

Correlation of Quality Measures With Estimates of Treatment Effect in Meta-analyses of Randomized Controlled Trials

Ethan M. Balk, MD, MPH

Peter A. L. Bonis, MD

Harry Moskowitz, MD, MS

Christopher H. Schmid, PhD

John P. A. Ioannidis, MD

Chenchen Wang, MD, MSc

Joseph Lau, MD

SEVERAL STUDIES HAVE SUGGESTED that specific measures of trial quality, such as concealment of random allocation, blinding of patients and outcome assessors, and handling of dropouts, may significantly influence observed treatment effects in single studies,^{1,2} specific clinical areas,^{3,4} and meta-analyses from a mixture of clinical areas.^{5,6} Proposed quality measures have been incorporated into a growing number of scales that attempt to quantify overall trial quality.⁷ These findings have led to recommendations that investigators conducting meta-analyses should take into account the quality measures and scales when drawing conclusions.⁸⁻¹²

This approach can have a major impact on inferences drawn. In one study, Jüni et al³ found a wide range of estimates for the effectiveness of low-molecular-weight heparin for treatment of deep vein thrombosis by using different quality scales to divide “high-quality” from “low-quality” studies in a single meta-analysis. The summary odds ratio (OR), or the OR calculated by quantitatively combining indi-

Context Specific features of trial quality may be associated with exaggeration or shrinking of the observed treatment effect in randomized studies. Therefore, assessment of trial quality is often used in meta-analysis. However, the degree to which specific quality measures are associated with treatment effects has not been well established across a broad range of clinical areas.

Objective To determine if quality measures are associated with treatment effect size in randomized controlled trials (RCTs).

Design Quality measures from published quality assessment scales were evaluated in RCTs included in meta-analyses from 4 medical areas (cardiovascular disease, infectious disease, pediatrics, and surgery). Included meta-analyses incorporated at least 6 RCTs, examined dichotomous outcomes, and demonstrated significant between-study heterogeneity in the odds ratio (OR) scale.

Main Outcome Measures Relative ORs comparing overall treatment effect (summary OR) of high- vs low-quality studies, as determined by each quality measure, with relative ORs less than 1 indicating larger treatment effect in low-quality studies.

Results Twenty-four quality measures were analyzed for 276 RCTs from 26 meta-analyses. Relative ORs of high- vs low-quality studies for these quality measures ranged from 0.83 to 1.26; none was statistically significantly associated with treatment effect. The proportion of studies fulfilling specific quality measures varied widely in the 4 medical areas. In analyses limited to specific medical areas, placebo control, multicenter studies, study country, caregiver blinding, and statistical methods were significantly associated with treatment effect on 7 occasions. These relative ORs ranged from 0.40 to 1.74. However, the directions of these associations were not consistent.

Conclusions Individual quality measures are not reliably associated with the strength of treatment effect across studies and medical areas. Although use of specific quality measures may be appropriate in specific well-defined areas in which there is pertinent evidence, findings of associations with treatment effect cannot be generalized to all clinical areas or meta-analyses.

JAMA. 2002;287:2973-2982

www.jama.com

vidual ORs from similar studies, varied depending on which studies were determined to be of high quality and

were thus included in meta-analysis. In a controversial recent meta-analysis, Gøtzsche and Olsen¹³ found that screen-

Author Affiliations: Evidence-based Practice Center, Division of Clinical Care Research, Tufts University School of Medicine, New England Medical Center, Boston, Mass (Drs Balk, Bonis, Moskowitz, Schmid, Wang, and Lau); and the Biomedical Research Institute, Foundation for Research and Technology Hellas, Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology,

University of Ioannina School of Medicine, Ioannina, Greece (Dr Ioannidis). Dr Moskowitz is now with the Division of General Pediatrics, Mount Sinai Hospital, Mount Sinai School of Medicine, New York, NY.

Corresponding Author and Reprints: Joseph Lau, MD, Division of Clinical Care Research, New England Medical Center, 750 Washington St, No. 63, Boston, MA 02111 (e-mail: JLau1@lifespan.org).

ing mammography did not reduce breast cancer deaths in 2 studies with “adequate randomization,” while a highly significant effect was found among the 5 studies in which randomization was “not adequate.” However, the analysis was criticized for its definition of inadequate randomization and failure to consider other explanations, including other quality measures.^{14,15} Furthermore, the quality measures found to be associated with treatment

Table 1. Quality Measure Definitions

Quality Measure	Description
Study question well defined in introduction/methods	Study needed to clearly define intervention studied, population studied, condition of interest, and outcome of interest in introduction or methods sections of main body of text or abstract.
Study question well defined anywhere in article	As above, but criteria could be met from any section of article.
Placebo control	Required term <i>placebo</i> or description of placebo (eg, saline).
Appropriate outcome studied	Were study outcomes appropriate based on study design, condition, and intervention studied?
Multicenter study	Did study include more than 1 site?
Study country	Study considered to be from United States or other “research country” if any of the investigators were based in that country. Analyzed 2 ways.*
Adequate selection criteria	Were inclusion and exclusion criteria clearly and completely reported?
Randomization methods described	Was any description given of how randomization (allocation among treatment arms) was achieved, or did the article say only “randomized”?
Central randomization site	Was randomization performed by researchers at a site separate from the patients and caregivers (central) or at a site where caregivers could be involved in patient allocation (local)? Both single-center and multicenter studies could have central or local randomization. Randomization by pharmacy or laboratory staff was assumed to be central unless there was indication that these staff may have been directly involved in patient care. Randomization methods such as use of envelopes, cards, or registration numbers were assumed to be local unless explicitly stated.
Allocation concealment	Was allocation fully concealed? If randomization site was central or randomization method was performed using computers, blinded code or blinded medicine vials, or opaque envelopes, allocation was adequately concealed. Tables, cards, etc, were not adequately concealed. Randomization by birth year or registration number was not adequately concealed regardless of where randomization was performed.
Patients blinded	Were patients reported to have been blinded? If not stated explicitly, infants and patients receiving identical-appearing treatments (active or placebo) were considered to have been blinded.
Caregivers blinded	Caregivers included physicians, nurses, and other health care practitioners in direct patient care or parents (or equivalent) of outpatient infants.
Outcome assessors blinded	Outcome assessors included physicians or other health care practitioners or researchers who evaluated either patients, their records, or their laboratory or radiology tests to determine study outcomes.
Data analysts blinded	Data analysts were considered to be blinded in studies that explicitly reported that the analysis of data was performed by individuals who were unaware of the treatment assignment.
Double blinded	Were both patients and either caregiver or outcome assessor blinded?
Valid statistical methods	Were the statistical methods used considered valid and appropriate, based on study design and outcomes of interest?
Statistician author or acknowledged	The degrees and department affiliations of the study authors were examined. If any author had an MPH or PhD or equivalent, or if any author was a member of a department of statistics, epidemiology, or equivalent, that person was considered to be a statistician (or to have statistical knowledge). In addition, the acknowledgment section was reviewed for mention of a statistician.
Intention-to-treat analysis	Are all analyzed patients analyzed in the group to which they were originally allocated? Dropouts were allowable so long as the reasons for withdrawal were not related to the group to which they were assigned (bias). ⁶⁴
Power calculation reported	Was a power calculation reported for any outcome evaluated in the study?
Stopping rules described	Did the article report and describe rules for stopping the study, such as excess mortality? (This does not include the rules for dropping patients from the study.)
Baseline characteristics reported	Were any baseline characteristics reported that compared the treatment and control groups?
Groups similar at baseline	Were the treatment and control groups similar in the characteristics reported?
Confounders accounted for	If there were baseline differences in the groups that could be confounders, were these examined?
Dropouts recorded	Were the number of dropouts recorded (either explicitly or by reporting the number enrolled and the number evaluated)?
Percentage dropouts	What percentage of subjects dropped out?
Reason for dropouts given	If there were dropouts, were the reasons for dropouts reported?
Findings support conclusions	Were the conclusions valid based on the findings, study design, and power?

