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, Juni 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 individual 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 screening individual 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 screening

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.
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. Furthermore, the quality measures found to be associated with treatment

Table 1. Quality Measure Definitions

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study question well defined in introduction/methods</td>
<td>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.</td>
</tr>
<tr>
<td>Study question well defined anywhere in article</td>
<td>As above, but criteria could be met from any section of article.</td>
</tr>
<tr>
<td>Placebo control</td>
<td>Required term placebo or description of placebo (eg, saline).</td>
</tr>
<tr>
<td>Appropriate outcome studied</td>
<td>Were study outcomes appropriate based on study design, condition, and intervention studied?</td>
</tr>
<tr>
<td>Multicenter study</td>
<td>Did study include more than 1 site?</td>
</tr>
<tr>
<td>Study country</td>
<td>Study considered to be from United States or other “research country” if any of the investigators were based in that country. Analyzed 2 ways.*</td>
</tr>
<tr>
<td>Adequate selection criteria</td>
<td>Were inclusion and exclusion criteria clearly and completely reported?</td>
</tr>
<tr>
<td>Randomization methods described</td>
<td>Was any description given of how randomization (allocation among treatment arms) was accomplished, or did the article say only “randomized”?</td>
</tr>
<tr>
<td>Central randomization site</td>
<td>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.</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>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.</td>
</tr>
<tr>
<td>Patients blinded</td>
<td>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.</td>
</tr>
<tr>
<td>Caregivers blinded</td>
<td>Caregivers included physicians, nurses, and other health care practitioners in direct patient care or parents (or equivalent) of outpatient infants.</td>
</tr>
<tr>
<td>Outcome assessors blinded</td>
<td>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.</td>
</tr>
<tr>
<td>Data analysts blinded</td>
<td>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.</td>
</tr>
<tr>
<td>Double blinded</td>
<td>Were both patients and either caregiver or outcome assessor blinded?</td>
</tr>
<tr>
<td>Valid statistical methods</td>
<td>Were the statistical methods used considered valid and appropriate, based on study design and outcomes of interest?</td>
</tr>
<tr>
<td>Statistician author or acknowledged</td>
<td>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.</td>
</tr>
<tr>
<td>Intention-to-treat analysis</td>
<td>Were 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).*</td>
</tr>
<tr>
<td>Power calculation reported</td>
<td>Was a power calculation reported for any outcome evaluated in the study?</td>
</tr>
<tr>
<td>Stopping rules described</td>
<td>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.)</td>
</tr>
<tr>
<td>Baseline characteristics reported</td>
<td>Were any baseline characteristics reported that compared the treatment and control groups?</td>
</tr>
<tr>
<td>Groups similar at baseline</td>
<td>Were the treatment and control groups similar in the characteristics reported?</td>
</tr>
<tr>
<td>Confounders accounted for</td>
<td>If there were baseline differences in the groups that could be confounders, were these examined?</td>
</tr>
<tr>
<td>Dropouts recorded</td>
<td>Were the number of dropouts recorded (either explicitly or by reporting the number enrolled and the number evaluated)?</td>
</tr>
<tr>
<td>Percentage dropouts</td>
<td>What percentage of subjects dropped out?</td>
</tr>
<tr>
<td>Reason for dropouts given</td>
<td>If there were dropouts, were the reasons for dropouts reported?</td>
</tr>
<tr>
<td>Findings support conclusions</td>
<td>Were the conclusions valid based on the findings, study design, and power?</td>
</tr>
</tbody>
</table>

*See Table 4.
QUALITY MEASURES AND TREATMENT EFFECT IN RCTS

Effect vary among investigators. Schu-  
zel et al10 reported that poorly concealed al-  
location or lack of double blinding re-  
sulted in a significant overestimation of  
treatment effect by 41% and 17%, re-  
spectively, in 250 studies of perinatal  
medicine. Moher et al11 reported a sim-  
lar 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 differ-  
ent quality measures with treatment  
effect and the absence of a gold-  
standard quality assessment instru-  
ment has resulted in a proliferation of  
quality scales used in meta-analyses.  
Juni et al3 identified 37 meta-analyses  
that used 26 different instruments to  
assess trial quality. The number of spe-  
cific quality measures in these scales  
ranged from 3 to 34, and the weights  
assigned to 3 common measures (ran-  
domization, blinding, and dropouts)  
ranged from 0% to 100%.

