·方法学研究·

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

李柄辉[1, 2]，訾　豪[1]，李路遥[1, 3]，王朝阳[1, 3]，任学群[3]，郭　毅[1]，译

1. 武汉大学中南医院循证与转化医学中心（武汉 430071）
2. 武汉大学中南医院泌尿外科（武汉 430071）
3. 河南大学循证医学与临床转化研究院（河南开封 475000）

【摘要】方法学质量（偏倚风险）评估是医学研究开始前的重要步骤。因此，准确判断研究类型是前提，选择合适的评价工具也很重要。本综述介绍了用于随机对照试验（包括个体和整群）、动物实验、非随机干预性研究（包括随访研究、有对照组的前后对照研究、前后对照研究、非对照纵向研究、间断时间序列研究）、队列研究、病例－对照研究、横断面研究（包括分析和描述性研究）、观察性病例系列和病例报告、比较效果研究、诊断性研究、卫生经济学评价、预测研究（包括预测变量研究、预测模型影响研究、预后预测模型研究）、质性研究、结果测量工具（包括患者－报告的结果测量发展、内容真实性、结构真实性、内部一致性、跨文化真实性/测量不变性、信度、测量误差、校标真实性、结构真实性假设检验、反应度）、系统评价与 Meta 分析，以及临床实践指南的方法学质量评价工具。通过本综述，读者可以区分医学研究的类型并选择适当的工具。全面掌握相关知识并多加练习是正确评估方法学质量的基本要求。

【关键词】方法学质量；偏倚风险；质量评估；方法学清单；评价工具；干预性研究；动物实验；观察性研究；质性研究；结果测量工具；系统评价；Meta 分析；临床实践指南

Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?

Translated by Bing-Hui LI[1, 2], Hao ZI[1], Lu-Yao LI[1, 3], Chao-Yang WANG[1, 3], Xue-Qun REN[3], Yi GUO[1]

1. Center for Evidence-Based and Translational Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

2. Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

3. Institutes of Evidence-Based Medicine and Knowledge Translation, Henan University, Kaifeng 475000, China

Corresponding author: Xue-Qun REN, E-mail: renxuequn001@163.com; Yi GUO, E-mail: guoyi@whu.edu.cn

DOI: 10.12173/j.issn.1004-5511.2021.01.07

基金项目：国家重点研发计划重点专项（SQ2020YFF0426494，2016YFC0106300）；武汉市卫健委新冠肺炎疫情防控应急科研专项重点项目一类（EG20A02）

通信作者：任学群，教授，主任医师，博士研究生导师，E-mail：renxuequn001@163.com

郭毅，教授，博士研究生导师，E-mail：guoyi@whu.edu.cn

注：本文原文刊发于 Military Medical Research [Ma LL, Wang YY, Yang ZH, et al. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?[J]. Mil Med Res, 2020, 7(1): 7]，获其编辑部授权出版中文翻译版。本译文对原文略作调整，详细信息见原文。

翻译：李柄辉，王朝阳；回译：訾豪，李路遥；审校：任学群，郭毅。

在 20 世纪，著名教授 Cochrane A、Guyatt GH 和 Chalmers IG[1–3] 的开创性工作使我们进入了循证医学（evidence–based medicine，EBM）时代。在这个时代，如何检索、评价和使用最佳证据非常重要。系统评价与 Meta 分析是科学总结一次研究数据最常用的方法 [4–6]，也是制定临床实践指南（clinical practice guideline，CPG）的基础 [7]。因此，为了进行系统评价和 / 或 Meta 分析，评估原始研究的方法学质量非常重要。当然，在使用系统评价 /Meta 分析前评估其自身的方法学质量亦是关键。质量包括内部真实性和外部真实性，而方法学质量通常是指内部真实性 [8–9]。内部真实性也被 Cochrane 协作网称为“偏倚风险（risk of bias，RoB）”[9]。