*See Table 4.

effect vary among investigators. Schulz et al⁴ reported that poorly concealed allocation or lack of double blinding resulted in a significant overestimation of treatment effect by 41% and 17%, respectively, in 250 studies of perinatal medicine. Moher et al³ reported a similar bias for allocation concealment but no significant bias for double blinding. Others found generally larger bias for double blinding but no significant bias for allocation concealment.^{2,3,6}

The uncertain association of different quality measures with treatment effect and the absence of a gold-standard quality assessment instrument has resulted in a proliferation of quality scales used in meta-analyses. Jüni et al³ identified 37 meta-analyses that used 26 different instruments to assess trial quality. The number of specific quality measures in these scales ranged from 3 to 34, and the weights assigned to 3 common measures (randomization, blinding, and dropouts) ranged from 0% to 100%.

Adding to the uncertainty, quality is not consistently defined across specialties, nor have specific quality measures been shown to correlate with treatment effects in different clinical areas. A more detailed understanding of the relationship between specific features of study quality and estimates of treatment effect is needed. This study was designed to measure the degree to which study quality, as determined by a wide range of previously described measures of study design and conduct, is associated with combined estimates of treatment effect from a variety of meta-analyses that included randomized controlled trials (RCTs) from several medical and surgical areas.

METHODS

We selected meta-analyses from 4 medical areas (cardiovascular disease, infectious disease, pediatrics, and surgery), and extracted data on specific quality measures and outcomes from the RCTs that had been included in the meta-analyses. For each quality measure, we then calculated a relative OR for treat-

ment effect, defined as the ratio of the strength of the treatment effect in studies in which the quality measure was present to the strength of the effect in studies in which it was absent.

Quality Measures Used

We identified specific quality measures previously demonstrated or hypothesized to be associated with estimates of treatment effect by reviewing published studies of quality measures and quality assessment scales.^{3-5,7,16-33} These studies were compiled from a MEDLINE search for *quality* and *randomized controlled trials* and from reference lists of methodological articles. We used the definitions for each quality measure as described by authors. For quality measures not clearly described, we reached consensus on definitions. We aimed to establish definitions of study quality that could be applied most consistently across a variety of study types. Thus, we formalized a process that all researchers grading the quality of studies would have to perform. Definitions of quality measures are listed in TABLE 1.

Analyses were performed only on quality measures for which we could reach consensus on the definition and could dichotomize. Studies that did not report on a specific quality measure were assumed to be of low quality for that measure.

Selection of Meta-analyses

We selected meta-analyses in 4 areas (cardiovascular disease, infectious disease, pediatrics, and surgery) because they represent a variety of medical areas. We selected cardiovascular meta-analyses from among those used in a previous analysis by our group.³⁴ Meta-analyses for other areas were found by searching the MEDLINE database (1966-2000) and the Cochrane Database of Systematic Reviews (2000, issue 4).

Included meta-analyses incorporated at least 6 RCTs, examined dichotomous outcomes, and demonstrated significant between-study heterogeneity in the OR scale ($P < .10$ for the χ^2 statistic or a nonzero between-study vari-

ance, τ^2 , by the DerSimonian and Laird random-effects model).^{35,36} We required statistical heterogeneity of treatment effect across trials within each meta-analysis because meta-analyses with homogenous treatment effects across trials are unlikely to find that estimates of treatment effects are associated with quality measures (or other factors). We excluded abstracts, letters, unavailable articles, and those for which detailed outcomes data were not provided. Meta-analyses were selected without a priori knowledge of the quality of the studies used. All meta-analyses that met inclusion criteria were included.

Outcomes Evaluated

For cardiovascular studies, the outcome used was mortality. For studies in the other clinical areas, the outcome used varied across meta-analyses. Within meta-analyses, only outcomes with heterogeneous treatment effects were considered. If multiple outcomes were available for analysis, those examined by the largest number of studies or that were most clearly defined were used. Failure of treatment or control (eg, death) was considered a positive outcome in all studies.

Data Extraction

We developed the quality assessment form and extracted data in a 4-stage process. First, 4 clinicians (E.M.B., P.A.L.B., H.M., and C.W.) trained in clinical epidemiology and study design coded data from the same pilot set of 8 studies and discussed discrepancies. Second, the quality assessment form was revised and was again tested by having each investigator extract data from a different pilot set of 8 studies. Further refinements and clarifications were performed in the data extraction definitions of specific quality measures. Third, the 4 investigators independently extracted data from the remaining English-language RCTs. Data from each trial were extracted by 2 investigators. The studies were divided so that each investigator would be paired with each of the 3 other data extractors for approximately one third of the studies. This helped en-

sure uniform application of definitions and scaling of the quality items. When necessary, data were extracted from referenced articles that described a study's methods. Fourth, discrepancies were reviewed to achieve consensus between each pair of data ex-

tractors. A third investigator arbitrated disagreements. Data from 13 Spanish-, German-, French-, and Italian-language articles were extracted by single investigators in consultation with other investigators. Studies in other languages were excluded.

