·方法学指南解读·

《药物流行病学研究方法学指南（第2版）》系列解读（10）：质量评估工具介绍与案例分析

郑卿勇[1, 2#]，徐彩花[1, 2#]，周泳佳[1, 2]，唐　筱[3]，张梦君[4]，祁金智[3]，刘　明[5]，高　亚[6, 7]，孙　凤[8, 9, 10, 11, 12, 13]，田金徽[1, 2, 14]

1. 兰州大学循证医学中心，兰州大学基础医学院（兰州 730000）
2. 甘肃省循证医学重点实验室（兰州 730000）
3. 兰州大学第一临床医学院（兰州 730000）
4. 河北大学护理学院（河北保定 071000）
5. 新加坡南洋理工大学李光前医学院（新加坡 308232）
6. 山东大学齐鲁医学院公共卫生学院医学数据学系（济南 250012）
7. 国家健康医疗大数据研究院（济南 250003）
8. 北京大学公共卫生学院流行病与卫生统计学系（北京 100191）
9. 重大疾病流行病学教育部重点实验室（北京大学）（北京 100191）
10. 北京大学医学部药品上市后安全性研究中心（北京 100191）
11. 北京大学第三医院眼科（北京 100191）
12. 新疆医科大学中医学院（乌鲁木齐 830017）
13. 新疆石河子大学医学院（新疆石河子 832000）
14. 甘肃省中医院风湿骨病中心（兰州 730050）

【摘要】方法学质量评估是连接原始研究与可靠循证证据的关键环节，也是将《药物流行病学研究方法学指南（第 2 版）》核心原则付诸实践的必要路径。然而，实践中对研究方法学稳健性的判断（偏倚风险评估）常与对其报告透明度的核查（报告规范遵循）相混淆。本文系统性地回应了这一挑战，并深入辨析两者的内涵与分野。在此基础上，研究对主流质量评估工具进行了全面梳理，涵盖了针对干预性研究（如 RoB 2、ROBINS-I）、观察性研究（如 NOS、JBI/SIGN/NIH 系列）、二次研究（如 AMSTAR 2）及卫生经济学评价等特定类型工具的方法学特点与演进脉络，并通过一项完整的案例剖析了 ROBINS-I 工具的实践应用。本文倡导“以评促建”的方法学思想，研究者应将评估标准内化为设计准则，从对已有文献的“回溯性批判”升维至对新开展研究的“前瞻性质控”，同时探讨了人工智能在辅助评估中的新兴范式。本文可为研究者和实践者在繁多的工具中进行审慎选择、并对药物流行病学证据进行严谨的批判性评价，提供一份全面的方法学参考。

【关键词】药物流行病学；方法学；指南；质量评估工具；偏倚风险

【中图分类号】R 181.3+5　　**【文献标识码】**A

DOI: 10.12173/j.issn.1005-0698.202509141

[#] 共同第一作者

基金项目：国家自然科学基金面上项目（72474008）；国家自然科学基金国际（地区）合作与交流项目（72361127500）；甘肃省科技重大专项－社会发展领域（24ZD17FA003）

通信作者：孙凤，博士，研究员，博士研究生导师，Email：sunfeng@bjmu.edu.cn
田金徽，博士，教授，博士研究生导师，Email：tjh996@163.com

Guide on Methodological Standards in Pharmacoepidemiology (2nd edition) and their series interpretation (10): an overview and case study of quality assessment tools

ZHENG Qingyong[1,2#], XU Caihua[1,2#], ZHOU Yongjia[1,2], TANG Xiao[3], ZHANG Mengjun[4], QI Jinzhi[3], LIU Ming[5], GAO Ya[6,7], SUN Feng[8,9,10,11,12,13], TIAN Jinhui[1,2,14]

1. Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou 730000, China

2. Key Laboratory of Evidence-based Medicine of Gansu Province, Lanzhou 730000, China

3. The First School of Clinical Medicine, Lanzhou University, Lanzhou 730000, China

4. College of Nursing, Hebei University, Baoding 071000, Hebei Province, China

5. Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore

6. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, China

7. National Institute of Health and Medical Big Data, Jinan 250003, China

8. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China

9. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

10. Center of Post-Marketing Safety Evaluation of Drugs, Peking University Health Science Center, Beijing 100191, China

11. Department of Ophthalmology, Peking University Third Hospital, Beijing 100191, China

12. College of Traditional Chinese Medicine, Xinjiang Medical University, Urumqi 830017, China

13. School of Medicine, Shihezi University, Shihezi 832000, Xinjiang Uygur Autonomous Region, China

14. Center for Rheumatology and Musculoskeletal Health, Gansu Provincial Hospital of Traditional Chinese Medicine, Lanzhou 730050, China

#Co-first authors: ZHENG Qingyong and XU Caihua

Corresponding authors: SUN Feng, Email: sunfeng@bjmu.edu.cn; TIAN Jinhui, Email: tjh996@163.com

【Abstract】Methodological quality assessment is a pivotal link between primary studies and reliable evidence-based practice, and an essential pathway for operationalizing the core principles of the *Guide on Methodological Standards in Pharmacoepidemiology (2nd edition)*. A prevalent challenge in practice, however, is the conflation of appraising methodological robustness (risk of bias assessment) with verifying reporting transparency (adherence to reporting guidelines). This paper systematically addresses this fundamental challenge, beginning with a clear distinction between the essence and boundaries of these two concepts. On this basis, the article provides a comprehensive review of mainstream quality assessment tools, covering the methodological features and evolutionary trajectory of numerous instruments for interventional (e.g., RoB 2, ROBINS-I), observational (e.g., NOS, the JBI/SIGN/NIH series), secondary (e.g., AMSTAR 2), and other specific types of studies such as health economic evaluations. Furthermore, a complete case study is used to illustrate the practical application of the ROBINS-I tool. The paper's central thesis advocates for an "appraisal-informed design" philosophy, urging a conceptual shift from the retrospective critique of existing literature to the prospective quality control of new research by internalizing appraisal standards as design principles, while also exploring the emerging paradigm of artificial intelligence in assisting assessment. This paper provides a comprehensive methodological reference for researchers and practitioners to prudently select appropriate assessment tools and to conduct rigorous critical appraisals of pharmacoepidemiological evidence.

【Keywords】Pharmacoepidemiology; Methodology; Guildelines; Quality assessment tools; Risk of bias

高质量的药物流行病学研究是支撑临床决策、药品监管及卫生政策制定的证据基础。新近发布的《中国药物流行病学研究方法学指南（第2版）》（以下简称“指南第2版”）为提升我国该领域的研究质量，构建了一套系统的方法学框架[1-2]。通览指南第2版，其核心宗旨在于保障研究结论的真实性与可靠性，对各类偏倚的识别与控制贯穿了研究从设计到解读的全过程。这为研究者如何构思并执行一项严谨的研究，提供了明确指引。

然而，在研究证据的应用层面，文献的产出仅是起点。对于证据的审阅者——无论是系统评价的执行者、临床实践的决策者，还是新药价值的评估者——其面临的核心任务是对大量的已有研究进行筛选与评判。此时，一个关键的方法学问题便凸显出来：应如何将指南所倡导的原则转化为一套客观、可重复的标准化流程，用以评价一项研究的内在质量？这即是方法学质量评估的价值所在。

值得注意的是，指南第2版在强调研究设计严谨性的同时，也倡导遵循国际通行的报告规范以提升研究的透明度[3]。但必须明确，报告的完整性并不必然等同于方法学的稳健性。一项陈述清晰、要素齐全的研究，仍可能因其内在的设计缺陷而存在严重的偏倚风险。因此，在确保研究报告透明度的基础上，如何深入评价其方法学的严谨性，便成为一项更为关键的任务。

为回应这一需求，本文将从辨析方法学质量、偏倚风险及研究报告规范等核心概念入手，进而对国际上针对不同研究设计的主流质量评估工具进行系统梳理，并通过一个具体案例剖析评估工具在实践中的应用，以期为我国学者在研究设计时前瞻性地规避方法学缺陷，在文献审阅阶段批判性评估证据可靠性提供参考。

