Appendix S1. Deviations from the study protocol

Protocol method	Deviation from protocol method, with justification
We planned to include 40 systematic reviews (see sample size calculation in[1]).	Prior to undertaking any data analysis we increased the target sample size to 44 reviews as we had the resources to complete this additional work. *Type of deviation:* addition
We planned to retrieve unpublished review protocols from systematic reviewers.	We did not attempt to retrieve unpublished review protocols from systematic reviewers, as we assumed it would be challenging to verify that the review protocol was finalised prior to commencing the review *Type of deviation*: omission
We described the PBI as follows: "This index is based on the ordered effect estimates for each trial and the positioning (that is, rank) of the effect estimate selected within that order. A rank of 1 is assigned to the smallest effect estimate and a rank equal to the number of effect estimates is assigned to the largest effect estimate. Since the number of effect estimates varies across trials we rescale the ranks of the effect estimates to reflect their relative positioning (in ranking units) between the smallest and largest effect estimates. This is obtained by subtracting one from the rank of the selected effect estimate and dividing by the number of effect estimates minus one. The smallest effect estimate in a trial then has a location of zero and the largest effect estimate has a location of 1"[1].	We did not specifically state in the protocol that ranking was to be based on both the magnitude and *direction* of the effect estimates, although we intended this to be the case since we wished to test whether there was evidence that systematic reviewers selectively include effect estimates on the basis of both magnitude and direction of effect. Therefore, we clarified the terminology to indicate that a rank of 1 is assigned to the effect estimate that is least favourable to the intervention group (in terms of both magnitude and direction) and a rank equal to the number of effect estimates is assigned to the effect estimate that is most favourable to the intervention group. Further, we have clarified that after rescaling the ranks to reflect their relative positioning in ranking units, the effect estimate that is least favourable to the intervention group has a rank position of zero and the effect estimate that is most favourable to the intervention group has a rank position of 1. *Type of deviation*: clarification
We planned to calculate the PBI to assess the possible selection mechanism in which the smaller P-values of the effect estimates are chosen for inclusion.	We did not undertake this analysis, because reporting of an exact P-value in the trial report was only available for 20% (115/585) of all available effect estimates. Trialists were more likely to just state whether the P-value was less than or greater than 0.05. This infrequent reporting of exact P-values reduced the number of P-values for inclusion in the calculation of the PBI. *Type of deviation*: omission
We planned subgroup analyses to explore whether the	In addition to the pre-defined subgroup analyses, we undertook a post-hoc linear regression to explore whether the

Protocol method	Deviation from protocol method, with justification
availability of a systematic review protocol, and a core outcome measurement set for the clinical condition of the review, modified the PBI.	PBI was modified by the number of available effect estimates in a trial, because we wished to examine whether the rank position of the effect estimate selected for inclusion in the meta-analysis was dependent on the number of effect estimates available. *Type of deviation*: addition
We planned a sensitivity analysis to explore whether the PBI was modified when only trial effect estimates that were compatible with the eligibility criteria and decision rules in the methods sections of the review were included.	In addition to the pre-defined sensitivity analysis, we undertook the following post-hoc sensitivity analyses: 1. Converting all trial effect estimates to standardised mean differences (SMDs) allowed us to calculate the PBI in the circumstance where multiple effect estimates were available for the same outcome domain, but measured on different scales. However, there is not necessarily a one-to-one relationship between the rank positions of effect estimates based on the mean difference and SMD (because the SMD additionally depends on the pooled standard deviation). Therefore, in a post-hoc sensitivity analysis we calculated the PBI based on the rank positions of the mean difference for the subset of trial effect estimates that were measured on the same scale as the effect estimate included in the index meta-analysis. This allowed us to more accurately assess whether systematic reviewers had selectively included trial effect estimates based on the magnitude of the mean difference in raw measurement scale units (rather than in SMD units). 2. In some trial reports, only an effect estimate and its standard error or 95% CI were presented (that is, means and standard deviations per group were not available). In this circumstance, to include in a SMD meta-analysis, algebraic manipulation was required. Algebraic manipulation may be considered challenging by some systematic reviewers, so effect estimates requiring algebraic manipulation may not have been considered by reviewers in the set of effect estimates to potentially include in the meta-analysis. For the primary calculation of the PBI, we excluded trial effect estimates that required algebraic manipulation; however, we undertook a post-hoc sensitivity analysis to explore whether the PBI was modified when we included these trial effect estimates. 3. For meta-analyses comparing an active intervention with a placebo/no intervention control, our hypothesis was that if selective inclusion occurred, it would occur in the direction of selecting the trial effect estimate that was most favourable to the active intervention. For meta-analyses comparing two active interventions, we determined which intervention was the newer one from the text of the review, and ranked trial effect estimates based on their favourability to the newer intervention. We performed a post-hoc sensitivity analysis excluding meta-analyses of head-to-head comparisons to examine the impact on the PBI. *Type of deviation (applies to all three)*: addition
We planned to calculate the PBI at the meta-analysis level (we stated "The PBI described above will also be	We did not calculate the PBI at the meta-analysis level, because it only provides information about the rank position of the index meta-analytic effect, and not information on the potential impact of selective inclusion on the

