Empirical Evidence of Associations Between Trial Quality and Effect Size
Methods Research Report

Empirical Evidence of Associations Between Trial Quality and Effect Size

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
http://www.ahrq.gov

Contract No. HHSA 290-2007-10062-I

Prepared by:
Southern California Evidence-based Practice Center
Santa Monica, CA

Investigators:
Susanne Hempel, Ph.D.
Marika J. Suttorp, M.S.
Jeremy N.V. Miles, Ph.D.
Zhen Wang, M.S.
Margaret Maglione, M.P.P.
Sally Morton, Ph.D.
Breanne Johnsen, B.A.
Diane Valentine, J.D.
Paul G. Shekelle, M.D., Ph.D.

AHRQ Publication No. 11-EHC045-EF
June 2011
This report is based on research conducted by the Southern California Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10062-I). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.

Suggested Citation:
Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers; as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
Acknowledgements

We would like to thank the Tufts Medical Center and the Research Triangle Institute – University of North Carolina (RTI-UNC) Evidence-based Practice Center for their collaboration in identifying EPC reports and Kirstin Nyrop and Jonathan V. Todd for quality scoring of selected articles. We also thank Ethan Balk, Nancy Berkman, Isabelle Boutron, Tim Carey, Mark Helfand, David Moher, and Sydne Newberry for comments on an earlier draft of this report.
Empirical Evidence of Associations Between Trial Quality and Effect Sizes

Structured Abstract

Objectives. To examine the empirical evidence for associations between a set of proposed quality criteria and estimates of effect sizes in randomized controlled trials across a variety of clinical fields and to explore variables potentially influencing the association.

Methods. We applied quality criteria to three large datasets of studies included in a variety of meta-analyses covering a wide range of topics and clinical interventions consisting of 216, 165, and 100 trials. We assessed the relationship between quality and effect sizes for 11 individual criteria (randomization sequence, allocation concealment, similar baseline, assessor blinding, care provider blinding, patient blinding, acceptable dropout rate, intention-to-treat analysis, similar cointerventions, acceptable compliance, similar outcome assessment timing) as well as summary scores. Inter-item relationships were explored using psychometric techniques. We investigated moderators and confounders affecting the association between quality and effect sizes across datasets.

Results. Quality levels varied across datasets. Many studies did not report sufficient information to judge methodological quality. Some individual quality features were substantially inter-correlated, but a total score did not show high overall internal consistency (α 0.55 to 0.61). A factor analysis-based model suggested three distinct quality domains. Allocation concealment was consistently associated with slightly smaller treatment effect estimates across all three datasets; other individual criteria results varied. In dataset 1, the 11 individual criteria were consistently associated with lower estimated effect sizes. Dataset 2 showed some unexpected results; for several dimensions, studies meeting quality criteria reported larger effect sizes. Dataset 3 showed some variation across criteria. There was no statistically significant linear association of a summary scale or factor scores with effect sizes. Applying a cutoff of 5 or 6 criteria met (out of 11) differentiated high and low quality studies best. The effect size differences for a cutoff at 5 was -0.20 (95% confidence interval [CI]: -0.34, -0.06) in dataset 1 and the respective ratio of odds ratios in dataset #3 was 0.79 (95% CI: 0.63, 0.95). Associations indicated that low-quality trials tended to overestimate treatment effects. This observation could not be replicated with dataset 2, suggesting the influence of confounders and moderators. The size of the treatment effect, the condition being treated, the type of outcome, and the variance in effect sizes did not sufficiently explain the differential associations between quality and effect sizes but warrant further exploration in explaining variation between datasets.

Conclusions. Effect sizes of individual studies depend on many factors. The conditions where quality features lead to biased effect sizes warrant further exploration.
# Contents

**Executive Summary** .......................................................................................................................... ES-1  
**Background** ........................................................................................................................................... 1  
**Methods** .................................................................................................................................................. 4  
  Quality Criteria .......................................................................................................................................... 4  
  Study Pool Selection ................................................................................................................................. 4  
    Dataset 1: Back Pain Trials .................................................................................................................... 5  
    Dataset 2: EPC Reports .......................................................................................................................... 5  
    Dataset 3: Published “Pro-bias” Sample ............................................................................................... 6  
  Procedure .................................................................................................................................................... 6  
  Analysis ....................................................................................................................................................... 7  
    Data Description .................................................................................................................................... 8  
    Association Between Quality and Effect Sizes ...................................................................................... 10  
    Moderators and Confounders .............................................................................................................. 11  
    Variable 1: Size of Treatment Effect ................................................................................................. 13  
    Variable 2: Condition Being Treated .................................................................................................. 13  
    Variable 3: Type of Outcome .............................................................................................................. 13  
    Variable 4: Variance in Effect Sizes .................................................................................................... 14  
**Results** .................................................................................................................................................. 15  
  Data Description ....................................................................................................................................... 15  
  Relationship Between Total Quality Scores and Year of Publication .................................................. 15  
  Quality of the Reporting ......................................................................................................................... 17  
  Intercorrelations Quality Features ....................................................................................................... 18  
  Internal Consistency ............................................................................................................................... 21  
  Factor Analysis .......................................................................................................................................... 22  
  Association Between Quality and Effect Sizes ...................................................................................... 24  
    Dataset 1: Back Pain Trials .................................................................................................................... 24  
    Dataset 2: EPC Reports .......................................................................................................................... 26  
    Dataset 3: Published “Pro-bias” Dataset ............................................................................................... 29  
  Comparison Across Datasets .................................................................................................................. 33  
  Why the Association Between Quality Features and Effect Sizes Might Vary Across Datasets: Moderators and Confounders ........................................................................................................ 34  
    Variable 1: Size of Treatment Effect ................................................................................................. 34  
    Variable 2: Condition Being Treated .................................................................................................. 36  
    Variable 3: Type of Outcome .............................................................................................................. 39  
    Variable 4: Variance in Effect Sizes .................................................................................................... 41  
**Discussion** ............................................................................................................................................. 42  
  Quality of the Reporting ......................................................................................................................... 42  
  Psychometric Analysis ........................................................................................................................... 42  
  Associations Between Internal Validity and Effect Sizes ...................................................................... 44  
  Moderators and Confounders ............................................................................................................... 45  
  Implication for Practice ......................................................................................................................... 46  
  Future Research ....................................................................................................................................... 46  
  Conclusions ............................................................................................................................................... 47  
**References** ............................................................................................................................................. 48
Appendixes
Appendix A. References Dataset 1: Back Pain, 216 Trials
Appendix B. References Dataset 2: EPC Reports, 165 Trials
Appendix C. References Dataset 3: Published “Pro-bias” Dataset, 100 Trials
Appendix D. Comparison Fixed-Effects Model Results
Appendix E. Comparison Random Effects Meta-regression Results
Appendix F. Quality Rating Form
Executive Summary

Background

Trial design and execution factors are widely believed to be associated with bias. Bias is typically defined as a systematic deviation of an estimate, such as the estimated treatment effect from the true value. More factors have been proposed as associated with bias than have actually been empirically confirmed by systematic examination. There are some conflicting results regarding the association of quality features and effect sizes. Little is known about moderators and confounders that might predict when quality features (or the lack thereof) influence results of research studies.

Objective

The objective of this project was to examine the empirical evidence for associations between a set of proposed quality criteria and estimates of effect sizes in randomized controlled trials using multiple datasets representing a variety of clinical fields and to explore variables potentially influencing the association.

Methods

We applied a set of proposed quality criteria to three large datasets of studies included in a variety of systematic reviews covering a wide range of clinical fields. The first dataset was derived from all Cochrane Back Review Group reviews of nonsurgical treatment for nonspecific low back pain in the Cochrane Library 2005, issue 3; the set included 216 individual trials. For the second dataset we searched prior systematic reviews and meta-analyses conducted by Agency for Healthcare Research and Quality-funded Evidence-based Practice Centers with the goal of assembling a set with a wide range of clinical topics and interventions; this dataset included 165 trials. The third dataset was obtained by replicating a selection of trials used in a published meta-epidemiological study demonstrating associations of quality with the size of treatment effects; this set included 100 trials (79 percent of the original dataset).

The proposed set of 11 quality features comprised the following:

- Generation of the randomization sequence
- Concealment of treatment allocation
- Similarity of baseline values
- Blinding of outcome assessors
- Blinding of care providers
- Blinding of patients
- Acceptable dropout rate and stated reasons for withdrawals
- Intention-to-treat analysis
- Similarity of cointerventions
- Acceptable compliance
- Similar timing of outcome assessment.

In addition we applied the Jadad components and scale, and criteria suggested by Schulz, including allocation concealment, to one of the datasets. The inter-item relationships of the proposed quality criteria were explored using psychometric methods. A multiple indicator
multiple cause (MIMIC) factor analysis explored inter-item correlations as well as associations of quality features with reported effect sizes.

We assessed the relationship between quality and effect sizes for individual criteria as well as summary scores. In particular, the use of total quality scores per study with each item adding to a sum score, factor-analytically derived broad quality domains, and the application of different cutoffs for a total quality score was further explored.

We investigated moderators and confounders that affect the association between quality measures and the size of the treatment effect across datasets. In particular, we investigated whether (1) the overall size of the treatment effect of the intervention observed in datasets, (2) the condition being treated, (3) the investigated type of outcome, and (4) the variance in effect sizes across studies moderates or confounds the association between quality and effect sizes.

Results

The average quality levels varied across datasets. Many studies did not report sufficient information to judge the quality of the feature (although quality of reporting increased after the introduction of the Consolidated Standards of Reporting Trials statement). Some individual quality features were substantially intercorrelated, but a total score did not show high overall internal consistency of the 11 quality features ($\alpha$'s = 0.55 to 0.61). A MIMIC factor-analytic model suggested three distinct quality domains; randomization sequence generation and allocation concealment constituted the first factor, the blinding items constituted a second factor, and the third factor was primarily derived from the acceptable dropout rate item.

Allocation concealment was consistently associated with a slightly smaller treatment effect across all three datasets: Effect size differences were $-0.08$ (95% CI: $-0.23, 0.07$) in dataset 1 and $-0.06$ (95% CI: $-0.22, 0.11$) in dataset 2. The ratio of odds ratios was 0.91 (0.72, 1.14) in the third dataset where only categorical outcome measures were included; hence, we computed odds ratios rather than effect sizes. Other individual criteria results varied across datasets. In dataset 1 the 11 individual quality criteria were consistently associated with a lower effect size, indicating that low-quality studies overestimated treatment effects. Results in dataset 2 showed unexpected results: Higher quality studies reported larger effect sizes in this sample. The third dataset showed some variation across quality criteria.

There was no statistically significant linear association of a summary quality score (derived by equally weighing all 11 quality items) and effect sizes, which would have indicated that the effect size decreased linearly with increased quality. There was also no consistent linear association across datasets for the factor scores.

Applying a cutoff of 5 or 6 quality criteria met (out of a possible 11) differentiated high- and low-quality studies best. Effect size differences were $-0.20$ in dataset 1. In the third dataset, the ratio of odds ratios were 0.79 (cutoff at 5; 95% CI: 0.63, 0.95) and 0.77 (cutoff at 6; 95% CI: 0.63, 0.99). These associations indicated that low-quality trials tended to overestimate treatment effects. This effect could not be replicated in dataset 2, suggesting the influence of confounders and moderators of the association.

The specific moderators and confounders that were investigated in this report did not sufficiently explain the variation in associations across datasets. When controlling for the mean treatment effect obtained in each included meta-analysis, the differences across datasets in observed associations between quality and effect sizes remained. A stratified analysis for the condition being treated also failed to explain the contrary results observed in dataset 2 compared to the other two datasets; the clinical condition did not appear to confound the underlying
association between quality and effect sizes for individual quality criteria, and the interaction effect of condition with total quality score was also not statistically significant. When categorizing the different measures used to show a treatment effect into objective versus more subjective outcomes, the type of outcome did not show statistically significant interaction effects. The variance in effect sizes within datasets varied across the three datasets and may potentially explain differences observed in the association between quality and effect sizes across datasets; this finding should be investigated systematically. Several assumptions can be tested in meta-epidemiological datasets that may help determine when and which quality features lead to biased effect sizes.

Conclusions

The associations between quality features and effect sizes are complex. Effect sizes of individual studies depend on many factors. In two datasets, individual quality items and summary scores of items were associated with differences in effect sizes. This relationship was not found in the remaining dataset. Despite several exploratory analyses, we were not able to explain these differences. The conditions under which quality features and which features lead to biased effect sizes warrant further exploration and factors such as the variance in quality scores and effect sizes will be investigated in a subsequent project.
**Background**

Trial design and execution factors are widely believed to be associated with bias in randomized controlled trials (RCT). Bias is typically defined as a systematic deviation of an estimate, such as the estimated treatment effect from the true value. A number of individual quality criteria and quality checklists or scales for RCTs have been proposed (see e.g., Moja, Telaro D’Amico, et al. 2005; West, King, Carey, et al., 2002). These cover potential threats to the internal validity of the trial methodology.

Quality checklists typically provide a selection of quality features that are scored individually. Quality scales provide in addition a total quality score, either by summing up individual features giving equal weights to each feature or by putting more emphasis on selected features. Existing quality checklists and scales address the conduct or research methodology of the individual study, so they concern the internal validity of the research study, but they frequently also include other quality aspects of publications. Jadad and colleagues (Jadad, Moore, Carroll, et al., 1996) proposed a scale of 0 to 5 to evaluate RCTs with “low” and “high” internal validity in pain research. The Jadad scale, based on three criteria (randomization, double-blinding, and a description of dropouts), is widely used as a summary quality measure of randomized controlled trials (RCTs) and is one of the few tools where the psychometric properties have been evaluated and are acceptable. However, the Jadad scale has some limitations, e.g., the double blinding criterion is usually reported in fewer than 10 to 20 percent of studies. Many trials involve devices, surgery, or other interventions for which double blinding is either impractical or impossible and the double blinding criterion accounts for 40 percent of the Jadad score. An additional criterion, the concealment of treatment allocation, is not included in the Jadad scale but is widely used in addition to the criteria proposed by Jadad et al. (1996).

Verhagen, de Vet, de Bie, et al. (1998) developed a nine-item list of quality items specifically focused on internal validity, using a formal Delphi process of three rounds, which included leading experts from around the world. The 2008 Cochrane handbook (Higgins and Green, 2008) introduced a Risk of Bias tool based on the domains sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, and other sources of bias.

More factors have been proposed as related to bias than have actually been confirmed by systematic examination. Only a few researchers have published investigations of the association between selected trial quality and effect sizes obtained in individual trials. It is assumed that the conduct of the research methodology will influence the result that is obtained by the trial. The study methodology appears to distort the true value expected to be shown in the study. Typically, it is assumed that low-quality trials exaggerate treatment effects. Colditz, Miller, and Mosteller (1989) found RCTs to have smaller effect sizes than non-RCTs in studies of surgical therapy and RCTs that are double blind have smaller effect sizes than nonblinded trials of medical therapy. Schulz, Chalmers, Hayes, et al. (1995) assessed 250 trials in 33 meta-analyses and reported that inadequate concealment of allocation accounted for a 41 percent increase in effect sizes. The lack of double blinding showed a 17 percent increase in reported treatment effect. Contrarily, Emerson, Burdick, Hoaglin, et al. (1990) found no relationship between a consensus-developed quality scale (0–100 points) and treatment differences. Balk and colleagues (Balk, Bonis, Moskowitz, et al., 2002) applied 24 existing quality measures and assessed 26 meta-analyses involving 276 RCTs. The analysis focused on four conditions: cardiovascular disease, infectious disease, pediatrics, and surgery. The study found no indication of bias; individual quality measures were not reliably associated with the strength of treatment effect across studies and
clinical areas. Moher, Pham, Jones, et al. (1998) used Jadad’s scale and Schulz’s “concealment of allocation” in a large study assessing 11 meta-analyses including 127 RCTs. All trials were scored and the meta-analyses replicated. Low-quality studies were associated with an increased treatment estimate of 34 percent compared with high-quality trials. Studies with inadequate treatment allocation concealment showed a 37 percent increased effect compared to concealed trials. Juni, Altman, and Egger (2001) have summarized the data from Schulz et al. (1995), Moher et al. (1998), Kjaergard, Villumsen, and Gluud (1999) and Juni, Tallon, Egger, et al. (2000) in a pooled analysis and provide evidence for associations of effect sizes with allocation concealment (ratio of odds ratios [ROR] 0.70; 95% CI: 0.62, 0.80) and double blinding (ROR 0.86; 95% CI: 0.74, 0.99) while the generation of treatment allocation did not have a statistically significant effect across datasets (ROR 0.81; 95% CI: 0.60, 1.09). Pidal, Hrobjartsson, Jorgensen, et al. (2007) outline the potential consequences for meta-analysis conclusions. When only trials with adequate concealment were included in meta-analyses, two-thirds lost statistical significance of the primary result, primarily to loss of power (as a result of a smaller sample size) but also a shift in the point estimate towards a less beneficial effect. These studies provide data on quantifying the risk of bias associated with individual or sets of quality criteria.

The association between quality features and effect sizes may vary across datasets according to factors yet to be explored. Investigating moderators and confounders that may influence the association between quality and effect sizes and that may explain some of the conflicting results shown in the literature is an evolving field. Wood, Egger, Gluud, et al. (2009) used three sets of “meta-epidemiological studies,” that is, studies investigating the associations of quality features and effect sizes (Kjaergard, Villumsen & Gluud, 2001; Schulz, et al., 1995; Egger, Juni, Bartlett, Holenstein, and Sterne, 2003). The group investigated whether the nature of the intervention and the type of outcome measures influences the effect of allocation concealment and blinding. They found that studies using subjective outcomes showed exaggerated effect sizes when there was inadequate or unclear allocation concealment or lack of blinding. In studies using objective outcomes such as mortality, the association of quality with trial results was negligible. Differentiating drug interventions and nondrug interventions showed no significant differences on the effect on allocation concealment or blinding.

