

基于 COSMIN 指南对中医证候疗效评价量表的系统评价

苏凯林^{1,2,3}, 冯贞贞^{1,2,3}, 王佳佳^{1,2,3}, 王露^{1,2,3}, 赵贵香^{1,2,3}, 李建生^{1,2,3}✉

1. 河南中医药大学第一附属医院, 河南省郑州市金水区人民路 19 号, 450003; 2. 河南中医药大学第一临床医学院; 3. 河南中医药大学呼吸疾病中医药防治省部共建协同创新中心



[摘要] **目的** 系统评价中医证候疗效评价量表的方法学质量及测量学特性, 为中医临床实践选用高质量评估工具提供参考。**方法** 检索中国知网、万方数据库、维普网、中国生物医学文献数据库、PubMed、Cochrane Library、Embase 和 Web of Science 自建库至 2025 年 4 月 2 日收录的涉及中医证候疗效评价量表测量学特性评价的文献, 提取资料, 依据健康测量工具共识标准 (COSMIN) 评估纳入量表的方法学质量及测量学特性, 并结合证据分级标准形成相应推荐意见。**结果** 共纳入 46 项研究, 涵盖 22 个普适性证候疗效评价量表和 24 个疾病特异性证候疗效评价量表, 均未报告跨文化效度和测量误差。推荐等级显示, 2 个量表符合 A 级推荐, 建议临床使用; 38 个量表为 B 级推荐, 提示存在潜在应用价值但需进一步验证; 6 个量表为 C 级推荐, 建议进一步完善。**结论** 现有中医证候疗效评价量表存在方法学质量参差不齐、测量学特性报告不完整及修订重视度不足等问题, 未来应加强中医证候量表研制的规范化设计, 完善关键测量属性的验证流程, 注重对量表进行动态修订, 为证候疗效的精准评价提供高质量工具支持。

[关键词] 中医证候; 疗效评价; 量表; 测量学特性; COSMIN 指南; 系统评价

构建符合中医药特点的疗效评价体系, 是中医药现代化、国际化进程中亟待攻克的重大工程技术难题^[1]。证候是中医疗效评价的核心指标, 已成为中医疗效评价体系中不可或缺的一部分^[2]。量表具有客观、量化、标准化等特点, 是近年来最常用的中医证候疗效评价方法与载体^[3]。良好的测量学特性是量表应用的基本前提^[4], 但当前尚缺乏专门针对证候疗效评价量表测量学特性的标准化评价工具。健康测量工具共识标准 (consensus-based standards for the selection of health measurement instruments, COSMIN) 指南由多国心理测量学领域专家联合开发, 为测评工具的质量评估提供了标准化框架与规范流程, 其核心测量学特性评价维度 (信度、效度、反应度) 与当前证候疗效评价量表研发中涉及的测量学指标高度契合^[2, 5]。在国内, 已有学者将

COSMIN 指南用于中医临床疗效评价量表、中医生活质量量表心理测量学性能的评价^[6-7], 初步验证了其在中医特色测量工具评价领域的适用性及价值。因此, 本研究基于 COSMIN 指南, 系统评价现有中医证候疗效评价量表的方法学质量和测量学特性, 据此形成证据推荐建议以帮助研究者选择高质量量表应用于临床实践。同时, 系统梳理当前量表研制过程中的方法学实践现状与关键问题, 为后续相关研究的开展与优化提供参考。本研究严格按照结局测评工具系统评价报告规范 (PRISMA-COSMIN) 执行^[8], 已在国际前瞻性系统评价注册平台注册 (注册号: CRD420251026383)。

1 资料与方法

1.1 文献检索与筛选

检索中国知网、万方数据库、维普网、中国生物医学文献数据库、PubMed、Cochrane Library、Embase 和 Web of Science 收录的有关中医证候疗效评价量表测量学特性评价的文献。检索时限自各数据库建库至 2025 年 4 月 2 日, 参考 COSMIN 指南推

基金项目: 国家重点研发计划中医药现代化重点专项 (2024YFC3505302); 河南省国家中医药传承创新中心科研专项 (2023ZXZX1126); 河南省“双一流”创建科学研究专项 (HSRP-DFCTCM-2023-3-01, HSRP-DFCTCM-2023-4-09)

✉ 通讯作者: li_js8@163.com

荐检索策略^[5]，综合运用主题词与自由词，并对纳入文献的参考文献实施补充检索。由 2 名研究者严格遵循纳入、排除标准独立筛选并交叉核对。如有分歧，通过讨论或与第三名研究者协商解决。

文献纳入标准：1) 研究对象具备明确的中医证候诊断；2) 研究涉及中医证候疗效评价量表的开发或验证；3) 研究至少包含量表的 1 种测量学特性评估结果。

文献排除标准：1) 动物实验、会议论文、综述类文献；2) 无法获取全文及重复发表的文献。

1.2 资料提取

通过 Excel 2019 表格进行数据录入与统计，提取信息包括文献基本信息、研究对象的基本特征、量表条目数、量表完成时间及测量学特性等。

1.3 质量评价

1.3.1 方法学质量评价 基于 COSMIN 偏倚风险清单开展系统性评估^[5]。该清单围绕内容效度、内部结构和其他测量属性 3 个部分展开，内容效度包括工具开发与内容效度，内部结构包括结构效度、内部一致性与跨文化效度，其他测量属性包括稳定性、测量误差、效标效度、假设检验、反应度，共计 10 个核心模块，113 条具体评估条目。条目评价采用“很好”“良好”“模糊”“不良”4 点评分法，各模块质量评级遵循最低计分原则，即选取该模块内条目最低评分作为模块最终评级。

