

· 循证理论与实践 ·

Meta分析系列之七：非随机实验性研究、诊断性试验及动物实验的质量评价工具

曾宪涛, 庄丽萍, 杨宗国, 董圣杰

[中图分类号] R4 [文献标志码] A [文章编号] 1674-4055(2012)06-0496-04

非随机实验性研究 (non-randomized experimental study) 和诊断准确性试验的Meta分析是临床研究重要组成部分, 而临床前的动物实验的Meta分析也日渐得到重视, 其结论已成为循证决策的依据之一。在前面的章节中, 我们已对随机对照试验^[1]及观察性研究^[2]的质量评价工具进行了简介, 现对较为常用的或推荐使用的非随机实验性研究、诊断准确性试验及动物实验的评价工具进行简介。

1 非随机实验性研究

1.1 MINORS 条目 非随机对照试验方法学评价指标 (methodological index for non-randomized studies, MINORS)^[3]是由法国外科医师Slim等在2007年全面回顾文献及专家共识的基础上制定的临床干预研究的质量评价工具, 特别适用于外科非随机对照干预性研究 (non-randomized surgical studies) 质量的评价。评价指标共12条, 每一条分为0~2分。前8条针对无对照组的研究, 最高分为16分; 后4条与前8条一起针对有对照组的研究, 最高分共24分。0分表示未报道; 1分表示报道了但信息不充分; 2分表示报道了且提供了充分的信息 (表1)。

1.2 Reisch评价工具 1988年, 为了促进临床药物治疗性研究质量的评价, 美国德克萨斯大学医学部的Reisch等^[4]制作了相应的评价工具, 即Reisch评价工具。该工具以公认的标准和清单为基础, 归纳了12个类别共57个条目, 其中包括设计、样本量、随机化和对对照组的相关内容。清单的主要作用是评价研究的设计和绩效而不是数据的分析。该工具的每个条目都是以确定的问题形式出现, 回答方式包括“是”, “否”, “不知道或不清楚”或“没有合适的答案”。

Reisch评价工具主要是针对临床药物治疗质量的评价, 因此很多条目具有明显的临床药物特殊性。该工具对于非随机实验性研究的质量评价可用度较高, 但需结合实际的研究做一些调整以发挥该工具的评价功能。现被Cochrane肠道炎症疾病小组使用, 因其条目繁多, 使用耗时, 易用性不强, 在此不作详细介绍。

1.3 TREND 声明 TREND (the Transparent Reporting of

基金项目: 湖北省教育科学“十二五”规划2012年度重点课题(2012A050), 湖北医药学院2011年度优秀中青年科技创新团队项目(2011 CZX01)

作者单位: 442000 十堰, 湖北医药学院附属太和医院口腔医学中心(曾宪涛); 复旦大学附属肿瘤医院中西医结合科 上海医学院肿瘤学系(庄丽萍); 复旦大学附属公共卫生临床中心中医科(杨宗国); 苏州大学附属第一医院骨科(董圣杰)

通讯作者: 董圣杰, E-mail: dongshengjay@126.com

doi: 10.3969/j.1674-4055.2012.06.003

Evaluations with Nonrandomized Designs) 声明^[5]是美国疾病预防控制中心 (Centers for Disease Control and Prevention, CDC) HIV/AIDS综合防治研究 (PRS) 小组为提高对艾滋病防治行为干预研究的综合能力, 于2003年7月在亚特兰大召开的CDC下属期刊编辑会议上达成的共识。与会者认为更清晰和标准的研究评价报告不应只包括随机设计, 还要扩展到非随机对照设计, 由此提出非随机对照设计报告规范 (TREND), 会议等具体内容可参阅官网<http://www.cdc.gov/trendstatement/>。自然, 这一规范可以用来非随机实验性研究的质量, 因此不是专用工具, 在此不作详细列出。

2 诊断性研究

诊断性研究一般而言可以分为两类: 一是评价运用诊断试验后对改善患者的治疗/预后效果, 多采用随机对照设计, 此时质量评价方法与随机对照试验相同^[1]; 二是评价诊断试验的准确性, 此种最为常见及熟知。