Protocol method	Deviation from protocol method, with justification
used to compare the index meta-analytic effect estimates with all possible meta-analytic effects"[1].)	magnitude of the meta-analytic effect; the latter of which we consider is of most interest. *Type of deviation*: omission
We planned to investigate the impact of any potential selective inclusion of trial effect estimates on the meta-analytic effects as follows: "For each meta-analysis, all possible meta-analytic effects will be calculated from all combinations of available RCT effect estimates. The meta-analysis model used to combine the estimates (either fixed or random effects) will be the model that was used in the systematic review… For each meta-analysis, the difference between the index meta-analytic effect estimate and the median of all possible meta-analytic effect estimates will be calculated. These differences will be standardised (by dividing by the pooled baseline standard deviation of the outcome) and meta-analysed using a random effects model across reviews. The meta-analytic weights will be based on the standardised standard error of the median meta-analytic estimates, and between RCT variability estimated using DerSimonian and Laird's method of moments estimator[2]. Note that this approach ignores the correlation between the meta-analytic effects within meta-analysis, arising from correlated RCT effects"[1].	We modified our approach to investigating the impact of any potential selective inclusion of trial effect estimates on the meta-analytic effects in the following ways: 1. Instead of using the model that was used in the systematic review to combine the effect estimates, we used the random-effects model because this was the model used in the majority of index meta-analyses (n = 24/31). 2. We did not standardize the difference between the index meta-analytic effect estimate and the median of all possible meta-analytic effect estimates because we meta-analysed SMDs, not mean differences. 3. We calculated non-parametric statistics to describe the distribution of the differences between the index meta-analytic SMD and the median of all possible meta-analytic SMDs. 4. When we meta-analysed the differences (using a random-effects model, with between trial variability estimated using DerSimonian and Laird's method of moments estimator[2]), we weighted each difference by the standard error of the index meta-analytic SMD, instead of the median meta-analytic SMD. *Type of deviation (applies to all four)*: modification

References

1. Page MJ, McKenzie JE, Green SE, Forbes AB. An empirical investigation of the potential impact of selective inclusion of results in systematic reviews of interventions: study protocol. Systematic Reviews. 2013;2:21.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88.

Appendix S2. Search strategies

Cochrane Database of Systematic Reviews (Wiley Interscience (Online) interface) search strategy

1. degenerative arthritis[tw]
2. "Arthritis, Rheumatoid"[MeSH]
3. rheumatoid arthritis[tw]
4. rheumatism[tw]
5. "Arthritis, Juvenile Rheumatoid"[MeSH]
6. caplan's syndrome[tw]
7. felty's syndrome[tw]
8. rheumatoid[tw]
9. ankylosing spondylitis[tw]
10. arthrosis[tw]
11. sjogren*[tw]
12. "Osteoarthritis"[MeSH]
13. Osteoarthr*[tw]
14. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #13 or #14
15. "Depression"[MeSH]
16. "Anxiety"[MeSH]
17. "Anxiety Disorders"[MeSH]
18. depress*[tw]
19. dysthymi*[tw]
20. anxiety[tw] OR anxious[tw]
21. #15 or #16 or #17 or #18 or #19 or #20
22. #14 or #21
23. #22 Limits: English, Publication Date from 2010/01/01 to 2012/01/31

MEDLINE (PubMed interface) search strategy

1. degenerative arthritis[tw]
2. "Arthritis, Rheumatoid"[MeSH]
3. rheumatoid arthritis[tw]
4. rheumatism[tw]
5. "Arthritis, Juvenile Rheumatoid"[MeSH]
6. caplan's syndrome[tw]
7. felty's syndrome[tw]
8. rheumatoid[tw]
9. ankylosing spondylitis[tw]
10. arthrosis[tw]
11. sjogren*[tw]
12. "Osteoarthritis"[MeSH]
13. osteoarthr*[tw]
14. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #13 or #14
15. "Depression"[MeSH]
16. "Anxiety"[MeSH]
17. "Anxiety Disorders"[MeSH]
18. depress*[tw]
19. dysthymi*[tw]

