

## · 循证理论与实践 ·

## Meta分析系列之四：观察性研究的质量评价工具

曾宪涛, 刘慧, 陈曦, 冷卫东

[中图分类号] R4

[文献标志码] A

[文章编号] 1674-4055(2012)04-0297-03

观察性研究 (observational study) 又称非实验性研究 (non-experimental study), 是指没有加入研究人员的任何干预 (试验的或其他方面) 措施, 允许事件自然发展的研究过程<sup>[1]</sup>。医学研究中有很大部分是观察性的研究, 多用于评估教育项目和研究可能造成疾病或损害的危险因素, 由于人的内在特点或实施条件涉及医学伦理等原因, 这类研究通常不能实现随机化。因此, 与随机对照试验相比, 观察性研究更容易受到偏倚风险影响, 发生选择性偏倚的风险大于实验性研究。

由于观察性研究能够提供许多极为重要的信息, 并在医学研究中占很大的比例, 因此, 观察性研究的Meta分析具有重要意义<sup>[2]</sup>: ①对观察性研究进行定量分析, 以确定某因素是否为危险因素; ②提供更准确的效应量估计和增加统计效能, 分析量效关系; ③分析不同研究间的异质性; ④总结各单个研究的结果; ⑤研究稀有暴露因素及其相互作用; ⑥研究稀少疾病的危险因素。近年来观察性研究的Meta分析大量涌现, 并呈逐年增多的趋势, 已成为流行病学研究中的重要部分。

在上一讲中, 我们对观察性研究所包含的类型及随机对照试验的质量评价工具进行了介绍<sup>[3]</sup>, 本节将对观察性研究的质量评价工具做一简介 (诊断性研究属于较为特殊的观察性研究, 将在下一讲中介绍)。

### 1 NOS量表

纽卡斯尔-渥太华量表 (the Newcastle-Ottawa Scale, NOS) 适用于评价病例对照研究和队列研究<sup>[4-5]</sup>。它通过三大块共8个条目的方法评价队列研究 (表1)<sup>[6]</sup>和病例对照研究 (表2)<sup>[7]</sup>, 具体包括研究人群选择 (selection)、可比性 (comparability)、暴露 (exposure) 评价或结果 (outcome) 评价。NOS对文献质量的评价采用了星级系统的半量化原则, 满分为9颗星。NOS有自己专用的网站 ([http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.asp](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp)), 提供了量表的word及pdf版本, 可免费下载。

### 2 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) 除了制定了针对随机对照试验的质量评价清单<sup>[3]</sup>外, 也制定了针对观察性研究的评价清单<sup>[8]</sup>, 主要包括队列研究与病例-对照研究。CASP清单用于评价队列研究的清单包括12个问题, 其中

前2条是筛选问题, 后10条是细节问题; 1~7和10~12条均用“是”、“否”及“不知道”判定 (表3)<sup>[9]</sup>。用于评价病例-对照研究的清单包括11个问题, 其中前2条是筛选问题, 后9条是细节问题; 1~6和9~11条均用“是”、“否”及“不知道”判定 (表4)<sup>[10]</sup>。

### 3 AHRQ横断面研究评价标准

美国卫生保健质量和研究机构 (Agency for Healthcare Research and Quality, AHRQ) 对观察性研究的质量评价标准进行了推荐, 其中推荐NOS量表作为评价队列研究和病例-对照研究的标准; 推荐评价横断面研究 (cross-sectional study) 的标准包括11个条目, 分别用“是”、“否”及“不清楚”作答<sup>[11]</sup>: (1) 是否明确了资料的来源 (调查, 文献回顾)? (2) 是否列出了暴露组和非暴露组 (病例和对照) 的纳入及排除标准或参考以往的出版物? (3) 是否给出了鉴别患者的时间阶段? (4) 如果不是人群来源的话, 研究对象是否连续? (5) 评价者的主观因素是否掩盖了研究对象其他方面情况? (6) 描述了任何为保证质量而进行的评估 (如对主要结局指标的检测/再检测); (7) 解释了排除分析的任何患者的理由; (8) 描述了如何评价和 (或) 控制混杂因素的措施; (9) 如果可能, 解释了分析中是如何处理丢失数据的; (10) 总结了患者的应答率及数据收集的完整性; (11) 如果有随访, 查明预期的患者不完整数据所占的百分比或随访结果。