目前有三种类型的质量评价工具：量表、清单和条目 [10–11]。2015 年，Zeng 等 [11] 研究了用于随机对照试验（randomized controlled trial，RCT）、非随机临床干预研究、队列研究、病例 – 对照研究、横断面研究、病例系列、诊断准确性研究（也称为诊断准确性试验；diagnostic test accuracy，DTA）、动物实验、系统评价与 Meta 分析、以及 CPG 的方法学质量工具。由于，现有评价工具可能会发生一些变化、新工具也可能会出现，近年来也开发了新的研究方法，因此，有必要系统地研究用于评估方法学质量的常用工具，尤其是用于经济学评价、临床预测规则 / 模型和质性研究的工具。

本综述介绍了截至 2019 年 12 月的一次和二次医学研究的相关方法学质量（包括“RoB”）评价工具，表 1 列出了基本特征。希望本综述可以帮助证据的生产者、使用者和研究者。

表1　方法学质量（偏倚风险）评价工具的基本特征

Table 1. Basic characteristics of methodological quality (deviation risk) evaluation tools

序号	研发组织/团队	工具名称	研究类型
1	Cochrane协作网	Cochrane RoB工具和RoB 2.0工具	随机对照试验、诊断准确性研究
2	物理治疗证据数据库（PEDro）	PEDro量表	随机对照试验
3	照护的有效实践与组织（EPOC）	EPOC偏倚风险评估工具	随机对照试验、临床对照试验、前后对照研究、间断时间序列研究
4	文献严格评价项目（CASP）	CASP清单	随机对照试验、队列研究、病例–对照研究、横断面研究、诊断性研究、临床预测规则、经济学评价研究、质性研究、系统评价
5	美国国立卫生研究院（NIH）	NIH质量评价工具	对照干预研究、队列研究、横断面研究、病例–对照研究、前后对照研究（无对照组）、病例系列（干预性）系统评价与Meta分析
6	乔安娜·布里格斯学院（JBI）	JBI清单	随机对照试验、非随机试验研究、队列研究、病例–对照研究、横断面研究、患病率研究、病例报告、经济学评价研究、质性研究、文献和专家观点、系统评价
7	苏格兰国家指南小组（SIGN）	SIGN方法学清单	随机对照试验、队列研究、病例–对照研究、诊断性研究、经济学评价研究、系统评价与Meta分析
8	美国卒中治疗学术产业圆桌会议小组（STAIR）	CAMARADES工具	动物研究
9	荷兰动物实验系统评价研究中心（SYRCLE）	SYRCLE’s偏倚风险评估工具	动物研究
10	Sterne JAC等	ROBINS–I工具	非随机干预性研究
11	Slim K等	MINORS工具	非随机干预性研究
12	加拿大卫生经济学研究所（IHE）	IHE质量评价工具	病例系列（干预性）
13	Wells GA等	纽卡斯尔–渥太华量表（NOS）	队列研究 病例–对照研究
14	Downes MJ等	AXIS工具	横断面研究

续表1

序号	研发组织/团队	工具名称	研究类型
15	美国卫生保健和质量机构（AHRQ）	AHRQ方法学清单	横断面/患病率研究
16	Crombie I	Crombie条目	横断面研究
17	比较效果的优良研究（GRACE）团队	GRACE清单	比较效果研究
18	Whiting PF等	QUADAS工具和QUADAS-2工具	诊断准确性研究
19	英国国家健康和临床优化研究所（NICE）	NICE方法学清单	经济学评价研究
20	英国国家社会研究中心	质性研究评估框架	质性研究（社会研究）
21	Hayden JA等	QIPS工具	预测研究（预测因素研究）
22	Wolff RF等	PROBAST	预测研究（预测模型研究）
23	基于共识的健康测量工具选择标准（COSMIN）倡议	COSMIN RoB清单	患者报告的结果测量、内容真实性、结构真实性、内部一致性、跨文化真实性/测量不变性、信度、测量误差、校标真实性、结构真实性假设检验、反应度
24	Shea BJ等	AMSTAR和AMSTAR 2	系统评价
25	决策支持部门（DSU）	DSU网状Meta分析（NMA）方法学清单	网状Meta分析
26	Whiting P等	ROBIS工具	系统评价
27	Brouwers MC等	AGREE和AGREE II	临床实践指南