1.3.2 测量学特性评价 基于 COSMIN 良好测量学特性标准开展系统性评估^[5]。该标准针对上述 10 个模块中除“工具开发”外的 9 个模块所对应的测量学特性，各特性评价结果包括“充分”“不确定”“不充分”3 个等级。

1.3.3 证据等级评价 基于 COSMIN 改良版证据分级标准开展系统性评级^[9-10]。预设测量学特性证据质量为“高质量”，然后依据偏倚风险、不一致性、间接性、不精确性 4 个因素对其进行等级调整，划分为“高”“中”“低”“极低”4 个等级，最终结合测量学特性结果形成推荐意见。A 级推荐要求工具内容效度充分（不限证据等级）且内部一致性充分（至少低等级证据支持），推荐使用；C 级推荐适用于存在高质量证据证实工具任一测量学特性不充分，不建议使用；B 级推荐适用于未满足 A 级或 C 级判定，提示具备潜在适用性，但需通过后续研究完善验证。

2 结果

2.1 文献筛选结果

通过系统检索及参考文献追溯共获得文献 2476 篇，查重后剔除 528 篇，阅读标题、摘要排除 1874 篇，阅读全文后排除 28 篇，最终纳入文献 46 篇。文献筛选流程见附图 1（请扫描本文二维码获取）。

2.2 纳入研究基本特征

共纳入 46 个证候疗效评价量表^[11-56]，其中普适性量表 22 个，疾病特异性量表 24 个。普适性量表涵盖 16 个证候，出现频次排名前六位的依次为阴虚证（2 次）、阳虚证（2 次）、血瘀证（2 次）、血虚证（2 次）、气滞证（2 次）和气虚证（2 次）；疾病特异性量表涉及 14 个病种，频次排名前三位的依次为冠心病（9 次）、慢性阻塞性肺疾病（2 次）和慢性乙型肝炎（2 次）。研究实施范围呈现均衡分布特征，单中心研究与多中心研究各 23 项。各量表原始文献在中国知网数据库中的被引频次范围为 0~56 次，纳入研究的基本特征详见附表 1（请扫描本文二维码获取）。

2.3 纳入研究的方法学质量和测量学特性评价

纳入的 46 项研究中，跨文化效度与测量误差的相关信息均未报告，其余评价结果详见附表 2（请扫描本文二维码获取）。

2.3.1 工具开发 纳入研究中是否对小组会议或访谈进行录音转录与量性研究样本量大小是影响工具开发阶段方法学质量的主要原因。31 项研究^[14, 17, 19, 21-23, 26, 28, 31, 34-51, 53-56]未明确描述录音转录相关内容，方法学质量评为“模糊”；1 项研究^[18]样本量 ≥ 50 且 < 100 例，10 项研究^[11-13, 15, 20, 25, 29-30, 32-33]样本量 ≥ 30 且 < 50 例，3 项研究^[24, 27, 52]样本量 < 30 例，方法学质量分别评为“良好”“模糊”与“不良”。研制证候疗效评价量表应注重界定证候的诊断与疗效评价属性，进而筛选具有疗效评价属性的指标^[57]。1 项研究^[16]主要通过参考证候诊断标准构建条目池，方法学质量评为“模糊”。

2.3.2 内容效度 内容效度评价基于相关性、全面性和可理解性 3 个核心维度展开。40 项研究^[11-15, 17-27, 29-30, 32-53]报告了内容效度，其中 25 项研究^[14, 17-19, 21-24, 26, 34-35, 37-38, 40-45, 47-48, 50-53]详细阐述相关性维度信息，38 项研究^[11-15, 17-27, 29-30, 32-35, 37-45, 47-53]报告了全面性与可理解性维度内容。共 12 项研究^[18, 22, 35, 37, 40-44, 47-48, 52]在评估过程中综合考虑患

者与专家的意见,其中3项研究^[22, 35, 48]在上述3个维度均协同分析患者与专家观点,测量学特性为“充分”,其余研究均为“不确定”。患者访谈阶段中,11项研究^[11-13, 15, 18, 20, 25, 29-30, 32-33]因访谈者均接受专项研究培训,方法学质量为“良好”;3项研究^[22, 27, 52]访谈者样本量不足30例,13项研究^[35-37, 39-44, 46-49]未明确描述录音转录相关内容,故上述16项研究的方法学质量均为“模糊”。专家咨询环节中,7项研究^[22, 26, 34, 45, 50-52]因纳入专家超过30名且其专业背景覆盖研究目标领域所需学科构成,方法学质量为“良好”;16项研究^[14, 17-18, 21, 23-24, 37-38, 40-44, 47-48, 53]纳入专家人数不足30名,2项研究^[19, 35]未完整报告专家人数信息,因此上述18项研究的方法学质量均为“模糊”。

2.3.3 结构效度 4项研究^[18, 28, 46, 54]应用验证性因子分析(CFA),其中1项研究^[28]因样本量低于条目数的5倍,方法学质量降级为“不良”;4项研究^[18, 28, 46, 54]报告的拟合指数 <0.95 ,故测量学特性均为“不充分”。40项研究^[11-13, 15-17, 19-20, 22-27, 29-45, 47-53, 55-56]应用探索性因子分析(EFA),其中2项^[29, 52]样本量不足100,3项研究^[24, 39, 56]样本量低于条目数的5倍,因此方法学质量分别降级为“模糊”和“不良”;2项研究^[16, 48]样本量充足且因子分析结果与前期构想一致,测量学特性为“充分”,17项研究^[17, 19, 22-26, 29, 35-36, 38, 43, 45, 50, 52, 55-56]未提供完整的因子载荷矩阵,故测量学特性为“不确定”,21项研究^[11-13, 15, 20, 27, 30-34, 37, 39-42, 44, 47, 49, 51, 53]由于超过10%的条目在多个因子上的载荷 ≥ 0.30 或未满足所有条目在其主因子上的载荷量 ≥ 0.30 ,测量学特性为“不充分”。2项研究^[14, 20]未采用因子分析,结构效度评价分别为“不良”与“不确定”。

