Quality of Cochrane reviews: assessment of sample from 1998

Ole Olsen, Philippa Middleton, Jeanette Ezzo, Peter C Gøtzsche, Victoria Hadhazy, Andrew Herxheimer, Jos Kleijnen, Heather McIntosh

Abstract

Objective To assess the quality of Cochrane reviews.
Design Ten methodologists affiliated with the Cochrane Collaboration independently examined, in a semistructured way, the quality of reviews first published in 1998. Each review was assessed by two people; if one of them noted any major problems, they agreed on a common assessment. Predominant types of problem were categorised.
Setting Cyberspace collaboration coordinated from the Nordic Cochrane Centre.
Studies All 53 reviews first published in issue 4 of the *Cochrane Library* in 1998.
Main outcome measure Proportion of reviews with various types of major problem.
Results No problems or only minor ones were found in most reviews. Major problems were identified in 15 reviews (29%). The evidence did not fully support the conclusion in nine reviews (17%), the conduct or reporting was unsatisfactory in 12 reviews (23%), and stylistic problems were identified in 12 reviews (23%). The problematic conclusions all gave too favourable a picture of the experimental intervention.
Conclusions Cochrane reviews have previously been shown to be of higher quality and less biased on average than other systematic reviews, but improvement is always possible. The Cochrane Collaboration has taken steps to improve editorial processes and the quality of its reviews. Meanwhile, the Cochrane Library remains a key source of evidence about the effects of healthcare interventions. Its users should interpret reviews cautiously, particularly those with conclusions favouring experimental interventions and those with many typographical errors.

Introduction

In the late 1980s clinicians drew attention to the poor scientific quality of healthcare review articles.[1–3] Subsequently, the need for systematic reviews of the effects of healthcare interventions has been widely recognised and checklists and guidelines have been developed.[4–6] The Cochrane Collaboration has led the way in setting new standards for preparing systematic reviews,[4] which are published in electronic format (CD Rom issued quarterly) as part of the *Cochrane Library*. In contrast, few conventional medical journals provide specific guidelines for authors of systematic reviews.[7] Cochrane reviews should therefore be expected to use higher quality methods and should be less prone to bias than systematic reviews published in traditional medical journals. These expectations have been confirmed by four comparative studies of quality and one of bias.[8–12]

However, there is always room for improvement. All scientific reports, including Cochrane reviews, should be read critically. Errors occur, and potential biases may emerge. The comments and criticisms (electronic "letters to the editor" linked to the relevant review) published in the Cochrane Library and the ensuing changes show that some Cochrane reviews have needed correction and improvement.

A group of Cochrane methodologists collaborated to critically read a sample of Cochrane reviews in order to identify and characterise the most common methodological problems. We expected that Cochrane reviews would fulfil most of the criteria that were listed in the current version of the Cochrane handbook[4] and in relevant checklists, so we used a semistructured approach that allowed the assessors to note all kinds of problems they encountered. Our aim was to identify the aspects of Cochrane reviews that are most in need of improvement.

Methods

During the 1998 Cochrane Colloquium the lead researcher (OO) contacted 11 methodologists with various Cochrane affiliations, who subsequently volunteered to assess the methodological quality of the 53 reviews first published in issue 4 of the *Cochrane Library* in 1998.[w1-w53] The project was carried out in 1999 and was coordinated by the Nordic Cochrane Centre. Each review was independently examined by two assessors. We allocated the reviews to assessors by assigning a random number to each review, sorting the numbers in ascending order, and linking the sorted list to a prespecified list of the 55 possible pairs of assessors. Each assessor was assigned nine or 10 reviews, and this assignment was not subsequently changed.

