

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

Eveline Nuesch, research fellow,<sup>1,2</sup> Sven Trelle, associate director,<sup>1,2</sup> Stephan Reichenbach, senior research fellow,<sup>1,3</sup> Anne W S Rutjes, senior research fellow,<sup>1,4</sup> Beatrice Tschannen, research fellow,<sup>1</sup> Douglas G Altman, director and professor of statistics in medicine,<sup>5</sup> Matthias Egger, head of department and professor of epidemiology and public health,<sup>1</sup> Peter Juni, head of division and professor of clinical epidemiology<sup>1,2</sup>

<sup>1</sup>Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland

<sup>2</sup>CTU Bern, Bern University Hospital, Switzerland

<sup>3</sup>Department of Rheumatology, Immunology and Allergology, Bern University Hospital, Switzerland

<sup>4</sup>Laboratory of Clinical Epidemiology of Cardiovascular Disease, Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Santa Maria Imbaro, Italy.

<sup>5</sup>Centre for Statistics in Medicine, University of Oxford, Oxford

Correspondence to: P Juni  
juni@ispm.unibe.ch

Cite this as: *BMJ* 2010;341:c3515  
doi:10.1136/bmj.c3515

### 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 patients' reported pain as an outcome.

**Methods** We compared estimated benefits of treatment between large trials (at least 100 patients per arm) and small trials, explored funnel plots supplemented with lines of predicted effects and contours of 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% confidence interval  $-0.34$  to  $-0.08$ ,  $P=0.001$ ). Depending on criteria used, six to eight funnel plots indicated small study effects. In six of 13 meta-analyses, the overall pooled estimate suggested a clinically relevant, significant benefit of treatment, whereas analyses restricted to large trials and predicted effects in large trials yielded smaller non-significant estimates.

**Conclusions** Small study effects can often 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 often been suggested that small trials tend to report larger treatment benefits than larger trials.<sup>1,2</sup> Such small study effects can result from a combination of lower methodological quality of small trials and publication and other reporting biases<sup>2-8</sup> but could also reflect clinical heterogeneity if small trials were more careful in

selecting patients and implementing the experimental intervention.<sup>9</sup> The funnel plot is a scatter plot of treatment effects against standard error as a measure of statistical precision.<sup>9,10</sup> Imprecision of estimated treatment effects will increase as the sample size of component trials decreases. 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.<sup>9</sup> The plot can be enhanced by lines of the predicted treatment effect from meta-regression with the standard error as explanatory variable<sup>11,12</sup> and contours that divide the plot into areas of significance and non-significance.<sup>13,14</sup> A recent study of trials of anti-depressants<sup>15</sup> 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 methods and analysis.<sup>14</sup>

Small study effects are not uncommon in osteoarthritis research; several recent meta-analyses found pronounced asymmetry of funnel plots.<sup>16-18</sup> We previously studied the influence of methodological characteristics on estimated effects in a set of clinical osteoarthritis trials that used pain outcomes reported by patients and found that deficiencies in concealment of random allocation, blinding of patients, and analyses can distort the results in these trials.<sup>19,20</sup> Different components of inadequate trial methods often concur. A trial with adequate allocation concealment, for example, is more likely to report analyses according to the intention to treat principle.<sup>19,20</sup> Meta-epidemiological studies found that smaller trials are less likely to use adequate random sequence generation, adequate allocation concealment, and double blinding<sup>7,8,19</sup> and that different methodological components are associated with exaggerated benefits of treatment.<sup>7,8,19-23</sup>

We explored the presence and extent of small study effects in meta-analyses of osteoarthritis trials using three different approaches: analyses stratified according to sample size, inspection of funnel plots, and

prediction of treatment effects based on the standard error used as a measure of statistical precision of trials. We then determined 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 influenced 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 pain related outcome reported by patients for any intervention compared with 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.<sup>20</sup> Reports of all component trials from included meta-analyses were obtained. No language restrictions were applied.