1 质量评估的理论基础与核心概念辨析

1.1 药物流行病学研究中质量评估的必要性

药物流行病学研究的价值最终取决于其结论的可靠性[4-5]。然而在实践中，针对同一问题的研究常因方法学严谨性的差异而呈现异质性、甚至矛盾的结论[2]。这一现实，使得超越传统“证据金字塔”对研究设计类型的标签化评判，深入审视每一项研究的内在质量，成为甄别可信证据的根本要求。事实上，一项设计严谨、执行得当的大型观察性研究，其结论的可靠性可能远超一项存在严重偏倚风险的随机对照试验（randomized controlled trial，RCT）[6]。这一原则不仅适用于单篇文献的批判性阅读，其重要性在证据的系统性综合中更是被进一步放大。系统评价与Meta分析作为证据合成的核心方法，其结论的稳健性直接取决于被纳入原始研究的质量。在此过程中，方法学质量评估不仅是筛选研究的“准入门槛”，更是后续解释研究间异质性、开展敏感性分析的关键依据[7]。若忽视此步骤，将不同质量的研究混同分析，则合成证据的可靠性将从根本上受到动摇。

1.2 核心概念辨析：偏倚风险与报告规范

对研究证据进行批判性审阅，首先需明确区分报告规范与偏倚风险评估这两个既相互关联又存在本质差异的概念。报告规范（如针对观察性研究的STROBE声明[8]或随机试验的CONSORT声明[9]）的核心宗旨在于提升研究的透明度，其通过提供结构化的清单与流程图，来标准化研究方法的叙述与结果的呈现，旨在确保研究设计、执行与分析过程被完整披露，为后续的严谨评价提供必要的信息基础。因此，报告规范关注的是研究报告的完整性与清晰度，其角色是描述性的，而非评判性的。

然而，报告的透明度本身并不等同于方法学的稳健性。一项遵循了完备报告规范的研究，可能只是清晰地陈述了其内在的方法学缺陷[10-11]，而报告本身无法修正这些缺陷。识别并量化这些缺陷对研究结论的潜在影响，正是偏倚风险评估的核心任务。偏倚风险评估工具，如纽卡斯尔－渥太华量表（Newcastle-Ottawa Scale，NOS）[12]或非随机干预性研究偏倚风险评估（risk of bias in non-randomized studies-of interventions，ROBINS-I）[13-14]，其功能在于超越对信息呈现的核查，深入对方法学本身的批判性审视，以判断研究结论偏离“真实”的风险大小。简言之，在证据评价的实践中，报告规范是确保信息完整的前提，它使得后续的偏倚风险评估成为可能，但绝不能替代后者。前者旨在确保研究能够被准确理解，而后者则致力于判断其结论是否可信。

1.3 质量评估在研究与实践中的应用场景

方法学质量评估的应用贯穿于证据的审阅、合成与生产全过程，其价值在以下几个关键场景中逐层体现（表1）。

1.3.1 个体研究的批判性阅读

对单篇文献的批判性审阅是质量评估最基础、最直接的应用。评估工具为此提供了一套超越研究结论、直击方法学内核的系统化框架。以Cochrane手册为例，其推荐的RoB（risk of bias）2工具要求评价者必须检视一项RCT在随机化过程、干预偏离、数据缺失、结局测量及结果选择性报告等多个领域的偏倚风险[15]。因此，即便一项研究报告了统计学显著的“阳性”结果，若其在分配隐藏或盲法等关键环节存在严重缺陷，其结论的可信度也应受到质疑[16]。审阅的核心在于，需针对不同研究类型（如观察性研究、诊断试验）选用恰当的评估工具。例如，除了应用NOS对队列研究进行经典评价外，还可以借助美国国立卫生研究院（National Institutes of Health，NIH）制定的清单更细致地审视其参与率和混杂控制等问题[17]，以精准识别各自独特的偏倚来源。

1.3.2 证据合成中的关键作用

对原始研究的严谨评价构成了系统评价与Meta分析证据的基石。在证据合成过程中，对原始研究的偏倚风险评估，直接影响研究的纳入与排除、研究间异质性的解释以及敏感性分析的执行。在高影响力期刊发表的Meta分析中，相当一部分通过剔除“高偏倚风险”研究来检验其结论的稳健性[18]。不同的Meta分析常因其纳入研究的质量标准不同而得出迥异的结论，van Dijk等[19]曾剖析过2个结论截然相反的Meta分析，发现对原始研究质量的不同考量，是导致最终结论分歧的关键原因之一。

1.3.3 临床指南的证据评级

证据合成的一个关键应用，即为临床指南的证据评级与推荐意见制定提供依据。现代临床指南普遍采用建议评估、制定和评估分级（Grading of Recommendations Assessment, Development and Evaluation，GRADE）体系，其核心是对证据体进行系统的质量评价，其中原始研究的偏倚风险是导致证据最终降级的最主要因素。许多权威指南制定机构，如苏格兰国家指南小组（Scottish Intercollegiate Guidelines Network，SIGN）[17]，本身就提供了一套严格的方法学清单，其评级过程同样将内部真实性作为核心考量。

GRADE手册[20]明确规定，即便证据体全部来自RCT，若这些试验普遍存在严重的方法学缺陷（如分配隐藏不当或盲法失败），其证据质量等级也必须降级，从而直接削弱最终推荐意见的强度（如从“强推荐”降为“弱推荐”）。

1.3.4 新研究设计的前瞻性指导

最终，质量评估的价值由对已有研究的回溯性评价，延伸至对新研究设计的前瞻性指导，形成一个完整的闭环。研究者在构思研究方案时，可参照主流偏倚风险评估工具的评价条目进行自我审视。这种“以评促建”的思路，能够帮助研究者从源头提升研究的内部真实性。

例如，RoB 2和ROBINS–I等工具提供的一系列结构化设问，可被研究者用作设计阶段的自查清单，以系统性地识别方案中的潜在弱点。一些方法学指南已将这种前瞻性思维融入其中。NIH实用试验协作组便强调，在设计整群随机试

表1 方法学质量评估在研究与实践中的应用场景

Table 1. Applications of methodological quality assessment in research and practice

应用场景	具体作用	在证据体系中的角色
个体研究的批判性阅读	·提供系统化审阅框架 ·独立判断研究的内部真实性	证据评价的基础单元
系统评价与Meta分析	·决定原始研究的纳入资格 ·解释研究间的异质性 ·作为敏感性分析的依据	个体证据的系统性综合
临床指南的证据评级	·作为GRADE等体系的核心要素 ·偏倚风险是证据降级的重要因素	综合证据向临床实践的转化
新研究设计的前瞻性指导	·“以评促建”，自我审视研究方案 ·提前识别并规避潜在方法学缺陷	将评价经验反哺于新的研究实践

注：GRADE. 建议评估、制定和评估分级（Grading of Recommendations Assessment, Development and Evaluation）。

验时，需前瞻性地考虑并预设针对特定识别偏倚的分析调整方案，即为将质量评估标准融入研究设计前端、主动规避偏倚的典范[21]。

2 主流偏倚风险评估工具概述

在药物流行病学研究中，偏倚风险评估是保障研究内部效度的关键环节，旨在系统识别可能导致效应估计值偏离真实值的各类系统性误差[22]。不同研究设计固有的方法学特征与偏倚来源各异，因此必须选用针对性的评估工具，以保证评价过程的系统性、透明性与结果的可比性（图1）。这一节将依据证据产生的层级，逐步梳理针对原始研究（包括干预性与观察性研究）、二次研究（即证据合成）、卫生经济学评价研究与真实世界研究的方法学质量评价工具，以期为研究者提供一份清晰、实用的工具选择与应用指引。

2.1 原始研究的偏倚风险评估

原始研究是证据产生的源头，对其偏倚风险的评估是整个证据链条中最为基础也最为关键的一环[23]。根据研究者是否对暴露因素施加干预，原始研究可主要分为干预性研究与观察性研究两大类，两者面临的主要偏倚与评估的侧重点截然不同。

2.1.1 干预性研究的评估工具

干预性研究旨在评估特定干预措施的因果效应，其偏倚风险评估是判断结论可靠性的核心。根据分配干预时是否采用随机化这一根本性设计差异，其评估工具亦有明确的区分，并形成了各自的方法学体系。

