Bias and variation in meta-analyses of randomised clinical trials

Graduate School for Cellular and Biomedical Sciences

University of Bern

PhD Thesis

Submitted by

Eveline Bettina Nüesch

from Balgach SG

Thesis advisors

Prof. Dr. Peter Jüni and Dr. Sven Trelle
Institute of Social and Preventive Medicine
Faculty of Medicine of the University of Bern

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

Table of contents

Abstract

Decisions about the use of medical interventions are frequently based on results from meta-analyses of randomised clinical trials. Inadequate methodology in randomised trials, selective reporting of trials and results, and data extraction from published reports, however, may bias results of meta-analyses.

The objective of this thesis was to examine factors associated with bias and variation in meta-analyses of randomised clinical trials. These factors include methodological characteristics at the level of individual trials and problems with data extraction at the level of meta-analyses. This thesis is based on series of meta-epidemiological studies in component trials of different meta-analyses and on two clinical examples.

The first three articles report on a meta-epidemiological study including 21 meta-analyses with 190 trials that compared therapeutic interventions with placebo or non-intervention control in patients with hip or knee osteoarthritis and used patient reported pain as an outcome. Treatment effects tended to be less beneficial in trials with adequate allocation concealment, adequate blinding of patients, intention-to-treat analysis and in large trials with at least 100 patients per arm. Article 4 provides a clinical example of a systematic review and meta-analysis illustrating how dimensions of methodological quality of included trials and small study effects affect pooled results. Articles 5 and 6 are based on randomly selected Cochrane reviews that presented a result as a standardised mean difference, and the corresponding protocols and component trials. Frequently, trials reported multiple intervention groups, multiple time points and outcome data from multiple measurement scales. Multiplicity of data in trial reports had an impact on the agreement between different observers when extracting data from trial reports and on the variability of the pooled results in meta-analyses. Article 7 discusses how different methodological approaches including funnel plots, stratified analyses accompanied by interaction tests and heterogeneity-adjusted trial sequential analysis contribute to our understanding of bias and inconclusive results in meta-analyses.

This thesis suggests that flaws in the conduct and design of randomised trials and meta-analyses frequently result in biased estimates of treatment benefits and that the extent and direction of bias might be unpredictable for a specific situation. Methodological characteristics of randomised trials (allocation concealment, patient blinding and intention-to-treat analysis), small sample size, inter-observer variation related to data extraction and multiplicity of data in trial reports may affect results and impact on the validity of meta-

analyses. To avoid potential bias, trialists should ensure adequate concealment of allocation, attempt blinding of patients and describe results from intention-to-treat analyses. In systematic reviews and meta-analyses, a detailed protocol and independent data extraction by at least two investigators might improve the validity of results. Results of meta-analyses of methodologically questionable trials should be distrusted. Small study effects should be examined and the influence of inadequate allocation concealment, patient blinding and exclusions from the analysis should be routinely assessed.

Introduction

Meta-analyses of randomised clinical trials

Synthesis of relevant evidence, which has accumulated over time, is essential for informed decision making about the use of medical interventions by clinicians, researchers, and policy makers. Randomised clinical trials are generally considered the best study design for obtaining evidence of effectiveness and safety of medical interventions. Systematic reviews and meta-analyses are scientific rigorous approaches to synthesise the available evidence from randomised trials. Systematic reviews, a review that has been prepared using a documented and systematic approach, allow a more objective appraisal of the evidence than narrative reviews or expert opinions.[1] Meta-analyses, statistical combinations of several randomised trials, may enhance precision of estimated treatment effects and allow the exploration of bias and confounding as an explanation of variation between trials.

To combine results from individual trials in a meta-analysis, the treatment effects from individual studies need to be expressed as appropriate effect estimates. Estimates of treatment effects are usually expressed as risk ratios, odds ratios or risk differences for trials with a binary outcome, as rate ratios for trials with count data, as differences in means or standardised mean differences for trials with a continuous outcome and as hazard ratios for trials with a time-to-event outcome.[2] The primary focus of this thesis is on outcomes measured on continuous or numerical rating scales. Therefore, effect estimates for continuous outcomes are presented in more detail. In the i^{th} trial, the mean response in the experimental group of size n_{Ei} is denoted m_{Ei} with standard deviation SD_{Ei}, and the mean response in the control group of size n_{Ci} is denoted m_{Ci} with standard deviation SD_{Ci}. If the outcome in all individual trials is measured on the same scale, the treatment effect can be expressed as the difference in means between experimental and control groups in the i^{th} trial denoted as

(1) $m_{Ei} - m_{Ci}$

with variance $SD_{Ei}^2/n_{Ei} + SD_{Ci}^2/n_{Ci}$

If the outcome in individual trials is measured on different scales, however, it is necessary to express the treatment effects on a uniform scale. The calculation of effect sizes (or standardised mean differences) allows that estimates of treatment effects are expressed in a standardised way as standard deviation units.[2 3] Several methods of standardisation are available. The most frequently calculated effect size is Cohen's d, where the difference in means is divided by the pooled standard deviation SD_{pooled}:[4]

(2) $d = (m_E - m_C) / SD_{pooled}$

$$SD_{pooled}^2 = ((n_E - 1)\, SD_E^2 + (n_C - 1)\, SD_C^2) / (n_E + n_C - 2)$$

The variance of Cohen's d is calculated as follows:[3 4]

(3) $(n_E + n_C)/(n_E * n_C) + d^2/(2*(n_E + n_C - 2))$

An effect size of 0.20 standard deviation units can be considered a small difference between experimental and control group, an effect size of 0.50 a moderate difference, and 0.80 a large difference.[4] Other standardisation methods include Hedges' g, which accounts for underestimations of standard deviations in small studies,[5] and Glass' Δ, which uses the standard deviation of the control group for the standardisation rather than the pooled standard deviation.[6] The different methods of standardisation will produce similar results unless the sample is very small or the standard deviations vary substantially between the groups.[3]

In meta-analyses, effect estimates from individual trials i (i = 1,…,k), denoted as $\hat{y}_i$, are combined across k trials to calculate a pooled estimate of the treatment benefit $\hat{y}$. Generally, the pooled effect estimate is a weighted average of the effect estimates from each individual trial $\hat{y}_i$ with weights based on the variances $var(\hat{y}_i) = \sigma_i^2$ of these estimates.[7] Two approaches to combine effect estimates from individual trials exist: fixed-(or common) effect and random-effects meta-analysis.[3 8] Fixed-effect meta-analysis usually uses inverse-variance weights $w_i = 1/\sigma_i^2$, which gives a pooled effect estimate

(4) $\hat{y}_{fixed} = \Sigma\, w_i\, \hat{y}_i / \Sigma\, w_i$

Fixed-effect meta-analytic models assume a common (fixed) effect that underlies each trial and that the true variances σ_i^2 are known, although in practice they are estimated from the data. Random-effects meta-analytic models, however, assume that effects from individual studies are from a common normal distribution with an overall average treatment effect y and between-trial variance τ^2: $y_i \sim N(y, \tau^2)$.[9] An estimate of the between-trial variance τ^2 is incorporated into the weights: $w_i = 1/(\sigma_i^2 + \tau^2)$.[9-11] This implies that if $\tau^2 > 0$, random-effects weights $w_i = 1/(\sigma_i^2 + \tau^2)$ will be smaller and more similar across trials than fixed-effect weights $w_i = 1/\sigma_i^2$.[3] Therefore, smaller and less precise trials will receive more relative weight in random-effects meta-analyses than in fixed-effect meta-analyses. The variance of the pooled effect estimate $\hat{y}$ is given by Σw_i. Confidence intervals for $\hat{y}$ and z-statistics to test the hypothesis of no difference between groups can be derived.[3]

Generally, results of meta-analyses are presented in forest plots, where the pooled estimates with their corresponding confidence intervals are presented. Figure 1 shows an example of a forest plot from a meta-analysis comparing opioids with placebo or no control intervention in patients with knee or hip osteoarthritis.[12] Treatment effects of 10 included randomised trials are shown one below the other using red squares for the estimated standardised mean difference and horizontal lines for the corresponding 95% confidence intervals. The size of the red square symbol is inversely related to the variance of the individual trial. Diamonds denote pooled estimated effects with the corresponding 95% confidence intervals derived from inverse-variance random effects meta-analysis. This example shows that opioid treatment has a moderate effect on pain compared to placebo with a pooled standardised mean difference of -0.36 (95% CI -0.47 to -0.26). The plot is stratified according to type of opioids and display pooled standardised mean differences derived from random effects meta-analysis between different types of opioids and placebo separately.

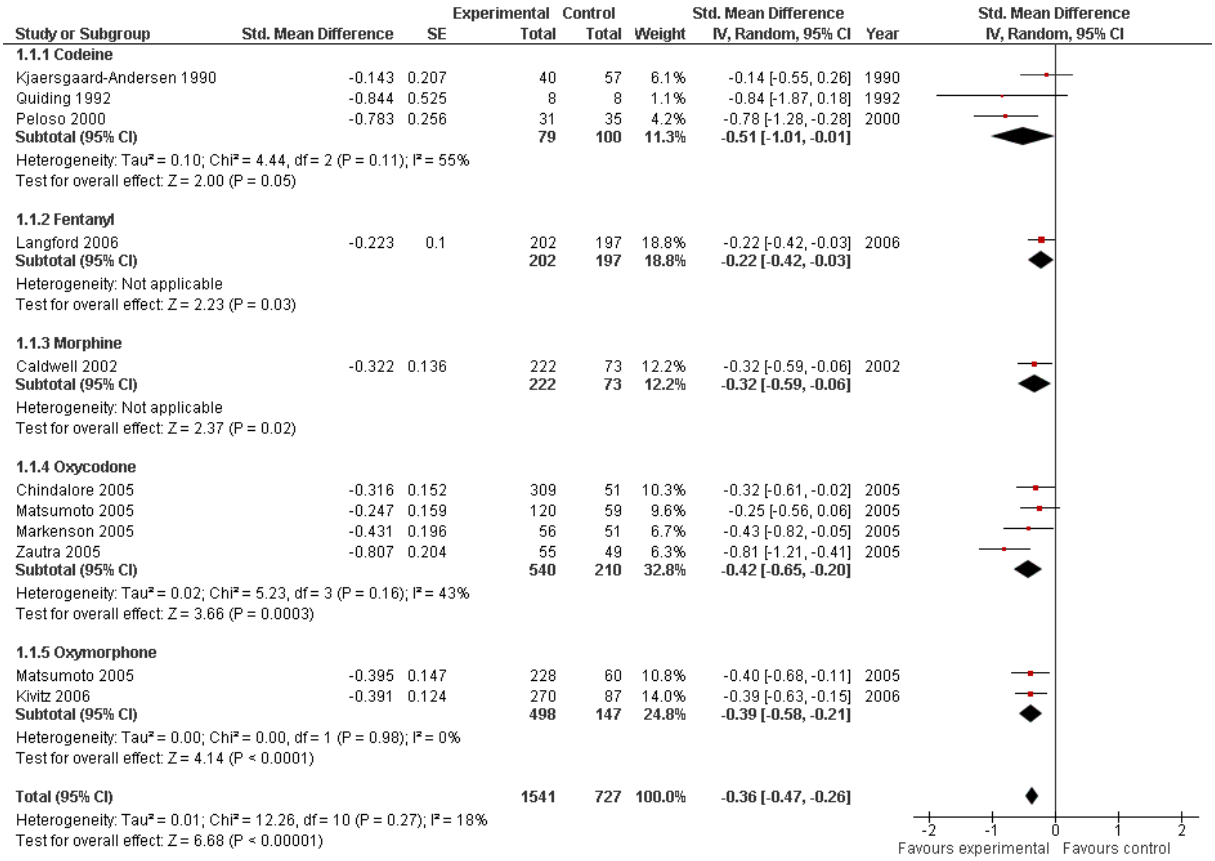

Figure 1: Forest plot of 10 trials comparing the effects of any type of opioids and control (placebo or no intervention) on knee or hip pain. Values on x-axis denote standardised mean differences. The plot is stratified according to type of opioids.[12]

Forest plots are a tool to visually examine the variation between trials. More elaborated methods for visual inspection of the presence of heterogeneity between trials have been proposed by Galbraith[13] and L'Abbé.[14] Statistical heterogeneity can be assessed using hypothesis tests of homogeneity of the $\hat{y}_i$. The most commonly used test is based on the Cochran's Q statistic:[15]

$$(5) \qquad Q = \Sigma\, w_i\, (\hat{y}_i - \hat{y}_{fixed})^2$$

that follows approximately a χ^2 distribution with k – 1 degrees of freedom under the null hypothesis of homogeneity.[10 16] The weights w_i are inverse-variance weights. However, quantification of between-trial heterogeneity is generally more informative than a formal test of homogeneity. The test has often inappropriate power and makes it difficult to decide whether heterogeneity is present and clinically meaningful.[16 17]

An important objective of random-effects meta-analysis is the estimation of the between-trial variance τ^2 including its uncertainty. The commonly used estimate of the variance τ^2 is based on the method of moments originally proposed for binary data by DerSimonian and Laird.[10] Several other methods for the estimation of τ^2 are available, including other non-iterative, maximum-likelihood based or Bayes estimation procedures.[18-20] Standard random-effects meta-analyses ignore the imprecision in the τ^2 estimate.[9 17] Special maximum likelihood methods or Bayesian approaches allow for the imprecision of τ^2 estimates and reflect more accurately the uncertainty in the between-trial heterogeneity variance.[9 21 22] A variance estimate $\tau^2 < 0.06$ might be interpreted as low and $\tau^2 \geq 0.06$ as high heterogeneity between trials in a random-effects meta-analysis. A τ^2 of 0.06 corresponds to a difference between smallest and largest effect sizes of about 1 standard deviation unit.[22] Other measures to quantify between-trial heterogeneity are also available.[17] A widely used measure is the I^2 statistic, which describes the percentage of variation across trials that is attributable to heterogeneity rather than to sampling error. I^2 can be derived from Cochran's Q statistic and approximate confidence intervals can be calculated.[17 23] In the absence of heterogeneity I^2 equals 0. I^2 values of 25%, 50%, and 75% may be interpreted as low, moderate, and high between trial heterogeneity, although its interpretation depends on the size and number of trials included.[24]

The variation between trials might arise from differences in studied populations, interventions, or outcomes and from differences in methodology and design between included trials. Analyses stratified by characteristics of included trials accompanied by appropriate interaction tests can be used to explore sources of between-trial heterogeneity. Meta-

regression analyses are widely used to explore sources of heterogeneity in meta-analyses:[25 26] weighted regression models to estimate the effect of trial-level covariates x_i on the overall effect, where trials with lower standard errors contribute with higher weights to the regression analysis. Random-effects meta-regression analyses are based on hierarchical models with normal error terms μ_i and corresponding variances σ_i^2 that vary between trials, and error terms ε_i with corresponding variance τ^2 assumed equal between trials.

(6) $$y_i = x_i\beta + \mu_i + \varepsilon_i$$

$$\mu_i \sim N(0,\sigma_i^2)$$

$$\varepsilon_i \sim N(0,\tau^2)$$

It is unlikely that the covariates included in the meta-regression models explain all of the variation between trials and therefore, meta-regression models must allow for residual heterogeneity.[26 27] Figure 2 shows an example of a random-effects meta-regression analysis to explore the dose-response relationship between estimates of treatment effects on pain and daily morphine equivalence doses in trials comparing opioids with control interventions in patients with knee or hip osteoarthritis.[12] The example shows no clear association between treatment benefit and daily morphine dose; predicted treatment effects from random-effects meta-regression do not vary according daily morphine dose.

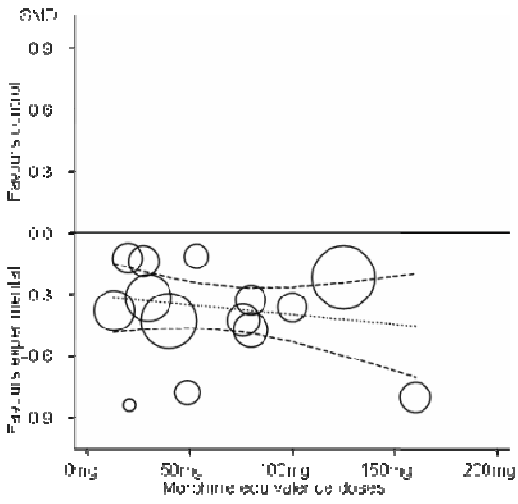

Figure 2: Standardised mean differences of knee or hip pain (y-axis) are plotted against total daily dose of morphine equivalents (x-axis). The size of the circles is proportional to the random-effects weights that were used in the meta-regression. The dotted line indicates predicted treatment effects (regression line) from univariable meta-regression by using daily morphine equivalence doses the explanatory variable, and dashed lines represent the 95% confidence intervals.[12]

The exploration of sources of heterogeneity between trials using meta-regression will be limited by the availability of data from trial reports and by the number of included trials in the meta-analysis. Meta-regression analyses are based on aggregate level data and may be affected by the 'ecological fallacy', if the relationship between the covariates and treatment effects are not the same within and across trials.[26] Ecological fallacy can be interpreted as

failure of associations seen at trial level to correspond to associations at patient level.[28] Meta-analyses based on individual patient data rather than on aggregate level data avoid these problems. Finally, a naïve use of meta-regression models to determine the association between estimated treatment effects and underlying risk can produce artefacts due to measurement error in the covariate and regression to the mean.[25]

Bayesian approaches to random-effects meta-analysis and meta-regression assuming exchangeability of underlying effects are based on hierarchical Bayesian models and allow to obtain inferences for the parameters of interest from their posterior distributions simulated using Markov Chain Monte Carlo methods.[22 29] Meta-analyses methods so far presented only considered pair-wise comparisons of treatments. However, clinicians, researchers and policy makers may be rather interested in the comparison of multiple treatments for decision making. In recent years, network meta-analyses for mixed treatment comparisons have been developed that allow comparison of several treatments while fully preserving the randomised treatment comparisons within trials.[30 31] An overview of Bayesian approaches to classical and network meta-analysis is provided elsewhere in more detail.[29 32]

Bias in randomised trials and meta-analyses

The objective of any randomised trial or meta-analysis is to obtain a precise and valid estimate of the effect of an intervention on the outcome of interest. Reducing random error in the data (sampling error) will result in more precise estimates of intervention effects, while avoiding systematic error or bias will result in valid estimates. Internal validity refers to methodological rigour of a randomised trial or meta-analysis, whereas external validity refers to the extent to which results can be generalised to other circumstances.[33] The focus of this thesis is on internal validity of randomised trials and meta-analyses. Bias might be defined as "any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth",[34] or more formally, the bias b_θ of an estimator T for the parameter θ is

(7) $$b_\theta (T) = E_\theta (T) - \theta,$$

where $E_\theta (T)$ denotes the expected value of the estimator T. Thus, we estimate θ without bias if $E_\theta (T) = \theta$. Different authors use different terminology for biases occurring in clinical trials. Throughout this thesis the following classification of bias will be used, which is explained below in more detail: selection bias (at study entry), performance bias, detection bias and attrition bias. In randomised trials, bias can occur at any stage during

the trial.[33] Figure 3 shows different stages of a randomised trial and the measures that can be taken to minimise bias at these different stages.

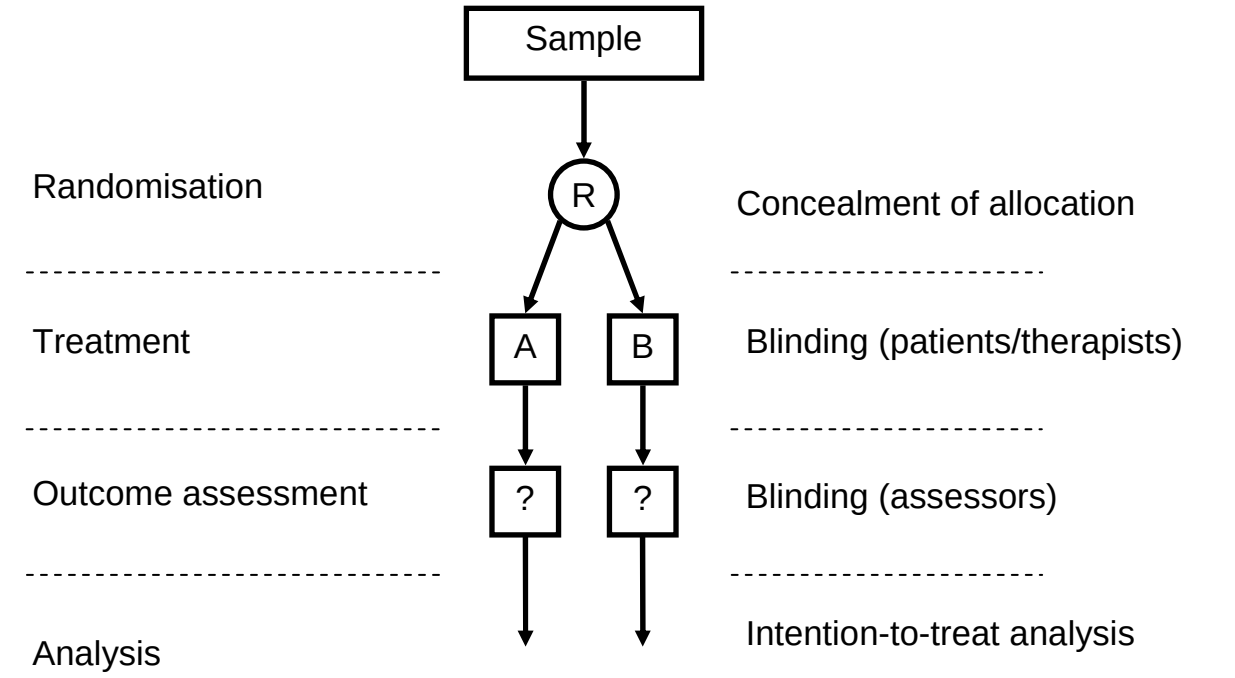

Figure 3: Measures to minimise bias at different stages of a randomised trial.

Selection bias in randomised trials arises from biased allocation of patients to comparison groups, i.e. if not all patients have the same probability to be assigned to either group. Minimisation of selection bias is obtained by two interrelated steps: The generation of a random allocation sequence and the concealment of this sequence to prevent personnel assigning patients to treatment groups from foreseeing allocation. A random allocation sequence can be generated for example by coin tossing, dice throwing or by using a computer algorithm.[33] Concealment of allocation is adequate if the investigators responsible for patient selection and inclusion are unable to know before allocation which treatment was next, for example, central randomisation; the use of sequentially numbered, sealed, and opaque assignment envelopes; or coded drug packs.[33] Performance bias results from an unequal provision of care apart from treatments under evaluation. Procedures that prevent therapists and patients from knowing which intervention was received, i.e. blinding of therapists and patients will minimise this type of bias. Blinding can be achieved by using identically looking placebo tablets or by using sham interventions that are indistinguishable from the experimental interventions. Detection bias results from biased assessment of outcomes, and will be minimised by procedures that prevent outcome assessors from knowing which intervention was received, i.e. blinding of outcome assessors. Outcome assessors can be

blinded if they are independent (not therapists) or – in the case of therapist- or patient-reported outcomes – if interventions are indistinguishable. Attrition bias will result from biased occurrence and handling of deviations from protocol and losses to follow-up. This type of bias will be minimised by procedures that prevent exclusion of randomised patients from the analyses and minimise protocol deviations and losses to follow-up. Adequate analyses include all randomised patients in the group they were originally allocated to, regardless of their adherence to the study protocol according to the intention-to-treat principle, avoiding any selective exclusion of patients after randomisation.[33 35]

Meta-analyses results can be distorted by biases in individual randomised trials as described above, but also by bias across trials. For example, trials that do not show a significant benefit of the experimental intervention are less likely to be published than trials that found a significant result (publication bias).[36 37] Bias can result from incomplete and selective reporting of outcomes[38-40] (outcome reporting bias). Failure to publish non-significant results may distort meta-analyses results:[41] publication and outcome reporting biases can result in overly optimistic estimates of treatment benefits, and can affect the precision of effect estimates from random-effects meta-analysis.[42 43]

Other potential threats to the validity meta-analyses based on aggregate data relate to extraction of data from trial reports and multiplicity of data in trial reports. As meta-analyses are usually based on data that have already been processed, interpreted, and summarised by other researchers, data extraction can be complicated and can lead to important errors.[44] There is often a multiplicity of data in trial reports that makes it difficult to decide which ones to use in a meta-analysis. Furthermore, data are often incompletely reported,[38 44] which makes it necessary to perform calculations or impute missing data, such as missing standard deviations. Different observers may get different results, but previous studies on observer variation have not been informative, because of few observers, few trials, or few data.[45 46]

Empirical evidence of bias

The presence of different types of bias in randomised trials has been empirically assessed in several studies. A classical study was published by Schulz et al, who examined dimensions of methodological quality associated with estimated treatment effects in 250 randomised trials from 33 meta-analyses in pregnancy and childbirth.[47] They found that estimates of treatment effects were larger in trials were associated with reporting of allocation concealment and double-blinding, but found no clear associations with generation of allocation sequence and exclusions from the analysis.[47] Until now, several additional studies have been published that

examined associations between allocation concealment and estimated treatment benefits in binary outcomes.[48-53] On average, trials with adequate allocation concealment showed more beneficial effects compared to trials without adequate concealment.[52] However, the different studies employed slightly different definitions of adequate allocation concealment, included trials from different clinical areas and the results varied between different studies and the associations were less pronounced in more recent studies.[52]

The association between blinding and estimated treatment benefits has been assessed in several studies using binary outcomes.[47-53] Results in individual studies were less clear, but in a meta-analysis of all published meta-epidemiological studies, estimated treatment effects were more beneficial in trials with double-blinding compared to trials without double blinding.[52] A pilot study including 35 randomised trials examined the associations between allocation concealment or blinding with estimated treatment effects in non-binary outcomes, but was underpowered to obtain conclusive results.[54] To my knowledge, a single methodological study nested in a randomised controlled trial compared the effect of blinded and unblinded outcome assessments on apparent treatment benefits in a randomised trial and found that unblinded physicians were more likely to provide a favourable rating of the outcome in experimental group as compared with the control group.[55]

The association of withdrawals, dropouts, and exclusions after randomisation with estimated treatment effects has been explored in four meta-epidemiological studies of binary outcomes.[47 49 50 53 56] Direction and magnitude of attrition bias varied between different studies according to different methods and definitions used and different clinical areas addressed: attrition bias may result in both overestimation and underestimation of treatment effects, and its magnitude is difficult to predict.[50 56 57]

Several studies reported evidence of within-trial selective reporting of outcomes. A comparison between trial protocols and published reports showed that reporting of outcomes is frequently incomplete, i.e. with insufficient information to be included in a meta-analysis and that statistically significant outcomes are more likely to be reported in trial publications.[38-40] Trials published in languages other than English tend to be of lower quality and show more beneficial treatment effects than trials published in English.[51 58] Broad literature searches and inclusion of unpublished trials and trials published in languages other than English will increase precision and may minimise bias.[51] In addition, trials with more beneficial results are more likely to be published as full reports and the time to publication is shorter than for trials with less beneficial results.[37 59 60] Meta-analyses including published

trials only, are therefore likely to overestimate treatment benefits. Funnel plots can be used as a tool to identify small study effects. Funnel plot asymmetry can arise from publication and reporting bias, poor methodological quality of smaller trials or true underlying heterogeneity due to differences in intensity of interventions or differences in characteristics of included patients between larger and smaller trials.[61-63] Figure 4 shows an example of a funnel plot, where there is evidence of small study effects.[64] Funnel plots are based on the fact that precision of an estimated treatment effect will increase as the sample size of component trials increases. Results from small trials with large standard errors will scatter widely at the bottom of a plot of treatment effect against standard error, with the spread narrowing among larger trials. In the absence of small study effects the plot will resemble a symmetrical inverted funnel. Conversely, if small study effects are present, funnel plots will often be asymmetrical.[61] The plot can be enhanced by lines of the predicted treatment effect from meta-regression using that standard error as explanatory variable,[65] a regression test of asymmetry[61 62] and contours that divide the plot into areas of statistical significance and non-significance.[66 67]

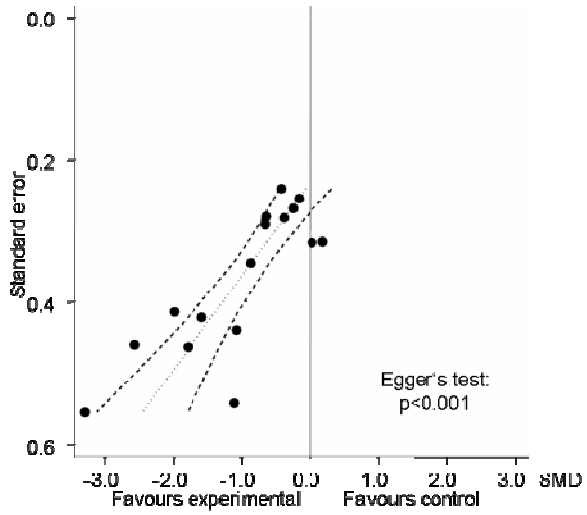

Figure 4: Funnel plot for effects on knee pain showing evidence of small study effects. Numbers on x-axis refer to standardised mean differences, on y-axis to standard errors of standardised mean differences.[64]

Based on the presented empirical evidence of bias, the Cochrane Collaboration recommends assessing risk of bias in included trials which involves the assessment and presentation of individual components, such as allocation concealment, blinding, incomplete outcome data and selective reporting.[2]

The presence and extent of different types of bias might be related to the type of outcome assessed. In general, randomised controlled trials using subjective outcomes are more susceptible to bias than trials using objective outcomes such as overall mortality. In a combined analysis of data from three meta-epidemiological studies of binary outcomes across

different medical fields, the authors recently found that overestimations of treatment benefits were more pronounced for subjective outcomes as compared with objective outcomes such as overall mortality.[68] To my knowledge, no studies so far have systematically assessed different types of bias in meta-analyses using patient-reported outcomes measured on continuous or rating scale.

Meta-epidemiological research

Meta-epidemiological studies examine the association of specific characteristics of a trial, such as concealment of allocation or blinding of patients, with estimated treatment effects in a collection of meta-analyses and their component trials.[69 70] An early example of a meta-epidemiological study using binary outcomes was published by Schulz et al (see above).[47] Different meta-meta-analytic approaches to the analysis of meta-epidemiological studies for binary outcomes expressed as odds ratios are presented in detail by Sterne and co-authors.[70] A general approach to the analysis of meta-epidemiological studies is to calculate effect estimates (ratios of odds ratios or differences in effect sizes) within each meta-analysis. These effect estimates are then combined across meta-analyses using fixed or random effects meta-analysis.[70] In a first step, effect sizes are calculated separately for trials with adequate methodology and trials with inadequate methodology using standard random-effects models within each meta-analysis. This step yields an estimate of bias in each meta-analysis. In a second step, differences in effect sizes between trials with and without adequate methodology from individual meta-analyses are pooled across meta-analyses using fixed- or random-effects models. If the effect of a trial characteristic varies between meta-analyses, then analyses based on the fixed-effect approach will underestimate the uncertainty in estimated differences in effect sizes.[70]

Meta-epidemiological studies are observational studies by design and results might therefore be affected by confounding. For example, trials with adequate allocation concealment may be more likely to use adequate blinding methods and the association between allocation concealment and estimated treatment benefits may be confounded by blinding. One approach is to control effect estimates for confounding factors separately in each meta-analysis by stratification in analogy to Mantel-Haenszel procedures or with random-effects meta-regression and then combine these across meta-analyses as described above. This approach allows the confounding effect to vary across meta-analyses.[70 71] In meta-regression models potential confounders can be added as regression terms, or interactions can be modelled to allow the average bias to vary according to another characteristic (e.g. type of outcome).[72]

References

1. Egger M, Davey Smith G, Altman D. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publishing Group 2001.
2. Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.2 [updated September 2009] The Cochrane Collaboration, 2008.
3. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analsis. In: Egger M, Davey Smith G, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publishing Group, 2001.
4. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum, 1988.
5. Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges LV, editors. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.
6. Glass GV. Primary, secondary and meta-analysis of research. *Educat Res* 1976;5:3-8.
7. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* 1991;10(11):1665-77.
8. Sutton AJ, Higgins JP. Recent developments in meta-analysis. *Stat Med* 2008;27(5):625-50.
9. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172(1):137-159.
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.
11. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;18(20):2693-708.
12. Nuesch E, Rutjes AW, Husni E, Welch V, Juni P. Oral or transdermal opioids for osteoarthritis of the knee or hip. *Cochrane Database Syst Rev* 2009(4):CD003115.
13. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889-94.
14. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;107:224-33.
15. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10:101-29.

16. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998;17(8):841-56.
17. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21(11):1539-58.
18. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996;15(6):619-29.
19. Knapp G, Biggerstaff BJ, Hartung J. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biom J* 2006;48(2):271-85.
20. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med* 2007;26(1):37-52.
21. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med* 1997;16(7):753-68.
22. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: Joh Wiley & Sons Ltd, 2004.
23. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.
24. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
25. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;16(23):2741-58.
26. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21(11):1559-73.
27. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995;14(4):395-411.
28. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008.
29. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;10(4):277-303.
30. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002;21(16):2313-24.
31. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;23(20):3105-24.
32. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005;331(7521):897-900.

33. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.
34. Sackett DL. Bias in analytic research. *J Chron Dis* 1979;32:51-63.
35. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999;319(7211):670-4.
36. Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316(7124):61-6.
37. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009(1):MR000006.
38. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291(20):2457-65.
39. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171(7):735-40.
40. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330(7494):753.
41. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000;320(7249):1574-7.
42. Jackson D. The implications of publication bias for meta-analysis' other parameter. *Stat Med* 2006;25(17):2911-21.
43. Jackson D. Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analysis. *Biometrics* 2007;63(1):187-93.
44. Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 2007;298(4):430-7.
45. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol* 2005;58(7):741-2.
46. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol* 2006;59(7):697-703.
47. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.

48. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999;354(9193):1896-900.
49. 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.
50. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287(22):2973-82.
51. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. 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.
52. Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 2007;36(4):847-57.
53. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2008;149(3):219.
54. Fenwick J, Needleman IG, Moles DR. The effect of bias on the magnitude of clinical outcomes in periodontology: a pilot study. *J Clin Periodontol* 2008;35(9):775-82.
55. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 1994;44(1):16-20.
56. Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. *Int J Epidemiol* 2005;34(1):79-87.
57. Juni P, Egger M. Commentary: Empirical evidence of attrition bias in clinical trials. *Int J Epidemiol* 2005;34(1):87-8.
58. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *Int J Epidemiol* 2002;31(1):115-23.
59. Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. *Cochrane Database Syst Rev* 2007(2):MR000011.
60. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev* 2007(2):MR000010.

61. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34.
62. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;25(20):3443-57.
63. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology* 2000;53(11):1119-1129.
64. Rutjes AW, Nuesch E, Sterchi R, Kalichman L, Hendriks E, Osiri M, et al. Transcutaneous electrostimulation for osteoarthritis of the knee. *Cochrane Database Syst Rev* 2009(4):CD002823.
65. Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, Sterne JA, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 2005;366(9487):726-32.
66. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol* 2008;61(10):991-6.
67. Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ* 2009;339:b2981.
68. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, 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.
69. Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. *BMJ* 1997;315(7109):617-9.
70. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
71. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959;22:719-748.
72. Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. *Stat Med* 2007;26(14):2745-58.

Objective

The overall objective of this thesis was to examine factors associated with bias and variation in meta-analyses of randomised clinical trials. These factors include methodological characteristics at the level of individual trials and problems with data extraction at the level of meta-analyses.

Specifically, the objectives were to examine:

- Whether excluding patients from the analysis of randomised trials is associated with biased estimates of treatment effects and higher heterogeneity between trials. (Article 1)
- The association of adequate allocation concealment and patient blinding with estimates of treatment benefits in osteoarthritis trials. (Article 2)
- The presence and extent of small study effects in clinical osteoarthritis research. (Article 3)
- Whether potential variation between trials can be explained by biases affecting individual trials or by publication bias in a meta-analysis comparing transcutaneous electrostimulation with sham or no specific intervention in patients with knee osteoarthritis. (Article 4)
- The inter-observer variation related to extraction of continuous and numerical rating scale data from trial reports for use in meta-analyses. (Article 5)
- The scope for multiplicity in a sample of meta-analyses using the standardised mean difference as an effect measure and the impact of the multiplicity on the results. (Article 6)
- How different methodological approaches such as funnel plots, stratified analyses accompanied by interaction tests and heterogeneity-adjusted trial sequential analysis contribute to our understanding of bias and inconclusive results in meta-analyses. (Article 7)

Article 1

The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study

Eveline Nüesch,[1,2] Sven Trelle,[1,2] Stephan Reichenbach,[1,3] Anne W.S. Rutjes,[1,4] Elizabeth Bürgi,[5] Martin Scherer,[6,7] Douglas G. Altman,[8] Peter Jüni[1,2]

From [1]Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2]CTU Bern, Bern University Hospital, Switzerland; [3]Department of Rheumatology, Immunology and Allergology Bern University Hospital, Switzerland; [4]Department of Clinical Pharmacology and Epidemiology Consorzio Mario Negri Sud, Santa Maria Imbaro, Chieti, Italy; [5]Department of Internal Medicine, Bern University Hospital, Switzerland; [6]Department of General Practice, University of Göttingen, Germany; [7]Institute of Social Medicine, University of Lübeck, Germany; [8]Centre for Statistics in Medicine, University of Oxford, United Kingdom

BMJ 2009 Sep 07; 339: b3244

Abstract

Objective To examine whether excluding patients from the analysis of randomised trials is associated with biased estimates of treatment effects and higher heterogeneity between trials.

Design Meta-epidemiological study based on a collection of meta-analyses of randomised trials.

Data sources 14 meta-analyses including 167 trials that compared therapeutic interventions with placebo or non intervention control in patients with osteoarthritis of the hip or knee and used patient reported pain as an outcome.

Methods Effect sizes were calculated from differences in means of pain intensity between groups at the end of follow-up, divided by the pooled standard deviation. Trials were combined by using random effects metaanalysis. Estimates of treatment effects were compared between trials with and trials without exclusions from the analysis, and the impact of restricting meta-analyses to trials without exclusions was assessed.

Results 39 trials (23%) had included all patients in the analysis. In 128 trials (77%) some patients were excluded from the analysis. Effect sizes from trials with exclusions tended to be more beneficial than those from trials without exclusions (difference −0.13, 95% confidence interval −0.29 to 0.04). However, estimates of bias between individual meta-analyses varied considerably (τ^2=0.07). Tests of interaction between exclusions from the analysis and estimates of treatment effects were positive in five meta-analyses. Stratified analyses indicated that differences in effect sizes between trials with and trials without exclusions were more pronounced in meta-analyses with high between trial heterogeneity, in meta-analyses with large estimated treatment benefits, and in meta-analyses of complementary medicine. Restriction of meta-analyses to trials without exclusions resulted in smaller estimated treatment benefits, larger P values, and considerable decreases in between trial heterogeneity.

Conclusion Excluding patients from the analysis in randomised trials often results in biased estimates of treatment effects, but the extent and direction of bias is unpredictable. Results from intention to treat analyses should always be described in reports of randomised trials. In systematic reviews, the influence of exclusions from the analysis on estimated treatment effects should routinely be assessed.

Introduction

In clinical trials, deviations from protocol and losses to follow-up often lead to the exclusion of some randomised patients from the analysis.[1 2] Patients excluded after randomisation are unlikely to be representative of patients remaining in the trial. For example, patients may not be available for follow-up because they have an acute exacerbation of their condition or severe side effects,[3] and patients with protocol deviations may have a worse prognosis than those adhering to the protocol.[4] The selective occurrence and biased handling of protocol deviations and losses to follow-up may lead to results that differ systematically from the true values. This is generally referred to as attrition bias.[2] To ensure that intervention and control groups are comparable and to prevent attrition bias, all randomised patients should be included in the analysis and kept in their original groups, regardless of their adherence to the study protocol. In other words, the analysis should be done according to the intention to treat principle, avoiding any selective exclusion of patients after randomisation.[2 5]

Meta-epidemiological studies examine the association of specific characteristics of a trial, such as concealment of allocation or blinding of patients, with estimated treatment effects in a collection of meta-analyses and their component trials.[6 7] The association of withdrawals, dropouts, and exclusions after randomisation with estimated treatment effects has been explored in four meta-epidemiological studies of binary outcomes.[1 6 8-10] The direction and magnitude of attrition bias varied between different studies according to different methods and definitions used and different clinical areas addressed: attrition bias may result in both overestimation and underestimation of treatment effects, and its magnitude is difficult to predict.[1 9 11 12] In general, randomised controlled trials using subjective outcomes are more susceptible to bias than trials using objective outcomes such as overall mortality. A recent study found that bias associated with the lack of allocation concealment and lack of blinding was restricted to trials using subjectively assessed outcomes.[12] In trials of osteoarthritis, treatment effects are often evaluated using subjective outcomes, such as intensity of pain or disability measured on visual analogue, numerical rating, or Likert scales, whereas objective binary outcomes, such as mortality, are addressed rarely. Meta-analyses of osteoarthritis trials may therefore be particularly prone to attrition bias associated with exclusions of patients from the analysis.[2]

We carried out a meta-epidemiological study to assess the impact of attrition bias in meta-analyses of non-binary patient reported outcomes, such as pain intensity. We examined whether excluding patients from the analysis were associated with biased estimates of

treatment effects and with increased heterogeneity between trials in meta-analyses of interventions used for the treatment of pain in osteoarthritis.

Methods

Search and selection of meta-analyses and component trials

We searched the Cochrane Library, Medline, Embase, and CINAHL using a combination of keywords and text words related to osteoarthritis. These were combined with validated filters for systematic reviews and meta-analyses.[13] The last update was carried out on 20 November 2007 (see web extra appendix table 1 for details of search strategy).

We included meta-analyses of randomised or quasi-randomised trials in patients with osteoarthritis of the knee or hip. Trials using an unpredictable allocation sequence were considered as randomised, trials using potentially predictable allocation mechanisms, such as alternation or the allocation of patients according to their date of birth, were considered as quasi-randomised. Meta-analyses were eligible if they assessed patient reported pain comparing any intervention with placebo, sham, or a non-intervention control. Two reviewers independently evaluated the reports for eligibility, and disagreements were resolved by discussion. If necessary a third reviewer was consulted to reach consensus. Reports of all component trials were obtained, and no language restrictions were applied.

Data collection and quality assessment

Two reviewers used a standardised form to independently extract data from the original reports of individual trials on interventions, funding, year of publication, publication language, design, study size, blinding of patients, losses to follow-up, exclusions, handling of missing data, and results. When necessary we approximated means and measures of dispersion from figures. For crossover trials we extracted data from the first period only.[14] Disagreements were resolved by discussion with a third reviewer and subsequent consensus.

Trials were classified to have had no exclusions of patients from the analysis if there was an explicit statement that all randomised patients were included in the analysis of the outcome we extracted or if the reported numbers of patients randomised and analysed on this outcome were identical. We classified trials to have had exclusions if they explicitly reported exclusions from the analysis, if the number of patients analysed was lower than the number of patients randomised, or if it was unclear whether exclusions from the analysis had occurred. Concealment of treatment allocation was considered adequate if the investigators responsible for patient inclusion were unable to suspect before allocation which treatment was next—

central randomisation or the use of sequentially numbered, sealed, and opaque randomisation envelopes was deemed adequate, for example. Blinding of patients was considered adequate if experimental and control interventions were described as indistinguishable or if a double dummy technique was used.[2]

Outcome measures

The primary outcome was patient reported pain. If different pain outcomes were reported we extracted one outcome per study according to a hierarchy described previously.[15 16] If more then one time point was reported we extracted the latest time point up to three months after the end of treatment for potentially structure modifying agents and up to 12 months after the end of treatment for behaviour changing interventions. For all other interventions we extracted the outcome at the end of treatment.

Statistical analysis

We expressed treatment effects as effect sizes by dividing the difference in mean values at the end of follow-up by the pooled standard deviation. Negative effect sizes indicate a beneficial effect of the experimental intervention. If some required data were unavailable we used approximations as previously described.[16] If a trial was included in more than one meta-analysis we inflated standard errors to avoid double counting of patients—for example, if the control group of a trial with three arms was included in two different metaanalyses, we inflated the standard error of the estimate for the control group by $\sqrt{2}$. We used standard inverse variance random effects meta-analyses to combine effect sizes across trials and calculated the variance estimate τ^2 as a measure of heterogeneity.[17]

Within each meta-analysis we used random effects meta-analysis to estimate effect sizes separately for trials with and trials without exclusions of patients from the analysis. Then we derived differences between estimates from trials with exclusions and trials without exclusions for each meta-analysis and combined these differences using random effects meta-analysis, which fully accounted for the variability in bias between meta-analyses.[7] A negative difference in effect sizes indicates that trials with exclusions show a more beneficial treatment effect. Meta-analyses including only trials with exclusions or only trials without exclusions did not contribute to the analysis. Formal tests of interaction between exclusions from the analysis and estimated treatment benefits were done separately for each meta-analysis based on z scores for estimated differences in effect sizes between trials with and trials without exclusions and the corresponding standard errors. We carried out stratified analyses accompanied by interaction tests according to the following characteristics: between trial

heterogeneity in the overall meta-analysis (low, $\tau^2<0.06$, v high, $\tau^2\geq0.06$), treatment benefit in the overall meta-analysis (small, effect sizes >−0.5, v large, effect sizes ≤−0.5),[15 18] and type of intervention assessed in the meta-analysis (drug v other interventions, conventional v complementary medicine). A τ^2 of 0.06 corresponds to a difference between smallest and largest effect sizes of about 1 standard deviation unit.[19] To control confounding by concealment of allocation and by patient blinding, we used stratification by these factors to derive differences between trials with and trials without exclusions adjusted for concealment of allocation and adjusted for patient blinding.

Finally, we compared pooled effect sizes, between trial heterogeneity, precision defined as the inverse of the standard error, and P values for pooled effect sizes between overall random effects meta-analyses including all trials and restricted meta-analyses including trials without exclusions only. Measures were compared using scatter plots and Wilcoxon rank tests for paired observations. P values were two sided. All analyses were done in STATA version 10.

Results

Characteristics of included meta-analyses

Overall, 354 reports of reviews of interventions in osteoarthritis were identified (fig 1). Seventeen reports including 21 meta-analyses were eligible. Of these, 14 meta-analyses included at least one trial with and one without exclusion of patients from the analysis and contributed to the study.[16 20-30]

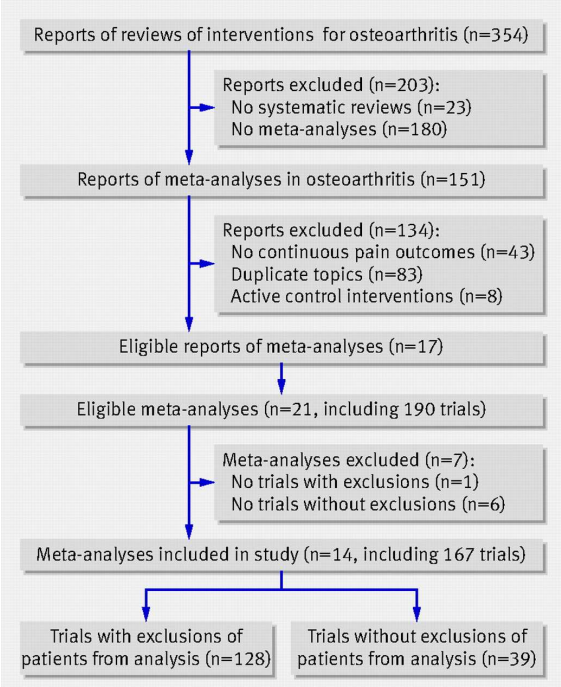

Figure 1: Identification of meta-analyses in osteoarthritis

Table 1 shows the characteristics of the included meta-analyses. The metaanalyses included 167 trials in 41 170 patients. Eight meta-analyses assessed the efficacy of drug interventions and five assessed interventions in complementary medicine. The number of trials per metaanalysis ranged from three to 24 (median 11) and the number of patients contributing to the meta-analysis from 278 to 13 659 (median 1731). The pooled effect sizes derived from random effects meta-analyses including all trials ranged from −0.07, indicating essentially no benefit, to −0.88, representing a large benefit (median −0.40). All meta-analyses favoured the experimental intervention and 11 of 14 showed statistically significant differences between experimental and control intervention at the conventional level of P=0.05. The variance τ^2 as a measure of between trial heterogeneity varied between 0.00 and 0.52 (median 0.04, table 1).

Table 1 | Characteristics of included meta-analyses

Interventions	Drug intervention	Complementary medicine	No of trials*	No of patients	Effect size (95% CI)	Heterogeneity τ^2 (P value)
Exercise *v* control[20]	No	No	16	2700	−0.29 (−0.38 to −0.21)	0.00 (0.41)
Viscosupplementation *v* placebo[21]	Yes	No	22	3046	−0.33 (−0.50 to −0.17)	0.11 (<0.001)†
Self management *v* control[22]	No	No	12	5812	−0.07 (−0.15 to 0.02)	0.01 (0.15)
Glucosamine *v* placebo[23]	Yes	Yes	15	1518	−0.61 (−0.94 to −0.28)‡	0.35 (<0.001)†
Diacerein *v* placebo[25]	Yes	No	6	1613	−0.24 (−0.35 to −0.13)	0.00 (0.33)
Acetaminophen (paracetamol) *v* placebo[24]	Yes	No	5	1849	−0.23 (−0.37 to −0.10)	0.01 (0.13)
Opioids *v* placebo[26]	Yes	No	13	3713	−0.39 (−0.47 to −0.31)	0.00 (0.26)
Oral NSAIDs *v* placebo[29]	Yes	No	24	13659	−0.40 (−0.49 to −0.31)	0.04 (<0.001)
Topical NSAIDs *v* placebo[29]	Yes	No	9	1302	−0.47 (−0.65 to −0.29)	0.04 (0.018)
Low-level laser therapy *v* placebo[28]	No	Yes	8	347	−0.47 (−0.98 to 0.04)	0.42 (<0.001)†
TENS *v* sham[28]	No	Yes	10	358	−0.88 (−1.36 to −0.39)‡	0.52 (<0.001)†
Weight reduction *v* control[27]	No	No	3	278	−0.12 (−0.33 to 0.09)	0.01 (0.34)
Acupuncture *v* control[30]	No	Yes	6	1540	−0.49 (−0.78 to −0.19)	0.12 (<0.001)†
Chondroitin *v* placebo[16]	Yes	Yes	20	3833	−0.72 (−0.95 to −0.49)‡	0.23 (<0.001)†

NSAIDs=non-steroidal anti-inflammatory drugs; TENS=transcutaneous electrical nerve stimulation. Effect sizes and corresponding 95% confidence intervals were derived from random effects meta-analyses of all trials. Negative effect sizes indicate a beneficial effect of experimental intervention. Meta-analyses are ordered according to year of publication.
*Number of trials totals 169 as two trials were included each in two different meta-analyses.
†Meta-analyses considered to have high heterogeneity between trials ($\tau^2 \geq 0.06$).
‡Meta-analyses considered to have large estimated treatment benefit according to overall meta-analysis including all trials (pooled effect size ≤−0.50).

