International Journal of Technology Assessment in Health Care, **17:4** (2001), 457–466.

REVIEWING THE REVIEWS

How Strong Is the Evidence? How Clear Are the Conclusions?

Jeannette Ezzo
Barker Bausell
University of Maryland

Daniel E. Moerman
University of Michigan – Dearborn

Brian Berman
University of Maryland

Victoria Hadhazy
Social and Scientific Systems

Abstract

Objectives: The objectives of this paper were: a) to determine what can be learned from conclusions of systematic reviews about the evidence base of medicine; and b) to determine whether two readers draw similar conclusions from the same review, and whether these match the authors' conclusions.
Methods: Three methodologists (two per review) rated 160 Cochrane systematic reviews (issue 1, 1998) using pre-established conclusion categories. Disagreements were resolved by discussion to arrive at a consensual score for each review. Reviews' authors were asked to use the same categories to designate the intended conclusion. Interrater agreements were calculated.
Results: Interrater agreement between two readers was 0.68 and 0.72, and between readers and authors, 0.32. The largest categories assigned by methodologists were "positive effect" (22.5%), "insufficient evidence" (21.3%), and "evidence of no effect" (20.0%). The largest categories assigned by authors were "insufficient evidence" (32.4%), "possibly positive" (28.6%), and "positive effect" (26.7%).
Conclusions: The number of reviews indicating that the modern biomedical interventions show either no effect or insufficient evidence is surprisingly high. Intterrater disagreements suggest a surprising degree of subjective interpretation involved in systematic reviews. Where patterns of disagreement emerged between authors and readers, authors tended to be more optimistic in their conclusions than the readers. Policy implications are discussed.

Keywords: Evidence-based medicine, Review literature, Outcomes assessment (health care), Meta-analysis, Randomized controlled trials

The growing popularity of evidence-based medicine (EBM) has seen a corresponding rise in the popularity of systematic reviews. Systematic reviews are considered the method par

The authors express appreciation to Diane Nickols for editorial assistance and to the many authors of Cochrane reviews for their commitment and diligence and for participating in this study. This project was funded by NIH grant no. 1 R21-RR09327-01.

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

excellence of summarizing the evidence of a given topic in medicine. They have been acknowledged for their usefulness to a variety of readerships including consumers of health care (4), clinicians (16), policy makers (4;6), researchers (3;17), and funding agencies (19).

The recent popularity of evidence-based medicine has led some to query, "Upon what was medicine based all those years before EBM?" Assessments of how much of medical practice is based on scientifically sound evidence have been the topic of recent debate (2;9;26). Various methods have been used to assess the strength of the evidence base of medical practice. The most common method, a prospective approach, requires physicians to document all patient contacts, treatments, and treatment rationales within a specified time period. These practices are subsequently evaluated by a gold standard such as an expert review panel (10) or a literature review (9). Another approach, the case study method (22;23), tracks a commonly used treatment retrospectively for the historical trail of evidence that presumably led to the treatment's wide acceptance and use. Results of these medical evidence studies range from conclusions that very little of medical practice is evidence-based (10;22;23) to a recent conclusion that most of medical practice is evidence-based (9). Now that numerous systematic reviews are available, another option for assessing the scientific foundation of a medical practice emerges: to assess the conclusions of these evidence-based summaries.

When weighing the evidence presented in systematic reviews, however, an additional issue must be addressed. Are the conclusions clear enough that two people reading the same review of the same evidence draw the same conclusions about the effectiveness and safety of a treatment? Also, does the reader's interpretation match what the authors intended?

OBJECTIVES

The following review of reviews has three major objectives:

1. To determine what can be learned from the aggregate conclusions of systematic reviews about the evidence base of medicine;
2. To determine whether two readers reading the same systematic review draw similar conclusions; and
3. To determine whether the readers' conclusions match the conclusions intended by the authors of the review and to identify specific review features that are associated with reader–author agreement.

