Annals of Oncology 9: 601–605, 1998.

Review on evidence-based cancer medicine

The pitfalls of narrative reviews in clinical medicine

C. J. Williams

Co-ordinator, Cochrane Cancer Network, Institute of Health Sciences, Oxford, UK

Key words: narrative reviews, publication bias, meta-analysis

Introduction

Though the need for reviews of data in clinical medicine is not new, the pressures to use such reviews are becoming increasingly apparent. Currently, there about 23,000 biomedical journals worldwide publishing some 2,000,000 peer reviewed papers a year. One study has estimated that a general physician needs to read 19 original articles a day, for 365 days a year, to keep abreast of their field [1]. Added to the physical and mental impossibility of reading so much material, is the problem of accessing these 6935 pertinent papers among the two million papers published in many journals around the world in a variety of languages. The need for good reviews is crystal clear – the question that this paper will examine is whether the classic narrative review that we have used up till now is of high enough quality.

Reviews, of whatever type, are entirely dependent on the quality of the primary research that they are based on. Although many questions in cancer have been addressed by randomised clinical trials (RCTs), there are still many other questions where it is difficult to carry out RCTs and where there is a paucity of reliable data. At present there are no tried and tested methods for systematically reviewing the data from anecdotal or uncontrolled studies, though good new reviews of this type suggest that useful methods can be developed [2]. The biases in systematic reviews of observational data are a topic of present research, it is suspected that the biases are greater than in systematic reviews of RCTs. This article will, therefore, concentrate mainly on the quality of reviews of RCTs in cancer and other areas of medicine.

Assessing the quality of narrative reviews

It is only recently that critical attention has been paid to the quality of narrative reviews in clinical medicine. This is perplexing since the first randomised clinical trial was carried out 50 years ago. The purpose of randomisation is simply to avoid bias in selection of patients in a trial, large numbers of patients reaching the end point of interest being required to avoid random errors (Table 1). During the last half century it is, therefore, suprising that we have allowed narrative reviews of RCTs to be written where there has been *no* systematic attempt to avoid bias. There has been a dual standard in the way that we have used data in RCTs and reviews. Cynthia Mulrow was one of the first to show that there was a real problem with narrative reviews in her analysis of 50 reviews in general medical journals which was published in 1987 [3].

Mulrow developed a simple set of criteria to test whether a review might be potentially biased. Deeks has boiled these down further in his comments that reviews should be *rigorous, informative, comprehensive and explicit* [4]. C. Mulrow's criteria fit these requirements, she asked whether, on reading a review carefully, it was clear that the following criteria were specified, unclear or not specified.

Specified purpose. A clearly specified purpose(s) is important as a frame of reference for the reader and to help determine strategies to select information and decide on the method of assessment

Data identification. In recent years it has become clear that publication bias is an all too real phenomena (Figure 1). This means, essentially, that the easier it is too find an RCT, the more likely it is to be 'positive' – showing the 'new' treatment to be more active than the

Table 1. Requirements to assess moderate effects reliably.

Avoidance of moderate biases:
Properly concealed randomisation.
Intention to treat analysis.
Avoid data-dependent emphasis on subgroups.
Systematic review of *all* relevant RCTs.
Avoidance of moderate random errors:
Large numbers in new trials (many endpoints).
Systematic review of *all* relevant RCTs to give the largest number of participants.

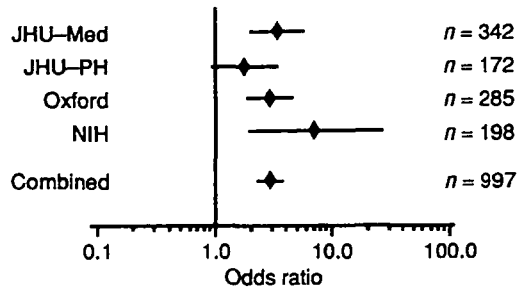

Figure 1. Meta-analysis of four studies (JHU-Med; JHU-PH; Oxford; NIH) examining association between significant results and publication: unadjusted odds ratios and 95% confidence intervals [5].

control. Added to the difficulty of identifying all pertinent published RCTs is the problem of unpublished RCTs.

