

- 20 Kish L. Survey sampling. London: Wiley, 1965.
- 21 Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardised assessment for use by lay interviewers. *Psychol Med* 1992; **22**: 465–86.
- 22 Last JM, ed. A dictionary of epidemiology. New York: Oxford University Press, 1988.
- 23 Weich S, Churchill R, Lewis G, Mann AH. Do socioeconomic risk factors predict the incidence and maintenance of psychiatric disorder in primary care? *Psychol Med* 1997; **27**: 73–80.
- 24 Weich S, Lewis G. Poverty, unemployment, and the common mental disorders: population based cohort study. *BMJ* 1998; **317**: 115–19.
- 25 Power C, Manor O. Explaining social class differences in psychological health among young adults: a longitudinal perspective. *Soc Psychiatry Psychiatr Epidemiol* 1992; **27**: 284–91.
- 26 Wilkinson RG. Unhealthy societies: the afflictions of inequality. London: Routledge, 1996.
- 27 Brown GW, Harris T. Social origins of depression. London: Tavistock, 1978.
- 28 Abramson LY, Seligman MEP, Teasdale JD. Learned helplessness in humans: critique and reformulation. *J Abnorm Psychol* 1978; **87**: 49–74.

## Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?

David Moher, Ba' Pham, Alison Jones, Deborah J Cook, Alejandro R Jadad, Michael Moher, Peter Tugwell, Terry P Klassen

### Summary

**Background** Few meta-analyses of randomised trials assess the quality of the studies included. Yet there is increasing evidence that trial quality can affect estimates of intervention efficacy. We investigated whether different methods of quality assessment provide different estimates of intervention efficacy evaluated in randomised controlled trials (RCTs).

**Methods** We randomly selected 11 meta-analyses that involved 127 RCTs on the efficacy of interventions used for circulatory and digestive diseases, mental health, and pregnancy and childbirth. We replicated all the meta-analyses using published data from the primary studies. The quality of reporting of all 127 clinical trials was assessed by means of component and scale approaches. To explore the effects of quality on the quantitative results, we examined the effects of different methods of incorporating quality scores (sensitivity analysis and quality weights) on the results of the meta-analyses.

**Findings** The quality of trials was low. Masked assessments provided significantly higher scores than unmasked assessments (mean 2.74 [SD 1.10] vs 2.55 [1.20]). Low-quality trials (score  $\leq 2$ ), compared with high-quality trials (score  $> 2$ ), were associated with an increased estimate of benefit of 34% (ratio of odds ratios [ROR] 0.66 [95% CI 0.52–0.83]). Trials that used inadequate allocation concealment, compared with those

that used adequate methods, were also associated with an increased estimate of benefit (37%; ROR=0.63 [0.45–0.88]). The average treatment benefit was 39% (odds ratio [OR] 0.61 [0.57–0.65]) for all trials, 52% (OR 0.48 [0.43–0.54]) for low-quality trials, and 29% (OR 0.71 [0.65–0.77]) for high-quality trials. Use of all the trial scores as quality weights reduced the effects to 35% (OR 0.65 [0.59–0.71]) and resulted in the least statistical heterogeneity.

**Interpretation** Studies of low methodological quality in which the estimate of quality is incorporated into the meta-analyses can alter the interpretation of the benefit of intervention, whether a scale or component approach is used in the assessment of trial quality.

*Lancet* 1998; **352**: 609–13

See Commentary page xxx–xx

### Introduction

The conduct of a meta-analysis is retrospective<sup>1</sup> and is therefore susceptible to several sources of bias.<sup>2</sup> Meta-analyses of randomised controlled trials (RCTs) include studies of variable methodological quality. Features of RCTs that confer the least biased estimates of treatment effect have been intensively studied lately. Differences in quality across trials may indicate that the results of some trials are more biased than others. Meta-analysts need to take this information into consideration to reduce or avoid bias whenever possible. Similarly, there are few data to guide reviewers as to whether any method of quality assessment provides a more biased estimate than any other. In this study, we addressed whether the method of quality assessment of RCTs by a validated scale approach rather than one involving individual components influences estimates of intervention efficacy.

