Methods Research Report

Detection of Associations Between Trial Quality and Effect Sizes

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Detection of Associations Between Trial Quality and Effect Sizes

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

Contract No. HHSA 290-2007-10056-I

Prepared by:
Southern California Evidence-based Practice Center
RAND Corporation
1776 Main Street
Santa Monica, CA 90407

Investigators:
Susanne Hempel, Ph.D.
Jeremy Miles, Ph.D.
Marika J. Suttorp, M.S.
Zhen Wang, M.S.
Breanne Johnsen, B.A.
Sally Morton, Ph.D.
Tanja Perry, B.H.M.
Diane Valentine, J.D.
Paul G. Shekelle, M.D., Ph.D.

AHRQ Publication No. 12-EHC010-EF
January 2012

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. HHSA 290 2007 10056 I). The findings and conclusions in this document are those of the author(s), who are responsible for its content; and do not necessarily represent the views of AHRQ. 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 clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

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.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested Citation: Hempel S, Miles J, Suttorp M, Wang Z, Johnsen B, Morton S, Perry T, Valentine D, Shekelle P. Detection of Associations between Trial Quality and Effect Sizes. Methods Research Report. Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I. AHRQ Publication No. 12-EHC010-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2012.

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 and strategies. 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 in 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, EPC Program
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Sydne Newberry for editorial support, Brett Ewing for statistical support, Aneesa Motala for administrative support, and Ethan Balk, Nancy Berkman, Isabelle Boutron, Stephanie Chang, Issa Dahabreh, Rochelle Fu, Jonathan Sterne, Meera Viswanathan for comments on earlier drafts of the report.

Peer Reviewers

Ethan Balk, M.D., M.P.H.
Tufts Medical Center
Boston, MA

Nancy Berkman, Ph.D., M.L.I.R.
RTI International
Research Triangle Park, NC

Isabelle Boutron, M.D., Ph.D.
University of Oxford
Oxford, United Kingdom

Rochelle Fu, Ph.D.
Oregon Health and Science University
Portland, OR

Jonathan Sterne, Ph.D.
University of Bristol
Bristol, United Kingdom

Meera Viswanathan, Ph.D.
RTI International
Research Triangle Park, NC

Detection of Associations Between Trial Quality and Effect Sizes

Structured Abstract

Objectives. To examine associations between a set of trial quality criteria and effect sizes and to explore factors influencing the detection of associations in meta-epidemiological datasets.

Data Sources. The analyses are based on four meta-epidemiological datasets. These datasets consist of a number of meta-analyses; each contained between 100 and 216 controlled trials. These datasets have "known" qualities, as they were used in published research to investigate associations between quality and effect sizes. In addition, we created datasets using Monte Carlo simulation methods to examine their properties.

Review Methods. We identified treatment effect meta-analyses and included trials and extracted treatment effects for four meta-epidemiological datasets. We assessed quality and risk of bias indicators with 11 Cochrane Back Review Group (CBRG) criteria. In addition, we applied the Jadad criteria, criteria proposed by Schulz (e.g., allocation concealment), and the Cochrane Risk of Bias tool. We investigated the effect of individual criteria and quantitative summary scores on the reported treatment effect sizes. We explored potential reasons for differences in associations across different meta-epidemiological datasets, clinical fields and individual meta-analyses. We investigated factors that influence the power to detect associations between quality and effect sizes in Monte Carlo simulations.

Results. Associations between quality and effect sizes were small, e.g. the ratio of odds ratios (ROR) for unconcealed (vs. concealed) trials was 0.89 (95% CI: 0.73, 1.09, n.s.), but consistent across the CBRG criteria. Based on a quantitative summary score, a cut-off of six or more criteria met (out of 11) differentiated low- and high-quality trials best with lower quality trials reporting larger treatment effects (ROR 0.86, 95% CI: 0.70, 1.06, n.s.). Results for evidence of bias varied between datasets, clinical fields, and individual meta-analyses. The simulations showed that the power to detect quality effects is, to a large extent, determined by the degree of residual heterogeneity present in the dataset.

Conclusions. Although trial quality may explain some amount of heterogeneity across trial results in meta-analyses, the amount of additional heterogeneity in effect sizes is a crucial factor in determining when associations between quality and effect sizes can be detected. Detecting quality moderator effects requires more statistically powerful analyses than are employed in most investigations.

Contents

Tables

Figures

Appendixes

Executive Summary

Background

Trial design and execution factors are potentially associated with bias in the effect sizes reported for randomized controlled trials. Bias is defined as a systematic deviation of the estimated treatment effect from the true (or population) value. Although a number of factors have been proposed to be associated with bias, an actual association has been empirically confirmed for only a few and the literature shows 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 and which factors moderate the detection of associations.

In previous research (Hempel et al., 2011), we investigated the effect of the individual criteria used by the Cochrane Back Review Group (CBRG) and a quantitative summary score derived from the criteria. The set covers established criteria internal validity and quality of the reporting criteria as well as quality indicators that have rarely been assessed or shown to be indicators of bias. Previous results showed that the criteria, in particular when combined as a quantitative summary score, differentiated high and low quality trials in two out of three meta-epidemiological datasets which comprised a number of individual meta-analyses and included trials. High- and low-quality trials, as measured by the selected criteria, showed a trend for differences in reported effect sizes, with low quality trials exaggerating treatment effects. In order to continue to test the generalizability of quality criteria and the situations where they may be most useful, we expanded our analytic capability by including a new meta-epidemiological dataset in our analyses and investigated factors that may explain when quality is associated with effect size and when these associations can be detected in datasets.

The association between quality features and effect sizes is complex, and the conditions in which quality is most likely to be associated with bias warrant further exploration. We expect our results to contribute empirical evidence to the continuing professional debate about the appropriate role of quality criteria in systematic reviews of randomized controlled trials.

Objectives and Key Questions

The objectives of the project were to examine associations between individual and summary indicators of trial quality and effect sizes and to explore factors influencing the detection of associations in meta-epidemiological datasets. The selected quality criteria address design and execution factors of the trial as well as the quality of the reporting. We are interested in the association between trial quality criteria and the size of the reported treatment effect; that is, whether trials meeting the quality criteria reported different treatment effects than trials not meeting quality criteria. The project aimed to answer the following questions:

- Are the selected quality criteria, individually as well as combined, useful as indicators of bias in diverse clinical contexts? The usefulness was operationalized as predictive validity – whether meeting or not meeting the quality criteria is associated with differential effect sizes in treatment effect trials.
- Which factors influence the presence and the detection of associations between quality and effect sizes? The question was investigated in empirical meta-epidemiological datasets as well as Monte Carlo simulation models.

Methods

Association Between Quality and Effect Sizes in Empirical Datasets

We applied 11 quality criteria to 4 large meta-epidemiological datasets. These datasets included a variety of meta-analyses covering a wide range of clinical fields. Each meta-analysis contributed between 3 and 45 individual trials. The first dataset was derived from all CBRG reviews of non-surgical treatment for non-specific low back pain in the Cochrane Library 2005, issue 3; the dataset included 216 individual trials. For the second dataset we searched prior systematic reviews and meta-analyses conducted by Agency for Healthcare Research and Quality (AHRQ)-funded Evidence-based Practice Centers (EPCs) with the goal of assembling a dataset 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 that demonstrated associations between quality and the size of treatment effects; this dataset included 100 trials. For the purpose of this report we assembled an additional dataset with 'known qualities'. This fourth dataset was based on another published meta-epidemiological dataset. One of the selection criteria for the trials that were included in this dataset was that meta-analyses used to establish the dataset had to report evidence of heterogeneity across trials.

We assessed quality criteria and risk of bias indicators for all included trials. We used the CBRG quality criteria, which cover 11 quality features (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 co-interventions, acceptable compliance, and similar timing of outcome assessment. In addition, we applied the Jadad criteria (randomization, blinding, withdrawals and dropouts, total score), criteria proposed by Schulz (concealment of treatment allocation, sequence generation, inclusion in the analysis of all randomized participants, double blinding), and the Cochrane Risk of Bias tool (sequence generation; allocation concealment; blinding of participants, personnel and outcome assessors; incomplete outcome data; selective outcome reporting; other sources of bias; overall risk of bias).

For all datasets we calculated odds ratios for datasets with dichotomous outcomes and effect sizes for datasets with continuous outcomes. We investigated associations between quality and reported treatment effects in meta-regression by calculating the differences in effect sizes or ratios of odds ratios (ROR) of trials meeting quality criteria compared with those that did not. A negative effect size difference indicated that trials meeting quality criteria (high-quality trials) reported smaller effect sizes and a ROR less than 1 indicates that high-quality trials reported a smaller treatment effect compared with those trials that did not meet the quality criteria.

Heterogeneity and Effect Size Distributions

In a further analysis we explored the heterogeneity and effect size distribution shown in the different datasets. The meta-epidemiological datasets comprise different meta-analyses that each contain individual trials. We used I^2 to measure the percent of variation across trials in each meta-analysis datasets that is due to heterogeneity. In addition, we computed I^2 estimates across all trials at the meta-epidemiological dataset level as an indicator of the variation represented by the datasets. In order to address the variation in distributions found across datasets, we aimed to

sample individual datasets to mirror plotted distributions found in the various analyzed datasets. We used a non-parametric approach and ranked the studies included in dataset 1 and dataset 2 from the smallest to the largest effect size. We then matched them and assigned dataset 2 effect sizes to the dataset 1 quality criteria based on the rankings of the effect sizes within the datasets.

Monte Carlo Simulations

Finally, we created datasets by Monte Carlo simulation methods to systematically explore factors that influence the power to detect an association between quality and effect sizes in meta-epidemiological datasets. We determined the properties of the sampling distribution of the meta-regression estimates under different conditions to determine the factors that influence the likelihood of detecting associations between quality and effect sizes. We generated data sampled from populations that matched three meta-epidemiological datasets in terms of number of trials, sample size per trial and level of heterogeneity in the overall dataset. For each of the datasets we systematically altered two parameters: First, we randomly generated populations with quality effects of 0.1 and 0.2 (effect size differences between high and low quality trials). We modeled heterogeneity by adding a variance parameter to the simulations – each meta-analysis was comprised of studies sampled from a population with a specified effect size. Effect sizes were generated at the study level, and then (for heterogeneity) a random parameter was added to the effect size, to introduce population level heterogeneity. We used a heterogeneity value that matched observed results in the empirical datasets. To model reduced heterogeneity we halved the value of the parameter that was added and in addition used a heterogeneity parameter of zero. Outcome data were then generated for individual trial participants in for intervention and control groups for each trial, with 50 percent of individuals assigned to each group.

This gave 3 (levels of heterogeneity) × 2 (quality effects) = 6 cells in the simulation for each dataset. For each of these cells in the simulation we generated and analyzed 1000 datasets. A function was written in R version 2.12 to generate data, analyze simulations, and aggregate results.

Results

Association Between Quality and Effect Sizes in Empirical Datasets

Results for evidence of bias varied between meta-epidemiological datasets, clinical fields, and individual meta-analyses. In the new empirical dataset (‘Heterogeneity set’) compiled specifically for this report, associations between CBRG criteria and effect sizes were small, for example the ROR between unconcealed and concealed trials was 0.89 (95% CI: 0.73, 1.09). None of the associations was statistically significant but the large majority of RORs indicated that trials that did not meet investigated quality criteria such as concealment of treatment allocation, similarity of baseline values, blinding of outcome assessors, blinding of care providers, blinding of patients, use of intention-to-treat analysis, similarity of co-interventions, and similar timing of outcome assessment reported slightly larger treatment effects. Results for other published quality criteria and risk of bias indicators applied in parallel showed similar results.

Based on a quantitative summary score for the CBRG criteria, the data for the new meta-epidemiological dataset showed that a cut-off of six or more criteria met (out of 11) differentiated high- and low-quality trials best (ROR 0.86; 95% CI: 0.70, 1.06). Similarly, in two

of the other meta-epidemiological datasets a cut-off of five or six criteria showed the largest differences in effect sizes (e.g., cut-off 5 criteria met, effect size difference -20, 95% CI: -0.34, -0.06 and ROR 0.79, 95% CI: 0.63, 0.95) while one dataset showed no effect of quality (effect size difference 0.02, 95% CI: -0.015, 0.062). The difference in reported treatment effects based on the quantitative summary score only marginally exceeded those of individually criteria and the difference was not statistically significant in this new dataset. Across datasets, most consistent associations between quality and effect sizes were found for allocation concealment with concealed trials showing smaller treatment effects.

In individual meta-analyses that constituted each meta-epidemiological datasets, high-quality trials were sometimes associated with smaller reported effect sizes and sometimes with the opposite effect (associated with larger effect sizes). A correction for clustering by meta-analysis had no noticeable effect on estimates of the associations between quality and effect sizes.

Based on these results, it cannot be determined whether the proposed extended list of quality criteria should be applied regularly when judging the quality of studies (key question 1). However, in order to evaluate this finding, it is important to know the quality of the applied test and whether the lack of observed effect indicates conclusively that trial quality is not associated with reported effect sizes.

Heterogeneity and Effect Size Distributions

The amount of heterogeneity in dataset 1 estimated by I^2 ranged from 8.9 to 85.3 percent in individual trials and the overall estimated heterogeneity for the entire dataset was 72.4 percent. In dataset 2 (EPC reports), individual I^2 estimates were generally higher and ranged from 26.2 to 99.4 percent and the overall dataset estimate was 97.5 percent. Several meta-analyses in dataset 3 'Pro-bias' showed no evidence of heterogeneity; the overall dataset heterogeneity estimate was 59.6 percent. Some individual estimates were not statistically significant in dataset 4 ('Heterogeneity set') and the overall database estimate was 60 percent, which was comparable to dataset 1 and 3, although heterogeneity across studies was one of the explicit inclusion criteria for the selection of meta-analyses that were compiled for this dataset.

The effect size distributions varied across meta-epidemiological datasets; in particular the dataset 2 (EPC reports) distribution was less symmetric and bell-shaped than those of the other datasets. To investigate whether the distribution characteristics are correlated with the difference in associations between quality and effect sizes between datasets, we used a non-parametric sampling approach to mirror the distribution of dataset 2 in other datasets. The effect size histograms showed that the sampling method was successful in creating a similar distribution shape. With regard to associations between quality and effect sizes, dataset 1 (Back pain) which had originally shown consistent effects of quality now showed conflicting results—some criteria were associated with effect sizes, some were not, and the direction of effects varied across quality criteria. To investigate whether this observation could be replicated in another dataset, we applied the process to dataset 3 ('Pro-bias'). The difference between the original data and the new data was less clear in this dataset. This finding may in part be due to inconsistencies across quality criteria that had appeared in the original data and in part to the fact that it is difficult to compare ratios of odds ratios and effect sizes.

Monte Carlo Simulations

Trial quality is typically explored as a potential source of heterogeneity across trials; however, the simulation analyses show that additional heterogeneity can reduce the power to detect statistically significant trial quality effects (key question 2).

In Monte Carlo simulations designed to reflect the characteristics of the empirical datasets, power to detect quality moderator effects in three sets of data was variable, and for many datasets of study parameters power was low. Even large quality effects mirroring substantial differences in reported treatment effects between high- and low-quality trials, in simulations set up to maximize statistical power (by assigning 50 percent of trials to high quality and 50 percent to low quality) could not be detected in the presence of a large amount of additional heterogeneity across trials, that is heterogeneity not due to quality.

These results indicate that failure to detect quality effects should not be taken as evidence that there are no quality effects. Furthermore, based on our analyses, individual meta-analyses should include steps to minimize heterogeneity through the inclusion of additional study level covariates. These refinements can reduce unexplained heterogeneity and thereby aid the investigation of quality effects and the potential for bias.

Future Research

Our analyses have shown that it is challenging to detect effects of study quality on reported treatment effects. This is the case for individual meta-analyses as well as meta-epidemiological datasets. From this it follows that the failure to detect a statistically significant quality effect should not be interpreted as meaning that a quality effect is not present.

Future studies that investigate the effects of quality as a moderator of outcomes in randomized trials should take steps to ensure that unexplained heterogeneity is minimized. In meta-epidemiological datasets, minimizing heterogeneity can be achieved though many means (e.g., utilizing a larger number of trials). Following our analyses, individual meta-analyses might achieve an adequate level of heterogeneity through the inclusion of additional study-level covariates when investigating the association between trial quality and effect sizes.

More empirical evidence is needed to determine which quality features are likely to influence reported effect sizes, and under which conditions. This question is of particular importance for the critical appraisal of systematic reviews when aiming to summarize the existing evidence appropriately.

Conclusion

Although trial quality may explain some amount of heterogeneity across trial results in meta-analyses, the amount of additional heterogeneity in effect sizes is a crucial factor determining when associations between quality and effect sizes can be detected. Detecting quality moderator effects requires more statistically powerful analyses than are employed in many investigations.

Introduction

Background and Scope

For evidence syntheses that systematic reviews provide, the critical appraisal of studies is an important process. The quality of the evidence and the potential for bias in reported results should be taken into account when evaluating individual studies or the overall existing evidence base. Variation in the quality of studies may also explain differences in individual study results across studies. The quality of studies is routinely assessed in meta-analyses as a potential source of heterogeneity; that is, variation in study results across different studies. For researchers preparing overviews and policy makers utilizing these evidence overviews, it is important to know which features, if any, are most likely to distort study results. However, associations between study quality and effect sizes are complex, empirical evidence of bias has been established only for selected quality criteria, and open questions regarding the presence and detection of bias should be explored further.

A large number of individual quality criteria and quality checklists or scales have been proposed for controlled trials (see e.g. Moja, Telaro D'Amico, et al. 2005; West, King, Carey, et al., 2002). These tools cover primarily potential threats to the internal validity of the trial methodology. Juni et al. (2001) differentiate dimensions related to selection bias (e.g. inadequate randomization), to performance and detection bias (e.g. outcome assessors not blind to the intervention allocation), and to attrition bias (e.g., deviations from the randomization protocol and analysis exclusions). The assessment of the methodological quality of a trial is closely linked to the quality of the reporting. Typically, only the information reported in the publication on the trial is available to the reader to judge the quality of the trial. Quality can also relate to the external validity of studies which refers to the generalizability of study results; quality is undoubtedly a multidimensional concept (Juni et al., 2001), definitions vary, and there is little consensus on its scope.

Quality checklists typically provide a selection of quality features that are scored individually. In addition, quality scales provide a total quality score, a quantitative summary score derived from the individual criteria 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 primarily the conduct of the individual study or its research methodology and concern the internal validity of the research study, but frequently include also other quality aspects such as the quality of the reporting of trial evaluation. 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 RCTs. A central criterion, the concealment of treatment allocation, was introduced by Schulz et al. (1995) and is widely used in addition to the criteria proposed by Jadad et al. (1996).

Design and execution factors of randomized controlled trials (RCTs) are widely believed to be associated with the treatment effects reported for those studies. This association is an indicator of bias. We define bias as a systematic deviation of an estimate, in this case the deviation of the estimated treatment effect from the true value. More factors have been proposed to be related to bias than have actually been confirmed by systematic examination of associations between quality and reported treatment effects. When assessing the quality of trials it is assumed that the conduct of the research methodology may 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. There is evidence for some methodological variables showing that low quality trials exaggerate treatment effects. Colditz, Miller, and Mosteller (1989) found RCTs to have smaller effect sizes than non-RCTs in trials of surgical therapy, and RCTs that were double-blind had smaller effect sizes than non-blinded trials of medical therapy. Schulz, Chalmers, Hayes et al. (1995) reported that inadequate concealment of allocation accounted for a substantial increase in effect sizes. The lack of double blinding was also shown to be associated with an increase in reported treatment effect. Moher, Pham, Jones, et al. (1998) used Jadad's scale and Schulz's "concealment of allocation" in a large study that assessed 11 meta-analyses (including 127 RCTs). All trials were scored and the meta-analyses replicated. Low-quality trials were associated with an increased treatment estimate compared with high-quality trials. Studies with inadequate treatment allocation concealment also showed an increased effect size 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 and double blinding, whereas the generation of treatment allocation did not show a statistically significant effect across datasets. Allocation concealment may show the most consistent associations with effect sizes (Hempel et al., 2011; Kjaergard, Villumsen, & Gluud, 2008).

The quality of individual trials is of particular importance to systematic reviews. Reviews that aim to summarize the available evidence adequately are particularly affected by results that depend on the quality of the trial. The methodological quality of studies included in a systematic review can have a substantial impact on treatment effect estimates (Verhagen, de Vet, de Bie, Boers, & van den Brandt, 2001). Pildal, 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 due to loss of power (as a result of a smaller sample size) but also due to 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 2008 Cochrane Handbook (Higgins and Green, 2008) introduced a Risk of Bias tool that suggested the assessment of the randomization sequence generation, the concealment of treatment allocation, blinding, the adequate handling of incomplete outcome data, selective outcome reporting, and other sources of bias. These criteria were selected based on the existing evidence for bias; that is, an association between quality criteria and effect sizes, and the tool is much more comprehensive than many existing quality checklists. The tool also refrains from using a quantitative summary score approach for an overall assessment of the risk of bias of the study and suggests that reviewers take the individual quality domains into account and use their judgment to decide which domains are crucial for the evaluation. However, the inter-rater reliability of these judgments may be limited (Hartling et al., 2009). Furthermore, the tool explicitly put more emphasis on individual outcomes and suggested that the potential for bias should be assessed for each individual outcome rather than assessing the study's overall quality or overall risk of bias.

The Cochrane Back Review Group (CBRG) Editorial Board developed an 11-item criteria list in an expert-guided process for the assessment of trials included in Cochrane reviews by the group. The items cover established quality criteria (allocation concealment, blinding) as well as criteria for which the potential for bias has rarely been investigated or existing investigations

showed conflicting results (e.g., similarity of co-interventions, compliance). The tool addresses aspects of internal validity of trials as well as the quality of the reporting. These quality criteria showed consistent influences on effect sizes of trials reporting interventions for back pain (van Tulder, Suttorp, Morton, et al., 2009). In addition, a quantitative summary score of 0 to 11, based on the 11-item list, was applied as a measure of overall internal validity. A cut-off of 5 or 6 criteria met (out of 11) differentiated high and low quality trials best (that is showing the largest difference in effect sizes between high- and low-quality studies). When applying the criteria to other datasets, we found mixed results (Hempel et al., 2011). Associations between individual as well as sum scores were found in one out of two additional datasets of interventions covering a wide range of clinical areas.

Presence and Detection of Associations Between Quality and Effect Sizes

The variation in associations between quality and reported effect sizes across datasets raises several questions. In our analyses, we found that in one meta-epidemiological dataset the extended list quality criteria (CBRG quality criteria set) did not appear to be useful in differentiating high and low quality studies, meaning that effect sizes were not consistently higher in studies not meeting quality criteria across the CBRG criteria compared with studies meeting the criteria. However, in this same dataset, even well established individual criteria such as blinding, as well as summary scores such as the Jadad score or the CBRG summary score, showed no associations with effect sizes. The dataset included meta-analyses of EPC reports covering a wide range of clinical fields (Hempel et al., 2011). From this result it can be concluded that either the CBRG criteria did not apply to the trials present in the dataset or that other factors inherent in the dataset influenced either the presence or the detection of associations between quality and effect sizes. Similarly, Emerson, Burdick, Hoaglin, et al. (1990) found no relationship between a consensus-developed quality scale (0 to 100 points) and treatment effect differences. Balk, Bonis, Moskowitz, et al. (2002) applied 24 existing quality measures and assessed a number of meta-analyses involving 276 RCTs. 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.

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 (or its detection) and that may explain some of the conflicting results shown in the literature, is a new and evolving field. The following two paragraphs outline factors that have been discussed in the literature and that may influence the shown association between quality and effect sizes in meta-analyses and meta-epidemiological datasets.

Among other factors, it has to be taken into account that critical appraisal in systematic reviews is based on reported information. The information reported depends on the information authors choose to report, which is likely influenced by the word limits that many journals impose, which may make it impossible for authors to fully explain the trial methodology. Reported characteristics depend to some extend on the convention at the time of publishing and journal requirements. The publication of the Consort Statement (Begg et al., 1996) provides much-needed guidance for authors to enable them to standardize and improve the reporting of RCTs.

Another factor to consider is that not much is known about the reliability of the assessment process. Few tools have been psychometrically evaluated. The Jadad scale is one of the few tools with known inter-rater reliability (Jadad et al., 1996). Recently, the Cochrane Risk of Bias tool has been evaluated (Hartling et al., 2009), and the results indicate that more guidance is needed in order for reviewers to agree about the risk of bias in studies, particularly when global assessments about the overall risk of bias are performed. However, the reliability of critical appraisals as part of systematic reviews is largely unknown and the inter-rater agreement between individual ratings may be a poor estimate. Typically, in systematic reviews, reconciled ratings are used, where discrepancies between two or more independent reviewer assessments are discussed and reconciled. These reconciled ratings should reduce individual reviewer errors, that is, random errors, and also, to some extent, individual reviewer bias.

Furthermore, whether reported results are prone to bias or whether bias can be shown might be a characteristic of the outcome or the intervention. Wood, Egger, Gluud, et al. (2008) used three datasets of meta-epidemiological studies, that is, studies investigating the associations of quality features and effect sizes (Schulz, Chalmers, 1995, Kjaergard, Villumsen, 2001, Egger, Juni, 2003). The group investigated whether the nature of the intervention and the type of outcome measures influence the effect of allocation concealment and blinding. They found that trials using subjective outcomes showed exaggerated effect sizes when there was inadequate or unclear allocation concealment or lack of blinding. In trials using objective outcomes such as mortality, the association of quality with trial results was negligible. Differentiating drug interventions and non-drug interventions, which was explored in a further analysis, indicated no significant differences on the effect on allocation concealment or blinding.

