BMJ open

Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2012-001368
Article Type:	Research
Date Submitted by the Author:	23-Apr-2012
Complete List of Authors:	Oremus, Mark; McMaster University, McMaster Evidencebased Practice Centre Oremus, Carolina Hall, Geoffrey McKinnon, Margaret Systematic Review Team, ECT & Cognition
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Mental health
Keywords:	Depression & mood disorders < PSYCHIATRY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, NEUROLOGY

SCHOLARONE™
Manuscripts

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales

Mark Oremus[1,2], Carolina Oremus[3,4], Geoffrey B.C. Hall[3,4], Margaret C. McKinnon[3,4], ECT & Cognition Systematic Review Team*[3,4]

[1]McMaster Evidence-based Practice Centre, McMaster University, Hamilton, Ontario, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[3]McMaster Integrative Neuroscience Discovery and Study (MINDS) Program, Hamilton, Ontario, Canada
[4]Department of Psychiatry and Behavioural Neuroscience, Hamilton, Ontario, Canada

*The ECT & Cognition Systematic Review Team includes Allyson Graham, Caitlin Gregory, Gagan Fervaha, Lindsay Hanford, Anthony Nazarov, Melissa Parlar, Maria Restivo, Erica Tatham, and Wanda Truong.

Corresponding author
Mark Oremus
Department of Clinical Epidemiology and Biostatistics
McMaster University
50 Main Street East – Room 308
Hamilton, Ontario, Canada
L8N 1E9
Tel.: 905-525-9140, x22437
Fax: 905-522-7681
E-mail: oremusm@mcmaster.ca

Keywords: systematic review; quality assessment; reliability; Jadad scale; Newcastle-Ottawa Scale.

Word Count: 3,143

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

ABSTRACT

Introduction: Quality assessment of included studies is an important component of systematic reviews.

Objective: We investigated interrater and test-retest reliability for assessments of study quality conducted by inexperienced student raters.

Design: Student raters received a training session in quality assessment using the Jadad scale for randomized controlled trials and the Newcastle-Ottawa Scale (NOS) for observational studies. Raters were randomly assigned into five pairs and they each independently rated the quality of 13-20 articles. Two months later, each rater re-assessed the quality of half of these articles.

Setting: University program (McMaster Integrative Neuroscience Discovery and Study [MINDS] Program).

Participants: 10 students (seven graduate, three undergraduate) taking MINDS Program courses.

Main Outcome Measures: Main outcomes included interrater reliability measured using Kappa and the intraclass correlation coefficient type 2,1, or ICC(2,1). We measured test-retest reliability using ICC(2,1). To assess differences in total score on the Jadad scale, we calculated mean differences in total score for each rater pair and individual rater.

Results: Interrater reliability was generally poor and test-retest reliability was fair to excellent. Mean differences in total scale score within rater pairs and within individual raters on the Jadad scale were minor and not statistically significantly different from zero, except for one rater's mean test-retest difference.

For peer review only

Conclusion: Agreement between raters was lower for 'interpretive' questions on the Jadad and NOS (e.g., questions asking about appropriateness of double-blinding or representativeness of exposed cases). A pilot rating phase following rater training may be one way to improve agreement.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

ARTICLE SUMMARY

Article focus

- To examine the interrater and test-retest reliability of inexperienced raters' quality assessments of articles included in a systematic review.

Key messages

- Among inexperienced raters, interrater reliability using the Jadad scale and Newcastle-Ottawa Scale was generally poor; test-retest reliability was fair to excellent.
- Systematic reviewers must pay special attention to training inexperienced quality raters; a pilot rating phase might be a helpful means of improving reliability among inexperienced raters, especially when rating observational study quality.

Strengths and limitations of this study

- No other study has examined the reliability of quality assessments in a group of inexperienced raters.
- Results may differ depending on rater background and experience, quality assessment instruments, and topic under study.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

INTRODUCTION

Systematic reviews summarize healthcare research evidence and they are useful for assessing whether treatment benefits outweigh risks.[1, 2] Accordingly, conclusions drawn from systematic reviews may impact clinical care and patient outcomes, thereby necessitating high standards of methodological rigour.

One critical component of conducting systematic reviews involves evaluation of the methodological quality of included studies. Study quality may influence treatment effect estimates and the validity of conclusions drawn from such estimates.[3] Through quality assessment, researchers identify strengths and weaknesses of existing evidence[4] and suggest ways to improve future research.

Careful work has identified key quality assessment domains.[1, 5] For randomized controlled trials (RCTs), these domains include appropriate generation of random allocation sequences, concealment of allocation sequences, blinding (of participants, health care providers, data collectors, and outcome assessors), and reporting of proportions of patients lost to follow-up.[1] For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment.[5]

Numerous scales exist to help raters assess study quality.[5-11] The majority of these scales list quality assessment domains and require raters to indicate whether each domain is present or absent from the studies under consideration. Some scales (e.g., Jadad,[6] Newcastle-Ottawa Scale [NOS][5]) assign points when quality domains are present, thus permitting the calculation of overall 'quality scores'. Other scales (e.g., risk of bias[8]) ask raters to rank the degree of bias (high, low, unclear) associated with each

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

quality domain.

Generally, quality scales demonstrate good interrater and test-retest reliability. Reliability coefficients such as Kappa (κ) are typically greater than 0.60,[9-17] although recent work reports κs of less than 0.50 for eight of nine questions on the NOS.[18]

Although quality assessment is now regarded as a standard component of systematic reviews, one issue that has received little attention in the literature is the effect of rater experience on the reliability of quality assessments. This issue is important because raters may be drawn from vast pools of persons with varying degrees of methods expertise, from experienced faculty to inexperienced students.

As part of an ongoing meta-analysis of electroconvulsive therapy (ECT) and cognitive impairment, we investigated interrater and test-retest reliability for student raters with no previous experience in quality assessment. To the best of our knowledge, no other study has examined this topic.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

METHODS

Study design

We retrieved 78 articles that contained data on cognitive impairment following the use of ECT to treat major depressive disorder. Fifty-five articles reported results of randomized controlled trials (RCTs), with one article containing results of five separate studies and two other articles each containing results of two separate studies, for a total of 61 RCTs. Fifteen articles reported on cohort studies and eight reported on case-control studies. Eleven articles were published prior to 1980, 17 between 1980 and 1989, 15 between 1990 and 1999, and 35 since 2000.

One author (MO) with systematic review experience trained 10 students (three undergraduate, seven graduate) to rate the methodological quality of published study reports using the 6-item Jadad scale for RCTs[6, 19] and the NOS for observational studies.[5] Training consisted of a 90-minute didactic session divided into two parts: part one highlighted the importance of quality assessment in systematic reviews and part two contained a question-by-question description of the Jadad and NOS instruments. We provided a standardized tabular spreadsheet for student raters to use during quality assessment.

We used a random number table to assign the student raters into five pairs and we subsequently distributed between 13 and 20 articles to each pair. No article was assigned to more than one pair; pairs received a mix of RCTs and observational studies. Articles fluctuated across pairs because of constraints on rater availability due to competing academic demands.

Raters determined the type of study design (i.e., RCT or observational) for each of

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

their assigned articles and one author (CO) verified their choices. Raters then independently rated their assigned articles to permit us to examine interrater reliability.

Statistical analysis

We used κ[20, 21] to measure interrater reliability for individual Jadad and NOS questions. We interpreted κ values as follows: greater than 0.80 was very good, 0.61 to 0.80 was good, 0.41 to 0.60 was moderate, 0.21 to 0.40 was fair, and less than 0.21 was poor.[22]

For test-retest reliability, each rater re-assessed half of the articles to which they had been assigned during the interrater reliability phase. The re-assessments took place two months after the interrater reliability phase[13] to minimize the possibility that recall of first assessments would influence the second assessments.

We employed the intraclass correlation coefficient-model 2,1, or ICC(2,1),[23] to measure interrater and test-retest reliability for the Jadad and NOS total scores. We computed separate ICC(2,1) values for consistency (systematic differences between raters are considered irrelevant) and absolute agreement (systematic differences between raters are considered relevant).[24] ICC(2,1) values were interpreted as follows: greater than 0.75 was excellent, 0.40 to 0.75 was fair to good, and less than 0.40 was poor.[25]

To investigate the differences in total Jadad scale scores within rater pairs (interrater) and within individual raters (test-retest), we calculated a mean difference in score for each rater pair and individual rater. We compared differences in score within rater pairs using the Wilcoxon rank-sum test and within individual raters using the Wilcoxon signed-rank test. We did not conduct this analysis for the NOS because of the small number of cohort and case-control studies.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

We did not pool mean differences since we did not expect to find a pooled estimate that would be different from zero. This is because the ordering within pairs was arbitrary (i.e. whether differences were calculated as rater1-rater2 or vice-versa).

SAS v9.2 (The SAS Institute, Cary, NC) was utilized to calculate κ and p-values for the Wilcoxon tests; SPSS v20 (IBM Corp., Armonk, NY) was used to calculate ICC(2,1). The level of significance was α=0.05.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

RESULTS

Interrater reliability

For interrater reliability, agreement between raters on individual questions was generally poor (Table 1). Half of the questions on the Jadad scale had moderate κs and the other half had poor κs. On the NOS, all κs were poor for the cohort study questions (NOS cohort) and six of eight κs were poor for the case-control study questions (NOS case-control).

Insert Table 1 Here

Examining total scale scores within rater pairs (Table 2), agreement was poor for the Jadad scale and NOS cohort and fair for the NOS case-control. However, point estimate ICC(2,1)s for the NOS cohort and case-control were not statistically significantly different from zero. Point estimate ICC(2,1)s and 95% confidence intervals did not appreciably differ according to calculation based on consistency or absolute agreement.

Insert Table 2 Here

The mean differences in total score on the Jadad scale within rater pairs ranged from 0.00 to 0.70; no difference was statistically significantly different from zero (Table 3).

Insert Table 3 Here

Test-retest reliability

Test-retest reliability following a two-month interval between assessments was fair to good for the Jadad scale and NOS cohort and excellent for the NOS case-control (Table 4). Point estimate ICC(2,1)s and 95% confidence intervals calculated for consistency

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

were similar to the results calculated for absolute agreement.

*****Insert Table 4 Here*****

The mean differences in total score on the Jadad scale within individual raters, subtracting scores at the second assessment from scores at the first assessment, ranged from 0.00 to 0.64 for nine of the raters (Table 3). None of these differences were statistically significantly different from zero. For one rater, the mean difference in total score was 3.38 (p=0.01).

Mean Differences in Total Score: Newcastle-Ottawa Scale

Although we did not apply formal statistical hypothesis testing to mean differences in total score on the NOS, the data suggest larger differences compared to the Jadad scale. Mean differences on the NOS cohort ranged from 0.25 to 3.00 (rater pairs) and 0.00 to 1.67 (individual raters). On the NOS case-control, mean differences spanned from 0.50 to 2.00 (rater pairs) and 0.00 to 1.00 (individual raters).

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

DISCUSSION

Overview and discussion of key findings

We investigated interrater and test-retest reliability for student raters with no previous experience in quality assessment. Our study is novel because, to the best of our knowledge, no other research has examined this issue. The raters used the Jadad scale and NOS to assess the quality of studies on the topic of ECT and cognitive impairment. Interrater reliability was generally poor and test-retest reliability was fair to excellent. Our results highlight the need for researchers to consider rater experience during the quality assessment of articles included in systematic reviews.

For interrater reliability, the poor κs on the Jadad scale pertained to the questions about appropriateness of double blinding and the clarity of reporting withdrawals, inclusion/exclusion criteria, and adverse effects. Often, authors did not report methods of blinding and raters had to make judgments about whether to award a point for the question on appropriateness of double blinding. Despite what we communicated during the training session, some raters may have given authors the benefit of the doubt and awarded the point for appropriateness if studies simply reported double blinding, even though another question on the Jadad scale already asked whether authors reported their studies as blinded. Similarly, differences in rater opinion regarding what constitutes an 'adequate' description of withdrawals, inclusion/exclusion criteria, or adverse effects led to poor agreement on these questions. To improve interrater agreement among inexperienced raters, we suggest a pilot phase wherein raters rate the quality of a subsample of articles to allow for identification and clarification of areas of ambiguity.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

With regard to the NOS, question-specific interrater reliability was poorer than that of the Jadad scale. We believe the NOS's poor reliability may be explained in part by differences in how raters answered interpretive questions, e.g., whether exposed cohorts are somewhat or truly representative of the average exposed person in the community (first question on NOS cohort).