Statistical Analyses

Quality measures were dichotomized to capture high quality vs low quality. We estimated the effect of quality measures by calculating relative ORs of treatment effect for each measure. The relative OR compares the OR of high-

Table 2. Descriptions and Summary Odds Ratios and Heterogeneity of Meta-analyses Examined

Meta-analysis Category	Treatment Examined	Outcome Examined	No. of Studies Evaluated	Summary OR (95% Confidence Interval) of Evaluated Studies*	χ^2 of Between-Study Heterogeneity
Cardiovascular Disease					
Antiplatelet Trialists' Collaboration, ³⁸ 1988	Aspirin	Mortality	10	0.88 (0.78-1.00)	12.53†
Hine et al, ³⁹ 1989‡	Class I antiarrhythmics	Mortality	10	1.22 (0.92-1.62)	13.23†
Leizorovicz and Boissel, ⁴⁰ 1983‡	Anticoagulants	Mortality	11	0.83 (0.68-1.01)	20.61
Yusuf et al, ⁴¹ 1985‡	β -Blockers	Mortality	13	0.83 (0.73-0.94)	17.33†
Yusuf et al, ⁴² 1985‡	IV Streptokinase	Mortality	27	0.76 (0.69-0.83)	32.11†
Yusuf et al, ⁴³ 1988‡	Nitrates	Mortality	9	0.73 (0.53-0.98)	8.42†
Rossouw et al, ⁴⁴ 1990	Cholesterol reduction	Mortality	7	0.85 (0.78-1.03)	8.12†
Teo and Yusuf, ⁴⁵ 1993	Magnesium	Mortality	6	0.31 (0.17-0.58)	20.44
Infectious Disease					
Colditz et al, ⁴⁶ 1994	BCG vaccine	Tuberculosis	10	0.48 (0.33-0.70)	62.89
Jefferson et al, ⁴⁷ 2000	Amantadine and rimantadine	Clinical or serological influenza	8	0.36 (0.23-0.57)	10.42†
Langley et al, ⁴⁸ 1993	Preoperative antibiotics	Shunt infection	10	0.51 (0.31-0.84)	9.89†
McIntosh and Olliaro, ⁴⁹ 2000	Artemisinin drug	Malaria mortality	10	0.63 (0.41-0.98)	17.79
Smaill, ⁵⁰ 2000	Antibiotics	Pyelonephritis	8	0.22 (0.12-0.41)	10.00†
Smieja et al, ⁵¹ 2000	Isoniazid	Active tuberculosis	10	0.37 (0.29-0.46)	15.24
Pediatrics					
Ausejo et al, ⁵² 1999	Glucocorticoids	Croup score improves <2	9	0.32 (0.16-0.65)	10.08†
Bhuta and Ohlsson, ⁵³ 1998	Dexamethasone	Chronic lung disease	6	0.32 (0.11-0.87)	18.40
Kellner et al, ⁵⁴ 1996	Bronchodilators	Unimproved distress score (bronchiolitis)	8	0.29 (0.12-0.69)	21.03
Kozyrskij et al, ⁵⁵ 1998	Short-course antibiotics	Acute otitis media failure to cure	27	1.09 (0.94-1.27)	37.23
Rosenfeld and Post, ⁵⁶ 1992	Antibiotics	Otitis media with effusion failure to cure	10	0.34 (0.18-0.62)	41.77
Surgery					
Chan et al, ⁵⁷ 1994	Highly selective vagotomy	Visick score	11	1.97 (1.13-3.44)	13.98†
Chung and Rowland, ⁵⁸ 1999	Laparoscopic hernia repair	Recurrence	6	0.80 (0.27-2.40)	11.87
MacRae and McLeod, ⁵⁹ 1998	Stapled colorectal anastomoses	Anastomotic leak	9	1.10 (0.68-1.78)	15.69
Martin-Hirsch et al, ⁶⁰ 2000	Laser ablation	Residual cervical intraepithelial neoplasia	7	1.00 (0.59-1.68)	20.11
Nelson, ⁶¹ 1999	Anal stretch	Anal fissure persistence	6	1.64 (0.48-5.63)	13.73
Pocock et al, ⁶² 1995	Percutaneous coronary intervention	Myocardial infarction or cardiac death	7	0.92 (0.66-1.29)	11.02
Sauerland et al, ⁶³ 1998	Laparoscopic appendectomy	Complications	21	0.87 (0.61-1.25)	30.84

*Summary random-effects model odds ratio (OR) of unfavorable outcome. Calculations performed in Meta-Analyst version 0.991 (Boston, Mass).
 †Of between-study heterogeneity not significant at $P < .10$ value, but between-study variance by DerSimonian and Laird random-effects model, $\tau^2 > 0$.³⁶
 ‡Updated by Lau et al.³⁴

quality studies to that of low-quality studies for each quality measure. Relative ORs greater than 1 indicate that high-quality studies had larger ORs than low-quality studies.

To estimate the relative OR, we used a Bayesian hierarchical model with random effects.³⁷ This multilevel structure accounted for the nesting of trials within meta-analyses as well as the variability across meta-analyses. For each trial, we assumed that the outcomes followed binomial distributions independently in the treatment and control groups. The log odds of the probability of an outcome in each control group was assumed to be normally distributed, centered around an average log odds for the meta-analysis. The log OR of an outcome, defined as the difference in log odds between the treatment and control groups, was assumed to be normally distributed, with mean $\alpha_j + \beta_j \times x_{ij}$, where x_{ij} is the quality measure in the *i*th study of the *j*th meta-analysis. For a dichotomous quality measure, β_j represented the relative log OR between the 2 levels of the measure. The exponential of β_j is the relative OR. Both the mean log odds in the control group and the regression slope and intercept for the log OR differed across meta-analyses.

These regression slopes and intercepts were assumed to be random effects drawn from a population of such slopes and intercepts. We used 2 different population models. One model assumed a single common mean intercept and slope for the population, around which the α_j and β_j varied according to a normal distribution with common variances τ_α^2 and τ_β^2 , respectively. The other model assumed different α_j and β_j by medical area so that there were 4 separate population intercepts and slopes corresponding to the cardiovascular disease, infectious disease, pediatric, and surgical areas. Non-informative prior distributions were chosen for all parameters to simulate the random-effects model.