We gave a letter A-E to the overall assessment for each review: A indicated no problems, B minor problems, C major problems, D lack of clarity, and E other types of comments (box 1). We collected and

See also editorial by Clarke and Langhorne

Nordic Cochrane Centre, Rigshospitalet, Dept 7112, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark
Ole Olsen *senior researcher*
Peter C Gøtzsche *director*

Australasian Cochrane Centre, Department of General Practice, Flinders Medical Centre, Adelaide, South Australia, Australia 5042
Philippa Middleton *assistant director*
Heather McIntosh *lecturer*

Cochrane Complementary Medicine Field, Complementary and Alternative Program, University of Maryland, School of Medicine, Baltimore, MD, USA
Jeanette Ezzo *systematic reviews coordinator*
Victoria Hadhazy *research associate*

continued over

BMJ 2001;323:829–32

UK Cochrane Centre, NHS R&D Programme, Oxford OX2 7LG
Andrew Herxheimer *emeritus fellow*
NHS Centre for Reviews and Dissemination, University of York, York YO1 5DD
Jos Kleijnen *director*
Correspondence to: O Olsen o.olsen@cochrane.dk

tabulated the individual scores at the Nordic Cochrane Centre. If one of the assessors had noted a major problem in a review, the two assessors decided whether to give feedback to the reviewers and editors by using the comments and criticisms system in the Cochrane Library. The lead researcher also used submitted comments and criticisms to identify common types of problems. This paper focuses on reviews that had major problems.

Results

One of the 11 methodologists withdrew from the project, and a few assessors did not manage to assess all of their allotted reviews. We initially received 91 (86%) sets of comments out of the expected 106; these related to 52 of the 53 reviews. The review that had not been assessed was removed from the analysis. Three assessors subsequently volunteered to read additional reviews, so that two people assessed any review in which a major problem had been identified.

The scores given in the 91 independent assessments were A, 24; B, 31; C, 19; D, 10; AB, 3; BD, 3; and DE, 1. The number of A scores (indicating no problems) given by individual assessors ranged from 0 to 6, with a median of 2, out of a possible maximum of 10. The number of C scores (major problems) ranged from 0 to 4, also with a median of 2 (table 1).

Of the 52 reviews, 39 (75%) were assessed independently by two reviewers (table 2). Pairs of assessors agreed completely for 13 (33%) reviews; they gave assessments in adjacent categories (A and B or B and C) for 14 (34%). Two (5%) reviews had contradictory assessments (A and C); for each of the remaining 10 (28%) reviews, one of the assessors felt it lacked clarity (D). The 13 reviews assessed by only one reviewer obtained the following scores: A, 1; AB, 2; B, 7; and C, 3. Nineteen (37%) of the 52 reviews had at least one A score, and 17 (33%) had at least one C score.

Pairs of assessors reached agreement on 13 comments and criticisms, which both reviewers wrote jointly; for various reasons an additional four were contributed by only one assessor (one additional assessor withdrew from the project). The full texts of the submitted comments and criticisms are on the *BMJ*'s website.

While classifying the comments and criticisms, we discovered that for two reviews the assessors seemed to have agreed on a C by mistake, leaving 15 reviews (29%, 95% confidence interval 17% to 43%) with major problems. There were three areas of concern. Firstly, the evidence did not support the conclusions in nine (17%) reviews (table 3). Secondly, the conduct or reporting of the reviews was unsatisfactory in 12 (23%) reviews (box 2). Thirdly, there were stylistic concerns with 12 (23%) reviews.

Box 1: Assessment format—the assessors were asked to choose one option for each review and give details as appropriate

A I encountered no problems in the review. I would be proud to show this review as an example of a Cochrane review
B I encountered minor problems in the review—namely, the following: …
C I encountered major problems in the review, and I think it deserves a comment and criticism along the lines of …
D The review might be OK, but I need clarification on …
E Any other unstructured comments

Table 1 Scoring of reviews: quality of 52 reviews as assessed by 10 assessors

Assessor	No of A scores (no problems)	No of C scores (major problems)
Andrew Herxheimer	6	1
Jeanette Ezzo	4	0
Ole Olsen	3	1
Heather McIntosh	2	1
Jos Kleijnen	2	2
Peter Gøtzsche	2	4
Phil Alderson	2	2
Victoria Hadhazy	2	3
Matthias Egger	1	2
Philippa Middleton	0	3
Median (range)	2 (0-6)	2 (0-4)

Maximum number of scores given by one assessor=10.

