

# The Medical Review Article Revisited: Has the Science Improved?

Finlay A. McAlister, MD, MSc; Heather D. Clark, MD; Carl van Walraven, MD, MSc; Sharon E. Straus, MD; Fiona M.E. Lawson, MB; David Moher, MSc; and Cynthia D. Mulrow, MD, MSc

**Background:** The validity of a review depends on its methodologic quality.

**Objective:** To determine the methodologic quality of recently published review articles.

**Design:** Critical appraisal.

**Setting:** All reviews of clinical topics published in six general medical journals in 1996.

**Measurements:** Explicit criteria that have been published and validated were used.

**Results:** Of 158 review articles, only 2 satisfied all 10 methodologic criteria (median number of criteria satisfied, 1). Less than a quarter of the articles described how evidence was identified, evaluated, or integrated; 34% addressed a focused clinical question; and 39% identified gaps in existing knowledge. Of the 111 reviews that made treatment recommendations, 48% provided an estimate of the magnitude of potential benefits (and 34%, the potential adverse effects) of the treatment options, 45% cited randomized clinical trials to support their recommendations, and only 6% made any reference to costs.

**Conclusions:** The methodologic quality of clinical review articles is highly variable, and many of these articles do not specify systematic methods.

*Ann Intern Med.* 1999;131:947-951.

For author affiliations and current addresses, see end of text.

Review articles are an important element of most medical journals and a popular source of information for clinicians (1). Given the increasing volume of medical literature and the limited time for reading that busy clinicians have, reliance on review articles is likely to increase. However, concerns have been raised that narrative, nonsystematic review articles may produce biased conclusions (2, 3). Mulrow (4) examined 50 review articles published in four major medical journals (*New England Journal of Medicine*, *Annals of Internal Medicine*, *JAMA*, and *Archives of Internal Medicine*) in 1985–1986 and found that none fulfilled 8 explicit criteria for scientifically sound summaries of the evidence. As a result, she and others (5, 6) proposed criteria for conducting and evaluating review articles that would improve their quality; these criteria are the first 10 listed in the **Appendix Table**. In a study of 36 review articles done by nine content experts and methodologists, these criteria were shown to yield reliable and valid estimates of the scientific quality of reviews (7, 8).

We sought to describe the methods used in recently published review articles and determine whether the attention paid to the methodologic shortcomings of review articles has led to improvements in their scientific quality.

## Methods

By using the Science Citation Index (9), we stratified the 12 general medicine journals that are considered “core journals” for *ACP Journal Club* (10) into those with high impact factors (scores  $\geq 5$ ) and those with lower impact factors. We randomly selected 3 (*New England Journal of Medicine*, *Annals of Internal Medicine*, and *JAMA*) of the 4 high-impact core journals and 3 (*British Medical Journal*, *American Journal of Medicine*, and *Journal of Internal Medicine*) of the 8 core journals with lower impact factors. The sampling frame included 3 of the 4 journals examined in Mulrow’s original study (4). All 6 journals were hand-searched by two of the investigators independently, and review articles published between January and December 1996 were retrieved. Review articles were defined as full-text articles published under the banner “review” that dealt with disease states; had the words “review,” “overview,” or “meta-analysis” in the title or abstract; or indicated in the text that the intention was to review or summarize the literature about a clinical topic. Editorials, correspondence, and conference summaries were excluded. Disagreements on article eligibility (which occurred in five cases) were resolved by consensus.

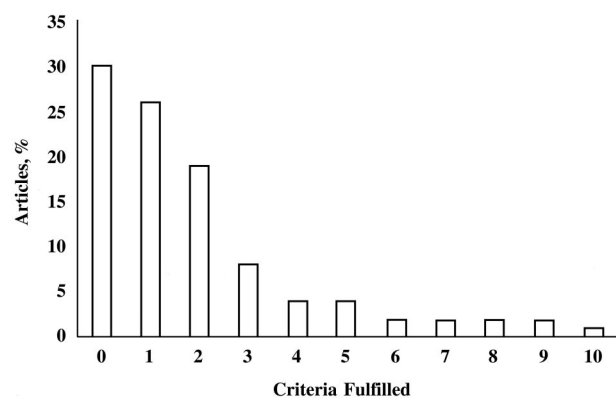
After training with a test set of articles, five of

the authors used explicit criteria (**Appendix Table**) to rate the identified review articles. So that the raters' assessments would be blinded, the articles' authors, author affiliations, and sources of articles were masked. The first 10 criteria in the **Appendix Table** (hereafter referred to as "methodologic criteria") have been previously validated (7) and assess methodologic rigor, whereas the last five criteria were developed for this study to evaluate the scientific basis of treatment recommendations. Because inter-rater agreement was excellent on the test set (overall agreement, 94%), each rater was assigned a random sample of 20% of the articles. One rater independently evaluated a random sample of five articles from each rater to assess inter-rater reliability for the criteria. The mean  $\kappa$  value was 0.79, indicating substantial agreement; overall agreement was 95% and ranged from 84% for criteria 1 and 6 to 100% for criteria 2, 3, and 5.