Cochrane 的偏倚风险评估工具是评估 RCT 偏倚风险公认的金标准，其最新版本 RoB 2 对评估逻辑进行了重大革新，实现了从“清单式检查”向“领域化判断”的理念转变。它通过一系列针对性的“信号问题”，引导评估者对 RCT 在随机化过程、干预偏离、结局数据缺失、结局测量及选择性报告等 5 个关键领域的偏倚风险进行系统性判断。RoB 2 对不同试验设计（如平行组、交叉、整群随机试验）提供了专门的评估模板，其评估结果能够与 GRADE 证据分级系统无缝衔接，是当前高质量系统评价的首选工具。

面对因伦理或可行性限制而广泛开展的非随机干预研究（non-randomised studies of the effects of interventions，NRSI），Cochrane 协作网开发了 ROBINS-I 工具[13]。该工具的理论核心建立在“目标试验模拟”（target trial emulation，TTE）框架之上，即要求评估者首先在概念上构建一个该

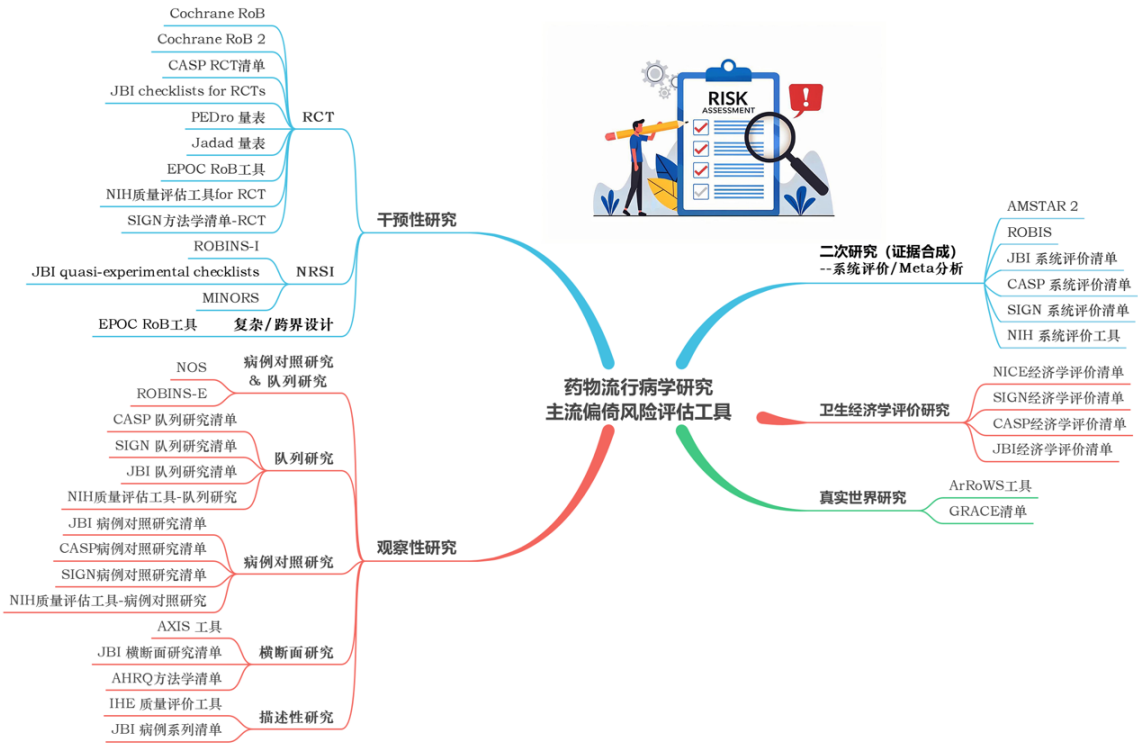

图1 药物流行病学研究主流偏倚风险评估工具概览

Figure 1. Overview of mainstream risk of bias assessment tools in pharmacoepidemiology research

NRSI所希望模拟的、无偏的理想化随机试验，然后通过7个密切相关的偏倚领域：混杂偏倚、研究对象选择偏倚、干预分类偏倚、偏离既定干预偏倚、缺失数据偏倚、结局测量偏倚、结果选择性报告偏倚，系统性地审视NRSI在模拟该理想试验的各个环节中所产生的偏倚风险。ROBINS-I严谨的理论框架和结构化的评估流程[13, 24]，为观察性证据的因果推断提供了坚实的方法学标尺。

除了上述两大核心工具，学界还发展了多种适用于不同场景的清单式与评分制工具，其共同构成了方法学演进的全貌（表2）。一些结构化的清单式工具，如关键质量评估技能项目（Critical Appraisal Skills Programme，CASP）[25]与乔安娜·布里格斯学院（Joanna Briggs Institute，JBI）系列[26]，因其清晰简便，在临床教学或快速评估中仍有应用价值；而针对特定研究类型（如卫生政策干预）的EPOC工具[27]，则提供了更具针对性的评价维度。与此同时，回顾方法学的发展历程，一些早期的评分制工具，如Jadad量表[28]、PEDro量表[29]及非随机研究的方法学项目（methodological index for non-randomized studies，MINORS）[30]等，也曾被广泛应用。然而，这些评分制工具的核心局限在于其倾向于提供一个过于简化的总分，这可能掩盖不同偏倚来源的影响，且部分条目存在将“报告质量”与“方法学质量”混淆的风险。这种从“评分制”到“清单式”，再到当前“领域化判断”的演变，反

表2 主要干预性研究偏倚风险评估工具概览与比较

Table 2. Comparison of key risk of bias assessment tools for interventional studies

类别	工具名称	评估形式/特点	主要评价维度
RCT	Cochrane RoB[15]	领域判断	7个维度：随机序列生成、分配隐藏、对研究者和受试验者的盲法、对结局评估者的盲法、结局数据完整性、选择性报告、其他偏倚
	Cochrane RoB 2[31]	领域判断，个体/整群/交叉RCT	5个核心领域：随机化过程、干预偏离、结局缺失数据、结局测量、选择性报告
	CASP RCT清单[25]	清单式	3部分（11个问题）：研究有效性（随机化、分配隐藏、盲法等）、结果（效应量、精确度）、临床适用性
	JBI checklists for RCTs[26]	清单式	13项条目：随机序列生成、分配隐藏、受试者盲法、治疗师盲法、评估者盲法、组间基线可比性、失访处理、结局测量、统计分析等
	PEDro量表[32]	评分制，适用物理治疗领域的RCT	11项条目：资格标准、随机分配、分配隐藏、基线可比性、受试者/治疗师/评估者的盲法、充分随访、意向性治疗分析、组间比较、点估计和变异性
	Jadad量表（牛津评分系统）[28]	评分制	3项核心条目：随机化、盲法、退出与失访描述
	NIH质量评估工具for RCTs[17]	清单式	14项条目：明确目的、随机化、分配隐藏、受试者/实施者/评估者的盲法、基线可比、失访/退出率、意向性治疗分析、干预依从性、结局测量、结局预指定、统计分析、多重结局、选择性报告等
	SIGN方法学清单-RCT[17, 33]	清单式	2部分（10个问题）：内部真实性（明确目的、随机序列产生、分配隐藏、盲法、基线可比性、干预特异性、结局测量、失访/退出率、意向性治疗分析、多重结局）、总体评估（偏倚风险评估、因果关系确定、结果适用性）
NRSI	ROBINS-I[13]	领域判断	7个核心领域：混杂、对象选择、干预分类、干预偏离、数据缺失、结局测量、选择性报告
	JBI quasi-experimental checklists[34]	清单式	9项条目：因果关系明确性、对照组选择适当性、混杂因素控制、干预措施描述、结局测量方法、随访完整性、统计分析适当性、结果解释合理性、研究局限性认识
	MINORS[30]	评分制，适用（外科）非随机研究	12项条目：前8条适用于所有NRSI（明确目的、纳入患者、数据收集、终点、无偏评估、随访期、失访率、样本量计算），后4条专用于比较性研究（对照组、同期、基线可比、统计分析）。
复杂/跨界设计	EPOC RoB工具[27]	跨界适用（RCT & NRSI），复杂干预试验（ITS，CBA）	9个标准领域：随机化（如适用）、分配隐藏、基线可比性、数据完整性、知识污染、选择性报告等

