·基础理论与方法·

诊断试验准确性比较研究的偏倚评估工具——QUADAS-C

杨秋玉[1]　陆瑶[2,3]　谢欣玲[4]　赖鸿皓[2,3]　田晨[2,3]　牛猛[5]　田金徽[6,7]　李霓[8]　李江[8]　葛龙[2,3,7]

[1]兰州大学护理学院循证护理研究中心，兰州 730000；[2]兰州大学公共卫生学院循证社会科学研究中心，兰州 730000；[3]兰州大学公共卫生学院社会医学与卫生事业管理研究所，兰州 730000；[4]兰州大学第二临床医学院，兰州 730000；[5]兰州大学第一医院放射科，兰州 730000；[6]兰州大学基础医学院循证医学中心，兰州 730000；[7]甘肃省循证医学与临床转化重点实验室，兰州 730000；[8]国家癌症中心/国家肿瘤临床医学研究中心/中国医学科学院北京协和医学院肿瘤医院癌症早诊早治办公室，北京 100021

杨秋玉和陆瑶对本文有同等贡献

通信作者：葛龙，Email：gelong2009@163.com；李江，Email：lij@cicams.ac.cn

【摘要】 本文介绍了诊断试验准确性比较研究的偏倚评估工具（QUADAS-C）的主要内容，阐述了QUADAS-C与QUADAS-2的联系与区别，同时介绍了如何将QUADAS-C与QUADAS-2结合使用及如何呈现评价结果。QUADAS-C在QUADAS-2的基础上，共扩展14个标志性问题，形成4个关键领域（病例选择、待评价试验、金标准、病例流程和待评价试验与金标准之间的时间间隔）。与QUADAS-2不同的是，QUADAS-C各领域第一个标志性问题的回答均需结合QUADAS-2的评价结果；此外，QUADAS-C仅评价原始研究的偏倚风险而不包含适用性评价。完成QUADAS-C评价最终得出原始研究每个领域偏倚风险为“低”“高”或“不清楚”的结果。

【关键词】 诊断试验准确性比较研究；　偏倚风险；　系统评价

基金项目：北京市科技新星计划（Z201100006820070）；甘肃省科技计划（20CX4ZA027，20CX9ZA112）

QUADAS-C—A tool for assessing risk of bias regarding Quality Assessment of Diagnostic Accuracy Studies-Comparative

Yang Qiuyu[1], Lu Yao[2,3], Xie Xinling[4], Lai Honghao[2,3], Tian Chen[2,3], Niu Meng[5], Tian Jinhui[6,7], Li Ni[8], Li Jiang[8], Ge Long[2,3,7]

[1]Evidence Based Nursing Centre, School of Nursing, Lanzhou University, Lanzhou 730000, China; [2]Evidence-Based Social Science Research Center, School of Public Health, Lanzhou University, Lanzhou 730000, China; [3]Department of Social Science and Health Management, School of Public Health, Lanzhou University, Lanzhou 730000, China; [4]The Second School of Clinical Medicine of Lanzhou University, Lanzhou 730000, China; [5]Department of Radiology, the First Hospital of Lanzhou University, Lanzhou 730000, China; [6]Evidence Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou 730000, China; [7]Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou 730000, China; [8]National Cancer Center/National Cancer Clinical Medical Research Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking

DOI：10.3760/cma.j.cn112338-20211101-00841

收稿日期　2021-11-01　　**本文编辑**　万玉立

引用格式：杨秋玉，陆瑶，谢欣玲，等. 诊断试验准确性比较研究的偏倚评估工具——QUADAS-C[J]. 中华流行病学杂志，2022，43(6)：938-944. DOI：10.3760/cma.j.cn112338-20211101-00841.

Yang QY, Lu Y, Xie XL, et al. QUADAS-C—A tool for assessing risk of bias regarding Quality Assessment of Diagnostic Accuracy Studies-Comparative[J]. Chin J Epidemiol, 2022, 43(6): 938-944. DOI: 10.3760/cma. j. cn112338-20211101-00841.