20. anxiety[tw] OR anxious[tw]
21. #15 or #16 or #17 or #18 or #19 or #20
22. #14 or #21
23. cochrane database syst rev[ta] or search[tw] or meta-analysis[pt] or medline[tw] or systematic review[tw]
24. (meta-analysis[pt] or meta-analysis[tw] or meta-analysis[mesh] or review[pt] or search*[tw]) and methods[ab]
25. #23 or #24
26. #22 and #25
27. #26 Limits: English, Publication Date from 2010/01/01 to 2012/01/31

Appendix S3. Methods used to extract outcome data from trial reports

There is variation in how systematic reviewers specify meta-analysis outcomes. Some systematic reviewers define broad outcomes such as "depression", while others define more specific outcomes, such as "Montgomery-Asberg Depression Rating Scale (MADRS) score at 12 weeks". We adhered to these specifications when deciding which data to extract. For the outcome "depression", we extracted all depression scales, time points and analyses (such as both unadjusted and covariate adjusted analyses) as long as they were compatible with the pre-defined eligibility criteria and decision rules to select effect estimates. For the outcome "MADRS score at 12 weeks", we only extracted data for the MADRS scale at 12 weeks, though extracted results arising from multiple analyses (such as both unadjusted and covariate adjusted analyses) if these were compatible with the pre-defined eligibility criteria and decision rules to select effect estimates.

When multi-arm trials are encountered and each group is eligible for inclusion in a meta-analysis, systematic reviewers need to use a method which avoids multiple counting of participants. For continuous outcomes, systematic reviewers may choose to (1) include data from only one of the active intervention arms and the control arm, (2) calculate the mean effect of the two active intervention arms and compare this to the control arm, or (3) include data from each active intervention arm as separate comparisons in the meta-analysis by dividing the sample size of the control arm by the number of comparisons[1,2]. If systematic reviewers pre-defined a method to deal with multi-arm trials, we followed that method when extracting data. If systematic reviewers did not pre-define a method to deal with multi-arm trials, and:

(1) selected one of the active intervention arms to include in the meta-analysis, we extracted the data required to calculate effect estimates for two comparisons: (a) active intervention A versus control, and (b) active intervention B versus control;
(2) calculated the mean effect of the two active intervention arms, we extracted the data required to calculate effect estimates for three comparisons: (a) active intervention A versus control, (b) active intervention B versus control, and (c) mean of active intervention A and B versus control;
(3) included multiple comparisons in the meta-analysis by dividing the control group in half, we extracted the data required to calculate effect estimates for two comparisons:

(a) active intervention A versus control, and (b) active intervention B versus control, where for both comparisons the control group sample size was halved.

The three methods above were used when dealing with three-arm trials. We extended these methods when there were more than three arms.

All trial effect estimates included in mean difference meta-analyses must be in units of one particular scale, though estimates can comprise a mixture of final values and change from baseline values. In contrast, the measurement scale units of trial effect estimates included in standardised mean difference (SMD) meta-analyses can vary, though it is recommended that all estimates are final values, or change from baseline values, not a mixture[3]. Therefore, for mean difference meta-analyses we only extracted data for the particular scale included by the systematic reviewer, but extracted final and change from baseline values when available. For SMD meta-analyses that included final values, we extracted final values only for any relevant measurement scale (and vice versa for SMD meta-analyses that included change from baseline values).

Example 1: Index meta-analysis was specified as "mean difference in Beck Depression Inventory scores", and the systematic reviewers did not pre-define any eligibility criteria or decision rules to select effect estimates. We extracted all results for the Beck Depression Inventory (for example, at all time points, final and change from baseline values, unadjusted and covariate adjusted analyses), but no results for any other depression scales.

Example 2: Index meta-analysis was specified as "SMD in pain scores at 12 weeks". The systematic reviewers only included change from baseline values, and did not pre-define any eligibility criteria or decision rules to select effect estimates. We extracted from each trial all results for pain (for example, based on any pain scale, analyses undertaken on intention-to-treat and per-protocol samples), but only if the values were change from baseline to 12 weeks (that is, no final values, and no other time points).