### Methods

#### Selection of meta-analyses

We randomly (random numbers table) selected 12 meta-analyses from our larger database of 491 meta-analyses of RCTs. Three inclusion criteria were used: that the report was published in English; that there was no formal incorporation of quality scores in the quantitative analysis; and that the outcomes were presented as binary data, reported as an overall quantitative

Thomas C Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute (D Moher MSc, B Pham MSc, A Jones BSc, T P Klassen MD); Departments of Medicine (P Tugwell MD), Pediatrics (D Moher, T P Klassen), and Epidemiology and Community Medicine (D Moher, P Tugwell, T P Klassen), University of Ottawa; Department of Clinical Epidemiology and Biostatistics, McMaster University (D J Cook MD); Health Information Research Unit, Department of Epidemiology and Biostatistics, McMaster University, Canada (A R Jadad MD); and Division of Public Health and Primary Health Care, Institute of Health Sciences, Oxford, UK (M Moher MB)

**Correspondence to:** Mr David Moher, Thomas C Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute, Room R226, 401 Smyth Road, Ottawa, Ontario, K1H 8L1, Canada (e-mail: 107656.3375@compuserve.com)

### Features used to assess quality of trial reports

#### Randomisation

Was the study described as randomised (this includes the use of words such as randomly, random, and randomisation)? An additional point was given if the method to generate the sequence of randomisation was described and it was appropriate (eg, table of random numbers, computer generated). However, a point was deducted if the method to generate the sequence of randomisation was described and it was inappropriate (eg, date of birth).

#### Double-blinding

Was the study described as double blind? An additional point was given if the method of masking was described and it was appropriate (eg, identical placebo). However, a point was deducted if the method of masking was described and it was inappropriate (eg, comparison of tablet versus injection with no double dummy).

#### Dropouts and withdrawals

Defined, on the scale, as trial participants who were included in the study but did not complete the observation period or who were not included in the analysis (but should have been described). The numbers and reasons for withdrawal in each group had to be stated for a point to be awarded. If there were no withdrawals, the report should have said so. If there was no statement on withdrawals, this item was given no point.

#### Generation of random numbers

Clinical trials that reported the following methods for generation of their allocation sequence were considered adequate: computer, random numbers table, shuffled cards or tossed coins, and minimisation. Inadequate methods included alternate assignment and assignment by odd/even birth date or hospital number.

#### Allocation concealment

Adequate concealment was that up to the point of treatment (eg, central randomisation). The other category consisted of trials in which allocation concealment was not reported or was inadequate (eg, alternation).

summary result. Meta-analyses were excluded if the report did not provide references for the included trials. Nine of the meta-analyses were randomly chosen from those on the three most frequently reported categories of the International Classification of Disease, 9th revision: three each on digestive diseases,<sup>3-5</sup> circulatory diseases,<sup>6-8</sup> and mental health.<sup>9-11</sup> The remaining three meta-analyses were randomly chosen from the Cochrane Database of Systematic Reviews—one on stroke<sup>12</sup> and two on pregnancy and childbirth.<sup>13,14</sup>

#### Selection of RCTs

Each meta-analysis was reviewed by two of the investigators to agree on the reported principal outcome or outcomes. Because most of the meta-analyses did not explicitly report the primary outcomes,<sup>15</sup> these outcomes were selected on the basis of the largest number of RCTs reporting data on that endpoint (eg, mortality). One meta-analysis<sup>14</sup> was excluded because the data were provided to the principal investigator solely for the purposes of his meta-analysis. This resulted in the selection of 22 independent outcomes (owing to non-overlapping trials) across 11 meta-analyses, from which 127 RCTs were identified and retrieved.

#### Quality assessment

The report of each RCT included in the selected meta-analyses was photocopied twice. On one copy authors, affiliations, any other identifiers such as funding sources, and references were concealed by means of a black marker. The quality of reporting of each of the resulting 254 RCTs was assessed by all of the investigators with an incomplete randomised Latin square design

(ie, each reviewer was randomly assigned both masked and unmasked RCTs but never both versions of the same RCT).

