Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study

Lorenzo P Moja, Elena Telaro, Roberto D'Amico, Ivan Moschetti, Laura Coe, and Alessandro Liberati on behalf of the Metaquality Study Group

Abstract

Objectives To describe how the methodological quality of primary studies is assessed in systematic reviews and whether the quality assessment is taken into account in the interpretation of results.

Data sources Cochrane systematic reviews and systematic reviews in paper based journals.

Study selection 965 systematic reviews (809 Cochrane reviews and 156 paper based reviews) published between 1995 and 2002.

Data synthesis The methodological quality of primary studies was assessed in 854 of the 965 systematic reviews (88.5%). This occurred more often in Cochrane reviews than in paper based reviews (93.9% v 60.3%, P < 0.0001). Overall, only 496 (51.4%) used the quality assessment in the analysis and interpretation of the results or in their discussion, with no significant differences between Cochrane reviews and paper based reviews (52% vs 49%, P = 0.58). The tools and methods used for quality assessment varied widely.

Conclusions Cochrane reviews fared better than systematic reviews published in paper based journals in terms of assessment of methodological quality of primary studies, although they both largely failed to take it into account in the interpretation of results. Methods for assessment of methodological quality by systematic reviews are still in their infancy and there is substantial room for improvement.

Introduction

Critical appraisal of the methodological quality of primary studies is an essential feature of systematic reviews. A recent review showed that lack of adherence to a priori defined validity criteria may help explain why primary studies on the same topic provide different results. Some key issues still remain unresolved: which checklist and scale is the ideal approach and how the results of quality assessment in a systematic review should be handled in the analysis and interpretation of results.

We compared the approaches used for quality assessment of primary studies by Cochrane systematic reviews with systematic reviews published in paper based journals. We determined how quality assessment is used and whether systematic reviews consider quality assessment in their results.

Methods

We sampled systematic reviews from two databases in the Cochrane Library: the Cochrane Database of Systematic Reviews, which includes reviews prepared by review groups of the Cochrane Collaboration; and the Database of Abstracts of Reviews of Effectiveness (DARE), which selects systematic reviews published in peer reviewed journals on the basis of their adherence to a few methodological requirements.

We stratified all 1297 Cochrane systematic reviews published in issue 1, 2002, of the Cochrane Library by type of intervention (six levels: drugs; rehabilitation or psychosocial; prevention or screening; surgery or radiotherapy; intervention; communication, organisational, or educational; other) and by Cochrane review group (50 levels). We used a computer generated randomisation scheme to select at least 50% of the systematic reviews in each cell. Our final sample represented 62.4% (n = 809) of the Cochrane reviews. The paper based systematic reviews were extracted from DARE, including all systematic reviews published in 2001 registered up to November 2002.

Data extraction form

We assessed the systematic reviews by using an ad hoc data extraction form. We developed this form by taking into account published reports on the quality assessment of trials included in systematic reviews.

We did not aim at standardised operational definitions of the quality measures but accepted at face value what was reported by the authors of individual studies. As a common taxonomy for quality assessment does not exist, we used a large number of descriptive quality components to capture as many of the different definitions as possible.

For each systematic review we sought general information (title, authors, publication date, type of intervention, and presence of a meta-analysis). We then evaluated what authors reported in the methods section of their review for quality assessment. In particular, we tried to ascertain whether authors stated they would have assessed the quality and how (scale or checklist, components studied, composite score) and in what way they planned to use the quality assessment (for example, as exclusion criteria, for sensitivity analysis). See bmj.com for a summary version of the data extraction form.

We then evaluated how authors assessed quality. We recorded if trials were combined in a quantitative meta-analysis; if the quality was evaluated; if scales, checklists, and scores were used; and how the quality was formally incorporated. Assessors judged whether an attempt had been made to incorporate the quality assessment in the results, either qualitatively or quantitatively.
We purposely did not make our operationalised definition of qualitative too stringent. If authors made some comments or discussed the results with reference to the quality of trials then we considered this sufficient to classify the systematic review as having incorporated qualitatively the quality assessment.

Our definition of quantitative was more stringent and included the carrying out of a sensitivity or subgroup analysis (with quality as a stratifying factor) and use of a quality score as a weight or factor for cumulative meta-analysis or metaregression.

