BMJ

RESEARCH

Risk of bias versus quality assessment of randomised controlled trials: cross sectional study

Lisa Hartling, assistant professor Maria Ospina, project manager Yuanyuan Liang, research scientist and biostatistician Donna M Dryden, assistant professor Nicola Hooton, project coordinator Jennifer Krebs Seida, project coordinator Terry P Klassen, professor

Alberta Research Centre for Health Evidence, Department of Pediatrics, University of Alberta, Aberhart Centre One, Edmonton, AB, Canada T6G 2J3

Correspondence to: L Hartling
hartling@ualberta.ca

Cite this as: *BMJ* 2009;339:b4012
doi:10.1136/bmj.b4012

ABSTRACT

Objectives To evaluate the risk of bias tool, introduced by the Cochrane Collaboration for assessing the internal validity of randomised trials, for inter-rater agreement, concurrent validity compared with the Jadad scale and Schulz approach to allocation concealment, and the relation between risk of bias and effect estimates.

Design Cross sectional study.

Study sample 163 trials in children.

Main outcome measures Inter-rater agreement between reviewers assessing trials using the risk of bias tool (weighted κ), time to apply the risk of bias tool compared with other approaches to quality assessment (paired *t* test), degree of correlation for overall risk compared with overall quality scores (Kendall's τ statistic), and magnitude of effect estimates for studies classified as being at high, unclear, or low risk of bias (metaregression).

Results Inter-rater agreement on individual domains of the risk of bias tool ranged from slight (κ=0.13) to substantial (κ=0.74). The mean time to complete the risk of bias tool was significantly longer than for the Jadad scale and Schulz approach, individually or combined (8.8 minutes (SD 2.2) per study *v* 2.0 (SD 0.8), P<0.001). There was low correlation between risk of bias overall compared with the Jadad scores (P=0.395) and Schulz approach (P=0.064). Effect sizes differed between studies assessed as being at high or unclear risk of bias (0.52) compared with those at low risk (0.23).

Conclusions Inter-rater agreement varied across domains of the risk of bias tool. Generally, agreement was poorer for those items that required more judgment. There was low correlation between assessments of overall risk of bias and two common approaches to quality assessment: the Jadad scale and Schulz approach to allocation concealment. Overall risk of bias as assessed by the risk of bias tool differentiated effect estimates, with more conservative estimates for studies at low risk.

INTRODUCTION

Systematic reviews are considered the most comprehensive way of judging whether a treatment "does more good than harm."[1] The methodological quality of studies included in a systematic review can have a substantial impact on estimates of treatment effect, which may affect the validity of the conclusions of a review.[2] Careful consideration and appraisal of the methodological characteristics of the primary studies is an essential feature of systematic reviews. This helps to identify areas of strength and weakness in the existing evidence[3] and to formulate recommendations to improve the conduct and value of future research.

The terms quality, validity, and bias[4] have been used interchangeably in the systematic review literature to describe methodological conditions that are associated with the validity of study results. Traditionally, quality assessment in systematic reviews has primarily involved the appraisal of internal validity—how well the study was designed and executed to prevent systematic errors or bias. Bias can result from flaws in the design, conduct, analysis, interpretation, or reporting of a study. In randomised controlled trials, bias has been classified into four general categories: selection, performance, detection, and attrition.[5]

Control of bias in randomised controlled trials is necessary to reduce the risk of making incorrect conclusions about treatment effects.[6] Several empirical studies have documented how the lack of adequate randomisation, concealment of allocation, double blinding, and differential losses to follow-up or dropouts per treatment group may affect the observed treatment effects.[5 7-11] Several meta-epidemiological studies have examined the effect of certain methodological characteristics and biases of individual randomised controlled trials on the pooled estimates of meta-analyses.[5 7 12] Although the findings have been inconsistent across individual studies, evidence shows that inadequate or unclear allocation concealment and lack of double blinding lead to exaggerated estimates of treatment effects.

