

## 4.5 高质量证据和低质量证据的区分

不是所有证据都是高质量和可靠的决策依据。许多（但不是所有）形式的证据都有相应的工具，可用于帮助判断证据（单个研究或证据体）的可靠性。正如我们所述，这些工具通过评分或评级的方式帮助使用者了解对证据的确信程度。现在许多期刊要求作者遵循报告规范，如适用于随机对照试验的CONSORT和适用于证据综合的PRISMA。大部分期刊并未要求审稿人使用特定工具评价研究质量或推荐强度；因此发表在同行评审期刊上的并不一定都是高质量文章。

问题	回复
研究（和指南）的质量（或可信度）各不相同	<ul style="list-style-type: none"> <li>已针对特定研究设计（如随机对照试验）、研究设计大类（如观察性研究、定性研究和证据综合）和指南开发了质量评价（或严格评价）工具——实例（RoB2、ROBINS-I、JBI清单、AMSTAR和AGREE II）见本章末尾附录（<a href="#">第4.16节</a>）</li> <li>工具可以产生的结果包括总结性判断（例如，RoB2或ROBINS-I中的低偏倚风险），根据得分进行范围分组（例如，AMSTAR中的高质量），一组得分（例如，AGREE II的六个领域），或一组可以为总结性判断提供信息的考虑因素（例如，JBI清单）</li> </ul>
证据体的可靠度（或对证据体的确信程度）各不相同	<ul style="list-style-type: none"> <li>为解决同一问题（例如，干预措施对特定结局的影响或公民依附于特定现象的意义）的证据体开发了可靠性评价工具——见<a href="#">第4.16节</a>的两个实例（GRADE和GRADE CERQual）</li> <li>工具可以得到一个关于可信度的总体判断，包括对真实效果与估计效果的相似程度（例如，GRADE中的高质量证据），或定性研究结果对研究内容代表程度的可信度（利用GRADE CERQual）</li> <li>对效果估计的可靠性进行总结性判断比证明干预措施“有效”或“无效”的统计学意义检验更有帮助（如果显著性水平为0.05，则20次试验中会有一次偶然情况发生）</li> </ul>
推荐强度各不相同	<ul style="list-style-type: none"> <li>已为指南推荐意见开发了推荐强度分级工具（例如，GRADE，除上述对证据体级别进行分级外）——见<a href="#">第4.16节</a>的示例</li> <li>工具可以产生一个总结性判断，即大部分决策者是否会选择继续进行干预（例如，GRADE中的强推荐）或者大多数人是否需要仔细权衡干预的利弊</li> </ul>
证据的来源（或用于产生证据的方法）可能很难判断	<ul style="list-style-type: none"> <li>目前还没有广泛接受的工具用以评估以下内容的可信度： <ul style="list-style-type: none"> <li>专家，尽管在预测方面确实存在像“良好判断项目”这样的例子（我们将在本章后文中讨论专家意见；关于模型参数的专家意见，则在<a href="#">第4.16节</a>讨论）</li> <li>用于生成某些形式证据的模型（我们在<a href="#">第4.7节</a>讨论气候变化模型时，同时在<a href="#">第4.16节</a>讨论这个问题）</li> <li>用于生成某些类型证据的人工智能算法，尽管已经出现了像TRIPOD这样的例子(3)</li> </ul> </li> </ul>

当证据被整合到数据平台、模型和其他格式中时，以及利益冲突起作用时，区分高质量和低质量的证据就格外具有挑战性。我们将在后面的[第4.12](#)、[4.14](#)和[4.16](#)讨论。虽然这不是本报告的重点，但区分高质量和低质量的“原始数据”也是一种挑战，比如UNICEP已经开发了数据质量框架来帮助解决这个问题（[bit.ly/3DQRRv](http://bit.ly/3DQRRv)）。

一些“一站式商店”，如社会系统证据（SSE）和COVID-19知证决策协作网（COVID-19 Evidence Network to support Decision-making, COVID-END）清单（在**第4.6节**中描述），使用了其中的一些工具，以便于决策者及其支持者可以关注高质量的证据综合，或者了解他们正在使用的是当前可用（如果不是高质量的）的最佳证据综合。

COVID-19大流行要求决策者在短时间内做出艰难的决策，最初的证据很少，且通常是间接证据。后来随着时间的推移，研究、证据体和推荐意见均通过一个稳健的过程生产或制订。为支持基于证据体（而非单个研究）的COVID-19决策，COVID-END在其“最佳”证据综合清单中列出了那些最新的（基于检索证据的日期）、高质量（基于AMSTAR工具）证据，并提供证据的可靠性评估结果（基于GRADE工具）。

并非所有证据均是高质量的，也并非所有全球证据均适用于特定情境。例如，证据综合只纳入在高收入国家开展的研究时，其可能在一些低收入国家中适用性有限。在基线条件、实际情况和制约因素以及本地系统的结构特征（例如，国家卫生系统或省/州教育系统）方面可能存在重大差异。SUPPORT工具可以同样帮助人们思考证据综合结果的本土适用性，并考虑即使发现不适用时，如何仍然得出见解。(4)

贝叶斯推理正在获得越来越多的关注，它可以重新绘制我们对于挑战和应对挑战的“心理地图”，不是用新的信息取代我们认为的所有知识，而是将我们的理解调整到一个适当的程度。这个程度取决于你对现有知识（某事为真的“先验”概率）的确信程度，以及你对新知识的确信程度。如果新知识来自于高质量的证据综合，其中包括了与你所处环境相似情境下开展的研究，则可以对新知识有更高的确信程度。



### **证据中介和生产者, Gillian Leng**

是一位经验丰富的主管，领导技术评估和指南机构，支持政府、服务提供商和患者的健康和社会照护决策

英国多年来一直致力于鼓励证据综合和使用的有关工作，从第一个预防海员坏血病的随机对照试验，到最近新提出的促进证据在一系列政策领域使用的“**What Works中心**”。作为这次循证运动的一部分，在过去20年间，英国国家卫生与临床优化研究所（NICE）改变了证据在医疗实践以及更广泛的公共卫生举措和社会照护中的使用。

COVID-19大流行极大地增强了需要高质量证据为政策和实践提供信息的必要性，还突出了社交媒体和相关错误信息的负面后果。在这种情况下，全球证据委员会应对社会挑战的工作非常重要，应该被视为世界各地所有政策制定者的必备读物。

