·系列讲座·

偏倚风险评估系列:(六)诊断试验

曲艳吉　杨智荣　孙凤　詹思延
510080 广州,广东省心血管病研究所,广东省人民医院,广东省医学科学院心外科(曲艳吉);CB1 8RN 英国剑桥大学临床医学院初级医疗中心(杨智荣);100191 北京大学公共卫生学院流行病与卫生统计学系,北京大学循证医学中心(孙凤、詹思延)
通信作者:孙凤, Email:sunfeng@bjmu.edu.cn
DOI:10.3760/cma.j.issn.0254-6450.2018.04.028

【摘要】 本讲座详细介绍了诊断试验准确性研究的偏倚评估工具(QUADAS-2)的主要内容,同时阐述了QUADAS-2的开发过程及与第一版QUADAS工具的区别,并举例说明QUADAS-2的使用方法和注意事项。QUADAS-2相比第一版QUADAS工具有巨大改进,例如QUADAS-2删除了QUADAS中易混淆的条目内容,仅通过对重叠度最小的4个关键领域(Domain)的描述和对每个领域内信号问题的回答完成偏倚风险和适用性两个核心方面的评价,最后得出原始研究每个领域的偏倚风险和适用性为高(High)、低(Low)或不清楚(Unclear)的结论,而不再是给出原始研究质量评价的总分,这与Cochrane干预措施系统综述中偏倚风险的评估一致。同时,QUADAS-2还可以应用于金标准中包括随访但不涉及预后问题的原始研究的偏倚风险评价。虽然开展QUADAS-2评价需要花费更多时间,但这对于诊断准确性研究的偏倚风险评价非常重要。QUADAS-2研究组后续还将在比较多种待评价诊断试验的原始研究中应用QUADAS-2进行偏倚风险评估,使用者可持续关注其进展,同时也可在线反馈使用体验或提供改进建议。

【关键词】 偏倚风险;评估工具;诊断试验;系统综述
基金项目:国家自然科学基金(71673003)

Risk on bias assessment: (6) A Revised Tool for the Quality Assessment on Diagnostic Accuracy Studies (QUADAS-2)　*Qu Yanji, Yang Zhirong, Sun Feng, Zhan Siyan*

Department of Cardiac Surgery, Guangdong Cardiovascular Institute, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China (Qu YJ); Primary Care Unit, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridgeshire CB1 8RN, UK (Yang ZR); Department of Epidemiology and Biostatistics, School of Public Health, Center of Evidence-based Medicine and Clinical Research, Peking University, Beijing 100191, China (Sun F, Zhan SY)
Corresponding author: Sun Feng, Email: sunfeng@bjmu.edu.cn

【Abstract】 This paper introduced the Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), including the development and comparison with the original QUADAS, and illustrated the application of QUADAS-2 in a published paper related to the study on diagnostic accuracy which was included in systematic review and Meta-analysis. QUADAS-2 presented considerable improvement over the original tool. Confused items that included in QUADAS had disappeared and the quality assessment of the original study replaced by the rating of risk on bias and applicability. This was implemented through the description on the four main domains with minimal overlapping and answering the signal questions in each domain. The risk of bias and applicability with 'high', 'low' or 'unclear' was in line with the risk of bias assessment of intervention studies in Cochrane, so to replace the total score of quality assessment in QUADAS. Meanwhile, QUADAS-2 was also applicable to assess the diagnostic accuracy studies in which follow-up without prognosis was involved in golden standard. It was useful to assess the overall methodological quality of the study despite more time consuming than the original QUADAS. However, QUADAS-2 needs to be modified to apply in comparative studies on diagnostic accuracy and we hope the users would follow the updates and give their feedbacks on line.

【Key words】 Risk of bias; Tool for assessment; Diagnostic accuracy studies; Systematic review
Fund program: National Natural Science Foundation of China (71673003)

诊断试验是对疾病进行诊断的试验方法，包括各种实验室检查、影像学诊断等。诊断试验的应用非常广泛，不仅用于选择可靠的诊断方法、正确判读诊断结果，还可用于预测预后疗效等。诊断试验结果的可靠性取决于其方法学质量，所以评价诊断试验的准确性是应用其结果的前提。同时，评价纳入研究的方法学质量也是开展系统综述至关重要的一步。梳理既往发表的诊断准确性研究的系统综述，可发现曾使用过一系列不同的质量评价工具和检查清单对诊断准确性原始研究进行质量评价[1]。其中，QUADAS（Quality Assessment of Diagnostic Accuracy Studies）是唯一经过验证的、最有效的质量评价工具。原版QUADAS于2003年问世，后于2011年进行升级改版，就是本文介绍的QUADAS-2[2-3]。

诊断准确性系统综述纳入的原始研究属于观察性研究，由于设计和实施不同，通常都存在显著异质性，故严格评价所有纳入研究的偏倚风险是非常必要的。同时，熟悉偏倚风险评价规则也有利于在开展诊断试验准确性原始研究时从源头上降低偏倚风险，提高研究质量。