Other factors that are inherent to datasets may influence our ability to detect effects of trial quality. In a previous AHRQ report on empirical evidence for associations between quality and effect sizes (Hempel et al., 2011) we outlined a number of factors, such as the size of the treatment effect, the condition being treated, the type of outcome measure, and the variation in effect sizes, within the dataset that may potentially influence when quality effects lead to bias and whether the association can be detected in a dataset. The role of some of these factors can be explored in datasets of "known quality"—published datasets where basic characteristics have already been established. Some of the factors and their effects can be tested by using simulations of meta-analyses. Simulations have rarely been applied to meta-analytic questions but can be a powerful tool in systematically assessing the effects of hypothesized factors (e.g., Field, 2001; Field, 2005; Morton, Adams, Suttorp et al., 2004).

To pursue these open questions we combined the use of different meta-epidemiological datasets with "known qualities" and simulation methods for this report.

Objectives and Key Questions

The objectives of the project were to examine associations between individual and summary indicators of trial quality and effect sizes and to explore factors influencing the detection of associations in meta-epidemiological datasets. The selected quality criteria addressed design and execution factors of the trial as well as the quality of the reporting. For this project we are interested in the association between trial quality criteria and the size of the reported treatment effect. The project aimed to answer the following questions:

- Are the selected quality criteria, individually as well as combined, useful as indicators of bias in diverse clinical contexts? The usefulness was operationalized as predictive

validity – whether meeting or not meeting the quality criteria is associated with differential effect sizes in treatment effect trials.

- Which factors influence the presence and the detection of associations between quality and effect sizes? The question was investigated in empirical meta-epidemiological datasets as well as Monte Carlo simulation models.

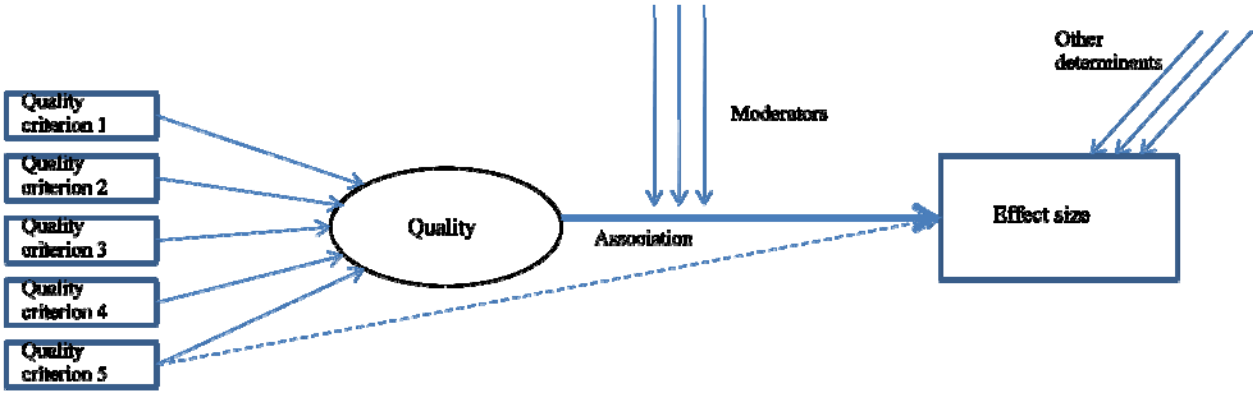
We expect our results to contribute empirical evidence to the continuing professional debate about the appropriate role of quality criteria in systematic reviews of RCTs.

Analytic Framework

We tested the hypothesis that the investigated quality criteria show an association with reported effect sizes in different meta-epidemiological datasets and individual clinical areas and meta-analyses. In addition, we investigated moderating factors (such as the effect of heterogeneity and the effect size distribution) that might influence the detection of theses associations.

Figure 1 represents the underlying assumptions.

Figure 1. Analytic framework: presence and detection of associations between trial quality and effect sizes



The figure shows a simplified diagrammatic representation of the assumption that there is an association between quality features of research studies and the size of the treatment effect reported for the study. This association is represented by the bold arrow. The arrows in the figure indicate the direction of effects, for example quality is assumed to influence the effect sizes (as opposed to an assumption of trial effect sizes influencing the quality of a trial). Each quality criterion is either individually linked to effect sizes or contributes to a quality score that is a composite of individual criteria.

Effect sizes are influenced by many variables in addition to the methodological quality of the research design and the way the study is conducted. The figure depicts the assumption that other variables apart from quality will influence effect sizes; in the regression equation, these are referred to as error. These other variables include the true effect of the intervention, as other potential influences and measurement error; these variables are termed other determinants. 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).

In addition, we assume that there are factors (moderators) that influence the detection of associations between methodological quality and the reported effect size, represented by the

vertical arrows. These factors determine whether an association can be observed in empirical investigations. Potential factors are the size of the treatment effect, issues related to the condition being treated, the type of outcome measure, the variance in effect sizes across included trials (heterogeneity) or distribution characteristics of quality criteria and effect sizes.

The diagram also shows assumptions that need to be made regarding the relationship between these variables, namely that both quality and other factors (such as efficacy) must not be related to the random error that is associated with each study. For example, if there were a factor that predicts both the effect size of a study and the quality of that study there would a relationship between quality and effect size; however, this would be caused by the fact that there was a common cause, and hence the analysis makes the assumption that these variables are unrelated.

Methods

Quality Criteria

We applied the CBRG Internal Validity criteria (van Tulder et al, 2003) to individual trials included in a dataset. In addition, we used the Jadad scale (Jadad et al., 1996), criteria proposed by Schulz et al. (1995), and the Cochrane Risk of Bias tool (Higgins & Green, 2008). The items and the scoring guideline are shown in the appendix. The criteria address the internal validity of the trials, design and execution factors, and the adequacy of the reporting of the trial.

CBRG Internal Validity 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 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 [ITT] analysis), similarity of co-interventions, adequacy of compliance, and similar assessment timing across groups.

The items are scored 'Yes,' 'No,' and 'Unclear.' There is guidance for the appropriateness of each answer category (see appendix for the full scoring instructions). For example, assessor blinding 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 typically defined according to the clinical field, that is, in order to select the most relevant variable to adequately judge baseline comparability or to determine thresholds for dropouts that are adequate for the individual clinical field.

Jadad Scale

In addition, we applied the Jadad scale (Jadad et al., 1996) as one established measure of study quality. Use of this scale entails the assessment of the presence and quality of the randomization procedure (0 to 2 points), the presence and quality of double-blinding procedure (0 to 2 points), and the description of withdrawals (0 to 1 point). The items are summed, with summary scores varying from 0 to 5.

Schulz's Criteria

For comparison reasons, we also used criteria proposed by Schulz et al. (1995), operationalized as in the original publications. Schulz introduced the assessment of the concealment of treatment allocation, a dimension that showed the most consistent differences in high- and low-quality studies in previous analyses. In addition, the generation of the allocation sequence, the inclusion in the analysis of all randomized participants, and the reporting of double blinding are scored as present or not.

Cochrane Risk of Bias Tool

Finally, we also applied the Cochrane Risk of Bias tool, a widely used tool to assess RCTs published in the 2008 version of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2008). The quality features that were assessed were whether the

allocation sequence was adequately generated, whether the allocation was adequately concealed, whether knowledge of the allocated treatment was adequately prevented during the study (blinding), whether incomplete outcome data were adequately addressed, whether the reports of the study are free of suggestion of selective outcome reporting, and whether the study was apparently free of other problems that could put it at a high risk of bias. Finally, the reviewers assessed the overall risk of bias of each study, expressed as high, low, or unclear. The Cochrane Risk of Bias Tool has since been revised and now distinguishes assessor blinding from blinding of participants and personnel (Higgins & Gree, 2011).

The table below (Table 1) shows the quality domains that are represented in the different critical appraisal instruments that were applied in parallel. The individual interpretation of the quality domains, that is the operationalization and scoring instructions, vary across instruments and are shown in full in the appendix. Some translations address the reporting of the trial details, the design and execution factors, or both.

Table 1. Represented quality domains

Quality Domain	CBRG	Jadad	Schulz	RoB
Randomization	x	x	x	x
Allocation concealment	x		x	x
Similar baseline	x			
Assessor blinding	x			
Care provider blinding	x			
Patient blinding	x			
Dropouts and withdrawals	x	x		
Original group (ITT)	x			
Similar co-interventions	x			
Acceptable compliance	x			
Similar timing	x			
Blinding (summary item)		x	x	x
Analysis of all pts / exclusions			x	
Incomplete outcome data				x
Selective outcome reporting				x
Other sources of bias (undefined)				x
Quantitative summary score	x	x		
Overall risk of bias assessment				x

CBRG = Cochrane Back Review Group; RoB = risk of bias; ITT = intention to treat
Note: The individual operationalizations and scoring instructions vary across instruments.

Several domains relate to the potential for selection bias (randomization, allocation concealment, similar baseline for treatment and control group), the potential for performance and detection bias (assessor blinding, patient blinding, care provider blinding or summary blinding judgment; similar co-interventions, similar timing of the outcome assessment in treatment and control groups), the potential for attrition bias (description, rate, and handling of dropouts and withdrawals, analyses of participants according to randomization / intention-to-treat analysis, exclusion from the analyses, incomplete outcome data), and unique dimensions (compliance, selective outcome reporting, other (undefined) sources of bias), or summary assessments (quantitative or qualitative).

In comparison, the CBRG criteria cover all general dimensions assessed by the other three measures apart from the selective outcome reporting in addition to unique dimensions (e.g., similar timing of outcome assessments). However, individual interpretations and scoring instructions vary across the instruments.

Study Pool Selection

This project drew on empirical study pools as well as Monte Carlo simulations to estimate effects.

We used four epidemiological datasets to investigate the research questions. The four datasets consisted of up to 12 meta-analyses each, and each meta-analysis included a varying number of individual trials ranging from 3 to 45 trials. There were no overlaps in included trials across datasets. Three datasets have been described in detail in previous work; one new dataset was added for the purpose of this report. Two datasets were available to us through previous work, two other datasets were assembled for their "known characteristics" with regard to associations between quality and effect sizes as outlined in detail below.

Dataset 1: Back Pain Trials

RCTs in this dataset were included in reviews of non-surgical treatment for non-specific low-back pain present in the Cochrane Library 2005, issue 3 (Assendelft, Morton, Yu, et al. 2004; Furlan, van Tulder, Tsukayama, et al., 2005; Furlan, Imamura, Dryden, et al., 2008; Hagen, Hilde, Jamtvedt et al., 2001; Hayden, van Tulder, Malmivaara, et al. 2005; Henschke, Ostelo, van Tulder, et al., 2005; Heymans, van Tulder, Esmail, et al., 2005; Karjalainen, Malmivaara, van Tulder, et al., 2001; 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). The reviews from eight topic areas assessed the effects of acupuncture, back schools, behavioral therapy, exercise therapy, spinal manipulative therapy, muscle relaxants, non-steroidal anti-inflammatory drugs (NSAIDs), and other approaches (bed rest, lumbar supports, massage, multidisciplinary bio-psycho-social rehabilitation, and transcutaneous electrical nerve stimulation) for the treatment of low-back pain. Comparisons were placebo, usual care, no treatment, or other treatments. The dataset included 216 trials. This dataset is described in detail elsewhere (van Tulder et al., 2009).

Dataset 2: EPC Reports

This dataset was assembled for a previous methods report on associations between quality and effect sizes (Hempel et al., 2011) and is 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 dataset of trials that represented a wide range of clinical topics and interventions. As outlined in the report, 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 needed only assess the quality of the included trials. The dataset was drawn from 12 evidence reports, the majority of which 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 diverse topics, and included pharmacological therapies as well as behavior modification interventions. All trials 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.

Dataset 3: "Pro-bias" set

This third dataset was obtained by replicating a selection of trials used by Moher et al. (1998). The dataset was chosen because it has shown evidence of bias for established quality criteria (see Moher et al., 1998) and is, therefore, designated in this report as "pro-bias." We replicated the methods described by Moher et al. for selection of trials. Two reviewers independently reviewed the 11 meta-analyses chosen by the authors. These meta-analyses covered digestive diseases (Marshall & 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 and Sandercock, 1995), and pregnancy and childbirth (Hughes, Collins, and Vanderkeckhove, 1996). We were able to retrieve, quality score, and abstract 100 RCTs of the originally published dataset (79 percent).

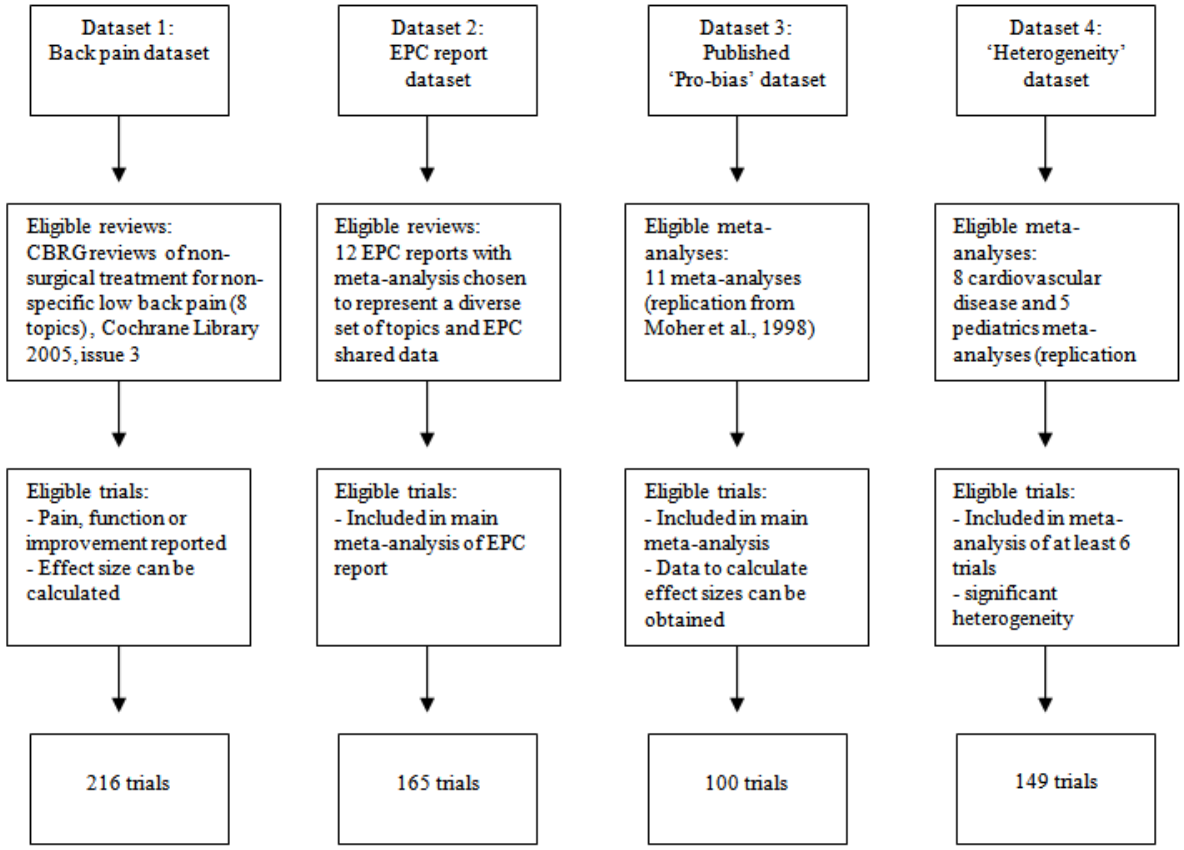
Dataset 4: "Heterogeneity" set

For the purpose of this report, we compiled a fourth dataset of meta-analyses and included trials. This dataset was obtained by replicating a selection used by Balk et al. (2002). The dataset was chosen because heterogeneity across studies was one of the inclusion criteria to select meta-analyses for the dataset. In addition, the "known quality" for this dataset was its demonstrated lack of reliable association between quality and effect sizes across studies and quality criteria. For this dataset meta-analyses were included that demonstrated significant heterogeneity in the odds ratio scale defined as $p<0.10$ for the chi-square statistic of between-study heterogeneity or a nonzero variance tau-squared (τ^2) in DerSimonian and Laird random-effects models. We selected the eight included cardiovascular disease meta-analyses and the five pediatric meta-analyses from this dataset. Since the original publication does not specify exactly which trials were included in the analysis, we replicated the methods described by Balk et al. (2002) for trial selection.

The cardiovascular disease trials were derived from published meta-analyses on treatment with aspirin (Antiplatelet Trialists' Collaboration, 1988), Class I antiarhythmics (Hine et al., 1989), anticoagulants (Leizorovicz and Boissel, 1983), beta-blockers (Yusuf et al., 1985), intravenous streptokinase (Yusuf et al., 1985), nitrates (Yusuf et al., 1988), cholesterol reduction agents (Rossouw et al., 1990), and magnesium (Teo and Yusuf, 1993), and an update search published by Lau et al. (1992). The pediatrics meta-analyses investigated glucocorticoids (Ausejo et al., 1999), dexamethasone (Bhuta and Ohlsson, 1998), bronchodilators (Kellner et al., 1996) , short-course antibiotics (Kozyrskyj et al., 1998), or antibiotics (Rosenfeld & Post, 1992). The individual trials reported on the outcome mortality, improved croup score, chronic lung disease, unimproved bronchiolitis distress score, acute otitis media failure to cure, and otitis media with effusion failure to cure. Our replication of these two clinical topics includes 149 trials; Balk et al. (2002) included 153 trials. The references of the included trials are shown in the appendix.

The flow diagram (Figure 2) summarizes the dataset composition.

Figure 2. Flow diagram



CBRG = Cochrane Back Review Group; EPC = Evidence Based Practice Center

Procedure

For all datasets, two reviewers independently rated each trial by applying the outlined criteria. For dataset 1 we used the published quality scores from the individual Cochrane reviews. For dataset 2 to 4 the majority of trials were rated by the same set of reviewers. We developed and pilot tested a standardized form to record decisions for the quality and risk of bias criteria. The reviewers used the full publications to score the studies and were not blinded to the identities of authors, journals, or other variables. The reviewers were experienced in critical appraisal of research studies 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 participants in each treatment group. The selected outcome per trial was determined by the meta-analyses the trial was part of. The outcome was either the primary outcome, or the outcome in a meta-analysis that included the largest number of trials, where more than one meta-analysis was presented and trials reported more than one outcome.

Most trials were compared against placebo. In trials with active comparisons, the coining of treatment and control group (that is the decision which group was considered the intervention

and which the control group for the analysis) was either guided by input from experts in the research field or applied the selection made in the original meta-epidemiological dataset (two of the utilized datasets are replications of previously reported datasets). Sensitivity analyses such as restricting to data from placebo-controlled only studies did not indicate effects of the coining or the use of absolute effect sizes on associations between quality and effect sizes. However, for dataset 1 (Back pain) absolute effect sizes were selected as the final measure because this dataset included more comparisons between treatment and placebo as well as comparisons between active treatments than the other datasets.

For dataset 1 (Back pain) and 2 (EPC reports), in order to be able to combine studies within datasets or potentially between datasets, standardized effect sizes (ES) were computed for each study. As all studies in dataset 3 (Pro-bias) and dataset 4 (Heterogeneity set) reported dichotomous outcomes, odds-ratios were calculated. As a quality check, the point estimate and 95 percent confidence interval (CI) of each meta-analysis included in each dataset was calculated and compared to the original meta-analytic result. To explore effects of coining in dataset 4, we calculated ratios of odds ratios separately for studies favoring the treatment group in a sensitivity analysis.

The Monte Carlo simulations and the effect size distribution analyses are based on datasets 1 to 3. Dataset 4 was compiled in parallel to this work and was primarily used to replicate associations between quality and effect sizes in a dataset that had been selected in parts based on the presence of heterogeneity in meta-analyses.

Analysis

Association Between Quality and Effect Sizes in Empirical Datasets

We investigated the association between quality and effect sizes in two ways. First, the differences between results in studies that met a quality criterion and those that did not were calculated for each quality feature. Secondly, we used a quantitative summary score and explored different cut-offs of quality scores according to the number of quality criteria met.

For all analyses, we differentiated quality items scored “yes” and those with the quality item scored “not yes,” which included the answers “no” and “unclear.”

Trials in two of the datasets used a continuous outcome, and two used a dichotomous outcome. For continuous outcomes we used the difference in effect sizes between two subgroups (studies with criterion met versus studies that did not meet the quality criterion) as a measure of bias. The difference was estimated using meta-regression (Berkey et al., 1995). A random effects meta-regression was conducted separately for each quality criterion. The coefficient from each regression estimates the difference in effect sizes between those studies with the quality feature scored “yes” (criterion met) versus “not yes” (criterion not met or unclear). No effect of quality would be shown as identical effect sizes between high- and low-quality trials, that is effect sizes would be independent from meeting or not meeting quality criteria. A difference with a significance level of $p<0.05$ was considered statistically significant.

In two other datasets (set 3 and 4), all studies used dichotomous outcomes. An odds ratio below 1 indicated the treatment group is doing better than the control group. For the analysis, we compared odds ratios (OR) of studies where the quality criterion was either met or not met and computed the ratio of the odds ratios (ROR). The ROR is OR(no)/OR(yes) where OR(no) is the pooled estimate of studies not meeting the quality criterion and OR(yes) is the pooled estimate of studies where the quality criterion is met.

We also aggregated across CBRG criteria and computed quantitative summary scores based on the number of criteria met. Different cut-offs (depending on the number of criteria met) were explored to differentiate high and low quality studies. The difference in effect sizes and ratios of odds ratios of studies above and below possible thresholds was investigated.

All meta-epidemiological datasets used in this project consisted of trials that were selected through meta-analyses. These meta-analyses then contributed individual trials that make up the total dataset. Different statistical techniques have been suggested to investigate the effects of study characteristics such as the quality of individual trials and approaches vary regarding the integration of the clustering (e.g., Sterne et al., 2002). We investigated the effects of clustering by contrasting the corrected and uncorrected associations between quality and effect sizes. Meta-regressions correcting for clustering with meta-analysis were analyzed using a Huber/White (sandwich) estimator (Hedges et al., 2010).

Heterogeneity and Effect Size Distributions

As outlined, for the purpose of this report we assembled one dataset based on a published meta-epidemiological dataset that was specifically designed to represent heterogeneity between trials. One of the selection criteria for the trials that were included in this dataset was that meta-analyses used to establish the dataset had to report evidence of heterogeneity across trials (see Balk et al., 2002). Heterogeneity in meta-analysis refers to the variance in the estimated effect sizes between studies. In a fixed-effects meta-analysis, it is assumed that all studies are sampled from the same population, and hence variation in e sizes is due only to random variation. The degree to which the effect size in each study varies from the estimated population effect size is a function of the sample size of that study. Larger studies would be expected to have point estimates of effect size which are closer to the pooled estimate than smaller studies. Thus if the fixed-effects assumption is true, a meta-analysis containing studies with smaller sample sizes would have greater variance in effect sizes than a meta-analysis containing larger sample sizes.

Heterogeneity in meta-analysis is quantified using several measures. Q is a test statistic which assesses the total variance of the effect sizes of the studies. When the fixed-effects assumption is true, the expected value of Q is approximately equal to the number of studies minus one. The value of the statistical significance of Q depends on the number of studies, hence interpretation is difficult. An alternative to Q is I^2. I^2 is calculated as: $100(Q-(r-1))/Q$, where r is the number of studies included in the meta-analysis. I^2 is the proportion of variance in the effect sizes that cannot be explained by chance. I^2 values close to zero indicate little or no heterogeneity, whereas those closer to 100 percent indicate higher levels of heterogeneity (Higgins, Thompson, Deeks, and Altman, 2003). Values of I^2 of 25 percent are considered small, 50 percent are considered moderate, and 75 percent are considered large. When Q is lower than r-1, values of I^2 below zero are possible. However, when this is the case, I^2 is bounded at zero, and is constrained to be zero.

The datasets comprise different meta-analyses that each contain individual trials. In the empirical datasets we used I^2 to measure the percent of variation across trials in each meta-analysis that is due to heterogeneity. In addition, we computed I^2 across all trials at the dataset level regardless of the meta-analyses the trial was originally part of as an estimate of the variation represented by the datasets.

Effect Size Distribution

In our previous report (Hempel et al., 2011), we discussed our observation that the three different datasets used to investigate associations between quality and effect sizes varied in a number of characteristics. In particular, the dataset distributions of both effect sizes and quality scores of the included trials varied considerably across analyzed datasets. Compared to the other datasets, dataset 1 (Back pain data, van Tulder et al., 2009) was more symmetric, and this dataset also showed consistent quality-effect size associations. In contrast, dataset 2 showed a noticeable skewed distribution of reported effect sizes and incidentally also no association between quality scores and effect sizes. The variation in effect sizes alone may explain the differential associations between quality and effect sizes observed across datasets. In order to address the variation, we aimed to sample datasets to mirror distributions found in the various analyzed datasets in an exploratory analyses.

First, we manually sampled effect sizes from dataset 1 in an attempt to replicate the distribution of effect sizes seen in dataset 2. However, the inspection of the resulting distributions showed that we were not able to match the shape of the distributions satisfactorily. Instead, we used a non-parametric approach in order to mirror the various distributions. First we ranked the trials included in dataset 1 and dataset 2 from the smallest to the largest effect size. Then we matched the rankings and assigned dataset 2 effect sizes to the dataset 1 quality criteria based on the rankings of the effect sizes within the datasets. Hence the quality scoring was kept the same from one dataset and we sought to find out what the associations between quality and effect sizes would be if the effect sizes and the resulting distribution were those of the second dataset.