2.1 QUADAS 工具 QUADAS (Quality Assessment of Diagnostic Accuracy Studies) 是由英国约克大学Whiting等遵照Delphi法于2003年制定的专用于系统评价中评价诊断准确性试验质量的工具^[6-7]。QUADAS工具是目前唯一一个经过严格评价和验证的诊断准确性试验质量评价标准, 共14个条目, 每一条目采用“是”、“否”或“不确定”评价 (表2)。2008年, Cochrane协作网推荐QUADAS作为Cochrane诊断性试验准确性系统评价中质量评价的标准, 并根据协作网的筛查和诊断性研究方法学组的意见, 将QUADAS的第3条、第8条和第9条列入非必须评价条目, 故Cochrane诊断性试验准确性系统评价中质量评价的标准最终为11条^[8]。

2.2 Cochrane DTA工作组标准 Cochrane DTA工作组 (Cochrane Diagnostic Test Accuracy Working Group) 除了将上述QUADAS工具作为推荐使用的评价工具外, 还在此基础上针对特定的研究设计制定了附加质量评价条目^[9]。特定的研究主要包括延迟验证 (需要对研究对象进行纵向随访) 研究和诊断比较 (避免选择性偏差和独立的多重测试评估), 可能附加的质量条目如下: ①是否在研究开始前已确定了临界值 (cut-off value)? ②研究开始后试验的技术指标是否未发生改变? ③研究是否提供了阳性结果的清晰定义? ④试验操作者是否接受了适当的培训? ⑤治疗是否在测试指标和参考标准执行后停止? ⑥是否报告了观察者间差异 (observer variation) 的数据, 以及是否在可接受的范围内? ⑦是否报告了仪器间差异 (instrument variation) 的数

表1 MINORS评价条目

序号	条目	提示
1	明确的给出了研究目的	所定义的问题应该是精确的且与可获得文献有关
2	纳入患者的连贯性	所有具有潜在可能性的患者（满足纳入标准）都在研究期间被纳入了（无排除或给出了排除的理由）
3	预期数据的收集	收集了根据研究开始前制定的研究方案中设定的数据
4	终点指标能恰当的反映研究目的	明确的解释用来评价与所定义的问题一致的结局指标的标准。同时，应在意向性治疗分析的基础上对终点指标进行评估
5	终点指标评价的客观性	对客观终点指标的评价采用评价者单盲法，对主观终点指标的评价采用评价者双盲法。否则，应给出未行盲法评价的理由
6	随访时间是否充足	随访时间应足够长，以使得能对终点指标及可能的不良事件进行评估
7	失访率低于5%	应对所有的患者进行随访。否则，失访的比例不能超过反映主要终点指标的患者比例
8	是否估算了样本量	根据预期结局事件的发生率，计算了可检测出不同研究结局的样本量及其95%可信区间；且提供的信息能够从显著统计学差异及估算把握度水平对预期结果与实际结果进行比较
9~12条用于评价有对照组的研究的附加标准		
9	对照组的选择是否恰当	对于诊断性试验，应为诊断的“金标准”；对于治疗干预性试验，应是能从已发表研究中获取的最佳干预措施
10	对照组是否同步	对照组与试验组应该是同期进行的（非历史对照）
11	组间基线是否可比	不同于研究终点，对照组与试验组起点的基线标准应该具有相似性。没有可能导致使结果解释产生偏倚的混杂因素
12	统计分析是否恰当	用于计算可信区间或相对危险度（RR）的统计资料是否与研究类型相匹配

表2 QUADAS条目

序号	条目	针对点
1	病例谱是否包含了各种病例及易混淆的疾病病例?	疾病谱组成
2	研究对象的选择标准是否明确?	选择标准
3*	金标准是否能准确区分有病、无病状态?	金标准
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?	疾病进展偏倚
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?	部分参照偏倚
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?	多重参照偏倚
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）?	混合偏倚
8*	待评价试验的操作是否描述的足够清楚且可进行重复?	待评价试验的实施
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?	金标准的实施
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?	试验解读偏倚
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?	金标准解读偏倚
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?	临床解读偏倚
13	是否报告了难以解释/中间试验结果?	难以解释的试验结果
14	对退出研究的病例是否进行解释?	退出病例