注：AMSTAR，a measurement tool to assess systematic reviews，一种评估系统评价的测量工具；AHRQ，Agency For Healthcare Research And Quality，美国卫生保健和质量机构；AXIS，appraisal tool for cross-sectional studies，横断面研究评价工具；CASP，critical appraisal skills programme，文献严格评价项目；CAMARADES，the collaborative approach to meta-analysis and review of animal data from experimental studies，实验研究动物数据的Meta分析和系统评价协作网；COSMIN，consensus-based standards for the selection of health measurement instruments，基于共识的健康测量工具选择标准；DSU，decision support unit，决策支持部门；EPOC，the effective practice and organisation of care group，照护的有效实践与组织；GRACE，the god research for comparative effectiveness initiative，比较效果的优良研究；IHE，Canada Institute Of Health Economics，加拿大卫生经济学研究所；JBI，Joanna Briggs Institute，乔安娜·布里格斯学院；MINORS，methodological index for non-randomized studies，非随机研究的方法学指标；NOS，Newcastle-Ottawa scale，纽卡斯尔-渥太华量表；NMA，network meta-analysis，网状Meta分析；NIH，National Institutes Of Health，美国国立卫生研究院；NICE，National Institute for Clinical Excellence，英国国家健康和临床优化研究所；PEDro，physiotherapy evidence database，物理治疗证据数据库；PROBAST，the prediction model risk of bias assessment tool，预测模型偏倚风险评估工具；QUADAS，quality assessment of diagnostic accuracy studies，诊断准确性研究质量评价；QIPS，quality in prognosis studies，预后研究质量；RoB，risk of bias，偏倚风险；ROBINS-I，risk of bias in non-randomised studies-of interventions，干预性非随机研究偏倚风险；ROBIS，risk of bias in systematic review，系统评价偏倚风险；SYRCLE，systematic review center for laboratory animal experimentation，动物实验系统评价研究中心；STAIR，Stroke Therapy Academic Industry Roundtable，美国卒中治疗学术产业圆桌会议小组；SIGN，The Scottish Intercollegiate Guidelines Network，苏格兰国家指南小组。

1 干预性研究

1.1 随机对照试验（个体或整群）

第一个 RCT 由 Hill BA（1897-1991）设计，并成为迄今为止实验研究设计的“金标准”[12-13]。如今，随机试验的 Cochrane 偏倚风险工具（于 2008 年制定并于 2011 年 3 月 20 日进行修订）是 RCT 中最常用的质量评价工具[9,14]，被称为“RoB”。2019 年 8 月 22 日（于 2016 年制定）发布了此工具的随机试验中评估 RoB 的修订版（RoB 2.0）[15]。RoB 2.0 工具适用于个体随机、平行组和整群随机试验，可从专用网站 https://www.riskofbias.info/welcome/rob-2-0-tool 上获取。RoB 2.0 工具包含五个偏倚领域，与原始 Cochrane RoB 工具相比有很大变化（附表 1-A-B 列出了这两个版本的主要条目）。

物理治疗证据数据库（PEDro）量表是专门用于 RCT 物理治疗的一种方法学评价工具[16-17]，可

在 http://www.pedro.org.au/english/downloads/pedro-scale/ 获取，内容涉及 11 个条目（附表 1-C）。照护的有效实践与组织（EPOC）是一个 Cochrane 评审小组，该小组开发了一种用于复杂干预随机试验的工具（称为“EPOC RoB 工具”）。该工具有 9 个条目（附表 1-D），可在 https://epoc.cochrane.org/resources/epoc-resources-review-authors 获取。文献严格评价项目（CASP）是牛津大学三重价值医疗保健中心（3V）的一部分，该产品组合提供资源以及学习和发展机会，以支持文献严格评价的发展（https://www.casp-uk.net/）[18-20]。RCT 的 CASP 清单由三个部分组成，涉及 11 个条目（附表 1-E）。美国国立卫生研究院（NIH）还开发了质量评价工具，用于对照干预研究（附表 1-F），以评估 RCT 的方法学质量（https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools）。

乔安娜·布里格斯学院（JBI）是一家独立的、国际性的、非营利性研究与开发组织，总部位于南澳大利亚州阿德莱德大学健康与医学学院（https://joannabriggs.org/）。它制定了许多重要的评估清单，涉及医疗保健干预措施的可行性、适当性、有意义性和有效性。附表 1-G 列出了针对 RCT 的 JBI 严格评估清单，其中包括 13 个条目。

