International Journal of Technology Assessment in Health Care, 12:2 (1996), 195–208.

ASSESSING THE QUALITY OF RANDOMIZED CONTROLLED TRIALS

Current Issues and Future Directions

David Moher
University of Ottawa

Alejandro R. Jadad
McMaster University

Peter Tugwell
University of Ottawa

Abstract

Assessing the quality of randomized controlled trials is a relatively new and important development. Three approaches have been developed: component, checklist, and scale assessment. Component approaches evaluate selected aspects of trials, such as masking. Checklists and scales involve lists of items thought to be integral to study quality. Scales, unlike the other methods, provide a summary numeric score of quality, which can be formally incorporated into a systematic review. Most scales to date have not been developed with sufficient rigor, however. Empirical evidence indicates that differences in scale development can lead to important differences in quality assessment. Several methods for including quality scores in systematic reviews have been proposed, but since little empirical evidence supports any given method, results must be interpreted cautiously. Future efforts may be best focused on gathering more empirical evidence to identify trial characteristics directly related to bias in the estimates of intervention effects and on improving the way in which trials are reported.

Since the randomized controlled trial (RCT) was introduced in its modern form approximately four decades ago (52), it has gained prominence with health care researchers because of its potential to control for bias. However, poorly conducted trials may yield misleading results. It is therefore important for all involved in health care to be bale to assess the reliability and validity of the research evidence.

Quality is a construct (concept) that can be defined in many ways, including the literary aspects of the report of a trial or its external validity, i.e., the degree to which it is possible to generalize trial results. Our focus is on one important aspect of methodologic quality (hereafter simply "quality"), internal validity, which we define as "the confidence that the trial design, conduct, analysis, and presentation

This work has benefited from the active support of the Cochrane Collaboration SORT (Standards of Reporting Trials) group. We also thank Drs. Graham Nichol and Sharon Walsh, and Marie Penman for their collaboration in assessing the quality of the 12 antithrombotic trials reported in this paper.

has minimized or avoided biases in its intervention comparisons." However, we recognize that this definition excludes other methodologic aspects of quality, for example, those concerned with the precision and reliability of measurements or estimation of compliance; we have not considered these here. In most instances, however, the only way to assess the quality of a trial is to rely on the information contained in the report. Therefore, a trial with a biased design that is well reported could be judged as having high quality, and a well-designed but poorly reported trial could be judged as having low quality.

Investigators assess quality because they wish to estimate the effects of bias on the results of a trial. Differences of quality across trials may indicate that the results of some are more biased than those of others. Systematic reviewers need to take this into account.

In this review we discuss some of the most important issues in assessing the quality of RCTs, focusing on the methods available for quality assessment, empirical evidence supporting such methods, and the future directions for research in this area.

METHODS TO ASSESS QUALITY

Assessment of Quality Using Individual Components

One approach to assessing quality has focused on "component" issues in trial reports. Components include randomization, masking (blinding), and sample size. In a review of 45 trials in three leading general medical journals published in 1985, Pocock and colleagues (42) reported that sample size was mentioned in only 11% of the reports, only 13% gave confidence intervals, and the use of statistical analysis tended to exaggerate treatment efficacy.

To avoid selection bias in assigning patients to intervention, concealment of assignment is essential and should be feasible in all trials. Chalmers and colleagues (13) reviewed 145 reports of RCTs in the treatment of acute myocardial infarction to assess whether concealment of patient assignment affected trail results. Their results indicated that trials that reported concealed assignment compared with unconcealed assignment had smaller treatment effects as defined by case-fatality rates. Schulz and colleagues (47) reviewed 250 reports of RCTs and found the odds ratios in the unclearly concealed trials were on average 30% (95% CI: 21%, 38%) lower than in the adequately concealed trials, i.e., estimating the intervention to be more effective than it really was.

Colditz and colleageus (16) have reported similar results concerning the level of masking. In a review of 113 reports of clinical trails, these authors noted that trials that reported a higher level of masking tended to show smaller treatment effects compared with those trials that used lower levels of masking (e.g., double blind versus single blind), i.e., the lower the level of masking, the greater the increase in treatment effectiveness.

These studies have provided important information on the quality of reporting of individual items and have highlighted how inadequate reporting should lead readers to be skeptical about the validity of trials results. Unfortunately, assessing one component of a trial report may provide only minimal information about its overall quality.