Four large studies have investigated the association between the outcome of trials and their publication record. These studies looked at a body of known RCTs and the association between trials with 'significant results' and their record of subsequent publication in a peer review journal. When combined these studies included 997 RCTs in a variety of fields in medicine [5]. The combined odds ratio was 2.88 (95% confidence intervals: 2.13–3.89) favouring 'significant' trials being published. Even when there was a very high rate of publication in a group (the NIH published 188 of 198 trials) there was still strong evidence of a bias not to publish trials that were *not* 'significant'.

It is fascinating to see that the sorts of odds ratios reported in these studies show much more marked degrees of difference than those usually seen in individual trials testing different anti-cancer therapies. Identification of unpublished trials is important since they are common (the Early Breast Cancer Trialists Group reported that 22% of the studies relevant to their meta-analysis were unpublished [6]) and are more likely not to favour the 'new treatment'. It is also clear that many RCTs reported as meeting abstracts are never published subsequently in peer review journals. In one study, 1193 of 2391 (49%) of abstracts were never published as a full paper [7]. Once again there is a bias towards not further publishing 'negative' trials.

Publication bias is more of a problem in narrative than systematic reviews. In some narrative reviews unpublished data, personal series and trials are selectively included in a biased fashion. The main issues in this area are *completeness* of the data and *lack of bias*. The latter, being guaranteed by the former.

Reliance on MEDLINE and other databases to identify published RCTs is also fraught with problems. For instance, MEDLINE only includes about 4000 journals and is weighted towards English language publications. Even when journals are indexed in MEDLINE, studies have shown that only a modest proportion of RCTs found by hand searching of the journals are identified by a MEDLINE search [8].

Data selection. One of the main purposes of the protocol written for an RCT is to spell out which patients will be included and which excluded from the trial. There is an obvious need to state which studies will be included in a review, otherwise there is a real risk that exclusion of data from some RCTs will result in bias. In fact the risk of producing an unreliable conclusion is much greater in a review – in an RCT the patients, regardless of how they are selected, are randomised. In a review the factors to be taken into account should be made clear in a protocol before the review begins – inclusion/exclusion criteria, end points and methodology should all be specified before the review is undertaken.

Validity assessment. Reports of RCTs are complex and there is a need to ensure that the data extracted is accurate and of high quality. This usually requires the co-operation of two or more individuals assessing the paper independently. They can cross check the data extracted and where subjective quality criteria are applied can ensure that this is done in an unbiased a fashion as possible. There is a need to check that good RCTs are included in the review. Often this is more a matter of whether good clinical trial practice has been adhered to in the RCTs, rather than how they were done in detail. Narrative reviews, in contrast, often use post-publication letters to discredit work, rather than setting quality criteria before the review is carried out. Such retrospective weeding out of trials on quality grounds, is likely to result in bias.

Data synthesis. Most reviews attempt some degree of qualitative integration by mentioning limitations and inconsistencies in existing data. In cancer reviews the presentation of response and survival data is often in tabular form, the data being synthesised in the discussion. It is however possible to carryout quantitative synthesis of data from a number of RCTs using the methods of meta-analysis. Pooling the results of RCTs is less likely to be biased than qualitative synthesis where the reviewer is free to pick and choose which bits of data to emphasise.

Summaries and future directions. Summaries are helpful in providing a compressed version of the review in an easily manageable form. However, there is a danger that conclusions may not be supported by a valid review process – caveats and confounding factors only being included in the text of the full paper. Readers who do not go beyond the summary being misled. One of the purposes of reviews of RCTs is to inform the research process and discussion of future directions is important.

The acid test of a good review is that it should be entirely transparent. The reader should be able to see where the data came from and how the reviewer handled it and reached their conclusions.

Table 2. Quality assessment according to Mulrow criteria [10].