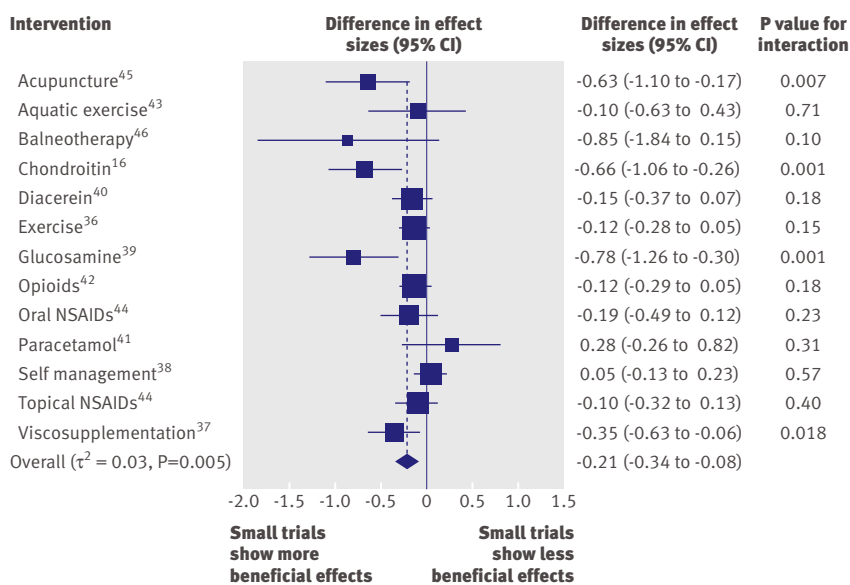
### Data extraction and quality assessment

Two reviewers used a standardised form to independently extract data from individual trials regarding design, interventions, year of publication, trial size, sample size calculation, exclusions, and results.<sup>20</sup> 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.<sup>16,19,24</sup> Concealment of treatment allocation was considered as adequate if investigators responsible for selection of patients were unable to suspect before allocation which treatment was next—for example, 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). We used a cut-off of an average of 100 randomised patients per treatment arm to distinguish between small and large trials, irrespective of the number of patients subsequently excluded from the analysis. A two arm trial with 110 patients in one arm and 95 patients in the second arm, for example, was classified as large. A sample size of  $2 \times 100$  patients will yield more than 80% power to detect a small to moderate effect size of  $-0.40$  at a two sided  $\alpha=0.05$ , 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 (SD). Negative effect sizes indicate a beneficial effect of the experimental intervention. If some required data were unavailable, we used approximations as previously described.<sup>16</sup> Within each meta-analysis, we estimated effect sizes of large ( $\geq 100$  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  $\tau^2$  as a measure of heterogeneity between trials,<sup>25,26</sup> 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.<sup>26,27</sup> Meta-analyses that included exclusively small or exclusively large trials did not contribute to the analysis. Negative differences in effect sizes indicate that small trials show more beneficial treatment effects than large trials. The variability between meta-analyses was quantified with the heterogeneity variance  $\tau^2$ . To account for the correlation between sample size and methodological quality, we used stratification by these components in analogy to Mantel-Haenszel procedures<sup>28</sup> and derived differences between small and large trials adjusted for concealment of allocation, blinding of patients, 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<sup>20</sup>: heterogeneity between trials 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  $\leq -0.5$ )),<sup>24,29</sup> and type of intervention assessed in the meta-analysis (drug *v* other interventions, conventional *v* complementary medicine). These stratified analyses were accompanied by interaction tests based on z scores, which are defined as the difference in effect sizes between strata divided by the standard error (SE) of the difference.



**Fig 1** | Difference in effect sizes between 95 small trials with fewer than 100 patients per arm and 58 large trials. Negative differences indicate 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

**Table 1** | Comparison of characteristics between small and large trials in meta-analyses in osteoarthritis research. Figures are numbers (percentages)