Recently, quality criteria suggested by the Cochrane Back Review Group (CBRG) were found to be associated with effects sizes in RCTs of interventions for back pain (van Tulder, Suttrop, Morton, et al., 2009). The CBRG Editorial Board developed an 11-item criteria list, based on the 9-item Delphi list (Verhagen et al., 1998) and the 3-item Jadad criteria, for evaluation of internal validity of RCTs (van Tulder, Furlan, Bombardier, et al., 2003). Modifications were made to tailor the criteria list to the expected needs of trials of treatments for back pain. The Delphi list was modified by adding three items that had been eliminated between rounds two and three of the Delphi (items about withdrawals and dropouts, compliance rate, and co-interventions), deleting one Delphi list criterion on specifying eligibility criteria and adding one item about the timing of measurement of outcomes. This 11-item list was then proposed by the CBRG editorial board as the standard measure for assessing quality of controlled trials and has been used in virtually all CBRG reviews. A summary score of 0 to 11, based on the 11-item list, was developed as a measure of overall internal validity. Results of applying this set of criteria on all trials of nonsurgical therapy in CBRG reviews showed consistent effects of the criteria on effect sizes.
We aim to assess the potential usefulness of the set of CBRG quality criteria to other clinical conditions by applying these criteria to large datasets of RCTs covering diverse clinical topics and diverse outcome measures. We examine the empirical evidence of associations between individual quality criteria as well as summary scores. In addition, factors influencing the association between quality and effect sizes are explored.
Methods

This project developed sequentially over time. The original study was part of a project for the Cochrane Back Review Group (CBRG). The additional work was funded by the Agency for Healthcare Research and Quality in steps as results of earlier analyses suggested fruitful areas for testing of new hypotheses.

Quality Criteria

We applied the 11 CBRG Internal Validity criteria (van Tulder et al, 2003) that appeared very promising in the quality scoring of Cochrane back reviews. The items cover established quality criteria (allocation concealment, blinding), as well as criteria for which no evidence on their potential for bias has been investigated or existing studies showed conflicting results.

The individual criteria address the adequacy of the randomization sequence generation, concealment of treatment allocation, baseline similarity of treatment groups, outcome assessor blinding, care provider blinding, patient blinding, adequacy and description of the dropout rate, analysis according to originally assigned group (intention-to-treat analysis), similarity of cointerventions, adequacy of compliance and similar assessment timing across groups. The items and the scoring guideline are shown in Appendix F.

The answer mode employed the following categories: “Yes,” “No,” and “Unclear.” The CBRG offers concrete guidance for each answer category. Assessor blinding for example is scored positively when assessors were either explicitly blinded or the assessor is clearly not aware of the treatment allocation (e.g., in automated test result analysis).

A number of items are topic specific and have to be defined individually. For each topic, a content expert (typically a clinician with trial research experience) was contacted to assist in the selection of baseline comparability variables and to establish reasonable dropout and compliance rates. The baseline comparability assessment requires that topic specific key prognostic predictors of the outcome are specified and the baseline comparability of the treatment groups has to be judged. For interventions that involve considerable patient commitment (e.g., presenting at multiple outpatient appointments) a dropout rate of about 25 percent was considered sufficient, while for other interventions a rate of 10 percent was considered sufficient in order to meet this criterion in the specific clinical area.

In addition, for one of the datasets the Jadad scale (Jadad et al., 1996) and criteria proposed by Schulz et al. (1995), operationalized as in the original publications, was applied. The Jadad scale (0 to 5 points) assesses randomization (0 to 2 points), blinding (0 to 2 points), and withdrawals (0 to 1 point). The applied Schulz criteria were allocation concealment, randomization sequence, analysis of all randomized participants, and double blinding. The items together with the scoring instructions can be found in Appendix F.

Study Pool Selection

This project drew on three different study pools. One was available from previous work for the Cochrane Back Review Group, the project has been described in detail elsewhere (van Tulder et al., 2009). Two datasets were obtained for the purpose of this project only (datasets 2 and 3). First results on the association between quality and effect sizes in dataset 1 have been published previously (van Tulder et al., 2009), all further analyses were prepared for this report only.
Dataset 1: Back Pain Trials

For the CBRG project the quality criteria were originally applied to all CBRG reviews of nonsurgical treatment for nonspecific low back pain present in the Cochrane Library 2005, issue 3. The study set was drawn from 12 reviews (Assendelft, Morton, Yu, et al., 2004; Furlan, van Tulder, Tsukayama, et al., 2005; Furlan, Imamura, Dryden, et al., 2005; Hagen, Hilde, Jamtvedt, et al., 2005; Hayden, van Tulder, Malmivaara, et al., 2005; Henschke, Ostelo, van Tulder, et al., 2005; Heymans, van Tulder, Esmail, et al., 2004; Karjalainen, Malmivaara, van Tulder, et al., 2003; Khadilkar, Odebiyi, Brosseau, et al., 2005; Roelofs, Deyo, Koes, et al., 2005; van Tulder, Touray, Furlan, et al., 2003; van Duijvenbode, Jellema, van Poppel, et al., 2005). Studies reported on pain, function, or other improvement measures. The reviews assessed the effect of acupuncture, back schools, behavioral treatment, exercise therapy, bedrest, lumbar supports, massage, multidisciplinary bio-psycho-social rehabilitation, muscle relaxants, spinal manipulative therapy, and transcutaneous electrical nerve stimulation (TENS) for the treatment of low-back pain. Comparisons were placebo, usual care, or no treatment or comparisons between treatments. The dataset included 216 trials.

Dataset 2: EPC Reports

In the first of two efforts supported by AHRQ, we assembled a second dataset of trials based on Evidence-based Practice Center (EPC) reports. We searched prior systematic reviews and meta-analyses conducted by AHRQ-funded EPCs with the goal of assembling a test set of studies that represented a wide range of clinical topics and interventions. The criteria for selection were that the EPC report had to include a meta-analysis and that the EPC had to be willing to provide us with the data on outcomes, such that we only needed to assess the quality of the included trials. The study set was drawn from 12 evidence reports, the majority were also published as peer review journal articles (Balk, Lichtenstein, Chung, et al., 2006; Balk, Tatsioni, Lichtenstein, et al., 2007; Chapell, Reston, Snyder, et al., 2003; Coulter, Hardy, Shekelle et al., 2003; Donahue, Gartlehner, Jonas, et al., 2007; Hansen, Gartlehner, Webb, et al., 2008; Hardy, Coulter, Morton, et al., 2002; Lo, LaValley, McAlindon, et al., 2003; Shekelle, Morton, Hardy, 2003; Shekelle, Maglione, Bagley, et al., 2007; Shekelle, Morton, Maglione, et al., 2004; Towfigh, Romanova, Weinreb, et al., 2008). The reports addressed a diverse set of topics, pharmacological therapies as well as behavior modification interventions. All studies included in the main meta-analysis of the report were selected; studies included in more than one report entered our analysis only once. The dataset included 165 trials.

The reports addressed pharmaceuticals (orlistat, vitamin E, drugs for arthritis, S-adenosylmethionine, chromium, atypical antipsychotics, omega-3 fatty acids); non-pharmacological studies such as self-monitoring of blood glucose (SMBG), diet and weight loss, chronic disease self-management (CDSM); interventions to manage and treat diabetes (chromium, SMBG, CDSM); complementary and alternative medicine/dietary supplements (vitamin E, chromium, omega-3); as well as mental health topics (Alzheimer’s, obsessive-compulsive-disorder [OCD]).

In each of the evidence reports one meta-analysis (in general the analysis with the largest number of trials) was selected and all studies included in that pooled analysis were chosen for the study pool. Only one comparison per study was included. Multiple publications and multiple outcomes were excluded so that each unique study entered the test set only once. In the majority, individual studies compared the intervention to placebo or usual care.
Dataset 3: Published “Pro-bias” Sample

Following the results of the analysis of the EPC reports, we obtained a third dataset of studies. This third dataset was obtained by replicating a selection of trials used by Moher et al. (1998). The dataset was chosen as it has shown evidence of bias for established quality criteria (see Moher et al., 1998) and is designated in this report as “pro-bias.” Since the original publication does not specify exactly which trials and which outcomes were included in this analysis, we replicated the methods described by Moher and colleagues for selection. Two reviewers independently reviewed the 11 meta-analyses chosen by Moher et al. and reconciled their assessment of the primary outcome and the main meta-analysis in the publication. Following the described approach, this designation of the primary outcome was based on the largest number of randomized controlled trials (RCTs) reporting data on that endpoint since many meta-analyses did not identify a primary outcome. Individual trials present in multiple meta-analyses were included only once so that a trial did not enter our analyses more than once. Where multiple comparisons were reported in original articles we included those data chosen in the main analysis of the 11 meta-analyses. We were able to retrieve, quality score, and abstract 100 RCTs of the originally published set (79 percent).

The trials came from meta-analyses on digestive diseases (Marshall and Irvine, 1995; Pace, Maconi, Molteni, et al., 1995; Sutherland, May, and Shaffer, 1993), circulatory diseases (Ramirez-Lasspas and Cipolle, 1988; Lensing, Prins, Davidson, et al., 1995; Loosemore, Chalmers, and Dormandy, 1994), mental health (Mari and Streiner, 1994; Loonen, Peer, and Zwanikken, 1991; Dolan-Mullen, Ramirez, and Groff, 1994), stroke (Counsell Sandercock, 1995) and pregnancy and childbirth (Hughes, Collins, and Vanderkeckhove, 1995).

The flow diagram in Figure 1 summarizes the dataset composition.

Procedure

We developed and pilot tested a standardized form to record decisions for the quality criteria. For all datasets, two reviewers independently rated the quality of each study by applying the outlined quality criteria. The reviewers used the full publications to score the studies and were not blinded to authors, journals or other variables. The reviewers were experienced in rating study quality in the context of evidence based medicine and underwent an additional training session for this study. The pair of reviewers reconciled any disagreement through consensus; any remaining disagreements were resolved by discussion in the research team.

The outcomes of the individual RCTs were extracted by a statistician together with measures of dispersion where available and the number of patients in each group. For dataset 1 (back pain) absolute effect sizes were used as this dataset included comparisons between treatment and placebo as well as comparisons between active treatments. For dataset 2 (EPC reports) in order to be able to combine studies within data sets or where possible between datasets, standardized effect sizes were computed for each study. As all studies in dataset 3 (pro-bias) reported dichotomous outcomes, odds-ratios (OR) were calculated. As a quality check, the point estimate and 95 percent confidence interval (CI) of each meta-analysis was calculated and compared to the original meta-analytic result.
Analysis

Figure 2 depicts the basic hypothesis of the project: the assumption that there is an association between quality features of research studies and the size of the reported treatment effect. The arrows indicate the direction of effects. The figure also depicts the assumption that other variables apart from quality will affect effect sizes, as represented by the arrow on the right. These other variables include the true effect of the intervention as well as other potential influences; quality variables may explain part of the reported effect sizes, but there are other and possibly more important factors that are not quality related (e.g., the efficacy of the treatment). The analysis covers descriptive information on the datasets, an evaluation of the association between quality and effect sizes, and an analysis of potential moderators and confounders to investigate which factors influence the association between quality criteria and effect sizes.

The three datasets were often used to replicate results obtained in one dataset to test the robustness of effects across datasets; some analyses were only possible in one or two datasets. The initial intention to combine all three datasets to allow more powerful analyses was not feasible due to differences in outcome measures (all RCTs in dataset 3 used dichotomous outcomes, to transform all outcomes into continuous measures was considered problematic).
Since this analysis plan involves multiple testing, we considered several methods for accounting for this; however these are not appropriate when tests are correlated. In addition, there is debate about the range to which multiple testing corrections should be employed (for an analysis, a study) and each of these would lead to different conclusions. All statistical multiple testing approaches lead to substantial loss of power (Bland and Altman, 1995). We therefore chose not to employ any of the methods to “correct” for multiple testings. Instead our results need to be interpreted with more caution, as a result of multiple testing.

Data Description

The three datasets were derived through different means and differed in a number of ways. First, we investigated if there were systematic differences related to the level of quality within the datasets. The level of quality may vary between clinical fields as the clinical areas may have different standards or awareness of quality features. The quality of published RCTs may have improved since the publication of the Consolidated Standards of Reporting Trials (CONSORT) statement in 1996 so another variable we explored further was the year of publication of studies included in each dataset.

To describe the internal consistency of the quality items, inter-item correlations and the Cronbach’s alpha statistic for an overall quality scale were computed in each dataset. The Pearson correlations across items were inspected for consistency (are the individual quality features positively correlated) but also to detect high inter-item correlations (e.g. above 0.5) as an indicator for conceptual overlap (the answer in one item lets us predict the answer in another item).

All of the items score quality features. It is possible that the features are independent of each other (blinding of outcome assessors is not necessarily related to the similarity of the co-interventions). However, empirically the presence of one quality indicator might increase the likelihood that a second quality criterion is also fulfilled. For example, a study that used an appropriate method for a randomization sequence may also be likely to have employed an appropriate method to guarantee allocation concealment. Finally, theoretically, it is also possible that the individual items are indicators of an underlying factor representing “quality.” A quality RCT is more likely to show several fulfilled quality criteria. Individual quality items may be indicators of this underlying quality factor.

We also used the individual quality items to create a sum scale. This overall quality score was computed by calculating the average quality scores across all items, with all items being...
weighted equally. Cronbach’s alpha values range from 0 to 1; alpha coefficients above 0.7 indicate internal consistency. The Cronbach’s alpha statistic was exploratory and was chosen as a measure with well-known properties, not because we assume a shared overarching latent quality factor. The included quality features may still be conceptually independent from another and may not represent items from the same item pool of a shared latent factor.

We also used factor analysis to explore the structure of the relationships between the 11 items. Conventional exploratory factor analysis attempts to find latent factors which explain the covariance between a set of items. Factor analysis assumes an underlying factor that is hypothesized to influence a number of observed variables, that is, the individual items. Factor analysis can show whether all included items can be explained through one underlying factor (e.g., “quality”), whether there are clusters of items representing different quality aspects, or whether all 11 items are unrelated and represent unique features. Conventional factor analysis only takes the pattern of quality scores across items into account; this approach does not incorporate the relationships with an outcome (such as effect size). We used an extension of factor analysis, a multiple indicator multiple cause (MIMIC) model, which allows us to model the relationships between the items in an exploratory fashion, and simultaneously examine the relationship between the latent variables that were identified and the outcome of interest (in this case, the effect size of the study). The factor analysis hence takes the inter-item relationships as well as the strength of association with effect sizes into account.

The path model shown in Figure 3 below is a simplified diagrammatic representation of the model assuming four indicators of quality ($x_1$ to $x_4$), that is, individual quality features. Single-headed arrows are regression paths—the four indicators of quality are hypothesized to be explained by two latent variables, $F_1$ and $F_2$. The two latent (unmeasured) variables represent distinct broad quality domains but are not necessarily completely independent from each other either; we assume in our model that they are correlated (the two-headed, curved arrow indicates this).

We hypothesize that the covariances between $x$ variables are accounted for by the factors, the latent variables. We assume the latent (unmeasured) factors ($F_1$, $F_2$) are responsible for the majority of variation in individual quality criteria, and that these latent variables are also predictors of effect size. The indicator variables, that is, the individual quality items, are not conceptualized as being correlated; they can be independent from another, such as blinding and similarity of cointerventions. The partial correlation between individual quality indicators and the effect size is diminished to zero when controlling for the latent factors.

In summary, the effect size reported in the trials is regressed on the latent variables—that is quality is indicated by the $x$-variables (individual quality features), but the latent variables (unmeasured, broad quality domains) are hypothesized to predict variance in the effect size. It has to be kept in mind that variables other than quality will affect effect sizes, as represented by the arrow on the right.

To identify the appropriate number of latent factors that are required to account for the data, we employed fit indices ($\chi^2$, comparative fit index, [CFI] and root mean square error of approximation [RMSEA]). We tested a series of models, each time increasing the number of factors and comparing the improvement of the model fit. This approach is used to determine the smallest number of factors that can account for the data.
The factor analysis solution is more parsimonious and enables a large number of items to be reduced to a smaller number of underlying factors. Factor analysis allows summarizing results across items without reducing the quality data to a summary score. However, the analysis should be considered descriptive as data are not weighted by standard error as is conventional in meta-analysis.

Association Between Quality and Effect Sizes

We investigated the association between quality and effect sizes in a number of ways. First, the differences between results in studies that met a quality criterion were calculated for each of the 11 quality features. Secondly, a summary score was calculated across all quality components and a linear relationship between quality and effect sizes was investigated. Third, the associations based on empirically derived factor scores was tested, the factor structure took the inter-correlations between items and their effects on outcomes into account. Fourth, we explored different cutoffs of quality scores according to the number of quality components met.

For all analyses we differentiated quality items scored “yes” and those with the quality item scored “not yes” which included the answer “no” and “unclear” unless otherwise stated.

Individual Criteria

In the first two datasets (back pain, EPC reports) we compared the effect sizes of studies with the quality item scored “yes” and those with the quality item scored “not yes” for each of the 11 quality features. The difference in effect sizes between these two subgroups per feature was used as a measure of bias. The difference was estimated using meta-regression (Berkey et al., 1995). A meta-regression was conducted separately for each quality feature. The coefficient from each regression estimates the difference in effect sizes between those studies with the quality feature scored “yes” versus “not yes.” A difference with a significance level of $p<0.05$ was considered statistically significant.

In the third dataset, the published “pro-bias” dataset, all studies used dichotomous outcomes. An odds ratio below 1 indicates the treatment group is doing better than the control. For the analysis we compared odds ratios (ORs) of studies where the quality criterion was either met or not met and computed the ratio of the odds ratios (ROR). The ROR is $\frac{OR_{no}}{OR_{yes}}$ where $OR_{no}$ is the pooled estimate of studies without the quality feature and $OR_{yes}$ is the pooled estimate of studies where the quality criterion is met.
Note that the interpretation of reported differences of the first two datasets differs from that of the third one. In the first two datasets a negative difference coefficient indicated that studies with the quality item scored “yes” have smaller effect sizes than those that scored “not yes.” Hence, a negative difference indicates that the higher quality RCTs report less pronounced treatment effects. In the third dataset a ROR of being less than 1 indicates that high quality studies reported smaller treatment effects (i.e., the OR closer to 1) than low quality studies.

We compared results based on a random effects meta-regression, and a fixed effects model in order to be able to match results reported in the literature.

**Sum Scores**

The sum of the quality items scored “yes” was calculated across all 11 items with all items contributing equally to the total score. To assess a linear relationship between overall quality and effect size, reported outcome results were regressed on the sum score. A simple linear relationship indicates that the reported treatment effects increase the lower the quality level. A level of p<0.05 was considered statistically significant.

