

PROBAST: 诊断或预后多因素预测模型研究偏倚风险的评估工具



陈香萍¹, 张奕¹, 庄一渝², 章仲恒³

1. 浙江大学医学院附属邵逸夫医院重症医学科 (杭州 310016)

2. 浙江大学医学院附属邵逸夫医院护理部 (杭州 310016)

3. 浙江大学医学院附属邵逸夫医院急诊医学科 (杭州 310016)

【摘要】 本文介绍采用 PROBAST (prediction model risk of bias assessment tool) 工具评估诊断或预后多因素预测模型研究的偏倚风险和适用性, 包括该工具的制定背景、适用范围和使用方法, 同时采用 PROBAST 工具进行了实例分析。该工具主要涉及研究对象、预测因子、结果和统计分析 4 个领域, 通过这 4 个领域对研究的偏倚风险进行评估, 适用性则由前 3 个领域进行判断。PROBAST 工具为诊断或预后多因素预测模型研究的评价提供了规范的评价方法, 从而筛选出合格的文献进行数据分析, 帮助临床决策建立科学基础。

【关键词】 PROBAST; 预测模型; 系统评价; 偏倚风险

PROBAST: a tool for assessing risk of bias in the study of diagnostic or prognostic multi-factorial predictive models

CHEN Xiangping¹, ZHANG Yi¹, ZHUANG Yiyu², ZHANG Zhongheng³

1. Department of Intensive Care Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, P.R.China

2. Department of Nursing, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, P.R.China

3. Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, P.R.China

Corresponding author: ZHUANG Yiyu, Email: zhuangyy@zju.edu.cn

【Abstract】 This study aims to introduce how to use the PROBAST (prediction model risk of bias assessment tool) to evaluate risk of bias and applicability of the study of diagnostic or prognostic predictive models, including the introduction of the background, the scope of application and use of the tool. This tool mainly involves the four areas of participants, predictors, outcomes and analyses. The risk of bias in the research is evaluated through the four areas, while the applicability is evaluated in the first three. PROBAST provides a standardized approach to evaluate the critical appraisal of the study of diagnostic or prognostic predictive models, which screens qualified literature for data analysis and helps to establish a scientific basis for clinical decision-making.

【Key words】 PROBAST; Predictive model; Systematic review; Risk of bias

预测模型是以疾病的预测因子为基础, 按照数学模型计算发生事件概率的方式^[1, 2]。临床预测模型包括诊断预测模型和预后预测模型。诊断预测模型是计算个体患有疾病的概率, 而预后预测模型是预测具有此类事件风险的个体未来发生特定结果或事件的概率或风险^[3]。预测模型有助于临床决策^[4]。目前已开发了越来越多的预测模型, 尤其在

心血管、癌症、糖尿病等领域^[5-8], 有的预测模型被推荐用于临床风险评估^[9, 10]。然而, 已有的预测模型研究存在报告质量差、预测结果不准确、证据相互矛盾和临床应用局限等缺点^[11-14], 导致预测模型存在开发多但应用少的特点^[15, 16]。因此, 对预测模型的研究进行评估和规范有助于提高研究质量, 从而为临床决策提供科学依据。最近, 荷兰wolff等学者开发了针对个体的预后或诊断多因素预测模型研究的偏倚风险评价工具 (prediction model risk of bias assessment tool, PROBAST)^[17], 其可被用来评估多因素预测模型研究的偏倚风险。本文介绍该

DOI: 10.7507/1672-2531.201910087

基金项目: 中华护理学会 2018 年度立项科研课题 (编号: ZHKY201820)

通信作者: 庄一渝, Email: zhuangyy@zju.edu.cn



工具的制定背景、评估方法和评估要点并进行解读,以期提高国内学者对该工具的理解,从而帮助研究者规范使用该工具。

1 PROBAST 工具制定背景和适用范围

PROBAST 是由荷兰乌得勒支大学 Moons Karel 等^[17]按照德尔菲(Delphi)法于 2019 年制定的专用于评价开发、验证或更新预后或诊断多因素预测模型基础研究的工具。该工具包含 4 个领域,涉及 20 个标志性问题,可评价预测模型研究的偏倚风险和适用性。评价者可在 www.probast.org 网站上查询和下载该工具。

PROBAST 工具可用于评估任何类型的诊断或预后预测模型的研究。无论该研究使用何种预测因子、预测何种结果、使用何种建模方法,均可使用该工具进行评价。但该工具不适用于预后因素研究、预测模型的比较研究和预测模型定性研究。