Adding to the uncertainty, quality is  
not consistently defined across spe-  
cialties, nor have specific quality mea-  
sures 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 previ-  
ously 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 medi-  
cal areas (cardiovascular disease, infec-  
tious 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 stud-  
ies 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 mea-  
sures previously demonstrated or hy-  
pothesized to be associated with esti-  
mates 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 ran-  
domized controlled trials and from re-  
ference lists of methodological articles.  
We used the definitions for each qual-  
ity measure as described by authors. For  
quality measures not clearly de-  
scribed, we reached consensus on defi-  
nitions. We aimed to establish defini-  
tions of study quality that could be  
ished most consistently across a va-  
riety of study types. Thus, we formal-  
ized a process that all researchers grad-  
ing the quality of studies would have  
to perform. Definitions of quality mea-  
sures 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 dis-  
eease, pediatrics, and surgery) because  
they represent a variety of medical ar-  
as. 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 dichoto-  
mous outcomes, and demonstrated  
significant between-study heterogene-  
ity in the OR scale (P<.10 for the χ² sta-  
tistic or a nonzero between-study vari-  
ance, r² by the DerSimonian and Laird  
random-effects model).35,36 We re-  
quired statistical heterogeneity of treat-  
ment effect across trials within each  
meta-analysis because meta-analyses  
with homogenous treatment effects  
across trials are unlikely to find that es-  
timates of treatment effects are associ-  
ated with quality measures (or other  
factors). We excluded abstracts, letters,  
available articles, and those for which  
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. Juni 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 special-  
ies, 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 treat-  
ment 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 medi-  
cal areas (cardiovascular disease, infec-  
tious 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 stud-  
ies 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 treat-  
ment 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 win-  
domized controlled trials and from ref-  
lence lists of methodological articles.  
We used the definitions for each qual-  
ity measure as described by authors. For  
quality measures not clearly defined,  
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 mea-  
sures 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 dis-  
eease, pediatrics, and surgery) because  
they represent a variety of medical ar-  
as. 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 dichoto-  
mous outcomes, and demonstrated  
significant between-study heterogene-  
ity in the OR scale (P<.10 for the χ² sta-  
