预测因子的信息且所有研究对象的结局以相同的方式定义和评估，临床结局的判定是在不知道预测因子的情况下判定的，因此问题3.1至3.3均判断为“是”。预测因子的测量和临床结局的判定之间的时间间隔设置合理，未发现潜在的偏倚风险，但对边缘化群体未明确说明，问题3.4判断为“可能是”。因此，领域3为“高质量”。同时结局的定义、评估以及评定时机与评估者的预期用途相符，该领域适用性评价为“高适用性”。

领域4，此模型的开发基于已有数据库，样本容量充足（n=489 850），问题4.1判断为“是”；连续变量（心率、收缩压、舒张压、呼吸频率、体温、血氧饱和度、年龄）没有额外归一化或离散化处理，问题4.2判断为“是”；未因预测因子的缺失而排除研究对象且采用了合理的缺失数据处理方法，问题4.3判断为“是”；使用了改变数据分布的校正方法但通过重新加权进行了重新校准，问题4.4判断为“是”；该研究样本量足够大，开发的模型过拟合的风险较小，问题4.5判断为“是”，因此领域4为“高质量”。

5.2 模型评估部分

该研究使用了十折交叉验证及简单划分法进行了内部验证性能评估，同时基于卡贝尔·亨廷顿医院数据集（Cabell Huntington Hospital Dataset）进行了外部验证。因此，模型评估时需同时报告内部验证性能和外部验证性能的评价结果。

5.2.1 内部验证性能

模型内部验证性能评估使用了与模型开发相同来源的数据，因此前3个领域的评价结果与模型开发部分相同。对于领域4，问题4.1判断为“是”；无论是十折交叉验证还是简单划分出的内部验证集，样本量均在数万例，模型评估样本容量充足，问题4.2判断为“是”；未因预测因子而排除研究对象且采用了合理的缺失数据处理方法，问题4.3判断为“是”；模型开发过程中解决了类别不平衡问题，但模型评估在未进行不平衡校正的数据集上评估，问题4.4判断为“是”；训

练集与内部验证集数据无重叠，未出现数据泄露，问题4.5判断为“是”；验证过程中没有采用重采样的方法，问题4.6判断为“不适用”；报告了区分度但没有报告校准度，问题4.7判断为“否”，因此领域4为高偏倚风险。综合4个领域，模型内部验证性能具有高偏倚风险。

5.2.2 外部验证性能

前3个领域虽采用不同的数据来源，但采用相同的预测因子和结局指标定义及测量方法，且两数据库均为回顾性研究的已有数据集，因此前3个领域结果与模型开发一致。对于领域4，外部验证性能评估使用了与模型开发完全不同的数据，而避免了只有表观性能的偏倚风险，问题4.1判断为“是”；外部验证数据集样本量为13 581例，样本量充足，问题4.2判断为“是”；未因预测因子缺失而排除研究对象且采用了合理的缺失数据处理方法，问题4.3判断为“是”；外部验证性能在未进行不平衡校正的数据集上评估，问题4.4判断为“是”；外部验证依旧报告了区分度但没有报告校准度，问题4.7判断为“否”，因此领域4为高偏倚风险。综合4个领域，模型外部验证性能具有高偏倚风险。

5.3 整体评价

模型开发部分4个领域的质量判定为“低、不清楚、高、高”，前3个领域适用性评价分别为“高、不清楚、高”，因此模型开发的整体研究质量低，模型适用性为不清楚；模型评估部分外部验证性能与内部验证性能在4个领域的偏倚风险判定均为“高、不清楚、低、高”，前3个领域适用性评价分别为“高、不清楚、高”，因此，模型整体适用性为不清楚，且存在高偏倚风险。

6 讨论

PROBAST+AI是在PROBAST基础上，经系统开发面向预测模型研究的专用评估工具，旨在评估原始研究在开发、验证或更新用于个体化预测的预测模型时存在的质量、偏倚风险及适用性问题[20]。该工具同时涵盖模型开发与模型评估两类场景，既适用于AI/ML预测模型研究，也适用于传统回归预测模型研究。