The approach to quality assessment in systematic reviews is inconsistent and often debated.[5] The uncertainty about how quality measures are associated with estimates of treatment effect and the absence of a gold standard to assess the validity of randomised controlled trials[13] have resulted in the development of a large number of quality assessment tools.[1 14] Only 12% of the available scales and checklists to assess the methodological

quality of randomised controlled trials have been empirically evaluated.[14] Furthermore, these tools[13 15] often contain elements attributable to reporting (for example, whether the study population was described) and design (for example, whether a sample size calculation was carried out) that are not related to bias.[4]

In February 2008 the Cochrane Collaboration introduced a risk of bias tool to assess the internal validity of randomised controlled trials.[4] The tool was developed to address some of the shortcomings of existing quality assessment instruments. Specifically, it was developed to assess the degree to which the results of a study "should be believed."[4] The choice of components for inclusion in the tool was based on empirical evidence showing their association with effect estimates.[7 8 16] The developers also aimed to distinguish between the actual methods used for carrying out the randomised controlled trials rather than the reporting.

The risk of bias tool is based on six domains: sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, and "other sources of bias." Critical assessments on the risk of bias (high, low, unclear) are made separately for each domain. A final overall assessment within or across studies is based on the responses to individual domains. The assessments are to be made on the basis of the trial report as well as additional documents, such as the study protocol. Those carrying out the assessments are required to record the reasons for their decisions. In this way the rationale for any judgments is documented and transparent.

Although the use of the risk of bias tool has been recommended for systematic reviews done within the Cochrane Collaboration, it has not been validated formally and it is unknown how the tool compares to other approaches currently available to assess the validity of a study. We evaluated the inter-rater agreement of the risk of bias tool, the concurrent validity of the tool compared with the Jadad scale[17] and Schulz[7] approach to allocation concealment, and the relation between overall risk of bias as assessed by the risk of bias tool and study effect estimates. We also compared the time required to apply the risk of bias tool compared with the Jadad scale and Schulz approach.

METHODS

This cross sectional analytical study was carried out on a convenience sample of 163 full manuscripts of randomised controlled trials in child health; these manuscripts resulted from abstracts that were presented at the annual scientific meetings of the Society for Pediatric Research between 1992 and 1995. The trials were part of a previously published project examining publication bias.[18] Their methodological quality had been assessed previously using the Jadad scale and Schulz approach to allocation concealment.[7 17] Likewise, effect estimates for the primary outcome in each trial had been extracted.

Two reviewers (LH, MO) independently evaluated a random sample of 80 randomised controlled trials to assess the time to complete the risk of bias tool. This preliminary evaluation also helped to develop some guidelines for application of the tool to the entire sample of trials. One reviewer (NH) recorded the time required to apply the Jadad scale and Schulz approach to the same sample of 80 trials. Two reviewers (LH, MO, DD, NH, or JS) independently applied the risk of bias tool on the remaining trials after pilot assessment and discussion of five trials among the group of reviewers.

The primary outcome selected for each trial was used for those items in the risk of bias tool that require an outcome focused evaluation—namely, blinding and incomplete outcome data. We applied the tool based on instructions in the *Cochrane Handbook*[4] and consulted one of the developers of the tool (David Moher) for clarification as needed. For the "other sources of bias" domain, we assessed potential bias due to baseline differences, inappropriate influence of the study sponsor, and early stopping for benefit. For crossover designs, we also considered whether such a design was appropriate and whether the wash-out period was sufficient.[4] Overall risk assessments (high, unclear, low) were based on the approach presented in the *Cochrane Handbook*.[4]

Analysis

We used weighted κ to assess inter-rater agreement for each domain of the risk of bias tool and for the final overall assessment.[19 20] We categorised agreement as poor (0.00), slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), or almost perfect (0.81-1.00).[19] Correlations between domains of the risk of bias tool, the Jadad scale, and Schulz approach were calculated using Kendall's τ statistic to assess the concurrent validity of the risk of bias tool. We also assessed the degree of correlation for the overall risk of bias assessment compared with the Jadad overall score, overall risk of bias assessment compared with the Schulz approach, and high or low risk as assessed by risk of bias compared with low or high quality as assessed by the Jadad overall score (<3 *v* ≥ 3[9 21]). Using the paired *t* test we compared the time to apply the risk of bias tool and time to apply the Schulz approach and the Jadad scale.