2 PROBAST 工具使用方法

2.1 PROBAST 工具评价步骤

2.1.1 步骤 1: 明确系统评价的问题 评价者首先应明确系统评价的问题,比较与原始研究问题的匹配性。采用 CHARMS (checklist for critical appraisal and data extraction for systematic reviews of prediction modeling studies) 可帮助评价者对系统评价的问题进行定义。

2.1.2 步骤 2: 对预测模型研究进行分类 由于不同的评价问题适用于不同类型的预测模型,因此评价者应先对研究进行分类。预测模型研究可分为开发、验证、开发和验证 3 种类型。建立新模型或者通过向已开发模型添加新预测因子来创建新模型的研究是模型开发研究;使用样本以外的数据评估现有预测模型的预测性能的研究是模型验证研究;通过其他数据进行外部验证的基础上对该模型进行调整或扩展的研究是模型开发和验证研究。

2.1.3 步骤 3: 评估偏倚风险和适用性 该步骤旨在对研究对象、预测因子、结果和统计分析 4 个领域进行评估,每个领域的评估结果采用低、高或不清楚进行判断。偏倚风险的 4 个评估领域每个领域涵盖了 2~9 个标志性问题,每个问题采用“是/可能是”、“可能不是/不是”或“没有信息”进行回答。如果评价者所获得的信息无法非常足够的判断为“是”或“否”,则可使用“可能是”或者“可能不是”进行判断。适用性的评价则由前 3 个领域进行评估,其评估方法与偏倚风险相似,但每个领域没有标志性问题。具体的等级判断标准和标志性问题评估要点分别见表 1 和表 2。

2.1.4 步骤 4: 整体情况判断 评价者应综合步骤 3 中每个领域评估的结果,对预测模型整体的偏倚风险和适用性判断为低、高或不清楚。具体的判断标准见表 1。

2.2 PROBAST 工具各领域条目解读

2.2.1 偏倚风险判断 领域 1: 研究对象。① 标志性问题 1: 对于研究数据来源于随机对照试验或注册数据或前瞻性队列研究、病例队列研究可判断为“是/可能是”。当研究数据来源于现有数据或回顾性队列研究或传统的病例-对照研究时将增加其偏倚风险;但采用巢式病例-对照研究或调整两组基线风险后认为其偏倚风险降低,也可评为“是/可能是”;若无相关信息则判断为“没有信息”。② 标志性问题 2: 评价者应关注纳入与排除标准或研究对象的招募策略是否可能使纳入对象无法代表预期的目标人群。若诊断模型研究的对象结局已知、或不是疑似患者、或预后模型研究的对象已具有患病的风险、或研究对象的选取中直接排除了可能改变预测模型性能的特定亚组,则均可判断为“不是/可能不是”;若纳入与排除标准合理且相对应可判断为“是/可能是”;若研究未报告相关信息,则评为“没有信息”。

领域 2: 预测因子。① 标志性问题 1: 如果对

表 1 PROBAST 偏倚风险和适用性的等级评定标准

评价标准	整体偏倚风险评价标准	整体适用性评价标准	风险等级
所有标志性问题评价结果均为“是/可能是(Y/PY)”	4 个领域的评价结果均为低偏倚风险	3 个领域的评价结果均为低适用性风险	低
≥1 个标志性问题评价结果为“不是/可能不是(N/PN)”	≥1 个领域的评价结果为高偏倚风险或 4 个领域的评价结果均为低偏倚风险但预测模型的开发无外部验证*	≥1 个领域的评价结果为高适用性风险	高
≥1 标志性问题评价结果为“没有信息(NI)”,其它标志性问题评价结果均为“是/可能是(Y/PY)”	≥1 个领域的评价结果为不清楚且其它领域为低偏倚风险	≥1 个领域的评价结果为不清楚且其他领域为低适用性偏倚风险	不清楚