|                              | No of allocated patients |                     | P value* |
|------------------------------|--------------------------|---------------------|----------|
|                              | <100 per arm (n=95)      | ≥100 per arm (n=58) |          |
| Concealment of allocation:   |                          |                     |          |
| Adequate                     | 19 (20)                  | 22 (38)             | 0.010    |
| Inadequate/unclear           | 76 (80)                  | 36 (62)             |          |
| Blinding of patients:        |                          |                     |          |
| Adequate                     | 41 (43)                  | 30 (52)             | 0.25     |
| Inadequate/unclear           | 54 (57)                  | 28 (48)             |          |
| Intention to treat analysis: |                          |                     |          |
| Yes                          | 16 (17)                  | 16 (28)             | 0.23     |
| No/unclear                   | 79 (83)                  | 42 (72)             |          |
| Sample size calculation:     |                          |                     |          |
| Reported                     | 37 (39)                  | 38 (66)             | <0.001   |
| Not reported                 | 58 (61)                  | 20 (34)             |          |
| Year of publication:         |                          |                     |          |
| 1980-98                      | 45 (47)                  | 11 (19)             | 0.001    |
| 1999-2007                    | 50 (53)                  | 47 (81)             |          |
| Drug intervention†:          |                          |                     |          |
| Yes                          | 70 (74)                  | 43 (74)             | 0.97     |
| No                           | 25 (26)                  | 15 (26)             |          |
| Complementary medicine‡:     |                          |                     |          |
| Yes                          | 30 (32)                  | 11 (19)             | 0.09     |
| No                           | 65 (68)                  | 47 (81)             |          |

\*Derived from logistic regression models adjusted for clustering of trials within meta-analyses.

†Includes chondroitin, diacerin, glucosamine, NSAIDs, opioids, paracetamol, and viscosupplementation.

‡Includes acupuncture, balneotherapy, chondroitin, and glucosamine.

We drew funnel plots, plotting effect sizes of individual trials on the x axis against their SEs on the y axis. Under the assumption that effect sizes of individual studies are normally distributed, significance of any point of the funnel plot can be derived directly from effect sizes and corresponding SE with Wald tests.<sup>13,30</sup> As previously described, we used this to enhance funnel plots by contours dividing the plot into areas of significance with a two sided  $P \leq 0.05$  and areas of non-significance with a  $P > 0.05$ .<sup>13,30</sup> If trials seem to be missing in areas of non-significance, this adds to the notion of the presence of bias.<sup>13,14</sup> We added lines to the funnel plots, which represented the predicted treatment effect derived from univariable random effects meta-regression models using the SE as explanatory variable.<sup>11,12</sup> Then, we assessed funnel plot asymmetry with regression tests, a weighted linear regression of the effect sizes on their SEs, using the inverse of the variance of effect sizes as weights.<sup>29</sup> Asymmetry coefficients, defined as the difference in effect size per unit increase in SE,<sup>10,11</sup> were combined with inverse variance random effects models, crude and adjusted for adequate concealment of allocation, blinding of patients, and intention to treat analysis. Negative asymmetry coefficients indicate that estimated treatment benefits increase with increasing SEs.

We compared three different approaches to estimate treatment effects: 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 SE as explanatory variable for trials with a SE of 0.1.<sup>12,14</sup> A SE 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  $-0.40$  SD units, which corresponds to the median minimal clinically important difference found in recent trials in patients with osteoarthritis.<sup>31-34</sup> Results were considered concordant if point estimates differed by less than 0.10 SD units<sup>35</sup> and if the status of significance at a two sided  $\alpha=0.05$  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, heterogeneity between trials, precision defined as the inverse of the SE, 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-Corp, College Station, TX).

## RESULTS

The study sample and its origin are described elsewhere.<sup>19,20</sup> Twenty one meta-analyses described in 17 reports were eligible. Of these, 13 meta-analyses<sup>16,36-46</sup> (153 trials with 41 605 patients) included both small and large trials and contributed to the current analyses. The median number of trials included per meta-analysis was 12 (range 3-24) and the median number of patients was 1849 (347-13 659). The pooled effect sizes ranged from  $-0.07$  to  $-1.11$  and the heterogeneity between trials from a  $\tau^2$  of 0.00 to 0.47. Eight meta-analyses assessed drug interventions, and five assessed non-drug interventions. Four assessed interventions in complementary medicine, and nine assessed 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 ( $P=0.001$ ) and were more likely to report adequate concealment of allocation ( $P=0.01$ ) and calculation of sample size ( $P<0.001$ ).