2.3.4 内部一致性 44项研究^[11-20, 22-38, 40-56]采用Cronbach's α 系数进行评估,故方法学质量为“很好”;2项研究^[17, 19]未具体报告量表Cronbach's α 系数,4项研究^[22, 32-33, 54]Cronbach's α 系数 <0.7 ,其余研究Cronbach's α 系数均 >0.7 ,因此测量学特性分别为“不确定”“不充分”与“充分”。

2.3.5 稳定性 23项研究^[11-15, 20-23, 25, 27-30, 32-33, 45, 48-50, 52-54]采用重复测量法以评估量表的稳定性,但均未详细描述重复测量期间受试者待测指标的稳定性、测量条件的一致性及时时间间隔的合理性,方法学质量为“模糊”;3项研究^[22-23, 45]未清晰说明两次测量间相关系数的具体分析,1项研究^[54]报告

Spearman <0.7 ,其余研究的组内相关系数(kappa/Pearson/Spearman)均 >0.7 ,故测量学特性分别为“不确定”“不充分”与“充分”。

2.3.6 效标效度 COSMIN指南规定在效标效度检验中新开发的量表只能以原版量表为“金标准”。10项研究^[20, 23, 27-29, 32-33, 38, 50, 52]采用理化指标、医生判断或其他常用量表为金标准进行效标效度评估,均未达到COSMIN指南要求,故效标效度评为“不良”和“不充分”。

2.3.7 假设检验 结果显示共有33项研究^[11-13, 15, 18, 20-21, 24-25, 27-34, 37-45, 47-49, 51-52, 55-56]完成了对量表的假设检验评估。2项研究^[18, 28]通过比较量表与其他测量工具报告了聚合效度,由于未明确阐述对照测量工具的测量属性且比较结果不佳,故方法学质量为“不良”,测量学特性为“不充分”。31项研究^[11-13, 15, 20-21, 24-25, 27, 29-34, 37-45, 47-49, 51-52, 55-56]使用量表进行亚组比较报告了已知群体效度或区分效度,其中7项研究^[38, 40-44, 47]未充分报告亚组的重要特征,方法学质量为“模糊”,其余研究方法学质量为“良好”;3项研究^[38-39, 49]未全面报告已知群体效度结果,测量学特性为“不明确”,其余研究均满足75%的结果与假设相符的条件,故测量学特性为“良好”。

2.3.8 反应度 结果显示共有22项研究^[12-16, 21-22, 25, 27-28, 30-34, 45, 48-49, 51, 53, 55-56]采用配对 t 检验或非参数检验对干预前后量表得分变化进行分析以评估反应度,两种分析方法均存在一定局限性^[4],因此反应度评为“不良”和“不充分”。

2.4 纳入工具的证据等级评价及推荐意见

偏倚风险中,内容效度维度:15个量表^[11-13, 15, 20, 25-26, 29-30, 32-34, 45, 50-51]因评级“良好”实施降1级,25个量表^[14, 17-19, 21-24, 27, 35-44, 46-49, 52-53]因评级“模糊”执行降2级。结构效度维度:3个量表^[18, 46, 54]因评级“很好”不予降级,35个量表^[11-13, 15-17, 19-20, 22-23, 25-27, 30-38, 40-45, 47-51, 53, 55]因评级“良好”实施降1级,“模糊”评级的2个量表^[29, 52]与“不良”评级的6个量表^[14, 21, 24, 28, 39, 56]分别执行降2级和降3级处理。44个量表^[11-20, 22-38, 40-56]的内部一致性均“很好”,不予降级。23个量表^[11-15, 20-23, 25, 27-30, 32-33, 45, 48-50, 52-54]的稳定性“模糊”,降2级。10个量表^[20, 23, 27-29, 32-33, 38, 50, 52]的效标效度“不良”,降3级。假设检验维度:24个量表^[11-13, 15, 20-21, 24-25, 27, 29-34, 37, 39, 45, 48-49, 51-52, 55-56]因评级

“良好”实施降 1 级，“模糊”评级的 7 个量表^[38, 40-44, 47]与“不良”评级的 2 个量表^[18, 28]分别执行降 2 级和 3 级处理。22 个量表^[12-16, 21-22, 25, 27-28, 30-34, 45, 48-49, 51, 53, 55-56]的反应度“不良”，降 3 级。不一致性：各量表均未重复应用于多项研究，故不存在不一致性。间接性：各量表研究中实际聚焦的群体及构念与预先设定内容保持一致，因此不存在间接性。不精确性：在重测信度检验阶段，1 个量表^[50]因样本量低于 100 导致评级

降 1 级，19 个量表^[11-15, 20-23, 25, 27-30, 32-33, 45, 48, 52]因样本量低于 50 导致评级降 2 级。

综合以上因素进行证据等级评价。2 个量表因内容效度“充分”且获高质量证据支持其内部一致性“充分”，为 A 级推荐；3 个量表因高质量证据显示结构效度“不充分”，另有 3 个量表因高质量证据表明内部一致性“不充分”，均列为 C 级推荐；其余 38 个量因未满足 A 级或 C 级标准，被归类为 B 级推荐，详见表 1。