*: 外部验证指利用模型开发中未使用过的数据来评估模型在新数据中的表现,常见形式包括时段验证、空间验证和领域验证^[18]。



表2 PROBAST 各个领域标志性问题的评估要点

评价领域	评价结果		
第一部分：偏倚风险评价			
1 研究对象	<input type="checkbox"/> 低偏倚风险	<input type="checkbox"/> 高偏倚风险	<input type="checkbox"/> 不清楚
1.1 数据来源是否合适？ 是/可能是：来源于随机对照试验、注册数据、前瞻性队列研究、巢式病例-对照研究或病例队列研究。不是/可能不是：来源于非巢式的病例-对照研究或回顾性队列研究。没有信息：未报告研究对象来源。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
1.2 研究对象的纳入与排除标准是否合理？ 是/可能是：纳入与排除标准合理且相对应。不是/可能不是：① 诊断模型：已知研究对象结局或不是疑似病例；② 预后模型：研究对象具有患病风险；③ 排除了可能改变预测模型性能的特定亚组。没有信息：未报告纳入与排除标准相关的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
2 预测因子	<input type="checkbox"/> 低偏倚风险	<input type="checkbox"/> 高偏倚风险	<input type="checkbox"/> 不清楚
2.1 预测因子的定义和评估对所有研究对象是否相同？ 是/可能是：预测因子的定义和评估对所有研究对象均相同。不是/可能不是：预测因子采用不同的定义和评估。没有信息：未报告预测因子定义和评估的相关信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
2.2 是否是在不清楚结果数据的情况下评估预测因子？ 是/可能是：评估者在评估预测因子时不清楚结果信息。不是/可能不是：评估预测因子时使用了结果信息。没有信息：未报告是否在不了解结果信息的情况下评估预测因子。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
2.3 预测模型所包含的预测因子是否有效？ 是/可能是：预测模型所包含的预测因子均有效。不是/可能不是：预测模型所包含的预测因子无效。没有信息：未报告预测因子在预测模型应用时是否有效的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3 结果	<input type="checkbox"/> 低偏倚风险	<input type="checkbox"/> 高偏倚风险	<input type="checkbox"/> 不清楚
3.1 结果的分类方法是否合理？ 是/可能是：指南或期刊认可的分类方法。不是/可能不是：采用了次优的分类方法而导致分类错误。没有信息：未报告结果分类方法的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3.2 结果的定义是否合理？ 是/可能是：结果的定义是客观的或使用标准的结果定义。不是/可能不是：结果的定义不是标准的且未预先设定。没有信息：未报告结果的定义是否标准的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3.3 结果的定义是否排除了预测因子？ 是/可能是：结果定义中没有包含预测因子。不是/可能不是：结果定义中包含了≥1个预测因子。没有信息：未报告结果的定义是否排除预测因子的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3.4 结果的定义对所有研究对象是否相同？ 是/可能是：结果的定义对研究对象均相同。不是/可能不是：结果的定义对研究对象不同。没有信息：未报告结果的定义对所有研究对象是否相同的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3.5 确定结果时是否不清楚预测因子的信息？ 是/可能是：确定结果时不清楚预测因子的信息。不是/可能不是：在确定结果时使用了预测因子信息。没有信息：未报告结果确定时是否不清楚预测因子的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
3.6 预测因子评估和结果确定的时间间隔是否合理？ 是/可能是：预测因子评估和结果确定的时间间隔合理。不是/可能不是：预测因子评估和结果确定的时间间隔太短或太长导致无法正确记录结果类别和样本量。没有信息：未报告预测因子评估和结果确定的时间间隔的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
4 统计分析	<input type="checkbox"/> 低偏倚风险	<input type="checkbox"/> 高偏倚风险	<input type="checkbox"/> 不清楚
4.1 样本量是否合理？ 是/可能是：① 预测模型开发研究：EPV ≥ 20；② 预测模型验证研究：样本量 ≥ 100。不是/可能不是：① 预测模型开发研究：EPV < 10；② 预测模型验证研究：样本量 < 100。没有信息：① 模型开发研究中无法计算 EPV 值；② 模型验证研究中未报告样本量信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
4.2 对连续和分类的自变量的处理是否合适？ 是/可能是：① 预测模型开发研究：连续变量未转化成 ≥ 2 个类别的变量或已检验连续变量的非线性拟合或分类变量采用标准定义；② 预测模型验证研究：连续变量采用相同的定义和转换方式或分类变量采用相同的切割点。不是/可能不是：① 模型开发研究：连续变量转化成 ≥ 2 类别的变量；② 模型验证研究：连续变量采用不同的定义和转化方式或分类变量采用不同的切割点。没有信息：未报告连续变量或分类变量的相关信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息
4.3 所有纳入对象是否均包含在统计分析中？ 是/可能是：所有纳入对象均包含在数据分析中。不是/可能不是：部分纳入对象被不合理的排除。没有信息：未报告所有纳入对象是否包含在统计分析中的信息。	<input type="checkbox"/> 是/可能是	<input type="checkbox"/> 不是/可能不是	<input type="checkbox"/> 没有信息