Poor question-specific interrater agreement on the NOS also reflects an inherent challenge with rating the quality of observational studies compared to RCTs. This challenge is exemplified by the multiplicity of tools that exist to assess observational study quality. Two systematic reviews[26, 27] each found over 80 such tools, which varied in design and content. Despite the cornucopia of tools, no gold standard scale exists to rate the quality of observational studies.[28]

Rater disagreements on interpretive questions and inherent challenges with assessing observational study quality explain the negative κs that were calculated for some NOS questions. Negative κs result when agreement occurs less often than predicted by chance alone. This suggests genuine disagreement between raters or an underlying issue with the instrument itself.[29] Indeed, Hartling et al. reported that raters had difficulty using the NOS because of uncertainty over the meaning of certain questions (e.g., representativeness of the exposed cohort, selection of non-exposed cohort) and response options (e.g., 'truly' versus 'somewhat' exposed).[18] These difficulties existed despite Hartling et al.'s use of a pilot training phase. Our raters' difficulties with the interpretative questions might have been a function of issues with the NOS, which could be related to the broader challenge of assessing the quality of observational studies.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Question-specific differences between raters also led to poor interrater agreement on total scores for the Jadad scale and NOS cohort. This may not be evident by comparing the κs and ICC(2,1)s calculated for Jadad. κs for four of eight Jadad questions were moderate yet the ICC(2,1) for total score was poor. However, since total scores are computed using raters' answers to all of the questions on a scale (some answers are awarded one point and others zero points), raters who disagree on small numbers of questions (e.g., two of eight questions) will nonetheless show poor agreement on total scores.

Conversely, for the NOS case-control, κs for six of eight questions were poor yet the ICC(2,1) was fair. In this situation, no 'reliability' relation exists between responses to questions and total scores. For example, rater 1 might answer 'yes' (one point per 'yes' response) and rater 2 might answer 'no' (zero points per 'no' response) to even-numbered questions. For odd-numbered questions, the pattern is reversed. Assuming eight questions, interrater reliability at the question level will be poor because the raters did not agree on their responses, but their overall scores will be equivalent.

Many authors base their discussions of study quality in systematic reviews on raters' responses to individual questions on quality assessment scales. Given that we found generally poor interrater reliability on answers to questions, the process of resolving conflicts between raters becomes important. Many reviews simply report that raters solved disagreements by consensus without describing specific procedures. We speculate that conflict resolution may occasionally be approached in an ad hoc nature or treated as a nuisance to be dealt with as expeditiously as possible. We suggest the process of conflict resolution should be more of a formalized endeavour requiring raters

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

to set aside some 'resolution time' and articulate their reasons for choosing specific answers. In the event the raters do not agree, a third party may be asked to listen to each rater's opinion and make a decision. Although space restrictions in journals might prevent authors from reporting such procedures (when they exist) in manuscripts, the move toward publication of systematic review protocols, for example as mandated by the United States Agency for Healthcare Research and Quality's Effective Health Care Program,[30] provides authors with an opportunity to elaborate on their consensus processes.

Test-retest reliability was better than interrater reliability. Individual raters appeared to adopt a uniform approach to assessing the quality of articles assigned to them. Each rater had her or his own understanding of the interpretive questions and applied this point-of-view consistently throughout the rating process. The issue was the difference in interpretations between raters.

Comparison with other studies

To the best of our knowledge, no other study has examined interrater and test-retest reliability for a group of novice student quality assessors. Two published studies[31, 32] of rater agreement included persons with different levels of experience, although the focus was on extraction of article data (e.g., info on study design, sample characteristics, length of follow-up, definition of outcome, and results) rather than quality assessment. Horton et al. classified rater experience as minimal, moderate, or substantial and asked raters to extract data from three studies on insomnia therapy.[31] They found no statistically significant differences in error rates according to experience. Hayward et al.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

trained two experienced raters and one inexperienced rater to independently extract data from seven studies.[32] Agreement between raters was largely perfect.

A recent AHRQ methods report had 16 raters assess the quality of 131 cohort studies using the NOS. Rater experience ranged from four months to 10 years; 13 raters had formal training in systematic reviews.[18] κs were less than 0.50 for eight of nine NOS questions, although the authors did not break down their results by rater experience.

Oremus et al. examined the interrater reliability of the Jadad scale using three raters (two experienced faculty members and one inexperienced PhD student), who read the methods and results of 42 Alzheimer's disease drug trials.[19] The ICC(2,1) for total scores on the Jadad scale was 0.90. Al-Harbi et al. engaged two paediatric surgeons to rate 46 cohort studies that were presented at Canadian Association of Pediatric Surgeons annual meetings and later published in the Journal of Pediatric Surgery.[12] The authors did not specify whether the surgeons received training in quality assessment. The ICC between surgeons, calculated on NOS total scores, was 0.94.

The lower interrater reliability of the novice student raters in this study, compared to the raters in the Oremus et al.[19] and Al-Harbi et al.[12] studies, may be explained by topic familiarity and similarity of expertise. The faculty raters in the Oremus et al. study had previously worked on a systematic review of Alzheimer's disease medications and their expertise lay in two domains of epidemiology, i.e., neuroepidemiology and pharmacoepidemiology. The paediatric surgeons in Al-Harbi et al. may have possessed at least a general familiarity with the types of cohort studies conducted in their speciality. These characteristics may have predisposed the raters to adopt more uniform opinions on the questions contained in Jadad and NOS. In contrast, the novice student raters in our

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

study had for the most part not been exposed to systematic reviews and quality assessment in the past. Also, seven of these raters were recent entrants to graduate school and they came from a variety of undergraduate backgrounds such as medicine, psychology, and basic science.

Limitations

Readers should exercise caution when generalising the results of our study to other types of raters or scenarios. Reliability could differ according to raters' disciplines and levels of training, even among groups of inexperienced students. Reliability is also partly a function of the instruments used in the quality assessment. Furthermore, the topic under study could influence reliability, as could certain parameters of the systematic review methodology. For example, the meta analysis on ECT and cognition, upon which we based this study, included 28 papers published prior to 1990. The style of reporting results in older papers does not always facilitate quality assessment or data extraction. Systematic reviews that include older papers could therefore present challenges for maintaining acceptable levels of interrater and test-retest reliability.

Conclusions

In conclusion, we asked a group of 10 novice students to rate the quality of 78 articles that contained data on cognitive impairment following the use of ECT to treat major depressive disorder. Overall, interrater reliability on the Jadad scale and NOS was poor, although test-retest reliability ranged from fair to excellent. We trained the raters prior to the quality assessment exercise yet interrater agreement was low for several questions that required a certain degree of interpretation to answer. This was especially so for the

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

NOS and underscores an inherent greater difficulty with assessing the quality of observational studies compared to RCTs.

In addition to standardized training prior to commencing quality assessment, a pilot rating phase may also be necessary to discuss scale questions that generate disagreement among novice student raters. This procedure could help the raters develop standardized interpretations to minimize disagreement.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Acknowledgements: Special thanks to Eleanor Pullenayegum and Harry Shannon for their helpful comments on an earlier draft of this manuscript.

Contributors: MO and CO conceived and designed the study. MO analysed the data. MO, CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team interpreted the data. MO drafted the manuscript. CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript.

Funding: This study did not receive funds from any sponsor. No person or organization beyond the authors had any input in study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication.

Competing interests: The authors have no competing interests to declare.

Ethics approval: Not required.

Provenance and peer review: Not commissioned; externally peer reviewed.

Data sharing: No additional data available.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

REFERENCES

1. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.

2. Agency for Healthcare Research and Quality (AHRQ). *Systems to Rate the Strength of Scientific Evidence.* Evidence Report/Technology Assessment No. 47. Rockville, MD: Agency for Healthcare Research and Quality, 2002.

3. Verhagen AP, de Vet HC, de Bie RA, *et al*. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651-4.

4. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.

5. Wells GA, Shea B, O'Connell D, *et al*. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (2 April 2012).

6. Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

7. Chalmers TC, Smith H, Jr., Blackburn B, *et al*. A method for assessing the quality of a randomized control trial. *Control Clin Trials* 1981;2:31-49.

8. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. www.cochrane-handbook.org (23 April 2012).

9. Maher CG, Sherrington C, Herbert RD, *et al*. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther* 2003;83:713-21.

10. Kocsis JH, Gerber AJ, Milrod B, *et al*. A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Compr Psychiatry* 2010;51:319-24.

11. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377-84.

12. Al-Harbi K, Farrokhyar F, Mulla S, *et al*. Classification and appraisal of the level of clinical evidence of publications from the Canadian Association of Pediatric Surgeons for the past 10 years. *J Pediatr Surg* 2009;44:1013-7.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

13. Berard A, Andreu N, Tetrault J, *et al*. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol* 2000;10:498-503.

14. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.

15. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6:e17242.

16. Tooth L, Bennett S, McCluskey A, *et al*. Appraising the quality of randomized controlled trials: inter-rater reliability for the OTseeker evidence database. *J Eval Clin Pract* 2005;11:547-55.

17. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;135:982-9.

18. Hartling L, Hamm M, Milne A, *et al*. *Validity and Inter-rater Reliability Testing of Quality Assessment Instruments*. Rockville, MD: Agency for Healthcare Research and Quality, 2012.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

19. Oremus M, Wolfson C, Perrault A, *et al*. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord* 2001;12:232-6.

20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Manag* 1960;20:37-46.

21. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd edn. Hoboken, NJ: John Wiley & Sons, 2003.

22. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.

24. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th edn. Oxford: Oxford University Press, 2008.

25. Fleiss J. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

26. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol* 2010;63:1061-70.

27. 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;36:666-76.

28. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. *Int J Evid Based Healthc* 2010;8:247.

29. Juurlink DN, Detsky AS. Kappa statistic. *CMAJ* 2005;173:16.

30. Agency for Healthcare Research and Quality (AHRQ). Effective Health Care Program. Rockville, MD: Agency for Healthcare Research and Quality. http://www.effectivehealthcare.ahrq.gov (2 April 2012).

