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 ≤2), compared with high-quality trials (score >2), were associated with an increased estimate of benefit of 34% (ratio of odds ratios [ROR] 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.


See Commentary page xxx–xx

Introduction

The conduct of a meta-analysis is retrospective and is therefore susceptible to several sources of bias. 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 MSc)

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)

ARTICLES


...
Features used to assess quality of trial reports

Randomisation
Was the study described as randomised (this includes the use of words such as random, randomly, 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, table of random numbers, computer generated). 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).

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 χ² 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). The data were ranked independently by two investigators (ALJ, DM) and consensus was achieved for any discrepancies before data entry.

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 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.

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. 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. 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.

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.

ARTICLES
The results of these analyses are reported in terms of a ratio of odds ratio (ROR) and odds ratios (OR). By our modeling convention, an OR and ROR below 1 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, 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 0.2, 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 (<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. 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 (<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). Our analyses did not result in any meaningful differences in terms of magnitude and direction of bias or statistical significance from those already reported.

**Table 2:** Quality of reporting of 127 RCTs assessed by a scale and individual quality components under masked and unmasked conditions.

<table>
<thead>
<tr>
<th>Component approach to quality assessment (%)</th>
<th>Masked (n=127)</th>
<th>Unmasked (n=127)</th>
<th>% difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomisation generation</td>
<td>15.0</td>
<td>14.3</td>
<td>0.07 (−2.65 to 3.45)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>14.3</td>
<td>10.7</td>
<td>0.76 (0.94 to 0.26)</td>
</tr>
<tr>
<td>Double-blinding</td>
<td>66.4</td>
<td>66.4</td>
<td>3.6 (−0.16 to 0.58)</td>
</tr>
</tbody>
</table>

*Paired t test for scale, p=0.005. † Adequate allocation concealment, p>0.05.

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

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

1 Including only trials with allocation concealment reported adequately.

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

3 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.
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 that different scales applied to the same RCT can provide widely differing estimates of quality in terms of absolute scores and rankings.

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.

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. 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. Both approaches offer advantages.

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. 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, 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.

Our results highlight the influence of low-quality trials in the conduct of systemic reviews. This effect has not gone unnoticed. M uch effort has been expended lately in developing evidence-based methods to help improve the quality of reporting of clinical trials. Several journals have endorsed these approaches 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 M, Deborah C, Alejandro J, Terry K, Michael M, and Peter T 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 K for Schultz 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 C is a career scientist of the Ontario Ministry of Health. Alejandro J is a National Health Research Scholar, Health Canada.

This work was funded (93/52/3) by the National Health Service (UK) and Development Programme, Health Technology Assessment, Ontario Ministry of Health. Alejandro R Jadad is a National Health Research Scholar, Health Canada.

This paper forms part of the Cochrane Collaboration, an international group that produces systematic reviews of interventions. The Cochrane Collaboration, based at Oxford, has been set up to provide a forum for those who conduct and use systematic reviews. The Cochrane Library, which contains systematic reviews, is available online through Ovid (Oxford: Update Software, 1995) and also as CD-ROMs. The CONSORT statement is available in the Cochrane Library. The CONSORT statement describes the minimum information that should be included in reports of clinical trials. Contributors

References