续表 2

评价领域	评价结果
4.4 缺失数据的纳入对象是否进行合适的处理? 是/可能是: ① 没有遗漏的预测因子或结果, 并且纳入对象未因数据缺失而被排除; ② 采用多重插补法处理缺失数据。不是/可能不是: ① 直接排除缺失数据的纳入对象; ② 缺失数据的处理方法存在明显缺陷; ③ 未明确提及处理缺失数据的方法。没有信息: 没有足够的信息来确定处理缺失数据的方法是否合适。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
*4.5 是否避免了采用单因素分析法筛选预测因子? 是/可能是: 未基于单因素分析法筛选预测因子。不是/可能不是: 基于单因素分析法筛选预测因子。没有信息: 未报告是否避免采用单因素分析法筛选预测因子的信息。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
4.6 数据的复杂性是否进行考虑? 是/可能是: 数据的复杂性都得到了合理的解释或者一些数据的复杂性被确认为不重要。不是/可能不是: 未对复杂性数据进行说明解释。没有信息: 未报告数据中的复杂性是否存在或适当说明的信息。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
4.7 是否对预测模型性能进行评估? 是/可能是: 正确评估预测模型的区分度和校准度。不是/可能不是: ① 校准度和区分度未进行评估; ② 仅使用 Hosmer-Lemeshow 拟合优度检验校准度。没有信息: 未报告预测模型的区分度或校准度的信息。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
*4.8 是否对预测模型过度拟合、欠拟合和最优拟合进行考虑? 是/可能是: 正确使用内部验证法且对模型性能的后续调整进行评估。不是/可能不是: ① 未使用内部验证; ② 内部验证仅包含数据的随机拆分验证法; ③ 未包括中介效应检验或交叉验证。没有信息: 未报告是否使用内部验证法的信息。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
*4.9 预测因子及其权重是否与报告的结果一致? 是/可能是: 预测因子的回归系数与报告结果一致。不是/可能不是: 预测因子的回归系数与报告结果不一致。没有信息: 未报告回归系数是否与报告结果一致的信息。	<input type="checkbox"/> 是/可能是 <input type="checkbox"/> 不是/可能不是 <input type="checkbox"/> 没有信息
整体偏倚风险评价: <input type="checkbox"/> 低偏倚风险 <input type="checkbox"/> 高偏倚风险 <input type="checkbox"/> 不清楚	
第二部分: 适用性评价	
1. 研究对象 低适用性风险: 原始研究纳入对象和临床设计与系统评价问题相符。高适用性风险: 原始研究纳入对象和临床设计与系统评价研究问题不同。不清楚: 未报告研究对象和临床设计的相关信息。	<input type="checkbox"/> 低适用性风险 <input type="checkbox"/> 高适用性风险 <input type="checkbox"/> 不清楚
2. 预测因子 低适用性风险: 原始研究预测因子的定义、评估和评估时间与系统评价问题相符。高适用性风险: 原始研究预测因子的定义、评估和评估时间与系统评价问题不同。不清楚: 未报告预测因子的信息。	<input type="checkbox"/> 低适用性风险 <input type="checkbox"/> 高适用性风险 <input type="checkbox"/> 不清楚
3. 结果 低适用性风险: 结果的定义、时间间隔和分析方法与系统评价问题相符。高适用性风险: 结果的定义、时间间隔和分析方法是系统评价问题所预期的其它结果。不清楚: 未报告结果的定义、时间间隔和分析方法的信息。	<input type="checkbox"/> 低适用性风险 <input type="checkbox"/> 高适用性风险 <input type="checkbox"/> 不清楚
整体适用性评价: <input type="checkbox"/> 低适用性风险 <input type="checkbox"/> 高适用性风险 <input type="checkbox"/> 不清楚	

★: 该条目仅限于模型开发研究; EPV: 每个自变量的事件数 (events per variable), 指研究对象中较少组的数量除以自变量的个数; 区分度: 是指模型区分是否患有待诊断的疾病 (诊断模型) 或是否发生预期的事件 (预后模型) 的能力, 也就是将患者按照风险的大小进行排序的能力; 校准度: 评估预测的概率与实际观察到的概率的一致性。