Models were fit using a Markov chain Monte Carlo algorithm with WinBUGS software version 1.3 (D. J. Spiegelhalter, A. Thomas, and N. G. Best, Medi-

cal Research Council Biostatistics Unit, Cambridge, England), with appropriate convergence of the Markov chains.

Assessment of the associations between quality measures and treatment effect were limited to quality measures that were present in 10% to 90% of the trials. These cutoffs were chosen to ensure sufficient heterogeneity in the quality measures for meaningful comparisons. Analysis of the percentage of dropouts was limited to studies that reported whether there were dropouts. Analyses of whether dropouts were explicitly recorded and whether the reasons for dropouts were recorded were

limited to meta-analyses that included 6 or more studies that provided information on dropouts.

RESULTS

Meta-analyses and RCTs Included in the Study

Twenty-six meta-analyses were included in the analysis (TABLE 2). These included 8 cardiovascular disease,³⁸⁻⁴⁵ 6 infectious disease,⁴⁶⁻⁵¹ 5 pediatric,⁵²⁻⁵⁶ and 7 surgical meta-analyses.⁵⁷⁻⁶³ We extracted data from 276 RCTs, which represented 85% of the trials from the meta-analyses (a list of the trials is available from the author). The

Table 3. Percentage of Studies With High-Quality Measures*

Quality Measures	Overall (N = 276)	Cardiovascular Disease (n = 93)	Infectious Disease (n = 56)	Pediatrics (n = 60)	Surgery (n = 67)
Study question well defined in introduction and/or methods	87	88	89	97	76
Study question well defined anywhere in article	96	99	95	98	91
Placebo control	41	61	54	43	1
Appropriate outcome studied	99	100	100	100	94
Multicenter study	47	68	43	38	28
Study country, United States	30	27	38	43	16
Research country†	89	98	70	92	90
Adequate selection criteria	92	96	88	97	87
Randomization methodology described	61	59	64	57	64
Central randomization site	24	30	29	27	7
Allocation concealment	39	51	34	35	31
Patients blinded	46	61	55	62	1
Caregivers blinded	38	52	38	60	0
Outcome assessors blinded	42	52	43	55	16
Data analysts blinded	7	11	2	10	3
Double blinded	40	53	43	62	0
Valid statistical methods	75	77	61	83	78
Statistician author or acknowledged	36	49	23	27	34
Intention-to-treat analysis	83	92	80	92	64
Power calculation reported	25	32	13	28	21
Stopping rules described	5	10	2	5	3
Baseline characteristics reported	88	95	71	88	91
Groups similar at baseline	77	83	59	82	81
Confounders accounted for	82	91	66	87	79
Dropouts recorded	89	98	66	100	85
Reason for dropouts given	77	88	59	91	62
Median percentage of dropouts	3	4	3	4	1
Findings support conclusions	91	96	84	92	90

*All data are presented as percentages.

†Australia, Canada, Israel, Japan, New Zealand, United States, and Western Europe.

remaining trials were generally reported in abstracts, letters, or unavailable journals.

Quality Measures

The final data extraction form included 28 quality measures (Table 1 and TABLE 3). These included questions on study definition and design, study location, randomization, blinding, statistical analysis, reporting, subject withdrawals, and conclusions.

Overall, interrater agreement of quality measures was high. Prior to recon-

ciliation of discrepancies, a median of 86% of responses agreed for each quality measure. Outcome assessor blinding, inclusion of a statistician, accounting for confounders, and randomization site had the poorest agreement, ranging from 69% to 78%. Study country and outcome appropriateness had the highest agreement at 97% and 96%, respectively. Determining whether the study was performed as an intention-to-treat analysis proved to be the most difficult question to clearly define. After data extraction was complete, all

studies were reviewed in conference to determine the type of analysis using the definition of the intention-to treat principle by Lachin.⁶⁴

Frequency of Quality Measures

Quality measures were present in different proportions of studies within each of the 4 clinical domains. Many of the differences were due to the inherent differences of studies within the 4 clinical areas. For example, patient and caregiver blinding and placebo control were rare among surgical trials but were com-

Table 4. Complete Relative Odds Ratio Results*

Quality Measures	Relative Odds Ratio (95% Confidence Interval)†				
	Overall (N = 276)	Cardiovascular Disease (n = 93)	Infectious Disease (n = 56)	Pediatrics (n = 60)	Surgery (n = 67)
Study question well defined in introduction or methods	0.85 (0.64-1.06)	0.98 (0.68-1.30)	0.58 (0.32-1.80)	NA	0.81 (0.42-1.40)
Study question well defined anywhere in paper	NA	NA	NA	NA	NA
Placebo control	0.85 (0.61-1.09)	1.03 (0.84-1.27)	0.62 (0.40-0.91)‡	0.40 (0.22-0.86)‡	NA
Appropriate outcome studied	NA	NA	NA	NA	NA
Multicenter study	1.06 (0.90-1.25)	1.30 (0.87-1.94)	0.96 (0.68-1.40)	1.74 (1.09-2.80)‡	0.71 (0.46-0.94)‡
Study country, United States	1.05 (0.93-1.19)	1.12 (0.94-1.38)	0.87 (0.55-1.29)	1.02 (0.55-1.58)	0.84 (0.49-1.38)
Research country§	0.95 (0.70-1.29)	NA	0.62 (0.44-0.92)‡	NA	0.99 (0.61-1.59)
Adequate selection criteria	0.94 (0.73-1.28)	NA	1.73 (0.81-5.49)	NA	0.65 (0.45-1.11)
Randomization methodology described	1.03 (0.89-1.20)	1.14 (0.95-1.40)	1.13 (0.73-1.67)	1.00 (0.63-1.46)	0.76 (0.53-1.08)
Central randomization site	1.01 (0.85-1.18)	1.14 (0.91-1.49)	0.93 (0.58-1.64)	0.88 (0.49-1.51)	NA
Allocation concealment	1.05 (0.91-1.21)	1.14 (0.96-1.42)	0.97 (0.68-1.42)	0.90 (0.58-1.28)	0.73 (0.36-1.24)
Patients blinded	0.95 (0.70-1.13)	1.08 (0.86-1.38)	0.70 (0.46-1.11)	0.79 (0.39-1.19)	NA
Caregivers blinded	0.98 (0.75-1.20)	1.09 (0.91-1.29)	0.62 (0.43-0.91)‡	1.13 (0.73-1.84)	NA
Outcome assessors blinded	1.02 (0.82-1.22)	1.11 (0.87-1.39)	0.84 (0.55-1.27)	1.02 (0.57-1.61)	0.87 (0.56-1.36)
Data analysts blinded	NA	NA	NA	NA	NA
Double blinded	1.02 (0.79-1.24)	1.10 (0.90-1.33)	0.71 (0.47-1.12)	1.05 (0.56-1.61)	NA
Valid statistical methods	1.11 (0.95-1.31)	1.03 (0.81-1.33)	1.17 (0.78-1.77)	0.97 (0.48-1.73)	1.63 (1.03-2.83)‡
Statistician author or acknowledged	1.04 (0.92-1.17)	1.05 (0.86-1.29)	0.85 (0.59-1.28)	1.12 (0.81-1.60)	1.13 (0.73-1.67)
Intention-to-treat analysis	0.91 (0.70-1.13)	0.89 (0.60-1.25)	0.80 (0.42-1.37)	NA	1.14 (0.73-2.05)
Power calculation reported	1.08 (0.95-1.23)	1.04 (0.89-1.22)	1.32 (0.88-1.95)	1.13 (0.76-1.62)	0.91 (0.55-1.43)
Stopping rules described	NA	NA	NA	NA	NA
Baseline characteristics reported	1.00 (0.68-1.52)	1.19 (0.69-1.96)	1.18 (0.66-2.21)	0.66 (0.35-1.29)	NA
Groups similar at baseline	1.06 (0.90-1.24)	1.03 (0.73-1.33)	1.15 (0.82-1.71)	1.13 (0.73-1.66)	0.77 (0.51-1.30)
Confounders accounted for	0.96 (0.79-1.23)	NA	0.94 (0.60-1.49)	0.96 (0.50-1.65)	1.20 (0.76-1.72)
Dropouts recorded¶	1.26 (0.87-2.05)	NA	1.11 (0.62-1.91)	NA	1.12 (0.54-2.33)
Reason for dropouts given¶	0.93 (0.77-1.13)	0.94 (0.75-1.17)	0.94 (0.55-1.52)	NA	0.70 (0.43-1.16)
Percentage of dropouts#	1.02 (0.94-1.12)	1.00 (0.89-1.14)	1.07 (0.84-1.34)	1.17 (0.92-1.57)	1.03 (0.86-1.27)
Findings support conclusions	0.83 (0.66-1.10)	NA	0.79 (0.57-1.11)	NA	0.71 (0.30-1.46)