苏格兰国家指南小组（SIGN）成立于 1993 年（https://www.sign.ac.uk/）。其目标是通过减少实践和结果的差异，基于当前证据的有效实践制定和传播国家临床指南，来提高苏格兰患者的卫生保健质量。它还制定了许多重要的评估清单，以评估包括 RCT（附表 1-H）在内的不同研究类型的方法学质量。

此外，Jadad 量表[21]、改良的 Jadad 量表[22-23]、Delphi 列表[24]、Chalmers 量表[25]、英国国家健康和临床优化研究所（NICE）方法学清单[11]、Downs & Black 清单[26]，以及 West 等在 2002 年总结的其他工具[27]，如今并不常用或不被被推荐使用。

1.2 动物实验

在开展临床试验之前，通常在动物模型中测试新药的安全性和有效性[28]，因此动物研究被视为临床前研究，具有重要意义[29-30]。同样，动物研究的方法学质量也需要评估[30]。1999 年，最初的“美国卒中治疗学术产业圆桌会议小组（STAIR）”推荐了他们评估卒中动物研究质量的标准[31]，该工具也称为“STAIR”。2009 年，STAIR 小组更新了标准，并制定了“确保高质量科学研究的推荐意见”[32]。此外，Macleod 等[33]在 2004 年提出了一种基于 STAIR 的工具来评估动物研究的方法学质量，总分为 10 分，也称为“CAMARADES（实验研究动物数据的 Meta 分析和评价的协作方法）”；其中“S”以前代表“卒中（Stroke）”，现在代表“研究（Studies）”（http://www.camarades.info/）。在 CAMARADES 工具中，每个条目的最高得分为 1 分，而该工具的总得分最高为 10 分（附表 1-J）。

2008 年，荷兰动物实验系统评价研究中心（SYRCLE）成立，该团队基于原始的 Cochrane RoB 工具[34]开发并发布了用于动物干预性研究的 RoB 工具——SYRCLE 的 RoB 工具。这个新工具包含 10 个条目，目前已成为评价动物干预性研究方法学质量最为推荐的工具（附表 1-I）。

1.3 非随机研究

在临床研究中，RCT 并不总是可行的[35]。因此，非随机设计仍然很重要。在非随机研究（也称为准实验研究）中，研究者控制参与者的分组，但未采用随机分组[36]，包括随访研究。根据是否进行比较，非随机临床干预研究可以分为比较和非比较两种亚型，非随机研究的偏倚风险—干预性（ROBINS-I）评价工具[37]是首选推荐工具。开发此工具的目的是评估非随机干预性研究的偏倚风险，这类研究评估干预措施的相对有效性（损害或获益），但未采用随机方法将受试者（个体或整群）分配给各组。此外，JBI 严格评估清单包括 9 个项目，也适用于准实验研究（非随机实验研究）。非随机研究的方法学指标（MINORS）[38]工具也可以被用于评价非随机研究，该工具共包含 12 个得分点；前 8 个条目可以用于非比较研究和比较研究，而后 4 个条目则适用于两组或多组的研究。每个项目的得分都从 0 到 2，总体质量得分为 16 或 24 分。附表 1-K-L-M 分别列出了这三个工具的主要条目。

具有单独对照组的非随机研究也可以称为临床对照试验或前后对照研究。对于这种设计类型，EPOC RoB 工具是合用的（附表 1-D）。使用此工具时，“随机序列生成”和“分配隐藏”应记为“高风险”，而其他项目的评分可能与随机试

验的评分相同。

没有单独对照组的非随机研究可以是前后对照研究、病例系列（非对照纵向研究）或间断时间序列研究。病例系列研究通常描述了一系列个体，接受相同的干预，并且没有对照组[9]。有几种工具可以评估病例系列研究的方法学质量。最新的一个工具是在 2012 年由 Moga C 等[39]使用加拿大卫生经济学研究所（IHE）研发的改良德尔菲技术开发的，因此，也被称为“IHE 质量评价工具”（附表 1–N）。此外，NIH 也为病例系列研究开发了质量评价工具，其中包括 9 个项目（附表 1–O）。对于间断时间序列研究，建议使用“EPOC 间断时间序列研究 RoB 工具”（附表 1–P）；对于前后研究，建议使用 NIH 无对照组的前后研究质量评价工具（附表 1–Q）。