31. Horton J, Vandermeer B, Hartling L, *et al*. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 2010;63:289-98.

32. Haywood KL, Hargreaves J, White R, *et al*. Reviewing measures of outcome: reliability of data extraction. *J Eval Clin Pract* 2004;10:329-37.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

TABLES

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Randomization	0.50 (-1.00 to 1.00)	Representative-ness of exposed cohort	-0.13 (-0.36 to 0.11)	Case definition adequate	1.00 (1.00 to 1.00)
Appropriate randomization	0.56 (0.29 to 0.83)	Selection of non-exposed cohort	-0.14 (-0.28 to 0.00)	Cases representative	-0.20 (-0.49 to 0.09)
Double-blind	0.41 (0.16 to 0.66)	Exposure ascertainment	0.00 (0.00 to 0.00)	Control selection	0.25 (-0.19 to 0.69)
Appropriate double-blind	0.17 (-0.07 to 0.41)	Outcome not present at baseline	0.20 (-0.33 to 0.73)	Control definition	0.14 (-0.54 to 0.82)
Description of withdrawals	0.21 (-0.02 to 0.45)	Comparability of cohorts	0.12 (-0.23 to 0.47)	Case and control comparability	0.00 (0.00 to 0.00)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question (continued)

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Description of inclusion / exclusion criteria	0.27 (-0.03 to 0.57)	Outcome assessment	0.31 (-0.08 to 0.69)	Exposure ascertainment	-0.11 (-0.68 to 0.46)
Description of adverse effects	0.13 (-0.11 to 0.37)	Follow-up long enough	-0.09 (-0.22 to 0.04)	Same ascertainment method for cases and controls	0.60 (-0.07 to 1.00)
Description of statistical analysis	0.49 (0.21 to 0.77)	Follow-up adequate	0.39 (-0.02 to 0.81)	Non-response rate	-0.11 (-0.65 to 0.43)

CI, confidence interval; NOS, Newcastle-Ottawa Scale.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 2 Interrater reliability for Jadad and Newcastle-Ottawa scales: total scale scores within rater pairs

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad	0.32 (0.08 to 0.53)	0.32 (0.08 to 0.52)
Newcastle-Ottawa – Cohort	-0.19 (-0.63 to 0.34)	-0.19 (-0.67 to 0.35)
Newcastle-Ottawa – Case-control	0.55 (-0.18 to 0.89)	0.46 (-0.13 to 0.92)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 3 Mean differences in total score on Jadad scale*

Pair	Mean Difference	Rater	Mean Difference
1	0.25 (p=0.46)	1	0.08 (p=1.00)
2	0.30 (p=0.24)	2	0.42 (p=0.81)
3	0.70 (p=0.46)	3	0.64 (p=0.45)
4	0.00 (p=1.00)	4	3.38 (p=0.01)
5	0.47 (p=0.39)	5	0.33 (p=0.25)
		6	0.00 (p=1.00)
		7	0.18 (p=0.81)
		8	0.00 (p=1.00)
		9	0.42 (p=0.26)
		10	0.00 (p=1.00)

*Score range=0-8.

eview only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 4 Test-retest reliability for Jadad and Newcastle-Ottawa scales: comparison of total scale scores for individual raters after two assessments

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad	0.56 (0.42 to 0.67)	0.55 (0.41 to 0.67)
Newcastle-Ottawa – Cohort	0.61 (0.24 to 0.82)	0.62 (0.25 to 0.83)
Newcastle-Ottawa – Case-control	0.85 (0.55 to 0.95)	0.83 (0.48 to 0.95)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

STROBE statement checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract		
	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract (√) (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found (√)
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported (√)
Objectives	3	State specific objectives, including any prespecified hypotheses (√)
Methods		
Study design	4	Present key elements of study design early in the paper (√)
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection (N/A)
Participants	6	(*a*) *Cohort study*?Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up*Case-control study*?Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls*Cross sectional study*?Give the eligibility criteria, and the sources and methods of selection of participants (N/A) (*b*) *Cohort study*?For matched studies, give matching criteria and number of exposed and unexposed*Case-control study*?For matched studies, give matching criteria and the number of controls per case (N/A)
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable (√)
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group (√)
Bias	9	Describe any efforts to address potential sources of

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

		bias (N/A)
Study size	10	Explain how the study size was arrived at (√)
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why (N/A)
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (√) (*b*) Describe any methods used to examine subgroups and interactions (N/A) (*c*) Explain how missing data were addressed (N/A) (*d*) *Cohort study*?If applicable, explain how loss to follow-up was addressed*Case-control study*?If applicable, explain how matching of cases and controls was addressed*Cross sectional study*?If applicable, describe analytical methods taking account of sampling strategy (N/A) (*e*) Describe any sensitivity analyses (N/A)
Results		
Participants	13*	(*a*) Report numbers of individuals at each stage of study?eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (N/A) (*b*) Give reasons for non-participation at each stage (N/A) (*c*) Consider use of a flow diagram (N/A)
Descriptive data	14*	(*a*) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (N/A) (*b*) Indicate number of participants with missing data for each variable of interest (N/A) (*c*) *Cohort study* Summarise follow-up time (eg average and total amount) (N/A)
Outcome data	15*	*Cohort study* Report numbers of outcome events or summary measures over time (N/A) *Case-control study* Report numbers in each exposure category, or summary measures of exposure (N/A) *Cross sectional study* Report numbers of outcome events or summary measures (N/A)
Main results	16	(*a*) Report the numbers of individuals at each stage of the study, eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (N/A) (*b*) Give reasons for non-participation at each stage (N/A)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

		(*c*) Consider use of a flow diagram (N/A)
Other analyses	17	Report other analyses done, eg, analyses of subgroups and interactions, and sensitivity analyses (N/A)
Discussion		
Key results	18	Summarise key results with reference to study objectives (√)
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias. (√)
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. (√)
Generalisability	21	Discuss the generalisability (external validity) of the study results. (√)
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based. (√)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2012-001368.R1
Article Type:	Research
Date Submitted by the Author:	12-Jun-2012
Complete List of Authors:	Oremus, Mark; McMaster University, McMaster Evidencebased Practice Centre Oremus, Carolina Hall, Geoffrey McKinnon, Margaret Systematic Review Team, ECT & Cognition
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Mental health
Keywords:	Depression & mood disorders < PSYCHIATRY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, NEUROLOGY

SCHOLARONE™
Manuscripts

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

June 12, 2012

Dr Trish Groves
Editor-in-chief, BMJ Open
Deputy Editor, BMJ
BMJ Group
London, UK

Subject: bmjopen-2012-001368.R1 – Resubmission of Manuscript Entitled "Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales"

Dear Dr. Groves:

Thank you for the opportunity to revise and resubmit our manuscript. We would also like to thank the reviewers for their helpful comments, which we believe helped us improve the quality of the manuscript.

Our responses to the reviewers' comments are in italicized text below. The page and line numbers listed in our responses refer to sections of the tracked changes version of the revised manuscript.

Reviewer: Tatyana A Shamliyan

1. The abstract does not describe a selection of the articles (13-20) and design distribution.

We added these descriptions to the abstract (lines 9-11) and revised a small section of the text (p. 7, line 22) to clarify how the articles were distributed to raters.

2. The abstract does provide sample size calculation (10 students) and definitions of poor, fair, or excellent reliability.

Thank you for the comment.

3. Conclusions about low reliability rating for specific questions are not supported by result section.

We re-wrote the results section of the abstract to include summaries of the calculated Kappas and intraclass correlation coefficients. We removed reference to specific types of questions from the abstract conclusion. We also revised the first bullet under 'Key messages' of the 'Article Summary'.

4. Limitations of the study do not address limitations of the used scales (Jadad scale and Newcastle-Ottawa Scale). The scales with judgmental and interpretive questions could have poor reliability despite several phases of training.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

We added mention of this issue to the discussion (p. 16, lines 18-20).

5. The training to clarify "ambiguity" in the methodological and reporting quality of the evaluated studies may not improve reliability testing that was due to scale content and structure.

We added mention of this issue to the discussion (p. 12, lines 1-4).

6. The recent publication by the Cochrane Bias Methods Group and the Cochrane Statistical Methods Group recommended :" Do not use quality scales.
Quality scales and resulting scores are not an appropriate way to appraise clinical trials. They tend to combine assessments of aspects of the quality of reporting with aspects of trial conduct, and to assign weights to different items in ways that are difficult to justify. Both theoretical considerations and empirical evidence suggest that associations of different scales with intervention effect estimates are inconsistent and unpredictable." BMJ 2011;343:d5928. This publication should be mentioned in the discussion.

We added mention of this issue to the conclusion (p. 17, lines 18-23). We also cited the BMJ publication (reference 33 in bibliography).

7. "For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment.[5]" Selection and attrition bias are also very important when evaluating internal validity of the observational studies of health c are interventions.

We added mention of this issue to the introduction (p. 5, line 17).

8. "Articles fluctuated across pairs because of constraints on rater availability due to competing academic demands". Please clarify what do you mean by article fluctuation and "competing academic demands".

We revised the sentence in question to enhance clarity (p. 8, lines 1-2).

9. Please clarify had senior reviewers evaluated quality of the articles before giving the articles to the students and had they compared own ranking with the ranking by non experienced raters?

Senior reviewers did not rate article quality. We added mention of this issue in the limitations section (p. 16, lines 14-16).

10. Please justify the same size and describe student invitation response rate and articles selection.

The 10 students in the study were a convenience sample and we added mention of this fact to the limitations section (p. 16, lines 12-13).

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

We added two sentences to the methods to describe student invitation and response rate (p. 7, lines 11-13).

We clarified article selection by editing the last paragraph of the introduction (p. 6, lines 10-13) and the first paragraph of the methods (p. 7, lines 3-5) to explain that the 78 articles in our study came from an ongoing systematic review of cognitive impairment and electroconvulsive therapy. These 78 articles were the included studies in the meta-analysis.

11. Please clarify that your goal was quality evaluation of observational studies of health care interventions.

Our goal was to examine the reliability of quality assessments done by inexperienced student raters. We included quality assessments of RCTs and observational studies in our examination. We clarified these points in the last paragraph of the introduction (p. 6, lines 10-13).

Reviewer: Arianne P Verhagen

1. I think the design is flawed in a way that the authors actually evaluate the effect of a training course on quality assessment rather than the reliability.

We agree that training programs may influence reliability and we added mention of this fact to the 'strengths and limitations' section of 'article summary' box and to the limitations section (p. 16, lines 11-12).

We disagree that the design is flawed. We patterned our study design on an approach used by several similar investigations, including references 12, 13, 14, 15, 16, 18, 31, and 32 from our bibliography. The primary objective of all of these studies was to calculate reliability, not to examine the impact of training programs on reliability.

2. [A]lso the statistics need to be discussed with a statistician.

Eleanor Pullenayegum and Harry Shannon, both listed in the acknowledgements, are statisticians who provided feedback on the manuscript prior to submission.

3. They [statistics] also do not seem to be very well documented on the topic.

We used standard statistics (Kappa and intraclass correlation coefficient) to calculate reliability. We referenced our sources for these statistics (references 20, 21, 23, and 24 in the bibliography). We also referenced the sources for our interpretations (e.g., 'poor', 'fair') of these statistics (references 22 and 25 in the bibliography).

4. [V]arious key references are missing.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

We would be happy to look into these references if the reviewer could provide us with a list.

5. [T]hey used a scale that it not very often used (modified Jadad scale).

The modified Jadad scale contains the original three questions proposed by Jadad et al. (http://ac.els-cdn.com/0197245695001344/1-s2.0-0197245695001344-main.pdf?_tid=7e939adc61c0566105605a2bc2682525&acdnat=1339428722_689c0e9c11502cca0c52cacb7a1e1d76). The modified scale also contains three additional questions considered by Jadad et al. in their original scale development work. The additional three questions were added to the Jadad scale for a systematic review of Alzheimer's disease medications (http://www.cadth.ca/media/pdf/106_alzheimers1_tr_e.pdf). The modified Jadad scale had excellent interrater reliability (ICC=0.90) in this systematic review (see reference 19 in the bibliography) and was subsequently used in a range of other systematic reviews, e.g., Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure (http://www.ahrq.gov/downloads/pub/evidence/pdf/bnp/bnp.pdf), Diagnosis and Treatment of Secondary Lymphedema (https://www.cms.gov/Medicare/Coverage/DeterminationProcess/downloads/id66aTA.pdf), Pharmacological Treatment of Dementia (http://www.ahrq.gov/downloads/pub/evidence/pdf/dempharm/dempharm.pdf).

6. I think the discussion lack clarity and does not discuss the main limitations of studies like these.

We would be happy to clarify any section of the discussion that may be lacking clarity. We encourage the reviewer to point out any sections that she feels may require more clarity.

We provided a limitations section in the initial manuscript and we added to this section in response to both reviewers' comments.

General comments

1. The authors do not explain why they study the reliability in (inexperienced) students. I cannot see what the rationale is to do so.

The introduction to our original manuscript contained two paragraphs that explained our rationale for studying reliability in inexperienced students. With some modifications, we retained these paragraphs in the current version of the manuscript (p. 6, lines 5-13).

What they actually do is to evaluate the output of the training course. I think this training course (of 90 minutes!) is not very good as the interrater reliability directly after the course is low.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

We agree that training programs may influence reliability and we added mention of this fact to the 'strengths and limitations' section of 'article summary' box and to the limitations section (p. 16, lines 11-12).

Some of the poor reliability scores may also result from the difficulty of using the NOS, which we addressed in the discussion of the original manuscript. We retained this section in the current manuscript (p. 12, lines 16-23; p. 13, lines 1-4).

In 2 months time the authors do not expect any recall of the first score, I assume most of the information of the course might also be forgotten.

The literature provided very little guidance on an adequate time frame for measuring test-retest reliability in studies such as ours. We based our two-month interval on a study that did utilize methods similar to ours (reference 13 in the bibliography).