METHODS

To achieve these objectives, systematic reviews from the Cochrane electronic library were selected. Cochrane reviews were selected over print reviews due to concerns that the latter might be prone to a similar publication bias documented for clinical trials (7), thereby leading to an overly optimistic view of the evidence. By contrast, the Cochrane Collaboration publishes all systematic reviews that have passed peer review, regardless of outcome. Furthermore, Cochrane reviews more often include features to minimize other biases, such as a rating system for trial methodologic quality and no language restrictions (14).

Pretest

The rating system for conclusions was based on a two-phase pretest. In phase I, the three readers (JE, DM, VH) were presented with the conclusion categories used in a previous review of reviews (15) of either positive, negative, or neutral (nonsignificant) and were asked to rate 10 reviews using the classification system in order to ascertain whether this

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

system was sufficiently adequate to use in the larger study. Readers (two per review) were instructed to base conclusions on the review's primary outcome. Discussion following the pretest revealed that although the readers agreed that these categories presented the statistical possibilities, the categories did not sufficiently capture clinically important distinctions. For example, readers expressed concern that an explicit category of "treatment does more harm than good" had not been included, and this was subsequently added. Also, readers suggested that some reviews tended to show positive findings, but a major methodologic issue remained unresolved, such as a small sample size or all trials being methodologically flawed. In these circumstances, readers agreed that "positive" was too definitive, whereas "insufficient evidence" was also inaccurate. A category of "possibly positive" was established to accommodate these cases. Finally, readers suggested that a statistically nonsignificant finding could have one of three very different clinical interpretations, and the conclusion categories should reflect these: a) insufficient evidence (findings were not significantly different between the two groups, and more evidence is needed); b) evidence of no effect (findings were not significantly different between the two groups, and there was enough evidence to say that there is not a treatment difference); and c) two active treatments were equivalent, and both were effective. The conclusions classification was revised to incorporate readers' suggestions, and the final classification included six possibilities (Table 1).

Readers rated eight more reviews and found they were in agreement approximately three-fourths of the time. Discussion of the disagreements appeared to be due to real differences in the interpretation of the study result. The readers agreed the six-item classification system adequately captured the reviews' clinical possibilities, and so this classification was used for the subsequent review of reviews.

The Review of Reviews

All completed reviews in the *Cochrane Library*, issue 1, 1998 (n = 326) were assessed for the number of included trials and descriptions of the methodologic quality of those trials. A subset of 160 reviews was alternately selected for inclusion except when alternate selection yielded a review already used in a pretest or a complementary medicine review, and then the next review was chosen. Complementary medicine treatments (i.e., acupuncture, spa therapy, herbals) were omitted from the selection process, so that reviews represented conventional medical practice. Two readers per article categorized the conclusions of 160 reviews using the categories in Table 1. When various outcomes were measured and there was a discrepancy in the results, the primary outcome was taken as the variable of interest. The interrater agreements were calculated and consensual scores arrived at by discussion. Authors of these reviews were then contacted and asked to use the same classification system to best select the conclusion that matched what they had intended for their review. Interrater agreements between the authors' scores and the consensual readers' scores were then calculated.

Additional data were extracted from the reviews that received an author response. These data were: a) whether side effects had been discussed in the abstract; b) whether the quality of the trials was discussed in relation to the authors' conclusions; c) whether data had been pooled and an effect size with confidence intervals presented; d) whether pooled effects were based on individual patient data; e) the sample size of the largest trial in the review; and f) whether a multicenter trial was represented. It was hypothesized that reviews that discussed side effects, incorporated quality of trials into conclusions, had pooled data, had used individual patient data, had at least one trial of over 500 patients, had a multicenter trial, or had a large number of included trials would add to the clarity of the review and therefore would be associated with reader-author agreement.