Union Medical College, Beijing 100021, China
Yang Qiuyu and Lu Yao contributed equally to the article
Corresponding authors: Ge Long, Email: gelong2009@163.com; Li Jiang, Email: lij@cicams.ac.cn

【Abstract】 This paper introduced the Quality Assessment of Diagnostic Accuracy Studies-Comparative (QUADAS-C), illustrated the comparison with the QUADAS-2, and using QUADAS-C together with QUADAS-2 to present QUADAS-C results through systematic reviews. Like the domain for QUADAS-2, QUADAS-C retained four domains, including patient selection, index test, reference standard, flow, and timing, and comprised additional questions for each QUADAS-2 part. Unlike the QUADAS-2 tool, the starting question of each domain for QUADAS-C was designed to summarize the risk of biased information captured by QUADAS-2. QUADAS-C only dealt with the risk of bias but did not include the part of concerns regarding applicability. The answers to signaling questions for each domain of QUADAS-C would lead to a 'low''high' or 'unclear' risk of biased judgment for the original study.

【Key words】 Comparative diagnostic test accuracy study; Risk of bias; Systematic review

Fund programs: Beijing Nova Program (Z201100006820070); Gansu Province Science and Technology Planning Project (20CX4ZA027, 20CX9ZA112)

诊断试验准确性(diagnostic test accuracy,DTA)研究在评估诊断指标的准确性中非常重要,不仅可评价某种诊断试验的准确性,还可评估多种诊断试验并比较其准确性。在同一诊断试验研究中比较2个及以上的DTA研究,即为DTA比较研究[1-2]。可靠的诊断试验可为临床决策提供重要依据,而诊断试验的可靠性来源于无偏倚的研究设计和方法。目前,已开发一系列工具和清单来评价诊断试验研究的偏倚风险[3]。其中诊断准确性研究质量评价工具(Quality Assessment of Diagnostic Accuracy Studies,QUADAS)是最常见的DTA偏倚风险评估工具,首个版本的QUADAS于2003年发布[4],并于2011年进行修订,形成QUADAS-2[5]。然而,QUADAS-2不能完全识别DTA比较研究的偏倚来源。因此,QUADAS-C(QUADAS-Comparative)工作组开发并于2021年10月在*Annals of Internal Medicine*发布了QUADAS-2的扩展版——QUADAS-C工具[6]。本文对QUADAS-C的主要内容和使用方法进行介绍和解读。

一、QUADAS-C的制定背景

在DTA比较研究中,试验间准确性的比较可产生偏倚,例如,在一项比较2个DTA研究中,试验A和试验B接受不同的金标准试验,或者在已知试验B结果的情况下解释试验A结果,这些问题均可使试验间准确性比较结果产生偏倚。研究表明,2017年已有238篇DTA比较研究的系统评价发表,但仅2篇系统评价计划或已经实施了偏倚风险评估,该研究也没有确定任何适用于评价DTA比较研究偏倚风险的工具,表明QUADAS-2并不完全适用于DTA比较研究的偏倚评价[2]。因此,为制定适用于DTA偏倚风险评价的工具,QUADAS-C工作组邀请了24名国际上诊断试验领域的专家,通过4轮德尔菲调查和1轮面对面专家共识,对QUADAS-2进行优化和设计,开发出了QUADAS-C工具。

二、工具解读

1. QUADAS-C与QUADAS-2的区别与联系:QUADAS-2工具主要用于单个诊断试验研究的偏倚风险和适用性评价,包括11个标志性问题,分为4个关键领域,包括病例选择、待评价试验、金标准、病例流程和待评价试验与金标准之间的时间间隔,其中适用性评价只包含前三者。每个问题可用“是”“否”或“不清楚”回答,最后根据每个问题的评价结果得出每个领域偏倚风险和适用性的评价结果,包括“高”“低”或“不清楚”。QUADAS-C工具用于2个及以上DTA比较研究的偏倚风险评价,包含14个标志性问题,也分为相同的4个领域,每个问题可用“是”“否”“不清楚”或“不适用”回答,其中每个领域第一个标志性问题的回答需结合QUADAS-2对每个诊断试验的评价结果,最终每个领域偏倚风险的评价结果可为“高”“低”或“不清楚”。该工具不包含适用性评价,但工作组建议使用者基于QUADAS-2工具评价准确性比较研究中准确性较差的诊断试验的适用性,以此判断诊断准确性比较研究的适用性。

2. QUADAS-C工具的介绍:完整版本的QUADAS-C可在QUADAS网站(www.quadas.org)获取。QUADAS-C工具主要适用于完全配对或随机设计的DTA比较研究,虽然也可用于评估其他设计类型的DTA比较研究,但需根据相应的研究设