**Data extraction**

We developed a draft of an extraction checklist and piloted it on 40 randomly selected Cochrane reviews. The checklist was revised and an instruction manual prepared. The checklist was further tested on another random sample of 130 systematic reviews, and further refinements were incorporated. Inter-rater agreement, based on a random sample of 5% of the Cochrane reviews and paper based reviews (48 reviews), was high. Inter-rater reliability was moderate to perfect (percentage mean agreement 94, range 71.1-100; prevalence and bias adjusted $\kappa$ statistic mean 0.80, range 0.40-1.00).

Twelve pairs of investigators independently extracted the data. Disagreements were resolved through discussion and, when necessary, centrally reviewed.

**Statistical analysis**

Owing to the frequent imbalance of marginals in our contingency matrix, we used a prevalence and bias adjusted $\kappa$ statistic to assess inter-rater reliability.\(^1\)\(^2\)\(^3\)\(^4\) We report confidence intervals of differences between proportions and $P$ values of $\chi^2$ tests.

**Results**

We analysed 965 systematic reviews: 809 Cochrane reviews and 156 reviews published in paper based journals (figure). Quality assessment was assessed in 854 (88.5%) of the reviews and was more often carried out in Cochrane reviews than in paper based reviews (93.9% vs 60.3%, $P < 0.001$; table 1). The same was true when we compared the proportions of reviews using quality assessment in an informal fashion (90.5% vs 51.9%; $P < 0.0001$; table 2). The formal approaches most used by both types of review were exclusion criteria (12%), sensitivity analysis (10%), exploration of heterogeneity (8%), and subgroup analysis (4%).

The quality components most frequently assessed were, in decreasing order, allocation concealment, blinding, and losses to follow-up. The difference between Cochrane reviews and paper based reviews for these and intention to treat analysis significantly favoured Cochrane reviews (table 3).

The most commonly used quality scale was the Jadad scale ($n = 113$, 11.7%; table 3). Cochrane reviews used the scale less often than paper based reviews. In 65.0% ($n = 526$) of Cochrane reviews and 48.1% ($n = 75$) of paper based reviews, authors carried out the quality assessment using single components rather than a formal scale.

No significant differences emerged when Cochrane reviews and paper based reviews were analysed separately by type of intervention assessed—for example, drug compared with non-drug interventions.

**Utilisation of quality assessment**

We found that 496 systematic reviews (51.4%) linked quality to the interpretation of results, with no difference in the proportions of Cochrane reviews and paper based reviews

\[ (51.8\% \pm 49.4\%, P = 0.58; \text{table 1}). \text{This also held true for the subgroup analysis of drug compared with non-drug interventions.} \]

**Is quality assessment carried out as stated?**

The authors of Cochrane reviews were more likely than those of paper based reviews to state that they would assess quality (93.7% vs 63.5%) yet did not always do so (table 1). About 5% of systematic reviews in each group carried out quality assessment despite not being explicitly stated in the methods. Finally, only 328 (33.9%) of the systematic reviews formally specified how they planned to use the quality assessment in the methods (for example, for sensitivity analysis, exclusion criteria): 36.0% ($n = 291$) of Cochrane reviews and 23.7% ($n = 37$) of paper based reviews ($P = 0.79$).

**Discussion**

More than 50% of systematic reviews (both Cochrane reviews and reviews based in paper articles) did not specify in the methods whether and how they would use quality assessment in the analysis and interpretation of results. Cochrane reviews fared better than paper based reviews in carrying out quality assessment but were equally unsuccessful in taking it into account.

Since 1987 several authors have explored the extent to which systematic reviews and meta-analyses incorporate quality assessment into the results, with unsatisfactory findings.\(^5\)\(^6\)\(^7\)\(^8\)\(^9\) Moher et al analysed 240 systematic reviews\(^5\); quality assessment was carried out more often in Cochrane reviews than in reviews published in paper based journals (36/36, 100% vs 78/204, 38%; $P < 0.001$), although only 29 took such assessment into account during their analysis (paper based reviews 27/78 34% vs Cochrane reviews 2/36 6%; $P < 0.001$).
During the past 15 years research has concentrated on two main issues: which components of the quality assessment (for example, allocation concealment) are predictive of valid results and what tool (scales or checklists) best assesses quality. In 2003 Egger et al found that allocation concealment and double blinding were strongly related to treatment effects. Despite the dozens of quality scales and checklists that have been proposed, the answer is still unclear and many doubt that a generic quality assessment tool that would prove valid in all cases can ever be found. In our study the most frequently used tool was the Jadad scale, a tool that has been criticised for its low sensitivity and that does not consider allocation concealment because it was developed before the importance of concealment was established. Moreover, less attention has been paid to explore how quality can be used in the interpretation of the results of systematic reviews.

