

The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis

Peter Jüni, MD

Anne Witschi, MD

Ralph Bloch, MD, PhD

Matthias Egger, MD, MSc

ALTHOUGH RANDOMIZED CONTROLLED trials provide the best evidence of the efficacy of medical interventions, they are not immune to bias. Studies relating methodological features of trials to their results have shown that trial quality influences effect sizes. For populations of trials examining treatments in myocardial infarction,¹ perinatal medicine,² and various disease areas,³ it has consistently been shown that inadequate concealment of treatment allocation, resulting, for example, from the use of open random-number tables, is associated on average with larger treatment effects. One of these studies² also found larger average effect sizes if trials were not double-blind. Analyses of individual trials suggest that in some instances effect sizes are also overestimated if some participants, for example, those not adhering to study medications, were excluded from the analysis.⁴⁻⁶ Informal qualitative research has indicated that investigators sometimes undermine the random allocation of study participants, for example, by opening assignment envelopes or holding translucent envelopes up to a light bulb.⁷ In response to this situation, guidelines on the conduct and reporting of clinical trials and scales to measure the quality of published trials have been developed.^{8,9}

The quality of trials is of obvious relevance to meta-analysis. If the raw ma-

Context Although it is widely recommended that clinical trials undergo some type of quality review, the number and variety of quality assessment scales that exist make it unclear how to achieve the best assessment.

Objective To determine whether the type of quality assessment scale used affects the conclusions of meta-analytic studies.

Design and Setting Meta-analysis of 17 trials comparing low-molecular-weight heparin (LMWH) with standard heparin for prevention of postoperative thrombosis using 25 different scales to identify high-quality trials. The association between treatment effect and summary scores and the association with 3 key domains (concealment of treatment allocation, blinding of outcome assessment, and handling of withdrawals) were examined in regression models.

Main Outcome Measure Pooled relative risks of deep vein thrombosis with LMWH vs standard heparin in high-quality vs low-quality trials as determined by 25 quality scales.

Results Pooled relative risks from high-quality trials ranged from 0.63 (95% confidence interval [CI], 0.44-0.90) to 0.90 (95% CI, 0.67-1.21) vs 0.52 (95% CI, 0.24-1.09) to 1.13 (95% CI, 0.70-1.82) for low-quality trials. For 6 scales, relative risks of high-quality trials were close to unity, indicating that LMWH was not significantly superior to standard heparin, whereas low-quality trials showed better protection with LMWH ($P < .05$). Seven scales showed the opposite: high quality trials showed an effect whereas low quality trials did not. For the remaining 12 scales, effect estimates were similar in the 2 quality strata. In regression analysis, summary quality scores were not significantly associated with treatment effects. There was no significant association of treatment effects with allocation concealment and handling of withdrawals. Open outcome assessment, however, influenced effect size with the effect of LMWH, on average, being exaggerated by 35% (95% CI, 1%-57%; $P = .046$).

Conclusions Our data indicate that the use of summary scores to identify trials of high quality is problematic. Relevant methodological aspects should be assessed individually and their influence on effect sizes explored.

JAMA. 1999;282:1054-1060

www.jama.com

terial used is flawed, then the conclusions of meta-analytic studies will be equally invalid. Following the recommendations of the Cochrane Collaboration and other experts in the field,¹⁰⁻¹² many meta-analysts assess the quality of trials and exclude trials of low methodological quality in sensitivity analyses. In a meta-analysis of trials comparing low-molecular-weight heparin (LMWH) with standard heparin for thromboprophylaxis in general surgery, Nurmohamed et al¹³ found a significant reduction of 21% in the risk of

deep vein thrombosis (DVT) with LMWH ($P = .012$). However, when the analysis was limited to trials with strong methods, as assessed by a scale consisting of 8 criteria, no significant difference between the 2 heparins re-

Author Affiliations: Clinical Epidemiology Study Group, Berne, Switzerland (Drs Jüni and Witschi), Institute for Medical Education, University of Berne (Drs Bloch and Jüni); and MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, Bristol, England (Dr Egger).

Corresponding Author and Reprints: Matthias Egger, MD, MSc, Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol, BS8 2PR England (e-mail: m.egger@bristol.ac.uk).

See also pp 1061 and 1083.

mained (relative risk [RR] reduction, 9%; $P = .38$). The authors therefore concluded that “there is at present no convincing evidence that in general surgery patients LMWHs, compared with standard heparin, generate a clinically important improvement in the benefit to risk ratio.”¹³ In contrast, another group of meta-analysts did not consider the quality of trials and concluded that “LMWHs seem to have a higher benefit to risk ratio than unfractionated heparin in preventing perioperative thrombosis.”¹⁴

Although widely recommended, the method of assessing and incorporating the quality of clinical trials is a matter of ongoing debate.¹⁵ This is reflected by the plethora of available instruments. In a search covering the years up to 1993, Moher et al⁹ identified 25 different quality assessment scales. Most of these scoring systems lack a focused theoretical basis and their objectives are unclear. The scales differ considerably in terms of dimensions covered, size, and complexity, and the weight assigned to the key domains most relevant to the control of bias (randomization, blinding, and withdrawals)¹⁶ varies widely (TABLE 1).