此外，对于非随机干预性研究，Reisch 工具（治疗研究评估清单）[11,40]、Downs & Black 清单[26]以及 Deeks 等总结的其他工具[36]，如今已不常用或不被推荐使用。

2 观察性研究和诊断研究

观察性研究包括队列研究、病例 – 对照研究、横断面研究、病例系列、病例报告和比较效果研究[41]，可分为分析性研究和描述性研究[42]。

2.1 队列研究

队列研究包括前瞻性队列研究、回顾性队列研究和双向性队列研究[43]。有一些评估队列研究质量的工具，例如 CASP 队列研究清单（附表 2–A）、SIGN 队列研究清单（附表 2–B）、NIH 观察性队列研究和横断面研究质量评价工具（附表 2–C），用于队列研究的纽卡斯尔 – 渥太华量表（NOS，附表 2–D）和用于队列研究的 JBI 清单（附表 2–E）。但是，Downs & Black 清单[26]和 NICE 清单[11]现已不常使用或不被推荐使用。

NOS 量表[44–45]由澳大利亚纽卡斯尔大学和加拿大渥太华大学之间合作研发，是目前评价队列研究最常用的工具，使用者可以根据特定主题进行修改。

2.2 病例–对照研究

病例 – 对照研究根据是否存在特定疾病或状况选择受试者，并寻找可能导致疾病或结局的早期暴露因素[42]。相比队列研究，它的优势在于不会出现受试者“脱落”或“失访”的问题。目前，有一些可评估病例 – 对照研究方法学质量的工具，包括 CASP 病例 – 对照研究清单（附表 2–F）、SIGN 病例 – 对照研究清单（附表 2–G）、NIH 病例 – 对照研究质量评价工具（附表 2–H），JBI 病例 – 对照研究清单（附表 2–I）和 NOS 病例 – 对照研究量表（附表 2–J）。其中，NOS 也是当今最常用于评价病例 – 对照研究的工具，并且可以根据特定的主题进行修改。

此外，Downs & Black 清单[26]和 NICE 清单[11]现已不常用或不被推荐使用。

2.3 横断面研究（分析性或描述性）

横断面研究是用来提供某个时间点特定人群中疾病和其他变量的研究。它可以分为分析性和描述性。描述性横断面研究仅描述特定人群在某个时间点或一段时间内的病例或事件的数量；而分析性横断面研究可用于推断疾病与其他变量之间的关系[46]。

为了评估分析性横断面研究的质量，目前推荐的工具有 NIH 质量评价工具（附表 2–C），JBI 分析性横断面研究评估清单（附表 2–K）和横断面研究评价工具（appraisal tool for cross–sectional studies，AXIS 工具，附表 2–L）[47]。AXIS 工具是一项评估研究设计和报告质量以及偏倚风险的工具，该工具于 2016 年开发，包含 20 个条目。在这三种工具中，JBI 清单是最常用的一种。

通常使用描述性横断面研究来描述疾病的患病率和发病率。因此，用于分析性横截面研究的评价工具并不适用。只有很少的质量评价工具适用于描述性横断面研究，例如用于报告流行率数据研究的 JBI 量表[48]（附表 2–M），美国卫生保健质量和研究机构（AHRQ）用于评估横断面研究 / 患病率研究的方法学量表（附表 2–N），以及 Crombie 用于评估横断面研究质量的条目[49]（附表 2–O）。其中，JBI 工具是最新的。

2.4 病例系列和病例报告

与上述干预性病例系列不同，病例报告和病例系列用于报告新发疾病或独特发现[50]。因此，它们属于描述性研究。仅有 JBI 清单这一种工具用于病例报告的方法学质量评价（附表 2–P）。

2.5 比较效果研究

比较效果研究（comparative effectiveness research，CER）比较了某医疗条件的替代治疗方案真实世界情况下的结果[51]。它的关键要素包括

效果的研究（在真实世界中的效果），而不是效力（理想效果），以及替代策略之间的比较[52]。2010 年，比较效果的优良研究（GRACE）团队成立并制定了相关原则，以帮助医疗保健提供者、研究人员、期刊读者和编辑者评估观察性比较效果研究的质量[41]。2016 年，发布了 GRACE 清单 5.0 版（附表 2–Q），用于评估 CER 的质量。