### 4 STROBE及STREGA声明

STROBE (the Strengthening the Reporting of Observational Studies in Epidemiology) 声明<sup>[12]</sup>是用来规范观察性研究报告格式的, 包括队列研究、病例-对照研究及横断面研究, 自然也可以用来评估这些类型研究的质量; 对于分子流行病学观察性研究, 亦可以采用其报告规范, STROBE的扩展版 (STROBE-ME)<sup>[13]</sup>。STREGA (Strengthening the Reporting of Genetic Association Studies) 声明<sup>[14]</sup>也属于STROBE的扩展版, 是专为规范遗传学关联研究的报告规范, 亦可以用来评估遗传关联性研究的质量。因其不是专用于评价观察性研究的质量, 故在此不详细列出。

### 5 结语

常用于meta分析的观察性研究包括队列研究、病例-对照研究及横断面研究三大类, 因这三类研究受偏倚影响的程度依次增加, 目前尚无一种通用的评价工具。NOS的制作很好地结合了病例对照研究和队列研究的实际, 从随机对照试验评价方法中得以引申, 能较好地应用于非随机对照试验的系统评价, 已被Cochrane协作网的非随机研究方法学组用于培训中并推荐使用。CASP目前应用最多的国家是英国及加拿大。AHRQ横断面研究评价标准较为客观。我们建议选择NOS评价队列研究和病例-对照研究; 选择AHRQ横断面研究评价标准评价横断面研究; 在临床循证时针对单个队列研究或病例-对照研究则推荐选择CASP清单, 因为它

基金项目: 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, 冷卫东); 浙江大学医学部公共卫生学院流行病学与卫生统计学教研室 (刘慧); 武汉大学口腔医学院口腔预防科 (陈曦)

通讯作者: 冷卫东, E-mail: lengtaihe@163.com.

doi:10.3969/j.1674-4055.2012.04.004

表1 队列研究的NOS评价标准<sup>[6]</sup>

栏目	条目 <sup>#</sup>	评价标准
研究人群选择	暴露组的代表性如何(1分)	①真正代表人群中暴露组的特征 <sup>*</sup> ; ②一定程度上代表了人群中暴露组的特征 <sup>*</sup> ; ③选择某类人群, 如护士、志愿者; ④未描述暴露组来源情况
	非暴露组的选择方法(1分)	①与暴露组来自同一人群 <sup>*</sup> ; ②与暴露组来自不同人群; ③未描述非暴露组来源情况
	暴露因素的确定方法(1分)	①固定的档案记录(如外科手术记录) <sup>*</sup> ; ②采用结构式访谈 <sup>*</sup> ; ③研究对象自己写的报告; ④未描述
	确定研究起始时尚无要观察的结局指标(1分)	①是 <sup>*</sup> ; ②否
组间可比性	设计和统计分析时考虑暴露组和未暴露组的可比性(2分)	①研究控制了最重要的混杂因素 <sup>*</sup> ; ②研究控制了任何其他的混杂因素 <sup>*</sup> (此条可以进行修改用以说明特定控制第二重要因素)
结果测量	研究对于结果的评价是否充分(1分)	①盲法独立评价; ②有档案记录; ③自我报告; ④未描述
	结果发生后随访是否足够长(1分)	①是(评价前规定恰当的随访时间) <sup>*</sup> ; ②否
	暴露组和非暴露组的随访是否充分(1分)	①随访完整 <sup>*</sup> ; ②有少量研究对象失访但不至于引入偏倚(规定失访率或描述失访情况) <sup>*</sup> ; ③有失访(规定失访率)但未行描述; ④未描述随访情况

注: #: 给分条目; \*: 给分点

表2 病例-对照研究的NOS评价标准<sup>[7]</sup>

栏目	条目 <sup>#</sup>	评价标准
研究人群选择	病例确定是否恰当(1分)	①恰当, 有独立的确定方法或人员 <sup>*</sup> ; ②恰当, 如基于档案记录或自我报告; ③未描述
	病例的代表性(1分)	①连续或有代表性的系列病例 <sup>*</sup> ; ②有潜在选择偏倚或未描述
	对照的选择(1分)	①与病例同一人群的对照 <sup>*</sup> ; ②与病例同一人群的住院人员为对照; ③未描述
组间可比性	对照的确定(1分)	①无目标疾病史(端点) <sup>*</sup> ; ②未描述来源
	设计和统计分析时考虑病例和对照的可比性(2分)	①研究控制了最重要的混杂因素 <sup>*</sup> ; ②研究控制了任何其他的混杂因素 <sup>*</sup> (此条可以进行修改用以说明特定控制第二重要因素)
暴露因素的测量	暴露因素的确定(1分)	①固定的档案记录(如外科手术记录) <sup>*</sup> ; ②采用结构式访谈且不知访谈者是病例或对照 <sup>*</sup> ; ③采用未实施盲法的访谈(即知道病例或对照的情况); ④未描述
	采用相同的方法确定病例和对照组暴露因素(1分)	①是 <sup>*</sup> ; ②否
	无应答率(1分)	①病例和对照组无应答率相同 <sup>*</sup> ; ②描述了无应答者的情况; ③病例和对照组无应答率不同且未描述