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

Table 1. Conclusion Categories, Definitions, and Proportions Classified by Readers and Authors

Conclusion category	Definition	Readers' consensual rating for all included reviews (n = 160) n (%)	Readers' consensual rating for reviews to which Authors responded (n = 105) n (%)	Authors' rating (n = 105) n (%)
Positive effect	Treatment is more beneficial/ effective than control for the primary outcome.	36 (22.5)	23 (21.9)	28 (26.7)
Possibly positive effect	Treatment may have a positive effect, but a major unresolved methodologic issue, such as all studies being very low quality, or findings based on only 1 study, precludes making a definitive statement.	30 (18.8)	18 (17.1)	30 (28.6)
Two active treatments are equal	Two biologically active treatments, such as two antibiotics, are equally as effective for the condition being treated. This category to be used only when comparing two active treatments, not placebo or no treatment.	15 (9.4)	8 (7.6)	1 (1.0)
Insufficient/inconclusive evidence	There is not sufficient evidence to determine effectiveness.	34 (21.3)	25 (23.8)	34 (32.4)
No effect	There *is* sufficient evidence and there is no effect.	32 (20.0)	20 (19.0)	3 (2.9)
Harmful effect	Treatment does more harm than good.	13 (8.1)	11 (10.5)	9 (8.6)

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

RESULTS

Descriptive Statistics

The number of primary studies included per Cochrane review was small (median = 5; range, 0–47). Approximately one-fifth of the 326 reviews commented that the quality of the included trials was either generally poor or methodologically limited.

Comparison of Readers' Ratings

Readers' interrater agreements on the reviews' conclusions were 0.68 and 0.72, respectively, for readers 1 and 3 and readers 1 and 2, indicating moderate agreement (21). "Evidence of positive effect" was the largest category for reviews (n = 36, 22.5%), followed by "insufficient evidence" (n = 34, 21.3%) and "evidence of no effect" (n = 32, 20.0%). The fewest number of reviews were classified as "harmful" (Table 1).

Areas where readers commonly disagreed were in interpreting statistically nonsignificant findings, such as when one reader thought a review did not have sufficient evidence to say the treatment was "not effective" and the other reader thought there was enough evidence presented to say that the treatment was "not effective." Another common area of disagreement occurred in areas where one reader thought the evidence was only strong enough to say "possibly positive" and the second reader thought there was enough good evidence to say "positive." Readers most commonly agreed on the conclusion that intervention did more harm than good.

Comparison of Authors' Ratings with Readers' Consensual Ratings

Of the 160 reviews included in this study, authors responded to 105; therefore, our findings are based on authors' responses to 105 reviews. The interrater agreement between the author of the review and the readers' consensual score (kappa = 0.32) was notably less than the agreement between the two readers. Because some cells (i.e., possibly positive and positive) might have substantially overlapped, these cells were collapsed and the kappa recalculated. It was only modestly improved (kappa = 0.43). It was suggested that the cell "two treatments appear equal" may have been unclear to the authors, since only one author made this selection. When this category was removed from the analysis and the collapsed cells retained, the kappa increased to 0.54.

"Insufficient evidence" was the largest category for authors' ratings (n = 34, 32.4%), followed by "possibly positive effect" (n = 30, 28.6%) and "evidence of positive effect" (n = 28, 26.7%) (Table 1). Authors tended to select the "positive effect" or "possibly positive effect" almost 15% more often than the readers. The most notable difference between authors' and readers' ratings in a single category was in the "evidence of no effect" category. "Evidence of no effect" was among the most frequently selected categories for readers, whereas authors rarely selected it (n = 3, 2.9%).

With one exception, authors and readers were in agreement for the greatest proportion of reviews in each category (Table 2, see diagonal). The one exception was a systematic disagreement in which many of the reviews rated by readers as "evidence of no effect" were rated by authors as "insufficient evidence" (Table 2). Another pattern in the disagreement was one in which many of the reviews rated by authors as "possibly positive" were rated by readers as either "insufficient evidence," "evidence of no effect," or "evidence of harm" (Table 2). Other disagreements did not follow such a clear pattern.

To estimate the proportion of reviews that represent practices based on weak or no evidence, the conclusion categories "evidence of no effect" and "insufficient evidence" were summed. According to readers' consensual scores, this was 41.3% (n = 66 of 160) of the reviews represented in this sample. According to authors' ratings, this was more than one-third (35.3%, n = 37 of 105).