Since our purpose was not to evaluate a training program, we did not assess recall of course content.

2. For the assessment the authors used a modified Jadad scale. I do not think A. Jadad will be very pleased that this scale is chosen instead of the original one. The modifications are all related to external validity and have no clear relation with actual quality of the study. I recommend sticking to the original scale when studying reliability in a general way as the authors aimed to do.

Table 1 already presents interrater reliability for the 'original' three Jadad questions (randomization, double-blinding, description of withdrawals), along with the follow-up questions on appropriateness of double-blinding and randomization, which are also part of Jadad et al.'s initial (3-item) scale. To account for the reviewer's comment, we calculated interrater and test-retest reliability for total scores based on the original 3-item Jadad scale (p. 8, lines 21-23; p. 10, lines 10 & 19-20; Table 3).

3. Concerning the analysis the authors not only assessed reliability using Kappa scores and ICC, but also calculated mean differences between rater-pairs. I do not see the rationale for this. This analysis does not add anything to the answer on the study question whether quality assessment done by inexperienced raters after a 90 min course is reliable. I should delete this from the manuscript as it confuses the reader and does not inform them.

We removed the mean differences comparison from the manuscript.

4. One of the key messages is that the reliability between inexperienced raters is low. You do not have to do a study to show this, every course teacher knows, so what's new?

This article is the first research to quantify test-retest and interrater reliability for inexperienced student raters. This study empirically tests what "every course teacher

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

knows". Prior to this study, no one could say for sure whether reliability for inexperienced raters was low, nor could anyone estimate 'how low' this reliability might be.

Sincerely,

Mark Oremus, PhD
McLaughlin Foundation Professor of Population and Public Health & Assistant Professor, Department of Clinical Epidemiology & Biostatistics
Co-Associate Director, McMaster Evidence-based Practice Centre
Associate Scientific Director, Canadian Longitudinal Study on Aging
McMaster University

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales

Mark Oremus[1,2], Carolina Oremus[3,4], Geoffrey B.C. Hall[3,4], Margaret C. McKinnon[3,4], ECT & Cognition Systematic Review Team*[3,4]

[1]McMaster Evidence-based Practice Centre, McMaster University, Hamilton, Ontario, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[3]McMaster Integrative Neuroscience Discovery and Study (MINDS) Program, Hamilton, Ontario, Canada
[4]Department of Psychiatry and Behavioural Neuroscience, Hamilton, Ontario, Canada

*The ECT & Cognition Systematic Review Team includes Allyson Graham, Caitlin Gregory, Gagan Fervaha, Lindsay Hanford, Anthony Nazarov, Melissa Parlar, Maria Restivo, Erica Tatham, and Wanda Truong.

Corresponding author
Mark Oremus
Department of Clinical Epidemiology and Biostatistics
McMaster University
50 Main Street East – Room 308
Hamilton, Ontario, Canada
L8N 1E9
Tel.: 905-525-9140, x22437
Fax: 905-522-7681
E-mail: oremusm@mcmaster.ca

Keywords: systematic review; quality assessment; reliability; Jadad scale; Newcastle-Ottawa Scale.

Word Count: 3,044

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

ABSTRACT

Introduction: Quality assessment of included studies is an important component of systematic reviews.

Objective: We investigated interrater and test-retest reliability for quality assessments conducted by inexperienced student raters.

Design: Student raters received a training session on quality assessment using the Jadad scale for randomized controlled trials and the Newcastle-Ottawa Scale (NOS) for observational studies. Raters were randomly assigned into five pairs and they each independently rated the quality of 13-20 articles. These articles were drawn from a pool of 78 papers examining cognitive impairment following electroconvulsive therapy to treat major depressive disorder. The articles were randomly distributed to the raters. Two months later, each rater re-assessed the quality of half of their assigned articles.

Setting: McMaster Integrative Neuroscience Discovery and Study (MINDS) Program.

Participants: 10 students taking MINDS Program courses.

Main Outcome Measures: We measured interrater reliability using Kappa and the intraclass correlation coefficient type 2,1, or ICC(2,1). We measured test-retest reliability using ICC(2,1).

Results: Interrater reliability varied by scale question. For the 6-item Jadad scale, question-specific Kappas ranged from 0.13 ([95% confidence interval] -0.11 to 0.37) to 0.56 (0.29 to 0.83). The ranges were -0.14 (-0.28 to 0.00) to 0.39 (-0.02 to 0.81) for the NOS cohort and -0.20 (-0.49 to 0.09) to 1.00 (1.00 to 1.00) for the NOS case-control. For overall scores on the 6-item Jadad scale, ICC(2,1)s for interrater and test-retest reliability (accounting for systematic differences between raters) were 0.32 (0.08 to 0.52) and 0.55

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

(0.41 to 0.67) respectively. Corresponding ICC(2,1)s for the NOS cohort were -0.19 (-0.67 to 0.35) and 0.62 (0.25 to 0.83), and for the NOS case-control the ICC(2,1)s were 0.46 (-0.13 to 0.92) and 0.83 (0.48 to 0.95).

Conclusion: Interrater reliability was generally poor to fair and test-retest reliability was fair to excellent. A pilot rating phase following rater training may be one way to improve agreement.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

ARTICLE SUMMARY

Article focus

- To examine the interrater and test-retest reliability of inexperienced raters' quality assessments of articles included in a systematic review.

Key messages

- Among inexperienced raters, interrater reliability using the Jadad scale and Newcastle-Ottawa Scale was generally poor to fair; test-retest reliability was fair to excellent.
- Systematic reviewers must pay special attention to training inexperienced quality raters; a pilot rating phase might be a helpful means of improving reliability among inexperienced raters, especially when rating observational study quality.

Strengths and limitations of this study

- No other study has examined the reliability of quality assessments in a group of inexperienced raters.
- Results may differ depending on rater background and experience, rater training, quality assessment instruments, and topic under study.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

INTRODUCTION

Systematic reviews summarize healthcare research evidence and they are useful for assessing whether treatment benefits outweigh risks.[1, 2] Accordingly, conclusions drawn from systematic reviews may impact clinical care and patient outcomes, thereby necessitating high standards of methodological rigour.

One critical component of conducting systematic reviews involves evaluation of the methodological quality of included studies. Study quality may influence treatment effect estimates and the validity of conclusions drawn from such estimates.[3] Through quality assessment, researchers identify strengths and weaknesses of existing evidence[4] and suggest ways to improve future research.

Careful work has identified key quality assessment domains.[1, 5] For randomized controlled trials (RCTs), these domains include appropriate generation of random allocation sequences, concealment of allocation sequences, blinding (of participants, health care providers, data collectors, and outcome assessors), and reporting of proportions of patients lost to follow-up.[1] For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment,[5] as well as selection and attrition biases.

Numerous scales exist to help raters assess study quality.[5-11] The majority of these scales list quality assessment domains and require raters to indicate whether each domain is present or absent from the studies under consideration. Some scales (e.g., Jadad,[6] Newcastle-Ottawa Scale [NOS][5]) assign points when quality domains are present, thus permitting the calculation of overall 'quality scores'. Other scales (e.g., risk of bias[8]) ask raters to rank the degree of bias (high, low, unclear) associated with each

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

quality domain.

Generally, quality scales demonstrate good interrater and test-retest reliability. Reliability coefficients such as Kappa (κ) are typically greater than 0.60,[9-17] although recent work reports κs of less than 0.50 for eight of nine questions on the NOS.[18]

Although quality assessment is now regarded as a standard component of systematic reviews, one issue that has received little attention in the literature is the effect of rater experience on the reliability of quality assessments. This issue is important because raters may be drawn from vast pools of persons with varying degrees of methods expertise, from experienced faculty to inexperienced students.

We investigated interrater and test-retest reliability for student raters with no previous experience in the quality assessment of randomized controlled trials (RCTs) and observational studies. To the best of our knowledge, no other study has examined this topic.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

METHODS

Study design

In an ongoing systematic review of cognitive impairment following electroconvulsive therapy (ECT) to treat major depressive disorder, 78 published articles passed title and abstract, and full-text, screening. These articles formed the basis of this study. Fifty-five of the articles reported the results of randomized controlled trials (RCTs), with one article containing results of five separate studies and two other articles each containing results of two separate studies, for a total of 61 RCTs. Fifteen articles reported on cohort studies and eight reported on case-control studies. Eleven articles were published prior to 1980, 17 between 1980 and 1989, 15 between 1990 and 1999, and 35 since 2000.

We invited all 10 students (three undergraduate, seven graduate) taking a 'special topics' course in the McMaster Integrative Neuroscience Discovery and Study Program to participate in this study. All 10 students accepted the invitation. One author (MO) with systematic review experience trained the students to rate the methodological quality of published study reports using the 6-item Jadad scale for RCTs[6, 19] and the NOS for observational studies.[5] Training consisted of a 90-minute didactic session divided into two parts: part one highlighted the importance of quality assessment in systematic reviews and part two contained a question-by-question description of the Jadad and NOS instruments. We provided a standardized tabular spreadsheet for student raters to use during quality assessment.

We used a random number table to assign the student raters into five pairs and we randomly distributed between 13 and 20 articles to each pair. None of the 78 articles was assigned to more than one pair; pairs received a mix of RCTs and observational studies.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

The number of articles assigned to the pairs depended on the amount of time each rater could devote to this study.

Raters determined the type of study design (i.e., RCT or observational) for each of their assigned articles and one author (CO) verified their choices. Raters then independently rated their assigned articles to permit us to examine interrater reliability.

Statistical analysis

We used κ[20, 21] to measure interrater reliability for individual Jadad and NOS questions. We interpreted κ values as follows: greater than 0.80 was very good, 0.61 to 0.80 was good, 0.41 to 0.60 was moderate, 0.21 to 0.40 was fair, and less than 0.21 was poor.[22]

For test-retest reliability, each rater re-assessed half of the articles to which they had been assigned during the interrater reliability phase. The re-assessments took place two months after the interrater reliability phase[13] to minimize the possibility that recall of first assessments would influence the second assessments.

We employed the intraclass correlation coefficient-model 2,1, or ICC(2,1),[23] to measure interrater and test-retest reliability for the Jadad and NOS total scores. We computed separate ICC(2,1) values for consistency (systematic differences between raters are considered irrelevant) and absolute agreement (systematic differences between raters are considered relevant).[24] ICC(2,1) values were interpreted as follows: greater than 0.75 was excellent, 0.40 to 0.75 was fair to good, and less than 0.40 was poor.[25]

We calculated two sets of ICC(2,1)s for the Jadad scale. The first set pertained to the 6-item Jadad scale [19] and the second set pertained to the original 3-item Jadad scale [6].

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

SAS v9.2 (The SAS Institute, Cary, NC) was utilized to calculate κ; SPSS v20 (IBM Corp., Armonk, NY) was used to calculate ICC(2,1). The level of significance was α=0.05.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

RESULTS

Interrater reliability

For interrater reliability, agreement between raters on individual questions was generally poor (Table 1). Half of the questions on the Jadad scale had moderate κs and the other half had poor κs. On the NOS, all κs were poor for the cohort study questions (NOS cohort) and six of eight κs were poor for the case-control study questions (NOS case-control).

Insert Table 1 Here

Examining total scale scores within rater pairs (Table 2), agreement was poor for the Jadad scale (6- and 3-item versions) and NOS cohort and fair for the NOS case-control. However, point estimate ICC(2,1)s for the NOS cohort and case-control were not statistically significantly different from zero. Point estimate ICC(2,1)s and 95% confidence intervals did not appreciably differ according to calculation based on consistency or absolute agreement.

Insert Table 2 Here

Test-retest reliability

Test-retest reliability following a two-month interval between assessments was fair to good for the Jadad scale and NOS cohort and excellent for the NOS case-control (Table 3). Test-retest reliability was slightly higher for the 3-item Jadad scale versus the 6-item Jadad scale. Point estimate ICC(2,1)s and 95% confidence intervals calculated for consistency were similar to the results calculated for absolute agreement.