表 1 中医证候疗效评价量表系统评价纳入文献的证据等级评价及推荐意见

评估工具	内容效度		结构效度		内部一致性		稳定性		效标效度		假设检验		反应度		推荐等级
	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据	
	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量	
血热证中医疗效评价量表 ^[11]	?	中	-	中	+	高	+	极低	NR	NR	+	中	NR	NR	B
内风证中医疗效评价量表 ^[12]	?	中	-	中	+	高	+	极低	NR	NR	+	中	?	极低	B
外风证中医疗效评价量表 ^[13]	?	中	-	中	+	高	+	极低	NR	NR	+	中	?	极低	B
广泛性焦虑障碍心脾两虚证评价量表 ^[14]	?	低	?	极低	+	高	+	极低	NR	NR	NR	NR	?	极低	B
实火证中医疗效评价量表 ^[15]	?	中	-	中	+	高	+	极低	NR	NR	+	中	?	极低	B
帕金森病肾虚血瘀证中医证候量表 ^[16]	NR	NR	+	中	+	高	NR	NR	NR	NR	NR	NR	?	极低	B
AECOPD 证候疗效评价量表 ^[17]	?	低	?	中	?	高	NR	NR	NR	NR	NR	NR	NR	NR	B
冠心病痰湿证 PRO 量表 ^[18]	?	低	-	高	+	高	NR	NR	NR	NR	-	极低	NR	NR	C
COPD 稳定期证候疗效评价量表 ^[19]	?	低	?	中	?	高	NR	NR	NR	NR	NR	NR	NR	NR	B
血虚证中医疗效评价量表 ^[20]	?	中	-	中	+	高	+	极低	-	极低	+	中	NR	NR	B
老年冠心病阳虚证 CRO 临床疗效评价量表 ^[21]	?	低	?	极低	NR	NR	+	极低	NR	NR	+	中	?	极低	B
小儿积滞病中医证候疗效评价量表 ^[22]	+	低	?	中	-	高	?	极低	NR	NR	NR	NR	?	极低	C
火热炽盛证评价标准 ^[23]	?	低	?	中	+	高	?	极低	-	极低	NR	NR	NR	NR	B
老年冠心病阳虚证 PRO 临床疗效评价量表 ^[24]	?	低	?	极低	+	高	NR	NR	NR	NR	+	中	NR	NR	B
气滞证中医疗效评价量表 ^[25]	?	中	?	中	+	高	+	极低	NR	NR	+	中	?	极低	B
慢性心力衰竭气虚证评价量表 ^[26]	?	中	?	中	+	高	NR	NR	NR	NR	NR	NR	NR	NR	B
气虚证中医疗效评价量表 ^[27]	?	低	-	中	+	高	+	极低	-	极低	+	中	?	极低	B
冠心病心绞痛痰瘀互结证疗效评价量表 ^[28]	NR	NR	-	极低	+	高	+	极低	-	极低	-	极低	?	极低	B
血瘀证中医疗效评价量表 ^[29]	?	中	?	低	+	高	+	极低	-	极低	+	中	NR	NR	B
阳虚证中医疗效评价量表 ^[30]	?	中	-	中	+	高	+	极低	NR	NR	+	中	?	极低	B
气虚血瘀证评价量表 ^[31]	NR	NR	-	中	+	高	NR	NR	NR	NR	+	中	?	极低	B
痰浊证中医疗效评价量表 ^[32]	?	中	-	中	-	高	+	极低	-	极低	+	中	?	极低	C
阴虚证中医疗效评价量表 ^[33]	?	中	-	中	-	高	+	极低	-	极低	+	中	?	极低	C
冠心病心绞痛血瘀证疗效评价标准 ^[34]	?	中	-	中	+	高	NR	NR	NR	NR	+	中	?	极低	B
气滞血瘀证评价量表 ^[35]	+	低	?	中	+	高	NR	NR	NR	NR	NR	NR	NR	NR	A
非酒精性脂肪性肝病中医证候量表 ^[36]	?	低	?	中	+	高	NR	NR	NR	NR	NR	NR	NR	NR	B
围绝经期综合征中医 PRO 量表 ^[37]	?	低	-	中	+	高	NR	NR	NR	NR	+	中	NR	NR	B
肝郁化热证评价量表 ^[38]	?	低	?	中	+	高	NR	NR	-	极低	?	低	NR	NR	B
冠心病痰热证中医四诊信息量表 ^[39]	?	低	-	极低	NR	NR	NR	NR	NR	NR	?	中	NR	NR	B
血虚证自评量表 ^[40]	?	低	-	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B

注：AECOPD，慢性阻塞性肺疾病急性加重期；PRO，患者报告结局；COPD，慢性阻塞性肺疾病；CRO，医生报告结局；测量学特性评价结果：+，充分；-，不充分；?，不确定；NR，未报告。