After we determined the number of criteria fulfilled by each article, the number of articles meeting each criterion was ascertained. The proportion of reviews from high-impact journals and from lower-impact journals that met each criterion were compared. Only the subset of reviews containing treatment recommendations were used for the comparisons for criteria 7 to 15.

The chi-square test was used and 95% CIs for the observed differences were calculated for all comparisons. In addition, articles identified as meta-analyses, systematic reviews, or overviews in their title, abstract, or text were compared with the remaining review articles for each criterion (again by using the chi-square test and by limiting analysis to those that included treatment recommendations for criteria 7 to 15). Logistic regression analysis was used to adjust for journal of publication.

The funding organizations were not involved in the design, conduct, analysis, or reporting of this study.



**Figure.** Percentage of 158 review articles published in 1996 that fulfilled specific methodologic criteria. The numbers on the x axis refer to the first 10 criteria listed in the Appendix Table.

## Results

A total of 158 review articles were identified: 60 from *New England Journal of Medicine*, 33 from *Annals of Internal Medicine*, 24 from *British Medical Journal*, 18 from *American Journal of Medicine*, 13 from *JAMA*, and 10 from *Journal of Internal Medicine*. (A full list of the included articles is available from Dr. McAlister on request.) Most reviews were written by more than one author (median number of authors, 2 [range, 1 to 10 authors]), and 111 reviews (70%) made treatment recommendations. Although most articles had sections reviewing the relevant pathophysiology and pharmacodynamics, 19 articles (12%) were largely basic science reviews.

Only 2 reviews met all 10 methodologic criteria; the median number of criteria fulfilled was one (**Figure**). Nineteen (12%) of the review articles were described as "meta-analysis," "systematic review," or "overview" in the title or abstract; 12 of these 19 articles were published in high-impact journals. A higher proportion of these articles met the methodologic criteria (**Appendix Table**); the comparisons for all but criterion 6 were statistically significant, even after adjustment for journal of publication ( $P < 0.005$  for all comparisons). Although there was significant heterogeneity among journals ( $P < 0.005$  for criteria 1 to 5, 7, and 8), review articles published in high-impact journals did not meet methodologic criteria more frequently than those published in other journals (**Table**). Exclusion of the reviews from one journal with methodologic scores that were significantly lower than the rest ( $P < 0.001$  for all comparisons) resolved much of the interjournal heterogeneity and revealed that reviews published in the remaining high-impact journals more often met some (but not all) of the methodologic criteria (**Table**).

The 44 articles that described how evidence was located used the following sources (which were not mutually exclusive): MEDLINE (42 articles), hand search of reference lists from published studies or review articles (30 articles), other electronic databases (14 articles), contact with experts in the field (10 articles), hand search of relevant journals (7 articles), and contact with the pharmaceutical industry (5 articles). Of the 22 reviews that included a description of the electronic search strategy, the search was restricted to English-language publications in 11.

Of the 111 reviews in which treatment recommendations were made, a median of 3 therapies (range, 1 to 23 therapies) were discussed. The 16 meta-analyses that included treatment recommendations were more focused, examining a median of 1 therapy (range, 1 to 6 therapies). The scientific basis