一、QUADAS-2的制定背景

QUADAS自发布以来，得到了广泛应用。在Database of Abstracts of Reviews of Effects中，超过200个系统综述的摘要中提到了QUADAS，其被引用次数超过500次。Cochrane协作组（Agency for Healthcare Research and Quality, Cochrane Collaboration）和英国国家卫生与临床优化研究所（U. K. National Institute for Health and Clinical Excellence, NICE）都推荐在诊断试验准确性研究的系统综述中使用QUADAS开展质量评价。但是，与此同时，使用者和Cochrane协作组也对进一步改进QUADAS提出了建议。使用者反馈的问题主要包括：条目13“是否报告了难以解释/中间试验结果”中的“难以解释/中间试验结果”不明确，导致条目评级不明；条目5部分参照偏倚和条目14退出病例之间可能存在重复；以及难以在金标准中包括随访的研究中使用QUADAS开展评价。于是，在第一版QUADAS工具的使用经验和关于诊断准确性研究中偏倚来源和变异的新证据的基础上，研究组对QUADAS工具重新进行优化和设计，从而诞生了QUADAS-2。

二、工具解读

1. QUADAS-2与QUADAS的区别与联系：第一版QUADAS工具包括14个条目，评估偏倚的风险、变异来源（适用性）和报告的质量。每个条目都评为“是（Yes）”、“否（No）”或“不清楚（Unclear）”，最后根据每个条目的评分得到一个质量评价的总分。QUADAS-2在QUADAS基础上删除了易混淆的条目内容，通过对4个关键领域（Domain）的描述和对每个领域内信号问题的回答完成偏倚风险和适用性两个核心方面的评价，最后得出原始研究每个领域的偏倚风险和适用性为“高（High）”、“低（Low）”或“不清楚（Unclear）”的结论。总之，QUADAS-2对比QUADAS的主要区别包括：①将原始研究的“质量”评价分成“偏倚的风险（诊断准确性的估计避免偏倚风险的程度）”和“适用性（原始研究对综述问题的适用程度）”的评价；②将评价内容限制为重叠度最小的4个关键领域，包括病例选择（patient selection）、待评价的诊断试验（index test）、金标准（reference standard）、病例流程和诊断试验与金标准的时间间隔（flow and timing），其中全部4个领域均用于评价偏倚风险。前3个领域还设置了适用性评价的条目，供评估者选择是否进行评价；③QUADAS-2在偏倚风险判断中增加了信号问题，旨在帮助评价者进行判断；④将QUADAS-2应用于评价比较多种诊断试验准确性的研究和金标准中包括随访、但不涉及预后问题的研究；⑤QUADAS-2将QUADAS工具中所使用的条目评级“是”、“否”、“不清楚”变更为偏倚风险为“低”、“高”和“不清楚”，这与Cochrane干预措施系统综述中对偏倚风险的评估一致。QUADAS-2与QUADAS的条目对比见表1。

2. QUADAS-2工具介绍：完整版的QUADAS-2工具可在QUADAS网站（www.quadas.org）上获得。同时，该网站还提供了相关培训信息、其他附加信号问题库、每个领域详细的评价指引、完成QUADAS-2评价的举例、提取数据用的Access数据库、产生结果图的Excel表格和总结结果的Word表格模板等。使用QUADAS-2工具进行诊断试验准确性研究的质量评价时，评价者都应该登录QUADAS网站了解QUADAS-2的详细信息，并选择性下载和使用其在线资源。QUADAS- 2工具用于评价诊断准确性原始研究的方法学质量时，其不能够替代系统综述中数据提取的过程，而应是在原始数据提取后使用。QUADAS-2评价信息的提取可登录QUADAS网站下载Access数据库（quadas2-database）模板，以表单形式逐步录入完成。QUADAS-2中4个关键领域及其偏倚风险和适用性评价介绍如下。