Insert Table 3 Here

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

DISCUSSION

Overview and discussion of key findings

We investigated interrater and test-retest reliability for student raters with no previous experience in quality assessment. Our study is novel because, to the best of our knowledge, no other research has examined this issue. The raters used the Jadad scale and NOS to assess the quality of studies on the topic of ECT and cognitive impairment. Interrater reliability was generally poor to fair and test-retest reliability was fair to excellent. Our results highlight the need for researchers to consider rater experience during the quality assessment of articles included in systematic reviews.

For interrater reliability, the poor κs on the Jadad scale pertained to the questions about appropriateness of double blinding and the clarity of reporting withdrawals, inclusion/exclusion criteria, and adverse effects. Often, authors did not report methods of blinding and raters had to make judgments about whether to award a point for the question on appropriateness of double blinding. Despite what we communicated during the training session, some raters may have given authors the benefit of the doubt and awarded the point for appropriateness if studies simply reported double blinding, even though another question on the Jadad scale already asked whether authors reported their studies as blinded. Similarly, differences in rater opinion regarding what constitutes an 'adequate' description of withdrawals, inclusion/exclusion criteria, or adverse effects led to poor agreement on these questions. To improve interrater agreement among inexperienced raters, we suggest a pilot phase wherein raters rate the quality of a subsample of articles to allow for the identification and clarification of areas of ambiguity.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

We recognize that any strategy to improve reliability will be limited by instrument content and structure. Scales with larger numbers of interpretive questions will likely have lower reliability than scales with fewer interpretive questions, regardless of the efforts made to improve reliability.

With regard to the NOS, question-specific interrater reliability was poorer than that of the Jadad scale. We believe the NOS's poor reliability may be explained in part by differences in how raters answered interpretive questions, e.g., whether exposed cohorts are somewhat or truly representative of the average exposed person in the community (first question on NOS cohort).

Poor question-specific interrater agreement on the NOS also reflects an inherent challenge with rating the quality of observational studies compared to RCTs. This challenge is exemplified by the multiplicity of tools that exist to assess observational study quality. Two systematic reviews[26, 27] each found over 80 such tools, which varied in design and content. Despite the cornucopia of tools, no gold standard scale exists to rate the quality of observational studies.[28]

Rater disagreements on interpretive questions and inherent challenges with assessing observational study quality explain the negative κs that were calculated for some NOS questions. Negative κs result when agreement occurs less often than predicted by chance alone. This suggests genuine disagreement between raters or an underlying issue with the instrument itself.[29] Indeed, Hartling et al. reported that raters had difficulty using the NOS because of uncertainty over the meaning of certain questions (e.g., representativeness of the exposed cohort, selection of non-exposed cohort) and response options (e.g., 'truly' versus 'somewhat' exposed).[18] These

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

difficulties existed despite Hartling et al.'s use of a pilot training phase. Our raters' difficulties with the interpretative questions might have been a function of issues with the NOS, which could be related to the broader challenge of assessing the quality of observational studies.

Question-specific differences between raters also led to poor interrater agreement on total scores for the Jadad scale and NOS cohort. This may not be evident by comparing the κs and ICC(2,1)s calculated for Jadad. κs for four of eight Jadad questions were moderate yet the ICC(2,1) for total score was poor. However, since total scores are computed using raters' answers to all of the questions on a scale (some answers are awarded one point and others zero points), raters who disagree on small numbers of questions (e.g., two of eight questions) will nonetheless show poor agreement on total scores.

Conversely, for the NOS case-control, κs for six of eight questions were poor yet the ICC(2,1) was fair. In this situation, no 'reliability' relation exists between responses to questions and total scores. For example, rater 1 might answer 'yes' (one point per 'yes' response) and rater 2 might answer 'no' (zero points per 'no' response) to even-numbered questions. For odd-numbered questions, the pattern is reversed. Assuming eight questions, interrater reliability at the question level will be poor because the raters did not agree on their responses, but their overall scores will be equivalent.

Many authors base their discussions of study quality in systematic reviews on raters' responses to individual questions on quality assessment scales. Given that we found generally poor interrater reliability on answers to questions, the process of resolving conflicts between raters becomes important. Many reviews simply report that

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

raters solved disagreements by consensus without describing specific procedures. We speculate that conflict resolution may occasionally be approached in an ad hoc nature or treated as a nuisance to be dealt with as expeditiously as possible. We suggest the process of conflict resolution should be more of a formalized endeavour requiring raters to set aside some 'resolution time' and articulate their reasons for choosing specific answers. In the event the raters do not agree, a third party may be asked to listen to each rater's opinion and make a decision. Although space restrictions in journals might prevent authors from reporting such procedures (when they exist) in manuscripts, the move toward publication of systematic review protocols, for example as mandated by the United States Agency for Healthcare Research and Quality's Effective Health Care Program,[30] provides authors with an opportunity to elaborate on their consensus processes.

Test-retest reliability was better than interrater reliability. Individual raters appeared to adopt a uniform approach to assessing the quality of articles assigned to them. Each rater had her or his own understanding of the interpretive questions and applied this point-of-view consistently throughout the rating process. The issue was the difference in interpretations between raters.

Comparison with other studies

To the best of our knowledge, no other study has examined interrater and test-retest reliability for a group of novice student quality assessors. Two published studies[31, 32] of rater agreement included persons with different levels of experience, although the focus was on extraction of article data (e.g., info on study design, sample characteristics, length of follow-up, definition of outcome, and results) rather than quality assessment.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Horton et al. classified rater experience as minimal, moderate, or substantial and asked raters to extract data from three studies on insomnia therapy.[31] They found no statistically significant differences in error rates according to experience. Hayward et al. trained two experienced raters and one inexperienced rater to independently extract data from seven studies.[32] Agreement between raters was largely perfect.

A recent AHRQ methods report had 16 raters assess the quality of 131 cohort studies using the NOS. Rater experience ranged from four months to 10 years; 13 raters had formal training in systematic reviews.[18] κs were less than 0.50 for eight of nine NOS questions, although the authors did not break down their results by rater experience.

Oremus et al. examined the interrater reliability of the Jadad scale using three raters (two experienced faculty members and one inexperienced PhD student), who read the methods and results of 42 Alzheimer's disease drug trials.[19] The ICC(2,1) for total scores on the Jadad scale was 0.90. Al-Harbi et al. engaged two paediatric surgeons to rate 46 cohort studies that were presented at Canadian Association of Pediatric Surgeons annual meetings and later published in the Journal of Pediatric Surgery.[12] The authors did not specify whether the surgeons received training in quality assessment. The ICC between surgeons, calculated on NOS total scores, was 0.94.

The lower interrater reliability of the novice student raters in this study, compared to the raters in the Oremus et al.[19] and Al-Harbi et al.[12] studies, may be explained by topic familiarity and similarity of expertise. The faculty raters in the Oremus et al. study had previously worked on a systematic review of Alzheimer's disease medications and their expertise lay in two domains of epidemiology, i.e., neuroepidemiology and pharmacoepidemiology. The paediatric surgeons in Al-Harbi et al. may have possessed at

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

least a general familiarity with the types of cohort studies conducted in their speciality. These characteristics may have predisposed the raters to adopt more uniform opinions on the questions contained in Jadad and NOS. In contrast, the novice student raters in our study had for the most part not been exposed to systematic reviews and quality assessment in the past. Also, seven of these raters were recent entrants to graduate school and they came from a variety of undergraduate backgrounds such as medicine, psychology, and basic science.

Limitations

Readers should exercise caution when generalising the results of our study to other types of raters. Reliability could differ according to raters' disciplines and levels of training. Reliability in our study also could have been affected by the specific training program we gave to the students. Additionally, the 10 student raters in this study were a convenience sample that might not represent all raters with similar disciplines and training.

We did not compare the students' rankings with the rankings of more experienced raters (e.g., faculty who conduct systematic reviews). Thus, we could not assess the relative differences in reliability between experienced raters and inexperienced students.

Reliability is also partly a function of the instruments used in the quality assessment. Indeed, instruments with many interpretive questions (e.g., appropriateness of randomization and double-blinding, representativeness of exposed cohort, or adequacy of case definition) could have poor reliability despite several phases of training.

Furthermore, the topic under study could influence reliability, as could certain methodological decisions related to the systematic review. For example, the systematic review of ECT and cognition, upon which we based this study, included 28 papers

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

published prior to 1990. Since the style of reporting in older papers does not always facilitate quality assessment or data extraction, systematic reviews that include older papers could present challenges for maintaining acceptable levels of interrater and test-retest reliability.

Conclusions

In conclusion, we asked a group of 10 novice students to rate the quality of 78 articles that contained data on cognitive impairment following the use of ECT to treat major depressive disorder. Overall, interrater reliability on the Jadad scale and NOS was poor to fair and test-retest reliability was fair to excellent. We trained the raters prior to the quality assessment exercise yet interrater agreement was low for several questions that required a certain degree of interpretation to answer. This was especially so for the NOS and underscores an inherent greater difficulty with assessing the quality of observational studies compared to RCTs.

In addition to standardized training prior to commencing quality assessment, a pilot rating phase may also be necessary to discuss scale questions that generate disagreement among novice student raters. This procedure could help the raters develop standardized interpretations to minimize disagreement.

While the Cochrane Collaboration has stated that quality scales and scale scores are inappropriate means of ascertaining study quality,[33] our results are relevant because many researchers continue to use the Jadad scale and NOS in their systematic reviews. Indeed, our work suggests an area of future research. The Cochrane Collaboration has proposed a 'risk of bias' tool to assess the quality of RCTs.[33] The reliability of the risk of bias tool should be assessed in raters with different levels of experience.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Acknowledgements: Special thanks to Eleanor Pullenayegum and Harry Shannon for their helpful comments on an earlier draft of this manuscript.

Contributors: MO and CO conceived and designed the study. MO analysed the data. MO, CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team interpreted the data. MO drafted the manuscript. CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript.

Funding: This study did not receive funds from any sponsor. No person or organization beyond the authors had any input in study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication.

Competing interests: The authors have no competing interests to declare.

Ethics approval: Not required.

Provenance and peer review: Not commissioned; externally peer reviewed.

Data sharing: No additional data available.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

REFERENCES

1. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.

2. Agency for Healthcare Research and Quality (AHRQ). *Systems to Rate the Strength of Scientific Evidence.* Evidence Report/Technology Assessment No. 47. Rockville, MD: Agency for Healthcare Research and Quality, 2002.

3. Verhagen AP, de Vet HC, de Bie RA, *et al*. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651-4.

4. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.

5. Wells GA, Shea B, O'Connell D, *et al*. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (2 April 2012).

6. Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

7. Chalmers TC, Smith H, Jr., Blackburn B, *et al*. A method for assessing the quality of a randomized control trial. *Control Clin Trials* 1981;2:31-49.

8. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. www.cochrane-handbook.org (23 April 2012).

9. Maher CG, Sherrington C, Herbert RD, *et al*. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther* 2003;83:713-21.

10. Kocsis JH, Gerber AJ, Milrod B, *et al*. A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Compr Psychiatry* 2010;51:319-24.

11. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377-84.

12. Al-Harbi K, Farrokhyar F, Mulla S, *et al*. Classification and appraisal of the level of clinical evidence of publications from the Canadian Association of Pediatric Surgeons for the past 10 years. *J Pediatr Surg* 2009;44:1013-7.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

13. Berard A, Andreu N, Tetrault J, *et al*. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol* 2000;10:498-503.

14. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.

15. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6:e17242.

16. Tooth L, Bennett S, McCluskey A, *et al*. Appraising the quality of randomized controlled trials: inter-rater reliability for the OTseeker evidence database. *J Eval Clin Pract* 2005;11:547-55.

17. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;135:982-9.

18. Hartling L, Hamm M, Milne A, *et al*. *Validity and Inter-rater Reliability Testing of Quality Assessment Instruments*. Rockville, MD: Agency for Healthcare Research and Quality, 2012.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

19. Oremus M, Wolfson C, Perrault A, *et al*. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord* 2001;12:232-6.

20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Manag* 1960;20:37-46.

21. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd edn. Hoboken, NJ: John Wiley & Sons, 2003.

22. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.

24. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th edn. Oxford: Oxford University Press, 2008.

25. Fleiss J. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

26. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol* 2010;63:1061-70.

