



4.16 质量评价工具的示例（第4.5节的附录）



证据形式	质量评价工具的示例
现存质量评价工具的证据类型	
数据分析	<p>非随机干预性研究偏倚评估工具RoB2 (riskofbias.info) 适用于明确特定因素（包括干预措施）和特定结果相关性的观察性研究，存在以下偏倚风险来源：</p> <ul style="list-style-type: none"> • 混杂偏倚（由于一个或多个未被考虑的其他因素，观察到的因素和结果之间的关系与真实关系不同） • 选择受试者偏倚 • 干预措施分类 • 意向干预偏离偏倚 • 缺失数据偏倚 • 结果偏倚 • 选择性报告偏倚
评价	<p>偏倚风险评估工具 (RoB)2 (riskofbias.info) 适用于随机对照试验，混杂偏倚较少，但偏倚风险与上述某些（尽管较少）来源相同：</p> <ul style="list-style-type: none"> • 随机化过程中的偏倚 • 偏离既定干预措施的偏倚 • 结局数据缺失的偏倚 • 结果测量的偏倚 • 选择性报告偏倚
行为/实施研究	请参阅其他部分以了解相关类型的研究或综合
定性研究	<p>JBI定性研究的关键评价清单 (bit.ly/31Lsib1)，有许多非常不同的考虑因素，例如：</p> <ul style="list-style-type: none"> • 研究方法与研究问题、数据收集方法、数据呈现和分析、结果解释之间的一致性，以及所述哲学观点与研究方法之间的一致性 • 研究者的反思性，如从文化和理论上定位研究者的表述，并讨论研究者对研究的影响，反之亦然 • 研究参与者的代表性及其意见 • 从分析和数据解释中得出结论的流程
证据综合	<p>证据综合中考虑的相关研究类型见上文</p> <p>系统评价/Meta分析方法学质量评价工具 (AMSTAR; amstar.ca) 用于评价证据综合的质量，偏倚风险可能来自：</p> <ul style="list-style-type: none"> • 通过全面检索已发表的文献和灰色文献，确定所有潜在的相关研究，并且不对语言进行限制 • 使用有关研究设计和受试者、干预措施/因素、对照和结局的明确标准，筛选出所有解决研究问题的研究，并且至少有两名评价人员应用该标准 • 对所有纳入的研究进行质量评价和数据提取 • 对所有纳入的研究结果进行综合 <p>请注意，AMSTAR有两个版本：1) 原始版本，适用于所有类型的综合研究，尽管从领域和条目中删除了一些标准；2) AMSTAR的第二个版本，与随机对照试验的综合更为相关</p> <p>推荐分级的评估、制订与评价 (GRADE; bit.ly/3C9pMrx)，用于评价干预结局的证据质量，其中影响证据质量的因素有：</p> <ul style="list-style-type: none"> • 降级因素：偏倚风险（来自随机对照试验的证据起始质量高，来自观察性研究的证据初始质量低，随后根据RoB2或ROBINS-I进行升降级）、不精确性（例如，一两个小型研究）、不一致性（例如，两项研究显示结果差距很大）、间接性（例如，使用替代措施或研究的非典型环境）、以及发表偏倚（例如，由于缺乏研究注册平台，此偏倚在观察性研究中更常见；或由于发表阳性结果研究的商业动机，此在企业资助的研究中更常见） • 升级因素：大效应量、存在剂量-效应关系，以及负偏倚 <p>GRADE CERQual (cerqual.org) 用于对感兴趣的现象做定性描述的证据分级，其中：</p> <ul style="list-style-type: none"> • 由于对方法学局限性的担忧（因使用类似上述JBI的关键评价工具发现研究设计或报告方式上的问题）、相关性（因主要研究的环境与综合问题的环境有实质性的不同）、结果一致性（因一些数据与研究结果矛盾或含糊不清）和数据充分性（因数据不充分或仅来自少数研究或受试者），证据质量则被降级



<p>技术评估/成本效果分析</p>	<p>国际卫生技术评估机构协作网 (International Network of Agencies for Health Technology Assessment, INAHTA) 技术评估质量清单 (bit.ly/2YIVMVK), 14个条目中有两个条目涉及证据综合的方法 (类似于AMSTAR), 另一个条目涉及评估是否通过附带的成本效果分析 (与当地, 即国家或次国家的成本核算数据) 进行适用性分析, 以及考虑当地的法律、伦理和社会影响</p> <p>成本效果分析的德拉蒙德 (Drummond) 清单 (bit.ly/3FbnB8R), 以及更普遍的经济评价, 包括研究设计、数据收集以及结果分析和解释的问题</p> <p>飞利浦 (Philips) 成本效果分析清单包括决策分析模型部分 (bit.ly/3FcWBGc), 其中包含涉及模型的结构 (例如, 明确的理由、合理的假设和适当的时间范围)、采用的数据 (例如, 来自观察研究的基线概率、来自随机对照试验的治疗效果以及对四种不精确性的评估, 即模型的结构、遵循的方法步骤、所研究人群的异质性和使用的参数), 以及一致性 (内部和外部)。TRUST工具同时可以用来评价决策分析模型的不精确性 (bit.ly/3quFSKp)</p>
<p>指南</p>	<p>AGREE II工具 (bit.ly/30qvFAB) 用于评价指南的制订、报告和评价 (或质量评价), 该工具包含6个领域的23个条目, 每个领域独立评分:</p> <ul style="list-style-type: none"> • 描述范围和目的 • 利益相关者 (公民/患者和专业人士) 的参与 • 制订严谨性 (将证据综合作为一种证据, 使用严格的推荐意见形成过程, 并将推荐意见与支持性证据联系起来) • 表达清晰性 • 适用性 • 编辑独立性 (与资助者和小组成员的利益冲突有关) <p>GRADE (bit.ly/3C9pMrx) 用于评估推荐意见的强度, 它有四个关键考虑因素:</p> <ul style="list-style-type: none"> • 平衡 (权衡) 理想和不理想的结果, 考虑到对理想和不理想结果效应量的最佳估计, 以及对这些结果的重要性 (评估偏好和价值观) • 干预措施对重要结局效应量的把握度 (见前一行) • 使用偏好和价值观及可变性资源的把握度
<p>暂无质量评价工具的证据类型</p>	
<p>建模</p>	<p>对于大多数类型的模型, 目前还没有广泛接受使用的工具, 然而, 存在一些关于模型的共性问题 (就像上述飞利浦清单中部分列出的问题), 例如:</p> <ul style="list-style-type: none"> • 模型的结构 (例如, 明确的理由、合理的假设和适当的时间范围) • 使用的数据 (例如, 来自观察性研究的基线概率, 来自一系列来源的干预效果*, 以及对四种不可靠性的评估, 即模型的结构、遵循的方法步骤、研究人群的异质性以及使用的参数) • 一致性 (内部和外部) • 软件或工具的可用性, 便于其他人可以对其进行评价 <p>*COVID-19面临的挑战之一是, 通常用于获取干预效果的研究设计, 如随机对照试验, 在伦理上或后勤上存在困难和 (或) 需要时间来完成, 因此需要使用其他研究设计, 并需要征求专家意见 (并且有一些系统和透明的方式进行, 如SHELF-见 bit.ly/30nteC4)</p>
<p>某些暂无质量评价工具的证据类型所使用的方法</p>	
<p>人工智能</p>	<p>暂无广泛认同的工具</p>