The Science of Reviewing Research[a]

ANDREW D. OXMAN[b,c,d] AND GORDON H. GUYATT[d,e]

Departments of [c]Family Medicine
[d]Clinical Epidemiology and Biostatistics, and [e]Medicine
Faculty of Health Sciences
McMaster University
Hamilton, Ontario, Canada L8N 3Z5

AUTHORITY, SUPERSTITION, AND SCIENCE

Medical practitioners enjoy positions of authority. Considerable economic resources are directed to the practice and improvement of medicine. Courts of law recognize qualified medical practitioners as being expert witnesses on many matters. Witchdoctors and other alternative health care providers do not receive comparable resources or recognition. Why is this? Modern medicine is based on science, whereas witchcraft and other alternatives are based on "superstition."[1] But what is it that distinguishes science from superstition, and to what extent can medical practitioners claim to be scientific?

Most medical practitioners are not likely to have given much thought to questions such as these. According to Thomas Kuhn, the same can be said of practitioners of other scientific disciplines.[2] Kuhn has described scientific paradigms as encompassing all that which the practitioners of a particular scientific discipline take for granted. The paradigm constitutes the framework within which the scientists reason when they try to solve their scientific problems. It represents the premises of scientific thinking and therefore is not usually considered a scientific problem in itself. As Kuhn points out, as a general rule, scientists do not learn concepts, laws, and theories in the abstract. Instead, they gradually learn to use these intellectual tools by reading and by listening. As a consequence, scientists may "learn easily and well about the particular individual hypotheses that underlie a concrete piece of current research," but in spite of that "they are little better than laymen at characterizing the established basis of their field."[2]

When defects in an existing paradigm accumulate to the extent that the paradigm is no longer tenable, the paradigm is challenged and replaced by a new way of looking at the world. Kuhn, a physicist, is particularly interested in physics, chemistry, and astronomy, and uses examples from the history of these sciences to illustrate his ideas. It is not clear that science always develops in leaps and bounds as described by Kuhn. It is particularly uncertain to what extent Kuhn's theory correctly describes the development of medicine, which encompasses both clinical research and the practice of medicine. Moreover, medicine comprises a variety of subdisciplines. The paradigm underlying medical thinking is likely to vary, for example, from pathologist to psychiatrist.[3]

[a] This work was supported by Grant No. 01969 from the Ontario Ministry of Health. Drs. Guyatt and Oxman are Career Scientists of the Ontario Ministry of Health.

[b] Address for correspondence: Dr. Andy Oxman, Department of Family Medicine, McMaster University Medical Centre 1200 Main Street West, Room 2V10, Hamilton, Ontario, Canada L8N 3Z5.

Nonetheless, it is tempting to describe changes that have occurred over the past 30 years as a paradigm shift.[3,4] In the 1960s, an increasing number of clinicians began to demand empirical proof of the effectiveness of medical interventions. Led by pioneers such as Archie Cochrane, Austin Bradford Hill, Richard Doll, and more recently by people like Richard Peto and Iain Chalmers, the randomized controlled trial has emerged as the ideal—or paradigm—of clinical research.

Kuhn's view is that scientists working within a scientific paradigm need not necessarily engage in philosophical debate regarding the paradigm so long as their activity within the paradigm is productive. Others, such as Popper, argue that this is dangerous. According to Popper, a scientist who is not critical of the paradigm within which s/he works "has been taught in a dogmatic spirit" and is "a victim of indoctrination."[3]

In this paper we will briefly describe the shift from "authoritative reviews" of medical problems to systematic reviews that place particular emphasis on the role of "experts" in the process of synthesizing the results of research. This shift can be viewed as an extension of the shift from a paradigm that relied heavily on unsystematic clinical experience and pathophysiologic rationales to one that stresses rigorous clinical evaluations of medical interventions. Finally, we would like to reiterate that we need to be critical of the "new paradigm," even though it is highly productive.

