Some Cochrane risk of bias items are not important in osteoarthritis trials: A meta-epidemiological study based on Cochrane reviews

Julie Bolvig,1,2 Carsten B. Juhl,1,3 Isabelle Boutron,4,5 Peter Tugwell,6,7,8 Elizabeth A. T. Ghogomu,7,8 Jordi Pardo Pardo,7,8 Tamara Rader,7,8 George A. Wells,9 Alain Mayhew,5,10,11 Lara Maxwell,7 Hans Lund,2,12 Henning Bliddal,1 Robin Christensen,1 and The Editorial Board of the Cochrane Musculoskeletal Group.

1: Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksberg, Copenhagen, Denmark.  
2: SEARCH research group, Research Unit of Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark.  
3: Department of Rehabilitation, Copenhagen University Hospital, Herlev and Gentofte  
4: METHODS Team, Centre of Research in Epidemiology and Statistics Sorbonne Paris Cité, UMR 1153, Institut National de la Santé et de la Recherche Médicale (INSERM), University Paris Descartes, Paris, France  
5: Cochrane Methods-Bias Group, Ottawa Hospital Research Institute, Centre for Practice-Changing Research, Ottawa, Canada  
6: Ottawa Hospital Research Institute, Joint Coordinating Editor CMSG  
7: Cochrane Musculoskeletal Group, Ottawa Hospital Research Institute, Centre for Practice-Changing Research, Ottawa, Canada  
8: Centre for Global Health, University of Ottawa  
9: Epidemiology and Community Medicine, University of Ottawa  
10: Knowledge Synthesis Group, Ottawa Hospital Research Institute, Centre for Practice-Changing Research, Ottawa, Canada  
11: C.T. Lamont Primary Health Care Research Centre, Bruyère Research Institute, Ottawa, Canada  
12: Centre for Evidence-Based Practice, Bergen University College, Norway

Correspondence to:  
Robin Christensen, BSc, MSc, PhD; senior biostatistician  
Professor, adj. of clinical epidemiology; Head of Musculoskeletal Statistics Unit,  
The Parker Institute,  
Bispebjerg and Frederiksberg Hospital,  
Nordre Fasanvej 57  
DK-2000 Copenhagen F  
Denmark  
Phone: (+45) 3816 4165  
Fax: (+45) 3816 4159  
E-mail: robin.christensen@regionh.dk

Keywords: osteoarthritis, meta-analysis, meta-epidemiology, risk of bias

Word count (excl. title page, abstract and references): 4059
<table>
<thead>
<tr>
<th>Author</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>JBH</td>
<td><a href="mailto:jbhansen@health.sdu.dk">jbhansen@health.sdu.dk</a></td>
</tr>
<tr>
<td>CBJ</td>
<td><a href="mailto:cjuhl@health.sdu.dk">cjuhl@health.sdu.dk</a></td>
</tr>
<tr>
<td>IB</td>
<td><a href="mailto:isabelle.boutron@htd.aphp.fr">isabelle.boutron@htd.aphp.fr</a></td>
</tr>
<tr>
<td>PT</td>
<td><a href="mailto:Tugwell.BB@uOttawa.ca">Tugwell.BB@uOttawa.ca</a></td>
</tr>
<tr>
<td>EG</td>
<td><a href="mailto:eghogomu@toh.on.ca">eghogomu@toh.on.ca</a></td>
</tr>
<tr>
<td>JPP</td>
<td><a href="mailto:jpardo@ohri.ca">jpardo@ohri.ca</a></td>
</tr>
<tr>
<td>TR</td>
<td><a href="mailto:trader@ohri.ca">trader@ohri.ca</a></td>
</tr>
<tr>
<td>GAW</td>
<td><a href="mailto:gawells@ottawaheart.ca">gawells@ottawaheart.ca</a></td>
</tr>
<tr>
<td>AM</td>
<td><a href="mailto:amayhew@bruyere.org">amayhew@bruyere.org</a></td>
</tr>
<tr>
<td>LM</td>
<td><a href="mailto:lmaxwell@uottawa.ca">lmaxwell@uottawa.ca</a></td>
</tr>
<tr>
<td>HL</td>
<td><a href="mailto:hlund@health.sdu.dk">hlund@health.sdu.dk</a></td>
</tr>
<tr>
<td>HB</td>
<td><a href="mailto:henning.bliddal@regionh.dk">henning.bliddal@regionh.dk</a></td>
</tr>
<tr>
<td>RC</td>
<td><a href="mailto:robin.christensen@regionh.dk">robin.christensen@regionh.dk</a></td>
</tr>
</tbody>
</table>
ABSTRACT

Objective To evaluate the impact of bias-related study characteristics on treatment effects in osteoarthritis (OA) trials.