**Table. Comparison of Methods Used in 1996 Review Articles, by Journal Impact Factor\***

Criterion	Reviews in High-Impact Journals (n = 106)	Reviews in Lower-Impact Journals (n = 52)	Proportional Difference (95% CI)†	Proportional Difference without Outlier‡
	n (%)			
1. The review addressed a focused clinical question	34 (32)	20 (38)	-0.06 (-0.22 to 0.10)	0.16 (-0.04 to 0.36)
2. The method of locating evidence was described	32 (30)	12 (23)	0.07 (-0.08 to 0.22)	0.47 (0.27 to 0.67)
3. Explicit criteria were used to select studies	15 (14)	7 (13)	0.01 (-0.11 to 0.13)	0.20 (0.04 to 0.36)
4. The methodologic validity or quality of the included studies was assessed	8 (8)	6 (12)	-0.04 (-0.13 to 0.05)	0.05 (-0.09 to 0.19)
5. Assessments of studies were reproducible (that is, they were done by more than one reviewer)	13 (12)	5 (10)	0.02 (-0.08 to 0.12)	0.18 (0.03 to 0.33)
6. Directives for future research initiatives were proposed in reviews that included treatment recommendations§:	41 (39)	20 (38)	0.01 (-0.15 to 0.17)	0.08 (-0.12 to 0.28)
7. Sources of heterogeneity (clinical or study design) in existing data were addressed	8 (10)	7 (22)	-0.12 (-0.26 to 0.02)	-0.01 (-0.21 to 0.19)
8. Quantitative synthesis of existing data was done	15 (19)	8 (25)	-0.06 (-0.23 to 0.11)	0.08 (-0.14 to 0.30)
9. The major clinically relevant outcomes (benefits and harms) were considered	27 (34)	11 (34)	0 (-0.19 to 0.19)	-0.01 (-0.24 to 0.22)
10. The generalizability of existing data was addressed	9 (11)	4 (13)	-0.02 (-0.15 to 0.11)	0.05 (-0.12 to 0.22)
11. Randomized clinical trials were cited as support for treatment recommendations	32 (41)	18 (56)	-0.15 (-0.35 to 0.05)	-0.04 (-0.28 to 0.20)
12. An estimate of treatment effect was provided	30 (38)	23 (72)	-0.38 (-0.59 to -0.17)	-0.08 (-0.31 to 0.15)
13. The treatment effect was expressed as:				
Relative risk or relative risk reduction	11 (14)	10 (31)	-0.17 (-0.33 to -0.01)	-0.16 (-0.36 to 0.04)
Odds ratio	3 (4)	5 (16)	-0.12 (-0.22 to -0.02)	-0.07 (-0.23 to 0.09)
Absolute risk reduction	2 (3)	2 (6)	-0.03 (-0.11 to 0.05)	0 (-0.12 to 0.12)
Number needed to treat	1 (1)	0	0.01 (-0.03 to 0.05)	0.03 (-0.04 to 0.10)
None of the above	13 (16)	6 (19)	-0.03 (-0.18 to 0.12)	0.11 (-0.10 to 0.32)
14. The precision of the treatment effect (confidence interval) was given	11 (14)	7 (22)	-0.08 (-0.23 to 0.07)	0.05 (-0.16 to 0.26)
15. The costs of the treatment options were considered	7 (9)	0	0.09 (-0.01 to 0.19)	0.09 (-0.02 to 0.20)

\* High-impact journals had an impact score  $\geq 5$ .

† Differences between proportions of reviews published in high-impact journals and those published in lower-impact journals that met each criterion. Negative values mean that the criterion was more often met by articles in lower-impact journals; positive values mean that the criterion was more often met by articles in high-impact journals.

‡ Differences between proportions of reviews published in high-impact journals and those published in lower-impact journals after exclusion of all articles from one journal with methodologic scores significantly lower than those from the other journals.

§ Includes 79 reviews from high-impact journals and 32 reviews from lower-impact journals.

of these recommendations is outlined in the last five criteria of the **Appendix Table**.

## Discussion

In summary, only a minority of the review articles published in six widely read general medical journals specified rigorous, systematic methods of identifying, evaluating, and synthesizing the evidence, thereby raising concerns about the validity of their conclusions and recommendations (11). Furthermore, although most review articles made recommendations for therapeutic options, their clinical relevance may be limited: Only one third discussed benefits and harms of the treatment options, only 6% mentioned the relative or absolute costs of the various options, less than half provided an estimate of the anticipated treatment effect, and very few presented the results in clinically relevant formats that would be easily understood by clinicians.

Although there is still substantial room for improvement, these results do represent progress from the situation in 1985–1986 (**Appendix Table**). A higher proportion of recently published reviews specify how the evidence was identified and synthesized. This same pattern was observed in a recent

study of oncology review articles published in a single journal between 1983 and 1995 (12). Similarly, our findings for meta-analyses published in these journals in 1996 are consistent with contemporaneous investigations in other fields and represent substantial improvement over the state of affairs in the mid-1980s (13, 14).