（1）领域1：病例选择：

①偏倚风险：病例的选择是否会带来偏倚？

表1　QUADAS-2的信号问题和QUADAS的条目对比

领域	QUADAS-2信号问题	QUADAS条目	解读
病例选择			
	1. 是否纳入了连续或随机病例? 2. 是否避免采用病例-对照设计? 3. 研究是否避免了不恰当的排除?	条目1　病例谱是否包含了各种极易混淆的疾病病例? 条目2　研究对象的选择标准是否明确?	QUADAS-2相对于QUADAS增加了研究设计的问题(信号问题2)
待评价的诊断试验			
	1. 是否在不知金标准结果的情况下解释诊断试验的结果? 2. 如果使用阈值,这个值是否预先设定?	条目8　待评价试验的操作是否描述的足够清楚且可进行重复? 条目10　待评价试验的结果判读是否是在不知晓金标准试验结果的情况下进行的? 条目12　当解释试验结果时可获得的临床资料是否与实际应用中可获得的临床资料一致?	QUADAS-2删除了QUADAS中待评价试验实施(条目8)和临床解读偏倚的问题(条目12);将试验实施和解释的问题在适用性评价中考虑;增加了关于阈值是否预先设定的问题(信号问题2)
金标准			
	1. 金标准是否能准确地区分目标疾病? 2. 是否在不知诊断试验结果的情况下解释金标准的结果?	条目3　金标准是否能准确区分有病、无病状态? 条目7　金标准试验是否独立于待评价试验(即待评价试验不包含在金标准中)? 条目9　金标准试验的操作是否描述的足够清楚且可以进行重复? 条目11　金标准试验的结果判读是否是在不知晓待评价试验结果的情况下进行的?	QUADAS-2删除了QUADAS中混合偏倚的问题(条目7)和金标准实施的问题(条目9);将金标准实施的问题在适用性中予以考虑
病例流程和诊断试验与金标准的时间间隔			
	1. 诊断试验和金标准之间是否有适当的时间间隔? 2. 是否所有患者都接受了金标准试验? 3. 是否所有患者都接受了同样的金标准试验? 4. 是否所有患者都纳入了分析?	条目4　金标准和待评价试验检测的间隔时间是否足够短,以避免出现疾病病情的变化? 条目5　是否所有的样本或随机选择的样本均接受了金标准试验? 条目6　是否所有病例无论待评价试验的结果如何,都接受了相同的金标准试验? 条目13　是否报告了难以解释/中间试验结果? 条目14　对退出研究的病例是否进行解释?	QUADAS-2删除了QUADAS中难以解释的试验结果的问题(条目13);将QUADAS中条目4"金标准和待评价试验检测的间隔时间是否足够短"修改为"诊断试验和金标准之间是否有适当的时间间隔",以便应用于金标准中包括随访的研究评价;将条目14"对退出研究的病例是否进行解释"修改为"是否所有患者都纳入了分析",避免重复

信号问题1:是否纳入了连续或随机病例?理想的研究应该纳入符合条件的所有可疑患者的一个连续或随机样本,避免潜在偏倚。如果原始研究纳入的是在一定时间范畴内的连续病例,则信号问题1就评为"是";如果是非连续或非随机病例,例如,专门选择重症的病例或按就诊时间选取病例,即评为"否";若研究虽然报告了纳入的时间范畴,但没有说明是否为连续或随机病例,则评为"不清楚"。

信号问题2:是否避免了病例-对照设计?事实上,病例-对照设计可能影响的不只是病例选择这一关键领域。在病例对照设计中,诊断试验组入选的参加者均为患者,而对照组的参加者均为非病例,结果很可能会夸大诊断的准确性(领域1:病例选择);由于事先明确了诊断的分组,待评价试验很大可能是在知晓临床诊断的情况下开展的,因此会影响到待评价试验结果的判读(领域2:待评价的诊断试验)。相反,在非病例-对照设计的研究中,尤其是在横断面研究中,疾病的临床诊断往往与待评价试验独立进行,即"金标准"的实施是在不知道待评价试验结果的情况下开展的,带来的偏倚风险较低(领域3:金标准)。具体评价时,如果研究采用非病例对照设计,就评为"是";若采用病例对照设计,即其中一组为病例,而另一组为疑似病例,那么就应评为"否";若提供的资料不足以判断,则判定为"不清楚"。

信号问题3:研究是否避免了不恰当的排除?如果研究存在不合理的排除标准,如排除了难以诊断(difficult-to-diagnose)的患者,结果可能会高估诊断的准确性。例如,在本文介绍的高敏感度肌钙蛋白hs-cTnT(high sensitivity troponin)检测排除急性心肌梗死的准确性研究的系统综述中,连续纳入心电图ST段抬高者的研究得到的敏感度很可能高于病例中包括了可疑但ST段未抬高者(即难以诊断患者)的研究结果。同样地,纳入已知疾病的患者和没有患病的对照组也会夸大诊断的准确性。相反,如果研究排除了更容易被诊断者,则可能会低估诊断

的准确性。在评价过程中，如果研究特别排除了难以诊断或易于诊断的病例，评价为“否”；否则评价为“是”；如果没有报告排除标准，则可认为是“不清楚”。

②适用性：原始研究中纳入的病例特征与系统综述中不符？

如果原始研究中纳入的病例与系统综述中的目标人群在某些方面，如所患疾病的严重程度、人口学特征、诊断或并发症、研究背景和先前的检查方法等存在差异，就要考虑适用性的问题。例如，大面积的心肌梗死相比小面积的心肌梗死在急性期通常会导致更高的心肌酶和肌钙蛋白水平，从而提高灵敏度的估计。

（2）领域2：待评价的诊断试验：

①偏倚风险：待评价试验的实施或解释是否会产生偏倚？

信号问题1：待评价试验的结果判读是否是在不知晓金标准试验结果的情况下进行的？这个条目与干预研究中的盲法类似。因为金标准的结果信息可能会影响对待评价试验结果的解释，所以评价者在判定待评价试验结果时应该不知道“金标准”的结果。如果待评价试验始终是在金标准制定之前实施和解释或文章中明确说明了评价者是在不知道“金标准”结果的情况下判读待评价试验的结果，那么该条目就可以评定为“是”，相反则为“否”；如果没有介绍就评为“不清楚”。