**Factor Scores**

We used the empirically derived factor scores representing broad quality domains and regressed effect sizes on these factors. The factor scores were based on the inter-item relationships as well as their effects, that is, the association with the study results that provides a description of distinct groups of items. The analysis was equivalent to the sum score analysis.

**Cutoffs**

Different cutoffs depending on the number of criteria met were explored to differentiate high and low quality studies. The difference in effect sizes of studies above and below possible thresholds was investigated. The statistical analysis followed the approach outlined under (1).

The different methods of establishing associations between quality and effect sizes were exploratory and we did not a priori assume consistent results across methods. For example, a simple linear relationship between a total quality scale and effect sizes will not necessarily be present even when individual quality features show large associations with effect sizes; the internal consistency across items was one of the issues under investigation.

The analyses were conducted separately in each of the three datasets. Each dataset consisted of trials included in up to 12 meta-analyses. We did not correct for clustering in analyses within datasets. We do not assume nonindependence of RCTs within meta-analyses since the selection into the meta-analysis happened after the event (when the study was already conducted and published).

**Moderators and Confounders**

Effect sizes are influenced by many variables, not just the methodological quality of the research study. In addition, we have to assume from conflicting literature results that there are factors that influence the relationship between methodological quality and the effect size. Figure 4 shows a model that assumes factors influencing the association between quality and effect sizes and indicates that effect sizes are also influenced by other variables independent from trial quality.
Two effects need to be considered: confounding effects and moderating effects. These are of particular relevance in dataset 2, where papers are selected from a wide range of clinical topics and interventions.

Confounding effects occur when the quality of trials is not equally distributed across areas of study, resulting in a correlation between quality and area of study. This correlation can lead to erroneous conclusions if the area of study is not incorporated as a covariate. In extreme cases, this correlation can lead to counter-intuitive results, an effect known as Simpson’s paradox. The example in Table 1 considers two areas of study, labeled A and B, and a measure of quality, such as randomization, which is either achieved or not achieved, giving four combinations. The effect sizes are given in the table below. Within study area A, the effect size is 0.1 higher when the quality measure is not achieved. Similarly, within study area B the effect size is 0.1 higher when the quality rating is not achieved. However, studies in Area B have higher effect sizes on average (0.25) than studies in Area A (0.15), and studies in Area B are much more likely to have achieved the quality rating. This confounding means for subpopulations of studies the result is in one direction, but for the whole population the result is in the other direction.

The second potential issue is one of moderation. In the case of moderation, the causal effect for a quality rating varies between different substantive areas. We illustrate a moderator effect in Table 2. This example shows that for substantive area A, quality does not influence the effect size; however for area B there is a substantial influence of quality on effect size. To take the average quality association would be inappropriate when the influence differs across substantive areas (and would therefore be influenced by the number of studies identified in each area).
The literature reports some conflicting results regarding the strength of association between quality features and effect sizes indicating that we have to assume factors that influence the relationship between the two variables. In this project we set out to investigate the influence of four variables: the size of the treatment effect, the condition that is being treated, the type of analyzed outcome and the variance in effect sizes across studies for the quality feature in question.

**Variable 1: Size of Treatment Effect**

We tested the hypothesis that the association of quality features and reported effect sizes varies according to the size of the overall treatment effect. Strong treatment effects may mask any effects of quality features on the individual study outcome. Similarly, an ineffective treatment may likewise yield the same result regardless of study quality. We computed the mean effect size for each included meta-analysis and added this variable to the regression models and compared results between two datasets.

**Variable 2: Condition Being Treated**

We tested the hypothesis that the association of quality features and effect size varies by condition. Under this hypothesis the selection of clinical conditions in a dataset determines whether or not an association between quality and effect size can, or cannot be shown. The underlying factors for this differential effect may remain unknown; we are only testing whether the association with quality features can be documented in one clinical area or groups of clinical areas but not in others.

The analysis was restricted to the large and diverse EPC report dataset (dataset 2, 165 trials). The back pain studies addressed a homogeneous condition. The third dataset was too small to investigate the effects for each of the 11 included conditions (most comparisons would be incomputable) and too unbalanced (only 2 out of 11 studies were not drug studies, only 1 meta-analysis was in pregnancy and childbirth).

**Variable 3: Type of Outcome**

We tested the hypothesis that the association of quality and effect sizes varies by type of outcome. Whether an association of quality and effect sizes can be shown may depend primarily on the investigated outcome. Some types of outcomes may be more susceptible to bias than others. More objective versus more subjective outcome measures may represent different kind of outcome types. Hypothesis 3 tests whether the association of quality features and effect size may vary by the type of analyzed outcome.

In the back pain dataset, the measured outcomes were all either subjective outcomes such as pain or outcomes involving clinical judgment such as “improvement,” so this set could not contribute to this moderator analysis. The outcomes in the EPC report dataset were more diverse. We distinguished automated data (hemoglobin A1c, high-density lipoprotein, and total cholesterol) versus other endpoints (Alzheimer’s Disease Assessment Scale cognition score, arthritis responders, reduction in seizures, pain, OCD improvement, weight loss, and depression scores). In the third dataset, we distinguished objective data such as death, pregnancy, and biochemical indicators of smoking cessation, from other endpoints of a more subjective nature or involving clinical judgment (response in ulcer healing or pain relief, bleeding complications,
schizophrenic relapse, ulcer healing rate, affective relapse, and maintenance of ulcerative colitis remission).

**Variable 4: Variance in Effect Sizes**

We tested the hypothesis that the association of quality features and effect sizes may depend on the variance in effect sizes across studies in a given dataset. In a dataset where there is a wide range of reported effect sizes across studies, quality may be more likely to explain differences in effect sizes across studies.
Results

Data Description

The years of publication of the included papers are shown in Figure 5.

Figure 5. Year of publication of included studies

![Graph showing the percentage of papers published over time for different datasets: Back pain, EPC reports, and Pro-bias.](image)

EPC = Evidence-based Practice Center

The three datasets showed some differences: dataset 3 (published “pro-bias” dataset) included many older papers with a peak in the 1990s compared to the other datasets and all studies were published before 1996. The dataset 1 (back pain data) included mainly newer publications, several published in the last decade. The studies included in the Evidence-based Practice Center (EPC) reports were published over a large period of time, with no particular peak.

Relationship Between Total Quality Scores and Year of Publication

We investigated in each dataset the relationship between a quality sum score based on the mean of the assessed quality features and the year of publication. Figure 6 plots both variables.

In addition, the difference in quality between pre- and post-Consolidated Standards of Reporting Trials (CONSORT) publications was tested (1996 used as cutoff). In the back pain dataset, the difference in total scores between pre- and post-CONSORT published trials was 0.58 (SE 0.32, p=0.07) on the 11-item scale. The quality of studies published after the introduction of the CONSORT statement was better but not statistically significantly higher. In the EPC report dataset the difference between pre- and post-CONSORT quality ratings was 1.35 (SE 0.32, p<0.001). All studies included in the third dataset were published before the introduction of CONSORT.

To ensure that the effect was not an artifact of the fact that quality of studies was improving over time anyway, regardless of CONSORT, we estimated the effect of time for papers published both pre- and post-CONSORT. These effects were not statistically significant.
Figure 6. Total quality and year of publication

Dataset 1: Back pain

Dataset 2: EPC reports

Dataset 3: “pro-bias”

Note: data points have been “jittered” to avoid overlap.
Quality of the Reporting

Figure 7 shows the distribution of answers to the quality items (yes, unclear, no). A “yes” is an indicator of high quality for all items (randomization sequence, allocation concealment, baseline similarity, outcome assessor blinding, care provider blinding, patient blinding, dropout rate and description, analysis in original group, intention to treat [ITT], cointerventions, compliance, and assessment timing), for example, the outcome assessors were blind.

Figure 7. Quality item answer distribution
In the back pain dataset, the presence or absence of quality features is relatively evenly distributed for most items. Patient and provider blinding was not very common in the included trials and presumably often impossible due to the nature of the interventions. Similar timing of outcome measure assessment in the treatment and the control group was common, but there were a few deviations. The studies included in the EPC reports showed less variation across items. Many quality features were either present or there was not enough information to judge the individual quality feature. The answer “no” was very rare. The “unclear” answer was very common in dataset 2 (EPC reports) and 3 (published “pro-bias” dataset) indicating that the original studies did not report enough information to judge the quality feature. Very few trials scored negatively for the assessed quality features, that is, the reviewer had enough information to know that the design feature was not adhered to (e.g., the patient was not blinded). In datasets 2 and 3 there was virtually no variance in the item “Was the timing of the outcome assessment similar in all groups?” across studies, indicating that this quality feature may be unique to back pain trials.

Figure 8 allows a comparison of “yes” answers across the three datasets.

Figure 8. Criterion met across datasets

![Graph showing the percentage of studies meeting criteria across datasets]

EPC = Evidence-based Practice Center

The level of criteria met was highest in the EPC report dataset for the blinding items, similarities of cointerventions, similar timing, and the analysis in the original group assignment (ITT analysis). The lowest quality level across quality criteria was generally observed in the third dataset, which contains older studies, all published before the CONSORT statement.

**Intercorrelations Quality Features**

Although conceptually presumably independent, in practice studies that pay attention to one quality feature (e.g., allocation concealment) often do so also for others (e.g., using an adequate method of generating a randomization sequence). To trace the empirical interrelatedness of the quality items, Tables 3–5 show the inter-item correlations of quality features in each of the three datasets.
Table 3. Dataset 2 Inter-item correlations dataset 1 (back pain)

<table>
<thead>
<tr>
<th></th>
<th>Randomization Adequate</th>
<th>Allocation Concealment</th>
<th>Similar Baseline</th>
<th>Assessor Blind</th>
<th>Care Provider Blind</th>
<th>Patient Blind</th>
<th>Acceptable Dropout Rate</th>
<th>Original Group (ITT)</th>
<th>Similar Cointerventions</th>
<th>Acceptable Compliance</th>
<th>Similar Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>0.53</td>
<td>0.19</td>
<td>0.01</td>
<td>-0.30</td>
<td>-0.20</td>
<td>0.10</td>
<td>0.23</td>
<td>0.20</td>
<td>0.28</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rand Adequate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concealed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor Blind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care Provider Blind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Blind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable Dropout Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Group (ITT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Cointerventions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable Compliance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ITT = intention to treat

The mean inter-item correlation in this dataset was 0.1. Some of the items were substantially inter-correlated, for example, was the treatment allocation concealed correlated highly with an adequate randomization sequence, and if studies reported patient blinding, the studies tended to also report provider and outcome assessor blinding. The majority of features showed coherence but did not indicate that items were redundant and the information for one item was contained in another. There were a few negative correlations; the only noteworthy correlation was that the studies that reported an adequate randomization procedure stated that the care providers were not blinded (often impossible in this dataset given the interventions).
Table 4. Dataset 2 Inter-item correlations dataset 2 (EPC reports)

<table>
<thead>
<tr>
<th></th>
<th>Rand Adequate</th>
<th>Concealed</th>
<th>Similar Baseline</th>
<th>Assessor Blind</th>
<th>Provider Blind</th>
<th>Patient Blind</th>
<th>Dropout Acceptable</th>
<th>Original Group</th>
<th>Co-Intervention</th>
<th>Compliance Accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization Adequate</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation Concealment</td>
<td>0.74</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Baseline</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor Blind</td>
<td>0.01</td>
<td>0.06</td>
<td>-0.07</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care Provider Blind</td>
<td>0.13</td>
<td>0.15</td>
<td>-0.17</td>
<td>0.36</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Blind</td>
<td>0.05</td>
<td>0.11</td>
<td>-0.15</td>
<td>0.15</td>
<td>0.78</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable Dropout Rate</td>
<td>0.12</td>
<td>0.08</td>
<td>-0.08</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.09</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Group (ITT)</td>
<td>0.15</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.01</td>
<td>0.41</td>
<td>0.36</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Cointerventions</td>
<td>0.09</td>
<td>0.02</td>
<td>0.09</td>
<td>-0.01</td>
<td>0.17</td>
<td>0.20</td>
<td>-0.07</td>
<td>0.13</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acceptable Compliance</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.08</td>
<td>0.09</td>
<td>0.19</td>
<td>0.07</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td>Similar Timing</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.12</td>
<td>0.15</td>
<td>0.17</td>
<td>0.20</td>
<td>0.03</td>
<td>0.03</td>
<td>0.38</td>
<td>0.15</td>
</tr>
</tbody>
</table>

ITT = intention to treat

The mean inter-item correlation in the EPC reports was $r = 0.11$. Adequate sequence of randomization and concealment of treatment allocation were highly intercorrelated, as were provider and patient blinding.

Table 5. Dataset 3 Inter-item correlations dataset 3 (“pro-bias”)

<table>
<thead>
<tr>
<th></th>
<th>Rand Adequate</th>
<th>Concealed</th>
<th>Similar Baseline</th>
<th>Assessor Blind</th>
<th>Provider Blind</th>
<th>Patient Blind</th>
<th>Dropout Acceptable</th>
<th>Original Group</th>
<th>Co-Intervention</th>
<th>Compliance Accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization Adequate</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation Concealment</td>
<td>0.49</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Baseline</td>
<td>0.34</td>
<td>0.13</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor Blind</td>
<td>0.02</td>
<td>0.15</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care Provider Blind</td>
<td>0.07</td>
<td>0.00</td>
<td>0.28</td>
<td>0.69</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Blind</td>
<td>-0.02</td>
<td>-0.04</td>
<td>0.24</td>
<td>0.53</td>
<td>0.79</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable Dropout Rate</td>
<td>0.13</td>
<td>0.04</td>
<td>0.16</td>
<td>-0.12</td>
<td>-0.03</td>
<td>-0.08</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Group (ITT)</td>
<td>0.01</td>
<td>-0.03</td>
<td>0.16</td>
<td>-0.14</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar Cointerventions</td>
<td>0.09</td>
<td>-0.08</td>
<td>0.34</td>
<td>0.10</td>
<td>0.19</td>
<td>0.15</td>
<td>-0.01</td>
<td>0.06</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acceptable Compliance</td>
<td>-0.03</td>
<td>0.00</td>
<td>-0.07</td>
<td>-0.14</td>
<td>-0.12</td>
<td>-0.14</td>
<td>0.06</td>
<td>0.03</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Similar Timing</td>
<td>0.05</td>
<td>0.06</td>
<td>0.26</td>
<td>0.12</td>
<td>0.32</td>
<td>0.42</td>
<td>0.05</td>
<td>0.01</td>
<td>0.17</td>
<td>0.07</td>
</tr>
</tbody>
</table>

ITT = intention to treat
The mean inter-item correlation in the third dataset was 0.11, and the pattern was very similar to the two other datasets.

In addition to the newly proposed quality features, we also applied the Jadad scale and the quality features suggested by Schulz et al. (1995) including allocation concealment. Table 6 shows the intercorrelations between the newly proposed items and the features scored according to the Jadad scale instructions and Schulz’s original instructions. Correlations in bold indicate corresponding quality domains. As expected, there were strong correlations between Cochrane Back Review Group (CBRG) Internal Validity items and corresponding items for Jadad and Schulz—items describing randomization, concealment, blinding, and dropouts.

### Table 6. Correlation of criteria with Jadad and measures proposed by Schulz

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Jadad Randomization</th>
<th>Jadad Blinding</th>
<th>Jadad Withdrawals</th>
<th>Jadad Total</th>
<th>Schulz Concealment</th>
<th>Schulz Sequence</th>
<th>Schulz Analysis</th>
<th>Schulz Blinding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was the method of randomization adequate?</td>
<td>0.86</td>
<td>0.13</td>
<td>0.04</td>
<td>0.50</td>
<td>0.39</td>
<td>0.91</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td>Was the treatment allocation concealed?</td>
<td>0.44</td>
<td>0.12</td>
<td>-0.01</td>
<td>0.29</td>
<td>0.84</td>
<td>0.46</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>Were the groups similar at baseline regarding the most important prognostic indicators?</td>
<td>0.31</td>
<td>0.33</td>
<td>0.16</td>
<td>0.43</td>
<td>0.08</td>
<td>0.28</td>
<td>0.05</td>
<td>0.32</td>
</tr>
<tr>
<td>Was the outcome assessor blinded?</td>
<td>0.13</td>
<td>0.65</td>
<td>-0.09</td>
<td>0.47</td>
<td>0.09</td>
<td>0.03</td>
<td>-0.20</td>
<td>0.74</td>
</tr>
<tr>
<td>Was the care provider blinded?</td>
<td>0.13</td>
<td>0.82</td>
<td>0.05</td>
<td>0.64</td>
<td>-0.05</td>
<td>0.06</td>
<td>0.03</td>
<td>0.93</td>
</tr>
<tr>
<td>Were patients blinded?</td>
<td>0.13</td>
<td>0.82</td>
<td>0.05</td>
<td>0.64</td>
<td>-0.05</td>
<td>0.06</td>
<td>0.03</td>
<td>0.93</td>
</tr>
<tr>
<td>Was the dropout rate acceptable?</td>
<td>-0.02</td>
<td>0.76</td>
<td>0.04</td>
<td>0.53</td>
<td>-0.09</td>
<td>-0.03</td>
<td>0.11</td>
<td>0.87</td>
</tr>
<tr>
<td>Were all randomized participants analyzed in the group to which they were originally assigned?</td>
<td>0.12</td>
<td>0.02</td>
<td>0.76</td>
<td>0.34</td>
<td>0.04</td>
<td>0.15</td>
<td>-0.27</td>
<td>-0.04</td>
</tr>
<tr>
<td>Were co-interventions avoided or similar?</td>
<td>0.14</td>
<td>0.22</td>
<td>0.08</td>
<td>0.25</td>
<td>-0.18</td>
<td>0.07</td>
<td>-0.04</td>
<td>0.23</td>
</tr>
<tr>
<td>Was the compliance acceptable in all groups?</td>
<td>0.02</td>
<td>0.00</td>
<td>0.18</td>
<td>0.08</td>
<td>0.09</td>
<td>-0.04</td>
<td>0.13</td>
<td>-0.10</td>
</tr>
<tr>
<td>Was the timing of the outcome assessment similar in all groups?</td>
<td>0.11</td>
<td>0.31</td>
<td>0.08</td>
<td>0.30</td>
<td>0.06</td>
<td>0.02</td>
<td>0.10</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Note: in bold are correlations reflecting similar constructs.