Characteristics of component trials

Table 2 shows the characteristics of included trials. In total, 39 of the 167 trials (23%) included all randomised patients in the analysis. In 114 trials (69%) there were exclusions, and in 14 trials (8%) it was unclear whether exclusions had occurred. Exclusions ranged from 0.1% to 40% (median 7.2%). Trials with exclusions were less likely to provide information on losses to follow-up (P=0.002). Data imputations using the last observation carried forward method were reported by 27% of trials with exclusions and 49% of trials without exclusions, multiple imputation by 4% and 15%, and for 68% and 15% it was unclear how the trialists dealt with missing data in the analysis. Trials with exclusions were published earlier (mean 1998, SD 6) than trials without exclusions (2001, SD 4; P=0.002) and tended to report adequate concealment of allocation and sample size calculations less often. No clear

Table 2 | Characteristics of component trials

Characteristics	No (%) of trials with exclusions (n=128)	No (%) of trials without exclusions (n=39)	P value
Losses to follow-up:			
None*	1 (1)	9 (23)	0.002
<10%	32 (26)	6 (15)	
10-20%	26 (20)	6 (15)	
>20%	26 (20)	17 (44)	
Information unavailable	43 (33)	1 (3)	
Imputation of missing data:			
Last observation carried forward	35 (27)	19 (49)	<0.001
Multiple imputation	5 (4)	6 (15)	
Explicitly no losses to follow-up*	1 (1)	8 (21)	
Information unavailable	87 (68)	6 (15)	
Adequate concealment of allocation:			
Yes	24 (19)	15 (38)	0.07
No or unclear	104 (81)	24 (62)	
Described as double blind:			
Yes	88 (69)	28 (72)	0.74
No	40 (31)	11 (28)	
Adequate blinding of patients:			
Yes	60 (46)	21 (54)	0.43
No or unclear	68 (54)	18 (46)	
Primary outcome:			
Reported	70 (55)	27 (69)	0.27
Not reported	58 (45)	12 (31)	
Sample size calculation:			
Reported	50 (39)	23 (59)	0.12
Not reported	78 (61)	16 (41)	
No of patients randomly assigned:			
>200	48 (38)	21 (54)	0.20
≤200	80 (62)	18 (46)	
Drug intervention:			
Yes	85 (66)	28 (72)	0.53
No	43 (34)	11 (28)	
Complementary medicine:			
Yes	44 (34)	15 (38)	0.62
No	84 (66)	24 (62)	
Funding by non-profit organisation:			
Yes	30 (24)	9 (23)	0.96
No or unclear	98 (76)	30 (77)	
Language of primary report:			
English	120 (94)	38 (97)	0.49
Non-English	8 (6)	1 (3)	
Year of publication:			
1980-9	17 (13)	0 (0)	0.003
1990-9	49 (38)	10 (26)	
2000-7	62 (49)	29 (74)	

P values are derived from logistic regression models adjusted for clustering of trials within meta-analyses. Comparisons of frequency of concealment of allocation, description of double blinding, adequate blinding of patients, trial size, type of intervention, funding, language of publication, and publication year were preplanned. *One trial reporting that no patient was lost to follow-up used the last observation carried forward approach to impute some missing outcome data.

differences were observed for blinding, reporting of primary outcome, type of intervention, source of funding, and language of publication.

Effect of exclusions on estimates of treatment effects

Figure 2 shows the forest plot of differences in effect sizes between trials with and trials without exclusions across the 14 meta-analyses. On average, treatment effects were more beneficial in trials with exclusions than in trials without exclusions (difference in effect sizes −0.13, 95% confidence interval −0.29 to 0.04, P=0.13), but the variability in bias between meta-analyses was considerable (τ^2=0.07, P<0.001). Differences in effect sizes ranged from −0.82 to 0.35. Tests of interaction between exclusions from the analysis and estimates of treatment effects were positive at the conventional level of P=0.05 in five meta-analyses: in four meta-analyses estimated effects were more beneficial in trials with exclusions from the analysis and in one meta-analysis estimated effects were more beneficial in trials without exclusions (fig 2).

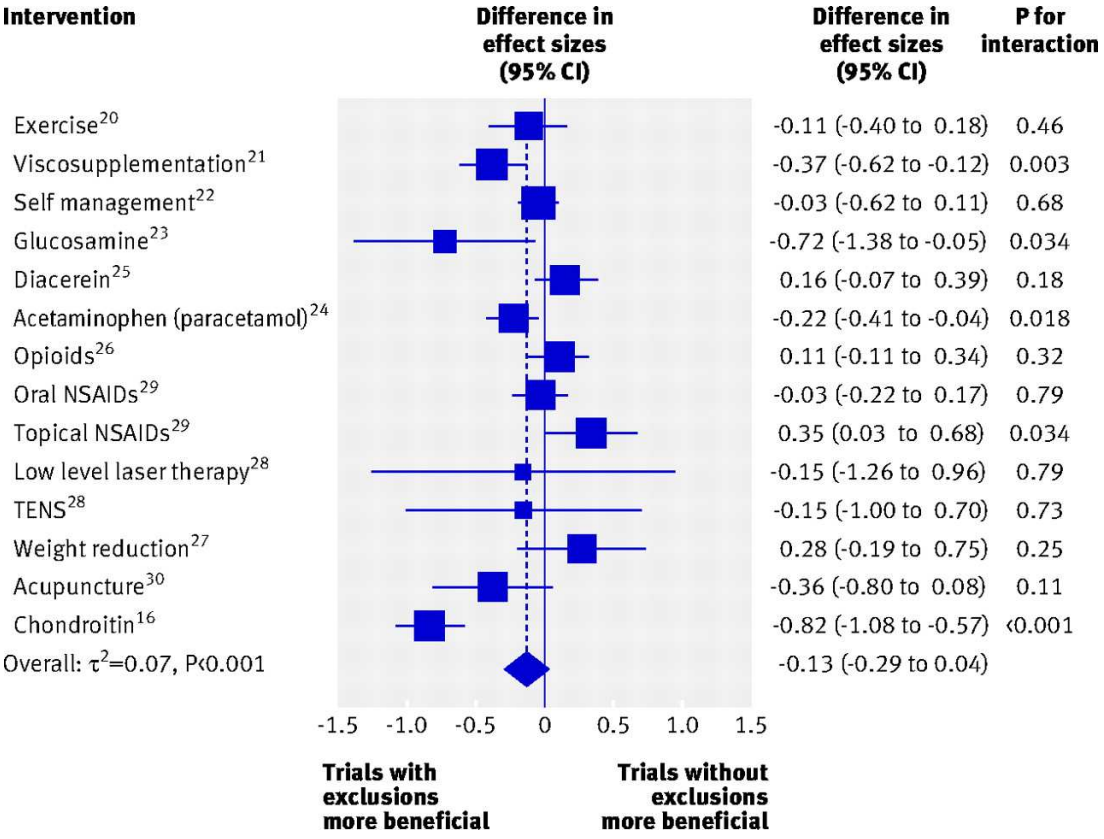

Figure 2 Difference in effect sizes between 128 trials with and 39 trials without exclusions of patients from analysis. A negative difference in effect sizes indicates that trials with exclusions of patients from analysis show more beneficial treatment effects. P values are for interaction between exclusions from analysis and effect sizes. Meta-analyses are ordered according to year of publication. NSAIDs=non-steroidal anti-inflammatory drugs; TENS=transcutaneous electrical nerve stimulation

Figure 3 presents results of stratified analyses. Differences between trials with and trials without exclusions were evident in meta-analyses with a high degree of between trial heterogeneity, but not in meta-analyses with low between trial heterogeneity (P for interaction <0.001). Similarly, differences were more pronounced in meta-analyses with large estimated treatment benefits in the overall meta-analysis compared with metaanalyses with small estimated benefits (P for interaction <0.001) and in meta-analyses of complementary interventions compared with conventional medicine (P for interaction <0.001). When stratifying for these characteristics, the variability in bias decreased considerably. For example, τ^2 was 0.03 in meta-analyses of complementary medicine and 0.02 in meta-analyses of conventional medicine. When adjusting for concealment of allocation (−0.11, 95% confidence interval −0.28 to 0.05, P=0.18) or patient blinding (−0.15, −0.30 to 0.00, P=0.047), average differences between trials with and trials without exclusions of patients were robust. In both adjusted analyses the variability in bias between meta-analyses was much the same as in the primary analysis, with variance estimates τ^2 of 0.08 (P<0.001) and 0.06 (P<0.001), respectively.

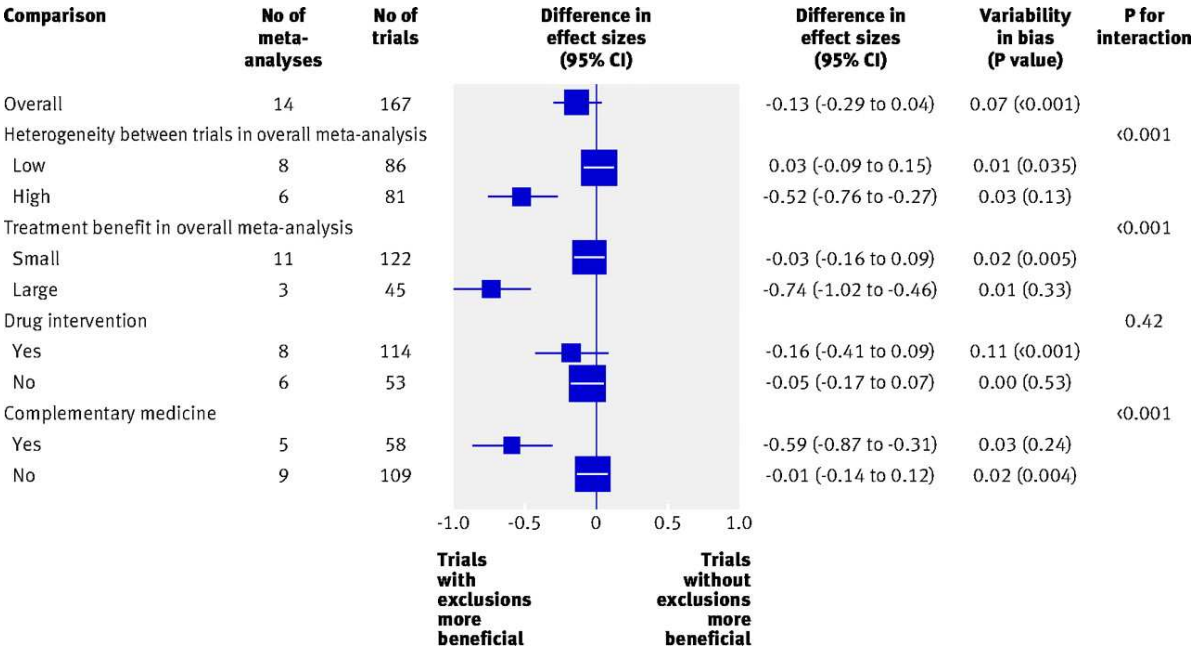

Figure 3 Differences in effect sizes between 128 trials with and 39 trials without exclusions of patients from analysis stratified according to four characteristics of meta-analyses. See table 1 for a description of meta-analyses according to these characteristics. A τ^2 <0.06 indicates low between trial heterogeneity and a τ^2 ≥0.06 high between trial heterogeneity. An effect size >–0.5 indicates a small benefit of the experimental intervention and an effect size ≤–0.5 a large benefit. A negative difference in effect sizes indicates that trials with exclusions of patients from analysis show a more beneficial treatment effect. Variability in bias between-meta-analyses is expressed as heterogeneity variance τ^2

Impact of restricting meta-analyses to trials without exclusions

Figure 4 presents comparisons of overall meta-analyses including all trials with restricted meta-analyses including trials without exclusions only. After the restriction the number of trials included in a single meta-analysis decreased from a median of 11 to a median of 2 and the number of patients from a median of 1731 to a median of 401. Estimates of treatment benefits decreased in 10 meta-analyses and increased in four (P=0.10). Between trial heterogeneity decreased in 12 meta-analyses and increased in one (P=0.006). For one meta-analysis only one trial had no exclusions from the analysis, and no between trial heterogeneity could be estimated after the restriction.[30] Precisions of pooled effect size estimates decreased in nine meta-analyses and increased in five (P=0.25). P values became larger in 10 meta-analyses and smaller in four (P=0.016). After the restriction to trials without exclusions only, six meta-analyses lost statistical significance at the conventional level of P=0.05.

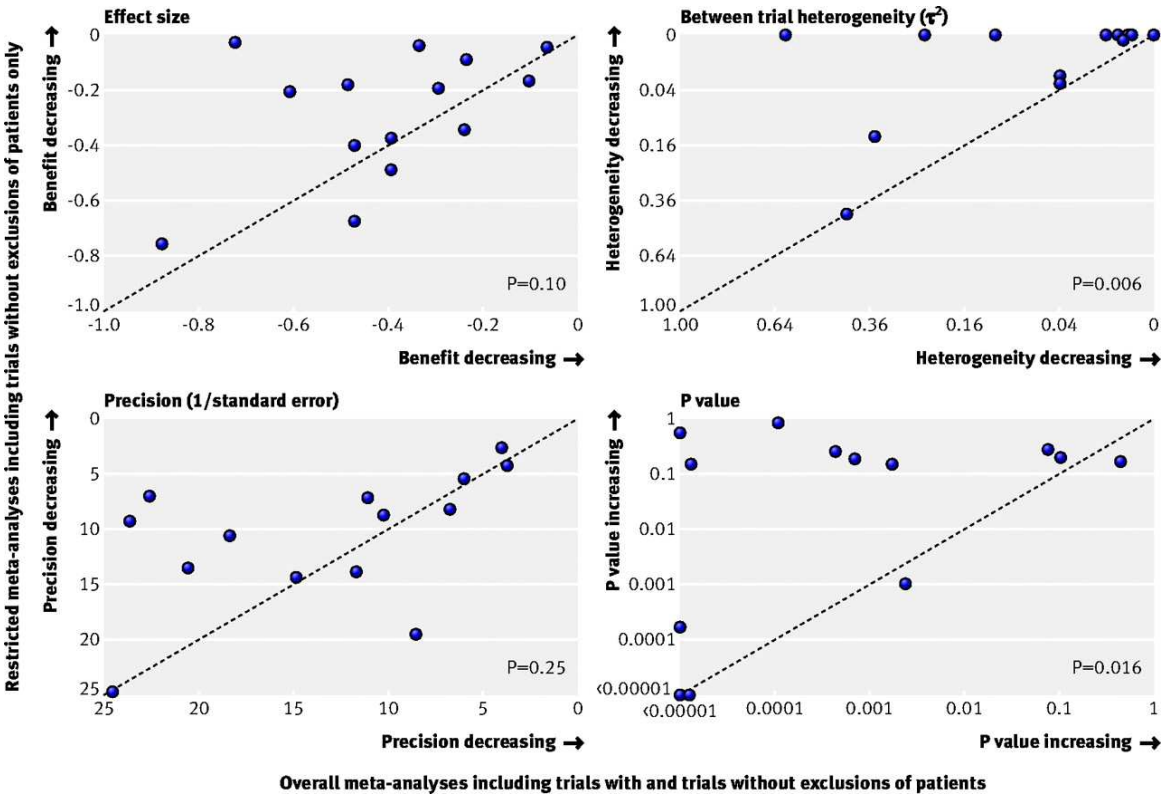

Figure 4 Effect sizes, between trial heterogeneity τ^2, precision, and P values of overall treatment benefits compared between overall meta-analyses including trials with and without exclusions of patients (x axis) and restricted meta-analyses including trials without exclusions of patients only (y axis). Dashed line indicates that estimates are identical. P values are derived from Wilcoxon rank tests for paired observations.

Discussion

In this meta-epidemiological study of 14 meta-analyses and 167 trials we found that excluding randomised patients from the analysis often resulted in biased estimates of treatment effects. The average estimate of bias of a difference in effect size of −0.13 may seem small (fig 2), however it corresponds to one quarter to one half of a typical treatment effect found for interventions in osteoarthritis.[15] The impact of exclusions on estimates of treatment effects seemed most pronounced in meta-analyses with large treatment benefits, metaanalyses on complementary interventions, and metaanalyses with a high degree of heterogeneity between trials, but the extent and direction of bias may be unpredictable in a specific situation. Tests of interaction between exclusions from the analysis and estimates of treatment effects were statistically significant in five meta-analyses; in four of these meta-analyses, estimated treatment effects were less beneficial in trials without exclusions.

When restricting meta-analyses to trials without exclusions, P values increased in most cases and six meta-analyses lost statistical significance at P=0.05 (fig 4). This increase in P values was not only due to a loss of statistical power.[31] As a result of the restriction the between trial heterogeneity τ^2 decreased considerably. Therefore the average loss of statistical precision of random effects meta-analyses was smaller than what could be expected after the exclusion of over half the trials. Only in five meta-analyses was there a relevant loss of precision after the restriction, in six meta-analyses the statistical precision remained much the same, and in three meta-analyses the precision increased.

Strengths and limitations of the study

In practice, various definitions of the intention to treat principle are used.[5 32] In our study we did not rely on statements in the trial reports on whether an intention to treat analysis was done or not. Rather we required explicit information about the flow of patients through the study[33 34] or clear statements that all randomised patients were included in the analysis. Some might argue that our distinction between trials with and trials without exclusions from the analysis was overly stringent. The exclusion of only a small proportion of patients from the analysis, for example, may be considered unlikely to have any impact on estimated treatment benefits. We would expect that any bias associated with exclusions from the analysis will increase with the number of exclusions. Therefore the overall estimate of bias might increase with the selection of a less rigorous cut off. Others may argue that our classification was not stringent enough and that we should have required an affirmative statement that no crossovers had occurred and that all randomised patients were included in the analysis in the group to

which they were originally allocated. Only seven of the 167 included trials (4%) explicitly provided this information, so we were unable to examine this issue.

As with other meta-epidemiological studies,[69] our study is based on published information and depends on the quality of the trial reports. Even though the quality of reporting is generally low,[35 36] we were able to determine in all but 14 trials whether exclusions from the analysis had occurred. Compared with previous meta-epidemiological studies,[1 6 8-10] misclassification of trials due to inadequate reporting[11] is therefore less likely to have introduced bias in our study. At least two thirds of the trials included in our study had incomplete outcome data. Two approaches towards imputation of missing data are generally used to replace missing data and allow an intention to treat analysis: the last observation carried forward method and multiple imputation. We were unable to examine whether the approach used for data imputation influences estimates of treatment effects because of the strikingly low quality of reporting.[32 34 35] Other types of bias that may affect the results of randomised trials include selection bias due to inadequate concealment of allocation, and performance and assessor bias due to a lack of blinding of patients and therapists.[2 12] In our study, the observed association between exclusions of patients from the analysis and estimates of treatment effects could be confounded by concealment of allocation: trials with exclusions tended to report adequate concealment of allocation less often than trials without exclusions. This correlation may have resulted in spurious associations. When accounting for concealment of allocation in a sensitivity analysis, however, we found our results to be robust. Finally, characteristics of meta-analyses were also correlated. For example, meta-analyses in complementary medicine were likely to show large treatment benefits and a high degree of heterogeneity between trials. Our understanding of the interplay of these characteristics is incomplete. Therefore the results of our stratified analyses (fig 3) should to be interpreted with caution. A detailed examination of that problem would require a much larger set of meta-analyses.

Context

As is the case for other types of bias,[12] the extent of attrition bias might depend on the type of outcome. Ours is the first meta-epidemiological study to investigate pain as a patient reported outcome, a measure extensively used in research on osteoarthritis.[15] This outcome is more prone to bias than objective binary outcomes such as mortality.[12] Four studies have examined the impact of attrition bias on estimates of treatment benefits in randomised controlled trials and meta-analyses on an odds ratio scale.[1 6 8-10] The direction and magnitude of attrition bias

varied between different studies according to different methods and definitions used, different clinical areas addressed, and the potential for misclassification of trials because of inadequate reporting.[6 8-10] A recent study used individual patient data and found that analyses with patients excluded showed more beneficial effects of the experimental treatment than analyses according to the intention to treat principle.[1] Another study of placebo controlled trials of serotonin reuptake inhibitors sponsored by the pharmaceutical industry found that the experimental intervention was favoured less in intention to treat analyses than in per protocol analyses. Many published reports of these trials ignored the results of intention to treat analyses and reported only the more favourable per protocol analyses.[37] Several authors pointed out that attrition bias can go in either direction and is difficult to predict for a specific situation,[1 2] which is in accordance with our findings of highly variable effects between meta-analyses. Previous meta-epidemiological studies, which examined the effect of exclusions from the analysis,[1 6 8-10] might be reanalysed in the light of our results to examine the variability in bias associated with exclusions.

Implications

The intention to treat principle aims to compare patients in the groups to which they were originally allocated. The most stringent interpretation of intention to treat includes the analysis of all patients, regardless of whether they were eligible, received treatment, and adhered to the protocol.[5] In practice, various interpretations are used, some of which allow for exclusions after randomisation. Many trialists exclude randomised patients who did not receive at least one dose of the allocated intervention, whereas others exclude patients found retrospectively to be ineligible.[5 38] Both approaches to excluding patients from the analysis may produce unbiased estimates if patients and treating doctors are unaware of the allocated intervention and if the decision to exclude patients is based solely on information collected before randomisation and unrelated to group assignment and clinical outcome.[38] In addition, exclusions from the analysis owing to randomly missing outcome data may be less problematic than the selective exclusion of patients owing to protocol violations. These assumptions are hardly ever verifiable: details on the flow of participants through the various stages of a trial and descriptions of procedures used to determine whether patients should be excluded from the analysis are often omitted from published reports of randomised trials.[5 34] Therefore it is difficult to determine from published information whether reported exclusions from the analyses resulted in bias,[2] and strict adherence to the intention to treat principle should be advocated.[33 39]

The purpose of an intention to treat analysis is to preserve an unbiased treatment allocation and the prognostic balance between treatment groups. In contrast, per protocol analyses include only those patients who received treatment as defined in the study protocol and provided outcome data. Patients excluded from per protocol analyses are likely to be different from those analysed: they may have had an acute exacerbation of the studied condition or experienced side effects of the evaluated intervention.[3] Trials without exclusions more often reported imputations of missing data than those with exclusions. The last observation carried forward approach was used most often: missing values were replaced by the last value observed. This method is popular for imputation of missing data in musculoskeletal research[40] [41] but leads to overly precise estimates and potential bias.[42] [43] Multiple imputation is more difficult to carry out but avoids those problems[44]: each missing value is replaced by multiple simulated values, and the analysis of the resulting multiple versions of the complete dataset can account for the uncertainty about missing data. The CONSORT statement urges transparent reporting of the flow of participants through the various stages of a trial, including a description of withdrawals and losses to follow-up and the reasons for exclusions from the analysis.[33] [39] In our view a detailed description of strategies used to handle missing outcome data is also essential.

Conclusions

The box summarises our recommendations for practice. Excluding patients from the analysis often results in biased estimates of treatment effects in randomised trials. To avoid potential attrition bias, trialists should ensure low dropout rates and high compliance rates and minimise missing outcome data. Results of intention to treat analyses, which are based on the inclusion of all patients in the analysis in the group to which they were originally allocated, should always be reported. Sensitivity analyses, which are restricted to patients adhering to the protocol, may be described in addition. In systematic reviews and meta-analyses, data extraction should be based on results from analyses of all randomised patients, whenever possible. The influence of exclusions from analysis on estimated treatment benefits should be routinely assessed in stratified analyses. This may be particularly important in complementary medicine, in the presence of high heterogeneity between trials, and when pooled effect sizes indicate a large benefit of evaluated interventions.

Recommendations for practice

- Since excluding patients from the analysis often results in biased estimates of treatment effects, trialists should ensure low dropout rates and high compliance rates and minimise missing outcome data
- Trialists should always report results of intention to treat analyses, including all randomised patients in the analysis in the group to which they were originally allocated. If data imputations are necessary to carry out an intention to treat analysis, multiple imputation should be used to replace missing data
- Those critically appraising trials should generally rely on results from intention to treat analyses
- Authors of systematic reviews should routinely examine the influence of exclusions from the analysis on estimated treatment effects. In case of discrepancies between trials with and trials without exclusions, trials without exclusions should be given precedence

We thank Sacha Blank, Liz King, Katharina Liewald, Linda Nartey, Rebekka Sterchi, and Beatrice Tschannen for help with data extraction; the authors who provided study data; Malcolm Sturdy for the development and maintenance of the database; and Fitore Sallahaj for data entry.

Contributors: PJ conceived the study and developed the protocol. EN, ST, SR, AWSR, EB, and MS collected the data. EN, ST, and PJ did the analysis and interpreted the analysis in collaboration with SR, AWSR, EB, MS, and DGA. EN, ST, and PJ drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version of the manuscript. PJ and SR obtained public funding. PJ provided administrative, technical, and logistic support. EN and PJ are the guarantors.

Funding: Swiss National Science Foundation (grant Nos 4053-40- 104762/3 and 3200-066378) to PJ and SR, and the Swiss Society of Internal Medicine to PJ. The study was part of the Swiss National Science Foundation's National Research Programme 53 on musculoskeletal health. SR was a recipient of a research fellowship funded by the Swiss National Science Foundation (grant No PBBEB-115067). MS was supported by a young investigators' award of the German Ministry of Education and Research (grant No 01 GK

0516). PJ was a PROSPER (programme for social medicine, preventive and epidemiological research) fellow funded by the Swiss National Science Foundation (grant No 3233-066377). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data of the study and had final responsibility for the decision to submit for publication. None of the authors is affiliated with or funded by any manufacturer of any intervention used for osteoarthritis.

Competing interests: None declared.

Ethical approval: Not required.

References

1. Tierney JF, Stewart LA. Investigating patient exclusion bias in metaanalysis. Int J Epidemiol 2005;34:79-87.
2. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42-6.
3. Sackett DL, GentM. Controversy in counting and attributing events in clinical trials. N Engl J Med 1979;301:1410-2.
4. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl JMed 1980;303:1038-41.
5. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-4.
6. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-12.
7. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, EggerM. Statistical methods for assessing the influence of study characteristics on treatment effects in ‘meta-epidemiological’ research. Stat Med 2002;21:1513-24.
8. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann InternMed 2001;135:982-9.
9. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973-82.

10. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2008;149:219.
11. Juni P, Egger M. Commentary: empirical evidence of attrition bias in clinical trials. Int J Epidemiol 2005;34:87-8.
12. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: metaepidemiological study. BMJ 2008;336:601-5.
13. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ 2005;330:68.
14. Elbourne DR, Altman DG, Higgins JP, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: methodological issues. Int J Epidemiol 2002;31:140-9.
15. Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006;20:721-40.
16. Reichenbach S, Sterchi R, Scherer M, Trelle S, Burgi E, Burgi U, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. Ann Intern Med 2007;146:580-90.
17. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88.
18. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum, 1988.
19. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: Wiley, 2004.
20. FransenM, McConnell S, Bell M. Exercise for osteoarthritis of the hip or knee. Cochrane Database Syst Rev 2003(3):CD004286.
21. Lo GH, LaValley M, McAlindon T, Felson DT. Intra-articular hyaluronic acid in treatment of knee osteoarthritis: a meta-analysis. JAMA 2003;290:3115-21.
22. Chodosh J, Morton SC, Mojica W, Maglione M, Suttorp MJ, Hilton L, et al. Meta-analysis: chronic disease self-management programs for older adults. Ann InternMed 2005;143:427-38.
23. Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, et al. Glucosamine therapy for treating osteoarthritis. Cochrane Database Syst Rev 2005(2):CD002946.
24. Towheed TE, Maxwell L, JuddMG, CattonM, HochbergMC, Wells G. Acetaminophen for osteoarthritis. Cochrane Database Syst Rev 2006(1):CD004257.

25. Rintelen B, Neumann K, Leeb BF. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. Arch Intern Med 2006;166:1899-906.
26. Avouac J, Gossec L, Dougados M. Efficacy and safety of opioids for osteoarthritis: a meta-analysis of randomized controlled trials. Osteoarthritis Cartilage 2007;15:957-65.
27. Christensen R, Bartels EM, Astrup A, Bliddal H. Effect of weight reduction in obese patients diagnosed with knee osteoarthritis: a systematic review and meta-analysis. Ann Rheum Dis 2007;66:433-9.
28. Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. BMC Musculoskelet Disord 2007;8:51.
29. Bjordal JM, Klovning A, Ljunggren AE, Slordal L. Short-term efficacy of pharmacotherapeutic interventions in osteoarthritic knee pain: a meta-analysis of randomised placebo-controlled trials. Eur J Pain 2007;11:125-38.
30. Manheimer E, Linde K, Lao L, Bouter LM, Berman BM. Meta-analysis: acupuncture for osteoarthritis of the knee. Ann InternMed 2007;146:868-77.
31. Pildal J,Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P. Impact of allocation concealment on conclusions drawn from metaanalyses of randomized trials. Int J Epidemiol 2007;36:847-57.
32. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin Trials 2007;4:350-6.
33. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-94.
34. Egger M, Juni P, Bartlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996-9.
35. Huwiler-Muntener K, Juni P, Junker C, EggerM. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA 2002;287:2801-4.
36. Soares HP,Daniels S, Kumar A, ClarkeM, Scott C, Swann S, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. BMJ 2004;328:22-4.
37. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326:1171-3.

38. FergussonD, Aaron SD, Guyatt G,Hébert P. Postrandomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002;325:652-4.
39. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. Ann InternMed 2001;134:657-62.
40. Baron G, Boutron I, Giraudeau B, Ravaud P. Violation of the intent-to-treat principle and rate ofmissing data in superiority trials assessing structural outcomes in rheumatic diseases. Arthritis Rheum 2005;52:1858-65.
41. KimM. Statistical methods in Arthritis&Rheumatism: current trends. Arthritis Rheum 2006;54:3741-9.
42. Julious SA, MulleeMA. Issues with using baseline in last observation carried forward analysis. Pharm Stat 2008;7:142-6.
43. Streiner DL. The case of the missing data: methods of dealing with dropouts and other research vagaries. Can J Psychiatry 2002;47:68-75.
44. Baron G, Ravaud P, Samson A, Giraudeau B. Missing data in randomized controlled trials of rheumatoid arthritis with radiographic outcomes: a simulation study. Arthritis Rheum 2008;59:25-31.

Appendix Table 1 Search strategy

Medline, Embase, Cinahl were searched via the Ovid platform (www.ovid.com) using the search strategy below. The Cochrane Library was searched using the search strategy below without the methodological filters for human studies and systematic reviews/meta-analyses. The search was last updated November 20, 2007.

1. osteoarthriti$.ti,ab,sh.
2. osteoarthro$.ti,ab,sh.
3. gonarthriti$.ti,ab,sh.
4. gonarthro$.ti,ab,sh.
5. coxarthriti$.ti,ab,sh.
6. coxarthro$.ti,ab,sh.
7. arthros$.ti,ab.
8. arthrot$.ti,ab.
9. ((knee$ or hip$ or joint$) adj3 (pain$ or ach$ or discomfort$)).ti,ab.
10. ((knee$ or hip$ or joint$) adj3 stiff$).ti,ab.
11. or/1-10
12. animal.sh.
13. human.sh.
14. 12 and 13
15. 12 not 14
16. Cochrane database of systematic reviews.jn.
17. search.tw.
18. meta-analysis.pt.
19. Medline.tw.
20. systematic review.tw.
21. or/16-20
22. 11 not 15
23. 21 and 22
24. remove duplicates from 23

Article 2

The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study

Eveline Nüesch,[1,2)] Stephan Reichenbach,[1,3)] Sven Trelle,[1,2)] Anne W.S. Rutjes,[1,4)] Katharina Liewald,[1)] Rebekka Sterchi,[1)] Douglas G. Altman,[5)] Peter Jüni[1,2)]

From [1)]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2)]CTU Bern, Bern University Hospital, Switzerland; [3)]Department of Rheumatology, Immunology and Allergology, Bern University Hospital, Switzerland; [4)]Laboratory of Clinical Epidemiology of Cardiovascular Disease, Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Santa Maria Imbaro, Italy; [5)]Centre for Statistics in Medicine, University of Oxford, United Kingdom

Arthritis Rheum. 2009 Nov 30;61(12):1633-1641.

Abstract

Objective To evaluate the association of adequate allocation concealment and patient blinding with estimates of treatment benefits in osteoarthritis trials.

Methods We performed a meta-epidemiologic study of 16 meta-analyses with 175 trials that compared therapeutic interventions with placebo or non intervention control in patients with hip or knee osteoarthritis. We calculated effect sizes from the differences in means of pain intensity between groups at the end of follow-up divided by the pooled SD and compared effect sizes between trials with and trials without adequate methodology.

Results Effect sizes tended to be less beneficial in 46 trials with adequate allocation concealment compared with 112 trials with inadequate or unclear concealment of allocation (difference -0.15; 95% confidence interval [95% CI] -0.31, 0.02). Selection bias associated with inadequate or unclear concealment of allocation was most pronounced in meta-analyses with large estimated treatment benefits ($P < 0.001$ for interaction), meta-analyses with high between-trial heterogeneity ($P = 0.009$), and meta-analyses of complementary medicine ($P = 0.019$). Effect sizes tended to be less beneficial in 64 trials with adequate blinding of patients compared with 58 trials without (difference -0.15; 95% CI -0.39, 0.09), but differences were less consistent and disappeared after accounting for allocation concealment. Detection bias associated with a lack of adequate patient blinding was most pronounced for non-pharmacologic interventions ($P < 0.001$ for interaction).

Conclusion Results of osteoarthritis trials may be affected by selection and detection bias. Adequate concealment of allocation and attempts to blind patients will minimize these biases.

Introduction

Inadequate methodology may flaw the results of randomized osteoarthritis trials.[1] Meta-epidemiologic studies examine the association of specific trial characteristics, such as concealment of allocation or patient blinding, with estimated treatment effects in a collection of meta-analyses and their component trials.[2–5] In meta-analyses of these meta-epidemiologic studies, inadequate concealment of allocation and a lack of double blinding were associated with exaggerated estimates of treatment benefits.[1,6] In a combined analysis of data from 3 meta-epidemiologic studies of binary outcomes across different medical fields, we recently found that overestimations of treatment benefits were more pronounced for subjective outcomes as compared with objective outcomes such as overall mortality.[4] Subjective outcomes, such as patient-reported pain intensity measured on visual analog, numeric rating, or Likert scales, are frequently used in osteoarthritis trials, whereas objective binary outcomes, such as mortality, are rarely addressed.[7–9]

We performed a meta-epidemiologic study in the field of clinical osteoarthritis research to determine whether components of methodologic quality are associated with overestimates of treatment effects. We previously reported that the exclusion of randomized patients from the analysis was associated with likely overestimates of treatment benefits in osteoarthritis trials, but the extent and direction of this attrition bias resulting from the biased exclusion of patients after entry into the trial remained unpredictable in a specific situation.[5] Bias may also occur at earlier stages of a trial: selection bias through the biased allocation of patients to comparison groups at trial entry if the allocation of patients is not adequately concealed, and detection and performance bias if blinding of patients is inadequate, which may result in biased assessment of self reported outcomes, differential placebo or nocebo effects across comparison groups, and the unequal intake of analgesic cointerventions apart from the treatment under evaluation.[1] Here we report on the association of estimates of treatment benefits with the adequacy of concealment of allocation and patient blinding in clinical osteoarthritis research.

Materials and Methods

Searches and selection of meta-analyses. We searched The Cochrane Library, Medline, EMBase, and CINAHL using a combination of keywords and text words related to osteoarthritis, which were combined with validated filters for controlled clinical trials and meta-analyses. Details of the search strategy are described elsewhere.[5] The last update was performed on November 20, 2007.

Meta-analyses of randomized or quasi-randomized trials in patients with osteoarthritis of the knee or hip were eligible if they evaluated patient-reported pain in patients allocated to any intervention compared with patients allocated to placebo, a sham intervention, or a non-intervention control group. If one topic was covered by several reports, the most recent report was included. Two reviewers independently evaluated the reports for eligibility and disagreements were resolved by discussion or by involvement of a third reviewer. Reports of all component trials were obtained and no language restrictions were applied.

Data extraction. Two reviewers independently extracted data from individual trials regarding interventions, funding, publication year, design characteristics, study size, and results on a standardized form. The primary outcome was pain intensity. If different pain-related outcomes were reported, we referred to a previously described hierarchy of outcomes[9,10] and extracted the outcome that was highest on this list. Global pain took precedence over pain on walking and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain subscores, for example. If a trial report provided data on global pain scores and WOMAC pain subscores, we only recorded data on global pain scores. If more than one time point was reported, we extracted the outcome at 3 months after the end of treatment for potentially structuremodifying agents, such as chondroitin, and at 12 months after the end of treatment for behavior-changing interventions, such as education. For all other interventions, we extracted the outcome at the end of treatment. When necessary, means and measures of dispersion were approximated from figures. For crossover trials, we extracted data from the first period only.[11] Disagreements were discussed with a third reviewer and subsequent consensus was reached.

Quality assessment. Concealment of allocation was considered adequate if the investigators responsible for patient selection and inclusion were unable to know before allocation which treatment was next, e.g., central randomization; the use of sequentially numbered, sealed, and opaque assignment envelopes; or coded drug packs. Concealment of allocation of trials, which lacked a specific statement, was classified as unclear. Patient blinding was considered

adequate if a placebo or sham control intervention was used and experimental and control interventions were described as indistinguishable or the use of a double dummy technique was reported. Analyses were considered to be performed adequately according to the intent-to-treat principle if all of the randomized patients were included in the analysis.[5] The definitions of different types of bias in randomized trials and measures to minimize them are provided in Table 1.

Table 1. Definitions of the terminology used in this study

Type of bias	Definition	Measure to minimize bias
Selection bias	Biased allocation to comparison groups	Concealment of allocation: Procedures that prevent personnel assigning patients to intervention groups from foreseeing allocation. Adequate if the investigators responsible for patient selection and inclusion were unable to know before allocation which treatment was next, e.g., central randomization; the use of sequentially numbered, sealed, and opaque assignment envelopes; or coded drug packs.
Performance bias	Unequal provision of care apart from intervention under evaluation	Blinding: Procedures that prevent therapists and patients from knowing which intervention was received. Adequate if allocated interventions were indistinguishable.
Detection bias	Biased assessment of outcome(s)	Blinding: Procedures that prevent outcome assessors from knowing which intervention was received. Adequate if independent, blind outcome assessors evaluated outcomes or, in the case of patient-reported outcomes, if allocated interventions were indistinguishable.
Attrition bias	Biased occurrence and handling of deviations from protocol and losses to followup	Procedures that prevent exclusion of randomized patients from the analyses and minimize protocol deviations and losses to followup. Adequate if all randomized patients were included in the analysis in the group they were originally allocated to, regardless of their adherence to the study protocol (intent-to-treat analysis).

Data synthesis. Treatment effects were expressed as effect sizes (ES), dividing the difference in mean values at the end of the trial by the pooled SD.[10] A negative ES indicates a beneficial effect of the experimental intervention. If some required data were unavailable, we used approximations as previously described.[10] We used standard random-effects meta-analyses to combine ES across trials and calculated the DerSimonian and Laird estimate of the variance τ^2 to determine heterogeneity between trials.[12,13]

Within each meta-analysis, we estimated the ES of trials with and without adequate allocation concealment separately using a random-effects meta-analysis. For each meta-analysis, we derived the difference between pooled estimates from trials with adequate allocation concealment and trials without adequate allocation concealment. Then we combined these differences using a random-effects meta-analysis fully allowing for heterogeneity between meta-analyses,[3] and measured the variability in bias estimates between meta-analyses using τ^2 as a measure of heterogeneity.[13] Formal tests of interaction between concealment of allocation and estimated treatment benefits were performed separately for each meta-analysis based on Z scores for the estimated difference in ES between trials with and without adequate

concealment of allocation and the corresponding SE. In sensitivity analyses, we additionally stratified by patient blinding and intent-to-treat analysis to account for potential confounding by these factors. The same procedure was followed for trials with and without adequate blinding of patients. A negative difference in ES indicates that trials with adequate allocation concealment or adequate patient blinding show a less beneficial treatment effect. Then we performed stratified analyses accompanied by interaction tests based on Z scores according to the following prespecified characteristics:[5] treatment benefit in overall meta-analysis (small [ES greater than -0.5] versus large [ES less than or equal to-0.5]), between-trial heterogeneity in overall metaanalysis (low [τ^2 <0.06] versus high [$\tau^2 \geq 0.06$]), and type of intervention assessed (pharmacologic versus nonpharmacologic interventions; conventional versus complementary medicine). The prespecified cutoff of $\tau^2 = 0.06$ corresponds to a difference between the smallest and largest ES of approximately 1 ES.

Finally, we compared pooled ES, between-trial heterogeneity, precision defined as 1/SE, and P values for pooled ES between random-effects meta-analyses including all trials and restricted to meta-analyses including trials with adequate concealment of allocation or adequate patient blinding only using Wilcoxon's rank tests for paired observations. All P values were 2-sided. All analyses were performed with Stata statistical software, version 10.1 (StataCorp, College Station, TX).

Results

Selection and characteristics of meta-analyses. We identified 151 reports of meta-analyses of osteoarthritis trials. A total of 134 reports were excluded because they either included no continuous pain outcome (n = 43), covered duplicate topics (n = 83), or used only active control interventions (n = 8).[5] One report described 4 meta-analyses and one report described 2 meta-analyses. Therefore, 21 meta-analyses described in 17 reports were eligible,[5] and 16 meta-analyses of 175 trials and 41,142 patients[10,14–24] contributed to the analyses. Characteristics of the included meta-analyses are shown in Supplementary Appendix A. The median treatment benefit in the 16 included meta-analyses was an ES of -0.43 (range -0.88 to -0.07) with a median between-trial heterogeneity variance of 0.04 (range 0.00–0.52). Four meta-analyses showed a large treatment effect[10,15,21] and 7 showed a high degree of between-trial heterogeneity.[10,15–17,21] Seven meta-analyses addressed pharmacologic interventions[10,14,15,17,22,24] and 9 addressed nonpharmacologic treatments.[16,18–21,23] Nine were related to conventional interventions[14,17–20,22–25] and 7 were related to complementary medicine.[10,15,16,21]

Concealment of allocation. Fourteen meta-analyses with 158 trials and 40,437 patients included both trials with and without adequate concealment of allocation and contributed to the analysis. Table 2 shows a comparison of the characteristics of these trials. Forty-six trials (29%) reported adequate allocation concealment and 111 trials (70%) were unclear about concealment of allocation. One trial (1%) was quasi-randomized using alternation, and allocation concealment was considered inadequate. Of trials with adequate concealment, 26

Table 2. Comparison of characteristics between trials with and trials without adequate concealment of allocation

	Adequate (n = 46), no. (%)	Inadequate or unclear (n = 112), no. (%)	P*
Adequate patient blinding			0.008
Yes	33 (72)	42 (37)	
No/unclear	13 (28)	70 (63)	
Intent-to-treat analysis			0.07
Yes	15 (33)	17 (15)	
No/unclear	31 (67)	95 (85)	
Number of allocated patients			0.021
>200	27 (59)	40 (36)	
≤200	19 (41)	72 (64)	
Pharmacologic intervention†			0.59
Yes	30 (65)	79 (71)	
No	16 (35)	33 (29)	
Complementary medicine‡			0.92
Yes	15 (33)	35 (31)	
No	31 (67)	77 (69)	
Year of publication			< 0.001
1980–1999	8 (17)	61 (54)	
2000–2007	38 (83)	51 (46)	

* Derived using logistic regression models adjusted for clustering of trials within meta-analyses.
† Pharmacologic interventions include chondroitin, diacerein, glucosamine, oral and topical nonsteroidal antiinflammatory drugs (NSAIDs), opioids, and viscosupplementation, and nonpharmacologic interventions include acupuncture, aquatic exercise, exercise, pulsed electromagnetic fields, self-management, static magnets, and weight reduction.
‡ Interventions in complementary medicine include acupuncture, chondroitin, glucosamine, pulsed electromagnetic fields, and static magnets, and interventions in conventional medicine include aquatic exercise, chondroitin, diacerein, exercise, glucosamine, oral and topical NSAIDs, opioids, self-management, viscosupplementation, and weight reduction.

(56%) used coded drug packs or devices; 15 (33%) used central randomization; 4 (9%) used sequentially numbered, sealed, and opaque assignment envelopes; and 1 (2%) used an onsite computer system with allocations kept in a locked unreadable computer file that could be accessed only after the characteristics of an enrolled participant were entered into the database. Four trials that reported the use of assignment envelopes were not deemed to have adequate concealment of allocation because they did not specify that the envelopes were sequentially numbered, sealed, and opaque. Trials with adequate allocation concealment were more likely to report adequate blinding of patients (P = 0.008) and to perform intent-to-treat analyses (P = 0.07), were larger (P = 0.021), and were published more recently (P < 0.001) than trials with inadequate or unclear concealment of allocation.

Figure 1A shows the forest plot of differences in ES between trials with and trials without adequate concealment. Trials with adequate allocation concealment tended to show smaller treatment benefits than trials with inadequate or unclear concealment, with a difference in ES of -0.15 (95% confidence interval [95% CI] -0.31, 0.02; P = 0.08). Differences in ES between trials with and trials without adequate concealment ranged from -1.07 to 0.46, and the variability in bias estimates between meta-analyses was moderate, with a τ^2 estimate of 0.06. Tests of interaction between allocation concealment and ES were positive in 3 of 14 meta-analyses at the conventional level of P = 0.05. The results of stratified analyses are shown in Figure 2. Differences in ES between trials with and without adequate concealment were larger in meta-analyses with a large treatment benefit as compared with meta-analyses with a small benefit (P for interaction < 0.001), meta-analyses with a high degree of between-trial heterogeneity (P for interaction = 0.009), and meta-analyses of complementary medicine as compared with conventional medicine (P for interaction = 0.019).

Figure 3A shows the comparisons of overall meta-analyses, including all trials with meta-analyses restricted to trials with adequate concealment of allocation. Estimates of treatment benefits became smaller in 9 and larger in 5 meta-analyses (P = 0.11). Between-trial heterogeneity decreased in 12 meta-analyses and increased in 2 (P = 0.003), and P values of pooled effects increased in 11 meta-analyses and decreased in 3 (P = 0.026). Statistical precision decreased in 9 meta-analyses and increased in 5 (P = 0.36).

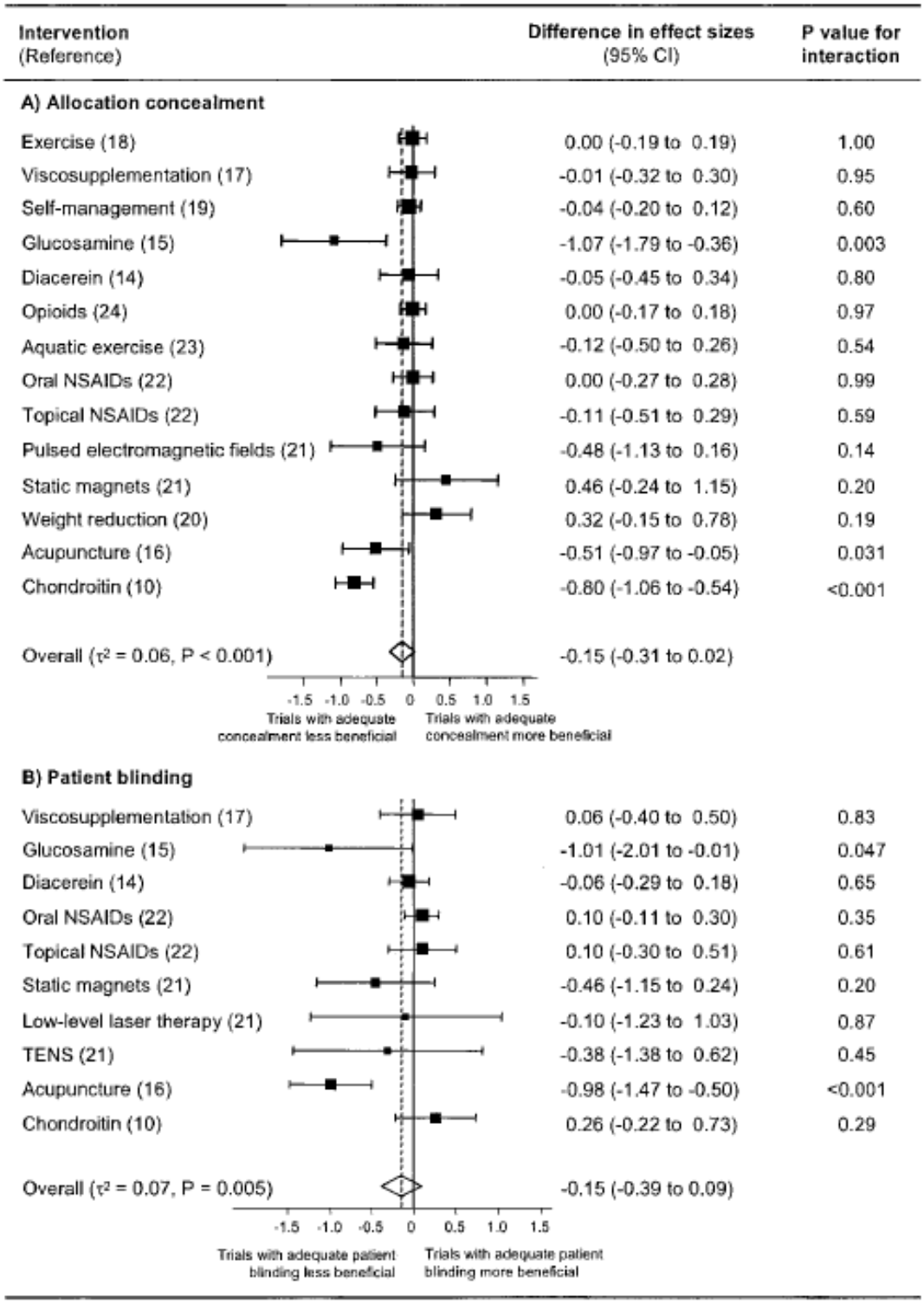

Figure 1 Forest plots of the differences in effect sizes between A, 46 trials with and 112 trials without adequate concealment of allocation, or B, 64 trials with and 58 trials without adequate patient blinding. P values are for the interaction between A, adequate concealment, or B, patient blinding and effect sizes. 95% CI = 95% confidence interval; NSAIDs = nonsteroidal antiinflammatory drugs; TENS = transcutaneous electrical nerve stimulation.