Checklists and Scales

Checklists provide a qualitative estimate of the overall quality of a trial using itemized criteria for comparing the trials. The development of checklists is a logical extension

of component assessment of quality. As such, checklist items do not have numerical scores attached to them.

Scales are similar to checklists except that each item of a scale is scored numerically and an overall quality score is generated.

Assessment of Quality with Checklists. In a recent systematic review of the literature, we identified nine checklists and 25 scales (38). The nine checklists identified varied from four to 57 items (Table 1). Four of the checklists were designed to assess the methodologic quality of the trials (4;6;54;55), three to assess the quality of reporting (17;25;35), and the remaining two to assess both methodologic quality and the quality of reporting (22;34). The selection of items for all the checklists used "accepted criteria," defined as items noted by textbooks on clinical trials to be important in defining the quality of clinical trials (37;41). Seven of the checklists included at least one item about patient assignment (6;17;22;25;35;54;55); eight had at least one item about masking (6;17;22;25;34;35;54;55); five had at least one item about patient follow-up (6;17;22;25;54); and eight had at least one item about statistical analysis (4;17;22;25;34;35;54;55). We have determined that quality assessment of a trial report using any of the checklists takes 30 minutes or less.

Assessment of Quality Using Scales. Scales have been developed increasingly because many systematic reviewers want to include a measure of study quality of individual trials in their reviews. Of the 25 scales identified (Table 2) 15 were designed to assess the quality of any trial (10;12;15;16;18;19;23;28;29;31;44;45;49). The remaining 10 scales assess the quality of trials in specific subject areas (e.g., contrast media, pain). Six of the scales defined "quality" as used in their scale development (5;10;15;23;29). Three of the scales were designed to assess the quality of the trial report (2;23;29), eight to assess methodologic quality of the study (5;7;12;28; 32;33;48), and the remaining 14 to assess both. The number of items in a scale ranged from 3 to 34.

Twenty-four of the 25 scale used "accepted criteria" to select the items for inclusion (2;5;7;10;12;15;16;18;19;23;24;28;31;32;33;39;40;44;45;48;49;53). The remaining scale (29) used a large number of items that were narrowed down to the final version of the scale, using standard scale development techniques. Twenty-two of the scales had at least one item on how patients were assigned to treatment (2; 5;7;10;12;15;16;18;19;23;24;28;29;31;32;33;39;45;48;49;53); 20 had at least one item on masking (2;5;12;15;18;19;23;24;29;31;32;33;40;44;45;49;53); 11 had at least one item on patient follow-up (5;12;16;32;33;39;44;45;48;63); and 21 had at least one item on statistical analysis (2;5;12;15;16;18;19;23;24;29;31;32;33;40;45;49;53).

Twelve of the scales had been tested for inter-rater reliability (2;7;15;25;28;29; 31;32;44;45;53); five of these reported percent agreement, six reported intra-class correlation or kappa, and one reported Pearson correlations.

The scoring method varies considerably among scales, as does the possible range of scores. Seventeen of the scales provided detailed instructions on how scores should be assigned to each item as well as how to compute the overall summary score (2; 7;10;12;15;23;29;31;32;33;39;40;44;45;53). Total scores for each of the scales ranged from 1 to 170 points. Eight of the scales used a gradation of scores within each item (e.g., score of 1 to 3).

Four of the scales recommended scoring procedures designed to minimize bias (12;15;29;32). For example, the quality assessor should not know the identify of the trial's author(s), journal, and outcome. Only one scale, constructed by one of us (29), described how items were initially selected, how and why the final items were

Table 1. Descriptive Characteristics of Published Checklists Used to Assess the Quality of Randomized Controlled Trials

Checklist name[a]	No. of items	Quality defined[b]	Type of quality assessed[c]	Items selected[d]	Patient assignment[e]	Masking[f]	Patient follow-up[g]	Statistical analysis[h]
Badgley (4)	5	p	m	ac	n	n	n	y
Bland (6)	18	p	m	ac	y	y	y	n
DerSimonian (17)	11	n	r	ac	y	y	y	y
Gardner (22)	26	n	m&r	ac	y	y	y	y
Grant (25)	28	n	r	ac	y	y	y	y
Lionel (34)	45	n	m&r	ac	n	y	n	y
Mahon (35)	4	n	r	ac	y	y	n	y
Thomson (54)	11	n	m	ac	y	y	y	y
Weintraub (55)	57	n	m	ac	y	y	n	y

[a] name of principal author.
[b] n = no; y = yes; p = partially defined.
[c] m = Methodologic quality; r = quality of report.
[d] ac = accepted criteria (see text for details).
[e] Was there an item on patient assignment? n = no; y = yes.
[f] Was there an item on masking? n = no; y = yes.
[g] Was there an item on patient follow-up? n = no; y = yes.
[h] Was there an item on statistical analysis? n = no; y = yes.

included, how the scale discriminated between trials of differing quality, and what range of quality scores was obtained during its development. In 11 of the 25 scales (2;5;7;10;16;18;28;29;39;44;48), all items can be scored in 10 minutes or less (range, 5 to 45 minutes).