Criteria	Reviews (no. 106)					
	Specified		Unclear		Not specified	
	No.	%	No.	%	No.	%
Purpose	105	99	1	1	0	0
Data identification	12	11	11	10	83	79
Data Selection	11	10	21	20	74	70
Validity assessment	9	8	15	14	82	78
Qualitative synthesis	106	100	0	0	0	0
Quantitative synthesis	1	1	1	1	104	98
Summary	101	95	4	4	1	1
Future directions	81	76	15	14	10	10

How well have Mulrow's criteria been applied in cancer reviews?

In a paper using Mulrow's criteria to assess the quality of cancer reviews, the short answer is – not very well [9]. In this paper, 106 narrative reviews in the Journal of Clinical Oncology were assessed between its inception in January 1983 through to December 1995. Although some items were specified in the great majority of reviews, others were only specified in a minority of the reviews (Table 2). Thus, authors rarely gave information on methods of data identification (11.3%), data selection (10.4%) and assessment of validity (8.4%). Quantitative synthesis of the data was also rarely undertaken (1%). Conversely, authors nearly always specified the purpose of the review, carried out qualitative synthesis and gave a summary and discussed future directions. These results are worrying since the items rarely covered satisfactorily are those specifically designed to avoid bias. Even worse results were found in a study of reviews in the primary chemotherapy of ovarian carcinoma [10, 11]. This study, in addition, found that only four of 49 reviews gave any raw data (table or figure) from original RCTs. The remainder included little or no data from the RCTs they were reviewing. Most reviews referred to very few of the available RCTs (mean number of references to RCTs 5.7, CI 0–30 – when more than 70 RCTs have been published) and were thus selective in their use of the literature. The literature was also potentially skewed by multiple publication of reviews. This is a major problem with RCTs where it is sometimes difficult to know whether several publications refer to different trials or just one. Of the 49 reviews of chemotherapy in ovarian cancer found in a MEDLINE search, one author wrote 10. This study also found that most reviews were not focussed, asking multiple questions rather than tackling one or two clearly defined topics.

Does it matter that reviews are not systematic?

If current narrative reviews are potentially biased there must be a major risk that they lead to the wrong conclusions. This may result in false positive and false negative conclusions or may delay the eventual discovery of the correct conclusion. Antman's key paper comparing the results of meta-analyses of RCTs with recommendations from clinical experts shows how damaging narrative reviews can be [12]. They meta-analysed the results of 182 RCTs of therapies for acute MI and examined the recommendations of 43 review articles and 100 textbook chapters. An important example from this study was that thrombolytic drugs did not begin to be recommended, even for specific indications, by more than half the experts until 13 years after they *could* have been shown to be effective (Figure 2). Six years elapsed between the time that the first meta-analysis was published showing an impressive benefit for thrombolytic therapy and the time when the majority of reviewers recommended it for specific or general use. Conversely, the majority of review authors recommended lidocaine for prophylaxis against ventricular fibrillation throughout the 25 years of the study, yet there is no evidence of a mortality reduction in controlled trials (Figure 3). Current thinking is that this therapy may not be effective and may actually be harmful.

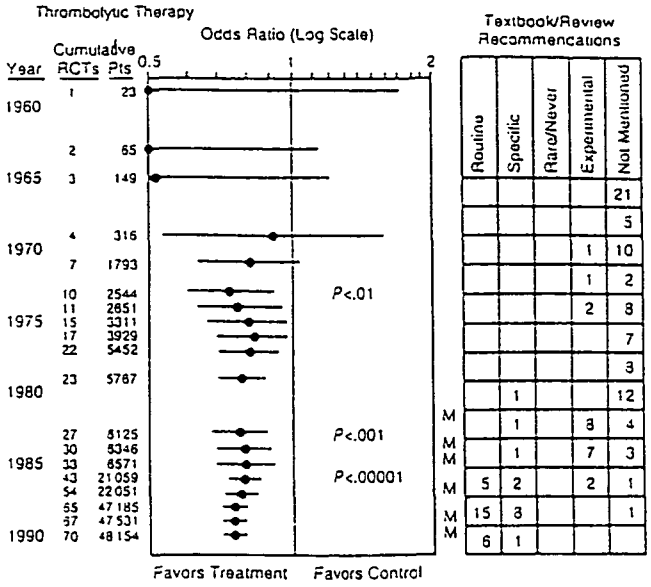

Figure 2. Cumulative meta-analysis of data from RCTs testing the effectiveness of thrombolytic therapy after MI compared with recommendations of experts [12].