不同的研究对象采用了不同的预测因子定义和评估方法, 可能将由于以上差异影响结果引起较高的偏倚风险。若对不同研究对象采用的预测因子的定义和评估方法均相同, 可判断为“是/可能是”; 相反则为“不是/可能不是”; 若研究未报告相关信息, 则评为“没有信息”。② 标志性问题 2: 未采用盲法会产生结果信息影响预测因子评估过程的风险, 尤其在诊断模型研究中, 这种偏倚情况较多, 而在前瞻性队列研究中较少。若评估者在评估预测因子时不清楚结果信息, 可判断为“是/

可能是”; 相反则为“不是/可能不是”; 若研究未报告相关信息, 则评为“没有信息”。③ 标志性问题 3: 要使预测模型在现实环境中可用, 所有纳入的预测因子必须是有效的, 因此要明确该模型纳入的所有预测因子是否有统计学意义及预测模型中是否包含了缺少数据的预测因子。若预测模型所包含的预测因子均有效, 可判断为“是/可能是”; 相反则为“不是/可能不是”; 若研究未报告分类方法的信息, 则为“没有信息”。

领域 3: 结果。① 标志性问题 1-2: 该问题旨

在评估是否存在结果分类错误和非标准的结果定义而导致的偏倚风险。结果分类错误和非标准的结果定义将可能导致回归系数、截距(逻辑回归和参数生存模型)或基线风险(Cox 回归模型)的偏差。尤其当数据是来源于登记数据或现有的研究数据时,结果分类方法和定义的不统一很可能产生偏倚风险。若研究中结果的分类方法和定义已经写入临床指南或被已发表的研究证实,可判断为“是/可能是”;若产生了错误地分类或者使用了非标准阈值来定义结果,则为“不是/可能不是”;若研究未报告相关信息,则评为“没有信息”。② 标志性问题 3: 如果模型所包含的预测因子是结果定义中相关因素的一部分,则预测因子和结果之间的关联性可能被高估。若预测的结果定义不包含预测因子,则可判断为“是/可能是”;相反则为“不是/可能不是”;若研究未报告相关信息,则判断为“没有信息”。③ 标志性问题 4-5: 该问题旨在评估不同对象对结果的定义是否相同及是否使用盲法在不清楚预测因子信息的情况下确定结果。其判断方法和领域 2 中标志性问题 1 和 2 的评估方法相似。④ 标志性问题 6: 这个问题的判断需要临床专业知识来明确适当的时间间隔,预测因子和预测结果的评估最好在第一时间点进行。但在研究实施中,预测因子和结果的评估之间会有一段时间间隔,这可能导致诊断结果改善或恶化;尤其在急性传染病中,即使短暂的时间延迟也将产生偏倚。若预测因子评估和结果确定的时间间隔合理,可判断为“是/可能是”;若由于时间间隔而无法获取具有代表性的样本量或正确记录结果类型,则评为“不是/可能不是”;若研究未报告相关信息,则判断为“没有信息”。

领域 4: 统计分析。由于其中一些评估内容需要统计方面的专业知识,因此建议至少由 1 名在预测模型研究中具有统计专业知识的研究人员共同完成这个领域的评估。① 标志性问题 1: 对于不同的预测模型研究类型,样本量评价方法不同。对于模型开发研究,通过每个自变量的事件数(events per variable, EPV),即研究对象中较少组的数量除以自变量的个数进行评价。若模型开发研究中 $EPV \geq 20$ 例或模型验证研究中样本量 ≥ 100 例,可判断为“是/可能是”;相反则为“不是/可能不是”;若研究未报告相关信息,则判断为“没有信息”。② 标志性问题 2: 对于不同的预测模型研究类型,自变量处理方法不同。若在模型开发研究中连续性变量未转化成 ≥ 2 个类别的变量、或采用

分数多项式等方法检验了连续变量的非线性拟合情况、或分类变量采用标准的定义,在模型验证研究中连续变量采用相同的定义和转换方式或分类变量采用相同的切割点进行分类,则可判断为“是/可能是”;若模型开发研究中连续性变量转化成 ≥ 2 类别的变量、或在模型验证研究中连续变量采用不同的定义和转化方式、或分类变量采用不同的切割点进行分类,则评为“不是/可能不是”。