The average difference in effect sizes between large and small trials across the 13 included meta-analyses was  $-0.21$  (95% confidence interval  $-0.34$  to  $-0.08$ ,  $P=0.001$ ), with more beneficial effects found in small trials (fig 1). At the level of individual meta-analyses, tests for interaction between treatment benefits and trial size were positive in four meta-analyses (31%).<sup>16,37,39,45</sup> The variability across meta-analyses was small to moderate, with a  $\tau^2$  estimate of 0.03 ( $P=0.005$ ). Table 2 shows the average difference in effect sizes between large and small trials, both crude and after adjustment for the methodological quality of trials. Differences in effect sizes between small and large trials were robust after adjustment for blinding of patients ( $-0.21$ ,  $-0.33$  to  $-0.09$ ,  $P=0.001$ ), slightly attenuated after adjustment for concealment of allocation ( $-0.16$ ,

**Table 2** | Estimates of small study effects in meta-analyses of osteoarthritis trials

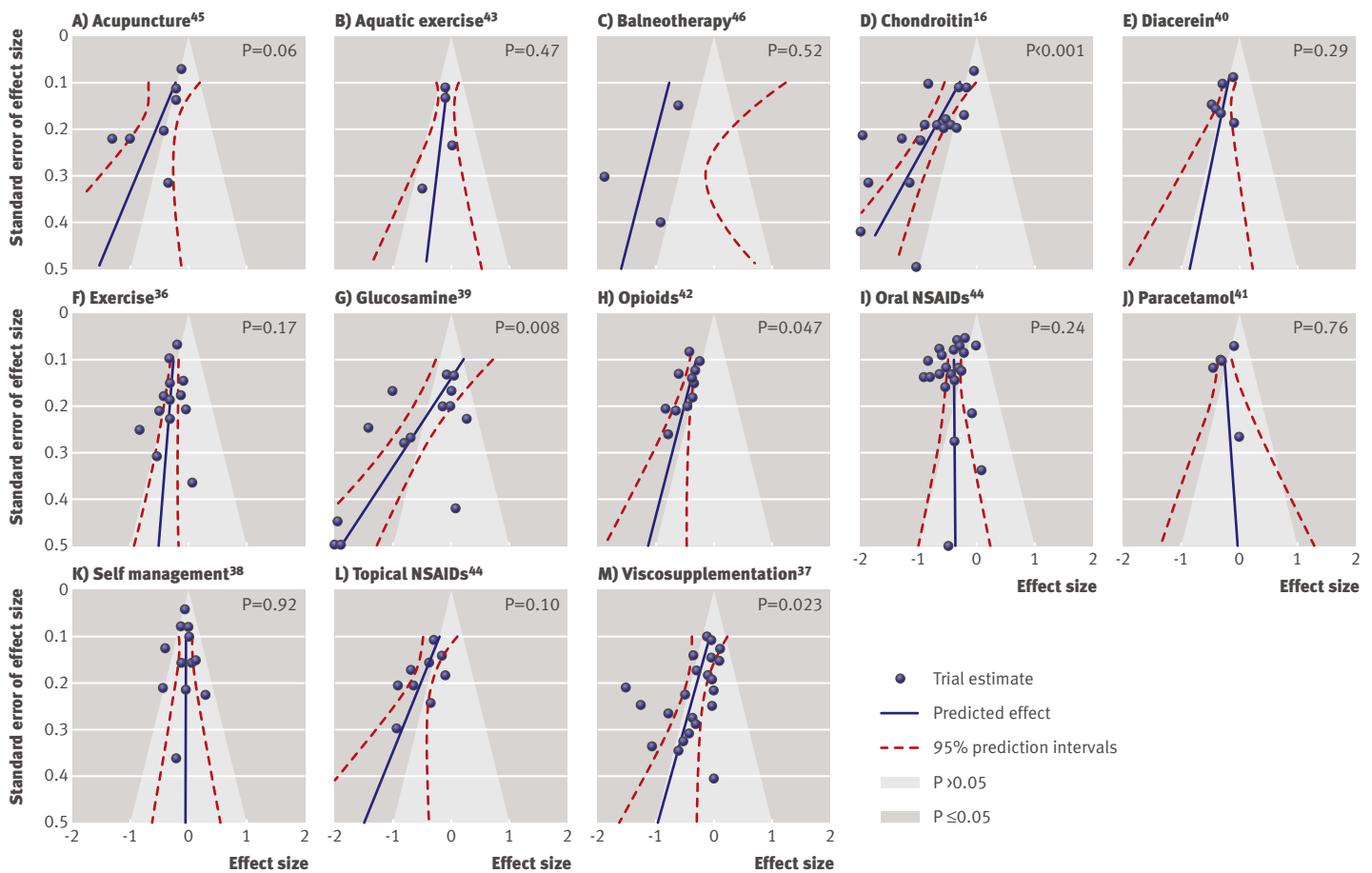
|  | Difference in effect sizes |         | Variability† (P value) |
|--|----------------------------|---------|------------------------|
|  | Difference* (95% CI)       | P value |                        |
| Overall, crude                         | -0.21 (-0.34 to -0.08)     | 0.001   | 0.03 (0.005)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