续表 1 中医证候疗效评价量表系统评价纳入文献的证据等级评价及推荐意见

Table 1 (continued)	Evidence level assessment and recommendation grades of included TCM syndrome efficacy evaluation scales															
评估工具	内容效度		结构效度		内部一致性		稳定性		效标效度		假设检验		反应度		推荐等级	
	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据	总体	证据		
	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量	评级	质量		
月虚证自评量表 ^[41]	?	低	-	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B	
日虚证自评量表 ^[42]	?	低	-	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B	
气虚证自评量表 ^[43]	?	低	?	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B	
气滞证自评量表 ^[44]	?	低	-	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B	
小儿厌食（喂养障碍）中医证候疗效评价量表 ^[45]	?	中	?	中	+	高	?	极低	NR	NR	+	中	?	极低	B	
毒品成瘾中医证候PRO量表 ^[46]	?	低	-	高	+	高	NR	NR	NR	NR	NR	NR	NR	NR	C	
血瘀证自评量表 ^[47]	?	低	-	中	+	高	NR	NR	NR	NR	+	低	NR	NR	B	
强直性脊柱炎肾虚督寒证中医疗效评价量表 ^[48]	+	低	+	中	+	高	+	极低	NR	NR	+	中	?	极低	A	
脾气虚PRO量表 ^[49]	?	低	-	中	+	高	+	低	NR	NR	?	中	?	极低	B	
冠心病稳定型心绞痛痰瘀互结证疗效评价计分表 ^[50]	?	中	?	中	+	高	+	极低	-	极低	NR	NR	NR	NR	B	
冠心病心绞痛中医证候疗效评价量表 ^[51]	?	中	-	中	+	高	NR	NR	NR	NR	+	中	?	极低	B	
冠心病稳定型心绞痛痰瘀互结证自评量表 ^[52]	?	低	?	低	+	高	+	极低	-	极低	+	中	NR	NR	B	
慢性乙型肝炎脾虚证候量表 ^[53]	?	低	-	中	+	高	+	低	NR	NR	NR	NR	?	极低	B	
经前期综合征肝气逆证疗效评价量表 ^[54]	NR	NR	-	高	-	高	-	低	NR	NR	NR	NR	NR	NR	C	
慢性乙型肝炎肝胆湿热证自评量表 ^[55]	NR	NR	?	中	+	高	NR	NR	NR	NR	+	中	?	极低	B	
高血压肝阳上亢证自评量表 ^[56]	NR	NR	?	极低	+	高	NR	NR	NR	NR	+	中	?	极低	B	

注：AECOPD，慢性阻塞性肺疾病急性加重期；PRO，患者报告结局；COPD，慢性阻塞性肺疾病；CRO，医生报告结局；测量学特性评价结果：+，充分；-，不充分；？，不确定；NR，未报告。

3 讨论

证候疗效是中医药临床疗效评估的核心依据^[58]。量表作为证候疗效评价的主要载体，良好的测量学属性是其临床应用的前提^[3-4]。本研究基于COSMIN指南对现有46个中医证候疗效评价量表进行了系统评价，发现现有量表在构建与应用方面已积累了一定基础，为证候疗效评价提供了重要支持，但整体上仍存在研究实施的关键细节报告不充分、部分测量学特性评估缺失及修订重视度不足等问题。本研究据此提出相关建议，以期为进一步完善证候疗效评价量表的研制与优化提供一定参考。

3.1 优化提升中医证候疗效评价量表的方法学质量

3.1.1 全面评估内容效度，系统整合专家与患者意见 内容效度作为工具评估中最关键的测量属性，直接反映工具内容与待测构念的匹配程度，其评估围绕相关性、全面性及可理解性3个维度展开^[5]。纳入研究中21项未评估相关性，反映出多数研究对内容效度的关键维度缺乏全面考量，未能充分论证条目与待测构念之间的映射关系，可能影响量表对证候变化的评估准确性。此外，内容效度

评估强调专家咨询法与患者访谈法的联合应用，此亦为当前证候疗效评价量表研制策略中的重要内容，前者通过多轮结构化咨询可优化量表内容覆盖度，后者借助深度访谈技术能有效提高条目的可理解性以控制应答偏差，二者协同可全面提升量表的科学性和实用性^[2, 5]。然而，多项研究未能整合这两类关键信息，可能导致量表缺乏足够中医理论支撑或与实际应用场景脱节。因此，为系统性提升量表的内容效度，建议后续研究参考COSMIN指南从框架构建与实施操作两个层面规范评估流程。在框架设计上应确立整合专家与患者双视角的基本原则，全面覆盖内容效度的3个核心维度。在实施过程中需着力提升信息采集质量：在满足研究范式所需样本量的基础上，专家咨询应广泛涵盖中医学、护理学及测量学等相关领域的专家，以提升专业意见的广度与深度；患者访谈可通过预先制定访谈大纲、对访谈者进行统一培训以规范访谈流程，从而保障数据真实可靠。同时，学术期刊可增设内容效度审核清单，鼓励研究者重视并完善相关内容。

3.1.2 重视CFA在结构效度检验中的规范应用

结构效度检验常采用因子分析法，其中EFA可

通过数据驱动为理论模型构建提供初步证据,但其因子结构易受样本波动影响,且无法验证预设理论模型的合理性^[59];而CFA则基于预设模型,通过拟合优度指数量化模型与观测数据的匹配程度,系统验证量表结构是否符合理论假设,在论证严谨性上更具优势^[60-61]。然而,当前研究对CFA的应用较为缺乏,这可能源于以下两方面原因:在理念认知层面,部分研究者尚未充分认识到CFA作为理论驱动型验证工具的核心价值;在方法实践层面,CFA对研究者的统计建模能力、数据质量及样本量的更高要求构成了一定的应用挑战。建议后续研究强化CFA的规范应用,从EFA主导的探索性研究转向EFA和CFA协同的验证性研究,以提升量表结构的科学性与认可度。

3.1.3 详细报告重测方法实施细节以充分评估稳定性 稳定性是指在相同条件下对同一对象重复测量所得结果的一致性^[62]。鉴于证候具有动态时空特征,易受时间、环境等因素影响发生演变^[63],其稳定性评估面临较大挑战。若未验证两次测量间隔期内证候保持相对稳定,则难以判断观测到的结果波动源于测量工具的误差还是证候自身的演变,进而可能引发对工具信度的误判。此外,重测间隔需权衡记忆效应与状态波动进行合理设定,间隔过短易受患者记忆效应干扰,过长则可能引入证候演变等混杂因素^[64]。因此,未来研究应注重详细报告证候状态的评估结果及重测间隔的设立依据,以保障稳定性评估结论的可靠性。