*NA indicates unable to analyze because too few (<10%) or too many (>90%) studies met quality criteria.
 †Relative odds ratios are the ratio of odds ratios of studies with quality measure to odds ratios of studies without quality measure, weighted by random-effects model method.
 ‡A larger number implies a larger treatment effect in those studies with the quality measure (treatment was associated with more bad outcomes than control).
 ‡Odds ratios are statistically significant at P = .05.
 §Australia, Canada, Israel, Japan, New Zealand, United States, and Western Europe.
 ||Patient and either caregiver or outcome assessor blinded.
 ¶N = 141. See text.
 #N = 261. Remaining studies did not report whether there were any dropouts. Data are relative odds ratios per 1 percentage-point increase in dropouts.

mon among cardiovascular disease studies. Four quality measures could not be reliably analyzed because either too few or almost all studies included the quality criteria (TABLE 4).

Quality Measure Associations With Treatment Effect

When all clinical domains were combined, point estimates for relative ORs of high-quality vs low-quality studies for the quality measures ranged from 0.83 to 1.26 (Table 4). However, none

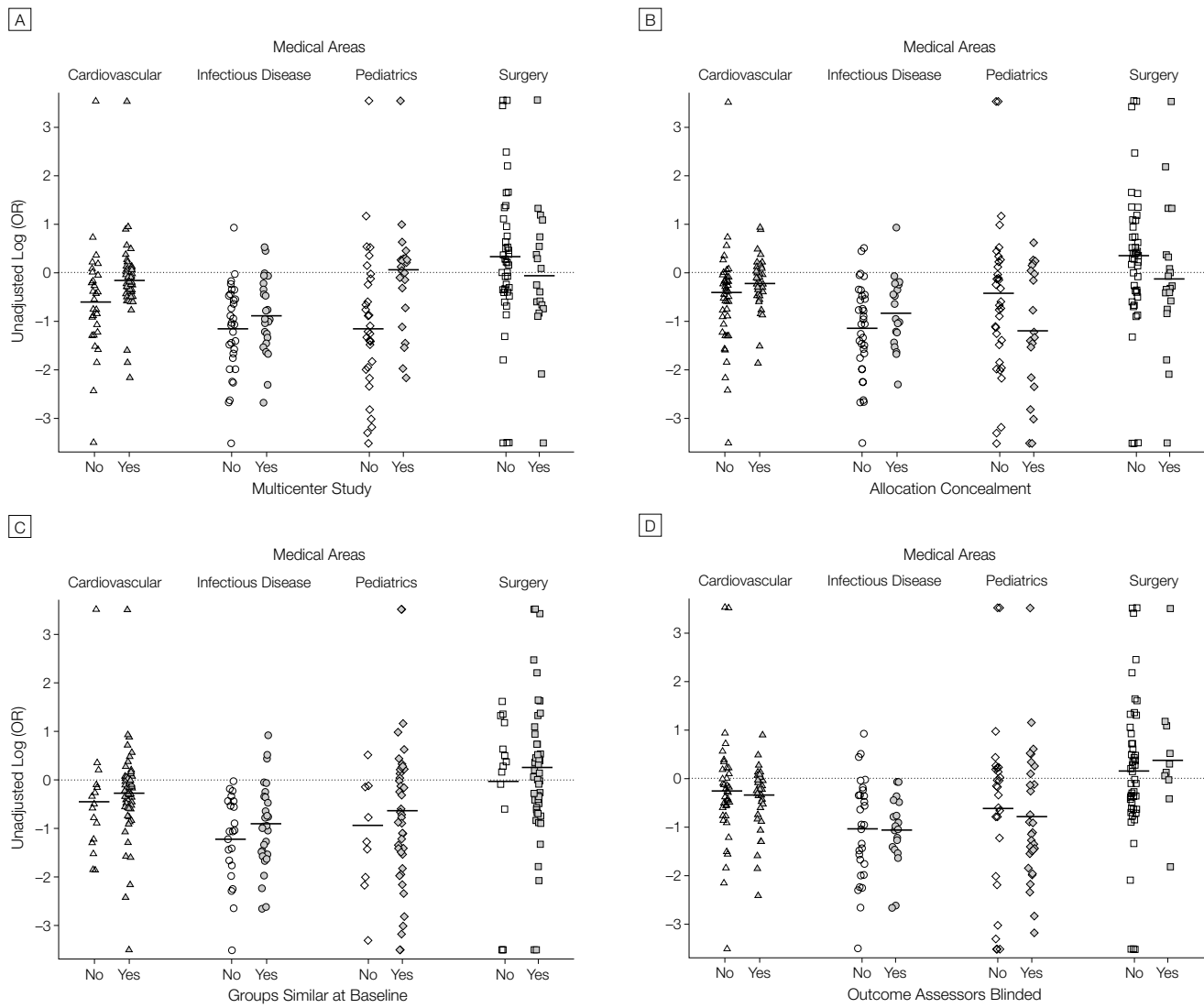
of the 24 tested quality measures was found to be significantly associated with treatment effect. Based on 95% confidence intervals, there were trends toward association of study quality and treatment effect for use of valid statistical methods and reporting of power calculations.