Quality assessments were made with a validated scale<sup>16</sup> and individual components known to affect estimates of intervention efficacy.<sup>17</sup> The scale consists of three items pertaining to descriptions of randomisation, masking, and dropouts and withdrawals in the report of an RCT. The scale ranges from 0 to 5, with higher scores indicating better reporting. The individual components assess the adequacy of reporting of randomisation, allocation concealment, and double-blinding and are described in detail elsewhere.<sup>17</sup> We pretested our methods by means of an interobserver reliability study, assessed with the intraclass correlation coefficient on a separate set of RCTs; values above 0.61 were taken to indicate substantial agreement.<sup>18</sup>

Quality was defined as the confidence that the study design, conduct, analysis, and presentation limited biased comparisons of the intervention under consideration. Quality was assessed by the features listed in the panel.

High-quality trials scored more than 2 out of a maximum possible score of 5. Low-quality trials scored 2 or less out of a maximum possible score of 5. These assignments were made before the start of the study.

#### Data extraction

In addition to the quality assessment of each RCT, we extracted the following data: the number of events and patients in the control group, and the number of events and patients in the intervention group. The data were extracted independently by two investigators (ALJ, DM) and consensus was achieved for any discrepancies before data entry.

#### Data analyses

To assess mean differences in quality scores between masked and unmasked RCTs we used a paired *t* test. To assess differences between masked and unmasked trials in the proportion with adequately reported components we used  $\chi^2$  analysis.

The point estimate and 95% CI from each meta-analysis were replicated by the same analytical procedures as reported by the authors of the original publication (full details available from *The Lancet*). To examine the impact of quality assessment on the combined point estimates, we replicated the methods used elsewhere.<sup>17</sup> Briefly, logistic-regression models were used to explore the relation between a binary outcome of an unwanted event (eg, death) and several independent factors. The independent variables included an overall intervention effect, trial indicators to allow for the variation among the trials, modified treatment effects to capture variation among the meta-analyses, and an estimate of quality. Quality scores were incorporated into the analysis in several ways: as a threshold, a quality weight, or individual component (eg, double-blinding). We also undertook a sensitivity analysis to compare further component assessment of quality and scale assessment.

*Threshold analysis*—For trials assessed on individual components, only the trials that adequately reported the characteristic were included in our analysis. With the scale approach, only the trials scoring above a prespecified score were included in the analysis.

*Sensitivity analysis*—For trials assessed on individual components, two data syntheses were done: analysis of the results for the trials in which the item was adequately reported, and also presentation of the result for the trials that inadequately reported the characteristic. With the scale approach, two analyses were done: analysis of the results for the trials in which the item scored above a prespecified score, and presentation of the results for the trials scoring below the prespecified score.

*Quality weight*—In the main meta-analysis, study estimates were combined after weighting proportionally to their precision to derive the pooled estimate. In the corresponding sensitivity analysis, we advocated the use of a quality weight that was a product of precision and the quality of reporting score. By weighting on precision and trial quality (in this study scaled by

	Masked (n=127)	Unmasked (n=127)	% difference (95% CI)
<b>Mean (SD) score on quality rating scale</b>			
Randomisation	1.09 (0.45)	1.08 (0.45)	0.02 (-0.05 to 0.08)
Double-blinding	1.10 (0.84)	1.00 (0.79)	0.10 (0.02 to 0.18)
Withdrawals/dropouts	0.59 (0.49)	0.50 (0.50)	0.09 (-0.002 to 0.18)
Total score*	2.74 (1.10)	2.55 (1.20)	0.19 (0.06 to 0.32)
<b>Component approach to quality assessment (%)</b>			
Randomisation generation	15.0	14.3	0.07 (-2.05 to 3.45)
Allocation concealment†	14.3	10.7	3.60 (0.94 to 6.26)
Double-blinding	66.4	64.3	2.1 (-1.60 to 5.80)

\*Paired *t* test for scale,  $p=0.005$ . †Adequate allocation concealment,  $p=0.004$ .

Table 1: Quality of reporting of 127 RCTs assessed by a scale<sup>16,17</sup> and individual quality components under masked and unmasked conditions<sup>3</sup>

the quality score), we can assess the effect of various bias-induced features of the trial design and reporting on the pooled estimates of treatment efficacy.

The results of these analyses are reported in terms of a ratio of odds ratio (ROR) and odds ratios (OR). By our modelling convention, an OR and ROR below 1.0 indicate an effective intervention in the subgroups of trials defined in the nominator compared with those in the denominator (eg, low-quality trials vs high-quality trials). Thus, the ROR can be interpreted as providing an estimate of the effects of quality on the point estimate and the precision of the result.