注: #: 给分条目; \*: 给分点

表3 评价队列研究质量的CASP清单<sup>[8]</sup>

条目	提示
第一部分 研究结果可靠吗?	
1. 研究是否提出了清晰明确的问题?	①研究的人群; ②研究的危险因素; ③可能的结局; ④可能的有益或有害的效应
2. 回答问题的方式是否合适?	①队列研究适合研究目的吗? ②队列研究能解决问题吗?
3. 队列研究人群的选择方式合适吗?	①是否可以代表研究的人群? ②样本人群有什么特别的特征吗? ③是否包含了所有应纳入的人群?
4. 是否准确的测量暴露因素以减少偏倚?	①使用的是主观还是客观的测量方法? ②测量结果的真实性如何(是否是被验证的)? ③测量方式是一样的吗?
5. 是否精确测量了研究结果以减少偏倚?	①使用的是主观还是客观的测量方法? ②测量结果的真实性如何(是否是被验证的)? ③有无可靠的系统方法来探查所有的病例(测量疾病的发生)? ④不同组的诊断方式是否相似? ⑤是否对研究对象及结果评价者采取盲法?
6. A 作者考虑到所有重要的混杂因素了吗?(列出作者忽略但您考虑到的因素); B 在设计和/或分析中对混杂因素采取措施了吗?	在设计阶段的严格控制; 在分析阶段使用技术手段如建模、分层、回归、敏感性分析来纠正、控制、调整混杂因素
7. A 对研究对象的随访是否完成? B 随访时间是否足够长?	①不管效应的好坏, 应该有足够的时间来显露; ②失访的人群可能具有不同的结局; ③在开放或动态队列中, 对于离开和加入队列的研究对象有无特殊要求?
第二部分 研究结果是什么?	
8. 研究结果如何?	①基线的结果? ②是否报道暴露组和非暴露组的比例或比率? 两者有区别吗? ③暴露因素与结局的关联强度如何(RR值为多少)? ④绝对危险度降低值(ARR值)是多少?
9. 研究结果的精确度如何?	可信区间是多少?
10. 结果是否可信?	①无法忽略的大效应量; ②有无偏倚、机遇或混杂因素的影响? ③研究的设计和方法是否有缺陷导致结果不可靠? ④考虑Bradford Hills标准(时间序列、剂量-效应梯度、生物学相似性、一致性)
第三部分 研究结果适用吗?	
11. 试验结果能否适用于当地人群?	①纳入试验的研究人群是否与你所研究的人群相似? ②当地的环境和研究中的是否相似? ③能否量化对当地人群的有益和有害效应?
12. 研究结果与其它证据是否符合?	--