计调整评价条目(例如增加新的或删除现有的标志性问题)。

(1)QUADAS-C与QUADAS-2结合使用:首先,从病例选择领域开始,使用者基于QUADAS-2工具独立评价每个待评价试验。假设研究只比较2个待评价试验的准确性,则分别得出待评价试验A和试验B的偏倚风险和适用性评价结果。接着,同样在病例选择领域,使用者基于QUADAS-C工具评价试验A与试验B准确性比较的偏倚风险。每个领域第一个标志性问题均是"在该领域中,每个待评价试验的偏倚风险是否被判断为'低'?",回答该问题需结合QUADAS-2对待评价试验A和试验B的评价结果,如果待评价试验A与试验B的偏倚风险均评价为"低",则该标志性问题的回答为"是"。然后,使用者基于QUADAS-C继续回答该领域的其他标志性问题,综合该领域标志性问题的评价结果,可得出待评价试验A与试验B准确性比较研究在该领域中偏倚风险的评价结果。对于其他领域,评估流程相同。见表1。

(2)标志性问题的回答:领域1:病例选择:

标志性问题1(C1.1):在该领域中,每个待评价试验的偏倚风险是否评为"低"?

QUADAS-C每个领域第一个问题均是从评价每个待评价试验的偏倚风险是否为"低"开始(结合QUADAS-2的评价结果)。在评价过程中,若每个待评价试验的偏倚风险(QUADAS-2中的问题1.4)均被评为"低",则应回答"是";若至少有一个待评价试验的偏倚风险被评为"高"或"不清楚",则应回答"否"。

回答为"否"的例子[7]:在一项比较磁共振成像(MRI)和磁共振关节造影(MRA)诊断肩关节盂唇损伤(SLAP)准确性的研究中,仅纳入接受MRI、MRA和金标准试验的参与者,因此只接受MRI或MRA或不接受金标准试验的参与者均被排除,本研究纳入的所有参与者很可能是从可疑患者中抽样,因此,在QUADAS-2中,MRI、MRA的偏倚风险均评价为"高",则QUADAS-C在该标志性问题的回答为"否"。

标志性问题2(C1.2):是否使用完全配对或随机研究设计?

该问题主要关注原始研究是否采用了完全配对研究设计(每位参与者接受所有待评价试验)或随机研究设计(参与者被随机分配到某一个待评价试验的研究设计)。由于这两种研究设计可使待评价试验间具有可比性,因此是DTA的理想研究设计。虽然部分配对随机子集设计(partially paired, random subset design)是为了减少待评价试验间的混杂,但在该标志性性问题中回答应为"否"。在评价过程中,若原始研究采用完全配对或随机设计,则回答为"是",若未采用则回答为"否";若无足够信息来判断,则回答为"不清楚"。

回答为"否"的例子[8]:在一项比较MRI和MRA诊断SLAP准确性的研究中,参与者根据医生的要求接受MRI或MRA检查。因此,该研究不属于完全配对或随机研究。

标志性问题3(C1.3):分配序列是否随机(仅适用于随机研究设计)?

如果使用随机化方法将参与者分配到每个待评价试验中,则需仔细审查分配序列的随机方法。在评价过程中,若原始研究的分配序列是随机生成,如通过计算机生成随机号、随机数字表或抽签等方法,则标志性问题3的回答应为"是";若非随机生成,如通过交替分配、基于日期(如出生日期或住院顺序)、临床医生或患者的选择等方法,则评为"否";若无足够信息判断,则回答为"不清楚";若原始研究并非随机研究设计,则回答为"不适用"。

回答为"否"的例子[9]:一项研究比较两种经皮肝活检针(Menghini和Tru-cut)诊断肝硬化的准确性,虽然该研究描述为随机研究,但研究者使用了交替分配方案:每个月更换活检针的类型来实现随机化。因此,分配序列的产生并非随机。

标志性问题4(C1.4):病例被纳入和分配到待评价试验前,随机分配序列是否被隐藏(仅适用于随机研究设计)?

分配隐藏的恰当方法包括中心随机方案(例如由独立的中心药房、电话或互联网的随机服务提供商执行)和不透明密闭信封。在评价过程中,若原始研究分配序列被隐藏,则回答为"是";若未被隐藏,则回答为"否";若无足够信息判断,则回答为"不清楚";若原始研究并非随机研究设计,则回答为"不适用"。

回答为"否"的例子[9]:在比较Menghini和Tru-cut活检针诊断肝硬化的准确性的研究中,分配顺序未被隐藏,因为使用了一种可预测的分配方法(交替分配)。

领域2:待评价试验:

标志性问题1(C2.1):在该领域中,每个待评价试验的偏倚风险是否评为"低"?