Study design Based on OA trials included in Cochrane reviews the impact of study characteristics on treatment effect estimates were evaluated. Characteristics included items of the risk of bias tool (RoB), trial size, single vs multi-site, and source of funding. Effect sizes were calculated as standardized mean differences (SMDs). Meta-regression was performed to identify “relevant study-level covariates” that decreases the between-study variance ($\hat{\tau}^2$).

Results Twenty reviews including 126 OA trials with a high degree of heterogeneity was included ($\hat{\tau}^2=0.1247$). Among RoB domains only patient blinding had an impact on the results (reducing heterogeneity according to $\hat{\tau}^2 <7\%$). Inadequate blinding of patients yielded larger effects (SMD$_{\text{Difference}} = 0.15; 95\% \text{ CI}: 0.11 \text{ to } 0.29, P=0.035$). The most important study characteristic was trial size (heterogeneity reduced by 25%), with small trials reporting larger effects (SMD$_{\text{Difference}} = 0.29; 95\% \text{ CI}: 0.16 \text{ to } 0.42, P<0.001$).

Conclusion In musculoskeletal reviews addressing pain, all the items included in the Cochrane risk of bias tool might not be equally important. OA trial results may be affected by bias constructs that are not yet fully elucidated.

Keywords Meta-epidemiology, Meta-Research, Meta-Analysis, bias, osteoarthritis, pain

Prospero (CRD42013006924)
WHAT IS NEW?

What is already known?
- Poor internal validity leading to bias in trials can lead to over- or under estimation of the true intervention effect
- Bias aspects that are considered important for systematic reviews are included in the Cochrane risk of bias tool

What is new?
- The Cochrane risk of bias tool include some domains that are not associated with evidence of bias in estimating treatment effects in OA trials
- OA trials with inappropriate blinding (over-) estimate the effect size by almost 50 percent compared to adequately blinded trials
- For patient–reported outcomes (such as pain), all the items included in the Cochrane risk of bias tool might not be equally important
- Results of OA trials are apparently affected by bias domains and study design characteristics that are not yet included in the Cochrane risk of bias tool (e.g., trial size and single vs multi-site status).
INTRODUCTION

Bias is defined as “a systematic error, or deviation from the truth, in results or inferences”(1). Poor internal validity in trials can lead to overestimation or underestimation of the intervention effect(2). Regardless of the tools used to assess risk of bias, the methods summarizing potential bias and incorporating bias assessments into meta-analyses vary greatly(3-6). Bias associated with particular characteristics of studies may be examined using meta-epidemiology, which analyze a cluster of meta-analyses where the influence of the trial characteristics—such as the judgments of risk of bias on treatment effects estimates—are explored(7-10). By using a meta-epidemiologic approach, it is possible to examine the association of specific trial characteristics in a collection of meta-analyses based on the included trials and the reported effect size(9;10).

Much research has been conducted over the years to understand bias in trials and how it may influence the results of systematic reviews(11). The Cochrane Collaboration’s methods groups developed the Cochrane risk of bias tool for assessing RCTs, using empirical validation based on the methodological contributions of meta-epidemiological studies (1;6). This tool has since been updated in 2011 and is soon to be updated again and is a widely used tool to assess RoB domains for RCTs; risk of bias assessment using Cochrane risk of bias tool is a mandatory component of Cochrane Reviews(6;12). Risk of bias identified within the domains from the Cochrane risk of bias tool (i.e. selection bias, performance bias, detection bias, attrition bias, and reporting bias), if indeed present, might “overestimate” or change the direction of the effect in favor of an experimental treatment. This might as well apply for single versus multi-site trials, trials size, and source of funding(9;13;14).

Following research from Nuesch et al showed that many of these bias items are important in osteoarthritis (OA), where most trials use pain as the primary outcome(15;16). We wanted to explore (and confirm) the findings of Nuesch et al, by replicating their meta-epidemiological approach on another dataset based on the data currently available in the Cochrane Library, to see whether the same applied. Further we were interested in exploring the importance of the many risk of bias domains (and items) in determining the quality of systematic reviews.
Our objective was to evaluate the impact of different study characteristics including both well-recognized domains and potentially new bias items on estimates of treatment effects in meta-analyses of interventions applied in OA trials.

**METHODS**

The protocol for this study (17) was registered on PROSPERO (CRD42013006924) and the study conforms to the ‘Preferred Reporting Items for Systematic Reviews and Meta-Analyses’ (PRISMA) statement (18).