3.2 补充完善中医证候疗效评价量表的测量学特性

3.2.1 完整报告EFA中的因子载荷矩阵 因子载荷矩阵是呈现条目与因子关联强度的核心数据。完整报告该矩阵对结构效度评估具有重要意义。一方面可通过条目在目标因子上的载荷值判断条目是否聚焦于预设构念;另一方面可通过跨因子载荷(即条目在多个因子上的载荷 ≥ 0.30)识别量表潜在的维度混杂问题^[5]。然而,多数研究未完整报告此关键信息,这削弱了研究结果的透明度,阻碍了其他研究者对其进行分析、评估,甚至可能导致对量表结构合理性的误判。因此,今后研究应高度重视完整报告因子载荷矩阵,为结构效度评估提供必要的支撑,进而为结构优化提供依据。

3.2.2 补充跨文化效度验证与测量误差评估 通过跨文化效度检验可有效识别并调整量表条目语义、应答模式或结构维度的文化特异性偏差,以增强量表在不同文化背景下的适用性。证候具有复杂

的理论内涵与鲜明的文化特性^[63],其文化负载概念易在跨文化传播中产生释义偏差,故更需通过跨文化效度检验验证核心概念的跨文化等效性,为量表在多元文化语境下的临床适用性提供实证支持。此外,测量误差指测量结果与真实值之间的偏差,通过系统评估测量误差可有效识别并量化潜在误差来源,进而针对性优化量表条目,从而提升测量结果的可靠性^[65]。然而,可能因中医证候理论的独特性,加之研究周期与资源投入的客观制约,本研究纳入的量表均未开展跨文化效度与测量误差评估。鉴于中医证候研究日益国际化,建议未来研究开展多阶段跨文化效度验证流程,以确保条目在语义、概念及操作层面的跨文化等效性,同时通过测量标准误、最小可测变化值等指标识别并控制误差来源,以增强量表的适用性与可靠性。

3.3 推广A级量表临床应用并差异化完善B级、C级量表工具

推荐等级分布显示,2个量表获A级推荐,临床应用价值较高;38个量表为B级推荐,其与A级的主要差距在于未结合专家咨询与患者认知访谈证明其内容效度“充分”;6个量表被列为C级,主要因结构效度或内部一致性评估结果欠佳。此分布特征表明,当前中医证候疗效评价量表的整体质量有待提升,高质量工具较为缺乏,多数B级和C级量表因关键测量学特性验证不充分,不同程度地限制了其临床应用,由此可见规范化量表测量学特性验证流程具有必要性与紧迫性。因此,建议未来研究在适宜条件下积极推广A级量表的临床应用;对于B级量表,可参照COSMIN指南围绕其薄弱或缺失的测量学特性开展针对性再验证与补充验证;对于C级量表,建议在系统评估其理论基础与结构合理性的基础上,考虑进行理论框架重构、条目修订或方法学优化,以提升其测量学性能。

3.4 注重对量表的动态修订以持续提升其临床应用价值

测评工具的研制是一个循环反复的过程,修订是优化测评工具的必要途径^[66]。然而,本研究纳入的量表主要聚焦于单次开发的测量学特性验证阶段,尚缺乏系统性的后续修订研究。这一现状或与修订工作的高资源需求及中医结局测评工具标准化修订指南的缺失有关。量表的单次考评可为其临床应用提供初步的质量保障,但难以全面反映量表在不同临床场景、人群特征及时间维度下的测量误差,且当前证候疗效评价指标体系尚未统一,可能

随医学技术进步或循证医学证据更新而发生动态调整^[58]。因此,建议未来研究注重对量表进行多阶段临床验证,结合临床实践需求与循证医学进展进行内容优化,并通过跨中心协作验证修订版量表的测量学特性,透明化报告版本迭代内容以实现量表价值的最大化。

当前尚无证候疗效评价量表质量评估标准,本研究借鉴 COSMIN 指南进行质量评估可能存在如下局限性:首先,在样本量评估方面, COSMIN 指南采用了无弹性调整的统一标准,这可能导致针对临床少见、病例来源受限证候的研究,在方法学评价中需额外应对其固有局限。其次,指南推荐的重测间隔(2周)对于证候演变速度差异较大的病证可能缺乏普适性。最后,效标效度验证所依赖的“原版量表”金标准,与当前领域多数量表属首创开发、修订机制尚未完善的现状相矛盾,导致多数量表在此项被机械地评为“不良”,可能难以客观反映其效度验证的真实水平。因此,未来在研制证候疗效评价量表评估标准中,除借鉴国际标准外,应注重结合中医证候特点,以推动该领域科学化、规范化发展。