表 1　QUADAS-C 与 QUADAS-2 结合使用的建议[6]

领域	评估内容	评估结果	
领域1：病例选择			
QUADAS-2		待评价试验A	待评价试验B
标志性问题	1.1 是否纳入了连续或随机的病例?	是/否/不清楚	是/否/不清楚
	1.2 是否避免采用病例—对照类研究设计?	是/否/不清楚	是/否/不清楚
	1.3 研究是否避免了不恰当的排除?	是/否/不清楚	是/否/不清楚
偏倚风险	1.4 选择的病例是否会产生偏倚?	低/高/不清楚	低/高/不清楚
临床适用性	1.5 原始研究中纳入的病例特征是否与系统评价不符?	低/高/不清楚	低/高/不清楚
QUADAS-C		待评价试验准确性的比较	
标志性问题	C1.1 在该领域中，每个待评价试验的偏倚风险是否被评为“低”?[a]	是/否	
	C1.2 是否使用完全配对或随机研究设计?	是/否/不清楚	
	C1.3 分配序列是否随机?[b]	是/否/不清楚/不适用	
	C1.4 病例被纳入和分配到待评价试验前，随机分配序列是否被隐藏?[b]	是/否/不清楚/不适用	
偏倚风险	C1.5 病例选择是否在比较中产生偏倚?	低/高/不清楚	
领域2：待评价试验			
QUADAS-2		待评价试验A	待评价试验B
标志性问题	2.1 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?	是/否/不清楚	是/否/不清楚
	2.2 若使用了阈值，那么该阈值是否预先确定?	是/否/不清楚	是/否/不清楚
偏倚风险	2.3 待评价试验的实施或解释是否会产生偏倚?	低/高/不清楚	低/高/不清楚
临床适用性	2.4 原始研究中的诊断试验的实施或解释是否与系统评价中不同?	低/高/不清楚	低/高/不清楚
QUADAS-C		待评价试验准确性的比较	
标志性问题	C2.1 在该领域中，每个待评价试验的偏倚风险是否被评为“低”?[a]	是/否	
	C2.2 是否在不知道其他待评价试验结果的情况下解释拟评价试验结果?[c]	是/否/不清楚/不适用	
	C2.3 正在进行的待评价试验是否不可能影响其他待评价试验?[c]	是/否/不清楚/不适用	
	C2.4 诊断试验的实施和解释是否不会有利于某个待评价试验?	是/否/不清楚	
偏倚风险	C2.5 待评价试验的实施和解释是否在比较中产生偏倚?	低/高/不清楚	
领域3：金标准			
QUADAS-2		待评价试验A	待评价试验B
标志性问题	3.1 金标准是否可以正确地区分目标疾病?	是/否/不清楚	是/否/不清楚
	3.2 金标准的解释是否在对待评价试验结果不知情的情况下做出的?	是/否/不清楚	是/否/不清楚
偏倚风险	3.3 金标准的实施及解释是否会产生偏倚?	低/高/不清楚	低/高/不清楚
临床适用性	3.4 金标准所定义的目标疾病是否与系统评价中不符?	低/高/不清楚	低/高/不清楚
QUADAS-C		待评价试验准确性的比较	
标志性问题	C3.1 在该领域中，每个待评价试验的偏倚风险是否被评为“低”?[a]	是/否	
	C3.2 金标准是否避免合并任何待评价试验?	是/否/不清楚	
偏倚风险	C3.3 金标准及其实施和解释是否在比较中产生偏倚?	低/高/不清楚	
领域4：病例流程和待评价试验与金标准之间的时间间隔			
QUADAS-2		待评价试验A	待评价试验B
标志性问题	4.1 待评价试验和金标准之间是否有恰当的时间间隔?	是/否/不清楚	是/否/不清楚
	4.2 是否所有患者都接受了金标准试验?	是/否/不清楚	是/否/不清楚
	4.3 是否所有患者都接受了相同的金标准试验?	是/否/不清楚	是/否/不清楚
	4.4 是否所有患者均被纳入分析?	是/否/不清楚	是/否/不清楚
偏倚风险	4.5 病例的流程是否会产生偏倚?	低/高/不清楚	低/高/不清楚
QUADAS-C		待评价试验准确性的比较	
标志性问题	C4.1 在该领域中，每个待评价试验的偏倚风险是否被评为“低”?[a]	是/否	
	C4.2 待评价试验间是否有恰当的时间间隔?	是/否/不清楚	
	C4.3 所有待评价试验是否使用相同的金标准?	是/否/不清楚	
	C4.4 待评价试验间缺失数据的比例和原因是否相似?	是/否/不清楚	
偏倚风险	C4.5 病例流动是否在比较中产生偏倚?	低/高/不清楚	