When the 4 clinical areas were considered separately, 5 quality measures had significant associations with treatment effect in 7 cases (Table 4). However, no consistent patterns emerged.

Multicenter studies appeared to be associated with either an increase or a decrease in treatment effect in pediatric and surgical studies, respectively.

FIGURE 1 and FIGURE 2 display 2 sets of complementary graphs for 4 quality measures chosen because they are commonly thought to be associated with treatment effect or because of inconsistent findings in different medical areas (ie, multicenter study). In Figure 1, the scatterplots of the unadjusted treatment effects of studies scoring as high

Figure 1. Relationship of Unadjusted Treatment Effect and Quality Measures



Relationship of unadjusted log odds ratio (OR) of individual studies (N=276) and 4 quality measures. Markers are arranged in matched columns by medical area. Horizontal bars represent unadjusted mean log OR of studies within each column.

quality compared with those of low quality is roughly the same. Even in the few cases in which apparently large differences in the mean treatment effects of high- and low-quality studies occur (eg, multicenter studies and allocation concealment in pediatric studies), the range of treatment effects across studies was generally similar.

Figure 2 displays the statistical analysis by adjusting the treatment effects for each clinical area and meta-analysis. The graphs directly compare the adjusted log OR of combined treatment effect estimates of high- and low-quality studies of each meta-analysis. Again, no quality measure consistently differentiated studies by treatment effect across medical areas, which would be observed in clustering of points to one side of the diagonal line of identity. Except for occasional outliers, the treatment effects of high- and low-quality studies were similar within each meta-analysis, regardless of the quality measure used.

Analyses were also performed using fixed-effects and random-effects linear regression models, controlling for meta-analysis and medical area. Results were similar.

COMMENT

Previous studies have described associations between specific quality measures and treatment effects.^{2,5,20} In contrast, our analysis did not reveal any consistent associations between quality measure and the magnitude of the treatment effect in 4 clinical areas. In particular, double blinding and allocation concealment, 2 quality measures that are frequently used in meta-analyses, were not associated with treatment effect.

Our sample included studies from heterogeneous meta-analyses in 4 medical areas. We might have found some of the quality measures to be statistically significant if we had analyzed a broader range of clinical areas. In particular, it is possible that various quality measures that trended toward significance could have been significant if they had been applied to a different set

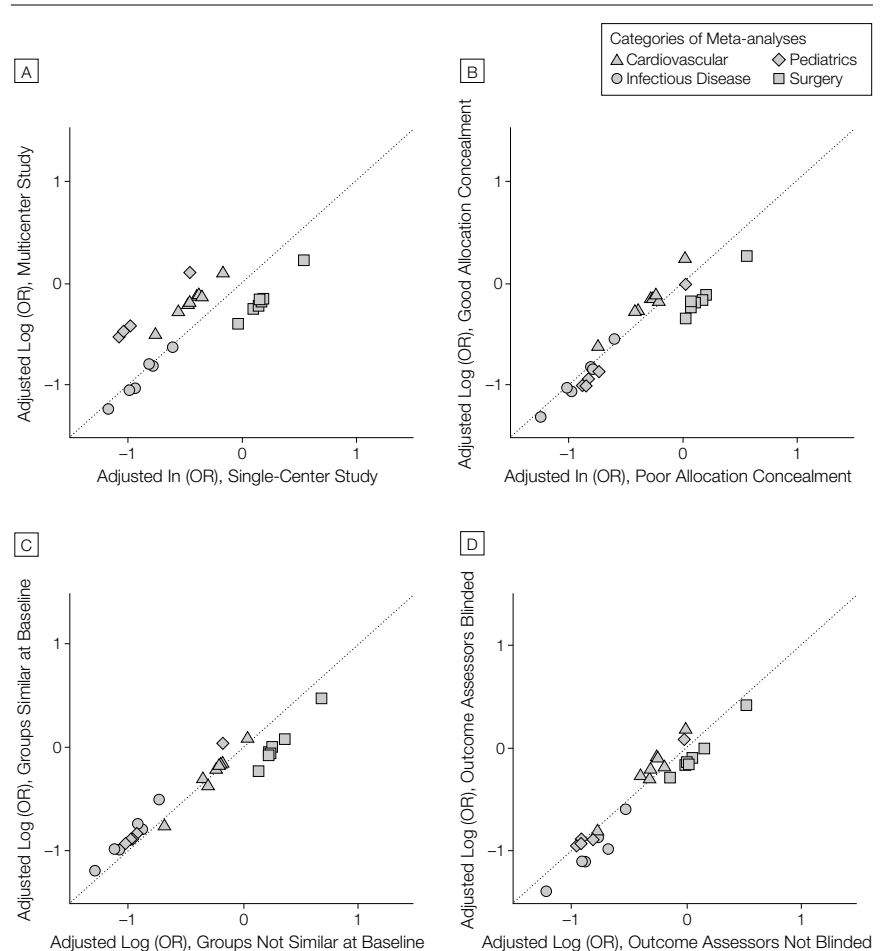
of clinical areas or if we had included an even larger number of RCTs. However, the small magnitude of the relative ORs (0.83-1.26, with most ranging from 0.93-1.08) and their lack of consistency suggest that quality effects are not as large as earlier reports have found. Furthermore, the observation that only 7 (7%) of 102 associations tested were statistically significant at the $P < .05$ level suggests that our positive findings may have been due to chance alone.