Effect sizes were calculated using Cohen's d for continuous outcomes; for dichotomous outcomes we converted the odds ratios into effect sizes using a method devised by Hasselblad and Hedges.[22] The effect sizes were combined under DerSimonian and Laird random effects model.[23] Statistical heterogeneity was quantified using the I^2 statistic.[24 25] We used meta-regression to evaluate the effect of risk of bias on the effect size while controlling for possible confounders at study level, including study type (efficacy *v* equivalence), study design (crossover, factorial, or parallel), and outcome type (binary *v* continuous, objective *v* subjective). Studies were defined as efficacy versus equivalence on the basis of the "authors' statements on the primary hypothesis."[18] To determine outcome type (objective *v* subjective), two reviewers (LH, SC) classified the outcomes according to published

Table 1 | Inter-rater agreement using risk of bias tool

Domain	Risk of bias assessments			Weighted κ (95% CI)
	High	Unclear	Low	
Sequence generation	4	107	52	0.74 (0.64 to 0.85)
Allocation concealment	5	105	53	0.50 (0.36 to 0.63)
Blinding	16	49	98	0.35 (0.22 to 0.47)
Incomplete data	25	52	86	0.32 (0.19 to 0.45)
Selective reporting	16	19	128	0.13 (−0.05 to 0.31)
Other sources of bias	15	85	63	0.31 (0.17 to 0.44)
Overall risk of bias	61	96	6	0.27 (0.13 to 0.41)

guidelines[26] and reached consensus through discussion. Analyses were done using SAS version 9.1, SPSS version 11.5, SPlus version 8.0 (Insightful, Seattle, WA), and Intercooled Stata version 7.0.

RESULTS

Inter-rater agreement—Table 1 shows the number of studies assessed as high, unclear, or low risk of bias for the different risk of bias domains (see the web extra for similar summary information for the Jadad scale and Schulz approach to allocation concealment). Inter-rater agreement for the individual domains of the risk of bias tool ranged from slight (κ=0.13 for selective reporting) to substantial (κ=0.74 for sequence generation, table 1). Discrepancies were largely driven by reliance on reporting compared with judgment on risk of bias. Hence domains that involved a greater degree of subjective judgment about potential risk of bias, such as blinding, tended to have poorer inter-rater agreement than domains that were more objective, such as sequence generation. For example, the same level of blinding in a study could yield more or less biased results for different outcomes: a hard end point such as mortality may always be at low risk of bias regardless of the extent of blinding; for a subjective outcome such as quality of life, bias may be more likely if blinding of patients and caregivers was inadequate. Table 2 itemises some of the sources of discrepancies and recommendations on how these might be tackled.

Time for risk of bias versus quality assessment—The mean total time to complete the risk of bias tool by two reviewers (including consensus) for a single outcome was 20.7 minutes (SD 7.6; range 11-58 minutes). Based on a sample of 80 trials the mean time for a single reviewer to complete the risk of bias tool was 8.8 minutes (SD 2.2) compared with 0.5 minutes (SD 0.3) for the Schulz approach ($P<0.001$), 1.5 minutes (SD 0.7) for the Jadad scale ($P<0.001$), and 2.0 minutes (SD 0.8) for the Schulz approach and Jadad scale combined ($P<0.001$).

Concurrent validity of risk of bias tool—A high degree of correlation was found between the domains of risk of bias sequence generation compared with Jadad randomisation, risk of bias allocation concealment compared with Schulz allocation concealment, and risk of bias blinding compared with Jadad double blinding (table 3). Correlation was low for the comparisons between the domains of risk of bias incomplete outcome data and the Jadad withdrawal item, risk of bias overall risk and total Jadad score, and risk of bias overall risk and Schulz allocation concealment (table 3).

Relation between risk of bias and magnitude of effect estimates—Effect estimates were larger for studies assessed as having high or unclear risk of bias (high, n=61, effect size 0.52, 95% confidence interval 0.37 to 0.66; unclear, n=96, effect size=0.52, 0.39 to 0.64) compared with those with low risk of bias (n=6, effect size 0.23, −0.16 to 0.62; figure). Several potential confounders were controlled for through metaregression. The only variable that was statistically significant was study type—efficacy versus equivalence. The trend for efficacy studies was similar to all studies combined, where studies with high and unclear risk of bias had larger effect sizes than those with low risk of bias (high, n=47, effect size 0.69, 0.50 to 0.87; unclear, n=79, effect size 0.64, 0.50 to 0.78; low, n=5, effect size 0.34, −0.10 to 0.78). A reverse pattern was observed for equivalence studies, where those with high or unclear risk of bias were closer to the null compared with those of low risk (high,

Table 2 | Sources of discrepancies and recommendations for selected domains of risk of bias tool