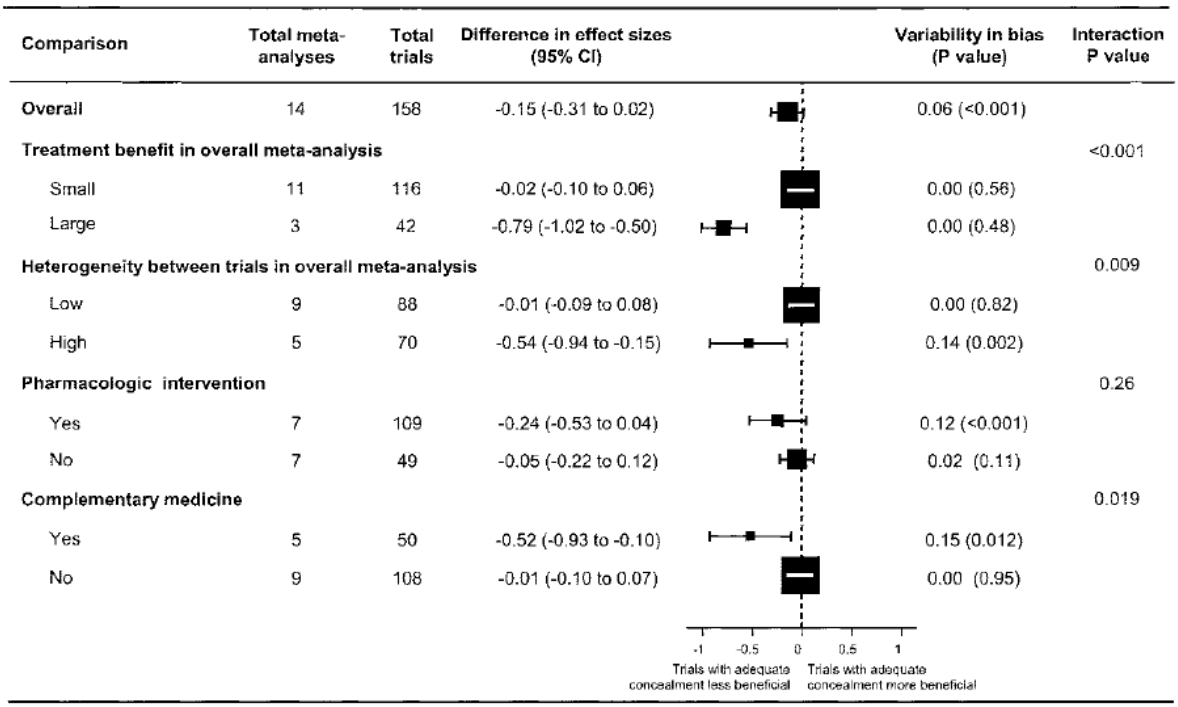

Figure 2. The differences in effect sizes (ES) between 46 trials with and 112 trials without adequate allocation concealment are shown, stratified according to the following characteristics: treatment benefit in the overall meta-analysis, degree of heterogeneity between trials in the overall meta-analysis, pharmacologic intervention (yes/no), and complementary medicine (yes/no). An ES greater than -0.5 indicates a small benefit and an ES less than or equal to -0.5 indicates a large benefit of the experimental intervention. A $\tau^2<0.06$ indicates low between-trial heterogeneity and a $\tau^2\geq0.06$ indicates high between-trial heterogeneity. Pharmacologic interventions include chondroitin, diacerein, glucosamine, oral and topical nonsteroidal antiinflammatory drugs, opioids, and viscosupplementation. Complementary medicine includes acupuncture, chondroitin, glucosamine, pulsed electromagnetic fields, and static magnets. A negative difference in ES indicates that trials with adequate allocation concealment show a less beneficial treatment effect. Variability in bias is shown as the between–meta-analysis heterogeneity variance τ^2 accompanied by P values for heterogeneity between meta-analyses. 95% CI = 95% confidence interval.

Patient blinding. Ten meta-analyses in 122 trials and 27,452 patients included both trials with and trials without adequate blinding of patients and contributed to the analysis. The characteristics of these trials are shown in Table 3. In 64 trials (52%) patients were adequately blinded, in 51 trials (42%) a placebo or sham intervention was used but adequacy of patient blinding remained unclear, and in 7 trials (6%) no placebo or sham intervention was used. Of all of the trials with adequate patient blinding, 55 (86%) reported indistinguishable interventions and 9 (14%) reported the use of double-dummy techniques. Trials with adequate patient blinding were more likely to adequately conceal treatment allocation (P = 0.006) and to evaluate complementary medical interventions (P = 0.023).

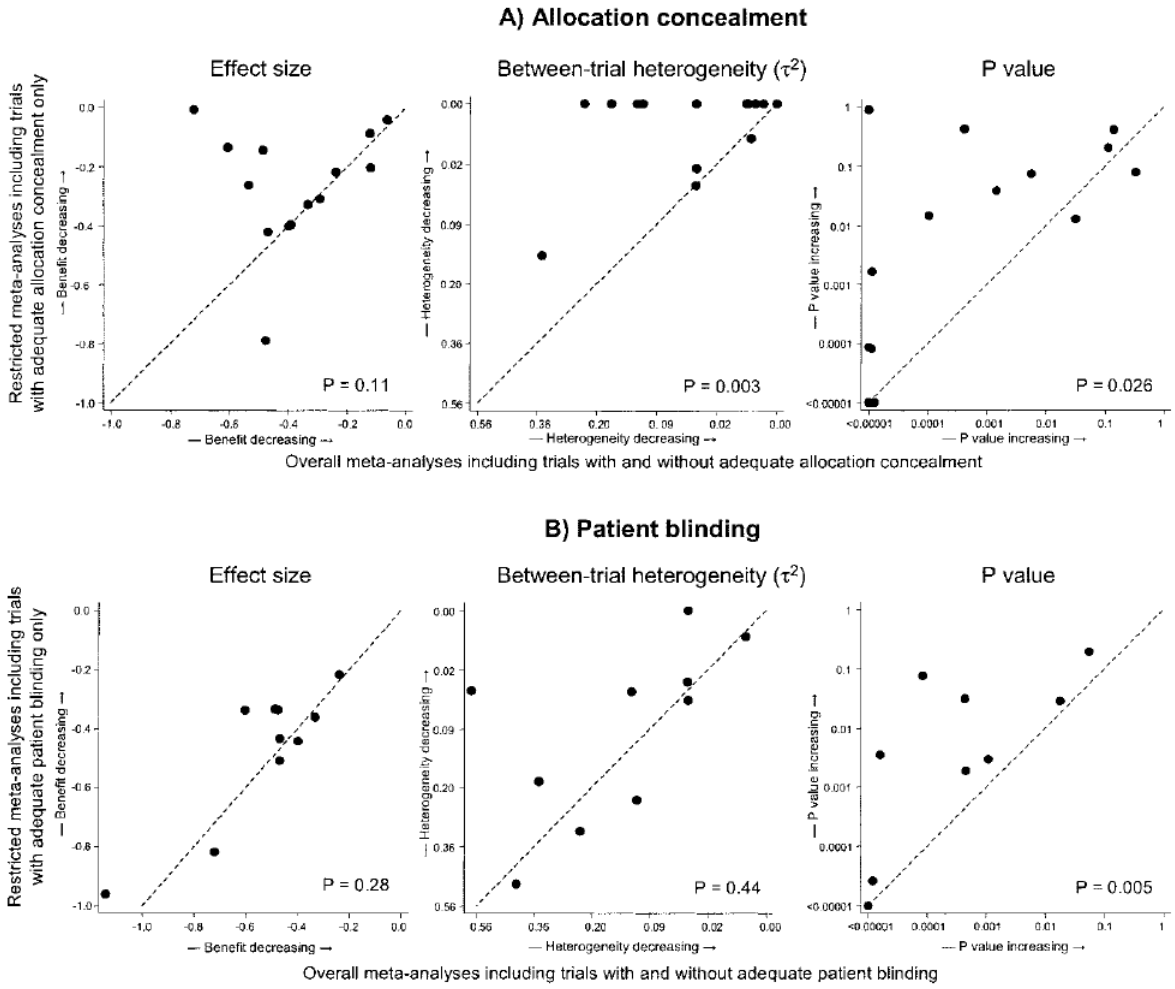

Figure 3. Effect sizes, between-trial heterogeneity variance τ^2, and P values of overall treatment benefits are compared between overall meta-analyses including all trials (x-axis) and restricted meta-analyses including A, trials with adequate allocation concealment only, or B, trials with adequate patient blinding only (y-axis). Broken lines indicate that the estimates are identical. P values are derived using Wilcoxon's rank tests for paired observations.

Figure 1B shows the forest plot of differences in ES between trials with and trials without adequate blinding. Again, estimated treatment effects in trials with adequate patient blinding tended to be smaller compared with treatment effects in trials with inadequate or unclear patient blinding, with a difference in ES of -0.15, but the corresponding CI was wide (95% CI -0.39, 0.09; P = 0.22). In 2 of 10 meta-analyses, tests of interaction between patient blinding and ES were positive. The variability in bias estimates between meta-analyses was high, with a τ^2 estimate of 0.07, and differences in ES ranged from -1.01 to 0.26 between individual meta-analyses. Results of stratified analyses are shown in Figure 4. Differences in ES between trials with and without adequate patient blinding were similar in meta-analyses with small and large treatment benefits (P for interaction = 0.75) and with high and low between-

trial heterogeneity (P for interaction = 0.19), but differences were more pronounced in meta-analyses of nonpharmacologic interventions as compared with metaanalyses of pharmacologic interventions (P for interaction < 0.001) and in meta-analyses of complementary medicine compared with conventional medicine (P for interaction = 0.07). Figure 3B shows the comparisons of overall meta-analyses including all trials with meta-analyses restricted to trials with adequate patient blinding. Estimates of treatment benefits decreased in 6 meta-analyses and increased in 4 (P = 0.28). Heterogeneity between trials decreased in 5 meta-analyses and increased in 5 (P = 0.44), and P values increased in 10 meta-analyses and decreased in none (P = 0.005). Statistical precision decreased in 6 meta-analyses and increased in 4 (P = 0.11).

Table 3. Comparison of characteristics between trials with and trials without adequate patient blinding

	Adequate (n = 64), no. (%)	Inadequate or unclear (n = 58), no. (%)	P*
Adequate allocation concealment			0.006
Yes	23 (36)	3 (5)	
No/unclear	41 (64)	55 (95)	
Intent-to-treat analysis			0.46
Yes	18 (28)	13 (22)	
No/unclear	46 (72)	45 (78)	
Number of allocated patients			0.13
>200	27 (42)	20 (34)	
≤200	37 (58)	38 (66)	
Pharmacologic intervention†			0.59
Yes	49 (77)	47 (81)	
No	15 (23)	11 (19)	
Complementary medicine‡			0.023
Yes	38 (59)	23 (40)	
No	26 (41)	35 (60)	
Year of publication			0.74
1980–1999	31 (48)	30 (52)	
2000–2007	33 (52)	28 (48)	

* Derived using logistic regression models adjusted for clustering of trials within meta-analyses.
† Pharmacologic interventions include chondroitin, diacerein, glucosamine, oral and topical nonsteroidal antiinflammatory drugs (NSAIDs), and viscosupplementation, and nonpharmacologic interventions include acupuncture, low-level laser therapy, static magnets, and transcutaneous electrical nerve stimulation.
‡ Interventions in complementary medicine include acupuncture, chondroitin, glucosamine, low-level laser therapy, static magnets, and transcutaneous electrical nerve stimulation, and interventions in conventional medicine include diacerein, oral and topical NSAIDs, and viscosupplementation.

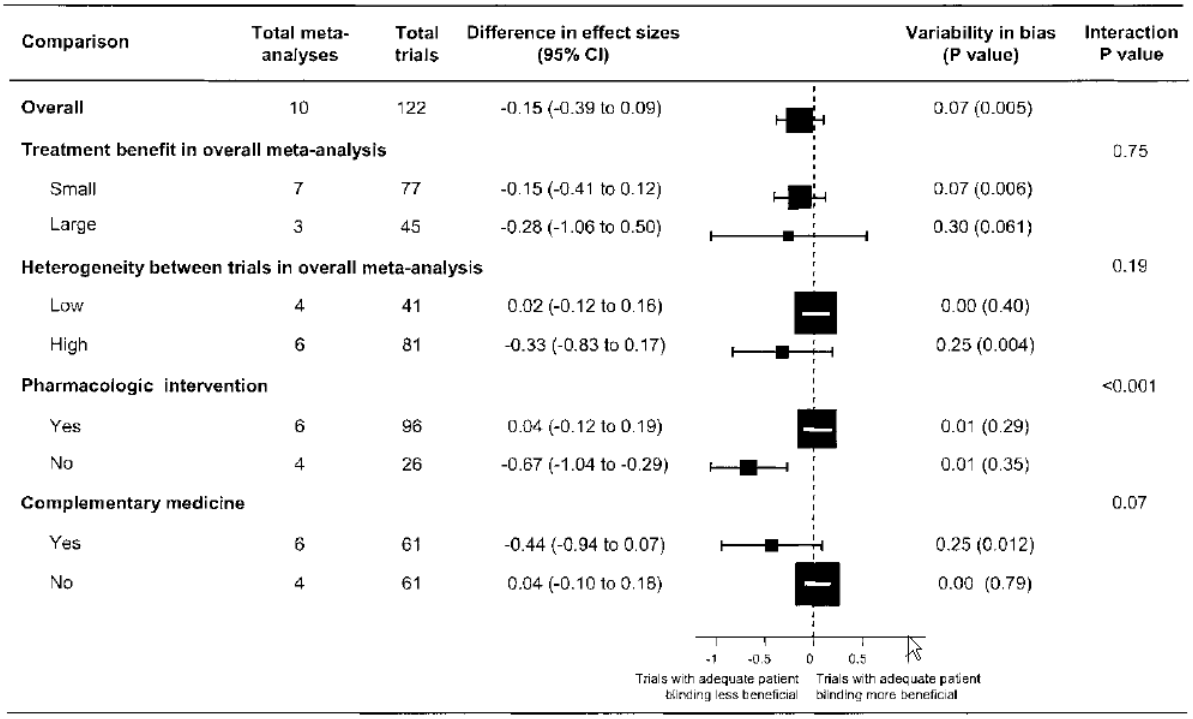

Figure 4. The differences in effect sizes (ES) between 64 trials with and 58 trials without adequate patient blinding are shown, stratified according to the following characteristics: treatment benefit in the overall meta-analysis, degree of heterogeneity between trials in the overall meta-analysis, pharmacologic intervention (yes/no), and complementary medicine (yes/no). An ES greater than -0.5 indicates a small benefit and an ES less than or equal to -0.5 indicates a large benefit of the experimental intervention. A $\tau^2<0.06$ indicates low between-trial heterogeneity and a $\tau^2\geq0.06$ indicates high between-trial heterogeneity. Pharmacologic interventions include chondroitin, diacerein, glucosamine, oral and topical nonsteroidal antiinflammatory drugs, and viscosupplementation. Complementary medicine includes acupuncture, chondroitin, glucosamine, low-level laser therapy, static magnets, and transcutaneous electrical nerve stimulation. A negative difference in ES indicates that trials with adequate patient blinding show a less beneficial treatment effect. Variability in bias is shown as the between–meta-analysis heterogeneity variance τ^2 accompanied by P values for heterogeneity between meta-analyses. 95% CI =95% confidence interval.

Sensitivity analyses. The effects of allocation concealment became more robust after accounting for the presence or absence of adequate patient blinding (difference in ES -0.24; 95% CI -0.41, -0.07), and more precise but attenuated after accounting for intent-to-treat analyses (difference in ES -0.08; 95% CI -0.21, 0.04). The variability in bias estimates in these analyses was similar after accounting for patient blinding ($\tau^2 = 0.07$, $P < 0.001$), but decreased after accounting for intent-to-treat analyses ($\tau^2 = 0.04$, $P = 0.002$). The effects of patient blinding entirely disappeared when accounting for allocation concealment (difference in ES 0.01; 95% CI -0.18, 0.18) and were attenuated when accounting for intent-to-treat analyses (difference in ES -0.06; 95% CI -0.20, 0.09). The variability in bias estimates

decreased after accounting for these characteristics: τ^2 estimates were 0.03 in both analyses (P = 0.08 and 0.05 for heterogeneity, respectively).

Discussion

In this meta-epidemiologic study of osteoarthritis trials, we found that trials with inadequate or unclear concealment of allocation showed larger estimates of treatment benefits than trials with adequate concealment. Evidence of bias was mainly seen in meta-analyses with large treatment effects, meta-analyses with high between-trial heterogeneity, and in meta-analyses of complementary medicine, with a pattern and magnitude of effects similar to what we found previously for bias associated with failure to perform an intent-to-treat analysis.[5] The average bias associated with a lack of concealment of allocation corresponds to one-fourth to one-half of a typical treatment effect found for interventions in osteoarthritis.[9] Evidence of bias associated with a lack of adequate patient blinding was found less consistently. Patients are difficult to blind if allocation is not adequately concealed: if patients and investigators enrolling patients are able to decipher the allocation schedule, subsequent blinding will be impossible. Unsurprisingly, the effects of blinding entirely disappeared after accounting for concealment of allocation in the overall analysis. However, stratified analyses suggested that adequate blinding of patients may be important for nonpharmacologic interventions. The average bias found for this group of interventions was an ES of -0.67, which is larger than the typical treatment effect found for most interventions used for osteoarthritis.[9] This effect was robust to the adjustment for concealment of allocation in a post hoc analysis (difference in ES after accounting for concealment -0.62; 95% CI -1.09, -0.16).

The assessment of the methodologic quality of a trial is intertwined with the quality of reporting: the extent to which a report provides information about the design, conduct, and analysis of the trial.[1] Unfortunately, reports often omit important methodologic details,[26] including who was actually blinded and whether blinding was successful at the time of patient-reported assessments of pain intensity.[25,27–29] A widely used approach to this problem is to assume that the quality is inadequate unless the information to the contrary is provided. This is often justified because faulty reporting generally reflects faulty methods.[1,2] A well-conducted but badly reported trial will, however, be misclassified. Misclassification may have been particularly prominent in the assessment of the adequacy of patient blinding in drug trials. Some of these trials could have adequately blinded patients using matching placebos without describing it. The resulting misclassification would explain the apparent lack of bias associated with patient blinding in pharmacologic trials.

The current study differs in 3 important aspects from previously published meta-epidemiologic studies that addressed the impact of allocation concealment and blinding on estimated treatment benefits.[2,6,30–34] First, we specifically estimated the extent of bias in trials using patient reported pain intensity as a subjective outcome. Subjective outcomes are likely to be more prone to bias due to unclear allocation concealment and inadequate blinding than objective outcomes such as mortality.[4] Second, almost all of the previous meta-epidemiologic studies have considered binary outcomes. To our knowledge, only one pilot study including 35 trials addressed the association between treatment benefits and allocation concealment or blinding in continuous outcomes, but was underpowered to obtain conclusive results.[35] Third, we provide a comprehensive assessment of the extent of unclear concealment of allocation and the lack of patient blinding in randomized osteoarthritis trials and the resulting biases. Less than one-third of the trials reported adequate concealment of allocation. On average, these trials suggested less beneficial treatment effects than the remaining trials. Random allocation of patients can be adequately concealed in any trial, irrespective of the types of interventions compared. Admittedly, patient blinding is not possible for some interventions, such as exercise or self-management. However, even in trials that evaluated interventions that were amenable to blinding, only approximately half reported adequate attempts to blind patients.

Selection bias at trial entry might be the underlying mechanism of an overestimation in trials with inadequate or unclear allocation concealment, whereas selection bias after entry is the likely mechanism resulting in overestimates of treatment benefits in trials that exclude randomized patients from the analysis.[1,5] Lack of adequate patient blinding might result in exaggerated treatment effects due to detection bias in patient-reported outcomes and performance bias introduced by the unequal intake of analgesic cointerventions apart from the treatment under evaluation.[1] Differential placebo or nocebo effects may also be important: patients who know that they receive active treatment may perceive less pain than patients in the inactive control group. In our study, these possible sources of bias introduced by the behavior and perception of patients appeared less important than the selection biases discussed above, which are mainly introduced by investigators.[1]

Only a combination of adequate allocation concealment and adequate analysis according to the intent-to-treat principle will avoid selection biases and render trial results valid and credible. Special caution should be taken when interpreting the results of meta-analyses indicating large benefits of experimental interventions, a high degree of between-trial heterogeneity, or in meta-analyses of complementary medicine. Here, stratified analyses

according to the presence or absence of adequate concealment of allocation and intent-to-treat analysis should be considered mandatory.[5] In case of discrepancies, trials that avoided selection biases should be given precedence.

Trialists should always ensure adequate concealment of allocation and take measures to minimize dropout rates and maximize compliance with the trial protocol to allow an analysis according to the intent-to-treat principle. Blinding of patients is desirable and should be attempted. Authors of reports of osteoarthritis trials should painstakingly follow the Consolidated Standards of Reporting Trials statement[36,37] to ensure fully transparent reporting of methods and results.

Acknowlegments We thank Sacha Blank, Elizabeth Bu¨ rgi, Liz King, Linda Nartey, Martin Scherer, and Beatrice Tschannen for contributing to data extraction. We are grateful to Malcolm Sturdy for the development and maintenance of the database.

Author contributions All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Jüni had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study conception and design. Jüni.
Acquisition of data. Nüesch, Reichenbach, Trelle, Rutjes, Liewald, Sterchi.
Analysis and interpretation of data. Nüesch, Reichenbach, Trelle, Rutjes, Altman, Jü ni.
Statistical analysis. Nüesch, Jüni.

References

1. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.
2. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
3. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21:1513–24.
4. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601–5.
5. Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Burgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ 2009;339:b3244.
6. Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol 2007;36:847–57.
7. Altman R, Brandt K, Hochberg M, Moskowitz R, Bellamy N, Bloch DA, et al. Design and conduct of clinical trials in patients with osteoarthritis: recommendations from a task force of the Osteoarthritis Research Society. Results from a workshop. Osteoarthritis Cartilage 1996;4:217–43.
8. Pham T, van der Heijde D, Altman RD, Anderson JJ, Bellamy N, Hochberg M, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. Osteoarthritis Cartilage 2004;12:389–99.
9. Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006;20:721–40.
10. Reichenbach S, Sterchi R, Scherer M, Trelle S, Burgi E, Burgi U, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. Ann Intern Med 2007;146:580–90.
11. Elbourne DR, Altman DG, Higgins JP, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: methodological issues. Int J Epidemiol 2002;31:140–9.
12. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

13. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999;18:2693–708.
14. Rintelen B, Neumann K, Leeb BF. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. Arch Intern Med 2006;166:1899–906.
15. Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, et al. Glucosamine therapy for treating osteoarthritis. Cochrane Database Syst Rev 2005;2:CD002946.
16. Manheimer E, Linde K, Lao L, Bouter LM, Berman BM. Metaanalysis: acupuncture for osteoarthritis of the knee. Ann Intern Med 2007;146:868–77.
17. Lo GH, LaValley M, McAlindon T, Felson DT. Intra-articular hyaluronic acid in treatment of knee osteoarthritis: a metaanalysis. JAMA 2003;290:3115–21.
18. Fransen M, McConnell S, Bell M. Exercise for osteoarthritis of the hip or knee. Cochrane Database Syst Rev 2003;3:CD004286.
19. Chodosh J, Morton SC, Mojica W, Maglione M, Suttorp MJ, Hilton L, et al. Meta-analysis: chronic disease self-management programs for older adults. Ann Intern Med 2005;143:427–38.
20. Christensen R, Bartels EM, Astrup A, Bliddal H. Effect of weight reduction in obese patients diagnosed with knee osteoarthritis: a systematic review and meta-analysis. Ann Rheum Dis 2007;66:433–9.
21. Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain: a systematic review and meta-analysis of randomised placebo-controlled trials. BMC Musculoskelet Disord 2007;8:51.
22. Bjordal JM, Klovning A, Ljunggren AE, Slordal L. Short-term efficacy of pharmacotherapeutic interventions in osteoarthritic knee pain: a meta-analysis of randomised placebo controlled trials. Eur J Pain 2007;11:125–38.
23. Bartels EM, Lund H, Hagen KB, Dagfinrud H, Christensen R, Danneskiold-Samsoe B. Aquatic exercise for the treatment of knee and hip osteoarthritis. Cochrane Database Syst Rev 2007; 4:CD005523.
24. Avouac J, Gossec L, Dougados M. Efficacy and safety of opioids for osteoarthritis: a meta-analysis of randomized controlled trials. Osteoarthritis Cartilage 2007;15:957–65.
25. Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. BMJ 2004;328:432.

26. Pildal J, Chan AW, Hrobjartsson A, Forfang E, Altman DG, Gotzsche PC. Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. BMJ 2005;330:1049.
27. Montori VM, Bhandari M, Devereaux PJ, Manns BJ, Ghali WA, Guyatt GH. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol 2002;55:787–90.
28. Haahr MT, Hrobjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. Clin Trials 2006;3:360–5.
29. Hrobjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. Int J Epidemiol 2007;36:654–63.
30. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352:609–13.
31. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001;135:982–9.
32. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2008;149:219.
33. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7:1–76.
34. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973–82.
35. Fenwick J, Needleman IG, Moles DR. The effect of bias on the magnitude of clinical outcomes in periodontology: a pilot study. J Clin Periodontol 2008;35:775–82.
36. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.
37. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134: 657–62.

Article 3

Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study

Eveline Nüesch,[1,2)] Sven Trelle[1,2)] Stephan Reichenbach,[1,3)] Anne W.S. Rutjes,[1,4)] Beatrice Tschannen,[1)] Douglas G. Altman,[5)] Matthias Egger,[1)] Peter Jüni,[1,2)]

From [1)]Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2)]CTU Bern, Bern University Hospital, Switzerland; [3)]Department of Rheumatology, Immunology and Allergology, Bern University Hospital, Switzerland; [4)]Laboratory of Clinical Epidemiology of Cardiovascular Disease, Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Santa Maria Imbaro, Italy; [5)]Centre for Statistics in Medicine, University of Oxford, United Kingdom

This article has been submitted.

Abstract

Objective To examine the presence and extent of small study effects in clinical osteoarthritis research.

Design Meta-epidemiological study.

Data sources 13 meta-analyses including 153 randomised trials (41,605 patients) that compared therapeutic interventions with placebo or non-intervention control in patients with osteoarthritis of the hip or knee and used patient reported pain as an outcome.

Methods We compared estimated treatment benefits between large trials of at least 100 patients per arm and small trials, explored funnel plots supplemented with lines of predicted effects and contours of statistical significance and used three approaches to estimate treatment effects: meta-analyses including all trials irrespective of sample size, meta-analyses restricted to large trials and treatment effects predicted for large trials.

Results On average, treatment effects were more beneficial in small than in large trials (difference in effect sizes, -0.21, 95%-CI -0.34 to -0.08, P=0.001). Depending on criteria used, six to eight funnel plots were suggestive of small study effects. In 6 of 13 meta-analyses, the overall pooled estimate suggested a clinically relevant, statistically significant treatment benefit, whereas analyses restricted to large trials and predicted effects in large trials yielded smaller, non-significant estimates.

Conclusions Small study effects may frequently distort results of meta-analyses. The influence of small trials on estimated treatment effects should be routinely assessed.

Introduction

The methodological quality and unbiased dissemination of clinical trials is crucial for the validity of systematic reviews and meta-analyses. It has been repetitively suggested that small trials tend to report larger treatment benefits than larger trials.[1 2] Such small study effects may result from a combination of lower methodological quality of small trials, publication and other reporting biases,[2-8] but could also reflect clinical heterogeneity if small trials were more careful in selecting patients and implementing the experimental intervention.[9] The funnel plot is a scatter plot of treatment effects against standard error as a measure of statistical precision.[9 10] Imprecision of estimated treatment effects will increase as the sample size of component trials increases. Thus, in the absence of small study effects, results from small trials with large standard errors will scatter widely at the bottom of a funnel plot while the spread narrows with increasing sample size and the plot will resemble a symmetrical inverted funnel. Conversely, if small study effects are present, funnel plots will be asymmetrical.[9] The plot can be enhanced by lines of the predicted treatment effect from meta-regression using the standard error as explanatory variable,[11 12] and contours that divide the plot into areas of statistical significance and non-significance.[13 14] A recent study of anti-depressant trials[15] found that these approaches increased the understanding of the interplay of several biases associated with small sample size, including publication bias, selective reporting of outcomes and inadequate methodology and analysis of trials.[14]

Small study effects may be particularly prominent in osteoarthritis research, where several recent meta-analyses found pronounced asymmetry of funnel plots.[16-18] We previously studied the influence of methodological characteristics on estimated effects in a set of clinical osteoarthritis trials using patient-reported pain outcomes and found that deficiencies in concealment of random allocation, patient blinding and analyses may distort the results in these trials.[19 20] Different components of inadequate trial methodology often concur. A trial with adequate allocation concealment for example, is more likely to report analyses according to the intention-to-treat principle.[19 20] Meta-epidemiological studies found that smaller trials are less likely to use adequate random sequence generation, adequate allocation concealment and double blinding,[7 8 19] and that different methodological components are associated with exaggerated treatment benefits.[7 8 19-23]

Here, we explore the presence and extent of small study effects in meta-analyses of osteoarthritis trials and determine whether sensitivity analyses based on a restriction of meta-analyses to large, appropriately powered trials or based on a prediction of treatment effects in large trials influences conclusions of meta-analyses.

Methods

Selection of meta-analyses and component trials

We included meta-analyses of randomised or quasi-randomised, controlled trials in patients with osteoarthritis of the knee or hip. Meta-analyses were eligible, if they included a patient-reported pain-related outcome for any intervention compared to placebo, sham or no control intervention. Two reviewers independently evaluated reports of meta-analyses for eligibility. Details of the search strategy and selection process are described elsewhere.[20] Reports of all component trials of included meta-analyses were obtained. No language restrictions were applied.

Data extraction and quality assessment

Data of individual trials regarding design, interventions, publication year, trial size, sample size calculation, exclusions, and results were extracted independently by two reviewers on a standardized form as previously described.[20] The primary outcome was pain. If different pain-related outcomes were reported, we extracted one pain-related outcome per study according to a pre-specified hierarchy.[16 19 24] Concealment of treatment allocation was considered adequate if investigators responsible for patient selection were unable to suspect before allocation which treatment was next, e.g. central randomisation or sequentially numbered, sealed, opaque envelopes. Blinding of patients was considered adequate if experimental and control interventions were described as indistinguishable or if a double-dummy technique was used. Handling of incomplete outcome data was considered adequate if all randomised patients were included in the analysis (intention-to-treat principle). A cut-off of 100 allocated patients per treatment arm was used to distinguish between small and large trials. A sample size of 2x100 patients will yield more than 80% power to detect a small to moderate effect size of -0.40 at a two-sided α=0.05, which corresponds to a difference of 1 cm on a 10 cm visual analogue scale between experimental and control intervention in a two-arm trial.

Data synthesis

We expressed treatment effects as effect sizes by dividing the difference in mean values at the end of follow-up by the pooled standard deviation. Negative effect sizes indicate a beneficial effect of the experimental intervention. If some required data were unavailable, we used approximations as previously described.[16] Meta-analyses including exclusively small or exclusively large trials did not contribute to the analysis. Within each meta-analysis, we estimated effect sizes of large (≥100 patients per trial arm) and small trials (<100 patients per

trial arm) separately using inverse-variance random-effects meta-analysis, calculated the DerSimonian and Laird estimate of the variance τ^2 as a measure of between-trial heterogeneity,[25 26] and derived differences between pooled estimates of large and small trials. We then combined these differences across meta-analyses using an inverse-variance random-effects model, which fully allowed for heterogeneity between meta-analyses.[26 27] Negative differences in effect sizes indicate that small trials show more beneficial treatment effects than large trials. The variability between meta-analyses was expressed as the heterogeneity variance τ^2. To account for the correlation between sample size and methodological quality, we used stratification by these components in analogy to Mantel-Haenszel procedures[28] and derived differences between small and large trials adjusted for concealment of allocation, patient blinding and intention-to-treat analysis. We performed analyses of associations between sample size and estimated treatment benefits stratified according to the following pre-specified characteristics:[20] between trial heterogeneity in the overall meta-analysis (low, $\tau^2<0.06$, v high, $\tau^2 \geq 0.06$), treatment benefit in the overall meta-analysis (small, effect sizes >-0.5, v large, effect sizes ≤ -0.5),[24 29] and type of intervention assessed in the meta-analysis (drug versus other interventions, conventional versus complementary medicine). These stratified analyses were accompanied by interaction tests.

We drew funnel plots (effect sizes of individual trials plotted against their standard errors) that were enhanced by contours that divide the plot into areas of statistical significance and non-significance at the traditional level of $\alpha=0.05$ based on standard Wald tests as previously described.[13 30] If trials seem to be missing in areas of statistical non-significance, then this adds to the notion of the presence of bias.[13 14] We added lines of the predicted treatment effect to the funnel plots derived from univariable random-effects meta-regression models using the standard error as explanatory variable.[11 12] Then, we assessed funnel plot asymmetry with regression tests, a weighted linear regression of the effect sizes on their standard errors, using the inverse of the variance of effect sizes as weights.[2 9]

We compared pooled effect sizes from overall random-effects meta-analyses, pooled effect sizes from random-effects meta-analyses restricted to large trials only, and predicted effect sizes from random-effects meta-regression models using the standard error as explanatory variable for trials with a standard error of 0.1.[12 14] A standard error of 0.1 is found in a large two-arm trial with 200 randomised patients per group, which will have more than 95% power to detect an effect size of about -0.40 standard deviation units, which corresponds to the median minimal clinically important difference found in recent trials in patients with osteoarthritis.[31-34] Results were considered concordant if point estimates differed by less than

0.10 standard deviation units[35] and if the status of statistical significance at a two-sided α=0.05 remained unchanged, as indicated by the presence or absence of an overlap of the 95% confidence interval with the null effect. Finally, we compared pooled effect sizes, between-trial heterogeneity, precision defined as the inverse of the standard error, and P values for pooled effect sizes between random-effects meta-analyses including all trials and meta-analyses including large trials only, using Wilcoxon's rank tests for paired observations. All P values are two-sided. All data analysis was performed in STATA version 10 (Stata Corporation, College Station, Texas).

Results

The study sample and its origin were described elsewhere.[19 20] 21 eligible meta-analyses described in 17 reports were eligible. Of these, 13 meta-analyses[16 36-46] (153 trials with 41,605 patients) included both, small and large trials and contributed to the analyses. The median number of trials included per meta-analysis was 12 (range 3 to 24) and the median number of patients 1849 (347 to 13659). The pooled effect sizes ranged from -0.07 to -1.11 and the heterogeneity between trials from a τ^2 of 0.00 to 0.47. Eight meta-analyses assessed drug interventions and 5 meta-analyses non-drug interventions. Four assessed interventions in complementary medicine and 9 interventions in conventional medicine.

Table 1 describes the characteristics of the 153 component trials. 58 (38%) trials included at least 100 patients per arm and 95 (62%) trials were smaller. The number of allocated patients ranged from 201 to 2957 in large trials, and from 8 to 362 in small trials. Large trials were published more recently (P=0.002), were more likely to report adequate concealment of allocation (P=0.010) and to report a sample size calculation (P<0.001).

Figure 1 shows a forest plot of differences in effect sizes between small and large trials across the 13 meta-analyses. The average difference in effect sizes between large and small trials across the 13 included meta-analyses was -0.21, with more beneficial effects found in small trials (95%-CI -0.34 to -0.08, P=0.001). At the level of individual meta-analyses, tests for interaction between treatment benefits and trial size were positive in 4 meta-analyses (31%).[16 37 39 45] The variability across meta-analyses was small to moderate, with a τ^2 estimate of 0.03 (P=0.005).

	Number of allocated patients		
	<100 per arm (n=95)	**≥ 100 per arm** (n=58)	**P value**
Concealment of allocation			0.010
Adequate	19 (20%)	22 (38%)	
Inadequate / unclear	76 (80%)	36 (62%)	
Blinding of patients			0.25
Adequate	41 (43%)	30 (52%)	
Inadequate / unclear	54 (57%)	28 (48%)	
Intention-to-treat analysis			0.23
Yes	16 (17%)	16 (28%)	
No / unclear	79 (83%)	42 (72%)	
Sample size calculation			<0.001
Reported	37 (39%)	38 (66%)	
Not reported	58 (61%)	20 (34%)	
Year of publication			0.002
1980 – 1999	55 (58%)	14 (24%)	
2000 – 2007	40 (42%)	44 (76%)	
Drug intervention			0.97
Yes	70 (74%)	43 (74%)	
No	25 (26%)	15 (26%)	
Complementary medicine			0.09
Yes	30 (32%)	11 (19%)	
No	65 (68%)	47 (81%)	

Table 1 Comparison of characteristics between small and large trials P values are derived from logistic regression models adjusted for clustering of trials within meta-analyses. Drug interventions include chondroitin, diacerein, glucosamine, NSAIDs, opioids, paracetamol and viscosupplementation. Interventions in complementary medicine include acupuncture, balneotherapy, chondroitin, and glucosamine.

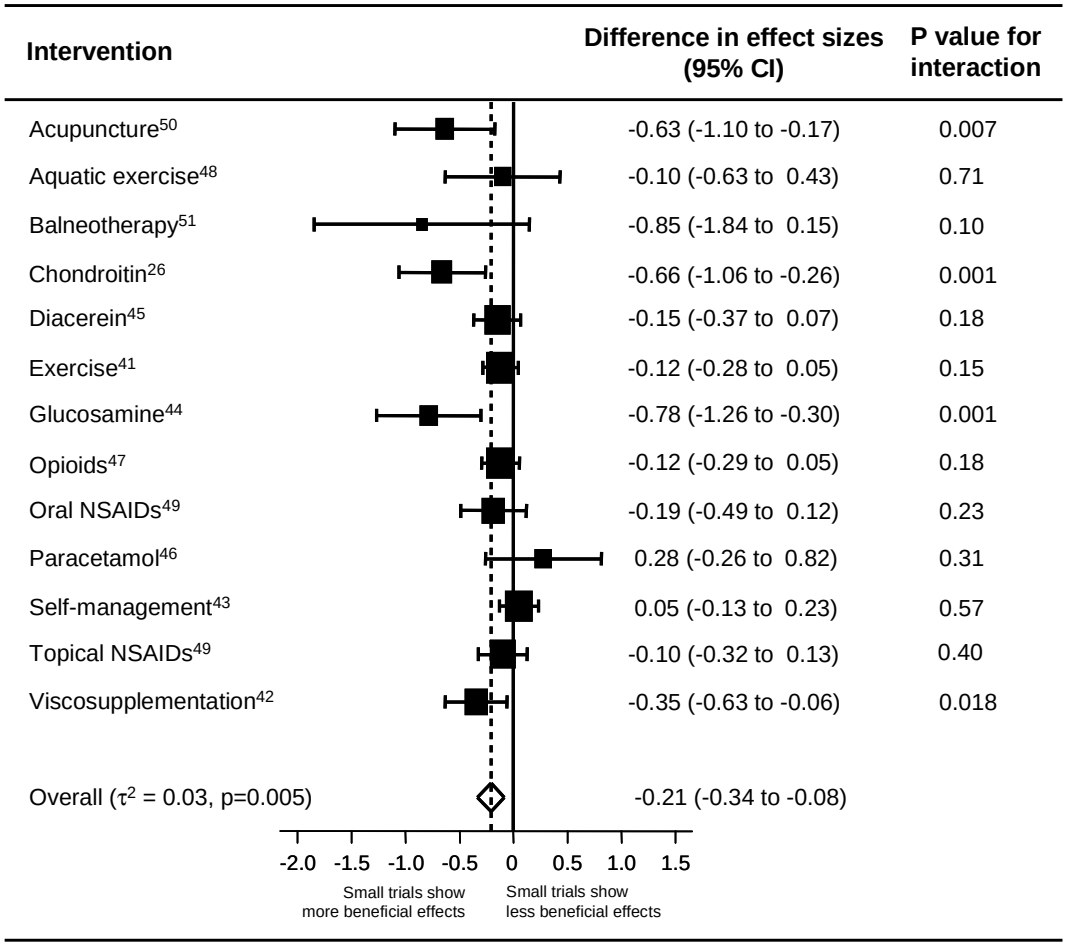

Figure 1 Difference in effect sizes between 95 small trials with less than 100 patients per arm and 58 large trials. A negative difference in effect sizes indicates that small trials show more beneficial treatment effects. P values are for interaction between sample size and effect sizes. NSAIDs=non-steroidal anti-inflammatory drugs.

	Estimates of small study effects		
	Difference in effect sizes		Variability
	ΔES (95% CI)	P value	τ^2 (P value)
Overall, crude	-0.21 (-0.34 to -0.08)	0.001	0.03 (P=0.005)
Adjusted for methodological component:			
Concealment of allocation	-0.16 (-0.27 to -0.06)	0.002	0.02 (P=0.06)
Blinding of patients	-0.21 (-0.33 to -0.09)	0.001	0.03 (P=0.010)
Intention-to-treat analysis	-0.12 (-0.21 to -0.02)	0.016	0.01 (P=0.18)

Table 2 95% CI: 95% confidence interval; ΔES: difference in effect size between 95 small and 58 large trials; τ^2: between meta-analyses heterogeneity variance estimate.

Table 2 presents average difference in effect sizes between large and small trials, crude (top) and after adjustment for the methodological quality of trials (bottom). Differences in effect sizes between small and large trials were robust after adjustment for blinding of patients (-0.21, 95% CI -0.33 to -0.09, P=0.001), slightly attenuated after adjustment for concealment of allocation (-0.16, 95% CI -0.27 to -0.06, P=0.002), but nearly halved after adjustment for intention-to-treat analysis (-0.12, 95% CI -0.21 to -0.02, P=0.016). The variability across meta-analyses was similar between crude and adjusted analyses.

Table 3 presents results from analyses stratified according to the magnitude of treatment effects, the between-trial heterogeneity found in overall meta-analyses, and according to type of experimental intervention. Differences in effect sizes between large and small trials were most pronounced in meta-analyses with large treatment benefits, meta-analyses with a high degree of between-trial heterogeneity and meta-analyses of complementary interventions (P for interaction all <0.001).

Figure 2 shows funnel plots of all 13 meta-analyses including prediction lines from meta-regression models with the standard error as an explanatory variable and 5% contour areas to display areas of significance and non-significance. For six funnel plots, both, the scatter of effect estimates and the prediction line indicated asymmetry (Panels A, D, G, H, L, M).[16 37 39 42 44 45] For another two funnel plots, mainly the prediction lines suggested asymmetry (Panels C and E),[40 46] whereas the remaining 5 funnel plots appeared symmetrical and prediction lines nearly upright (Panels B, F, I, J, K).[36 38 41 43 44] The regression test was statistically significant at P≤0.05 in four meta-analyses (Panels D, G, H, M)[16 37 39 42] and showed a statistical trend in another two (P≤0.10, Panels A, L).[44 45] In 5 funnel plots, the contours to distinguish between areas of statistical significance and non-significance at P=0.05, suggested missing trials in areas of non-significance (Panels A, C, D, H, L).[16 42 44-46]

Table 3 Stratified analyses

Comparison	No of meta-analyses	No of trials	Difference in effect sizes (95% CI)	Variability (P value)	P for interaction
Overall	13	153	-0.21 (-0.34 to -0.08)	0.03 (0.005)	
Treatment benefit in overall meta-analysis					<0.001
Small	10	115	-0.13 (-0.22 to -0.03)	0.01 (0.17)	
Large	3	38	-0.72 (-1.02 to -0.43)	0.00 (0.90)	
Heterogeneity between trials in overall meta-analysis					<0.001
Low	8	87	-0.08 (-0.16 to -0.00)	0.00 (0.66)	
High	5	66	-0.55 (-0.73 to -0.36)	0.00 (0.46)	
Pharmacological intervention					0.67
Yes	8	113	-0.23 (-0.39 to -0.08)	0.03 (0.021)	
No	5	40	-0.17 (-0.40 to 0.06)	0.03 (0.041)	
Complementary medicine					<0.001
Yes	4	41	-0.70 (-0.95 to -0.45)	0.00 (0.96)	
No	9	112	-0.10 (-0.18 to -0.03)	0.00 (0.43)	

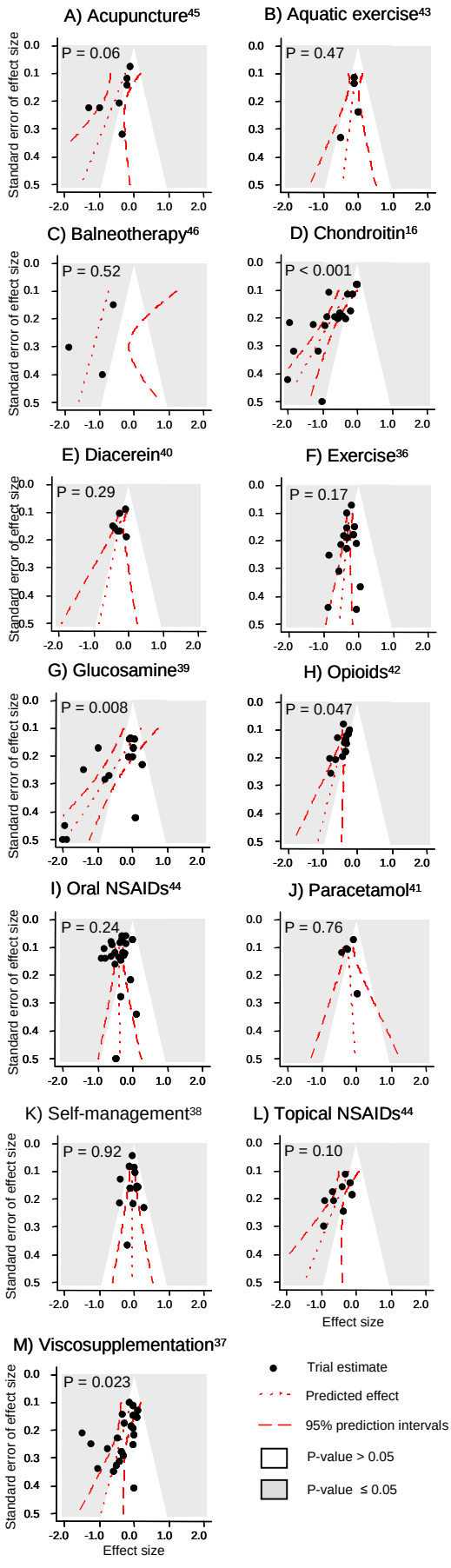

Figure 2 Funnel plots of 13 included meta-analyses including prediction lines from univariable meta-regression models with the standard error as explanatory variable (red) and 5% contour areas to display areas of significance (grey) and non-significance (white). P-values were derived from regression tests for asymmetry. NSAIDs=non-steroidal anti-inflammatory drugs.

Figure 3 presents a graphical summary of results of individual meta-analyses of all trials (black), meta-analyses restricted to large trials (blue) and predicted effect sizes for trials with a standard error of 0.1 (red). Results of all three analytical approaches were concordant in 7 meta-analyses (Figure 3, Panels B, E, F, H, I, J, K).[36 38 41-44] In the remaining 6, both approaches, the restricted analysis and the predicted effect were discordant to the overall analysis (Panels A, C, D, G, L, M).[16 37 39 44-46] In 3 of these, statistical significance at the conventional level of 0.05 was lost when restricting the analysis to large trials and when predicting the effect (Panels D, G, M), in the other 3, significance was lost when predicting the effect, but not when restricting the analysis (Panels A, C, L).[44-46] The median estimated treatment benefit decreased from -0.39 (range -1.11 to -0.06) in meta-analyses of all trials to -0.23 (range -0.59 to -0.04) in meta-analyses restricted to large trials (P=0.005) and the median between-trial heterogeneity decreased from a τ^2 of 0.20 (range 0.00 to 0.69) to a τ^2 of 0.04 (range 0.00 to 0.31, P=0.030). P-values of pooled effect sizes increased from a median of <0.001 (range <0.001 to 0.13) to 0.007 (range <0.001 to 0.61, P=0.016) in restricted meta-analyses, whereas precisions of pooled effect sizes were much the same (median 13 [range 2 to 24] versus 14 [range 7 to 21], P=0.70).

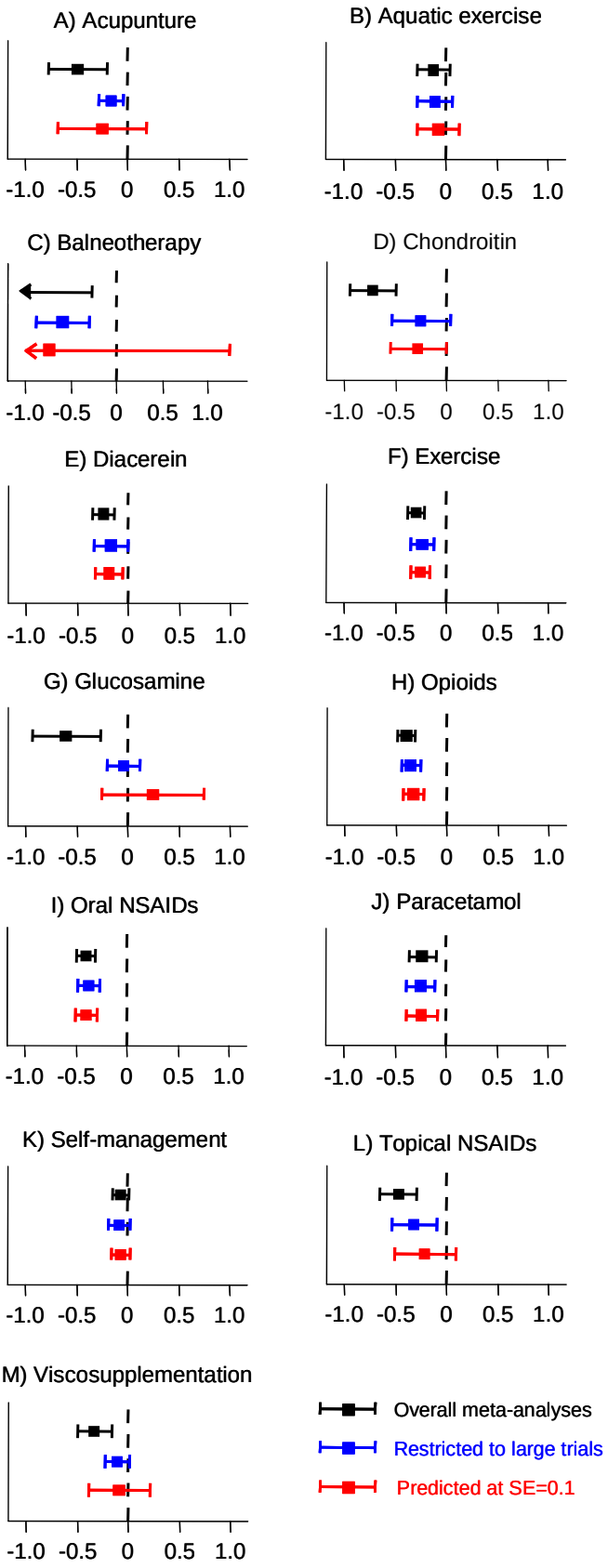

Figure 3. Results of individual random-effects meta-analyses of all trials (black), results of random-effects meta-analyses restricted to large trials with at least 100 patients per arm (blue), and effect sizes for trials with a standard error of 0.1 predicted from random-effects meta-regression models (red). NSAIDs = non-steroidal anti-inflammatory drugs, SE = standard error

Discussion

In this meta-epidemiological study in 13 meta-analyses of 153 osteoarthritis trials, we found larger estimated treatment benefits in small trials of less than 100 patients per trial arm as compared with larger trials. The average difference between small and large trials was about half the magnitude of a typical treatment effect found for interventions in osteoarthritis.[24] Small study effects were more prominent in 5 of the 13 meta-analyses, however. These showed a large extent of statistical heterogeneity, larger pooled estimates of treatment benefit than would typically be expected from effective intervention in osteoarthritis, and mainly covered complementary medical interventions. Taking into account contours used to distinguish between areas of statistical significance and non-significance, and lines of treatment effects predicted for different standard errors, we found 8 funnel plots suggestive of small study effects. Finally, we used three different approaches to estimate treatment effects of the 13 interventions included in this study: pooling all trials irrespective of sample size, restricting the analysis to large trials of more than 100 patients per trial arm, and predicting treatment effects for large trials using the corresponding standard error as independent variable. Estimates from these three approaches were discordant in 6 meta-analyses, with the overall pooled estimate suggesting a clinically relevant, statistically significant treatment benefit, which was not found in the other two approaches aimed at estimating the effect in large trials only.

