Eur J Epidemiol (2010) 25:603–605
DOI 10.1007/s10654-010-9491-z

COMMENTARY

Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses

Andreas Stang

Received: 2 December 2009 / Accepted: 8 July 2010 / Published online: 22 July 2010

The quality assessment of non-randomized studies is an important component of a thorough meta-analysis of non-randomized studies. Low quality studies can lead to a distortion of the summary effect estimate. Recent guidelines for the reporting of meta-analyses of observational studies recommend the assessment of the study quality (MOOSE) [1]. In principal, three categories of quality assessments tools are available: scales, simple checklists, or checklists with a summary judgment (for details see Sanderson et al. 2007 [2]). The results of the quality assessment can be used in several ways such as forming inclusion criteria for the meta-analysis, informing a sensitivity analysis or meta-regression, weighting studies, or highlighting areas of methodological quality poorly addressed by the included studies [3]. It has been criticized that the use of summary scores involve inherent weighting of component items including items that may not be related to the validity of the study findings [2].

Sanderson et al. [2] recently identified overall 86 tools for assessing the quality of non-randomized studies. Their review "highlighted the lack of a single obvious candidate tool for assessing quality of observational epidemiological studies" [2]. In the field of randomized trials, it has been shown that the choice of quality scale can dramatically influence the interpretation of meta-analyses, and can even reverse conclusions regarding the effectiveness of an intervention [4].

Wells et al. [5] proposed a scale for assessing the quality of published non-randomized studies in meta-analyses, called the Newcastle-Ottawa-Scale (NOS). This tool can either be used as a checklist or scale. The NOS was developed using a Delphi process and thereafter was tested on systematic reviews and further refined. Separate NOS scales were developed for cohort and case-control studies. The NOS contains eight items, categorized into three dimensions including selection, comparability, and—depending on the study type—outcome (cohort studies) or exposure (case-control studies). For each item a series of response options is provided. A star system is used to allow a semi-quantitative assessment of study quality, such that the highest quality studies are awarded a maximum of one star for each item with the exception of the item related to comparability that allows the assignment of two stars. The NOS ranges between zero up to nine stars.

To my knowledge, the NOS scales have not been published in peer-reviewed journals so far. The only reference I could find is a web-based link [5]. A Medline recherché ("Newcastle-Ottawa-Scale"[All Fields], November 10, 2009) showed that overall 14 articles used the term "Newcastle-Ottawa-Scale" in their abstracts (references are available on request) and used the NOS scale for assessing the quality of published non-randomized studies in meta-analyses. All articles quoted the web-link of the Ottawa Health Research Institute [5]. All articles presented the results of meta-analyses and were published between 2004 and 2009.

Although the authors of NOS stated that the validity assessment of the scale is under development, Li et al. [6] who used this scale in a meta-analysis recently remarked that "*The scale has been shown to be reliable and valid*". I believe that the NOS includes problematic items with an uncertain validity. Previously, Deeks et al. [3] concluded that with a few caveats (missing item for the appropriateness of the analysis, lack of information related to the

A. Stang (✉)
Institut für Klinische Epidemiologie, Medizinische Fakultät,
Martin-Luther-Universität Halle-Wittenberg,
Magdeburger Str. 8, 06097 Halle (Saale), Germany
e-mail: andreas.stang@medizin.uni-halle.de

reliability and validity), the NOS is "suitable for use in a systematic review" and is "easy to use". The growing use of the NOS as an apparently established "easy to use" quality score among meta-analysts may be problematic as sometimes far-reaching conclusions are drawn [7]. The aim of this commentary is to critically discuss items of the NOS in depth.

Case-control studies

The NOS defines independent validation of the case status as an assessment by "e.g. >1 person/record/time/process to extract information, or reference to primary record source such as x-rays or medical/hospital records" [5]. Some of these assessments do not address validation as the assessment of the case status by more than one person; it may be used for the measurement of interobserver variability of the case status measurement. Furthermore, the assessment of the case status more than once or by more than one record or data source does not necessarily imply validation. It could imply the measurement of intraobserver reliability. Obviously, the validity item of the NOS is a mixture of validation, intra- and interobserver variability measurement (reliability).

The NOS gives a higher score to population-based than hospital controls. Although many epidemiologists prefer community to hospital controls, epidemiologic methods teach us that there cannot be a general preference of community controls over hospital controls because the study base principle drives the decision to sample controls from hospitals or from communities [8].

Wells et al. state that control for the most important factor by design (matching) or by analysis (adjustment) results in a higher score. The adjustment for a "second important factor" results in a still higher score. An empirical investigation on matching in case-control studies showed that the vast majority of case-control studies in the mid 1990s use matching. Most frequently, matching on age and gender is used to increase the efficiency of the adjustment of confounding of these variables compared to unmatched case-control studies [9]. Therefore, the quality items have little—if any—discriminatory effect as the vast majority of case-control studies are assigned stars on the NOS scale. Furthermore, the meaning of "important factor" is undefined and therefore arbitrary. Confounding is specific to the research question and the importance of confounding (e.g. change in estimate) of the NOS scale remains undefined.