注：[a]C1.1、C2.1、C3.1、C4.1 的判断需结合 QUADAS-2 的评价结果；[b]仅适用于随机设计；[c]仅用于接受多种诊断试验的参与者（完全配对或部分配对设计）

评价方法与C1.1相同。

标志性问题2(C2.2):是否在不知道其他待评价试验结果的情况下解释拟评价试验结果(仅适用于配对研究设计)?

该标志性问题只适用于完全配对或部分配对研究设计,判断偏倚风险时应考虑3个因素:①结果解释的主观性程度。与明确输出结果的待评价试验相比,需主观解释的待评价试验更容易产生偏倚。②待评价试验实施和解释的顺序。假设待评价试验A总是在待评价试验B之前实施和解释,那么可推断出待评价试验A是在不知道待评价试验B的情况下解释结果,解释待评价试验B的结果时则需对待评价试验A的结果施盲。③在单个待评价试验与诊断策略试验(包含多个诊断试验)的比较中,回答为"否"并不意味着高偏倚风险,判断时需考虑是否符合临床实践。在评价过程中,若试验A的解释在不知试验B结果的情况下进行,则应回答"是",否则应回答为"否";若无足够信息判断,则回答为"不清楚";若每位参与者只接受一种待评价试验,则应回答"不适用"。

回答为"否"的例子[10]:一项研究比较简易智力状况检测量表(MMSE)与全科医生认知功能评估量表(GPCOG)在相同参与者中筛查痴呆的准确性。测试由同一护士在同一阶段进行(先进行MMSE测试,后进行GPCOG测试),因此护士在进行GPCOG测试时已经知道MMSE的测试结果。

标志性问题3(C2.3):正在进行的待评价试验是否不可能影响其他待评价试验(仅适用于配对研究设计)?

该标志性问题只适用于完全配对和部分配对研究设计。如果某一个待评价试验影响或干扰后续待评价试验的诊断准确性,则可能发生偏倚。标志性问题C2.2侧重于结果解释时引起的偏倚,标志性问题C2.3强调待评价试验的相互影响,例如接受多轮问卷调查的参与者因疲劳而影响后续问卷调查的效果。理想情况下,如果某一待评价试验可能影响后续待评价试验的效果,那么每位参与者应只接受一种待评价试验。在评价过程中,若某一试验不影响其他试验的准确性,则应回答"是";若影响则应回答"否";若无够信息判断,则应回答"不清楚";若每位参与者只接受一种待评价试验,则应回答"不适用"。

回答为"否"的例子[11]:一项研究比较了3种认知功能测试量表在相同参与者中筛查痴呆的准确性,包括MMSE、罗兰通用痴呆量表(RUDAS)、改良的金伯利土著认知评估量表(KICA)。由于这些量表包含类似的测试(例如绘画、手锻炼),参与者可能得出相同的测试效果。

标志性问题4(C2.4):诊断试验的实施和解释是否不会有利于某个待评价试验?

在平等的比较中,所有待评价试验都应在相同的情况下实施和解释,避免不恰当的有利于某个待评价试验的因素。例如在同一研究中,一项生物标志物检查使用新鲜的样本,而另一生物标志物检查使用冷冻样本,那么比较的结果很可能存在偏倚。如果这种差异符合临床实践,则是可接受的。在评价过程中,若待评价试验间的实施和解释无差异或差异符合临床实践,则应回答"是";若存在差异,则应回答"否";若无足够信息判断,则应回答"不清楚"。

回答为"否"的例子[12]:在一项非配对研究中,比较正常剂量CT和低剂量CT诊断儿童阑尾炎的准确性,正常剂量CT使用老一代的CT扫描仪进行测试,而低剂量CT使用现代CT扫描仪进行测试。

领域3:金标准:

标志性问题1(C3.1):在该领域中,每个待评价试验的偏倚风险是否被评为"低"?