tistic or a nonzero between-study vari-  
ance, r² by the DerSimonian and Laird  
random-effects model).35,36 We re-  
quired statistical heterogeneity of treat-  
ment effect across trials within each  
meta-analysis because meta-analyses  
with homogenous treatment effects  
across trials are unlikely to find that es-  
timates of treatment effects are associ-  
ated with quality measures (or other  
factors). We excluded abstracts, letters,  
available articles, and those for which  
detailed outcomes data were not pro-  
vided. Meta-analyses were selected with-  
out 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 out-  
come used was mortality. For studies in  
the other clinical areas, the outcome  
used varied across meta-analyses. Within  
meta-analyses, only outcomes with het-  
erogeneous treatment effects were con-  
sidered. 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 pro-  
cess. First, 4 clinicians (E.M.B., P.A.L.B.,  
H.M., and C.W.) trained in clinical epide-  
miology 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 inves-  
tigator extract data from a different  
pi lot set of 8 studies. Further refine-  
ments and clarifications were performed  
in the data extraction definitions of spe-  
cific quality measures. Third, the 4 in-  
vestigators independently extracted data  
from the remaining English-language  
RCTs. Data from each trial were ex-  
tracted 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 extractors. 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>
<thead>
<tr>
<th>Table 2. Descriptions and Summary Odds Ratios and Heterogeneity of Meta-analyses Examined</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meta-analysis Category</strong></td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Cardiovascular Disease</td>
</tr>
<tr>
<td>Antiplatelet Trialists'</td>
</tr>
<tr>
<td>Collaboration,38 1988</td>
</tr>
<tr>
<td>Hine et al,39 1989‡</td>
</tr>
<tr>
<td>Leizorovicz and Boissel,40</td>
</tr>
<tr>
<td>Yusuf et al,41 1985†</td>
</tr>
<tr>
<td>Yusuf et al,42 1985†</td>
</tr>
<tr>
<td>Yusuf et al,43 1988‡</td>
</tr>
<tr>
<td>Rossouw et al,44 1990</td>
</tr>
<tr>
<td>Teo and Yusuf,45 1993</td>
</tr>
<tr>
<td>Infectious Disease</td>
</tr>
<tr>
<td>Colditz et al,46 1994</td>
</tr>
<tr>
<td>Jefferson et al,47 2000</td>
</tr>
<tr>
<td>Preoperative antibiotics</td>
</tr>
<tr>
<td>McIntosh and Olliaro,48 2000</td>
</tr>
<tr>
<td>Small,49 2000</td>
</tr>
<tr>
<td>Smieja et al,51 2000</td>
</tr>
<tr>
<td>Pediatrics</td>
</tr>
<tr>
<td>Ausejo et al,52 1999</td>
</tr>
<tr>
<td>Bhuta and Orhsson,53 1998</td>
</tr>
<tr>
<td>Kelner et al,54 1996</td>
</tr>
<tr>
<td>Kozyrskyj et al,55 1998</td>
</tr>
<tr>
<td>Rosenfeld and Post,56 1992</td>
</tr>
<tr>
<td>Surgery</td>
</tr>
<tr>
<td>Chan et al,57 1994</td>
</tr>
<tr>
<td>Chung and Rowland,58 1999</td>
</tr>
<tr>
<td>MacRae and McLeod,59 1998</td>
</tr>
<tr>
<td>Martin-Hirsch et al,60 2000</td>
</tr>
<tr>
<td>Nelson,61 1999</td>
</tr>
<tr>
<td>Pocock et al,62 1995</td>
</tr>
<tr>
<td>Sauerland et al,63 1998</td>
</tr>
</tbody>
</table>