\*Difference in effect size between 95 small and 58 large trials.  
†Estimate of between meta-analyses heterogeneity variance ( $\tau^2$ ).

-0.27 to -0.06,  $P=0.002$ ), but nearly halved after adjustment for intention to treat analysis (-0.12, -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 heterogeneity between trials found in overall meta-analyses, and the 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 heterogeneity between trials, 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 SE as an explanatory variable and 5% contour areas to display areas of significance and non-significance. For six funnel plots, the scatter of effect estimates and the prediction line indicated asymmetry (A, D, G, H, L, M).<sup>16 37 39 42 44 45</sup> For two other funnel plots, the prediction lines mainly suggested asymmetry (C, E),<sup>40 46</sup> whereas the remaining five funnel plots seemed symmetrical and prediction lines nearly upright (B, F, I, J, K).<sup>36 38 41 43 44</sup> The regression test was significant at  $P \leq 0.05$  in four meta-analyses (D, G, H, M)<sup>16 37 39 42</sup> and showed a statistical trend in another two ( $P \leq 0.10$ , A, L).<sup>44 45</sup> In five funnel plots, the contours to distinguish between areas of significance and non-significance at  $P=0.05$  suggested missing trials in areas of non-significance (A, C, D, H, L).<sup>16 42 44 46</sup> The weighted average of asymmetry coefficients across all meta-analyses was -1.79 (-2.81 to -0.78). This indicates that, on average, the estimated treatment benefit increases by 1.79 SD units for each unit increase in the SE. It was much the same after adjustment for concealment of allocation (-1.86, -2.98 to -0.74), slightly more pronounced after adjustment for blinding (-2.22, -3.28 to -1.17), but slightly less pronounced



**Fig 2** | Funnel plots of 13 included meta-analyses including prediction lines from univariable meta-regression models with SE as explanatory variable (dashed red) and 5% contour areas to display areas of significance (blue) and non-significance (pale blue). Lines for predicted effects should be interpreted independently of contours delineated by shaded areas. Extent of deviation of lines for predicted effects from vertical line should be considered, irrespective of relation of lines to contours. P values derived from regression tests for asymmetry. NSAIDs=non-steroidal anti-inflammatory drugs

**Table 3** Analyses stratified according to characteristics of meta-analyses of osteoarthritis trials

| 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:            |                     |              |                                     |                        |                    |
| Small benefit  | 10                  | 115          | -0.13 (-0.22 to -0.03)              | 0.01 (0.17)            | <0.001             |
| Large benefit  | 3                   | 38           | -0.72 (-1.02 to -0.43)              | 0.00 (0.90)            |                    |
| Heterogeneity between trials in overall meta-analysis: |                     |              |                                     |                        |                    |
| Low heterogeneity                                      | 8                   | 87           | -0.08 (-0.16 to -0.00)              | 0.00 (0.66)            | <0.001             |
| High heterogeneity                                     | 5                   | 66           | -0.55 (-0.73 to -0.36)              | 0.00 (0.46)            |                    |
| Drug intervention‡:                                    |                     |              |                                     |                        |                    |
| Yes  | 8                   | 113          | -0.23 (-0.39 to -0.08)              | 0.03 (0.021)           | 0.67               |
| No   | 5                   | 40           | -0.17 (-0.40 to 0.06)               | 0.03 (0.041)           |                    |
| Complementary medicine§:                               |                     |              |                                     |                        |                    |
| Yes  | 4                   | 41           | -0.70 (-0.95 to -0.45)              | 0.00 (0.96)            | <0.001             |
| No   | 9                   | 112          | -0.10 (-0.18 to -0.03)              | 0.00 (0.43)            |                    |

\*Estimate of between meta-analyses heterogeneity variance ( $\tau^2$ ).