参考文献

- [1] 中华中医药学会. 2023 年度中医药重大科学问题、工程技术难题和产业技术问题[J]. 中医杂志, 2023, 64(14):1405-1421.
- [2] 李建生,冯贞贞,谢洋,等. 病证结合模式下中医证候疗效评价量表研制的实践与策略[J]. 中国全科医学, 2022, 25(20):2513-2519.
- [3] 王建新,任毅铭,丰雪,等. 中医证候疗效评价方法的研究进展[J]. 中国中药杂志, 2024, 49(6):1467-1473.
- [4] PRINSEN CAC, MOKKINK LB, BOUTER LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures [J]. Qual Life Res, 2018, 27(5): 1147-1157.
- [5] MOKKINK LB, ELSMAN EB, TERWEE CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0 [J]. Qual Life Res, 2024, 33(11):2929-2939.
- [6] 朱秀媛,罗宇茵,李婧菱,等. 脑卒中中医临床疗效评估工具测量学特性的系统评价[J]. 中国医药导报, 2025, 22(5):118-126.
- [7] 杨洋,孙亚男,于长禾. 冠心病中医临床疗效评价量表的测量学性能及研究方法学质量评价[J]. 中医杂志, 2021, 62(3):217-223.
- [8] ELSMAN EBM, MOKKINK LB, TERWEE CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024 [J]. J Clin Epidemiol, 2024, 173: 111422. doi: 10.1016/j.jclinepi. 2024. 111422.
- [9] ALHAZZANI W, GUYATT G. An overview of the GRADE approach and a peek at the future [J]. Med J Aust, 2018, 209(7):291-292.
- [10] 陈祎婷,沈蓝君,彭健,等. 改良版定量系统评价证据分级方法对患者报告结局测量工具的评价[J]. 解放军护理杂志, 2020, 37(10):57-60.
- [11] 朱昱林. 血热证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2024.
- [12] 任立松. 内风证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2024.
- [13] 白燕来. 外风证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2024.
- [14] 杨慧慧. 基于中医学特征和症状学特点的 GAD 心脾两虚证疗效评价量表的研究[D]. 哈尔滨:黑龙江省中医药科学院, 2023.
- [15] 腾颖. 实火证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2023.
- [16] 李敏,齐小荣,李明伟,等. 帕金森病肾虚血瘀证中医证候量表评价[J]. 中华中医药杂志, 2023, 38(1): 434-436.
- [17] 冯贞贞,何雯青,谢洋,等. 慢性阻塞性肺疾病急性加重期证候疗效评价量表初步研制[J]. 中国中西医结合杂志, 2023, 43(5):564-571.
- [18] 徐桂琴. 冠心病痰湿证 PRO 量表的编制与初步考评[D]. 南京:南京中医药大学, 2022.
- [19] 李建生,冯贞贞,谢洋. 基于临床调查的慢性阻塞性肺疾病稳定期证候疗效评价量表的初步形成[J]. 中医杂志, 2022, 63(13):1235-1242.
- [20] 刘竞男. 血虚证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2022.
- [21] 李苏凝. 老年冠心病(阳虚证)CRO 临床疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2022.
- [22] 李瑞本,马延宁,胡思源,等. 小儿积滞病中医证候疗效评价量表的编制和北方地区检验[J]. 天津中医药, 2022, 39(7):846-853.
- [23] 胡海殷,王虎城,王辉,等. 火热炽盛证(实热火毒)评价标准及其解读[J]. 中华中医药杂志, 2022, 37(6): 2998-3002.
- [24] 赵洪欣. 老年冠心病(阳虚证)患者报告结局(PRO)量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2021.
- [25] 崔博涵. 气滞证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学, 2021.
- [26] 章轶立. 慢性心力衰竭气虚证辨证与评价量表研制及其常模的初步构建[D]. 北京:北京中医药大学, 2021.

- [27]于莉,张会永,王佳楠,等. 基于患者报告结局的气虚证中医疗效评价量表考评[J]. 上海中医药杂志, 2020,54(9):60-64.
- [28]张宏. 冠心病心绞痛痰瘀互结病证结合疗效评价量表的考评[D]. 沈阳:辽宁中医药大学,2020.
- [29]于莉,杨可鑫,张会永,等. 血瘀证中医疗效评价量表的考评研究[J]. 时珍国医国药, 2019, 30(5): 1277-1280.
- [30]王姝琦. 阳虚证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学,2019.
- [31]邱怡. 基于气虚血瘀证评价量表临床验证的证候类中药疗效评价方法研究[D]. 北京:北京中医药大学, 2019.
- [32]李芹,于莉,张会永,等. 痰浊证中医疗效评价量表的考评[J]. 中华中医药杂志, 2019, 34(9): 3970-3976.
- [33]杨可鑫. 阴虚证中医疗效评价量表的研制与考评[D]. 沈阳:辽宁中医药大学,2018.
- [34]王阶,何庆勇. 冠心病心绞痛血瘀证疗效评价标准[J]. 中国实验方剂学杂志, 2018, 24(15): 1-3.
- [35]王阶,安宇,何庆勇,等. 基于患者报告结局的气滞血瘀证评价量表研制[J]. 中国实验方剂学杂志, 2018, 24(15): 21-28.
- [36]丁斯斯. 非酒精性脂肪性肝病中医证候量表的研制与考核[D]. 长沙:湖南中医药大学,2018.
- [37]李慧. 基于证素原理的围绝经期综合征中医 PRO 量表的研制和考评[D]. 福州:福建中医药大学,2017.
- [38]郭盛楠,刘强,王少卿,等. 肝郁化热证证候评价量表的研制[J]. 中国中医基础医学杂志, 2017, 23(3): 351-353.
- [39]赵利. 冠心病痰热证中医四诊信息量表研制与应用研究[D]. 北京:北京中医药大学,2016.
- [40]张昌秦. 血虚证自评量表的初步研制及考评[D]. 福州:福建中医药大学,2016.
- [41]熊丽枚. 阴虚证自评量表的初步研制及考评[D]. 福州:福建中医药大学,2016.
- [42]万信. 阳虚证自评量表的初步编制及考评[D]. 福州:福建中医药大学,2016.
- [43]陈玉招. 气虚证自评量表的初步研制及考评[D]. 福州:福建中医药大学,2016.
- [44]姚明龙. 气滞证自评量表的初步研制及考评[D]. 福州:福建中医药大学,2015.
- [45]孔秀路,徐田华,胡思源,等. 小儿厌食(喂养障碍)中医证候疗效评价量表的信度、效度和反应度测评[J]. 中医杂志, 2015, 56(9): 753-758.
- [46]居睿. 毒品成瘾中医证候 PRO 量表的条目筛选和考核[D]. 福州:福建中医药大学,2015.
- [47]陈湘筠. 血瘀证自评量表的初步研制及考评[D]. 福州:福建中医药大学,2015.
- [48]赵国青. 强直性脊柱炎肾虚督寒证中医疗效评价量表的建立与评价[D]. 广州:广州中医药大学,2013.
- [49]林建东. 脾气虚 PRO 量表的研制与考核[D]. 福州:福建中医药大学,2013.
- [50]王鸿琳. 冠心病稳定型心绞痛(痰瘀互结证)疗效评价证候计分表的研制与初步考评[D]. 沈阳:辽宁中医药大学,2012.
- [51]何庆勇. 冠心病心绞痛病证结合疗效评价体系研究[D]. 北京:北京中医药大学,2011.
- [52]刘丽星. 冠心病稳定型心绞痛(痰瘀互结证)自评量表研制与考评[D]. 沈阳:辽宁中医药大学,2011.
- [53]乐敏,黄杏,高月求. 慢性乙型肝炎脾虚证候量表研制初探[J]. 上海中医药杂志, 2008, 42(3): 6-9.
- [54]王海军. 经前期综合征肝气逆证辨证规范及疗效评价量表研究[D]. 济南:山东中医药大学,2007.
- [55]郭全,陈泽奇,刘小珍,等. 慢性乙型肝炎肝胆湿热证评定量表的初步编制及考评[J]. 中国中医药信息杂志, 2007, 14(7): 9-12.
- [56]郭全,陈泽奇,刘小珍,等. 原发性高血压肝阳上亢证评定量表的初步编制及考评[J]. 中国临床康复, 2006, 10(43): 20-23.
- [57]冯贞贞,李建生. 中医证候疗效评价量表研制现状[J]. 中医学报, 2022, 37(12): 2628-2635.
- [58]王建新,代欣玥,孙明月,等. 中医证候疗效评价指标体系的构建与思考[J]. 中国新药杂志, 2025, 34(4): 341-345.
- [59]SCHREIBER JB. Issues and recommendations for exploratory factor analysis and principal component analysis [J]. Res Social Adm Pharm, 2021, 17(5): 1004-1011.
- [60]ROGERS P. Best practices for your confirmatory factor analysis: a JASP and lavaan tutorial [J]. Behav Res Methods, 2024, 56(7): 6634-6654.
- [61]彭健,沈蓝君,陈祎婷,等. COSMIN-RoB 清单简介及测量工具内部结构研究的偏倚风险清单解读[J]. 中国循证医学杂志, 2020, 20(10): 1234-1240.
- [62]MOKKINK LB, TERWEE CB, PATRICK DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes [J]. J Clin Epidemiol, 2010, 63(7): 737-745.
- [63]田金洲,王永炎,时晶,等. 证候的概念及其属性[J]. 北京中医药大学学报, 2005, 28(5): 6-8.
- [64]李灿,辛玲. 调查问卷的信度与效度的评价方法研究[J]. 中国卫生统计, 2008, 25(5): 541-544.
- [65]彭健,沈蓝君,陈祎婷,等. 对 COSMIN-RoB 清单中测量工具稳定性、测量误差和效标效度研究偏倚风险的清单解读[J]. 中国循证医学杂志, 2020, 20(11): 1340-1344.