In the field of cancer, the finding by the Early Breast Cancer Trialists' Collaborative Group that oophorectomy prolongs survival in premenopausal women came as a surprise, since most clinicians had long since concluded that this was an ineffective treatment [13]. Failure to systematically review the literature not only leads to wrong conclusions, it also leads to inappropriate research. There are a number of instances where research has continued when the question had already been answered. For example, a large series of RCTs on the use of prophylactic antibiotics after caesarian section were carried out in the two decades after 1970 [14]. Throughout most of this period there was sufficient

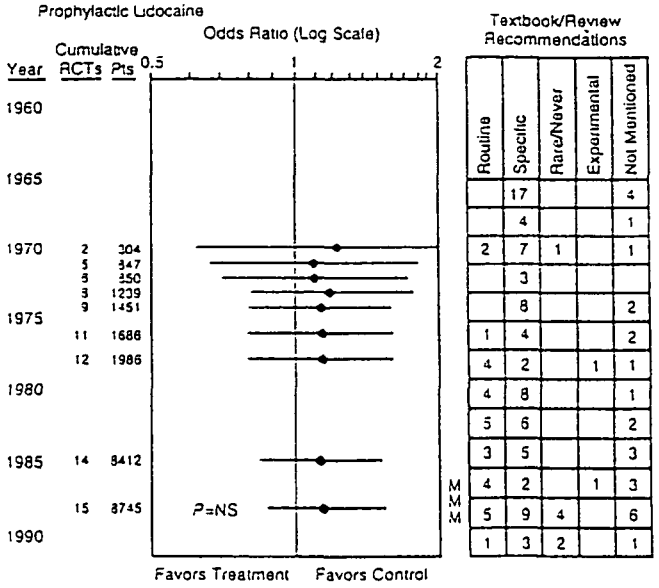

Figure 3. Cumulative meta-analysis of data from RCTs testing the effectiveness of lidocaine therapy after MI compared with recommendations of experts [12].

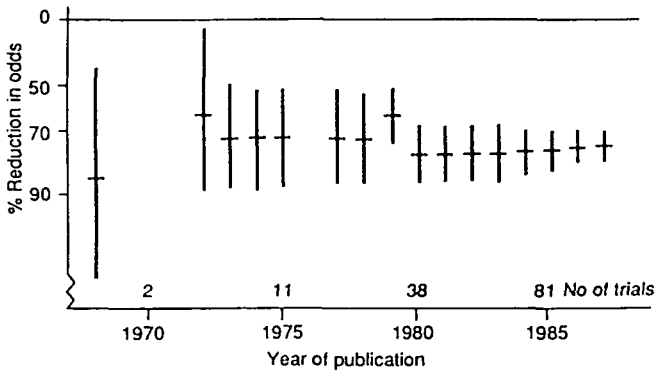

Figure 4. Cumulative estimates of the extent to which prophylactic antibiotics reduce the odds of serious postoperative infection after caesarian section [14].

evidence to show that severe infections were significantly less common if antibiotics were used (Figure 4). Despite this, new RCTs testing this question are still being conducted in the 1990s.

Where do we go from here?

The current state of affairs is patently unsatisfactory, but this does not mean that all reviews should be meta-analyses. However, there is a need for many of the questions currently being addressed in narrative reviews to now be answered in systematic reviews that quantitatively synthesis the data. Thus, a question such as does chemotherapy improve survival when used as primary therapy in ovarian cancer, should in general be approached by a systematic review rather than a narrative review. Where there are few or no RCTs, meta-analytic techniques are not appropriate, but reviewers should in general take a more systematic approach to the preparation of narrative-like reviews.