AUTHORITATIVE REVIEWS VERSUS SYSTEMATIC REVIEWS

Traditionally, review articles which survey an area of scientific inquiry or clinical practice have been written by experts in the field. When seeking critiques of review articles (peer review), editors have looked to other experts in the field for help. These policies are illustrated by the two following responses to a small survey of editors of medical journals undertaken in 1986:[5]

> From the point of view of an editor of a journal, the acceptability of research reviews depends greatly on the advice given by experts in the field. Reviews should give an adequate coverage of the literature and it is only those who know the field who would be able to advise on whether this had been done. The assessment of original articles of course is another matter and criteria for this activity have been published in a number of places.

> The people chosen to serve on the Editorial Committee are those who we believe have a good grasp on their particular subfields within medicine. At a yearly meeting, each Committee member proposes certain topics with specific authors to be invited. An invitation to prepare a review on a particular topic is then sent to the chosen author. From this procedure I think you can see that we rely heavily on the expertise of our individual Committee members as guided by the whole group to choose qualified reviewers. We thus monitor the quality of the manuscripts before they are even written!

These policies may appear intuitively reasonable and appropriate. However, there are reasons for serious skepticism. It is possible that experts might lack the objectivity desirable in preparing or critiquing a review article. For example, personal experience in primary research is highly salient and considerably more vivid than the research of others, and therefore likely to be overweighted in judgments.[6] This is also true for personal clinical experience.

When the consistency of expert ratings of journal articles has been examined, it has been found to be poor. Ten articles from which quantitative estimates of

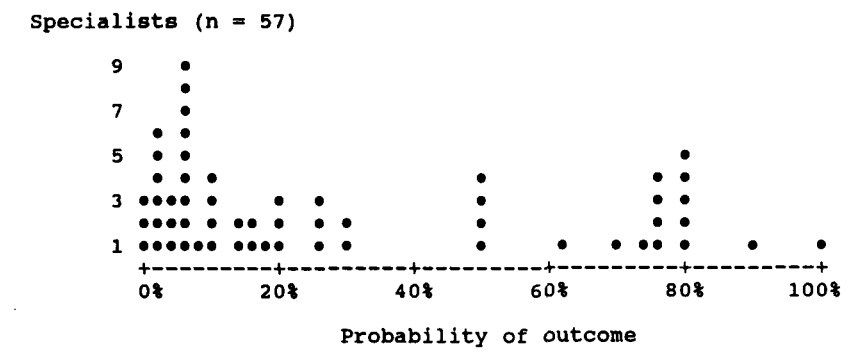

FIGURE 1. Specialists' beliefs about the probability of a particulaly important outcome for a common and important intervention. (The specialty society that convened the meeting at which the estimates were obtained requested to remain anonymous.) (Adapted from Eddy.[17])

consistency (inter-judge agreement) are available found correlation coefficients of from 0.19 to 0.54, with most results clustering around the lower values.[7-16] While the need for agreement among peer reviewers has been challenged, and its desirability has been questioned,[19,20] agreement is important, though not sufficient, to ensure the quality of the peer review process. If peer reviewers cannot agree, the quality of their judgments must be considered unreliable.

Similarly, experts often disagree about the results of a review. Eddy has provided the following example of this problem: A group of medical specialists met to develop a guideline for a common and important intervention. When they were asked to write down their beliefs about the probability of a particularly important outcome in patients receiving this intervention their answers ranged from 0% to 100% (FIG. 1).[17] Whatever mental processes the experts used to arrive at their beliefs, they had very different perceptions.

The problem with not knowing the reasoning that was used is that it is impossible to critique the methods that were used. There is no way to tell which answer is most correct. Moreover, when information is synthesized and probabilities are estimated informally, there are a number of factors that can lead to systematic errors in the judgments that are made.[18] One such bias is a tendency to overlook small but clinically important effects when research is synthesized subjectively. Cooper and Rosenthal demonstrated this experimentally by randomly assigning reviewers to either use or not use meta-analysis to combine the results of several studies. The studies, which included some that did not show significant results, demonstrated an overall significant effect ($p = 0.016$). The reviewers not using meta-analysis were significantly more likely to find little or no support for the hypothesis being tested.