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

Table 2. Readers' Conclusions Compared with Authors' Conclusions

Authors' conclusions	Readers' conclusions						
	Positive effect	Possibly positive effect	Two active treatments are equal	Insufficient evidence	No effect	Harmful effect	Row (authors') totals
Positive effect	**15**	6	4	2	1		28
Possibly positive effect	6	**10**	1	5	5	3	30
Two active treatments are equal			**1**				1
Insufficient evidence	1	1	2	**18**	9	3	34
No effect		1			**2**		3
Harmful effect	1				3	**5**	9
Author did not respond	13	12	7	9	12	2	55
Column (readers' totals)	36	30	15	34	32	13	160

Bolded diagonal depicts number of reviews with agreement between readers and authors.

Of the 105 reviews, 49 (46.7%) mentioned side effects in the abstract, 19 (18.1%) incorporated the quality of the trials into explaining the conclusions, 84 (80%) presented pooled data, 25 (23.8%) contained at least one trial of greater than 500 patients, none meta-analyzed individual patient data, and seven (6.7%) mentioned a multicenter trial. None of these factors was associated with reader–author agreement.

Using the median number of trials ($n = 5$) to dichotomize data, an association was found between reader–author agreement and number of total trials in the review ($n \leq 5$ vs. $n > 5$) ($X^2 = 6.23$, $p = .01$); however, contrary to what was expected, there was greater agreement among reviews with less than or equal to five trials, mostly driven by reader–author's agreement that there was insufficient evidence.

DISCUSSION

The low agreement between readers and authors is striking. A question arises as to why authors' ratings differed so much from the readers' conclusions. There are several possibilities: first, the classification system may not have been sufficiently clear to the authors even though definitions were provided to authors along with the classification options; however, collapsing cells for which there could be substantial overlap failed to substantially improve the interrater agreement. Although a categorization system with fewer categories may have produced greater reliability, it is likely such a system, such as the one we started with, is likely to omit important, clinically relevant interpretations.

Second, many authors may be clinicians, whereas all readers were methodologists, and it may be more difficult for clinicians to separate other sources of information, including their actual experience with a treatment, from the quantitative review of that treatment. Disagreement between clinicians who are experts in their fields and methodologists has been observed in previous research on systematic reviews (20). One way this bias might operate is that the results of a systematic review could be positive, supporting the existence of a treatment effect, and therefore rated as positive by a clinician, but this effect could be judged to be too little by the methodologists to support the use of the intervention.

A third explanation may be that authors, regardless of whether they are clinicians, may tend to be overly optimistic about their review's conclusions. It is noteworthy that

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

when patterns of reader–author disagreement emerged, disagreement was in the direction of authors being more optimistic about the reviews' conclusions than were the readers. Authors' optimism has been previously demonstrated in reports of clinical trials (5) and in another review of Cochrane reviews (18).

A fourth reason for the reader–author disagreement may be that conclusions of reviews are ambiguously worded; indeed, some (25) have criticized the wording of Cochrane reviews as being overly vague. By contrast, others have argued (13) that imprecise wording reflects the state of the science since only "large randomized trials with a complete absence of methodological flaws" can prove something with certainty.

The fact that we could identify only one feature of the reviews associated with reader–author agreement, and that pertained to agreement of insufficient evidence rather than any definitive effectiveness category, illustrates that this issue of interpretation is a complex one. The difficulty is likely associated with the fact that many of the trials upon which medical practice is based are low quality and poorly powered. Theoretically, one would expect that there is less ambiguity about clinical treatments for which there are many high-quality randomized controlled trials, but for most of the treatments in this study, there is no such bank of studies. Clinicians and researchers, therefore, must do their best to interpret what is available, which is usually less than optimal. Given this state of the science, more definitive wording of reviews' conclusions or suggesting that authors use a forced-choice set of conclusions, such as the roster used in this study, would be of only limited value. The bank of poorly conducted trials remains unchanged.

POLICY IMPLICATIONS

How then, given the substantial weight of systematic reviews in guiding policy and practice guidelines, can information be clarified and optimized in systematic reviews? From our analysis, we have identified three areas wherein the conduct and/or reporting of systematic reviews can be more informative: a) report side effects; b) clearly state how the quality of trials has influenced the overall conclusions of the review; and c) when possible, analyze individual patient data.