We repeated the meta-analysis of Nurmohamed et al¹³ using different scales and thus examined whether the type of scale used for assessing the quality of trials affects the conclusions of meta-analytic studies.

METHODS

Scales Used to Assess Trial Quality

We used the 25 scales described by Moher et al.⁹ When necessary, we adapted items that were developed for specific situations. For example, 1 item in a scale developed to assess trials of corticosteroids in alcoholic hepatitis²⁸ considered the similarity of the prognostic variables total bilirubin, prothrombin time, and hepatic encephalopathy at baseline. We included this item, but considered the variables that Nurmohamed et al¹³ identified as of prognostic importance (for example, the sex of patients, the duration of operation, and the presence of malignancies). If instructions for

use of a quality assessment instrument were unclear, we obtained additional published information or contacted the authors. Ambiguities remained for some scales and we developed our own a priori rules to deal with these situations.

Assessment of Trials

All 17 general surgery trials comparing LMWH with standard heparin that were included in the original meta-analysis were assessed with each of the 25 scales. To maintain comparability with the original meta-analysis,¹³ we included a single-center report⁴⁰ from a multicenter study,⁴¹ which had erroneously been included in addition to the main report.⁴² Information on authors, author affiliation, study centers, drugs, results, conclusions, source, year of publication, references, funding, and acknowledgments was concealed by a person unrelated to the study, using a

black marker and subsequent photocopying. A few items made unblinding of some of this information (in most cases the results section) necessary. Reports were independently assessed by 2 of the authors (P.J., A.W.). One assessor rated trials in the opposite order, also reversing the sequence of scales. Interobserver reliability was determined for each scale using the intraclass correlation coefficient.⁴³ Disagreements were resolved by consensus.

Data Analysis

We repeated the original meta-analysis using each of the 25 quality assessment scales, keeping all other aspects constant. The same fixed-effects model was used for combining trials, with data being provided by the authors of the original meta-analysis.¹³ Effect estimates were weighted according to the inverse of their variance. The primary end point

Table 1. Characteristics of 25 Scales for Quality Assessment of Clinical Trials

Scale	No. of Items	Weight Given to Methodological Key Domains, %*		
		Randomization	Blinding	Withdrawals
Andrew, ¹⁷ 1984	11	9.1	9.1	9.1
Beckerman et al, ¹⁸ 1992	24	4.0	12.0	16.0
Brown, ¹⁹ 1991	6	14.3	4.8	0
Chalmers et al, ²⁰ 1990	3	33.3	33.3	33.3
Chalmers et al, ²¹ 1981	30	13.0	26.0	7.0
Cho and Bero, ²² 1994	24	14.3	8.2	8.2
Colditz et al, ²³ 1989	7	28.6	0	14.3
Detsky et al, ²⁴ 1992	14	20.0	6.7	0
Evans and Pollock, ²⁵ 1985	33	3.0	4.0	11.0
Goodman et al, ²⁶ 1994	34	2.9	2.9	5.9
Gøtzsche, ²⁷ 1989	16	6.3	12.5	12.5
Imperiale and McCullough, ²⁸ 1990	5	0	0	0
Jadad et al, ²⁹ 1996	3	40.0	40.0	20.0
Jonas et al, 1993†	18	11.1	11.1	5.6
Kleijnen et al, ³⁰ 1991	7	20.0	20.0	0
Koes et al, ³¹ 1991	17	4.0	20.0	12.0
Levine, ³² 1991	29	2.5	2.5	3.1
Linde et al, ³³ 1997	7	28.6	28.6	28.6
Nurmohamed et al, ¹³ 1992	8	12.5	12.5	12.5
Onghena and Van Houdenhove, ³⁴ 1992	10	5.0	10.0	5.0
Poynard, ³⁵ 1988	14	7.7	23.1	15.4
Reisch et al, ³⁶ 1989	34	5.9	5.9	2.9
Smith et al, ³⁷ 1992	8	0	25.0	12.5
Spitzer et al, ³⁸ 1990	32	3.1	3.1	9.4
ter Riet et al, ³⁹ 1990	18	12.0	15.0	5.0

*Weight of methodological domains most relevant to the control of bias, expressed as percentage of maximum scores.

†Unpublished.

was DVT and major bleeding was the secondary end point. We performed stratified analyses dividing trials into high-quality and low-quality groups, using the definitions given by the authors of the scales. If no definitions were available from authors, we considered trials with scores above the median as high quality. We also performed analyses using the median as the cutoff point for all scales. To determine whether the restriction of scales to domains most relevant to the control of bias (randomization, blinding, and withdrawals)¹⁶ affects results, we deleted all items that were unrelated to these domains, recalculated scores, and repeated stratified analyses using the median as the cutoff point.