Since the datasets varied in size, we sampled from the largest dataset (without replacement) to get to the number of studies in the smallest dataset. Hence, we sampled 165 studies from dataset 1 to map to the 165 in dataset 2. For dataset 3, we sampled 100 dataset 2 studies to map to the data.

The treatment effects are expressed as effect sizes in dataset 1 and 2 but as odds ratios in dataset 3, as dataset 3 included trials reporting on categorical outcomes. Thus, for dataset 3 ("Pro-bias"), we computed the ratio of odds ratios between high and low quality trials.

Monte Carlo Simulations

We created datasets by Monte Carlo simulation methods to systematically explore the effect of factors that influence the ability to detect the association between quality and effect sizes in meta-epidemiological datasets. We investigated the effect of sample size, heterogeneity, and the size of the quality effect operationalized as a difference in effect sizes between low and high quality studies. We determined the properties of the sampling distribution of the meta-regression estimates under different conditions to determine the factors that influence the likelihood of detecting associations between quality and effect sizes.

Meta-analyses combine the results of multiple studies testing the same hypothesis and estimate a pooled effect. The reported effect sizes are estimates of the true (population) treatment effect. Effect sizes across individual studies may show variation. As outlined, study quality has been found to influence effect sizes in a number of meta-epidemiological studies (but not all) using datasets consisting of a number of meta-analyses and their included trials. Typically, lower quality studies are being associated with larger effect sizes. In a meta-analysis, failure to account for study quality may lead to additional heterogeneity and bias in the parameter estimates. However, not all meta-epidemiological studies or individual meta-analyses find an association

between quality and effect sizes. The circumstances under which quality is associated with bias (a deviation from the true treatment effect) are largely unknown.

The power of a statistical test of a null hypothesis is the probability that the null hypothesis will be rejected – that is, that the test is statistically significant – given certain population parameters, and a sample size for that study. For many statistical tests, it is relatively straightforward to estimate the power of the test: For example, for a *t*-test comparing the means of two samples with equal variance, knowing the size of the sample, the standardized difference between the means (in the population), and the value to be used as a cut-off for alpha (almost always 0.05), we can estimate the probability that the null hypothesis will be rejected. For more complex statistical tests the distributional characteristics of the test statistics are not necessarily known and hence power cannot be calculated.

We undertook Monte Carlo simulations to determine the power of the meta-regressions in datasets to detect quality effects, given different levels of quality effects and sample sizes. In a first step, we investigated the power and average parameter estimates for different models in Monte Carlo simulations, e.g., in models where quality was not considered compared to models where quality was added to the model. In these basic models, we also investigated moderator effects such as the size of the treatment effect. The basic models assumed an arbitrary number of trials per meta-analysis and participants per treatment. However, the final results presented in this report are based on models that correspond to existing empirical datasets to increase their external validity. Three datasets were used (Back pain, EPC reports, Pro-bias), all described in detail elsewhere (Hempel et al., 2011).

Design

We used Monte Carlo simulation to examine the effects of heterogeneity on the ability of meta-analyses to detect quality effects in datasets that correspond to existing empirical meta-epidemiological datasets. We generated data sampled from populations that matched three meta-epidemiological datasets in terms of number of trials, sample size per trial and level of heterogeneity in the overall dataset. We then randomly assigned 50 percent of trials to be high quality and 50 percent to be low quality. Use of 50 percent maximizes the power to detect quality effects. In fact, the proportion of high-quality studies varied across the datasets and between measures of quality; however the effects of variation in the proportion of categorical predictor variables are well known, hence we did not investigate this factor. Given that the additional structure of trials being included in about one dozen meta-analyses per dataset had only a negligible effect on the associations between quality and effect sizes, the additional structure was not added to the simulation model and for each dataset only the number of trials, not the specific number of meta-analyses, was integrated into the model.

The simulations proceeded in the following manner. First, we generated a vector of effect sizes (d_j) for each trial. Trials were randomly assigned to be high quality or low quality. For high quality trials, $B = 0$, for low quality trials, $B > 0$, indicating a larger effect size for lower quality trials. Then for J trials sampled from a normal distribution with mean of D and variance v.

$$d_j \sim N\left(D + Q_j, v\right)$$

The second stage was to generate individual trial participant data (y_{ij}) for n_j participants within each trial, by multiplying the treatment effect for that trial (d_j) by the condition of the

individual in that trial (x_{ij}), giving a value of either 0 or d_j, the individual participant value was then sampled from a normal distribution with mean equal to either 0 or d_j.

$$y_{ij} \sim N(d_j x_{ij}, 1)$$

Thus:

$$y_{ij} = (d_j + L_J B + \zeta_j) x_{ij} + \varepsilon_{ij}$$

Where:

y_{ij} is the outcome variable for individual i in study j.

d_j is the effect size for high quality studies.

L_J is a dichotomous (0, 1) indicator of low study quality.

B is the increase in effect size associated with a low quality study.

ζ_j is the random heterogeneity parameter, with standard deviation equal to ν.

x_{ij} is a dichotomous individual level indicator of intervention group status (0, 1).

ε_{ij} is a random error term, with standard deviation equal to 1.

When the data had been generated, we calculated the mean and standard deviation of the intervention and control groups for each study.

For each of the three datasets we systematically altered two parameters: First, we generated populations with quality effects (B) of 0.1 and 0.2 (on the standardized effect scale), reflecting the approximate range of the quality effects within the datasets. Outcome data were generated for individual trial participants for intervention ($x = 1$) and control ($x = 0$) groups for each trial, with 50 percent of individuals assigned to each group.

We modeled heterogeneity by adding a variance parameter (ν) to the simulations—each meta-analysis was comprised of studies sampled from a population with a specified effect size (d), effect sizes were generated at the study level, and then (for heterogeneity) a random parameter was added to the effect size, to introduce population level heterogeneity. We used a value for ν such that the τ^2 from the simulation matched the τ^2 for the sample. To model reduced heterogeneity we halved the value of the parameter that was added.

To explore the effects of heterogeneity, we ran simulations with three separate values for the heterogeneity parameter: First, with a heterogeneity parameter that gave a value for τ^2 that matched the dataset; second we halved the value of the variance parameter, and third we used a heterogeneity parameter of zero (i.e. no heterogeneity at the population level). This gave 3 (levels of heterogeneity) × 2 (quality effects) = 6 cells in the simulation for each dataset. For each of these cells in the simulation we generated and analyzed 1000 datasets.

A function was written in R version 2.12 (R Development Core Team, 2010) to generate data, analyze simulations, and aggregate results. The R code is shown in the appendix (See Description of Monte Carlo Simulation). Random effects normally distributed with a mean of 0 were simulated via the rnorm() function. The effect size for each study was calculated using the es() function in the metafor package (version 1.5; Viechtbauer, 2010), and the results were then pooled using the rma() function of the metafor package, using the DerSimonian-Laird random effects estimator.

Results

This section describes the empirical datasets and the observed association between quality and effect sizes, the heterogeneity and effect size distribution represented in the datasets, and the Monte Carlo simulation results. Throughout, the different dataset results are compared and similarities and differences are highlighted.

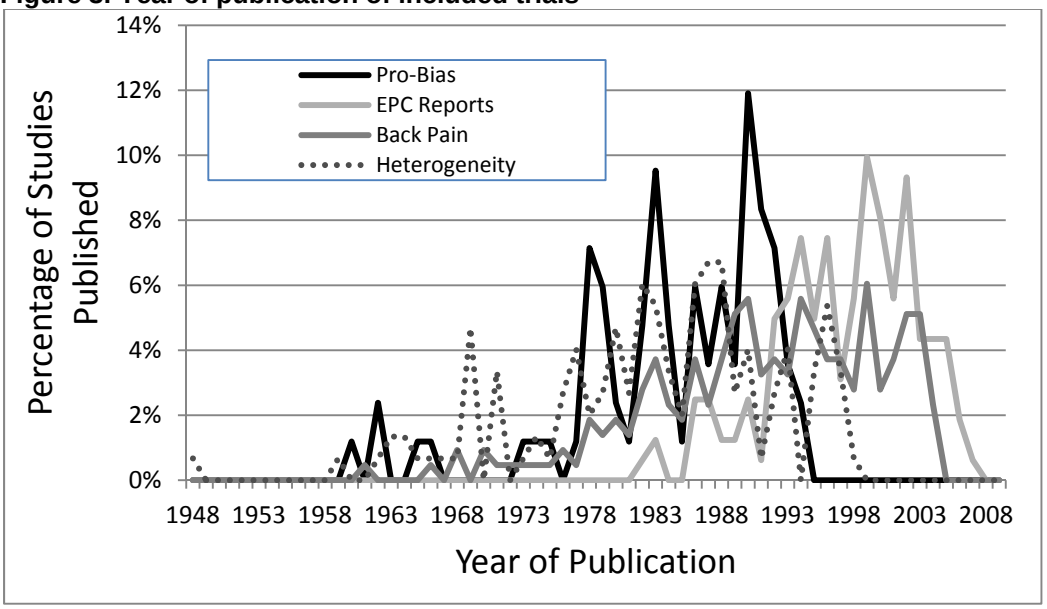
Empirical Datasets

Three of the datasets (Back pain, EPC reports, 'Pro-bias') we used for our analyses have been described in detail elsewhere (Hempel et al., 2011). As outlined, we added an additional meta-epidemiological dataset to explore the associations between quality and effect sizes for the purpose of this study. Dataset 4 is henceforth called 'Heterogeneity set,' to indicate that this dataset comprised meta-analyses selected to show heterogeneity between studies.

Dataset Description

Figure 3 shows the years of publication of the included papers for all four datasets.

Figure 3. Year of publication of included trials



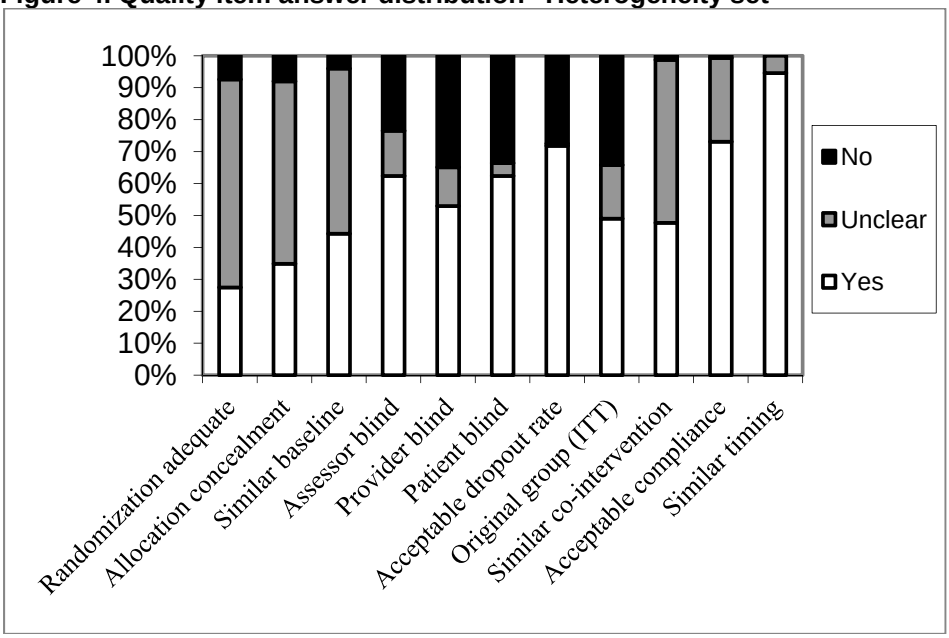
EPC = Evidence Based Practice Center

The studies included in the new dataset 'Heterogeneity set' were older than those included in the Back pain, EPC reports and 'Pro-bias' datasets.

Quality of the Reporting

The figure below (Figure 4) shows the distribution of answers to the quality items (yes, unclear, no) for this new empirical dataset, "Heterogeneity set." A "yes" is an indicator of high quality for each of the items (randomization sequence, allocation concealment, baseline similarity, outcome assessor blinding, care provider blinding, patient blinding, dropout rate and description, analysis in original group (ITT), co-interventions, compliance, and assessment timing); for example, that the outcome assessors were blinded.

Figure 4. Quality item answer distribution "Heterogeneity set"

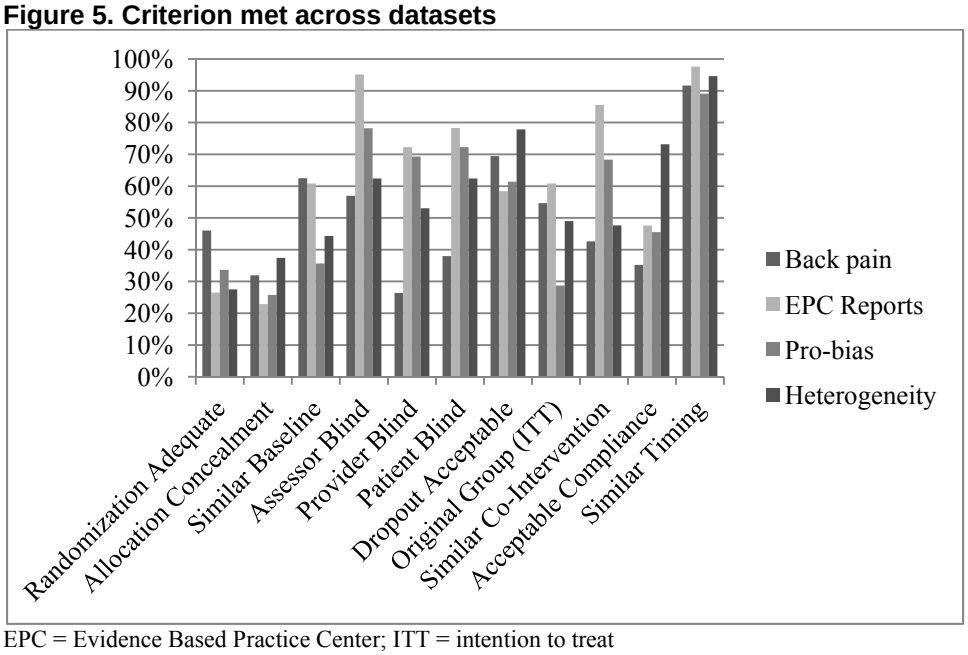


ITT = intention to treat

For many quality criteria, information was insufficient to judge the individual quality feature. Although the large majority of trials were described as randomized, many publications did not report on the generation of the randomization sequence and whether a truly random sequence was adhered to.

The figure below (Figure 5) allows a comparison of "yes" answers across the four datasets, i.e. an indication that the feature was reported in the publication and the criterion was met.

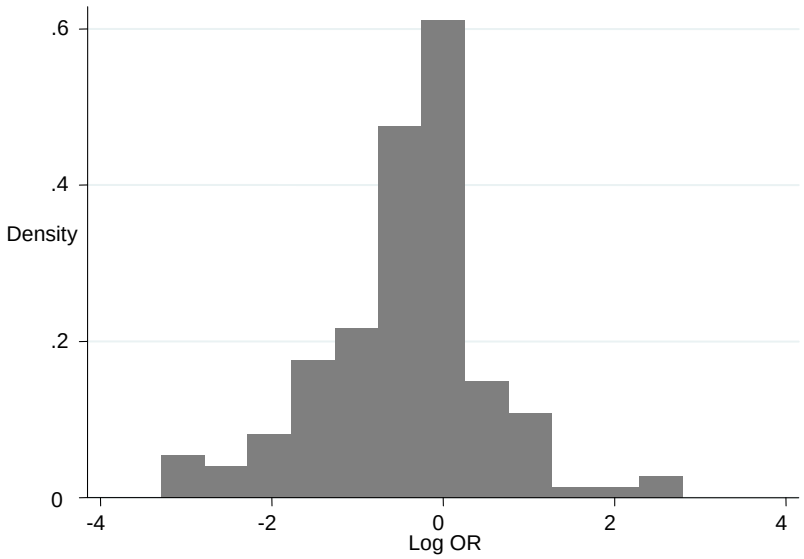
Figure 5. Criterion met across datasets



EPC = Evidence Based Practice Center; ITT = intention to treat

The distribution of treatment effects for all 149 trials included in the 'Heterogeneity' dataset is shown in figure below (Figure 6) (the log odds ratios are displayed).

Figure 6. Treatment effect distribution of "Heterogeneity set"



OR = odds ratio

Observed Association Between Quality and Effect Sizes

As outlined in the methods section, we investigated whether there was an association between the quality of the trial and the reported treatment effect of the trial. The following tables show the odds ratios in trials meeting a quality criterion, the odds ratio of trials not meeting the criterion, and the ratio of odds ratios between these trial groups for the new dataset specifically compiled for this report (Dataset 4, 'Heterogeneity set'). This set is based on a previously

published datasets with known qualities as outlined in the methods section. In addition, the tables report the number of trials meeting each criterion and the number of trials not meeting the criterion, to demonstrate how common each quality aspect was in the sample and to show the sample size of trials in each group of trials (that is the number of low- and high-quality trials applying each respective quality criterion).

Table 2. Ratio of odds ratios between studies fulfilling criteria: "Heterogeneity set," CBRG criteria

CBRG Criteria	# Criterion met	# Criterion not met	OR (met)	95% CI (met)	OR (not met)	95% CI (not met)	ROR	95% CI
Randomization adequate	41	108	0.65	(0.53, 0.79)	0.76	(0.68, 0.86)	1.18	(0.94, 1.49)
Allocation concealment	52	97	0.79	(0.67, 0.92)	0.70	(0.61, 0.80)	0.89	(0.73, 1.09)
Similar baseline	66	83	0.75	(0.65, 0.86)	0.71	(0.62, 0.82)	0.95	(0.78, 1.16)
Assessor blind	93	56	0.73	(0.65, 0.83)	0.73	(0.62, 0.86)	0.99	(0.80, 1.22)
Care provider blind	79	70	0.75	(0.66, 0.86)	0.71	(0.61, 0.82)	0.94	(0.77, 1.15)
Patient blind	93	56	0.77	(0.68, 0.87)	0.67	(0.57, 0.80)	0.88	(0.71, 1.08)
Acceptable dropout rate	107	42	0.73	(0.65, 0.82)	0.74	(0.61, 0.90)	1.02	(0.81, 1.27)
Original group (ITT)	73	76	0.76	(0.66, 0.87)	0.71	(0.61, 0.82)	0.93	(0.76, 1.14)
Similar co-interventions	71	78	0.80	(0.69, 0.92)	0.68	(0.59, 0.78)	0.85	(0.69, 1.03)
Acceptable compliance	109	40	0.71	(0.63, 0.80)	0.80	(0.65, 0.98)	1.12	(0.89, 1.41)
Similar timing	141	8	0.74	(0.67, 0.82)	0.64	(0.43, 0.97)	0.87	(0.57, 1.32)
Summary Score								
≥9 vs <9	25	124	0.77	(0.61, 0.97)	0.73	(0.65, 0.81)	0.94	(0.73, 1.22)
≥8 vs <8	44	105	0.80	(0.67, 0.96)	0.71	(0.63, 0.79)	0.88	(0.71, 1.09)
≥7 vs <7	72	77	0.78	(0.68, 0.89)	0.69	(0.60, 0.79)	0.88	(0.72, 1.08)
≥6 vs <6	91	58	0.77	(0.69, 0.87)	0.66	(0.56, 0.79)	0.86	(0.70, 1.06)
≥5 vs <5	114	35	0.73	(0.66, 0.82)	0.73	(0.58, 0.92)	1.00	(0.77, 1.28)
≥4 vs <4	132	17	0.74	(0.67, 0.82)	0.66	(0.46, 0.94)	0.89	(0.61, 1.29)

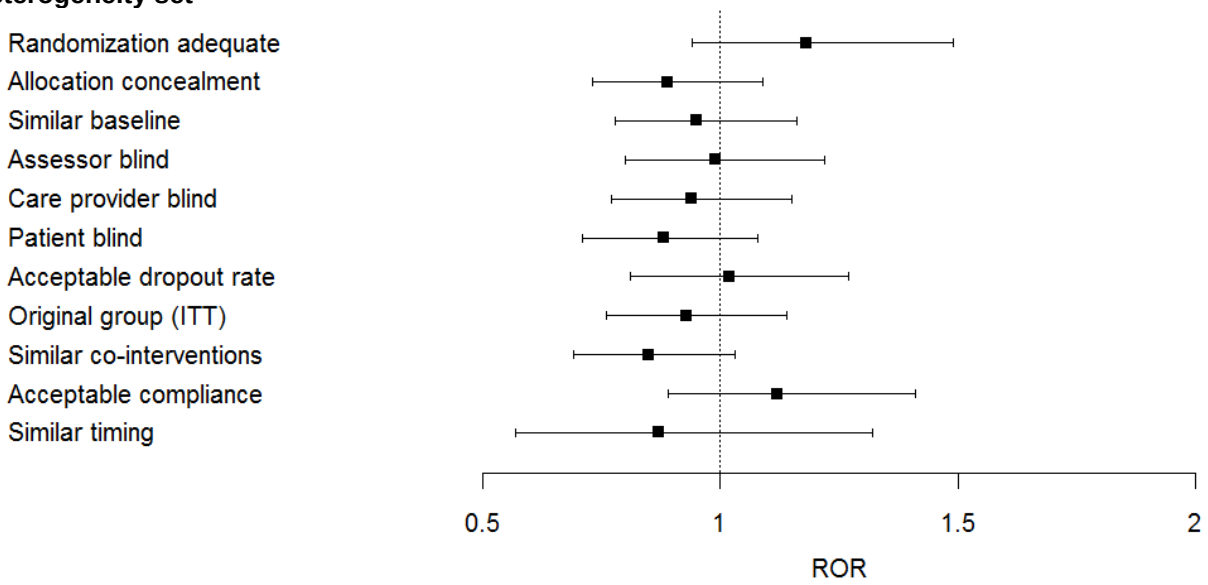
CBRG = Cochrane Back Review Group; CI = confidence interval, OR = odds ratio, ROR = ratio of odds ratios

The associations between the CBRG quality criteria and effect sizes were small and none achieved statistical significance in this dataset. For example, unconcealed compared to concealed trials showed an ROR of 0.89, 95% CI: 0.73, 1.09. However, the majority of RORs indicated that trials without meeting a quality criterion reported slightly larger treatment effects across investigated criteria (i.e. concealment of treatment allocation, similarity of baseline values, blinding of care providers, blinding of patients, intention-to-treat analysis, similarity of co-interventions, and similar timing of outcome assessment). In this dataset, adequate randomization and acceptable compliance were associated with larger effects, while acceptable dropout rates and assessor blinding did not show any or only extremely small differences in effect sizes.

Based on a quantitative summary score derived from all quality criteria, a cut-off of six or more criteria met differentiated low- and high-quality trials best with an ROR of 0.86 (95% CI: 0.70, 1.06).

The figure below (Figure 7) graphically represents the association between quality and effect sizes for the individual quality criteria in the "Heterogeneity" dataset.

Figure 7. Ratio of odds ratios between studies fulfilling CBRG criteria versus not: "Heterogeneity set"



CBRG = Cochrane Back Review Group; ROR = ratio of odds ratios

The figure displays the ratio of odds ratios of reported effect sizes for low- and high-quality trials. In addition, the 95 percent confidence interval for the effect is shown.

For trials included in this fourth dataset we have also applied the Jadad scale, criteria suggested by Schulz et al. (1995), and the Cochrane risk of bias tool. Although all represented quality domains are also included in the 11-item CBRG quality criteria list, individual operationalizations vary slightly across measures. The table below (Table 3) shows the results for the individual Jadad criteria, the total Jadad score, the Schulz criteria, and the individual Risk of Bias tool criteria as well as the overall Cochrane risk of bias assessment.

Table 3. Ratio of odds ratios between studies fulfilling criteria; "Heterogeneity set," other criteria

Quality Criteria	# Criterion met	# Criterion not met	OR (met)	95% CI (met)	OR (not met)	95% CI (not met)	ROR (met)	95% CI (not met)
Jadad								
Randomization=2	34	115	0.61	(0.49, 0.76)	0.77	(0.69, 0.86)	1.27	(0.99, 1.62)
Blinding=2	41	108	0.76	(0.63, 0.91)	0.72	(0.64, 0.81)	0.95	(0.76, 1.18)
Withdrawal=1	119	30	0.75	(0.67, 0.84)	0.68	(0.54, 0.85)	0.91	(0.71, 1.17)
Jadad: total ≥3	83	66	0.70	(0.61, 0.80)	0.78	(0.67, 0.91)	1.11	(0.90, 1.36)
Schulz								
Concealment	52	97	0.79	(0.68, 0.92)	0.70	(0.61, 0.79)	0.89	(0.72, 1.08)
Sequence	34	115	0.61	(0.49, 0.76)	0.77	(0.69, 0.86)	1.27	(0.99, 1.62)
Analysis	77	72	0.72	(0.63, 0.83)	0.74	(0.64, 0.86)	1.03	(0.84, 1.26)
Blinding	78	71	0.75	(0.66, 0.87)	0.71	(0.61, 0.82)	0.94	(0.77, 1.15)

Table 3. Ratio of odds ratios between studies fulfilling criteria; "Heterogeneity set," other criteria (continued)

Quality Criteria	# Criterion met	# Criterion not met	OR (met)	95% CI (met)	OR (not met)	95% CI (not met)	ROR (met)	95% CI (not met)
Cochrane Risk of Bias								
Sequence generation	35	114	0.60	(0.48, 0.75)	0.77	(0.69, 0.86)	1.28	(1.00, 1.64)
Allocation concealment	51	98	0.79	(0.68, 0.93)	0.69	(0.61, 0.79)	0.87	(0.72, 1.07)
Blinding	79	70	0.75	(0.66, 0.86)	0.71	(0.61, 0.82)	0.94	(0.77, 1.15)
Incompl. outcome data	74	75	0.75	(0.65, 0.86)	0.72	(0.62, 0.83)	0.96	(0.78, 1.17)
Sel. outcome reporting	123	26	0.75	(0.67, 0.83)	0.65	(0.50, 0.84)	0.86	(0.65, 1.14)
Other sources of bias	38	111	0.74	(0.61, 0.89)	0.73	(0.65, 0.82)	0.99	(0.80, 1.24)
RoB overall risks	43	106	0.76	(0.62, 0.92)	0.72	(0.64, 0.81)	0.96	(0.76, 1.20)

CI = confidence interval; OR = odds ratio; RoB = risk of bias; ROR = ratio of odds ratios
Note: The Jadad randomization and blinding score ranges from 0 to 2, the withdrawal score from 0 to 1, the total score from 0 to 5.

