Reliability and Validity of the Newcastle Ottawa Scale

Lisa Hartling,1 Michele Hamm,1 Andrea Milne,1 Ben Vandermeer,1 Mohammed Ansari,2 Alexander Tsertsivadze,2 Donna Dryden1
1 Alberta Research Centre for Health Evidence, Department of Pediatrics, University of Alberta, Canada
2 The Ottawa Hospital Research Institute, Ottawa, Canada

Background

- The Newcastle Ottawa Scale (NOS) is used to assess methodological quality of cohort and case-control studies in systematic reviews.
- With respect to cohort studies, the NOS developers showed high intra-class correlation (0.94) based on assessments of 10 studies by two raters.
- The tool developers also showed criterion validity by comparing the NOS with another commonly used scale (Downs and Black, intra-class correlation=0.88) based on 10 cohort studies.
- There has been limited independent evaluation of the inter-rater reliability and validity of the NOS tool.

Objectives

1. Assess the reliability of the NOS between individual raters.
2. Assess the reliability of the NOS by examining effect estimates or measures of association across the different NOS items (Table).

Methods

- Two reviewers independently assessed 131 cohort studies from 8 meta-analyses (Table).
- Disagreements were resolved through discussion to produce single assessments for each study.
- Inter-rater agreement was calculated using weighted or unweighted Cohen’s kappa statistic, as appropriate.
- For each meta-analysis we calculated a ratio of odds ratios (ROR) for studies assessed as meeting or not meeting each NOS item.
- The RORs for each meta-analysis were combined to give an overall estimate of differences in effect estimates using meta-analytic techniques with inverse variance weighting and a random effects model.
- The feedback from reviewers in applying the tool and identification of items that are particularly problematic provide valuable information for revisions and more detailed guidance.

Key Messages

- Inter-rater reliability between reviewers on the Newcastle Ottawa Scale 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. This may have been due to low power.
- The results of systematic reviews and their interpretation could be misleading if they are based on unreliable assessments of quality.
- The feedback from reviewers in applying the tool and identification of items that are particularly problematic provide valuable information for revisions and more detailed guidance.

Results

Reliability

- Inter-rater agreement varied from poor to substantial across the different NOS items (Table).
- Interviews with the reviewers provided input on challenges using the tool, for example:
  - Whether to assess each study based on the individual report, or as it related to the systematic review question.
  - Difficulty interpreting some terminology (“selected” population).
  - Unclear distinction between some response options (“truly” vs. “somewhat” representative population; “structured interview” vs. “written self-report”).
  - Lack of clarity about whether selection domain assessed bias regarding how participants were selected or applicability of the study population to the population in general.
  - Uncertainty on how to assess comparability, e.g., some studies discussed testing different confounders but only included confounders that showed a significant difference in the final model.
  - Item about adequacy of followup of cohorts is missing a response option that includes a larger number lost to followup and a description is provided (indicating no imbalance between groups).
  - Reviewers commented that “unclear” or “no description” options were needed for some items.
- No associations were found between individual NOS items or overall NOS score and magnitude of effect estimates (Table).

<table>
<thead>
<tr>
<th>Domain within the Newcastle Ottawa Scale</th>
<th>Agreement, κ (95% CI)</th>
<th>Interpretation (based on Landis and Koch 1977)</th>
<th>Ratio of Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representativeness of the exposed cohort</td>
<td>0.23 (0.09, 0.41)</td>
<td>Fair</td>
<td>1.01 (0.85, 1.20)</td>
</tr>
<tr>
<td>Selection of the non-exposed cohort</td>
<td>-0.03 (-0.06, 0.00)</td>
<td>Poor</td>
<td>1.83 (0.92, 3.64)</td>
</tr>
<tr>
<td>Ascertainment of exposure</td>
<td>0.43 (0.25, 0.61)</td>
<td>Moderate</td>
<td>1.13 (0.93, 1.37)</td>
</tr>
<tr>
<td>Demonstration that the outcome was not present at the outset of study</td>
<td>-0.06 (-0.20, 0.07)</td>
<td>Poor</td>
<td>0.72 (0.49, 1.07)</td>
</tr>
<tr>
<td>Comparability</td>
<td>0.18 (-0.12, 0.47)</td>
<td>Slight</td>
<td>0.86 (0.56, 1.31)</td>
</tr>
<tr>
<td>Assessment of outcome</td>
<td>0.49 (0.28, 0.70)</td>
<td>Moderate</td>
<td>1.04 (0.79, 1.38)</td>
</tr>
<tr>
<td>Length of followup sufficient</td>
<td>0.68 (0.47, 0.89)</td>
<td>Substantial</td>
<td>0.84 (0.55, 1.27)</td>
</tr>
<tr>
<td>Adequacy of participant followup</td>
<td>0.29 (0.12, 0.46)</td>
<td>Fair</td>
<td>0.99 (0.91, 1.08)</td>
</tr>
<tr>
<td>Total stars</td>
<td>0.29 (0.10, 0.47)</td>
<td>Fair</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Total stars