Meta-regression analyses were performed to examine the association of global quality scores with estimated ef-

fects on the risk of DVT. The random-effects regression model, described in detail elsewhere,⁴⁴ relates the treatment effect to study quality, assuming a normal distribution for the residual errors with both a within-study and an additive between-studies component of variance. The between-studies variance was estimated by an iterative procedure, using an estimate that is based on a restricted maximum likelihood method.

We standardized scores by subtracting the median from individual values and dividing the result by the interquartile range. All scores had a standardized distribution with a median of 0 and an interquartile range of 1 and regression coefficients were therefore comparable. For each scale we calculated the expected RR for hypothetical trials with the highest and lowest pos-

sible score. As a measure of overall agreement between the 25 scales, we calculated the intraclass correlation coefficient for the standardized scores.⁴³ Finally, we examined the separate influence of the 3 key domains that have been shown to be associated empirically with bias: concealment of treatment allocation,¹⁻³ blinding of outcome assessments,^{2,45} and handling of dropouts and withdrawals in the analysis.⁴⁻⁶ Finally, we performed sensitivity analyses using a random-effects model⁴⁶ for combining trials, inspecting funnel plots, and testing for the presence of publication bias.⁴⁷

We used Meta-Analyst software (Joseph Lau, Boston, Mass) for fixed-effects meta-analysis and the program Metareg⁴⁸ in Stata (Stata Corporation, College Station, Tex) for meta-regression analysis. Results are given as RRs with 95% confidence intervals (CIs). All *P* values are 2-sided.

Table 2. Median Scores From 25 Quality Assessment Scales for 17 Trials Comparing Heparins for Thromboprophylaxis in General Surgery, and Thresholds for Definition of High Quality*

Scale	Median Score (Range), %	Threshold for High Quality, %†
Poynard, ³⁵ 1988	38.5 (15.4-76.9)	50.0
Chalmers et al, ²¹ 1981	39.8 (8.6-76.8)	NA
Spitzer et al, ³⁸ 1990	48.1 (25.9-78.8)	NA
Beckerman et al, ¹⁸ 1992	50.0 (25.0-75.0)	52.0
Linde et al, ³³ 1997	50.0 (14.3-92.9)	71.4
Chalmers et al, ²⁰ 1990	55.6 (11.1-88.9)	66.7
Cho and Bero, ²² 1994	55.6 (37.8-75.6)	NA
Detsky et al, ²⁴ 1992	56.7 (23.3-89.3)	NA
Colditz et al, ²³ 1989	57.1 (14.3-85.7)	NA
Götzsche, ²⁷ 1989	57.1 (7.1-71.4)	NA
Smith et al, ³⁷ 1992	57.1 (25.7-85.7)	50.0
Jonas et al, 1993‡	58.3 (33.3-88.9)	76.0
Imperiale and McCullough, ²⁸ 1990	60.0 (20.0-100)	80.0
Jadad et al, ²⁹ 1996	60.0 (0-100)	60.0
Koes et al, ³¹ 1991	60.0 (20.0-78.6)	50.0
Reisch et al, ³⁶ 1989	62.5 (37.5-87.5)	NA
Onghena and Van Houdenhove, ³⁴ 1992	62.9 (34.3-100)	NA
Evans and Pollock, ²⁵ 1985	63.8 (32.5-88.2)	NA
Levine, ³² 1991	64.4 (26.8-79.5)	60.0
Goodman et al, ²⁶ 1994	67.7 (31.0-83.2)	60.0
Kleijnen et al, ³⁰ 1991	70.0 (30.0-98.0)	55.0
Nurmohamed et al, ¹³ 1992	75.0 (25.0-100)	87.5
Andrew, ¹⁷ 1984	77.3 (45.5-90.9)	72.7
Brown, ¹⁹ 1991	81.0 (52.4-95.2)	81.0
ter Riet et al, ³⁹ 1990	82.9 (48.6-91.4)	50.0

*Scores and thresholds are expressed as percentage of the maximum score. Scales are arranged in increasing order of the median score.

†NA indicates data not available because thresholds were not provided.

‡Unpublished.

RESULTS

Trial Quality

Interrater reliability was excellent for most scales. Intraclass correlation coefficients were above 0.9 for 12 scales (48%), 0.8 to 0.9 for 10 scales (40%), and less than 0.8 for 3 scales (12%). The median quality of the 17 trials as assessed by the 25 scales ranged from 38.5% to 82.9% of the maximum score (TABLE 2). The authors of 16 scales defined a threshold for high quality, with the median threshold corresponding to 60% of the maximum score. Agreement for standardized scores between the 25 scales was substantial (intraclass correlation coefficient, 0.72 [95% CI, 0.59-0.86]).