注：CASP. 关键质量评估技能项目（Critical Appraisal Skills Programme）；JBI. 乔安娜·布里格斯学院（Joanna Briggs Institute）；PEDro. 物理治疗证据数据库（The Physiotherapy Evidence Database）；EPOC. 有效的实践和组织护理组偏倚风险工具（Cochrane Effective Practice and Organization of Care Group）；NIH. 美国国立卫生研究院（National Institutes of Health）；SIGN. 苏格兰国家指南小组（Scottish Intercollegiate Guidelines Network）；RCT. 随机对照试验（randomized controlled trial）；NRSI. 非随机干预性研究（non-randomized studies of interventions）；ROBINS-I. 非随机干预性研究偏倚风险评估（risk of bias in non-randomized studies-of interventions）；MINORS. 非随机研究的方法学项目（methodological index for non-randomized studies）；ITS. 中断时间序列研究（interrupted time series）；CBA. 有对照的前后研究（controlled before-after study）。

映了学界对偏倚风险评估的认知在不断深化。

2.1.2 观察性研究的评估工具

观察性研究作为药物流行病学研究的核心方法论，尤其在评价药物的长期效应、罕见不良反应以及真实世界用药模式时，其价值无可替代。然而，由于缺乏随机干预，这类研究自然地更易受到混杂、选择偏倚及信息偏倚的影响。因此，借助专用的评估工具对其进行严格、系统地质量审阅，是判断其结论可靠性的前提。

表 3 展示了目前主流观察性研究评估工具的核心特征。在分析性观察性研究中，队列研究与病例对照研究是最核心的设计类型。针对这 2 类研究，NOS 是当前应用最广泛的评估工具 [35]。该量表通过对研究人群的选择、组间的可比性以及结局 / 暴露的确定 3 个维度进行星级评分，其优势在于结构清晰、直观便捷。然而，其评分机制的主观性与对混杂偏倚评估的相对简化也备受讨论。为应对更复杂的暴露效应评估，Cochrane 协作组开发了暴露类非随机研究偏倚风险评价（risk of bias in non-randomized studies of exposure，ROBINS-E）工具 [36]，它同样基于 TTE 理论，为该领域提供了更严谨的评估范式。此外，JBI[37]、CASP[38]、SIGN 及 NIH 等 [17] 机构也分别针对这 2 种设计，提供了各具侧重的清单式评估工具。

表3 主要观察性研究偏倚风险评估工具

Table 3. Key risk of bias assessment tools for observational studies

亚类	研究设计	评估形式/特点	主要评价维度
队列研究&病例对照研究	NOS[35]	评分制	3个维度：研究人群选择、组间可比性、结局/暴露的确定
	ROBINS-E[36]	领域判断	7个领域：混杂、对象选择、暴露测量、暴露后干预、数据缺失、结局测量、选择性报告
队列研究	JBI队列研究清单[37]	清单式	11项条目：暴露组/非暴露组相似性、暴露测量、混杂控制、随访完整性、失访处理等
	CASP队列研究清单[25]	清单式	3个部分（12个问题）：研究有效性（明确目的、对象选择、暴露测量、结局测量、混杂控制、随访完整性/时长等）、结果（效应量、精确度、可信度）、临床适用性
	SIGN队列研究清单[16]	清单式	2个部分（14个问题）：内部真实性（选择偏倚、信息偏倚、混杂偏倚、统计分析等）、总体评估（总体偏倚风险评估、因果关系确定、结果适用性）
	NIH质量评估工具-队列研究[17]	清单式	14项条目：明确目的/人群、样本量计算、对象选择、暴露测量、随访期、暴露评估、结局测量、释放率、混杂控制等
病例对照研究	JBI病例对照研究清单[37]	清单式	10项条目：病例/对照组可比性、匹配策略、暴露测量、混杂控制、暴露时间窗界定等
	CASP病例对照研究清单[25]	清单式	3个部分（11个问题）：研究有效性（明确目的、病例/对照人群选择、暴露测量、组件可比、混杂控制等）、结果（效应量、精确度、可信度）、临床适用性
	SIGN病例对照研究清单[17]	清单式	2个部分（11个问题）：内部真实性（选择偏倚、信息偏倚、混杂偏倚、统计分析等）、总体评估（总体偏倚风险评估、因果关系确定、结果适用性）
	NIH质量评估工具-病例对照研究[17]	清单式	12项条目：明确目的/人群、样本量计算、对象选择、暴露测量、随访期、暴露评估、结局测量、释放率、混杂控制等
横断面研究	AXIS工具[39]	清单式	20项条目：研究目的、抽样方法、人群、偏倚处理、响应率、结果报告、讨论的局限性等
	JBI横断面研究清单[40]	清单式	8项条目：纳入标准、研究对象/背景描述、暴露/结局测量、混杂识别与处理、统计分析等
	AHRQ方法学清单[41]	清单式	11项条目：信息来源、研究时间、纳排标准、盲法、混杂控制/评估、数据缺失、随访完整性、失访处理等
描述性研究	IHE质量评价工具[38]	清单式	20项条目：研究问题、病例选择/代表性、干预/结局测量、随访完整性、统计与报告等
	JBI病例系列清单[43]	清单式	10项条目：纳入标准、连续纳入、病例人口学/临床信息、结局评估、统计分析等

注：NOS. 纽卡斯尔-渥太华量表（Newcastle-Ottawa Scale）；ROBINS-E. 暴露类非随机研究偏倚风险评价（risk of bias in non-randomized studies of exposure）；CASP. 关键质量评估技能项目（Critical Appraisal Skills Programme）；JBI. 乔安娜·布里格斯学院（Joanna Briggs Institute）；NIH. 美国国立卫生研究院（National Institutes of Health）；SIGN. 苏格兰国家指南小组（Scottish Intercollegiate Guidelines Network）；AXIS. 横断面研究评价工具（appraisal tool for cross-sectional studies）；IHE. 加拿大卫生经济研究所量表（Institute of Health Economics）；AHRQ. 美国卫生保健质量和研究机构（Agency for Healthcare Research and Quality）。

对于横断面研究，其主要偏倚风险的威胁在于无法确定因果时序。为此，AXIS 工具作为首个专用评估工具[39]，通过 20 个条目对研究的方法学与报告质量进行全面审视。JBI[40] 和 AHRQ[41] 也提供了更为简洁的清单，以辅助快速评估。

而在描述性观察性研究中，如病例系列与病例报告，其评估重点则转向报告的系统性与信息的可靠性。为此，加拿大卫生经济研究所量表（Institute of Health Economics，IHE）质量评价工具[42] 和 JBI 病例系列批判性评估清单[42] 为评估这类缺乏比较组的特殊研究类型，提供了规范化的评估框架。

2.2 二次研究（证据合成）的方法学质量评估

系统评价与 Meta 分析通过对特定临床问题的已有原始研究进行系统性地检索、筛选、评价与合成，旨在提供比任何单个研究更为可靠与全面的证据。然而，证据合成过程本身也是一项严谨的研究，其每一个环节——从方案制定、文献检索到数据提取与合并——都可能引入偏倚，从而影响最终结论的可靠性。因此，对系统评价这类二次研究进行方法学质量的审阅，即“对评价的评价”，是判断其结论可信度的关键步骤。

表 4 展示了二次研究（证据合成）主流应用工具的核心模块。在这一领域，系统评价评估测量工具（a measurement tool to assess systematic reviews 2，AMSTAR 2）是当前国际上评估系统评价（尤其是包含干预性研究的系统评价）方法学质量最权威、应用最广泛的工具[44–45]。该工具共包含 16 项评估条目，全面覆盖了系统评价制作的全流程，从方案是否预注册、检索策略是否全面，到纳入 / 排除标准、数据提取、原始研究的偏倚风险评估、Meta 分析方法的适当性，以及对发表偏倚和利益冲突的考量。其中，AMSTAR 2 创新性地将 7 项条目定义为“关键条目（critical domains）”，并摒弃了简单的总分制，转而根据关键条目中存在的缺陷数量，对系统评价的整体可信度给出高、中、低或极低的分级结论。

另一个核心工具 ROBIS（risk of bias in systematic reviews）则提供了不同的评估视角[46]，其更聚焦于评价过程本身是否存在偏倚风险，即审视评价者的行为是否可能导致其结论偏离真相。

ROBIS 通过对研究资格标准的制定、研究的识别与筛选、数据的收集与研究的评估、证据的综合与发现是个关键领域的判断，最终对该系统评价的总体偏倚风险得出“低风险”“高风险”“不清楚”的结论。