The mean residual deviance of the fitted models reflects the degree of heterogeneity between trials after adjustment for the independent factors. As suggested elsewhere,<sup>17</sup> we used an approximate *F* test to assess the effects of heterogeneity. For all analyses, probability values of 5% or less were taken to be statistically significant.

## Results

### Trials

The 127 RCTs included in the 11 meta-analyses involved 10 492 patients. The 11 meta-analyses were published between 1988 and 1995 in ten journals or the Cochrane Database of Systematic Reviews. The trials on which they were based were published between 1960 and 1995, in 57 journals and three books. One study was unpublished. The majority of outcomes (15/22 [68%]) included can be defined as objective (eg, histological remission, major amputation, overall mortality, conception rate, smoking cessation assessed biochemically).

### Effect of masked assessment

An assessment of the quality of reports of RCTs under masked and unmasked conditions by the scale and component evaluations is given in table 1. The overall quality of reporting of RCTs with the masked scale assessment was 2.74 (SD 1.10), which corresponds to 54.8% of the maximum possible value (5.00). There were significant differences between masked and unmasked evaluation of the quality of reporting of RCTs (table 1). Masked assessment resulted in higher scores than unmasked assessments (2.74 vs 2.55; difference 3.8%;  $p=0.005$ ). We based all further reported analyses on masked assessments only. With the component approach to quality assessment, few RCTs reported on either the methods used to generate the randomisation schedule (15.0% by masked assessment) or the methods used to conceal the randomisation sequence until the point of randomisation occurred (14.3%). Allocation concealment was identified more frequently as adequate under masked than under unmasked assessment (14.3 vs 10.7%,

$p=0.004$ ). With the scale approach, 121 (95%) trials were described as randomised or reported on the methods used to generate participant assignment (or both). Of these trials, only 19 (16%) adequately described allocation concealment.

### Influence of different quality-assessment methods

We were able to replicate closely the results of the published meta-analyses for all 22 selected outcomes. Table 2 shows the influence of quality assessments of the primary trials on the results of the meta-analyses. Trials with a low quality score ( $\leq 2$ ), compared with high-quality trials (score  $> 2$ ), resulted in a 34% greater estimate of the treatment effect (ROR 0.66 [95% CI 0.52–0.83]).

To illustrate the effect of quality-assessment method on an individual meta-analysis, we give the example of Lensing and colleagues' meta-analysis of the efficacy of low-molecular-weight (LMW) heparin.<sup>7</sup> Five RCTs were included in this meta-analysis, resulting in a statistically beneficial effect of LMW heparin on mortality related to deep-vein thrombosis (mortality reduction 47%; OR 0.53 [95% CI 0.32–0.90], 2p for heterogeneity=0.71). When quality assessments were incorporated into the analysis, the beneficial effect of LMW heparin was no longer apparent. For the two RCTs with low quality scores ( $\leq 2$ ), the OR was not significant (0.42 [0.15–1.17], 2p for heterogeneity 0.52), although the point estimate suggests a greater efficacy of LMW heparin. The result was similar for the three high-quality (score  $> 2$ ) trials (OR 0.57 [0.30–1.10], 2p for heterogeneity=0.47). Use of a quality weight resulted in almost no exaggeration of the point estimate and the precision of the statistical result was maintained (OR 0.52 [0.27–0.98], 2p=0.71).

We did a threshold analysis to find out whether the exaggerated intervention effects reported above in relation to the quality scores could be explained by those RCTs in which allocation concealment was inadequately done and inadequately reported, as has been previously suggested (table 2).<sup>17</sup> Our analyses did not result in any meaningful differences in terms of magnitude and direction of bias or statistical significance from those already reported.