†Derived from interaction tests based on z scores.

‡Includes chondroitin, diacerein, glucosamine, NSAIDs, opioids, paracetamol, and viscosupplementation.

§Includes acupuncture, balneotherapy, chondroitin, and glucosamine.

after adjustment for intention to treat analysis (-1.41, -2.27 to -0.54). Confidence intervals of adjusted and the unadjusted estimates overlapped widely.

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

## DISCUSSION

In this meta-epidemiological study of 13 meta-analyses of 153 osteoarthritis trials, we found larger estimated benefits of treatment in small trials with fewer than 100 patients per trial arm 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.<sup>24</sup> Small study effects, however, were more prominent in five of the 13 meta-analyses. These showed a large extent of

statistical heterogeneity, larger pooled estimates of treatment benefit than would typically be expected from an effective intervention in osteoarthritis, and mainly covered complementary medical interventions. Taking into account contours used to distinguish between areas of significance and non-significance and lines of treatment effects predicted for different standard errors, we found eight 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 at least 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 six meta-analyses, with the overall pooled estimate suggesting a clinically relevant, significant benefit of treatment, which was not found in the other two approaches aimed at estimating the effect in large trials only.

## Strengths and limitations of the study

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.<sup>7,8,47</sup> When accounting for blinding of patients, we found the association between sample size and treatment effect to be completely robust. Adjustment for concealment of allocation resulted in a slight attenuation, whereas adjustment 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 might contribute to the observed small study effects, which is in line with the findings of a recent study of trials of anti-depressants.<sup>15</sup> In addition to publication and reporting biases, switching from an intention to treat to a per

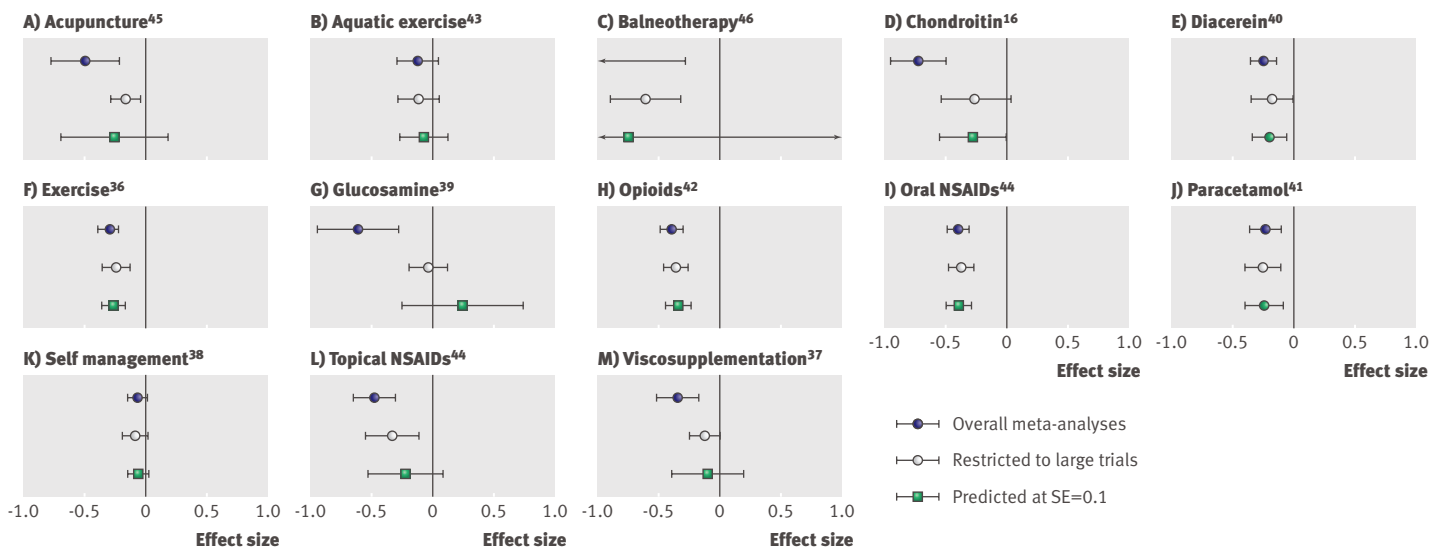
protocol analysis seemed to contribute to discrepancies between published and unpublished results.<sup>14</sup> The assessment of components of methodological quality will depend strongly on reporting quality<sup>48</sup> and might be affected by misclassification, whereas sample size or standard error might be extracted more easily. Sample size or statistical precision might therefore be the best single proxy for the cumulative effect of the different sources of bias in randomised trials in osteoarthritis and probably also in other fields: selection, performance, detection, and attrition bias<sup>49</sup>; selective reporting of outcomes<sup>34</sup>; and publication bias.<sup>6,50</sup>