Analyses Stratified by Trial Quality

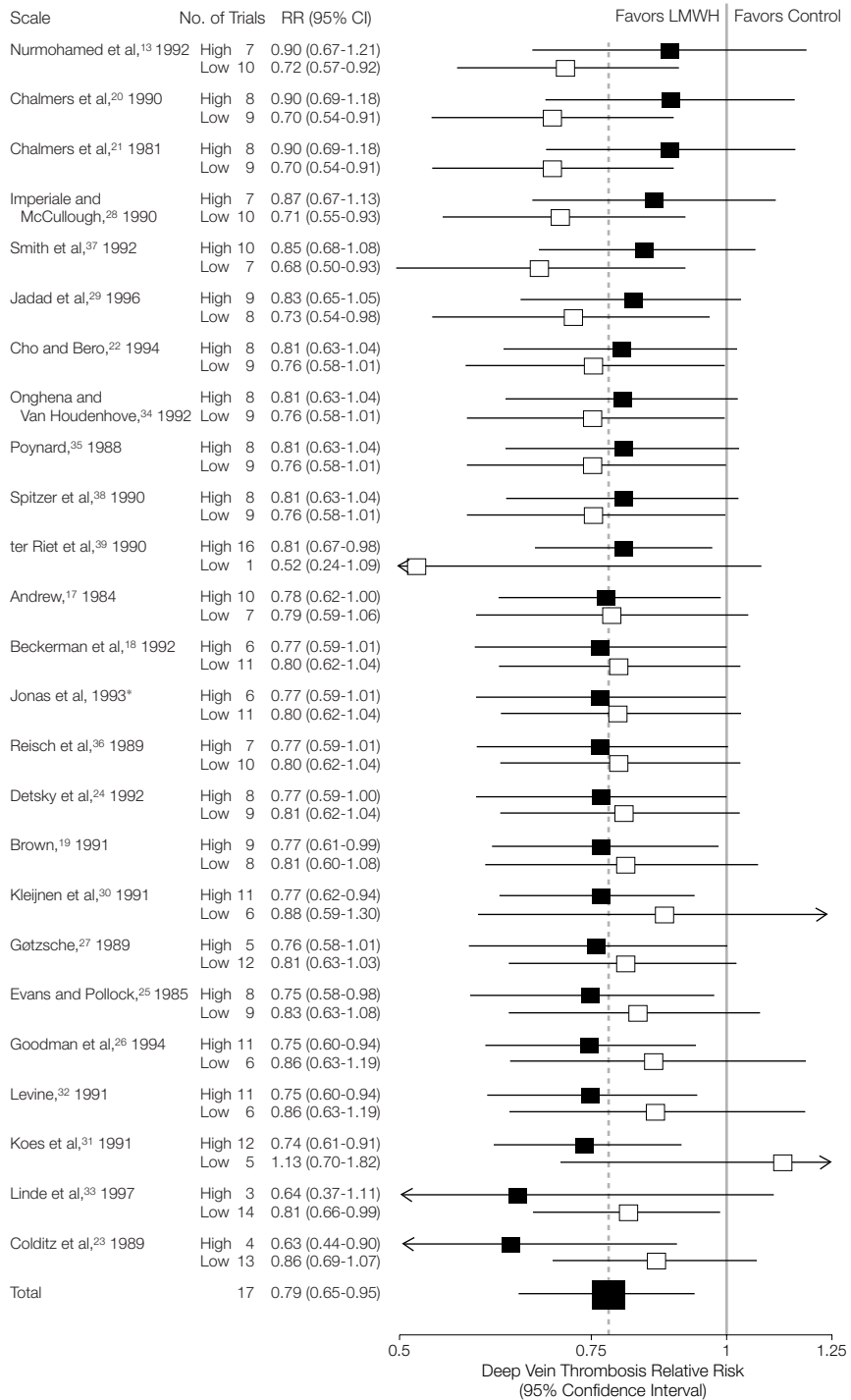
For all trials combined, the RR of DVT comparing LMWH with standard heparin was 0.79 (95% CI, 0.65-0.95) and thus identical to the results of the original analysis.¹³ FIGURE 1 shows the results of analyses stratified by quality using the 25 scales. Pooled RRs ranged from 0.63 to 0.90 for high-quality trials, and from 0.52 to 1.13 for low-quality trials. Six scales with pooled RRs

of high-quality trials more than 0.79 and CIs overlapping 1 indicated that LMWH was not significantly superior to standard heparin, whereas low-quality trials assessed by these scales showed significantly better protection with LMWH ($P < .05$). Seven scales showed the opposite: high-quality trials indicated that LMWH was beneficial ($P < .05$) with RRs of less than 0.79, whereas low-quality trials showed no significant difference. For the remaining 12 scales, pooled results from low-quality and high-quality strata indicated similar effects. Results were not materially altered when using the median score as the cutoff point for high-quality trials throughout. In meta-regression no significant difference in effect estimates between high-quality and low-quality trials was evident for any of the scales used ($P > .10$). Significant differences in the risk of major bleeding were not observed between the 2 heparins overall or when stratified by quality.