2.6 诊断性研究

临床医生使用诊断试验（也称为“诊断准确性试验，DTA”）来确定患者是否存在某种状况，从而制定适当的治疗计划[53]。DTA 在设计方面具有一些独特的特征，这些特征不同于标准的干预性研究和观察性研究。2003 年，Whiting 等[53–54]开发了一种评估 DTA 质量的工具，即诊断准确性研究质量评估（quality assessment of diagnostic accuracy studies，QUADAS）工具。2011 年，推出了修订的“QUADAS–2”工具（附表 2–R）[55–56]。此外，该领域常用的评价工具有 CASP 诊断清单（附表 2S）、SIGN 诊断研究清单（附表 2–T）、JBI 诊断准确性试验评估清单（附表 2–U）和 Cochrane 诊断准确性试验偏倚风险评估工具（附表 2–V）。

其中，Cochrane 偏倚风险工具（https://methods.cochrane.org/sdt/）基于 QUADAS 工具，而 SIGN 清单和 JBI 清单则基于 QUADAS–2 工具。目前 QUADAS–2 工具是最为推荐的工具。在 2004 年 Whiting 等[53]综述中提及的其他相关工具如今已不再使用。

3 其他类型一次研究

3.1 卫生经济学评价

卫生经济学评价研究比较了替代干预措施的资源使用、成本和健康影响[57]。它着重于确定、衡量、评估和比较两种或多种替代干预方案的资源使用、成本和效益 / 效果[58]。如今，卫生经济学研究越来越受欢迎。当然，其方法学质量也需要在使用之前进行评估。Drummond 和 Jefferson 在 1996 年开发了第一个进行此类评估的工具[59]，之后，许多工具根据 Drummond 的条款或其修订版[60]被开发出来。例如 SIGN 经济学评价清单（附表 3–A），CASP 经济学评价清单（附表 3–B）、JBI 经济学评价清单（附表 3–C）和 NICE 用于经济学评价的方法清单（附表 3–D）等。

我们认为综合健康经济评估报告标准（consolidated health economic evaluation reporting standards，CHEERS）声明[61]属于报告工具，而不是方法学质量评价工具，因此不建议使用它来评估卫生经济学评价研究的方法学质量。

3.2 质性研究

在医疗保健领域，质性研究旨在理解和解释个人经历、行为、互动和社会环境，以解释感兴趣的现象，例如患者和临床医生的态度、信念和观点，照顾者与患者之间的人际关系，疾病经历，以及患者痛苦的影响[62]。与定量研究相比，用于质性研究的评价工具更少。如今，CASP 质性研究清单（附表 3–E）是最常用的工具。此外，JBI 质性研究清单[63–64]（附表 3–F）和英国国家社会研究中心的质性研究评估框架[65]（附表 3–G）也同样适用。

3.3 预测研究

临床预测研究包括预测因子发现（预后因素）研究，预测模型研究（研发、验证以及扩展或更新）和预测模型影响研究[66]。对于预测因子发现研究，可以使用预后研究质量（quality in prognosis studies，QIPS）工具[67]评估其方法学质量（附表 3–H）。对于预测模型影响研究，如果使用随机比较设计，则可以使用 RCT 评价工具，如 RoB 2.0 工具。如果使用非随机比较设计，则可以使用非随机研究工具，如 ROBINS–I 工具。对于诊断和预后预测模型研究，可以使用预测模型偏倚风险评价工具（prediction model risk of bias assessment tool，PROBAST；附表 3–I）[68]和 CASP 临床预测规则清单（附表 3–J）。

3.4 文献和专家观点

基于文献和专家观点的证据（也称为“非研究证据”）来自各种期刊、杂志、专著和报告中出现的专家观点、共识、论述、评论以及假说[69–71]。如今，只有 JBI 清单可用于评估文献和专家观点的质量（附表 3–K）。