The most important limitation of our study is that we cannot exclude alternative explanations of small study effects other than bias: smaller trials might 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.<sup>2,9,51</sup> Selection of patients and implementation of interventions might be particularly important for complex interventions. For example, in a meta-analysis of inpatient geriatric consultations, some differences in observed effects between small and large trials could be explained by more careful implementation of the intervention by experienced consultants.<sup>9,52</sup> The low quality of reporting, however, makes it difficult to examine this issue systematically, and we are unaware of any methodological study to have dealt with this. Interestingly, for four out of five meta-analyses of complex interventions in our study (aquatic exercise, balneotherapy, exercise, and self management), we found little evidence for asymmetrical funnel plots, and only for acupuncture, as a complex complementary intervention, was there evidence of asymmetry. Investigators should be careful to report the selection of patients and the implementation of interventions in sufficient detail, particularly in trials of complex interventions,

to allow a more systematic appraisal of this issue in the future. 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 could have been too superficial and some of the selection criteria too narrow to include a large proportion of unpublished trials. The meta-analyses included in our study, however, are probably representative, and we believe therefore that our results are generalisable. Another limitation is that our analysis was based on published information only and depends on the quality of reporting, which is often unsatisfactory.<sup>49</sup>

**Comparison with other studies**

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<sup>7,8</sup> found more beneficial treatment effects in small trials with inadequate methodology compared with large trials. In an analysis of homoeopathy trials Shang et al found that smaller trials and those of lower quality show more beneficial treatment effects than larger and higher quality trials.<sup>11</sup> 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.<sup>14</sup> Application of 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, Moreno et al regressed treatment effects against their variance, which performed well in a simulation study but has been shown to give similar results to using the standard



**Fig 3** | Results of individual random effects meta-analyses of all trials (blue circle), results of random effects meta-analyses restricted to large trials with at least 100 patients per arm (open circle), and effect sizes for trials with SE of 0.1 predicted from random effects meta-regression models (green square). NSAIDs=non-steroidal anti-inflammatory drugs

**WHAT IS ALREADY KNOWN ON THIS TOPIC**

Small study effects refer to a tendency of small trials to report larger benefits of treatment than larger trials

Such effects can result from a combination of lower methodological quality of small trials and publication and other reporting biases or from true clinical heterogeneity

Contour enhanced funnel plots and regression tests can be used to identify small study effects in meta-analyses

**WHAT THIS STUDY ADDS**

Small study effects often affect results of meta-analyses in osteoarthritis research

Sample size or statistical precision might be the best single proxy for the cumulative effect of the different sources of bias in randomised osteoarthritis trials and probably also in other specialties

In the presence of small study effects, restriction of analyses to large trials or predictions of treatment benefits observed in large trials might provide more valid estimates than overall analyses of trials, irrespective of sample size

error as an explanatory variable.<sup>12</sup> In funnel plots, however, treatment effects will typically be plotted against their standard error, and significance tests will be generally based on *z* or *t* values, which again are directly derived from standard errors. 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.<sup>12</sup> 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 SE of 0.1 will typically be found in a large two arm trial with a continuous primary outcome including 200 patients per group. Such a trial will yield more than 95% power to detect an effect size of  $-0.40$  SD units and still more than 80% power to detect an effect size of about  $-0.30$  SD units. Trials considerably larger than that will probably not be needed for continuous primary outcomes.