信号问题2：如果使用了阈值，那么这个阈值是否预先设定？如果诊断试验准确性的原始研究中所选择的阈值是根据灵敏度和/或特异度而选择的最优结果，那么很可能会高估准确性。评价中，如果研究所使用的阈值是在研究实施前就确定的，判定为“是”，相反即为“否”，信息不足以判断则为“不清楚”。

②适用性：原始研究中诊断试验的实施或解释是否与系统综述中不同？

诊断试验的技术、实施和解释都可能会影响对其准确性的估计。所以如果诊断试验的方法与系统综述中说明的不同，那么就要考虑适用性的问题。例如，在本文介绍的高敏感度肌钙蛋白hs-cTnT排除急性心肌梗死的准确性研究的系统综述中，不同肌钙蛋白水平的界值都会影响排除急性心肌梗死患者的准确性，较低水平的界值，如3 ng/L或5 ng/L相比14 ng/L有更高的准确性。

（3）领域3：金标准：

①偏倚风险：金标准的实施及解释是否会产生偏倚？

信号问题1：金标准是否可以准确地区分目标疾病？选择恰当的金标准在诊断准确性研究中至关重要。因为对诊断试验进行准确性评估就是建立在金标准的灵敏度和特异度为100%，待评价试验与金标准的结果差异是由于待评价试验对目标疾病的不正确分类而产生的前提假设基础上的。评价中，如果研究所用金标准可以正确区分目标疾病或已是现有疾病诊断的最佳方法，判定为“是”，否则即为“否”，判断依据不足就评为“不清楚”。

信号问题2：金标准的解释是否在对待评价诊断试验结果不知情的情况下做出的？该条目与待评价诊断试验的信号问题类似，即金标准结果判读是否使用了盲法。如果预先知道待评价诊断试验的结果可能会影响对金标准试验结果的解释，从而带来潜在偏倚。金标准结果的判读是在不了解待评价试验结果时进行的即评价为“是”，相反为“否”，难以判断则为“不清楚”。

②适用性：金标准所定义的目标疾病是否与系统综述中不符？

这里需要考虑两点：一是原始研究中所用的金标准与系统综述中所定义的纳入研究的金标准是否相同；二是即使采用相同的金标准，其对目标疾病的定义是否相同，即是否采用相同的阈值判断患者与非患者。例如，本文所举示例中，系统综述纳入了使用标准肌钙蛋白Ⅰ检测、标准肌钙蛋白T检测或其与高敏感度肌钙蛋白联合检测作为金标准的原始研究，该综述在结果分析时考虑到了不同金标准对准确性估计的影响，并通过回归分析确定这种影响是否有统计学意义。

（4）领域4：病例流程和诊断试验与金标准之间的时间间隔：

①偏倚风险：病例的流程是否会产生偏倚？

信号问题1：待评价试验和金标准之间是否有恰当的时间间隔？最理想的状态是同时收集同一患者的诊断试验和金标准试验的结果。因为诊断试验或金标准试验延迟，或在诊断试验和金标准试验间隔期间开始治疗，疾病的康复或恶化都可能会造成结果的错分。导致高风险偏倚的时间间隔因疾病的状态不同而异，对于慢性病，短期间隔可能不会有问题，但对于急性感染性疾病而言，就可能有问题。如果金标准是要通过随访获得结果的，则可能需要一个最短随访期来评估是否发生了目标疾病。例如，评估利用磁共振成像（MRI）早期诊断多发性硬化，至少需要约10年的随访期确定所有符合诊断标准

的患者都得以诊断。因此，这个问题的判定取决于目标疾病，结合疾病种类对“恰当的时间间隔”进行判定是关键，而且这个时间间隔应在正式使用QUADAS-2前确定。若待评价试验和金标准实施的间隔在所规定的范围内，则评价为“是”；若超出了该时间间隔，那么就评价为“否”；信息不足则为“不清楚”。

信号问题2：是否所有患者都接受了金标准试验，且是相同的金标准试验？虽然在QUADAS-2的原始表格中这个问题是分开两个条目的，即“是否所有患者都接受了金标准试验”和“是否所有患者都接受了相同的金标准试验”，但在QUADAS-2配套的解读文件中，对这两个条目是一起进行解释说明的。如果研究中只有一部分研究对象接受了金标准，或者有一些患者接受了不同的金标准，那么就可能会发生验证偏倚（Verification bias）。如果待评价诊断试验的结果影响了是否执行金标准或使用哪个金标准的判断，那么诊断准确性的评估也可能发生偏倚。例如，在高敏感度肌钙蛋白检测排除急性心肌梗死的例子中，如果对阳性者进一步行标准的肌钙蛋白检测和心电图检查（金标准1），而对阴性排除者通过临床随访确定是否发生心肌梗死（金标准2），则可能会将诊断试验的假阴性结果错分为真阴性，因为临床随访可能会漏掉那些诊断试验结果为阴性的急性心肌梗死患者，从而高估高肌钙蛋白检测排除急性心肌梗死的准确性。实际评估中，若可以清晰地判断研究中所有病例均接受了同一个金标准验证其疾病状态，那么就判定为“是”，相反则为“否”，若研究未报告该信息则评价为“不清楚”。