传统预测模型（如逻辑回归）与现代AI/ML模型（如深度学习、集成方法）在开发和验证流程上存在显著差异[10]。AI/ML模型虽在某些任务中表现出色，但其高复杂性、黑箱特性、对数据的高度依赖以及验证不足等问题，导致其在临床转化中面临较大挑战[8, 21]。同时原始PROBAST未充分覆盖以下问题：复杂数据预处理（如图像标准化、文本嵌入）、自动特征选择或表示学习、重抽样过程中信息泄露的高风险、模型透明度与可解释性问题[22-23]。在此基础上，PROBAST+AI逐渐形成并不断完善。

相较于PROBAST-2019，PROBAST+AI工具的显著改进之一为明确区分模型开发过程中的方法学质量概念与模型性能评估中的偏倚风险概念[12]。方法学质量关注的是模型开发流程，决定模型的本身质量；偏倚风险则关注评估过程，决定性能估计的真实性。通过分离这2个维度，解决了以往评估中概念模糊的问题，使研究人员能够更精准地诊断问题所在，即1个模型可能本身开发质量高，但其报告的性能却因不当的验证方法导致偏倚风险高；反之亦然。当前评估框架有助于全面判断、甄别出真正可靠、可用的预测模型。同时PROBAST+AI通过模型的适用性评价，可保证开发出来的模型与研究和应用目的契合，提高模型在目标人群和特定场景中应用的匹配度，优化资源配置，避免资源浪费。

新兴的PROBAST+AI工具还明确强调了模型开发与评估的公平性要求，以确保模型在不同人群亚组（特别是边缘化群体）中均能保持公平和有效。公平性要求之一在于确保建模和验证数据具有充分的多样性和代表性，否则可能导致模型对特定亚组无效。研究显示，当AI模型用于医疗时算法可能会系统性地歪曲并加剧少数群体的健康问题，导致健康不平等问题，因此提出了“促进数据多样性、包容性与普适性标准”的倡议[24]。然而需要注意的是，仅靠工具的理论分析可能不足以全面识别算法偏倚与公平性问题，这些问题将在模型的真实世界应用中得到检验。这要求研究人员、使用者及监管机构共同推动预测模型在临床实践中透明、负责且合乎伦理的应用。

随着AI/ML的普及，Collins团队在2024年对临床预测模型研究的报告规范进行了更新。既往报告规范TRIPOD-2015[25]重点关注回归模型，并不能完美地适用于所有的预测模型。在此背景下，针对AI/ML预测模型研发出的TRIPOD-AI报告规范合理地解决了这一问题[26]。AI/ML的应用

日益广泛，PROBAST+AI与TRIPOD+AI形成了一个“评价+报告”的闭环，通过对模型开发、评估的规范评价和透明报告，共同助力于临床预测模型的质量改善。

尽管PROBAST+AI工具功能强大，但其有效应用仍面临一些挑战。例如，评估过程对评估者的方法学素养要求较高；对于文献中报告不充分的信息如何判定也较依赖主观判断，可能影响一致性；此外，仅凭条目式工具进行理论层面的判断，未必能够完全揭示算法偏倚或公平性问题，仍需结合具体应用场景加以验证。此外，本文相关解读是建立在对原始文献、附件及官网解释说明文件等进行系统梳理、学习的基础上，因作者团队未实际参与PROBAST+AI工具的开发过程，因此部分解读可能存在一定的主观理解，例如对于信号问题修改的原因，原始材料中未明确提及，因此作者团队基于自身理解对信号问题变化的逻辑进行了解读。

综上所述，PROBAST+AI作为评估预测模型质量、偏倚风险与适用性的更新版工具，全面取代了2019版PROBAST，考虑了AI/ML算法中的关键问题。推荐相关研究者积极学习、广泛采纳该工具，以提升临床预测模型研究质量。

附件见《医学新知》官网（https://yxxz.whuznhmedj.com/futureApi/storage/appendix/202601172.pdf）

伦理声明：不适用
作者贡献：研究设计、论文审阅：靳英辉、阎思宇；资料翻译、图表绘制、示例分析、论文撰写：邹汶廷、郑起程；资料翻译、内容讨论、论文审阅：黄桥、王永博、任相颖、肖时烽；基金支持：靳英辉、王永博
数据获取：本研究中使用和（或）分析的所有数据均包含在文本中
利益冲突声明：无
致谢：不适用