**Eligibility criteria**

Only Cochrane Reviews with meta-analyses of randomized or controlled trials in patients with OA published in the Cochrane Library were included. Meta-analyses were only included if they reported results from a patient-reported pain outcome for any intervention compared with sham, placebo, or no intervention control (19). Moreover, only trials where patients in both the intervention and the control group received the same treatment except for an “add-on” in the intervention group were eligible for inclusion. A single reviewer evaluated the eligibility of the meta-analysis. Sufficient trial data was available from the report of the meta-analyses to perform all the analyses. There was no language restriction.

**Search for meta-analyses and inclusion of trials**

In the Cochrane Database of Systematic Reviews, eligible trials were identified from published Cochrane Reviews (i.e., meta-analyses) after a thorough search, using a combination of keywords and text words related to "osteoarthritis" (17). The most recent version of the Cochrane Review was used.

**Risk of bias in individual studies**

The RoB within each full-text trial was assessed using the domains of the risk of bias tool as recommended by the Cochrane Collaboration (6), which comprise methods for sequence generation
and maintaining allocation concealment, blinding, and management of incomplete outcome data. Each item was rated as high risk, low risk, or unclear risk of bias. The following additional bias sources were also assessed: a] single versus multi-site trials (20), b] small vs. large trials(13), and c] source of funding(21).

A trial was considered a multi-site trial if more than one site was involved. In case of missing information, the trial was classified as multi-site if it reported several ethics committees. If the report stated both a single ethics committee and a single author affiliation, the trial was classified as a single site. Information about single- and multi-site trials was extracted from the text, statements, author’s affiliations, and acknowledgement in every included trial.

To determine whether a trial was small or large, a threshold of 128 participants was used: If the total number of randomized participants was less than 128 patients, the study was regarded as a small trial; consequently trials with ≥128 participants were referred to as large trials. The threshold of 64 participants in each group corresponds to a reasonable statistical power (80%) to detect a standardized mean difference (SMD) ≥ 0.5(22).

Trials were specified either as being funded by for-profit, non-profit, or unclear source of funding. Non-profit funding included money received from both non-profit organizations (e.g., internal institutional funding and governmental funding) and not funded trials. For-profit organizations were defined as companies that might acquire financial gains or losses depending on the outcome of the trial. Further, trials with a mix of for-profit and non-profit were considered as for-profit funded. Deficient information about the source of funding was reported as unclear. Funding was defined as including provision of human resources (authorship, statistical analysis, or other assistance), study materials (drug, placebo, assay kits, or similar materials), or grants(21). Sources of funding were extracted from the text, statements of sources of support, authors’ affiliations, acknowledgments, and trial registration, if available. Two reviewers (JB and CBJ) independently completed all the risk of bias assessment Disagreements were solved by discussion or by contacting a third reviewer (RC).

**Data collection process and data items**

A systematic, standardized data extraction approach was used to gather information from all the eligible OA studies. For systematic reviews, the review ID, author, year of publication, and accumulated trial size combined in the meta-analysis (i.e. total number [N_{Total}] number in
intervention group \([N_I]\), and number in control group \([N_C]\) were abstracted. The type of intervention was categorized into non-pharmacological modalities (NP), pharmacological modalities (P) and surgical modalities (S) according to the "suggested sequential, pyramid approach to management of OA" by Dieppe & Lohmander(23), the Osteoarthritis Research Society International (OARSI) recommendations(24) and the definitions from Cochrane Musculoskeletal Group (25). The primary outcome (pain) was abstracted from overall pain reported in the Cochrane Review’s first pain measure forest plot. The outcome was collected at the time point closest to the 12 weeks follow-up. At the trial level, studies were assigned a trial ID and information on the name of the author, year of publication, type of pain measure, type of intervention, and type of OA condition was extracted. From the forest plot available in the included reviews, mean values \((m_{\text{Intervention}}\text{ and } m_{\text{Control}})\), standard deviations \((SD_{\text{Intervention}}\text{ and } SD_{\text{Control}})\) and size of the trials were extracted. In meta-analyses where only SMDs and confidence intervals (CI) were reported, these were extracted in order to obtain an effect size. Further, we assessed the following characteristics from the trials: attrition rate during the trial \((a_{\text{Total}}, a_{\text{Intervention}}, \text{ and } a_{\text{Control}})\), and trial duration (weeks). Finally, ‘treatment’ provided to the comparator groups was extracted, whether it was placebo, sham, waiting-list or nothing.