信号问题3：是否所有患者均纳入分析？研究招募的所有参加者都应该纳入分析。因为失访者与随访到的患者之间会存在系统性的差异，所以如果纳入研究的患者数与结果中2×2表里的患者数不同，会存在潜在偏倚。评价中，如果所有病例都纳入研究即评价为“是”；如果在结果分析时有病例遗漏则为“否”；未说明或无法判断则为“不确定”。

3. 应用QUADAS-2开展诊断准确性研究质量评价的4个阶段：

（1）第1阶段：提出研究问题：评估者首先要提出所要研究的问题，明确纳入的病例（P）、待评价的诊断试验（I）、金标准（R）、结局疾病（O）、纳入研究类型（S）等，类似于PICOS的原则。因为诊断试验的准确性会受到其在诊断路径上的应用位置的影响，故评估者应注意待评价试验与当前实际应用的试验之间的关系[分流（triage）、附加（add-on）、替代（replace）]，此外还要描述病例所处的环境、患者的临床表现和以往接受的诊断方法及结果等。

（2）第2阶段：根据需要建立综述专用的QUADAS-2及评价指南：QUADAS-2工具要求综述者根据自己的研究“剪裁”工具，即对工具进行调整，包括增加或删减信号问题和建立综述专用的评价指南指导如何评估每个信号问题和使用这些信息判断总的偏倚风险。图1为“剪裁”系统综述专用QUADAS-2工具的4个步骤。步骤1：根据实际情况，评估者若发现QUADAS-2有信号问题不适用于所开展的系统综述，可对这些问题进行删减。若信号问题不足以覆盖综述所研究的问题，评估者应该适当增加，但要避免增加过多的信号问题使工具过于复杂。无论进行上述何种调整，都应该明确报告理由。步骤2：评估者对所调整的内容达成共识，然后建立评估专用的偏倚风险评价指南。步骤3：由至少两位研究者独立使用调整好的QUADAS-2工具预评价少数研究。如果所得一致性好，则此工具可被用于评价所纳入的全部研究；相反，如果一致性不好，就需要进一步修改工具，重复步骤1～3，直至得到一致性较好的QUADAS-2工具，用于所有纳入原始研究的评价，即完成步骤4。

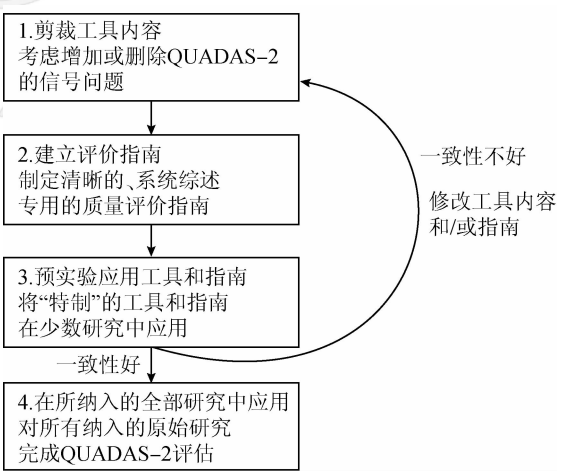

图1　“裁剪”系统综述专用的QUADAS-2工具的流程

（3）第3阶段：流程图：评估者要准备每个纳入原始研究的流程图便于评价中判断偏倚的风险。如果原始研究已发表详细的流程图，可以直接使用；如果原始研究中没有报告或已发表的流程图不够，则综述者可自行绘制流程图。需要说明的是，因为流程图不需要作为QUADAS-2评估的一部分报告，所以手绘流程图就可以了。

（4）第4阶段：判断偏倚风险和适用性：偏倚风险：QUADAS-2工具的4个关键领域都要进行偏倚风险的评估。偏倚风险评估包括3个部分：支持偏倚风险判断的信息、信号问题和偏倚风险的判断。记录支持判断偏倚风险的信息，目的是使评价透明化并方便独立的评价者之间进行讨论。信号问题用于辅助偏倚风险的判断，每个问题需要回答“是”、“否”或“不清楚”。偏倚风险的判断以信号问题的回答为依据，分为“低”、“高”或“不清楚”。此时，评估者需使用在第2阶段产生的评价标准判断偏倚风险。例如，如果所有信号问题的答案都是“是”，表示低偏倚风险；如果有某个信号问题的答案是“否”，则表示存在偏倚风险；其他情况则判断为“不清楚”。评估者可合理制定自己的评价标准，但要注意的是，该标准应在正式评估前（第2阶段）制定，任何根据信号问题的回答而对评价标准进行调整都是不合适的。适用性：判断适用性部分的组织设计与偏倚风险评估部分相似，但不设置信号问题。评价者记录用于判断适用性的信息，然后对纳入研究与系统综述研究问题的匹配程度进行评级。根据匹配程度不同，适用性对应的评级分为“低”、“高”或“不清楚”。需要特别强调的，适用性是要跟系统综述的研究问题结合起来评价的，不属于偏倚风险的内容，可由评价者选择评价或者不评价，但应在正式评估前确定，避免选择性报告结果。适用性评价应该参照第1阶段的内容，因为综述的研究问题是在第1阶段记录的。同样，“不清楚”选项只能在研究报告的数据不足以进行质量评价的情况下选用。