Our study had several potential limitations. First, the authors of articles in which systematic methods were not specified may have used such methods but not reported them (or the journal may have removed the details of the methods in prepublication). However, preliminary evidence suggests that there is little difference between the conduct of studies and their published methods (7, 15). In other words, “if what was done is not reported, there is a good chance that it was not done rigorously” (7). Second, our finding that meta-analyses are more scientifically rigorous than other reviews may be seen as tautological because the criteria used in this study may be used by meta-analysts in performing their studies. However, meta-analysts disagree over some issues of methodology, and published meta-analyses often vary in the quality of their reports (13). Third, the practical importance of the methodologic flaws identified in these reviews is difficult to ascertain. We lack empirical evidence

that reviews with explicit, systematic methods yield results closer to the truth, but it is known that nonsystematic reviews are more prone to random and systematic errors and often provide conflicting summaries of the evidence (6, 16). Although meta-analyses may also provide conflicting results, closer examination of their explicit methods sections more readily allows resolution of discordant conclusions (17). Our conclusions are based on reviews from only 6 journals, thereby raising concerns about the generalizability of our results; however, these journals are all widely read and were selected from the 12 general medical journals identified as "core journals" for *ACP Journal Club*. Thus, we believe that our results are generalizable for the general medical reader. Finally, it must be recognized that only the first 10 criteria in the **Appendix Table** have been empirically validated; because the remaining 5 criteria are opinion-based, they may not adequately reflect or measure the scientific basis for treatment recommendations.

We recognize that there is heterogeneity in review article types and that for the many gray areas in medicine where strong evidence is lacking, narrative review articles can provide useful guidance for the clinician (18). However, because biased reports can lead to inappropriate estimates of treat-

ment effects (19, 20), the reports of reviews should be as free of bias as possible and narrative reviews should be held to the same standards as primary research reports or systematic reviews. In the words of Oxman and colleagues (8), "The fact that a review article is published in a peer reviewed journal, even a prestigious one, is no guarantee of scientific quality." Our findings reveal some improvement over the situation in the mid-1980s, but it is important to emphasize that only a minority of published reviews are based on rigorous, systematic methods. Preliminary evidence suggests that the use of explicit criteria in the conduct of reviews (such as those used in the Cochrane Collaboration) may enhance scientific quality (12, 14); we therefore call for more attention to rigorous methods on the part of authors and editors of review articles. Only by such vigilance will review articles "accomplish the task of advancing scientific knowledge" (4).

From University of Oxford, Oxford, United Kingdom; University of Alberta, Edmonton, Alberta, Canada; University of Ottawa, Loeb Health Research Institute, and The Thomas C. Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada; University of Toronto, Toronto, Ontario, Canada; and Audie L. Murphy Veterans Affairs Hospital, San Antonio, Texas.

**Appendix Table. Methods Specified in Review Articles Published in 1985–1986 and 1996**

Criterion	1985–1986 Review Articles (n = 50)*	1996 Review Articles (n = 158)	1996 Meta-Analyses (n = 19)†
	←————— n (%) —————→		
1. The review addressed a focused clinical question	40 (80)	54 (34)	18 (95)
2. The method of locating evidence was described‡	1 (2)	44 (28)	18 (95)
3. Explicit criteria were used to select studies	1 (2)	22 (14)	13 (68)
4. The methodologic validity or quality of the included studies was assessed	1 (2)	14 (9)	13 (68)
5. Assessments of studies were reproducible (that is, they were done by more than one reviewer)	0	18 (11)	14 (74)
6. Directives for future research initiatives were proposed in reviews that included treatment recommendations§:	21 (42)	61 (39)	10 (53)
7. Sources of heterogeneity (clinical or study design) in existing data were addressed	Not assessed	15 (14)	12 (75)
8. Quantitative synthesis of existing data was done	3 (6)	23 (21)	16 (100)
9. The major clinically relevant outcomes (benefits and harms) were considered	Not assessed	38 (34)	10 (63)
10. The generalizability of existing data was addressed	Not assessed	13 (12)	5 (31)
11. Randomized clinical trials were cited as support for treatment recommendations	Not assessed	50 (45)	15 (94)
12. An estimate of treatment effect was provided	Not assessed	53 (48)	16 (100)
13. The treatment effect was expressed as:	Not assessed		
Relative risk or relative risk reduction		21 (19)	7 (44)
Odds ratio		8 (7)	3 (19)
Absolute risk reduction		4 (4)	2 (13)
Number needed to treat		1 (1)	1 (6)
None of the above		19 (17)	3 (19)
14. The precision of the treatment effect estimate (confidence interval) was given	Not assessed	18 (16)	14 (74)
15. The costs of the treatment options were considered	Not assessed	7 (6)	1 (6)

\* Adapted from reference 4.

† The 19 meta-analyses include articles described as "meta-analyses," "systematic reviews," or "overviews" and represent a subset of the 158 review articles from 1996.

‡ This criterion was considered to be fulfilled only if the article described where the evidence was retrieved from (electronic databases, contact with experts, contact with pharmaceutical industry, or hand search of journals or personal files).