As Cochrane reviews are preceded by a published protocol and the Cochrane handbook mandates that some form of quality assessment of primary studies is to be done, it is not surprising that authors of Cochrane reviews state that they will carry out quality assessment and do so. Yet when it comes to the more complex, yet potentially relevant, aspect of incorporating quality into the results, Cochrane reviews fared no better than their paper based counterpart.

These findings may have several explanations. That Cochrane reviews provide more details may be due to the absence of limitations on space in electronic publications; however, most of the medical journals now publish a web version of their papers, with different space restrictions. We investigated this potential of the electronic versions but found that none of the paper based reviews was supplemented with an electronic appendix of quality assessment. It is also possible that authors are unaware of, or that editors are not interested in, publishing extensive electronic versions. Moreover, most authors of paper based articles may be aware of space limitations imposed by journals and thus omit details of quality assessment because they believe that they not are relevant to their results. This could be a reflection of a bad practice: a more common use of unplanned outcome dependent subgroup analyses. Where subgroup analyses are not predefined, the risk exists that data may have been

### Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cochrane systematic reviews (n=809)</th>
<th>Paper based systematic reviews (n=156)</th>
<th>Difference between proportions (95% CI)</th>
<th>P value (χ² test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors would carry out quality assessment</td>
<td>758 (93.7)</td>
<td>99 (63.5)</td>
<td>30.2 (22.5 to 38.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Quality assessment carried out</td>
<td>760 (93.9)</td>
<td>94 (60.3)</td>
<td>33.6 (25.8 to 41.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Quality assessment linked to results</td>
<td>419 (51.8)</td>
<td>77 (49.4)</td>
<td>2.4 (-6.1 to 11.0)</td>
<td>0.5777</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Approaches used to incorporate quality assessment*</th>
<th>Cochrane systematic reviews (n=809)</th>
<th>Paper based systematic reviews (n=156)</th>
<th>Cochrane and paper based reviews</th>
<th>Difference between proportions (95% CI)</th>
<th>P value (χ² test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General comment</td>
<td>732 (90.5)</td>
<td>81 (51.9)</td>
<td>813 (95)</td>
<td>38.6 (30.5 to 46.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Formal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exclusion criteria</td>
<td>102 (12.6)</td>
<td>9 (5.8)</td>
<td>111 (12)</td>
<td>6.8 (2.5 to 11.2)</td>
<td>0.0142</td>
</tr>
<tr>
<td>Exploration of heterogeneity</td>
<td>63 (7.8)</td>
<td>11 (7.1)</td>
<td>74 (8)</td>
<td>0.7 (-3.7 to 5.2)</td>
<td>0.7571</td>
</tr>
<tr>
<td>Subgroup analysis</td>
<td>29 (3.6)</td>
<td>8 (5.1)</td>
<td>37 (4)</td>
<td>-1.5 (-5.2 to 2.1)</td>
<td>0.3588</td>
</tr>
<tr>
<td>Sensitivity analysis</td>
<td>89 (11.0)</td>
<td>9 (5.8)</td>
<td>98 (10)</td>
<td>5.2 (1.0 to 9.5)</td>
<td>0.0476</td>
</tr>
<tr>
<td>Weighting of estimates</td>
<td>1 (0.1)</td>
<td>0</td>
<td>1 (0.1)</td>
<td>-0.1 (-0.1 to 0.4)</td>
<td>0.6604</td>
</tr>
<tr>
<td>Cumulative meta-analysis</td>
<td>1 (0.1)</td>
<td>0</td>
<td>1 (0.1)</td>
<td>-0.1 (-0.1 to 0.4)</td>
<td>0.6604</td>
</tr>
</tbody>
</table>

*More than one answer possible.

### Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cochrane systematic reviews (n=809)</th>
<th>Paper based systematic reviews (n=156)</th>
<th>Difference between proportions (95% CI)</th>
<th>P value (χ² test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality components**:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>639 (79.0)</td>
<td>37 (23.7)</td>
<td>55.3 (48.6 to 62.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Any type of blinding</td>
<td>604 (74.7)</td>
<td>64 (41.6)</td>
<td>33.7 (23.4 to 41.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Generation of allocation sequence</td>
<td>209 (25.8)</td>
<td>26 (16.7)</td>
<td>9.1 (2.6 to 15.8)</td>
<td>0.0146</td>
</tr>
<tr>
<td>Similarity of groups at baseline</td>
<td>142 (17.6)</td>
<td>17 (10.8)</td>
<td>6.7 (1.1 to 12.3)</td>
<td>0.0402</td>
</tr>
<tr>
<td>Description of outcomes</td>
<td>110 (13.6)</td>
<td>14 (9.0)</td>
<td>4.6 (-0.5 to 9.7)</td>
<td>0.1142</td>
</tr>
<tr>
<td>Intention to treat analysis</td>
<td>240 (29.7)</td>
<td>19 (12.2)</td>
<td>17.5 (11.5 to 23.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Sample size</td>
<td>99 (12.2)</td>
<td>24 (15.4)</td>
<td>-3.2 (-9.2 to 3.0)</td>
<td>0.2805</td>
</tr>
<tr>
<td>Losses to follow-up</td>
<td>561 (69.3)</td>
<td>52 (33.3)</td>
<td>38.0 (28.0 to 44.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Quality scale:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schulz</td>
<td>37 (4.6)</td>
<td>2 (1.3)</td>
<td>3.3 (1.0 to 5.6)</td>
<td>0.0590</td>
</tr>
<tr>
<td>Jadad</td>
<td>84 (10.4)</td>
<td>29 (18.6)</td>
<td>-8.4 (-14.9 to -2.0)</td>
<td>0.0025</td>
</tr>
<tr>
<td>Cochrane Group</td>
<td>133 (16.4)</td>
<td>0 (0)</td>
<td>16.4 (13.9 to 19.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other</td>
<td>35 (4.3)</td>
<td>20 (12.8)</td>
<td>-8.5 (-13.9 to -3.1)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

*More than one answer possible.
dredged in search of a significant result. Cochrane systematic reviews are, at least in principle, protected against post hoc analyses through the preliminary publication of a detailed study protocol.

Limitations of the study
One limitation of our study is that the Cochrane reviews were published between 1995 and 2002, whereas the paper based articles were first published in 2001. Although older, the Cochrane reviews still fared better for frequency of quality assessment. The Cochrane reviews were, however, remiss in their efforts to incorporate the quality assessment findings in the presentation and interpretation of results. We believe that this difference in publication dates did not affect our results because between 1995 and 2000 no major methodological advances or new consensus emerged in the literature on systematic reviews. Another limitation is whether DARE was an appropriate source from which to sample paper based articles. This is a legitimate concern as it may have led to the selection of a control group with better than average quality. Any selection bias would move our estimate towards the null effect or would minimise the difference between the two types of reviews. Thus, our results could be understated.

A third possible limitation is that incomplete reporting might have influenced our assessment. Evidence, however, suggests that what is reported about important aspects of the conduct of a study typically do reflect what is actually done. Finally, we assessed quality assessment using a checklist that had been developed ad hoc. Although the lack of validation may be criticised, we believe that the items have good face validity: we attempted to reduce assessors’ subjectivity, and the inter-rater reliability was acceptably high. We did have trouble with the classification of quality items, tools, and approaches, as there are innumerable ways to define study quality. It is possible that we recorded quality data with slightly different meanings from those intended by the authors of the studies. Given that almost all systematic reviews were incomplete in their reporting and given the narrative nature of the quality assessment, it is likely that our checklist may have had decreased ability to reflect what authors truly wanted to do and did. It was beyond the scope of our study to assess the appropriateness of the methods chosen for quality assessment by the authors.

Conclusions
Within the Cochrane Collaboration there is room for better standardisation of approaches to quality assessment. The Cochrane handbook should provide clearer guidelines on how to do it. Less clear is how to improve quality assessment in articles published in paper based journals. For such reviews, improvement may only come once a consensus of methodology for systematic reviews has been decided. Peer reviewers and editors should scrutinise systematic reviews to ensure consistency among the various sections and to avoid outcome dependent analyses.