Large trials tend to be of higher quality than small trials and the observed association between sample size and treatment effect could be confounded by methodological quality.[7 8 47] When accounting for patient blinding, we found the association between sample size and treatment effect completely robust. Accounting for concealment of allocation resulted in a slight attenuation, whereas adjusting for the presence or absence of an intention-to-treat analysis nearly halved the association between sample size and treatment effect. This suggests that problems with exclusions from the analysis after randomisation may contribute to the observed small study effects, which is in line with a recent study of anti-depressant trial.[15] This study suggested that, in addition to publication and reporting biases, switching from an intention-to-treat to a per protocol analysis contributed to discrepancies between published and unpublished results.[14] The assessment of components of methodological quality will depend strongly on reporting quality[48] and may be affected by misclassification, whereas sample size or standard error may be extracted more easily. Sample size or statistical precision may therefore be the best single proxy for the cumulative effect of the different sources of bias in randomised osteoarthritis trials and probably also in other fields: selection,

performance, detection and attrition bias,[49] selective reporting of outcomes[3 4] and publication bias.[6 50]

The most important limitation of our study is that we cannot exclude alternative explanations of small study effects other than bias: smaller trials may have been more careful in implementing the intervention or in including patients who are particularly likely to benefit from the intervention, which could result in larger treatment effects and true clinical heterogeneity.[2 9 51] In addition, the selection of component trials was based on the literature searches and selection criteria of published meta-analyses. Some of the searches in these meta-analyses may have been too superficial and some of the selection criteria too narrow to include a large proportion of unpublished trials. However, the meta-analyses included in our study are likely to be representative of the field and we believe therefore that our results are generalisable. Another limitation is that our analysis is based on published information only and depends on the quality of reporting, which is often unsatisfactory.[49]

To our knowledge, this is the first meta-epidemiological study to systematically assess small study effects in a series of meta-analyses with continuous clinical outcomes. In an analysis of trials with binary outcomes, Kjaergard et al[7 8] found more beneficial treatment effects in small trials with inadequate methodology as compared with large trials. Shang et al, in an analysis of homeopathy trials, found smaller trials and those of lower quality to show more beneficial treatment effects than larger and higher-quality trials.[11] Moreno et al recently assessed the performance of contour enhanced funnel plots and a regression based adjustment method to detect and adjust for small study effects in placebo-controlled antidepressant trials previously submitted to the US Food and Drug Administration (FDA) and matching journal publications.[14] Applying the regression based adjustment method to the journal data produced a similar pooled effect to that observed by a meta-analysis of the complete unbiased FDA data. In contrast to our study, they regressed treatment effects against their variance. In funnel plots, treatment effects will typically be plotted against their standard error, however, and significance tests will be generally based on z or t values, which again are calculated by dividing treatment effects by their standard error. Therefore, we deem it preferable to regress treatment effects against the standard error rather than the variance. A second discrepancy is that Moreno et al predicted effects for infinitely large trials of a variance of zero. By definition, such a trial would be overpowered to detect a minimally clinically relevant difference between groups and we deem it preferable to predict treatment effects for large trials with adequate power to detect small, albeit relevant effects. The chosen standard error of 0.1 at which treatment effects will be found in a large two-arm trial with a continuous primary

outcome including 200 randomised patients per group to yield more than 95% power to detect an effect size of -0.40 standard deviation units and still more than 80% power to detect an effect size of about -0.30 standard deviation units. Trials considerably larger than that will hardly be needed in case of a continuous primary outcome.

The meta-regression model used to predict effects incorporates residual heterogeneity unexplained by regressing treatment effect against standard error. In case of large unexplained heterogeneity, it will appropriately indicate uncertainty in the predicted estimate as reflected by a wide 95% prediction interval, even though an analysis restricted to large trials may yield precise estimates. This was observed in 5 meta-analyses of our study[37 39 44-46] and taken as an indication of residual uncertainty necessitating additional explorations of sources of heterogeneity or additional, appropriately designed large scale trials. For continuous outcomes, definitions of large trials and methodologies used for assessing funnel plot asymmetry may be generally suitable as reported here. Trials with an average of 100 patients per trial arm will yield about 80% power to detect a small to moderate effect size of -0.40 standard deviation units, which corresponds to the median minimal clinically important difference found in recent studies in patients with osteoarthritis.[32-34] For binary outcomes, the definition of large trials will depend on control group event rates and a definition of what constitutes a moderate, but clinically relevant effect. In addition, the regression test for funnel plot asymmetry originally reported[9] may be associated with an inappropriately high rate of false positives if odds ratios or risk ratios are used. Therefore, a modification of the test should be considered as reported by Harbord et al.[51] Non-parametric tests will result in lower power than the regression tests discussed here and may be less appropriate.

An inspection of funnel plots and stratified analyses according to sample size accompanied by appropriate interaction tests should be considered routine procedures in any meta-analysis, possibly accompanied by a regression test for funnel plot asymmetry and prediction of effects in large trials using meta-regression.[47] In the presence of asymmetry of funnel plots, systematic reviewers should also report meta-analyses restricted to large trials or effects predicted for large trials. Readers and clinicians should be careful in interpreting results of small trials of low methodological quality and meta-analyses including mainly such trials.

Funding: Swiss National Science Foundation (grant Nos 4053-40-104762/3 and 3200-066378) to PJ and SR. The study was part of the Swiss National Science Foundation's National Research Programme 53 on musculoskeletal health. SR was a recipient of a research fellowship funded by the Swiss National Science Foundation (grant No PBBEB-115067). DGA was supported by Cancer Research UK. PJ was a PROSPER (programme for social medicine, preventive and epidemiological research) fellow funded by the Swiss National Science Foundation (grant No 3233-066377). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all data of the study and had final responsibility for the decision to submit for publication. None of the authors is affiliated with or funded by any manufacturer of any intervention used for osteoarthritis.

Acknowledgments: We thank Sacha Blank, Elizabeth Bürgi, Liz King, Katharina Liewald, Linda Nartey, Martin Scherer and Rebekka Sterchi for contributing to data extraction. We are grateful to Malcolm Sturdy for the development and maintenance of the database.

Author Contributions: EN and PJ conceived the study and developed the protocol. EN, ST, SR, AWSR, and BT were responsible for the acquisition of the data. EN and PJ did the analysis and interpreted the analysis in collaboration with ST, SR, AWSR, BT, DGA and ME. EN and PJ wrote the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version of the manuscript. PJ and SR obtained public funding. PJ provided administrative, technical, and logistic support. EN and PJ are the guarantors for the study.

References

1. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001;323(7304):101-5.
2. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;53(11):1119-29.
3. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330(7494):753.
4. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291(20):2457-65.
5. Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316(7124):61-6.
6. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009(1):MR000006.
7. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2008;149(3):219.
8. 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.
9. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34.
10. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54(10):1046-55.
11. Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, Sterne JA, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 2005;366(9487):726-32.
12. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009;9:2.
13. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol* 2008;61(10):991-6.

14. Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ* 2009;339:b2981.
15. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358(3):252-60.
16. Reichenbach S, Sterchi R, Scherer M, Trelle S, Burgi E, Burgi U, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med* 2007;146(8):580-90.
17. Rutjes AW, Nuesch E, Sterchi R, Kalichman L, Hendriks E, Osiri M, et al. Transcutaneous electrostimulation for osteoarthritis of the knee. *Cochrane Database Syst Rev* 2009(4):CD002823.
18. Vlad SC, LaValley MP, McAlindon TE, Felson DT. Glucosamine for pain in osteoarthritis: why do trial results differ? *Arthritis Rheum* 2007;56(7):2267-77.
19. Nuesch E, Reichenbach S, Trelle S, Rutjes AWS, Liewald K, Sterchi R, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum* 2009;61(12):1633-1641.
20. Nuesch E, Trelle S, Reichenbach S, Rutjes AWS, Bürgi E, Scherer M, et al. The effects of the excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ* 2009;339:b3244.
21. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-412.
22. Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 2007;36(4):847-57.
23. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, 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.
24. Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. *Best Pract Res Clin Rheumatol* 2006;20(4):721-40.
25. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.
26. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;18(20):2693-708.

27. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
28. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959;22:719-748.
29. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum, 1988.
30. Palmer TM, Peters JL, Sutton AJ, Moreno SG. Contour-enhanced funnel plots for meta-analysis. *The Stata Journal* 2008;8(2):242-254.
31. Eberle E, Ottillinger B. Clinically relevant change and clinically relevant difference in knee osteoarthritis. *Osteoarthritis Cartilage* 1999;7(5):502-3.
32. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum* 2001;45(4):384-91.
33. Angst F, Aeschlimann A, Michel BA, Stucki G. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol* 2002;29(1):131-8.
34. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain* 2004;8(4):283-91.
35. Tendal B, Higgins JP, Juni P, Hrobjartsson A, Trelle S, Nuesch E, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ* 2009;339:b3128.
36. Fransen M, McConnell S, Bell M. Exercise for osteoarthritis of the hip or knee. *Cochrane Database Syst Rev* 2003(3):CD004286.
37. Lo GH, LaValley M, McAlindon T, Felson DT. Intra-articular hyaluronic acid in treatment of knee osteoarthritis: a meta-analysis. *JAMA* 2003;290(23):3115-21.
38. Chodosh J, Morton SC, Mojica W, Maglione M, Suttorp MJ, Hilton L, et al. Meta-analysis: chronic disease self-management programs for older adults. *Ann Intern Med* 2005;143(6):427-38.
39. Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, et al. Glucosamine therapy for treating osteoarthritis. *Cochrane Database Syst Rev* 2005(2):CD002946.

40. Rintelen B, Neumann K, Leeb BF. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. *Arch Intern Med* 2006;166(17):1899-906.
41. Towheed TE, Maxwell L, Judd MG, Catton M, Hochberg MC, Wells G. Acetaminophen for osteoarthritis. *Cochrane Database Syst Rev* 2006(1):CD004257.
42. Avouac J, Gossec L, Dougados M. Efficacy and safety of opioids for osteoarthritis: a meta-analysis of randomized controlled trials. *Osteoarthritis Cartilage* 2007;15(8):957-65.
43. Bartels EM, Lund H, Hagen KB, Dagfinrud H, Christensen R, Danneskiold-Samsoe B. Aquatic exercise for the treatment of knee and hip osteoarthritis. *Cochrane Database Syst Rev* 2007(4):CD005523.
44. Bjordal JM, Klovning A, Ljunggren AE, Slordal L. Short-term efficacy of pharmacotherapeutic interventions in osteoarthritic knee pain: A meta-analysis of randomised placebo-controlled trials. *Eur J Pain* 2007;11(2):125-38.
45. Manheimer E, Linde K, Lao L, Bouter LM, Berman BM. Meta-analysis: acupuncture for osteoarthritis of the knee. *Ann Intern Med* 2007;146(12):868-77.
46. Verhagen AP, Bierma-Zeinstra SM, Boers M, Cardoso JR, Lambeck J, de Bie RA, et al. Balneotherapy for osteoarthritis. *Cochrane Database Syst Rev* 2007(4):CD006864.
47. Nuesch E, Juni P. Commentary: Which meta-analyses are conclusive? *Int J Epidemiol* 2009;38(1):298-303.
48. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;287(21):2801-4.
49. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.
50. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev* 2007(2):MR000010.
51. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;25(20):3443-57.

Article 4

Transcutaneous electrical nerve stimulation for osteoarthritis of the knee: a systematic review and meta-analysis

Anne W.S. Rutjes,[1)] Eveline Nüesch,[1,2)] Rebekka Sterchi,[1)] Leonid Kalichman,[3)] Erik Hendriks,[4)] Stephan Reichenbach,[1,5)] Manathip Osiri,[6)] Lucie Brosseau,[7)] Peter Jüni[1,2)]

From [1)]Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2)]CTU Bern, Bern University Hospital, Switzerland; [3)]Department of Physical Therapy, Recanati School for Community Health Professions, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel; [4)]Epidemiology Department, Maastricht University, Maastricht, The Netherlands; [5)]Department for Rheumatology, Clinical Immunology, and Allergology, University Hospital, Bern, Switzerland; [6)]Department of Medicine, Faculty of Medicine, Bangkok, Thailand; [7)]School of Rehabilitation Sciences, Faculty of Health Sciences, University of Ottawa, Ottawa, Canada

The first two authors contributed equally to the manuscript.

Cochrane Database of Systematic Reviews 2009(4):CD002823.

Abstract

Background Osteoarthritis is the most common form of joint disease and the leading cause of pain and physical disability in the elderly. Transcutaneous electrical nerve stimulation (TENS), interferential current stimulation and pulsed electrostimulation are used widely to control both acute and chronic pain arising from several conditions, but some policy makers regard efficacy evidence as insufficient.

Objectives To compare transcutaneous electrostimulation with sham or no specific intervention in terms of effects on pain and withdrawals due to adverse events in patients with knee osteoarthritis.

Search strategy We updated the search in CENTRAL, MEDLINE, EMBASE, CINAHL and PEDro up to 5 August 2008, checked conference proceedings and reference lists, and contacted authors.

Selection criteria Randomised or quasi-randomised controlled trials that compared transcutaneously applied electrostimulation with a sham intervention or no intervention in patients with osteoarthritis of the knee.

Data collection and analysis We extracted data using standardised forms and contacted investigators to obtain missing outcome information. Main outcomes were pain and withdrawals or dropouts due to adverse events. We calculated standardised mean differences (SMDs) for pain and relative risks for safety outcomes and used inverse-variance random-effects meta-analysis. The analysis of pain was based on predicted estimates from meta-regression using the standard error as explanatory variable.

Main results In this update we identified 14 additional trials resulting in the inclusion of 18 small trials in 813 patients. Eleven trials used TENS, four interferential current stimulation, one both TENS and interferential current stimulation, and two pulsed electrostimulation. The methodological quality and the quality of reporting was poor and a high degree of heterogeneity among the trials (I^2 = 80%) was revealed. The funnel plot for pain was asymmetrical (P < 0.001). The predicted SMD of pain intensity in trials as large as the largest trial was -0.07 (95% CI -0.46 to 0.32), corresponding to a difference in pain scores between electrostimulation and control of 0.2 cm on a 10 cm visual analogue scale. There was little evidence that SMDs differed on the type of electrostimulation (P = 0.94). The relative risk of being withdrawn or dropping out due to adverse events was 0.97 (95% CI 0.2 to 6.0).

Authors' conclusions In this update, we could not confirm that transcutaneous electrostimulation is effective for pain relief. The current systematic review is inconclusive, hampered by the inclusion of only small trials of questionable quality. Appropriately designed trials of adequate power are warranted.

Background

Osteoarthritis is an age-related condition, occurring more frequently in women than in men. Its prevalence, causal associations and outcomes vary markedly according to the joint site affected.[1] Osteoarthritis is characterised by focal areas of loss of articular cartilage in synovial joints, accompanied by subchondral bone changes, osteophyte formation at the joint margins, thickening of the joint capsule and mild synovitis.[2] The objectives of management of knee osteoarthritis are to relieve pain and to maintain or improve function. Different modalities in physiotherapy have been suggested to improve the clinical course of knee osteoarthritis, with potentially fewer adverse effects than medical treatment,[3 4] but some policymakers consider the evidence for effectiveness to be insufficient.[5]

Transcutaneous electrostimulation, the application of any electrical current through the skin with the aim of pain modulation, is a frequently used modality in knee osteoarthritis.[6 7] It is based on the 'Gate-Control Theory' of pain perception as described by Melzack and Wall.[8] The theory suggests that the stimulation of large diameter, (A-beta) primary sensory afferent cutaneous fibres activates inhibitory interneurons in the spinal cord dorsal horn and, thereby,may attenuate the transmission of nociceptive signals from small diameter A-delta and C fibres. Other suggested mechanisms include a stimulation of β endorphin production.[9-11] and even the potential for articular cartilage repair.[12 13]

Several types of electrostimulation are available. Conventional transcutaneous electrical nerve stimulation (TENS), in its narrow sense, uses moderate to high frequency current of 40 to 150 Hz and 50 to 100 μsec pulse width, typically at a low intensity, to stimulate sensory fibres. Several other types of TENS were subsequently developed, which differ in intensity, pulse width or frequency. Acupuncture-like TENS (AL TENS) uses a low frequency current of 0.5 to 10 Hz and a pulse width of > 150 μsec at a high intensity to stimulate both motor and sensory fibres. The stimulation may be painful, and the intensity of the current will depend on the patient's individual pain tolerance. Burst TENS was developed to minimise patients' discomfort, as experienced with AL TENS. It uses short bursts of high frequency current of typically 80 to 100 Hz, which are repetitively applied at low intensity and a burst frequency of around 5 Hz, to stimulate motor and sensory fibres. The intensity used is slightly higher than used with conventional TENS. Brief TENS uses a high frequency current of more than 100 Hz and 150 to 250 μsec pulse width at the maximal intensity tolerated by the patient to stimulate not only motor and sensory, but also nociceptor fibres. Modulation TENS combines several of the modalities above, typically using alternations of low and high frequency

currents.[14 15] Classical interferential current stimulation simultaneously uses two non-modulated biphasic pulsed currents applied with two sets of electrodes with four electrical poles; one current is fixed at approximately 4000 Hz and the other ranging typically from 4000 to 4100 Hz. The superimposition of the two currents results in a new frequency with a range from 1 to 100 Hz.[16] Modulated interferential current stimulation uses directed currents between two electrical poles and vectorially sums currents in the tissue, with a carrier frequency typically set at 4000 Hz, a beat frequency at 80 Hz, and a modulation frequency set between 0 to 150 Hz. The effective frequency is defined by the sum of beat and modulation frequency and varies between 80 and 230 Hz. The high frequency of the carrier currents in inferential current stimulation leads to a considerably lower impedance of skin and subcutaneous tissue as compared with conventional TENS and minimises patients' discomfort. Lastly, pulsed electrostimulation applies high frequency current of 100 Hz and a pulse width of 640 to 1800 μsec, typically using knee garments with flexible, embedded electrodes and a small battery-operated generator, allowing application times of several hours rather than 15 to 60 minutes, as is the case for any other of the modalities described above.

Objectives

We set out to compare transcutaneous electrostimulation with sham or no specific intervention in terms of effects on pain and function and safety outcomes in patients with knee osteoarthritis and to explore whether potential variation between trials could be explained by characteristics of the electrostimulation, by biases affecting individual trials or by publication bias.

Methods

Criteria for considering studies for this review

Types of studies Randomised or quasi-randomised controlled trials with a control group receiving a sham intervention or no intervention.

Types of participants Studies including at least 75% of patients with clinically and/or radiologically confirmed osteoarthritis of the knee.

Types of interventions Any type of transcutaneous electrostimulation with electrodes set to stimulate nerves supplying the knee joint area aiming at pain relief. We did not consider transcutaneous electrostimulation aiming at muscle strength enhancement, such as neuromuscular electrostimulation, and electrostimulation not directly aimed at stimulating nerves of the knee joint area, such as transcranial applications or transcutaneous spinal electroanalgesia. There were no restrictions related to the type of electrode used.

Types of outcome measures

Main outcomes Main outcomes were pain intensity as the effectiveness outcome[17 18] and withdrawals or drop-outs because of adverse events as the safety outcome. If data on more than one pain scale were provided for a trial, we referred to a previously described hierarchy of pain-related outcomes[1 19] and extracted data on the pain scale that is highest on this hierarchy:

1. Global pain
2. Pain on walking
3. WOMAC osteoarthritis index pain subscore
4. Composite pain scores other than WOMAC
5. Pain on activities other than walking
6. Rest pain or pain during the night
7. WOMAC global algofunctional score
8. Lequesne osteoarthritis index global score
9. Other algofunctional scale
10. Patient's global assessment
11. Physician's global assessment

If pain outcomes were reported at several time points, we extracted the estimate at the end of the treatment period.

Secondary outcomes Secondary outcomes were function, the number of patients experiencing any adverse event and patients experiencing any serious adverse events. We defined serious adverse events as events resulting in hospitalisation, prolongation of hospitalisation, persistent or significant disability, congenital abnormality/birth defect of offspring, life-threatening events or death. If data on more than one function scale were provided for a trial, we extracted data according to the hierarchy presented below.

1. Global disability score
2. Walking disability
3. WOMAC disability subscore
4. Composite disability scores other than WOMAC
5. Disability other than walking
6. WOMAC global scale
7. Lequesne osteoarthritis index global score
8. Other algofunctional scale
9. Patient's global assessment

10. Physician's global assessment

If function outcomes were reported at several time points, we extracted the estimate at the end of the treatment period. For safety outcomes, we extracted end of trial data.

Search methods for identification of studies

Electronic searches We searched the Cochrane Central Register of Controlled Trials (CENTRAL) (The Cochrane Library 2008, issue 3), MEDLINE and EMBASE through the Ovid platform (www.ovid.com), CINAHL through EBSCOhost, Physiotherapy EvidenceDatabase (PEDro, http:/ / www.pedro.fhs.usyd.edu.au/, from 1929 onwards), all from implementation to 5 August 2008, using a combination of keywords and text words related to electrostimulation combined with keywords and text words related to osteoarthritis and a validated filter for controlled clinical trials.[20] The search strategy is presented in Appendix 1 and Appendix 2.

Searching other sources We manually searched conference proceedings, used Science Citation Index to retrieve reports citing relevant articles, contacted content experts and trialists and screened reference lists of all obtained articles, including related reviews. Finally, we searched several clinical trial registries (www.clinicaltrials.gov, www.controlledtrials.com, www.actr.org.au, www.umin.ac.jp/ ctr) to identify ongoing trials. The last update of the manual search was on 2 February 2009.

Data collection and analysis

Selection of studies Two review authors evaluated independently all titles and abstracts for eligibility (see Figure 1). We resolved disagreements by discussion. We applied no language restrictions. If multiple reports described the same trial, we considered all.

Data collection Two review authors (AR and EN, RS or LK) extracted trial information independently using a standardised, piloted data extraction form accompanied by a codebook. We resolved disagreements by consensus or discussion with a third author (SR or PJ).We extracted the type of electrostimulation, including the mode of function (types of stimulator and electrode), the pulse form (intensity, rate and width), the electrode placement site and the frequency and duration of treatment. Other data extracted included the type of control intervention used, patient characteristics (gender, average age, duration of symptoms, type of joint), characteristics of pain, function and safety outcomes, design, trial size, trial duration (defined as time from randomisation until end of follow up), type and source of financial support and publication status. When necessary, we approximated means and measures of dispersion from figures in the reports. For cross-over trials, we extracted data from the first

period only. Whenever possible, we used results from an intention-to-treat analysis. If effect sizes could not be calculated, we contacted the authors for additional data.

Quality assessment Two review authors (AR and EN, RS or LK) independently assessed randomisation, blinding, selective outcome reporting and handling of incomplete outcome data in the analyses.[21 22] We resolved disagreements by consensus or discussion with a third author (SR or PJ). We assessed two components of randomisation: generation of allocation sequences and concealment of allocation. We considered generation of sequences adequate if it resulted in an unpredictable allocation schedule; mechanisms considered adequate included random-number tables, computer-generated random numbers, minimisation, coin tossing, shuffling of cards and drawing of lots. Trials using an unpredictable allocation sequence were considered randomised; trials using potentially predictable allocation mechanisms, such as alternation or the allocation of patients according to date of birth, were considered quasi-randomised. We considered allocation concealment adequate if the investigators responsible for patient selection were unable to suspect before allocation which treatment was next; methods considered adequate included central randomisation and sequentially numbered, sealed, opaque envelopes. We considered blinding of patients adequate if a sham intervention was used that was identical in appearance from the control intervention. Transcutaneous electrostimulation generally does not allow blinding of therapists, whereas pain as the main effectiveness outcome is patient-reported by definition. Therefore, we did not assess blinding of therapists and outcome assessors. We considered handling of incomplete outcome data adequate if all randomised patients were included in the analysis (intention-to-treat principle). Finally, we used GRADE to describe the quality of the overall body of evidence,[22 23] defined as the extent of confidence in the estimated treatment benefits and harms.

Data synthesis We summarised continuous outcomes using standardised mean differences (SMD), with the differences in mean values at the end of treatment across treatment groups divided by the pooled standard deviation. If differences in mean values at the end of the treatment were unavailable, we used differences in mean changes. If some of the required data were unavailable, we used approximations as previously described.[19] A SMD of -0.20 standard deviation units can be considered a small difference between experimental and control group, a SMD of -0.50 a moderate difference, and -0.80 a large difference.[1 24] SMDs can also be interpreted in terms of the percent of overlap of the experimental group's scores with the scores of the control group. A SMD of -0.20 indicates an overlap in the distributions of pain or function scores in about 85% of cases, a SMD of -0.50 in approximately 67% and a SMD of -0.80 in about 50% of cases.[1 24] On the basis of a median pooled SD of 2.5 cm found

in large-scale osteoarthritis trials that assessed pain using a 10 cm visual analogue scale (VAS),[25] SMDs of -0.20, -0.50 and -0.80 correspond to approximate differences in pain scores between experimental and control groups of 0.5, 1.25 and 2.0 cm on a 10 cm VAS. SMDs for function were back transformed to a standardised WOMAC disability score[26] ranging from 0 to 10 on the basis of a median pooled SD of 2.1 units observed in large-scale osteoarthritis.[25] We expressed binary outcomes as relative risks.

We used standard inverse-variance random-effects meta-analysis[27] to combine trials overall and stratified according to gross categories of electrostimulation (TENS, interferential current stimulation or pulsed electrostimulation).We quantified heterogeneity between trials using the I^2 statistic,[28] which describes the percentage of variation across trials that is attributable to heterogeneity rather than to chance and the corresponding χ^2 test. I^2 values of 25%, 50% and 75% may be interpreted as low, moderate and high between-trial heterogeneity, although the interpretation of I^2 depends on the size and number of trials included.[29] The association between trial size and treatment effects was investigated in funnel plots, plotting effect sizes on the vertical axis against their standard errors on the horizontal axis. We assessed asymmetry by the asymmetry coefficient: the difference in effect size per unit increase in standard error,[30] which is mainly a surrogate for sample size, and used uni-variable meta-regression analysis to predict treatment effects in trials as large as the largest trials included in the meta-analysis, using the standard error as the explanatory variable.[31] In view of the biased nature of the predominantly small trials included in the meta-analysis of pain intensity, we considered the predicted estimates of effectiveness more reliable than the pooled estimates.

For the analysis on the effectiveness outcomes pain and function, we differentiated between TENS, interferential current stimulation and pulsed electrostimulation. Then, we performed effectiveness analyses stratified by the following trial characteristics: concealment of allocation, use of a sham intervention in the control group, blinding of patients, analysis in accordance with the intention-to-treat principle, trial size, difference in the use of analgesic cointerventions, specific type of electrostimulation, duration of stimulation per session, number of sessions per week, duration of electrostimulation per week as an overall measure of treatment intensity, and duration of treatment period. A cut-off of 200 patients was used to distinguish between small and large trials; a sample size of 100 patients per group will yield more than 80% power to detect a small to moderate SMD of -0.40 at a two-sided P of 0.05.

For the analysis according to specific type of stimulation, we distinguished between high frequency TENS, burst TENS, modulation TENS, low frequency TENS, interferential current stimulation or pulsed electrostimulation. We classified conventional TENS and brief TENS as

high frequency TENS. Cut-offs of 20 and 60minutes were used for the duration of electrostimulation per session, corresponding to the typical treatment duration in physical therapy, and the optimum stimulation duration suggested by Cheing 2003. A cut-off of four weeks was used for the overall duration of the treatment period (time from randomisation to last session), in line with the previous version on this review. Cut-offs of three and seven were used for the number of sessions per week; one and five hours for the duration of electrostimulation per week, corresponding to the distribution of tertiles. We used uni-variable random-effects meta-regression models to determine whether treatment effects were affected by these factors.[32] Then, we converted SMDs of pain intensity and function to odds ratios[33] to derive numbers needed to treat (NNT) to cause one additional treatment response on pain or function as compared with control, and numbers needed to harm (NNH) to cause one additional adverse outcome. We defined treatment response as a 50% improvement in scores,[34] which corresponds to an average decrease of 1.2 standard deviation units. Based on the median standardised pain intensity at baseline of 2.4 standard deviation units and the median standardised decrease in pain scores of 0.72 standard deviation units observed in large osteoarthritis trials,[25] we calculated that a median of 31% of patients in the control group would achieve an improvement of pain scores of 50% or more. This percentage was used as the control group response rate to calculate NNTs for treatment response on pain. Based on the median standardised WOMAC function score at baseline of 2.7 standard deviation units and the median standardised decrease in function scores of 0.58 standard deviation units,[25] 26% of patients in the control group would achieve a reduction in function of 50% or more. Again, this percentage was used as the control group response rate to calculate NNTs for treatment response on function. We used median risks of 150 patients with adverse events per 1000 patient-years, four patients with serious adverse events per 1000 patient-years and 17 drop-outs due to adverse events per 1000 patient-years observed in placebo groups in large osteoarthritis trials[25] to calculate NNHs for safety outcomes. We performed analyses in RevMan version 5 (RevMan 2008) and STATA version 10.1 (StataCorp, College Station,Texas). All P values are two-sided.

Results

Description of studies We identified 1697 references to articles and considered 85 to be potentially eligible (Figure 1). Twenty-two reports describing 18 completed trials in 813 patients and two protocols describing uncompleted trials[13 35] met our inclusion criteria. Six trials evaluated high frequency TENS,[36-43] one high frequency and burst TENS,[11] one high frequency TENS and interferential current stimulation,[44] one low frequency, high frequency

and modulation TENS with alternating low and high frequency current,[45] one burst TENS,[46] two low frequency TENS,[47 48] four interferential current stimulation,[49-52] and three evaluated pulsed electrostimulation.[13 53 54] The protocol of Palmer 2007 did not specify which type of TENS would be used.

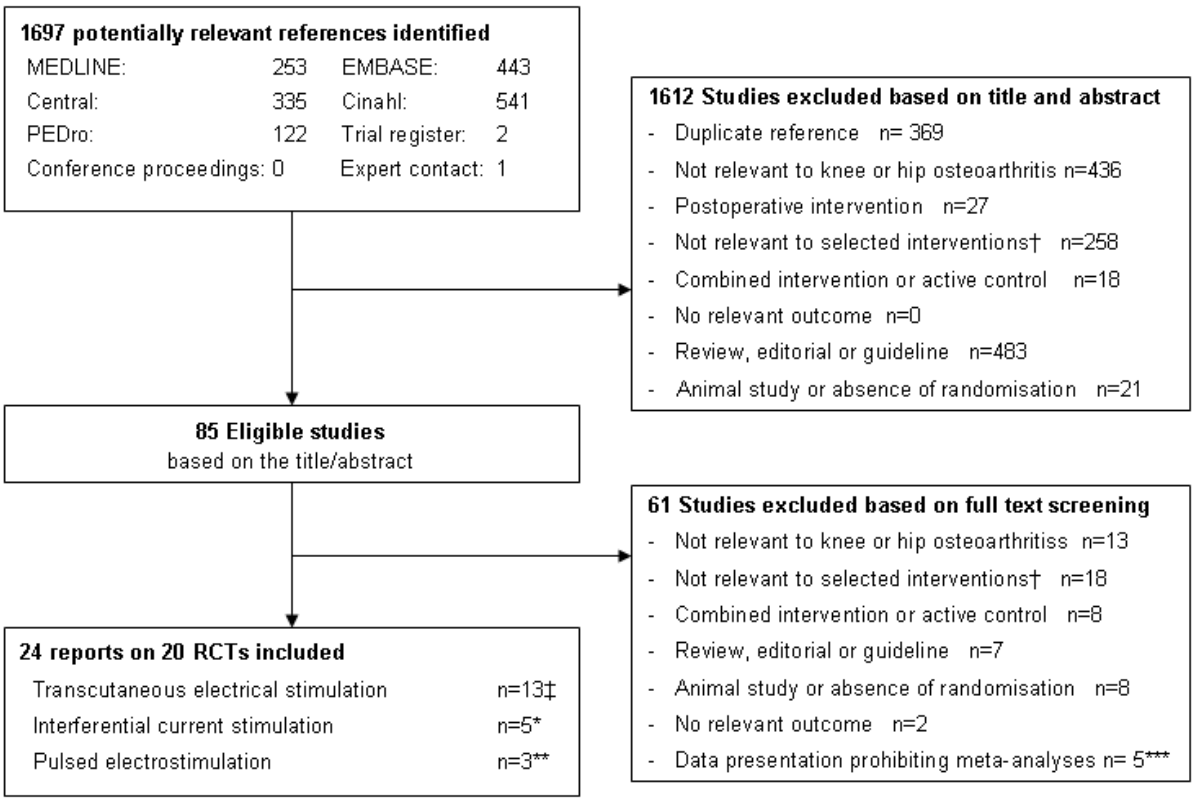

Figure 1 Flow chart. † interventions: any type of transcutaneous applied electrical stimulation primarily aiming at pain relief, with electrode placement involving knee innervation, ‡ described in 17 reports, including a protocol of an ongoing trial, *one trial including both interferential current stimulation and TENS versus control, ** including a protocol of an ongoing trial, *** primary authors of 5 reports on 4 cross-over trials were unable to provide data before cross-over

The description of the uncompleted trials can be found in the 'Characteristics of ongoing studies' table. Of the completed trials, 17 trials used a parallel group and one a 2 x 2 factorial design.[51] Twelve trials used a sham intervention in the control group, five used no intervention[37 44 47 51 52] and one trial had both a sham and a no intervention control.[38-40] Standardised cointerventions, provided in both experimental and control groups, were used in five trials with no intervention controls[37-40 44 47 52] and in two trials with a sham intervention.[36 49] Cetin 2008 used hot packs and exercise, Adedoyin 2002 dietary advice and exercise, Quirk 1985,Cheing 2002 and Adedoyin 2005 exercise, Bal 2007 used infra-red therapy and Ng 2003 an educational pamphlet. In addition, Itoh 2008 assigned 50% of patients to acupuncture using a factorial design.

Characteristics of the currents varied considerably, even within a specific type of electrostimulation. In the three trials evaluating low frequency TENS, pulse width and pulse frequency ranged from 200 μsec and 2 Hz to 1000 μsec and 4 Hz, with intensities set to reach a comfortable level in one,[45] and resulting in muscle contraction in two trials.[47 48] In trials of high frequency TENS, pulse width and pulse frequency ranged from 80 μsec and 32 Hz[43] to 200 μsec and 100 Hz,[41] with the majority of intensities described as strong but comfortable. In trials of burst TENS, Fargas-Babjak 1989 used a pulse frequency of 200 Hz, a train length of 125 μsec and a repetition frequency of 4 Hz with intensity increased up to the patients' limits of tolerability, while Grimmer 1992 used a pulse frequency of 80 Hz, an unclear train length and pulse width and a repetition frequency of 3 Hz, with the intensity resulting in a strong, tolerable tingling sensation and visible, but comfortable muscle contraction. In the five trials of interferential current stimulation, the beat frequency ranged from30 to 130 Hz and intensities resulted typically in tingling sensations in four trials,[44 49 51 52] and pain in one.[50] The two trials of pulsed electrostimulation were the only ones to use intensities below the sensory threshold.[53 54] The trials used the same device, which produces monophasic, spike-shaped pulses in a frequency of 100 Hz. The intensity of the current was initially increased until a tingling sensation was felt and subsequently reduced until this sensation disappeared.
The trials differed in type, number and localisation of electrodes used (see 'Characteristics of included studies'). The median duration of electrostimulation per session was 25 minutes (range 15 minutes to 8.2 hours), with a duration of 15 to 20 minutes in 10 trials,[37 41 43 44 47-52] 30 to 40 minutes in six;[11 36 41 42 45 46] and 60 minutes or more in 4 trials.[38-41 53 54] The median number of treatment sessions per week was 3.5 (range 1 to 14), with up to three sessions per week in eight trials,[11 37 43 44 49-52] four to six in seven,[36 38-42 45 47 48] and seven or more in three trials.[46 53 54] This resulted in a median duration of electrostimulation of 1.5 hours per week (range 15 minutes to 57.4 hours). The median length of the treatment period was four weeks (range one day to 12 weeks). All but one trial explicitly included patients with knee osteoarthritis only, with the diagnosis based on clinical and/or radiographic evidence. Fargas-Babjak 1989 included patients with either knee or hip osteoarthritis, and failed to report the percentage of patients with knee osteoarthritis, but it was considered likely that this percentage was above 75%. The majority of patients had a clinical severity requiring simple non-surgical treatments.[1] In one trial of pulsed electrostimulation, the majority of patients (41 out of 58) were candidates for total knee arthroplasty, however.[53] The description of patient characteristics was generally poor. Only four trials[36 42 48 53] reported the average disease duration, which ranged from two to 8.4 years.

Four cross-over trials could not be included because of incomplete reporting, which did not allow the distinction between treatment phases.[55-57] All but Lewis 1985 were included in the previous version of this review.[6] Three other trials were excluded because of an active control intervention using another type of electrostimulation.[58-60] Detailed reasons for exclusion are displayed in 'Characteristics of excluded studies'

Risk of bias in included studies Figure 2 summarises the methodological characteristics and source of funding of included trials. One trial reported both adequate sequence generation and adequate concealment of allocation,[53] five trials reported only adequate sequence generation,[42 43 45 47 51] and one trial reported adequate concealment, but provided insufficient detail on the generation of allocation sequence.[11] Two trials were quasi-randomised, one used alternation to allocate patients to experimental and control intervention,[49] the other allocated patients according to hospital registration number.[36] In the remaining nine trials, low quality of reporting hampered any judgement regarding sequence generation and concealment of allocation.

Six trials[11 42 45 46 53 54] were described as double-blind. Thirteen trials used sham interventions, all using identical devices in experimental and control groups.[11 36 38-43 45 46 48-50 53 54] In 10 out of 13 trials, sham devices had broken leads so that no current could pass, whereas the indicator light or digital display of intensity control functioned normally. In the two pulsed electrostimulation trials, all patients were instructed to increase the intensity until a tingling sensation was felt, after which they were asked to reduce intensity just below the perception (sensory) level. Pulsed electrostimulation sham devices were adapted with an automatic shut-off as soon as the amplitude was reduced.[53 54] Only the sham device used in Defrin 2005 was not considered to lead to adequate patient blinding, as the sham device was described as shut off. Only the two trials of pulsed electrostimulation, however, which used currents below the sensory threshold, were deemed to have fully credible blinding of patients.[53 54]

Sixteen out of 18 completed trials contributed to the analysis of pain outcomes. Of these, only three trials,[11 36 49] which had analysed all randomly assigned patients, were considered to have an intention-to-treat analysis of pain outcomes at end of treatment. In three trials[37 47 50] it was unclear whether exclusions of randomised patients from the analysis had occurred, in five trials[42 45 46 48 53] exclusions were reported, but their percentage remained unclear and in the remaining six trials the median reported exclusion rate was 7% in the experimental and 11.5% in the control groups (range 0% to 25% in both experimental and control groups). Two out of nine trials contributing to the analysis of function outcomes were considered to have an intention-to-treat analysis.[36 52] In one trial[37] it was unclear whether exclusions of randomised

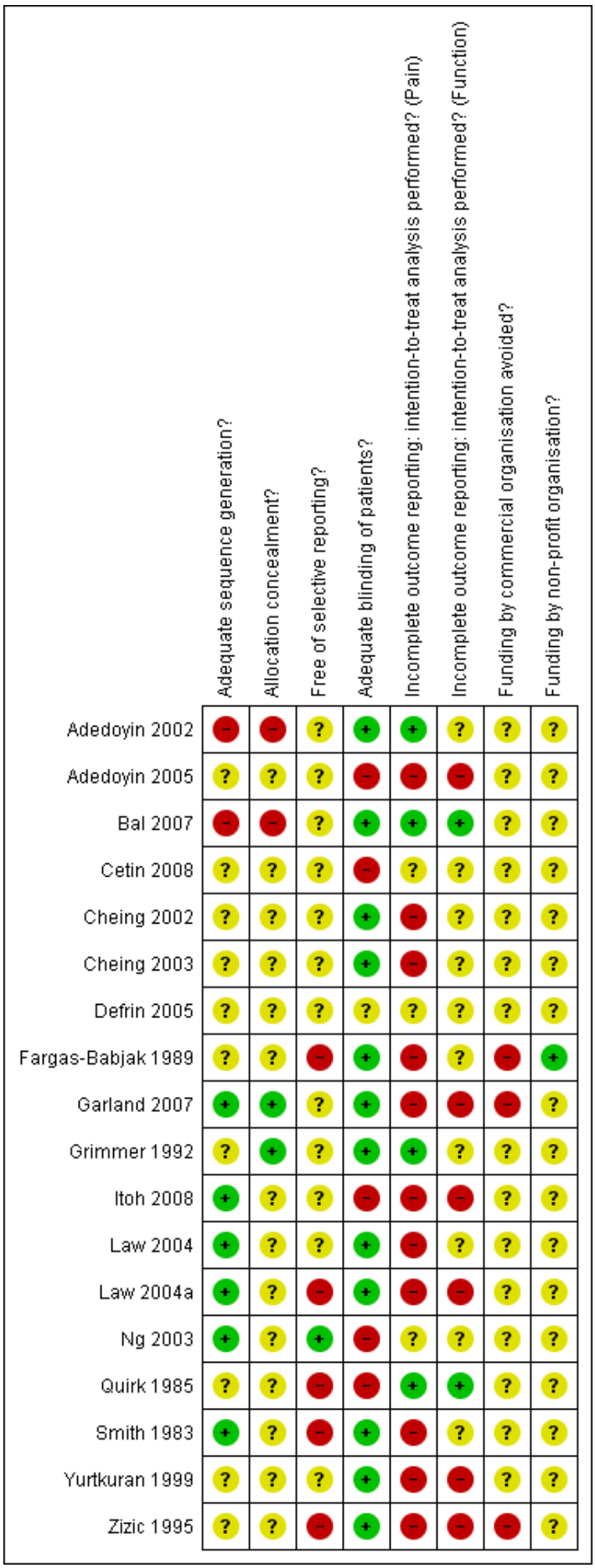

Figure 2. Methodological characteristics and source of funding of included trials. (+) indicates low risk of bias, (?) unclear and (-) a high risk of bias on a specific item.

patients from the analysis had occurred, in three trials[42 48 53] exclusions were reported, but their percentage remained unclear and in the remaining three trials the median reported exclusion rate was 11.5% in experimental and 12% in control groups (range 0% to 25% in experimental, and 11% to 25% in control groups, respectively).

Only three trials explicitly specified primary outcomes,[49 51 54] although one of these specified more than two.[54] Only one trial reported a sample size calculation.[44] None of the trials had a sufficient sample size of at least 200 patients overall to achieve sufficient power for detecting a small to moderate SMD. Only three trials reported their source of funding: one was supported by a nonprofit organisation and a commercial body,[46] the other two by a commercial body only.[53 54]

For the effectiveness outcomes pain and function, the quality of the evidence [23] was classified as very low in view of the risk of bias in the included, predominantly small trials of questionable quality, the large heterogeneity between trials, the potential for selective reporting of function outcomes and the exploratory nature of the model used to predict SMDs of pain in trials as large as the largest trials ('Summary of findings for the main comparison'). For the safety outcomes, the quality of the evidence[23] was classified as moderate to low, again because of the predominantly small trials of questionable quality, the small number of trials reporting the outcomes and the small number of events resulting in imprecise estimates.

Effects of interventions

Knee pain Sixteen trials with 18 comparisons (726 patients) contributed to the meta-analysis of pain outcomes (Figure 3). The analysis suggested an overall large SMD of -0.86 (95% CI -1.23 to -0.49), which corresponds to a difference in pain scores of 2.1 cm on a 10 cm VAS between electrostimulation and control, favouring electrostimulation. Within the types of electrostimulation, a very large effect was found for interferential current stimulation (SMD -1.20, 95% CI -1.99 to -0.42), a large effect in TENS (SMD -0.85, 95%CI -1.36 to -0.34) and a moderate effect in pulsed electrostimulation (SMD -0.41, 95% CI -0.77 to -0.05). However, interaction tests provided little evidence for differences between different types. Pooling all types of electrostimulation, an I^2 of 80% indicated a high degree of between-trial heterogeneity (P for heterogeneity < 0.001), which was not substantially reduced when pooling types of electrostimulation separately. Four trials[41 42 45 50] showed unrealistically large SMDs of twice to three times the magnitude of what would be expected for total joint replacement.[1] The funnel plot appeared asymmetrical (Figure 4, P for asymmetry < 0.001) and the corresponding asymmetry coefficient was -7.6 (95% CI -10.6 to -4.5).

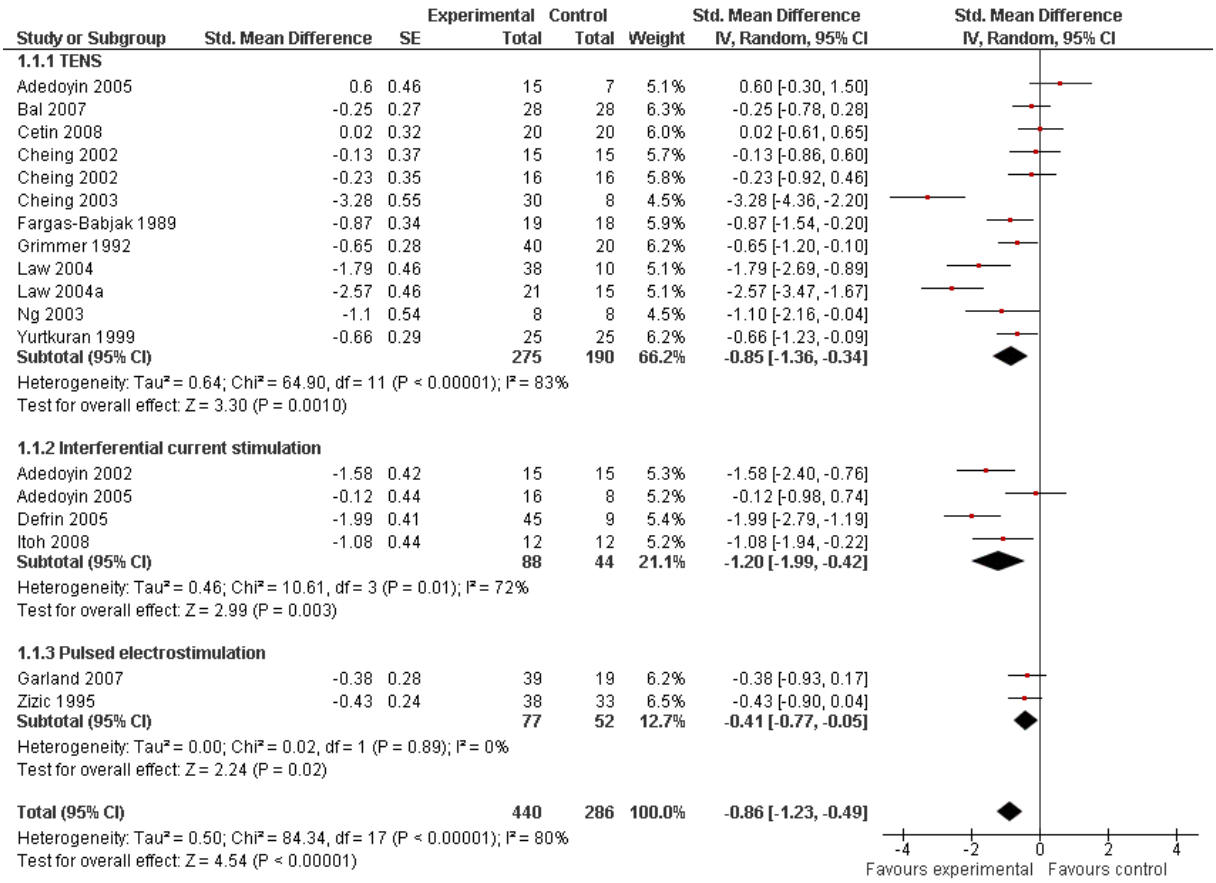

Figure 3. Forest plot of 16 trials comparing the effects of any type of transcutaneous electrostimulation and control (sham or no intervention) on knee pain. Values on x-axis denote standardised mean differences. The plot is stratified according to type of electrostimulation. Law 2004 reported on knee level, we inflated the standard error with sqrt(number knees)/sqrt(number patients) to correct for clustering of knees within patients. Adedoyin 2005 and Cheing 2002 contributed with two comparisons each. In Adedoyin 2005, the standard error was inflated and the number of patients in the control group was halved to avoid duplicate counting of patients when including 2 both comparisons in the overall meta-analysis. Data relating to the 3, 2, 3 and 4 active intervention arms in Cheing 2003, Grimmer 1992, Law 2004 and Defrin 2005, respectively, were pooled.