Thus, scales vary in size, complexity, and level of development. It might be useful to know whether different scales yield similar results when applied to the same trial. Such information could guide quality assessors in their choice of scale. For example, there would be little advantage in using a 15-item scale to assess quality if similar results could be obtained by using a three-item scale.

Assessment of Quality Across Scales

Additional research suggests that different scales are bound to generate discrepant results. A study was undertaken at the Clinical Epidemiology Unit, Loeb Medical Research Institute, to establish whether different scales gave different quantitative and qualitative assessments of the quality of RCTs. The members of the research team first trained themselves in assessing quality, using six published scales (2;7;12;18;24;45). Each member independently assessed the quality of 12 of 15 trials (the remaining three trials were either available only in a technical report or not published in English) included in a systematic review of antithrombotic therapy in acute ischemic stroke (46) using at least two scales. After scoring was completed, the results were reviewed and differences were resolved through consensus and arbitration. The results (Table 3) show that overall quality scores for each trial varied greatly across scales, ranging from 23% to 74% of the maximum possible value. Similarly discrepant results were obtained using rank scores of individual trials (Table 3). These results suggest that different trials might be included or excluded from a systematic review, depending on the scale used to assess quality and the methods of including quality scores into the review.

Powe and colleagues (43) assessed the quality of 100 contrast media trials published between 1982 and 1987, using a scale developed by Chalmers and coworkers (12), and reported a mean quality score of 39% (SD = 12). Andrew (2) assessed the quality of 49 contrast media trials published during the 1980s in five leading radiology journals using his own customized scale. He reported a mean quality score of 70% (SD = 14.6) (3). Although some of the trials reviewed by both groups differed, it is unlikely this is the explanation for the wide variation in their assessed quality. In contrast to these results, Detsky and colleagues (18) assessed the quality of 18 trials used in a parenteral nutritional systematic review using two of the scales (12;18) included in the study described above. They reported only minor differences in raw scores of quality, and rankings of quality remained similar across trials.

INCORPORATING QUALITY SCORES INTO SYSTEMATIC REVIEWS

Four ways (11;21;30) of incorporating quality scores into a systematic review have been suggested: (a) only trials meeting a threshold score; (b) using the quality score as a weight in estimating the effect sized; (c) performing cumulative systematic reviews using quality scores as the input sequence; and (d) visualizing the effect of scores using plots. There is little evidence supporting the validity or any of these methods. The use of the threshold approach, perhaps the most commonly recommended method, may profoundly affect the number of trials included in the analysis of a systematic review. This approach has been used as a decision aid for the inclusion of trials in the antithrombotic systematic review previously discussed (46). If the

Table 2. Descriptive Characteristics of Published and Unpublished Scales Used to Assess the Quality of Randomized Controlled Trials