**Data synthesis**

Treatment effect sizes (ES) were expressed as SMDs by dividing the difference in mean values available from the forest plots by the standard deviation (SD) reported. In studies where means were not given, SMDs and confidence intervals (CIs) were used to calculate the overall ES. Positive ES indicated a beneficial effect of the experimental intervention (i.e. pain reduction). We noted whether the data was expressed as values from follow-up or changes from baseline(26).

**Statistical analysis**

To empirically assess the effect of risk of bias assessment on the overall treatment effects in reviews, stratified analyses across trials according to all the different extracted trial characteristics were performed and P-values for interaction between trial characteristics and treatment effect were derived. All the statistical meta-regression analyses were based on mixed-effects Restricted Maximum Likelihood (REML) models (i.e., where both random and fixed factors
apply) (27;28); computations were developed and analyzed using Stata Statistical Software: Release 14.2. College Station, TX: StataCorp LP. The ES according to the RoB assessment for each RoB item was estimated by using a random effects meta-analysis model fully allowing for heterogeneity, to measure the variability in bias estimates, expressed as an estimate of \( \tau^2 \) (i.e., \( \tau^2 \)). In principle within each review, the ES for trials according to the different RoB assessments is estimated by using a random effects meta-analysis model. The next level is then the level of systematic review which is included as a fixed factor. From this basic “meta-meta-analysis” model(7) we are able to explore whether different study level characteristics can reduce heterogeneity as a consequence of being included as a covariate in the model. A “relevant study-level covariate” was defined as one that decreases the between-study variance (\( \tau^2 \), estimated as Tau-squared [\( \tau^2 \)]) as a consequence of inclusion in the mixed-effects statistical model(29). All P-values and 95% confidence intervals (95% CI) were two sided.

RESULTS

Characteristics

The search was carried out on 31 January 2014. As illustrated in Figure 1, 78 relevant Cochrane Reviews were identified. Twenty meta-analyses(30-49), including 126 trials, corresponding to 140 intervention comparisons in 19,052 (n=10,580 for intervention groups and n=8,472 for control groups) patients were included. Table 1 describes the included comparisons. Data was expressed as values from follow up in one review(48), for another review data was mixed final- and change from baseline values(35) and for 18 reviews data was expressed as change from baseline values. The median number of studies included per meta-analysis was 5 trials (range 2 to 32) and the median number of patients per meta-analysis was 611 (109 to 3,600). The individual SMD’s derived from the included trials ranged from -5.33(50) to -0.73(51). Eight (30;36;37;39-41;44;45) of the included meta-analyses assessed efficacy of non-pharmacological interventions, ten (31-35;38;42;46-48) assessed efficacy of pharmacological interventions and two meta-analyses (43;49) assessed the efficacy of surgical interventions. For 91 (65%) of the randomized comparisons, we were able to derive the mean scores from the intervention- and control group followed by the corresponding SD.
In the remaining comparisons we derived the SMD and corresponding standard error (SE). The 140 comparisons were categorized into three treatment groups; non-pharmacological modalities (NP) 76 (54%), pharmacological modalities (P) 54 (39%) and surgical modalities (S) 10 (7%).

The OA condition examined was primarily knee 65 (46%), and a mix of different conditions (e.g., knee and hip) 68 (49%). OA of the hip was examined in 4 (3%) and other conditions (e.g., thumb) in 3 (2%).

**Estimates of treatment effects**

The overall analysis of change in pain showed that there was an effect size of 0.42 (95% CI, 0.34 to 0.49), indicating moderate all-in-all effect of interventions for pain reduction in OA. However as anticipated from the very heterogeneous group of trials (in terms of very different pain interventions and population) included, the overall meta-analysis revealed a large heterogeneity ($I^2 = 0.01247$) indicating a high between-study heterogeneity (7,52).

The stratified analyses shown in Table 2 presents the estimates for all assessed bias items, including trial size, single versus multi-site and source of funding. These analyses show that specific bias items could reduce the between study variability seen across the OA trials, with $I^2$ varying according to different covariates included in the meta-regression models. The largest reduction in heterogeneity was found in the ‘trial size’ variable. This yielded a reduction in $I^2$ at 24.86%, supported by a statistically significant P-value ($P < 0.001$) for interaction between small and large trials ($ES_{large} = 0.26$, 95%CI [0.16 to 0.35] vs. $ES_{small} = 0.55$, 95%CI [0.46 to 0.64]); i.e., small studies revealed an overestimated effect size (relative difference 112%). ‘Blinding of patients’ reduced the heterogeneity with respectively 7.30%, $P = 0.035$ with a difference between adequately performed patient blinding ($ES = 0.38$, 95%CI [0.30 to 0.46]) and inadequately/unclear performed patient blinding ($ES = 0.50$, 95%CI [0.37 to 0.63]) revealed an overestimated effect size (relative difference 32%). Further, ‘single versus multi-site’ reduced the heterogeneity with 1.92% , $P = 0.018$ with a difference in effect sizes between multi-site ($ES = 0.33$, 95%CI [0.24 to 0.43]) and single site ($ES = 0.50$, 95%CI [0.40 to 0.61]).