Domain	Source of discrepancy	Recommendation
Blinding	Previous tools judge this domain on basis of reporting. In risk of bias tool, reviewers make judgment on potential risk of bias associated with level of blinding depending on nature of outcome	Identify outcomes (or groups of outcomes) to be assessed by this domain a priori; develop guides for interpretation and application of this domain on basis of nature of intervention and outcomes chosen for review
Incomplete data	Previous tools judge this domain largely on reporting. In risk of bias tool, reviewers make judgment on extent of withdrawals, reasons, and whether these two factors are likely to yield biased results	Identify outcomes (or groups of outcomes) to be assessed by this domain a priori; develop guides for interpretation and application of several factors—proportion of withdrawals or dropouts from overall sample, reasons for withdrawals or dropouts, and whether reasons and extent of withdrawals or dropouts were different across study groups
Selective reporting	Ideally, outcomes planned for study (in study protocol) would be compared with those that were analysed and reported. The search and identification of study protocols may not be fruitful or feasible	In absence of protocols or resources to locate protocols for each included trial, outcomes described in methods section to be compared with those reported in results. Studies that report few outcomes may also be at risk of selective reporting bias. A priori, identify key outcomes that should be reported for particular intervention and patient population
Other sources of bias	Some of these include early stopping, baseline imbalance, differential diagnostic activity, contamination; some are based on trial design (for example, crossover, cluster, factorial). These items will vary according to context and studies relevant to given systematic review	Reviewers should decide a priori which "other sources of bias" will be assessed and develop guides for interpretation. Consideration should always be given to whether there were differences across groups in important variables at baseline; whether authors declared their source of funding; and whether trial was stopped early because benefit was shown

Table 3 | Correlation between domains and overall risk as assessed by risk of bias tool compared with Jadad scale and Schulz approach to allocation concealment

Comparison	Kendall's τ
Comparison of domains:	
Risk of bias sequence generation (yes/no/unclear) *v* Jadad randomisation (bonus/deduction)	0.788
Risk of bias allocation concealment (yes/no/unclear) *v* Schulz allocation concealment (adequate/inadequate/unclear)	0.729
Risk of bias blinding (yes/no/unclear) *v* Jadad double blinding (bonus/deduction)	0.219
Incomplete outcome data (yes/no/unclear) *v* Jadad withdrawals	−0.09
Comparison of overall risk or "quality":	
Risk of bias overall risk (high/unclear/low) vs. Jadad (0-5)	0.059
Risk of bias overall risk (high or unclear/low) *v* Jadad (0-2/3-5)	0.085
Risk of bias overall risk (high/unclear/low) *v* Schulz allocation concealment (adequate/inadequate/unclear)	0.138

n=14, effect size 0.06, −0.06 to 0.17; unclear, n=17, effect size −0.08, −0.30 to 0.15; low: n=1, effect size −0.32, −0.88 to 0.25).

DISCUSSION

We applied the Cochrane risk of bias tool to a sample of 163 randomised controlled trials in children. Despite guidance from the *Cochrane Handbook* on how to apply the tool, the overall inter-rater agreement was fair. Our results stemmed from application of the tool by reviewers working in the same institution and review team. More variability across different research groups might be expected. This highlights the need for clear and detailed instructions to optimise reliability.

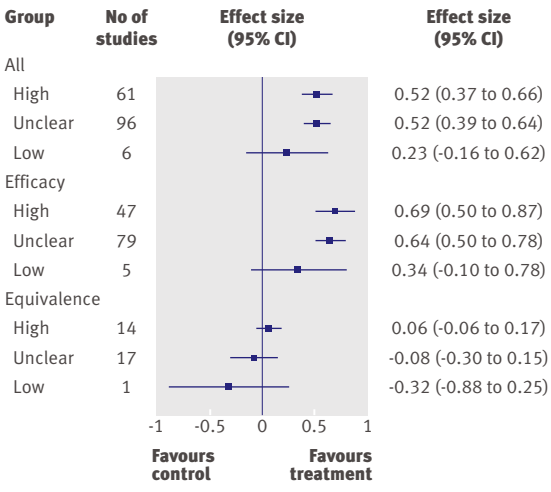
Much of the disagreement arose from items requiring judgment on the potential risk of bias given the methods or approaches described in a study. This underscores the need to establish clear guidelines at the outset of a review and to carry out pilot testing with a sample of studies that are representative of the review question or clinical area. In future research we will examine whether decision rules can reduce inter-rater variability.