The correspondence was not perfect, since although assessing similar quality domains, the scoring rules are not identical across quality scoring systems. For example, the CBRG provides a number of rules for when blinding can be assumed when blinding was not explicitly reported in the study (e.g., when automated data are concerned). Schulz’s item “Inclusion in the Analysis of All Randomized Participants” instructs that the item should be answered in the affirmative when the publication reports “or gives the impression” that no exclusions have taken place (“often not explicit”), whereas the corresponding CBRG item requires an explicit statement in the text or explicit data; otherwise the item will be scored “unclear.”

### Internal Consistency

In the back pain dataset, the Cronbach’s alpha coefficient of a summary scale was 0.56, in the EPC reports alpha was 0.55, and in the third dataset alpha was 0.61, indicating in all three datasets only moderate internal consistency. The level of consistency does not indicate that all items are measuring the same construct.
To investigate whether the internal consistency was mainly influenced by one or two outlying items, the alpha coefficient was computed excluding each item in turn (alpha if item deleted analysis), depicted in Table 7. This analysis can also show whether there are items that do not add any information that is already captured through other items (in that case, the scale alpha does not drop although the item is removed from the scale).

Table 7. Alpha if item deleted

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Scale α if Item Deleted</th>
<th>Scale α if Item Deleted</th>
<th>Scale α if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was the method of randomization adequate?</td>
<td>0.53</td>
<td>0.49</td>
<td>0.54</td>
</tr>
<tr>
<td>Was the treatment allocation concealed?</td>
<td>0.50</td>
<td>0.49</td>
<td>0.56</td>
</tr>
<tr>
<td>Were the groups similar at baseline regarding the most important prognostic indicators?</td>
<td>0.53</td>
<td>0.61</td>
<td>0.48</td>
</tr>
<tr>
<td>Was the outcome assessor blinded?</td>
<td>0.52</td>
<td>0.55</td>
<td>0.53</td>
</tr>
<tr>
<td>Was the care provider blinded to the intervention?</td>
<td>0.56</td>
<td>0.46</td>
<td>0.49</td>
</tr>
<tr>
<td>Were patients blinded?</td>
<td>0.57</td>
<td>0.47</td>
<td>0.51</td>
</tr>
<tr>
<td>Was the dropout rate described and acceptable?</td>
<td>0.53</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>Were all randomized participants analyzed in the group to which they were originally assigned?</td>
<td>0.52</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Were cointerventions avoided or similar?</td>
<td>0.51</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>Was the compliance acceptable in all groups?</td>
<td>0.52</td>
<td>0.55</td>
<td>0.61</td>
</tr>
<tr>
<td>Was the timing of the outcome assessment in all groups similar?</td>
<td>0.55</td>
<td>0.54</td>
<td>0.52</td>
</tr>
</tbody>
</table>

EPC = Evidence-based Practice Center

The “alpha if item removed” scores indicated in the back pain dataset that one of the blinding items may be unnecessary since its absence does not decrease alpha. All other items did not affect the total score substantially, there was also no indication that one particular item was the “odd one out,” not related to an overall quality score constituted by these 11 quality features. In the EPC projects, removing the items that concern the similarity of the baseline would raise alpha slightly, and the outcome assessor blinding item does not add any information that is not already captured by (presumably) the other blinding items. The removal of the compliance item would not lower the Cronbach’s alpha value in the third dataset, indicating that this item does not contribute to increased reliability of a total scale.

Factor Analysis

To investigate the structure of relationships between quality features we fitted one, two, and three factor models in a multiple indicator multiple cause (MIMIC) model to each dataset. The factors group related items, in terms of intercorrelations as well as in their strength of association with effect sizes. For each dataset, we estimated three models—a single quality factor model, two quality factors, and three quality factors. (Figure 3 shows a two quality factor model). Each model was assessed using a range of fit measures, which indicate the degree of misfit between the model and the data. The \( \chi^2 \) test should be nonsignificant in a well-fitting model, the
comparative fit index should be over 0.95, and the root mean square error of approximation should be less than 0.05.

In meta-analysis, the studies are weighted by the standard error of the estimate. In fitting a MIMIC model, techniques have not been developed that allow us to estimate weights for studies separately; hence, the relationships between effect size and quality factors should be interpreted in this light.

The fit indices for the datasets are shown in Table 8. In the back pain dataset, the patient blinding item and similar timing had to be removed due to collinearity, in the EPC report set the assessor blinding and similar timing had to be removed, and in the third dataset the assessor blinding item was dropped due to lack of variance. A model assuming three factors gave a good fit to the data in all three datasets.

<table>
<thead>
<tr>
<th>Fit Indices</th>
<th>Dataset 1: Back Pain</th>
<th>Dataset 2: EPC Reports</th>
<th>Dataset 3: “Pro-bias”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 Factor</td>
<td>2 Factor</td>
<td>3 Factor</td>
</tr>
<tr>
<td>$\chi^2$ (df)</td>
<td>169 (23)</td>
<td>N/C</td>
<td>19.1 (13) N/C</td>
</tr>
<tr>
<td>$p$</td>
<td>&lt;0.001</td>
<td>N/C</td>
<td>&lt;0.001 N/C</td>
</tr>
<tr>
<td>CFI</td>
<td>0.59</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.171</td>
<td>0.47</td>
<td>0.15</td>
</tr>
</tbody>
</table>

df = degrees of freedom; N/C = model failed to converge; CFI = comparative fit index; RMSEA = root mean square error of approximation

The factor loadings of the individual quality features on the latent factors are shown in Table 9. The largest loading of each item in each dataset is highlighted in bold. Factor loadings are the correlations of each quality feature with the factor. Factor loadings constitute the factors.

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Back Pain</th>
<th>EPC Reports</th>
<th>Published Set “Pro-bias”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>0.82</td>
<td>-0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>0.92</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>0.41</td>
<td>0.00</td>
<td>0.18</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>0.22</td>
<td>1.01</td>
<td>-0.08</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>-0.42</td>
<td>0.83</td>
<td>0.14</td>
</tr>
<tr>
<td>Patient blind</td>
<td>-0.07</td>
<td>0.99</td>
<td>0.06</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>0.31</td>
<td>-0.08</td>
<td>0.58</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>0.43</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>0.59</td>
<td>-0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>Similar timing</td>
<td>0.00</td>
<td>0.42</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: loadings in bold are significant (p<0.05).
EPC = Evidence-based Practice Center; F = factor; ITT = intention to treat

Factor one in the back pain dataset consisted mainly of the randomization sequence item, allocation concealment, and the compliance item. Baseline similarity and the similarity of co-interventions also loaded on this item. Factor two represented blinding. Factor three was made up of the acceptable dropout item across datasets and the ITT item (original group) in two out of the
three datasets. The correlations between factors were not statistically significant (F1, F2: -0.14 [p=.224]; F1, F3: 0.13 [p=.281]; F2, F3: 0.11 [p=0.252]).

In the EPC report, dataset factor one consisted of randomization sequence and allocation concealment. Items relating to blinding loaded on factor 2 as did the original group, co-intervention and similar baseline items. The only item with a high loading on factor 3 was the dropout measure. The correlations between factors were not statistically significant (F1, F2: 0.24 [p=.056]; F1, F3: 0.15 [p=.230]; F2, F3: 0.09 [p=.446]).

In the third dataset, the randomization item and allocation concealment loaded again on factor 1. Factor 3 was the blinding factor, but factor 2 consisted mainly of similar baseline and similar cointervention reporting and a couple of other items also loaded on this factor. The correlations between factors were not statistically significant (F1, F2: -0.22 [p=.553]; F1, F3: -0.20 [p=.392]; F2, F3: -0.42 [p=.019]).

Similarities across all three datasets were that the randomization and concealment items formed one “treatment allocation” factor. Blinding constituted another robust factor across datasets, independent from the treatment allocation factor. In each dataset a third factor had to be assumed accounting for additional variance not covered by the two other factors.

**Association Between Quality and Effect Sizes**

**Dataset 1: Back Pain Trials**

As reported previously (van Tulder et al., 2009), studies included in the CBRG that scored positive for a quality item reported smaller effect sizes compared with trials that did not fulfill the criterion. The differences were not statistically significant but the included features showed consistency across domains with 10 out of 11 features indicating this effect, as depicted in Table 10.

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ES</td>
<td>95% CI</td>
<td>ES</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>104</td>
<td>112</td>
<td>0.51 (0.41, 0.61)</td>
<td>0.49 (0.40, 0.59)</td>
<td>0.02 (-0.12, 0.16)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>69</td>
<td>147</td>
<td>0.45 (0.33, 0.57)</td>
<td>0.53 (0.45, 0.62)</td>
<td>-0.08 (-0.23, 0.07)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>135</td>
<td>81</td>
<td>0.47 (0.38, 0.55)</td>
<td>0.57 (0.45, 0.68)</td>
<td>-0.10 (-0.24, 0.05)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>123</td>
<td>93</td>
<td>0.46 (0.37, 0.55)</td>
<td>0.56 (0.46, 0.67)</td>
<td>-0.10 (-0.25, 0.04)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>57</td>
<td>159</td>
<td>0.43 (0.30, 0.56)</td>
<td>0.53 (0.45, 0.61)</td>
<td>-0.10 (-0.26, 0.06)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>82</td>
<td>134</td>
<td>0.48 (0.37, 0.60)</td>
<td>0.52 (0.43, 0.60)</td>
<td>-0.03 (-0.18, 0.11)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>150</td>
<td>66</td>
<td>0.46 (0.38, 0.55)</td>
<td>0.60 (0.47, 0.73)</td>
<td>-0.13 (-0.29, 0.02)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>118</td>
<td>98</td>
<td>0.46 (0.37, 0.55)</td>
<td>0.56 (0.45, 0.67)</td>
<td>-0.10 (-0.24, 0.04)</td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>92</td>
<td>124</td>
<td>0.45 (0.35, 0.56)</td>
<td>0.54 (0.45, 0.63)</td>
<td>-0.09 (-0.23, 0.05)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>76</td>
<td>140</td>
<td>0.50 (0.39, 0.61)</td>
<td>0.51 (0.42, 0.59)</td>
<td>-0.01 (-0.15, 0.14)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>198</td>
<td>18</td>
<td>0.49 (0.42, 0.56)</td>
<td>0.66 (0.40, 0.92)</td>
<td>-0.17 (-0.43, 0.10)</td>
</tr>
</tbody>
</table>

ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat
Figure 9 depicts the difference in effect sizes between studies meeting the individual quality criterion and those that do not. A negative effect size difference indicates possible bias; high-quality studies (those that meet the quality criterion) reported smaller effect sizes.

**Figure 9. Difference in effect sizes based on quality features dataset 1 (back pain)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Yes</th>
<th>Yes</th>
<th>Effect Size Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>104</td>
<td>112</td>
<td>0.02 (-0.12, 0.16)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>69</td>
<td>147</td>
<td>-0.08 (-0.23, 0.07)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>135</td>
<td>81</td>
<td>-0.10 (-0.24, 0.05)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>123</td>
<td>93</td>
<td>-0.10 (-0.25, 0.04)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>57</td>
<td>159</td>
<td>-0.10 (-0.26, 0.06)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>82</td>
<td>134</td>
<td>-0.03 (-0.18, 0.11)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>150</td>
<td>66</td>
<td>-0.13 (-0.29, 0.02)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>118</td>
<td>98</td>
<td>-0.10 (-0.24, 0.04)</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>92</td>
<td>124</td>
<td>-0.09 (-0.23, 0.05)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>76</td>
<td>140</td>
<td>-0.01 (-0.15, 0.14)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>198</td>
<td>18</td>
<td>-0.17 (-0.43, 0.10)</td>
</tr>
</tbody>
</table>

CI = confidence interval; ITT = intention to treat

In terms of an established quality measure, differences in effect sizes between low and high quality studies were -0.14 (p<0.05, random effects meta-regression) for items constituting the Jadad scale. High-quality studies reported a mean estimated effect size of 0.45, while low-quality studies reported a mean estimated effect size of 0.60. The fixed-effects model effect size difference was -0.09 (p<0.05).

**Summary and Factor Scores**

To explore a linear effect of quality on effect size, we regressed the effect sizes on a total quality score value which we had computed for each study. The total quality score was obtained by equally weighing each of the 11 quality components to contribute to an overall quality score. The linear effect of the total quality scores across studies was negligible and not statistically significant (-0.04, SE 0.018, p=0.053, 95% CI: -0.073, 0.005).

When effect size was regressed on each of the three factors established in the factor analysis (each factor representing empirical clusters of quality features), the results were also not statistically significant. Standardized effects were for factor 1: 0.07 (p=0.699), factor 2: -0.23 (p=0.077) and factor 3: -0.15 (p=0.245).

Statistically significant results of quality were shown when applying a cutoff of 5 or 6 quality items fulfilled, the difference in effect size between low and high quality studies was 0.20 (van Tulder et. al., 2009). For this analysis, we used a total quality score per study and applied a cut-off empirically distinguishing high- and low-quality studies, depicted in Table 11.
Table 11. Comparison of different quality cutoffs using a total score dataset 1 (back pain)

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Number Equal or Above Cutoff</th>
<th>Number Below Cutoff</th>
<th>High Quality</th>
<th>Low Quality</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>95% CI</td>
<td>ES</td>
<td>95% CI</td>
<td>ESDiff</td>
</tr>
<tr>
<td>≥9 vs &lt;9</td>
<td>0.46 (0.23, 0.69)</td>
<td>0.51 (0.43, 0.58)</td>
<td>-0.04 (-0.29, 0.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥8 vs &lt;8</td>
<td>0.43 (0.28, 0.58)</td>
<td>0.52 (0.44, 0.60)</td>
<td>-0.09 (-0.26, 0.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥7 vs &lt;7</td>
<td>0.44 (0.33, 0.56)</td>
<td>0.54 (0.45, 0.63)</td>
<td>-0.10 (-0.24, 0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥6 vs &lt;6</td>
<td>0.42 (0.33, 0.51)</td>
<td>0.62 (0.51, 0.73)</td>
<td>-0.20 (-0.34, -0.06)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥5 vs &lt;5</td>
<td>0.44 (0.36, 0.52)</td>
<td>0.64 (0.52, 0.76)</td>
<td>-0.20 (-0.35, -0.05)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥4 vs &lt;4</td>
<td>0.48 (0.41, 0.56)</td>
<td>0.61 (0.44, 0.77)</td>
<td>-0.13 (-0.31, 0.06)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05
ES = effect size; CI = confidence interval; ESDiff = effect size difference

Dataset 2: EPC reports

The mean treatment effect across all studies in the EPC report dataset was 0.43 (95% CI: 0.34, 0.53). Few quality features showed differences in effect sizes according to whether these criteria were met as depicted in Table 12. A negative difference indicates that the effect size for the studies fulfilling the criterion is smaller than the effect size for the studies not meeting the criterion. The “no” and “unclear” answers were combined for all initial analyses to increase the number of analyzable studies.

Table 12. Difference in effect sizes dataset 2 (EPC Report)

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>95% CI</td>
<td>ES</td>
<td>95% CI</td>
<td>ESDiff</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>44</td>
<td>121 0.44 (0.30, 0.57)</td>
<td>0.43 (0.34, 0.51)</td>
<td>0.01 (-0.15, 0.17)</td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>38</td>
<td>127 0.39 (0.25, 0.53)</td>
<td>0.45 (0.36, 0.53)</td>
<td>-0.05 (-0.22, 0.11)</td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>100</td>
<td>65 0.40 (0.31, 0.49)</td>
<td>0.49 (0.37, 0.61)</td>
<td>-0.09 (-0.24, 0.05)</td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>157</td>
<td>8 0.43 (0.36, 0.51)</td>
<td>0.37 (0.04, 0.71)</td>
<td>0.06 (-0.28, 0.41)</td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>120</td>
<td>45 0.48 (0.40, 0.56)</td>
<td>0.29 (0.15, 0.43)</td>
<td>0.19 (0.03, 0.35)*</td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>130</td>
<td>35 0.47 (0.40, 0.55)</td>
<td>0.26 (0.11, 0.42)</td>
<td>0.21 (0.04, 0.39)*</td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>96</td>
<td>69 0.50 (0.40, 0.59)</td>
<td>0.35 (0.24, 0.45)</td>
<td>0.15 (0.01, 0.29)*</td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>101</td>
<td>64 0.45 (0.36, 0.54)</td>
<td>0.40 (0.27, 0.52)</td>
<td>0.05 (-0.10, 0.20)</td>
<td></td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>142</td>
<td>23 0.44 (0.36, 0.52)</td>
<td>0.39 (0.20, 0.58)</td>
<td>0.05 (-0.15, 0.28)</td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>79</td>
<td>86 0.44 (0.34, 0.55)</td>
<td>0.42 (0.32, 0.52)</td>
<td>0.02 (-0.12, 0.17)</td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>161</td>
<td>4 0.44 (0.37, 0.51)</td>
<td>0.19 (-0.24, 0.62)</td>
<td>0.25 (-0.19, 0.69)</td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05
EPC = Evidence-based Practice Center; ES = effect size; CI = confidence interval; ESDiff = effect size difference; ITT = intention to treat

High- and low-quality studies showed no difference in effect sizes for several quality features and the direction of possible bias was not consistent across dimensions. This concerned newly proposed quality features as well as established quality criteria such as blinding. For three
criteria, a significant difference was found but the effect was discordant with previous studies: 
when care provider and patients were explicitly blinded, the average effect size in those trials 
was 0.48 and 0.47 compared to 0.30 and 0.27 in low or unclear quality trials. Studies that 
reported an acceptable dropout rate had an average effect size of 0.50; studies without 
description or adequate rate showed a mean effect of 0.35. Figure 10 depicts the direction of 
effects graphically. The difference in effect size is shown; a negative difference indicates that the 
high-quality studies in this dataset reported smaller effect sizes.