Applying the Jadad criteria, higher quality studies reported slightly larger treatment effects, primarily triggered by the randomization item ("Was the study described as randomized and was the method to generate the sequence appropriate?"). Double-blinding and a description of withdrawals showed effects corresponding with the CBRG quality criteria (i.e., criteria met are associated with smaller reported effect sizes). The Schulz criteria showed similar results to the Jadad criteria, and allocation concealment showed the largest difference between low- and high-quality studies (0.89; 95% CI: 0.72, 1.08). The Cochrane risk of bias tool also showed small effects of quality, with lower quality studies reporting larger effect sizes; however the sequence generation item again showed the opposite effect. None of the effects are statistically significant.

Sensitivity Analyses: Stratification by Meta-analysis, Stratification by Clinical Area, and Correction for Clustering

In order to investigate whether the associations between quality and effect sizes are consistent or notably different across individual meta-analyses and clinical fields, we stratified the "Heterogeneity" dataset accordingly. As outlined previously, the dataset includes trials from 13 meta-analyses from two very different fields: pediatric interventions and cardiovascular disease interventions.

The table below (Table 4) shows the effect size difference expressed as the ratio of odds ratio for low (criterion not met) and high (criterion met) quality studies for each meta-analysis.

Table 4. Difference in odds ratios for studies fulfilling CBRG criteria by individual meta-analyses

CBRG Criteria	ROR (a)	ROR (b)	ROR (c)	ROR (d)	ROR (e)	ROR (f)	ROR (g)	ROR (h)	ROR (i)	ROR (j)	ROR (k)	ROR (l)	ROR (m)
Randomization adequate	NC	0.35	1.99	0.83	2.09	1.13	NC	0.96	NC	0.88	1.39	NC	0.47
Allocation concealment	NC	0.97	NC	1.69	NC	1.11	0.55	0.59	1.01	0.88	0.70	0.89	0.98
Similar baseline	NC	0.97	1.33	0.85	0.55	0.82	1.13	0.83	1.07	1.00	0.49	NC	1.15
Assessor blind	NC	1.00	NC	1.06	0.42	0.98	0.80	0.89	NC	1.07	0.65	0.74	NC

Table 4. Difference in odds ratios for studies fulfilling CBRG criteria by individual meta-analyses (continued)

CBRG Criteria	ROR (a)	ROR (b)	ROR (c)	ROR (d)	ROR (e)	ROR (f)	ROR (g)	ROR (h)	ROR (i)	ROR (j)	ROR (k)	ROR (l)	ROR (m)
Care provider blind	NC	NC	NC	0.66	0.58	0.98	0.80	0.89	NC	1.60	0.67	0.76	NC
Patient blind	NC	NC	NC	0.66	0.58	NC	NC	0.71	NC	1.05	0.48	0.76	NC
Acceptable dropout rate	NC	NC	0.69	0.69	1.34	1.26	1.21	1.07	1.07	1.13	1.36	NC	NC
Original group (ITT)	NC	3.04	2.97	1.38	2.84	0.70	0.69	NC	0.80	1.11	0.81	1.41	1.47
Similar co-interventions	NC	0.31	NC	0.78	0.72	NC	1.29	0.89	0.85	0.93	NC	NC	1.74
Acceptable compliance	NC	NC	NC	1.25	2.02	NC	0.73	0.87	0.91	NC	NC	NC	NC
Similar timing	NC	NC	NC	0.64	NC	0.91	NC	NC	NC	0.83	NC	NC	NC
Summary Score													
≥9 vs <9	NC	0.73	NC	NC	NC	1.12	0.80	NC	1.08	NC	NC	NC	1.98
≥8 vs <8	NC	0.97	3.11	0.59	NC	0.86	0.80	NC	1.08	1.04	NC	NC	1.23
≥7 vs <7	NC	NC	0.69	0.53	NC	0.98	0.80	0.70	0.71	1.09	0.56	0.89	NC
≥6 vs <6	NC	NC	0.69	0.60	NC	NC	0.80	0.89	0.72	1.10	0.27	1.07	NC
≥5 vs <5	NC	NC	NC	0.97	0.97	NC	NC	0.86	0.72	0.99	NC	NC	NC
≥4 vs <4	NC	NC	NC	1.94	0.77	NC	NC	NC	NC	0.75	NC	NC	NC

ATC = Antiplatelet Trialists' Collaboration; CBRG = Cochrane Back Review Group; ITT = intention to treat; NC = could not be computed (less than 3 trials in the high or low quality trial group); ROR = ratio of odds ratios

Note: List of meta-analyses (see reference list): (a) Ausejo;(b) Bhuta; (c) Kellner; (d) Kozyrskyj; (e) Rosenfeld; (f) ATC; (g) Hine; (h) Leizorovicz; (i)Yusuf, beta-blockade; (j)Yusuf, streptokinase; (k) Yusuf, nitrates; (l) Rossouw; (m) Teo.

Note: An ROR less than 1 indicates that high-quality trials reported a smaller treatment effect compared to those trials that met the quality criteria.

The number of included trials varied across meta-analyses, and for several individual meta-analyses the ratio of odds ratio of high and low quality studies could not be calculated because there were fewer than three trials present in the group of high-quality trials and/or low-quality trials. In meta-analyses that included a sufficient number of trials, effects of quality varied, in terms of both size and direction of effects. Reported concealment of treatment allocation was associated with smaller treatment effects in seven individual meta-analyses. (The ratio of odds ratios ranged from 0.55 to 0.98.) In three meta-analyses the opposite effect was found, and in three meta-analyses the effect could not be studied due to the small number of included trials.

As mentioned, the individual meta-analyses represented two very different clinical fields. Five meta-analyses investigated interventions in pediatric samples. The remaining were cardiovascular disease meta-analyses. The table below (Table 5) shows the effect size difference for high (criterion fulfilled) and low (criterion not fulfilled) quality studies for each of the two general clinical fields.

Table 5. Difference in odds ratios for studies fulfilling CBRG criteria by clinical field

CBRG Criteria	ROR Pediatrics N=56	95% CI Pediatrics	ROR Cardiovascular Disease N=93	95% CI Cardiovascular Disease
Randomization adequate	1.86	(0.93, 3.70)	1.02	(0.85, 1.23)
Allocation concealment	1.52	(0.75, 3.07)	0.85*	(0.73, 0.99)*
Similar baseline	1.15	(0.61, 2.19)	0.93	(0.80, 1.09)

Table 5. Difference in odds ratios for studies fulfilling CBRG criteria by clinical field (continued)

CBRG Criteria	ROR Pediatrics N=56	95% CI Pediatrics	ROR Cardiovascular Disease N=93	95% CI Cardiovascular Disease
Assessor blind	1.17	(0.58, 2.37)	0.91	(0.77, 1.07)
Care provider blind	0.98	(0.51, 1.86)	0.90	(0.78, 1.05)
Patient blind	0.98	(0.51, 1.86)	0.87	(0.74, 1.03)
Acceptable dropout rate	0.83	(0.40, 1.72)	1.12	(0.95, 1.33)
Original group (ITT)	1.45	(0.75, 2.79)	0.91	(0.78, 1.07)
Similar co-interventions	0.51*	(0.28, 0.94)*	0.96	(0.82, 1.11)
Acceptable compliance	1.87	(0.98, 3.57)	0.99	(0.81, 1.20)
Similar timing	0.78	(0.21, 2.94)	0.81	(0.56, 1.16)
Summary Score				
≥9 vs <9	1.15	(0.52, 2.59)	0.93	(0.76, 1.13)
≥8 vs <8	1.23	(0.61, 2.49)	0.84*	(0.72, 0.98)*
≥7 vs <7	1.09	(0.57, 2.10)	0.91	(0.78, 1.06)
≥6 vs <6	0.90	(0.47, 1.70)	0.95	(0.79, 1.13)
≥5 vs <5	1.46	(0.72, 2.93)	0.88	(0.71, 1.08)
≥4 vs <4	1.59	(0.62, 4.08)	0.71*	(0.53, 0.94)*

CBRG = Cochrane Back Review Group; CI = confidence interval; ITT = intention to treat; ROR = ratio of odds ratios
* $p<0.05$
Note: An ROR less than 1 indicates that high-quality trials reported a smaller treatment effect compared to those trials that met the quality criteria.

Across clinical fields, the effects associated with individual quality criteria varied. Meeting quality criteria was sometimes associated with a smaller treatment effect and sometimes with a smaller reported effect. Based on a quantitative summary score, a cut-off of 6 quality criteria met was most consistent across clinical fields.

As a further sensitivity analysis we estimated meta-regressions correcting for clustering with meta-analysis using a Huber/White (sandwich) estimator. Table 6 presents corrected and uncorrected point estimates and confidence intervals.

Table 6. Ratio of odds ratios for CBRG criteria corrected and uncorrected for clustering

CBRG Criteria	ROR	95% CI	ROR (Corrected)	95% CI (Corrected)
Randomization adequate	1.18	(0.94, 1.49)	1.18	(0.83, 1.67)
Allocation concealment	0.89	(0.73, 1.09)	0.89	(0.66, 1.18)
Similar baseline	0.95	(0.78, 1.16)	0.95	(0.78, 1.17)
Assessor blind	0.99	(0.80, 1.22)	0.98	(0.79, 1.22)
Care provider blind	0.94	(0.77, 1.15)	0.94	(0.78, 1.13)
Patient blind	0.88	(0.71, 1.08)	0.88	(0.69, 1.12)
Acceptable dropout rate	1.02	(0.81, 1.27)	1.13	(0.84, 1.51)
Original group (ITT)	0.93	(0.76, 1.14)	0.92	(0.68, 1.26)
Similar co-interventions	0.85	(0.69, 1.03)	0.86	(0.73, 0.99)*
Acceptable compliance	1.12	(0.89, 1.41)	1.10	(0.83, 1.45)
Similar timing	0.87	(0.57, 1.32)	0.85	(0.67, 1.08)

CBRG = Cochrane Back Review Group; CI = confidence interval; ITT = intention to treat; ROR: ratio of odds ratios, corrected for clustering using a Huber/White (sandwich) estimator
* $p<0.05$

We found that this alternative estimation procedure made only small changes to estimates and standard errors and corresponding confidence intervals.

Comparison Across Meta-epidemiological Datasets

The following table (Table 7) compares the association between quality and effect sizes across the four epidemiological datasets that we compiled in the course of the project. The individual CBRG criteria are shown; in addition, the results for the quantitative summary score exploring different cut-offs are reported. Note the dataset 1 and 2 are displayed as effect size differences, for dataset 3 and 4 the ratio of odds ratio is shown.

Table 7. CBRG criteria across datasets

CBRG Criteria	Dataset 1 Back Pain		Dataset 2 EPC Reports		Dataset 3 Pro-bias		Dataset 4 Heterogeneity set	
	ESdiff	95% CI	ESdiff	95% CI	ROR	95% CI	ROR	95% CI
Randomization adequate	0.02	(-0.12, 0.16)	0.01	(-0.15, 0.17)	0.94	(0.75, 1.17)	1.18	(0.94, 1.49)
Allocation concealment	-0.08	(-0.23, 0.07)	-0.05	(-0.22, 0.11)	0.91	(0.72, 1.14)	0.89	(0.73, 1.09)
Similar baseline	-0.10	(-0.24, 0.05)	-0.09	(-0.24, 0.05)	0.98	(0.80, 1.21)	0.95	(0.78, 1.16)
Assessor blind	-0.10	(-0.25, 0.04)	0.06	(-0.28, 0.41)	1.35	(1.05, 1.73)	0.99	(0.80, 1.22)
Care provider blind	-0.10	(-0.26, 0.06)	0.19	(0.03, 0.35)*	0.83	(0.67, 1.02)	0.94	(0.77, 1.15)
Patient blind	-0.03	(-0.18, 0.11)	0.21	(0.04, 0.39)*	0.97	(0.78, 1.21)	0.88	(0.71, 1.08)
Acceptable dropout	-0.13	(-0.29, 0.02)	0.15	(0.01, 0.29)*	0.72	(0.59, 0.88)	1.02	(0.81, 1.27)
Original group (ITT)	-0.10	(-0.24, 0.04)	0.05	(-0.10, 0.20)	0.91	(0.74, 1.12)	0.93	(0.76, 1.14)
Similar co-interventions	-0.09	(-0.23, 0.05)	0.05	(-0.15, 0.28)	1.50	(1.22, 1.85)	0.85	(0.69, 1.03)
Acceptable compliance	-0.01	(-0.15, 0.14)	0.02	(-0.12, 0.17)	0.72	(0.59, 0.88)	1.12	(0.89, 1.41)
Similar timing	-0.17	(-0.43, 0.10)	0.25	(-0.19, 0.69)	1.33	(0.94, 1.88)	0.87	(0.57, 1.32)
Summary Score								
≥9 vs <9	-0.04	(-0.29, 0.20)	0.15	(-0.01, 0.30)	1.09	(0.81, 1.46)	0.94	(0.73, 1.22)
≥8 vs <8	-0.09	(-0.26, 0.08)	0.03	(-0.11, 0.18)	0.85	(0.68, 1.07)	0.88	(0.71, 1.09)
≥7 vs <7	-0.10	(-0.24, 0.05)	0.01	(-0.14, 0.16)	1.05	(0.86, 1.29)	0.88	(0.72, 1.08)
≥6 vs <6	-0.20	(-0.34, -0.06)*	0.16	(-0.03, 0.35)	0.77	(0.63, 0.95)*	0.86	(0.70, 1.06)
≥5 vs <5	-0.20	(-0.35, -0.05)*	0.27	(0.02, 0.52)*	0.79	(0.63, 0.99)*	1.00	(0.77, 1.28)
≥4 vs <4	-0.13	(-0.31, 0.06)	0.41	(-0.02, 0.84)	0.83	(0.65, 1.07)	0.89	(0.61, 1.29)

CBRG = Cochrane Back Review Group; CI = confidence interval, ESdiff = effect size difference; ITT = intention to treat; ROR = ratio of odds ratios

* $p<0.05$

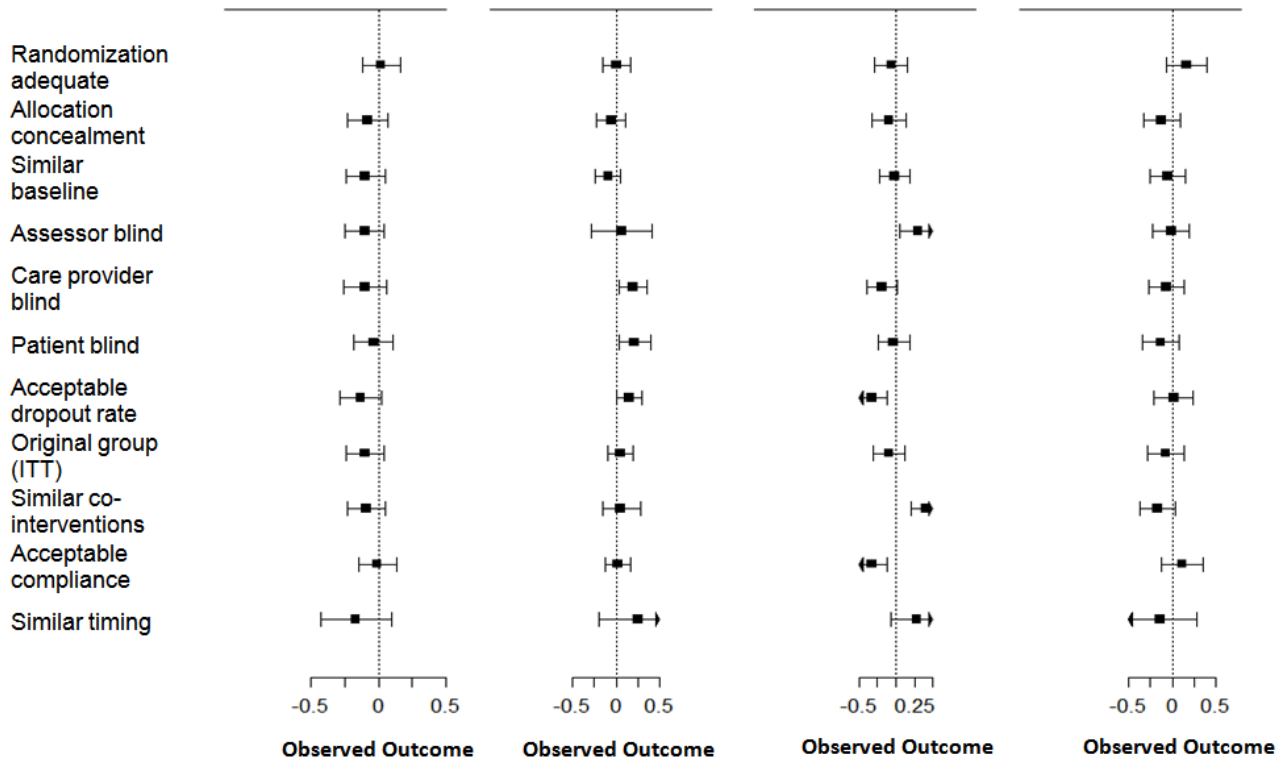
Note: "Pro-bias" data are based on fixed-effects model.

Observed associations between quality and effect sizes varied for individual quality criteria across datasets primarily due to results in dataset 2 (EPC reports). Most consistent results were shown for allocation concealment: across datasets, concealed trials reported smaller effect sizes compared to unconcealed trials.

Applying a quantitative summary score, a cut-off of 6 met criteria differentiated studies best across three out of four datasets (i.e., the difference between low and high quality studies was most distinct); however, this effect could not be found for dataset 3 (the EPC reports). In only one of the datasets (Dataset 1 Back pain) was the difference statistically significantly different.

The figure (Figure 8) allows a graphic overview for the CBRG criteria across the four epidemiological datasets that we compiled in the course of the project.

Figure 8. Associations between CBRG criteria and reported treatment effects across datasets



CBRG = Cochrane Back Review Group; ITT = intention to treat
Note: From left to right: dataset 1 (Back pain trials), dataset 2 (EPC reports), dataset 3 ('Pro-bias', fixed-effects model), dataset 4 ("Heterogeneity set")

Results are based on effect size differences in the first two datasets and on the ratio of odds ratios in dataset 3 and 4. In addition, the 95 percent confidence interval for the effect is shown. The direction of effects is displayed consistently across datasets—point estimates on the left indicate that low-quality studies reported larger treatment effects.

Heterogeneity and Effect Size Distributions

The different datasets comprise individual meta-analyses, each of which contributes trials to the total dataset. The following table (Table 8) shows the amount of between-study variance in effect sizes expressed as I^2. In addition, the I^2 of the entire dataset is shown.

Table 8. Heterogeneity

Interventions	# Studies	Pooled RE ES	I^2
Dataset 1: Back Pain			
Acupuncture	17	0.45	71.5%*
Back School	14	0.40	68.6%*
Behavior Therapy	18	0.62	28.9%
Exercise Therapy	45	0.58	68.1%*
Manipulation	23	0.50	62.1%*
Muscle Relax	23	0.53	85.3%*
NSAID	33	0.39	70.7%*
Other	43	0.51	8.9%
Overall	216	0.50	72.4%*
Dataset 2: EPC Reports			
Alzheimer's	14	0.50	26.2%
Arthritis	3	1.45	99.4%*
CDSM	19	0.34	49.7%*
Chromium	6	0.41	84.6%*
Epilepsy	25	0.52	52.7%*
Glucosamine	20	0.37	71.7%*
OCD	9	0.95	98.2%*
Omega 3	17	0.04	57.0%*
Orlistat	17	0.47	65.2%*
S-AMe	11	0.62	69.1%*
SMBG	10	0.14	33.5%
Vitamin E	14	0.10	51.9%*
Overall	165	0.43	97.5%*
Dataset 3: Pro-bias			
Dolan	11	0.49	74.1%*
Hughes	3	0.90	0.0%
Lensing	5	0.19	0.0%
Loonen	9	0.25	0.0%
Loosemore	6	0.37	49.7%
Mari	5	0.19	0.0%
Marshall	5	0.11	48.9%
Pace	22	0.35	48.5%*
Ramirez	10	0.88	0.0%
Sandercock	9	1.67	3.9%
Sutherland	15	0.83	52%*
Overall	100	0.46	59.6%*
Dataset 4: Heterogeneity set			
Ausejo	5	0.59	88.3%*
Bhuta	6	0.32	72.7%*
Kellner	8	0.26	61.8%*
Kozyrskyj	27	1.08	31.0%
Rosenfield	10	0.32	77.9%*
ATC	10	0.89	16.0%
Hine	10	1.28	32.4%

Table 8. Heterogeneity (continued)

Interventions	# Studies	Pooled RE ES	I^2
Leizorovicz	11	0.78	39.1%
Yusuf – Beta-blocker	13	0.8	41.4%
Yusuf - Streptokinase	28	0.76	0.0%
Yusuf - Nitrates	8	0.67	41.7%
Rossouw	7	0.82	24.1%
Teo	6	0.31	31.0%
Overall	149	0.75	60.0%*

ATC = Antiplatelet Trialists' Collaboration; CDSM = chronic disease self-management; I^2 = amount of between-study variance in effect sizes; OCD = obsessive-compulsive disorder; NSAID = non-steroidal anti-inflammatory drugs; RE ES = effect size (random effects meta-analysis); RE OR = odds ratio (random effects meta-analysis);
S-AMe: = S-adenosylmethionine; SMBG = self-monitoring of blood glucose
* $p<0.05$, chi-square test.

The amount of heterogeneity in dataset 1 estimated by I^2 ranged from 8.9 to 85.3 percent in individual trials. The overall estimated heterogeneity for the entire dataset was 72.4 percent.

In dataset 2, the EPC reports, individual I^2 estimates were generally higher and ranged from 26.2 to 99.4 percent and the overall dataset estimate was 97.5 percent.

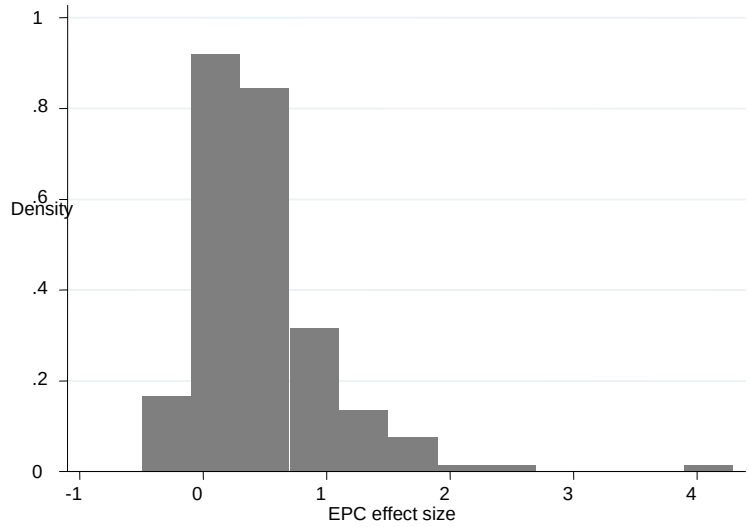
Several meta-analyses in dataset 3 (Pro-bias) showed no evidence of heterogeneity; the overall dataset heterogeneity estimate was 59.6 percent.

Dataset 4 showed some heterogeneity, but some estimates were not statistically significant and the overall database estimate was 60 percent. The overall estimate was not higher but comparable to dataset 1 and 3, although heterogeneity across studies was one of the inclusion criteria for the selection of meta-analyses that were compiled for this dataset.

Dataset Distributions

The effect size distributions varied across meta-epidemiological datasets that we assembled in the course of the project. In particular, the effect size distribution in one of the datasets (Dataset 2, EPC reports) was less symmetric (bell-shaped) than those of the other datasets. The figure below (Figure 9) shows the distribution of the effect sizes reported for each included study in this dataset.

Figure 9. Dataset 2 distribution



EPC = Evidence Based Practice Center

As previously reported, the mean treatment effect across all studies in this dataset was 0.43 (95% CI: 0.34, 0.53). Few quality features were associated with differences in effect sizes according to whether these criteria were met. The table (Table 9) shows the effect sizes for studies meeting the particular criterion and studies not meeting the criterion in dataset 2 (EPC reports) as previously published. The last column shows the difference between studies that met and did not meet criteria. A negative difference indicates that the effect size for the studies fulfilling the criterion was smaller than the effect size for the studies not meeting the criterion.