[66]王佳佳,谢洋,李建生. 临床结局测评工具修订的现状
及流程思考:以慢性阻塞性肺疾病患者报告结局量表

修订为例[J]. 中国全科医学, 2020, 23(29): 3644-
3648.

A Systematic Review of Traditional Chinese Medicine Syndrome Efficacy Evaluation Scales Based on the COSMIN Guidelines

SU Kailin^{1,2,3}, FENG Zhenzhen^{1,2,3}, WANG Jiajia^{1,2,3}, WANG Lu^{1,2,3}, ZHAO Guixiang^{1,2,3}, LI Jiansheng^{1,2,3}

1. The First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou, 450003; 2. The First Clinical Medical College of Henan University of Chinese Medicine; 3. Co-construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan and Education Ministry of P. R. China, Henan University of Chinese Medicine

ABSTRACT Objective To systematically evaluate the methodological quality and measurement properties of traditional Chinese medicine (TCM) syndrome efficacy evaluation scales, and to provide evidence-based references for selecting high-quality assessment tools in TCM clinical practice. **Methods** China National Knowledge Infrastructure (CNKI), Wanfang Data, VIP Database, Chinese Biomedical Literature Database (CBM), PubMed, the Cochrane Library, Embase, and Web of Science were searched from inception to April 2, 2025, for studies evaluating the measurement properties of TCM syndrome efficacy evaluation scales. Data were extracted, and the methodological quality and measurement properties of the included scales were assessed according to the consensus-based standards for the selection of health measurement instruments (COSMIN). Recommendation levels were formulated based on the grading of evidence. **Results** A total of 46 studies were included, involving 22 generic syndrome efficacy evaluation scales and 24 disease-specific syndrome efficacy evaluation scales. None of the scales reported cross-cultural validity or measurement error. According to the recommendation grades, 2 scales met Grade A recommendations and are suggested for clinical use; 38 scales were classified as Grade B, indicating potential applicability but requiring further validation; and 6 scales were classified as Grade C, suggesting the need for further refinement. **Conclusion** Existing TCM syndrome efficacy evaluation scales exhibit substantial variability in methodological quality, incomplete reporting of measurement properties, and insufficient attention to scale revision. Future efforts should emphasize standardized design in the development of TCM syndrome scales, strengthen validation procedures for key measurement properties, and prioritize dynamic revision of scales, thereby providing high-quality tools to support the precise evaluation of syndrome efficacy.

Keywords traditional Chinese medicine syndromes; efficacy evaluation; scales; measurement properties; consensus-based standards for the selection of health measurement instruments guidelines; systematic review

(收稿日期: 2025-07-29; 修回日期: 2025-12-17)

[编辑: 焦 爽]