在实践中，AMSTAR 2 与 ROBIS 形成了有力的互补[45–46]。二者结合使用，能为系统评价提供一个更为全面、立体的“质量评估”。此外，JBI[47]、CASP[38]、SIGN、NIH 等[17, 33] 机构也提供了结构清晰的清单式工具，它们在结构和侧重点上各有不同，共同构成了该领域的工具体系。

2.3 其他特定类型研究的方法学质量评估

除了传统的干预性与观察性研究，药物流行病学领域还涉及多种特定类型的研究，其质量评估亦需借助专门的工具（表 5）。

表4 主要二次研究（证据–合成）方法学质量评估工具

Table 4. Key methodological quality assessment tools for secondary research (evidence synthesis)

工具名称	评估形式/特点	主要评价维度
AMSTAR 2	清单式	16项条目（含7项关键条目）：PICO、方案注册、检索策略、研究筛选、数据提取、纳入研究列表、偏倚风险评估、Meta分析方法、发表偏倚等
ROBIS	领域判断	4个核心领域：研究资格标准、研究的识别与筛选、数据的收集与研究的评估、证据的综合与发现
JBI系统评价清单	清单式	11项条目：审查问题清晰性、纳入标准、检索策略、研究筛选流程、偏倚风险评估、数据提取、数据合成、结论与证据一致性等
CASP系统评价清单	清单式	3部分（10个问题）：研究有效性（问题明确、研究类型、检索策略等）、结果（结果合并、精确度等）、临床适用性
SIGN系统评价清单	清单式	2部分（12项核心条目）：内部真实性（问题明确、纳排标准、文献检索、研究筛选、数据提取、纳入/排除研究列表、偏倚风险评估、发表偏倚、Meta分析方法、利益冲突）、研究的整体评价
NIH系统评价工具	清单式	8项条目：问题明确、纳排标准、文献检索、研究筛选、偏倚风险评估、数据提取、发表偏倚、异质性评估

注：AMSTAR 2. 系统评价评估测量工具（a measurement tool to assess systematic reviews 2）；ROBIS. 系统评价偏倚风险评价工具（risk of bias in systematic reviews）；CASP. 关键质量评估技能项目（Critical Appraisal Skills Programme）；JBI. 乔安娜·布里格斯学院（Joanna Briggs Institute）；NIH. 美国国立卫生研究院（National Institutes of Health）；SIGN. 苏格兰国家指南小组（Scottish Intercollegiate Guidelines Network）；PICO. 患者群体（population）、干预措施（intervention）、对照措施（comparison）和结局指标（outcome）。

表5 主要其他特定类型研究方法学质量评估工具

Table 5. Key methodological quality assessment tools for other specific study types

大类	工具名称	评估形式/特点	主要评价维度
卫生经济学评价[53]	NICE经济学评价清单	清单式	2个部分（18个问题）：适用性（判断研究是否与指南问题相关，如人群、干预、成本角度等是否匹配NICE标准）、局限性（对通过筛选的研究进行深入方法学质量评估，关注模型结构、数据来源、分析方法和利益冲突）
	CASP经济学评价清单	清单式	2个部分（9个问题）：内部真实性（问题、设计、成本与结果的识别/测量/评估、贴现、假设与敏感性分析、增量分析）、总体评估（总体质量评估、结果适用性）
	JBI经济学评价清单	清单式	3个部分（12个问题）：研究有效性（问题、替代品、基本有效性、识别、测量和评估）、成本与结果（贴现、增量分析和敏感性分析）、适用性（有效性和成本可转化性）
	SIGN 经济学评价清单	清单式	11个问题：问题与替代品定义、成本与结果的识别、测量、评估、贴现、增量分析、敏感性分析、结果的适用性和推广性
真实世界研究	ArRoWS工具[49–50]	领域判断	6大领域：数据来源、研究设计、混杂、其他偏倚、分析方法、敏感性分析
	GRACE清单[51–52]	报告规范&清单	2部分（11个问题）：数据（暴露/结局描述、测量客观性/对等性、协变量）、方法（新用药者设计、对照组可比性、混杂控制、不朽时间偏倚、敏感性分析）

注：NICE. 英国国家卫生与临床优化研究所（National Institute for Health and Care Excellence）；CASP. 关键质量评估技能项目（Critical Appraisal Skills Programme）；JBI. 乔安娜·布里格斯学院（Joanna Briggs Institute）；SIGN. 苏格兰国家指南小组（Scottish Intercollegiate Guidelines Network）；ArRoWS. 真实世界观察性研究评估工具（assessment of real-world observational studies）；GRACE. 比较效果的优良研究（The good research for comparative effectiveness）。

卫生经济学评价旨在比较不同医疗干预的成本与效果，其方法学质量直接关系到卫生资源的合理配置。针对此类研究，目前主流的评估工具以清单式为主。其中，英国国家卫生与临床优化研究所（National Institute for Health and Care Excellence，NICE）的清单因其全面性而备受推崇，它不仅关注研究的内部真实性（如模型结构、数据来源、分析方法），还特别强调了研究结果与决策问题的适用性。此外，CASP、JBI、SIGN 等机构也提供了结构相似的清单[38]。而早期的 QHES 量表[48]则是一个评分制工具，但因其条目更侧重于报告质量而非方法学质量，在当前实践中已较少作为首选。

真实世界研究（real world study，RWS）的质量评估是当前方法学领域的热点与难点。针对这一需求，新兴的 ArRoWS 工具为评价真实世界观察性研究的偏倚风险[49–50]，提供了一个专门的评估框架，其涵盖了数据来源、研究设计、混杂控制及分析方法等六大核心领域。此外，GRACE 清单[51–52]则通过提供一套包含 11 个关键问题的结构化列表，旨在系统性地提升比较效果研究的报告透明度与完整性，从而为后续运用 ArRoWS 等工具进行严谨的方法学质量评价提供坚实的信息基础。

2.4 评估工具的选择策略与总结

偏倚风险评估工具的选择并非简单的按图索骥，而是一项需在评估目的、研究设计与现实可行性之间进行综合权衡的方法学决策。如旨在为临床指南提供高级别证据的系统评价，因其对方法学严谨性的要求极高，故而建议选用 RoB 2 或 ROBINS–I 这类能够对偏倚来源进行深度剖析的领域化判断工具；而在临床教学或快速文献审阅的场景下，以 CASP 为代表的清单式工具或因其简洁高效而更具优势。

此外，待评价研究的设计特征直接决定了工具的适用性。通用工具虽能应对标准的平行组 RCT，但当面对整群随机试验或中断时间序列等复杂设计时，则需借助 Cochrane EPOC 等专用标准，以识别其独特的偏倚威胁。同样，对于无对照组的病例系列研究，旨在进行“比较研究”的 ROBINS–I 等工具或不适用，而 IHE 工具则提供了更佳的评估方案。

理想的严谨性与现实的可行性之间往往存在张力。例如，一次严谨的 ROBINS–I 评估可能耗时甚久，其资源投入并非所有场景适用。因此，审慎地选择与评估目的及资源相匹配的工具，并在报告中清晰阐明其方法学局限性，本身即是科学严谨性的体现。

3 案例分析：偏倚风险评估工具的应用实践

3.1 案例背景与评估准备

为具体阐释偏倚风险评估工具的应用，本文

以2024年Cancian等[54]发表于*Ophthalmology and Therapy*的研究为例，系统演示如何应用ROBINS-I工具进行评估。该研究是一项前瞻性、单中心队列研究，旨在评价法瑞西单抗用于既往抗血管内皮生长因子（vascular endothelial growth factor，VEGF）治疗效果不佳而换药的新生血管性年龄相关性黄斑变性（neovascular age-related macular degeneration，nAMD）患者的真实世界疗效。研究的核心设计是通过比较患者自身在换用新药（法瑞西单抗）前后1年内视力及光学相干断层扫描（optical coherence tomography，OCT）等临床指标的变化，以评价干预的有效性。

ROBINS-I工具的评估逻辑是将观察性研究与其旨在模拟的理想RCT（即“目标试验”）进行比较。尽管原研究为单臂设计，其内在的科学问题（评价一种新干预的增量效果）决定了其理想的“目标试验”必须包含一个平行的活性对照组，因为只有通过比较才能分离出干预的净效应，并将其与疾病自然病程、向均值回归等时变混杂因素区分开。该案例研究与其理想“目标试验”的核心要素对比见表6。