THE RELATIONSHIP BETWEEN EXPERTISE AND METHODOLOGIC RIGOR

In the process of developing criteria for formal evaluation of the methodologic rigor of review articles,[19-21] we addressed two issues related to the role of expertise

TABLE 1. Criteria for Assessing the Methodologic Rigor of Research Reviews

1. Were the search methods reported?
2. Was the search comprehensive?
3. Were the inclusion criteria reported?
4. Was selection bias avoided?
5. Were the validity criteria reported?
6. Was validity assessed appropriately?
7. Were the methods used to combine studies reported?
8. Were the findings combined appropriately?
9. Were the conclusions supported by the reported data?
10. What was the overall scientific quality of the review?

in the process of preparing and critiquing review articles. We compared the consistency of assessments of the methodologic rigor of review articles using our criteria by experts in the field (given instructions, but no training) with that of non-experts (trained to apply our criteria in a standardized way). In addition, we examined the relationship between the expertise of the author and the methodologic rigor of the review article.

Methods

The results of our assessment of the reliability and validity of the criteria have been reported elsewhere.[20,21] Twelve judges evaluated the methodologic rigor of 36 published review articles using the criteria shown in TABLE 1. The review articles were drawn from three sampling frames: articles highly rated by criteria external to the study; meta-analyses; and articles selected from a broad spectrum of medical journals. Four categories of judges assessed the articles: research methodologists, clinicians with research training, research assistants, and content-area experts, with three judges in each category. The non-experts all received training in the application of our criteria.

Authors of the review articles were surveyed. Respondents were asked to categorize their level of expertise using the following seven-point scale:

1	2	3	4	5	6	7
Limited Background		Knowledgeable		Very Knowledgeable		Expert

Expert = Prior to writing the review, you had already read extensively in this area, *and* done research or written articles on the same topic.

Very Knowledgeable = You had already read extensively on this topic, but *not* done research in this area.

Knowledgeable = You kept up with the literature in this general area routinely, *and* were familiar with *most* of the primary research in this area.

Limited Background = You had *not* read most of the primary literature directly relevant to the topic of this review.

Respondents were also asked to estimate the amount of time they spent preparing their reviews, and to rate the strength of their prior opinions on the topic of the reviews as follows:

Please indicate the *strength of your opinion prior* to preparing this review, with respect to the primary question that the review addresses.

Respondents were asked to categorize the strength of their opinions using the following seven-point scale:

1	2	3	4	5	6	7
Decided		Strong Opinion		Week Opinion		Undecided

An intraclass correlation coefficient (ICC), which is the ratio of the variance between review articles to the total variance, was used to measure agreement among judges.[22] The ICC's and their 95% confidence intervals (CIs) were calculated according to Shrout and Fleiss' guidelines.[23] The analyses were done using BMDP.[24]

Spearman rank order correlations between the degree of expertise of the author and the author's strength of prior opinion, the amount of time spent preparing the review article, and the quality of the review were calculated. For the correlation of degree of expertise with the quality of the review, the quality of the review was determined by taking the mean global rating of the reviewers (groups 1, 2, and 3) who reviewed all 36 articles (there is a summary question which asks for a global rating of the methodologic quality of the review).

Results

The intraclass correlation coefficients for each of the four groups of raters are summarized in FIGURE 2. Consistency of ratings was higher for groups 1 to 3 than for the experts on each of ten questions. The gradient between groups 1 to 3 and the experts was considerably greater for questions which required substantial judgment (questions 2, 4, 6, 8, 9, 10) than for questions which did not require as much judgment (questions 1, 3, 5, 7). For the overall rating of methodologic rigor, the intraclass correlations were 0.79 (95% CI, 0.65–0.87) for group 1, 0.77 (95% CI, 0.51–0.79) for group 2, 0.69 (95% CI, 0.38–0.78) for group 3, and 0.23 (95% CI, 0.03–0.45) for the experts.