**Figure 10. Difference in effect sizes based on quality features dataset 2 (EPC reports)**

<table>
<thead>
<tr>
<th>Study</th>
<th>Number</th>
<th>Effect Size Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>44 121</td>
<td>0.01 (-0.15, 0.17)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>38 127</td>
<td>-0.05 (-0.22, 0.11)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>100 65</td>
<td>-0.09 (-0.24, 0.05)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>157 8</td>
<td>0.06 (-0.28, 0.41)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>120 45</td>
<td>0.19 (0.03, 0.35)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>130 35</td>
<td>0.21 (0.04, 0.39)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>96 69</td>
<td>0.15 (0.01, 0.29)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>101 64</td>
<td>0.05 (-0.10, 0.20)</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>142 23</td>
<td>0.05 (-0.15, 0.28)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>79 86</td>
<td>0.02 (-0.12, 0.17)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>161 4</td>
<td>0.25 (-0.19, 0.69)</td>
</tr>
</tbody>
</table>

CI = confidence interval; ITT = intention to treat

Using the items constituting the Jadad scale (randomization, blinding, and dropouts),
differences in effect sizes between low- and high-quality studies were 0.09 (n.s., random effects 
meta-regression). High-quality studies reported an effect size of 0.45 while low quality reported 
an effect size of 0.35 across studies. Using a fixed-effects model, the mean effect sizes were 1.07 
versus 0.13 (overall effect size difference 0.94, p<0.05). The fixed-effects analysis is particularly 
fected by outliers. When excluding those three studies with extremely high and unmatched 
results, effect size differences were still 0.29. All analyses indicated that in this dataset there was 
o statistically significant difference in high- and low-quality studies, and often high-quality 
studies reported somewhat larger treatment effects, hence opposite to what we expected to find.

As shown earlier, in this dataset there was a high number of “unclear” answers. To 
investigate whether the combination of “no’s” and “unclear’s” may have distorted the effects of 
quality, we estimated the results separately, and compared the explicit negative and the unclear 
cases to the cases where the feature was explicitly present, that is, the criterion was clearly met, 
as depicted in Table 13.
### Table 13. Criterion met versus not met and versus unclear (EPC reports)

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Met</th>
<th>Not</th>
<th>Unclear</th>
<th>Difference Criterion Met Versus Not Met</th>
<th>Difference Criterion Met Versus Unclear</th>
<th>Difference Criterion Met or Unclear Versus Not Met</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ESdiff</td>
<td>95% CI</td>
<td>ESdiff</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>44</td>
<td>7</td>
<td>114</td>
<td>0.23</td>
<td>(-0.19, 0.65)</td>
<td>0.00</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>38</td>
<td>4</td>
<td>123</td>
<td>0.35</td>
<td>(-0.20, 0.90)</td>
<td>-0.07</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>100</td>
<td>17</td>
<td>48</td>
<td>-0.10</td>
<td>(-0.35, 0.14)</td>
<td>-0.09</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>157</td>
<td>2</td>
<td>6</td>
<td>-0.93</td>
<td>(-1.57, -0.30)*</td>
<td>0.40</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>120</td>
<td>33</td>
<td>12</td>
<td>0.11</td>
<td>(-0.07, 0.30)</td>
<td>0.38</td>
</tr>
<tr>
<td>Patient blind</td>
<td>130</td>
<td>31</td>
<td>4</td>
<td>0.19</td>
<td>(0.01, 0.36)*</td>
<td>0.38</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>96</td>
<td>65</td>
<td>4</td>
<td>0.14</td>
<td>(-0.01, 0.28)</td>
<td>0.33</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>101</td>
<td>10</td>
<td>54</td>
<td>0.08</td>
<td>(-0.24, 0.39)</td>
<td>0.05</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>142</td>
<td>5</td>
<td>19</td>
<td>-0.17</td>
<td>(-0.59, 0.25)</td>
<td>0.11</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>79</td>
<td>8</td>
<td>78</td>
<td>0.25</td>
<td>(-0.06, 0.47)</td>
<td>-0.01</td>
</tr>
<tr>
<td>Similar timing</td>
<td>161</td>
<td>4</td>
<td>0</td>
<td>0.25</td>
<td>(-0.19, 0.69)</td>
<td>NA</td>
</tr>
</tbody>
</table>

* p<0.05

EPC = Evidence-based Practice Center; ESdiff = effect size difference; CI = confidence interval; ITT = intention to treat

When combining high-quality studies and studies rated as unclear, thereby giving the benefit of the doubt, the sample sizes for negative studies are very small. This stratified analysis showed that the lack of association of the feature with effect sizes cannot be generally explained by the combination of explicitly negative and unclear answers.

We also investigated in this dataset whether the conversion of dichotomous outcomes to effect sizes that was necessary in some studies may have influenced the associations between quality features and study results. Only considering original continuous outcomes, the differences between low and high quality studies ranged from -0.12 (similar baseline) to 0.33 (similar timing of outcome assessment) where a negative difference indicates that the studies with the feature showed smaller effect sizes. The 0.33 difference was based on 3 studies only where the similar outcome assessment criterion was not met or remained unclear.

### Summary and Factor Scores

To explore a linear effect of quality on effect size, we again regressed the effect sizes on the total quality score values. The effect of the total quality scores weighting each item equally was negligible and not significant (0.02; p=0.233, 95% CI: -0.015, 0.062).

We also used the established factor scores that group similar items in terms of inter-item correlations as well as associations with effect sizes. Differential trends shown for the individual items as seen in the table above should become apparent using these factor empirically derived item clusters. Raw effect size (not accounting for precision with any weights) was regressed on the established factors. High-loading items from factor 1 (adequate randomization sequence, allocation concealment) were not statistically significantly associated with effect sizes (-0.172, 95% CI: -0.41, 0.06; p=0.15), nor was describing an acceptable dropout rate which mainly contributed factor 3 (0.14; 95% CI: -0.06, 0.49; p=0.16). The two blinding items and the ITT
item that constitute factor 2 showed a statistically significant influence (but unexpected directionality) on effect sizes (0.24, 95% CI: 0.04, 0.44; p=0.02).

As depicted in Table 14, when comparing studies with high or low quality and applying different cutoffs, we found a marginal significant difference in effect sizes for five or more quality criteria met. However, the direction of effects was opposite to what we found in the back pain trials: high quality studies reported larger treatment effects. There was no indication that low-quality studies overestimated treatment effects; in this dataset the high-quality RCTs reported larger effects. Appendix D shows the results based on a fixed-effects model (this analysis results in smaller confidence intervals and several significant results, but the analysis is more affected by outliers).

### Table 14. Comparison of different quality cutoffs using a total score (EPC reports)

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Number Equal or Above Cutoff</th>
<th>Number Below Cutoff</th>
<th>High Quality</th>
<th>Low Quality</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ESdiff 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥9 vs &lt;9</td>
<td>42  (0.40, 0.67)</td>
<td>0.39 (0.31, 0.47)</td>
<td>0.15 (-0.01, 0.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥8 vs &lt;8</td>
<td>65  (0.34, 0.56)</td>
<td>0.42 (0.32, 0.51)</td>
<td>0.03 (-0.11, 0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥7 vs &lt;7</td>
<td>103 (0.35, 0.52)</td>
<td>0.43 (0.31, 0.55)</td>
<td>0.01 (-0.14, 0.16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥6 vs &lt;6</td>
<td>135 (0.38, 0.53)</td>
<td>0.46 (0.12, 0.47)</td>
<td>0.16 (-0.03, 0.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥5 vs &lt;5</td>
<td>149 (0.38, 0.53)</td>
<td>0.45 (0.18, 0.42)</td>
<td>0.27 (0.02, 0.52)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥4 vs &lt;4</td>
<td>160 (0.37, 0.51)</td>
<td>0.44 (0.04, 0.46)</td>
<td>0.41 (-0.02, 0.84)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05

EPC = Evidence-based Practice Center; ES = effect size; CI = confidence interval; ESdiff = effect size difference

The above table also clearly demonstrates the imbalance of this dataset. There were very few low-quality studies in this dataset. When comparing studies that only reached 4 out of 11 possible quality scores, only 5 studies could be included in the analysis.

**Dataset 3: Published “Pro-bias” Dataset**

Given the discrepant results between the analyses of the back pain dataset and the EPC reports dataset, we decided to add a third dataset. We were struck, in particular, by the observation that in the EPC dataset for the majority of quality features we actually found results in the opposite direction as expected from prior research (high-quality studies reported larger effect sizes). As outlined in the method section for our third dataset, we therefore determined that we should use one where established criteria such as the Jadad and Schulz items had known values in the expected direction. For that reason, we decided to use a replication of the dataset used by Moher and colleagues in their original validation of these quality features.

As opposed to the prior two datasets, all the outcomes in this dataset used dichotomous outcomes. Therefore, instead of a difference in effect sizes we use, as Moher and colleagues did in their original analysis, the odds ratio as the measure of effect and the ratio of odds as assessment of the effect of a quality criterion across studies. The overall odds ratio across studies was 0.47 (95% CI: 0.42, 0.52). An odds ratio below 1 indicates the treatment group is doing better than the control.

In this dataset we also applied the full Jadad scale and the criteria proposed by Schulz (1995), which included concealment of allocation using the original scoring instructions. We compared adequate randomization (score=2 versus <2), blinding (score=2 versus <2) and withdrawals
(score 1 versus 0) and a total score of 3 or more (out of 5) compared to less than 3 for the total Jadad score. For the Schulz scores, we compared criterion met versus not met or unclear. Table 15 shows the results for these established quality criteria.

**Table 15. Difference in odds ratios for Jadad and Schulz criteria**

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR 95% CI</td>
<td>OR 95% CI</td>
<td>ROR 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Random Effects Meta-regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jadad: randomization=2</td>
<td>33 67</td>
<td>0.48 (0.36, 0.57)</td>
<td>0.45 (0.35, 0.57)</td>
<td>0.95 (0.62, 1.44)</td>
<td></td>
</tr>
<tr>
<td>Jadad: blinding=2</td>
<td>36 64</td>
<td>0.43 (0.31, 0.61)</td>
<td>0.47 (0.37, 0.60)</td>
<td>1.08 (0.72, 1.64)</td>
<td></td>
</tr>
<tr>
<td>Jadad: withdrawals=1</td>
<td>74 26</td>
<td>0.48 (0.38, 0.60)</td>
<td>0.41 (0.29, 0.60)</td>
<td>0.87 (0.56, 1.34)</td>
<td></td>
</tr>
<tr>
<td>Jadad: total ≥3</td>
<td>62 38</td>
<td>0.46 (0.36, 0.60)</td>
<td>0.45 (0.33, 0.61)</td>
<td>0.97 (0.65, 1.45)</td>
<td></td>
</tr>
<tr>
<td>Schulz: concealment</td>
<td>26 74</td>
<td>0.58 (0.40, 0.86)</td>
<td>0.42 (0.34, 0.53)</td>
<td>0.72 (0.46, 1.13)</td>
<td></td>
</tr>
<tr>
<td>Schulz: sequence</td>
<td>30 70</td>
<td>0.45 (0.31, 0.65)</td>
<td>0.46 (0.36, 0.58)</td>
<td>1.01 (0.66, 1.56)</td>
<td></td>
</tr>
<tr>
<td>Schulz: analysis</td>
<td>41 59</td>
<td>0.50 (0.37, 0.68)</td>
<td>0.43 (0.33, 0.55)</td>
<td>0.85 (0.57, 1.26)</td>
<td></td>
</tr>
<tr>
<td>Schulz: blinding</td>
<td>66 34</td>
<td>0.44 (0.34, 0.56)</td>
<td>0.50 (0.35, 0.69)</td>
<td>1.13 (0.74, 1.71)</td>
<td></td>
</tr>
<tr>
<td><strong>Fixed Effects Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jadad: randomization=2</td>
<td>33 67</td>
<td>0.51 (0.43, 0.61)</td>
<td>0.45 (0.40, 0.51)</td>
<td>0.88 (0.70, 1.09)</td>
<td></td>
</tr>
<tr>
<td>Jadad: blinding=2</td>
<td>36 64</td>
<td>0.45 (0.37, 0.54)</td>
<td>0.47 (0.42, 0.53)</td>
<td>1.05 (0.85, 1.31)</td>
<td></td>
</tr>
<tr>
<td>Jadad: withdrawals=1</td>
<td>74 26</td>
<td>0.52 (0.46, 0.59)</td>
<td>0.39 (0.33, 0.46)</td>
<td>0.74 (0.60, 0.92)*</td>
<td></td>
</tr>
<tr>
<td>Jadad: total ≥3</td>
<td>62 38</td>
<td>0.51 (0.44, 0.58)</td>
<td>0.43 (0.37, 0.49)</td>
<td>0.85 (0.69, 1.03)</td>
<td></td>
</tr>
<tr>
<td>Schulz: concealment</td>
<td>26 74</td>
<td>0.57 (0.46, 0.69)</td>
<td>0.44 (0.39, 0.49)</td>
<td>0.77 (0.61, 0.97)*</td>
<td></td>
</tr>
<tr>
<td>Schulz: sequence</td>
<td>30 70</td>
<td>0.50 (0.41, 0.60)</td>
<td>0.46 (0.41, 0.51)</td>
<td>0.92 (0.74, 1.15)</td>
<td></td>
</tr>
<tr>
<td>Schulz: analysis</td>
<td>41 59</td>
<td>0.51 (0.44, 0.59)</td>
<td>0.43 (0.38, 0.50)</td>
<td>0.85 (0.70, 1.04)</td>
<td></td>
</tr>
<tr>
<td>Schulz: blinding</td>
<td>66 34</td>
<td>0.48 (0.42, 0.55)</td>
<td>0.44 (0.38, 0.52)</td>
<td>0.92 (0.75, 1.12)</td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios

As the original paper by Moher and colleagues does not specify whether they used a fixed-effects or a random-effects model, we present the results in Table 15 using both methods. Our fixed-effect results come closest to the original results presented by Moher et al. (1998). The dimensions overall showed consistent results, with the ratio of odds ratios (ROR) for low-quality studies compared to high-quality studies being less than 1, meaning high-quality studies reported smaller treatment effects (i.e., larger ORs) than did low-quality studies. Using a fixed-effects model, the concealment criterion item of Schulz is statistically significantly associated with smaller treatment effects, and the Jadad scale is nearly so (ROR = 0.85, 95% CI: 0.69, 1.03). Using the random effects model, both point estimates go in the expected direction, but neither result is statistically significantly different from an ROR of 1.

Table 16 shows the odds ratios of studies where the criterion is met and odds ratios for studies where the criterion is not met, either because of poor reporting or due to the design, conduct, or analysis of the individual study, using a fixed-effects model. The corresponding results using a meta-regression model assuming random effects are shown in Appendix E. To assess the difference between these two study types, we estimated an ROR. Again, an ROR less
than 1 indicated that the studies that did not meet the quality criteria had a better treatment effect than those studies that met the quality criteria.

Table 16. Difference in odds ratios for proposed quality criteria dataset 3 ("pro-bias")

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR 95% CI</td>
<td>OR 95% CI</td>
<td>ROR 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>34 66</td>
<td>0.49 (0.41, 0.59)</td>
<td>0.46 (0.41, 0.52)</td>
<td>0.94 (0.75, 1.17)</td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>26 74</td>
<td>0.50 (0.41, 0.62)</td>
<td>0.46 (0.41, 0.51)</td>
<td>0.91 (0.72, 1.14)</td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>36 64</td>
<td>0.47 (0.40, 0.56)</td>
<td>0.46 (0.41, 0.52)</td>
<td>0.98 (0.80, 1.21)</td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>78 22</td>
<td>0.44 (0.39, 0.49)</td>
<td>0.59 (0.47, 0.74)</td>
<td>1.35 (1.05, 1.73)</td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>69 31</td>
<td>0.50 (0.41, 0.57)</td>
<td>0.41 (0.35, 0.49)</td>
<td>0.83 (0.67, 1.02)</td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>72 28</td>
<td>0.47 (0.42, 0.53)</td>
<td>0.46 (0.38, 0.55)</td>
<td>0.97 (0.78, 1.21)</td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>62 38</td>
<td>0.54 (0.47, 0.62)</td>
<td>0.39 (0.34, 0.46)</td>
<td>0.72 (0.59, 0.88)</td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>29 71</td>
<td>0.49 (0.42, 0.58)</td>
<td>0.45 (0.40, 0.51)</td>
<td>0.91 (0.74, 1.12)</td>
<td></td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>68 32</td>
<td>0.40 (0.35, 0.46)</td>
<td>0.60 (0.51, 0.71)</td>
<td>1.50 (1.22, 1.85)</td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>46 54</td>
<td>0.56 (0.48, 0.66)</td>
<td>0.41 (0.36, 0.46)</td>
<td>0.72 (0.59, 0.88)</td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>89 11</td>
<td>0.45 (0.41, 0.50)</td>
<td>0.60 (0.43, 0.84)</td>
<td>1.33 (0.94, 1.88)</td>
<td></td>
</tr>
</tbody>
</table>

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

The direction of effects for the 11 quality features was not uniform but for the majority of quality domains the low quality studies reported smaller odds ratios thereby overestimating the treatment effect. Differences between low- and high-quality studies were statistically significant for the quality items assessor blinding, acceptable dropout rate, similar co-interventions, and acceptable compliance. Figure 11 below displays the effects graphically.

Figure 11. Ratio of odds ratio based on quality features dataset 3 ("pro-bias"), FE

<table>
<thead>
<tr>
<th>Study</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Ratio of Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>34 66</td>
<td></td>
<td>0.94 (0.75, 1.17)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>26 74</td>
<td></td>
<td>0.91 (0.72, 1.14)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>36 64</td>
<td></td>
<td>0.98 (0.80, 1.21)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>78 22</td>
<td></td>
<td>1.35 (1.05, 1.73)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>69 31</td>
<td></td>
<td>0.83 (0.67, 1.02)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>72 28</td>
<td></td>
<td>0.97 (0.78, 1.21)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>70 30</td>
<td></td>
<td>0.72 (0.59, 0.88)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>29 71</td>
<td></td>
<td>0.91 (0.74, 1.12)</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>68 32</td>
<td></td>
<td>1.50 (1.22, 1.85)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>46 54</td>
<td></td>
<td>0.72 (0.59, 0.88)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>89 11</td>
<td></td>
<td>1.33 (0.94, 1.88)</td>
</tr>
</tbody>
</table>

FE = based on fixed-effects model; CI = confidence interval; ITT = intention to treat
Summary and Factor Scores

We also computed total quality scores for each study based on the 11 assessed quality features. Using this summary score and regressing effect size on quality, thereby assuming a linear relationship between the two variables, we find no statistically significant effect (estimate = 0.033, p=0.486, 95% CI: -0.061, 0.127) for total quality, regardless of the employed meta-regression model.