表6 案例研究与“目标试验”核心要素对比

Table 6. Comparison of core elements between the case study and the "target trial"

要素	案例研究	理想的“目标试验”
P（人群）	对标准抗VEGF治疗反应不佳的nAMD患者	对标准抗VEGF治疗反应不佳的nAMD患者
I（干预）	换用法瑞西单抗	随机分配至换用法瑞西单抗
C（对照）	无平行对照组（采用自身前后对照）	随机分配至继续原标准抗VEGF治疗
O（结局）	1年内的视力及OCT指标变化	1年内的视力及OCT指标变化

3.2 基于ROBINS-I的7个偏倚领域评估

本部分将依据ROBINS-I工具的框架，逐一评估示例研究[54]在7个偏倚领域的风险，每个领域的评估都将围绕该研究与其前述“目标试验”的差异展开。

3.2.1 混杂偏倚（bias due to confounding）

最终判断：严重偏倚风险（critical risk of bias）。依据：该研究为单臂自身前后对照设计，虽避免了不同患者间的基线混杂，但其核心缺陷在于完全无法控制关键的时变混杂因素。在长达1年的观察期内，诸如疾病的自然病程波动、因患者被筛选为“疗效不佳”而极易出现的“向均值回归”统计学趋势以及换用新药带来的安慰剂效应等，都可能独立影响结局。由于缺乏一个平行的对照组，研究设计本身无法将观察到的效应从这些混杂因素的影响中分离出来。因此，其效应估计值很可能严重偏离真实的因果效应。

3.2.2 研究对象选择的偏倚（bias in selection of participants）

最终判断：低偏倚风险（low risk of bias）。依据：研究纳入了所有符合预设标准并在特定时间窗内就诊的连续患者，并将研究的“时间零点”明确定义为换用新药的时刻。这种策略确保了在队列形成的过程中，参与者的选择未受到干预开始后事件的影响。需要注意的是，虽然整个队列是经过高度筛选的“疗效不佳者”，但对于“评价这类患者换药后会发生什么”这一特定问题而言，其入组过程的偏倚风险较低。

3.2.3 干预分类的偏倚（bias in classification of interventions）

最终判断：低偏倚风险（low risk of bias）。依据：研究为单臂设计，所有参与者均接受了明确定义的法瑞西单抗治疗。由于不存在平行对照组，因此不可能发生因干预信息测量不准确或不完整而导致的组间错分。

3.2.4 偏离既定干预的偏倚（bias due to deviations from intended interventions）

最终判断：中等偏倚风险（moderate risk of bias）。依据：原文结果部分明确报告“three patients (9%) had been switched from faricimab to an alternative medication”。这意味着存在偏离既定干预的情况。然而，文章的统计分析似乎是基于完成治疗并有完整数据的患者进行的，并未明确说明如何处理这些偏离者的结局数据（如采用意向性治疗分析）。若这些因疗效不佳而换药的患者被排除在最终分析之外，将可能导致对法瑞西单抗疗效的高估。

3.2.5 缺失数据偏倚（bias due to missing data）

最终判断：信息不足（no information）。依据：

原文中纳入了 38 例，排除 5 例，最终分析了 33 例。排除原因包括“lack of pre-switch clinical documentation”。此外，随访过程中提到“one patient (3%)... died”。文章的分析是基于最终可用的 33 例患者的数据，但并未明确对因死亡或其他原因导致的潜在数据缺失进行处理（如多重插补）。这种基于完整病例的分析，在数据缺失与结局相关时可能引入偏倚。

3.2.6 结局测量的偏倚（bias in measurement of outcomes）

最终判断：中等偏倚风险（moderate risk of bias）。依据：研究的结局指标中，如最佳矫正视力，其测量过程包含一定的主观成分。根据研究描述，这些结局是由知晓患者正在接受新药治疗的临床医生或技术人员进行测量的。由于缺乏盲法，评估者对新药疗效的预期可能会无意识地影响其测量行为（即观察者偏倚），从而可能导致对疗效的系统性高估。

3.2.7 结果选择性报告的偏倚（bias in selection of the reported result）

最终判断：中等偏倚风险（moderate risk of bias）。依据：研究方案未在公共平台预先注册，这带来了基于结果进行选择性报告的理论风险。然而，该研究报告了包括重要功能结局“最佳矫正视力”在内的多个预设结局，且并未回避其中一些指标未达到统计学显著性的“阴性”结果。这种对阴性结果的呈现，在一定程度上削弱了选择性报告的可能性。因此综合考虑，该领域仍存在一定风险。

3.3 总体偏倚风险判断及其对证据解读的意义

根据 ROBINS-I 工具规则，一项研究的总体偏倚风险由所有单个领域中最高的风险等级决定。由于混杂偏倚被判定为“严重风险”，因此，Cancian 等[54]的这项研究总体偏倚风险被判定为“严重风险”。

这一判断意味着该研究的效应估计值信赖度极低。这并非否定研究者观察到的临床现象，而是从方法学上指出，该研究的设计无法将药物的真实药理效应与多种重要的偏倚来源有效分离。其单臂自身前后对照的设计，使结论极易受到时变混杂的严重影响，如疾病自然病程的波动、向均值回归的统计学趋势及安慰剂效应等。因此，研究者无法将观察到的改善清晰地归因于药物的真实效应，还是多种偏倚的混合效应。

在证据应用层面，这一“严重偏倚风险”的判断具有明确的指导意义。首先，这代表其证据强度不足以视为对法瑞西单抗疗效的可靠证实，故不应直接用于更新临床实践指南或影响临床决策。在进行证据合成时，此类研究通常会被系统评价所排除，或仅在敏感性分析中考察其影响；若依据 GRADE 体系对证据体进行评级，来自此类研究的证据将可能被直接判定为极低质量。

尽管如此，该研究在假设生成阶段仍具价值。尤其在药物上市初期，此类快速、便捷的真实世界研究能够为新药在特定人群（如难治性患者）中的应用提供初步的疗效信号与安全性线索，为后续开展更严谨的研究提供依据和方向。未来的研究中建议采用能够有效模拟“目标试验”的设计，例如设立平行的、可比的阳性药物对照组。在方案设计阶段，研究者即可参照 ROBINS-I 的 7 个领域，或 ArRoWS 工具所提出的六大框架，进行前瞻性的风险识别与规避，才能得出更可信的结论。

4 结语

尽管偏倚风险评估工具为证据评价提供了结构化的框架，但其应用本身并非一劳永逸的解决方案。研究者必须清晰地认识到，任何工具的产出，其质量上限都取决于使用者的专业判断。此外，评估工具本身的多样性也带来了新的挑战。面对从经典的 NOS 到新兴的 ArRoWS 等众多工具，如何根据具体场景做出最优选择，本身就考验着研究者的素养。更需警惕的是，“低偏倚风险”的标签可能会给研究者一种虚假的安全感，掩盖了该结论在推广至不同人群或临床情境时（即外部效度）的潜在局限性。因此，审慎地将工具的结构化判断与深入的方法学思考相结合，是避免工具被滥用或误读的关键。

对于我国药物流行病学研究者而言，熟练应用评估工具的更深远意义在于“以评促建”，这意味着将评估思维从对他人研究的“回溯性批判”，升维为对自己研究的“前瞻性质控”。研究者在构思方案的阶段，即可参照 RoB 2 或 ROBINS-I，对自身设计进行系统性的“预评估”，从而在早期识别并修正潜在的方法学缺陷。这种

将评价标准内化为设计准则的实践，是从源头上提升我国药物流行病学研究原始创新质量与国际竞争力的根本路径之一。

与此同时，偏倚风险评估的方法学本身也在不断演进。除了开发新工具，或许更重要的方向在于对现有工具进行整合与提炼，如开发一套适用于所有观察性研究的“核心偏倚条目集”，并辅以针对特定设计的扩展模块，以平衡评估的深度与效率。此外，随着人工智能，特别是大语言模型（large language model，LLM）技术的发展，其在辅助偏倚风险评估中的应用潜力已初步显现。近期证据[55-56]表明，先进的LLM在识别RCT的偏倚风险时，已能达到与人类专家相当的一致性，且效率呈几何级数提升。尽管目前人工智能在深刻理解复杂的观察性研究设计、评估字里行间的未测量混杂方面仍有局限，但其无疑为解决评估过程中的效率瓶颈问题提供了全新路径。未来，开发人机协同的评估流程，将人类专家的深度方法学判断与人工智能的高效信息提取和初步筛选相结合，或将成为提升证据评价质量与效率的新范式。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 吴昀效，颜济南，聂晓璐，等.《中国药物流行病学研究方法学指南（第2版）》及其系列解读（1）：概述[J]. 药物流行病学杂志，2025, 34(1): 2-11. [Wu YX, Yan JN, Nie XL, et al. *Guide on Methodological Standards in Pharmacoepidemiology in China (2nd edition)* and their series interpretation (1): an overview[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(1): 2-11.] DOI: 10.12173/j.issn.1004-5511.202412131.