27. 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;36:666-76.

28. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. *Int J Evid Based Healthc* 2010;8:247.

29. Juurlink DN, Detsky AS. Kappa statistic. *CMAJ* 2005;173:16.

30. Agency for Healthcare Research and Quality (AHRQ). Effective Health Care Program. Rockville, MD: Agency for Healthcare Research and Quality. http://www.effectivehealthcare.ahrq.gov (2 April 2012).

31. Horton J, Vandermeer B, Hartling L, *et al*. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 2010;63:289-98.

32. Haywood KL, Hargreaves J, White R, *et al*. Reviewing measures of outcome: reliability of data extraction. *J Eval Clin Pract* 2004;10:329-37.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

33. Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

TABLES

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Randomization	0.50 (-1.00 to 1.00)	Representative-ness of exposed cohort	-0.13 (-0.36 to 0.11)	Case definition adequate	1.00 (1.00 to 1.00)
Appropriate randomization	0.56 (0.29 to 0.83)	Selection of non-exposed cohort	-0.14 (-0.28 to 0.00)	Cases representative	-0.20 (-0.49 to 0.09)
Double-blind	0.41 (0.16 to 0.66)	Exposure ascertainment	0.00 (0.00 to 0.00)	Control selection	0.25 (-0.19 to 0.69)
Appropriate double-blind	0.17 (-0.07 to 0.41)	Outcome not present at baseline	0.20 (-0.33 to 0.73)	Control definition	0.14 (-0.54 to 0.82)
Description of withdrawals	0.21 (-0.02 to 0.45)	Comparability of cohorts	0.12 (-0.23 to 0.47)	Case and control comparability	0.00 (0.00 to 0.00)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question (continued)

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Description of inclusion / exclusion criteria	0.27 (-0.03 to 0.57)	Outcome assessment	0.31 (-0.08 to 0.69)	Exposure ascertainment	-0.11 (-0.68 to 0.46)
Description of adverse effects	0.13 (-0.11 to 0.37)	Follow-up long enough	-0.09 (-0.22 to 0.04)	Same ascertainment method for cases and controls	0.60 (-0.07 to 1.00)
Description of statistical analysis	0.49 (0.21 to 0.77)	Follow-up adequate	0.39 (-0.02 to 0.81)	Non-response rate	-0.11 (-0.65 to 0.43)

CI, confidence interval; NOS, Newcastle-Ottawa Scale.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 2 Interrater reliability for Jadad and Newcastle-Ottawa scales: total scale scores within rater pairs

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad – 6-item	0.32 (0.08 to 0.53)	0.32 (0.08 to 0.52)
Jadad – 3-item	0.35 (0.11 to 0.56)	0.35 (0.11 to 0.56)
Newcastle-Ottawa – Cohort	-0.19 (-0.63 to 0.34)	-0.19 (-0.67 to 0.35)
Newcastle-Ottawa – Case-control	0.55 (-0.18 to 0.89)	0.46 (-0.13 to 0.92)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 3 Test-retest reliability for Jadad and Newcastle-Ottawa scales: comparison of total scale scores for individual raters after two assessments

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad – 6-item	0.56 (0.42 to 0.67)	0.55 (0.41 to 0.67)
Jadad – 3-item	0.67 (0.55 to 0.76)	0.67 (0.55 to 0.76)
Newcastle-Ottawa – Cohort	0.61 (0.24 to 0.82)	0.62 (0.25 to 0.83)
Newcastle-Ottawa – Case-control	0.85 (0.55 to 0.95)	0.83 (0.48 to 0.95)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales

Mark Oremus[1,2], Carolina Oremus[3,4], Geoffrey B.C. Hall[3,4], Margaret C. McKinnon[3,4], ECT & Cognition Systematic Review Team*[3,4]

[1]McMaster Evidence-based Practice Centre, McMaster University, Hamilton, Ontario, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[3]McMaster Integrative Neuroscience Discovery and Study (MINDS) Program, Hamilton, Ontario, Canada
[4]Department of Psychiatry and Behavioural Neuroscience, Hamilton, Ontario, Canada

*The ECT & Cognition Systematic Review Team includes Allyson Graham, Caitlin Gregory, Gagan Fervaha, Lindsay Hanford, Anthony Nazarov, Melissa Parlar, Maria Restivo, Erica Tatham, and Wanda Truong.

Corresponding author
Mark Oremus
Department of Clinical Epidemiology and Biostatistics
McMaster University
50 Main Street East – Room 308
Hamilton, Ontario, Canada
L8N 1E9
Tel.: 905-525-9140, x22437
Fax: 905-522-7681
E-mail: oremusm@mcmaster.ca

Keywords: systematic review; quality assessment; reliability; Jadad scale; Newcastle-Ottawa Scale.

Word Count: 3,~~143~~044

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

ABSTRACT

Introduction: Quality assessment of included studies is an important component of systematic reviews.

Objective: We investigated interrater and test-retest reliability for quality assessments ~~of study quality~~ conducted by inexperienced student raters.

Design: Student raters received a training session o~~i~~n quality assessment using the Jadad scale for randomized controlled trials and the Newcastle-Ottawa Scale (NOS) for observational studies. Raters were randomly assigned into five pairs and they each independently rated the quality of 13-20 articles. These articles were drawn from a pool of 78 papers examining cognitive impairment following electroconvulsive therapy to treat major depressive disorder. The articles were randomly distributed to the raters. Two months later, each rater re-assessed the quality of half of ~~these~~ their assigned articles.

Setting: ~~University program (~~McMaster Integrative Neuroscience Discovery and Study ~~[~~(MINDS~~]~~) Program~~)~~.

Participants: 10 students ~~(seven graduate, three undergraduate)~~ taking MINDS Program courses.

Main Outcome Measures: ~~Main outcomes included~~We measured interrater reliability ~~measured~~ using Kappa and the intraclass correlation coefficient type 2,1, or ICC(2,1). We measured test-retest reliability using ICC(2,1). ~~To assess differences in total score on the Jadad scale, we calculated mean differences in total score for each rater pair and individual rater.~~

Results: ~~Interrater reliability was generally poor and test-retest reliability was fair to excellent.~~ Interrater reliability varied by scale question. For the 6-item Jadad scale,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

question-specific Kappas ranged from 0.13 ([95% confidence interval] -0.11 to 0.37) to 0.56 (0.29 to 0.83). The ranges were -0.14 (-0.28 to 0.00) to 0.39 (-0.02 to 0.81) for the NOS cohort and -0.20 (-0.49 to 0.09) to 1.00 (1.00 to 1.00) for the NOS case-control. For overall scores on the 6-item Jadad scale, ICC(2,1)s for interrater and test-retest reliability (accounting for systematic differences~~Mean differences in total scale score within rater pairs and within individual~~ between raters) were 0.32 (0.08 to 0.52) and 0.55 (0.41 to 0.67) respectively. Corresponding ICC(2,1)s for the NOS cohort were -0.19 (-0.67 to 0.35) and 0.62 (0.25 to 0.83), and for the NOS case-control the ICC(2,1)s were 0.46 (-0.13 to 0.92) and 0.83 (0.48 to 0.95)~~raters on the Jadad scale were minor and not statistically significantly different from zero, except for one rater's mean test-retest difference~~.

Conclusion: Interrater reliability was generally poor to fair and test-retest reliability was fair to excellent. ~~Agreement between raters was lower for 'interpretive' questions on the Jadad and NOS (e.g., questions asking about appropriateness of double-blinding or representativeness of exposed cases).~~ A pilot rating phase following rater training may be one way to improve agreement.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

ARTICLE SUMMARY

Article focus

- To examine the interrater and test-retest reliability of inexperienced raters' quality assessments of articles included in a systematic review.

Key messages

- Among inexperienced raters, interrater reliability using the Jadad scale and Newcastle-Ottawa Scale was generally poor to fair; test-retest reliability was fair to excellent.
- Systematic reviewers must pay special attention to training inexperienced quality raters; a pilot rating phase might be a helpful means of improving reliability among inexperienced raters, especially when rating observational study quality.

Strengths and limitations of this study

- No other study has examined the reliability of quality assessments in a group of inexperienced raters.
- Results may differ depending on rater background and experience, rater training, quality assessment instruments, and topic under study.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

INTRODUCTION

Systematic reviews summarize healthcare research evidence and they are useful for assessing whether treatment benefits outweigh risks.[1, 2] Accordingly, conclusions drawn from systematic reviews may impact clinical care and patient outcomes, thereby necessitating high standards of methodological rigour.

One critical component of conducting systematic reviews involves evaluation of the methodological quality of included studies. Study quality may influence treatment effect estimates and the validity of conclusions drawn from such estimates.[3] Through quality assessment, researchers identify strengths and weaknesses of existing evidence[4] and suggest ways to improve future research.

Careful work has identified key quality assessment domains.[1, 5] For randomized controlled trials (RCTs), these domains include appropriate generation of random allocation sequences, concealment of allocation sequences, blinding (of participants, health care providers, data collectors, and outcome assessors), and reporting of proportions of patients lost to follow-up.[1] For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment,-[5] as well as selection and attrition biases.

Numerous scales exist to help raters assess study quality.[5-11] The majority of these scales list quality assessment domains and require raters to indicate whether each domain is present or absent from the studies under consideration. Some scales (e.g., Jadad,[6] Newcastle-Ottawa Scale [NOS][5]) assign points when quality domains are present, thus permitting the calculation of overall 'quality scores'. Other scales (e.g., risk of bias[8]) ask raters to rank the degree of bias (high, low, unclear) associated with each

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

quality domain.

Generally, quality scales demonstrate good interrater and test-retest reliability. Reliability coefficients such as Kappa (κ) are typically greater than 0.60,[9-17] although recent work reports κs of less than 0.50 for eight of nine questions on the NOS.[18]

Although quality assessment is now regarded as a standard component of systematic reviews, one issue that has received little attention in the literature is the effect of rater experience on the reliability of quality assessments. This issue is important because raters may be drawn from vast pools of persons with varying degrees of methods expertise, from experienced faculty to inexperienced students.

~~As part of an ongoing meta-analysis of electroconvulsive therapy (ECT) and cognitive impairment, w~~We investigated interrater and test-retest reliability for student raters with no previous experience in the quality assessment of randomized controlled trials (RCTs) and observational studies. To the best of our knowledge, no other study has examined this topic.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

METHODS

Study design

In an ongoing systematic review of cognitive impairment following electroconvulsive therapy (ECT) to treat major depressive disorder, ~~We retrieved~~ 78 published articles ~~that contained data on cognitive impairment following the use of ECT to treat major depressive disorder~~passed title and abstract, and full-text, screening. These articles formed the basis of this study. Fifty-five of the articles reported the results of randomized controlled trials (RCTs), with one article containing results of five separate studies and two other articles each containing results of two separate studies, for a total of 61 RCTs. Fifteen articles reported on cohort studies and eight reported on case-control studies. Eleven articles were published prior to 1980, 17 between 1980 and 1989, 15 between 1990 and 1999, and 35 since 2000.

We invited all 10 students (three undergraduate, seven graduate) taking a 'special topics' course in the McMaster Integrative Neuroscience Discovery and Study Program to participate in this study. All 10 students accepted the invitation. One author (MO) with systematic review experience trained ~~10~~ the students ~~(three undergraduate, seven graduate)~~ to rate the methodological quality of published study reports using the 6-item Jadad scale for RCTs[6, 19] and the NOS for observational studies.[5] Training consisted of a 90-minute didactic session divided into two parts: part one highlighted the importance of quality assessment in systematic reviews and part two contained a question-by-question description of the Jadad and NOS instruments. We provided a standardized tabular spreadsheet for student raters to use during quality assessment.

We used a random number table to assign the student raters into five pairs and we

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

~~subsequently~~ randomly distributed between 13 and 20 articles to each pair. None of the 78 articles was assigned to more than one pair; pairs received a mix of RCTs and observational studies. The number of a~~A~~rticles assigned to the pairs depended ~~fluctuated across pairs because of constraints~~ on the amount of time each rater could devote to this study~~rater availability due to competing academic demands~~.