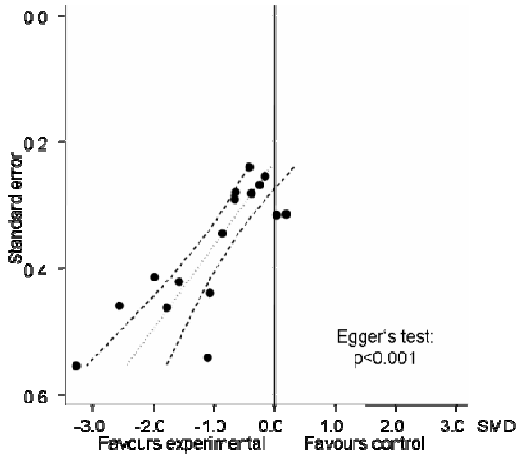

Figure 4. Funnel plot for effects on knee pain. Numbers on x-axis refer to standardised mean differences (SMDs), on y-axis to standard errors of SMDs.

Variable	N of trials	N of patients (experimental)	N of patients (control)	Pain intensity	Heterogeneity	P for interaction
	n	n	n	SMD (95% CI)	I^2 (%)	
All trials	16	440	286	-0.86 (-1.23 to -0.49)	80%	
Allocation concealment						0.47
Adequate	2	79	39	-0.52 (-0.91 to -0.13)	0%	
Inadequate or unclear	14	361	247	-1.03 (-1.49 to -0.57)	84%	
Type of control intervention*						0.12
Sham intervention	12	354	216	-1.13 (-1.59 to -0.67)	82%	
No control intervention	5	86	70	-0.31 (-0.80 to 0.19)	58%	
Blinding of patients						0.37
Adequate	11	309	205	-1.05 (-1.52 to -0.59)	82%	
Inadequate or unclear	6	131	79	-0.63 (-1.31 to 0.05)	81%	
Use of analgesic cointerventions						0.36
Similar between groups	4	124	83	-0.57 (-1.16 to 0.02)	74%	
Not similar or unclear	12	316	23	-1.10 (-1.60 to -0.59)	84%	
Intention-to-treat analysis						0.73
Yes	3	83	63	-0.76 (-1.43 to -0.09)	72%	
No or unclear	13	357	223	-1.00 (-1.48 to -0.53)	84%	
Type of ES**						0.94
High frequency TENS	8	177	139	-0.82 (-1.51 to -0.12)	86%	
Burst TENS	2	39	38	-0.85 (-1.32 to -0.38)	0%	
Modulation TENS	1	13	3	-1.41 (-2.92 to 0.10)	N/A	
Low frequency TENS	3	46	40	-0.82 (-1.29 to -0.34)	0%	
Interferential current stimulation	4	88	44	-1.20 (-1.99 to -0.42)	71%	
Pulsed ES	2	77	52	-0.41 (-0.77 to -0.05)	0%	
Duration of ES per session†						0.69‡
≤ 20 minutes	8	166	112	-0.95 (-1.55 to -0.35)	78%	
30 to 40 minutes	6	156	99	-1.45 (-2.28 to -0.62)	85%	
≥ 60 minutes	4	118	91	-0.47 (-0.96 to 0.02)	58%	
Number of sessions per week						0.90‡
≤ 3	6	163	91	-0.81 (-1.48 to -0.14)	82%	
4 to 6	7	182	125	-1.33 (-2.11 to -0.54)	88%	
≥ 7	3	96	70	-0.51 (-0.83 to -0.19)	0%	
Duration of ES per week***						0.74‡
≤1 hour	5	123	71	-0.85 (-1.72 to 0.01)	86%	
>1 to 5 hours	8	180	122	-1.42 (-2.11 to -0.74)	81%	
> 5 hours	5	137	109	-0.53 (-0.96 to -0.11)	55%	
Duration of treatment period						0.14
< 4 weeks	7	190	114	-1.39 (-2.13 to -0.66)	86%	
≥ 4 weeks	9	250	172	-0.64 (-1.06 to -0.22)	75%	

Table 1 Results of stratified analyses of pain outcomes ES: electrostimulation; *In Cheing 2002, two independent comparisons contributed in the two different strata. **Adedoyin 2005, Grimmer 1992 and Law 2004 contributed to two, two and three different strata: high-frequency TENS and interferential current stimulation, high-frequency TENS and burst, and high-, low-frequency and modulation TENS, respectively. †= Cheing 2003 contributed to all three different strata, with the same 8 control patients displayed in each stratum. ‡= p-values from test for trend.

This coefficient indicates that the benefit of electrostimulation increases by 7.6 standard deviation units for each unit increase in the standard error of the SMD, which is mainly a surrogate for sample size. The predicted SMD in trials as large as the largest trial (Zizic 1995, n = 71, standard error = 0.24) was -0.07 (95%CI -0.46 to 0.32), which corresponds to a difference in pain scores of 0.2 cm on a 10 cm VAS between electrostimulation and control. Referring to a median pain intensity of 6.1 cm in placebo groups at baseline, this corresponds to a difference of 4% improvement (95% CI -13% to +20%) between electrostimulation and control ('Summary of findings for the main comparison').

Table 1 presents results from stratified analyses. Estimates of SMD varied to some degree depending on concealment of allocation, adequacy of patient blinding, use of analgesic cointerventions and characteristics of electrostimulation, but 95% CIs of SMDs were wide and tests of interaction and tests for trend not statistically significant. There was little evidence to suggest that SMDs depended on the type of electrostimulation used (P for interaction = 0.94). Contrary to what would be expected in the presence of relevant placebo effects, we found some evidence towards larger benefits of electrostimulation in trials with a sham intervention as compared with trials without (P for interaction = 0.12). In addition, there was some evidence for larger benefits of electrostimulation associated with short durations of the overall treatment period of less than four weeks as compared with four weeks or more (P for interaction = 0.14). The analysis could not be stratified according to sample size, because none of included trials reached the prespecified sample size of 200 patients to be considered as adequately sized.

Withdrawals or drop-outs because of adverse events Eight trials (348 patients) contributed to the meta-analysis of patients withdrawn or dropped out because of adverse events (Figure 5). Of these, four TENS trials and one interferential current stimulation trial reported that no withdrawals or drop-outs due to adverse events had occurred, neither in experimental nor in control groups, therefore relative risks could not be estimated. In the remaining three trials, there was no evidence that transcutaneous electrostimulation is unsafe (relative risk 0.97), but 95% confidence intervals were wide and ranged from 0.16 to 6.00. Pooling all types of electrostimulation, an I^2 of 20%indicated a low degree of between-trial heterogeneity (P for heterogeneity = 0.29).

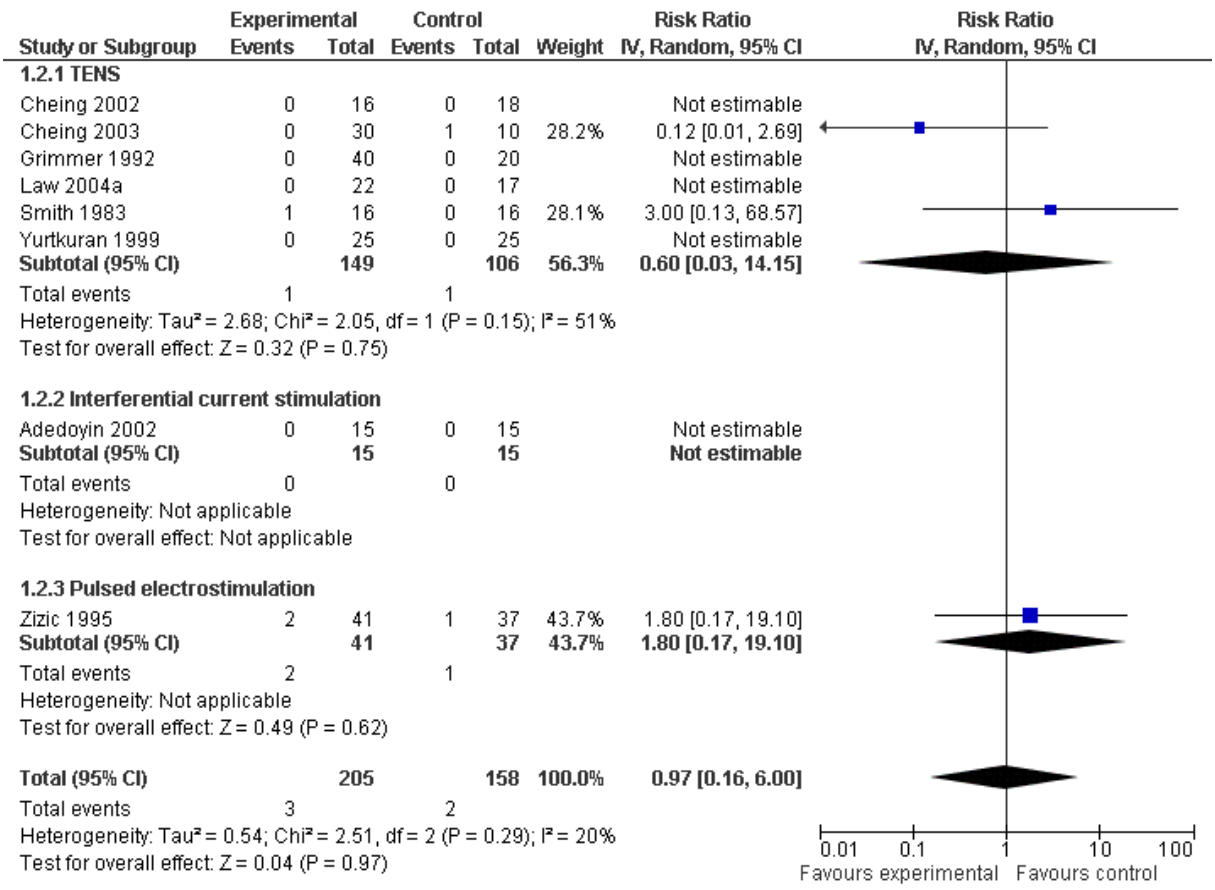

Figure 5. Forest plot of 8 trials comparing patients withdrawn or dropped out because of adverse events between any transcutaneous electrostimulation and control (sham or no intervention). Values on x-axis denote risk ratios. Risk ratios could not been estimated in 5 trials, because no dropout occurred in either group. The plot is stratified according to type of electrostimulation. Data relating to the 3 and 2 active intervention arms in Cheing 2003 and Grimmer 1992, respectively, were pooled.

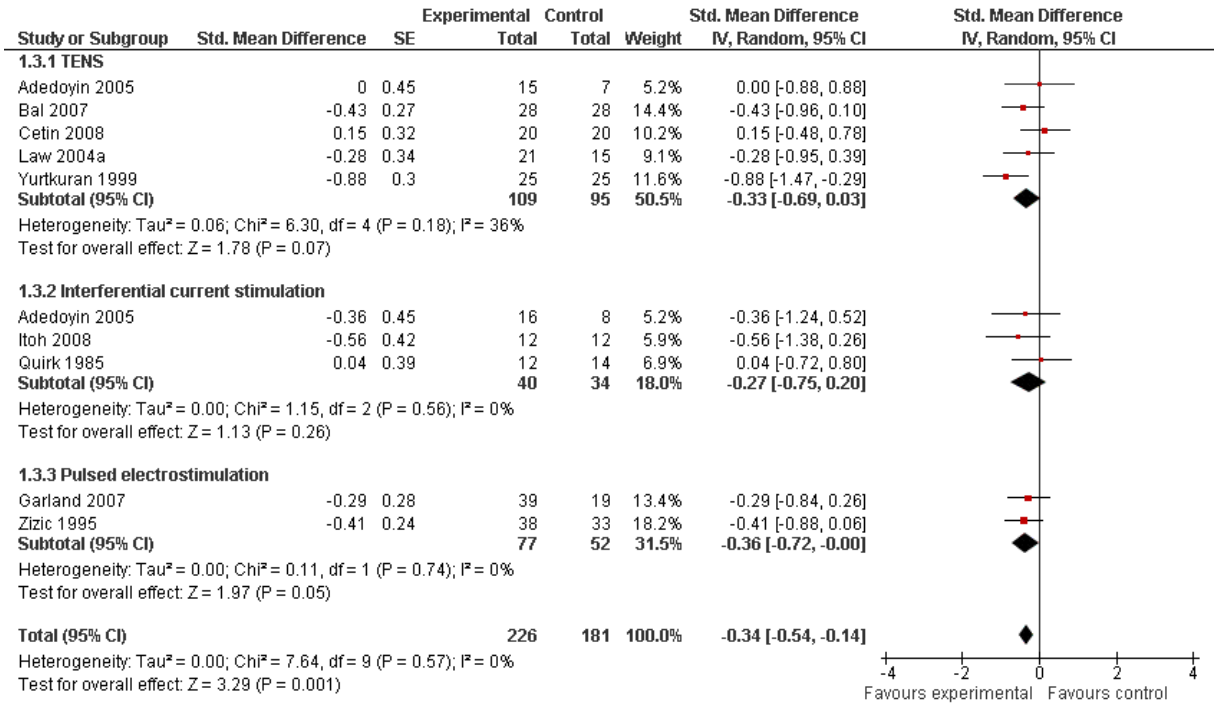

Figure 6. Forest plot of 9 trials comparing the effects of any type of transcutaneous electrostimulation and control (sham or no intervention) on function. Values on x-axis denote standardised mean differences. The plot is stratified according to type of electrostimulation. In Adedoyin 2005, the standard error was inflated and the number of patients in the control group was halved to avoid duplicate counting of patients when including both comparisons in the overall meta-analysis.

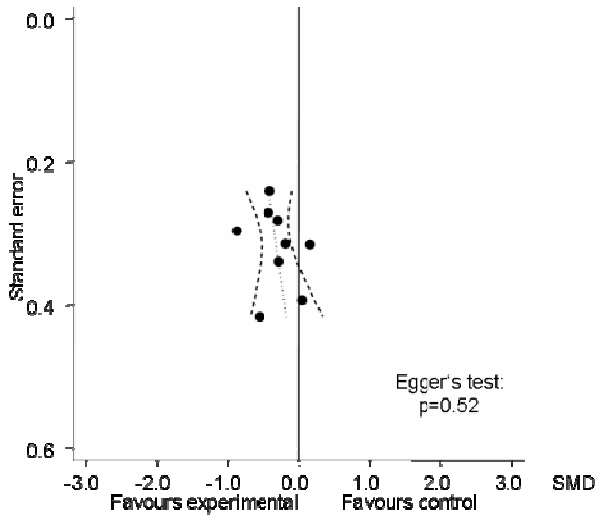

Figure 7. Funnel plot for effects on functioning of the knee.Numbers on x-axis refer to standardised mean differences (SMDs), on y-axis to standard errors of SMDs.

Function Nine trials (407 patients) contributed to the meta-analysis of function. The analysis suggested a small SMD of -0.34 (95% CI -0.54 to -0.14, Figure 6), which corresponds to a difference in function scores of 0.7 units on a standardised WOMAC disability scale ranging from 0 to 10, favouring electrostimulation. Referring to a median function score of 5.6 units in placebo groups at baseline, this corresponds to a difference of 20% improvement (95% CI +11% to +29%) between electrostimulation and control ('Summary of findings for the main comparison'). The estimated difference in the percentage of treatment responders between patients allocated to electrostimulation and patients allocated to placebo of 3% translated into an NNT to cause one additional treatment response on function of 29 (95% CI 19 to 69) ('Summary of findings for the main comparison').Differences between types of electrostimulation were not statistically significant. An I2 of 0% suggested no between-trial heterogeneity (P for heterogeneity = 0.57). The funnel plot did not appear asymmetrical (Figure 7, P for asymmetry = 0.52). The corresponding asymmetry coefficient was 1.4 (95% CI, -3.5 to 6.3).

Table 2 presents results from stratified analyses. Estimates of SMD varied to some degree depending on type of control intervention, adequacy of patient blinding, characteristics of electrostimulation and overall treatment period, but 95%CIs of SMDs were wide and tests for interaction and tests for trend not statistically significant. There was little evidence to suggest that SMDs depended on the type of electrostimulation used (P for interaction = 0.32). Again, the analysis could not be stratified according to sample size, because none of included trials reached the pre-specified sample size of 200 patients to be considered as adequately sized.

Other safety outcomes Three trials (175 patients) contributed to the meta-analysis of patients experiencing any adverse event (Figure 8) and four trials (195 patients) to the meta-analysis of patients experiencing any serious adverse event (Figure 9). In general, there was no evidence to suggest that electrostimulation is unsafe, but 95% CIs were wide and results inconclusive.

Variable	N of trials	N of patients (experimental)	N of patients (control)	Function	Heterogeneity	P for interaction
				SMD (95% CI)	I^2 (%)	
All trials	9	226	181	-0.34 (-0.54 to -0.14)	0%	
Allocation concealment						0.88
adequate	1	39	19	-0.29 (-0.85 to 0.26)	N/A	
inadequate or unclear	8	187	162	-0.34 (-0.56 to -0.12)	5%	
Type of control intervention						0.14
sham intervention	5	151	120	-0.46 (-0.70 to -0.21)	0%	
no control intervention	4	75	61	-0.10 (-0.45 to 0.24)	0%	
Blinding of patients						0.14
adequate	5	151	120	-0.46 (-0.70 to -0.21)	0%	
inadequate or unclear	4	75	61	-0.10 (-0.45 to 0.24)	0%	
Use of analgesic cointerventions						0.95
similar between groups	2	69	48	-0.33 (-0.70 to 0.05)	0%	
not similar or unclear	7	157	133	-0.34 (-0.60 to -0.08)	15%	
Intention-to-treat analysis						0.76
Yes	2	40	42	-0.28 (-0.71 to 0.16)	0%	
No or unclear	7	186	139	-0.35 (-0.58 to -0.12)	5%	
Type of ES**						0.32
High frequency TENS	4	84	70	-0.18 (-0.50 to 0.14)	0%	
Burst TENS	0					
Modulation TENS	0					
Low frequency TENS	1	25	25	-0.88 (-1.46 to -0.30)	N/A	
Interferential current stimulation	3	40	34	-0.27 (-0.75 to 0.20)	0%	
Pulsed ES	2	77	52	-0.36 (-0.72 to -0.00)	0%	
Duration of ES per session						0.80‡
≤ 20 minutes	5	100	86	-0.29 (-0.69 to 0.11)	44%	
30 to 40 minutes	2	49	43	-0.37 (-0.79 to 0.04)	0%	
≥ 60 minutes	2	77	52	-0.36 (-0.72 to -0.00)	0%	
Number of sessions per week						0.32‡
≤ 3	4	75	61	-0.10 (-0.45 to 0.24)	0%	
4 to 6	3	74	68	-0.54 (-0.88 to -0.20)	2%	
≥ 7	2	77	52	-0.36 (-0.72 to -0.00)	0%	
Duration of ES per week						0.32‡
≤ 1 hour	4	75	61	-0.10 (-0.45 to 0.24)	0%	
> 1 to 5 hours	3	74	68	-0.54 (-0.88 to -0.20)	2%	
> 5 hours	2	77	52	-0.36 (-0.72 to -0.00)	0%	
Duration of treatment period						0.18
< 4 weeks	3	74	68	-0.54 (-0.88 to -0.20)	2%	
≥ 4 weeks	6	152	113	-0.23 (-0.47 to 0.02)	0%	

Table 2. Results of stratified analyses of function. ES: electrostimulation; **Adedoyin 2005 contributed to two different strata: high-frequency TENS and interferential current stimulation; ‡= p-values from test for trend.

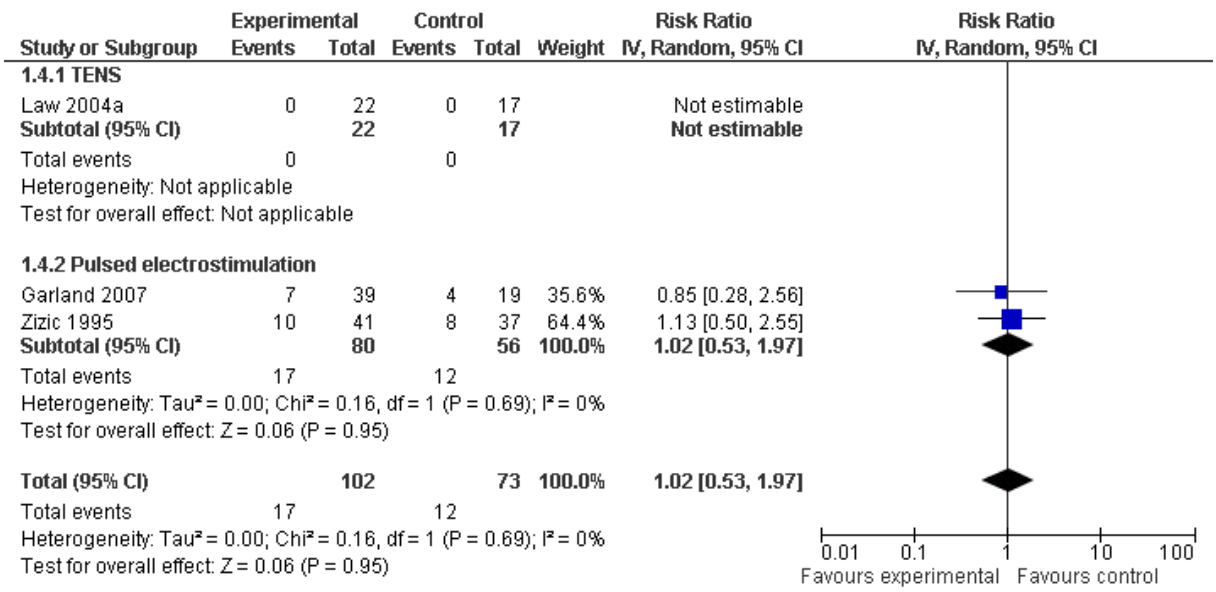

Figure 8 Forest plot of 3 trials comparing patients experiencing any adverse event between any transcutaneous electrostimulation and control (sham or no intervention). Values on x-axis denote risks ratios. The risk ratio in one TENS trial could not be estimated because no adverse event occurred in either group. The plot is stratified according to type of electrostimulation.

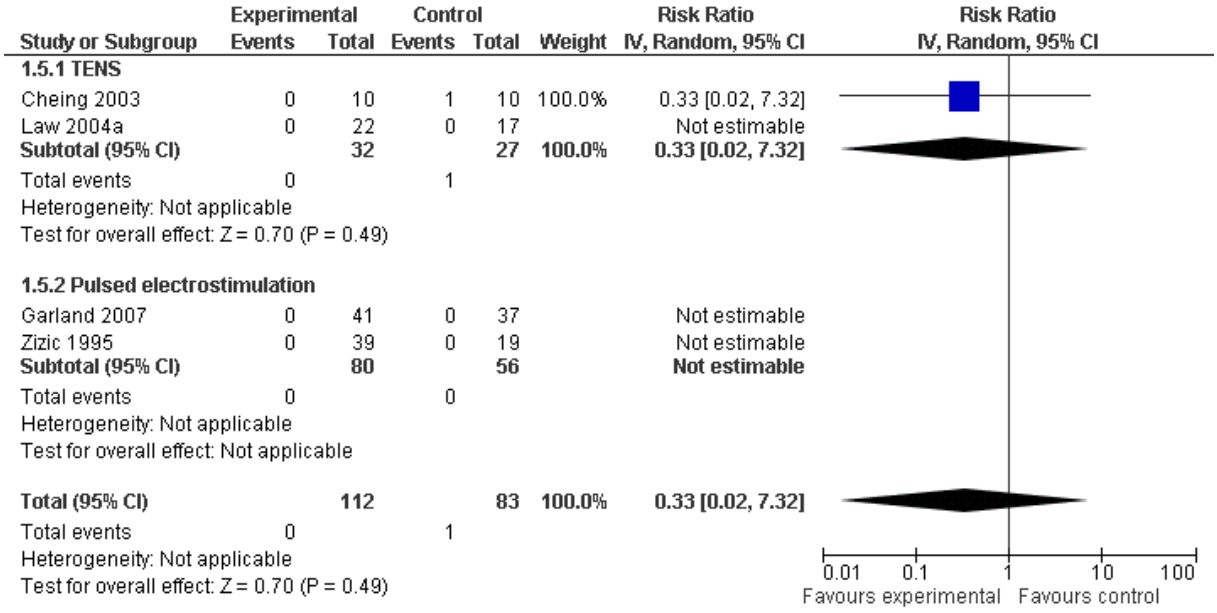

Figure 9 Forest plot of 4 trials comparing patients experiencing any serious adverse event between any transcutaneous electrostimulation and control (sham or no intervention). Values on x-axis denote risk ratios. Risk ratios could not been estimated in 3 trials, because no serious adverse event occurred in either group. The plot is stratified according to type of electrostimulation. Data relating to the 3 active intervention arms in Cheing 2003 were pooled.

Discussion

Summary of main results Our systematic review of trials comparing any type of transcutaneous electrostimulation with a sham or non-intervention control revealed a lack of adequately sized, methodologically sound and appropriately reported trials and a moderate to high degree of heterogeneity between trials, which made the interpretation of results difficult, particularly for joint pain as the primary therapeutic target of transcutaneous electrostimulation. In an attempt to minimise biases associated with small trials of questionable quality, we used meta-regression to predict effects of transcutaneous electrostimulation on pain and found the predicted effect sizes for pain negligibly small. The rates of withdrawals or drop-outs due to adverse events were comparable in experimental and control groups, but 95% CIs were wide and therefore inconclusive.

Quality of the evidence An inspection of funnel plots and a formal analysis of asymmetry indicated asymmetry for knee pain, but not for function, which suggested the presence of biases associated with small sample size particularly when estimating the effects of electrostimulation on knee pain. Asymmetrical funnel plots should be seen not only as an indication of publication bias, but as a generic tool for examination of small study effects: the tendency for the smaller studies to show larger treatment effects, possibly due to a combination of publication bias, selective reporting of outcomes and methodological problems particularly in small trials.[61 62] If reporting is inadequate, as was the case in our systematic review, then the standard error as a proxy for study size may be a more precise measure of trial quality than formal assessments of methodological quality. When modelling effects expected in trials as large as the largest trial included in our systematic review, we found effects on pain near null -0.07 (95% CI -0.46 to 0.32), which were clearly smaller than the pooled SMD actually found for pain in the meta-analysis -0.86 (95% CI -1.23 to -0.49). The effect of electrostimulation on function was small, but potentially clinically relevant, and the accumulated evidence appeared less affected by biases associated with small sample size. The methodological quality and the quality of reporting was poor. Insufficient information was noted in several randomised controlled trials about the treatment assignment procedure and concealment of allocation. Primary outcomes were specified in only three trials. Although several studies reported blinding of patients, complete blinding is difficult to achieve due to the sensory differences between treatment and placebo, as well as unintended communication between patient and evaluator.[63] Only Grimmer 1992 and Bal 2007 mentioned the inclusion of patients to be restricted to those without prior TENS experience; another two trials were likely to have achieved adequate blinding of patients with currents below the sensory

threshold used in the experimental group, which were likely to be indistinguishable from the sham intervention also for patients with treatment experience.[53 54] The majority of papers did not provide adequate information regarding withdrawals, drop-outs and losses to follow up, nor indicated whether patients with incomplete clinical data were included in the data analysis. Several trials omitted to describe adverse events, which is of concern.

Potential biases in the review process Our review is based on a broad literature search, and it seems unlikely that we missed relevant trials. Trial selection and data extraction, including quality assessment, were done independently by two authors to minimise bias and transcription errors. Components used for quality assessment are validated and reported to be associated with bias.[21 64]

As with any systematic review, our study is limited by the quality of included trials. As indicated above, trials generally suffered from poor methodological quality, inadequate reporting and small sample size. Some trials[41 42 50] showed unrealistically large SMDs of twice to three times the magnitude of what would be expected for total joint replacement.[1] Including these trials in the meta-analysis is likely to result in an overestimation of the benefits of transcutaneous electrostimulation.

Agreements and disagreements with other studies or reviews Interestingly, there are nearly as many systematic reviews and meta-analyses on transcutaneous electrostimulation in osteoarthritis as randomised trials. Here, we will focus mainly on the similarities and differences between ours and the previous version of this review,[6] which included seven transcutaneous electrical nerve stimulation (TENS) trials. We updated the search and used broader selection criteria, which resulted in 14 additional trials; 11 trials used TENS as the experimental treatment, four interferential current stimulation, one both TENS and interferential current stimulation, and two pulsed electrostimulation. As in the review of Osiri 2000, both parallel group and cross-over RCTs were included. For the cross-over studies, we only collected data from the first intervention phase in order to eliminate carry-over effects, whereas Osiri and colleagues included pooled data over all phases. We excluded three previously included cross-over trials, because the investigators were unable to provide data from the first phase only. In this update, we performed a more detailed quality assessment of component trials, followed by a detailed exploration of sources of variation between trials, including concealment of allocation, blinding, intention-to-treat analysis, characteristics of analyse continuous data, Osiri and colleagues used weighted mean differences or SMDs of the change from baseline scores, whereas we used SMDs of end of treatment scores and based our conclusions on treatment effects on pain predicted in uni-variable metaregression models

by using the standard error as the explanatory variable. In addition, fixed-effect models were used in the previous version unless there was statistically significant heterogeneity between trials based on $\chi 2$ testing. Model selection based on the mechanistic application of heterogeneity tests should be avoided, however. Here, we used random-effects models, which will generally be more conservative in terms of the estimated precision, but will be more affected by small study effects than a fixed-effect model, which makes an exploration of sources of variation, including different types of bias, mandatory. Results from the previous and current versions are therefore not directly comparable. Nevertheless, pooled SMDs for pain were favourable in our and the previous review,[6] with us reporting a pooled SMD of -0.86 (95% CI -1.23 to -0.49), whereas Osiri 2000 reported a SMD of -0.45 (95% CI -0.70 to -0.19), with confidence intervals overlapping widely. Although both Osiri and we acknowledge the risk of bias in summary estimates, Osiri concluded that transcutaneous electrostimulation is "shown to be effective in pain control over placebo". We disagree with these conclusions: when modelling effects expected in trials as large as the largest trial included, we found the SMD of pain near null and clinically irrelevant (-0.07, 95% CI -0.46 to 0.32). Osiri 2000 recorded function separately for the outcomes 'stiffness of the knee', '50-foot walking time', 'quadriceps muscle strength' and 'knee flection' with only one trial contributing to each of the categories. We choose a different approach, using a hierarchy developed to minimise the impact of selective reporting of outcomes and to allow for a synthesis of evidence across different studies using divergent definitions of function. Our effect sizes and conclusion concerning function are less favourable compared to those made by Osiri 2000. In this version, we also summarised safety data and found no evidence to suggest that electrostimulation is unsafe. Finally, unlike Osiri 2000, we also included trials of interferential current stimulation and pulsed electrostimulation. One of the two trials of pulsed electrostimulation[54] is covered in another Cochrane Review by Hulme 2002 on electromagnetic fields, even though the device used (BioniCare BIO-1000) does not generate electromagnetic fields, but electric currents.[65]

Authors' conclusions

Implications for practice Despite more than 20 years of clinical research, there is a lack of adequate evidence to support the use of any type of transcutaneous electrostimulation in patients with knee osteoarthritis. The effects on both knee pain and function are potentially clinically relevant and deserve further clinical evaluation.

Implications for research The current systematic review is inconclusive, hampered by the inclusion of only small trials of questionable quality.[62]. Adequately sized randomised parallel-group trials in about 2 x 100 patients with knee osteoarthritis are necessary to determine whether a specific type of transcutaneous electrostimulation is indeed associated with a clinically relevant benefit on pain. A sample size of 2 x 100 patients will yield more than 80% power to detect a small to moderate SMD of -0.40 at a two-sided P of 0.05, which corresponds to a difference of 1 cm on a 10 cm visual analogue scale (VAS) between experimental and control intervention. The trials should enrol patients without prior experience of any type of transcutaneous electrostimulation or evaluate success of blinding at the end of trial, use adequate concealment of allocation, experimental and sham interventions that are close to indistinguishable and an intention-to-treat analysis. Transcutaneous electrical nerve stimulation (TENS) devices are marketed as small, inexpensive, easy-to-use home units, but in the majority of trials TENS was administered by a therapist in a practice or hospital setting. Future research may focus on the effectiveness of self-administered TENS, with accurate recording of the duration of electrostimulation per day to assess compliance and enable the exploration of possible dose-effect relationships.

Acknowledgments We thank the Cochrane Musculoskeletal editorial team and Henk van Zutphen for valuable comments, and Malcolm Sturdy for database support. The authors are grateful to Serpil Bal, Gladys Cheing and Pearl Law for providing additional information concerning design and outcome data. We thank Beverly Lewis, Daniel Lewis and Mark Hallett who replied to our queries and attempted to locate files of trials published approximately 20 years ago, but were unable to provide additional outcome data.

Contributions of authors

Study conception: Rutjes, Jüni

Protocol development: Rutjes, Nüesch, Hendriks, Kalichman, Reichenbach, Jüni

Acquisition of data: Rutjes, Nüesch, Sterchi, Kalichman, Hendriks, Osiri, Brosseau, Reichenbach, Jüni

Analysis and interpretation of data: Rutjes, Nüesch, Sterchi, Hendriks, Kalichman, Osiri, Brosseau, Reichenbach, Jüni

Drafting of the manuscript: Rutjes

Critical revision of the manuscript for important intellectual content: Rutjes, Nüesch, Sterchi, Hendriks, Kalichman, Reichenbach, Jüni

Statistical analysis: Nüesch, Jüni, Rutjes

Obtained funding: Reichenbach, Jüni

Dr Rutjes and Mrs Nüesch contributed equally to this article.

Declarations of interest None.

Sources of support Internal sources: Institute of Social and Preventive Medicine, University or Bern, Switzerland. Intramural grants. External sources: Swiss National Science Foundation, Switzerland. National Research Program 53 on musculoskeletal health (grant numbers 4053-40-104762/3 and 3200-066378)

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON *[Explanation]*

Any type of transcutaneous electrostimulation compared with sham or no intervention for osteoarthritis of the knee

Patient or population: patients with osteoarthritis
Settings: physical therapy practice of outpatient clinic
Intervention: any type of transcutaneous applied electrostimulation
Comparison: sham or no specific intervention

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	No of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk* Sham or no specific intervention	Corresponding risk Any type of transcutaneous electrostimulation				
Pain Various pain scales Median follow-up: 4 weeks	**-1.8 cm change** on 10 cm VAS[1] 29% improvement	**-2.0 cm change** (Δ -0.2 cm, -1.2 to 0.8 cm)[2] 33% improvement (Δ +4%, -13% to +20%)[3]	SMD -0.07 (-0.46 to 0.32)	726 (16 studies)	+000 very low[4]	Little evidence of beneficial effect (NNT: not statistically significant) The estimated pain in the intervention group of large trials was derived from meta-regression using the standard error as independent variable
Function Various validated function scales Median follow-up: 4 weeks	**-1.2 units on WOMAC** (range 0 to 10)[1] 21% improvement	**-2.3 units on WOMAC** (Δ -1.1, -1.6 to -0.6)[5] 41% improvement (Δ +20%, +11% to +29%)[6]	SMD -0.34 (-0.54 to -0.14)	407 (9 studies)	+000 very low[7]	NNT: 29 (95% CI 19 to 69)[8]
Number of patients experiencing any adverse event Median follow-up: 4 weeks	150 per 1000 patient-years[1]	153 per 1000 patient-years (80 to 296)	RR 1.02 (0.53 to 1.97)	175 (3 studies)	++00 low[9]	No evidence of harmful effect (NNH: not statistically significant)
Number of patients withdrawn or dropped out because of adverse events Median follow-up: 4 weeks	17 per 1000 patient-years[1]	16 per 1000 patient-years (3 to 102)	RR 0.97 (0.16 to 6.00)	363 (8 studies)	+++0 moderate[10]	No evidence of harmful effect (NNH: not statistically significant)
Number of patients experiencing any serious adverse event Median follow -up: 4 weeks	4 per 1000 patient-years[1]	1 per 1000 patient-years (0 to 29)	RR 0.33 (0.02 to 7.32)	195 (4 studies)	++00 low[11]	No evidence of harmful effect (NNH: not statistically significant)

*The basis for the **assumed risk** in the safety outcomes (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: confidence interval; GRADE: GRADE Working Group grades of evidence (see explanations); NNT: number needed to treat; NNH: number needed to harm; RR: risk ratio; SMD: standardised mean difference

GRADE Working Group grades of evidence
High quality (++++): Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality (+++0): Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality (++00): Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality (+000): We are very uncertain about the estimate.

[1] Median reduction as observed across control groups in large osteoarthritis trials (Nuesch 2009).

[2] Standardised mean differences (SMDs) were back-transformed onto a 10 cm visual analogue scale (VAS) on the basis of a typical pooled SD of 2.5 cm in trials that assessed pain using a VAS, and expressed as change based on an assumed standardised reduction of 0.72 standard deviation units in the control group.

[3] The median observed pain score at baseline across control groups in large osteoarthritis trials was 6.1 cm on a 10 cm VAS (Nuesch 2009).

[4] Downgraded (3 levels) because the effect was estimated from a meta-regression model using the standard error as independent variable and because included trials were generally of low quality and small sample size: only 2 out of 16 trials used adequate concealment of allocation, only 3 performed analyses according to the intention-to-treat principle, and the presence of large between trial heterogeneity.

[5] Standardised mean differences (SMDs) were back-transformed onto a 0 to 10 standardised WOMAC function score on the basis of a typical pooled SD of 2.1 in trials that assessed function on WOMAC function scale and expressed as change based on an assumed standardised reduction of 0.58 standard deviation units in the control group.

[6] The median observed standardised WOMAC function score at baseline across control groups in large osteoarthritis trials was 5.6 units (Nuesch 2009).

[7] Downgraded (3 levels) because included trials were generally of low quality and small sample size: 1 out of 9 studies used adequate concealment of allocation methods, only 2 performed analyses according to the intention-to-treat principle, presence of moderate between trial heterogeneity, 9 out of 18 studies reported this outcome, likely leading to selective outcome reporting bias.

[8] Absolute response risks for function in the control groups were assumed 26% (see Methods section).

[9] Downgraded (2 levels) because the confidence interval crosses no difference in the pooled estimate, 1 out of 3 studies included all patients in this analysis, 3 out of 18 studies reported this outcome, likely leading to selective outcome reporting bias.

[10] Downgraded (1 level) because the confidence interval of the pooled estimate is wide and crossed no difference, 8 out of 18 studies reported this outcome, possibly leading to selective outcome reporting bias.

[11] Downgraded (2 levels) because 4 out of 18 studies reported this outcome, possibly leading to selective outcome reporting bias, the confidence interval of the pooled estimate is wide and crossed no difference.

Characteristics of included studies

Adedoyin 2002

Methods	Quasi-randomised trial using alternation for the allocation of patients 2-arm parallel group design Trial duration: 4 weeks No power calculation reported
Participants	30 patients randomised 30 patients with knee OA reported at baseline Study joints: 30 knees Number of females: 20 of 30 (67%) Average age: 59 years Average BMI: 28 kg/m^2
Interventions	Experimental intervention: interferential current stimulation, dietary advice and exercise, twice per week Control intervention: Sham interferential current stimulation, dietary advice and exercise, twice per week Duration of treatment period: 4 weeks Analgesics not allowed Device: Enraf-Nonius Endomed 5921 (4 pole) Self-administered: no Waveform: interferential Pulse width: not applicable Pulse frequency: amplitude-modulated frequency of 100 Hz for 15 min (beat frequency), 80 Hz for last 5 min (beat frequency) Amplitude: above sensory threshold, up to appreciable sensation Duration of stimulation per session: 20 minutes Electrodes: 4 electrodes covered with padding Placement: 2 latero-medial, 2 antero-posterior
Outcomes	Extracted pain outcome: global pain after 4 weeks, described as "Pain perception (VAS)" No function outcome reported Primary outcome: global pain (VAS)
Notes	All subjects from black Nigerian population

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	No	Alternation

Adedoyin 2002 (*Continued*)

Allocation concealment?	No	Alternation
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes
Adequate blinding of patients?	Yes	Sham device: identical in appearance, not increasing intensity, flash light on, patient in position unable to read level of intensity
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Yes	-
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable, no function outcome reported
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Adedoyin 2005

Methods	Randomised controlled trial 3-arm parallel group design Trial duration: 4 weeks Power calculation reported
Participants	51 patients randomised 46 patients with knee OA reported at baseline Study joints: 46 knees Number of females: 28 of 46 (61%) Average age: 55 years Average BMI: 28 kg/m2
Interventions	*Comparison 1* Experimental intervention: TENS and exercise twice per week Control intervention: exercise, twice per week *Comparison 2* Experimental intervention: interferential current stimulation and exercise, twice per week Control intervention: exercise, twice per week Duration of treatment period: 4 weeks Analgesics not allowed, patients confirmed not to take analgesics TENS Device: Endomed 5921D

Adedoyin 2005 (*Continued*)

	Self-administered: no Waveform: not reported Pulse width: 200 ms Pulse frequency: 80 Hz Amplitude: above sensory threshold, strong but comfortable Duration of stimulation per session: 20 minutes Electrodes: 2 electrodes 8 x 6 cm Placement: Each side of affected knee joint, aligned longitudinally along length of limb Interferential Current Stimulation Device: Endomed 5921D (2 pole) Waveform: interferential Pulse width: not applicable Pulse frequency: 80 Hz (beat) Amplitude: above sensory threshold: strong but comfortable, strong tingling sensation without muscle contraction Duration of stimulation per session: 20 minutes Electrodes: 2 electrodes 8 x 6 cm Placement: each side of affected knee joint, aligned longitudinally along length of limb
Outcomes	Extracted pain outcome: pain on activities other than walking after 4 weeks, described as "Pain recorded while standing (10-point pain rating scale with 0 "no pain", 5 "moderate pain" and 10 "worst pain imaginable")" Extracted function outcome: WOMAC global scale after 4 weeks (Likert) No primary outcome reported
Notes	-

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes
Adequate blinding of patients?	No	No sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	15 out of 15 (100%) in TENS group, 16 out of 19 (84%) in interferential current stimulation group, 15 out of 17 (88%) in control group analysed

Adedoyin 2005 (*Continued*)

Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Bal 2007

Methods	Quasi-randomised single centre controlled trial with allocation according to hospital registration number 2-arm parallel group design Trial duration: 13 weeks No power calculation reported
Participants	56 patients randomised 56 patients with knee OA reported at baseline Study joints: 56 knees Number of females: 50 of 56 (89%) Average age: 57 years Average BMI: 31 kg/m2 Average disease duration: 2 years
Interventions	Experimental intervention: TENS and infra-red therapy, 5 times per week Control intervention: sham TENS and infra-red therapy, 5 times per week Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and the intake was assessed Device: PlusMED 1-904 Self-administered: no Waveform: not reported Pulse width: 140 µsec Pulse frequency: 80 Hz Amplitude: above sensory threshold, not up to maximum tolerance, no muscle contractions observed* Duration of stimulation per session: 40 minutes Electrodes: 4, type unclear Placement: acupuncture points: ST36, GB34, SP10, SP9, ST34
Outcomes	Extracted pain outcome: WOMAC pain subscore after 13 weeks (Likert) Extracted function outcome: WOMAC disability subscore after 13 weeks (Likert) No primary outcome reported

Bal 2007 (*Continued*)

Notes	Article in Turkish, outcome assessment done by AR and RS assisted by a native Turkish researcher. Serpil Bal verified all extracted data. *as indicated by Serpil Bal in personal communication.