Scale name[a]	Type of scale[b]	Quality de-fined[c]	Type of quality as-sessed[d]	Items se-lected[e]	Patient assign-ment[f]	Mask-ing[g]	Patient follow-up[h]	Statis-tical analy-sis[i]	Number of items	Scale de-velop-ment[j]	Inter-rater reliabil-ity[k]	Time to com-plete[l]	Scoring range[m]	Detailed instruc-tions for scoring items[n]
Andrew (2)	s	n	r	ac	y	y	n	y	11	nr	0.95[q]	10	0–22	y
Goodman (23)	g	y	r	ac	y	y	n	y	34	nr	0.12[r]	15	34–170	y
Becker-man (5)	s	y	m	ac	y	y	y	y	25	nr	nr	10	0–25	n
Brown (7)	s	n	m	ac	y	n	n	n	6	nr	0.89[q]	10	0–21	y
Chalmers, I (10)	g	y	m&r	ac	y	y	n	y	3	nr	nr	<10	0–9	y
Chalmers, TC (12)	g	n	m	ac	y	y	y	y	27	nr	nr	45	0–100	y
Cho (15)	g	y	m&r	ac	y	y	n	y	24	nr	0.89[r]	30	0–1	y
Colditz (16)	g	n	m&r	ac	y	n	y	y	8	nr	nr	10	0–8	n
Detsky (18)	g	n	m&r	ac	y	y	n	y	5	nr	nr	10	0–15	n
Evans (19)	g	n	m&r	ac	y	y	n	y	33	nr	nr	15	0–100	n
Gøtzsche (24)	s	n	m&r	ac	y	y	n	y	8.8	nr	nr	15	0–8	n
Imperiale (28)	g	n	m	ac	y	n	n	n	5	nr	.79[r]	<10	0–5	n
Jadad (29)	g	y	r	pool	y	y	n	n	3	y	.66, 77[r]	<10	0–5	6
Jonas[o]	g	n	m	ac	y	y	y	y	20	nr	.6[r]	20	0–100	y
Kleijnen (31)	g	p	m&r	ac	y	y	n	y	7	nr	0.87[q]	15	0–100	y
Koes (32)	s	p	m	ac	y	y	y	y	17	nr	0.8[q]	15	0–100	y

Linde[p]	g	y	m&r	ac	y	y	y	y	24	mr	nr	30	0-100	y
Nurmo-hamed (39)	s	p	m&r	ac	y	y	n	n	8	nr	nr	10	0-8	y
Onghenia (40)	s	n	m&r	ac	n	n	n	y	10	nr	nr	15	0-10	y
Poynard (44)	g	p	m&r	ac	n	y	y	y	14	nr	>.66[r]	10	−2-26	y
Reisch (45)	g	n	m&r	ac	y	y	y	y	34	nr	0.99, 0.7[s]	30	0-34	y
Smith (48)	s	n	m	ac	y	n	y	n	8	nr	nr	10	0-40	n
Spitzer (49)	s	n	m	ac	y	y	n	y	5	nr	nr	25	0-5	n
Levine (33)	g	n	m&r	ac	y	y	y	y	29	nr	nr	30	0-100	y
Ter Riet (53)	s	p	m&r	ac	y	y	y	y	18	nr	0.93[q]	15	1-100	y

[a] Name of scale or principal author.
[b] g = Generic scale; s = specific scale (e.g., contrast media, pain).
[c] n = No; y = yes; p = partially defined.
[d] m = Methodological quality; r = quality of report.
[e] ac = Accepted criteria (see text for details); pool = pool of items (see text for details).
[f] Was there an item on patient assignment? n = no; y = yes.
[g] Was there an item on masking? n = no; y = yes.
[h] Was there an item on patient follow-up? n = no; y = yes.
[i] Was there an item on statistical analysis? n = no; y = yes.
[j] Was the scale rigorously developed? (see text for details); nr = Not reported; y = yes.
[k] nr = Not reported.
[l] Approximate time (in minutes) to complete scoring a trial; if it was not stated by the authors, we estimated the time by scoring trials.
[m] The range of potential scores using the scale; higher scores indicate superior quality.
[n] n = No; y = yes.
[o] Jonas, W.B. The likelihood of validity evaluation method. Unpublished manuscript, 1993.
[p] Linde K., Clausius N., Melchart D., Brandmaier R., Jonas W.B., & Eitel F. Controlled clinical trials on the efficacy of treatment strategies using homeopathic preparations: A systematic review. Unpublished manuscript, 1993.
[q] Percent agreement.
[r] Intraclass correlation or kappa.
[s] Pearson correlation.

Table 3. Quality Scores (and Ranks) of Each of 12 Randomized Controlled Trials (RCTs), Included in a Systematic Review of Antithrombotic Therapy in Acute Ischemic Stroke, Across Six Published Quality Assessment Scales