*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, r² > .36. ‡Updated by Lau et al.34

©2002 American Medical Association. All rights reserved.
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. 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 of the treatment and control groups. The log odds in the control group and the relative OR, defined as the difference in log odds between the treatment and control groups, was assumed to be normally distributed, with mean $\alpha_i + \beta_j \times x_{ij}$, where $x_{ij}$ is the quality measure in the $i$th study of the $j$th meta-analysis. For a dichotomous quality measure, $\beta_j$ represented the relative log OR between the 2 levels of the measure. The exponential of $\beta_j$ 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 $\alpha_i$ and $\beta_j$ varied according to a normal distribution with common variances $\tau_0^2$ and $\tau_0^2$, respectively. The other model assumed different $\alpha_i$ and $\beta_j$ by medical area so that there were 4 separate population intercepts and slopes corresponding to the cardiovascular disease, infectious disease, pediatric, and surgical areas. Noninformative 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, Medical 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, 6 infectious disease, 5 pediatric, and 7 surgical meta-analyses. 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

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

*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 reconciliation 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 highest agreement, ranging from 96% 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.94

**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***

<table>
<thead>
<tr>
<th>Quality Measures</th>
<th>Overall (N = 276)</th>
<th>Cardiovascular Disease (n = 93)</th>
<th>Infectious Disease (n = 56)</th>
<th>Pediatrics (n = 60)</th>
<th>Surgery (n = 67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study question well defined in introduction or methods</td>
<td>0.85 (0.64-1.06)</td>
<td>0.98 (0.68-1.30)</td>
<td>0.58 (0.32-1.80)</td>
<td>NA</td>
<td>0.81 (0.42-1.40)</td>
</tr>
<tr>
<td>Study question well defined anywhere in paper</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Placebo control</td>
<td>0.85 (0.61-1.09)</td>
<td>1.03 (0.84-1.27)</td>
<td>0.62 (0.40-0.91)†</td>
<td>0.40 (0.22-0.86)†</td>
<td>NA</td>
</tr>
<tr>
<td>Appropriate outcome studied</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Multicenter study</td>
<td>1.06 (0.90-1.25)</td>
<td>1.30 (0.87-1.94)</td>
<td>0.96 (0.68-1.40)</td>
<td>1.74 (1.09-2.80)†</td>
<td>0.71 (0.46-0.94)†</td>
</tr>
<tr>
<td>Study country, United States</td>
<td>1.05 (0.93-1.19)</td>
<td>1.12 (0.94-1.38)</td>
<td>0.87 (0.55-1.39)</td>
<td>1.02 (0.55-1.58)</td>
<td>0.84 (0.49-1.38)</td>
</tr>
<tr>
<td>Research country‡</td>
<td>0.95 (0.70-1.29)</td>
<td>NA</td>
<td>0.62 (0.44-0.92)‡</td>
<td>NA</td>
<td>0.99 (0.61-1.59)</td>
</tr>
<tr>
<td>Adequate selection criteria</td>
<td>0.94 (0.73-1.28)</td>
<td>NA</td>
<td>1.73 (0.81-5.49)</td>
<td>NA</td>
<td>0.65 (0.45-1.11)</td>
</tr>
<tr>
<td>Randomization methodology described</td>
<td>1.03 (0.89-1.20)</td>
<td>1.14 (0.95-1.40)</td>
<td>1.13 (0.73-1.67)</td>
<td>1.00 (0.63-1.46)</td>
<td>0.76 (0.53-1.08)</td>
</tr>
<tr>
<td>Central randomization site</td>
<td>1.01 (0.85-1.18)</td>
<td>1.14 (0.91-1.49)</td>
<td>0.93 (0.58-1.64)</td>
<td>0.88 (0.49-1.51)</td>
<td>NA</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>1.05 (0.91-1.21)</td>
<td>1.14 (0.96-1.42)</td>
<td>0.97 (0.68-1.42)</td>
<td>0.90 (0.58-1.28)</td>
<td>0.73 (0.36-1.24)</td>
</tr>
<tr>
<td>Patients blinded</td>
<td>0.95 (0.70-1.13)</td>
<td>1.08 (0.86-1.38)</td>
<td>0.70 (0.46-1.11)</td>
<td>0.79 (0.39-1.19)</td>
<td>NA</td>
</tr>
<tr>
<td>Caregivers blinded</td>
<td>0.98 (0.75-1.20)</td>
<td>1.09 (0.91-1.29)</td>