③ 标志性问题 3: 该问题旨在评价是否将合格的研究对象从研究中不合理地排除。但这与不恰当的纳入排除标准或结果中缺失数据的处理无关。除了由于无法解释的原因产生异常值或结果中的数据缺失而剔除部分研究对象外,还存在部分对象被剔除,则可判断为“不是/可能不是”。④ 标志性问题 4: 当研究过程中简单地处理纳入对象的缺失数据,将导致一定的偏倚风险。无论是在模型开发或验证研究中,多重插补在偏差和精度方面均优于其他方法。若研究中没有遗漏的预测因子或结果,且纳入对象未因数据缺失而被排除或者采用了多重插补法处理缺失数据,则可判断为“是/可能是”;若研究中直接剔除缺失数据、或缺失数据的处理方法存在明显缺陷、或未明确提及处理缺失数据的方法,则判断为“不是/可能不是”。⑤ 标志性问题 5: 仅对模型的开发研究需要进行该问题的判断。由于单因素分析法筛选的预测因子是仅根据其统计学意义($P < 0.05$)选择的,并未与其他自变量结合进行分析,而其中一些变量需调整其他变量才具有意义,因此将会因自变量的遗漏而产生偏倚。若研究中未基于单因素分析法筛选预测因子,则可判断为“是/可能是”;相反则为“不是/可能不是”;若研究未报告相关信息,则判断为“没有信息”。⑥ 标志性问题 6: 该问题旨在评估是否对研究中关键数据的复杂性问题进行了考虑,例如终检、竞争风险、控制参与者抽样等问题。若数据的复杂性都得到了合理解释或者一些数据的复杂性被确认为不重要,则可判断为“是/可能是”;相反则为“不是/可能不是”;若研究未报告相关信息,则判断为“没有信息”。⑦ 标志性问题 7: 要全面评估预测模型的预测性能,评价者必须评估研究中模型的校准度和区分度。若研究中带有校准图或表格,需要判断图或表格是否正确,同时要关注用于评估模型校准度和区分度的方法是否适用于该模型研究;若研究中没有校准图或表格,仅报告校准的统计数据或仅使用 Hosmer-Lemeshow 拟合优度检验校准度,可判断为“不是/可能不

是”。⑧ 标志性问题 8-9：仅对模型的开发研究需要进行该问题的判断。评价者需要对模型开发研究的拟合情况进行评估，若研究中正确使用内部验证法且对模型性能的后续调整进行评估，可判断为“是/可能是”；若未使用内部验证或内部验证仅包含数据的随机拆分验证法或未包括中介效应检验或交叉验证，可判断为“是/可能是”。此外，我们需要评估开发研究中最终模型预测因子的系数或截距是否与多变量分析报告的结果对应。若两者相对应，则可判断为“是/可能是”，相反则为“不是/可能不是”；若研究未报告相关信息，则判断为“没有信息”。

2.2.2 适用性评价 领域 1：研究对象。该领域的适用性是指原始研究中纳入的研究对象与系统评价需评价的研究对象的匹配程度。基于随机试验的预测模型研究，由于严格的纳入与排除标准和较少的预测因子可能导致其低适用性，相比之下来源于登记数据的研究其适用性更高。若原始研究纳入对象和临床设计与系统评价问题相符，则可判断为“低适用性风险”；相反则为“高适用性风险”；若研究未报告相关信息，则判断为“不清楚”。

领域 2：预测因子。如果纳入研究使用不同的定义或评估方法或评估时间，将导致某些模型研究的预测性能可能会与其它研究结果不同，而引起适用性风险问题。若三者均相符，可判断为“低适用性风险”；相反则为“高适用性风险”；若研究未报告相关信息，则判断为“不清楚”。

领域 3：结果。该领域的适用性评价方法和领域 2 的判断方法相似。若原始研究结果的定义、评估方法和评估时间与系统评价问题均相符，则可判断为“低适用性风险”；相反则为“高适用性风险”；若研究未报告相关信息，则判断为“不清楚”。

3 实例分析

本研究以《慢性阻塞性肺疾病急性加重期患者短期预后预测模型的建立》（以下简称“AECOPD 研究”）^[19]为例，采用 PROBAST 工具对其进行偏倚风险和适用性评价。