注：*：Cochrane协作网的筛查和诊断性研究方法学组（DTA）建议的非必需条目^[8]

据，以及是否在可接受的范围内？⑧目标是否是先行指明
的？⑨研究是否未受商业资助？

2.3 CASP清单 英国牛津循证医学中心文献严格评价项目（critical appraisal skill program, CASP, 2004）制定了多种规范，除了前述介绍的针对随机对照试验^[1]和分析性研究^[2]的质量评价工具之外，还制定了针对诊断性试验的质量评价清单^[10]（表3）。对于诊断试验，CASP采用了12个条目，仍然分为3个部分，1~2条为筛选问题，3~12条为细节问题，1~6和9~11均采用“是”、“否”及“不知道”判定。

2.4 STARD声明 STARD（the Standards for Reporting of Diagnostic Accuracy）声明^[11]是由荷兰阿姆斯特丹大学的Bossuyt等组成的诊断准确性研究报告标准（STARD）筹委会在2000年9月荷兰阿姆斯特丹举行的共识会议上，为了改进诊断准确性研究报告质量而形成的一个由25项条目组成的清单。如前所述，STARD也可以用来评价诊断性研究的质量，本处不做详细列出。

3 动物实验

3.1 STAIR清单 STAIR（the initial Stroke Therapy Academic Industry Roundtable）清单最早发表于1999年，由美国马萨诸塞大学医学院的Fisher M领头成立的STAIR小组制定^[12]。2009年，小组对STAIR进行了更新，制定了美国“确保高质量科学研究的推荐意见（Recommendations for Ensuring Good Scientific Inquiry）”，其清单的7条内容可以作为评价动物实验质量的标准^[13]：①样本量计算，②纳入与排除标准，③随机序列产生，④隐藏实验动物分组方案，⑤报道将动物排除分析的原因，⑥结局的盲法评价，⑦声明潜在的利益冲突及研究资助。

3.2 CAMARADES清单 CAMARADES（Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Stroke）清单是目前缺血性卒中动物试验Meta分析中最常用的质量评价清单，其官方网站为<http://www.camarades.info/>，评价内容包括10条，其他动物实验的评价标准多数是在此标准基础上修订的^[14,15]（表4）。

表3 CASP清单

条目	描述
第一部分 研究结果可靠吗?	
1.研究是否定义了一个清晰的问题?	问题需包括以下信息:受试者,检测方法,试验环境,结局指标
2.是否与合适的(参考试验作为金标准)进行对比?	参考试验是当前存在的最佳诊断方法吗?
3.是否对所有的受试者均进行了待评价试验与参考试验的检测?(参照偏倚)	所有患者均客观的接受了待评价试验与参考试验的检测;核查2×2表格
4.目标结果是否受到了参考试验结果的影响?(解读偏倚)	是否盲法评价;待评价试验是否独立进行的
5.是否清晰描述了待检测人群的疾病状态?(疾病谱)	主要症状;疾病分期或严重程度;共患病;鉴别诊断
6.是否提供了进行待检测试验方法的足够的细节信息?	是否严格遵照试验方案?
第二部分 研究结果是什么?	
7.研究结果是什么?	考虑:是否报告了敏感度、特异度、和/或似然比;结果是否通过一种我们可以算出的方式呈现
8.我们对结果的确信度如何?	考虑:它们可能是偶然发生的吗;可信区间是多少;它们是什么
第三部分 研究结果的适用性如何?	
9.研究结果能否用于你的目标患者/人群?	你认为你的目标患者/人群与研究中受试对象如此的不同,以至于不能应用吗?如年龄、性别、种族和疾病谱
10.试验能够用于你的目标患者/人群吗?	考虑来源和机会成本;对试验的期望要求的级别和可用性;当前临床实践及服务的可行性
11.是否所有的结局对目标个体/人群都重要吗?	应用试验结果是否能够改善患者健康;应用试验结果是否能够导致患者管理的改变
12.应用试验还能够对你的目标患者/人群产生哪些影响?	-