RCT no.	Andrew (2)	Brown (7)	Chalmers (12)	Detsky (18)	Gøtzsche[a] (24)	Reisch (45)	% Range (difference)	Rank range
1	72 (5)	62 (11)	30 (9)	61 (8)	43, 57 (4, 7)	61 (10	30–72 (42)	4–11
2	89 (3)	86 (1)	80 (1)	96 (1)	71, 71 (2, 3)	94 (1)	71–96 (25)	1–3
3	89 (3)	86 (1)	47 (7)	73 (3)	86, 71 (1, 3)	91 (3)	47–91 (44)	1–7
4	72 (5)	71 (10)	28 (11)	68 (6)	29, 86 (6, 1)	52 (12)	29–72 (43)	1–12
5	56 (9)	76 (6)	60 (5)	60 (9)	29, 14 (6, 12)	67 (9)	14–76 (62)	5–9
6	94 (1)	86 (1)	71 (3)	71 (4)	57, 71 (3, 3)	79 (5)	57–94 (37)	1–5
7	72 (5)	86 (1)	65 (4)	71 (4)	14, 43 (12, 11)	79 (5)	14–86 (72)	1–12
8	94 (1)	81 (5)	74 (2)	77 (2)	43, 71 (4, 3)	94 (1)	71–94 (23)	1–5
9	50 (10)	76 (6)	38 (8)	53 (11)	14, 57 (10, 7)	88 (4)	14–88 (74)	6–11
10	50 (10)	76 (6)	25 (12)	57 (10)	29, 57 (6, 7)	73 (8)	29–76 (47)	6–12
11	39 (12)	57 (12)	28 (10)	53 (11)	14, 57 (10, 7)	56 (11)	14–57 (43)	7–12
12	72 (5)	76 (6)	55 (6)	64 (7)	29, 86 (6, 1)	79 (5)	29–79 (50)	1–7

[a] The scale developed by Gøtzsche has two parts, methods and analysis, which are scored separately.

mean quality score is used as the threshold score, approximately 50% of the trials, depending on the scale used to assess quality, are not included in the analysis (Table 4). This proportion increases to about 75% if the mean quality score plus one standard deviation is used. When the median quality score is used, approximately 40% of the trials are not included in the analysis. These results pose serious problems for systematic reviewers; if quality scores influence the number of trials included in the quantitative analysis part of systematic review, they can easily affect the quantitative result.

There is evidence that trial quality can affect the results of systematic reviews. Nurmohamed and colleagues (39) reviewed trials comparing low-molecular-weight heparin (LMWH) with standard heparin in proximal deep-vein thrombosis (DVT). They reported a statistically significant beneficial effect of LMWH in reducing DVT when all trials were used in the analysis. When the analysis was limited to those trials having "strong" methodologic quality, both treatments appeared to be less effective in preventing DVT and the difference between them was not statistically significant.

Results similar to these, but in the opposite direction, have also been reported. In a systematic review (8) of education programs for patients with diabetes, no significant beneficial effect of the programs was found when all trials were included in the analysis. When only reports of "good" methodologic quality were analyzed, a significant benefit was observed.

Because of the potential impact of limiting an analysis to trials of minimum quality, the methods used to assess quality should be described in detail and analyses examining the effect of the quality score should be performed whenever possible.

MASKING THE ASSESSMENTS OF QUALITY

It was suggested more than 10 years ago that the quality of clinical reports should be assessed under masked conditions (13), that is, without the knowledge of the authors, institutions, or study results. Empirical evidence to support this recommendation has recently been published (29). Two groups of judges allocated randomly to conduct the assessments under masked or unmasked conditions assigned scores to the same set of articles and found that masked assessments of the reports produced significantly lower and more consistent scores than open assessments. This work is being extended by Berlin and colleagues (J. Berlin, personal communication, 1994), who are assessing whether the results of masked assessments of trial quality affect the overall results of systematic reviews. Although research on masked assessment of trial quality is still ongoing, our preliminary results lead us to suggest that trial quality should be assessed under masked conditions in any context in which quality judgments play a role in decision making.

FUTURE DIRECTIONS AND RECOMMENDATIONS

Over the last few years the number of published systematic reviews has increased remarkably (14). This is likely to continue with the development of international collaborative efforts such as the Cochrane Collaboration (9). Even though the assessment of the validity of the primary studies being reviewed is regarded as one of the key components of a systematic review, it is still unclear how it can be achieved reliably. Further research should concentrate primarily on the generation of empirical evidence to identify trial characteristics directly related to bias and on studies of how to improve the quality of trial reporting.