Method of quality assessment	ROR (95% CI)	Ratio of heterogeneity between trials (p from a test of similar degree of heterogeneity between trials)‡
<b>Scale</b>		
Low vs high*	0.66 (0.52–0.83)	1.06; <i>F</i> test with 49, 71 df, 2p=0.41
Low vs high†	0.73 (0.56–0.94)	1.01; <i>F</i> test with 49, 51 df, 2p=0.49
<b>Components§</b>		
Randomisation generation	0.89 (0.67–1.20)	1.36 ( <i>F</i> test with 102, 18 df, 2p=0.23)
Allocation concealment	0.63 (0.45–0.88)	1.17 ( <i>F</i> test with 101, 18 df, 2p=0.36)
Double-blinding	1.11 (0.76–1.63)	1.02 ( <i>F</i> test with 39, 81 df, 2p=0.46)

The analysis used the convention that treatment was more effective to prevent an adverse outcome. An OR below 1 indicates an effective intervention. An ROR of less than 1 also indicates an exaggeration of treatment effect.

\*Allowing for summary OR to vary according to quality (ie, quality by treatment interaction) in a base model consisting of intervention, trials, and modified OR according to meta-analyses.

†Including only trials with allocation concealment reported inadequately.

‡Residual deviance reflects degree of heterogeneity between trials derived from a base model. An approximate *F*-distribution was assumed for the ratio of residual deviances to compare the heterogeneity between different ways of incorporating quality. A larger degree of heterogeneity between trials results in a ratio larger than 1.

§Allowing for summary ORs to vary simultaneously according to the components (ie, component by treatment interactions).

Table 2: Influence of different method of quality assessment on treatment-effect estimates

	OR (95% CI)*	Estimated heterogeneity between trials†
Main analysis	0.61 (0.57–0.65)	2.99 ( $\chi^2$ with 121 df)
Sensitivity analysis		
Low quality	0.48 (0.43–0.54)	2.88 ( $\chi^2$ with 49 df)
High quality	0.71 (0.65–0.77)	2.73 ( $\chi^2$ with 71 df)
Quality weight	0.65 (0.59–0.71)	1.59 ( $\chi^2$ with 121 df)

\*Average intervention effect estimated from a base model consisting of intervention, trial, and modified OR according to meta-analyses. †Expected degree of heterogeneity (ie, residual deviance) is 1: large value indicates large heterogeneity between trials.

Table 3: Relation between different methods of incorporating quality scale into meta-analyses and resulting estimates of intervention effects

By incorporating estimates of quality based on individual components, we also detected exaggerated estimates of treatment effect (table 2). Clinical trials reporting allocation concealment inadequately, compared with those trials reporting it adequately, produced statistically exaggerated estimates of treatment effects of 37% (ROR 0.63 [95% CI 0.45–0.88]).

We did not find any significant differences in treatment effects for RCTs according to whether their reports adequately described how the randomisation sequence was generated. Similarly, we did not find an exaggerated treatment effect in relation to the adequacy with which RCT reports described how double-blinding was achieved.

#### *Influence of quality scale method (table 3)*

The average treatment benefit across all trials was 39% (OR 0.61 [0.57–0.65]). Quantitative analysis of only the trials with low quality scores resulted in an average treatment benefit of 52% (OR 0.48 [0.43–0.54]), whereas analysis of only the trials with high quality scores resulted in an average treatment benefit of 29% (OR 0.71 [0.65–0.77]). Use of all the trial scores as quality weights resulted in an average intervention benefit of 35% (OR 0.65 [0.59–0.71]). Use of a quality weight, rather than low quality scores or high quality scores, to incorporate estimates of quality into the quantitative analysis also produced the least statistical heterogeneity (table 3).

## Discussion

Assessment of the quality of reports of RCTs included in a meta-analysis adds another layer of complexity to the reviewing process. Our results suggest, however, that incorporation of an estimate of the quality of RCTs is important. We found a clinically important and statistically significant 30–50% exaggeration of treatment efficacy when results of lower-quality trials were pooled. Inflated estimates of treatment efficacy were found whether the trial quality assessments were made by a scale approach or by an individual component approach.

These results are consistent with the work of Schulz and colleagues,<sup>17</sup> who examined clinical trials on obstetrics and childbirth and found that trials with inadequate allocation concealment exaggerated treatment efficacy by 30–40% compared with trials that had adequate allocation concealment. Our work is based on analysis of studies on four clinical areas, and adds to the evidence that failure to consider trial quality may introduce bias in the results of meta-analysis. This effect is likely to vary somewhat according to how the treatment effect is summarised (eg, relative risk, risk difference) and the control-group event rate (eg, mortality, quality of life).