We found that the ratings for many domains of the risk of bias tool were "unclear." This may reflect the nature of the domain or the insufficient reporting of study methods and procedures. In some cases the assessment of "unclear" resulted from poor reporting at the individual study level. While reporting may improve for more recent studies as journals and authors adopt the consolidated standards of reporting trials (CONSORT) guidelines,[27] systematic reviewers will continue to face issues arising from poor reporting when they include studies from the era before the guidelines.

On average it took experienced reviewers less than 10 minutes to independently apply the tool for a single, predetermined outcome. The time required to complete the assessments may decrease with increased familiarity and use of the tool. However, more time will be required to apply the risk of bias tool in the context of a full systematic review, as assessments should be made for all main outcomes or classes of outcomes.[4] Furthermore, the *Cochrane Handbook* recommends that study protocols are sought to inform or verify judgments.[4] This would further increase the time required to complete the risk of bias assessment.

The correlation between the risk of bias tool and the Schulz or Jadad approaches was significant in some domains (sequence generation or randomisation, allocation concealment, blinding) but not others (missing data, overall scores). Higher correlations were obtained in domains that were most similar among the different tools. For example, the Jadad item evaluating whether the randomisation sequence was adequately generated is similar to the sequence generation domain of the risk of bias tool. The lack of correlation for the missing data domain seems to be due to the emphasis on reporting in the Jadad instrument compared with conduct in the risk of bias tool—that is, how missing data were handled.

The lack of a significant correlation between the overall risk of bias and Jadad scale, and the risk of bias and Schulz approach, may reflect the different dimensions evaluated by the instruments. The risk of bias tool measures several domains that contribute to the overall assessment of risk of bias, including allocation concealment, and also incorporates selective outcome reporting and "other sources" of bias, domains that are not assessed by the Jadad scale. The lack of correlation could also be explained by the difference in how assessments are made—that is, the reliance on



Effect size estimates according to risk of bias

WHAT IS ALREADY KNOWN ON THIS TOPIC

In February 2008 the Cochrane Collaboration introduced the risk of bias tool to assess the internal validity of randomised controlled trials

The tool is based on six domains, most of which have empirical evidence showing an association with biased results

WHAT THIS STUDY ADDS

Inter-rater agreement was fair but varied substantially across domains

The time to complete the risk of bias tool was significantly longer than other commonly used approaches in systematic reviews

A significant difference in effect sizes was observed between studies with a high or unclear risk of bias and those with a low risk of bias and these patterns were consistent within the subgroups of efficacy and equivalence

reporting for Jadad and Schulz approaches compared with the risk for biased results given the methods that were used. The lack of correlation suggests that the different tools are measuring different constructs; hence, the risk of bias tool may be more appropriate for assessing a trial's internal validity.

Several studies have provided empirical evidence showing that trials with methodological flaws may overestimate treatment effects. This has been observed for allocation concealment,[7 9 11 28] sequence generation,[7 29] double blinding,[7] handling of missing data,[30 31] and selective reporting of outcomes.[16 32 33] We evaluated the risk of bias tool and showed its ability to differentiate between trials that may have overestimated treatment effects. Our results show that studies assessed as at high or unclear risk of bias have larger effect estimates than studies with a low risk of bias. The pattern was consistent for efficacy studies, whereas the reverse pattern was observed for equivalence studies. These results should be considered cautiously given the small number of studies, particularly in the reference category. More rigorous statistical methods that minimise confounding due to intervention and disease are required to confirm these findings. Nevertheless, the results provide some preliminary validation on the usefulness of the risk of bias tool to identify studies that may exaggerate treatment effects. This is particularly relevant to systematic reviewers as well as any practitioner who wants to assess the potential impact of an intervention.