Table 9. Dataset 2 results associations between quality and effect sizes

CBRG Criteria	ES Criterion met	95% CI Criterion met	ES Criterion not met	95% CI Criterion not met	ESdiff	95% CI
Randomization adequate	0.44	(0.30, 0.57)	0.43	(0.34, 0.51)	0.01	(-0.15, 0.17)
Allocation concealment	0.39	(0.25, 0.53)	0.45	(0.36, 0.53)	-0.05	(-0.22, 0.11)
Similar baseline	0.40	(0.31, 0.49)	0.49	(0.37, 0.61)	-0.09	(-0.24, 0.05)
Assessor blind	0.43	(0.36, 0.51)	0.37	(0.04, 0.71)	0.06	(-0.28, 0.41)
Care provider blind	0.48	(0.40, 0.56)	0.29	(0.15, 0.43)	0.19	(0.03, 0.35)*
Patient blind	0.47	(0.40, 0.55)	0.26	(0.11, 0.42)	0.21	(0.04, 0.39)*
Acceptable dropout rate	0.50	(0.40, 0.59)	0.35	(0.24, 0.45)	0.15	(0.01, 0.29)*
Original group (ITT)	0.45	(0.36, 0.54)	0.40	(0.27, 0.52)	0.05	(-0.10, 0.20)
Similar co-interventions	0.44	(0.36, 0.52)	0.39	(0.20, 0.58)	0.05	(-0.15, 0.28)
Acceptable compliance	0.44	(0.34, 0.55)	0.42	(0.32, 0.52)	0.02	(-0.12, 0.17)
Similar timing	0.44	(0.37, 0.51)	0.19	(-0.24, 0.62)	0.25	(-0.19, 0.69)
Summary Score						
≥6 vs <6	0.46	(0.38, 0.53)	0.30	(0.12, 0.47)	0.16	(-0.03, 0.35)
≥5 vs <5	0.45	(0.38, 0.53)	0.18	(-0.06, 0.42)	0.27	(0.02, 0.52)*

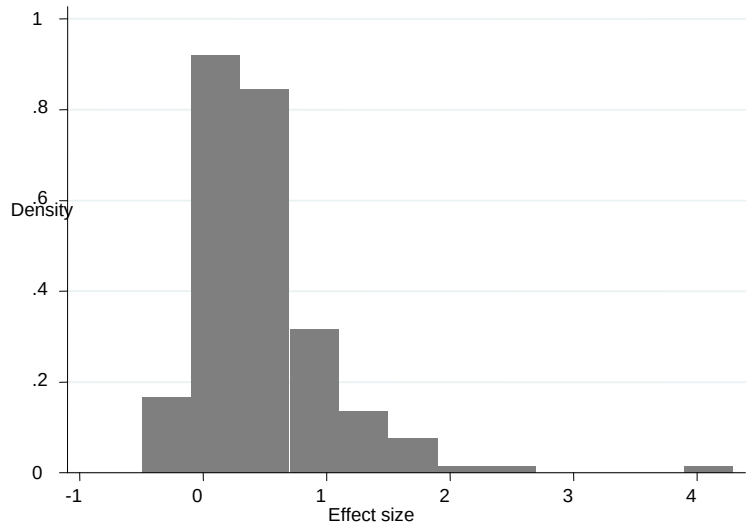
CBRG = Cochrane Back Review Group; CI = confidence interval; ES = effect size; ESdiff = effect size difference; ITT = intention to treat

* $p<0.05$

The results show that for only two criteria (allocation concealment and baseline similarity), lower-quality studies reported larger effect sizes than higher-quality studies. The other criteria showed no effect or the opposite effect, the quantitative summary score that statistically significantly differentiated high- and low-quality studies in other datasets showed no significant effect, and the direction of the effect trend was reversed.

As outlined in the methods section, we wanted to investigate whether the distribution characteristics are correlated with the difference in associations between quality and effect sizes; hence we used a non-parametric sampling approach to mirror the distribution of dataset 2 in other datasets. The following histogram (Figure 10) shows dataset 1 using dataset 2 effect sizes.

Figure 10. Dataset 1 using dataset 2 effect sizes



The histogram shows that the sampling method was successful in creating a similar distribution shape.

In a further step, we investigated how the association between quality and effect sizes would be affected by this process. The table below (Table 10) shows the associations between quality features and effect sizes based on these specifications. To allow a direct comparison, the table displays the new results as well as the original research results as published previously. Differences between high- and low-quality studies (defined as meeting a particular criterion or not) were expressed as effect size differences. A negative difference indicates that the group of studies not meeting the quality criterion reported larger effect sizes than the group of studies that met the individual criterion.

Table 10. Effect size difference comparison dataset 1

CBRG Criteria	New ESdiff	New 95% CI	Original ESdiff	Original 95% CI
Randomization adequate	0.14	(-0.01, 0.28)	0.02	(-0.12, 0.16)
Allocation concealment	-0.02	(-0.17, 0.13)	-0.08	(-0.23, 0.07)
Similar baseline	-0.03	(-0.18, 0.12)	-0.10	(-0.24, 0.05)
Assessor blind	0.09	(-0.05, 0.24)	-0.10	(-0.25, 0.04)
Care provider blind	0.07	(-0.09, 0.22)	-0.10	(-0.26, 0.06)
Patient blind	0.05	(-0.09, 0.20)	-0.03	(-0.18, 0.11)

Table 10. Effect size difference comparison dataset 1 (continued)

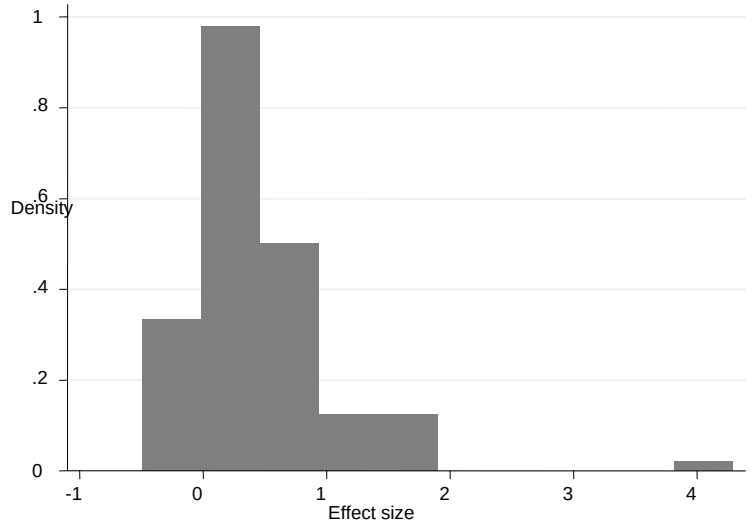
CBRG Criteria	New ESdiff	New 95% CI	Original ESdiff	Original 95% CI
Acceptable dropout rate	0.05	(-0.11, 0.20)	-0.13	(-0.29, 0.02)
Original group (ITT)	-0.08	(-0.23, 0.06)	-0.10	(-0.24, 0.04)
Similar co-interventions	0.00	(-0.15, 0.14)	-0.09	(-0.23, 0.05)
Acceptable compliance	-0.02	(-0.16, 0.13)	-0.01	(-0.15, 0.14)
Similar timing	-0.13	(-0.41, 0.15)	-0.17	(-0.43, 0.10)
Summary Score				
≥6 vs <6	-0.02	(-0.17, 0.12)	-0.20	(-0.34, -0.06)*
≥5 vs <5	0.08	(-0.08, 0.23)	-0.20	(-0.35, -0.05)*

CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat
* $p<0.05$

The table shows that dataset 1, which had originally shown consistent effects of quality by indicating that lower quality studies exaggerated treatment effects, now shows conflicting results. Some criteria were associated with effect sizes, some were not, and the direction of effects varied across effect sizes.

To investigate whether this observation could be replicated in another dataset, we applied the process to dataset 3 (Pro-bias). This dataset was a replication of parts of a larger dataset that had shown clear effects of quality in previous investigations (Moher et al., 1998). The following histogram (Figure 11) shows dataset 3 using dataset 2 effect sizes.

Figure 11. Dataset 3 using dataset 2 effect sizes



The histogram indicates that the sampling process was also successful in replicating a distribution that resembled the original distribution of dataset 2.

The table below (Table 11) shows the associations between quality features and effect sizes based on these specifications and also lists the original effects. The original results are reported as the ratio of odds ratios between high- and low-quality studies, whereas the new results are shown as effect size differences since the sampled dataset was based on continuous outcomes. A negative effect size difference means that the group of trials meeting the criterion reported

smaller treatment effects than the group of trials not meeting the criterion. A ROR less than one also indicates that studies meeting the criterion reported smaller treatment effects than studies not meeting the criterion.

Table 11. Effect size difference comparison dataset 3

CBRG Criteria	New ESdiff	New 95% CI	Original ROR	Original 95% CI
Randomization adequate	-0.13	(-0.32, 0.07)	1.05	(0.69, 1.60)
Allocation concealment	-0.03	(-0.23, 0.18)	0.88	(0.56, 1.39)
Similar baseline	-0.08	(-0.27, 0.11)	1.40	(0.92, 2.12)
Assessor blind	-0.01	(-0.23, 0.21)	1.55	(0.95, 2.51)
Care provider blind	-0.08	(-0.27, 0.11)	1.02	(0.66, 1.57)
Patient blind	-0.14	(-0.34, 0.06)	1.20	(0.77, 1.87)
Acceptable dropout rate	-0.08	(-0.27, 0.10)	0.83	(0.57, 1.22)
Original group (ITT)	-0.06	(-0.27, 0.14)	1.00	(0.65, 1.52)
Similar co-interventions	0.03	(-0.17, 0.23)	1.51	(1.00, 2.27)*
Acceptable compliance	0.01	(-0.17, 0.20)	0.70	(0.47, 1.03)
Similar timing	0.06	(-0.22, 0.34)	1.35	(0.71, 2.58)
Summary Score				
≥6 vs <6	-0.18	(-0.36, 0.01)	0.99	(0.66, 1.49)
≥5 vs <5	-0.11	(-0.32, 0.09)	1.05	(0.66, 1.67)

CBRG = Cochrane Back Review Group; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat; ROR = ratio of odds ratios
* $p<0.05$

The difference between the original data and the new data was less clear in this dataset. This finding may in part be due to inconsistencies across quality criteria that was apparent in the original data—while the majority of criteria trials not meeting criteria overestimated treatment effects, this effect was not shown for all criteria—and in part due to the fact that it is difficult to compare ratios of odds ratios and effect sizes.

Monte Carlo Simulations

We report simulations for three different sets of parameters. The parameters selected closely resemble realistic effects observed in previously obtained empirical datasets.

Dataset 1 (Back Pain Dataset Specifications)

The first analysis used the characteristics observed in dataset 1 (Back pain). This dataset included 216 individual trials with a mean sample size of 80. The observed difference between effect sizes of trials that did and did not attain a quality criterion ranged from 0.02 (adequate randomization) to -0.17 (similar timing), with most criteria indicating that lower-quality studies reported larger treatment effects, thereby exaggerating treatment effects compared to higher-quality studies. The difference, applying the summary score in effect size between low- and high-quality studies was 0.20 (van Tulder et. al., 2009).

We systematically varied the effect of quality and set it to be 0.1 or 0.2 (standardized effect size difference between high-and low-quality trials). We varied the amount of heterogeneity present in the dataset and such an additional variance parameter of 0.14 was used (giving an I^2 that matched the sample of 72.4 percent), a variance parameter set to half of that value, or a

variance parameter of zero. The table (Table 12) shows the effect of these specifications on the power.

Table 12. Power to detect quality moderator effects determined by Monte Carlo simulation under varying effects of quality and heterogeneity, with simulation parameters matching dataset 1 (Back pain trials)

Quality Effect	Additional Variance = 0.14	Additional Variance = 0.07	Additional Variance = 0.00
Median I^{2*}	72.4	55.5	0.00
Median τ^{2**}	0.13	0.07	0.00
Quality effect = 0.1	0.38	0.50	0.85
Quality effect = 0.2	0.91	1.00	1.00

*Observed heterogeneity (I-squared)
**Observed heterogeneity (tau-squared)

In this simulation, the power to detect quality effects as large as 0.2 was found to be high (0.91) in datasets that replicated the heterogeneity of dataset 1 (Back pain): Note that the distribution of quality was set up to ensure the maximum power to detect that effect. For smaller quality effects (0.1), power was considerably lower when heterogeneity was high. As heterogeneity decreased, the power to detect effects increased. With no additional heterogeneity (i.e. assumptions of a fixed-effects model being true), the power to detect a quality effect of 0.1 was acceptably high (0.85).

Dataset 2 (EPC Reports Dataset Specifications)

A further dataset, the EPC reports, was the basis of another set of simulations. This dataset contained 165 individual trials with a mean sample size of 286. The level of heterogeneity found in this study was very high (97.5 percent). The treatment effect was approximately 0.5. In the empirical dataset, study quality showed no consistent effect and several unexpected results in that lower study quality was associated with smaller reported treatment effects compared to higher-quality studies. The effect size differences ranged from -0.09 to 0.25.

The table (Table 13) shows the effects of three different levels of quality on power, assuming three different levels of heterogeneity. Again, the effect of quality was set at 0.1 or 0.2. An additional variance parameter of 0.70 was required to match the sample I^2 of 97.5, and 0.35 was used for half variance.

Table 13. Power to detect quality moderator effects determined by Monte Carlo simulation under varying effects of quality and heterogeneity, with simulation parameters matching dataset 2 (EPC reports)

Quality Effect	Additional Variance = 0.70	Additional Variance = 0.35	Additional Variance = 0.00
Median I^{2*}	97.5	95.5	0.00
Median τ^{2**}	0.61	0.32	0.00
Quality effect = 0.1	0.12	0.20	1.00
Quality effect = 0.2	0.37	0.60	1.00

*Observed heterogeneity (I-squared)
**Observed heterogeneity (tau-squared)

In this dataset, the power to detect effects of quality was low, when the effects of quality were 0.2. When heterogeneity was reduced, the power increased, although even with half the additional variance power was still lower than ideal, at 0.6. When there was no additional unexplained heterogeneity power was very high. This dataset contained a larger number of trials,

with large sample sizes, which further increase the power to detect quality effects, hence the high power when heterogeneity is low.

Dataset 3 ("Pro-bias" Dataset Specifications)

The final simulations were based on the pro-bias dataset, dataset 3. This dataset contained 100 trials with a mean sample size of 119. Between-study variance was moderate, with I^2 of approximately 60 percent. In the empirical dataset, the effect of study quality ranged from 0.01 to 0.25. As before, we investigated setting the quality moderator effect to 0.1 or 0.2. An additional variance parameter of 0.05 was used, matching the value of I^2 in the empirical dataset (Table 14).

Table 14. Power to detect quality moderator effects determined by Monte Carlo simulation under varying effects of quality and heterogeneity, with simulation parameters matching dataset 3 ('Pro-bias')

Quality Effect	Additional Variance = 0.05	Additional Variance = 0.025	Additional Variance = 0.00
Median I^{2*}	59.6	44.8	0.00
Median τ^{2**}	0.05	0.03	0.01
Quality effect = 0.1	0.42	0.58	0.73
Quality effect = 0.2	0.92	0.99	1.00

*Observed heterogeneity (I-squared)
**Observed heterogeneity (tau-squared)

In a dataset with these specifications, the power to detect quality effects was acceptable when the quality effect was as high as 0.2, however, for smaller quality effects, the power was reduced. Again, with reduced heterogeneity, the power increases, although even with zero additional heterogeneity, the power to detect a quality effect of 0.1 still does not reach generally acceptable levels.

Summary and Discussion

Summary

In this report we explored associations between a set of quality criteria and reported effect sizes in treatment effect trials, and explored factors influencing the presence and detection of associations in meta-epidemiological datasets.

In an analyzed empirical dataset, associations between quality and effect sizes were small, e.g., the ROR between unconcealed and concealed trials was 0.89 (95% CI: 0.73, 1.09; "Heterogeneity set") but was overall consistent across the CBRG criteria. Based on a quantitative summary score, a cut-off of six or more criteria met (out of 11) differentiated low- and high-quality trials best in three out of four datasets (e.g., ROR 0.86, 95% CI: 0.70, 1.06; "Heterogeneity set") but the effect was small and not statistically significant. Results for evidence of bias varied between meta-epidemiological datasets, individual meta-analyses within datasets, and clinical fields. Across all four datasets, most consistent associations between quality and effect sizes were found for allocation concealment with concealed trials showing smaller treatment effects.

Observed heterogeneity between studies varied across meta-epidemiological datasets. In an exploratory analysis, one dataset suggested a correlation between the effect size distribution and observed associations between quality and effect sizes; however, the effect could not be replicated in a second dataset.

The simulations showed that the power to detect quality effects is to a large extent determined by the degree of residual heterogeneity present in the dataset. Even large quality effects in simulations set up to maximize statistical power could not be detected in the presence of a large amount of additional heterogeneity across trials, that is, heterogeneity not due to quality.

Observed Quality Effects

We set out to explore whether the proposed extended list of quality criteria is useful in different clinical fields (key question 1). Validating criteria and scales used to assess the quality of trials is challenging. The concept of quality is not easy to define, and there is no widely accepted gold standard. In addition, while the conceptual, underlying quality domains may be critical dimensions in the evaluation of trials, the applied quality criteria are a translation of the theoretical concepts. The theoretical concept is not identical with the applied test or measure, here the operationalization of the concept as individual quality criteria with scoring instructions for reviewers. The operationalization may or may not be successful and the validity of each operationalization needs to be assessed. We have previously reported evidence of convergent validity for the CBRG criteria (Hempel et al., 2011), based on the Jadad items and scale and the criteria suggested by Schulz et al. (1995) that we applied in parallel. However, evidence of bias would represent predictive validity – is meeting or not meeting the quality criteria associated with differential effect sizes?

To judge the quality of published studies, we depend on the features the authors chose to report in publications. The information depends in part on space restrictions of journals as well as changing reporting standards. The datasets we employed consisted of trials published over a long time period. Concerted efforts have been made to improve and standardize the reporting of empirical studies, particularly RCTs, resulting in the widely accepted Consort Statement, first

published in 1996 (Begg et al., 1996). Generally, the quality of reporting has improved since the publication of the Consort statement (Kane, Wang & Garrard, 2007), and we, too, were able to show a significant trend in this direction in previously analyzed datasets (Hempel et al., 2011). The dataset assembled for this report was based on older meta-analyses and trials published earlier than previously analyzed datasets and is therefore particularly likely to suffer from a lack of reporting rather than lack of quality in the trial design or execution. In accordance with other investigations of quality effects (e.g., Hartling, Ospina, Liang, et al., 2009), we compared studies where criteria were reported as having been met with studies where the quality feature was not reported as having been met. We were concerned primarily with demonstrating the effects of high-quality studies; studies that reported a feature and the expression of the feature were an indicator of high quality. The publication year may have been a confounding factor in our analyses.

Similar to results reported by Balk et al. (2002), none of the individual quality criteria assessed in a meta-epidemiological dataset that was a replication of part of their dataset was associated with a statistically significant difference in effect sizes between trials meeting and trials not meeting the quality criteria. Balk and colleagues concluded from this (and the lack of consistency across clinical areas) that individual quality criteria are not reliably associated with reported treatment effect sizes. In our meta-epidemiological datasets we found that for CBRG criteria associations between quality and effect sizes were small; for example, the ROR between unconcealed and concealed trials was 0.89 (95% CI: 0.73, 1.09; "Heterogeneity set") and not statistically significant but there seemed to be evidence of consistency across the CBRG criteria, indicating that low quality trials tended to overestimate treatment effects. We also found this trend across all four meta-epidemiological datasets, suggesting a non-random association between quality and effect sizes. In particular, concealed trials were consistently associated with smaller treatment effects in each of the four datasets.

We investigated the effects of individual quality criteria as well as a quantitative summary score and explored different cut-offs and their ability to differentiate between low- and high-quality studies (shown as differences in reported effect sizes between trials meeting a number of quality criteria versus trials not meeting these). Based on a quantitative summary score derived from the individual CBRG criteria, a cut-off of six or more criteria met (out of 11) differentiated low- and high-quality trials best in three out of four datasets (e.g., ROR 0.86, 95% CI: 0.70, 1.06; "Heterogeneity set") suggesting an association between quality and effect sizes. However, this difference in effect size differences was only statistically significant in one out of four datasets, in the Back pain dataset for which clinical area the criteria were originally designed (see van Tulder et al., 2009).

Results for evidence of bias varied between meta-epidemiological datasets, broad clinical fields, and individual meta-analyses within datasets. All four meta-epidemiological datasets were derived from different meta-analyses representing diverse clinical fields. In addition, all included trials were selected based on their inclusion in specific meta-analyses rather than randomly selected from a pool of trials. Sterne et al. (2002) describe methods to allow for clustering on meta-analyses within a meta-epidemiological dataset, and we investigated the effects of clustering by contrasting the corrected and uncorrected standard errors between quality and effect sizes. In our analyses, the correction had no noticeable effect on estimates of the associations between quality and effect sizes. The use of the sandwich estimator is appropriate when any correlation of effect sizes within studies is considered to be a nuisance to be accounted for. Two alternative approaches were considered, a three-level meta-analysis, in which a

multilevel modeling framework is used to account for clustering of studies, or a two stage meta-analysis, in which groups of studies are meta-analyzed and a meta-analysis of those results is carried out. The distribution of quality criteria across meta-analyses within each dataset was unbalanced—sometimes extremely so—hence the multilevel or the three level approach would absorb any quality effects into the random intercept.

While results clearly showed some variation between meta-epidemiological datasets, clinical fields, and meta-analyses, we also found that aggregating from individual criteria to summary scores resulted in less variation across the two broad clinical fields, pediatrics and cardiovascular disease, assessed as part of the "Heterogeneity" meta-epidemiological dataset. The use of checklists of individual criteria when scoring quality versus the application of a quantitative summary score has been extensively discussed in the literature. Juni and colleagues (Juni, Witschi, Bloch, et al., 1999; Juni, Altman, et al., 2001) raised serious concerns about the use of quantitative summary scores. However, treating all quality items as truly independent variables also does not appear appropriate, as outlined in Hempel et al. (2011). Individual CBRG quality items showed substantial inter-item correlations.

The equal weighting of each item as applied in our summary score approach is commonplace but has not been 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. Individual quality criteria could be used to trigger an overall assessment of quality, which is more qualitatively than quantitatively derived, by adding individual item scores. 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 & Green, 2009). However, the reliability of qualitative overall evaluations has to be questioned, as Hartling et al. (2009) reported a kappa of 0.27 for reviewers to agree on the Cochrane Overall Risk of Bias dimension (Higgins & Green, 2008). Instead, a combined qualitative and quantitative approach might be useful: quality features could be ranked by importance for the clinical field *a priori* and weighted accordingly for a summary score.

Based on results shown in this report, it cannot be determined whether the proposed extended list of quality criteria should be applied regularly when judging the quality of studies. Individual criteria point to a uniform trend, but no statistically significant effect of quality on effect sizes was seen for any of the individual criteria. Based on a summary score that takes all 11 items into account, the largest difference in effect sizes was seen when applying a cut-off of six or more quality criteria met. However, the difference of combined measure did only marginally exceed those of individually criteria and statistically significant results were not shown in this dataset either. In order to evaluate this result, it is important to know the quality of the applied test as outlined in the section that follows on the detection of quality effects.

Detection of Quality Effects

We used Monte Carlo simulations to systematically investigate the factors that influence the power to detect quality effects (key question 2). In these simulations, we set the effect of quality to a particular level and investigated its influence on the power to detect these effects. Monte Carlo simulations have been used to investigate the properties of meta-analysis parameter estimates under different conditions (e.g. Field, 2001; Field, 2005; Morton et al., 2004). No other studies have investigated (to our knowledge) the power to detect quality effects (or other study level moderators) under varying levels of study heterogeneity. In the simulations presented here,

we investigated the effect of unexplained heterogeneity on the power to detect moderator effects of study quality.

In these simulations, power to detect quality moderator effects in the three datasets was variable. The simulations were set up to maximize the power. It should be noted that we assumed a best-case scenario by assigning 50 percent of studies to high quality and 50 percent to low quality. If this ratio were to move away from 50 percent, as is typical for empirical datasets, power would be reduced further. Furthermore, given the effect size for the main effect of the studies of approximately 0.5, a quality moderator effect of 0.2 means that the size of the effect would be overestimated by 60 percent in low quality studies. Yet even this large amount of bias due to quality proved difficult to detect in some of the analyses. We found that unexplained heterogeneity, at the levels estimated in our three datasets, dramatically reduced the power to detect a quality effect. The simulations therefore revealed that the estimates of the quality effects that failed to achieve significance are to be expected, even when quality effects are large relative to the size of the treatment effect. Where we also simulated datasets with reduced heterogeneity, we found improvements in power. Power can be increased through greater consistency of effect sizes between trials, a larger number of trials, larger trials, or larger effects of quality than are typically found in individual meta-analyses and some meta-epidemiological datasets.

While little research has directly examined the effects of heterogeneity on power in meta-analysis, there is a larger literature on the equivalent problem in cluster randomized trials, which are frequently analyzed with random effects multilevel models (Murray, 1998; Hayes and Moulton, 2010). In cluster randomized trials, the intra-class correlation describes the degree of similarity of patients within a cluster, relative to patients between clusters. In meta-analysis, patients are nested within trials, so heterogeneity means that patients within a trial are more closely related to patients in other trials than to each other.