首先假定构建“慢性阻塞性肺疾病急性加重

期患者危重情况预测模型的系统评价”，原始研究问题与系统评价问题见表 3，由表 3 可知，问题匹配。其次，AECOPD 研究中创建了新的预后模型，但未通过其他数据进行外部验证，因此属于“模型开发”研究。第三，对其进行偏倚风险评价和适用性评价。在研究对象领域中，因为 AECOPD 研究的数据来源于前瞻性队列研究，纳入与排除标准合理且对应，所以该领域的 2 个标志性问题均判断为“Y”，该领域偏倚风险判断为低偏倚风险。在预测因子领域中，AECOPD 研究中评估人员经过一对一的 3 轮统一培训，评估方法相同，并采用 3 种相同的评估工具（一般资料调查表、英国国家早期预警评分、Charlson 合并症评估表）对研究对象进行评估，且在评估过程中并不清楚结果数据；模型所纳入的 3 个预测因子均具有统计学意义（ $P < 0.05$ ）。因此在该领域中的 3 个标志性问题均判断为“Y”，该领域偏倚风险为低偏倚风险。在结果领域中，AECOPD 研究中未报告病情危重和病情平稳的结果分类方法和结果定义方法的相关信息；未报告结果中是否包含预测因子；对所有研究对象均采用了相同的结果定义；研究者在确定结果时不清楚预测因子的信息；未报告预测因子评估和结果确定的时间间隔的信息。因此该领域的标志性问题 1、2、3、6 判断为“NI”，标志性问题 4、5 判断为“Y”，该领域判断为高偏倚风险。在统计分析领域中，AECOPD 研究中 EPV 值为 15（病情危重组样本量 $n=45$ /预测因子数 $n=3$ ）小于 20；5 个连续性变量均转化为了二分类变量；统计分析中包含了所有纳入对象和研究数据；采用了单因素分析法进行预测因子的筛选；未报告数据是否存在复杂性；仅使用了 Hosmer-Lemeshow 拟合优度检验评估校准度；未报告内部验证方法是否已经采用的信息；3 个预测因子的回归系数（2.014、2.738、3.353）与结果报告中（2 分、3 分、3 分）相对应。因此该领域的标志性问题 1、2、5、7 判断为“N”，标志性问题 3、4、9 判断为“Y”，标志性问题 6、8 判断为“NI”，该领域判断为高偏倚风险。由于系统评价内容不明确，对于 AECOPD 研究无法进行适用性风险评估。第四，对该研究进行

表 3 AECOPD 研究的问题分析

类别	系统评价的问题	原始研究的问题
预期用途	预后	短期的预后；急诊中
研究对象	慢性阻塞性肺疾病急性加重期患者	急诊科慢性阻塞性肺疾病急性加重期患者
预测因子	包含临床检查、人口学资料、生化指标等类别	年龄、英国国家早期预警评分、Charlson 合并症指数
预测结果	病情危重情况	短期内病情危重程度



整体情况判断。对 AECOPD 研究的 4 个偏倚风险领域分别判断为低、低、高和高,因此该研究整体偏倚风险为高偏倚风险。但由于系统评价内容不明确,对于 AECOPD 研究无法进行整体的适用性风险评估。

4 讨论

评估纳入研究的偏倚风险是系统评价的重要组成部分。随着目前越来越多的预测模型开发和应用,则需要相关工具进行更为规范化的评价来提高研究质量和循证决策的科学性。PROBAST 是第一个专门用于评价诊断或预后模型的开发、验证或更新研究的偏倚风险的工具^[20],其经过严格的过程开发,但在评价过程中还存在一些需要注意的问题。

由于预测模型研究可分为不同的类型,故使用 PROBAST 工具评价时存在一定的差异性。虽然均可使用 PROBAST 工具对 3 种类型的研究进行评价,但模型的开发应包括采用自举法或交叉验证等进行内部验证,而模型的验证是不同或相同研究者收集不同参与者的数据进行外部验证。同时存在以下差别:①表 2 中的 3 个标志性问题(4.5、4.8 和 4.9)对“模型验证”研究是不需要评价的。②在“统计分析”领域的标志性问题 1 和 2 中,不同的研究类型采用不同的评价方法。

对于诊断或预后两种不同作用的预测模型,在 PROBAST 评价上也具有一定区别:①采用了不同术语来表示预测因子、结果和缺失值评估。诊断和预后模型对于预测因子分别采用诊断试验和预后因素进行描述,对于结果分别采用参考标准(金标准)和事件(例如死亡,疾病复发,疾病并发症或治疗反应等)来描述,对于缺失值评估分别采用部分证实和失访来描述。②对于一些标志性问题的判断方法不同。尤其在研究对象领域和统计分析领域,由于不同类型的预测模型其纳入对象的研究设计方案和结局指标具有较大差异,需要针对该类型进行具体评价。