Thirty of thirty-six authors (83%) responded to our survey concerning their methods. The correlation between expertise and strength of prior opinion was 0.55 ($p = 0.03$); the more expertise, the stronger the prior opinion. The correlation between expertise and the amount of time spent preparing a review was -0.40 ($p = 0.045$); the more expertise, the less time. The correlation between expertise and the quality of a review was -0.52 ($p = 0.004$); the more expertise, the lower quality.

DISCUSSION

At least two possible explanations for the poorer agreement among the experts are possible: either lack of training, or expertise itself. In either case, these results cast doubt on the wisdom of relying exclusively on experts without specific training to assess the methodologic rigor of review articles.

These results also cast doubt on the wisdom of relying on experts to be solely responsible for preparation of review articles. Our data suggest that experts, on average, write reviews of inferior quality; that the greater the expertise the more likely the quality is to be poor; and that the poor quality may be related to the

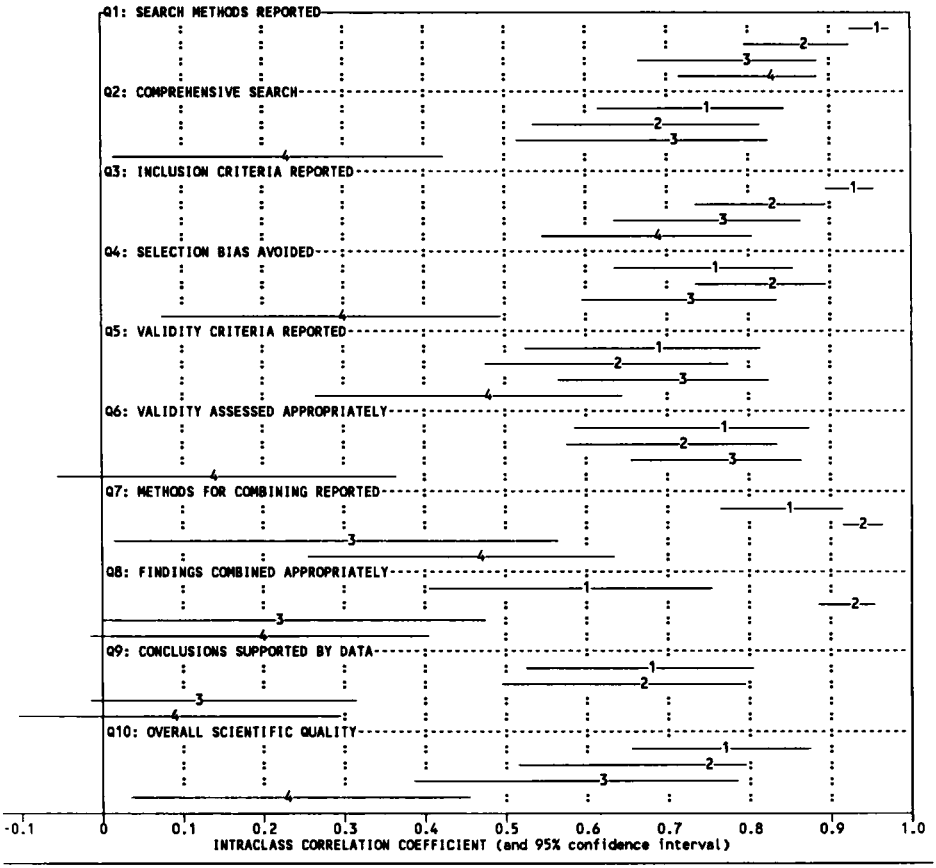

FIGURE 2. Agreement within groups of judges. 1 = experts in research methodology (group 1); 2 = MDs with research training (group 2); 3 = research assistants (group 3); and 4 = content-area experts (group 4).

strength of their prior opinions and the amount of time they spend preparing a review article.