Side Effects

The absence of the mention of side effects in most reviews in this study leaves the reader with unanswered questions. Were side effects not reported in the original trials, or were they reported in the trials but not in the review? Even a review that states that no adverse effects were reported in the original trials is more informative than a review that fails to mention adverse effects at all. Patients clearly weigh adverse effects against the benefits of a treatment and will even select a less effective treatment if there are also fewer adverse effects (11); therefore, failing to mention adverse effects in a review does a disservice to the reader. Clearly, the reporting of adverse effects can only be as good as the reporting in the original trials. One way to optimize reporting of side effects is for authors to read all publications on a given trial, since side effects are more likely to be reported in the earliest articles (12).

Quality of the Trials

Although it is a part of every Cochrane protocol to evaluate the quality of the trials, seldom do reviewers indicate how the quality of the trials influenced their final conclusions. Some authors have used methods that incorporate trial quality into the conclusions (24), but this remains the exception. Clinical guidelines and recommendations are increasingly using defined criteria for a qualitative summary of effect, and this should also be explicit in systematic reviews.

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

Individual Patient Data

For consensus statements as well as clinical guidelines and recommendations, the advantages of meta-analysis of individual patient data supercede pooling data of published results. Meta-analysis of individual patient data permits researchers to identify subgroups of patients for whom the treatment may be particularly effective, eliminate reporting bias of outcomes and adverse events, calculate adverse event rates, identify subgroups at risk for adverse events, and compare dosing regimens that may not even be part of the same clinical trial.

As a case in point, the recent NIH Consensus Conference on adjuvant therapy for breast cancer (1) relied heavily on meta-analyzed individual patient data to formulate guidelines for the use of tamoxifen (8). Through meta-analysis of individual patient data, reviewers were able to identify the subgroups of women who were most likely to respond to adjuvant tamoxifen (i.e., estrogen receptor–positive women), the subgroup that was most at risk for adverse events (i.e., women over 50), the more effective dosing regimen (5 years on the drug conferred greater benefit than 2 years), and adverse event rates.

Limitations of This Study

Our analysis of Cochrane reviews reflects the reviews in the Cochrane Library and therefore may not be generalizable to all medical practice or to a particular subspecialty. The Cochrane Library mirrors the efforts of Collaborative Review Groups, the work groups that are organized around a health condition within the collaboration and as yet are not representative of all medical practice. The intent of our analysis was to sample a cross-section of the Cochrane Library. No effort was made to select the reviews that were the most representative of medical practice; therefore, certain conditions may be overrepresented and others underrepresented. The findings of this study, however, are still remarkable—that there is a substantial number of fairly common, widely utilized medical interventions that are based on weak or no evidence.

Another potential limitation of using the Cochrane Library is that reviews with few and poor studies may have been the first to be completed, thereby overrepresenting topics with insufficient or no evidence. A similar study of print reviews may help address this issue.

It is noteworthy that one of the categories of the highest agreement between two readers, and also between readers and authors, was the case of interventions that did more harm than good. An understated value of systematic reviews is the ability to inform readers when an intervention is harmful.

Implications for Research

The findings of this study raise an important question for the practice of EBM. Would clinicians whose practice decisions may be influenced by the finding of a systematic review interpret the conclusions in the way the authors have intended? This should be a topic of future research. Also, when the author/reader disagreement followed a pattern, the pattern suggested that authors tended to be more optimistic about an intervention than did the readers. This potential source of bias in reviews warrants further investigation, such as a blinded study in which peer reviewers are blinded to the conclusions of a review until after reading the review and formulating a conclusion.

CONCLUSIONS

Both the number and quality of the primary studies on which much contemporary medical practice stands are remarkably weak. Readers all noted that the most common statement in Cochrane reviews was the complaint of the very poor quality of many clinical trials. The

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

number of reviews indicating that modern biomedical procedures show no effect (32/160; 20%) or insufficient evidence (34/160; 21%) seems very high, and the number indicating significant evidence of desirable effect (36/160; 23%) seems very low. If these reviews represent medical practice, it is much closer to White's estimate (26) than Ellis and colleagues (9).