2 颜济南，吴昀效，聂晓璐，等.《中国药物流行病学研究方法学指南（第2版）》的制订/修订过程[J]. 药物流行病学杂志，2025, 34(2): 121-135. [Yan JN, Wu YX, Nie XL, et al. Revision process of the *Guide on Methodological Standards in Pharmacoepidemiology in China (2nd edition)*[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(2): 121-135.] DOI: 10.12173/j.issn.1005-0698.202502028.

3 程静茹，陈锐娜，李嘉蕊，等.《药物流行病学研究方法学指南（第2版）》系列解读（9）：研究报告规范与结果可视化[J]. 药物流行病学杂志，2025, 34(9): 1004-1016. [Cheng JR, Chen RN, Li JR. *Guide on Methodological Standards in Pharmacoepidemiology (2nd edition)* and their series interpretation (9): research report standards and results visualization[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(9): 1004-1016.] DOI: 10.12173/j.issn.1005-0698.202508056.

4 Sabaté M, Montané E. Pharmacoepidemiology: an overview[J]. J Clin Med, 2023, 12(22): 7033. DOI: 10.3390/jcm12227033.

5 Weinstein EJ, Ritchey ME, Lo Re V, 3rd. Core concepts in pharmacoepidemiology: validation of health outcomes of interest within real-world healthcare databases[J]. Pharmacoepidemiol Drug Saf, 2023, 32(1): 1-8. DOI: 10.1002/pds.5537.

6 Gershon AS, Lindenauer PK, Wilson KC, et al. Informing healthcare decisions with observational research assessing causal effect. An official American Thoracic Society research statement[J]. Am J Respir Crit Care Med, 2021, 203(1): 14-23. DOI: 10.1164/rccm.202010-3943ST.

7 胡志德，主编. 系统评价与Meta分析论文撰写规范PRISMA解读[M]. 北京：中南大学出版社，2024: 1-245.

8 Dewidar O, Shamseer L, Melendez-Torres GJ, et al. Improving the reporting on health equity in observational research (STROBE-Equity): extension checklist and elaboration[J]. JAMA Netw Open, 2025, 8(9): e2532512. DOI: 10.1001/jamanetworkopen.2025.32512.

9 Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 explanation and elaboration: updated guideline for reporting randomised trials[J]. BMJ, 2025, 389: e081124. DOI: 10.1136/bmj-2024-081124.

10 Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews[J]. BMJ, 2021, 372: n160. DOI: 10.1136/bmj.n160.

11 高亚，刘明，杨珂璐，等. 系统评价报告规范：PRISMA 2020与PRISMA 2009的对比分析与实例解读[J]. 中国循证医学杂志，2021, 21(5): 606-616. [Gao Y, Liu M, Yang KL, et al. Reporting guideline for systematic review: comparative analysis of PRISMA 2020 and PRISMA 2009[J]. Chinese Journal of Evidence-Based Medicine, 2021, 21(5): 606-616.] DOI: 10.7507/1672-2531.202104143.

12 Carra MC, Romandini P, Romandini M. Risk of bias evaluation of cross-sectional studies: adaptation of the Newcastle-Ottawa Scale[J/OL]. J Periodontal Res, (2025-04-28) [2025-10-16]. DOI: 10.1111/jre.13405.

13 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions[J]. BMJ, 2016, 355: i4919. DOI: 10.1136/bmj.i4919.

14 胡植乔，王了，黄椿棚，等. 2024年新版干预类非随机研究偏倚风险评价工具（ROBINS-I V2）中文解读[J]. 中国循证医学杂志，2025, 25(6): 704-709. [Hu ZQ, Wang L, Huang CP, et al. Chinese introduction to risk of bias in nonrandomized studies of interventions version 2 (ROBINS-I V2) in 2024[J]. Chinese Journal of Evidence-Based Medicine, 2025, 25(6): 704-709.] DOI: 10.7507/1672-2531.202412066.

15 Nejadghaderi SA, Balibegloo M, Rezaei N. The Cochrane risk of bias assessment tool 2 (RoB 2) versus the original RoB: a perspective on the pros and cons[J]. Health Sci Rep, 2024, 7(6): e2165. DOI: 10.1002/hsr2.2165.

16 Higgins JP, Altman DG, Gøtzsche PC, et al. The Cochrane

collaboration's tool for assessing risk of bias in randomised trials[J]. BMJ, 2011, 343: d5928. DOI: 10.1136/bmj.d5928.

17 李柄辉 , 訾豪 , 李路遥 , 等 . 医学领域一次研究和二次研究的方法学质量（偏倚风险）评价工具 [J]. 医学新知 , 2021, 31(1): 51–58. [Li BH, Zi H, Li LY, et al. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?[J]. New Medicine, 2021, 31(1): 51–58.] DOI: 10.12173/j.issn.1004–5511.2021.01.07.

18 Marušić MF, Fidahić M, Cepeha CM, et al. Methodological tools and sensitivity analysis for assessing quality or risk of bias used in systematic reviews published in the high–impact anesthesiology journals[J]. BMC Med Res Methodol, 2020, 20(1): 121. DOI: 10.1186/s12874–020–00966–4.

19 van Dijk A, Almoghrabi N, Leijten P. One research question, two Meta–analyses, three conclusions: commentary on "A systematic review with Meta–analysis of Cognitive Bias Modification interventions for anger and aggression"[J]. Behav Res Ther, 2024, 173: 104475. DOI: 10.1016/j.brat.2024.104475.

20 Piggott T, Morgan RL, Cuello–Garcia CA, et al. Grading of Recommendations Assessment, Development and Evaluations (GRADE) notes: extremely serious, GRADE's terminology for rating down by three levels[J]. J Clin Epidemiol, 2020, 120: 116–120. DOI: 10.1016/j.jclinepi.2019.11.019.

21 Heagerty PJ. Experimental designs and randomization schemes[R/OL]. (2021–09–21) [2025–10–15]. https://rethinkingclinicaltrials.org/chapters/design/experimental–designs–and–randomization–schemes/experimental–designs–introduction/.

22 杨智荣 , 孙凤 , 詹思延 . 偏倚风险评估系列：（一）概述 [J]. 中华流行病学杂志 , 2017, 38(7): 983–987. [Yang ZR, Sun F, Zhan SY. Risk of bias assessment: (1) overview[J]. Chinese Journal of Epidemiology, 2017, 38(7): 983–987.] DOI: 10.3760/cma.j.issn.0254–6450.2017.07.027.

23 郑卿勇 , 许建国 , 刘明 , 等 . 网状 Meta 分析中的偏倚风险评估工具概览与思考 [J]. 中国循证医学杂志 , 2025, 25(5): 584–594. [Zheng QY, Xu JG, Liu M, et al. Overview and perspectives on risk of bias assessment tools in network Meta–analysis[J]. Chinese Journal of Evidence–Based Medicine, 2025, 25(5): 584–594.] DOI: 10.7507/1672–2531.202411077.

24 Zhang Y, Huang L, Wang D, et al. The ROBINS–I and the NOS had similar reliability but differed in applicability: a random sampling observational studies of systematic reviews/Meta–analysis[J]. J Evid Based Med, 2021, 14(2): 112–122. DOI: 10.1111/jebm.12427.

25 Critical Appraisal Skills Programme. CASP randomised controlled trial checklist[R/OL]. (2023–12–31) [2025–10–15]. https://casp–uk.net/casp–checklists/CASP–checklist–randomised–controlled–trials–RCT–2024.pdf.