What is it about expertise that might predispose to such difficulties in making judgments of methodologic rigor, and such problems in preparing high-quality reviews? It is natural that investigators will be biased so that they give more weight to their own work versus that of others. In addition, many investigators will have strong opinions about their area which will lead them to judge evidence differently according to whether it supports their beliefs. A third issue is the personal competitiveness and antagonism that unfortunately often plays a role in scientific endeavor and scientific debate.

An extreme interpretation of the results of this investigation might be that experts should occupy themselves with the task of producing new data, or retire from the topics of their expertise,[25] and so in either case leave judgments and summaries of their efforts to those who have specific training in the science of research reviews. A more reasonable interpretation might be to acknowledge the value of expertise in recognizing subtle, but important clues that someone without expertise might overlook while, at the same time, appreciating the risks of blind faith in the subjective thought processes of experts or anyone else.

As Louis Pasteur once wrote: "Chance favors the prepared mind."[26] Expertise might be of great value in this regard, provided experts are able to follow the injunction cited by Iain Chalmers in an article published a decade ago on scientific inquiry and authoritarianism: "Teach thy tongue to say 'I do not know' and thou shalt progress."[27]

CONCLUSION

Systematic and explicit approaches to reviewing research are essential, but not sufficient to ensure the validity of the results. "Science, no less than painting, cannot be done by numbers."[28] And, in the words of J. M. Ziman:[29]

> Our present system of rewards and incentives in science does not encourage individuals to devote themselves for years on end to these critical synthesizing activities. "Recognition," by way of professional advancement and prestige, is given solely for primary research; has any academy ever mentioned that the hero was the author of a valuable treatise or of the authoritative review that has since determined the course of research in his field?
>
> The trouble is, quite simply, a matter of philosophy. We are so obsessed with the notions of discovery and individual originality that we fail to realize that scientific research is essentially a corporate activity, in which the community achieves far more than the sum of the efforts of its members.

We are delighted that the New York Academy of Sciences has chosen to present the L. W. Frohlich award to Richard Peto and Iain Chalmers for their pioneering work as proponents and practitioners of systematic reviews of randomized controlled trials rather than for their "authoritative reviews." The successes that they have had, and the vision they have shown, in organizing international efforts to critically synthesize and keep up-to-date the scientific basis of medical practice are inspiring.

We would, nevertheless, like to come back to the need to remain critical of the very paradigm that they have helped to pioneer. There is still an enormous amount of productive work to be done within this paradigm. However, to capitalize on the solutions that can be derived within this paradigm, knowledge that is derived from other "paradigms" is also needed. In particular, on the level of clinical practice "hermeneutics" (interpretative reflection) is necessary to appreciate the subjective reality of individual patients; and on the level of policy, we clearly have a long way to go to ensure that the results of good clinical research actually lead us to do more good than harm.

ACKNOWLEDGMENTS

We would like to express our appreciation to the judges for their contribution to this study, particularly Drs. Charlie H. Goldsmith, Brian G. Hutchison, Ruth

A. Milner and David L. Streiner. We would also like to thank Dr. Joel Singer for his assistance with the analyses, and the authors of the review articles who responded to our survey.