Table 2 Combinations of scores within pairs of assessors

Combination of scores	No of occurrences
AA	6
AB*	6
AC	2
BB*	5
BC*	8
CC	2
DX†	10
Total‡	39

A=no problems; B=minor problems; C=major problems; D=lack of clarity; E=other comments.
*Counting one BD as B.
†DX denotes various combinations with D: AD, 4; BD, 2; CD, 2; other, 2.
‡13 reviews were assessed by only one reviewer.

The problematic conclusions all described the effect of the experimental treatment in terms that we judged to have been too optimistic (table 3). None of these nine reviews indicated a bias towards the control treatment.

The most common problems with methods (box 2) concerned inclusion and exclusion of trials (six reviews), concealment of allocation (five), loss to follow up (four), choice of outcome measures (four), and statistics (three). Other problems were found only once—for example, a conclusion based largely on a single trial that seemed to have major weaknesses.

Stylistic problems were indicated by statements such as "many spelling and grammatical errors," "a few typographical errors," "seems to be an unfinished draft," "needs to be edited to be more readable and comprehensible." Four reviews had many spelling and typographical errors, and these four also had problems relating to their methods and conclusions.

Discussion

Fifteen (29%) of 52 Cochrane reviews first published in 1998 were judged to have major problems, among which biased conclusions, problems with methods, and insufficient typographical or stylistic editing were the most common. Thus, even though Cochrane reviews

Table 3 Conclusions not supported by the evidence

Reviewers' conclusion	Assessors' comment
"Studies are of insufficient duration to identify a reduction in mortality" [w18]	"The studies . . . suggest an increased mortality (13 *v* 7)"
"X is the cornerstone of treatment" [w51]	"The comparisons with Y and Z are not that convincing; see graphs"
Treatment X might promote healing[w20]	"The study that suggested this . . . may be flawed in design, conduct, and analysis"
Treatment X "is associated with a substantially reduced risk" [w11]	"Too strong . . . as the confidence interval crossed 1"
"This review found a significant decrease in . . ." (one variable)[w31]	"Many small trials. . .with many outcome variables . . . implying a high risk of reporting bias"
Treatment X "is associated with a reduction in death or oxygen requirement" [w53]	"Significant reduction was seen only for the combined outcome, not for death or reduction in oxygen requirement separately"
"There is evidence to support the early use of X in disease Y" [w47]	"This conclusion . . .is unwarranted and misleading" as "the review might concern disease Z" (which is known to respond to treatment X)
"X is associated with short term improvements" [w29]	"Given the limitations of the included studies we would recommend that the conclusions be modified to be more cautious"
"There are too few data to draw any reliable conclusions" [w50]	"Potential adverse effect is so important that it should be mentioned in the conclusions"

are based on specific guidelines[4] and have higher quality methods on average than systematic reviews published in conventional journals,[8–11] problems were still common. The problems we identified in these reviews were brought to the attention of their authors through the electronic comments and criticism system; revised versions of some reviews have subsequently appeared in the Cochrane Library (in some instances with a changed title), and other revisions are being prepared.

Box 2: Some examples of problems with methods

Inclusion and exclusion criteria were not well defined,[w31,w51] inconsistently transferred from title to methods section (another category of patients or another type of intervention),[w9,w31] or applied inconsistently,[w31] subjectively,[w29] or with retrospective rationalisation.[w27] The problems partly related to choice of terminology—for example, "experimental or quasi-experimental designs" (meaning randomised or quasi-randomised or non-randomised studies?) resulting in different interpretations or enumerations in the "selection criteria," "description of studies," "main results," meta-analytic graphs, and "table of included studies" and in conflicting numbers[w31,w51]

Concealment of allocation was mixed up with double blinding,[w4,w20,w29,w31] or discrepancies existed between the quality of the trials as described in the table of included trials, the graphs, and the main text.[w31,w52] These problems may occur much more often than we found as most occurrences were noted by a single assessor