The variation in the direction of the treatment effects significantly associated with quality measures further calls

into question whether any of these associations could provide a general rule for evaluating the quality of RCTs across clinical areas. For example, multicenter studies were associated with a stronger treatment effect in cardiovascular and pediatric trials but a weaker treatment effect in infectious disease and surgical trials. Relative ORs were less than 1 for 10 quality measures and greater than 1 for 13 measures.

Other studies that have examined this issue have generally focused on individual meta-analyses or on single clinical categories of meta-analyses. Furthermore, the majority based their

Figure 2. Relationship of Adjusted Mean Treatment Effect and Quality Measures



Relationship of adjusted mean summary log odds ratio (OR) of high- and low-quality studies within meta-analyses (N=26) for 4 quality measures. If the adjusted log ORs of high- and low-quality studies are equal within a meta-analysis, marker falls on diagonal line of identity.

conclusions on a relatively small number of RCTs. An exception is the study by Moher et al,⁵ which also included multiple meta-analyses from various clinical categories. An association was found between treatment effect and both Jadad score¹⁶ and adequacy of allocation concealment. Although the associations were statistically significant, the differences were small.⁶⁵ Our findings do not discount the possibility that certain quality measures may be associated with treatment effect in specific clinical disciplines and for specific questions of interest. However, our analysis does call into question whether previous findings of quality-related modification of the treatment effect can be generalized across different medical disciplines or even across meta-analyses within a discipline.⁸

Another factor that may have contributed to the differences in our conclusions compared with previous studies may be the definitions used for the quality measures. There are innumerable ways to define study quality and specific quality measures. Thus, interpretation of the meaning of certain quality measures may have differed from previous reports. Although we met frequently to define and redefine quality measures to ensure consistency and clarity and analyzed only those measures that could be clearly defined, our definitions probably differ slightly from those of other authors. Furthermore, the influence of quality on treatment effect is frequently difficult to assess because the details of a trial's methods may not be fully reported. For a variety of reasons, almost all articles are incomplete in their reporting of various study aspects. It is frequently difficult to distinguish between methodologically poor studies and omissions in reporting the methods used.⁸ Hopefully, the publication of the original and revised CONSORT statements will lead to more complete reporting.^{28,66}

We found that the proportion of studies rated as high quality using the different quality measures varied considerably across the different medical areas, an observation consistent with

previous studies.^{1-5,29} A possible contributing factor is that certain quality measures may be easiest to apply within particular types of studies. For example, we found that the assessment of whether the caregiver was blinded was generally straightforward in studies of surgical interventions compared with some of the other types of studies.

Many factors can explain imprecision of treatment effects and heterogeneity of study findings found in meta-analyses. In addition to study quality, other factors include heterogeneity of study populations, treatments, outcomes, and study design⁶⁷; biases due to study design, meta-analysis inclusion criteria, publication bias⁶⁸ and evolving treatment effects⁶⁹; and random error.³⁵ Thus, quality is only one component of heterogeneity and has an uncertain role in explaining any treatment effect differences.

When evaluating the validity of studies, readers should continue to assess the quality of the study methods and reporting. This information is useful for understanding potential shortcomings and biases and for judging the generalizability of the results. However, it should not be assumed that any given quality measure will necessarily explain the treatment effect found. It is reasonable for researchers performing meta-analysis to continue using quality measures to examine heterogeneity among studies; however, the use of a given list of quality measures for all meta-analyses is probably not appropriate. Furthermore, one should consider that any quality measure that is found to partly explain heterogeneity in a given meta-analysis may do so purely by chance. Quality-related differences in the treatment effect should be treated as hypothesis-generating observations.

Our analysis also documents that the appraisal of quality in RCTs and meta-analyses is not straightforward. Unless definitions of quality measures are robustly constructed and validated, interrater agreement may often be unacceptably low. Subtle clarifications may be essential. We used a stringent ap-

proach to define quality measures, with 2 successive pilot phases, to ensure that quality measures were explicitly defined and clarified. Studies using less-rigorous methods would probably find even more variability in determination of study quality than we found.

Our study indicates that it would be inappropriate to quantitatively adjust the treatment effect of a given study or meta-analysis by using the average effects of specific quality measures discerned from prior meta-analyses.^{5,65} Assessment of quality may be useful in better understanding qualitative aspects of RCTs and meta-analyses on a case-by-case basis, but their translation to overarching, quantitative adjusting factors is precarious and should be avoided.

Author Contributions: *Study concept and design:* Balk, Bonis, Moskowitz, Schmid, Ioannidis, Wang, Lau.

Acquisition of data: Balk, Bonis, Moskowitz, Ioannidis, Wang, Lau.

Analysis and interpretation of data: Balk, Bonis, Moskowitz, Schmid, Ioannidis, Wang, Lau.

Drafting of the manuscript: Balk, Bonis, Moskowitz, Schmid, Ioannidis, Wang, Lau.

Critical revision of the manuscript for important intellectual content: Balk, Bonis, Moskowitz, Schmid, Ioannidis, Wang, Lau.

Statistical expertise: Balk, Moskowitz, Schmid, Ioannidis, Lau.

Obtained funding: Lau.

Administrative, technical, or material support: Balk, Bonis, Ioannidis, Wang, Lau.

Study supervision: Balk, Bonis, Lau.

Funding/Support: This article was produced under Agency for Healthcare Research and Quality contract 290-97-0019. Additional support included National Research Service Award training grant T32 HS00060.

Acknowledgment: We thank Bonnie MacLeod, BS, and Caroline McFadden, MD, for assistance with translation, data extraction, and organizational assistance.