评价方法与C1.1相同。

标志性问题2(C3.2):金标准是否避免合并任何待评价试验?

如果某一待评价试验是金标准的一部分,那么该试验的准确性偏高,试验间准确性的比较则存在偏倚。理想情况下,待评价试验不应成为金标准的一部分。在评价过程中,若待评价试验不属于金标准的一部分,则应回答"是";若属于则应回答"否";若无足够信息判断,则应回答"不清楚"。

回答为"否"的例子[10]:一项研究比较两份量表(MMSE和GPCOG)筛查痴呆的准确性,两份问卷均采用剑桥老年认知量表(CAMCOG)作为金标准,其中MMSE是CAMCOG中的一个分量表。

领域4:病例流程和待评价试验与金标准之间的时间间隔:

标志性问题1(C4.1):在该领域中,每个待评价试验的偏倚风险是否被判断为"低"?

评价方法与C1.1相同。

标志性问题2(C4.2):待评价试验间是否有恰当的时间间隔?

一般来说,当参与者入组后,同时实施所有待

评价试验，则待评价试验间的时间间隔被认为恰当。然而，对于不同目标疾病和待评价试验，恰当的时间间隔有所不同。例如，对于慢性病，短期的时间间隔可能是可以接受的，但对于急性疾病，则时间间隔不应过长。在临床实践中，待评价试验通常在不同的时间点进行，同时进行多个试验反而是不必要甚至不可取，因此需要结合临床专业知识判断时间间隔是否恰当。在评价过程中，若待评价试验同时进行或虽不同时进行但时间间隔恰当，则回答为“是”，否则应回答为“否”；若无足够信息判断，则回答“不清楚”。

回答为“否”的例子[13]：一项完全配对的研究中，比较了胸部超声与胸部 CT 诊断肺实质脓肿或坏死的准确性。超声和 CT 实施的平均时间间隔为 2.7 d（范围为 0~8 d）。由于时间间隔较长，肺部病变已经恶化，超声和 CT 的检查结果差异较大，二者准确性的比较存在偏倚。

标志性问题 3（C4.3）：所有待评价试验是否使用相同的金标准？

如果待评价试验 A 和试验 B 使用不同的金标准，那么二者间准确性的比较存在偏倚。在评价过程中，若所有参与者使用相同的金标准，则应回答为“是”，否则应回答为“否”，若待评价试验间的金标准可以互换（以同样的方式检测同一目标疾病），则该问题虽回答为“否”，但不意味着高偏倚风险；若无足够信息判断，则应回答“不清楚”。

回答为“否”的例子[14]：在一项完全配对研究中，使用 Hemoccult Ⅱ、Hemoccult Ⅱ Sensa 和 HemeSelect 筛查大肠癌，其中 Hemoccult Ⅱ 和/或 HemeSelect 检测阳性者接受结肠镜检查（金标准）；若仅 Hemoccult Ⅱ Sensa 检测为阳性，则建议在 6 个月和 12 个月时重复 Hemoccult Ⅱ 检测并进行乙状结肠镜检查（金标准）。因此，3 种试验的金标准不完全相同。

标志性问题 4（C4.4）：待评价试验间缺失数据的比例和原因是否相似？

如果待评价试验结果不可用、无效、或参与者在结果分析时被排除，就会出现数据缺失的情况。研究者应仔细考虑缺失数据的比例和原因是否对待评价试验间准确性的比较产生影响。在评价过程中，若无数据缺失或待评价试验间缺失比例或原因相同，则应回答“是”；若不同，则应回答“否”；若无足够信息判断，则应回答“不清楚”。

回答为“否”的例子[15]：在一项完全配对研究中，比较醋酸试验肉眼观察（VIA）与 HPV 检测筛查宫颈癌的准确性，VIA 组排除 30/4 039（0.7%）女性，HPV 组排除 493/4 039（12.2%）女性，作者报告这些女性被排除的原因是“不完整或不确定的调查”，但未解释 VIA 组和 HPV 组缺失参与者比例是否存在差异。