表4 CAMARADES清单

序号	条目
1	样本量计算
2	随机序列的产生
3*	盲法缺血诱导
4	结果的盲法评估
5	适当的动物模型
6*	应用无明显内在神经保护活性的麻醉药
7*	有温度控制的说明
8	遵守动物保护法
9	论文经同行评审后发表
10	声明潜在利益冲突

注: *表示此条目在其它动物实验质量评价中可相应的调整

3.3 ARRIV指南 ARRIV (Animals in Research: Reporting In Vivo Experiments) 指南是在国际实验动物3R中心 (National Centre for the Replacement, Refinement and Reduction of Animals in Research, NC3Rs) 的资助下, 由Kilkenny领头制定的动物实验研究报告指南^[16]。ARRIV是在充分借鉴CONSORT声明^[17]的基础上, 结合动物实验的特殊性制定的, 共包括6大部分20个条目。其不仅可以用来作为报告动物实验的规范, 还可以用来评价动物实验的质量, 此处亦不做详细介绍。

4 小结

在实际中, 常提的非随机化研究包括了非随机实验性研究及观察性研究^[18], 因此, 常产生将观察性研究的质量评价标准应用于评价非随机实验性研究的情况 (多为NOS)。我们认为, 这是不妥当的。针对非随机实验性研究, 我们推荐使用MINORS条目评价质量, 因为其包括了有对照组和无对照组两部分, 且可通过条目10判定对照组的设置类型。对于诊断准确性试验, 早期推荐使用的还有贝叶斯图书馆 (Bayes Library) 评价指标^[19], 此标准较为繁琐, 且随着贝叶斯图书馆合并到Cochrane图书馆中及其成员成为Cochrane DTA工作组的一部分, 后期未再更新及推荐; 现Cochrane协作网等推荐使用的是QUADAS工具, Cochrane

DTA工作组并根据特定的研究需求制定了补充的标准^[9]; CASP在动物实验方面, 尚无得到公认的质量评价标准, Cochrane协作网亦无相关推荐及工具, 当前应用较多的是Macleod等^[14]自行开发并经多次修订完善的CAMARADES清单, 使用者可以根据研究实际进行调整。

参考文献

- [1] 曾宪涛, 包翠萍, 曹世义, 等. Meta分析系列之三: 随机对照试验的质量评价工具[J]. 中国循证心血管医学杂志, 2012, 4(3): 183-5.
- [2] 曾宪涛, 刘慧, 陈曦, 等. Meta分析系列之四: 观察性研究的质量评价工具[J]. 中国循证心血管医学杂志, 2012, 4(4): 297-9.
- [3] Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (minors): development and validation of a new instrument[J]. ANZ J Surg, 2003, 73(9): 712-6.
- [4] Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies[J]. Pediatrics, 1989, 84(5): 815-27.
- [5] Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement[J]. Am J Public Health, 2004, 94(3): 361-6.
- [6] 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.
- [7] Whiting P, Rutjes AW, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies[J]. Health Technol Assess, 2004, 8(25): iii, 1-234.
- [8] Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011[J]. <http://www.cochrane-handbook.org>.
- [9] Deeks JJ, Bossuyt PM, Gatsonis C. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2009[J]. <http://srdta.cochrane.org>.
- [10] CASP (Critical Skills Appraisal Programme) [EB/OL]. [2012-10-01]. http://www.casp-uk.net/wp-content/uploads/2011/11/CASP_Diagnostic_Appraisal_Checklist_14oct10.pdf.
- [11] Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration [J]. Ann Intern Med, 2003, 138(1): W1-12.