Limitations of the study

This study has several limitations. For efficiency we used information that was generated as part of a previous study—namely, data on effect size; selection of a single, prespecified outcome; and previous assessments using the Jadad and Schulz approaches.[18] As such a delay occurred between application of the three instruments; moreover, the tools were applied by a different team of researchers. This may have contributed to some variability in the application and interpretation of these assessment tools and possibly attenuated the observed correlations; however, it is likely that this more closely resembles the use of these tools in real settings. We applied the risk of bias tool to a single outcome, which is not the recommended approach. This may have resulted in some studies being rated differently for overall risk of bias than if we had considered all of the main or important outcomes. Although we found significant differences in effect sizes when comparing studies at high or unclear risk of bias with those at low risk, these were based on small numbers of low risk studies (six in total) and the confidence interval for such studies was wide. Assessing a more recent sample of studies after the implementation of the CONSORT guidelines may increase the number of low risk studies and may provide a more certain estimate of the impact of risk of bias on effect size. Furthermore, the studies in our sample were published before the release of the CONSORT statement, which may have resulted in more "unclear" assessments than may be the case for more recently published studies. The sample of trials was heterogeneous for outcomes, interventions, and diseases; this differs from the hallmark meta-epidemiological studies in this area that have evaluated the relation between methodological characteristics and effect estimates.[34 35] We used effect sizes to standardise the measures of effect so that we could look at general patterns across studies with different risks of bias. Finally, the sample included only trials in children and therefore the results may not be generalisable to other areas of health care.

Conclusions

We found substantial variation in agreement across domains of the risk of bias tool. Generally the items with poor inter-rater agreement were those that required substantial judgment about the potential for the study methods to yield biased results. There was low correlation between overall assessments using the risk of bias tool compared with the two commonly used tools: the Jadad scale and the Schulz approach to allocation concealment. Overall risk as assessed by the risk of bias tool differentiated effect estimates, with more conservative estimates for low risk studies. Careful training and clear guidelines are required when applying the tool.

We thank Robin Leicht for help with retrieving the articles; Sarah Curtis for help with classification of study outcomes; Natasha Wiebe, Kelly Russell, and Kelly Stevens for their contributions to data extraction, quality assessment, and data analysis, and David Moher for instruction and guidance in applying the risk of bias tool.

Contributors: LH, MO, and TPK designed the study. LH coordinated the project and is guarantor. MO contributed to the conception of the study along with LH. LH, MO, DD, NH, and JS carried out the risk of bias assessments. YL analysed the data. LH, MO, DD, and TPK interpreted the data. NH carried out quality assessments. LH, MO, YL, DD, NH, and JS drafted and critically reviewed the manuscript. TPK critically revised the manuscript. All authors read and approved the manuscript.

Funding: None.

Competing interests: None declared.

Ethical approval: Not required.

1 Agency for Healthcare Research and Quality. Systems to rate the strength of scientific evidence. Evidence report/technology assessment No 47. 2002. www.ahrq.gov/clinic/epcsums/strengthsum.htm.

2 Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651-4.
3 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
4 Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.0.0. Cochrane Collaboration, 2008.
5 Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
6 Gluud LL. Bias in clinical intervention research. *Am J Epidemiol* 2006;163:493-501.
7 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
8 Schulz KF. Assessing allocation concealment and blinding in randomised controlled trials: why bother? *Evid Based Nurs* 2001;4:4-6.
9 Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609-13.
10 Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358-61.
11 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.
12 Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang, C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973-82.
13 Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VSS, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Med Res Methodol* 2004;4:22.
14 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16:62-73.
15 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-60.
16 Chan A, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
17 Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.
18 Klassen TP, Wiebe N, Russell K, Stevens K, Hartling L, Craig WR, et al. Abstracts of randomized controlled trials presented at the Society for Pediatric Research meeting: an example of publication bias. *Arch Pediatr Adolesc Med* 2002;156:474-9.
19 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
20 Cichetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *Am J EEG Technol* 1971;11:101-9.
21 Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999;3:1-98.
22 Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167-78.
23 DerSimonian R, Laird N. Meta analysis in clinical trials. *Contr Clin Trials* 1986;7:177-88.
24 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
25 Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. *Am J Epidemiol* 1992;135:571-8.
26 Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;336:601-5.
27 Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987-91.
28 Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? *Health Technol Assess* 2003;7:1-76.
29 Als-Nielsen B, Chen W, Gluud L, Siersma V, Hilden J, Gluud C. Are trial size and reported methodological quality associated with treatment effects? Observational study of 523 randomised trials. [Abstract.] Canadian Cochrane Network and Centre, University of Ottawa, 2004. www.cochrane.org/colloquia/abstracts/ottawa/P-003.htm.
30 Tierney JF, Stewart LA. Investigating patient exclusion bias in metaanalysis. *Int J Epidemiol* 2005;34:79-87.
31 Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. *J Clin Epidemiol* 2007;60:663-9.
32 Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
33 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753-6.
34 Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in "meta-epidemiological" research. *Stat Med* 2002;21:1513-24.
35 Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2008;149:219.

Accepted: 26 December 2008