The NOS gives a higher score to studies that had blinded exposure assessment. Blinding is sometimes impossible as the case-control status can be easily discerned due to visual or acoustic signs of the disease (e.g. larynx or pharynx cancer patients with hoarseness or fuzzy speech; patients with a loss of an eye due to enucleation of an uveal melanoma). Therefore, it is important to perform highly standardized interviews undertaken by trained study personnel that is regularly monitored throughout the study. For example, the INTERPHONE case-control study on the risk of mobile phone use and brain cancer did not blind the interviewers. However, the investigators organized highly standardized interviews by trained interviewers [10].

The NOS gives a higher score to case-control studies with comparable nonresponse among cases and controls than case-control studies with different response proportions. This item is in conflict with the concept of valid selection as response proportions have to be identical by exposure status within subgroups of cases and controls in order to minimize nonresponse bias as can be mathematically shown [11]. Table 1 exemplifies that identical response proportions among cases and controls (each group has a response of 50%) do not allow any conclusion about selection bias as long as the specific response proportions of all four cells (exposed and unexposed cases and controls) within the 2-by-2 table are unknown. Although the response proportions are 0.50 in the case and control group in my example, the response proportions of exposed and unexposed controls differ and thus produce a bias of the exposure-disease odds ratio estimate. Identical response proportions of the case and control group is no safeguard against selection bias.

Cohort studies

Three dimensions contribute to the overall quality score including assessment of selection of the exposed and unexposed cohort, comparability of the two cohorts, and outcome assessment.

The NOS assigns a higher score to cohort studies with community representativeness of the exposed cohort. Famous prospective cohort studies like the British Doctors Study, Physician's Health Study, and Nurses Health Study will get a lower quality score as compared to the Framingham Heart Study because the former cohorts are not representative of the general population. In theory, community representativeness of an exposed cohort has the advantage of better generalizibility of the study findings compared to an unrepresentative exposed cohort. However, cohort studies that aim to assemble a representative exposed cohort frequently suffer from low baseline response resulting in a questionable generalizibility of the study findings. Unrepresentative exposed cohorts may have the advantage of a higher baseline response, better exposure assessment and better follow-up response of cohort members that may result in a higher internal validity of the study findings compared to a cohort study with a representative exposed cohort.

Table 1 Identical response proportions of the cases and controls and selection bias

	Cases	Controls	Exposure-disease odds ratio
Complete participation (truth)			
Exposed subjects (N)	500	500	1.0
Unexposed subjects (N)	500	500	1.0 (reference)
Response proportions			
Exposed	0.5	0.4	
Unexposed	0.5	0.6	
Total[a]	0.5	0.5	
Observed distribution of exposed and unexposed subjects by case-control status (given response proportions)			
Exposed subjects (N)	250	200	1.5
Unexposed subjects (N)	250	300	1.0 (reference)

[a] Response proportion of the case and control group (regardless of exposure status); both groups have the same response proportion of 0.50

Similarly as for the case-control studies, Wells et al. state that control for the most important factor by design or by analysis results in a higher score for cohort studies. Again, the meaning of "most important factor" is undefined and therefore highly arbitrary.

The NOS gives the same score to studies that did independent or blind outcome assessment or record linkage (outcome identification through database records). Imagine two studies: one study did a detailed blinded outcome assessment based on all available follow-up documents and included a panel of several experts for the cancers of interest, the other study did a record linkage of all cohort members with routinely available records from a regional population-based cancer registry. The NOS assigns the same quality value to these two studies.

In conclusions, I believe that Wells et al. provide a quality score that has unknown validity at best, or that includes quality items that are even invalid. The current version appears to be inacceptable for the quality ranking of both case-control studies and cohort studies in meta-analyses. The use of this score in evidence-based reviews and meta-analyses may produce highly arbitrary results.

Acknowledgments I am very grateful for many helpful comments by an anonymous reviewer on an earlier version of this manuscript.

References

1. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology (MOOSE) group. JAMA. 2000;283(15):2008–12.
2. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):666–76.
3. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii–173.
4. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282(11):1054–60.
5. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses. Available from: URL: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm [cited 2009 Oct 19].
6. Li W, Ma D, Liu M, Liu H, Feng S, Hao Z, et al. Association between metabolic syndrome and risk of stroke: a meta-analysis of cohort studies. Cerebrovasc Dis. 2008;25(6):539–47.
7. Myung SK, Ju W, McDonnell DD, Lee YJ, Kazinets G, Cheng CT, et al. Mobile phone use and risk of tumors: a meta-analysis. J Clin Oncol. 2009;27:5565–72.
8. Miettinen OS. Theoretical epidemiology. Principles of occurrence research in medicine. Albany, New York: Delmar Publishers Inc; 1985.
9. Gefeller O, Pfahlberg A, Brenner H, Windeler J. An empirical investigation on matching in published case-control studies. Eur J Epidemiol. 1998;14(4):321–5.
10. Schüz J, Böhler E, Berg G, Schlehofer B, Hettinger I, Schlaefer K, et al. Cellular phones, cordless phones, and the risks of glioma and meningioma (Interphone Study Group, Germany). Am J Epidemiol. 2006;163(6):512–20.
11. Austin MA, Criqui MH, Barrett-Connor E, Holdbrook MJ. The effect of response bias on the odds ratio. Am J Epidemiol. 1981;114(1):137–43.