There was no significant change of $I^2$ in the association between effect size and adequate or inadequate/unclear performed for the following: ‘generation of sequence’,...
‘concealment of allocation’, ‘personnel blinding’, ‘handling of incomplete outcome data’ or ‘funding’.

As a post hoc analysis we also tested whether the influence of ‘trial size’ (reducing the between-trial heterogeneity) was independent of any of the existing RoB items. From these secondary multivariable meta-regression analyses, we found that both ‘blinding of participants’ (reducing $\chi^2$ further from 0.108 to 0.09154) and ‘blinding of caregivers’ (reducing $\chi^2$ to 0.1014) could further explain (i.e. reduce) the heterogeneity between trials. Furthermore concealment of allocation reduces ‘concealment of allocation’ reduced heterogeneity slightly (reducing $\chi^2$ to 0.1073). All other RoB items (and thus domains), could not reduce heterogeneity more than already observed from ‘Trial size’ (see Appendix 4).

**DISCUSSION**

Overall the effect of interventions included in the Cochrane Reviews presented a high level of between-trial heterogeneity. The most important study characteristic reducing heterogeneity was trials considered small (<128 participants in total) compared with trials considered large. Further, the following study characteristics (in prioritized order) reduced heterogeneity: ‘Patients blinding’ and ‘Single versus multi-site’.

There was no significant change of $\chi^2$ in the association between effect size and adequate or inadequate/unclear performed for the following: ‘generation of sequence’, ‘concealment of allocation’, ‘personnel blinding’, ‘handling of incomplete outcome data’ or ‘funding’. Furthermore, adding ‘blinding of participants’ or ‘personnel blinding’ as a covariate to the analysis of ‘Trial size’ could further reduce (i.e. explain) the heterogeneity between trials. All other RoB items (and thus domains), could not reduce heterogeneity more than already observed from ‘Trial size’.

**Comparison with other studies**

These results are consistent with findings of previous meta-epidemiological studies(13;15;16;20;53-55). Nüesch et al. (2010) found that an average difference in effect sizes between large and small
trials with more beneficial effects found in small trials(15). Our finding of overestimated effect size in small trials compared to large trials is in agreement. This difference might be caused by the fact, that large trials tend to be of higher methodological quality than small trials and that the observed association between sample size and treatment effect could be confounded by methodological quality(56,57). The overestimated effect estimates from the small studies could be caused by publication bias. Small studies showing insignificant results are less likely to be published than statistically significant small studies and large studies with insignificant results(58). Herbert and Bø (2005) pointed out an interesting debate of the quality of the interventions and its effect on the results(59). We know the importance of assessing the methodological quality of the trials before including them into a systematic review – but the quality of the intervention as well should be assessed if the systematic reviews examine complex interventions. In this systematic review we examined complex interventions as well, but we did not assess the intervention quality(59). We know that small trials tend to perform more efficient interventions and in better teams meaning a difference in the intervention between large and small studies – this together with the fact that small trials often are in limitation of a “likely limited power” to detect statistically significant changes(60) could as well explain the overestimated effect in small trials compared to large trials. Further, results of smaller studies are subject to greater sampling variation and hence are less precise – this is also called imprecision which is one of the key criteria in GRADE (Grading of Recommendations Assessment, Development and Evaluation).

Hrobjartsson et al. (2011) found studies with non-blinded patients showed larger effect estimates than studies with blinded patients. This is in agreement with our findings. The overestimated effect could be caused by the fact that pain is very sensitive to the lack of blinding among patients, as it is patient reported(16). Non-blinded patients report symptoms differently from the blinded patients because of the psychological effect of being treated with an intervention versus being treated with nothing (i.e. placebo) (61). Further, studies have shown that patients tends to report their symptoms in a manner they think could please the investigator(53) – this could result in patients in the intervention group giving overestimated report of pain relief because the investigator seems to consider and care about the patients’ health and welfare(53).