Table 4. Effect of Threshold Scores on Number of Trials Included in a Systematic Review of Antithrombotic Therapy in Acute Ischemic Stroke

Scale name	Mean (SD) quality scores of all (n = 12) RCTs included in the analysis	Median quality scores of all (n = 12) RCTs included in analysis	Number of RCTs remaining when mean is used as threshold score	Number of RCTs remaining when mean + ISD is used as threshold score	Number of RCTs remaining when median is used as threshold score
Andrew	70.8	72	8	2	8
Brown	76.6	76	5	0	9
Chalmers	50.1	51	6	3	6
Detsky	67.0	66	6	1	6
Gøtzsche[a]	38.2, 61.8	29, 71	5, 6	2, 2	9, 6
Reisch	76.1	79	7	3	7

[a] The scale developed by Gøzsche has two parts, methods and analysis, which are scored separately.

Table 5. Guidelines for Developing a Scale to Assess Trial Quality

1. Define the construct "quality."
2. Define the scope of the scale.
3. Define the population of end-users.
4. Select the targets.
5. Select the raters.
6. Score trials.
7. Measure the discriminatory power of items.
8. Measure interobserver reliability.
9. Propose a pilot version of the scale.
10. Encourage other researchers to replicate your findings.

GENERATION OF EMPIRICAL EVIDENCE

Investigations have shown that certain characteristics, such as the level of concealment in patient assignment and masking, may have an important influence on estimation of an intervention's effect. Investigations will need to continue on other markers of quality, such as the adequacy of patient follow-up.

One issue that has not been studied sufficiently, yet is often a criterion for exclusion from a review, is publication in a language other than English. Grégoire and colleageus (26) reviewed 36 systematic reviews published between January 1, 1991, and April 1, 1993, in eight leading general internal medicine journals to examine whether exclusion of non-English language trials affected the results of systematic reviews. Seventy-eight percent of the 36 reviews had language restrictions. The results of one review would have been significantly different (from no change in mortality to a significant decrease in mortality) had reports in all languages been included in the analysis. Clearly, a systematic review should consider all relevant trials regardless of language. Further work is needed to determine whether trials published in non-English journals differ in quality from those published in English-language journals. Given the potential influence of this practice on the effect estimate, authors should report justification for excluding non-English trials and include their citations.

More empirical evidence is required to establish whether component assessments and scale assessments result in different outcomes for the same trials and to determine whether one approach is more reliable. Even if reviewers decide to use a scale, the best method to use to incorporate the scores is not known.

In the future meaningful scales will incorporate items common to all trials, be easy to use, and focus on items related to the reduction of bias. Scale developers are advised to evaluate and revise only scales that fulfill these requirements and not create new instruments. A full discussion of the process of scale development is beyond the scope of this paper. We suggest some guidelines (Table 5) to help ensure that new scales, if deemed necessary, are developed appropriately. Interested readers can refer to several other additional sources (20;36;51).

Guidelines for assessing study quality need to be developed specifically for systematic reviewers. We believe systematic reviewers should formally assess quality, using either scores or components. Regardless of how quality is assessed, reviewers should perform sensitivity analyses using at least two different methods of incorporating quality, with and without quality assessments, and an analysis in which the quality assessor was masked while assessing trial quality. At one extreme, investigators may decide that the quality of trials is so low, a meta-analysis is impossible. Gøtzsche (24) has suggested the quality of RCT reports in rheumatology is so weak that one cannot place confidence in their conclusions.

IMPROVING THE REPORTING OF RCTS

In 1987 a proposal was made for more informative abstracts (1) to improve the quality of reporting of clinical research. The "structured" abstracts proposed, and since this time broadly used, provide readers with both a series of headings pertaining to the design, conduct, and analysis of a trial and standardized information within each heading. Evidence to date indicates that more informative abstracts have had a positive impact on how the results of abstracts are communicated (27).

More recently, there has been a call to extend this approach to reporting of the text of all RCTs (50). Structured reporting requires providing sufficiently detailed information about the design, conduct, and analysis of the trial for the reader to have confidence that the report is an accurate reflection of what occurred during the various stages of the trial. Structured reporting outlines the items to be included in the report of a trial, reasons for inclusion, and ways they can be included. It is hoped that structured reporting will improve overall quality, provide readers with essential information about what happened during the trial, and facilitate the conduct of systematic reviews.