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 might yield precise estimates. This was observed in five meta-analyses in our study<sup>37,39,44-46</sup> and was 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 methods used for assessing funnel plot asymmetry might 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  $-0.40$  SD units, which corresponds to the median minimal clinically important difference found in recent studies in patients with osteoarthritis.<sup>32-34</sup> For binary outcomes,

the definition of large trials will depend on event rates in the control group and a definition of what constitutes a moderate but clinically relevant effect. In addition, the regression test for funnel plot asymmetry originally reported<sup>9</sup> might 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.<sup>51</sup> Non-parametric tests will result in lower power than the regression tests discussed here and might be less appropriate. Similarly, funnel plots and analyses stratified according to sample size might be inconclusive if the range of sample sizes or standard errors of included trials is restricted. For example, the meta-analysis of diacerein trials in our study included only moderately sized to large trials, but no small trials, and firm conclusions about the presence or absence of small study effects might not be possible.

**Conclusions and implications**

An inspection of funnel plots and stratified analyses according to sample size accompanied by appropriate interaction tests should be considered as routine procedures in any meta-analysis, possibly accompanied by a regression test for funnel plot asymmetry and prediction of effects in large trials with meta-regression.<sup>47</sup> In the presence of asymmetry in funnel plots, systematic reviews 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.

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.

**Contributors:** 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 guarantors.

**Funding:** The study was funded by the Swiss National Science Foundation (grant No 4053-40-104762/3 and 3200-066378) through grants to PJ and SR and was part of the Swiss National Science Foundation's National Research Programme 53 on musculoskeletal health. SR's research fellowship was 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.

**Competing interests:** All authors have completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://www.icmje.org/coi_disclosure.pdf) (available on request from the corresponding author) and declare that all authors had: (1) No financial support for the submitted work from anyone other than their employer; (2) No financial relationships with commercial entities that might have an interest in the submitted work; (3) No spouses, partners, or children with relationships with commercial entities that might have an interest in the submitted work; (4) No Non-financial interests that may be relevant to the submitted work.

**Ethical approval:** Not required.

**Data sharing:** Data sharing: no additional data available.

- 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;323:101-5.
- 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;53:1119-29.
- 3 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
- 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;291:2457-65.
- 5 Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316:61-6.
- 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;149:219.
- 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;135:982-9.
- 9 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.
- 10 Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54:1046-55.
- 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;366:726-32.
- 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;61:991-6.
- 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;358:252-60.
- 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;146:580-90.
- 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;56:2267-77.
- 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;61:1633-41.
- 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;273:408-12.
- 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;36:847-57.
- 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;336:601-5.
- 24 Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. *Best Pract Res Clin Rheumatol* 2006;20:721-40.
- 25 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-88.
- 26 Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;18:2693-708.
- 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;21:1513-24.
- 28 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719-48.
- 29 Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum, 1988.
- 30 Palmer TM, Peters JL, Sutton AJ, Moreno SG. Contour-enhanced funnel plots for meta-analysis. *Stata J* 2008;8:242-54.
- 31 Eberle E, Ottlinger B. Clinically relevant change and clinically relevant difference in knee osteoarthritis. *Osteoarthritis Cartilage* 1999;7:502-3.
- 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;45:384-91.
- 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;29:131-8.
- 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;8:283-91.
- 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;290:3115-21.
- 38 Chodosh J, Morton SC, Mojica W, Maglione M, Suttrop MJ, Hilton L, et al. Meta-analysis: chronic disease self-management programs for older adults. *Ann Intern Med* 2005;143:427-38.
- 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;166:1899-906.
- 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;15:957-65.
- 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;11:125-38.
- 45 Manheimer E, Linde K, Lao L, Bouter LM, Berman BM. Meta-analysis: acupuncture for osteoarthritis of the knee. *Ann Intern Med* 2007;146:868-77.
- 46 Verhagen AP, Bierna-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;38:298-303.
- 48 Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;287:2801-4.
- 49 Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
- 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;25:3443-57.
- 52 Stuck A, Rubenstein L, Wieland D. Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity. *BMJ* 1998;316:469.

Accepted: 30 April 2010