参考文献

1 Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review[J]. BMJ, 2016, 353: i2416.
2 Bellou V, Belbasis L, Konstantinidis AK, et al. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal[J]. BMJ, 2019, 367: l5358.
3 de Munter L, Polinder S, Lansink KW, et al. Mortality prediction models in the general trauma population: a systematic review[J]. Injury, 2017, 48(2): 221-229.
4 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal[J]. BMJ, 2020, 369: m1328.
5 Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations[J]. Nat Med, 2021, 27(12): 2176-2182.
6 Yang Y, Zhang H, Gichoya JW, et al. The limits of fair medical imaging AI in real-world generalization[J]. Nat Med, 2024, 30(10): 2838-2848.
7 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies[J]. Ann Intern Med, 2019, 170(1): 51-58.
8 Schinkel M, Paranjape K, Nannan Panday RS, et al. Clinical applications of artificial intelligence in sepsis: a narrative review[J]. Comput Biol Med, 2019, 115: 103488.
9 Hurkmans C, Bibault JE, Clementel E, et al. Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example[J]. Radiother Oncol, 2024, 194: 110196.
10 Cai Y, Cai YQ, Tang LY, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review[J]. BMC Med, 2024, 22(1): 56.
11 Kaul T, Damen JA, Wynants L, et al. Assessing the quality of prediction models in healthcare using the Prediction Model Risk of Bias Assessment Tool (PROBAST): an evaluation of its use and practical application[J]. J Clin Epidemiol, 2025, 181: 111732.
12 Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods[J]. BMJ, 2025, 388: e082505.
13 Finlayson SG, Beam AL, van Smeden M. Machine learning and statistics in clinical research articles-moving past the false dichotomy[J]. JAMA Pediatr, 2023, 177(5): 448-450.
14 Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions version 6.5 (updated August 2024)[M]. Cochrane: Cochrane Collaboration, 2024.
15 Burdick H, Pino E, Gabel-Comeau D, et al. Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals[J]. BMC Med Inform Decis Mak, 2020, 20(1): 276.
16 刘俊平 . 诊断试验偏倚来源的研究进展[J]. 中国循证医学杂志, 2011, 11(7): 835-840.[Liu JP. Advances in research on bias sources in diagnostic test[J]. Chinese Journal of Evidence-Based Medicine, 2011, 11(7): 835-840.]
17 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II-binary and time-to-event outcomes[J]. Stat Med, 2019, 38(7): 1276-1296.
18 Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I-continuous outcomes[J]. Stat Med, 2019, 38(7): 1262-1275.
19 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model[J]. BMJ, 2020, 368: m441.
20 Moons KGM, Wolff RF, Riley RD, 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-W33.
21 Mäenpää SM, Korja M. Diagnostic test accuracy of externally validated

convolutional neural network (CNN) artificial intelligence (AI) models for emergency head CT scans-a systematic review[J]. Int J Med Inform, 2024, 189: 105523.

22 Elahmedi M, Sawhney R, Guadagno E, et al. The state of artificial intelligence in pediatric surgery: a systematic review[J]. J Pediatr Surg, 2024, 59(5): 774-782.

23 Carrasco-Ribelles LA, Llanes-Jurado J, Gallego-Moll C, et al. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review[J]. J Am Med Inform Assoc, 2023, 30(12): 2072-2082.

24 Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING together initiative[J]. Nat Med, 2022, 28(11): 2232-2233.

25 Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement[J]. BMJ, 2015, 350: g7594.

26 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods[J]. BMJ, 2024, 385: e078378.

收稿日期：2026 年 01 月 30 日 修回日期：2026 年 03 月 15 日
本文编辑：杨室淞 曹 越

引用本文：邹汶廷, 郑起程, 黄桥, 等. PROBAST+AI:基于传统或人工智能方法的预测模型研究质量、偏倚风险及适用性评价工具解读[J]. 医学新知, 2026, 36(7): 721-733. DOI: 10.12173/j.issn.1004-5511.202601172.
Zou WT, Zheng QC, Huang Q, et al. PROBAST+AI: an interpretation of the tool for assessing the quality, risk of bias and applicability of prediction models based on traditional or artificial intelligence methods[J]. Yixue Xinzhi Zazhi, 2026, 36(7): 721-733. DOI: 10.12173/j.issn.1004-5511.202601172.