Raters determined the type of study design (i.e., RCT or observational) for each of their assigned articles and one author (CO) verified their choices. Raters then independently rated their assigned articles to permit us to examine interrater reliability.

Statistical analysis

We used κ[20, 21] to measure interrater reliability for individual Jadad and NOS questions. We interpreted κ values as follows: greater than 0.80 was very good, 0.61 to 0.80 was good, 0.41 to 0.60 was moderate, 0.21 to 0.40 was fair, and less than 0.21 was poor.[22]

For test-retest reliability, each rater re-assessed half of the articles to which they had been assigned during the interrater reliability phase. The re-assessments took place two months after the interrater reliability phase[13] to minimize the possibility that recall of first assessments would influence the second assessments.

We employed the intraclass correlation coefficient-model 2,1, or ICC(2,1),[23] to measure interrater and test-retest reliability for the Jadad and NOS total scores. We computed separate ICC(2,1) values for consistency (systematic differences between raters are considered irrelevant) and absolute agreement (systematic differences between raters are considered relevant).[24] ICC(2,1) values were interpreted as follows: greater than 0.75 was excellent, 0.40 to 0.75 was fair to good, and less than 0.40 was poor.[25]

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

We calculated two sets of ICC(2,1)s for the Jadad scale. The first set pertained to the 6-item Jadad scale [19] and the second set pertained to the original 3-item Jadad scale [6].

~~To investigate the differences in total Jadad scale scores within rater pairs (interrater) and within individual raters (test-retest), we calculated a mean difference in score for each rater pair and individual rater. We compared differences in score within rater pairs using the Wilcoxon rank-sum test and within individual raters using the Wilcoxon signed-rank test. We did not conduct this analysis for the NOS because of the small number of cohort and case-control studies.~~

~~We did not pool mean differences since we did not expect to find a pooled estimate that would be different from zero. This is because the ordering within pairs was arbitrary (i.e. whether differences were calculated as rater1-rater2 or vice-versa).~~

SAS v9.2 (The SAS Institute, Cary, NC) was utilized to calculate κ ~~and p-values for the Wilcoxon tests~~; SPSS v20 (IBM Corp., Armonk, NY) was used to calculate ICC(2,1). The level of significance was α=0.05.

Formatted: Comment Text, Indent: First line: 0.39", Widow/Orphan control

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

RESULTS

Interrater reliability

For interrater reliability, agreement between raters on individual questions was generally poor (Table 1). Half of the questions on the Jadad scale had moderate κs and the other half had poor κs. On the NOS, all κs were poor for the cohort study questions (NOS cohort) and six of eight κs were poor for the case-control study questions (NOS case-control).

*****Insert Table 1 Here*****

Examining total scale scores within rater pairs (Table 2), agreement was poor for the Jadad scale (6- and 3-item versions) and NOS cohort and fair for the NOS case-control. However, point estimate ICC(2,1)s for the NOS cohort and case-control were not statistically significantly different from zero. Point estimate ICC(2,1)s and 95% confidence intervals did not appreciably differ according to calculation based on consistency or absolute agreement.

*****Insert Table 2 Here*****

~~The mean differences in total score on the Jadad scale within rater pairs ranged from 0.00 to 0.70; no difference was statistically significantly different from zero (Table 3).~~

~~*****Insert Table 3 Here*****~~

Test-retest reliability

Test-retest reliability following a two-month interval between assessments was fair to good for the Jadad scale and NOS cohort and excellent for the NOS case-control (Table ~~4~~3). Test-retest reliability was slightly higher for the 3-item Jadad scale versus the 6-item

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Jadad scale. Point estimate ICC(2,1)s and 95% confidence intervals calculated for consistency were similar to the results calculated for absolute agreement.

*****Insert Table ~~4~~ 3 Here*****

~~The mean differences in total score on the Jadad scale within individual raters, subtracting scores at the second assessment from scores at the first assessment, ranged from 0.00 to 0.64 for nine of the raters (Table 3). None of these differences were statistically significantly different from zero. For one rater, the mean difference in total score was 3.38 (p=0.01).~~

~~**Mean Differences in Total Score: Newcastle-Ottawa Scale**~~

~~Although we did not apply formal statistical hypothesis testing to mean differences in total score on the NOS, the data suggest larger differences compared to the Jadad scale. Mean differences on the NOS cohort ranged from 0.25 to 3.00 (rater pairs) and 0.00 to 1.67 (individual raters). On the NOS case-control, mean differences spanned from 0.50 to 2.00 (rater pairs) and 0.00 to 1.00 (individual raters).~~

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

DISCUSSION

Overview and discussion of key findings

We investigated interrater and test-retest reliability for student raters with no previous experience in quality assessment. Our study is novel because, to the best of our knowledge, no other research has examined this issue. The raters used the Jadad scale and NOS to assess the quality of studies on the topic of ECT and cognitive impairment. Interrater reliability was generally poor to fair and test-retest reliability was fair to excellent. Our results highlight the need for researchers to consider rater experience during the quality assessment of articles included in systematic reviews.

For interrater reliability, the poor κs on the Jadad scale pertained to the questions about appropriateness of double blinding and the clarity of reporting withdrawals, inclusion/exclusion criteria, and adverse effects. Often, authors did not report methods of blinding and raters had to make judgments about whether to award a point for the question on appropriateness of double blinding. Despite what we communicated during the training session, some raters may have given authors the benefit of the doubt and awarded the point for appropriateness if studies simply reported double blinding, even though another question on the Jadad scale already asked whether authors reported their studies as blinded. Similarly, differences in rater opinion regarding what constitutes an 'adequate' description of withdrawals, inclusion/exclusion criteria, or adverse effects led to poor agreement on these questions. To improve interrater agreement among inexperienced raters, we suggest a pilot phase wherein raters rate the quality of a subsample of articles to allow for the identification and clarification of areas of ambiguity.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

We recognize that any strategy to improve reliability will be limited by instrument content and structure. Scales with larger numbers of interpretive questions will likely have lower reliability than scales with fewer interpretive questions, regardless of the efforts made to improve reliability.

With regard to the NOS, question-specific interrater reliability was poorer than that of the Jadad scale. We believe the NOS's poor reliability may be explained in part by differences in how raters answered interpretive questions, e.g., whether exposed cohorts are somewhat or truly representative of the average exposed person in the community (first question on NOS cohort).

Poor question-specific interrater agreement on the NOS also reflects an inherent challenge with rating the quality of observational studies compared to RCTs. This challenge is exemplified by the multiplicity of tools that exist to assess observational study quality. Two systematic reviews[26, 27] each found over 80 such tools, which varied in design and content. Despite the cornucopia of tools, no gold standard scale exists to rate the quality of observational studies.[28]

Rater disagreements on interpretive questions and inherent challenges with assessing observational study quality explain the negative κs that were calculated for some NOS questions. Negative κs result when agreement occurs less often than predicted by chance alone. This suggests genuine disagreement between raters or an underlying issue with the instrument itself.[29] Indeed, Hartling et al. reported that raters had difficulty using the NOS because of uncertainty over the meaning of certain questions (e.g., representativeness of the exposed cohort, selection of non-exposed cohort) and response options (e.g., 'truly' versus 'somewhat' exposed).[18] These

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

difficulties existed despite Hartling et al.'s use of a pilot training phase. Our raters' difficulties with the interpretative questions might have been a function of issues with the NOS, which could be related to the broader challenge of assessing the quality of observational studies.

Question-specific differences between raters also led to poor interrater agreement on total scores for the Jadad scale and NOS cohort. This may not be evident by comparing the κs and ICC(2,1)s calculated for Jadad. κs for four of eight Jadad questions were moderate yet the ICC(2,1) for total score was poor. However, since total scores are computed using raters' answers to all of the questions on a scale (some answers are awarded one point and others zero points), raters who disagree on small numbers of questions (e.g., two of eight questions) will nonetheless show poor agreement on total scores.

Conversely, for the NOS case-control, κs for six of eight questions were poor yet the ICC(2,1) was fair. In this situation, no 'reliability' relation exists between responses to questions and total scores. For example, rater 1 might answer 'yes' (one point per 'yes' response) and rater 2 might answer 'no' (zero points per 'no' response) to even-numbered questions. For odd-numbered questions, the pattern is reversed. Assuming eight questions, interrater reliability at the question level will be poor because the raters did not agree on their responses, but their overall scores will be equivalent.

Many authors base their discussions of study quality in systematic reviews on raters' responses to individual questions on quality assessment scales. Given that we found generally poor interrater reliability on answers to questions, the process of resolving conflicts between raters becomes important. Many reviews simply report that

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

raters solved disagreements by consensus without describing specific procedures. We speculate that conflict resolution may occasionally be approached in an ad hoc nature or treated as a nuisance to be dealt with as expeditiously as possible. We suggest the process of conflict resolution should be more of a formalized endeavour requiring raters to set aside some 'resolution time' and articulate their reasons for choosing specific answers. In the event the raters do not agree, a third party may be asked to listen to each rater's opinion and make a decision. Although space restrictions in journals might prevent authors from reporting such procedures (when they exist) in manuscripts, the move toward publication of systematic review protocols, for example as mandated by the United States Agency for Healthcare Research and Quality's Effective Health Care Program,[30] provides authors with an opportunity to elaborate on their consensus processes.

Test-retest reliability was better than interrater reliability. Individual raters appeared to adopt a uniform approach to assessing the quality of articles assigned to them. Each rater had her or his own understanding of the interpretive questions and applied this point-of-view consistently throughout the rating process. The issue was the difference in interpretations between raters.

Comparison with other studies

To the best of our knowledge, no other study has examined interrater and test-retest reliability for a group of novice student quality assessors. Two published studies[31, 32] of rater agreement included persons with different levels of experience, although the focus was on extraction of article data (e.g., info on study design, sample characteristics, length of follow-up, definition of outcome, and results) rather than quality assessment.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Horton et al. classified rater experience as minimal, moderate, or substantial and asked raters to extract data from three studies on insomnia therapy.[31] They found no statistically significant differences in error rates according to experience. Hayward et al. trained two experienced raters and one inexperienced rater to independently extract data from seven studies.[32] Agreement between raters was largely perfect.

A recent AHRQ methods report had 16 raters assess the quality of 131 cohort studies using the NOS. Rater experience ranged from four months to 10 years; 13 raters had formal training in systematic reviews.[18] κs were less than 0.50 for eight of nine NOS questions, although the authors did not break down their results by rater experience.

Oremus et al. examined the interrater reliability of the Jadad scale using three raters (two experienced faculty members and one inexperienced PhD student), who read the methods and results of 42 Alzheimer's disease drug trials.[19] The ICC(2,1) for total scores on the Jadad scale was 0.90. Al-Harbi et al. engaged two paediatric surgeons to rate 46 cohort studies that were presented at Canadian Association of Pediatric Surgeons annual meetings and later published in the Journal of Pediatric Surgery.[12] The authors did not specify whether the surgeons received training in quality assessment. The ICC between surgeons, calculated on NOS total scores, was 0.94.

The lower interrater reliability of the novice student raters in this study, compared to the raters in the Oremus et al.[19] and Al-Harbi et al.[12] studies, may be explained by topic familiarity and similarity of expertise. The faculty raters in the Oremus et al. study had previously worked on a systematic review of Alzheimer's disease medications and their expertise lay in two domains of epidemiology, i.e., neuroepidemiology and pharmacoepidemiology. The paediatric surgeons in Al-Harbi et al. may have possessed at

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

least a general familiarity with the types of cohort studies conducted in their speciality. These characteristics may have predisposed the raters to adopt more uniform opinions on the questions contained in Jadad and NOS. In contrast, the novice student raters in our study had for the most part not been exposed to systematic reviews and quality assessment in the past. Also, seven of these raters were recent entrants to graduate school and they came from a variety of undergraduate backgrounds such as medicine, psychology, and basic science.

Limitations

Readers should exercise caution when generalising the results of our study to other types of raters ~~or scenarios~~. Reliability could differ according to raters' disciplines and levels of training. Reliability in our study also could have been affected by the specific training program we gave to the students. Additionally, the 10 student raters in this study were a convenience sample that might not represent all raters with similar disciplines and training.

