ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 982−993

Journal of Clinical Epidemiology

Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers

Lisa Hartling[a,*], Andrea Milne[a], Michele P. Hamm[a], Ben Vandermeer[a], Mohammed Ansari[b], Alexander Tsertsvadze[c], Donna M. Dryden[a]

[a]*Department of Pediatrics, Alberta Research Centre for Health Evidence and the University of Alberta Evidence-based Practice Center, University of Alberta, 4-472 Edmonton Clinic Health Academy, 11405-87 Avenue, Edmonton, Alberta, Canada T5G 1C9*

[b]*Clinical Epidemiology Program, University of Ottawa Evidence-based Practice Center, The Ottawa Methods Centre, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

[c]*University of Ottawa Evidence-based Practice Center and Centre for Practice-Changing Research, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

Accepted 20 March 2013; Published online 16 May 2013

Abstract

Objectives: To assess inter-rater reliability and validity of the Newcastle Ottawa Scale (NOS) used for methodological quality assessment of cohort studies included in systematic reviews.

Study Design and Setting: Two reviewers independently applied the NOS to 131 cohort studies included in eight meta-analyses. Inter-rater reliability was calculated using kappa (κ) statistics. To assess validity, within each meta-analysis, we generated a ratio of pooled estimates for each quality domain. Using a random-effects model, the ratios of odds ratios for each meta-analysis were combined to give an overall estimate of differences in effect estimates.

Results: Inter-rater reliability varied from substantial for *length of follow-up* (κ = 0.68, 95% confidence interval [CI] = 0.47, 0.89) to poor for *selection of the nonexposed cohort* and *demonstration that the outcome was not present at the outset of the study* (κ = −0.03, 95% CI = −0.06, 0.00; κ = −0.06, 95% CI = −0.20, 0.07). Reliability for overall score was fair (κ = 0.29, 95% CI = 0.10, 0.47). In general, reviewers found the tool difficult to use and the decision rules vague even with additional information provided as part of this study. We found no association between individual items or overall score and effect estimates.

Conclusion: Variable agreement and lack of evidence that the NOS can identify studies with biased results underscore the need for revisions and more detailed guidance for systematic reviewers using the NOS.

Keywords: Methodological quality; Internal validity; Reliability; Validity; Systematic reviews; Cohort studies

1. Introduction

The internal validity of a study reflects the extent to which the design and conduct of the study have minimized the impact of bias [1]. One of the key steps in a systematic review is the assessment of internal validity (or risk of bias, RoB) of all studies included for evidence synthesis. This assessment serves to identify the strengths and limitations of the included studies; investigate and explain heterogeneity of findings across a priori defined subgroups of studies based on RoB; and grade the quality or strength of evidence for a given outcome.

With the increase in the number of published systematic reviews [2] and development of systematic review methodology over the past 15 years [1], close attention has been paid to methods of assessing internal validity of individual primary studies. Until recently, this has been referred to as ''quality assessment'' or ''assessment of methodological quality'' [1]. In this context, ''quality'' refers to ''the confidence that the trial design, conduct, and analysis has minimized biases in its treatment comparisons'' [3]. To facilitate the assessment of methodological quality, a plethora of tools has emerged [3−6]. Some of these tools are applicable to specific study designs, whereas other more generic tools may be applied to more than one design. The tools

Funding disclosure and disclaimer: This manuscript is based on a project conducted by the University of Alberta Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290−2007−10021). The findings and conclusions in this manuscript are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. No statement in this manuscript should be construed as an official position of AHRQ or of the US Department of Health and Human Services.

* Corresponding author. Tel.: 780-492-6124; fax: 780-248-5627.
E-mail address: hartling@ualberta.ca (L. Hartling).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.03.003

What is new?

- Inter-rater reliability between reviewers on the Newcastle Ottawa Scale (NOS) ranged from poor to substantial but was poor or fair for most domains.
- No association was found between individual quality domains or overall quality score and effect estimates.
- These findings underscore the need for revisions and more detailed guidance to apply the NOS in systematic reviews.

usually incorporate items associated with bias (e.g., blinding, baseline comparability of study groups) and items related to reporting (e.g., was the study population described, was a sample size calculation performed) [1].