The results of our sensitivity analysis show that substantial exaggeration of treatment effects remains even when trials with adequate reporting of allocation concealment are removed from the analysis. Unfortunately, we found that few trials reported on methods of allocation concealment despite its importance. We hope that efforts to improve the quality of reporting of RCTs will better this situation. Reviewers should not interpret our results as indicating that a choice must be made between a component or a scale approach to quality assessment. Both approaches offer advantages.

We used both the individual component approach and a scale approach for quality assessment, including items derived from empirical studies showing that they can overestimate the effectiveness of an intervention. Whether these results remain stable with different criteria is uncertain. We have previously shown<sup>19</sup> that different scales applied to the same RCT can provide widely differing estimates of quality in terms of absolute scores and rankings. Use of less empirically based criteria for quality assessment may provide different estimates of the exaggeration of results from those reported here.

Our results suggest use of quality as a weight produces less statistical heterogeneity, a result that could have been expected. Statistical examination of whether the reduction in statistical heterogeneity is an artifact or a real effect associated with quality assessment is difficult and beyond the scope of this study. We do not believe that our results could be explained by artifact alone. Use of only high-quality trials or greater weighting of trials of higher quality is likely to result in a higher signal/noise ratio, thus reducing heterogeneity. Nonetheless, there may be certain conceptual advantages to use of a quality weight rather than a threshold approach. For example, with use of quality weight all trials can be included rather than a selected sample, as would be common with a threshold approach. One limitation of our study is that we did not explore the influence of other ways to incorporate quality weights into the quantitative analysis.<sup>20</sup>

The component approach to quality assessment may have the advantage that new evidence can be incorporated more quickly than with the approach using scales developed by accepted standards.<sup>21</sup> Scale developers will find it difficult to incorporate new evidence into their tools quickly. For this reason, many meta-analysts may prefer to use a component approach to quality assessment.

In using a scale approach to assess quality we found that masked assessments provided statistically higher scores than unmasked assessments. Whether this small absolute difference (3.8%) is important, in terms of additional efforts required by reviewers, is debatable. Many reviewers may see this difference as too small to be important. Several studies have examined the effects of masking on quality assessments of clinical trials.<sup>16,22,23</sup> The results show little consistency in direction or magnitude. A systematic review of these studies would shed light on this issue.

Our study is limited in that we did not explore the relation between unmasked quality assessments and estimates of treatment effects. In addition, the use of a quality score as a weight is based on an assumption that there is a linear relation between the estimates of quality and the weights assigned to the response options (eg, 1, 2, or 3). It is possible that the scaling relation is not linear

and the weighting system is more complex. If data appeared to suggest an indirect relation, our results might not be valid. Our study is also limited in that we used an abbreviated two-response option, rather than the three-response one reported by Schulz and colleagues,<sup>17</sup> to assess allocation concealment. This difference may explain the observed differences in the proportion of trials reporting adequate allocation concealment between masked and open quality assessment. This categorisation might also explain why there is less overlap between the component approach and the scale one. Despite our categorisation, our results are remarkably consistent with those of Schulz and colleagues.<sup>17</sup>

Our results highlight the influence of low-quality trials in the conduct of systemic reviews. This effect has not gone unnoticed. Much effort has been expended lately in developing evidence-based methods to help improve the quality of reporting of clinical trials.<sup>24-26</sup> Several journals have endorsed these approaches<sup>27-30</sup> and incorporated them into their instructions to authors. We hope that improvement in the quality of reporting of RCTs will also help reduce bias when such trials are included in systematic reviews.

#### Contributors

David Moher, Deborah Cook, Alejandro Jadad, Terry Klassen, Michael Moher, and Peter Tugwell developed the grant application to complete this project. All participated in every phase in the conduct of the study. We were assisted by Alison Jones, who coordinated the project, participated in the quality assessments, data extraction, database development, and data entry. Ba' Pham completed the statistical analysis, which included writing all the computer programs. All members of the team read earlier versions of the paper and provided feedback.

#### Acknowledgments

We thank Ken Schulz for reading earlier drafts of this paper and revisions; and Iain Chalmers for reviewing an earlier version of the paper and providing valuable feedback. Deborah Cook is a career scientist of the Ontario Ministry of Health. Alejandro R Jadad is a National Health Research Scholar, Health Canada.