The moderate scores on interrater agreements suggest that in several instances the conclusions of the review may be ambiguous, leading two readers to interpret the same review differently. The low interrater agreement scores between readers and authors suggest that the conclusions intended by the authors may, in many instances, not be the ones drawn by the readers. Authors tend to be more optimistic in their conclusions than the readers. Further research needs to be done to confirm whether reviews' conclusions may be biased toward being overly optimistic. Maximizing the information that can be gleaned from systematic reviews is a particularly timely issue because systematic reviews are increasingly being used to guide practice and policy.

REFERENCES

1. Adjuvant therapy for breast cancer. *NIH consensus statement*. Bethesda, Md: National Institutes of Health; 2000.
2. Aveyard P. Evidence-based medicine [letter]. *Lancet*. 1995;346:840.
3. Bausell RB. After the meta-analytic revolution. *Evaluation and the Health Profession*. 1993;16: 2-12.
4. Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. *Ann Intern Med*. 1997;127:37-42.
5. Cherkin D. Efficacy of acupuncture in treating low back pain: A systematic review of the literature. Paper presented at the NIH Consensus Conference on Acupuncture, Bethesda, Md, November 3–5, 1997.
6. Dickersin K, Manheimer E. The Cochrane Collaboration: Evaluation of health care and services using systematic reviews of the results of randomized controlled trials. *Clin Obstet Gynecol*. 1998;41:315-331.
7. Dickersin K, Min Y, Meinert CL. Factors influencing publication of research results. *JAMA*. 1992;267:374-378.
8. Early Breast Cancer Trialists' Collaboration Group. Tamoxifen for early breast cancer: An overview of the randomised trials. *Lancet*. 1998;351:1451-1467.
9. Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. *Lancet*. 1995;346:407-409.
10. Forsyth G. An enquiry into the drug bill. *Med Care*. 1963;1:10-16.
11. Gotzsche PC. Patients' preferences in indomethacin trials: An overview. *Lancet*. 1989;1:88-91.
12. Gotzsche PC. Multiple publication of reports of drug trials. *Eur J Clin Pharm*. 1989;36:429-432.
13. Jacobs A. Imprecise conclusions may be inevitable. *BMJ*. Available at: www.bmj.com/cgi/content/full/322/2/8/83.
14. Jadad AR, Cook DJ, Jones A, 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;280:278-280.
15. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: A systematic qualitative review of their methodology. *J Clin Epidemiol*. 1996;49:235-243.
16. Lehmann HP, Goodman SN. Specifications for formalizing clinical significance. *Med Decis Making*. 1995;15:424.
17. Meinert C. Meta-analysis: Science or religion. *Control Clin Trials*. 1989;10:257S-263S.
18. Olson O, Alderson P, Ezzo J, et al. Quality of Cochrane Reviews. *BMJ*. In press.
19. O'Toole LB. Using systematically synthesized evidence to inform the funding of new clinical trials—The UK Medical Research Council Approach. Paper presented at the Sixth International Cochrane Colloquium, Baltimore, Md, October 22–26, 1998.
20. Oxman AD, Guyatt GH, Singer J, et al. Agreement among reviewers of review articles. *J Clin Epidemiol*. 1991;44:91-98.
21. Rosner B. *Fundamentals of biostatistics*. London: Duxbury Press; 1990:458.

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press

22. Smith R. Where is the wisdom. . .? The poverty of medical evidence. *BMJ*. 1991;303:798-799.
23. Smith R. The ethics of ignorance. *J Med Ethics.* 1992;18:117-118.
24. Van Tulder MW, Cherkin CD, Berman BM, Lao L, Koes BW. Conservative treatment of acute and chronic lower back pain: A systematic review of the most common interventions. *Spine*. 1997;22:2128-2156.
25. Vickers A. On breakfast and randomised trials. *BMJ*. 2001;322:85.
26. White K. Evidence-based medicine [letter]. *Lancet*. 1995;346:837-838.

https://doi.org/10.1017/S0266462301107014 Published online by Cambridge University Press