ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 973−981

Journal of Clinical Epidemiology

Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs

Lisa Hartling[a,*], Michele P. Hamm[a], Andrea Milne[a], Ben Vandermeer[a], P. Lina Santaguida[b], Mohammed Ansari[c], Alexander Tsertsvadze[c], Susanne Hempel[d], Paul Shekelle[d], 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]*McMaster Evidence-based Practice Center, McMaster University, Edmonton, Alberta, Canada*

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

[d]*RAND Corporation, 1776 Main Street, Santa Monica, CA 90407, USA*

Accepted 23 July 2012; Published online 13 September 2012

Abstract

Objectives: To assess the reliability of the Cochrane Risk of Bias (ROB) tool between individual raters and across consensus agreements of pairs of reviewers and examine the impact of study-level factors on reliability.

Study Design and Setting: Two reviewers assessed risk of bias for 154 randomized controlled trials (RCTs). For 30 RCTs, two reviewers from each of four centers assessed risk of bias and reached consensus. We assessed interrater agreement using kappas and the impact of study-level factors through subgroup analyses.

Results: Reliability between two reviewers was fair for most domains ($\kappa = 0.24-0.37$), except sequence generation ($\kappa = 0.79$, substantial). Reliability results across reviewer pairs: sequence generation, moderate ($\kappa = 0.60$); allocation concealment and ''other sources of bias,'' fair ($\kappa = 0.37-0.27$); and other domains, slight ($\kappa = 0.05-0.09$). Reliability was influenced by the nature of the outcome, nature of the intervention, study design, trial hypothesis, and funding source. Variability resulted from different interpretation of the tool rather than different information identified in the study reports.

Conclusion: Low agreement has implications for interpreting systematic reviews. These findings suggest the need for detailed guidance in assessing the risk of bias.

Keywords: Risk of bias; Internal validity; Reliability; Systematic reviews; Meta-Analysis; Randomized controlled trials

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) of all studies included for evidence synthesis. This assessment serves to (1) identify the strengths and limitations of individual studies, (2) investigate and explain heterogeneity of findings across a priori defined subgroups of studies based on risk of bias, and (3) 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 the internal validity. 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 or avoided biases in its treatment comparisons'' [3]. To facilitate the assessment of methodological quality, a plethora of tools has emerged [3−6]. Although some of these tools are applicable to specific study

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 U.S. 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.2012.07.005

What is new?

- Interrater reliability between two reviewers applying the Risk of Bias (ROB) tool was low.
- This is the first study to examine interrater reliability for the ROB tool between consensus assessments across pairs of reviewers, and reliability was low.
- Reliability was influenced by study-level variables, including the nature of the outcome, nature of the intervention, study design, trial hypothesis, and funding source.
- Variability resulted more often from different interpretation of the tool rather than different information identified in the study reports.
- The findings provide direction for more detailed guidance in applying the ROB tool, which is essential to ensure appropriate interpretation of the evidence by decision makers who use systematic reviews.

designs, other more generic tools may be applied to more than one design. The tools usually incorporate items associated with bias (e.g., blinding and baseline comparability of study groups) and items related mainly to reporting (e.g., was the study population described and was a sample size calculation performed) [1]. The Cochrane Collaboration developed a new tool, released in 2008, to assess the potential for risk of bias in randomized controlled trials (RCTs). The Risk of Bias (ROB) tool [1] was developed to address some of the shortcomings of existing quality assessment instruments, including overreliance on reporting rather than methods.

The ROB tool was based on six domains: (1) sequence generation, (2) allocation concealment, (3) blinding, (4) incomplete outcome data, (5) selective outcome reporting, and (6) ''other sources of bias'' (e.g., design-specific risks of bias, early stopping for benefit, severe baseline imbalances, and inappropriate influence of funders). The developers of the tool aimed to distinguish between actual methods of conducting the trials vs. reporting. Furthermore, the choice of components for inclusion in the tool was based on the empirical evidence demonstrating their association with effect estimates.

Previous research examined the original Cochrane ROB tool in a sample of trials with a number of treatment conditions and showed that interrater agreement ranged from slight to substantial across the different domains, with the overall risk of bias assessment having ''fair'' agreement [7]. The authors identified sources of discrepancy and made recommendations to enhance the degree of consistency of the ROB tool. One of the stated limitations of this research was that the sample to which the tool was applied included only trials in children, the results of which may not be generalizable to trials conducted in other populations. A subsequent study by the same researchers showed improved interrater agreement on risk of bias assessments within the context of a specific systematic review [8]. The authors suggested that the improved agreement may have resulted from review-specific guidelines and pilot testing.

There is a clear need for interrater reliability testing of risk of bias assessment tools to enhance consistency in their application and interpretation across different systematic reviews. Furthermore, there is a need to determine the interrater reliability and validity to support the uptake and use of individual tools that are recommended by the systematic review community.

We undertook this project to assess the reliability of the ROB tool. We were interested in the reliability of risk of bias/quality assessments across individual raters and between consensus agreements of individual raters. The specific objectives were to (1) assess the reliability of the Cochrane ROB tool for RCTs between individual raters and between the consensus agreements of pairs of reviewers (i.e., comparing consensus agreements across four centers) and (2) examine the impact of study-level factors (e.g., outcomes, interventions, and conditions) on reliability of the Cochrane ROB tool.

2. Methods

This manuscript is part of a larger technical report conducted for the Agency for Healthcare Research and Quality. 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

A sample of 154 recently conducted RCTs involving adults was randomly selected, using a computer-generated randomization sequence, from a sample of 616 trials that were previously examined for quality of reporting by Hopewell et al. [9]. We chose this sample as it presented several advantages including the fact that it was a representative sample of all RCTs in the published literature, efficiencies in sample identification, and the potential for validation of assessments for key variables (e.g., allocation concealment, blinding, and attrition) by comparing them with those of another independent study team. The original sample included all primary reports of RCTs that were indexed in PubMed in December 2006 [9] and is likely representative of RCTs published in the medical literature.

Conducting sample size calculations for this type of research is challenging and cannot be determined using