- [12] Stroke therapy academic industry roundtable(Fisher M,Chair). Recommendations for standards regarding preclinical neuroprotective and restorative drug development [J]. Stroke,1999,30(12):2752-8.
- [13] Fisher M,Feuerstein G,Howells DW,*et al.* Update of the stroke therapy academic industry roundtable preclinical recommendations [J]. Stroke,2009, 40(6):2244-50.
- [14] Macleod MR,O'Collins T,Howells DW,*et al.* Pooling of animal experimental data reveals influence of study design and publication bias [J]. Stroke,2004,35(5):1203-8.
- [15] Crossley NA,Sena E,Goehler J,*et al.* Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach [J]. Stroke,2008,39(3):929-34.
- [16] Kilkenny C,Browne WJ,Cuthill IC,*et al.* Improving bioscience research reporting:the ARRIVE guidelines for reporting animal research [J]. PLoS Biol,2010,8(6):e1000412.
- [17] Schulz KF,Altman DG,Moher D,*et al.* CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials [J]. Int J Surg,2011,9(8):672-7.
- [18] Britton A,McKee M,Black N,*et al.* Choosing between randomised and non-randomised studies:a systematic review [J]. Health Technol Assess,1998,2(13):i-iv,1-124.
- [19] Battaglia M,Bucher H,Egger M,*et al.* The Bayes Library of Diagnostic Studies and Reviews [M]. 2nd edition,2002.

(收稿日期: 2012-09-12; 修回日期: 2012-11-10)

· 循证视窗 ·

心肌肌钙蛋白T可预测结构性心脏病结局

《美国医学会杂志》(JAMA)在2010年12月8日发表了一项关于心肌肌钙蛋白T(cTnT)预测慢性疾病价值的研究。该研究显示,随着高敏感性检测方法的应用,现已证明cTnT水平的升高与多种慢性疾病有关,包括冠状动脉疾病(CAD)、心衰及慢性肾病(CKD)等[JAMA, 2012, 308(22): 2317-22]。

研究发现多种慢性疾病都可导致cTnT浓度的增加,可在数种慢性疾病状态中被检测到,包括冠状动脉疾病(CAD)、心衰及慢性肾病(CKD)。根据文章的背景资料,某些研究提示,肌钙蛋白可能对发现亚临床的心血管疾病及在一般人群中评估心血管疾病的风险有用;然而,标准化检测的低敏感度会限制人们对肌钙蛋白测定的临床应用。而最近,一种对cTnT的高敏感检测方法已经研发出

来,它可以检测到比标准测试所能探测到的浓度低大约10倍的cTnT;通过此种方法在罹患慢性心衰和慢性冠心病的患者均可检测到循环血液中的cTnT,而较高浓度的cTnT与心血管死亡率的增加有着很强的关联性。研究人员在3546名有着多种族裔背景的年龄30~65岁的人中用标准测试法及高敏感性测试法对心肌肌钙蛋白T进行了检测,结果发现,在达拉斯县的成年人中,高敏感性检测的检测结果显示cTnT的流行率为25%,而在标准检测中,其流行率仅为0.7%,而且男性cTnT阳性的机会要比女性高3倍(37.1% vs. 12.9%),此外还与种族和年龄有关。特别是对于疑似心肌梗死表现的患者中,高敏感性检测可提高诊断的敏感性,特别是在症状出现的早期,有助于临床早期确诊高度疑似心肌梗死的患者。

(闫睿编译)

读者·作者·编者

本刊启用科技期刊学术不端文献检测系统公告

学术不端行为是指违反学术规范、学术道德的行为,国际上一般用来指捏造数据(fabrication)、篡改数据(falsification)和剽窃(plagiarism)三种行为。为了提高稿件质量,防止抄袭、伪造、剽窃、一稿多投等学术不端行为的发生,本刊现已启用“科技期刊学术不端文献检测系统”(简称AMLC),该系统是一个以《中国学术文献网络出版总库》和大量国际学术文献为全文对比资源,辅助检测抄袭、不当署名、伪造、篡改学术不端行为的智能系统,对检测出有严重不端行为的稿件,编辑部将一律退稿。

AMLC系统由中国知识资源总库所收录的数千万条中

文文献、数百万条英文文献支持,通过将检测文章与数据库内的文献进行比对,不仅可以检测文献总的文字复制比例,还可详细列出检测文献中每一段雷同文字的详细出处,并准确定位每一段文字的具体位置,并能给出一个完整的重复率比对报告。

因此,希望广大作者在撰写论文时,一定要本着实事求是的科学精神,自觉抵制学术不端行为,引用他人的科研成果务必标引参考文献。本刊希望借助此工具,与广大专家、读者、作者一道,共同遏制学术不端之风,构建公平公正的学术交流平台,营造健康的学术环境。