Summary Quality Scores and Effect Sizes

Meta-regression analysis confirmed the differences between scales that were observed in the stratified analysis. The coefficients (per point increase of standardized scores) ranged from -0.177 to 0.169, demonstrating that depending on the scale used, the effect size either increased or decreased with increasing trial quality. FIGURE 2 illustrates the relationship between RRs and quality scores for 3 scales.^{17,29,31} None of the 25 scales yielded a statistically significant association between summary scores and effect sizes. For hypothetical trials of maximum quality the RR of DVT comparing LMWH with standard heparin ranged from 0.57 to 0.95, whereas for hypothetical trials of minimum quality RRs ranged from 0.51 to 1.32. Heterogeneity between scales was reduced little when restricting scales to the domains most relevant to the control of bias (randomization, blinding, and withdrawals): RRs for DVT ranged from 0.63 to 0.98 for high-quality trials and from 0.68 to 0.89 for low-quality trials.

Figure 1. Results From Sensitivity Analyses Dividing Trials in High- and Low-Quality Strata, Using 25 Different Quality Assessment Scales



Relative risks (RRs) for deep vein thrombosis with 95% confidence intervals (CIs) are shown. LMWH indicates low-molecular-weight heparin. Black squares indicate estimates from high-quality trials and open squares indicate estimates from low-quality trials. Arrows indicate that the values are outside the range of the x axis. Broken line indicates combined estimate from all 17 trials. Solid line indicates null effect line. The scales are arranged in decreasing order of the RRs in trials deemed to be of high quality. Asterisk indicates unpublished scale.

Key Domains and Effect Sizes

The association between methodological key domains and estimates of treatment effects on the risk of DVT is explored in TABLE 3. There was no significant association with allocation concealment and handling of dropouts and withdrawals. Trials with open assessment of the outcome, however, overestimated treatment effects by 35% (95% CI, 1%-57%; $P = .046$). This association remained when all 3 key domains were included in a multivariate analysis ($P = .030$). Meta-analysis of the 6 trials

with open assessment of DVT suggested that LMWH was superior to standard heparin with an RR of 0.59 (95% CI, 0.42-0.84; $P = .004$). Conversely, the 11 trials with blinded outcome assessment showed no significant difference (RR, 0.89; 95% CI, 0.71-1.11; $P = .29$).

COMMENT

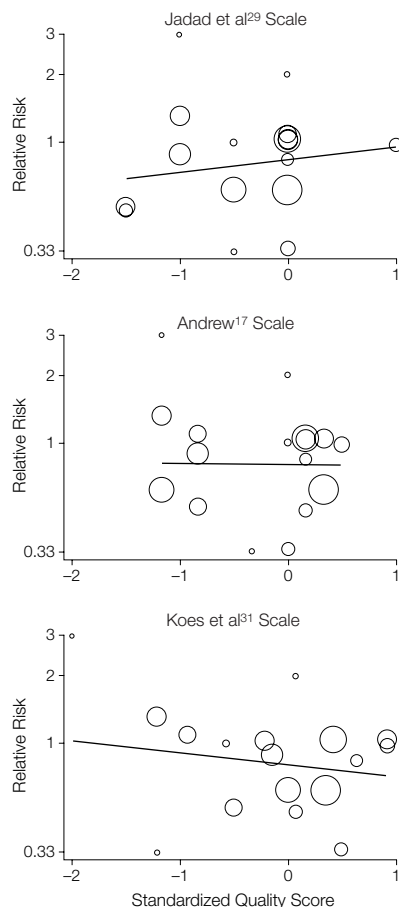
Meta-analysis is widely used to summarize the evidence on the benefits and risks of medical interventions. However, the findings of several meta-analyses of small trials have been contradicted subsequently by large controlled trials.^{47,49-51} The fallibility of meta-analysis is not surprising, considering the various biases that may be introduced by the process of locating and selecting studies, including publication bias,⁵² language bias,⁵³ and citation bias.⁵⁴ Low methodological quality of component studies is another potential source of systematic error. The critical appraisal of trial quality is therefore widely recommended and a large number of different instruments are currently in use. In a hand search of 5 general medicine journals dating 1993 to 1997 (*Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine*) we identified 37 meta-analyses using 26 different instruments to assess trial quality.