4. 使用QUADAS-2工具进行质量评估的结果报告：系统综述应该总结QUADAS-2评价所有纳入研究的结果，使用QUADAS-2进行偏倚风险评价不需要计算一个总结性的“质量分数”，因为该工具不是一个量表。对评价结果可以从如下几个方面报告：①交代总的偏倚风险。如果原始研究偏倚风险都判断为“低”，那么总体可评为“低偏倚风险”。如果研究的一个或多个领域评级为“高”或“不清楚”，那么总体结论就应为“存在偏倚风险”；②总结每个领域偏倚风险考虑评级分别为“低”、“高”、“不清楚”的研究数；③着重描述原始研究普遍评级较差或较好的信号问题和领域；④可同时报告适用性评级，所报告的方面跟偏倚风险评价的相似。

QUADAS网站提供了展示质量评价结果的Excel图（图2）和总结质量评价结果的Word表格模板（表2）下载。此外，评价者也可以根据自己调整的QUADAS-2内容参照表3格式展示对每个信号问题的评价后，再按照表2的格式展示对每个领域的评价结果。

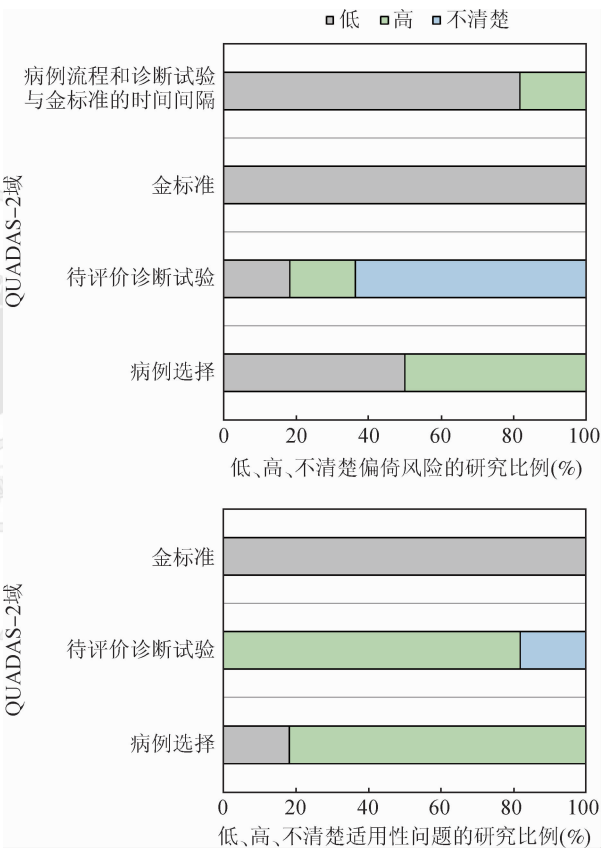

图2 QUADAS-2结果示意图

5. QUADAS-2质量评价结果的应用：对QUADAS-2得到的偏倚风险评价结果的应用包括如下几种方式：评价者可以限制性纳入所有或某些

表2 QUADAS-2各关键域评价结果示意表

研究	偏倚风险				适用性问题		
	病例选择	待评价的诊断试验	金标准	病例流程和诊断试验与金标准的时间间隔	病例选择	待评价的诊断试验	金标准
研究1	☺	☺	☺	☺	☹	☺	☺
研究2	☹	☹	☺	☺	☹	☺	☺
研究3	☹	☹	☺	☺	☹	☺	☺
研究4	☹	?	☺	☺	☹	☺	☺
研究5	☹	?	☺	☺	☹	?	☺

注：☺为低风险，☹为高风险，?不清楚

表3　QUADAS-2条目评分情况示意表

研究	偏倚风险											适用性考虑		
	病例选择			待评价的诊断试验		金标准		病例流程和诊断试验与金标准的时间间隔				病例选择	待评价的诊断试验	金标准
	Q1	Q2	Q3	Q1	Q2	Q1	Q2	Q1	Q2	Q3	Q4			
研究1	U	Y	U	U	U	Y	U	U	N	Y	N	Y	N	Y
研究2	Y	Y	U	U	U	Y	U	U	N	Y	N	Y	N	Y
研究3	Y	U	U	Y	U	Y	U	Y	Y	Y	Y	Y	N	N
研究4	U	Y	U	U	U	Y	U	U	N	Y	N	Y	N	N
研究5	Y	Y	U	U	U	Y	U	U	Y	Y	N	Y	Y	N