Assessors were particularly concerned about loss to follow up when rates of dropout were high (29%, 43%, 50%),[w29,w51] dropouts were treatment failures,[w29] or the reviewers claimed to have performed an intention to treat analysis but the numbers on the graphs did not fit with the number of included patients[w52]; the assessors asked for subgroup analyses including only trials with full follow up, more thorough documentation and discussion of loss to follow up,[w4] or more cautious conclusions[w29,w51]

Outcome measures were inappropriately combined (death plus surrogate outcome),[w53] inappropriately split (different lengths of follow up),[w20] too numerous without precautions against multiple testing,[w31,w52] unblinded subjective assessments,[w53] or derived from only a single small trial[w31]

Statistical problems related to extremely different standard deviations between experimental and control group[w18] or between trials[w31] or to a confidence interval with zero length[w51]

Strengths and weaknesses of the study

We studied Cochrane reviews that were first published nearly three years ago. A range of quality initiatives has been implemented by the Cochrane Collaboration since then; we hope that a study of current Cochrane reviews would reveal a smaller proportion of reviews with major problems.

Assessing only reviews first published in issue 4 of the *Cochrane Library* in 1998 was a decision of convenience and may have led to a sample that was not representative of all the Cochrane reviews available at that time. For example, if these new reviews had been contributed mainly by relatively inexperienced reviewers and editorial groups, our overall findings may have overestimated the proportion of reviews with major problems. On the other hand, even in 1998, three years after the first Cochrane reviews had been published, new reviews might be expected to have been of higher quality than older reviews, thus leading to a bias in the opposite direction.

Our individual assessments were done in a semistructured way, without a checklist, by experienced, selected volunteer methodologists from the Cochrane Collaboration, who were advised to spend not more than 10 hours in total on the exercise. The time constraint and lack of a checklist may have led to some errors going undetected. Conversely, the use of experienced methodologists may have led to detection of unexpected errors. The way assessors were recruited may have led us to be particularly mild or particularly hard in our assessments. Because agreement was reached on most assessments, disagreements in the individual assessments probably reflected oversights by one of the assessors rather than true disagreement. On the other hand, four of the 15 comments and criticisms that were submitted had been written by only one assessor. Thus the number of identified problems might be an underestimate. Alternatively, the high number of problems described as major may partly reflect a few very demanding assessors. The fact that agreements were reached despite the variability in individual assessments indicates that selection of another set of methodologically trained assessors would probably not have greatly altered the final assessments.

One person (OO), while reading the 17 comments, created the three categories by grouping the problems identified. No hypotheses regarding these categories were stated at the outset. Prespecified categories and two data extractors would have strengthened the validity of the findings.

Relation to other studies of bias

Empirical studies have identified several important biases, such as publication bias and bias related to poor randomisation, related to individual trials that all tend to exaggerate the estimated beneficial effects of new treatments.[13–15] Important bias also arises in the step from results to conclusions. In a study of 196 drug trials, bias in the conclusions or abstracts favoured the new drug in 81 of 82 reports.[16] In accordance with these findings, the bias we found favoured the experimental treatment in all of the nine reviews with problematic conclusions. Thus the same type of bias seems to occur in the conclusions of systematic reviews as in reports of trials. This bias in systematic reviews, which is unrelated to bias in the individual trials, seems not to have been reported before. However, although dubious or invalid statements were found in 76% of the conclusions or abstracts of drug trial reports,[16] they occurred in only 17% of the Cochrane reviews.

Steps to improve quality

Solutions to most of the methodological problems we met were described in the Cochrane handbook and other checklists for systematic reviews.[4 17] This indicates the need for better use of guidelines in scientific editing and peer review. The conclusion bias we observed has led to improved advice on how to write conclusions in Cochrane reviews in the most recent version of the Cochrane reviewers' handbook.[18]

Since 1998 the Cochrane Collaboration has taken several additional steps to improve the quality of its reviews. A quality advisory group has been established and the post of quality improvement manager has been created. The Cochrane reviewers' handbook is improved regularly, and tools for assessing the quality of reviews are being developed. Several new courses for reviewers, statisticians, and editors have been developed, and two centralised editing projects are running. It is also important that users of the Cochrane Library participate in the process by submitting any relevant comments and criticisms they might have.