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	No	The published report only stated that there was a random allocation of patients to comparison groups. In personal communication, investigator Serpil Bal stated that the patients were allocated according to last digit of their hospital registration number. Patients with even numbers were assigned to TENS group, patients with odd numbers to a sham intervention.
Allocation concealment?	No	No, the same investigator responsible of randomisation was giving interventions, as indicated by Serpil Bal in personal communication
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes, we have been unable to sort out this item with investigator Serpil Bal
Adequate blinding of patients?	Yes	Trial is described as single blind study using sham device PlusMED 1-904, indistinguishable from real TENS unit. Sham device had broken leads, no current passed but flashing light was on. None of the patients had prior experience with TENS.
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Yes	All subjects were available for end of treatment measurements, as indicated by Serpil Bal in personal communication
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Yes	All subjects were available for end of treatment measurements, as indicated by Serpil Bal in personal communication
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Cetin 2008

Methods	Randomised controlled trial 5-arm parallel group design Trial duration: 8 weeks No power calculation reported
Participants	100 patients randomised 100 patients with knee OA reported at baseline Study joints: 100 knees Number of females: 100 of 100 (100%) Average age: 60 years Average BMI: 28 kg/m2
Interventions	Experimental intervention: TENS + hot packs + isokinetic exercise, 3 times per week Control intervention: hot packs + isokinetic exercise, 3 times per week Duration of treatment period: 8 weeks Analgesics allowed, unclear whether intake was similar between groups Device: MED911 Self-administered: no Waveform: not reported Puls width: 60 msecs Pulse frequency: 60-100 Hz Amplitude: above sensory threshold, increased to point of seeing no contraction, while patient felt comfortable Duration of stimulation per session: 20 minutes Electrodes: not reported Electrode placement: around painful areas
Outcomes	Extracted pain outcome: pain on walking after 8 weeks, described as "Knee pain severity after a 50-m walk (VAS)" Extracted function outcome: Lequesne OA index global score after 8 weeks (Likert) No primary outcome reported
Notes	Only 2 arms qualified for inclusion in this review

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes

Cetin 2008 (*Continued*)

Adequate blinding of patients?	No	No sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Unclear	No information provided
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	No information provided
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Cheing 2002

Methods	Randomised controlled trial 4-arm parallel group design Trial duration: 8 weeks Randomisation stratified according to age, gender, BMI No power calculation reported
Participants	66 patients randomised 62 patients with knee OA reported at baseline Study joints: 62 knees Number of females: 53 of 62 (85%) Average age: 64 years Average BMI: 28 kg/m^2
Interventions	*Comparison 1* Experimental intervention: 60 min TENS, 5 times per week Control intervention: sham TENS, 5 times per week *Comparison 2* Experimental intervention: TENS plus exercise, 5 times per week Control intervention: exercise alone, 5 times per week Duration of treatment period: 4 weeks Analgesics allowed, unclear whether intake was similar between groups Device: MAXIMA III (dual channel) Self-administered: unclear, most likely not Waveform: square Pulse width: 140 μsec Pulse frequency: 80 Hz Amplitude: above sensory threshold, tingling sensation, 3 to 4 times above sensory threshold

Cheing 2002 (*Continued*)

	Duration of stimulation per session: 60 minutes Electrodes: 4 electrodes of 4 x 4 cm Placement: at acupuncture points: ST35, SP9, GB34, extra 31,32 (one electrode covering both extra 32 and ST35)
Outcomes	Extracted pain outcome: global pain after 8 weeks, described as "Intensity of subjective pain sensation (Baseline score on 0-10 cm VAS was standardised to be 100% in each of the groups. Follow up values were expressed as mean decrease in % from baseline)". No function outcome reported No primary outcome reported
Notes	-

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes
Adequate blinding of patients?	Yes	Comparison 1: Yes, sham device identical in appearance to real TENS unit, no current passed but indicator light was lit up Comparison 2: No, no sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	Comparison 1: 16 out of 16 (100%) randomised to experimental and 16 out of 18 (89%) randomised to control group were analysed Comparison 2: 15 out of 17 (88%) randomised to experimental and 15 out of 15 (100%) randomised to control group were analysed
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Cheing 2003

Methods	Randomised controlled trial 4-arm parallel group design Trial duration: 4 weeks Randomisation stratified according to gender No power calculation reported
Participants	40 patients randomised 38 patients with knee OA reported at baseline Study joints: 38 knees Number of females: 34 of 38 (89%) Average age: 66 years
Interventions	Experimental intervention: 20 min TENS in group 1, 40 min TENS in group 2, 60 min TENS in group 4, 5 times per week Control intervention: sham TENS, 5 times per week Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and whether intake was similar between groups Device: ITO 120Z TENS (dual channel) Self-administered: no Waveform: not reported Pulse width: 200 µsec Pulse frequency: 100 Hz Amplitude: above sensory threshold, strong but comfortable Duration of stimulation per session: 20 minutes Electrodes: 4 of 2 x 3 cm rubber electrodes Placement: 4 acupuncture points extra 31,32, ST35, GB34, SP9
Outcomes	Extracted pain outcome: pain on walking after 4 weeks, described as "pain during walking (VAS)" No function outcome reported No primary outcome reported
Notes	-

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes

Cheing 2003 (*Continued*)

Adequate blinding of patients?	Yes	Sham device: electronic circuit disconnected, no current passed, but indicator light on
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	30 out of 30 (100%) randomised to experimental and 8 out of 10 (80%) randomised to control group were analysed
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Defrin 2005

Methods	Randomised controlled trial 6-arm parallel group design Trial duration: 4 weeks No power calculation reported
Participants	62 patients randomised 62 patients with knee OA reported at baseline Study joints: 62 knees Average age: 67 years
Interventions	Experimental intervention: noxious adjusted interferential current stimulation in group 1, noxious unadjusted interferential current stimulation in group 2, innocuous adjusted interferential current stimulation in group 3, innocuous unadjusted interferential current stimulation in group 4, 3 times per week Control intervention: sham interferential current stimulation, 3 times per week Duration of treatment period: 4 weeks Analgesics allowed, unclear whether intake was similar between groups. Device: Uniphy: Phyaction electrical stimulator Self-administered: no Waveform: interferential Pulse width: not applicable Pulse frequency: 30 to 60 Hz (beat) Amplitude: above sensory threshold, 2 groups 30% above pain threshold; 2 groups 30% below pain threshold Duration of stimulation per session: 20 minutes Electrodes: 2 of 8 x 6 cm wet sponge electrodes Placement: medial and lateral aspects of the knee, 2 cm from outer margins of patella

Defrin 2005 (*Continued*)

Outcomes	Extracted pain outcome: global pain after 4 weeks, described as "chronic pain intensity (VAS)" No function outcome reported No primary outcome reported
Notes	1 out of 6 trial arms, the no-intervention control group was excluded in the review

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes
Adequate blinding of patients?	Unclear	Use of sham device: Uniphy-Phyaction electrical stimulator, however the device described as shut-off
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Unclear	No information provided
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Fargas-Babjak 1989

Methods	Randomised controlled trial 2-arm parallel group design Trial duration: 13 weeks No power calculation reported
Participants	56 patients randomised 56 patients with knee OA reported at baseline Study joints: 56 joints, most likely > 75% knees

Fargas-Babjak 1989 (*Continued*)

	Average age; gender, BMI: not reported
Interventions	Experimental intervention: burst TENS, twice per day Control intervention: sham TENS, twice per day Duration of treatment period: 6 weeks Analgesics allowed, but change of dosage prohibited. Unclear whether analgesics were assessed and whether intake was similar between groups. Device: Codetron Self-administered: yes Waveform: square Pulse width: 1000 μsec Pulse frequency: 200 Hz, train length of 125 ms, repetition frequency of 4 Hz (25 pulses per train) Amplitude: above sensory threshold, highest intensity that could be tolerated without inducing frank pain Duration of stimulation per session: 30 minutes Electrodes: 7 carbon rubber (self-adhesive) Karaya Pads electrodes of 2 x 3 cm Placement: 10 acupuncture points: GV14, GV4, GB30, GB34, SP13, B1 60, ST36, B1 40, SP9, LI4 and 3 extra tender points
Outcomes	Extracted pain outcome: global pain after 13 weeks described as "Pain improvement (percentage pain improvement based on VAS)" No function outcome reported No primary outcome reported
Notes	*Investigators named their intervention AL-TENS, but we coded it burst TENS in the analyses

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	No	Quote: "Full details of this (Percent Improvement Pain Scale) are reported elsewhere". Investigators however failed to provide reference.
Adequate blinding of patients?	Yes	Use of sham device: Codetron, identical in appearance, set at frequency of 0.2 Hz with a threshold electrical stimulus of 0.5 mA, which caused a sensation on the skin but failed causing the deep muscle afferent stimulation

Fargas-Babjak 1989 (*Continued*)

Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	56 patients randomised but only 19 analysed in the experimental, and 18 analysed in the control group
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	No	Sponsor: Electronic Health Machines
Funding by non-profit organisation?	Yes	NRC grant no: 689

Garland 2007

Methods	Randomised multicentre controlled trial 2-arm parallel group design Number of participating centres: 3 Trial duration: 12 weeks Randomisation stratified according to study site No power calculation reported
Participants	100 patients randomised 58 patients with knee OA reported at baseline; 41 out of 58 candidates for total knee arthroplasty Study joints: 58 knees Number of females: 38 of 58 (66%) Average age: 66 Disease duration: 8.4 years
Interventions	Experimental intervention: pulsed electrical stimulation Control intervention: sham intervention Duration of treatment period: 12 weeks Analgesics allowed and intake assessed, but unclear whether intake was similar. Device: BIO-1000 Self-administered: yes Waveform: unclear Pulse width: unclear Pulse frequency: 100 Hz Amplitude: below sensory threshold, initial increase of amplitude up to 12 Volt until a tingling sensation was felt then reduction of the amplitude until this sensation disappeared Duration of stimulation per session: 8.2 hours in active group, 7.8 hours in sham group (mean daily application time) Electrodes: flexible electrodes embedded in garment, type not reported

Garland 2007 (*Continued*)

	Electrode placement: negative electrode at patella, positive over anterior distal thigh
Outcomes	Extracted pain outcome: global pain after 12 weeks, described as "Considering your pain and symptoms in your study joint how are you doing today? (VAS)" Extracted function outcome: WOMAC disability subscore after 12 weeks (VAS) No primary outcome reported
Notes	*Due to major protocol violations, all 42 randomised patient of one site were excluded by Garland et al

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Yes	Random number table
Allocation concealment?	Yes	Central randomisation
Free of selective reporting?	Unclear	Quote: "Total WOMAC scores were not a defined outcome in the protocol, but are shown in Tables II(a)-(d)."
Adequate blinding of patients?	Yes	Use of sham device: BIO-1000, indistinguishable from active device, with automatic shut-off as soon as amplitude is reduced (all patients were instructed to reduce intensity just below perception level). Further adjustments required all devices to be restarted.
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	Due to major protocol violations, all 42 randomised patient of 1 site were excluded by original authors. From the other site, all patients randomised were included in the analysis.
Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above
Funding by commercial organisation avoided?	No	Sponsor: BioniCare Medical Technologies
Funding by non-profit organisation?	Unclear	No information provided

Grimmer 1992

Methods	Randomised controlled trial 3-arm parallel group design Trial duration: 1 day No power calculation reported
Participants	60 patients randomised 60 patients with knee OA reported at baseline Study joints: 60 knees Number of females: 37 of 60 (62%) Average age: 66 years
Interventions	Experimental intervention: high frequency TENS, once only in group 1, burst TENS, once only in group 2 Control intervention: sham TENS, once only Duration of treatment period: 1 day Analgesics not allowed Device: Medtronic Neuromed Selectra (dual channel) Self-administered: no Waveform: unclear Pulse width: unclear Pulse frequency: 80 Hz in group 1, 3 Hz trains of 7 80 Hz pulses in group 2 Amplitude: above sensory threshold, strong tolerable tingling paraesthesia Duration of stimulation per session: 30 minutes Electrodes: 4 carbon rubber silicone electrodes, 2 x 3 cm Placement: 4 acupuncture points around the knee: medial (SP9), lateral (GB33), posterior (UB40), anterior (SP10)
Outcomes	Extracted pain outcome: global pain immediately after first and only application, described as "Immediate pain relief (VAS)" No function outcome reported No primary outcome reported
Notes	-

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	Quote: "randomly allocated (by dice) into three groups of 20"
Allocation concealment?	Yes	By a person independent of the study
Free of selective reporting?	Unclear	Insufficient information provided; no access to study protocol

Grimmer 1992 (*Continued*)

Adequate blinding of patients?	Yes	Sham device: Medtronic Neuromed Selectra, with non-functioning leads. Patient were told that a very high frequency current was being tested and that no skin sensation would be felt.
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Yes	Degrees of freedom reported indicate that all randomised patients were included in the analysis
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Itoh 2008

Methods	Randomised controlled trial 2 x 2 factorial design Trial duration: 10 weeks No power calculation reported
Participants	32 patients randomised 32 patients with knee OA reported at baseline Study joints: 32 knees Number of females: 21 of 32 (66%)
Interventions	Experimental intervention: interferential current stimulation*, once per week Control intervention: no intervention, optional use of poultice 16 out of 32 patients (50%) allocated to acupuncture using a factorial design; no evidence for an interaction between treatments Duration of treatment period: 5 weeks Analgesics allowed and intake assessed, but unclear whether intake was similar. Device: HV-F3000 (single channel, 2 pole) Self-administered: no Waveform: sinusoidal Pulse width: not applicable Pulse frequency: amplitude-modulated frequency of 122 Hz (beat frequency) Amplitude: above sensory threshold, up to a tingling sensation, 2 to 3 times above sensory threshold Duration of stimulation per session: 15 minutes Placement: site of tenderness and opposite site Electrodes: 2 disposable electrodes different in size, 809 mm^2 and 5688 mm^2

Itoh 2008 (*Continued*)

Outcomes	Extracted pain outcome: global pain after 10 weeks, described as "Pain intensity (VAS)" Extracted function outcome: WOMAC global scale after 10 weeks (VAS) Primary outcomes: pain intensity, WOMAC global scale	
Notes	*The investigators used the label TENS in their report, but from their description of the intervention it was clear that interferential current stimulation was applied	
Risk of bias		
Item	**Authors' judgement**	**Description**
Adequate sequence generation?	Yes	Computer generated block randomisation. Quote "According to a block randomised allocation table (generated by Sample Size, version 2.0, Int), the enrolled patients were allocated to (1) the control (CT) group, (2) the acupuncture (ACP) group, (3) the transcutaneous electrical nerve stimulation (TENS) group or (4) the acupuncture and TENS (A&T) group."
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Insufficient information provided, no access to study protocol
Adequate blinding of patients?	No	No sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	12 out of 16 (75%) randomised to experimental and 12 out of 16 (75%) randomised to control group were analysed
Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Law 2004

Methods	Randomised controlled trial 4-arm parallel group design Trial duration: 4 weeks No power calculation reported
Participants	36 patients randomised 36 patients with knee OA reported at baseline Study joints: 48 knees* Number of females: 35 of 36 (97%) Average age: 82 years
Interventions	Experimental intervention: 2 Hz TENS in group 1, 100 Hz TENS in group 2, modulation TENS with alternations between 2 to 100 Hz in group 3, 5 times per week in all groups Control intervention: sham TENS, 5 times per week Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and whether intake was similar between groups Device: Han Acupoint Nerve Stimulation LH204H Self-administered: no Waveform: unclear Pulse width and frequency: 576 µsec and 2 Hz in group 1, 200 µsec and 100 Hz in group 2, 576/200 µsec and 2/100 Hz alternation in group 3 Amplitude: above sensory threshold, up to comfortable level, range 25 to 35 mA Duration of stimulation per session: 40 minutes Electrodes: 4 rubber electrodes of 4.5 x 3.8 cm Placement: 4 acupuncture points: ST35, LE4, SP9, GB34
Outcomes	Extracted pain outcome: pain on walking after 4 weeks, described as "intensity of pain felt while walking (VAS)" No function outcome reported No primary outcome reported
Notes	Outcome data were reported on knee level.

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Yes	Quote: "Randomization was carried out by drawing lots from the randomization envelope."
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Insufficient information provided; no access to study protocol

Law 2004 (*Continued*)

Adequate blinding of patients?	Yes	Use of sham device: identical in appearance, internal circuit disconnected, no current passed, indicator light on, digital display of intensity control functioned normally. Quote: "Only therapists who administered treatment to the subjects knew the group allocation, while the subjects and the assessor were not given this information."
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	In total, 3 patients dropped out and were excluded from analysis, as indicated by Gladys Cheing and Pearl Law in personal communication
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Law 2004a

Methods	Randomised controlled trial 2-arm parallel group design Trial duration: 2 weeks Unstratified randomisation Multicentre trial with 2 centres No power calculation reported
Participants	39 patients randomised 39 patients with knee OA reported at baseline Study joints: 39 knees Number of females: 37 of 39 (95%) Average age: 75 years Average BMI: 27 kg/m2 Average disease duration: 7.6 years
Interventions	Experimental intervention: TENS, 5 times per week Control intervention: sham TENS, 5 times per week Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and whether intake was similar between groups Device: ITO model 120Z (dual channel) Self-administered: no Waveform: unclear Pulse width: 200 μsec

Law 2004a (*Continued*)

	Pulse frequency: 100 Hz Amplitude: above sensory threshold, up to a comfortable level, range 25-35 mA Duration of stimulation per session: 40 minutes Electrodes: 4 rubber electrodes, 4.5 x 3.8 cm^2 Placement: acupuncture points: ST35, LE4, SP9, GB34
Outcomes	Extracted pain outcome: pain on walking after 2 weeks, described as "intensity of pain felt while walking (VAS)"** Extracted function outcome: walking disability after 2 weeks, described as "Timed-Up-and-Go test over 3 meters (seconds)" No primary outcome reported
Notes	**Only baseline values reported in the report. Contact established with investigators Law and Cheing, who provided end of treatment and follow-up data.

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Yes	Quote: "'by drawing lots from the randomization envelope without replacement"
Allocation concealment?	Unclear	Quote : "(...) carried out by physiotherapists who performed the treatment"
Free of selective reporting?	No	No results reported for some outcomes mentioned in the methods section, including pain intensity on VAS
Adequate blinding of patients?	Yes	Use of sham device: ITO model 120Z, no current delivered but flashing light on. Quote: "The assessors and subjects were blind to the group allocation. All subjects were told that when the indicator light of the TENS was blinking, it meant the machine was working properly. They might or might not feel any tingling sensation during treatment because the intensity of the current was small."
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	In total, 3 patients dropped out and were excluded from analysis, as indicated by Gladys Cheing and Pearl Law in personal communication
Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above

Law 2004a (*Continued*)

Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Ng 2003

Methods	Randomised controlled trial 3-arm parallel group design Trial duration: 4 weeks Unstratified randomisation No power calculation reported
Participants	24 patients randomised 24 patients with knee OA reported at baseline Study joints: 24 knees Number of females: 23 of 24 (96%) Average age: 85 years
Interventions	Experimental intervention: TENS, 4 times per week, with a total of 8 applications and educational pamphlet Control intervention: educational pamphlet Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and whether intake was similar between groups Device: ITO model F-2 (dual channel) Self-administered: no Waveform: unclear Pulse width: 200 μsec Pulse frequency: 2 Hz Amplitude: above sensory threshold, until strong, tolerable, stroking sensation, preferably evoking phasic muscle contraction Duration of stimulation per session: 20 minutes Electrode placement: acupuncture points ST35, EX-LE-4 Electrodes: 50 x 35 mm^2
Outcomes	Extracted pain outcome: global pain after 4 weeks, described as "pain (Numeric rating scale (NRS))" No function outcome reported No primary outcome reported
Notes	2 out of 3 trial arms qualified for inclusion in this review
Risk of bias	

Ng 2003 (*Continued*)

Item	Authors' judgement	Description
Adequate sequence generation?	Yes	Drawing lots. Quote: "Subjects were randomly assigned by drawing a piece of paper that designated each person to the EA, TENS, and control groups"
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Yes	Quote: "In each evaluation session, three outcome measures were collected." The authors present results of all these 3 outcomes.
Adequate blinding of patients?	No	No sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Unclear	No information provided
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Quirk 1985

Methods	Randomised controlled trial 3-arm parallel group design* Trial duration: 26 weeks No power calculation reported
Participants	38 patients randomised 38 patients with knee OA reported at baseline Study joints: 38 knees Number of females: 29 of 38 (76%) Average age: 63 years
Interventions	Experimental intervention: interferential current + exercise, interferential current stimulation: 3 times per week, exercise twice daily Control intervention: exercise twice daily Duration of treatment period: 4 weeks Analgesics allowed, unclear whether intake was similar between groups

Quirk 1985 (*Continued*)

	Device: Endomed 433 and Vacutron 423 (unclear whether 2 or 4 pole) Self-administered: no Waveform: interferential Pulse width: not applicable Pulse frequency: 0 to 100 Hz 10 minutes, 130 Hz last 5 minutes Amplitude: not reported Duration of stimulation per session: 15 minutes Electrodes: suction electrodes Placement: not reported
Outcomes	Extracted pain outcome: other after 26 weeks, described as "Pain composite score with items rest, post-exercise and night pain (approach unclear; either VAS or verbal scoring technique modified after Newland)"** Extracted function outcome: other algofunctional scale after 26 weeks, described as "Overall clinical condition scale developed by authors, which was based on 3 items for pain; rest-, post-exercise-, night pain and 3 for function; gait, method of climbing stairs and using walking aids (most likely Likert)". No primary outcome reported
Notes	*1 trial arm, in which shortwave diathermy was given, was excluded, **only baseline values with standard error and P values for change from baseline per group reported. No contact could be established with the investigators.

Risk of bias

Item	**Authors' judgement**	**Description**
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	No	No results reported for some outcomes mentioned in the methods section, including maximum knee girth
Adequate blinding of patients?	No	No sham intervention
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	Yes	Quote: "All patients completed their therapy and the first two assessments (baseline and end of treatment), while 92% completed the final assessment (3-6 months after treatment)"
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Yes	See above

Quirk 1985 (*Continued*)

Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Smith 1983

Methods	Randomised sham controlled trial 2-arm parallel group design Trial duration: 8 weeks Randomisation stratified according to gender Multicentre trial with 2 centres No power calculation reported
Participants	32 patients randomised 30 patients with knee OA reported at baseline Study joints: 30 knees Number of females: 20 of 30 (67%) Average age: 68 years
Interventions	Experimental intervention: TENS, twice per week* Control intervention: sham TENS, twice per week* Duration of treatment period: 4 weeks Analgesics intake assessed and found to be similar between groups Device: RDG Tiger Pulse Self-administered: no Waveform: square Pulse width: 80 μsec Pulse frequency: 32 to 50 Hz Amplitude: above sensory threshold, adjusted up to a comfortable tingling sensation Duration of stimulation per session: 20 minutes Electrodes: 4 Lec Tec pads applied with electrode jelly Placement: tender knee points or acupuncture points (SP9, xiyan and UB40)
Outcomes	Extracted pain outcome: global pain after 8 weeks, described as "Weekly pain score derived from daily pain recording (linear 7-point scale)"** No function outcome reported No primary outcome reported
Notes	*Preceded by 1 'standard' week without any treatment, **No pain outcome data presented, investigators were contacted, but we did not receive any reply. This study only contributed in safety analysis.

Risk of bias

Smith 1983 (*Continued*)

Item	Authors' judgement	Description
Adequate sequence generation?	Yes	Computer generated. Quote: "(...) assigned by random computer programme and effected by using sealed envelopes containing cards which defined the treatment (...)".
Allocation concealment?	Unclear	Sealed assignment envelopes, but unclear whether these were opaque and sequential
Free of selective reporting?	No	No results reported for some outcomes mentioned in the methods section, including sleep disturbance
Adequate blinding of patients?	Yes	Use of sham device: RDG Tiger Pulse with broken electrode connection at jack point, no current passed but flashing light on. Quote: "Exactly the same procedure were followed for both the treatment and control groups".
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	15 out of 16 (0.94) randomised to experimental and 15 out of 16 (0.94) randomised to control group were analysed
Incomplete outcome reporting: intention-to-treat analysis performed? Function	Unclear	Not applicable
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Yurtkuran 1999

Methods	Randomised controlled trial 4-arm parallel group design Trial duration: 2 weeks No power calculation reported
Participants	100 patients randomised, 25 per group 100 patients with knee OA reported at baseline Study joints: 100 knees Number of females: 91 of 100 (91%) Average age: 58 years

Yurtkuran 1999 (*Continued*)

Interventions	Experimental intervention: TENS, 5 times per week Control intervention: sham TENS, 5 times per week Duration of treatment period: 2 weeks Unclear whether analgesics were allowed and whether intake was similar between groups Device: MEA-TENS (dual channel) Self-administered: no Waveform: rectangular Pulse width: 1000 μsec Pulse frequency: 4 Hz* Amplitude: above sensory threshold, up to muscle contraction, just below pain tolerance threshold Duration of stimulation per session: 20 minutes Electrodes: 4 small MEA rubber electrodes Placement: 4 acupuncture points SP-9, GB-34, ST-34, ST-35
Outcomes	Extracted pain outcome: global pain after 2 weeks described as "Overall present pain intensity at rest (Likert)" Extracted function outcome: walking disability after 2 weeks, described as "50 foot walking time (in minutes)" No primary outcome reported
Notes	Two out of 4 groups, the electroacupuncture and ice massage groups, were excluded in this review. *Investigators named their intervention AL-TENS, but we coded it low frequency TENS in our analysis.

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	Unclear	Trial protocol not accessible, methods section not explicit about pre-specified outcomes
Adequate blinding of patients?	Yes	Sham device: MEA-TENS with broken lead at jack plug, no current passed but red indicator light on. Quote: "(...) treatment appeared to be done in the same way as the other groups without the subjects suspecting the nature of the stimulation".
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	Investigators reported that "no subject was withdrawn either active or placebo groups". However, the reported degrees of freedom indicate that 5 out of 100 patients were not included. It

Yurtkuran 1999 (*Continued*)

		remained unclear to which of the 4 groups the excluded patients belonged.
Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above
Funding by commercial organisation avoided?	Unclear	No information provided
Funding by non-profit organisation?	Unclear	No information provided

Zizic 1995

Methods	Randomised controlled trial 2-arm parallel group design Trial duration: 34 weeks Multicentre trial with 5 centres No power calculation reported
Participants	78 patients randomised 71 patients with knee OA reported at baseline Study joints: 71 knees Number of females: 33 of 71 (46%)
Interventions	Experimental intervention: pulsed electrostimulation stimulation, daily application Control intervention: sham pulsed electrostimulation, daily application Duration of treatment period: 4 weeks Analgesics allowed, intake assessed and found to be similar between groups. Device: Bionicare Stimulator BIO-1000 Self-administered: yes Waveform: monophasic, spiked Pulse width: unclear Pulse frequency: 100 Hz Amplitude: below sensory threshold, initial increase of amplitude until a tingling sensation was felt then reduction of the amplitude until this sensation disappeared Duration of stimulation: 6 to 10 hours per day Electrodes: 2, unclear whether positioned in knee garment Placement: one on knee, other on thigh directly above that knee
Outcomes	Extracted pain outcome: global pain after 34 weeks described as "Patient evaluation of pain of treated knee (Baseline based on 0-10 VAS, follow-up based on % change from baseline)"

Zizic 1995 (*Continued*)

	Extracted function outcome: patient's global assessment after 34 weeks, described as "Patient evaluation of function of treated knee (Baseline based on 0-10 VAS, follow-up based on % change from baseline)" More than 2 primary outcomes reported (1 physician global evaluation; 2) VAS pain; 3) VAS function)
Notes	-

Risk of bias

Item	Authors' judgement	Description
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Free of selective reporting?	No	No results reported for some outcomes mentioned in the methods, including walking time, tenderness and swelling
Adequate blinding of patients?	Yes	Sham device: BIO-1000, identical in appearance to active device, with automatic shut-off as soon as amplitude is reduced (all patients were instructed to reduce intensity just below perception level)
Incomplete outcome reporting: intention-to-treat analysis performed? Pain	No	38 out of 41 (0.93) randomised to experimental and 33 out of 37 (0.89) randomised to control group were analysed
Incomplete outcome reporting: intention-to-treat analysis performed? Function	No	See above
Funding by commercial organisation avoided?	No	Sponsor: Murray Electronics
Funding by non-profit organisation?	Unclear	No information provided

BMI = body mass index
min = minutes
OA = osteoarthritis
VAS = visual analogue scale

Characteristics of excluded studies

Barr 2004	Less than 50% of patients diagnosed with osteoarthritis of the knee
Bernau 1981	Not a randomised controlled trial, use of active control groups. Additional description: comparing diadynamic electrostimulation df, diadynamic electrostimulation cf and galvanic current
Burch 2008	Use of active control group. Additional description: randomised controlled trial comparing interferential current stimulation followed by patterned muscle stimulation and low-current transcutaneous electrical nerve stimulation (TENS).
Cauthen 1975	Not concerning osteoarthritis
Commandre 1977	No randomised controlled trial (review)
Cottingham 1985a	Not transcutaneous but subcutaneous application
Cottingham 1985b	Not transcutaneous but subcutaneous application. Abstract referring to same RCT as described in Cottingham 1985a.
Durmus 2005	Use of active control group (exercise)
Gaines 2001	Neuromuscular electrostimulation primarily aiming at muscle strengthening
Gaines 2004	Neuromuscular electrostimulation primarily aiming at muscle strengthening
Gibson 1989	Most likely not a randomised controlled trial; percutaneous electrostimulation primarily aiming at muscle strengthening
Godfrey 1979	Faradic electrostimulation with parameters set to increase muscle strength and use of active control (exercise plus low intensity (sham) faradic electrostimulation)
Grigor'eva 1992	No relevant pain or function outcomes
Guven 2003	High voltage galvanic electrostimulation for muscle strengthening
Hamilton 1959	Only 34% of patients suffered OA; use of active controls. Additional description: cross-over design evaluating faradic electrostimulation.
Huang 2000	TENS as part of a combined experimental intervention. Additional description design: 3 groups, Group A receiving auricular acupuncture, diet control and aerobic exercise, Group B like A with addition of TENS and ultrasound, Group C receiving TENS and ultrasound; unclear whether allocation was at random.
Jensen 1991	Use of active control: high frequency TENS versus low frequency TENS

(*Continued*)

Kang 2007	Percutaneous electrostimulation
Katsnelson 2004	Electrode placement not involving knee innervation: transcranial electrostimulation
Komarova 1998	Electrode placement not involving knee innervation: transcranial electrostimulation
Lewis 1984	Cross-over RCT reporting pooled results after completion of all phases. Contact established with Daniel and Beverly Lewis, who were unable to provide results for the first phase (before cross-over)
Lewis 1985	RCT reporting P values of effect only. Contact established with Daniel and Beverly Lewis, who could not provide any additional outcome data, nor could they indicate whether the design concerned a cross-over or a parallel RCT
Lewis 1988	Published abstract addressing the same cross-over RCT reported by Lewis 1994
Lewis 1994	Cross-over RCT reporting pooled results after completion of all phases. Contact established with Daniel and Beverly Lewis, who were unable to provide results for the first phase (before cross-over)
Lone 2003	Not a randomised controlled study. Additional description: before-after study design that was incorrectly labelled as randomised study by original authors.
Lund 2005	Not concerning osteoarthritis
Macchione 1995	Not a randomised controlled trial (review)
Matti 1987	Not concerning osteoarthritis, not a randomised clinical trial. Tetanus-like faradisation electrostimulation with exercise after surgical removal of meniscus, primarily aiming at muscle enhancement. Active control with 10 Hz sinusoidal current application and exercise.
Miranda-Filloy 2005	Electrical muscle stimulation using sport400 (Complex), primarily aiming at muscle strengthening
Mont 2006	Not a randomised clinical trial. Description: comparative study with historical control evaluating pulsed electrostimulation.
Oldham 1995	Neuromuscular electrostimulation primarily aiming at muscle strengthening
Oldham 1997	Electrostimulation primarily aiming at muscle strengthening
Oosterhof 2008	Mixed population, only 4 out of 163 patients reported to have knee, hip or ankle OA
Paillard 2005	Not concerning osteoarthritis (healthy volunteers)
Picaza 1975	Not concerning osteoarthritis and not a randomised controlled trial

(*Continued*)

Salaj 2001	Not a randomised controlled trial, combined multiple interventions in both interventions and control group
Salim 1996	Not a randomised controlled trial (review)
Sluka 1998	Animal study
Sok 2007	Concerns chronic knee pain. First author was contacted by email to verify how many patients had osteoarthritis. No response received. Additional description: article in Korean, using a TENS device, abstract however suggests that parameters were set to strengthen muscles.
Svarcova 1988a	Use of active control groups. Additional description: controlled trial with groups receiving either galvanic electrostimulation or YES ultrasound or pulsed shortwaves. Within these groups, half of the patients received ibuprofen, half received placebo ibuprofen. It was unclear whether allocation was at random.
Svarcova 1988b	See Svarcova 1988a. Double publication of the same study, including the same number of patient and outcome data.
Svarcova 1990	Use of active control group. Additional description: galvanic electrostimulation versus electroacupuncture.
Talbot 2003	Neuromuscular electrostimulation primarily aiming at muscle strengthening
Tam 2004	No relevant pain or function outcomes used
Taylor 1981	Incomplete presentation of data. Additional description: cross-over randomised clinical trial presenting pooled results only. Contact established with Mark Hallett, who was unable to provide data concerning the first phase, before cross-over. We were unable to contact the other authors.
Tulgar 1991	Not concerning osteoarthritis
Volklein 1990	Use of active control group. Additional description: random allocation of patients to 4 different types of diadynamic current.
Weiner 2007	Not transcutaneous but periosteal (needle) application
Zivkovic 2005	Use of active control group. Additional description: the combination of low-energy laser, pulsed electromagnetic field and kinesitherapy was compared to the combination of electrotherapy, pulsed electromagnetic field and kinesitherapy.

OA = osteoarthritis
RCT = randomised controlled trial
TENS = transcutaneous electrical nerve stimulation

Characteristics of ongoing studies

Fary 2008

Trial name or title	ACTRNI2607000492459
Methods	Double-blind, randomised placebo-controlled trial Randomisation method: computer-generated block randomisation with stratification for gender, age and intensity of pain Concealment of allocation: by independent administrator Blinding: patients, those administering treatment/s, those assessing outcomes, those analysing results/data Sample size calculation: reported Analyses based on intention-to-treat principle Trial duration: 26 weeks Sponsored by: non-profit organisation Arthritis Australia and Physiotherapy Research Foundation
Participants	70 patients with primary knee OA to be randomised Study joints: 70 knees Selection criteria: persistent, stable pain for minimum of 3 months, at least 25 mm on a 100 mm VAS
Interventions	Experimental intervention: pulsed electrostimulation, daily Control intervention: sham pulsed electrostimulation, daily Duration of treatment period: 26 weeks Analgesics allowed and measured with diary Device: Metron Digi-10s, adapted by engineer Self-administered: yes Waveform: pulsed, exponentially declining Pulse width: not reported Pulse frequency: 100 Hz Amplitude: below sensory threshold Duration of stimulation: minimally 7 hours per day Electrodes: not reported Electrode placement: not reported Sham device: identical in appearance
Outcomes	Primary outcomes: conflicting information reported in Australian/New Zealand clinical trial register (ANZCTR) and subsequent publication in BMC. In ANZCR reported as pain on VAS, in BMC more than 2 primary outcomes are reported; pain (VAS and WOMAC), function (WOMAC), and patient global assessment (VAS). Main time points of interest are reported consistently as baseline, 4, 16 and 26 weeks. Secondary outcomes: in ANZCTR reported as function (WOMAC) and patient global assessment (VAS); in BMC reported as stiffness (WOMAC 3.1), quality of life (SF-36), global perceived effect scale (GPES), physical activity (Human Activity Profile (HAP) questionnaire plus accelerometers Safety outcomes: in BMC, the recording of adverse events was reported
Starting date	26th of September 2007

Fary 2008 (*Continued*)

Contact information	Robyn E Fary Curtin University of Technology, School of Physiotherapy, Kent Street, Bentley, WA, 6102, Australia Tel: 08 9266 3667 Email: R.Fary@curtin.edu.au
Notes	Status at 17 July 2009: open to recruitment

Palmer 2007

Trial name or title	ISRCTN12912789
Methods	A randomised, sham-controlled trial with 3 parallel arms Randomisation method: not reported Concealment of allocation: not reported Blinding: not reported Sample size calculation: not reported Analyses: not reported whether is based on intention-to-treat principle Trial duration: 6 weeks Sponsored by: not reported
Participants	261 (87 in each arm) patients with primary knee OA to be randomised Study joints: knees Selection criteria: knee pain, radiographic (X-ray) evidence of osteophytes, and at least 1 of the following 3 criteria: 50 years or older, morning stiffness that lasts for less than 30 minutes, crepitus on active movement
Interventions	Experimental intervention: TENS, as much as needed and group education including self-efficacy and exercise training, once per week Control intervention 1: Sham TENS, as much as needed and group education once per week, as described above Control intervention 2: group education once per week, as described above Duration of treatment period: 6 weeks Analgesics: unclear wether analgesic intake is allowed and is measured Device: not reported Self-administered: yes Waveform: not reported Pulse width: not reported Pulse frequency: not reported Amplitude: “strong but comfortable” tingling sensation Duration of stimulation: defined as “as much as needed” Electrodes: not reported Electrode placement: within or close to the site of pain Sham device: identical in appearance, displays are active but there is no current output

Palmer 2007 (*Continued*)

Outcomes	Primary outcome: WOMAC function subscale (at baseline, 3, 6, 12 and 24 weeks) Secondary outcomes: 1. Total WOMAC score and WOMAC pain and stiffness subscale scores (at baseline, 3, 6, 12 and 24 weeks) 2. Knee extensor torque (quadriceps strength) (at baseline, 3, 6, 12 and 24 weeks) 3. Patient global assessment of change (at 3, 6, 12 and 24 weeks) 4. Self-efficacy for exercise (at baseline and 24 weeks) 5. Self-reported exercise adherence (at baseline, 3, 6, 12 and 24 weeks) 6. Logged TENS usage time (at 6 weeks)
Starting date	1 October 2007
Contact information	Dr Shea Palmer Faculty of Health and Social Care University of the West of England Blackberry Hill Bristol BS16 1DD United Kingdom Tel +44 (0)117 328 8919 Email Shea.Palmer@uwe.ac.uk
Notes	Status at 17 July 2009: completed at 30 June 2009

OA = osteoarthritis
TENS = transcutaneous electrical nerve stimulation
VAS = visual analogue scale

Appendix 1 MEDLINE, EMBASE and CINAHL search strategy

OVID MEDLINE	OVID EMBASE	CINAHL through EBSCOhost
search terms for design 1. randomized controlled trial.pt. 2. controlled clinical trial.pt. 3. randomized controlled trial.sh. 4. random allocation.sh. 5. double blind method.sh. 6. single blind method.sh. 7. clinical trial.pt. 8. exp clinical trial/ 9. (clin$ adj25 trial$).ti,ab. 10. ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti,ab. 11. placebos.sh. 12. placebo$.ti,ab. 13. random$.ti,ab. 14. research design.sh. 15. comparative study.sh. 16. exp evaluation studies/ 17. follow up studies.sh. 18. prospective studies.sh. 19. (control$ or prospectiv$ or volunteer$).ti,ab.	*search terms for design* 1. randomized controlled trial.sh. 2. randomization.sh. 3. double blind procedure.sh. 4. single blind procedure.sh. 5. exp clinical trials/ 6. (clin$ adj25 trial$).ti,ab. 7. ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti,ab. 8. placebo.sh. 9. placebo$.ti,ab. 10. random$.ti,ab. 11. methodology.sh. 12. comparative study.sh. 13. exp evaluation studies/ 14. follow up.sh. 15. prospective study.sh. 16. (control$ or prospectiv$ or volunteer$).ti,ab.	*Search terms for design* 1. (MH "Clinical Trials+") 2. (MH "Random Assignment") 3. (MH "Double-Blind Studies") or (MH "Single-Blind Studies") 4. TX (clin$ n25 trial$) 5. TX (sing$ n25 blind$) 6. TX (sing$ n25 mask$) 7. TX (doubl$ n25 blind$) 8. TX (doubl$ n25 mask$) 9. TX (trebl$ n25 blind$) 10. TX (trebl$ n25 mask$) 11. TX (tripl$ n25 blind$) 12. TX (tripl$ n25 mask$) 13. (MH "Placebos") 14. TX placebo$ 15. TX random$ 16. (MH "Study Design+") 17. (MH "Comparative Studies") 18. (MH "Evaluation Research") 19. (MH "Prospective Studies+") 20. TX (control$ or prospectiv$ or volunteer$) 21. S1 or S2 or (…….) or S20
Search terms for Osteoarthritis 20. osteoarthriti$.ti,ab,sh. 21. osteoarthro$.ti,ab,sh. 22. gonarthriti$.ti,ab,sh. 23. gonarthro$.ti,ab,sh. 24. coxarthriti$.ti,ab,sh. 25. coxarthro$.ti,ab,sh. 26. arthros$.ti,ab. 27. arthrot$.ti,ab. 28. ((knee$ or hip$ or joint$) adj3 (pain$ or ach$ or discomfort$)).ti,ab. 29. ((knee$ or hip$ or joint$) adj3 stiff$).ti,ab.	*Search terms for Osteoarthritis* 17. osteoarthriti$.ti,ab,sh. 18. osteoarthro$.ti,ab,sh. 19. gonarthriti$.ti,ab,sh. 20. gonarthro$.ti,ab,sh. 21. coxarthriti$.ti,ab,sh. 22. coxarthro$.ti,ab,sh. 23. arthros$.ti,ab. 24. arthrot$.ti,ab. 25. ((knee$ or hip$ or joint$) adj3 (pain$ or ach$ or discomfort$)).ti,ab. 26. ((knee$ or hip$ or joint$) adj3 stiff$).ti,ab.	*Search terms for Osteoarthritis* 22. osteoarthriti$ 23. (MH "Osteoarthritis") 24. TX osteoarthro$ 25. TX gonarthriti$ 26. TX gonarthro$ 27. TX coxarthriti$ 28. TX coxarthro$ 29. TX arthros$ 30. TX arthrot$ 31. TX knee$ n3 pain$ 32. TX hip$ n3 pain$ 33. TX joint$ n3 pain$ 34. TX knee$ n3 ach$ 35. TX hip$ n3 ach$

Appendix 1. MEDLINE, EMBASE and CINAHL search strategy (continued)

		36. TX joint$ n3 ach$ 37. TX knee$ n3 discomfort$ 38. TX hip$ n3 discomfort$ 39. TX joint$ n3 discomfort$ 40. TX knee$ n3 stiff$ 41. TX hip$ n3 stiff$ 42. TX joint$ n3 stiff$ 43. S22 or S23 or S24….or S42
Search terms for TENS 30. exp electric stimulation therapy/ 31. (electric$ adj (nerve or therapy)).tw. 32. (electric$ adj (stimulation or muscle)).tw. 33. electrostimulation.tw. 34. electroanalgesia.tw. 35. (tens or altens).tw. 36. electroacupuncture.tw. 37. neuromusc$ electric$.tw. 38. high volt.tw. 39. pulsed.tw. 40. (electric$ adj25 current).tw. 41. (electromagnetic or electrotherap$).tw. 42. iontophoresis.tw. 43. transcutaneous nerve stimulation.tw.	*Search terms for TENS* 27. exp electric stimulation therapy/ 28. (electric$ adj (nerve or therapy).tw. 29. (electric$ adj (stimulation or muscle)).tw. 30. electrostimulation.tw. 31. electroanalgesia.tw. 32. (tens or altens).tw. 33. electroacupuncture.tw. 34. neuromusc$ electric$.tw. 35. high volt.tw. 36. pulsed.tw. 37. electric current.sh. 38. (electric$ adj25 current).tw 39. (electromagnetic or electrotherap$).tw. 40. iontophoresis.tw. 41. transcutaneous nerve stimulation.tw.	*Search terms for TENS* 44. (MH "Electric Stimulation+") 45. TX (electric$ n1 nerve) 46. TX (electric$ n1 therapy) 47. TX (electric$ n1 stimulation) 48. TX (electric$ n1 muscle) 49. TX electrostimulation 50. TX electroanalgesia 51. TX tens 52. TX altens 53. TX electroacupuncture 54. TX neuromusc$ electric$ 55. TX high volt 56. TX pulsed 57. TX (electric$ n25 current) 58. TX ((electromagnetic or electrotherap$)) 59. TX iontophoresis 60. TX transcutaneous nerve stimulation 61. S44 or S45 or …. S60
Combining terms 44. or/1-19 45. or/20-29 46. or/30-40 47. and/44-46 48. animal/ 49. animal/ and human/ 50. 48 not 49 51. 47 not 50	*Combining terms* 42. or/1-16 43. or/17-26 44. or/27-37 45. and/42-44 46. animal/ 47. animal/ and human/ 48. 46 not 47 49. 45 not 48	*Combining terms* S21 and S43 and S61

Appendix 2. CENTRAL and PEDro search strategy

CENTRAL	PEDro
Search terms for Osteoarthritis	
#1. (osteoarthritis* OR osteoarthro* OR gonarthriti* OR gonarthro*	1. Electro in title or abstract
OR coxarthriti* OR coxarthro* OR arthros* OR arthrot* OR	2. Method: clinical trial
((knee* OR hip* OR joint*) near/3 (pain* OR ach* OR discomfort*))	3. Body part: thigh or hip
OR ((knee* OR hip* OR joint*) near/3 stiff*)) in Clinical Trials	4. Body part lower leg or knee
#2. MeSH descriptor Osteoarthritis explode all trees	Combination 1. and 2. and 3.
Search terms for TENS	Combination 1. and 2. and 4.
#3. MeSH descriptor Electric Stimulation Therapy explode all trees	1. TENS in title or abstract
#4. electric* near/ (nerve or therapy) in Clinical Trials	2. Method: clinical trial
#5. electric* near/ (stimulation or muscle) in Clinical Trials	3. Body part: thigh or hip
#6. electrostimulation in Clinical Trials	4. Body part lower leg or knee
#7. electroanalgesia in Clinical Trials	Combination 1. and 2. and 3.
#8. tens or altens in Clinical Trials	Combination 1. and 2. and 4.
#9. electroacupuncture in Clinical Trials	Combine all
#10. neuromusc* electric* in Clinical Trials	
#11. high volt in Clinical Trials	
#12. pulsed in Clinical Trials	
#13. (electric* near/25 current) in Clinical Trials	
#14. (electromagnetic or electrotherap*) in Clinical Trials	
#15. iontophoresis in Clinical Trials	
#16. transcutaneous nerve stimulation in Clinical Trials	
Combining terms	
#17. (#3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR	
#11 OR #12 OR #13 OR #14 OR #15 OR #16)	
#18. (#1 OR #2)	
#19. (#17 AND #18) in Clinical Trials	

References

1. Jüni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. *Best practice & research. Clinical rheumatology* 2006;20(4):721-740.
2. Solomon L, Kelly Wn HEDJRSSCB. Clinical features of osteoarthritis. *Textbook of Rheumatology*. Philadelphia: WB Saunders, 1997:1383-1393.
3. Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. *BMC musculoskeletal disorders* 2007;8:51-51.
4. Jamtvedt G, Dahm KT, Christie A, Moe RH, Haavardsholm E, Holm I, et al. Physical therapy interventions for patients with osteoarthritis of the knee: an overview of systematic reviews. *Physical therapy* 2008;88(1):123-136.
5. Gezondheidsraad. Efficacy of physical therapy: electrostimulation, laser therapy, ultrasound therapy (own translation). *Den Haag: Gezondheidsraad*, 1999.
6. Osiri M, Welch V, Brosseau L, Shea B, McGowan J, Tugwell P, et al. Transcutaneous electrical nerve stimulation for knee osteoarthritis. *Cochrane Database of Systematic Reviews* 2000(4).
7. Carroll D, Moore RA, McQuay HJ, Fairman F, Tramer M, Leijon G. Transcutaneous electrical nerve stimulation (TENS) for chronic pain. *Cochrane Database of Systematic Reviews* 2001(3).
8. Melzack R, Wall P. Pain mechanisms: A new theory. *Science* 1965;150:971-977.
9. Andersson SA, Hansson G, Holmgren E, Renberg O. Evaluation of the pain suppression effect of different frequencies of peripheral electrical stimulation in chronic pain conditions. *Acta Orthopaedica Scandinavia* 1976;47:149-157.
10. Mayer DJ, Prince DD, Snyder-Mackper L RA. The neurobiology of pain. *Clinical Electrophysiology, Electrotherapy and Electrophysiologic Testing*. Baltimore, MD: Williams & WIlkins, 1989:141-201.
11. Grimmer K. A controlled double blind study comparing the effects of strong burst mode TENS and High Rate TENS on painful osteoarthritic knees. *Australian Journal of Physiotherapy* 1992;38(1):49-56.
12. Haddad JB, Obolensky AG, Shinnick P. The biologic effects and the therapeutic mechanism of action of electric and electromagnetic field stimulation on bone and cartilage: new findings and a review of earlier work. *Journal of alternative and complementary medicine (New York, N.Y.)* 2007;13(5):485-490.

13. Fary RE, Carroll GJ, Briffa TG, Gupta R, Briffa NK, Fary Robyn E, et al. The effectiveness of pulsed electrical stimulation (E-PES) in the management of osteoarthritis of the knee: a protocol for a randomised controlled trial. *BMC Musculoskeletal Disorders* 2008;9:18-18.
14. Sluka KA, Walsh D. Transcutaneous electrical nerve stimulation: basic science mechanisms and clinical effectiveness. *The journal of pain : official journal of the American Pain Society* 2003;4(3):109-121.
15. Brosseau L, Yonge K, Marchand S, Robinson V, Osiri M, Wells G, et al. Efficacy of transcutaneous electrical nerve stimulation for osteoarthritis of the lower extremities: a meta-analysis. *Physical Therapy Reviews* 2004;9:213-233.
16. Wadsworth H, Chanmugan A. *Electrophysical agents in physiotherapy*. Marrickville, Australia: Science Press, 1980.
17. Altman R, Brandt K, Hochberg M, Moskowitz R, Bellamy N, Bloch DA, et al. Design and conduct of clinical trials in patients with osteoarthritis: recommendations from a task force of the Osteoarthritis Research Society. Results from a workshop. *Osteoarthritis Cartilage* 1996;4(4):217-243.
18. Pham T, van der Heijde D, Altman RD, Anderson JJ, Bellamy N, Hochberg M, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. *Osteoarthritis Cartilage* 2004;12(5):389-399.
19. Reichenbach S, Sterchi R, Scherer M, Trelle S, Burgi E, Burgi U, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med* 2007;146(8):580-590.
20. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews [see comments]. [Review] [31 refs]. *BMJ* 1994;309(6964):1286-1291.
21. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.
22. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*: The Cochrane Collaboration, 2008.
23. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336(7650):924-926.
24. Cohen, J. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

25. Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Burgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ* 2009;339:b3244.
26. Bellamy N. Outcome measurement in osteoarthritis clinical trials. *J Rheumatol* 1995;22(suppl.43):49-51.
27. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.
28. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.
29. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
30. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54(10):1046-1055.
31. Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, Sterne JA, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 2005;366(9487):726-732.
32. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;18(20):2693-708.
33. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 2000;19(22):3127-3131.
34. Clegg DO, Reda DJ, Harris CL, Klein MA, O'Dell JR, Hooper MM, et al. Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. *New England Journal of Medicine* 2006;354(8):795-808.
35. Palmer S. Effects of transcutaneous electrical nerve stimulation (TENS) and exercise on knee osteoarthritis (OA): a randomised controlled trial. *ISTRCTN*, 2007.
36. Bal S, Turan Y, Gurgan A. The effectiveness of transcutaneous electrical nerve stimulation in patients with knee osteoarthritis. *Journal of Rheumatology and Medical Rehabilitation* 2007;18(1):1-5.
37. Cetin N, Aytar A, Atalay A, Akman MN, Cetin Nuri, Aytar Aydan, et al. Comparing hot pack, short-wave diathermy, ultrasound, and TENS on isokinetic strength, pain, and functional status of women with osteoarthritic knees: a single-blind, randomized, controlled trial. *American Journal of Physical Medicine & Rehabilitation* 2008;87(6):443-451.

38. Cheing GL, Hui-Chan CW, Chan KM, Cheing Gladys LY, Hui-Chan Christina WY. Does four weeks of TENS and/or isometric exercise produce cumulative reduction of osteoarthritic knee pain? *Clinical Rehabilitation* 2002;16(7):749-760.
39. Cheing GL, Hui-Chan CW, Cheing Gladys LY, Hui-Chan Christina WY. Would the addition of TENS to exercise training produce better physical performance outcomes in people with knee osteoarthritis than either intervention alone? *Clinical Rehabilitation 2004* 2004;18(5):487-497.
40. Cheing GLY, Hui-Chan CWY, Chan KM. Does four weeks of TENS and/or isometric exercise produce cumulative reduction of osteoarthritic knee pain? *Pain Reviews* 2002;9(3/4):141-151.
41. Cheing GL, Tsui AY, Lo SK, Hui-Chan CW, Cheing Gladys LY, Tsui Amy YY, et al. Optimal stimulation duration of tens in the management of osteoarthritic knee pain. *Journal of Rehabilitation Medicine* 2003;35(2):62–68-62–68.
42. Law PPW, Cheing GLY, Tsui AYY. Does transcutaneous electrical nerve stimulation improve the physical performance of people with knee osteoarthritis? *Journal of Clinical Rheumatology* 2004;10(6):295–299-295–299.
43. Smith CR, Lewith GT, Machin D. TNS and osteo-arthritic pain. Preliminary study to establish a controlled method of assessing transcutaneous nerve stimulation as a treatment for the pain caused by osteo-arthritis of the knee. *Physiotherapy* 1983;69(8):266-268.
44. Adedoyin RA, Olaogun MOB, Oyeyemi AL. Transcutaneous electrical nerve stimulation and interferential current combined with exercise for the treatment of knee osteoarthritis: a randomised controlled trial. *Hong Kong Physiotherapy Journal* 2005;23:13-19.
45. Law PP, Cheing GL, Law Pearl PW, Cheing Gladys LY. Optimal stimulation frequency of transcutaneous electrical nerve stimulation on people with knee osteoarthritis. *Journal of Rehabilitation Medicine* 2004;36(5):220–225-220–225.
46. Fargas-Babjak A, Rooney P, Gerecz E. Randomized trial of Codetron for pain control in osteoarthritis of the hip/knee. *Clinical Journal of Pain* 1989;5(2):137-141.
47. Ng MM, Leung MC, Poon DM, Ng MML, Leung Mason CP, Poon DMY. The effects of electro-acupuncture and transcutaneous electrical nerve stimulation on patients with painful osteoarthritic knees: a randomized controlled trial with follow-up evaluation. *Journal of Alternative & Complementary Medicine* 2003;9(5):641–649-641–649.
48. Yurtkuran M, Kocagil T. TENS, electroacupuncture and ice massage: comparison of treatment for osteoarthritis of the knee. *American Journal of Acupuncture* 1999;27(3-4):133-140.

49. Adedoyin RA, Olaogun MOB, Fagbeja OO. Effect of interferential current stimulation in management of osteo-arthritic knee pain. *Physiotherapy* 2002;88(8):493-499.
50. Defrin R, Ariel E, Peretz C, Defrin Ruth, Ariel Efrat, Peretz Chava. Segmental noxious versus innocuous electrical stimulation for chronic pain relief and the effect of fading sensation during treatment. *Pain* 2005;115(1-2):152–160-152–160.
51. Itoh K, Hirota S, Katsumi Y, Ochi H, Kitakoji H, Itoh Kazunori, et al. A pilot study on using acupuncture and transcutaneous electrical nerve stimulation (TENS) to treat knee osteoarthritis (OA). *Chinesische Medizin* 2008;3:2-2.
52. Quirk AS, Newman RJaNKJ. An evaluation of interferential therapy, shortwave diathermy and exercise in the treatment of osteoarthrosis of the knee. *Physiotherapy* 1985;71:55-57.
53. Garland D, Holt P, Harrington JT, Caldwell J, Zizic T, Cholewczynski J, et al. A 3-month, randomized, double-blind, placebo-controlled study to evaluate the safety and efficacy of a highly optimized, capacitively coupled, pulsed electrical stimulator in patients with osteoarthritis of the knee. *Osteoarthritis & Cartilage* 2007;15(6):630-637.
54. Zizic TM, Hoffman KC, Holt PA, Hungerford DS, O'Dell JR, Jacobs MA, et al. The treatment of osteoarthritis of the knee with pulsed electrical stimulation. *Journal of Rheumatology* 1995;22(9):1757-1761.
55. Lewis B. Analgesic efficacy of transcutaneous electrical nerve stimulation compared with a non-steroidal anti-inflammatory drug in osteoarthritis [abstract]. *Aust.NZ.J Med.Suppl* 1985;15:189-189.
56. Lewis B, Lewis D, Cumming G. The analgesic efficacy of transcutaneous electrical nerve stimulation (TENS) compared with a non-steroidal anti-inflammatory drug (naprosyn) in painful osteoarthritis (OA) of the knee. [abstract]. *Aust.NZ.J Med.Suppl* 1988;18:224-224.
57. Taylor P, Hallett M, Flaherty L. Treatment of osteoarthritis of the knee with transcutaneous electrical nerve stimulation. *Pain* 1981;11(2):233-240.
58. Burch FX, Tarro JN, Greenberg JJ, Carroll WJ. Evaluating the benefits of patterned stimulation in the treatment of osteoarthritis of the knee A multi-center, randomized, single-blind, controlled study with an independent masked evaluator. *Osteoarthritis & Cartilage* 2008;16(8):865-872.
59. Jensen H, Zesler R, Christensen T. Transcutaneous electrical nerve stimulation (TNS) for painful osteoarthrosis of the knee. *International Journal of Rehabilitation Research* 1991;14(4):356-358.