Using the factor scores that group related items, in terms of intercorrelations as well as in their strength of association with study results (the MIMIC model) we can show differences in groups of quality measures. The regression effect of log odds ratios on the randomization factor (high loadings in this dataset: adequate randomization sequence generation, concealed treatment allocation; Factor 1) was 0.05, and was not significant (p=0.792). The effect of the blinding factor (high loadings in this dataset: provider blind, patient blind, assessor blind) was also not statistically significant (-0.06, p = 0.71), again regardless of the employed method. Factor 3 (high loadings in this dataset: acceptable dropout rate, similar baseline, original group (ITT), similar cointerventions, and similar timing) had a marginally nonsignificant, and negative effect on the log odds ratios (-0.22, p=0.07).

When comparing studies with high or low quality, and applying different cutoffs, and applying the proposed quality criteria of the Cochrane back review group to compute a total score, we find that based on a fixed-effects model, a cutoff of 5 or 6 differentiates the studies statistically significant as shown in Table 17.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Number Equal or Above Cutoff</th>
<th>Number Below Cutoff</th>
<th>High Quality</th>
<th>Low Quality</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR 95% CI</td>
<td>OR 95% CI</td>
<td>ROR 95% CI</td>
</tr>
<tr>
<td>≥9 vs &lt;9</td>
<td>14</td>
<td>86</td>
<td>0.43 (0.32, 0.57)</td>
<td>0.47 (0.42, 0.52)</td>
<td>1.09 (0.81, 1.46)</td>
</tr>
<tr>
<td>≥8 vs &lt;8</td>
<td>26</td>
<td>74</td>
<td>0.53 (0.43, 0.64)</td>
<td>0.45 (0.40, 0.50)</td>
<td>0.85 (0.68, 1.07)</td>
</tr>
<tr>
<td>≥7 vs &lt;7</td>
<td>44</td>
<td>56</td>
<td>0.45 (0.39, 0.53)</td>
<td>0.48 (0.42, 0.54)</td>
<td>1.05 (0.86, 1.29)</td>
</tr>
<tr>
<td>≥6 vs &lt;6</td>
<td>62</td>
<td>38</td>
<td>0.52 (0.45, 0.59)</td>
<td>0.40 (0.34, 0.47)</td>
<td>0.77 (0.63, 0.95)*</td>
</tr>
<tr>
<td>≥5 vs &lt;5</td>
<td>76</td>
<td>24</td>
<td>0.50 (0.44, 0.56)</td>
<td>0.39 (0.32, 0.48)</td>
<td>0.79 (0.63, 0.99)*</td>
</tr>
<tr>
<td>≥4 vs &lt;4</td>
<td>86</td>
<td>14</td>
<td>0.48 (0.43, 0.54)</td>
<td>0.40 (0.32, 0.50)</td>
<td>0.83 (0.65, 1.07)</td>
</tr>
</tbody>
</table>

* p<0.05

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios

High-quality studies show less pronounced treatment effects compared to low-quality studies. The equivalent analysis using a random effects meta-regression is reported in Appendix E. Using this model, a cutoff of 7 differentiates high- and low-quality studies best, but this result does not reach statistical significance.
Comparison Across Datasets

Figure 12 allows a comparison of observed indicators of bias for each individual quality feature across datasets.

**Figure 12. Differences in effect sizes across datasets**

Left to right: back pain data, EPC reports, “pro-bias” (fixed-effects model), “pro-bias” (random-effects model)
ITT = intention to treat

While the back pain dataset shows small but consistent results across quality criteria indicating that studies fulfilling the quality criterion report smaller effect sizes, the EPC dataset indicate for the majority of quality dimensions that high-quality studies reported larger treatment effects. The third dataset shows the most variation across quality criteria. In the fixed-effects analysis, the differences across high- and low-quality studies reach statistical significance.

Only allocation concealment showed consistent results across datasets. Unconcealed trials reported larger effect sizes in the back pain dataset (effect size difference in random effects meta-regression -0.08 (95% CI: -0.23, 0.07) and the EPC report dataset (effect size difference in random effects meta-regression -0.06 (95% CI: -0.22, 0.11), and the ratio of odds ratios in the third dataset also showed that unconcealed trials reported larger treatment effects (ROR=0.91, 95% CI: 0.72, 1.14). This analysis indicates that unconcealed trials tend to overestimate the treatment effect.

Across all three datasets, there was no statistically significant linear effect of quality on effect sizes. Regression models could not show that the effect size decreased linearly with increasing total quality scores.

The factors derived through factor analysis showed no statistically significant association between quality and effect sizes. One exception was the blinding/ITT factor (factor 2 in the EPC report data), but here with unexpected directionality (larger effects observed in high-quality studies).

Comparing different cutoffs shows that five or more fulfilled criteria differentiate high- and low-quality studies best across datasets. In dataset 1 (back pain), the difference between effect sizes was -0.20 for both, five criteria fulfilled or more (95% CI: -0.34, -0.06), and six criteria
fulfilled or more (95% CI: -0.35, -0.05). In the third dataset, the ratio of odds ratios was 0.79 (95% CI: 0.63, 0.95) and 0.77 (95% CI: 0.63, 0.99) respectively for five and six criteria met (based on a fixed-effects model). In both cases, low-quality studies overestimated treatment effects. However, in the EPC report dataset, a cutoff of five or more quality criteria met also resulted in a statistically significant result (effect size difference 0.27, 95% CI: 0.02, 0.52), but the direction was opposite to our expectations. Low-quality studies did not overestimate treatment effects but reported smaller effect sizes than high-quality trials in this dataset.

**Why the Association Between Quality Features and Effect Sizes Might Vary Across Datasets: Moderators and Confounders**

As seen above, the association between quality features and effect sizes varies across the three employed datasets. In two of the datasets, the Jadad and Schulz criteria show associations that are consistent with that found by others, namely that higher quality studies have smaller estimates of effect. In these two datasets, the CBRG internal validity items also, in general, show the predicted relationships between quality and effect size, and in each a summary score of the 11 items is useful for distinguishing high- and low-quality trials (cutoff 5 to 7 quality items met). However, in the EPC dataset, the majority of quality features show either no relationship or a paradoxical relationship with effect size. To try and understand why these differences might exist, we undertook an analysis looking at moderators and confounders based on our conceptual models outlined in the method section.

We wanted to investigate the effect of the size of the treatment effect, the effect of the condition being treated, the effect of the type of outcome, and the effect of the observed variance in quality features.

**Variable 1: Size of Treatment Effect**

One variable we pursued was the size of the treatment effect reported in each individual meta-analysis. Our hypothesis was that if a treatment is very effective, this may minimize any associations between quality and outcomes. Depending on the type of intervention, the achieved effect can vary systematically across studies, thereby possibly confounding an effect of the association of quality and study results across studies. We added the treatment effect observed in each individual meta-analysis to the regression model.

Table 18 shows the differences between low- and high-quality studies for the EPC report dataset when controlling for the mean treatment effect in each meta-analysis.
There is no indication that controlling for treatment effect size reveals associations of quality and effect sizes. In fact, controlling for this variable eliminates differences between high- and low-quality studies, effect size differences range around zero. The differential effect of possible bias (sometimes indicating that low-quality studies show larger effect sizes than high-quality studies, sometimes indicating that high-quality studies show larger effect sizes) that characterizes the EPC report dataset appears to be primarily based on this treatment effect variable.

A similar result was observed in the third dataset, which showed similar results overall to the original back pain results. The differences in effect sizes comparing high- and low-quality studies are negligible for several quality domains with the exception of reported provider blinding, acceptable dropout rate and the ITT item (original group) when comparing for size of treatment effect, as shown in Table 19.

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ESdiff 95% CI</td>
<td>ESdiff 95% CI</td>
<td></td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>44 (0.31, 0.51)</td>
<td>0.41 (0.35, 0.48)</td>
<td>0.00 (-0.12, 0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>38 (0.29, 0.50)</td>
<td>0.42 (0.36, 0.48)</td>
<td>-0.02 (-0.14, 0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>100 (0.31, 0.46)</td>
<td>0.44 (0.35, 0.53)</td>
<td>-0.04 (-0.16, 0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>157 (0.35, 0.46)</td>
<td>0.53 (0.28, 0.79)</td>
<td>-0.13 (-0.39, 0.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>120 (0.34, 0.47)</td>
<td>0.44 (0.33, 0.55)</td>
<td>-0.03 (-0.17, 0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>130 (0.35, 0.47)</td>
<td>0.42 (0.30, 0.55)</td>
<td>-0.01 (-0.16, 0.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>96 (0.34, 0.48)</td>
<td>0.42 (0.33, 0.50)</td>
<td>0.00 (-0.12, 0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>101 (0.34, 0.47)</td>
<td>0.43 (0.33, 0.52)</td>
<td>-0.02 (-0.14, 0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>142 (0.35, 0.47)</td>
<td>0.43 (0.28, 0.57)</td>
<td>-0.02 (-0.18, 0.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>79 (0.33, 0.48)</td>
<td>0.42 (0.35, 0.50)</td>
<td>-0.02 (-0.13, 0.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>161 (0.36, 0.47)</td>
<td>0.30 (-0.02, 0.62)</td>
<td>0.12 (-0.21, 0.44)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat
### Table 19. Controlling for size of treatment effect dataset 3 (“pro-bias”)

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>OR in Trials WithCriterion Met</th>
<th>OR in Trials WithCriterion Not Met</th>
<th>OR Difference</th>
<th>95% CI</th>
<th>95% CI</th>
<th>ROR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>34</td>
<td>66</td>
<td>0.44 (0.31, 0.63)</td>
<td>0.46 (0.37, 0.59)</td>
<td>1.02</td>
<td>(0.72, 1.43)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>26</td>
<td>74</td>
<td>0.42 (0.31, 0.58)</td>
<td>0.44 (0.37, 0.52)</td>
<td>1.04</td>
<td>(0.72, 1.49)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>36</td>
<td>64</td>
<td>0.44 (0.34, 0.57)</td>
<td>0.44 (0.36, 0.53)</td>
<td>0.99</td>
<td>(0.72, 1.35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>78</td>
<td>22</td>
<td>0.44 (0.37, 0.52)</td>
<td>0.44 (0.31, 0.62)</td>
<td>0.99</td>
<td>(0.67, 1.47)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>69</td>
<td>31</td>
<td>0.48 (0.40, 0.57)</td>
<td>0.36 (0.28, 0.47)</td>
<td>0.75</td>
<td>(0.55, 1.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>72</td>
<td>28</td>
<td>0.44 (0.37, 0.520</td>
<td>0.43 (0.33, 0.58)</td>
<td>0.99</td>
<td>(0.70, 1.39)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>62</td>
<td>38</td>
<td>0.50 (0.41, 0.61)</td>
<td>0.38 (0.30, 0.47)</td>
<td>0.75</td>
<td>(0.56, 1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>29</td>
<td>71</td>
<td>0.53 (0.41, 0.67)</td>
<td>0.39 (0.33, 0.47)</td>
<td>0.74</td>
<td>(0.55, 1.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>68</td>
<td>32</td>
<td>0.42 (0.35, 0.51)</td>
<td>0.47 (0.36, 0.61)</td>
<td>1.12</td>
<td>(0.81, 1.55)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>46</td>
<td>54</td>
<td>0.46 (0.37, 0.58)</td>
<td>0.42 (0.34, 0.51)</td>
<td>0.91</td>
<td>(0.67, 1.24)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>89</td>
<td>11</td>
<td>0.44 (0.37, 0.52)</td>
<td>0.42 (0.25, 0.70)</td>
<td>0.96</td>
<td>(0.56, 1.64)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

Using summary quality scores and regressing study results on quality, we find in both new datasets no significant results that indicate a significant linear relationship between these two variables. When controlling for the mean treatment effect of each of the 12 meta-analyses in the EPC reports dataset, the regression results are still -0.02 (p=0.38; 95% CI: -0.06, 0.02) as opposed to 0.02. When controlling for the mean effect size of each of the 11 meta-analyses in the third dataset, results are also unchanged: 0.04 (p=0.376; 95% CI: -0.05, 0.13), previously 0.03.

### Variable 2: Condition Being Treated

In the EPC report dataset, we found no clear associations between quality and effect size across all studies. Hence, we wanted to investigate whether pooling across meta-analyses masks associations between the quality features and effect sizes. In order to see whether the associations between quality and effect sizes are consistent or notably different across clinical fields, we stratified the studies by the condition being treated or the clinical field. Only this dataset was considered suitable for this analysis (see method section).

Table 20 shows the effect size difference for high- (criterion fulfilled) and low- (criterion not fulfilled) quality studies for each meta-analysis individually for the EPC report dataset studies. Each cell had to have at least three trials with the feature present versus absent or unclear to be estimated.
Table 20. Effect size differences studies fulfilling criterion versus not by clinical field (EPC reports)

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Alzheimer's</th>
<th>Arthritis</th>
<th>CDSM</th>
<th>Chromium</th>
<th>Epilepsy</th>
<th>Glucosamine</th>
<th>OCD</th>
<th>Omega 3</th>
<th>Orlistat</th>
<th>SAMe</th>
<th>SMBG</th>
<th>Vitamin E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>.04</td>
<td>n/a</td>
<td>-0.08</td>
<td>n/a</td>
<td>.00</td>
<td>-10</td>
<td>-12</td>
<td>-.01</td>
<td>.32</td>
<td>-.03</td>
<td>n/a</td>
<td>-.17</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>.07</td>
<td>n/a</td>
<td>-0.26</td>
<td>n/a</td>
<td>.00</td>
<td>-19</td>
<td>-53</td>
<td>.01</td>
<td>.32</td>
<td>n/a</td>
<td>n/a</td>
<td>.3</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>.08</td>
<td>n/a</td>
<td>-0.04</td>
<td>-.47</td>
<td>-.11</td>
<td>-.20</td>
<td>.28</td>
<td>-.10</td>
<td>n/a</td>
<td>-.53</td>
<td>.18</td>
<td>-.41</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>-.02</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>-.36</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.14</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.14</td>
</tr>
<tr>
<td>Patient blind</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>-.12</td>
<td>n/a</td>
<td>-.03</td>
<td>.36</td>
<td>-.05</td>
<td>-.09</td>
<td>n/a</td>
<td>.16</td>
<td>n/a</td>
<td>-.13</td>
<td>-.01</td>
<td>.38</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>n/a</td>
<td>n/a</td>
<td>0.31</td>
<td>-.45</td>
<td>-.04</td>
<td>-.12</td>
<td>-.02</td>
<td>-.04</td>
<td>-.85</td>
<td>-.05</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>n/a</td>
<td>n/a</td>
<td>0.04</td>
<td>n/a</td>
<td>.17</td>
<td>-.02</td>
<td>.79</td>
<td>-.23</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.13</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>n/a</td>
<td>n/a</td>
<td>0.01</td>
<td>.45</td>
<td>-.03</td>
<td>-.37</td>
<td>.18</td>
<td>-.15</td>
<td>-.02</td>
<td>.17</td>
<td>.1</td>
<td>-.02</td>
</tr>
<tr>
<td>Similar timing</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.22</td>
<td>n/a</td>
</tr>
</tbody>
</table>

EPC = Evidence-based Practice Center; n/a = not available (fewer than three trials in each group); ITT = intention to treat; CDSM: chronic disease self-management; OCD: obsessive-compulsive-disorder; SAMe: S-adenosylmethionine; SMBG: self-monitoring of blood glucose

The table shows that in each meta-analysis, many comparisons could not be computed due to lack of variance across studies in these smaller units. Very often there was not enough information to judge the criterion, meaning fewer than three studies in each meta-analysis scored definitely positive. For some quality features (e.g., was the timing of outcome assessment similar in both groups) there were no studies not meeting this criterion, so again the difference in effect sizes could not be computed.

There was no clear support for the hypothesis of the condition being treated masking the associations of quality through pooling. The quality effects were not confounded as outlined in the method section; the pooling appeared to cancel out conflicting effects across fields resulting in the negligible pooled effects seen. The effect of the quality features varied across clinical fields in that a quality criterion that was met sometimes indicated smaller effect sizes and sometimes larger effect sizes. For example, in most fields, the adequacy of randomization sequence showed a small difference between high- and low-quality studies in the direction that the high-quality studies reported smaller effect sizes, but the direction of effect was reversed for the orlistat trials. Most consistency in the expected effect (smaller effect sizes when quality criterion met) was found for randomization sequence adequate, similar baseline, ITT analysis, and acceptable compliance.

In addition, we repeated this analysis using a total quality score that considered all assessed quality features. Again, it is possible that the type of condition acts as a moderator or confounder. In some studies quality may have an effect on study results, but these effects are masked by other studies where there is no association between quality and study results. The effect of the total quality sum score on study results was calculated for each condition. The meta-regression slopes for each condition are shown in Table 21 and Figure 13.
Table 21. Total quality regressed on effect size (EPC reports)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Alzheimer’s</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>2. Arthritis</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>3. CDSM</td>
<td>-0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>4. Chromium</td>
<td>-0.35</td>
<td>0.18</td>
</tr>
<tr>
<td>5. Epilepsy</td>
<td>-0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>6. Glucosamine</td>
<td>-0.10</td>
<td>0.07</td>
</tr>
<tr>
<td>7. OCD</td>
<td>-0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>8. Omega 3</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>9. Orlistat</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>10. SAMe</td>
<td>-0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>11. SMBG</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>12. Vitamin E</td>
<td>0.09</td>
<td>0.12</td>
</tr>
</tbody>
</table>

EPC = Evidence-based Practice Center; SE = standard error; CDSM: chronic disease self-management; OCD: obsessive-compulsive-disorder; SAMe: S-adenosylmethionine; SMBG: self-monitoring of blood glucose

Figure 13. Meta-regression slopes showing relationship between total quality and effect size in each type of study

Circle is proportional to the size of the study; the line represents the effect on effect size.
The interaction effect of condition by total quality was not statistically significant \((p=0.574)\), meaning that slopes are not significantly different from each other indicating that condition is not a moderator (confounder) of the association between quality and effect sizes.