（3）使用 QUADAS-C 工具评估偏倚风险的 4 个阶段：工作组建议使用者基于 4 个阶段完成 QUADAS-C 评估，与 QUADAS-2 评估流程相似，均包括：①陈述系统评价问题；②根据需要调整评价工具及评价指南；③评估者准备或构建每个纳入原始研究的流程图；④判断偏倚风险。具体步骤见表 2[6]。

（4）QUADAS-C 偏倚风险评价结果的呈现：在诊断准确性比较研究的系统评价中，使用者应总结 QUADAS-2 和 QUADAS-C 的评价结果。QUADAS 网站提供了多种 QUADAS-C 偏倚风险评价结果呈现的图表，本文呈现其中一种模板（表 3），其他呈现方式详见 www.quadas.org。

表 2　QUADAS-C 评价的 4 个阶段[6]

阶　段	具 体 内 容
①陈述系统评价问题	使用者应明确比较待评价试验、目标疾病、病例 待评价试验可以是包含多个诊断试验的诊断策略
②根据需要调整评价工具及评价指南	使用者应根据需要在 QUADAS-C 中添加或删除标志性问题，同时建立评估专用的偏倚风险评价指南。若评价者间一致性不好，则需完善评估工具和评价指南
③评估者准备或构建每个纳入原始研究的流程图	建议使用者查看已发布的流程图。若原始研究中未报告或者报告不充分，则需自行绘制流程图。DTA 比较研究的流程图应该包括以下信息：如何招募参与者，参与者如何分配到每个待评价试验，每个参与者接受待评价试验顺序（适用于配对研究设计），以及缺失的或未接受金标准试验的参与者数量
④判断偏倚风险	各领域偏倚风险评估：若领域的每个标志性问题均评为“是”，则该领域的偏倚风险判断为“低”；若某一个标志性问题评为“否”，则判断为“高”。若无足够的信息判断偏倚风险，则判断为“不清楚” 整体偏倚风险评估：虽然整体偏倚风险的评估不是 QUADAS-C 正式组成的部分，但使用者可根据每个领域的偏倚风险结果做出判断。例如，全部领域的偏倚风险均评价为“低”，则可判断为“整体低偏倚风险”；若至少一个领域的偏倚风险评价为“高”，则可判断为“整体高偏倚风险”；若至少一个领域的偏倚风险评价为“不清楚”（无判断为高偏倚风险的领域），则可判断为“整体偏倚风险不明确”

注：DTA：诊断试验准确性

表 3　QUADAS-2 和 QUADAS-C 评价结果示意表[6]

研究	试验	偏倚风险（QUADAS-2）				适用性（QUADAS-2）			偏倚风险（QUADAS-C）			
		P	I	R	FT	P	I	R	P	I	R	FT
作者，年份	A	√	×	√	√	√	√	√	×	×	√	√
	B	√	√	√	√	√	√	×				
作者，年份	A	?	√	√	×	√	?	√	?	×	√	×
	B	?	√	√	×	√	√	√				
作者，年份	A	√	√	√	√	?	√	√	√	?	?	√
	B	√	?	√	√	?	√	√				

注：P 为病例选择；I 为待评价试验；R 为金标准；FT 为病例流程和待评价试验与金标准之间的时间间隔；√为低；×为高；?为不清楚

三、讨论

DTA 比较研究的偏倚风险评估，不仅是制作系统评价/Meta 分析的重要一步，也可为原始研究设计提供参考，减少偏倚的产生；另外，DTA 比较研究质量的可靠性是决定将哪个诊断试验用于临床实践的重要依据，因此，评价研究设计、实施、分析是否存在偏倚非常必要。QUADAS-C 在 QUADAS-2 的基础上开发，在 4 个关键领域共增加了 14 个标志性问题，在评价过程中需结合 QUADAS-2 的评价结果；由于 4 个标志性问题（C1.3、C1.4、C2.2 和 C2.3）适用于特定的研究设计，因此增加“不适用”这一选项。值得注意的是，QUADAS-C 适用于诊断试验间准确性的比较，而不适用于评价诊断治疗对患者重要结局有效性的研究。