评价者还应注意每个问题和领域对于整体评价的影响是不同的。尤其在工具的评价涉及多个领域,而每个领域下又有多个问题时,每个问题又包含多个选项的情况下,这方面的问题就显得更加重要了。如偏倚风险中对结果领域的评估,评价者应仔细评价那些决定结果的人是否可获得预测信息,如果信息出现在结果确定过程中或者如果不清楚,则在对该领域的偏倚风险进行总体判断时,该

信息的影响更大,而不能盲目地根据表 2 中常规的判断方法进行评价。

通过本文的实例分析可看出,该工具还存在不足之处:①使用工具评价时,每个领域需回答 2~9 个标志性问题,根据问题的评价结果判断出每个领域结果后还需对偏倚风险和适用性进行总体判断,使得评价过程过于复杂和耗时。②由于每个问题和领域对于整体评价的影响大小不同,而评价者需要综合多项结果进行整体判断,而这个过程缺乏清晰的指导供评价者参考。③由于该工具是第一个开发的工具,其信度效度、适用性及推广情况还需进一步检验。

总之,本文对 PROBAST 工具进行了介绍,希望能够提高研究者对该工具的理解,从而帮助研究者规范使用该工具。

参考文献

- 1 张蕊,郑黎强,潘国伟. 疾病发病风险预测模型的应用与建立. 中国卫生统计, 2015, 32(4): 724-726.
- 2 Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model. *Eur J Cardiothorac Surg*, 2018, 54(2): 203-208.
- 3 Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*, 2014, 11(10), e1001744.
- 4 Haider S, Sadiq SN, Moore D, et al. Prognostic prediction models for diabetic retinopathy progression: a systematic review. *Eye (Lond)*, 2019, 33(5): 702-713.
- 5 陈金宏,何耀,秦银河. 心血管病预测模型研究进展. *中华流行病学杂志*, 2010, 31(11): 1311-1314.
- 6 张佳妮,于诗嘉,史铁梅. 风险预测模型诊断子宫内膜癌的研究进展. *中国临床医学影像杂志*, 2009, 20(2): 115-116.
- 7 廖文圣,邵彦翔,孙光曦,等. 预后预测模型在临床非转移肾细胞癌中的应用价值. *四川大学学报(医学版)*, 2019, 50(5): 720-724.
- 8 徐泉喻,王梅新,汪惠才. 糖尿病前期人群筛查方法与风险预测模型的研究进展. *护理研究*, 2017, 31(35): 4468-4471.
- 9 Rabar S, Lau R, O'Flynn N, et al. Risk assessment of fragility fractures: summary of NICE guidance. *BMJ*, 2012, 345: e3698.
- 10 Jr GD, Lloydjones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*, 2014, 63(25): 2935-2959.
- 11 Grady D, Berkowitz SA. Why is a good clinical prediction rule so hard to find? *Arch Intern Med*, 2011, 171(19): 1701-1702.
- 12 Meier K, Parrish J, D'Souza R. Prediction models for determining the success of labor induction: a systematic review. *Acta Obstet Gynecol Scand*, 2019, 98(9): 1100-1112.
- 13 Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*, 2017, 356: i6460.
- 14 Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting

- of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg*, 2015, 102(3): 148-158.
- 15 陶立元, 刘珏, 曾琳, 等. 针对个体的预后或诊断多因素预测模型报告规范(TRIPOD)解读. *中华医学杂志*, 2018, 98(44): 3556-3560.
- 16 郑黎强, 张蕊. 疾病发病风险预测模型拟合度评价方法的研究进展. *中国卫生统计*, 2015, 32(3): 544-546.
- 17 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*, 2019, 170(1): 51-58.
- 18 王俊峰, 章仲恒, 周支瑞, 等. 临床预测模型: 模型的验证. *中国循证心血管医学杂志*, 2019, 11(2): 141-144.
- 19 安莹, 王艳玲. 慢性阻塞性肺疾病急性加重期患者短期预后预测模型的建立. *中华护理杂志*, 2019, 54(1): 42-46.
- 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. *Ann Intern Med*, 2019, 170(1): W1-W33.

收稿日期: 2019-10-18 修回日期: 2020-02-08

本文编辑: 张洋/蔡羽嘉