注：Y为是；U为不清楚；N为否；Q1～Q4为问题1，问题2，问题3，问题4

关键领域评级为低偏倚风险或低适用性考虑的原始研究进行初步分析，但目前来说，通常更倾向于对所有相关证据进行综述后再研究异质性来源；评价者也可以通过分析在所有或关键领域评级为“高”、“低”或“不清楚”的原始研究中，待评价诊断试验的准确性差异，作为亚组分析和敏感性分析结果呈现；此外，评价者还可以将QUADAS-2评价的关键领域或信号问题结果纳入Meta回归分析中探讨其与诊断准确性估计之间的关联。

三、实例分析

以一个诊断准确性的原始研究举例说明如何在系统综述中应用QUADAS-2开展偏倚风险和适用性评估。该研究于2011年6月26日发表在JACC（Journal of the American College of Cardiology）上，其通过开展队列研究和临床实践验证评价了以3 ng/L作为高敏感度肌钙蛋白hs-cTnT检测的界值排除急性心肌梗死的准确性[4]。2015年发表于The BMJ（The British Medical Journal）的1篇诊断试验准确性研究的系统综述和Meta分析纳入了上述原始研究[5]。这篇系统综述的研究问题是，在急诊室就诊患者中应用hs-cTnT的基线单次测量结果诊断急性心肌梗死的准确性（适用性评价将以此研究问题为依据）。综述者使用QUADAS-2对所有纳入文献进行了方法学质量评价，并将QUADAS-2条目作为异质性来源考虑加入到Meta回归的模型中。综述中对上述原始文献方法学质量的偏倚风险和适用性评价结果见表4。值得注意的是，在开展偏倚风险评价前，综述者首先根据研究内容和目的对QUADAS-2进行了调整，删除了病例选择中“是否避免采用病例-对照设计”、待评价诊断试验中“是否在不知金标准结果的情况下解释诊断试验的结果”、金标准中“金标准是否能正确地区分目标疾病”和病例流程和诊断试验与金标准的时间间隔中“是否所有患者都接受了金标准试验”、“是否所有患者都接受了同样的金标准试验”等5个信号问题，同时在金标准中新增了“金标准是否独立于待评价诊断试验（例如待评价诊断试验不是构成金标准的组成部分）”和“最终的诊断是否由2位医生独立裁定”2个信号问题。同时，综述者提前对每个信号问题及总的偏倚风险的评级标准进行了定义，见表4。

四、讨论

对纳入的诊断准确性原始研究开展严格的偏倚风险评价是系统综述中至关重要的一步。QUADAS-2是在广泛使用的QUADAS工具基础上经过严格、循证的过程产生的诊断准确性研究的质量评价工具。QUADAS-2相比第一版QUADAS工具有了巨大改进，但也存在局限。首先，虽然最初在设计时，制定者考虑到要将QUADAS-2扩展应用于评价比较多种诊断试验准确性研究的质量，但因为在试用时得到的评价者间的信度较差[3]，故认为在比较多种诊断试验准确性研究中使用QUADAS-2的证据基础不足，制定者特别说明QUADAS-2暂不能用于评价比较多种诊断试验的研究，并将在下一步的工作中着力解决这一问题。另外，已有综述者尝试使用改良的QUADAS-2评价比较诊断准确性研究的质量[6]。其次，完成QUADAS-2比原来的QUADAS工具更加费时，为了判断偏倚的风险和适用性评估，需要记录更多的自由文本，但记录的细节对于判断研究的偏倚风险非常有用。最后，QUADAS-2评价中信号问题的答案与偏倚风险和适用性的评级无法一一匹配，例如同一领域的3个信号问题的答案分别为“是”、“否”和“不清楚”时，如何定级偏倚风险和适用性，目前需要评估者事先制定评价标准。对此，制定者最好在使用手册中明确定义，并举例说明。