Our study shows that the type of scale used to assess trial quality can dramatically influence the interpretation of meta-analytic studies. Using 25 dif-

ferent scales, we reanalyzed a meta-analysis which, based on trials considered to be of high methodological quality, found little difference between LMWH and standard heparin in the prevention of postoperative thrombosis. Whereas for some scales these findings were confirmed, the use of others would have led to opposite conclusions, indicating that the beneficial effect of LMWH was particularly robust for trials deemed to be of high quality. Similarly, in meta-regression analysis effect size was negatively associated with some quality scores, but positively associated with others. Accordingly, RRs estimated for hypothetical trials of maximum or minimum quality varied widely between scales.

These discrepant results are not surprising when considering the heterogeneous nature of the instruments.¹⁵ Many scales include items that are more closely related to reporting quality, ethical issues, or to the interpretation of results rather than to the internal validity of trials. For example, some scales assessed whether the rationale for conducting the trial was clearly stated, whether the trialists' conclusions were compatible with the results obtained, or whether the report stated that participants provided written informed consent. Important differences also exist between scales that focus on internal validity. For example, the scale developed by Jadad et al,²⁹ which has been widely advocated,^{3,9,15} gives more weight

Figure 2. Examples of Weighted Regression Analyses of 17 Trials Comparing Low-Molecular-Weight Heparin With Standard Heparin



Relative risks (on logarithmic scale) are regressed against standardized scores from the scales described by Jadad et al²⁹ (top), Andrew¹⁷ (middle), and Koes et al³¹ (bottom). The size of the circle is proportional to the weighting factor (inverse of the variance).

Table 3. Results From Univariate Meta-Regression Analysis Relating Methodological Key Domains to Effect Sizes in 17 Trials Comparing Heparins for Thromboprophylaxis in General Surgery*

Methodological Domain	No. of Trials	Ratio of Relative Risks (95% CI)	P Value
Concealment of randomization			
Yes	6	1.00 (Referent)	.58
Unclear	11	1.12 (0.76-1.65)	
Blinding of outcome assessments			
Yes	11	1.00 (Referent)	.046
No	6	0.65 (0.43-0.99)	
Handling of dropouts and withdrawals			
Intention-to-treat analysis performed	7	1.00 (Referent)	.12
Intention-to-treat analysis not performed	10	1.37 (0.92-2.03)	

*CI indicates confidence interval. A ratio of relative risks of less than 1 indicates that methodologically inferior trials exaggerate the benefits of low-molecular-weight heparins compared with the referent group. A ratio of relative risks above 1 indicates the opposite.

to the quality of reporting than to actual methodological quality. A statement on withdrawals and dropouts will earn the point allocated to this domain, independently of whether the data were analyzed according to the intention-to-treat principle. The instrument addresses randomization but does not assess allocation concealment. The use of an open random-number table would thus be considered equivalent to concealed randomization using a telephone or computer system and earn the maximum points foreseen for randomization. Conversely, the scale developed by Chalmers et al²⁰ allocates 0 points for unconcealed but the maximum of 3 points for concealed randomization. The authors of the different scales clearly had different perceptions of trial quality, but definitions were rarely given, and the ability of the scales to measure what they are supposed to measure remains unclear.

Our study was based on a single meta-analysis and, strictly speaking, the results are only applicable to the 17 trials examined. It is unlikely, however, that agreement across scales would be better in other situations. Interestingly, in a recent review of treatment effects from trials deemed to be of high or low quality, Kunz and Oxman⁵⁵ found that in some meta-analyses there were no differences whereas in other meta-analyses high-quality trials showed either larger or smaller effects. In 1 analysis evaluating the effect of antiestrogen treatment in male infertility, the results were reversed with adverse effects on pregnancy rates in studies of high quality.⁵⁶ Different scales had been used for assessing quality and, in light of our study, it is possible that the choice of the scale contributed to the discrepant associations observed in these meta-analyses.

In our sample of trials we found that blinding of outcome assessment was the only factor significantly associated with effect size, with RRs on average being exaggerated by 35% if outcome assessment was open. When restricting the analysis to 11 trials with blinded outcome assessment, no significant difference between the 2 heparins was evi-

dent indicating that in general surgery patients, the 2 heparins may be equally effective. This was recently confirmed in an updated meta-analysis that included 25 double-blind trials.⁵⁷ The combined odds ratio for DVT was 0.99 (95% CI, 0.83-1.18). It is now generally agreed that the advantages of LMWH (reduced risk of heparin-induced thrombocytopenia⁵⁸ and convenience of once daily dosing) must be balanced against the greater cost of LMWH.⁵⁹