表4 评价病例-对照研究质量的CASP清单<sup>[10]</sup>

条目	提示
第一部分 研究结果可靠吗?	
1. 研究是否提出了清晰明确的问题?	①研究的人群; ②研究的危险因素; ③研究是为了检测有益或有害的效应?
2. 回答问题的方式是否合适?	①在目前的情况下, 病例-对照研究是否符合研究目的(结局是否罕见或有害)? ②病例对照研究能否解决研究问题?
3. 病例的选择方法是否合适?	①是否准确的定义了病例? ②病例组具有代表性吗(地理学上的和/或暂时的)? ③有无建立可靠的系统来选择病例? ④是研究发病率还是患病率? ⑤病例组有无特殊特征? ⑥研究时间范围是否与疾病/暴露有关? ⑦样本量充足吗? ⑧计算把握度了吗?
4. 对照组的选择方式是否合适?	①对照组具有代表性吗(人口学上和/或暂时的)? ②对照组有无特殊特征? ③应答率高吗? 不应答的人群是否具有不同特征? ④使用匹配选择、人群来源还是随机选择? ⑤样本量充足吗?
5. 是否准确测量暴露因素以减少偏倚?	①暴露因素是否有明确的定义? 测量方法是否准确? ②研究者使用的是主观还是客观的测量方法? ③测量方法的真实性如何(是否是被验证的)? ④病例组和对照组使用的测量方法是否相似? ⑤在适合使用盲法的地方是否使用了盲法? ⑥时间顺序正确吗(研究的暴露因素是否在结局前)?
6. A: 作者考虑了哪些混杂因素吗? (列出作者忽略但您考虑到的因素, 如基因、环境及社会经济的); B: 在设计和/或分析中, 研究者对潜在混杂因素采取措施了吗?	在设计阶段的严格控制; 在分析阶段使用技术手段如建模、分层、回归、敏感性分析来纠正、控制、调整混杂因素
7. 研究结果如何?	①基线的结果? ②分析方法合适吗? ③暴露因素与结局的关联强度如何(OR值为多少)? ④调整混杂因素后, 混杂因素是否还起作用? ⑤调整混杂因素是否对OR值有很大的影响?
第二部分 研究结果是什么?	
8. 研究结果的精确度如何? 危险效应的估计值精确度如何?	①P值是多少? ②可信区间是多少? ③研究者是否考虑所有重要的变量? ④如何评估排除的人群的研究效应?
9. 结果是否可信?	①无法忽略的大效应量; ②有无偏倚、机遇或混杂因素的影响? ③研究的设计和方法是否有缺陷导致结果不可靠? ④考虑Bradford Hills标准(时间序列、剂量-效应梯度、生物学相似性、一致性)
第三部分 研究结果适用吗?	
10. 试验结果能否适用于当地人群?	①纳入试验的研究人群是否与你所研究的人群相似? ②当地的环境和研究中的是否相似? ③能否量化对当地人群的有益和有害效应?
11. 研究结果与其它证据是否符合?	考虑所有可得到的、来自随机对照试验、系统评价、队列研究及病例-对照研究的一致性较好的证据

考虑了研究的外部适用性。

### 参考文献

- Grimes DA, Schulz KF. An overview of clinical research: the lay of the land[J]. Lancet, 2002, 359(9300):57-61.
- Egger M, Smith GD, Altman DG. Systematic reviews in health care. Meta-analysis in context [M]. 2<sup>nd</sup> ed. London: BMJ Publishing Group, BMA House, Tavistock Square, WC1H 9JR, 2001.
- 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.
- Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses[J]. Eur J Epidemiol, 2010, 25(9):603-5.
- Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses [EB/OL]. [2012-06-15]. [http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.htm](http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm).
- Wells G, Shea B, O'Connell D, et al. Newcastle-Ottawa Quality Assessment Scale -- Cohort Studies [EB/OL]. [2012-06-15]. [http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.asp](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp).
- Wells G, Shea B, O'Connell D, et al. Newcastle-Ottawa Quality Assessment Scale -- Case Control Studies [EB/OL]. [2012-06-15]. [http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.asp](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp).
- Ibbotson T, Grimshaw J, Grant A. Evaluation of a programme of workshops for promoting the teaching of critical appraisal skills [J]. Med Educ, 1998, 32(5):486-91.
- CASP(Critical Skills Appraisal Programme)[EB/OL]. [2012-06-15]. [http://www.casp-uk.net/wp-content/uploads/2011/11/CASP\\_Cohort\\_Appraisal\\_Checklist\\_14oct10.pdf](http://www.casp-uk.net/wp-content/uploads/2011/11/CASP_Cohort_Appraisal_Checklist_14oct10.pdf).
- CASP(Critical Skills Appraisal Programme)[EB/OL]. [2012-06-15]. [http://www.casp-uk.net/wp-content/uploads/2011/11/CASP\\_Case-Control\\_Appraisal\\_Checklist\\_14oct10.pdf](http://www.casp-uk.net/wp-content/uploads/2011/11/CASP_Case-Control_Appraisal_Checklist_14oct10.pdf).
- Rostom A, Dube C, Cranney A, et al. Celiac Disease. Rockville (MD): Agency for Healthcare Research and Quality (US); 2004 Sep. (Evidence Reports/Technology Assessments, No. 104.) Appendix D. Quality Assessment Forms. <http://www.ncbi.nlm.nih.gov/books/NBK35156>.
- Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration [J]. PLoS Med, 2007, 4(10):e297.
- Gallo V, Egger M, McCormack V, et al. Strengthening the Reporting of Observational Studies in Epidemiology -- Molecular Epidemiology (STROBE-ME): an extension of the STROBE Statement [J]. PLoS Med, 2011, 8(10):e1001117.
- Little J, Higgins JP, Ioannidis JP, et al. Strengthening the Reporting of Genetic Association Studies (STREGA): an extension of the STROBE statement [J]. PLoS Med, 2009, 6(2):e22.

(收稿日期:2012-06-06;修回日期:2012-07-10)