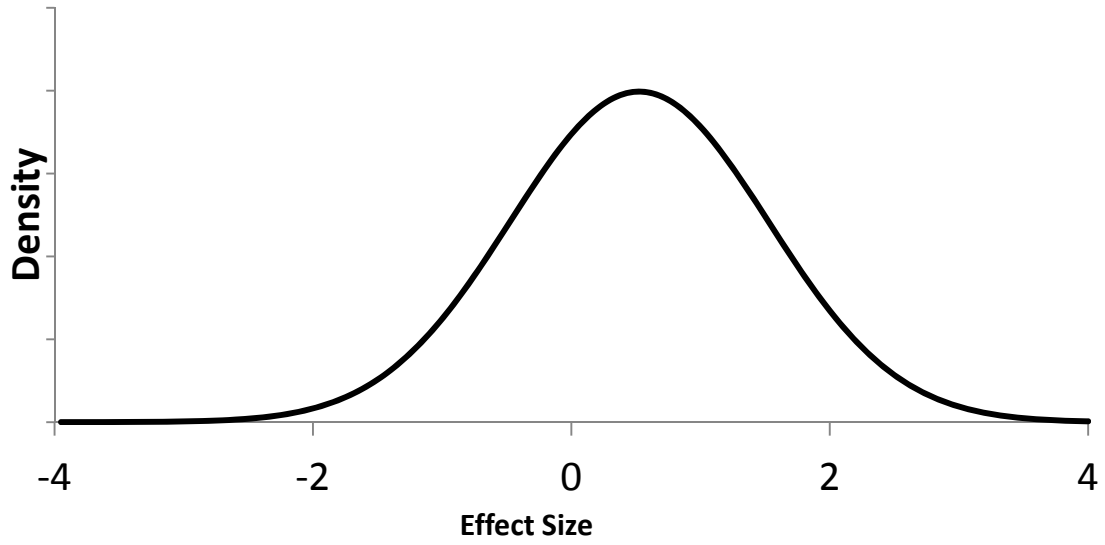
The reported results provide assistance in interpreting the differences in results for associations between quality and effect sizes observed in previous datasets, documented in this and a previous report (Hempel et al., 2011). The results will also assist in planning future studies as outlined in the future research section.

Causes and Implications of Heterogeneity

The simulation results may appear confusing with regards to heterogeneity. Trial quality is typically assessed as one source of heterogeneity. However, the simulations show that heterogeneity can also hinder the detection of quality effects. Quality effects can be a source of heterogeneity across trial results; however, unexplained heterogeneity (that is heterogeneity not due to quality effects), dramatically reduces the power to detect a quality effect. It is therefore useful to carefully examine the causes and implications of heterogeneity.

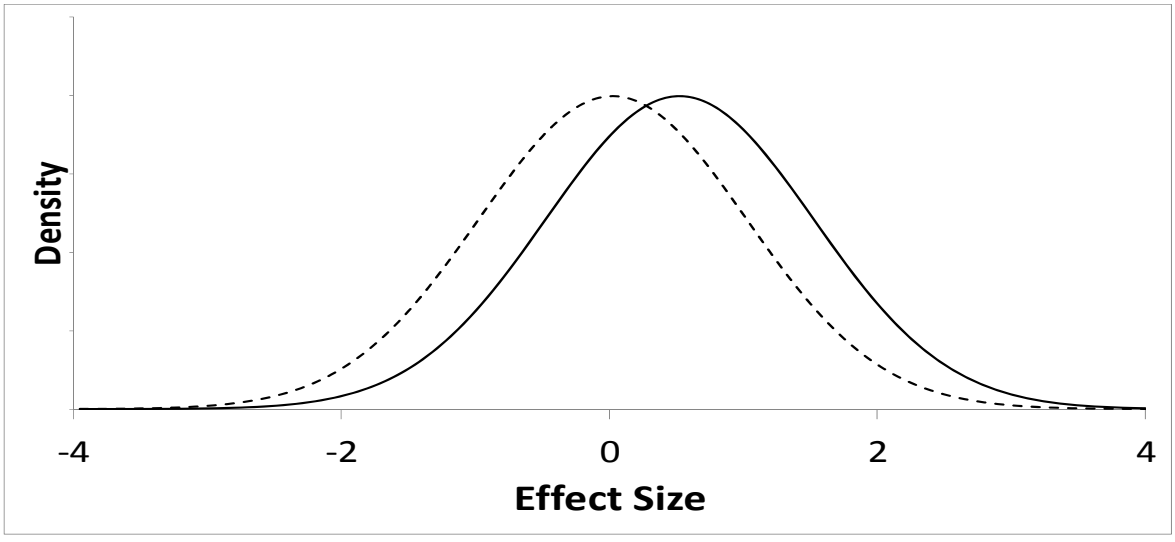
Although meta-analyses deal with heterogeneity in a variety of ways (Schroll et al, 2011), meta-regressions or subgroup analyses are conducted to explore whether differences in trial quality explain differences in reported trial results. However, the causes and implications of heterogeneity can be confused. If studies are considered to be samples from distinct populations – for example, of high and low quality, this will necessarily introduce a degree of heterogeneity. For example, when the studies are homogenous, the distribution of effect sizes is expected to follow a normal distribution, as shown in the figure (Figure 12). We have deliberately chosen a large effect size in the following examples to illustrate the causes and implications of heterogeneity.

Figure 12. Distribution of effect sizes where pooled effect size = 0.5



Where there are two populations (for example, trials with and without blinding) there will be two overlapping distributions. The following figure (Figure 13) shows two populations of trials in which one population (solid line) has an effect size = 0.5, the second population has an effect size of 0.0. In this chart, there is a clear separation of the two populations, and this is therefore a source of heterogeneity.

Figure 13. Distributions of effect sizes from two populations where pooled effect sizes are = 0.5 – (solid line) and 0 (dashed line)



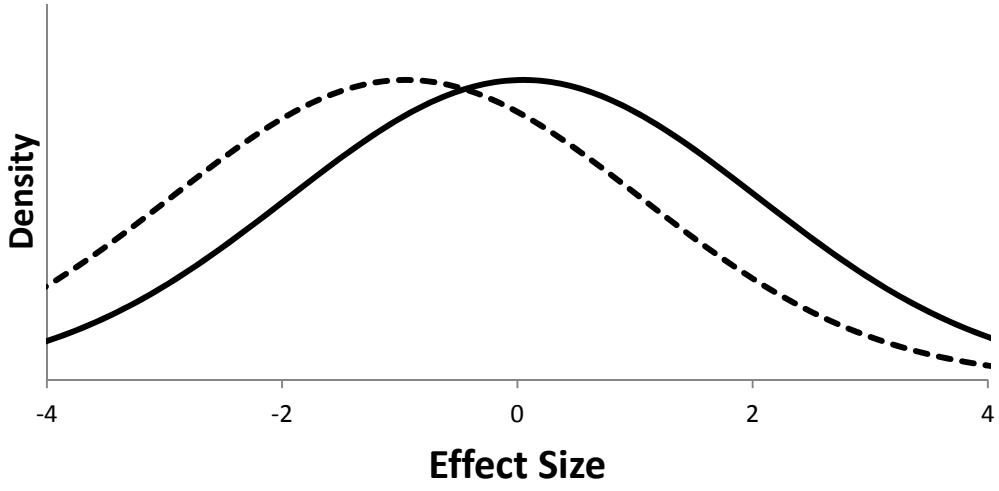
The sum of the two distributions gives a mixture of distributions, as shown in the next figure (Figure 14), where (because the sample sizes were equal) there is a pooled estimate of the effect of 0.25.

Figure 14. Mixture of distributions from two populations, effect size = 0 and effect size = 0.5

This distribution shown in this figure has heavier tails and is wider and flatter than a normal distribution shown in the previous figures, although it is unlikely that this would be detectable in practice. However, the variance of the distribution of this pooled sample is larger than would be expected which would signal that there was heterogeneity which could be explored.

When additional sources of heterogeneity are also present, in addition to heterogeneity due to the quality of studies, the heterogeneity is harder to discern; both by eye, and statistically. In the following figure (Figure 15), we have increased the degree of heterogeneity (the standard deviation is doubled), whilst maintaining the difference between effect sizes (at 0.5). This has the effect of halving the relative effect size, and thereby reducing the power to detect a difference between the two populations.

Figure 15. Difference between effect sizes of 0.5 with SD of effect sizes equal to 2



As shown in the figure, heterogeneity has therefore increased the difficulty of detecting the difference between high and low quality studies.

In our empirical data, a comparison between meta-epidemiological datasets showed that observed heterogeneity between trials within the meta-epidemiological datasets varied across datasets. In an exploratory analysis we attempted to correlate distribution (shape) characteristics and the strength of associations between quality and effect sizes but results were inconclusive. Quality effects are one source of heterogeneity which can be accounted for in the meta-analysis; however there are a very wide range of other factors that can introduce heterogeneity which are not quantified in the published papers, and the effects of which may be impossible to quantify. For example, in a study investigating quality effects in a diverse set of studies such as this, the effectiveness of the intervention and the sensitivity of the outcome may vary across trials. Even if these factors were accounted for, heterogeneity may still exist, Glasziou and Sanders (2002) discussed factors which cause heterogeneity, and suggested that they included features of the population such as age, severity of illness, gender, intervention factors (does, timing, duration of treatment) and co-interventions.

Limitations

As outlined, we tested a specific set of quality criteria and our conclusions for individual quality criteria are specific to the applied operationalization. The criteria tried to capture conceptual quality domains; however, we only have empirical data on the applied quality criteria scored by individual reviewers, the translation of the theoretical concepts.

For the current analyses we set out to use different meta-epidemiological datasets with known qualities. For the two published datasets that were selected (Moher et al., 1998; Balk et al., 2001) we tried to closely match study selections and analytical results. However, since the publications reported only limited detail, the replication was close but not identical. In addition, different quality measures and scoring interpretations were used to assess the quality of trials.

The different datasets were compiled for unique reasons and selection criteria. Among the four datasets, dataset 2 (the EPC report sample) showed the most unusual results for all included quality assessment instruments, CBRG criteria and others. In particular, the exploration of the effect size distribution points to the dataset as an outlier compared to the other compiled datasets. Most notable were the skewed distribution and extreme differences in effect sizes across conditions, as previously discussed (Hempel et al., 2011). This sample was a convenience sample based on availability of data. In addition, the included meta-analyses may have been too diverse with regard to clinical fields, interventions, and outcomes, which, in turn, may have introduced too many significant confounders to be useful for analyzing the research question.

The data showed that a cut-off of six or more criteria met (out of 11) differentiated low- and high-quality trials best (ROR 0.86, 95% CI: 0.70, 1.06), however, the difference between study groups was not statistically significant and two other cut-offs were only marginally different in their performance (cut-off at 7 and 8 criteria met showed a ROR or 0.88).

We did not assess the inter-rater reliability as part of this project, but other reliability measures have been reported in a previous report (Hempel et al., 2011). The reproducibility of quality judgments across independent raters is a valuable method for estimating the reliability of proposed items, but for evidence reviews it is standard to reconcile independent rater decisions. This approach helps to avoid individual reviewer bias and errors and the reconciled decision should be more reliable than the individual decision. Testing the rater agreement of reconciled decisions requires more than one pair of raters.

Detecting and interpreting quality effects are hampered by the observational nature of the data that are available to researchers. When studies differ by quality, it is likely that they differ in

other ways as well, both measured and unmeasured, which also might affect the effect size found in the study. For example, it may be the case that researchers running trials that are better funded are likely to carry out blinding more effectively, which may reduce effect sizes. However, these same researchers may have the resources to pay attention to treatment fidelity, which may improve effect sizes, and increase the effect size.

All of the presented analyses were exploratory in nature. The research field is developing, there are no established standards, and many open questions remain. Consequently, we have formulated cautious conclusions.

Future Research

Our analyses have shown that it is challenging for individual meta-analyses to detect effects of study quality on reported treatment effects. The modeling results showed that under most circumstances it is difficult to detect a quality effect, even when it is present and even when it is substantial. Although they included hundreds of individual trials, the datasets we have compiled in the course of the project did not necessarily have the power to detect influences of quality on effect sizes. From this observation, it follows that the failure to detect a statistically significant quality effect should not be interpreted as meaning that a quality effect is not present.

Future studies that investigate the effects of quality as a moderator of outcome in randomized controlled trials should take steps to ensure that unexplained heterogeneity, that is heterogeneity unrelated to potential quality effects, is minimized. In meta-epidemiological studies, power to detect quality effects can be increased through greater consistency of effect sizes between trials, possibly the careful selection of research areas, or a larger number of trials or larger trials. Theoretically, power would increase as well with larger effects of quality, larger effects than are typically found in individual meta-analyses and some meta-epidemiological datasets. Selection of trials to be incorporated into analyses may allow researchers to construct datasets that maximize power to detect quality effects. This may be done by careful selection of meta-analyses which incorporate both balance in covariates; that is,, with approximately equal numbers of high- and low-quality studies, and without excessive heterogeneity. Power analyses may then be carried out to determine whether the selected studies are likely to have power to detect quality effects.

Recent methodological advances have allowed researchers to make stronger causal conclusions from correlational or observational data. One approach that we believe may be fruitful is the use of propensity based matching (Pearl, 2009; Guo & Fraser, 2010; Rosenbaum & Rubin, 1983; Rosenbaum & Rubin, 1984), in which samples are selected and matched based on observed differences in the variables of interest.

As outlined, failure to detect a statistically significant quality effect should not be interpreted as meaning that a quality effect is not present in individual meta-analyses. Furthermore, based on our analyses, individual meta-analyses should include steps to minimize heterogeneity through the inclusion of additional study level covariates. These refinements can reduce unexplained heterogeneity and thereby aid the investigation of quality effects and the potential for bias.

More empirical evidence is needed to determine which quality features are likely to influence reported effect sizes, and under which conditions. This question is of particular importance for the critical appraisal of systematic reviews when aiming to summarize the existing evidence appropriately. Datasets and quality assessments assembled for the course of this project provide a wealth of information and may assist in answering remaining questions in this developing research field. In addition, the Bias in Randomized and Observational Studies (BRANDO)

dataset, with 3,477 trials, may prove to be useful for future meta-epidemiological analyses (see Savovic et al., 2010).

Conclusion

Although trial quality may explain some amount of heterogeneity across trial results in meta-analyses, the amount of additional heterogeneity (not due to trial quality) in effect sizes is a crucial factor determining when associations between quality and effect sizes can be detected. Detecting quality moderator effects requires more statistically powerful analyses than are employed in many investigations.

References

Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment.. Br Med J (Clin Res Ed). 1988;296(6618):320-31.PMID: 3125883

Assendelft WJ, Morton SC, Yu EI, et al. Spinal manipulative therapy for low back pain. A meta-analysis of effectiveness relative to other therapies. Ann Intern Med. 2003;138(11):871-81.PMID: 12779297

Ausejo M, Saenz A, Pham B, et al. The effectiveness of glucocorticoids in treating croup: meta-analysis. BMJ. 1999;319(7210):595-600.PMID: 10473471

Balk E, Tatsioni A, Lichtenstein A, et al. Effect of chromium supplementation on glucose metabolism and lipids: a systematic review of randomized controlled trials. Diabetes Care. 2007;30(8):2154-63.PMID: 17519436

Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA. 2002;287(22):2973-82.PMID: 12052127

Balk EM, Lichtenstein AH, Chung M, et al. Effects of omega-3 fatty acids on serum markers of cardiovascular disease risk: a systematic review. Atherosclerosis. 2006;189(1):19-30.PMID: 16530201

Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276(8):637-9.PMID: 8773637

Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. Stat Med. 1995;14(4):395-411.PMID: 7746979

Bhuta T, Ohlsson A. Systematic review and meta-analysis of early postnatal dexamethasone for prevention of chronic lung disease. Arch Dis Child Fetal Neonatal Ed. 1998;79(1):F26-33.PMID: 9797621

Chapell R, Reston J, Snyder D. Management of Treatment-Resistant Epilepsy. May 2003.;Evidence Report/Technology Assessment No. 77.(AHRQ Publication No. 03-0028).

Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med. 1989;8(4):441-54.PMID: 2727468

Coulter I, Hardy M, Shekelle P, et al. Effect of the Supplemental Use of Antioxidants Vitamin C, Vitamin E, and Coenzyme Q10 for the Prevention and Treatment of Cancer. 2003;Evidence Report/Technology Assessment Number 75.(AHRQ Publication No. 04-E003).

Counsell C, Sandercock P. Use of anticoagulants in patients with acute ischemic stroke. Stroke. 1995;26(3):522-3.PMID: 7533952

Dolan-Mullen P, Ramirez G, Groff JY. A meta-analysis of randomized trials of prenatal smoking cessation interventions. Am J Obstet Gynecol. 1994;171(5):1328-34.PMID: 7977542

Donahue K, Gartlehner G, Jonas D, et al. Comparative Effectiveness of Drug Therapy for Rheumatoid Arthritis and Psoriatic Arthritis in Adults. November 2007.;Comparative Effectiveness Review No. 11.(AHRQ Publication No. 08-EHC004-EF.).

Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess. 2003;7(1):1-76.PMID: 12583822

Emerson JD, Burdick E, Hoaglin DC, et al. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials. 1990;11(5):339-52.PMID: 1963128

Field AP. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychol Methods. 2001;6(2):161-80.PMID: 11411440

Field AP. Is the meta-analysis of correlation coefficients accurate when population correlations vary? Psychol Methods. 2005;10(4):444-67.PMID: 16392999

Furlan AD, Imamura M, Dryden T, et al. Massage for low-back pain. Cochrane Database Syst Rev. 2008;(4):CD001929.PMID: 18843627

Furlan AD, van Tulder M, Cherkin D, et al. Acupuncture and dry-needling for low back pain: an updated systematic review within the framework of the cochrane collaboration. Spine (Phila Pa 1976). 2005;30(8):944-63.PMID: 15834340

Glasziou P P & Sanders S L. Investigating causes of heterogeneity in systematic reviews. Statistics in Medicine. 2002, 21(11), 1503-1511.

Guo SY & Fraser MW. Propensity Score Analysis: Statistical Methods and Applications. 2010. Thousand Oaks, Ca: Sage.

Hagen KB, Hilde G, Jamtvedt G, et al. Bed rest for acute low back pain and sciatica. Nurs Times. 2001;97(31):40.PMID: 11957537

Hansen RA, Gartlehner G, Webb AP, et al. Efficacy and safety of donepezil, galantamine, and rivastigmine for the treatment of Alzheimer's disease: a systematic review and meta-analysis. Clin Interv Aging. 2008;3(2):211-25.PMID: 18686744

Hardy M, Coulter I, Morton SC, et al. S-Adenosyl-L-Methionine for Treatment of Depression, Osteoarthritis, and Liver Disease. October 2002;AHRQ Evidence Report No. 64.(AHRQ Publication No. 02-E034).

Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ. 2009;339:b4012.PMID: 19841007

Hayden JA, van Tulder MW, Malmivaara A, et al. Exercise therapy for treatment of non-specific low back pain. Cochrane Database Syst Rev. 2005;(3):CD000335.PMID: 16034851

Hayes RH and Moulton LH. Cluster randomised trials. 2010. Chapman and Hall.

Hedges, L. V., E. Tipton, and M. C. Johnson. 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 1(1), 39-65.

Hempel S, Suttorp MJ, Miles JNV, Wang Z, Maglione M, Morton S, Johnsen B, Valentine D, Shekelle PG. Empirical Evidence of Associations Between Trial Quality and Effect Sizes. Methods Research Report (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 11-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: http://effectivehealthcare.ahrq.gov.

Henschke N, Ostelo RW, van Tulder MW, et al. Behavioural treatment for chronic low-back pain. Cochrane Database Syst Rev. 2010;7:CD002014.PMID: 20614428

Heymans MW, van Tulder MW, Esmail R, et al. Back schools for nonspecific low back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976). 2005;30(19):2153-63.PMID: 16205340

Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.2. [updated September 2009]. Cochrane Collaboration. 2008.

Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0. [updated March 2011]. Cochrane Collaboration. 2011.

Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.PMID: 12958120

Hine LK, Laird NM, Hewitt P, et al. Meta-analysis of empirical long-term antiarrhythmic therapy after myocardial infarction. JAMA. 1989;262(21):3037-40.PMID: 2509746

Hughes E, Collins J, P. V. WITHDRAWN: Bromocriptine for unexplained subfertility in women. Cochrane Database Syst Rev. 1996;4.

Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1-12.PMID: 8721797

Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42-6.PMID: 11440947

Juni P, Tallon D, Egger M. Garbage in - garbage out'? Assessment of the quality of controlled trials in meta-analyses published in leading journals. Proceedings of the 3rd symposium on systematic reviews: beyond the basics, St Catherine's College, Oxford Oxford: Centre for Statistics in Medicine. 2000:19.

Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282(11):1054-60.PMID: 10493204

Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. J Clin Epidemiol. 2007;60(3):241-9.PMID: 17292017

Karjalainen K, Malmivaara A, van Tulder M, et al. Multidisciplinary biopsychosocial rehabilitation for subacute low back pain in working-age adults: a systematic review within the framework of the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976). 2001;26(3):262-9.PMID: 11224862

Kellner JD, Ohlsson A, Gadomski AM, et al. Efficacy of bronchodilator therapy in bronchiolitis. A meta-analysis. Arch Pediatr Adolesc Med. 1996;150(11):1166-72.PMID: 8904857

Khadilkar A, Milne S, Brosseau L, et al. Transcutaneous electrical nerve stimulation (TENS) for chronic low-back pain. Cochrane Database Syst Rev. 2005;(3):CD003008.PMID: 16034883

Kjaergard LL, Villumsen J, Gluud C. Quality of randomised clinical trials affects estimates of intervention efficacy. Proceedings of the 7th Cochrane colloquium Universita STommaso D'Aquino, Rome Milan: Centro Cochrane Italiano;. 1999:p. 57. (poster B10).

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2001;135(11):982-9.PMID: 11730399

Kjaergard LL, Villumsen J, Gluud C. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2008;149(3):219.PMID: 18942172

Kozyrskyj AL, Hildes-Ripstein GE, Longstaffe SE, et al. Treatment of acute otitis media with a shortened course of antibiotics: a meta-analysis. JAMA. 1998;279(21):1736-42.PMID: 9624028

Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med. 1992;327(4):248-54.PMID: 1614465

Leizorovicz A, Boissel JP. Oral anticoagulant in patients surviving myocardial infarction. A new approach to old data. Eur J Clin Pharmacol. 1983;24(3):333-6.PMID: 6345177

Lensing AW, Prins MH, Davidson BL, et al. Treatment of deep venous thrombosis with low-molecular-weight heparins. A meta-analysis. Arch Intern Med. 1995;155(6):601-7.PMID: 7887755

Lo GH, LaValley M, McAlindon T, et al. Intra-articular hyaluronic acid in treatment of knee osteoarthritis: a meta-analysis. JAMA. 2003;290(23):3115-21.PMID: 14679274

Loonen AJ, Peer PG, Zwanikken GJ. Continuation and maintenance therapy with antidepressive agents. Meta-analysis of research. Pharm Weekbl Sci. 1991;13(4):167-75.PMID: 1834986

Loosemore TM, Chalmers TC, Dormandy JA. A meta-analysis of randomized placebo control trials in Fontaine stages III and IV peripheral occlusive arterial disease. Int Angiol. 1994;13(2):133-42.PMID: 7525794

Mari JJ, Streiner DL. An overview of family interventions and relapse on schizophrenia: meta-analysis of research findings. Psychol Med. 1994;24(3):565-78.PMID: 7991739

Marshall JK, Irvine EJ. Rectal aminosalicylate therapy for distal ulcerative colitis: a meta-analysis. Aliment Pharmacol Ther. 1995;9(3):293-300.PMID: 7654892

Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609-13.PMID: 9746022

Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ. 2005;330(7499):1053.PMID: 15817526

Morton SC, Adams JL, Suttorp M, et al. Meta-regression Approaches: What, Why, When and How? Agency for Healthcare Research and Quality. 2004.PMID:

Murray D Design and analysis of group randomized trials. 1998. New York: Oxford University Press.

Pace F, Maconi G, Molteni P, et al. Meta-analysis of the effect of placebo on the outcome of medically treated reflux esophagitis. Scand J Gastroenterol. 1995;30(2):101-5.PMID: 7732329

Pearl J. Causality, 2nd edition. 2009. Cambridge: Cambridge University Press.

Pildal J, Hrobjartsson A, Jorgensen KJ, et al. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol. 2007;36(4):847-57.PMID: 17517809

Ramirez-Lassepas M, Cipolle RJ. Medical treatment of transient ischemic attacks: does it influence mortality? Stroke. 1988;19(3):397-400.PMID: 3354028

Roelofs PD, Deyo RA, Koes BW, et al. Nonsteroidal anti-inflammatory drugs for low back pain: an updated Cochrane review. Spine (Phila Pa 1976). 2008;33(16):1766-74.PMID: 18580547

Rosenbaum P & Rubin D B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika; 1983;70(1);41.55.

Rosenbaum P & Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association; 1984;79;516-524.

Rosenfeld RM, Post JC. Meta-analysis of antibiotics for the treatment of otitis media with effusion. Otolaryngol Head Neck Surg. 1992;106(4):378-86.PMID: 1533027

Rossouw JE, Lewis B, Rifkind BM. The value of lowering cholesterol after myocardial infarction. N Engl J Med. 1990;323(16):1112-9.PMID: 2215579

Savovic J, Harris RJ, Wood L, et al. Development of a combined database for meta-epidemiological research. Research Synthesis Methods. 2010;2(1):78

Schroll JB, Moustgaard R & Gotzsche P. Dealing with substantial heterogeneity in Cochrane reviews. Cross-sectional study. BMC Medical Research Methodology. 2011;(11:22).

Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408-12.PMID: 7823387

Shekelle P, Hardy ML, Coulter I, et al. Effect of the supplemental use of antioxidants vitamin C, vitamin E, and coenzyme Q10 for the prevention and treatment of cancer. Evid Rep Technol Assess (Summ). 2003;(75):1-3.PMID: 15523748

Shekelle PG, Maglione M, Bagley S, et al. Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics. January 2007;Comparative Effectiveness Review No. 6.

Shekelle PG, Morton SC, Maglione M, et al. Pharmacological and surgical treatment of obesity. Evid Rep Technol Assess (Summ). 2004;(103):1-6.PMID: 15526396

Sterne JA, Juni P, Schulz KF et al. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in Medicine, 2002; 21:1513-1524.