**Variable 3: Type of Outcome**

Table 22 shows differences of studies fulfilling a quality criterion compared to studies not meeting the criterion when controlling for the type of outcome, that is, objective or less prone to measurement error, versus other outcomes. In dataset 2 (EPC reports), 47 studies were classified as having an objective outcome as opposed to other endpoints; in the third dataset there were 35 studies (the back pain dataset did not include objective outcomes).

**Table 22. Difference in effect sizes between high- and low-quality studies, controlled for type of outcome**

<table>
<thead>
<tr>
<th>Quality feature</th>
<th>EPC reports</th>
<th>Published dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ESdiff</td>
<td>95% CI</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>0.01</td>
<td>(-0.14, 0.15)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>-0.05</td>
<td>(-0.20, 0.10)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>-0.01</td>
<td>(-0.15, 0.12)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>-0.06</td>
<td>(-0.38, 0.25)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>-0.14</td>
<td>(-0.32, 0.05)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>-0.16</td>
<td>(-0.37, 0.05)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>0.08</td>
<td>(-0.05, 0.21)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>-0.10</td>
<td>(-0.24, 0.05)</td>
</tr>
<tr>
<td>Similar cointerventions</td>
<td>-0.05</td>
<td>(-0.24, 0.13)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>-0.02</td>
<td>(-0.15, 0.11)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>0.09</td>
<td>(-0.32, 0.49)</td>
</tr>
</tbody>
</table>

EPC = Evidence-based Practice Center; ESdiff = effect size difference; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

For most individual quality domains differences between high- and low-quality studies were not more pronounced when controlling for the type of outcome. Controlling for the type of outcome, the association between a total quality score and reported effect size is 0.01 (SE 0.02) in the EPC reports and 0.02 (SE 0.05) in the “pro-bias” dataset.

In a moderator analysis, an interaction effect between total quality and the type of outcome measure was investigated. The slope of total quality for studies with nonobjective outcomes in the EPC report sample was 0.02 (95% CIs -0.03, 0.06; \(p=0.446\)). For studies with objective outcomes, the slope was -0.01 (95% CIs: -0.08, 0.05; \(p=0.639\)). The difference in slopes was not statistically significant (0.03; 95% CIs: -0.04, 0.10; \(p=0.411\)). Figure 14 shows the difference between these effects.
In the third dataset, the slope for nonobjective studies was -0.026 (SE 0.06; p=0.686); the slope for objective studies was 0.00 (SE 0.08; p=0.997). The interaction effect, the difference between these slopes was not statistically significant (0.026; SE 0.10; p=0.80). Figure 15 graphically depicts the effects of the type of outcome as a moderator, indicating a slightly different, that is, not parallel slope for the two types of studies.
Variable 4: Variance in Effect Sizes

In order to test the hypothesis that the variance in effect sizes explains differences between datasets, we compared the effect size distribution across datasets. Figures 16 and 17 show the distribution of the effect sizes reported for each included study in the three datasets.

**Figure 16. Effect size distribution dataset 1**

![](image1)

Back pain data, absolute values on the right

The distribution of effect sizes in the original back pain articles was approaching a normal distribution. The analyses presented in this report are based on absolute values but the quality effect size association results are very similar when using the original reported effect sizes. The effect size distribution of the other publications is shown in Figure 17.

**Figure 17. Effect size distribution dataset 2 (EPC reports) and dataset 3 (“pro-bias”)**

![](image2)

The three datasets show different distributions of effect sizes. The back pain and the published “pro-bias” data stem from datasets with substantial variation in reported effect sizes and approaching a symmetric distribution. The effect size distribution in the EPC report dataset was restricted; most reported results were small and very few were negative.
Discussion

This report quantifies the risk of bias associated with selected quality criteria across three different datasets. Our analyses show that the association between quality features and effect sizes is complex and may vary according to factors yet to be explored.

Quality of the Reporting

We found in all datasets that the quality of the reporting was lacking. Many studies in our datasets reported insufficient information to know whether a quality feature was met or not. Poor reporting in publications does not necessarily mean poor study quality. The majority of our analyses compared studies reporting positive features with studies where the quality feature was not reported, such as concealed treatment allocation. Hence, we were primarily concerned with demonstrating the effects of high-quality studies that reported a feature, and the expression of the feature was an indicator of high quality. Here, quality included the reporting, as well as quality in the design, conduct, and analysis of the study. Similarly, a recent study by Hartling, Ospina, Liang, et al. (2009) applying the Cochrane risk of bias (2008 version) compared studies with a high risk of bias with studies of low quality or unclear quality. We could also show in the analyzed datasets that the reporting has improved since publication of the Consort statement, in accordance with the observation of other reviewers (Kane, Wang, and Garrard, 2007).

However, it has to be considered that all included randomized controlled trials within datasets were identified through meta-analyses. The trials were all considered adequate for inclusion in a published meta-analysis. Moja et al. (2005) showed that 12 percent of Cochrane and the Database of Abstracts of Reviews of Effects (DARE) selected reviews used quality as an inclusion criterion. For Evidence-based Practice Center (EPC) reports in particular, this approach is also not uncommon, especially when sufficient high-quality studies are available. The quality of randomly selected trials may be lower still than encountered in our selected datasets.

Psychometric Analysis

First, our psychometric analysis indicates that the quality criterion “similar timing of outcome assessment” should be reassessed for inclusion in the 11-item list. This criterion is usually met; only 2 to 11 percent of studies across datasets indicate the possibility or evidence of differential outcome assessment.

Furthermore, we explored the interrelationship between quality features psychometrically and through the use of a multiple indicator multiple cause (MIMIC) model factor analysis. We were able to show that across all three datasets, several individual quality features showed substantial intercorrelations but the complete set of 11 items did not show marked internal consistency.

Although conceptually presumably independent, in practice we find that studies that observe good practice for selected quality features often also do so for other features. Studies that reported an adequate method of randomization sequence generation tended to also report the use of adequate treatment allocation concealment (intercorrelations ranging from 0.49 to 0.74 across datasets). Furthermore, the Cochrane Back Review Group (CBRG) criteria list differentiates patient, provider, and assessor blinding, but our analyses did not provide support for this differentiation. In our empirical data samples these features are substantially intercorrelated, and
in particular the provider blinding does not appear to contribute unique information. In two of the datasets, one of the three blinding items had to be excluded from the factor analyses due to collinearity. Therefore, while appealing conceptually, the distinction between blinding the patient, provider, or outcome assessor may not all independently assess study quality. However, when treating the items as the indicator of the same underlying construct, we found insufficient internal consistency to indicate a homogenous quality construct. The Cronbach’s alpha values in all three datasets were below values expected for a psychometric scale (alphas ranging from 0.55 to 0.61).

A factor analysis taking into account the intercorrelations between items as well as their effect on the reported treatment effect did not favor a one-factor solution. Best fit was achieved through three factors in all three datasets. The factors were similar, but the factorial validity was not perfect either, with some items loading on different factors across datasets. The randomization sequence and the allocation concealment item consistently formed a factor, and the blinding items also consistently formed a second factor across datasets. The other items were not represented by these factors, indicating an additional source of variance. A third factor consistently showed significant loading for acceptable dropout rates. Other items such as original group (intention to treat [ITT]) did also load on this factor but with less consistency across datasets.

The use of checklists when scoring quality versus the application of a summary score has been extensively discussed in the literature. Juni and colleagues (Juni, Witschi, Bloch, et al., 1999; Juni, Douglas, Altman, et al., 2001) raised serious concerns about the use of quantitative sum scores. However, treating all quality items as completely independent does also not appear appropriate either following our analyses. The Cochrane review handbook currently suggests the use of a domain-based evaluation of quality in which critical assessments are made separately for different domains (Higgins and Green, 2009). The equal weighing of each item as applied in our approach is common place but not validated. Depending on the intervention and the clinical field, some internal validity threats may be more pertinent than others; however there are as yet no data to guide what these associations may be. Quality criteria could be used to trigger an overall assessment of quality which is more qualitatively derived than quantitatively by adding individual item scores. The reliability of qualitative overall evaluations have to be considered though, as Hartling et al. (2009) reported a kappa of 0.27 for reviewers to agree on the Cochrane Overall Risk of Bias dimension (Higgins and Green, 2008). A combined qualitative and quantitative approach may be useful: quality features could be ranked by importance for the clinical field a priori and weighted accordingly for a summary score.

Validating scales used to assess the quality of trials is very difficult. The concept of quality is not easy to define and there is no widely accepted gold standard. In one of the datasets we applied the Jadad items and scale and the criteria suggested by Schulz et al. (1995) parallel to our proposed criteria following the original scoring guidelines. We were able to show convergent validity across quality domains. The correlations with the Jadad and the Schulz scales were satisfying throughout and ranged between 0.63 and .093. However, the crucial validity test for quality items is the predictive validity of the quality features and possibly scales—is there evidence of bias, and is meeting or not meeting the quality criteria associated with differential effect sizes.
Associations Between Internal Validity and Effect Sizes

This report provides empirical data on the impact of fulfilling or not fulfilling the quality criteria sequence of randomization, concealment of treatment allocation, similarity of baseline values, assessor blinding, care provider blinding, patient blinding, dropout rate, ITT analysis, similarity of cointerventions, acceptable compliance, and similar timing of outcome assessment across three datasets. Although the majority of systematic reviews assess the quality of included studies, and meta-regression analyses trying to trace the effects of quality are often undertaken, there are relatively few published studies showing an effect of quality on effect sizes, that is, empirical evidence of bias in reported study results that can presumably be attributed to the quality of the reporting or the conduct of the research study. For many suggested quality criteria and potential threats to the internal validity of RCTs (see e.g., West et al., 2002; Moja et al., 2005) there is still a dearth of published evidence on the extent of bias, that is, does not meeting the quality criteria show associations with the observed treatment effect.

The 11 proposed quality features contribute information to the evaluation of back pain trials as previously reported (Van Tulder et al., 2009). Although not statistically significant, individual features showed consistently associations with effect sizes depending on the quality of the trial. High-quality studies reported smaller effect sizes, indicating that low-quality studies tended to overestimate treatment effects. A dataset consisting of trials included in EPC reports showed a different pattern. The EPC dataset analysis showed for the majority of individual quality dimensions that the high quality studies in the dataset tended to reported larger treatment effects than the low quality trials that did not meet the quality criterion. The third dataset showed the most variation across quality criteria, but was more similar to the back pain dataset, in that meeting most individual criteria where associated with smaller effect sizes.

The feature allocation concealment showed the most consistent results across datasets. In all three datasets, allocation concealment was associated with effect sizes, and the direction of effect did not vary. Unconcealed trials reported smaller effect sizes in the back pain dataset (effect size difference -0.08, 95% CI: -0.23, 0.07) and the EPC report dataset (effect size difference -0.06, 95% CI: -0.22, 0.11), and the ratio of odds ratios (ROR) in the third dataset also showed that unconcealed trials reported larger treatment effects (ROR=0.91, 95% CI: 0.72, 1.14). This analysis indicated that unconcealed trials tend to overestimate the treatment effect. Similarly, Pidal et al. (2007) reported an ROR of 0.90 (0.81, 1.01); Wood et al. (2008) found a ROR of 0.91 (95% CI 0.80, 1.03) for objective and 0.69 (95% CI: 0.59, 0.82) for subjective outcomes. A pooled analysis using data from Schulz et al. (1995), Moher et al. (1998), Kjaergard et al. (2000), and Juni et al. (2000) showed an ROR of 0.70 (95% CI: 0.62, 0.80) across datasets (Juni et al., 2001). The influence of concealment of allocation on effect size seems to be the most consistent quality criteria.

When applying a total sum score derived from the mean item scores and regressing effect sizes on the sum score, we found no statistically significant linear effect. A simple linear relationship indicates that the reported treatment effects increase the lower the quality level. A similar approach was described by Emerson et al. (1990), who also found no linear relation between quality score and variation in treatment differences.

When using factor scores, rather than individual quality features or a simple sum score, we also did not find that these quality factors predicted effect sizes. The factor structure takes the inter-item correlations as well as their individual association with effect sizes into account. MIMIC models are generally a promising approach to describe complex relations between
multiple predictors and has been applied in a variety of fields (e.g., Hartford & Muthén, 2001; Urban & Demetrovics, 2010).

One power maximizing approach showed consistently statistically significant effects. Cutoff values based on the number of fulfilled quality criteria were able to differentiate high- and low-quality studies. In all three datasets, 5 or 6 (out of 11) fulfilled quality criteria differentiated the studies best. In dataset 1 (back pain) and dataset 3 (“pro-bias”), the replication of the set used by Moher et al. (1998) high-quality studies showed smaller treatment effects. Effect size differences were -0.20 in dataset #1 and the RORs were 0.79 (cutoff at 5) and 0.77 (cutoff at 6) in the third dataset when taking several quality criteria into account. However, in the EPC report dataset, we found unexpected results: studies with five or six fulfilled quality feature reported larger effect sizes than the low-quality trials (effect size difference 0.27). Moher et al. (1998) used the Jadad scale, which takes three individual quality features into account (randomization, blinding, withdrawals) to differentiate high- and low-quality studies, and reported a ratio of odds ratio of 0.66 (95% CI: 0.52, 0.83). Our results are similar in direction for two of the three datasets, but we found less-pronounced results. The difference in results may be due to the selected statistical approach and presumably in part due to the nonperfect overlap of studies since we were unable to obtain a quarter of the original dataset.

Comparing results across methods to test associations between quality criteria and effect sizes, the differences between the originally analyzed back pain dataset and the EPC report dataset are most striking. The quality criteria were developed for the CBRG and they also appear to be most useful in this dataset (see also van Tulder et al., 2003). However, it is possible that the tendency of EPC reports to consider quality as an inclusion criterion, potentially excluding fatally flawed studies from the review, is partly responsible for the different results in this dataset. The quality scores indicated that for several criteria such as blinding, similarity of cointerventions, and ITT analysis, the trials included in EPC reports scored higher.

**Moderators and Confounders**

The identified associations between quality and effect sizes varied across our datasets presented in this report which reflects also the extent of conflicting results reported in the literature (e.g., Moher et al., 1998, Balk et al., 2002; Juni et al., 2001). Research on the association between quality and effect sizes should focus on factors that can help predict when lack of quality is likely to result in a distorted estimate of treatment effects.

In this report, we investigated a number of differences across datasets. The systematically investigated moderators and confounders were the size of the treatment effect within meta-analyses, the condition being treated, the type of outcome, and the variance in effect sizes. These moderators did not sufficiently explain diverging results across our employed datasets.

The condition being treated or the clinical field the study was conducted in was not sufficient to explain differential effects of the association between quality and effect sizes. The size of the treatment effect could also not be shown as a significant moderator between the two variables. Unlike Wood et al. (2009), we could not show that the type of outcome explained differences in effects of associations; there was no statistically significant difference between slopes. However, it has to be noted that the type of outcome is to some extent already been taking into account in the CBRG guidance: assessor blinding is assumed for studies with automated test result analysis, it is assumed that the assessor is clearly not aware of the treatment allocation, regardless of whether the publication states that the assessor was blind.
The variance in effect sizes across a dataset was one factor that should be explored further in future research to see if this factor contributes to the question of when quality features are likely to influence effect sizes. Balk et al. (2002) used existing heterogeneity in odds ratios as an inclusion criterion for their meta-epidemiological study but concluded that the investigated quality measure are not reliable associated with the strength of treatment effect. Whether the variance in effect sizes is indeed a sufficient moderator to explain variation across datasets is a testable question and could be assessed with Monte Carlo simulations systematically investigating the effect of moderators or confounders that influence the association between quality and effect sizes.

Implication for Practice

In two of these last datasets, we showed that quality features can affect reported treatment effects. The 11 proposed quality features developed for the CBRG contribute information to the evaluation of back pain trials as previously reported (Van Tulder et al., 2009). Whether this extended list of quality features can be proposed for a more general use was one of the principal questions of this research project. Their general applicability has not been supported, as their effect was not uniform across datasets.

We conclude from our analyses that the association between quality features and reported treatment effects should be explored in systematic reviews. Regardless of whether quality criteria are assessed individually, through empirically derived factor scores or the use of a total score, regardless of whether the summary score was quantitatively, quantitative or through a combination derived, and regardless of whether low quality studies are excluded from the analysis or studies are pooled weighted by quality (e.g. Juni, Altman, and Egger, 2001; Welton, Ades, Altman, et al., 2009) quality should be taken into account when evaluating the existing evidence and the potential bias should be quantified. For situations where the Jadad criteria may be insufficiently applicable, our data provide some support for the use of the II-item CBRG list.

Future Research

Applying psychometric principles to the field of quality criteria is rarely explored but can provide useful insight into empirical associations of quality items. In this report we explored the reliability of the proposed quality items only through item and scale analysis. Future work should include an analysis of agreement between raters. The reproducibility of quality judgments across independent raters is another valuable method for estimating the reliability of proposed items or scales. Previous research has shown that also carefully developed tools may show disappointing rater agreement when scoring agreement is tested. Hartling et al. (2009) reported kappas ranging from 0.13 to 0.74 for domains of the Cochrane Risk of Bias tool. For evidence reviews, however, we suggest an additional approach for testing the reliability of tools. In systematic reviews, it is now standard to employ two independent raters when scoring the quality of included studies and to reconcile independent decisions for a final score. This approach helps to avoid individual reviewer bias and errors and the reconciled decision should be more reliable than the individual decision. The rater agreement of reconciled decisions across pairs of raters is a better indicator of the reliability of the tool because it mirrors more closely how the tool will be used in practice.

There is a need for more information on individual quality features and empirical evidence of bias. This concerns the many suggested quality criteria for which no empirical evidence is available yet or at least no summary across individual meta-analyses exists (see West et al., 2002; Moja et al., 2005). There are other quality criteria such as selective outcome reporting
(Kirkham, Dwan, Altman, et al., 2010) that may be difficult to operationalize and a replication of the effect in a different dataset would be useful. The scarcity of evidence is disappointing considering the fact that quality scoring is a standard method in systematic reviews. Many different quality criteria are used by reviewers and many represent possible and plausible threats to the validity of the study. Future reviews should report data on these associations in order to advance the evidence base for quality assessments.