In accordance with previous results, Dechartres et al (2011)(20) and Bafeta et al (2012)(55) found treatment effects on average were more beneficial in single site trials, than in multi-site trials which is in agreement with our findings. We found larger difference in effect
estimates between single versus multi-site trials. The difference between multi- and single site effect sizes might be due to ‘small study effects’, meaning the tendency for smaller studies in a meta-analysis to show larger intervention effects and because of the non-publication of studies with negative or small effects(15). Further, larger effect size in single site trials could be caused by a selected population and a higher expertise of teams(20;55;62).

Lundh et al (2012) found that industry-sponsored drug and medical device studies are more often favourable to the sponsor’s products than non-industry-sponsored studies (14). Our findings are not in agreement with the findings of Lundh et al. Unlike Lundh et al (2012), the results of this study are based on the reported information only. Potentially this could lead to commercially funded foundation being taken for non-profit and potentially mask the overestimated effect of for-profit bias. Further we did not consider the role of the for-profit sponsor in the design conduct, analyses and report. Further research should look into this topic.

**Strengths and limitations of the study**

Our study has limitations. Firstly, our study was conducted on the basis of published data only. Because of possible risk of publication bias, our results may be conservative, as the unpublished trials tend to have higher risk of bias(63). The selection of component trials was based on the literature included in meta-analysis from the Cochrane Library. Even though Cochrane Reviews are known for the quality of their searches – each one across several databases, our included studies are only those identified in those literature searches, which we had no control over. The meta-analyses included in this study, however, are probably representative why our results may be generalizable. Furthermore, in some cases, we included more than one comparison from a single trial. This might affect our results, as risk of bias assessment of these trials will count twice.

We used meta-analyses to explore the influence of the study characteristics on treatment effect estimates of OA trials. In theory, aggregation of data from multiple trials should enhance the accuracy and precision of any pooled result. Unfortunately data combination requires a leap of faith: it presumes that the differences among studies are primarily due to chance. In fact, it is not unknown that differences in the direction or size of treatment effects may be caused by other factors, such as methodological quality of the included trials(64). Still, review authors seem reluctant or fail to incorporate the risk of bias assessment in their analysis and conclusions(65), compromising the trustworthiness of results from meta-analysis derived from these reviews. Although the Cochrane Collaboration has improved the way we assess bias in included trials(11)
owing to Ken Schulz’s work on bias assessment(66), the Cochrane risk of bias tool in 2008(1), AMSTAR (Assessing the methodological quality of systematic reviews checklist)(67), the Cochrane Editorial requirement (MECIR standard)(12) and GRADE’(68), meta-analyses might still over- or underestimate the effect estimates from biased studies. Thus, meta-analyses may generate misleading results by ignoring meaningful heterogeneity among studies, entrenching the biases in individual studies, and introducing further biases through the process of finding studies and selecting results to be pooled.

We performed post hoc sensitivity analyses to explore whether our findings were robust; we found that four (of the eight) bias items resulted in the apparently logical trend, where highest ES’ are observed in the ‘inadequate’ category whereas the ‘adequate’ corresponded to the lowest ES (i.e., with the ‘unclear’ category being at the intermediate level). For the remaining four bias items (Blinding of patients, Blinding of personnel, Incomplete outcome data and Funding) the ‘unclear’ category presented the highest ES. The dichotomization of Trial size could for good reasons be criticized as being arbitrary as others have suggested the use of at least 100 in each group to imply a “large trial”. For the purpose of transparency we now presented a scatter plot (i.e. Bobble plot) illustrating the association between trial size and the corresponding effect size (appendix 5).

**Future research**

Future research should confirm our study’s recommendations of including trial size and single versus multi-site as new items for the next update of a risk of bias tool. Further, studies that examine whether a multiple-meta-regression could provide the true effect size for treatment of pain in OA trials would be very interesting to conduct. As well, it could be interesting to examine the interactions between the bias items and if these could lead to an additive effect between some of the items and through that influence the effect estimates. As our data was collected from twenty smaller Cochrane Reviews it was not possible for us to run analyses adjusting for the other risk of bias domains. Due to the limited number of included trials we might lack power to run mixed effect regression models (69;70). Future research should investigate the same objective on data from one large review as that would have the power to run adjusted analyses and test for interaction between RoB domains.
Our study looked into already known bias domains and further included three new items. There are multiple other potential bias domains that could have been investigated and discussed in this study, e.g. ‘year of publication’ (71).

This study examined the consequence of not blinding the patients. But we did not stratify for interventions that actually could be blinded from those where blinding was impossible (e.g. physical training or educational programs). This would be interesting to research, to see whether there is a difference between possible blinding and impossible blinding. The reason for the lack of transparency in trials should be explored or a tool to assess the meaning of unclear methodology should be prepared.