We believe that more research is needed to understand how best to assess and incorporate the methodological quality of primary studies into the results of systematic reviews. Progress towards the necessary improvements highlighted by the results of this study may come from two international meetings planned in May and June 2005 that will be dedicated to, respectively, an improvement of the current section of the Cochrane handbook on quality assessment in systematic reviews and a revision of the QUOROM statement.2 We thank Sabrina Bidoli and Sheila Iacono for their help with data entry and organisational activities; Simone Mangano and Luca Clivio for their support in developing the database; the other members of the Cochrane Collaboration who commented on the study and our interpretation of it when it was presented at the International X and XI Cochrane Colloquia (Stavanger, Norway, 2002 and Barcelona, Spain, 2003); and David Moher, Peter Tugwell, Keith O'Rourke, and Dave Sackett for commenting on an earlier version of this paper.

Contributors: LPM, ET, RD'A, and AL wrote the protocol of the study and RD'A and AL developed the grant application to complete this project. ET and LPM coordinated the project, participated in the quality assessment, data extraction, database development, and data entry. LC, IM, and RD'A participated in the quality assessment, data extraction, and data entry; RD'A, LPM, and IM completed the data analyses. All authors provided feedback on earlier versions of the paper. LPM, RD'A, and AL took primary responsibility for writing the manuscript, which was revised by coauthors. AL is guarantor for the paper. Metaquality Study Group: the following researchers were members of the Milan master course in systematic reviews and contributed to the study in the pilots, narrative and quality assessment of systematic reviews, data extraction, and data entry: Battaglia Alessandro, Bianco Elvira, Calderan Alessandro, Colli Agostino, Ferri Marica, Fraquelli Mirella, Girolami Bruno, Marchionni Enrico, Mezza Elisabetta, Piccoli Giorgia, Vignatelli Luca, Monaco Giuseppe, Morganti Carla, Franceschini Roberta, Bermond Francesca, Cevoli Sabina, Francoz Federico, Ruggeri Marco, Joppi Roberta, Minesso Elisabetta, Lomolino Grazia, and Belli Roberto.

Funding: Ministero della Istruzione, dell’Università e della Ricerca, Rome, Italy (protocol grant No 2802061749 COPIN 2002). Competing interests: All members of the research team are members of the Cochrane Collaboration. ET, IM, RD'A, LC, and AL are authors of Cochrane reviews published in the Cochrane Library. The views and opinions expressed are those of the authors and do not necessarily reflect those of the Cochrane Collaboration. Ethical approval: Not required.

What is already known on the topic
Appraisal of the methodological quality of primary studies is essential in systematic reviews
No consensus exists on the ideal checklist and scale for assessing methodological quality
The Cochrane Collaboration endorses a simple approach to quality assessment based on individual components such as allocation concealment

What this study adds
Approaches to quality assessment of primary studies by systematic reviews are heterogeneous and reflect a lack of consensus on best practice
Cochrane reviews assess methodological quality more often than paper based reviews
Both types of review failed to link the quality assessment to the interpretation of results in almost half of cases

3 Oman AL, Guptail GH. Guidelines for reading literature reviews. CMAJ 1988;133:697-703.
39 DARE—a database of high quality reviews or ‘quality assured reviews?* University of York: Centre for Reviews and Dissemination, 2004.

(Accepted 2 March 2005)
doi 10.1136/bmj.38414.515938.8F

Papers

Corinavo Cochrane Italiano, Istituto Mario Negri, Via Eritrea 62, 20157 Milan, Italy
Lorenzo P Moja researcher
Elena Telaro researcher
Laura Coe researcher
Alessandro Liberati director

Centro Cochrane Italiano, Istituto Mario Negri, Via Eritrea 62, 20157 Milan, Italy
Lorenzo P Moja researcher
Elena Telaro researcher
Ivan Moschetti researcher
Laura Coe researcher
Alessandro Liberati director

Università degli Studi di Modena e Reggio Emilia, Modena, Italy
Roberto D’Amico researcher

Correspondence to: A Liberati, Università degli Studi di Modena e Reggio Emilia, Modena, Italy aleislib@tin.it

BMJ Online First bmj.com