§ Includes 111 review articles from 1996 and 16 meta-analyses.

|| Includes treatment effects expressed as incremental changes in surrogate outcomes (such as reduction in blood pressure expressed as mm Hg or reduction in serum cholesterol level expressed as mmol/L) or changes in survival curves or cure rates without reference to contemporaneous controls.

*Acknowledgments:* The authors thank David Sackett for helpful comments on an earlier version of this manuscript, Brian Haynes for details on how *ACP Journal Club* selects journals for review, and Bridget Burchill for secretarial support.

*Grant Support:* Dr. McAlister is a Population Health Investigator of the Alberta Heritage Foundation for Medical Research. Dr. van Walraven is an Arthur Bond Scholar of the PSI Foundation of Ontario, Canada.

*Requests for Reprints:* Finlay McAlister, MD, 2E3.24 WMC, University of Alberta Hospital, 8440 112 Street, Edmonton, Alberta T6G 2R7, Canada; e-mail, Finlay.McAlister@ualberta.ca. For reprint orders in quantities exceeding 100, please contact Barbara Hudson, Reprints Coordinator; phone, 215-351-2657; e-mail, bhudson@mail.acponline.org.

*Current Author Addresses:* Drs. McAlister and Lawson: Department of Medicine, 2E3.24 Walter Mackenzie Centre, University of Alberta Hospital, 8440 112 Street, Edmonton, Alberta T6G 2R7, Canada.

Dr. Clark: Room 405, 737 Parkdale Avenue, Ottawa Hospital—Civic Campus, Ottawa, Ontario K1Y 1J8, Canada.

Dr. van Walraven: F660 Clinical Epidemiology Unit, Ottawa Hospital—Civic Campus, 1053 Carling Avenue, Ottawa, Ontario K1Y 4E9, Canada.

Dr. Straus: Department of Medicine, Mount Sinai Hospital, Suite 427, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada.

Mr. Moher: The Thomas C. Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute, 451 Smyth Road, Ottawa, Ontario, Canada.

Dr. Mulrow: Audie L. Murphy Memorial Veterans Hospital (11C6), 7400 Merton Minter Boulevard, San Antonio, TX, 78284.

## References

1. **McAlister FA, Graham I, Karr GW, Laupacis A.** Evidence-based medicine and the practicing clinician. *J Gen Intern Med.* 1999;14:236-42.
2. **Haynes RB.** Clinical review articles. *BMJ.* 1992;304:330-1.

3. **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.
4. **Mulrow CD.** The medical review article: state of the science. *Ann Intern Med.* 1987;106:485-8.
5. **Oxman AD, Cook DJ, Guyatt GH.** Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA.* 1994;272:1367-71.
6. **Oxman AD, Guyatt GH.** Guidelines for reading literature reviews. *CMAJ.* 1988;138:697-703.
7. **Oxman AD, Guyatt GH.** Validation of an index of the quality of review articles. *J Clin Epidemiol.* 1991;44:1271-8.
8. **Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, et al.** Agreement among reviewers of review articles. *J Clin Epidemiol.* 1991;44:91-8.
9. Science Citation Index. v. 26. In: *Journal Citation Reports, 1994.* Philadelphia: Institute for Scientific Information; 1995.
10. **Haynes RB.** Where's the meat in clinical journals? [Editorial] *ACP Journal Club.* 1993;119:A22-3.
11. **Felson DT.** Bias in meta-analytic research. *J Clin Epidemiol.* 1992;45:885-92.
12. **Bramwell VH, Williams CJ.** Do authors of review articles use systematic methods to identify, assess and synthesize information? *Ann Oncol.* 1997;8:1185-95.
13. **Sacks HS, Reitman D, Pagano D, Kupelnick B.** Meta-analysis: an update. *Mt Sinai J Med.* 1996;63:216-24.
14. **Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, et al.** Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *JAMA.* 1998;280:278-80.
15. **Liberati A, Himel HN, Chalmers TC.** A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol.* 1986;4:942-51.
16. **Cooper HM, Rosenthal R.** Statistical versus traditional procedures for summarizing research findings. *Psychol Bull.* 1980;87:442-9.
17. **Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, Buckingham L, et al.** Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. *JAMA.* 1996;275:308-14.
18. **Naylor CD.** Grey zones of clinical practice: some limits to evidence-based medicine. *Lancet.* 1995;345:840-2.
19. **Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al.** Does the quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* 1998;352:609-13.
20. **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.

© 1999 American College of Physicians—American Society of Internal Medicine