QUADAS-C 的使用存在局限性。首先，QUADAS-C 需要结合 QUADAS-2 使用，若使用者不熟悉 QUADAS-2，则会认为标志性问题数量非常多，评价耗时。其次，虽然可用于评估 2 个及以上 DTA 比较的研究，但仍存在较大挑战。此外，QUADAS-C 主要针对完全配对或随机研究设计开发，虽然可用于其他设计类型准确性比较研究，但需要对工具进行“裁剪”。由于 QUADAS-C 基于专家共识和理论形成，可能未完全识别比较研究中所有偏倚，因此，随着对 DTA 比较研究偏倚风险的认识，将继续更新完善 QUADAS-C。目前工作组正在计划开发基于网络的工具，可以解决用户提出的一些问题，例如自动完成标志性问题的评价，可选择查看标志性问题的解释，结合 QUADAS-2 和 QUADAS-C 的评价结果自动构建并输出偏倚风险评价的图表。

综上所述，QUADAS-C 工具可帮助使用者评价偏倚风险，避免原始研究的潜在偏倚，随着对 DTA 比较研究中偏倚来源的进一步认识，工作组将继续更新完善 QUADAS-C。

利益冲突　所有作者声明无利益冲突

作者贡献声明　杨秋玉、陆瑶：文献查阅、论文撰写；谢欣玲、赖鸿皓、田晨：资料整理、文献翻译；牛猛、田金徽、李霓：写作指导、经费支持；李江、葛龙：论文设计、论文修改和审校

参 考 文 献

[1] Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy[J]. Ann Intern Med, 2013, 158(7): 544-554. DOI:10.7326/0003-4819-158-7-201304020-00006.

[2] Yang BD, Vali Y, Sharifabadi AD, et al. Risk of bias assessment of test comparisons was uncommon in comparative accuracy systematic reviews: an overview of reviews[J]. J Clin Epidemiol, 2020, 127: 167-174. DOI: 10.1016/j.jclinepi.2020.08.007.

[3] 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.

[4] Whiting P, Rutjes AW, 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.

[5] 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.

[6] Yang BD, Mallett S, Takwoingi Y, et al. QUADAS-C:a tool for assessing risk of bias in comparative diagnostic accuracy studies[J]. Ann Intern Med, 2021, 174(11): 1592-1599. DOI:10.7326/M21-2234.

[7] Magee T. 3-T MRI of the shoulder: is MR arthrography necessary?[J]. Am J Roentgenol, 2009, 192(1):86-92. DOI: 10.2214/ajr.08.1097.

[8] Tuite MJ, Rutkowski A, Enright T, et al. Width of high signal and extension posterior to biceps tendon as signs of superior labrum anterior to posterior tears on MRI and MR arthrography[J]. Am J Roentgenol, 2005, 185(6): 1422-1428. DOI:10.2214/AJR.04.1684.

[9] Colombo M, del Ninno E, de Franchis R, et al. Ultrasound-assisted percutaneous liver biopsy: superiority of the Tru-Cut over the Menghini needle for diagnosis of cirrhosis[J]. Gastroenterology, 1988, 95(2): 487-489. DOI:10.1016/0016-5085(88)90509-4.

[10] Brodaty H, Connors MH, Loy C, et al. Screening for dementia in primary care:a comparison of the GPCOG and the MMSE[J]. Dement Geriatr Cogn Disord, 2016, 42(5/6): 323-330. DOI:10.1159/000450992.

[11] Radford K, Mack HA, Draper B, et al. Comparison of three cognitive screening tools in older urban and regional aboriginal Australians[J]. Dement Geriatr Cogn Disord, 2015, 40(1/2):22-32. DOI:10.1159/000377673.

[12] Didier RA, Vajtai PL, Hopkins KL. Iterative reconstruction technique with reduced volume CT dose index:diagnostic accuracy in pediatric acute appendicitis[J]. Pediatr Radiol, 2015, 45(2):181-187. DOI:10.1007/s00247-014-3109-7.

[13] Kurian J, Levin TL, Han BK, et al. Comparison of ultrasound and CT in the evaluation of pneumonia complicated by parapneumonic effusion in children[J]. AJR Am J Roentgenol, 2009, 193(6): 1648-1654. DOI: 10.2214/AJR.09.2791.

[14] Allison JE, Tekawa IS, Ransom LJ, et al. A comparison of fecal occult-blood tests for colorectal-cancer screening[J]. N Engl J Med, 1996, 334(3): 155-160. DOI: 10.1056/NEJM199601183340304.

[15] Shastri SS, Dinshaw K, Amin G, et al. Concurrent evaluation of visual, cytological and HPV testing as screening methods for the early detection of cervical neoplasia in Mumbai, India[J]. Bull World Health Organ, 2005, 83(3):186-194.