REFERENCES

1. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine,* 1987, 106, 598-604.
2. Andrew, E. Method for assessment of the reporting standard of clinical trials with Roentgen contrast media. *Acta Radiologica Diagnosis,* 1984, 25, 55-58.
3. Andrew, E., Eide, H., Fuglerud, P., et al. Publications on clinical trials with x-ray contrast media: Differences in quality between journals and decades. *European Journal of Radiology,* 1990, 10, 92-97.
4. Badgley, R. F. An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *Canadian Medical Association Journal,* 1961, 85, 246-50.
5. Beckerman, H., de Bie, R. A., Bouter, L. M., et al. The efficacy of laser therapy for musculoskeletal and skin disorders. In H. Beckerman & L. Bouter (eds.), *Effectiviteit van fysiotherapie: Een literatuuronderzoek.* Maastricht: Rijksuniversiteit Limburg, 1990, xx-xx.
6. Bland, J. M., Jones, D. R., Bennett, S., et al. Is the clinical trial evidence about new drugs statistically adequate? *British Journal of Clinical Pharmacology,* 1985, 19, 155-60.
7. Brown, S. A. Measurement of quality of primary studies for meta-analysis. *Nursing Research,* 1991, 40, 352-55.
8. Brown, A. S. Meta-analysis of diabetes patient education research: Variations in intervention effects across studies. *Research in Nursing & Health,* 1992, 15, 409-19.
9. Chalmers, I. Preparing, maintaining, and disseminating systematic review of the effects of health care. *Annals of the New York Academy of Sciences,* 1993, 703, 156-65.
10. Chalmers, I., Adams, M., Dickersin, K., et al. A cohort study of summary reports of controlled trials. *Journal of the American Medical Association,* 1990, 263, 1401-05.
11. Chalmers, I., Hetherington, J., Elbourne, D., et al. Materials and methods used in synthesizing evidence to evaluate the effects of care during pregnancy and childbirth. In I. Chalmers, M. Enkin, & M.J.N.C. Keirse (eds.), *Effective care in pregnancy and childbirth.* Oxford: Oxford University Press, 1989, xx-xx.
12. Chalmers, T. C., Adams, M., Dickersin, K., et al. A method for assessing the quality of randomized control trial. *Journal of the American Medical Association,* 1981, 2, 31-49.
13. Chalmers, T. C., Celano, P., Sacks, H. S., & Smith, H. Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine,* 1983, 309, 1358-61.
14. Chalmers, T. C., & Lau, J. Meta-analytic stimulus for changes in clinical trails. *Statistical Methods in Medical Research,* 1993, 2, 161-72.

15. Cho, M. K., & Bero, L. A. Instruments for reassessing the quality of drug studies published in the medical literature. *Journal of the American Medical Association,* 1994, 272, 101-04.
16. Colditz, G. A., Miller, J. N., & Mosteller, F. How study design affects outcomes in comparison of therapy. *Medical Statistics in Medicine,* 1989, 8, 441-54.
17. DerSimonian, R., Charette, L. J., McPeek, B., & Mosteller, F. Reporting on methods in clinical trials. *New England Journal of Medicine,* 1982, 306, 1332-37.
18. Detsky, A. S., Naylor, C. D., O'Rourke, K., et al. Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology,* 1992, 45, 225-65.
19. Evans, M., & Pollack, A. V. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *British Journal of Surgery,* 1985, 72, 256-60.
20. Feinstein, A. R. *Clinicmetrics.* New Haven, CT: Yale University Press, 1987.
21. Fleiss, J. L., & Gross, A. J. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology,* 1991, 44, 127-39.
22. Gardner, M. J., Machin, D., & Campbell, M. J. Use of checklists in assessing the statistical content of medical studies. In *Statistics with confidence: Confidence intervals and statistical guidelines.* London: BMJ Publications, 1989, 101-08.
23. Goodman, S. N., Berlin, J., Fletcher, R. H., & Fletcher, S. W. Manuscript quality before and after peer review and editing. *Annals of Internal Medicine,* 1994, 121, 11-21.
24. Gøtzsche, P. Methodology and overt and hidden bias: Reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials,* 1989, 10, 31-56 (erratum:356).
25. Grant, A. Reporting controlled trials. *British Journal of Obstetrics and Gynaecology,* 1989, 96, 397-400.
26. Grégoire, G., Derderian, F., & Le Lorier, J. Selecting the language of the publications included in a meta-analysis: Is there a tower of babel bias? *Journal of Clinical Epidemiology,* 1995, 48, 158-63.
27. Haynes, R. B., Mulrow, C. D., Huth, E. J., et al. More informative abstracts revisited. *Annals of Internal Medicine,* 1990, 113 69-76.
28. Imperiale, T. F., & McCulloughm, A. J. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomised trials. *Annals of Internal Medicine,* 1990, 113, 299-307.
29. Jadad-Bechara, A. R., Moore, R. A., & Carrol, D. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials,* 1996, 17, 1-12.
30. Jenicek, M. Meta-analysis in medicine: Where we are and where we want to go. *Journal of Clinical Epidemiology,* 1989, 42, 35-44.
31. Kleijnen, J., Knipschild, P., & ter Riet, G. Clinical trials of homeopathy. *British Medical Journal,* 1991, 302, 316-23.
32. Koes, B. W., Assendelft, W. J. J., van der Heijden, G. J. M. G., et al. Spinal manipulation and mobilisation for back and neck pain: A blinded review. *British Medical Journal,* 1991, 303, 1298-303.
33. Levine, J. *Trial assessment procedure scale (TAPS).* Bethesda, MD: Department of Health and Human Services, National Institute of Mental Health, 1980.
34. Lionel, N. D. W., & Herxheimer, A. Assessing reports of therapeutic trials. *British Medical Journal,* 1970, 3, 637-40.
35. Mahon, W. A., & Daniel, E. E. A method for the assessment of reports of drug trials. *Canadian Medical Association Journal,* 1964, 90, 565-69.
36. McDowell, I., & Newell, C. *Measuring health: A guide to rating scales and questionnaires.* New York: Oxford University Press, 1987.
37. Meinert, C. L. *Clinical trials, design, conduct, and analysis.* New York: Oxford University Press, 1986.