60. Volklein R, Callies R. Changes in pain by different types of diadynamic current in gonarthrosis and lumbar syndrome. *Zeitschrift fur Physiotherapie* 1990;42(2):113-118.
61. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology* 2000;53(11):1119-1129.
62. Nüesch E, Juni P. Commentary: Which meta-analyses are conclusive? *International journal of epidemiology* 2009;38(1):298-303.
63. Deyo RA, Wash NE, Schoenfeld LS, Ramamurthy S. Can trials of physical treatments be blinded: the example of transcutaneous electrical nerve stimulation for chronic pain. *American Journal of Physical Medicine and Rehabilitation* 1990;69:6-10.
64. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, 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.
65. Regence Medical Policy, Assessed J. Durable medical equipment section. Electrical stimulation for the treatment of arthritis. *Available from: http://blue.regence.com/trgmedpol/dme/dme70.html*.

Article 5

Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study

Britta Tendal,[1)] Julian P. Higgins,[2)] Peter Jüni,[3,4)] Asbjørn Hróbjartsson,[1)] Sven Trelle,[3,4)] Eveline Nüesch,[3,4)] Simon Wandel,[3,4)] Anders W. Jørgensen,[1)] Katarina Gesser,[5)] Søren Ilsøe-Kristensen,[5)] Peter C. Gøtzsche[1)]

From [1)]The Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark; [2)]MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, United Kingdom; [3)]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Switzerland; [4)]CTU Bern, Bern University Hospital, Switzerland; [5)]The Faculty of Pharmaceutical Sciences, University of Copenhagen, Denmark

BMJ. 2009 Aug 13; 339: b3128

Abstract

Objective To study the inter-observer variation related to extraction of continuous and numerical rating scale data from trial reports for use in meta-analyses.

Design Observer agreement study.

Data sources A random sample of 10 Cochrane reviews that presented a result as a standardised mean difference (SMD), the protocols for the reviews and the trial reports (n=45) were retrieved.

Data extraction Five experienced methodologists and five PhD students independently extracted data from the trial reports for calculation of the first SMD result in each review. The observers did not have access to the reviews but to the protocols, where the relevant outcome was highlighted. The agreement was analysed at both trial and meta-analysis level, pairing the observers in all possible ways (45 pairs, yielding 2025 pairs of trials and 450 pairs of meta-analyses). Agreement was defined as SMDs that differed less than 0.1 in their point estimates or confidence intervals.

Results The agreement was 53% at trial level and 31% at meta-analysis level. Including all pairs, the median disagreement was SMD=0.22 (interquartile range 0.07-0.61). The experts agreed somewhat more than the PhD students at trial level (61%v 46%), but not at meta-analysis level. Important reasons for disagreement were differences in selection of time points, scales, control groups, and type of calculations; whether to include a trial in the meta-analysis; and data extraction errors made by the observers. In 14 out of the 100 SMDs calculated at the meta-analysis level, individual observers reached different conclusions than the originally published review.

Conclusions Disagreements were common and often larger than the effect of commonly used treatments. Meta-analyses using SMDs are prone to observer variation and should be interpreted with caution. The reliability of meta-analyses might be improved by having more detailed review protocols, more than one observer, and statistical expertise.

Introduction

Systematic reviews of clinical trials, with meta-analyses if possible, are regarded as the most reliable resource for decisions about prevention and treatment. They should be based on a detailed protocol that aims to reduce bias by pre-specifying methods and selection of studies and data.[1] However, as meta-analyses are usually based on data that have already been processed, interpreted, and summarised by other researchers, data extraction can be complicated and can lead to important errors.[2]

There is often a multiplicity of data in trial reports that makes it difficult to decide which ones to use in a meta-analysis. Furthermore, data are often incompletely reported,[2 3] which makes it necessary to perform calculations or impute missing data, such as missing standard deviations. Different observers may get different results, but previous studies on observer variation have not been informative, because of few observers, few trials, or few data.[4 5] We report here a detailed study of observer variation that explores the sources of disagreement when extracting data for calculation of standardised mean differences.

Methods

Using a computer generated list of random numbers, we selected a random sample of 10 recent Cochrane reviews published in the Cochrane Library in issues 3 or 4 in 2006 or in issues 1 or 2 in 2007. We also retrieved the reports of the randomised trials that were included in the reviews and the protocols for each of the reviews. Only Cochrane reviews were eligible, as they are required to have a pre-specified published protocol.

We included reviews that reported at least one result as a standardised mean difference (SMD). The SMD is used when trial authors have used different scales for measuring the same underlying outcome—for example, pain can be measured on a visual analogue scale or on a 10-point numeric rating scale. In such cases, it is necessary to standardise the measurements on a uniform scale before they can be pooled in a meta-analysis. This is typically achieved by calculating the SMD for each trial, which is the difference in means between the two groups, divided by the pooled standard deviation of the measurements.[1] By this transformation, the outcome becomes dimensionless and the scales become comparable, as the results are expressed in standard deviation units.

The first SMD result in each review that was not based on a subgroup result was selected as our index result. The index result had to be based on two to 10 trials and on published data

only (that is, there was no indication that the review authors had received additional outcome data from the trial authors).

Five methodologists with substantial experience in meta-analysis and five PhD students independently extracted the necessary data from the trial reports for calculation of the SMDs. The observers had access to the review protocols but not to the completed Cochrane reviews and the SMD results. An additional researcher (BT) highlighted the relevant outcome in the protocols, along with other important issues such as pre-specified time points of interest, which intervention was the experimental one, and which was the control. If information was missing regarding any of these issues, the observers decided by themselves what to select from the trial reports. The observers received the review protocols, trial reports, and a copy of the Cochrane Handbook for Systematic Reviews[6] as PDF files.

The data extraction was performed during one week when the 10 observers worked independently at the same location in separate rooms. The observers were not allowed to discuss the data extraction. If the data were available, the observers extracted means, standard deviations, and number of patients for each group; otherwise, they could calculate or impute the missing data, such as from an exact P value. The observers also interpreted the sign of the SMD results—that is, whether a negative or a positive result indicated superiority of the experimental intervention. If the observers were uncertain, the additional researcher retrieved the paper that originally described the scale, and the direction of the scale was based on this information. All calculations were documented, and the observers provided information about any choices they made regarding multiple outcomes, time points, and data sources in the trial reports. During the week of data extraction the issue of whether the observers could exclude trials emerged, as there were instances where the observers were unable to locate any relevant data in the trial reports or felt that the trial did not meet the inclusion criteria in the Cochrane protocol. It was decided that observers could exclude trials, and the reasons for exclusion were documented.

Based on the extracted data, the additional researcher calculated trial and meta-analysis SMDs for each observer using Comprehensive Meta-Analysis Version 2. To allow comparison with the originally published meta-analyses, the same method (random effects or fixed effect model) was used as that in the published meta-analysis. In cases where the observers had extracted two sets of data from the same trial—for example, because there were two control groups—the data were combined so that only a single SMD resulted from each trial.[1]

Agreement between pairs of observers was assessed at both meta-analysis and trial level, pairing the 10 observers in all possible ways (45 pairs). This provides an indication of the likely agreement that might be expected in practice, since two independent observers are recommended when extracting data from papers for a systematic review.[1 2 5 6] Agreement was defined as SMDs that differed less than 0.1 in their point estimates and in their confidence intervals. The cut point of 0.1 was chosen because many commonly used treatments have an effect of 0.1 to 0.5 compared with placebo2; furthermore, an error of 0.1 can be important when two active treatments have been compared, for there is usually little difference between active treatments. Confidence intervals were not calculated, as the data from the pairings were not independent.

To determine the variation in meta-analysis results that could be obtained from the multiplicity of different SMD estimates across observers, we conducted a Monte Carlo simulation for each meta-analysis. In each iteration of the simulation, we randomly sampled one observer for each trial and entered his or her SMD (and standard error) for that trial into a meta-analysis. Thus each sampled meta-analysis contained SMD estimates from different observers. If the sampled observer excluded the trial from his or her meta-analysis, the simulated meta-analysis also excluded that trial. We examined the distribution of meta-analytic SMD estimates across 10 000 simulations.

Results

The flowchart for inclusion of meta-analyses is shown in figure 1. Out of 32 potentially eligible meta-analyses, the final sample consisted of 10.[7-16] The 10 meta-analyses comprised 45 trials, which yielded 450 pairs of observers at the meta-analysis level and 2025 pairs at the trial level.

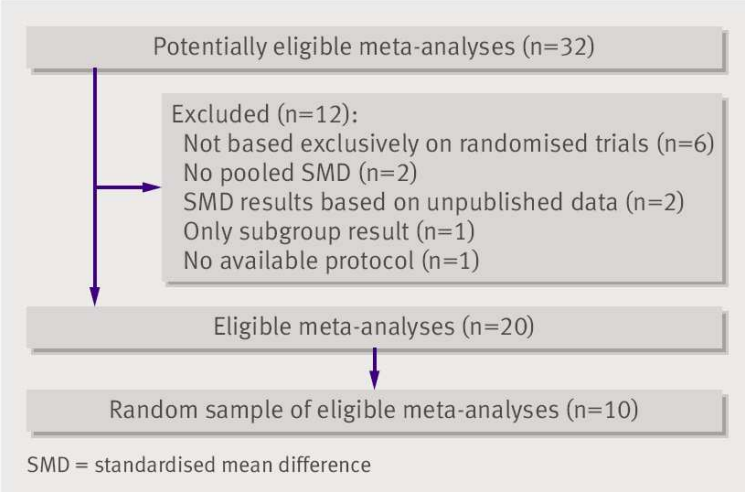

Figure 1 Flowchart for selection of meta-analyses

The level of information in the review protocols is given in table 1. None of the review protocols contained information on which scales should be preferred. Three protocols gave information about which time point to select and four mentioned whether change from baseline or values after treatment should be preferred. Nine described which type of control group to select, but none reported any hierarchy among similar control groups or any intentions to combine such groups.

	Meta-analysis									
Information	Gava et al[7]	Woodford et al[8]	Martinez et al[9]	Orlando et al[10]	Buckley et al[11]	Ipser et al[12]	Mistiaen et al[13]	Afolabi et al[14]	Uman et al[15]	Moore et al[16]
Possible control group(s)	√		√	√	√	√	√	√	√	√
Hierarchy of control groups								√*		√*
Which time point to select					√		√			√
Whether to use change from baseline or values after treatment		√		√	√	√				
Hierarchy of measuring methods or scales										
*Only one possible control group stated										

Table 1 Level of information provided in the 10 meta-analysis protocols used in this study for data extraction

The outcomes analysed in the 10 meta-analyses were diverse: in six, the outcome was a clinician reported score (three symptom scores, one general functioning score, one hepatic density score, and one neonatal score); in one, it was objective (range of movement in ankle joints); and in three, it was self reported (pain, tinnitus, and patient knowledge).

Agreement at trial level

In table 2 the different levels of agreement are shown. Across trials, the agreement was 53% for the 2025 pairs (61% for the 450 pairs of methodologists, 46% for the 450 pairs of PhD students, and 52% for the 1125 mixed pairs). The agreement rates for the individual trials ranged from 4% to 100%. Agreement between all observers was found for four of the 45 trials.

Observer pairs	No (%) of pairs in agreement
Trial level	
All pairs (n=2025):	1068 (53)
Methodologists (n=450)	273 (61)
PhD students (n=450)	209 (46)
Mixed pairs (n=1125)	586 (52)
Meta-analysis level	
All pairs (n=450):	138 (31)
Methodologists (n=100)	33 (33)
PhD students (n=100)	27 (27)
Mixed pairs (n=250)	78 (31)

*Agreement defined as SMDs that differed less than 0.1 in their point estimates and in their 95% confidence intervals.

Table 2 Levels of overall agreement between observer pairs in the calculated standardised mean differences (SMDs)* from 10 meta-analyses (which comprised a total of 45 trials)

Table 3 presents the reasons for disagreement, which fell into three broad categories: different choices, exclusion of a trial, and data extraction errors. The different choices mainly concerned cases with multiple groups to choose from when selecting the experimental or the control groups (15 trials), which time point to select (nine trials), which scale to use (six trials), and different ways of calculating or imputing missing numbers (six trials). The most common reasons for deciding to exclude a trial was that the trial did not meet the inclusion criteria described in the protocol for the review (14 trials) and that the reporting was so unclear that data extraction was not possible (14 trials). Data extraction errors were less common but involved misinterpretation of the direction of the effect in four trials.

Reason for disagreement	No of trials*
Different choices regarding:	
Groups, pooling, splitting	15
Timing	9
Scales	6
Different calculations or imputations	6
Dropouts	4
Use of change from baseline or values after treatment	4
Individual patient data	1
Exclusion of trials because:	
Did not meet protocol inclusion criteria	14
Reporting unclear	14
Missing data	7
Could not or would not calculate	2
Only change from baseline or only values after treatment	2
Errors due to:	
Misreading or typing error	4
Direction of effect	4
Standard error taken as standard deviation	2
Rounding	1
Calculation error	1

*There may be more than one reason for disagreement per trial.

Table 3 Reasons for disagreement among the 41 trials on which the observer pairs disagreed in the calculated standardised mean differences

The importance of which standard deviation to use was underpinned in a trial that did not report standard deviations.[17] The only reported data on variability were F test values and P values from a repeated measure, analysis of variance, performed on changes from baseline. The five PhD students excluded the trial because of the missing data, whereas the five experienced methodologists imputed five different standard deviations. One used a standard deviation from the report originally describing the scale, another used the average standard deviation reported in the other trials in the meta-analysis, and the other three observers calculated standard deviations based on the reported data, using three different methods. In

addition, one observer selected a different time point from the others. The different standard deviations resulted in different trial SMDs ranging from −1.82 to 0.34 in their point estimates.

Agreement at meta-analysis level

Across the meta-analyses, the agreement was 31% for the 450 pairs (33% for the 100 pairs of methodologists, 27% for the 100 pairs of PhD students, and 31% for the 250 mixed pairs) (table 2). The agreement rates for the individual meta-analyses ranged from 11% to 80% (table 4). Agreement between all observers was not found for any of the 10 meta-analyses.

	No (%) of pairs in agreement			
Meta-analysis	All pairs (n=45)	Methodologist (n=10)	Students (n=10)	Mixed pairs (n=25)
Gava et al[7]	6 (13)	1 (10)	0 (0)	5 (20)
Woodford et al[8]	11 (24)	2 (20)	1 (10)	8 (32)
Martinez et al[9]	7 (16)	3 (30)	1 (10)	3 (12)
Orlando et al[10]	5 (11)	1 (10)	2 (20)	2 (8)
Buckley et al[11]	6 (13)	1 (10)	1 (10)	4 (16)
Ipser et al[12]	13 (29)	4 (40)	2 (20)	7 (28)
Mistiaen et al[13]	16 (36)	6 (60)	2 (20)	8 (32)
Afolabi et al[14]	28 (62)	6 (60)	6 (60)	16 (64)
Uman et al[15]	36 (80)	6 (60)	10 (100)	20 (80)
Moore et al[16]	10 (22)	3 (30)	2 (20)	5 (20)

*Agreement defined as SMDs that differed less than 0.1 in their point estimates and in their 95% confidence intervals.

Table 4 Levels of agreement at the meta-analysis level between observer pairs in the calculated standardised mean differences (SMDs) from 10 meta-analyses*

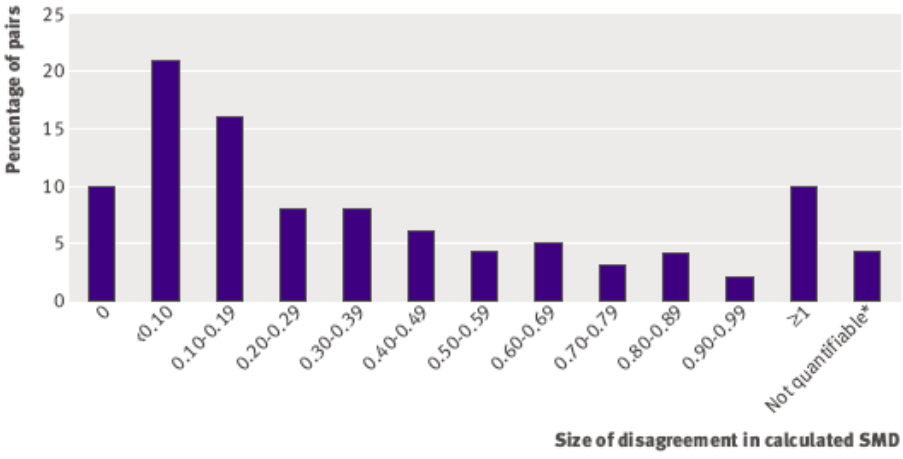

Figure 2 Sizes of the disagreements between observer pairs in the calculated standardised mean differences (SMDs) from 10 meta-analyses. Comparisons are at the meta-analysis level. (*All the underlying trials were excluded)

The distribution of the disagreements is shown in figure 2. Ten per cent agreed completely, 21% had a disagreement below our cut point of 0.1, 38% had a disagreement between 0.1 and

0.49, and 28% disagreed by at least 0.50 (including 10% that had disagreements of ≥1). The last 18 pairs (4%) were not quantifiable since one observer excluded all the trials from two meta-analyses. The median disagreement was SMD=0.22 for the 432 quantifiable pairs with an interquartile range from 0.07 to 0.61. There were no differences between the methodologists and the PhD students (table 2).

Figure 3 shows the SMDs calculated by each of the 10 observers for the 10 meta-analyses, and the results from the originally published meta-analyses. Out of the total of 100 calculated SMDs, seven values corresponding to significant results in the originally published meta-analyses were now non-significant, three values corresponding to non-significant results were now significant, and four values, which were related to the same published meta-analysis, showed a significantly beneficial effect for the control group whereas the original publication reported a significantly beneficial effect for the experimental group.[11] The SMDs for this meta-analysis had particularly large disagreements, partly because only two trials were included, leaving less possibility for the pooled result to average out. The reasons for the large disagreements were diverse and included selection of different time points, control groups, intervention groups, measurement scales, and whether to exclude one of the trials.

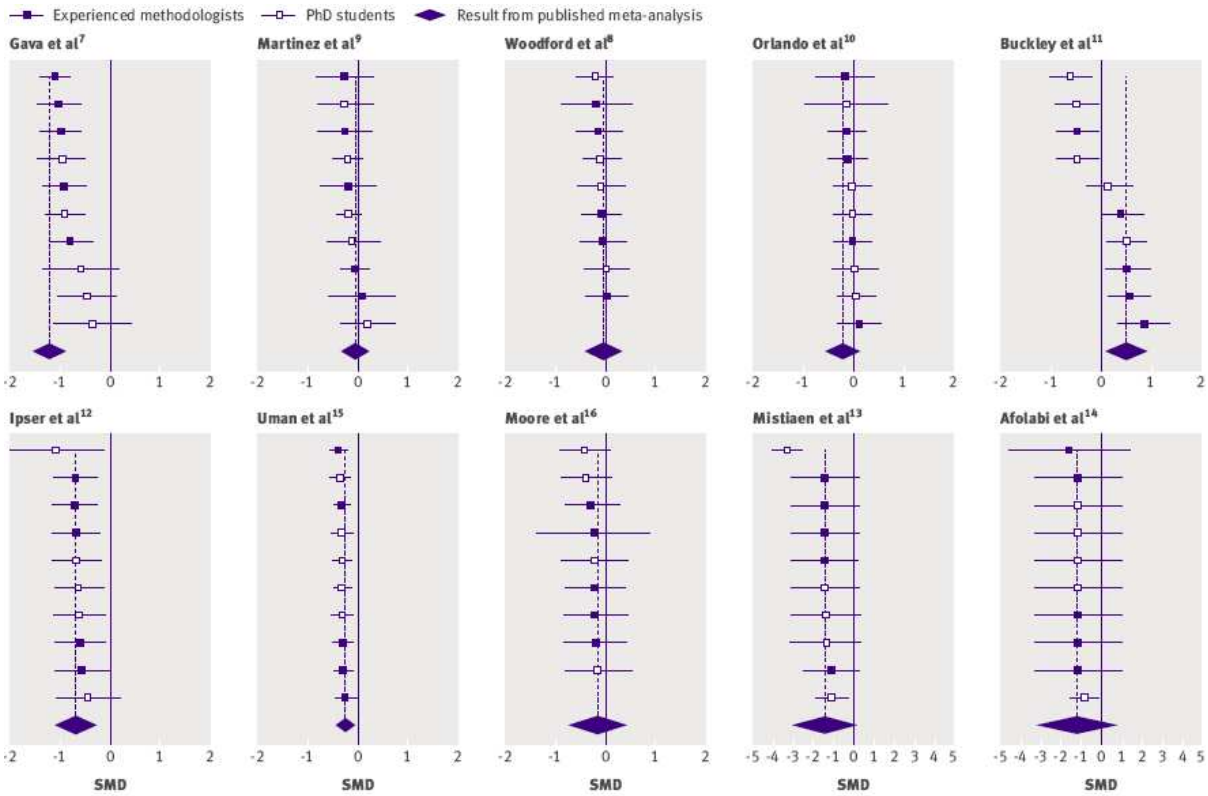

Figure 3 Forest plots of standardised mean differences (SMDs) and 95% confidence intervals calculated from data from each of the 10 observers for the 10 meta-analyses

The results of the Monte Carlo investigation are presented in figure 4. For four of the 10 meta-analyses[7 11 13 14] there was considerable variation in the potential SMDs, allowing for differences in SMDs of up to 3. In two of these, around half of the distribution extended beyond even the confidence interval for the published result of the meta-analysis.711 The other meta-analyses had three and two trials respectively, and the distributions reflect the wide scatter of SMDs from these trials.

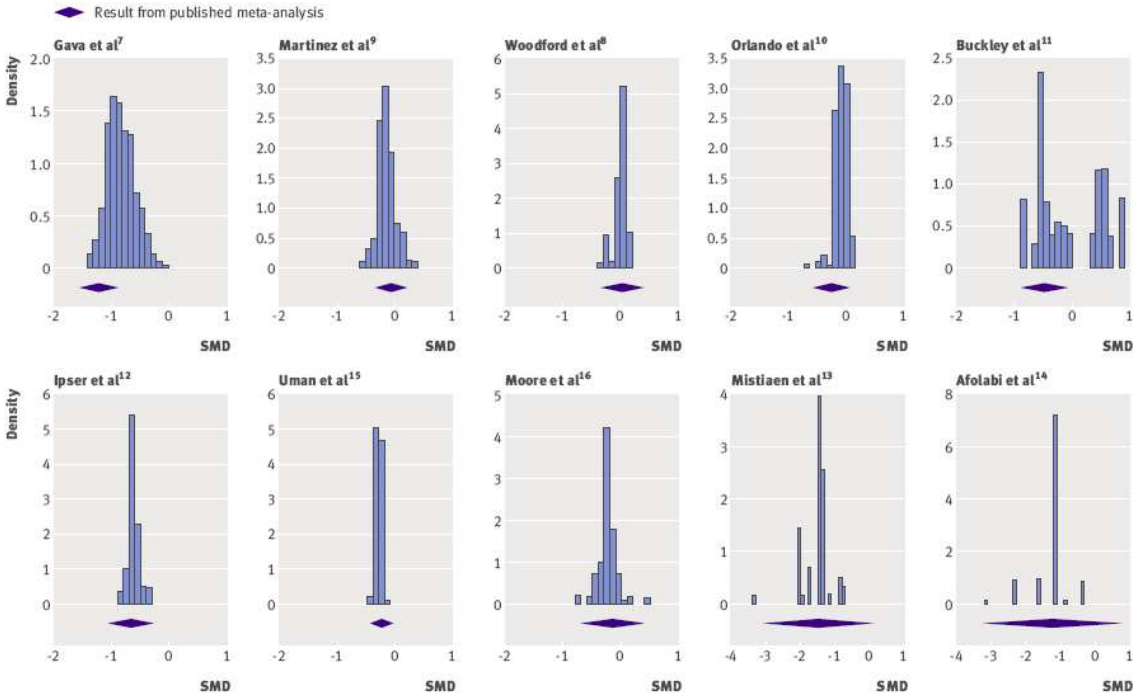

Figure 4 Histograms of standardised mean differences (SMD) estimated in the Monte Carlo simulations for each of 10 meta-analyses

Discussion

We found that disagreements between observers were common and often large. Ten per cent of the disagreements at the meta-analysis level amounted to an SMD of at least 1, which is far greater than the effect of most of the treatments we use compared with no treatment. As an example, the effect of inhaled corticosteroids on asthma symptoms, which is generally regarded as substantial, is 0.49.[18] Important reasons for disagreement were differences in selection of time points, scales, control groups, and type of calculations, whether to include a trial in the meta-analysis, and finally data extraction errors made by the observers.

The disagreement depended on the reporting of data in the trial reports and on how much room was left for decision in the review protocols. One of the reviews exemplified the variation arising from a high degree of multiplicity in the trial reports combined with a review protocol leaving much room for choice.[11] In the review protocol, the time point was described

as “long term (more than 26 weeks),” but in the two trials included in the meta-analysis there were several options. For one trial,[19] there were two: end of treatment (which lasted 9 months) or three month follow up. For the other,[20-22] there were three: 6, 12, and 18 month follow-up (treatment lasted 3 weeks). The observers used all the different time points, and all had a plausible reason for their choice: in concordance with the time point used in the other trial, the maximum period of observation, and the least drop out of patients.

Strengths and weaknesses

The primary strength of our study is that we took a broad approach and showed that there are other important sources of variation in meta-analysis results than simple errors. Furthermore, we included a considerable number of experienced as well as inexperienced observers and a large number of trials to elucidate the sources of variation and their magnitude. Finally, the study setup ensured independent observations according to the blueprint laid out in the review protocols and likely mirrored the independent data extraction that ideally should happen in practice.

The experimental setting also had limitations. Single data extraction produces more errors than double data extraction.[5] In real life, some of the errors we made would therefore probably have been detected before the data were used for meta-analyses, as it is recommended for Cochrane reviews that there should be at least two independent observers and that any disagreement should be resolved by discussion and, if necessary, arbitration by a third person.[1] We did not perform a consensus step, as the purpose of our study was to explore how much variation would occur when data extraction was performed by different observers. However, given the amount of multiplicity in the trial reports and the uncertainties in the protocols, it is likely that even pairs of observers would disagree considerably with other pairs.

Other limitations were that the observers were under time pressure, although only one person needed more time, as he fell ill during the assigned week. The observers were presented with protocols they had not developed themselves, based on research questions they had not asked, and in disease areas where they were mostly not experts. Another limitation is that, even though one of the exclusion criteria was that the authors of the Cochrane review had not obtained unpublished data from the trial authors, it became apparent during data extraction that some of the trial reports did not contain the data needed for the calculation of an SMD. It would therefore have been helpful to contact trial authors.

Other similar research

The SMD is intended to give clinicians and policymakers the most reliable summary of the available trial evidence when the outcomes have been measured on different continuous or numeric rating scales. Surprisingly, the method has not previously been examined in any detail for its own reliability. Previous research has been sparse and has focused on errors in data extraction.[2 4 5] In one study, the authors found errors in 20 of 34 Cochrane reviews, but, as they gave no numerical data, it is not possible to judge how often these were important.[4] In a previous study of 27 meta-analyses, of which 16 were Cochrane reviews,[2] we could not replicate the SMD result for at least one of the two trials we selected for checking from each meta-analysis within our cut point of 0.1 in 10 of the meta-analyses. When we tried to replicate these 10 meta-analyses, including all the trials, we found that seven of them were erroneous; one was subsequently retracted, and in two a significant difference disappeared or appeared.[2] The present study adds to the previous research by also highlighting the importance of different choices when selecting outcomes for meta-analysis. The results of our study apply more broadly than to meta-analyses using the SMD, as many of the reasons for disagreement were not related to the SMD method but would be important also when analysing data using the weighted mean difference method, which is the method of choice when the outcome data have been measured on the same scale.

Conclusions

Disagreements were common and often larger than the effect of commonly used treatments. Meta-analyses using SMDs are prone to observer variation and should be interpreted with caution. The reliability of meta-analyses might be improved by having more detailed review protocols, more than one observer, and statistical expertise. Review protocols should be more detailed and made permanently available, also after the review is published, to allow other researchers to check that the review was done according to the protocol. In February 2008, the Cochrane Collaboration updated its guidelines and recommended that researchers in their protocols list possible ways of measuring the outcomes—such as using different scales or time points—and specify which ones to use. Our study provides strong support for such precautions. Reports of meta-analyses should also follow published guidelines[1 23] to allow for sufficient critical appraisal. Finally the reporting of trials needs to be improved, according to the recommendations in the CONSORT statement,[24] reducing the need for calculations and imputation of missing data.

Contributors: All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: BT, PCG, JPTH, PJ. Acquisition of data: all authors. Analysis and interpretation of data: BT, PCG, JPTH, PJ, EN. Drafting of the manuscript: BT. Critical revision of the manuscript for important intellectual content: all authors. Statistical analysis: BT, PCG, JPTH, EN. Administrative, technical, or material support: BT, PCG. Study guarantor: PCG.

Funding: This study is part of a PhD funded by IMK Charitable Fund and the Nordic Cochrane Centre. The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. The researchers were independent from the funders.

Competing interests: None declared.

Ethical approval: Not required

References

1. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Version 5.0.0. 2008. www.cochrane-handbook.org
2. Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298:430-7.
3. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457-65.
4. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol 2005;58:741-2.
5. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generatedmore errors than double data extraction in systematic reviews. J Clin Epidemiol 2006;59:697-703.
6. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Version 4.2.6. 2006. www.cochrane-handbook.org
7. Gava I, Barbui C, Aguglia E, Carlino D, Churchill R, De VM, et al. Psychological treatments versus treatment as usual for obsessive compulsive disorder (OCD). Cochrane Database Syst Rev 2007;(2):CD005333.

8. Woodford H, Price C. EMG biofeedback for the recovery of motor function after stroke. Cochrane Database Syst Rev 2007;(2):CD004585.
9. Martinez DP, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. Cochrane Database Syst Rev 2007;(1):CD005233.
10. Orlando R, Azzalini L, Orando S, Lirussi F. Bile acids for non-alcoholic fatty liver disease and/or steatohepatitis. Cochrane Database Syst Rev 2007;(1):CD005160.
11. Buckley LA, Pettit T, Adams CE. Supportive therapy for schizophrenia. Cochrane Database Syst Rev 2007;(3):CD004716.
12. Ipser JC, Carey P, Dhansay Y, Fakier N, Seedat S, Stein DJ. Pharmacotherapy augmentation strategies in treatment-resistant anxiety disorders. Cochrane Database Syst Rev 2006;(4):CD005473.
13. Mistiaen P, Poot E. Telephone follow-up, initiated by a hospitalbased health professional, for postdischarge problems in patients discharged from hospital to home. Cochrane Database Syst Rev 2006;(4):CD004510.
14. Afolabi BB, Lesi FE, Merah NA. Regional versus general anaesthesia for caesarean section. Cochrane Database Syst Rev 2006;(4):CD004350.
15. Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents. Cochrane Database Syst Rev 2006;(4):CD005179.
16. Moore M, Little P. Humidified air inhalation for treating croup. Cochrane Database Syst Rev 2006;(3):CD002870.
17. Jones MK, Menzies RG. Danger ideation reduction therapy (DIRT) for obsessive-compulsive washers. A controlled trial. Behav Res Ther 1998;36:959-70.
18. Adams NP,Bestall JC, Lasserson TJ, Jones PW, Cates C. Fluticasone versus placebo for chronic asthma in adults and children. Cochrane Database Syst Rev 2005;(4):CD003135.
19. Durham RC, Guthrie M, Morton RV, Reid DA, Treliving LR, Fowler D, et al. Tayside-Fife clinical trial of cognitive-behavioural therapy for medication-resistant psychotic symptoms. Results to 3-month follow-up. Br J Psychiatry 2003;182:303-11.
20. Kemp R, Kirov G, Everitt B, Hayward P, David A. Randomised controlled trial of compliance therapy. 18-month follow-up. Br J Psychiatry 1998;172:413-9.
21. Healey A, Knapp M, Astin J, Beecham J, Kemp R, Kirov G, et al. Cost effectiveness evaluation of compliance therapy for people with psychosis. Br J Psychiatry 1998;172:420-4.

22. Kemp R, Hayward P, Applewhaite G, Everitt B, David A. Compliance therapy in psychotic patients: randomised controlled trial. BMJ 1996;312:345-9.
23. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.
24. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials 2001. Explore (NY) 2005;1(1):40-5.

Article 6

Multiplicity of data in trial reports creates an important challenge for the reliability of meta-analyses: an empirical study

Britta Tendal,[1)] Eveline Nüesch,[2,3)] Julian P. T. Higgins,[4)] Peter Jüni,[2,3)] Peter C. Gøtzsche[1)]

From [1)]The Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark; [2)]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Switzerland; [3)]CTU Bern, Bern University Hospital, Switzerland; [4)]MRC Biostatistics Unit, Cambridge, United Kingdom

This article has been submitted.

Abstract

Context: Authors performing meta-analyses of clinical trials often face a multiplicity of data in the trial reports. There may be several possible follow-up times, and the same outcome can be measured on different, but similar scales. The challenge of data multiplicity has not yet been examined in relation to meta-analyses.

Objectives: To examine the scope for multiplicity in a sample of meta-analyses using the standardised mean difference (SMD) as an effect measure, and to examine the impact of the multiplicity on the results.

Data source and study selection: We selected all Cochrane reviews published in The Cochrane Library in the issues 3, 2006 to 2, 2007 that presented a result as an SMD. We retrieved the trial reports that corresponded to the first SMD result in each review and retrieved the review protocols. These index SMDs were used to identify a specific outcome for each meta-analysis from its protocol.

Data extraction: Based on the protocols and the index outcome, two observers independently extracted the data necessary to calculate SMDs from the trial reports for any outcome measures or time points compatible with the protocol. Any information on which control groups to select was also used. Based on the extracted data, all possible SMDs were calculated in Monte Carlo simulations.

Results: Nineteen meta-analyses (83 trials) were included. The review protocols in many instances lacked information about which data to choose. Twenty-four (29%) trials reported data on multiple intervention groups, 30 (36%) provided data on multiple time points and 28 (34%) trials reported the index outcome measured on multiple scales. In 18 out of 19 meta-analyses, we found multiplicity of data in trial reports in at least one trial. Pooled SMD results were affected in 17 of 19 (89%) meta-analyses. In 18 meta-analyses including trials with multiple data, the median variability across meta-analyses was a median difference between two randomly selected SMDs within the same meta-analysis of 0.11 standard deviation units (range 0.03 to 0.41).

Conclusions: Multiplicity can impact importantly on meta-analyses. To reduce the risk of bias in reviews, protocols should pre-specify which results are preferred in relation to time points, intervention groups and scales.

Introduction

Meta-analyses of randomized controlled trials are pivotal for making evidence-based decisions. Multiple eligible data in reports of included trials is a challenge to systematic reviewers, but has not yet received much attention. There is often multiplicity of data in trial reports regarding multiple outcomes, multiple time points, multiple treatment groups, and subgroup analyses.[1] The choice of the outcome of interest is generally based on clinical judgement. However, a fundamentally similar outcome can be measured on several different scales and standardization to a common metric is required before the outcome can be combined in the meta-analysis. This is typically achieved by calculating the standardized mean difference (SMD) for each trial, which is the difference in means between the two groups, divided by the pooled standard deviation of the measurements.[2] By this transformation, the outcome becomes dimensionless and the scales become comparable, as the results are expressed in standard deviation units. For example, a meta-analysis addressing the pain as an outcome might include some trials that measured pain on a visual analogue scale and some trials that measured pain on a 20-point numeric rating scale. This possibility of combining outcomes measured on different scales potentially adds a layer of multiplicity, as the outcome of interest may be measured on more than one scale not only across trials but also within the same trial. Multiplicity of data in trial reports might lead to data driven decisions about what data are included in the meta-analysis and hence is a potential threat to the validity of meta-analysis results.

In this study, we empirically assessed the effect of multiple time points, multiple scales and multiple treatment groups on SMD results in a randomly selected sample of Cochrane reviews.

Methods

Material: We selected all new Cochrane reviews, published in The Cochrane Library during one year (Issues 3, 2006 to 2, 2007) that presented a result as an SMD. We retrieved the reports of all randomised trials that contributed to the first SMD result in each review, and retrieved the latest protocols for all reviews (downloaded in June 2007). Reviews were eligible if they reported at least one result as a standardized mean difference (SMD), if the SMD result were based on two to ten randomized controlled trials and if the outcome was included in the review protocol. Reviews were excluded if only subgroup results were presented. The first pooled SMD result in each review that was not based on a subgroup result was selected as our index SMD result. The index SMD result had to be based on published

data only, i.e. there was no indication in the review that the review authors had received additional outcome data from the trial authors. These index SMD results identified a single outcome for each meta-analysis. Following the published protocol, two observers (BT, EN) independently extracted all possible and reasonable data from the trial reports that could be used to calculate the desired SMD for this outcome. If some required data were unavailable, we used approximations as previously described.[3] Interim analyses were not included. Disagreements were resolved by discussion. We did not contact trial authors for unpublished data.

Data synthesis: For each meta-analysis, we assessed the extent of observed multiplicity by calculating absolute numbers and percentages of trials that reported more than one experimental or control group, more than one time point, and more than one measurement scale that were specified for the outcome of the index SMD result.

We conducted Monte Carlo simulations for each meta-analysis: To estimate the impact of overall multiplicity, in each trial we randomly sampled one SMD and its corresponding standard error from all possible SMDs generated by all multiple reported data to calculate pooled SMDs using fixed- or random-effects models, as originally done in the published reviews. In each meta-analysis we examined the distribution of pooled SMDs across 10,000 simulations using histograms. To estimate the impact of a single source of multiplicity (time points, intervention groups, measurement scales), we allowed only one source of multiplicity to vary at a time when randomly sampling SMDs for each trial. The other sources of multiplicity were standardized at pre-specified standard values (time point: post treatment, scale: first scale mentioned in text, groups: pooled groups). For example in the analysis regarding multiplicity originating from scales, the analysis is based on post treatment values and pooled groups (if there were several possible groups). The values of the different scales for this time point and these groups were then randomly sampled for the calculations of the pooled SMD results. The variability of SMD results due to multiplicity across possible variants of a meta-analysis was expressed as the empirical standard deviation of the distributions of pooled SMDs results obtained from the Monte Carlo simulations. Meta-analyses only including trials without multiple data did not contribute to these analyses.

Results

Figure 1 shows the flowchart for the selection of meta-analyses. Of 32 potentially eligible systematic reviews, we excluded 8 because no pooled SMD index result could be selected, 2 because all SMD results were based on unpublished data, 1 because only subgroup results were reported, 1 because no protocol was available and 1 because the SMD result was not described in the protocol. The 19 eligible meta-analyses included 83 trials that contributed to the study. [4-22] Table 1 shows characteristics of included systematic reviews. 8 systematic reviews addressed an intervention for a psychiatric condition, 2 an intervention for a musculoskeletal condition, 2 an intervention for a neurological condition, 1 an intervention for a gynaecologic, hepatologic and respiratory condition, respectively, and 4 interventions for other conditions. Psychological interventions were studied in 10 meta-analyses, pharmacological interventions in 4, physical interventions in 3, pharmacological interventions, and other interventions in 2 meta-analyses (exercise and humidified air). The outcomes analyzed in the 19 meta-analyses were diverse: in 3 meta-analyses the index outcome was pain, in 13 the index outcome was a symptom severity scale and for 3 meta-analyses, other index outcomes were selected.

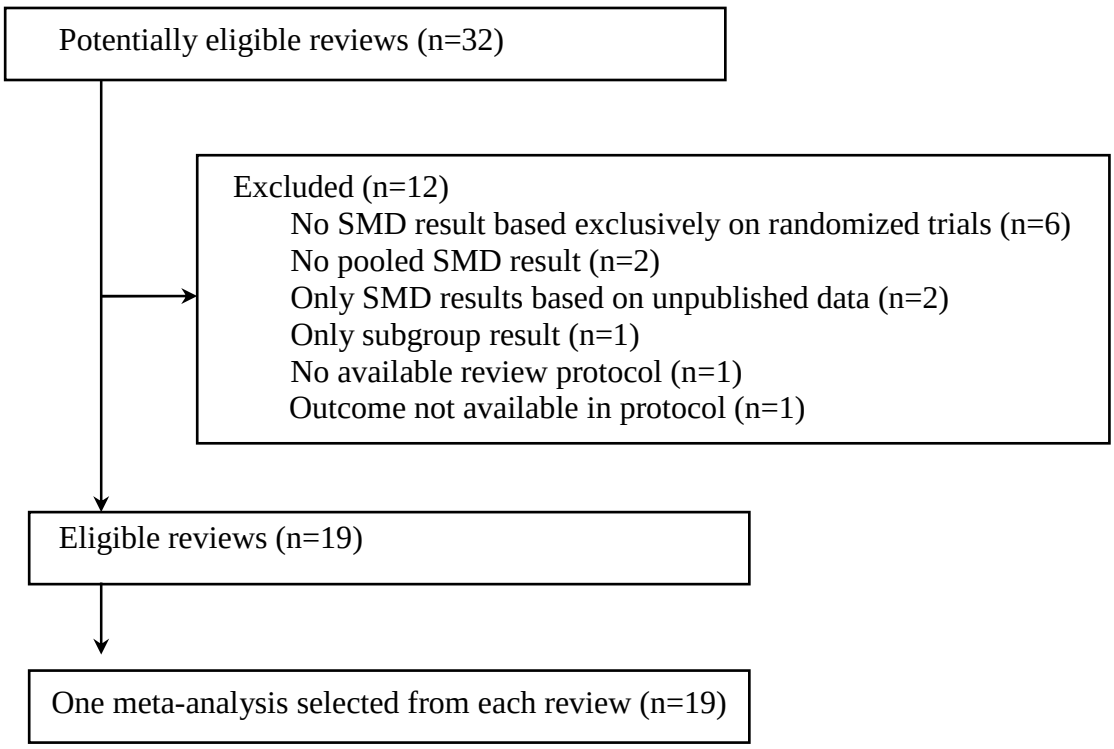

Figure 1 Flowchart for selection of meta-analyses.

Table 1 Characteristics of included systematic reviews

Author	Outcome	Disease	Intervention	Group
Yousefi-Nooraie	Low-back-related disability	Low-back pain	Low level laser therapy	Cochrane Back Group
Ahmad	Pain	Hysterosalpingography (tubal patency)	Analgesic	Cochrane Menstrual Disorders and Subfertility Group
Gava	Symptom level	Obsessive compulsive disorder	Psychological treatment	Cochrane Depression, Anxiety and Neurosis Group
Woodford	Range of movement	Stroke	EMG biofeedback	Cochrane Stroke Group
Martinez	Subjective tinnitus loudness	Tinitus	Cognitive behavioural therapy	Cochrane Ear, Nose and Throat Disorders Group
Orlando	Radiological response	Non-alcoholic fatty liver disease	Bile acids	Cochrane Hepato-Biliary Group
Furukawa	Global judgement	Panic disorders	Combined treatment Psychotherapy and antidepressant	Cochrane Depression, Anxiety and Neurosis Group
Hunot	Worry/fear symptoms	Generalised anxiety disorder	Psychological therapies	Cochrane Depression, Anxiety and Neurosis Group
Buckley	General functioning score	Schizophrenia	Supportive therapy	Cochrane Schizophrenia Group
Ipser	Symptom severity scales	Treatment-resistant anxiety disorders	Pharmacotherapeutic augmentation	Cochrane Depression, Anxiety and Neurosis Group
O'Kearney	Depression	Obsessive compulsive disorder	Behavioural/cognitive-behavioural therapy	Cochrane Depression, Anxiety and Neurosis Group
Mistaen	Patient knowledge regarding disease or symptom management	Postdischarge problem	Telephone follow-up	Cochrane Consumers and Communication Group
Abbass	Anxiety/depression	Common mental disorders	Psychotherapy	Cochrane Depression, Anxiety and Neurosis Group
Afolabi	Neonatal neurological and adaptive score	Caesarean section	Epidural	Cochrane Pregnancy and Childbirth Group
Uman	Pain	Needle-related procedural pain and distress	Psychological interventions	Cochrane Pain, Palliative and Supportive Care Group
Larun	Anxiety	Anxiety	Exercise	Cochrane Depression, Anxiety and Neurosis Group
Trinh	Pain	Neck disorder	Acupuncture	Cochrane Back Group
Moore	Symptom severity or symptom score	Croup	Humidified air	Cochrane Acute Respiratory Infections Group
Mytton	School responses	Agression/violence	Violence prevention program	Cochrane Injuries Group

Information in the review protocols: The level of information in the review protocols is given in Table 2. None of the review protocols contained information on which scales should be preferred. Eight protocols gave information about which time point to select. One gave enough information regarding time point to fully avoid multiplicity, as the outcome was post treatment. A typical statement leaving much room for data-driven decisions regarding the selection of a time point was: "All outcomes were reported for the short term (up to 12 weeks), medium term (13 to 26 weeks), and long term (more than 26 weeks)".[7] Another example was a review regarding humidified air for treating croup, [15] which stated, "The outcomes will be separately recorded for the week following treatment." The selected outcome was croup symptom score and the three included trials had time points from 20 min to 12 hours to choose between. In such a case the protocol does not help. Eighteen protocols described which type of control group to select but none reported any hierarchy among similar control groups or any intentions to combine such groups.

	Yousefi-Nooraie et al.	Ahmad et al.	Gava et al.	Woodford and Price	Martinez et al.	Orlando et al.	Furukawa et al.	Hunot et al.	Buckley and Pettit	Ipser et al.	O'Kearney et al.	Mistiaen and Poot	Abbass et al.	Afolabi et al.	Uman et al.	Larun et al.	Trinh et al.	Moore and Little	Mytton et al.
Possible control group(s)	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hierarchy of control groups														✓*				✓*	
Which time point/period to select		✓					✓	✓	✓			✓	✓					✓	✓
Time point *precisely* defined								✓											
Hierarchy of measuring methods or scales																			

Table 2 Content of review protocols

Observed multiplicity in trial reports: Table 3 presents the extent of multiplicity observed in the 19 reviews including 83 trials. Across all reviews 55 (66%) trials had multiple data from one or more of the three sources. Twenty-four (29%) trials reported data on more than one intervention or more than one control group, 30 (36%) trials provided data on more than one eligible time point and 28 (34%) trials reported the index outcome using more than one eligible measurement scale. In 11 of 19 (58%) meta-analyses, we found at least one trial that provided data on more than one intervention or more than one control group. 13 (68%) meta-

analyses included at least one trial that reported more than one eligible time point and 11 (58%) meta-analyses at least one trial that reported the index outcome using more than one eligible measurement scale. We found one meta-analysis, where all 3 included trials did only report data of one intervention and control group, one eligible time point and one measurement scale for the index outcome.[18]

		No trials with multiplicity regarding:			
	No trials included	**Any source**	**Intervention groups**	**Time points**	**Measurement scales**
Yousefi-Nooraie et al.	3	2 (67%)	0 (0%)	1 (33%)	1 (33%)
Ahmad et al.	5	3 (60%)	1 (20%)	2 (40%)	1 (20%)
Gava et al.	7	6 (86%)	4 (57%)	3 (43%)	5 (71%)
Woodford & Price	5	4 (80%)	2 (40%)	2 (40%)	3 (60%)
Martinez Devesa et al.	4	4 (100%)	3 (75%)	3 (75%)	1 (25%)
Orlando et al.	3	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Furukawa et al.	7	6 (86%)	6 (86%)	2 (29%)	4 (57%)
Hunot et al.	9	5 (56%)	2 (22%)	0 (0%)	5 (56%)
Buckley & Pettit	2	2 (100%)	0 (0%)	2 (100%)	0 (0%)
Ipser et al.	7	6 (86%)	0 (0%)	5 (71%)	3 (43%)
O'Kearney et al.	2	1 (50%)	1 (50%)	0 (0%)	0 (0%)
Mistaen & Poot	3	1 (33%)	1 (33%)	0 (0%)	0 (0%)
Abbass et al.	2	2 (100%)	0 (0%)	1 (50%)	2 (100%)
Afolabi et al.	2	1 (50%)	0 (0%)	1 (50%)	0 (0%)
Uman et al.	9	2 (22%)	2 (22%)	0 (0%)	0 (0%)
Larun et al.	5[§]	4 (80%)	1 (20%)	3 (60%)	2 (40%)
Trinh et al.	3	3 (100%)	0 (0%)	3 (100%)	1 (33%)
Moore & Little	3	2 (67%)	0 (0%)	2 (67%)	0 (0%)
Mytton et al.	2	1 (50%)	1 (50%)	0 (0%)	0 (0%)
All included reviews	83	55 (66%)	24 (29%)	30 (36%)	28 (34%)

Table 3 Observed multiplicity in the meta-analyses. §One trial from Larun et al. were excluded because lack of data in trial reports.

Effects of multiplicity on results of meta-analyses: Figure 2 presents distributions of possible pooled SMDs in each meta-analysis, when randomly selecting one possible SMD result per trial. The dots below the distributions indicate how many trials were included in the meta-analyses, open dots are trials without multiplicity, and filled dots are trials with multiplicity. We found that pooled SMD results were affected by any type of multiplicity of data in the included trials in 17 of 19 (89%) meta-analyses, in 1 meta-analysis we did not find multiple data in the trial reports[18] and in 1 meta-analysis the observed multiplicity had no effect on the pooled SMD results.[7] In all 11 (58%) meta-analyses including at least one trial with more than one experimental or control group, we found variability in the pooled SMD results due to this type of multiplicity. In 12 (63%) meta-analyses there was variability in the pooled SMD results due to multiplicity of data regarding time points (Figure 2, 3rd column). In one meta-analysis with two trials that reported more than one eligible time point, we did not find multiple possible pooled SMDs due to these different time points.[7] In 9 (47%) meta-analyses we found variability in pooled SMD results from trial data of multiple measurement scales used for the index outcome. In two meta-analyses, one trial in each meta-analysis reported data on more than one measurement scale for the index outcome, but this multiplicity did not affect the pooled SMD results.[6, 22]

Table 4 presents the variability of pooled SMD results according to different sources of multiplicity. We found that in 18 meta-analyses including trials with multiple data reported for any of the three sources evaluated. The median standard deviation was 0.11 (range 0.03 to 0.41), which corresponds to a median difference between two randomly selected SMDs within the same meta-analysis. The median difference across the 11 meta-analyses that included trials with multiple data regarding intervention groups was 0.05 standard deviation units (range 0.01 to 0.23) between two randomly selected SMDs calculated from different eligible intervention groups. The median standard deviation across 13 meta-analyses that included trials with data on multiple eligible time points was 0.06 (range 0.02 t 0.41) and across 11 meta-analyses including trials that provided data of multiple measurement scale for the index outcome was 0.09 (range 0.01 to 0.15).