There is a need for inter-rater reliability testing of quality assessment tools to enhance consistency in their application and interpretation across different systematic reviews. Furthermore, validity testing is essential to ensure that the tools being used can identify studies with biased results. Finally, there is a need to determine inter-rater reliability and validity to support the use of individual tools that are recommended by those developing methods for systematic reviews.

We undertook this project to assess the reliability and validity of the Newcastle Ottawa Scale (NOS). The NOS is a quality assessment tool for use on nonrandomized studies included in systematic reviews, specifically cohort and case−control studies. The tool was produced by the combined efforts of the Universities of Newcastle, Australia, and Ottawa, Canada [7], and was first reported at the Third Symposium for Systematic Reviews in Oxford, United Kingdom, in 2000 [8]. It has been endorsed for use in systematic reviews of nonrandomized studies by The Cochrane Collaboration [1].

The NOS includes separate assessment criteria for case−control and cohort studies covering the following domains: the selection of participants, comparability of study groups, and the ascertainment of exposure (for case−control studies) or outcome of interest (for cohort studies). A star rating system is used to indicate the quality of a study, with a maximum of nine stars [8]. Each criterion receives a single star if appropriate methods have been reported. The selection domain is subdivided to evaluate the selection of the exposed and nonexposed cohorts, the ascertainment of exposure, and whether the study demonstrated that the outcome of interest was not present at the start of the study. Comparability is the only category that may receive two stars: one if the most important confounders have been adjusted for in the analysis and a second star if any other adjustments were made. Outcome of interest is made up of three questions: the appropriateness of the methods used to evaluate the outcome, the length of follow-up, and the degree of the loss to follow-up [7].

The developers of the NOS have examined face and criterion validity, inter-rater reliability, and evaluator burden for the NOS. Face validity has been evaluated as strong by comparing each individual assessment item to their stem question. Criterion validity has shown a strong agreement with the Downs and Black assessment tool [9] on a series of 10 cohort studies evaluating hormone replacement therapy in breast cancer, with an intraclass correlation coefficient (ICC) of 0.88. Inter-rater reliability for the NOS on cohort studies was high with an ICC of 0.94. Evaluator burden, as assessed by the time required to complete the NOS evaluation, was shown to be significantly less than the Downs and Black tool ($P < 0.001$) [10]. The authors state that further assessment of the construct validity and the relationship between the external criterion of the NOS and its internal structures are under consideration [7]. These studies have been presented as abstracts.

The objectives of this study were to further assess the reliability of the NOS for cohort studies between individual raters and assess the validity of the NOS by examining whether effect estimates vary according to quality.

2. Methods

This article is part of a larger technical report conducted for the Agency for Healthcare Research and Quality (AHRQ). We followed a protocol that was developed a priori with input from experts in the field. Further details on methodology and results are available in the technical report (http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/).

2.1. Study selection

We used an iterative approach to identify a sample of cohort studies based on meta-analyses of cohort studies. Our operational definition of a cohort study was one in which individuals are grouped according to exposure status at baseline (exposed or unexposed) and are followed over time to determine if the development of the outcome of interest is different in the exposed and unexposed groups. Data may be collected prospectively or retrospectively. Initially, we searched reports completed through the Evidence-based Practice Center (EPC) Program of AHRQ to identify meta-analyses of cohort studies. We found three EPC reports [11−13] including 36 cohort studies that met the inclusion criteria. We subsequently conducted searches in MEDLINE using search terms to capture systematic reviews (meta-analys?s.mp, review.pt, and search.tw), cohort studies (exp Cohort Studies/, cohort$.tw, (observation$ adj stud$).tw) and meta-analyses (exp meta-analysis/, (analysis adj3 (group$ or pool$)).tw, (forest adj plot$).mp). Results