Where space does not allow the author to take a focussed approach to a single question, such as in a textbook, an alternative is to review evidence from systematic reviews or one or more large RCTs, rather than a few selected small individual trials.

Reviews are often written to stress a particular viewpoint or for 'political' reasons. It is unlikely that all such reviews will be systematic, but readers and preferably authors should be made aware of the existence of data from large RCTs and systematic reviews.

Conclusions

Narrative reviews are potentially biased and unreliable. It is disappointing that after more than a decade of exhortation to improve the quality of reviews, journals and their readers have continued to accept potentially biased and misleading reviews as acceptable [15]. Where possible, narrative reviews should be replaced by reviews using systematic methods. Such reviews do not constitute a gold standard (that should be data from one or more very large RCTs that give clear answers), but they should be more reliable and transparent than a narrative review. The quality of systematic reviews is variable and such reviews can still be biased. Readers should become more aware of systematic review methods and the need to learn appraisal skills in general. The development of quality checklists for RCTs [16] is, not suprisingly, mirrored by the development of similar checklists for reviews [17]. Trialists/reviewers, as well as readers, should use these.

References

1. Evidence-based medicine notebook. On the need for evidence-based medicine. EBM 1995; 1: 5–6.
2. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal replacement therapy: Collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Lancet 1997; 350: 1047–59.
3. Mulrow CD. The medical review article· State of the science. Ann Intern Med 1987; 106: 485–8.
4. Deeks J. Do antibiotics help children with acute otitis media? How to use an overview. In Risdale M (ed): Evidence-Based General Practice. 2nd edition. Edinburgh. Churchill-Livingstone 1998 (in press).
5. Dickersin K, Min Y-I. NIH clinical trials and publication bias. Online J Curr Clin Trials (serial online) 1993: Apr 28 (Doc No. 50).
6. Early Breast Cancer Trialists' Collaborative Group. Treatment of early breast cancer. Volume 1. Worldwide evidence 1985–1990. Oxford: Oxford University Press 1990.
7. Scherer RW, Dickersin K, Langenberg P. Full publication of results initially presented in abstracts. A meta-analysis. JAMA 1994; 272: 158–62.
8. Adams CE, Power A, Frederick K, Lefebrve C. An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. Psychol Med 1994; 24: 741-8.
9. Bramwell VHC, Williams CJ. Do authors of review article use

systematic methods to identify, assess and synthesize information? Ann Oncol 1997; 8: 1185–96.
10. Williams C, Lodge M, Bramwell V. Cancer reviews need to be systematic if they are to be reliable. Proc ECCO 1997; 9.
11. Williams C, Lodge M. Poor quality of reviews of randomised clinical trials of chemotherapy for advanced ovarian cancer: Need for action. Proc Am Soc Clin Oncol 1997; 16: Abstr 1534.
12. Antman EM, Lau J, Kupelnick B et al. A comparison of results of meta-analysis of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. JAMA 1992; 268: 240–8.
13. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Lancet 1992; 339: 1–15 and 71–85.
14. Chalmers I. Improving the quality and dissemination of reviews of clinical research. In Lock S (ed): The Future of Medical Journals. London: BMJ 1991; 127–46.
15. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. CMAJ 1988: 138: 697–703.
16. Begg C, Cho M, Eastwood S et al. Improving the quality of reporting randomized controlled trials. The CONSORT statement. JAMA 1996; 276: 637–9.
17. Oxman AD. Checklists for review articles. BMJ 1994; 309: 648–51.

Received 3 April 1998; accepted 7 April 1998.

Correspondence to:
C. J. Williams, MD
Cochrane Cancer Network
Institute of Health Sciences
P.O. Box 777
Oxford, OX3 7LF
UK
E-mail: cwilliams@canet org