3.5 结局测量工具

结果测量工具用于收集测量结果。“工具”一词涵盖的范围很广，可以指问卷（例如患者报告的生活质量结果、观察（例如临床检查的结果）、量表（例如视觉模拟量表）、实验室检查（例如血液检查）和影像学（例如超声或其他医学成像）[72–73]。测量可以是主观的或客观的，可以是一维的（例如态度）或多维的。目前，只有基于共识标准选

择健康测量工具（COSMIN）偏倚风险量表[74-76]（https://www.cosmin.nl/）这一种工具适合评估结果测量工具的方法学质量，附表 3-L 列出了其主要项目，包括患者报告结果测量（patient-reported outcome measure，PROM）的发展（附表 3-LA）、内容真实性（附表 3-LB）、结构真实性（附表 3-LC）、内部一致性（附表 3-LD）、跨文化真实性/测量不变性（附表 3-LE）、信度（附表 3-LF）、测量误差（附表 3-LG）、校标真实性（附表 3-LH）、结构真实性假设检验（附表 3-LI）和反应度（附表 3-LJ）。

4 二次研究

4.1 系统评价与Meta分析

系统评价与 Meta 分析是总结当前医学文献的科学方法[4-6]，其最终目的和价值在于促进医疗保健[6,77-78]。Meta 分析是一个将多项研究结果合并的统计过程，通常是系统评价的一部分[11]。当然，在使用系统评价与 Meta 分析之前，必须进行严格的评价。

1988 年，Sacks 等开发了第一个评估基于 RCT 的 Meta 分析的质量工具——Sack 质量评估清单（sack's quality assessment checklist，SQAC）[79]；1991 年，Oxman 和 Guyatt 开发了另一个工具——概述质量评估问卷（overview quality assessment questionnaire，OQAQ）[80-81]。为了克服这两个工具的缺点，2007 年，基于这两个工具开发了一种评估系统评价的评价工具（a measurement tool to assess systematic reviews，AMSTAR）[82]（https://www.amstar.ca/）。但是，初始版本的 AMSTAR 工具并未包含对非随机研究偏倚风险的评估，专家组认为修订应针对系统评价的所有方面。因此，2017 年发布了用于评价随机和非随机研究的新工具——AMSTAR 2[83]，附表 4-A 列出了其主要项目。

此外，CASP 的系统评价清单（附表 4-B）、SIGN 的系统评价与 Meta 分析评价清单（附表 4-C）、JBI 的系统评价与研究整合评价清单（附表 4-D）、NIH 的系统评价与 Meta 分析质量评价工具（附表 4-E）、英国约克大学决策支持单位（decision support unit，DSU）网状 Meta 分析（NMA）方法学清单（附表 4-F）和系统评价偏倚风险（risk of bias in systematic review，ROBIS）[84] 工具（附表 4-G）都可用于系统评价与 Meta 分析的质量评价。其中，最常用的是 AMSTAR 2、最常建议使用的是 ROIBS。

在这些工具中，AMSTAR 2 适用于评估随机或非随机干预性研究的系统评价和 Meta 分析，DSU-NMA 方法清单适用于网状 Meta 分析，而 ROBIS 适用于干预性研究、诊断准确性试验、临床预测和预后研究的 Meta 分析。

4.2 临床实践指南

CPG 很好地融入了临床医生和专业临床组织的观念或经验[85-87]；并将科学证据纳入临床实践[88]。但是，并非所有的 CPG 都是基于证据的[89-90]，其质量也参差不齐[91-93]。到目前为止，已经开发了 20 多种评价工具[94]。其中，指南研究与评价工具（AGREE）可作为开发临床途径评价工具的基础[94]。AGREE 工具于 2003 年首次发布[95]，并于 2009 年更新为 AGREE II 工具[96]（https://www.agreetrust.org/）。现在，AGREE II 工具是用于评价 CPG 最为推荐的工具（附表 4-H）。

此外，基于 AGREE II，开发了 AGREE 全球评级量表（agree global rating scale，AGREE GRS）工具[97]，作为评估 CPG 质量和报告的简要工具。