Implications for clinicians and policymakers

Seekers of the best available evidence on treatment and prevention should continue to look to the Cochrane Library as a key source of information, despite the deficiencies that we found in a minority of Cochrane reviews in the 1998 sample. Reliance on unsystematic reviews, textbooks, and anecdotal evidence is likely to be far more problematic.[1 19] No matter which sources of evidence are being used, users of the evidence need to learn the skills of critical appraisal. Guides and courses on critical appraisal are now widely accessible.[20] As with any scientific report, readers should themselves assess the reliability of individual Cochrane reviews. They should be particularly cautious of reviews with conclusions that favour experimental interventions when relatively little evidence is available for the review and of reviews with many typographical errors.

What is already known on this topic

Cochrane reviews are, on average, more systematic and less biased than systematic reviews published in paper journals

Errors and biases also occur in Cochrane reviews

What this study adds

Too often, reviewers' conclusions over-rated the benefits of new interventions

Readers of Cochrane reviews should remain cautious, especially regarding conclusions that favour new interventions

The Cochrane Collaboration has taken steps to improve the quality of reviews

We thank Phil Alderson and Matthias Egger for their contribution to the study and Iain Chalmers and Mike Clarke for useful comments on the manuscript.

Contributors: OO planned, coordinated, and reported the work and is the guarantor of the paper. All authors, Phil Alderson, and Matthias Egger participated in the assessments. All authors gave input to, and some commented on, the draft versions of the paper. All authors approved the manuscript.

Funding: OO was funded by the Danish Institute for Health Technology Assessment.

Competing interests: All assessors are associated with the Cochrane Collaboration.

1 Mulrow CD. The medical review article: state of the science. *Ann Intern Med* 1987;106:485-8.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *Can Med Assoc J* 1988;138:697-703.
3 Yusuf S, Simon R, Ellenberg S, eds. Proceedings of "Methodologic issues in overviews of randomized clinical trials." *Stat Med* 1987;6:217-409.
4 Mulrow CD, Oxman AD, eds. Cochrane Collaboration handbook. In: *Cochrane Library.* Issue 4. Oxford: Update Software, 1997.
5 Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597-9.
6 Oxman AD. Checklists for review articles. *BMJ* 1994;309:648-51.
7 Moher D, Jadad AR. How to peer review a manuscript. In: Godlee F, Jefferson T, eds. *Peer review in health sciences.* London: BMJ Books, 1999: 146-56.
8 Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M. Methodology and reports of systematic reviews and meta-analyses. *JAMA* 1998;280:278-80.
9 Shea B, Moher D, Pham B, Tugwell P. Assessing the quality of reporting meta-analyses of randomized controlled trials [abstract]. VII Cochrane Colloquium, Rome, 5-9 October 1999:A39.
10 Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews and meta-analyses: A systematic review of checklists and scales [abstract]. VII Cochrane Colloquium, Rome, 5-9 October 1999:A40.
11 Jadad AR, Moher M, Browman GP, Booker L, Sigouin C, Fuentes M, et al. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *BMJ* 2000;320:537-40.
12 Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.
13 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.
14 Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;315:645-5.
15 Tramér MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997;315:6335-40.
16 Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989;10:31-56.
17 Oxman AD, Cook DJ, Guyatt GH for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI: How to use an overview. *JAMA* 1994;272:1367-71.
18 Clarke M, Oxman AD, eds. Common errors in reaching conclusions. Cochrane reviewers' handbook 4.1.3, section 9.7. In: *Cochrane Library.* Issue 3. Oxford: Update Software, 2001.
19 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 1992;268:240-8.
20 EBH calendar of events. http://cebm.jr2.ox.ac.uk/docs/calendar.html (accessed 17 July 2001).

(Accepted 29 August 2001)