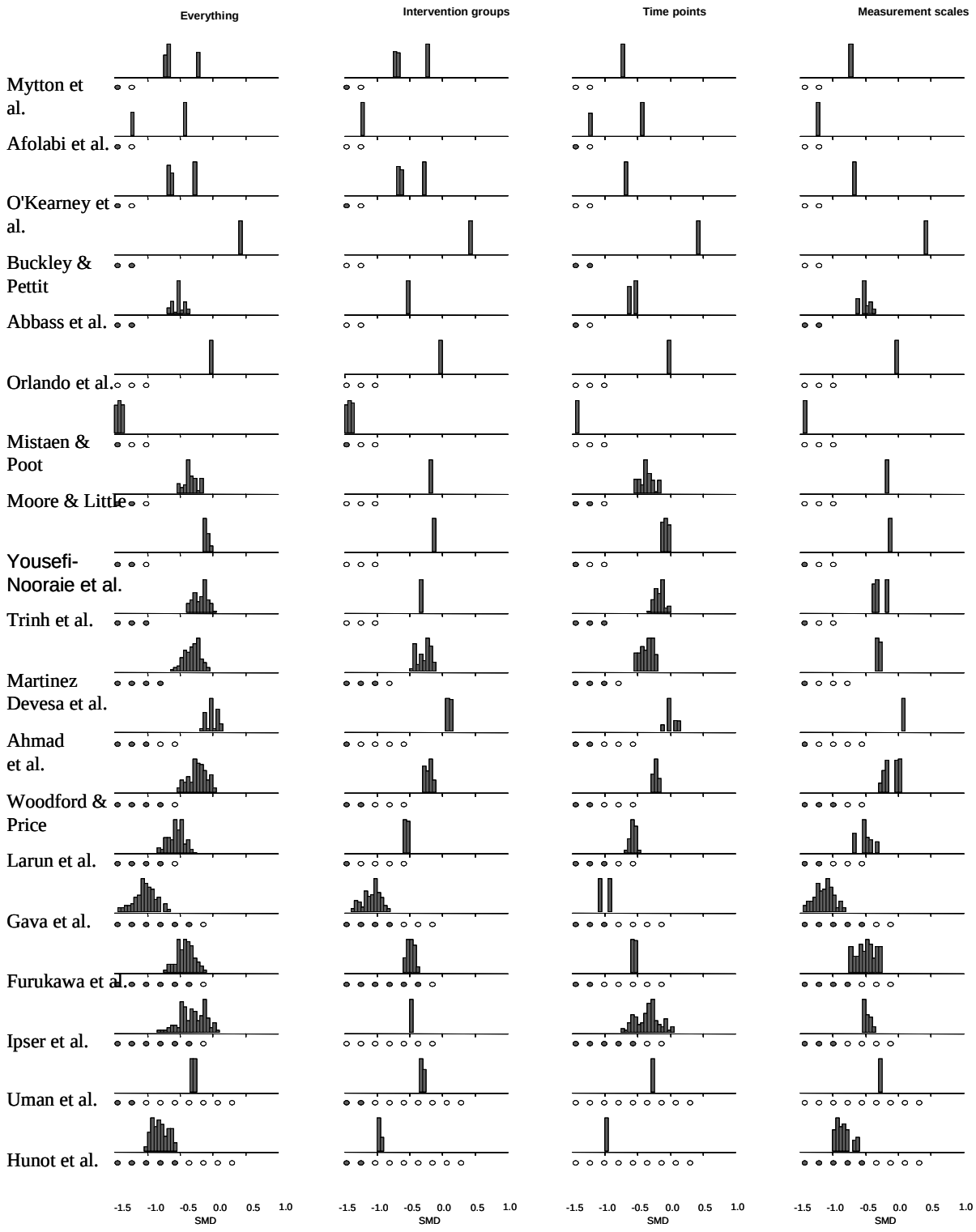

Figure 2 Monte Carlo distributions of possible pooled SMDs in each meta-analysis. The dots below the distributions indicate how many trials were included in the meta-analyses, open dots are trials without multiplicity, and filled dots are trials with multiplicity.

Source of multiplicity	Number of meta-analyses with multiple data	Variability in SMD results across meta-analyses (standard deviation [range])
Any source	18 of 19 (95%)	0.11 (0.03 to 0.41)
Intervention groups	11 of 19 (58%)	0.05 (0.01 to 0.23)
Time points	13 of 19 (68%)	0.06 (0.02 to 0.41)
Measurement scales	11 of 19 (58%)	0.09 (0.01 to 0.15)

Table 4 Variability in meta-analyses results

Comment

In 17 out of 19 meta-analyses included in our study, we found multiplicity of data in trial reports in at least one trial, which frequently resulted in substantial variabilities of pooled SMD results. The magnitude of impact of multiple data in trial reports regarding intervention groups, time points or measurement scales on meta-analyses results varied considerably across meta-analyses ranging from essentially no impact to an impact of multiple data corresponding to a small to moderate treatment benefit (a standard deviation of 0.2 across possible meta-analysis results). In our study we were able to estimate the impact of individual sources of multiple data in trial reports on the meta-analyses results, enabling us to judge whether the three different sources individually impacted on the pooled SMD results.

We randomly selected Cochrane reviews in our study and therefore, included a broad selection of interventions and outcomes that were expressed as SMDs. The variability of pooled SMD results due to multiple trial data did not seem to be particular for certain types of interventions or outcomes, although it varied substantially across meta-analyses. To estimate the impact of multiplicity on meta-analyses results, we randomly selected one SMD per trial from a pool of eligible SMDs that were calculated from multiple data in trial reports with equal probability. This process explores the magnitude of what is possible due to multiple reported data. However, there might be implicit rules regarding data-extraction within specialties. For example reviewers might find that one scale is more commonly used, e.g. Hamilton, than another and therefore select this scale if possible. This unwritten hierarchy of scales naturally reduces the perceived multiplicity, but should be made explicit to enhance transparency.

We relied on published information in the trial reports. Our results are transparent as we limited ourselves to published results. However, selective reporting of outcomes in trials [23-26] might have distorted our results: Positive, statistically significant results are more likely to be published than non-significant results.[27] In the presence of publication bias we might have underestimated the overall multiplicity. Our study is only able to provide an estimate of multiplicity among published results. Effective multiplicity might be even higher as published and unpublished results are likely to be different from each other. Our study was only possible because Cochrane Reviews are required to publish their protocols. We believe that for most meta-analyses published outside the Cochrane Library, no protocol is available[28] and the choice of multiple data possibly extracted is even larger than we observed in our study.

We examined three frequent sources of multiplicity of data in trial reports: time points, intervention groups and measurement scales. However, there are other types of multiple data in trial reports. For example, results might be reported from different types of analyses: intention-to-treat analyses might be reported alongside with per-protocol analyses. We were unable to explore this issue, because only few included trials provided results from more than one analysis. For each meta-analysis we specified an index outcome and could therefore not examine the impact of the selection of different outcomes for the reliability of meta-analyses results.

Our study provides an estimation of the extent and impact of multiplicity of data in trial reports on the results of meta-analyses. To our knowledge, our study is the first to show empirically that reliability of meta-analyses results might be compromised due to multiple data on time points, measurement scales and intervention groups provided in trial reports. We have previously reported results from an observer agreement study performed on a sample of the meta-analyses included in this study.[2] We found that disagreements among observers were common and often large, the main reasons for disagreement being: different choices (groups, time points, scales and calculations) whether to include certain trials and data extraction errors.[2] A recent paper by Bender et al. describes the problem of multiple comparisons in systematic reviews.[1] The authors identified common reasons for multiplicity in reviews, but did not estimate the impact on the meta-analytic results.[1]

Multiplicity due to selection of time points and groups is not unique to SMD; future research could therefore be done into whether multiplicity also is an issue for effect measures like the mean difference, for which the outcomes have to be measured on the same scale, or binary outcomes.

The extent of multiplicity of data found in trial reports reflects the information provided in the review protocols: A badly specified outcome in the review protocol will have led to a larger extent of observed multiplicity for this outcome than a precisely specified outcome in the review protocol, if an equal amount of data is found in the reports of included trials. Some might argue that data extraction for a meta-analysis is dependent on what is reported in trials and cannot be entirely specified in advance without knowledge of the included trials. However, systematic reviewers are usually not completely unaware of the potentially included trials at the protocol writing stage. In addition, we argue that to minimise data-driven selection of time points, measurement scales or intervention groups included in the meta-analyses, researchers should specify these decisions at protocol stage. If amendments to the protocol are indicated, these should be transparently reported.[29, 30] Whether more detailed protocols increase the reliability of meta-analyses results remains to be shown.

Implications: This study demonstrates that multiplicity is a real problem, two solutions come to mind: one might be to report and analyse everything another to make the protocols for systematic reviews more detailed. The first solution presents two large challenges, first how to interpret the results? If for example one scale in a trial shows a positive effect of an intervention and another scale in the same trial shows a negative effect. The other challenge is that this approach would involve multiple testing of the same outcome, as there would be multiple comparisons involving the same outcome. A possible way of dealing with observed multiplicity could be a multivariate meta-analysis, accounting for correlations among outcomes, time points and comparisons. The second solution regarding more detailed protocols would imply that reviewers specified which time points, scales and groups to consider and presented a hierarchy for scales and groups. It is however difficult to foresee everything in a protocol; this makes the obligation of the reporting to be clear of the systematic so much greater. The reporting needs to allow the reader to follow the process leading to the results; this also includes descriptions of choices made during the data extraction process. It is not possible to report everything, so what really matters and what issues are less important? Our study shows that time points, scales and groups have an impact on the results and therefore are important to report.

As for randomised trials, systematic reviews should have a detailed protocol. Only Cochrane reviews are required to have a published protocol. A descriptive study performed by Moher et al showed that only around ten percent of non-Cochrane reviews stated working from a protocol. [30] As pointed out in the PRISMA statement, protocol amendments should not necessarily be considered inappropriate but should definitely be acknowledged as such and

published. [29, 30] Our study suggests that protocol amendments are likely to produce differences in results und thus, protocol amendments should be discouraged unless clearly justified.

Conclusions: Variability in meta-analyses results is substantial due to multiplicity in trial reports paired with protocols lacking details defining what time points, scales and treatment groups ought to be included. Reviews are study designs in their own right and reviewers should anticipate multiplicity of data in trial reports and take this into account when writing protocols. To enhance reliability of meta-analyses results, we suggest that protocols should clearly define time points to be extracted, give a hierarchy of scales and clearly define eligible treatment and control groups and give strategies for handling multiple groups. Clinical judgment will be important to define at protocol stage, which time points and scales to be included. Ideally, the choice of time points and scales should be evidence based, but empirical evidence for the most interesting time points and a hierarchy of scales according to good validity and responsiveness are rarely available.

Author Contributions: All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Tendal, Nüesch, Gøtzsche
Acquisition of data: Tendal, Nüesch
Analysis and interpretation of data: Higgins, Nüesch, Tendal
Drafting of the manuscript: Tendal, Nüesch
Critical revision of the manuscript for important intellectual content: All authors
Administrative, technical, or material support: Gøtzsche
Study supervision: Gøtzsche
Financial Disclosures: None reported.

Funding/Support: This study is part of a PhD funded by IMK Charitable Fund.

Role of the Sponsors: The funding organizations played no role in the study design and conduct of the study, in the data collection, management, analysis, and interpretation of the data, or in the preparation, review, or approval of the manuscript.

References

1. Bender R, Bunce C, Clarke M, et al. Attention should be given to multiplicity issues in systematic reviews. *J Clin Epidemiol.* Sep 2008;61(9):857-865.
2. Tendal B, Higgins JP, Juni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ.* 2009;339:b3128.
3. Reichenbach S, Sterchi R, Scherer M, et al. Meta-analysis: chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med.* Apr 17 2007;146(8):580-590.
4. Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database Syst Rev.* 2006(4):CD004687.
5. Afolabi BB, Lesi FE, Merah NA. Regional versus general anaesthesia for caesarean section. *Cochrane Database Syst Rev.* 2006(4):CD004350.
6. Ahmad G, Duffy J, Watson AJ. Pain relief in hysterosalpingography. *Cochrane Database Syst Rev.* 2007(2):CD006106.
7. Buckley LA, Pettit T, Adams CE. Supportive therapy for schizophrenia. *Cochrane Database Syst Rev.* 2007(3):CD004716.
8. Furukawa TA, Watanabe N, Churchill R. Combined psychotherapy plus antidepressants for panic disorder with or without agoraphobia. *Cochrane Database Syst Rev.* 2007(1):CD004364.
9. Gava I, Barbui C, Aguglia E, et al. Psychological treatments versus treatment as usual for obsessive compulsive disorder (OCD). *Cochrane Database Syst Rev.* 2007(2):CD005333.
10. Hunot V, Churchill R, Silva de Lima M, Teixeira V. Psychological therapies for generalised anxiety disorder. *Cochrane Database Syst Rev.* 2007(1):CD001848.
11. Ipser JC, Carey P, Dhansay Y, Fakier N, Seedat S, Stein DJ. Pharmacotherapy augmentation strategies in treatment-resistant anxiety disorders. *Cochrane Database Syst Rev.* 2006(4):CD005473.
12. Larun L, Nordheim LV, Ekeland E, Hagen KB, Heian F. Exercise in prevention and treatment of anxiety and depression among children and young people. *Cochrane Database Syst Rev.* 2006;3:CD004691.

13. Martinez Devesa P, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. *Cochrane Database Syst Rev.* 2007(1):CD005233.

14. Mistiaen P, Poot E. Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. *Cochrane Database Syst Rev.* 2006(4):CD004510.

15. Moore M, Little P. Humidified air inhalation for treating croup. *Cochrane Database Syst Rev.* 2006;3:CD002870.

16. Mytton J, DiGuiseppi C, Gough D, Taylor R, Logan S. School-based secondary prevention programmes for preventing violence. *Cochrane Database Syst Rev.* 2006;3:CD004606.

17. O'Kearney RT, Anstey KJ, von Sanden C. Behavioural and cognitive behavioural therapy for obsessive compulsive disorder in children and adolescents. *Cochrane Database Syst Rev.* 2006(4):CD004856.

18. Orlando R, Azzalini L, Orando S, Lirussi F. Bile acids for non-alcoholic fatty liver disease and/or steatohepatitis. *Cochrane Database Syst Rev.* 2007(1):CD005160.

19. Trinh KV, Graham N, Gross AR, et al. Acupuncture for neck disorders. *Cochrane Database Syst Rev.* 2006;3:CD004870.

20. Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents. *Cochrane Database Syst Rev.* 2006(4):CD005179.

21. Woodford H, Price C. EMG biofeedback for the recovery of motor function after stroke. *Cochrane Database Syst Rev.* 2007(2):CD004585.

22. Yousefi-Nooraie R, Schonstein E, Heidari K, et al. Low level laser therapy for nonspecific low-back pain. *Cochrane Database Syst Rev.* 2007(2):CD005107.

23. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ.* Apr 2 2005;330(7494):753.

24. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA.* May 26 2004;291(20):2457-2465.

25. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ.* Sep 28 2004;171(7):735-740.

26. Vedula SS, Bero L, Scherer RW, Dickersin K. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *N Engl J Med.* Nov 12 2009;361(20):1963-1971.

27. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev.* 2009(1):MR000006.

28. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med.* Mar 27 2007;4(3):e78.

29. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ.* 2009;339:b2700.

30. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* 2009;339:b2535.

Article 7

Which meta-analyses are conclusive?

Eveline Nüesch[1,2)] and Peter Jüni[1,2)]

From [1)]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Switzerland; [2)]CTU Bern, Bern University Hospital, Switzerland

Int J Epidemiol. 2009 Feb;38(1):298-303

Abstract

Objective To examine how different methodological approaches such as funnel plots, stratified analyses accompanied by interaction tests and heterogeneity-adjusted trial sequential analysis contribute to our understanding of bias and inconclusive results in meta-analyses.

Methods We re-analysed the trials of intravenous magnesium in acute myocardial infarction using funnel plots accompanied by tests for asymmetry, analyses stratified for allocation concealment and sample size, and heterogeneity-adjusted trial sequential analysis.

Results Visual inspection of funnel plots and regression lines suggested asymmetry at all stages of the meta-analysis, but tests for funnel plot asymmetry became statistically significant only after the inclusion of LIMIT-2, the only adequately sized trial at that time. Differences in pooled effects between trials with and without allocation concealment and between large and small trials were apparent, but interaction tests for allocation concealment were positive only in fixed-effect meta-analyses. In heterogeneity-adjusted trial sequential analysis, the z-curve didn't cross the boundary before ISIS-4, a large scale trial in 58'050 patients, became available and the necessary information size of nearly 25 000 patients was reached, suggesting that the results of both, random- and fixed-effect meta-analyses were inconclusive.

Conclusions Funnel plots with statistical tests of asymmetry, stratified analyses accompanied by tests of interaction and heterogeneity-adjusted trial sequential analyses will all contribute to our understanding of which meta-analyses can be considered conclusive.

In 1991, a meta-analysis of 7 small-scale trials of intravenous magnesium in a total of 1266 patients with suspected acute myocardial infarction indicated a more than 50% reduction in the risk of death associated with magnesium (relative risk 0.48, 95% CI 0.26 to 0.88).[1] Yusuf et al updated this meta-analysis in 1993[2] to include LIMIT-2,[3] at the time the only adequately sized trial, with a power of 80% to detect a moderate to large relative reduction in the risk of death of 33% associated with magnesium. Based on a total of 8 trials in 3617 patients with a pooled relative risk of 0.59 (95% CI 0.38 to 0.91), the authors concluded that "intravenous magnesium is a safe, effective, widely practicable, and inexpensive intervention that has the potential of making an important impact on the management of patients with myocardial infarction".[2] In 1995, ISIS-4 became available,[4] a large scale trial in 58,050 patients, which had nearly 95% power to detect a small, but potentially clinically relevant reduction in the relative risk of death of 10% associated with magnesium. ISIS-4 clearly refuted the earlier meta-analyses and showed a trend towards more deaths in the patients allocated to magnesium, with the lower limit of the 95% confidence interval excluding any relevant benefit of the intervention (relative risk 1.05, 95% CI 0.99 to 1.12).

The case of magnesium in acute myocardial infarction cast serious doubts on the trustworthiness of meta-analyses. Which meta-analyses were conclusive and which were likely to be refuted by subsequent large-scale trials? Intrigued by the magnesium example, Egger and Davey Smith[5] suggested in 1995 that funnel plots could have been used as a diagnostic tool, in which estimates of treatment effect obtained in trials included in the magnesium meta-analyses[1 2] are plotted against a measure of sample size or statistical precision, to detect bias associated with small trials. In the absence of bias, the plot will typically resemble a symmetrical inverted funnel with the results of smaller trials more widely scattered than those of larger, more precise trials. Publication bias,[6] and poor design, execution and analysis of small trials[7] may result in skewed funnel plots. Visual inspection of the funnel plot of magnesium trials and a formal statistical test of its asymmetry indicated that the funnel plot was clearly asymmetrical before ISIS-4 became available.[1 2]

In 1997, Pogue and Yusuf[8 9] took a different approach and suggested that multiple looks in meta-analyses of randomised trials may be interpreted similarly to interim looks in a single trial. The problem of interim looks in a single trial was originally addressed by Armitage[10] and Pocock[11] by group sequential analysis. Lan and DeMets[12] extended the suggested concept with an alpha-spending function to allow flexible unplanned monitoring in a trial. They introduced the cumulative z-curve modelled as a Brownian motion and an alpha-spending

function according to O'Brien and Flemming[13] for the construction of monitoring boundaries. If a treatment effect larger than expected occurs, a trial should be terminated early when the cumulative z-curve for this treatment effect crossed the constructed sequential monitoring boundary. In early stages of a trial when data are sparse only very extreme results corresponding to extreme z-values are accepted to indicate premature termination of a trial. The monitoring boundaries become less stringent as more data accumulate and the planned sample size of the trial is approached. The same principle could be applied to meta-analyses to determine when a meta-analysis is conclusive. Only extreme results leading to z-values that cross highly stringent boundaries should be accepted if little information was accrued in a meta-analysis of few, small scale trials. Boundaries should become less stringent as more information accumulates.[8 9] In a cumulative meta-analysis of ten magnesium trials, Pogue and Yusuf found that the cumulative z-curve of the meta-analysis did not cross the specified monitoring boundary for overall mortality and suggested that the meta-analysis was not conclusive.[8] However, Egger et al identified 15 trials of magnesium in myocardial infarction published before ISIS-4.[4] When based on all 15 trials, rather than the ten trials selected by Pogue and Yusuf, the meta-analysis crossed the monitoring boundary and became conclusive, although the results were still contradicted by ISIS-4.[14] The approach failed to become widely adopted.

Recently, Wetterslev et al coined the term "trial sequential analysis" for an extension of Pogue and Yusuf's approach, which reflects an increase in uncertainty if heterogeneity between trials is present in a meta-analysis.[15] In this issue, two articles by the same group use trial sequential analysis to determine whether results of published meta-analyses in neonatology[16] and across different fields[17] are conclusive. Using trial sequential analyses, which account for the observed heterogeneity between trials, they find a substantial proportion of published meta-analyses potentially inconclusive. In both articles,[16 17] the authors point out that trial sequential analysis does not deal with systematic errors resulting from the inclusion of flawed trials[18] and outcome reporting[19] or publication biases[20] and that these sources of systematic errors should be appropriately examined using funnel plots[21] and analyses stratified according to methodological characteristics of trials accompanied by appropriate tests for interaction between trial characteristic and effect estimates.[22]

Here, we re-analyse the trials of intravenous magnesium in acute myocardial infarction to determine how the different diagnostic measures – funnel plots, stratified analyses according to methodological characteristics of trials and heterogeneity-adjusted trial sequential analysis – contribute to our understanding of bias and inconclusive results at four stages of the meta-

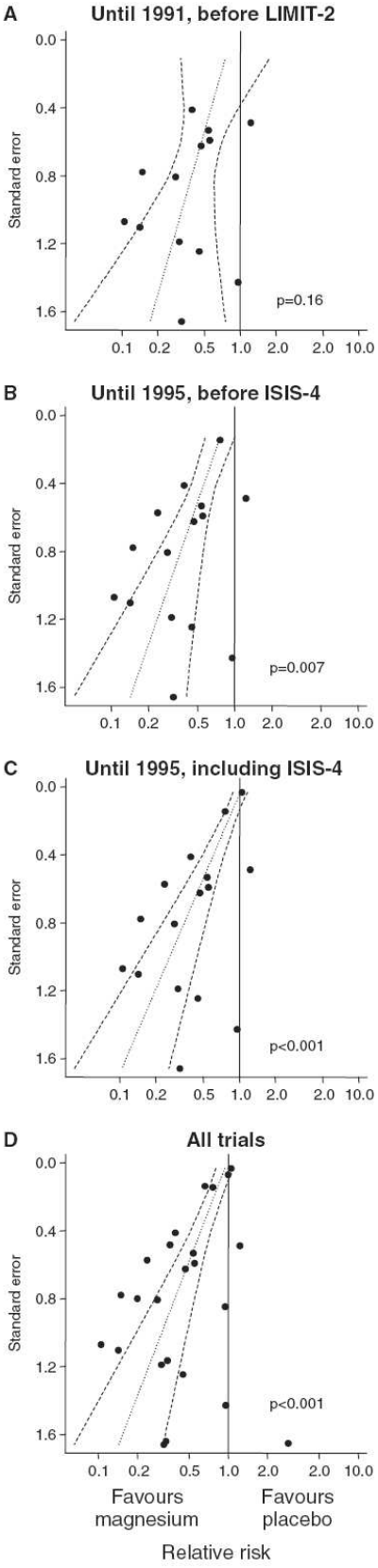

Figure 1 Funnel plots. Funnel plots are presented (A) for trials published until 1991, before LIMIT-2 became available; (B) until 1995, before ISIS-4 became available; (C) until 1995, including ISIS-4; and (D) up to 2004. Dotted lines indicate predicted treatment effects (regression line) from univariable meta-regression by using standard error as explanatory variable; dashed lines represent 95% CI. Regression lines are truncated at standard errors typically found in adequately sized trials with sufficient power to detect a moderate to large relative risk reduction of 30–40% (stages A and B) and at the standard error found in the largest trial included in the meta-analysis (stages C and D). P-values are derived from Egger's test for funnel plot asymmetry.

analysis: (A) trials available until 1991, before LIMIT-2,[3] (B) trials until 1995, before ISIS-4[4] became available, (C) all trials until 1995, including ISIS-4[4] and (D) all trials available to date.[14,23] Figure 1 presents funnel plots of effect sizes on the horizontal axis against their

standard errors on the vertical axis, displaying asymmetry as regression lines with 95% confidence bands derived from predicting the treatment effect from univariable meta-regression analysis with the standard error as the explanatory variable.[21] Visual inspection of funnel plot and regression line suggest asymmetry at all four stages A to D of the meta-analysis, but Egger's test for funnel plot asymmetry[23] becomes positive only at stage B, after the inclusion of LIMIT-2,[3] the only adequately sized trial at that time. In subsequent stages, the shape of the funnel plot remains essentially unchanged and Egger's test for asymmetry positive, suggesting bias.

			Random-effects meta-analysis		Fixed-effect meta-analysis		
	Number of trials	Number of patients	Relative risk (95% CI)	*P*-value for interaction	Relative risk (95% CI)	*P*-value for interaction	Heterogeneity I^2 (%)
A. Until 1991, before LIMIT-2[14]							
Overall	13	2028	0.46 (0.32–0.65)		0.42 (0.30–0.59)		0.0
Concealment of allocation				0.99		0.89	
Adequate	2	551	0.45 (0.20–1.02)		0.44 (0.20–0.98)		0.0
Inadequate or unclear	11	1477	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				–		–	
⩾2200 patients	0	0	–		–		–
<2200 patients	13	2028	0.46 (0.32–0.65)		0.42 (0.30–0.59)		0.0
B. Until 1995, before ISIS-4[14]							
Overall	15	4559	0.48 (0.34–0.67)		0.57 (0.47–0.70)		30.6
Concealment of allocation				0.64		0.036	
Adequate	4	2531	0.50 (0.28–0.91)		0.67 (0.52–0.85)		50.0
Inadequate or unclear	11	2028	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				0.094		0.002	
⩾2200 patients	1	2316	0.76 (0.59–0.99)		0.76 (0.59–0.99)		0.0
<2200 patients	14	2243	0.43 (0.31–0.60)		0.39 (0.29–0.54)		0.0
C. Until 1995, including ISIS-4[14]							
Overall	16	62609	0.53 (0.38–0.75)		1.01 (0.95–1.06)		66.8
Concealment of allocation				0.25		<0.001	
Adequate	5	60581	0.69 (0.46–1.03)		1.03 (0.97–1.09)		77.3
Inadequate or unclear	11	2028	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				0.007		<0.001	
⩾2200 patients	2	60366	0.92 (0.67–1.26)		1.04 (0.98–1.10)		82.4
<2200 patients	14	2243	0.43 (0.31–0.60)		0.39 (0.29–0.54)		0.0
D. All trials[14,24]							
Overall	24	72920	0.65 (0.53–0.80)		0.98 (0.93–1.03)		65.8
Concealment of allocation				0.40		<0.001	
Adequate	9	67945	0.80 (0.65–0.98)		1.02 (0.97–1.07)		71.7
Inadequate or unclear	15	4795	0.56 (0.43–0.74)		0.58 (0.47–0.71)		7.6
Sample size				0.001		<0.001	
⩾2200 patients	4	69758	0.89 (0.75–1.06)		1.01 (0.97–1.07)		83.0
<2200 patients	20	2982	0.42 (0.32–0.57)		0.39 (0.30–0.52)		0.0

Table 1 Stratified analyses

Results from stratified analysis according to allocation concealment and sample size are presented using fixed- and random-effects models including trials published until 1991 and before LIMIT-2; until 1995 and before ISIS-4, until 1995 including ISIS-4 and up to 2004. P-values for interaction between treatment effect and trial characteristics were derived using meta-regression for random-effects models and z-tests for fixed-effect models.

Table 1 presents the results from corresponding stratified analyses according to concealment of allocation and sample size. At stage A, stratified analyses using a fixed-effect and a random-effects models indicate no relevant differences between trials with adequate concealment and the remaining trials, whereas no adequately sized trials with sample sizes of $\geq$ 2200 patients were available. At stage B, after LIMIT-2[3] became available, differences become apparent between trials with and without concealment of allocation and between large and small trials, but pooled effects are statistically significant in all stratified analyses and interaction tests are positive only in fixed-effect meta-analyses. With the inclusion of ISIS-4,[4] the between trial heterogeneity becomes prominent. Therefore, random-effects models attribute considerably more weight to smaller studies than fixed-effect models and results from fixed-effect and random-effects meta-analyses including all trials are discordant: there is still a clinically relevant mortality reduction according to the random-effects, but a clear-cut null-result according to the fixed-effect meta-analysis. Even in the presence of high between-trial heterogeneity, random and fixed-effect models show concordant results if stratified according to trial size: no effect in adequately sized trials and an unrealistically large beneficial effect on overall mortality in small trials. Positive tests of interaction in both, random and fixed-effect analyses indicate that these differences between adequately sized and small trials are unlikely to have occurred by chance alone.

Figure 2 presents results from trial sequential analysis using fixed-effect meta-analysis (top) and random-effects meta-analysis (bottom). The dashed horizontal line represents the monitoring boundaries to be reached by the z-value of a meta-analysis to indicate that results are conclusive before the number of 24,899 patients is reached, which is necessary to detect a relative risk reduction of 15% with 80% power at a two-sided α of 0.01. The boundary becomes less stringent with more patients accruing and will converge to a z-value of 2.58 corresponding to the α-level of 0.01 indicating conclusive results when sufficient numbers of patients have been accumulated. Neither in random-effects, nor in fixed-effect meta-analyses, the z-curve crosses the boundary before ISIS-4 becomes available and the necessary information size of nearly 25,000 patients is reached, suggesting that the results of both, random and fixed-effect meta-analyses were inconclusive. After inclusion of ISIS-4,[4] however, results are conflicting: evidence of a null effect according to the fixed-effect model, but evidence of a benefit of magnesium according to the random-effects model, which vanishes only after the analysis is restricted to trials with adequate sample size (data available on request).

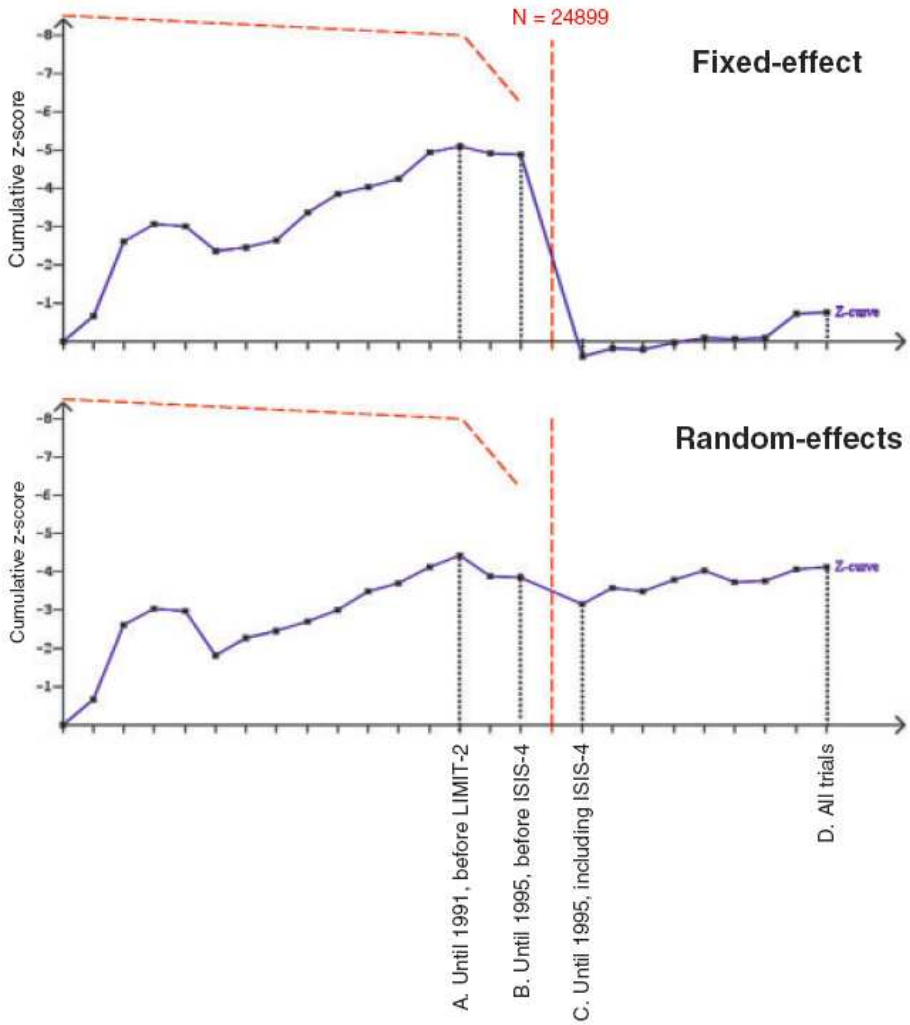

Figure 2 Heterogeneity-adjusted trial sequential analysis

Trial sequential analysis of trials of intravenous magnesium using fixed-effect (top) and random-effects meta-analysis (bottom). The dashed vertical line indicates that the number of patients necessary to detect a relative risk reduction of 15% with 80% power at α=0.01 is 24 899 if a baseline risk of 10% and a heterogeneity between trials of I^2=30% are assumed. The dashed horizontal line represents the monitoring boundaries to be reached by the z-value of a meta-analysis to indicate that results are conclusive before the necessary number of 24 899 patients is reached. The boundary becomes less stringent when more trials and patients are included and will converge to a z-value of 2.58, corresponding to the α-level of 0.01, to indicate conclusive results when sufficient numbers of patients are accumulated.

It is the overall pattern found in funnel plots, stratified analyses, and heterogeneity-adjusted trial sequential analysis, which provides a clear-cut insight into the trustworthiness of the different stages of the meta-analysis of magnesium in acute myocardial infarction.[1 2 14 23] At stage A, formal tests of funnel plot asymmetry and interaction tests accompanying stratified

analyses were still negative due to a lack of power, and some would have concluded that the evidence accumulated was unbiased and trustworthy. Heterogeneity-adjusted trial sequential analysis unequivocally indicates, however, that the evidence was inconclusive at this stage. At stage B, trial sequential analysis suggests that the accumulated evidence is still unconvincing when LIMIT-2[3] was included. In addition, the test for funnel plot asymmetry becomes positive. At stages C and D, after the inclusion of ISIS-4,[4] heterogeneity-adjusted trial sequential analyses of random-effects and fixed effects meta-analyses are discordant. Here, the appropriately powered tests of funnel plot asymmetry and tests of interaction between sample size and treatment effect indicate that the inclusion of trials of inadequate size leads to a severe distortion of results.

Egger and Davey Smith concluded in 1995 that "results of meta-analyses that are exclusively based on small trials should be distrusted - even if the combined effect is statistically highly significant. Several medium-sized trials of high quality seem necessary to render results trustworthy."[5] These conclusions still hold in 2009. If appropriately used and interpreted, funnel plots with formal statistical tests of asymmetry, stratified analyses accompanied by tests of interaction and heterogeneity-adjusted trial sequential analyses will all contribute to our understanding about when to consider a meta-analysis conclusive.

Acknowledgments: We are grateful to Kristian Thorlund, Jørn Wetterslev and Christian Gluud for help with trial sequential analysis of the magnesium trials and for stimulating discussions.

References

1. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ 1991;303(6816):1499-503.
2. Yusuf S, Teo K, Woods K. Intravenous magnesium in acute myocardial infarction. An effective, safe, simple, and inexpensive intervention. Circulation 1993;87(6):2043-6.
3. Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). Lancet 1992;339(8809):1553-8.
4. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. Lancet 1995;345(8951):669-85.

5. Egger M, Smith GD. Misleading meta-analysis. BMJ 1995;310(6982):752-4.
6. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337(8746):867-72.
7. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629-34.
8. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet 1998;351(9095):47-52.
9. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997;18(6):580-93; discussion 661-6.
10. Armitage P. Sequential analysis in therapeutic trials. Annu Rev Med 1969;20:425-30.
11. Pocock S. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191-9.
12. Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.
13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35(3):549-56.
14. Egger M, Smith GD, Sterne JA. Meta-analysis: is moving the goal post the answer? Lancet 1998;351(9114):1517.
15. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61(1):64-75.
16. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2008.
17. Thorlund K, Devereaux PJ, Wetterslev J, Gyuatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2008.
18. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, 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.
19. Chan AW, Hrobjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291(20):2457-65.

20. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE 2008;3(8):e3081.
21. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001;54(10):1046-55.
22. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001;323(7303):42-6.
23. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med 2006;25(20):3443-57.
24. Li J, Zhang Q, Zhang M, Egger M. Intravenous magnesium for acute myocardial infarction. Cochrane Database Syst Rev 2007(2):CD002755.

Discussion and outlook

This thesis suggests that flaws in the conduct and design of randomised clinical trials and meta-analyses frequently result in biased estimates of treatment benefits. Methodological trial characteristics, including allocation concealment and exclusions of randomised patients from the analysis were associated with estimated treatment benefits and may have biased results of individual trials and meta-analyses. In addition, small study effects made the interpretation of results from several included meta-analyses difficult. The impact of these characteristics on estimated treatment benefits in a specific situation was unpredictable, however. I also assessed the impact of methodological quality and sample size on the between-trial heterogeneity by restricting meta-analyses to trials with high methodological quality or to large trials. The variability between trials was substantially reduced when meta-analyses were restricted. In two additional studies, inter-observer variation in data extraction and multiplicity of outcome data presented in trial reports frequently hampered the validity of results in the studied meta-analyses. Monte Carlo simulations showed that disagreements between different observers when extracting data from trial reports and multiplicity of data in trial reports resulted in a substantial variability in pooled estimates of treatment benefit.

To my knowledge, this thesis provides the first systematic examination of bias and variation in randomised trials and meta-analyses of patient-reported outcomes measured on a continuous or rating scale, and of the impact of multiple choices in data extraction on the validity of results from meta-analyses. Most previous meta-epidemiological studies have concentrated on the associations of methodological trial characteristics such as allocation concealment, double-blinding and dropouts or exclusions with estimated treatment benefits measured on an odds ratio scale.[1-7] Meta-epidemiological approaches were used in this thesis, which allowed assessing the variability in effects between meta-analyses.[8 9]

The thesis is based on information extracted from published trial reports and depends on the quality of reporting, which is generally low.[10] The assessment of methodological characteristics such as allocation concealment will depend more on the quality of reporting than sample size of a trial.[11 12] Trial misclassification, if it is non-differential, will result in an underestimation of the true associations between methodological characteristics and treatment benefits. However, low quality of reporting and low quality of trial conduct are often intertwined: faulty reporting may represent faulty methods.[6 13] Therefore, misclassification might not be a frequent problem. Because this thesis is based on published trial results, the results will also be affected by selective reporting of outcomes and analyses. Selective

reporting of different analyses, e.g. reporting of changes rather than absolute values, or preferential reporting of more favourable per-protocol analyses rather than more conservative intention-to-treat analyses, might have affected the results in this study. I was unable to disentangle bias resulting from selective reporting and methodological quality. The variability of meta-analyses results due to different observers and multiple outcome data presented in trial reports might have been underestimated. The actual multiplicity might be even be more pronounced if both, published and unpublished outcome data had been available.

Meta-epidemiological studies are observational by nature and the associations between estimates of treatment benefits and methodological components might be confounded by several factors. In this-epidemiological study of 190 trials, confounding by disease and type of intervention was minimised by a restriction to meta-analyses of osteoarthritis trials and by stratification according to type of intervention.[8] Flaws in methodological conduct are likely to cluster in trials and therefore, confounding by different methodological components was controlled by stratification and reported as sensitivity analyses. Differences between trials with and trials without adequate methodology (concealment of allocation, patient blinding and intention-to-treat analysis) diminished or disappeared entirely after accounting for sample size of the trials. Conversely, the association between sample size and treatment effects were completely robust, when accounting for methodological components. Sample size of a trial might therefore be the best single proxy for the cumulative impact of methodological deficiencies, selective reporting and publication bias. Alternatively, smaller studies may be more careful in implementing the intervention or may include patients who are particularly likely to benefit from the intervention, both aspects resulting in larger treatment effects and true clinical heterogeneity.[14-16]

The results presented in this thesis have several implications for researchers performing randomised trials and meta-analyses. To avoid potential bias, trialists should always ensure adequate concealment of allocation and take measures to minimise dropout rates, maximise compliance and minimise missing outcome data. Blinding of patients is desirable and should be attempted. Results from intention to treat analyses should always be described in reports of randomised trials. The CONSORT statement urges transparent reporting of concealment of allocation, measures taken to blind study participants, the flow of participants through the various stages of a trial including withdrawals and losses to follow-up and the reasons for exclusions from the analysis.[17 18] Authors of reports of randomised trials should follow the CONSORT statement[17 18] to ensure fully transparent reporting of methods and results.

In systematic reviews and meta-analyses, a detailed protocol might improve the reliability of results. Data extraction should be done by more than one observer, and should be based on results from analyses including all randomised patients, whenever possible. Results of meta-analyses based on methodologically questionable trials should be distrusted. Even a meta-analysis that includes a large number of patients reaching the required information size to get adequate power[19] should be interpreted with caution, if mainly trials at high risk of bias contributed to the analysis, which may have distorted results.[20] The influence of allocation concealment, patient blinding, exclusions form the analysis, and sample size should be routinely assessed in stratified analyses.

The Cochrane Collaboration now advocates reporting the risk of bias for each included randomised trial in a Cochrane review by assessing individual methodological components such as sequence generation, allocation concealment, blinding, incomplete outcome data and selective reporting bias, which should help readers to judge the extent of bias in the reported meta-analysis.[21] Recently, Bayesian hierarchical models have been discussed to adjust treatment effects in a meta-analysis for bias.[22] These models use empirical prior information about the extent and direction of bias in randomised trials and meta-analyses. Meta-epidemiological studies might provide information that can be used to calculate bias-adjusted treatment effects in meta-analyses.[22] Further studies that disentangle the interplay between different dimensions of methodological quality will provide better understanding of the underlying mechanisms.

References

1. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287(22):2973-82.
2. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2008;149(3):219.
3. 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.
4. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999;354(9193):1896-900.
5. Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 2007;36(4):847-57.
6. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.
7. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, 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.
8. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
9. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ (Clinical research ed.)* 2008;336(7644):601-605.
10. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;287(21):2801-4.
11. Pildal J, Chan AW, Hrobjartsson A, Forfang E, Altman DG, Gotzsche PC. Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. *BMJ* 2005;330(7499):1049.

12. Hrobjartsson A, Pildal J, Chan AW, Haahr MT, Altman DG, Gotzsche PC. Reporting on blinding in trial protocols and corresponding publications was often inadequate but rarely contradictory. *J Clin Epidemiol* 2009;62(9):967-73.
13. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.
14. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134(8):663-94.
15. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134(8):657-62.
16. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol* 2009;38(1):276-86.
17. Nüesch E, Juni P. Commentary: Which meta-analyses are conclusive? *Int J Epidemiol* 2009;38(1):298-303.
18. Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.2 [updated September 2009] The Cochrane Collaboration, 2008.
19. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JA. Models for potentially biased evidence in meta-analysis using empirically based priors. *J R Stat Soc Ser A Stat Soc* 2009;172(1):119-136.

Acknowledgments

First of all, I would like to thank my supervisors Prof. Dr. Peter Jüni and Dr. Sven Trelle (ISPM, University of Bern, Switzerland) who gave me the opportunity for my PhD studies. They introduced me to biostatistics and clinical epidemiology, and to clinical and methodological research through scientific discussions, constructive critics and productive team work. They had always an open ear for my questions despite their full calendars. I am very grateful to my external co-referee Prof. Dr. Doug Altman (University of Oxford, UK) for his constructive ideas and encouragements for my research project. I would also like to thank Prof. Dr. Matthias Egger (ISPM, University of Bern), who was the mentor during my PhD studies.

Special thanks go to Dr. Anne Rutjes (ISPM, University of Bern) who taught me a lot about systematic reviews and meta-analyses, about methodological research projects and about how to get things done despite difficulties, to Dr. Stephan Reichenbach (ISPM, University of Bern) from whom I learned a lot about osteoarthritis and epidemiological studies and to Britta Tendal (The Nordic Cochrane Centre, Copenhagen, Denmark) for our interesting scientific discussions and for our productive collaboration.

I am also grateful to Prof. Dr. Peter Gøtzsche (The Nordic Cochrane Centre, Copenhagen, Denmark), Dr. Julian Higgins (University of Cambridge, UK) and to all other co-authors for the productive collaborations, to the editorial team of the Cochrane Musculoskeletal Group for substantial input and support to our Cochrane reviews, to Dr. Nicole Bender (ISPM, University of Bern) for help with regulations of the various Graduate Schools, to Malcolm Sturdy (ISPM, University of Bern) for database development and maintenance and to Madeleine Dähler and Natalie Studer (both ISPM, University of Bern) for administrative support. I would like to thank Madeleine, Rebekka, Simon and Sven who shared office with me, and all other members of the ISPM University of Bern and CTU Bern, Bern University Hospital. These people contributed substantially to my interesting and enjoyable time during the three years of my PhD studies.

During my PhD studies I received financial support from the Swiss National Science Foundation’s National Research Program 53 on musculoskeletal health, other grants of the Swiss National Science Foundation, the Swiss School of Public Health and the Graduate School of Cellular and Biomedical Sciences at University of Bern.

Last but not least I am very grateful to my family for their support and interest during my studies, and especially to my dear friend Michael for his continuous love, encouragement and patience. He encouraged me to start this thesis despite all my doubts and concerns, supported me when I was going through difficulties in my work, and shared the joy over my successes until termination of my studies.

Curriculum vitae

Personal data

Name	Eveline Bettina Nüesch
Residence	Saint-Louis-Strasse 2 CH-4056 Basel
E-Mail	enueesch@ispm.unibe.ch
Date of birth	15 August 1978
Origin	Balgach SG

Education

2007 – 2010	PhD in Biostatistics and Clinical Epidemiology Faculty of Medicine, University of Bern
2006 – 2008	Master Degree in Statistics Faculty of Economics, University of Neuchâtel
1998 – 2003	Diploma in Biology (MSc Degree) Faculty of Science, University of Basel
1993 – 1998	Matura Typus B Kantonsschule Heerbrugg (SG)

Professional experience

2008 – present	Statistician, Clinical Trials Unit (CTU), Bern University Hospital
2007 – present	Research Fellow, Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine (ISPM), University of Bern
2006 – 2007	Drug Safety Associate, Section Pharmacovigilance, Quintiles AG, Basel
2004 – 2005	Research Fellow in Genetics, Institute of Plant Biology, University of Zürich

List of publications

1. Rutjes AWS, **Nüesch E,** Sterchi R, Jüni P: Therapeutic ultrasound for osteoarthritis of the knee or hip. *Cochrane Database Syst Rev* 2010 Jan 20 (1):CD003132.

2. **Nüesch E,** Reichenbach S, Trelle S, Rutjes AWS, Liewald K, Sterchi R, Altman DG, Jüni P: The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum*. 2009 Nov 30;61(12):1633-1641.

3. Reichenbach S, Jüni P, **Nüesch E,** Frey F, Ganz R, Leunig M: An examination chair to measure internal rotation of the hip in routine settings: a validation study. *Osteoarthritis Cartilage*. 2009 Oct 8. [Epub ahead of print]

4. **Nüesch E,** Rutjes AWS, Trelle S, Reichenbach S, Jüni P: Doxycycline for osteoarthritis of the knee or hip. *Cochrane Database Syst Rev* 2009 Oct 7 (4):CD007323.

5. Rutjes AWS, **Nüesch E,** Reichenbach S, Jüni P: S-Adenosylmethionine for osteoarthritis of the knee or hip. *Cochrane Database Syst Rev* 2009 Oct 7 (4):CD007321.

6. **Nüesch E,** Rutjes AWS, Husni E, Welch V, Jüni P: Oral or transdermal opioids for osteoarthritis of the knee or hip. *Cochrane Database Syst Rev* 2009 Oct 7 (4):CD003115.

7. *Rutjes AWS, ***Nüesch E,** Sterchi R, Kalichman L, Hendriks E, Reichenbach S, Osiri M, Brosseau L, Jüni P: Transcutaneous electrical nerve stimulation for osteoarthritis of the hip or knee. *Cochrane Database Syst Rev* 2009 Oct 7 (4):CD002823.

8. **Nüesch E,** Trelle S, Reichenbach S, Rutjes AWS, Bürgi E, Scherer M, Altman DG, Jüni P: The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ.* 2009 Sep 07;339:b3244.

9. Tendal B, Higgins JP, Jüni P, Hróbjartsson A, Trelle S, **Nüesch E,** Wandel S, Jørgensen AW, Gesser K, Ilsøe-Kristensen S, Gøtzsche PC: Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ*. 2009 Aug 13;339:b3128.

10. **Nüesch E,** Jüni P: Commentary: Which meta-analyses are conclusive? *Int J Epidemiol.* 2009 Feb;38(1):298-303.

11. Jüni P, Battaglia M, **Nüesch E,** Hämmerle G, Eser P, van Beers R, Vils D, Bernhard J, Ziswiler H, Dähler M, Reichenbach S, Villiger PM.: A randomised controlled trial of spinal manipulative therapy in acute low back pain. *Ann Rheum Dis.* 2009 Sep;68(9):1420-7. Epub 2008 Sep 5.

12. Brand L, Hörler M, **Nüesch E,** Vassalli S, Barrell P, Yang W, Jefferson RA, Grossniklaus U, Curtis MD.: A versatile and reliable two-component system for tissue-specific gene induction in Arabidopsis. *Plant Physiol.* 2006 Aug;141(4):1194-204.

13. *Altenbach D, ***Nüesch E,** Ritsema T, Boller T, Wiemken A.: Mutational analysis of the active center of plant fructosyltransferases: Festuca 1-SST and barley 6-SFT. *FEBS Lett.* 2005 Aug 29;579(21):4647-53.

14. Altenbach D, **Nüesch E,** Meyer AD, Boller T, Wiemken A.: The large subunit determines catalytic specificity of barley sucrose:fructan 6-fructosyltransferase and fescue sucrose:sucrose 1-fructosyltransferase. *FEBS Lett.* 2004 Jun 4;567(2-3):214-8.

* indicates equally contributing authors.

Declaration of originality

Last name, first name: Nüesch, Eveline Bettina

Matriculation number: 98-055-700

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.

I am aware that in case of non-compliance, the Senate is entitled to divest me of the doctorate degree awarded to me on the basis of the present thesis, in accordance with the “Statut der Universität Bern (Universitätsstatut; UniSt)”, Art. 20, of 17 December 1997.

Place, date Signature

…… ……