5 结语

目前，循证医学已被广泛接受，医护人员的主要注意力在于“从证据到建议”[98-99]。因此，在使用之前对证据进行严格的评价是该过程的关键[100-101]。1987 年，Mulrow CD[102] 指出，医学综述需要常规使用科学的方法来识别、评估和综合信息。故在使用研究结果之前，必须对研究进行方法学质量评估。尽管自第一个质量评价工具问世以来已经过去了 20 多年，但许多用户仍然误解了方法学质量和报告质量。其中，有人使用报告清单来评估方法学质量，例如使用报告临床试验的统一标准（consolidated standards of reporting trials，CONSORT）声明[103] 来评估 RCT 的方法学质量，使用流行病学观察性研究增强报告（strengthening the reporting of observational studies in epidemiology，STROBE）评价队列研究方法学质量[104]。这种现象表明，需要对医学生和专业人员进行更多的临床流行病学普及教育。

方法学质量工具应根据不同研究类型的特征开发。本文中，我们使用“methodological

quality”“risk of bias”“critical appraisal”“checklist”“scale”“items”和“assessment tool”在 NICE 网站、SIGN 网站、Cochrane 图书馆网站和 JBI 网站中进行搜索。在此基础上，在 PubMed 中检索了“systematic review”“meta-analysis”“overview”和“clinical practice guideline”。与本团队之前的系统评价[11]相比，我们发现一些工具仍被推荐和使用，某些工具仍在使用但未被推荐，而有些则被淘汰[10,29-30,36,53,94,105-107]。这些工具极大地推动了临床实践的发展[108-109]。

此外，相较本团队既往研究成果[11]，本文列出了更多的工具，尤其是 2014 年之后开发的新工具和最新修订版。当然，我们还调整了研究类型分类的方法。首先，2014 年，NICE 提供了 7 个方法学检查清单，但现在仅保留和更新了经济学评价清单。此外，Cochrane RoB 2.0 工具、AMSTAR 2 工具、CASP 清单和大多数 JBI 清单都是最新修订版；NIH 的质量评价工具、ROBINS-I 工具、EPOC RoB 工具、AXIS 工具、GRACE 清单、PROBAST、COSMIN 偏倚风险清单和 ROBIS 工具都是新发布的工具。其次，本文还介绍了用于评价网状 Meta 分析、结果测量工具、文献和专家观点、预测研究、质性研究、卫生经济学评估和 CER 的工具。第三，我们将干预研究分为随机和非随机两种亚型，然后将非随机研究进一步分为有对照组和无对照组；此外，还将横断面研究分为分析性和描述性两种亚型，病例系列分为干预性和观察性两种亚型。这些分类更加客观和全面。

显然，适用于 RCT 的评价工具数量最多，其次是队列研究。JBI 的适用范围最广[63-64]，CASP 紧随其后。但是，仍需进一步努力来开发评价工具。对于某些研究类型，仅有一种适用的评价工具，例如 CER、结果测量工具、文献和专家观点、病例报告和 CPG。此外，对于许多研究类型，例如概述、遗传关联研究和细胞研究，都没有合适的评价工具。而且现有些工具尚未被学界完全认可。将来，如何开发公认的工具仍然是一项具有重要意义的工作[11]。

本综述可以帮助系统评价、Meta 分析、指南和证据使用者等专业人员在产生或使用证据时选择最佳工具。而且，方法学家可以获得开发新工具的研究主题。最重要的是，我们必须明白，所有评价工具都是主观的，使用这些工具时会受到用户的技能和知识水平的影响。因此，用户必须接受正规培训（必须具备相关的流行病学知识），具有严谨的学术态度，并且至少应由两名独立的审阅者参与评估和交叉检查，以最大限度地避免出现实施偏倚[110]。

附表中文版位置说明

附表 1：干预性研究评估工具及其主要组成部分

附表 2：观察性研究和诊断研究评估工具及其主要组成部分

附表 3：医学领域其他的一次研究评估工具及其主要组成部分

附表 4：医学二次医学研究评估工具及其主要组成部分

读者可自《医学新知》官网（http://www.jnewmed.com/）相对应文章中获取。

参考文献

见原文。

收稿日期：2020 年 10 月 16 日 修回日期：2020 年 12 月 28 日
本文编辑：桂裕亮 杨智华

引用本文：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. DOI: 10.12173/j.issn.1004-5511.2021.01.07.
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]. Yixue Xinzhi Zazhi, 2021, 31(1): 51-58. DOI: 10.12173/j.issn.1004-5511.2021.01.07.