REFERENCES

- Chalmers TC, Celano P, Sacks HS, Smith H, Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-1361.
- Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. *J Clin Epidemiol.* 1999;52:631-636.
- Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282:1054-1060.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408-412.
- Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* 1998;352:609-613.
- Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and

- small randomized trials in meta-analyses. *Ann Intern Med*. 2001;135:982-989.
7. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials*. 1995;16:62-73.
 8. Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt A. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol*. 2001;54:651-654.
 9. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *J Clin Epidemiol*. 1995;48:167-171.
 10. Mulrow CD, Oxman AD. *Cochrane Collaborative Handbook 4*. Oxford, England: Cochrane Library, Update Software; 1998.
 11. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 2001;323:42-46.
 12. Berlin JA, Rennie D. Measuring the quality of trials: the quality of quality scales. *JAMA*. 1999;282:1083-1085.
 13. Gøtzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet*. 2000;355:129-134.
 14. de Koning HJ. Assessment of nationwide cancer-screening programmes. *Lancet*. 2000;355:80-81.
 15. Cates C, Senn S. Screening mammography re-evaluated. *Lancet*. 2000;355:750.
 16. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996;17:1-12.
 17. Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. *Control Clin Trials*. 1981;2:31-49.
 18. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA*. 1994;272:101-104.
 19. Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *Br J Surg*. 1985;72:256-260.
 20. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med*. 1989;8:441-454.
 21. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of non-steroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials*. 1989;10:31-56.
 22. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nurs Res*. 1991;40:352-355.
 23. Beckerman H, de Bie RA, Bouter LM, De Cuyper HJ, Oostendorp RA. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. *Phys Ther*. 1992;72:483-491.
 24. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol*. 1992;45:255-265.
 25. Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Ann Intern Med* 1994;121:11-21.
 26. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA*. 1994;272:125-128.
 27. Assendelft WJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA*. 1995;274:1942-1948.
 28. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA*. 1996;276:637-639.
 29. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ*. 1996;312:742-744.
 30. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol*. 1998;51:1235-1241.
 31. Clark HD, Wells GA, Huet C, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials*. 1999;20:448-452.
 32. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Br J Surg*. 2000;87:1448-1454.
 33. Moher D, Cook DJ, Jadad AR, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess*. 1999;3:i-iv.
 34. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med*. 1992;327:248-254.
 35. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127:820-826.
 36. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177-188.
 37. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. New York, NY: Chapman & Hall; 1995.
 38. Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ*. 1988;296:320-331.
 39. Hine KL, Laird NM, Hewitt P, Chalmers TC. Meta-analysis of empirical long-term antiarrhythmic therapy after myocardial infarction. *JAMA*. 1989;262:3037-3040.
 40. Leizorovicz A, Boissel JP. Oral anticoagulant in patients surviving myocardial infarction: a new approach to old data. *Eur J Clin Pharmacol*. 1983;24:333-336.
 41. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27:335-371.
 42. Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J*. 1985;6:556-585.
 43. Yusuf S, Collins R, MacMahon S, Peto R. Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomized trials. *Lancet*. 1988;1:1088-1092.
 44. Rossouw JE, Lewis B, Rifkind BM. The value of lowering cholesterol after myocardial infarction. *N Engl J Med*. 1990;323:1112-1119.
 45. Teo KK, Yusuf S. Role of magnesium in reducing mortality in acute myocardial infarction: a review of the evidence. *Drugs*. 1993;46:347-359.
 46. Colditz GA, Brewer TF, Berkey CS, et al. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *JAMA*. 1994;271:698-702.
 47. Jefferson TO, Demicheli V, Deeks JJ, Rivetti D. Amantadine and rimantadine for preventing and treating influenza A in adults. *Cochrane Database Syst Rev*. 2000;2:CD001169.
 48. Langley JM, LeBlanc JC, Drake J, Milner R. Efficacy of antimicrobial prophylaxis in placement of cerebrospinal fluid shunts: meta-analysis. *Clin Infect Dis*. 1993;17:98-103.
 49. McIntosh HM, Olliaro P. Artemisinin derivatives for treating severe malaria. *Cochrane Database Syst Rev*. 2000;2:CD000527.
 50. Small F. Antibiotics for asymptomatic bacteriuria in pregnancy. *Cochrane Database Syst Rev*. 2000;2:CD000490.
 51. Smieja MJ, Marchetti CA, Cook DJ, Small FM. Isoniazid for preventing tuberculosis in non-HIV infected persons. *Cochrane Database Syst Rev*. 2000;2:CD001363.
 52. Ausejo M, Saenz A, Pham B, et al. The effectiveness of glucocorticoids in treating group: meta-analysis. *BMJ*. 1999;319:595-600.
 53. Bhuta T, Ohlsson A. Systematic review and meta-analysis of early postnatal dexamethasone for prevention of chronic lung disease. *Arch Dis Child Fetal Neonatal Ed*. 1998;79:F26-F33.
 54. Kellner JD, Ohlsson A, Gadomski AM, Wang EE. Efficacy of bronchodilator therapy in bronchiolitis: a meta-analysis. *Arch Pediatr Adolesc Med*. 1996;150:1166-1172.
 55. Kozyrskyj AL, Hildes-Ripstein GE, Longstaffe SE, et al. Treatment of acute otitis media with a shortened course of antibiotics: a meta-analysis. *JAMA*. 1998;279:1736-1742.
 56. Rosenfeld RM, Post JC. Meta-analysis of antibiotics for the treatment of otitis media with effusion. *Otolaryngol Head Neck Surg*. 1992;106:378-386.
 57. Chan VM, Reznick RK, O'Rourke K, Kitchens JM, Lossing AG, Detsky AS. Meta-analysis of highly selective vagotomy versus truncal vagotomy and pyloroplasty in the surgical treatment of uncomplicated duodenal ulcer. *Can J Surg*. 1994;37:457-464.
 58. Chung RS, Rowland DY. Meta-analyses of randomized controlled trials of laparoscopic vs conventional inguinal hernia repairs. *Surg Endosc*. 1999;13:689-694.
 59. MacRae HM, McLeod RS. Handsewn vs stapled anastomoses in colon and rectal surgery: a meta-analysis. *Dis Colon Rectum*. 1998;41:180-189.
 60. Martin-Hirsch PL, Paraskevaidis E, Kitchener H. Surgery for cervical intraepithelial neoplasia. *Cochrane Database Syst Rev*. 2000;2:CD001318.
 61. Nelson RL. Meta-analysis of operative techniques for fissure-in-ano. *Dis Colon Rectum*. 1999;42:1424-1428.
 62. Pocock SJ, Henderson RA, Rickards AF, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet*. 1995;346:1184-1189.
 63. Sauerland S, Lefering R, Holthausen U, Neugebauer EA. Laparoscopic vs conventional appendectomy: a meta-analysis of randomised controlled trials. *Langenbecks Arch Surg*. 1998;383:289-295.
 64. Lachin JM. Statistical considerations in the intent-to-treat principle. *Control Clin Trials*. 2000;21:167-189.
 65. Ioannidis JP, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet*. 1998;352:590-591.
 66. Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285:1987-1991.
 67. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351:123-127.
 68. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279:281-286.
 69. Ioannidis JP, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative meta-analyses. *Proc Natl Acad Sci U S A*. 2001;98:831-836.