REFERENCES

1. BRISKMAN, L. 1988. Doctors and witchdoctors: Which doctors are which? *In* Logic in Medicine. C. I. Phillips, Ed.: 1–16. British Medical Journal. London.
2. KUHN, T. S. 1970. The Structure of Scientific Revolutions, 2nd ed.: 46. The University of Chicago Press. Chicago.
3. WULFF, H. R., S. A. PEDERSEN & R. ROSENBERG. 1990. The paradigm of medicine. *In* Philosophy of Medicine, 2nd ed.: 1–12. Blackwell Scientific Publications. Oxford, U.K.
4. EVIDENCE-BASED MEDICINE WORKING GROUP. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. JAMA **268:** 2420–2425.
5. OXMAN, A. D. 1987. A Methodological Framework for Research Overviews. M.Sc. thesis.: 212–218. McMaster University. Hamilton, Ontario.
6. COOPER, H. M. 1986. On the social psychology of using research reviews: The case of desegregation and the black achiever. *In* Social Psychology of Education. R. S. Feldman, Ed.: 341–363. Cambridge University Press. Cambridge, U.K.
7. SMIGEL, E. O. & H. L. ROSS. 1970. Factors in the editorial decision. Am. Sociologist. **25:** 19–21.
8. INGELFINGER, F. J. 1974. Peer review in biomedical publication. Am. J. Med. **56:** 686–692.
9. SCOTT, W. A. 1974. Interreferee agreement on some characteristics of manuscripts. Am. Psychologist. **29:** 698–702.
10. CICCETTI, D. V. & H. CONN. 1976. A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. Yale J. Biol. Med. **49:** 373–383.
11. HENDRICK, C. 1976. Editorial comment. Person. Soc. Psychol. Bull. **2:** 207–208.
12. LINDER, D. E. 1977. Evaluation of the Personality and Social Psychology Bulletin by its readers and authors. Person. Soc. Psychol. Bull. **3:** 583–591.
13. GOTTFREDSON, S. D. 1978. Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgements. Am. Psychologist. **33:** 920–934.
14. SCARR, S. & B. L. R. WEBER. 1978. The reliability of reviews for the American Psychologist. Am. Psychologist. **33:** 935.
15. CICCETTI, D. V. & L. D. ERON. 1979. The reliability of manuscript reviewing for the Journal of Abnormal Psychology. J. Abnorm. Psychol. **22:** 596–600.
16. MARSH, H. W. & S. BALL. 1981. Interjudgemental reliability of reviews for the Journal of Educational Psychology. J. Ed. Psychol. **73:** 872–880.
17. EDDY, D. M., V. HASSELBLAD & R. SHACHTER. 1992. Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence: 3. Academic Press. San Diego, CA.
18. DAWSON, N. V. & H. R. ARKES. 1987. Systematic errors in medical decision making: Judgment limitations. J. Gen. Intern. Med. **2:** 183–187.
19. OXMAN, A. D. & G. H. GUYATT. 1988. Guidelines for reading literature reviews. Can. Med. Assoc. J. **138:** 697–703.
20. OXMAN, A. D., G. H. GUYATT, J. SINGER, *et al.* 1991. Agreement among reviewers of review articles. J. Clin. Epidemiol. **44:** 91–98.
21. OXMAN, A. D. & G. H. GUYATT. 1991. Validation of an index of the quality of review articles. J. Clin. Epidemiol. **44:** 1271–1278.
22. STREINER, D. L. & G. R. NORMAN. 1989. Health Measurement Scales: A Practical Guide to their Development and Use. Oxford University Press. Oxford, U.K.
23. SHROUT, P. E. & J. L. FLEISS. 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. **86:** 420–428.
24. DIXON, W. J., ED. 1983. BMDP Statistical Software. University of California Press. Berkeley, CA.

25. SACKETT, D. L. 1983. Proposals for the health sciences—I. Compulsory retirement for experts. J. Chron. Dis. **36:** 545–547.
26. CONANT, J. B. 1951. Science and Common Sense.: 109. Yale University Press. New Haven, CN.
27. CHALMERS, I. 1983. Scientific inquiry and authoritarianism in perinatal care and education. Birth **10:** 151–162
28. GJERTSEN, D. 1989. Science and Philosophy: Past and Present.: 113. Penguin Books. London.
29. ZIMAN, J. M. 1969. Information, communication, knowledge. Nature **224:** 318–324.