综上所述，我们建议评估者持续关注QUADAS-2的更新，并积极在线反馈使用经验及问题和建议，使得QUADAS-2更加完善。

表4 QUADAS-2应用实例：偏倚风险和适用性判断

领域	信号问题（是、否或不清楚）以及方法学质量评价标准	问题回答及总体评级
偏倚风险		
病例选择	问题1：是否纳入了病例的连续或随机样本？如果原始研究明确报告连续或随机纳入样本，则为“是”；如果使用非连续或方便样本为“否”；信息不足以判断为“不清楚”	不清楚 原始研究未交代
	问题2：研究是否避免了不恰当的排除？如果所有通常都接受心肌肌钙蛋白检测的疑似急性冠脉综合征患者都纳入研究，为“是”；如果有关患者，如无冠状动脉疾病史的患者被排除为“否”；报告的数据不足以判断为“不清楚”	是
	偏倚风险整体评价[a]	不清楚
待评价诊断试验	问题：如果使用阈值，这个值是否预先设定？如果至少报告了一个预先设定界值的结果为“是”；如果使用ROC曲线或其他方法产生界值，则为“否”	是
	偏倚风险整体评价[a]	低偏倚风险
金标准	问题1：是否在不知诊断试验结果的情况下解释金标准的结果？根据明确报告的情况，评为“是”或“否”，如果没有报告，则为“不清楚”	是
	问题2：金标准是否独立于待评价诊断试验（例如待评价诊断试验不是构成金标准的组成部分）？如果使用罗氏hs-cTnT序列检测作为金标准，“否”；使用其他现行标准的高敏感度心肌肌钙蛋白检测作为金标准为“是”；提供的信息不足以判断为“不清楚”	是
	问题3：最终的诊断是否由两位医生独立裁定？根据具体报告选择“是”或“否”，如果没有报告，则为“不清楚”	是
	偏倚风险整体评价[a]	低偏倚风险
病例流程和诊断试验与金标准的时间间隔	问题1：诊断试验和金标准之间是否有适当的时间间隔？如果金标准的检测是在完成诊断试验的24 h之内进行为“是”；如果超过24 h为“否”；没有报告，则为“不清楚”	是
	问题2：分析中是否纳入了所有患者？如果所有患者均纳入为“是”；如果在分析中排除了研究纳入的患者为“否”，没有足够的数据以供判断为“不清楚”	是
	偏倚风险整体评价[a]	低偏倚风险
适用性考虑		
病例选择	是否存在原始研究纳入患者和环境与系统综述研究问题不符的考虑？如果原始研究纳入的是急诊室就诊的有疑似非ST段抬高急性冠脉综合征症状的非选择性成年患者为“否”；如果患者或环境与系统综述研究问题不符（例如原始研究纳入有ST段抬高或左束支传导阻滞的患者）为“是”；没有详细说明为“不清楚”	是 原始研究未排除ST段抬高患者
	适用性整体评价	高不适用性考虑
待评价诊断试验	是否存在原始研究待评价诊断试验的操作或解释与系统综述中综述问题不同的考虑？如果使用的是商业获得的罗氏hs-cTnT，并按照生产者的推荐使用为“否”；如果不是为“是”	否
	适用性整体评价	低不适用性考虑
金标准	是否存在金标准定义的目标状态（疾病）与综述问题不符的考虑？如果目标状态是非ST段抬高的心肌梗死，即排除初诊心电图提示ST段抬高的患者为“否”；除此之外为“是”	是 原始研究的终点疾病患者未排除初诊心电图提示ST段抬高者
	适用性整体评价	高不适用性考虑

注：[a]Rules for producing an overall risk of bias rating for each domain 每个领域偏倚风险总体评级的规则：1. If all signalling questions within the domain are answered ‘yes’ then the risk of bias for this domain is rated ‘low’; 如果领域内所有信号问题的答案都为“是”，则领域的偏倚风险评级为“低”; 2. If at least one signalling question within the domain is answered ‘no’ then the risk of bias for this domain is rated ‘high’; 如果领域内至少有1个信号问题的答案是“否”，则领域的偏倚风险评级为“高”; 3. If at least one signalling question within the domain is answered ‘unclear’ while the remaining signalling questions are answered ‘yes’ then the risk of bias is rated ‘unclear’。如果领域内至少有1个信号问题的答案是“不清楚”，且其他信号问题的答案都为“是”，则领域的偏倚风险评级为“不清楚”

利益冲突 无

参 考 文 献

[1] Whiting P, Rutjes AWS, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools[J]. J Clin Epidemiol, 2005, 58(1): 1-12. DOI: 10.1016/j.jclinepi.2004.04.008.

[2] Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews[J]. BMC Med Res Methodol, 2003, 3: 25. DOI: 10.1186/1471-2288-3-25.

[3] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies[J]. Ann Intern Med, 2011, 155(8): 529-536. DOI: 10.7326/0003-4819-155-8-201110180-00009.

[4] Body R, Carley S, McDowell G, et al. Rapid exclusion of acute myocardial infarction in patients with undetectable troponin using a high-sensitivity assay[J]. J Am Coll Cardiol, 2011, 58(13): 1332-1339. DOI: 10.1016/j.jacc.2011.06.026.

[5] Zhelev Z, Hyde C, Youngman E, et al. Diagnostic accuracy of single baseline measurement of Elecsys Troponin T high-sensitive assay for diagnosis of acute myocardial infarction in emergency department: systematic review and Meta-analysis[J]. BMJ, 2015, 350: h15. DOI: 10.1136/bmj.h15.

[6] Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool[J]. Res Synth Methods, 2013, 4(3): 280-286. DOI: 10.1002/jrsm.1080.

（收稿日期：2017-08-01）

（本文编辑：王岚）