38. Moher, D., Jadad, A. R., Nichol, G., et al. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials,* in press.
39. Nurmohame, M. T., Rosendaal, F. R., Buller, H. R., et al. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: A meta-analysis. *Lancet,* 1992, 340, 152-56.
40. Onghenia, P., & Van Houdenhove, B. Antidepressants-induced analgesia in chronic non-malignant pain: A meta-analysis of 39 placebo-controlled studies. *Pain,* 1992, 49, 205-19.
41. Pocock, S. J. *Clinical trials: A practical approach.* Chichester, UK: John Wiley & Sons, 1983.
42. Pocock, S. J., Hughes, M. D., & Lee, R. J. Statistical problems in the reporting of clinical trials: A survey of three medical journals. *New England Journal of Medicine,* 1987, 317, 426-32.
43. Powe, N. R., Kinnison, M. L., & Steinberg, E. P. Quality assessment of randomized controlled trials of contrast media. *Radiology,* 1989, 170, 377-80.
44. Poynard, T. Evaluation de la qualité méthodologique des essais therapeutiques randomisés. *La Presse Medicale,* 1988, 17, 315-18.
45. Reisch, J. S., Tyson, J. E., & Mize, S. G. Aid to the evaluation of therapeutic studies. *Pediatrics,* 1989, 84, 815-27.
46. Sandercock, P .A. G., van den Belt, A. G. M., Lindley, R. I., & Slattery, J. Antithrombotic therapy in acute ischaemic stroke: An overview of the completed randomized trials. *Journal of Neurology, Neurosurgery, and Psychiatry,* 1993, 56, 17-25.
47. Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. Failure to conceal intervention allocation schedules in trials influenced estimates of treatment effects. *Controlled Clinical Trials,* 1994, 15, 63S.
48. Smith, K., Cook, D., Guyatt, G. H., et al. Respiratory muscle training in chronic airflow limitation: A meta-analysis. *American Review of Respiratory Disease,* 1992, 145, 533-39.
49. Spitzer, W. O., Lawrence, V., Dales, R., et al. Links between passive smoking and disease: A best evidence synthesis. *Clinical and Investigative Medicine,* 1990, 13, 17-42.
50. Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. *Journal of the American Medical Association,* 1994, 272, 1926-31.
51. Steiner, D. L., & Norman, G. R. *Health measurement scales: A practical guide to their development and use.* Oxford: Oxford University Press, 1989.
52. Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. *British Medical Journal,* 1948, II, 769-82.
53. Ter Riet, G., Kleijnen, J., & Knipschild, P. Acupuncture and chronic pain: A criteria-based meta-analysis. *Journal of Clinical Epidemiology,* 1990, 43, 1191-99.
54. Thomson, M. E., & Kramer, M. S. Methodologic standards for controlled clinical trials of early contact and maternal-infant behavior. *Pediatrics,* 1984, 73, 294-200.
55. Weintraub, M. How to critically assess clinical drug trials. *Drug Therapy,* 1982, 12, 131-48.