Sutherland LR, May GR, Shaffer EA. Sulfasalazine revisited: a meta-analysis of 5-aminosalicylic acid in the treatment of ulcerative colitis. Ann Intern Med. 1993;118(7):540-9.PMID: 8095128

Teo KK, Yusuf S. Role of magnesium in reducing mortality in acute myocardial infarction. A review of the evidence. Drugs. 1993;46(3):347-59.PMID: 7693427

Towfigh A, Romanova M, Weinreb JE, et al. Self-monitoring of blood glucose levels in patients with type 2 diabetes mellitus not taking insulin: a meta-analysis. Am J Manag Care. 2008;14(7):468-75.PMID: 18611098

van Duijvenbode IC, Jellema P, van Poppel MN, et al. Lumbar supports for prevention and treatment of low back pain. Cochrane Database Syst Rev. 2008;(2):CD001823.PMID: 18425875

van Tulder M, Furlan A, Bombardier C, et al. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine (Phila Pa 1976). 2003;28(12):1290-9.PMID: 12811274

van Tulder MW, Suttorp M, Morton S, et al. Empirical evidence of an association between internal validity and effect size in randomized controlled trials of low-back pain. Spine (Phila Pa 1976). 2009;34(16):1685-92.PMID: 19770609

van Tulder MW, Touray T, Furlan AD, et al. Muscle relaxants for nonspecific low back pain: a systematic review within the framework of the cochrane collaboration. Spine (Phila Pa 1976). 2003;28(17):1978-92.PMID: 12973146

Verhagen AP, de Vet HC, de Bie RA, et al. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol. 2001;54(7):651-4.PMID: 11438404

West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ). 2002;(47):1-11.PMID: 11979732

Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601-5.PMID: 18316340

Yusuf S, Collins R, MacMahon S, et al. Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomised trials. Lancet. 1988;1(8594):1088-92.PMID: 2896919

Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. Eur Heart J. 1985;6(7):556-85.PMID: 3899654

Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985;27(5):335-71.PMID: 2858114

Abbreviations and Acronyms

AHRQ	Agency for Healthcare Research and Quality
CBRG	Cochrane Back Review Group
CI	Confidence Interval
EPC	Evidence-based Practice Center
ES	Effect Size
ITT	Intention-to-treat
NSAID	Non-steroidal Anti-inflammatory Drugs
OR	Odds Ratio
RCT	Randomized Controlled Trial
ROR	Ratio of the Odds Ratio

Appendix A. Quality Rating Form and References of Included Trials

Article ID:	Reviewer:
First Author, Year: (Last Name Only)	
Meta-analysis:	

Original Quality Items

(adapted from Cochrane Back Review Group)

1. Was the study described as randomized?

Yes ..... ❑
No..... ❑

2. Treatment Allocation

a. Was the method of randomization adequate

Yes ..... ❑
No..... ❑
Don't know ..... ❑

b. Was the treatment allocation concealed?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

3. Were the groups similar at baseline regarding the most important prognostic indicators?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

4. Was the outcome assessor blinded?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

5. Was the care provider blinded?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

6. Were patients blinded?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

7. Were point estimates and measures of variability presented for the primary outcome measures?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

8. Was the drop-out rate described and the reason given?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

9. Was the drop-out rate acceptable?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

10. Were all randomized participants analyzed in the group to which they were originally assigned?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

11. Other sources of potential bias:

a. Were co-interventions avoided or similar?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

b. Was the compliance acceptable in all groups?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

c. Was the timing of the outcome assessment similar in all groups?

Yes ..... ❑
No..... ❑
Don't know ..... ❑

Scoring Guidelines Cochrane Back Pain Group

Table 3. Criteria for a judgment of 'yes' for the sources of risk of bias

1 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, computer-generated 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

2 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.

3 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.

4 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

5 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

6 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 drop-outs does not exceed 20% for short-term follow-up and 30% for long-term follow-up and does not lead to substantial bias a 'yes' is scored. (N.B. these percentages are arbitrary, not supported by literature).

7 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.

8 Selective outcome reporting

In order to receive a ‘yes’, the review author determines if all the results from all pre-specified outcomes have been adequately reported in the published report of the trial. This information is either obtained by comparing the protocol and the report, or in the absence of the protocol, assessing that the published report includes enough information to make this judgment.

9 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).

10 Co-Interventions

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

11 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.

12 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 Interventions2;5;9. 2008 Updated Guidelines for Systematic Reviews 9April 2008

Jadad Scale

Instrument to Measure the Likelihood of Bias in Pain Research Reports
This is not the same as being asked to review a paper. It should not take more than 10 minutes to score a report and there are no right or wrong answers. Please read the article and try to answer the following questions (see attached instructions): Scoring the items: Either give a score of 1 point for each "yes" or 0 points for each "no." There are no in-between marks.

Dimension			Sub Score
Randomization	1. Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)? = 1 point	Give 1 additional point if: For question 1, the method to generate the sequence of randomization was described and it was appropriate (table of random numbers, computer generated, etc.) 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.)	
Blinding	2. Was the study described as double blind? = 1 point	Give 1 additional point: If for question 2 the method of double blinding was described and it was appropriate (identical placebo, active placebo, dummy, etc.) 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)	
Withdrawals and dropouts	3. Was there a description of withdrawals and dropouts? = 1 point		
		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 (1995) Scoring

(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

Cochrane Risk of Bias Tool

Domain	Criteria	Review authors' judgment
Sequence generation	Yes: The investigators describe a random component in the sequence generation process such as: Referring to a random number table; Using a computer random number generator; Coin tossing; Shuffling cards or envelopes; Throwing dice; Drawing of lots; Minimization*. *Minimization may be implemented without a random element, and this is considered to be equivalent to being random. No: The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: Sequence generated by odd or even date of birth; Sequence generated by some rule based on date (or day) of admission; Sequence generated by some rule based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned above and tend to be obvious. They usually involve judgement or some method of non-random categorization of participants, for example: Allocation by judgement of the clinician; Allocation by preference of the participant; Allocation based on the results of a laboratory test or a series of tests; Allocation by availability of the intervention. Unclear: Insufficient information about the sequence generation process to permit judgement of 'Yes' or 'No'.	Was the allocation sequence adequately generated? YES / NO / UNCLEAR
Allocation concealment	Yes: Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation: Central allocation (including telephone, web-based and pharmacy-controlled randomization); Sequentially numbered drug containers of identical appearance; Sequentially numbered, opaque, sealed envelopes. No: Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: Using an open random allocation schedule (e.g. a list of random numbers); Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or nonopaque or not sequentially numbered); Alternation or rotation; Date of birth; Case record number; Any other explicitly unconcealed procedure. Unclear: Insufficient information to permit judgement of 'Yes' or 'No'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.	Was allocation adequately concealed? YES / NO / UNCLEAR
Blinding of participants, personnel and outcome assessors, *Outcome:*	Yes: Any one of the following: No blinding, but the review authors judge that the outcome and the outcome measurement are not likely to be influenced by lack of blinding; Blinding of participants and key study personnel ensured, and unlikely that the blinding could have been broken; Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the non-blinding of others unlikely to introduce bias. No: Any one of the following: No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding; Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken; Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias. Unclear: Any one of the following: Insufficient information to permit judgment of 'Yes' or 'No'; The study did not address this outcome.	Was knowledge of the allocated intervention adequately prevented during the study? YES / NO / UNCLEAR
Incomplete outcome data, *Outcome:*	Yes: Any one of the following: No missing outcome data; Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a clinically relevant impact on the intervention effect estimate; For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a clinically	Were incomplete outcome data adequately addressed? YES / NO / UNCLEAR

	relevant impact on observed effect size; Missing data have been imputed using appropriate methods. No: Any one of the following: Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate; For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size; 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization; Potentially inappropriate application of simple imputation. Unclear: Any one of the following: Insufficient reporting of attrition/exclusions to permit judgement of 'Yes' or 'No' (e.g. number randomized not stated, no reasons for missing data provided); The study did not address this outcome.	
Selective outcome reporting	Yes: Any of the following: The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon). No: Any one of the following: Not all of the study's pre-specified primary outcomes have been reported; One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; The study report fails to include results for a key outcome that would be expected to have been reported for such a study. Unclear: Insufficient information to permit judgement of 'Yes' or 'No'. It is likely that the majority of studies will fall into this category.	Are reports of the study free of suggestion of selective outcome reporting? YES / NO / UNCLEAR
Other sources of bias	Yes: The study appears to be free of other sources of bias. No: There is at least one important risk of bias. For example, the study: Had a potential source of bias related to the specific study design used; or Stopped early due to some data-dependent process (including a formal-stopping rule); or Had extreme baseline imbalance; or Has been claimed to have been fraudulent; or Had some other problem. Unclear: There may be a risk of bias, but there is either: Insufficient information to assess whether an important risk of bias exists; or Insufficient rationale or evidence that an identified problem will introduce bias.	Was the study apparently free of other problems that could put it at a high risk of bias? YES / NO / UNCLEAR
Overall risk of bias	Low: Plausible bias unlikely to seriously alter the results; low risk of bias for all key domains. Unclear: Plausible bias that raises some doubt about the results; unclear risk of bias for one or more key domains. High: Plausible bias that seriously weakens confidence in the results; high risk of bias for one or more key domains.	HIGH / LOW / UNCLEAR

References of Trials Included in the “Heterogeneity”- Set

[no author] A prospective trial of intravenous streptokinase in acute myocardial infarction (I.S.A.M.). Mortality, morbidity, and infarct size at 21 days. The I.S.A.M. Study Group. N Engl J Med. 1986 Jun 5;314(23):1465-71.

[no author] A randomized trial of aspirin and sulfinpyrazone in threatened stroke. The Canadian Cooperative Study Group. N Engl J Med. 1978 Jul 13;299(2):53-9.

[no author] A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA. 1982 Mar 26;247(12):1707-14.

[no author] A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. JAMA. 1980 Feb 15;243(7):661-9.

[no author] Aspirin in coronary heart disease. The Coronary Drug Project Research Group. J Chronic Dis. 1976 Oct;29(10):625-42.

[no author] Clofibrate and niacin in coronary heart disease. JAMA. 1975 Jan 27;231(4):360-81.
Controlled trial of soya-bean oil in myocardial infarction. Lancet. 1968 Sep 28;2(7570):693-9.

[no author] Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Lancet. 1986 Feb 22;1(8478):397-402.

[no author] Effects of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS. The Cardiac Arrhythmia Pilot Study (CAPS) Investigators. Am J Cardiol. 1988 Mar 1;61(8):501-9.

[no author] European Infarction Study (E.I.S.). A secondary prevention study with slow release oxprenolol after myocardial infarction: morbidity and mortality. Eur Heart J. 1984 Mar;5(3):189-202.

[no author] International mexiletine and placebo antiarrhythmic coronary trial: I. Report on arrhythmia and other findings. Impact Research Group. J Am Coll Cardiol. 1984 Dec;4(6):1148-63.

[no author] Ischaemic heart disease: a secondary prevention trial using clofibrate. Report by a research committee of the Scottish Society of Physicians. Br Med J. 1971 Dec 25;4(5790):775-84.
Low-fat diet in myocardial infarction: A controlled trial. Lancet. 1965 Sep 11;2(7411):501-4.

[no author] Persantine and aspirin in coronary heart disease. The Persantine-Aspirin Reinfarction Study Research Group. Circulation. 1980 Sep;62(3):449-61.

[no author] Phenytoin after recovery from myocardial infarction. Controlled trial in 568 patients. Lancet. 1971 Nov 13;2(7733):1055-7.

[no author] Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med. 1989 Aug 10;321(6):406-12.

[no author] Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Lancet. 1988 Aug 13;2(8607):349-60.

[no author] Reduction in mortality after myocardial infarction with long-term beta-adrenoceptor blockade. Multicentre international study: supplementary report. Br Med J. 1977 Aug 13;2(6084):419-21.

[no author] Streptokinase in acute myocardial infarction. European Cooperative Study Group for Streptokinase Treatment in Acute Myocardial Infarction. N Engl J Med. 1979 Oct 11;301(15):797-802.

[no author] Streptokinase in recent myocardial infarction: a controlled multicentre trial. European working party. Br Med J. 1971 Aug 7;3(5770):325-31.

[no author] The effect of pindolol on the two years mortality after complicated myocardial infarction. Eur Heart J. 1983 Jun;4(6):367-75.

[no author] Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. N Engl J Med. 1981 Apr 2;304(14):801-7.

[no author] Trial of clofibrate in the treatment of ischaemic heart disease. Five-year study by a group of physicians of the Newcastle upon Tyne region. Br Med J. 1971 Dec 25;4(5790):767-75.

[no author] United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: interim results. UK-TIA Study Group. Br Med J (Clin Res Ed). 1988 Jan 30;296(6618):316-20.

Aber CP, Bass NM, Berry CL, Carson PH, Dobbs RJ, Fox KM, et al. Streptokinase in acute myocardial infarction: a controlled multicentre study in the United Kingdom. Br Med J. 1976 Nov 6;2(6044):1100-4.

Abraham AS, Rosenmann D, Kramer M, Balkin J, Zion MM, Farbstien H, et al. Magnesium in the prevention of lethal arrhythmias in acute myocardial infarction. Arch Intern Med. 1987 Apr;147(4):753-5.

Adam D. Five-Day therapy with cefpodoxime versus ten-day treatment with cefaclor in infants with acute otitis media. Infection. 1995;23:398-9.

Alario AJ, Lewander WJ, Dennehy P, Seifer R, Mansell AL. The efficacy of nebulized metaproterenol in wheezing infants and young children. Am J Dis Child. 1992 Apr;146(4):412-8.

Amery A, Roeber G, Vermeulen HJ, Verstraete M. Single-blind randomised multicentre trial comparing heparin and streptokinase treatment in recent myocardial infarction. Acta Med Scand Suppl. 1969;505:1-35.

Andersen MP, Bechsgaard P, Frederiksen J, Hansen DA, Jurgensen HJ, Nielsen B, et al. Effect of alprenolol on mortality among patients with definite or suspected acute myocardial infarction. Preliminary results. Lancet. 1979 Oct 27;2(8148):865-8.

Arguedas A, Loaiza C, Herrera M, Mohs E. Comparative trial of 3-day azithromycin versus 10-day amoxycillin/clavulanate potassium in the treatment of children with acute otitis media with effusion. Int J Antimicrob Agents. 1996 Apr;6(4):233-8.

Arguedas A, Loaiza C, Rodriguez F, Herrera ML, Mohs E. Comparative trial of 3 days of azithromycin versus 10 days of clarithromycin in the treatment of children with acute otitis media with effusion. J Chemother. 1997 Feb;9(1):44-50.

Aronovitz G. A multicenter, open label trial of azithromycin vs. amoxicillin/ clavulanate for the management of acute otitis media in children. Pediatr Infect Dis J. 1996 Sep;15(9 Suppl):S15-9.

Aspenstroem G, Korsan-Bengtsen K. A Double Blind Study of Dicumarol Prophylaxis in Coronary Heart Disease. Acta Med Scand. 1964 Nov;176:563-75.

Baber NS, Evans DW, Howitt G, Thomas M, Wilson T, Lewis JA, et al. Multicentre post-infarction trial of propranolol in 49 hospitals in the United Kingdom, Italy, and Yugoslavia. Br Heart J. 1980 Jul;44(1):96-100.

Barber JM, Boyle DM, Chaturvedi NC, Singh N, Walsh MJ. Practolol in acute myocardial infarction. Acta Med Scand Suppl. 1976;587:213-9.

Barnett ED, Teele DW, Klein JO, Cabral HJ, Kharasch SJ. Comparison of ceftriaxone and trimethoprim-sulfamethoxazole for acute otitis media. Greater Boston Otitis Media Study Group. Pediatrics. 1997 Jan;99(1):23-8.

Baroffio R. [Efficacy of a randomized treatment with intravenous streptokinase in evolving myocardial infarction. In-hospital clinical and radionuclide evaluation and a follow-up]. Minerva Cardioangiol. 1986 Oct;34(10):607-14.

Bassand JP, Faivre R, Becque O, Habert C, Schuffenecker M, Petiteau PY, et al. Effects of early high-dose streptokinase intravenously on left ventricular function in acute myocardial infarction. Am J Cardiol. 1987 Sep 1;60(7):435-9.

Bastian BC, Macfarlane PW, McLauchlan JH, Ballantyne D, Clark P, Hillis WS, et al. A prospective randomized trial of tocainide in patients following myocardial infarction. Am Heart J. 1980 Dec;100(6 Pt 2):1017-22.

Benda L, Haider M, Ambrosch F. [Results of the Austrian myocardial infarction study on the effects of streptokinase (author's transl)]. Wien Klin Wochenschr. 1977 Dec 9;89(23):779-83.

Bett JH, Castaldi PA, Hale GS, Isbister JP, McLean KH, O'Sullivan EF, et al. Australian multicentre trial of streptokinase in acute myocardial infarction. Lancet. 1973 Jan 13;1(7794):57-60.

Boulesteix J, Dubreuil C, Moutot M, Rezvani Y, Rosembaum. Cefpodoxime proxteil 5 jours versus cefixime 8 jours, dans le traitment des otites moyennes aigues de l'enfant. Med Mal Infect. 1995;25:534-9.

Breddin K, Ehrly AM, Fechler L, Frick D, Konig H, Kraft H, et al. [Short-term fibrinolytic treatment in acute myocardial infarction]. Dtsch Med Wochenschr. 1973 Apr 27;98(17):861-73.
Breddin K, Loew D, Lechner K, Uberla K, Walter E. Secondary prevention of myocardial infarction. Comparison of acetylsalicylic acid, phenprocoumon and placebo. A multicenter two-year prospective study. Thromb Haemost. 1979 Feb 28;41(1):225-36.

Britton M, Helmers C, Samuelsson K. High dose acetylsalicylic acid after cerebral infarction: a Swedish co-operative study. Stroke. 1987;18:325-34.

Brozanski BS, Jones JG, Gilmour CH, Balsan MJ, Vazquez RL, Israel BA, et al. Effect of pulse dexamethasone therapy on the incidence and severity of chronic lung disease in the very low birth weight infant. J Pediatr. 1995 May;126(5 Pt 1):769-76.

Bussmann WD, Passek D, Seidel W, Kaltenbach M. Reduction of CK and CK-MB indexes of infarct size by intravenous nitroglycerin. Circulation. 1981 Mar;63(3):615-22.

Cairns JA, Gent M, Singer J, Finnie KJ, Froggatt GM, Holder DA, et al. Aspirin, sulfinpyrazone, or both in unstable angina. Results of a Canadian multicenter trial. N Engl J Med. 1985 Nov 28;313(22):1369-75.

Carlson LA, Rosenhamer G. Reduction of mortality in the Stockholm Ischaemic Heart Disease Secondary Prevention Study by combined treatment with clofibrate and nicotinic acid. Acta Med Scand. 1988;223(5):405-18.

Ceremuzynski L, Jurgiel R, Kulakowski P, Gebalska J. Threatening arrhythmias in acute myocardial infarction are prevented by intravenous magnesium sulfate. Am Heart J. 1989 Dec;118(6):1333-4.

Chamberlain DA, Jewitt DE, Julian DG, Campbell RW, Boyle DM, Shanks RG. Oral mexiletine in high-risk patients after myocardial infarction. Lancet. 1980 Dec 20-27;2(8208-8209):1324-7.

Chamberlain JM, Boenning DA, Waisman Y, Ochsenschlager DW, Klein BL. Single-dose ceftriaxone versus 10 days of cefaclor for otitis media. Clin Pediatr (Phila). 1994 Nov;33(11):642-6.

Chaput de Saintonge DM, Levine DF, Savage IT, Burgess GW, Sharp J, Mayhew SR, et al. Trial of three-day and ten-day courses of amoxycillin in otitis media. Br Med J (Clin Res Ed). 1982 Apr 10;284(6322):1078-81.

Cohen R, de La Rocque F, Boucherat M, al. E. Etude randomisee cefpodoxime proxetil 5 jours versus amoxilline-acide calvulanique 8 jours dans le traitment de l'otite moyenne aigue de l'enfant. Med Mal Infect. 1997;27:596-602.

Cohn JN, Franciosa JA, Francis GS, Archibald D, Tristani F, Fletcher R, et al. Effect of short-term infusion of sodium nitroprusside on mortality rate in acute myocardial infarction complicated by left ventricular failure: results of a Veterans Administration cooperative study. N Engl J Med. 1982 May 13;306(19):1129-35.

Conrad LL, Kyriacopoulos JD, Wiggins CW, Honick GL. Prevention of Recurrences of Myocardial Infarction; a Double-Blind Study of the Effectiveness of Long-Term Oral Anticoagulant Therapy. Arch Intern Med. 1964 Sep;114:348-58.

Corwin MJ, Weiner LB, Daniels D. Efficacy of oral antibiotics for the treatment of persistent otitis media with effusion. Int J Pediatr Otorhinolaryngol. 1986 Apr;11(2):109-12.

Cribier A, Berland J, Saoudi N, Redonnet M, Moore N, Letac B. Intracoronary streptokinase, OK! ... Intravenous streptokinase, first? Heparin or intravenous streptokinase in acute infarction: preliminary results of a prospective randomized trial with angiographic evaluation in 44 patients. Haemostasis. 1986;16 Suppl 3:122-9.

Daly K, Giebink G, Lindgren B, Anderson R. Controlled clinical trial for prevention of chronic otitis media with effusion. In: Lim, DJ, eds, Recent advances in otitis media. 1988(Toronto: BC Decker):247-50.

Daniel RR. Comparison of azithromycin and co-amoxiclav in the treatment of otitis media in children. J Antimicrob Chemother. 1993 Jun;31 Suppl E:65-71.

Dewar HA, Stephenson P, Horler AR, Cassells-Smith AJ, Ellis PA. Fibrinolytic therapy of coronary thrombosis. Controlled trial of 75 cases. Br Med J. 1963 Apr 6;1(5335):915-20.

Dioguardi N, Lotto A, Levi GF, Rota M, Proto C, Mannucci PM, et al. Controlled trial of streptokinase and heparin in acute myocardial infarction. Lancet. 1971 Oct 23;2(7730):891-5.

Durand M, Sardesai S, McEvoy C. Effects of early dexamethasone therapy on pulmonary mechanics and chronic lung disease in very low birth weight infants: a randomized, controlled trial. Pediatrics. 1995 Apr;95(4):584-90.

Durand P, Asseman P, Pruvost P, Bertrand ME, LaBlanche JM, Thery C. Effectiveness of intravenous streptokinase on infarct size and left ventricular function in acute myocardial infarction. Prospective and randomized study. Clin Cardiol. 1987 Jul;10(7):383-92.

Durrer JD, Lie KI, van Capelle FJ, Durrer D. Effect of sodium nitroprusside on mortality in acute myocardial infarction. N Engl J Med. 1982 May 13;306(19):1121-8.

Ebert RV. Long-term anticoagulant therapy after myocardial infarction. Final report of the Veterans Administration cooperative study. JAMA. 1969 Mar 24;207(12):2263-7.

Elwood PC, Sweetnam PM. Aspirin and secondary mortality after myocardial infarction. Lancet. 1979 Dec 22-29;2(8156-8157):1313-5.

Ernstson S, Anari M. Cefaclor in the treatment of otitis media with effusion. Acta Otolaryngol Suppl. 1985;424:17-21.

Fields WS, Lemak NA, Frankowski RF, Hardy RJ. Controlled trial of aspirin in cerebral ischemia. Stroke. 1977 May-Jun;8(3):301-14.

Flaherty JT, Becker LC, Bulkley BH, Weiss JL, Gerstenblith G, Kallman CH, et al. A randomized prospective trial of intravenous nitroglycerin in patients with acute myocardial infarction. Circulation. 1983 Sep;68(3):576-88.

Fletcher AP, Sherry S, Alkjaersig N, Smyrniotis FE, Jick S. The maintenance of a sustained thrombolytic state in man. II. Clinical observations on patients with myocardial infarction and other thromboembolic disorders. J Clin Invest. 1959 Jul;38(7):1111-9.

Giebink GS, Batalden PB, Le CT, Lassman FM, Buran DJ, Seltz AE. A controlled trial comparing three treatments for chronic otitis media with effusion. Pediatr Infect Dis J. 1990 Jan;9(1):33-40.
Gooch WM, 3rd, Blair E, Puopolo A, Paster RZ, Schwartz RH, Miller HC, et al. Effectiveness of five days of therapy with cefuroxime axetil suspension for treatment of acute otitis media. Pediatr Infect Dis J. 1996 Feb;15(2):157-64.

Gottlieb SH, Achuff SC, Mellits ED, Gerstenblith G, Baughman KL, Becker L, et al. Prophylactic antiarrhythmic therapy of high-risk survivors of myocardial infarction: lower mortality at 1 month but not at 1 year. Circulation. 1987 Apr;75(4):792-9.

Green SM, Rothrock SG. Single-dose intramuscular ceftriaxone for acute otitis media in children. Pediatrics. 1993 Jan;91(1):23-30.

Hansteen V, Moinichen E, Lorentsen E, Andersen A, Strom O, Soiland K, et al. One year's treatment with propranolol after myocardial infarction: preliminary report of Norwegian multicentre trial. Br Med J (Clin Res Ed). 1982 Jan 16;284(6310):155-60.

Harvald B, Hilden T, Lund E. Long-term anticoagulant therapy after myocardial infarction. Lancet. 1962 Sep 29;2(7257):626-30.

Healy GB. Antimicrobial therapy of chronic otitis media with effusion. Int J Pediatr Otorhinolaryngol. 1984 Oct;8(1):13-7.

Heikinheimo R, Ahrenberg P, Honkapohja H, Iisalo E, Kallio V, Konttinen Y, et al. Fibrinolytic treatment in acute myocardial infarction. Acta Med Scand. 1971 Jan-Feb;189(1-2):7-13.

Hendrickse WA, Kusmiesz H, Shelton S, Nelson JD. Five vs. ten days of therapy for acute otitis media. Pediatr Infect Dis J. 1988 Jan;7(1):14-23.