The importance of blinding could have been anticipated considering that the interpretation of the test (fibrinogen leg scanning) used to detect DVT can be subjective.⁶⁰ In other situations, blinding of outcome assessment may be irrelevant, such as when examining the effect of an intervention on overall mortality. In contrast to studies including large numbers of trials,¹⁻³ we did not find a significant association of concealment of treatment allocation with effect estimates. Our meta-analysis could have been too small to show this effect. Alternatively, concealment of treatment allocation may not have been relevant in the context of our study. The importance of allocation concealment may to some extent depend on whether strong beliefs exist among investigators regarding the benefits or risks of assigned treatments or whether equipoise of treatments is accepted by all investigators involved.⁷ Strong beliefs are probably more prevalent in trials comparing an intervention with placebo than in trials comparing 2 similar, active interventions. Other aspects may need to be considered in specific situations, such as whether the trial was terminated prematurely, the tests used for measuring end points, or the type of statistical model used. The importance of individual quality domains and, possibly, the direction of potential biases associated with these domains will thus vary according to the context, and the mechanistic application of scales with fixed weights allocated to a standard set of items may dilute, or entirely miss, associations.⁶¹ Indeed, in our sample of 17 trials none of the 25 composite scales was significantly associated with treatment effect. Many meta-analysts probably would have dis-

missed trial quality as a source of bias based on the analysis of summary scores and concluded that there was robust evidence favoring LMWH.

Although improved reporting practices should facilitate the assessment of methodological quality in the future, incomplete reporting continues to be an important problem when assessing trial quality.⁸ Because small single-center studies may be more likely to be of inadequate quality and more likely to be reported inadequately than large multicenter studies, the sample size and number of study centers may sometimes be useful proxy variables for study quality. Analyzing the effect of sample size will also shed light on the possible presence of publication bias.⁴⁷ One should not forget, however, that associations between domains of trial quality and treatment effects are subject to the potential biases of observational studies. Confounding could exist between measures of trial quality and other characteristics of trials, such as the setting, the characteristics of the participants, or the treatments.

The use of quality scores as weights when pooling studies has also been advocated, such as multiplying scores with the precision of effect estimates.³ Such procedures will affect both the combined effect estimate and its CI. As could be expected, we obtained different results depending on the scale used as the weighting factor when performing such analyses (data not shown). As pointed out previously by Detsky et al,²⁴ the incorporation of quality scores as weights lacks statistical or empirical justification.

CONCLUSIONS

The assessment of the methodological quality of randomized trials and the conduct of sensitivity analyses should be considered routine procedures in meta-analysis. Although composite quality scales may provide a useful overall assessment when comparing populations of trials, for example, trials published in different languages or disciplines, such scales should not generally be used to identify trials of apparent low quality or high quality in a given

meta-analysis.⁶¹ Rather, the relevant methodological aspects should be identified, ideally a priori, and assessed individually. This should always include the key domains of concealment of treatment allocation, blinding of outcome assessment or double blinding, and handling of withdrawals and dropouts. Finally, the lack of well-performed and adequately sized trials cannot be remedied by statistical analyses of small trials of questionable quality.

Acknowledgment: We are indebted to Bettina Lässer, RN, for obtaining and concealing primary trials, data entering, and checking; to Christoph Minder, PhD, for statistical advice; and to Michael Nurmohamed, MD, for providing unpublished data. We thank Erik Andrew, PhD, Sharon Brown, PhD, Mildred Cho, PhD, Allan Detsky, PhD, Steven Goodman, PhD, Peter Gøtzsche, MD, Wayne Jonas, MD, Jos Kleijnen, PhD, Bart Koes, PhD, Klaus Linde, MD, Patrick Onghena, PhD, Alan Pollock, ChB, Thierry Poynard, PhD, and Gerben ter Riet, PhD, for additional information on their scales. Finally, we are grateful to Iain Chalmers, MBBS, Peter Gøtzsche, MD, Nicola Low, MBBS, and 2 anonymous reviewers for helpful comments on an earlier draft of this article.

Author Contributions: Dr Jüni had main responsibility for the study protocol and performed statistical analyses and quality assessments. Dr Witschi participated in protocol development and assessed the quality of primary trials. Dr Bloch contributed to protocol development and supervised the study. Dr Egger reviewed the protocol and performed statistical analyses. Drs Jüni and Egger wrote the first draft of the article. All authors were involved in the writing of the final draft. Drs Jüni and Egger are the guarantors of this study.