This work was funded (93/52/3) by the National Health Service (UK) Research and Development Programme, Health Technology Assessment Programme. The views expressed in this article do not necessarily represent the views of the UK National Health Service.

#### References

- Chalmers TC. Problems induced by meta-analyses. *Stat Med* 1991; **10**: 971-80.
- Felson DT. Bias in meta-analytic research. *J Clin Epidemiol* 1992; **45**: 885-92.
- Marshall JK, Irvine EJ. Rectal aminosalicylate therapy for distal ulcerative colitis: a meta-analysis. *Aliment Pharmacol Ther* 1995; **9**: 293-300.
- Pace F, Maconi G, Molteni P, Minguzzi M, Bianchi Porro G. Meta-analysis of the effect of placebo on the outcome of medically treated reflux esophagitis. *Scand J Gastroenterol* 1995; **30**: 101-05.
- Sutherland LR, May GR, Shaffer EA. Sulfasalazine revisited: a meta-analysis of 5-aminosalicylic acid in the treatment of ulcerative colitis. *Ann Intern Med* 1993; **118**: 540-49.
- Ramirez-Lassepas M, Cipolle RJ. Medical treatment of transient ischemic attacks: does it influence mortality? *Stroke* 1988; **19**: 397-400.
- Lensing AW, Prins MH, Davidson BL, Hirsh J. Treatment of deep venous thrombosis with low-molecular-weight heparins: a meta-analysis. *Arch Intern Med* 1995; **155**: 601-07.
- Loosemore TM, Chalmers TC, Dormandy JA. A meta-analysis of randomized placebo control trials in Fontaine stages III and IV peripheral occlusive arterial disease. *Int Angiol* 1994; **13**: 133-42.
- Mari JJ, Streiner DL. An overview of family interventions and relapse on schizophrenia: meta-analysis of research findings. *Psychol Med* 1994; **24**: 565-78.
- Loonen AJ, Peer PG, Zwanikken GJ. Continuation and maintenance therapy with antidepressive agents: meta-analysis of research. *Pharmaceutisch Weekblad* 1991; **13**: 167-75.
- Dolan-Mullen P, Ramirez G, Groff JY. A meta-analysis of randomized trials of prenatal smoking cessation interventions. *Am J Obstet Gynecol* 1994; **171**: 1328-34.
- Counsell C, Sandercock P. Anticoagulants in acute stroke. In: Cochrane Library. [CDRom and online]: Cochrane Collaboration, issue 2. Oxford: Update Software, 1995.
- Hughes E, Collins J, Vanderkeckhove P. Bromocriptine, unexplained infertility. In: Cochrane Library [CDRom and online]: Cochrane Collaboration, issue 2. Oxford: Update Software, 1995.
- Grant A. Elective versus selective caesarean delivery of the small baby. In: Cochrane Library [CDRom and online]: Cochrane Collaboration, issue 2. Oxford: Update Software, 1995.
- Jadad AR, Cook DJ, Jones A, Klassen T, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998; **280**: 178-280.
- Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; **17**: 1-12.
- 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.
- Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159-74.
- Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. *Int J Technol Assess Health Care* 1996; **12**: 195-208.
- Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992; **45**: 255-65.
- Berlin JA, on behalf of the University of Pennsylvania Meta-analysis Blinding Study Group. Does blinding of readers affect the results of meta-analyses? *Lancet* 1997; **350**: 185-86.
- 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.
- McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review. *JAMA* 1990; **263**: 1371-76.
- The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. *JAMA* 1994; **272**: 1926-31.
- The Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Checklist of information for inclusion in reports of clinical trials. *Ann Intern Med* 1996; **124**: 741-43.
- Begg CB, Cho MK, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996; **276**: 637-39.
- Rennie D. How to report randomized controlled trials: the CONSORT statement. *JAMA* 1996; **276**: 649.
- McNamee D, Horton R. Lies, damn lies, and reports of RCTs. *Lancet* 1996; **348**: 562.
- Altman DG. Better reporting of randomized controlled trials: the CONSORT statement. *BMJ* 1996; **313**: 570-71.
- Freemantle N, Mason JM, Haines A, Eccles MP. CONSORT: an important step toward evidence-based health care. *Ann Intern Med* 1997; **126**: 81-83.