Henry RL, Milner AD, Stokes GM. Ineffectiveness of ipratropium bromide in acute bronchiolitis. Arch Dis Child. 1983 Nov;58(11):925-6.

Hjalmarson A, Herlitz J, Holmberg S, Ryden L, Swedberg K, Vedin A, et al. The Goteborg metoprolol trial. Effects on mortality and morbidity in acute myocardial infarction. Circulation. 1983 Jun;67(6 Pt 2):I26-32.

Hoberman A, Paradise JL, Burch DJ, Valinski WA, Hedrick JA, Aronovitz GH, et al. Equivalent efficacy and reduced occurrence of diarrhea from a new formulation of amoxicillin/clavulanate potassium (Augmentin) for treatment of acute otitis media in children. Pediatr Infect Dis J. 1997 May;16(5):463-70.

Hockings BE, Cope GD, Clarke GM, Taylor RR. Randomized controlled trial of vasodilator therapy after myocardial infarction. Am J Cardiol. 1981 Aug;48(2):345-52.

Husby S, Agertoft L, Mortensen S, Pedersen S. Treatment of croup with nebulised steroid (budesonide): a double blind, placebo controlled study. Arch Dis Child. 1993 Mar;68(3):352-5.

Ingvarsson L, Lundgren K. Penicillin treatment of acute otitis media in children. A study of the duration of treatment. Acta Otolaryngol. 1982 Sep-Oct;94(3-4):283-7.

Jaffe AS, Geltman EM, Tiefenbrunn AJ, Ambos HD, Strauss HD, Sobel BE, et al. Reduction of infarct size in patients with inferior infarction with intravenous glyceryl trinitrate. A randomised study. Br Heart J. 1983 May;49(5):452-60.

James JA. Dexamethasone in croup. A controlled study. Am J Dis Child. 1969 May;117(5):511-6.

Julian DG, Prescott RJ, Jackson FS, Szekely P. Controlled trial of sotalol for one year after myocardial infarction. Lancet. 1982 May 22;1(8282):1142-7.

Kafetzis DA, Astra H, Mitropoulos L. Five-day versus ten-day treatment of acute otitis media with cefprozil. Eur J Clin Microbiol Infect Dis. 1997 Apr;16(4):283-6.

Kari MA, Heinonen K, Ikonen RS, Koivisto M, Raivio KO. Dexamethasone treatment in preterm infants at risk for bronchopulmonary dysplasia. Arch Dis Child. 1993 May;68(5 Spec No):566-9.

Kennedy JW, Martin GV, Davis KB, Maynard C, Stadius M, Sheehan FH, et al. The Western Washington Intravenous Streptokinase in Acute Myocardial Infarction Randomized Trial. Circulation. 1988 Feb;77(2):345-52.

Khurana CM. A multicenter, randomized, open label comparison of azithromycin and amoxicillin/clavulanate in acute otitis media among children attending day care or school. Pediatr Infect Dis J. 1996 Sep;15(9 Suppl):S24-9.

Klassen TP, Craig WR, Moher D, Osmond MH, Pasterkamp H, Sutcliffe T, et al. Nebulized budesonide and oral dexamethasone for treatment of croup: a randomized controlled trial. JAMA. 1998 May 27;279(20):1629-32.

Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of salbutamol in acute bronchiolitis. J Pediatr. 1991 May;118(5):807-11.

Klassen TP, Watters LK, Feldman ME, Sutcliffe T, Rowe PC. The efficacy of nebulized budesonide in dexamethasone-treated outpatients with croup. Pediatrics. 1996 Apr;97(4):463-6.

Klein W, Pavek P, Brandt D, al. E. Resultate einer doppelblindstudie beim myokardinfarkt. In: Sailer S ed Die Fibriolysebehandlung des Akuten Myokardinfarktes. Wien: Verlag Bruder Hollinek; 1976. p. 65-73.

Lasierra Cirujeda L, Vilades Juan E, Fernandez Clemente JJ, Dulin Verde JJ, Moreno Resina B, Munilla Garcia A, et al. [Streptokinase in myocardial infarct]. Rev Clin Esp. 1977 Feb 28;144(4):251-7.

Leren P. The effect of plasma cholesterol lowering diet in male survivors of myocardial infarction. A controlled clinical trial. Acta Med Scand Suppl. 1966;466:1-92.

Lines D, Bates M, Rechtnan A, Sammartino L. Efficacy of nebulized ipratropium bromide in acute bronchiolitis Pediatr Rev Commun. 1992;6:161-7.

Lines D, JS K, P L. Efficacy of nebulized salbutamol in bronchiolitis. Pediatr Rev Commun. 1990;5:121-9.

Lis Y, Bennett D, Lambert G, Robson D. A preliminary double-blind study of intravenous nitroglycerin in acute myocardial infarction. Intensive Care Med. 1984;10(4):179-84.

Loeliger EA, Hensen A, Kroes F, van Dijk LM, Fekkes N, de Jonge H, et al. A double-blind trial of long-term anticoagulant treatment after myocardial infarction. Acta Med Scand. 1967 Nov;182(5):549-66.

Lowell DI, Lister G, Von Koss H, McCarthy P. Wheezing in infants: the response to epinephrine. Pediatrics. 1987 Jun;79(6):939-45.

Mallol J, Barrueto L, Girardi G, Munoz R, Puppo H, Ulloa V, et al. Use of nebulized bronchodilators in infants under 1 year of age: analysis of four forms of therapy. Pediatr Pulmonol. 1987 Sep-Oct;3(5):298-303.

Mandel EM, Rockette HE, Bluestone CD, Paradise JL, Nozza RJ. Efficacy of amoxicillin with and without decongestant-antihistamine for otitis media with effusion in children. Results of a double-blind, randomized trial. N Engl J Med. 1987 Feb 19;316(8):432-7.

McLinn S. A multicenter, double blind comparison of azithromycin and amoxicillin/ clavulanate for the treatment of acute otitis media in children. Pediatr Infect Dis J. 1996 Sep;15(9 Suppl):S20-3.

Meistrup-Larsen KI, Sorensen H, Johnsen NJ, Thomsen J, Mygind N, Sederberg-Olsen J. Two versus seven days penicillin treatment for acute otitis media. A placebo controlled trial in children. Acta Otolaryngol. 1983 Jul-Aug;96(1-2):99-104.

Meuwissen OJ, Vervoorn AC, Cohen O, Jordan FL, Nelemans FA. Double blind trial of long-term anticoagulant treatment after myocardial infarction. Acta Med Scand. 1969 Nov;186(5):361-8.

Mohs E, Rodriguez-Solares A, Rivas E, el Hoshy Z. A comparative study of azithromycin and amoxycillin in paediatric patients with acute otitis media. J Antimicrob Chemother. 1993 Jun;31 Suppl E:73-9.

Moller P, Dingsor G. Otitis media with effusion: can erythromycin reduce the need for ventilating tubes? J Laryngol Otol. 1990 Mar;104(3):200-2.

Morton BC, Smith FM, McKibbon TG, Nair RC, Poznanski WJ. Magnesium therapy in acute myocardial infarction. Magnesium Bulletin 1. 1981:192-94.

Nelson GI, Silke B, Ahuja RC, Hussain M, Taylor SH. Haemodynamic advantages of isosorbide dinitrate over frusemide in acute heart-failure following myocardial infarction. Lancet. 1983 Apr 2;1(8327):730-3.

Ness PM, Simon TL, Cole C, Walston A. A pilot study of streptokinase therapy in acute myocardial infarction: observations on complications and relation to trial design. Am Heart J. 1974 Dec;88(6):705-12.

Nielsen BL, Clausen J, Nielsen JS. Can procainamide improve the prognosis of patients with ventricular arrhythmias after myocardial infarction? Dan Med Bull. 1978 Apr;25(3):121-5.

Olson HG, Butman SM, Piters KM, Gardin JM, Lyons KP, Jones L, et al. A randomized controlled trial of intravenous streptokinase in evolving acute myocardial infarction. Am Heart J. 1986 Jun;111(6):1021-9.

Pestalozza G, Cioce C, Facchini M. Azithromycin in upper respiratory tract infections: a clinical trial in children with otitis media. Scand J Infect Dis Suppl. 1992;83:22-5.

Peter T, Ross D, Duffield A, Luxton M, Harper R, Hunt D, et al. Effect on survival after myocardial infarction of long-term treatment with phenytoin. Br Heart J. 1978 Dec;40(12):1356-60.

Ploussard JH. Evaluation of five days of cefaclor vs ten days of amoxicillin therapy in acute otitis media. Curr Ther Res. 1984;36:641-5.

Podoshin L, Fradis M, Ben-David Y, Faraggi D. The efficacy of oral steroids in the treatment of persistent otitis media with effusion. Arch Otolaryngol Head Neck Surg. 1990 Dec;116(12):1404-6.

Poliwoda H, Schneider B, Avenarius HJ. [Investigations of the clinical course of acute myocardial infarction. I. The fibrinolytic treatment of acute myocardial infarction with streptokinase (author's transl)]. Med Klin. 1977 Mar 18;72(11):451-8.

Principi N. Multicentre comparative study of the efficacy and safety of azithromycin compared with amoxicillin/clavulanic acid in the treatment of paediatric patients with otitis media. Eur J Clin Microbiol Infect Dis. 1995 Aug;14(8):669-76.

Puczynski MS, Stankiewicz JA, O'Keefe JP. Single dose amoxicillin treatment of acute otitis media. Laryngoscope. 1987 Jan;97(1):16-8.

Rasmussen HS, Gronbaek M, Cintin C, Balslov S, Norregard P, McNair P. One-year death rate in 270 patients with suspected acute myocardial infarction, initially treated with intravenous magnesium or placebo. Clin Cardiol. 1988 Jun;11(6):377-81.

Rastogi A, Akintorin SM, Bez ML, Morales P, Pildes RS. A controlled trial of dexamethasone to prevent bronchopulmonary dysplasia in surfactant-treated infants. Pediatrics. 1996 Aug;98(2 Pt 1):204-10.

Ritland S, Lygren T. Comparison of efficacy of 3 and 12 months' anticoagulant therapy after myocardial infarction. A controlled clinical trial. Lancet. 1969 Jan 18;1(7586):122-4.

Rodriguez AF. An open study to compare azithromycin with cefaclor in the treatment of children with acute otitis media. J Antimicrob Chemother. 1996 Jun;37 Suppl C:63-9.

Ryden L, Arnman K, Conradson TB, Hofvendahl S, Mortensen O, Smedgard P. Prophylaxis of ventricular tachyarrhythmias with intravenous and oral tocainide in patients with and recovering from acute myocardial infarction. Am Heart J. 1980 Dec;100(6 Pt 2):1006-12.

Sainsous J, Bonnet JL, Serradimigni A. Intravenous streptokinase versus heparin in the acute stage of myocardial infarction. A prospective randomised trial in south-east France. Haemostasis. 1986;16 Suppl 3:140-7.

Salathia KS, Barber JM, McIlmoyle EL, Nicholas J, Evans AE, Elwood JH, et al. Very early intervention with metoprolol in suspected acute myocardial infarction. Eur Heart J. 1985 Mar;6(3):190-8.

Schaad UB. Multicentre evaluation of azithromycin in comparison with co-amoxiclav for the treatment of acute otitis media in children. J Antimicrob Chemother. 1993 Jun;31 Suppl E:81-8.

Schreiber TL, Miller DH, Silvasi DA, Moses JW, Borer JS. Randomized double-blind trial of intravenous streptokinase for acute myocardial infarction. Am J Cardiol. 1986 Jul 1;58(1):47-52.

Schwartz RH, Rodriguez WJ. Trimethoprim-sulfamethoxazole treatment of persistent otitis media with effusion. Pediatr Infect Dis. 1982 Sep-Oct;1(5):333-5.

Seaman AJ, Griswold HE, Reaume RB, Ritzmann L. Long-term anticoagulant prophylaxis after myocardial infarction. N Engl J Med. 1969 Jul 17;281(3):115-9.

Shechter M, Hod H, Marks N, Behar S, Kaplinsky E. Magnesium therapy and mortality in acute myocardial infarction. American Journal of Cardiology. 1990;66:271-74.

Shinwell ES, Karplus M, Zmora E, Reich D, Rothschild A, Blazer S, et al. Failure of early postnatal dexamethasone to prevent chronic lung disease in infants with respiratory distress syndrome. Arch Dis Child Fetal Neonatal Ed. 1996 Jan;74(1):F33-7.

Smith LF, Heagerty AM, Bing RF, Barnett DB. Intravenous infusion of magnesium sulphate after acute myocardial infarction: effects on arrhythmias and mortality. Int J Cardiol. 1986 Aug;12(2):175-83.
Sorensen OH, Friis T, Jorgensen AW, Jorgensen MB, Nissen NI. Anticoagulant treatment of acute coronary thrombosis. Acta Med Scand. 1969 Jan-Feb;185(1-2):65-72.

Sorensen PS, Pedersen H, Marquardsen J, Petersson H, Heltberg A, Simonsen N, et al. Acetylsalicylic acid in the prevention of stroke in patients with reversible cerebral ischemic attacks. A Danish cooperative study. Stroke. 1983 Jan-Feb;14(1):15-22.

Super DM, Cartelli NA, Brooks LJ, Lembo RM, Kumar ML. A prospective randomized double-blind study to evaluate the effect of dexamethasone in acute laryngotracheitis. J Pediatr. 1989 Aug;115(2):323-9.

Tal A, Bavilski C, Yohai D, Bearman JE, Gorodischer R, Moses SW. Dexamethasone and salbutamol in the treatment of acute wheezing in infants. Pediatrics. 1983 Jan;71(1):13-8.

Taylor SH, Silke B, Ebbutt A, Sutton GC, Prout BJ, Burley DM. A long-term prevention study with oxprenolol in coronary heart disease. N Engl J Med. 1982 Nov 18;307(21):1293-301.

Thomsen J, Sederberg-Olsen J, Balle V, Vejlsgaard R, Stangerup SE, Bondesson G. Antibiotic treatment of children with secretory otitis media. A randomized, double-blind, placebo-controlled study. Arch Otolaryngol Head Neck Surg. 1989 Apr;115(4):447-51.

Varsano I, Frydman M, Amir J, Alpert G. Single intramuscular dose of ceftriaxone as compared to 7-day amoxicillin therapy for acute otitis media in children. A double-blind clinical trial. Chemotherapy. 1988;34 Suppl 1:39-46.

White HD, Norris RM, Brown MA, Takayama M, Maslowski A, Bass NM, et al. Effect of intravenous streptokinase on left ventricular function and early survival after acute myocardial infarction. N Engl J Med. 1987 Oct 1;317(14):850-5.

Witchitz S, Kolsky H, Moisson P, Chiche P. [Streptokinase and myocardial infarction: can fibrinolysis curb necrosis (author's transl)]. Ann Cardiol Angeiol (Paris). 1977 Jan-Feb;26(1):53-6.

Wright IS, Marple CD, Beck DF. Report of the Committee for the Evaluation of Anticoagulants in the Treatment of Coronary Thrombosis with Myocardial Infarction; a progress report on the statistical analysis of the first 800 cases studied by this committee. Am Heart J. 1948 Dec;36(6):801-15.

Yeh TF, Torre JA, Rastogi A, Anyebuno MA, Pildes RS. Early postnatal dexamethasone therapy in premature infants with severe respiratory distress syndrome: a double-blind, controlled study. J Pediatr. 1990 Aug;117(2 Pt 1):273-82.

Description of Monte Carlo Simulation

The Monte Carlo simulation was carried out by generating individual patient data for each study (trial). Individual patient data was then used to calculate summary statistics for each study, and these are then subjected to meta-analysis. The results of the meta-analyses are then summarized and returned.

An R function, presented below, runs the simulations.

Example: To run a simulation of a meta-analysis with an additive treatment effect variance of 0.28 (population variance) of 0.28, a low quality effect of zero, 216 studies per meta-analysis, an average additive treatment effect for each study of 0.1, 38 patients per arm:

```
atev0.28_lqe0.1 <- runMetaSim(ATEVariance=0.28,
lowQualityEffect=.1, nStudies=20, additiveTreatmentEffect=0.1,
nPatients = 100, nMetas = 1000)
```

This produces the following output:

```
  c(1:nMetas)     noModeratorEst     noModeratorSig
noModeratorI2
 Min.   :   1.0   Min.   :-0.52504   Min.   :0.000   Min.
:78.96
 1st Qu.: 250.8   1st Qu.:-0.22592   1st Qu.:0.000   1st
Qu.:90.72
 Median : 500.5   Median :-0.15397   Median :0.000   Median
:92.50
 Mean   : 500.5   Mean   :-0.14912   Mean   :0.247   Mean
:92.04
 3rd Qu.: 750.2   3rd Qu.:-0.06858   3rd Qu.:0.000   3rd
Qu.:93.90
 Max.   :1000.0   Max.   : 0.22088   Max.   :1.000   Max.
:96.52
 noModeratorTau2   lowQualityModEst   lowQualityModSig
 Min.   :0.07644   Min.   :-0.84262   Min.   :0.000
 1st Qu.:0.20129   1st Qu.:-0.25805   1st Qu.:0.000
 Median :0.25594   Median :-0.10680   Median :0.000
 Mean   :0.26746   Mean   :-0.10629   Mean   :0.094
 3rd Qu.:0.32325   3rd Qu.: 0.04945   3rd Qu.:0.000
 Max.   :0.59768   Max.   : 0.69177   Max.   :1.000
```

noModeratorEst is the treatment effect, the mean treatment effect is -0.099, when the population treatment effect was -0.149. The population ATE for high quality studies was 0.1, however, the effect of low quality studies was also 0.1, and half of the studies were low quality, giving a population mean of 0.15.

noModeratorSig is whether or not the treatment effect was found to be significant in the meta-analysis, the mean of this variable gives the power to detect the effect, hence the power to detect an effect when quality was not taken into account was 0.247.

noModeratorI2 is the I-squared for the analysis with no moderator. The median is 92.5.

noModeratorTau2 is the value for Tau-squared when a moderator is not included. The value here is 0.27.
lowQualityModEst is the estimate of the effect of a low quality study. The mean is -0.107, when the population mean entered into the simulation is 0.10.
lowQualityModEst is the effect of the moderator. The mean here is 0.106, compared with a population value of 0.10.
lowQualityModSig is the significance of the low quality moderator. The power of the study to detect a low quality effect is given by the mean of this variable – in this case, the power to detect an effect of quality is 0.094 (9.4%).

Steps in the analysis

1. For each study, in each meta-analysis, an effect size is generated. In the case of zero treatment effect variance, this effect is equal for all studies. In the case of a simulation with additional heterogeneity, a random normal variable with mean 0 and variance equal to the treatment effect variance is calculated, and this is added to the treatment effect of each study. In this way, the mean treatment effect is unchanged, but the variance of the treatment effect is increased.

2. For each study, in each meta-analysis, control and intervention samples are generated. For each study a normally distributed random variable is created, with mean=0 and sd=1 for the control group. For the intervention group, a normally distributed random variable with mean = treatment effect for that study and sd = 1. These data generated using the rnorm() function, in R BASE.

3. The effect size for each study is generated using escalc(), in the the R metafor package.

4. Studies are grouped into meta-analyses. Each meta-analysis is analyzed twice using rma(), once with low quality as a moderator, and once without. The parameter estimates are stored for both analyses, as is whether or not the result was statistically significant. The I-squared and tau-squared for the analysis with no moderator are stored.

5. The function returns a summary of these stored data.

```
rm(list=ls())
library(metafor)
library(compiler)


            #Define a function to return mean and SD from the same vector.
            meanSD <- function(x) {
                  sampleMean <- mean(x)
                  sampleSD <- sd(x)
                  x <- c(sampleMean, sampleSD)
                  return(x)
                  }
```

```
#First define function and set defaults - defaults are overridden by arguments passed
by function call.

runMetaSim <- function(    nMetas = 1000,
       #Number of meta-analyses to be simulated
                                    nStudies = 100,
       #Number of studies per meta-analysis
                                    nPatients = 100,
       #Number of patients per study
                                    additiveTreatmentEffect = 0.1,          #This
is the additive effect, so it's in log odds, gamma_0
                                    ATEVariance = 1,
       #A random variable to introduce study level heterogeneity - nu_1,
                                    method = "DL" ,
              #Method to combine data, FE for fixed, DL, etc for DerSimonian-Laird, etc
                                    lowQualityEffect= 0.1  ,
       #increase in effect size associated with low
                                    lowQualityRatio = 0.5      )
       #Set proportion of studies that have low quality


  {


        metaChars <- data.frame(c(1:nMetas))          #Create data frame with number of
rows equal to number of simulations


       #Now generate study parameters

       metaStudies <- data.frame(c(1:(nMetas*nStudies)))           #Generate data
frame containing one row for each study in meta-analysis

       metaStudies$metaAnalysisRepetitionNumber <- rep((c(1:nMetas)), nStudies)

       names(metaChars)[1] <- "metaAnalysisRepetitionNumber"

       metaStudies <- metaStudies[ order(metaStudies$metaAnalysisRepetitionNumber) ,]


       metaStudies  <- merge(metaStudies, metaChars, sort=TRUE,
by="metaAnalysisRepetitionNumber")
       metaStudies$rowNumber <- c(0:(nStudies*nMetas-1))
       metaStudies$lowQualityEffect<- lowQualityEffect

       metaStudies$studyNumberWithinMeta <- (metaStudies$rowNumber %% nStudies )  + 1
       metaStudies$lowQuality <- ifelse((metaStudies$studyNumberWithinMeta / nStudies
) > lowQualityRatio, 1, 0)



       #Generate additive treatment effect for each study.
       ##############################

       metaStudies$additiveTreatmentEffect     <- additiveTreatmentEffect
#That's the additive treatment effect
       metaStudies$studyATE <-    rnorm((nMetas*nStudies), mean=0, sd=ATEVariance**0.5
)          #That's where the random effect gets added.  The size of the randome effect

                                                                    #depends on ATE
variance
       metaStudies$additiveTreatmentEffect <- metaStudies$additiveTreatmentEffect +
metaStudies$studyATE
```

```
      metaStudies$additiveTreatmentEffect <- metaStudies$additiveTreatmentEffect +
      metaStudies$lowQuality * lowQualityEffect                          #Low quality
studies have larger effect sizes.

      metaStudies$n <-    metaStudies$n <-  nPatients

      meanNs <- aggregate(metaStudies$n,
by=list(metaStudies$metaAnalysisRepetitionNumber), mean)

      names(meanNs)[1] <- "metaAnalysisRepetitionNumber"
      names(meanNs)[2] <- "studyMeanNs"


      metaStudies <- merge(metaStudies, meanNs, by="metaAnalysisRepetitionNumber")

      metaStudies$n <- round(( nPatients / metaStudies$studyMeanNs * metaStudies$n) ,
digits=0)


      ####################################



      metaStudies$controlMean <- apply(metaStudies, 1, function(x) mean(rnorm(mean=0,
sd=1, n=nPatients)))   #Generate raw data form control group for each study, and find
mean and SD
      metaStudies$controlSD <- apply(metaStudies, 1, function(x) sd(rnorm(mean=0,
sd=1, n=nPatients)))


      metaStudies$interMean <- apply(metaStudies, 1, function(x) mean(rnorm(mean=0,
sd=1, n=nPatients)))  + metaStudies$additiveTreatmentEffect

                    #Generate raw data from treatment group for each study, and find
mean and SD

      metaStudies$interSD <- apply(metaStudies, 1, function(x) sd(rnorm(mean=0, sd=1,
n=nPatients)))

      es <- escalc( m1i =        controlMean,                     #Use escalc()
function, in metafor library, to generate effect size for each study
                        m2i =        interMean,
                        sd1i = controlSD,
                        sd2i = interSD,
                        n1i =  n,
                        n2i =  n,
                        measure =    "SMD",
                        data=metaStudies)

      metaStudies <- cbind(metaStudies, es)
      resultsNoModerator <- as.list(c(1:nMetas))
      resultsWithModerator <- as.list(c(1:nMetas))
      pValuesAndEsts <- WithModerator <- as.data.frame(c(1:nMetas))

      for(loop in c(1:nMetas)) {

            noMods <- with( subset(metaStudies,
metaStudies$metaAnalysisRepetitionNumber==loop),  {   #Run meta-analyses with no
moderator
                  rma(yi, vi, method=method ) }
                )
```

```
              lowQualityMod <- with( subset(metaStudies,
metaStudies$metaAnalysisRepetitionNumber==loop),  { #Run meta-analyses with moderator
                        rma(yi, vi, mods=~lowQuality, method=method ) }
                      )

              #Extract estimates, and significance of estimates, from data.
              pValuesAndEsts$noModeratorEst[loop] <- noMods[[1]]
              pValuesAndEsts$noModeratorSig[loop] <- ifelse( noMods[[4]] < 0.05, 1, 0)
              pValuesAndEsts$noModeratorI2[loop] <- noMods$I2
              pValuesAndEsts$noModeratorTau2[loop] <- noMods$tau2

              pValuesAndEsts$lowQualityModEst[loop] <-  lowQualityMod[[1]][[2]]
              pValuesAndEsts$lowQualityModSig[loop] <- ifelse( lowQualityMod[[4]][[2]]
< 0.05, 1, 0)


       }


       return( summary(pValuesAndEsts ))
}
#########################End of runMetaSim function.



start.t <- proc.time()           #Use proc.time functions to record time taken by
simulation

atev0.28_lqe0.1 <- runMetaSim(ATEVariance=0.28, lowQualityEffect=.1, nStudies=20,
       additiveTreatmentEffect=0.1, nPatients = 100, nMetas = 100)

end.t <- proc.time()
end.t[3] - start.t[3]
start.t <- proc.time()
atev0.28_lqe0.1
```