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.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. *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.
- Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *N Engl J Med.* 1979;301:1410-1412.
- May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. *Circulation.* 1981;64:669-673.
- Peduzzi P, Wittes J, Detre K, Holford T. Analysis as-randomized and the problem of non-adherence. *Stat Med.* 1993;12:1185-1195.
- Schulz KF. Subverting randomization in controlled trials. *JAMA.* 1995;274:1456-1458.
- 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.
- Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials. *Control Clin Trials.* 1995;16:62-73.
- Mulrow CD, Oxman AD, eds. *Cochrane Collaboration handbook [Cochrane Review on CD-ROM]*. Oxford, England: Cochrane Library, Update Software; 1998: issue 4.
- 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.
- Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet.* 1998;351:47-52.
- Nurmohamed MT, Rosendaal FR, Buller HR, et al. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *Lancet.* 1992;340:152-156.
- Leizorovicz A, Haugh MC, Chapuis FR, Samama MM, Boissel JP. Low molecular weight heparin in prevention of perioperative thrombosis. *BMJ.* 1992;305:913-920.
- Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. *Int J Technol Assess Health Care.* 1996;12:195-208.
- Chalmers I. Applying overviews and meta-analyses at the bedside. *J Clin Epidemiol.* 1995;48:67-70.
- Andrew E. Method for assessment of the reporting standard of clinical trials with roentgen contrast media. *Acta Radiol Diagn (Stockh).* 1984;25:55-58.
- Beckerman H, de Bie RA, Bouter LM, de Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders. *Phys Ther.* 1992;72:483-491.
- Brown SA. Measurement of quality of primary studies for meta-analysis. *Nurs Res.* 1991;40:352-355.
- Chalmers I, Adams M, Dickersin K, et al. A cohort study of summary reports of controlled trials. *JAMA.* 1990;263:1401-1405.
- 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.
- Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA.* 1994;272:101-104.
- Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med.* 1989;8:441-454.
- Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol.* 1992;45:255-265.
- 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.
- Goodman SN, Berlin JA, 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.
- 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.
- Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? *Ann Intern Med.* 1990;113:299-307.
- Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials.* 1996;17:1-12.
- Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homoeopathy. *BMJ.* 1991;302:316-323.
- Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. *BMJ.* 1991;303:1298-1303.
- Levine J. Trial assessment procedure scale (TAPS). In: Spilker B, ed. *Guide to Clinical Trials*. New York, NY: Raven Press; 1991:780-786.
- Linde K, Clausius N, Ramirez G, et al. Are the clinical effects of homoeopathy placebo effects? *Lancet.* 1997;350:834-843.
- Onghena P, Van Houdenhove B. Antidepressant-induced analgesia in chronic non-malignant pain. *Pain.* 1992;49:205-219.
- Poynard T. Evaluation de la qualité méthodologique des essais thérapeutiques randomisés. *Presse Med.* 1988;17:315-318.
- Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics.* 1989;84:815-827.
- Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis.* 1992;145:533-539.
- Spitzer WO, Lawrence V, Dales R, et al. Links between passive smoking and disease. *Clin Invest Med.* 1990;13:17-42.
- ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *J Clin Epidemiol.* 1990;43:1191-1199.
- Verardi S, Cortese F, Baroni B, Boffo V, Casciani CU, Palazzini E. Deep vein thrombosis prevention in surgical patients. *Curr Ther Res.* 1989;46:366-372.
- Verardi S, Casciani CU, Nicora E, et al. A multicentre study on LMW-heparin effectiveness in preventing postsurgical thrombosis. *Int Angiol.* 1988;7:19-24.
- Leizorovicz A, Haugh M, Boissel JP. Meta-analysis and multiple publication of clinical trial reports. *Lancet.* 1992;340:1102-1103.
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420-428.
- Thompson SG, Sharp S. Explaining heterogeneity in meta-analysis. *Stat Med.* In press.
- Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology.* 1994;44:16-20.
- DerSimonian R, Laird N. Meta analysis in clinical trials. *Control Clin Trials.* 1986;7:177-188.
- Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* 1997;315:629-634.
- Sharp S. Meta-analysis regression. *Stata Techn Bull.* 1998;42:16-22.
- Villar J, Carroli G, Belizán JM. Predictive ability of meta-analyses of randomised controlled trials. *Lancet.* 1995;345:772-776.
- Cappelleri JC, Ioannidis JPA, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? *JAMA.* 1996;276:1332-1338.
- Le Lorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med.* 1997;337:536-542.
- Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet.* 1991;337:867-872.
- Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet.* 1997;350:326-329.
- Gøtzsche PC. Reference bias in reports of drug trials. *BMJ.* 1987;295:654-656.
- Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ.* 1998;317:1185-1190.
- Khan KS, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med.* 1996;156:661-666.
- Koch A, Bouges S, Ziegler S, Dinkel H, Daures JP, Victor N. Low molecular weight heparin and unfractionated heparin in thrombosis prophylaxis after major surgical intervention: update of previous meta-analyses. *Br J Surg.* 1997;84:750-759.
- Warkentin TE, Levine MN, Hirsh J, et al. Heparin-induced thrombocytopenia in patients treated with low-molecular-weight heparin or unfractionated heparin. *N Engl J Med.* 1995;332:1330-1335.
- Weitz JI. Low-molecular weight heparins. *N Engl J Med.* 1997;337:688-698.
- Lensing AW, Hirsh J. ¹²⁵I-fibrinogen leg scanning. *Thromb Haemost.* 1993;69:2-7.
- Greenland S. Quality scores are useless and potentially misleading [reply to re: a critical look at some popular analytic methods]. *Am J Epidemiol.* 1994;140:300-302.