DISCUSSION

RICHARD PETO (*University of Oxford, Oxford, U.K.*): Dr. Oxman, even on reviewing evidence, you've got to start by saying what sort of relative risk you are talking about. These problems arise where you're likely to have moderate relative risks. You know you may get things wrong in terms of two-fold errors, but you're less likely to make 10-fold errors. For example, if you were reviewing the evidence as to whether smoking is massacreing vast numbers of people, then you wouldn't need the kind of methodologic quality that you've described, and it wouldn't necessarily be a fair criticism of review of evidence on smoking to say that the review didn't go into the kind of detail that you described. But if you were trying to study something, such as beta blockers in acute myocardial infarction, then you would need a great deal of attention to detail. In reviews, as in trials, the critical question is: What is the relative risk? That is, what sort of differences are we trying to discriminate between? You've shown very nicely that quite substantial biases will result if the reviewing process isn't done properly. But there are circumstances where reviews that are not done according to any of the criteria that you've described might nevertheless be scientifically sufficient. It is the relative risk that always matters in determining what is plausible.

ANDREW OXMAN (*McMaster University, Hamilton, Ontario, Canada*): For any research endeavor it's how important the problem is that determines how many resources we put into the effort. And I agree entirely with your comments earlier that a marker of importance is how common the problem is and how severe its outcomes are. From the point of view of a reviewer of research it is also important to take into consideration the quality of the available information. One needs to consider how thoroughly to look for research, what to include and what to exclude, and how to validate the information. All of this is going to be determined by the effect sizes one is looking at as well as the availability of the research, the amount of one's own resources, and the importance of the problem. I think that it's a mistake to say that these criteria don't apply to any type of problem—I think they do. What differs from problem to problem is the decisions that are made relative to each of these criteria about how many resources you use to identify research, how you assess the information, what effort you put into validating the data, and what analytic techniques you use to put them together at the end of the day.

PETO: For the evaluation of interventions, which is what this meeting is about, I think your material is totally relevant. It's just that there are other areas of

review that aren't so much concerned with the evaluation of interventions, such as the effects of smoking, where you don't need to randomize to work out that it causes lung cancer.

THOMAS CHALMERS (*New England Medical Center, Boston, Mass.*): I gathered from your presentation, Dr. Oxman, that you distinguished between a systematic review and a meta-analysis by pointing out that you didn't want to put the emphasis on statistics. If one assumes that a good meta-analysis includes all of the things that you've pointed out are necessary for doing a review—such as control of bias and duplicate determination, adequate searching of the literature, and a quantitative analysis of the data as presented in the literature—is there any place for a systematic review without the quantitative analysis? Is there any place for a review of data in the literature where the data have not been statistically analyzed? Should not all qualitative reviews that do not analyze the data as data be replaced by adequately done meta-analyses?

OXMAN: This relates to the point that Richard Peto just made—that there are situations in which the applications of statistical techniques aren't warranted either because the data are so obvious that you don't need them or because the data aren't sufficient for statistical analysis. A good example of this is seen by the knee replacement PORT that was discussed earlier. If the data you have are not good enough to be combined in order to draw conclusions it is worthwhile to know that so that investigators can go out then and collect good evidence. It's not worth the sort of efforts that Richard Peto's group puts into reviews, but it's worth having gone through the steps up to the point of saying that this is the best evidence we have, this is the quality of the evidence, this is what we know at this point in time, and now let's go out and try and learn some more.

DIXIE SNIDER (*Centers for Disease Control and Prevention, Atlanta, Ga.*): The systematic reviews and meta-analyses are properly emphasized as being very important. On the other hand, as someone who's recently retired as a content expert and moved into the area of methodology, I have also seen bias on the part of persons who do systematic reviews, particularly a tendency toward iconoclastic destruction of the status quo. Perhaps that's good, but I would also like to point out that many times the content experts are knowledgeable about information that's not necessarily available in the literature; having been on some data safety and monitoring boards I've experienced that on several occasions. So instead of having two different groups—the authorities and the methodologic experts—I would suggest that teams of people with an expertise in both areas "duke it out," so to speak. Such an approach could produce a better review than one coming at it from either extreme.

OXMAN: I agree entirely; in fact, that's the approach that we've been taking with the reviews with which I've been involved. Meta-analysis or systematic reviews in general are tools for putting together what we know and like any tool can be misused. It's important that these methods be used properly, and the insights that experts can bring to a review are important and should be kept in perspective.