We did not compare the students' rankings with the rankings of more experienced raters (e.g., faculty who conduct systematic reviews). Thus, we could not assess the relative differences in reliability between experienced raters and inexperienced students.

~~even among groups of inexperienced students.~~ Reliability is also partly a function of the instruments used in the quality assessment. Indeed, instruments with many interpretive questions (e.g., appropriateness of randomization and double-blinding, representativeness of exposed cohort, or adequacy of case definition) could have poor reliability despite several phases of training.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Furthermore, the topic under study could influence reliability, as could certain ~~parameters of the systematic review~~ methodological decisions related to the systematic review~~s~~. For example, the ~~meta analysis on~~systematic review of ECT and cognition, upon which we based this study, included 28 papers published prior to 1990. ~~The~~ Since the style of reporting ~~results~~ in older papers does not always facilitate quality assessment or data extraction, s~~. S~~ystematic reviews that include older papers could ~~therefore~~ present challenges for maintaining acceptable levels of interrater and test-retest reliability.

Conclusions

In conclusion, we asked a group of 10 novice students to rate the quality of 78 articles that contained data on cognitive impairment following the use of ECT to treat major depressive disorder. Overall, interrater reliability on the Jadad scale and NOS was poor to fair and ~~, although~~ test-retest reliability ~~ranged from~~was fair to excellent. We trained the raters prior to the quality assessment exercise yet interrater agreement was low for several questions that required a certain degree of interpretation to answer. This was especially so for the NOS and underscores an inherent greater difficulty with assessing the quality of observational studies compared to RCTs.

In addition to standardized training prior to commencing quality assessment, a pilot rating phase may also be necessary to discuss scale questions that generate disagreement among novice student raters. This procedure could help the raters develop standardized interpretations to minimize disagreement.

While the Cochrane Collaboration has stated that quality scales and scale scores are inappropriate means of ascertaining study quality,[33] our results are relevant because many researchers continue to use the Jadad scale and NOS in their systematic reviews.

Formatted: Indent: First line: 0.39"

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Indeed, our work suggests an area of future research. The Cochrane Collaboration has proposed a 'risk of bias' tool to assess the quality of RCTs.[33] The reliability of the risk of bias tool should be assessed in raters with different levels of experience.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

Acknowledgements: Special thanks to Eleanor Pullenayegum and Harry Shannon for their helpful comments on an earlier draft of this manuscript.

Contributors: MO and CO conceived and designed the study. MO analysed the data. MO, CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team interpreted the data. MO drafted the manuscript. CO, MCM, GBCH, and the ECT & Cognition Systematic Review Team critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript.

Funding: This study did not receive funds from any sponsor. No person or organization beyond the authors had any input in study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication.

Competing interests: The authors have no competing interests to declare.

Ethics approval: Not required.

Provenance and peer review: Not commissioned; externally peer reviewed.

Data sharing: No additional data available.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

REFERENCES

1. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.

2. Agency for Healthcare Research and Quality (AHRQ). *Systems to Rate the Strength of Scientific Evidence.* Evidence Report/Technology Assessment No. 47. Rockville, MD: Agency for Healthcare Research and Quality, 2002.

3. Verhagen AP, de Vet HC, de Bie RA, *et al*. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651-4.

4. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.

5. Wells GA, Shea B, O'Connell D, *et al*. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (2 April 2012).

6. Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

7. Chalmers TC, Smith H, Jr., Blackburn B, *et al*. A method for assessing the quality of a randomized control trial. *Control Clin Trials* 1981;2:31-49.

8. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. www.cochrane-handbook.org (23 April 2012).

9. Maher CG, Sherrington C, Herbert RD, *et al*. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther* 2003;83:713-21.

10. Kocsis JH, Gerber AJ, Milrod B, *et al*. A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Compr Psychiatry* 2010;51:319-24.

11. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377-84.

12. Al-Harbi K, Farrokhyar F, Mulla S, *et al*. Classification and appraisal of the level of clinical evidence of publications from the Canadian Association of Pediatric Surgeons for the past 10 years. *J Pediatr Surg* 2009;44:1013-7.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

13. Berard A, Andreu N, Tetrault J, *et al*. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol* 2000;10:498-503.

14. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.

15. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6:e17242.

16. Tooth L, Bennett S, McCluskey A, *et al*. Appraising the quality of randomized controlled trials: inter-rater reliability for the OTseeker evidence database. *J Eval Clin Pract* 2005;11:547-55.

17. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;135:982-9.

18. Hartling L, Hamm M, Milne A, *et al*. *Validity and Inter-rater Reliability Testing of Quality Assessment Instruments*. Rockville, MD: Agency for Healthcare Research and Quality, 2012.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only

19. Oremus M, Wolfson C, Perrault A, *et al*. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord* 2001;12:232-6.

20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Manag* 1960;20:37-46.

21. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd edn. Hoboken, NJ: John Wiley & Sons, 2003.

22. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.

24. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th edn. Oxford: Oxford University Press, 2008.

25. Fleiss J. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

26. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol* 2010;63:1061-70.

27. 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;36:666-76.

28. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. *Int J Evid Based Healthc* 2010;8:247.

29. Juurlink DN, Detsky AS. Kappa statistic. *CMAJ* 2005;173:16.

30. Agency for Healthcare Research and Quality (AHRQ). Effective Health Care Program. Rockville, MD: Agency for Healthcare Research and Quality. http://www.effectivehealthcare.ahrq.gov (2 April 2012).

31. Horton J, Vandermeer B, Hartling L, *et al*. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 2010;63:289-98.

32. Haywood KL, Hargreaves J, White R, *et al*. Reviewing measures of outcome: reliability of data extraction. *J Eval Clin Pract* 2004;10:329-37.

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

33. Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

TABLES

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Randomization	0.50 (-1.00 to 1.00)	Representative-ness of exposed cohort	-0.13 (-0.36 to 0.11)	Case definition adequate	1.00 (1.00 to 1.00)
Appropriate randomization	0.56 (0.29 to 0.83)	Selection of non-exposed cohort	-0.14 (-0.28 to 0.00)	Cases representative	-0.20 (-0.49 to 0.09)
Double-blind	0.41 (0.16 to 0.66)	Exposure ascertainment	0.00 (0.00 to 0.00)	Control selection	0.25 (-0.19 to 0.69)
Appropriate double-blind	0.17 (-0.07 to 0.41)	Outcome not present at baseline	0.20 (-0.33 to 0.73)	Control definition	0.14 (-0.54 to 0.82)
Description of withdrawals	0.21 (-0.02 to 0.45)	Comparability of cohorts	0.12 (-0.23 to 0.47)	Case and control comparability	0.00 (0.00 to 0.00)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 1 Interrater reliability for Jadad and Newcastle-Ottawa scales: by question (continued)

Question – Jadad Scale	Kappa (95% CI)	Question – NOS Cohort	Kappa (95% CI)	Question – NOS Case-control	Kappa (95% CI)
Description of inclusion / exclusion criteria	0.27 (-0.03 to 0.57)	Outcome assessment	0.31 (-0.08 to 0.69)	Exposure ascertainment	-0.11 (-0.68 to 0.46)
Description of adverse effects	0.13 (-0.11 to 0.37)	Follow-up long enough	-0.09 (-0.22 to 0.04)	Same ascertainment method for cases and controls	0.60 (-0.07 to 1.00)
Description of statistical analysis	0.49 (0.21 to 0.77)	Follow-up adequate	0.39 (-0.02 to 0.81)	Non-response rate	-0.11 (-0.65 to 0.43)

CI, confidence interval; NOS, Newcastle-Ottawa Scale.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 2 Interrater reliability for Jadad and Newcastle-Ottawa scales: total scale scores within rater pairs

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad – 6-item	0.32 (0.08 to 0.53)	0.32 (0.08 to 0.52)
Jadad – 3-item	0.35 (0.11 to 0.56)	0.35 (0.11 to 0.56)
Newcastle-Ottawa – Cohort	-0.19 (-0.63 to 0.34)	-0.19 (-0.67 to 0.35)
Newcastle-Ottawa – Case-control	0.55 (-0.18 to 0.89)	0.46 (-0.13 to 0.92)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 3 Mean differences in total score on Jadad scale*

Pair	Mean Difference	Rater	Mean Difference
1	0.25 (p=0.46)	1	0.08 (p=1.00)
2	0.30 (p=0.24)	2	0.42 (p=0.81)
3	0.70 (p=0.46)	3	0.64 (p=0.45)
4	0.00 (p=1.00)	4	3.38 (p=0.01)
5	0.47 (p=0.39)	5	0.33 (p=0.25)
		6	0.00 (p=1.00)
		7	0.18 (p=0.81)
		8	0.00 (p=1.00)
		9	0.42 (p=0.26)
		10	0.00 (p=1.00)

*Score range=0-8.

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

Table 4 3 Test-retest reliability for Jadad and Newcastle-Ottawa scales: comparison of total scale scores for individual raters after two assessments

Scale	ICC(2,1) (95% CI) – Consistency*	ICC(2,1) (95% CI) – Absolute Agreement†
Jadad – 6-item	0.56 (0.42 to 0.67)	0.55 (0.41 to 0.67)
Jadad – 3-item	0.67 (0.55 to 0.76)	0.67 (0.55 to 0.76)
Newcastle-Ottawa – Cohort	0.61 (0.24 to 0.82)	0.62 (0.25 to 0.83)
Newcastle-Ottawa – Case-control	0.85 (0.55 to 0.95)	0.83 (0.48 to 0.95)

*ICC(2,1) where systematic differences between raters are irrelevant.

†ICC(2,1) where systematic differences between raters are relevant.

CI, confidence interval; ICC, intraclass correlation coefficient.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

STROBE statement checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract		
	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract (√) (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found (√)
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported (√)
Objectives	3	State specific objectives, including any prespecified hypotheses (√)
Methods		
Study design	4	Present key elements of study design early in the paper (√)
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection (N/A)
Participants	6	(*a*) *Cohort study*?Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up*Case-control study*?Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls*Cross sectional study*?Give the eligibility criteria, and the sources and methods of selection of participants (N/A) (*b*) *Cohort study*?For matched studies, give matching criteria and number of exposed and unexposed*Case-control study*?For matched studies, give matching criteria and the number of controls per case (N/A)
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable (√)
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group (√)
Bias	9	Describe any efforts to address potential sources of

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

		bias (N/A)
Study size	10	Explain how the study size was arrived at (√)
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why (N/A)
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (√) (*b*) Describe any methods used to examine subgroups and interactions (N/A) (*c*) Explain how missing data were addressed (N/A) (*d*) *Cohort study*?If applicable, explain how loss to follow-up was addressed*Case-control study*?If applicable, explain how matching of cases and controls was addressed*Cross sectional study*?If applicable, describe analytical methods taking account of sampling strategy (N/A) (*e*) Describe any sensitivity analyses (N/A)
Results		
Participants	13*	(*a*) Report numbers of individuals at each stage of study?eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (N/A) (*b*) Give reasons for non-participation at each stage (N/A) (*c*) Consider use of a flow diagram (N/A)
Descriptive data	14*	(*a*) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (N/A) (*b*) Indicate number of participants with missing data for each variable of interest (N/A) (*c*) *Cohort study* Summarise follow-up time (eg average and total amount) (N/A)
Outcome data	15*	*Cohort study* Report numbers of outcome events or summary measures over time (N/A) *Case-control study* Report numbers in each exposure category, or summary measures of exposure (N/A) *Cross sectional study* Report numbers of outcome events or summary measures (N/A)
Main results	16	(*a*) Report the numbers of individuals at each stage of the study, eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (N/A) (*b*) Give reasons for non-participation at each stage (N/A)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.

		(*c*) Consider use of a flow diagram (N/A)
Other analyses	17	Report other analyses done, eg, analyses of subgroups and interactions, and sensitivity analyses (N/A)
Discussion		
Key results	18	Summarise key results with reference to study objectives (√)
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias. (√)
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. (√)
Generalisability	21	Discuss the generalisability (external validity) of the study results. (√)
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based. (√)

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2012-001368 on 31 July 2012. Downloaded from http://bmjopen.bmj.com/ on April 10, 2024 by guest. Protected by copyright.