26 Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials[J]. JBI Evid Synth, 2023, 21(3): 494–506. DOI: 10.11124/jbies–22–00430.

27 王敏 , 许培扬 , 王小万 , 等 . Cochrane EPOC 系统评价原始研究证据来源数据库及其检索策略分析 [J]. 医学信息学杂志 , 2008, 29(5): 13–19. [Wang M, Xu PY, Wang XW, et al. Analysis on the source database and search strategy of original evidence for Cochrane EPOC systematic reviews[J]. Journal of Medical Informatics, 2008, 29(5): 13–19.] DOI: 10.3969/j.issn.1673–6036.2008.05.003.

28 Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary?[J]. Control Clin Trials, 1996, 17(1): 1–12. DOI: 10.1016/0197–2456(95)00134–4.

29 Moseley AM, Herbert RD, Sherrington C, et al. Evidence for physiotherapy practice: a survey of the Physiotherapy Evidence Database (PEDro)[J]. Aust J Physiother, 2002, 48(1): 43–49. DOI: 10.1016/s0004–9514(14)60281–6.

30 Slim K, Nini E, Forestier D, et al. Methodological index for non–randomized studies (minors): development and validation of a new instrument[J]. ANZ J Surg, 2003, 73(9): 712–716. DOI: 10.1046/j.1445–2197.2003.02748.x.

31 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials[J]. BMJ, 2019, 366: l4898. DOI: 10.1136/bmj.l4898.

32 Cashin AG, Mcauley JH. Clinimetrics: Physiotherapy Evidence Database (PEDro) scale[J]. J Physiother, 2020, 66(1): 59. DOI: 10.1016/j.jphys.2019.08.005.

33 Habour R, 陈霖 . 临床实践指南的制定 : 苏格兰地区学院之间指南网络 [J]. 中国循证医学杂志 , 2003, 3(2): 153–156. https://d.wanfangdata.com.cn/periodical/ChVQZXJpb2RpY2FsQ0hJMjAyNTA2MjISD3pneHp5eDIwMDMwMjAxNRoIczlvNDJpajU%3D.

34 Barker TH, Habibi N, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi–experimental studies[J]. JBI Evid Synth, 2024, 22(3): 378–388. DOI: 10.11124/jbies–23–00268.

35 Stang A. Critical evaluation of the Newcastle–Ottawa Scale for the assessment of the quality of nonrandomized studies in Meta–analyses[J]. Eur J Epidemiol, 2010, 25(9): 603–605. DOI: 10.1007/s10654–010–9491–z.

36 Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non–randomized follow–up studies of exposure effects (ROBINS–E)[J]. Environ Int, 2024, 186: 108602. DOI: 10.1016/j.envint.2024.108602.

37 Aromataris E, Munn Z, editors. JBI manual for evidence synthesis[R/OL]. (2024–12–31) [2025–10–15]. https://synthesismanual.jbi.global.

38 Zeng X, Zhang Y, Kwong JS, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and Meta–analysis, and clinical practice guideline: a systematic review[J]. J Evid Based Med, 2015, 8(1): 2–10. DOI: 10.1111/jebm.12141.

39 Downes MJ, Brennan ML, Williams HC, et al. Development of a critical appraisal tool to assess the quality of cross–sectional studies (AXIS)[J]. BMJ Open, 2016, 6(12): e011458. DOI: 10.1136/bmjopen–2016–011458.

40 Barker TH, Hasanoff S, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for analytical cross-sectional studies[J/OL]. JBI Evid Synth, (2025-06-10) [2025-10-15]. DOI: 10.11124/JBIES-24-00523.

41 Ip S, Paulus JK, Balk EM, et al. Role of single group studies in Agency for Healthcare Research and Quality comparative effectiveness reviews[M]. Rockville (MD): Agency for Healthcare Research and Quality (US), 2013.

42 Moga C, Guo B, Schopflocher D, et al. Development of a quality appraisal tool for case series studies using a modified Delphi technique[R/OL]. (2012-06-02) [2025-10-15]. https://cobe.paginas.ufsc.br/files/2014/10/MOGA.Case-series.pdf.

43 Munn Z, Barker TH, Moola S, et al. Methodological quality of case series studies: an introduction to the JBI critical appraisal tool[J]. JBI Evid Synth, 2020, 18(10): 2127-2133. DOI: 10.11124/jbisrir-d-19-00099.

44 Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews[J]. BMC Med Res Methodol, 2007, 7: 10. DOI: 10.1186/1471-2288-7-10.

45 Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both[J]. BMJ, 2017, 358: j4008. DOI: 10.1136/bmj.j4008.

46 Whiting P, Savović J, Higgins JP, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed[J]. J Clin Epidemiol, 2016, 69: 225-234. DOI: 10.1016/j.jclinepi.2015.06.005.

47 Munn Z, Stone JC, Aromataris E, et al. Assessing the risk of bias of quantitative analytical studies: introducing the vision for critical appraisal within JBI systematic reviews[J]. JBI Evid Synth, 2023, 21(3): 467-471. DOI: 10.11124/jbies-22-00224.

48 赵丽婷, 何耀, 张素华, 等. 运用QHES量表评价我国消化系统疾病经济学研究质量[J]. 中国药物评价, 2013, 30(5): 309-312. [Zhao LT, He Y, Zhang SH, et al. Quality assessment of economics research of diseases of the digestive system published in Chinese by QHES Checklist[J]. Chinese Journal of Drug Evaluation, 2013, 30(5): 309-312.] DOI: 10.3969/j.issn.2095-3593.2013.05.015.

49 Coles B, Tyrer F, Hussein H, et al. Development, content validation, and reliability of the Assessment of Real-World Observational Studies (ArRoWS) critical appraisal tool[J]. Ann Epidemiol, 2021, 55: 57-63. e15. DOI: 10.1016/j.annepidem.2020.09.014.

50 张强, 卢存存, 王志飞, 等. 真实世界观察性研究评估（ArRoWS）工具的介绍[J]. 中国中医基础医学杂志, 2024, 30(10): 1691-1696. [Zhang Q, Lu CC, Wang ZF, et al. Introduction to the ArRoWS tool for assessing real-world observational studies[J]. Chinese Journal of Basic Medicine in Traditional Chinese Medicine, 2024, 30(10): 1691-1696.] DOI: 10.19945/j.cnki.issn.1006-3250.2024.10.030.

51 Esposito G, Dilaghi E, Costa-Santos C, et al. The Gastroscopy RAte of Cleanliness Evaluation (GRACE) Scale: an international reliability and validation study[J]. Endoscopy, 2025, 57(4): 312-320. DOI: 10.1055/a-2422-0856.

52 廖星, 章轶立, 谢雁鸣. 真实世界研究标准:RECORD清单和GRACE清单的解读[J]. 中国中药杂志, 2015, 40(24): 4734-4738. [Liao X, Zhang YL, Xie YM. Standards for real-world research: interpretation of the RECORD and GRACE checklists[J]. China Journal of Chinese Materia Medica, 2015, 40(24): 4734-4738.] DOI: 10.19540/j.cnki.cjcmm.2015.24.003.

53 田金徽, 张俊华, 郑丽, 主编. 医学研究检索与评价[M]. 北京: 科学出版社, 2025: 1-148.

54 Cancian G, Paris A, Agliati L, et al. One-year real-world outcomes of intravitreal faricimab for previously treated neovascular age-related macular degeneration[J]. Ophthalmol Ther, 2024, 13(11): 2985-2997. DOI: 10.1007/s40123-024-01036-4.

55 Rose CJ, Bidonde J, Ringsten M, et al. Using a large language model (ChatGPT-4o) to assess the risk of bias in randomized controlled trials of medical interventions: interrater agreement with human reviewers[J]. Cochrane Evid Synth Methods, 2025, 3(5): e70048. DOI: 10.1002/cesm.70048.

56 Nashwan AJ, Jaradat JH. Streamlining systematic reviews: harnessing large language models for quality assessment and risk-of-bias evaluation[J]. Cureus, 2023, 15(8): e43023. DOI: 10.7759/cureus.43023.

收稿日期：2025年09月30日 修回日期：2025年10月16日
本文编辑：冼静怡 杨 燕