The association between quality features and effect sizes is complex, and the conditions, when lack of quality is most likely to lead to bias, should be explored further in future research. Factors such as the variance in quality scores and effect sizes across studies could be systematically studied in “virtual datasets,” that is, by creating datasets employing Monte Carlo simulation methods. Using datasets of “known” properties would be useful to further study associations between the proposed quality criteria and effect sizes. An increase in sample size and thereby statistical power would enable researchers to detect small but systematic effects and shed further light on the question of when quality features are most useful to be taken into account when assessing treatment effects in published research.

**Conclusions**

The associations between quality features and effect sizes are complex. Effect sizes of individual studies depend on many factors. In two datasets, individual quality items and summary scores of items were associated with differences in effect sizes. This relationship was not found in the remaining dataset. Despite several exploratory analyses, we were not able to explain these differences. The conditions under which quality features and which features lead to biased effect sizes warrant further exploration, and factors such as the variance in quality scores and effect sizes will be investigated in a subsequent project.
References


Appendix A. References Dataset 1: Back Pain, 216 Trials


Hildebrandt VH, Proper KI, van den Berg R, et al. [Cesar therapy is temporarily more effective in patients with chronic low back pain than the standard treatment by family practitioner: randomized, controlled and blinded clinical trial with 1 year follow-up]. Ned Tijdschr Geneeskd 2000;144:2258–64.


Transcutaneous Electrical Nerve Stimulation (TENS) For Chronic Low Back Pain. Annual meeting of the American Academy of Orthopedic Surgeons; 1997; San Francisco.


Appendix B. References Dataset 2: EPC Reports, 165 Trials


Appendix C. References Dataset 3: Published “Pro-bias” Dataset, 100 Trials


Pince J. Thromboses veineuses des membres inferieurs et embolies pulmonaires au cours des accidents vasculaires cervebraux. A propos d'un essai comparatif de traitement preventif (These pour le doctorat d'etat en medecine). In press 1981.


Rutgeerts P. Comparative efficacy of coated, oral 5-
aminosalicylic acid (Claversal) and sulphasalazine for


## Appendix D. Comparison Fixed-Effects Model Results

### Table D1. Difference in effect sizes (EPC reports), FE

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ESdiff 95% CI</td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>44 121</td>
<td>0.59 (0.56, 0.61)</td>
<td>1.08 (1.07, 1.10)</td>
<td>-0.50 (-0.53, -0.47)*</td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>38 127</td>
<td>0.56 (0.53, 0.59)</td>
<td>1.09 (1.07, 1.10)</td>
<td>-0.53 (-0.56, -0.49)*</td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>100 65</td>
<td>1.00 (0.99, 1.02)</td>
<td>0.87 (0.83, 0.90)</td>
<td>0.14 (0.10, 0.17)*</td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>157 8</td>
<td>1.00 (0.98, 1.01)</td>
<td>0.41 (0.31, 0.51)</td>
<td>0.59 (0.49, 0.68)*</td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>120 45</td>
<td>1.09 (1.07, 1.10)</td>
<td>0.14 (0.10, 0.18)</td>
<td>0.95 (0.91, 0.99)*</td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>130 35</td>
<td>1.09 (1.07, 1.10)</td>
<td>0.08 (0.04, 0.12)</td>
<td>1.01 (0.96, 1.05)*</td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>96 69</td>
<td>1.15 (1.13, 1.16)</td>
<td>0.24 (0.21, 0.27)</td>
<td>0.90 (0.87, 0.93)*</td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>101 64</td>
<td>1.03 (1.02, 1.04)</td>
<td>0.40 (0.35, 0.45)</td>
<td>0.63 (0.58, 0.68)*</td>
<td></td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>142 23</td>
<td>1.02 (1.00, 1.03)</td>
<td>0.25 (0.18, 0.31)</td>
<td>0.77 (0.71, 0.83)*</td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>79 86</td>
<td>1.18 (1.17, 1.20)</td>
<td>0.29 (0.27, 0.32)</td>
<td>0.89 (0.86, 0.92)*</td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>161 4</td>
<td>0.99 (0.98, 1.01)</td>
<td>0.28 (0.15, 0.42)</td>
<td>0.71 (0.58, 0.84)*</td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05

EPC = Evidence-based Practice Center; FE = based on fixed-effects model; ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat

### Table D2. Comparison of different quality cutoffs using a total score (EPC reports), FE

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Number Equal or Above Cut-off</th>
<th>Number Below Cutoff</th>
<th>High Quality</th>
<th>Low Quality</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ES 95% CI</td>
<td>ESdiff 95% CI</td>
<td></td>
</tr>
<tr>
<td>≥9 vs &lt;9</td>
<td>42 123</td>
<td>1.22 (1.21, 1.24)</td>
<td>0.29 (0.26, 0.31)</td>
<td>0.93 (0.91, 0.96)*</td>
<td></td>
</tr>
<tr>
<td>≥8 vs &lt;8</td>
<td>65 100</td>
<td>1.16 (1.15, 1.18)</td>
<td>0.27 (0.24, 0.30)</td>
<td>0.89 (0.86, 0.93)*</td>
<td></td>
</tr>
<tr>
<td>≥7 vs &lt;7</td>
<td>103 62</td>
<td>1.03 (1.02, 1.04)</td>
<td>0.41 (0.37, 0.46)</td>
<td>0.62 (0.57, 0.67)*</td>
<td></td>
</tr>
<tr>
<td>≥6 vs &lt;6</td>
<td>135 30</td>
<td>1.01 (1.00, 1.02)</td>
<td>0.21 (0.14, 0.28)</td>
<td>0.80 (0.73, 0.88)*</td>
<td></td>
</tr>
<tr>
<td>≥5 vs &lt;5</td>
<td>149 16</td>
<td>1.00 (0.99, 1.02)</td>
<td>0.10 (0.00, 0.20)</td>
<td>0.90 (0.80, 1.00)*</td>
<td></td>
</tr>
<tr>
<td>≥4 vs &lt;4</td>
<td>160 5</td>
<td>0.99 (0.98, 1.00)</td>
<td>0.06 (-0.16, 0.27)</td>
<td>0.93 (0.72, 1.15)</td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05

EPC = Evidence-based Practice Center; FE = based on fixed effects model; ES = effect size; CI = confidence interval; ESdiff = effect size difference
Appendix E. Comparison Random Effects
Meta-regression Results

Table E1. Difference in odds ratios for proposed quality criteria (“pro-bias”), R

<table>
<thead>
<tr>
<th>Quality Feature</th>
<th>Number Criterion Met</th>
<th>Number Criterion Not Met</th>
<th>Effect Size in Trials With Criterion Met</th>
<th>Effect Size in Trials With Criterion Not Met</th>
<th>Effect Size Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR 95% CI</td>
<td>OR 95% CI</td>
<td>ROR 95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomization adequate</td>
<td>34 0.44 (0.31, 0.59)</td>
<td>66 0.46 (0.37, 0.59)</td>
<td>1.05 (0.69, 1.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>26 0.50 (0.34, 0.75)</td>
<td>74 0.44 (0.35, 0.56)</td>
<td>0.88 (0.56, 1.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar baseline</td>
<td>36 0.37 (0.26, 0.52)</td>
<td>64 0.51 (0.40, 0.66)</td>
<td>1.40 (0.92, 2.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor blind</td>
<td>78 0.42 (0.34, 0.52)</td>
<td>22 0.65 (0.42, 1.00)</td>
<td>1.55 (0.95, 2.51)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care provider blind</td>
<td>69 0.45 (0.36, 0.58)</td>
<td>31 0.46 (0.32, 0.66)</td>
<td>1.02 (0.66, 1.57)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient blind</td>
<td>72 0.44 (0.35, 0.55)</td>
<td>28 0.52 (0.36, 0.76)</td>
<td>1.20 (0.77, 1.87)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>62 0.50 (0.38, 0.64)</td>
<td>38 0.41 (0.30, 0.55)</td>
<td>0.83 (0.57, 1.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>29 0.46 (0.32, 0.65)</td>
<td>71 0.46 (0.36, 0.58)</td>
<td>1.00 (0.65, 1.52)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>68 0.40 (0.32, 0.51)</td>
<td>32 0.60 (0.43, 0.84)</td>
<td>1.51 (1.00, 2.27)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>46 0.56 (0.42, 0.75)</td>
<td>54 0.39 (0.30, 0.51)</td>
<td>0.70 (0.47, 1.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar timing</td>
<td>89 0.44 (0.36, 0.55)</td>
<td>11 0.60 (0.32, 1.11)</td>
<td>1.35 (0.71, 2.58)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05
R = random effects meta-regression; OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios;
ITT = intention to treat
Figure E1. Differences in effect sizes based on quality criteria (“pro-bias”), random effects meta-regression

<table>
<thead>
<tr>
<th>Study</th>
<th>Ratio of Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization adequate</td>
<td>1.05 (0.69, 1.60)</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>0.88 (0.56, 1.39)</td>
</tr>
<tr>
<td>Similar baseline</td>
<td>1.40 (0.92, 2.12)</td>
</tr>
<tr>
<td>Assessor blind</td>
<td>1.55 (0.95, 2.51)</td>
</tr>
<tr>
<td>Care provider blind</td>
<td>1.02 (0.66, 1.57)</td>
</tr>
<tr>
<td>Patient blind</td>
<td>1.20 (0.77, 1.87)</td>
</tr>
<tr>
<td>Acceptable dropout rate</td>
<td>0.83 (0.56, 1.22)</td>
</tr>
<tr>
<td>Original group (ITT)</td>
<td>1.00 (0.65, 1.52)</td>
</tr>
<tr>
<td>Similar co-interventions</td>
<td>1.51 (1.00, 2.27)</td>
</tr>
<tr>
<td>Acceptable compliance</td>
<td>0.70 (0.47, 1.03)</td>
</tr>
<tr>
<td>Similar timing</td>
<td>1.35 (0.71, 2.58)</td>
</tr>
</tbody>
</table>

CI = confidence interval; ITT = intention to treat

Table E2. Comparison of different quality cutoffs using a total score (“pro-bias”), R

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Number</th>
<th>Number</th>
<th>High Quality</th>
<th>Low Quality</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equal or Above Cut-off</td>
<td>Below Cut-off</td>
<td>OR 95% CI</td>
<td>OR 95% CI</td>
<td>ROR 95% CI</td>
</tr>
<tr>
<td>≥9 vs &lt;9</td>
<td>14</td>
<td>86</td>
<td>0.39 (0.23, 0.67)</td>
<td>0.47 (0.38, 0.58)</td>
<td>1.20 (0.68, 2.10)</td>
</tr>
<tr>
<td>≥8 vs &lt;8</td>
<td>26</td>
<td>74</td>
<td>0.46 (0.32, 0.67)</td>
<td>0.46 (0.36, 0.58)</td>
<td>0.99 (0.64, 1.55)</td>
</tr>
<tr>
<td>≥7 vs &lt;7</td>
<td>44</td>
<td>56</td>
<td>0.39 (0.29, 0.53)</td>
<td>0.51 (0.40, 0.67)</td>
<td>1.31 (0.88, 1.95)</td>
</tr>
<tr>
<td>≥6 vs &lt;6</td>
<td>62</td>
<td>38</td>
<td>0.46 (0.36, 0.59)</td>
<td>0.46 (0.32, 0.63)</td>
<td>0.99 (0.66, 1.49)</td>
</tr>
<tr>
<td>≥5 vs &lt;5</td>
<td>76</td>
<td>24</td>
<td>0.45 (0.36, 0.57)</td>
<td>0.47 (0.32, 0.71)</td>
<td>1.05 (0.66, 1.67)</td>
</tr>
<tr>
<td>≥4 vs &lt;4</td>
<td>86</td>
<td>14</td>
<td>0.46 (0.36, 0.55)</td>
<td>0.53 (0.32, 0.88)</td>
<td>1.19 (0.69, 2.06)</td>
</tr>
</tbody>
</table>

R = random effects meta-regression; OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios
### Appendix F. Quality Rating Form

<table>
<thead>
<tr>
<th>CBRG Quality Items</th>
<th>Yes</th>
<th>No</th>
<th>Don’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was the method of randomization adequate?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the treatment allocation concealed?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Were the groups similar at baseline regarding the most important prognostic indicators?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the outcome assessor blinded?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the care provider blinded?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Were patients blinded?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the dropout rate described and acceptable?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Were all randomized participants analyzed in the group to which they were originally assigned?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Were co-interventions avoided or similar?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the compliance acceptable in all groups?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the timing of the outcome assessment similar in all groups?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Article ID:**

**Reviewer:**

**First Author, Year:**

**Meta-analysis:**

_Last Name Only_
**Scoring Guidelines Cochrane Back Review Group**

**Randomization sequence**
A random (unpredictable) assignment sequence. Examples of adequate methods are coin toss (for studies with two groups), rolling a dice (for studies with two or more groups), drawing of balls of different colours, drawing of ballots with the study group labels from a dark bag, computergenerated random sequence, pre-ordered sealed envelops, sequentially-ordered vials, telephone call to a central office, and pre-ordered list of treatment assignments Examples of inadequate methods are: alternation, birth date, social insurance/security number, date in which they are invited to participate in the study, and hospital registration number

**Allocation concealment**
Assignment generated by an independent person not responsible for determining the eligibility of the patients. This person has no information about the persons included in the trial and has no influence on the assignment sequence or on the decision about eligibility of the patient.

**Patient blinding**
This item should be scored “yes” if the index and control groups are indistinguishable for the patients or if the success of blinding was tested among the patients and it was successful.

**Care provider blinding**
This item should be scored “yes” if the index and control groups are indistinguishable for the care providers or if the success of blinding was tested among the care providers and it was successful

**Assessor blinding**
Adequacy of blinding should be assessed for the primary outcomes. This item should be scored “yes” if the success of blinding was tested among the outcome assessors and it was successful or:
- **for patient-reported outcomes** in which the patient is the outcome assessor (e.g., pain, disability): the blinding procedure is adequate for outcome assessors if participant blinding is scored “yes”
- **for outcome criteria assessed during scheduled visit and that supposes a contact between participants and outcome assessors** (e.g., clinical examination): the blinding procedure is adequate if patients are blinded, and the treatment or adverse effects of the treatment cannot be noticed during clinical examination
- **for outcome criteria that do not suppose a contact with participants** (e.g., radiography, magnetic resonance imaging): the blinding procedure is adequate if the treatment or adverse effects of the treatment cannot be noticed when assessing the main outcome
- **for outcome criteria that are clinical or therapeutic events** that will be determined by the interaction between patients and care providers (e.g., co-interventions, hospitalization length, treatment failure), in which the care provider is the outcome assessor: the blinding procedure is adequate for outcome assessors if item “E” is scored “yes”
- **for outcome criteria that are assessed from data of the medical forms**: the blinding procedure is adequate if the treatment or adverse effects of the treatment cannot be noticed on the extracted data

**Dropouts**
The number of participants who were included in the study but did not complete the observation period or were not included in the analysis must be described and reasons given. If the percentage of withdrawals and dropouts does not exceed 20% for short-term follow-up and 30% for long-term followup and does not lead to substantial bias a 'yes' is scored. (N.B. these percentages are arbitrary, not supported by literature).

**ITT**
All randomized patients are reported/analyzed in the group they were allocated to by randomization for the most important moments of effect measurement (minus missing values) irrespective of noncompliance and co-interventions.

**Baseline comparability**
In order to receive a “yes”, groups have to be similar at baseline regarding demographic factors, duration and severity of complaints, percentage of patients with neurological symptoms, and value of main outcome measure(s).

**Co-Interventions**
This item should be scored “yes” if there were no co-interventions or they were similar between the index and control groups.

**Compliance**
The reviewer determines if the compliance with the interventions is acceptable, based on the reported intensity, duration, number and frequency of sessions for both the index intervention and control intervention(s). For example, physiotherapy treatment is usually administered over several sessions; therefore it is necessary to assess how many sessions each patient attended. For single-session interventions (for ex: surgery), this item is irrelevant.

**Timing**
Timing of outcome assessment should be identical for all intervention groups and for all important outcome assessments.

*Note: These instructions are adapted from van Tulder 2003, Boutron et al. 2005 (CLEAR NPT) and the Cochrane Handbook of Reviews of Interventions 2008 Updated Guidelines for Systematic Reviews 9April 2008*
### Jadad Scale

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sub Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Randomization</strong></td>
<td>1. Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)? = 1 point</td>
</tr>
<tr>
<td></td>
<td>Deduct 1 point if: For question 1, the method to generate the sequence of randomization was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc.)</td>
</tr>
<tr>
<td><strong>Blinding</strong></td>
<td>2. Was the study described as double blind? = 1 point</td>
</tr>
<tr>
<td></td>
<td>Deduct 1 point: If for question 2 the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy)</td>
</tr>
<tr>
<td><strong>Withdrawals and dropouts</strong></td>
<td>3. Was there a description of withdrawals and dropouts? = 1 point</td>
</tr>
</tbody>
</table>

**TOTAL JADAD SCORE**

---

### Jadad Guidelines for Assessment

1. **Randomization**
   A method to generate the sequence of randomization will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should be not regarded as appropriate.

2. **Double blinding**
   A study must be regarded as double blind if the word "double blind" is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.

3. **Withdrawals and dropouts**
   Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.
Schulz’s (1995) quality dimensions
(circle appropriate category)

1. Concealment of Treatment Allocation
a) Adequately concealed trial (i.e. central randomization; numbered or coded bottles or containers; drugs prepared by the pharmacy; serially numbered; opaque, sealed envelopes; or other description that contained elements convincing of concealment
b) Inadequately concealed trial (i.e. alternation or reference to case record numbers or dates of birth
c) Unclearly concealed trial (authors did either not report an allocation concealment approach at all or reported an approach that did not fall into the categories above

2. Generation of Allocation Sequence
a) Adequately sequence generation (random-number table, computer random-number generator, coin tossing, or shuffling)
b) Publication does not report one of the adequate approaches, those with inadequate sequence generation

3. Inclusion in the Analysis of All Randomized Participants
a) Publication reports or gives the impression that no exclusions have taken place (often not explicit)
b) Publication reports exclusions (e.g., protocol deviation, withdrawals, dropouts, loss to follow-up)

4. Double Blinding
a) Double-blinding reported
b) Double-blinding not reported