<td>0.62 (0.43-0.91)†</td>
<td>1.13 (0.73-1.84)</td>
<td>NA</td>
</tr>
<tr>
<td>Outcome assessors blinded</td>
<td>1.02 (0.82-1.22)</td>
<td>1.11 (0.87-1.39)</td>
<td>0.84 (0.55-1.27)</td>
<td>1.02 (0.57-1.61)</td>
<td>0.87 (0.56-1.36)</td>
</tr>
<tr>
<td>Data analysts blinded</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Double blinded‡</td>
<td>1.02 (0.79-1.24)</td>
<td>1.10 (0.90-1.33)</td>
<td>0.71 (0.47-1.12)</td>
<td>1.05 (0.56-1.61)</td>
<td>NA</td>
</tr>
<tr>
<td>Valid statistical methods</td>
<td>1.11 (0.95-1.31)</td>
<td>1.03 (0.81-1.33)</td>
<td>1.17 (0.78-1.77)</td>
<td>0.97 (0.48-1.73)</td>
<td>1.63 (1.03-2.83)†</td>
</tr>
<tr>
<td>Statistician author or acknowledged</td>
<td>1.04 (0.92-1.17)</td>
<td>1.05 (0.86-1.29)</td>
<td>0.85 (0.59-1.28)</td>
<td>1.12 (0.81-1.60)</td>
<td>1.13 (0.73-1.67)</td>
</tr>
<tr>
<td>Intention-to-treat analysis</td>
<td>0.91 (0.70-1.13)</td>
<td>0.89 (0.60-1.25)</td>
<td>0.80 (0.42-1.37)</td>
<td>NA</td>
<td>1.14 (0.73-2.05)</td>
</tr>
<tr>
<td>Power calculation reported</td>
<td>1.08 (0.95-1.23)</td>
<td>1.04 (0.89-1.22)</td>
<td>1.32 (0.88-1.96)</td>
<td>1.13 (0.76-1.62)</td>
<td>0.91 (0.55-1.43)</td>
</tr>
<tr>
<td>Stopping rules described</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Baseline characteristics reported</td>
<td>1.00 (0.68-1.52)</td>
<td>1.19 (0.69-1.96)</td>
<td>1.18 (0.66-2.21)</td>
<td>0.66 (0.35-1.29)</td>
<td>NA</td>
</tr>
<tr>
<td>Groups similar at baseline</td>
<td>1.06 (0.90-1.24)</td>
<td>1.03 (0.73-1.33)</td>
<td>1.15 (0.82-1.71)</td>
<td>1.13 (0.73-1.66)</td>
<td>0.77 (0.51-1.30)</td>
</tr>
<tr>
<td>Confounders accounted for</td>
<td>0.96 (0.79-1.23)</td>
<td>NA</td>
<td>0.94 (0.60-1.49)</td>
<td>0.96 (0.50-1.65)</td>
<td>1.20 (0.76-1.72)</td>
</tr>
<tr>
<td>Dropout recorded‡</td>
<td>1.26 (0.87-2.05)</td>
<td>NA</td>
<td>1.11 (0.62-1.91)</td>
<td>NA</td>
<td>1.12 (0.54-2.33)</td>
</tr>
<tr>
<td>Reason for dropouts given§</td>
<td>0.93 (0.77-1.13)</td>
<td>0.94 (0.75-1.17)</td>
<td>0.94 (0.55-1.52)</td>
<td>0.70 (0.43-1.16)</td>
<td>§</td>
</tr>
<tr>
<td>Percentage of dropouts‡</td>
<td>0.12 (0.94-1.12)</td>
<td>1.00 (0.89-1.14)</td>
<td>1.07 (0.84-1.34)</td>
<td>1.17 (0.92-1.57)</td>
<td>0.13 (0.86-1.27)</td>
</tr>
<tr>
<td>Findings support conclusions</td>
<td>0.83 (0.66-1.10)</td>
<td>NA</td>
<td>0.79 (0.57-1.11)</td>
<td>NA</td>
<td>0.71 (0.30-1.46)</td>
</tr>
</tbody>
</table>

*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.

©2002 American Medical Association. All rights reserved.
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. 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
conclusions on a relatively small number of RCTs. An exception is the study by Moher et al, which also included multiple meta-analyses from various clinical categories. An association was found between treatment effect and both Jadad score and adequacy of allocation concealment. Although the associations were statistically significant, the differences were small. 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 analyses are robustly constructed and validated, in- terpretation of the meaning of certain 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 approach 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. 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.
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.
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 a grant for Healthcare Research and Quality contract 290-97-0019. Additional support included National Research Service Award training grant T32 HD00060.
Acknowledgment: We thank Bonnie MacLeod, BS, and Caroline McFadden, MD, for assistance with translation, data extraction, and organizational assistance.
QUALITY MEASURES AND TREATMENT EFFECT IN RCTS

46. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998;279:281- 285.

©2002 American Medical Association. All rights reserved.