**Conclusions**

Results of OA trials may be affected by other domains than those already included in the Cochrane risk of bias tool (i.e. trial size and single versus multi-site). According to this meta-epidemiological study, bias associated with study design characteristics such as trial size, blinding of patients and single versus multi-site may lead to overestimation of intervention effect estimates and increases the between-trial heterogeneity in trials reporting subjectively assessed outcomes.

**ACKNOWLEDGEMENTS**

We thank the Parker Institute, Frederiksberg Hospital, and The Oak Foundation for creating the possibility and right settings for this research. Additionally, we thank the Cochrane Musculoskeletal Group Editorial and the Cochrane Methods – Bias Group, both based at the Ottawa Hospital Research Institute.

**FUNDING**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This research was funded by The Oak Foundation, which had no role in study
design, data collection, data synthesis, data interpretation, writing the report, or the decision to submit the manuscript for publication. The Cochrane Musculoskeletal Group is grateful for the financial support of the Canadian Institutes of Health Research (CIHR), the University of Ottawa, the National Health and Medical Research Council in Australia (NHMRC), and the Cabrini Institute, Melbourne, Australia.

**COMPETING INTERESTS**

This study had no financial competing interests. The Parker Institute's Musculoskeletal Statistics Unit is grateful for the financial support received from public and private foundations, companies, and private individuals over the years. The Parker Institute is supported by a core grant from the Oak Foundation; The Oak Foundation is a group of philanthropic organizations that, since its establishment in 1983, has given grants to not-for-profit organizations around the world. All the authors are involved with different health care initiatives and research (including Cochrane, OMERACT, and the GRADE Working Group) that could benefit from wide uptake of this publication. Isabelle Boutron is Co-convenor of the Cochrane Methods Bias group.
REFERENCE LIST


Ref Type: Online Source


FIGURE LEGENDS

Figure 1
M = identified Cochrane Reviews; M* = Possible eligible Cochrane Reviews; m = included meta-analysis; K = Trials from included Cochrane Reviews; k = Trials included in meta-analysis; k* = Intervention comparisons.
Table 1 – Characteristics of the included comparisons

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Comparison, k*</th>
<th>OA condition</th>
<th>Duration, weeks</th>
<th>Sequence Generation</th>
<th>Allocation Concealment</th>
<th>Blinding, participants</th>
<th>Blinding, caregivers</th>
<th>Outcome data</th>
<th>Trial size (&lt;128)</th>
<th>Single-site status</th>
<th>Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Knee</td>
<td>Hip</td>
<td>Mix</td>
<td>Other</td>
<td>n&lt;sub&gt;total&lt;/sub&gt;</td>
<td>n&lt;sub&gt;i&lt;/sub&gt;</td>
<td>n&lt;sub&gt;c&lt;/sub&gt;</td>
<td>A</td>
<td>U/I</td>
<td>A</td>
<td>U/I</td>
</tr>
<tr>
<td>NP</td>
<td>76</td>
<td>56</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td>7932</td>
<td>4451</td>
<td>3481</td>
<td>8</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>(54)</td>
<td>(86)</td>
<td>(100)</td>
<td>(24)</td>
<td>(0)</td>
<td>(42)</td>
<td>(42)</td>
<td>(41)</td>
<td>(6;12)</td>
<td>(57)</td>
<td>(43)</td>
</tr>
<tr>
<td>P</td>
<td>54</td>
<td>2</td>
<td>0</td>
<td>52</td>
<td>0</td>
<td>10405</td>
<td>5764</td>
<td>4641</td>
<td>8</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>(39)</td>
<td>(3)</td>
<td>(0)</td>
<td>(76)</td>
<td>(0)</td>
<td>(55)</td>
<td>(54)</td>
<td>(55)</td>
<td>[4;13]</td>
<td>(51)</td>
<td>(48)</td>
</tr>
<tr>
<td>S</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>715</td>
<td>365</td>
<td>350</td>
<td>13</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(0)</td>
<td>(0)</td>
<td>(100)</td>
<td>(4)</td>
<td>(3)</td>
<td>(4)</td>
<td>[12;42]</td>
<td>(50)</td>
<td>(50)</td>
<td>(40)</td>
</tr>
<tr>
<td>Total</td>
<td>140</td>
<td>65</td>
<td>4</td>
<td>68</td>
<td>3</td>
<td>19052</td>
<td>10580</td>
<td>8472</td>
<td>8</td>
<td>76</td>
<td>64</td>
</tr>
</tbody>
</table>

No.(%); Median [IQ], NP = Non-pharmacological modalities; P = Pharmacological modalities; S = Surgical modalities; n<sub>total</sub> = total number of participants; n<sub>i</sub> = number of participants in intervention group; n<sub>c</sub> = number of participants in control group; A = Adequate; I = Inadequate; U= Unclear; NF = Non-profit funding; FF = For-profit funding.
Table 2 - Results of the stratified Meta-analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Trials*</th>
<th>ES</th>
<th>95% CI</th>
<th>$\tau^2$</th>
<th>P-value for interaction</th>
<th>Relative difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>140</td>
<td>0.42</td>
<td>(0.34 to 0.49)</td>
<td>0.1247</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Sequence generation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>76</td>
<td>0.40</td>
<td>(0.30 to 0.49)</td>
<td>0.1268</td>
<td>0.504</td>
<td>1.68%</td>
</tr>
<tr>
<td>Inadequate/Unclear</td>
<td>64</td>
<td>0.44</td>
<td>(0.34 to 0.55)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>62</td>
<td>0.35</td>
<td>(0.24 to 0.45)</td>
<td>0.1218</td>
<td>0.066</td>
<td>-2.33%</td>
</tr>
<tr>
<td>Inadequate/Unclear</td>
<td>78</td>
<td>0.48</td>
<td>(0.38 to 0.57)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding (Patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>75</td>
<td>0.35</td>
<td>(0.26 to 0.44)</td>
<td>0.1156</td>
<td>0.035</td>
<td>-7.30%</td>
</tr>
<tr>
<td>Inadequate/Unclear</td>
<td>65</td>
<td>0.50</td>
<td>(0.39 to 0.60)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinding (Personnel)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>97</td>
<td>0.38</td>
<td>(0.30 to 0.46)</td>
<td>0.1240</td>
<td>0.125</td>
<td>-0.56%</td>
</tr>
<tr>
<td>Inadequate/Unclear</td>
<td>43</td>
<td>0.50</td>
<td>(0.37 to 0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete outcome data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate</td>
<td>103</td>
<td>0.38</td>
<td>(0.46 to 0.30)</td>
<td>0.1199</td>
<td>0.096</td>
<td>-3.85%</td>
</tr>
<tr>
<td>Inadequate/Unclear</td>
<td>37</td>
<td>0.52</td>
<td>(0.38 to 0.67)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>53</td>
<td>0.26</td>
<td>(0.16 to 0.35)</td>
<td>0.0937</td>
<td>&lt;0.001</td>
<td>-24.86%</td>
</tr>
<tr>
<td>Small</td>
<td>87</td>
<td>0.55</td>
<td>(0.46 to 0.64)</td>
<td></td>
<td></td>
<td>112%</td>
</tr>
<tr>
<td>Single- vs multi-site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi</td>
<td>60</td>
<td>0.33</td>
<td>(0.24 to 0.43)</td>
<td>0.1223</td>
<td>0.018</td>
<td>-1.92%</td>
</tr>
<tr>
<td>Single</td>
<td>80</td>
<td>0.50</td>
<td>(0.40 to 0.61)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-profit</td>
<td>58</td>
<td>0.36</td>
<td>(0.25 to 0.47)</td>
<td>0.1269</td>
<td>0.197</td>
<td>1.76%</td>
</tr>
<tr>
<td>For-profit/Unclear</td>
<td>82</td>
<td>0.46</td>
<td>(0.36 to 0.55)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trials*= number of comparisons; ES = effect size; $\tau^2$ = tau-squared (between-study variance); 95% CI = 95% confidence interval; Relative difference = relative difference in effect size between inadequate/unclear and adequate performed trial (%).
Reviews identified and reviewed on basis of title and abstract $M = 78$

Reviews excluded based on title and abstract $M = 41$
- Not RCTs = 1
- Meta-analysis with <2 trials = 1
- No meta-analysis = 1
- No treatment for OA = 35
- Withdrawals = 3

Reviews retrieved for full text review $M^* = 37$

Reviews excluded based on full text review $M^* = 17$
- Not RCTs = 1
- Not sham, placebo, or no-intervention control = 3
- No patient reported outcome = 1
- No meta-analysis from OA overall pain = 8
- Meta-analysis with <2 trials = 3
- Other = 1

Eligible meta-analysis $m = 20$

Trials from meta-analysis $K = 136$

Trials excluded $K = 10$
- Duplicates of same publication = 2
- No full text available = 8

Trials included in meta-epidemiological analysis $k = 126$
Included intervention comparisons $k^* = 140$

Figure 1 - Flow chart