References

1. Hasselblad V. Meta-analysis of multitreatment studies. Medical Decision Making. 1998;18(1):37-43.

2. Higgins JPT, Deeks JJ. Chapter 7: Selecting studies and collecting data. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.
3. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Appendix S4: Worked example of the Potential Bias Index

Consider a review consisting of 10 trials. The number of effect estimates for each trial is provided as well as the rank of the effect estimate chosen among those available for each trial. Suppose the data are the following:

Trial	Number of effect estimates (n)	Rank of chosen effect estimate (X)	Location of chosen effect estimate (Y)
1	3	2	0.50
2	7	6	0.83
3	4	3	0.67
4	6	5	0.80
5	4	2	0.33
6	5	4	0.75
7	6	6	1
8	2	1	0
9	4	4	1
10	5	4	0.75

For Trial #1, there were 3 effect estimates and the rank of the chosen effect estimate was 2, that is the middle or median value, and its location is therefore halfway between the lowest and highest rank. For Trial #2 there were 7 effect estimates and the chosen estimate had rank 6. There are a total of 6 units of rank between 1 and 7 (that is, 1 to 2, 2 to 3, 3 to 4, 4 to 5, 5 to 6 and 6 to 7) and the chosen rank of 6 is therefore 5/6ths = 83% of the distance between lowest and highest rank. In general, the rank location Y is calculated as (X-1)/(n-1).

The statistic PBI = $PBI = \dfrac{3\times 0.50 + 7\times 0.83 + \cdots + 4\times 1 + 5\times 0.75}{3+7+4+\cdots+4+5} = \dfrac{33.61}{46} = 0.73$

Therefore on average the effect estimates chosen were 73% of the distance between the smallest and largest rank, that is, approximately halfway between the middle rank and the maximum.

The standard error of the PBI can be calculated to be 0.118, and therefore the Z-statistic equals (0.73-0.50)/0.118 = 1.97, with a two-tailed p-value of 0.049. This indicates some evidence that the effect estimate selection is systematically higher than that expected by random selection.

A 95% confidence interval for the PBI is obtained from 1000 bootstrap replications as 0.58 to 0.85.

Appendix S5: Sensitivity analyses

A series of sensitivity analyses were undertaken to investigate whether the Potential Bias Index (PBI) was robust to certain assumptions. For systematic reviews without protocols, we could not determine whether the eligibility criteria and decision rules to select results in the methods section of the review were developed prior to or while undertaking the review. Therefore, in these reviews, our primary calculation of the PBI was based on the set of trial effect estimates that were compatible with the assumption of no pre-specified eligibility criteria or decision rules. However, we also performed a pre-specified sensitivity analysis where only trial effect estimates that were compatible with the eligibility criteria and decision rules in the methods sections of the review were included, so as to examine if this affected the PBI.

Converting all trial effect estimates to SMDs allowed us to calculate the PBI in the circumstance where multiple effect estimates were available for the same outcome domain, but measured on different scales. However, there is not necessarily a one-to-one relationship between the rank positions of effect estimates based on the mean difference and SMD (because the SMD additionally depends on the pooled standard deviation). Therefore, in a post-hoc sensitivity analysis we calculated the PBI based on the rank positions of the mean difference for the subset of trial effect estimates that were measured on the same scale as the effect estimate included in the index meta-analysis. This allowed us to more accurately assess whether systematic reviewers had selectively included trial effect estimates based on the magnitude of the mean difference in raw measurement scale units.

In some trial reports, only an effect estimate and its standard error or 95% CI were presented (that is, means and standard deviations per group were not available). In this circumstance, to include in a SMD meta-analysis, algebraic manipulation was required. Algebraic manipulation may be considered challenging by some systematic reviewers, so effect estimates requiring algebraic manipulation may not have been considered by reviewers in the set of effect estimates to potentially include in the meta-analysis. For the primary calculation of the PBI, we excluded trial effect estimates that required algebraic manipulation; however, we undertook a post-hoc sensitivity analysis to explore whether the PBI was modified when we included these trial effect estimates.

For meta-analyses comparing an active intervention with a placebo/no intervention control, our hypothesis was that if selective inclusion occurred, it would occur in the direction of selecting the trial effect estimate that was most favourable to the active intervention. For meta-analyses comparing two active interventions, we determined which intervention was the newer one from the text of the review, and ranked trial effect estimates based on their favorability to the newer intervention. We performed a post-hoc sensitivity analysis excluding meta-analyses of head-to-head comparisons to examine the impact on the PBI.

Finally, in our primary analysis of investigating the impact of any potential selective inclusion of trial effect estimates on meta-analytic SMDs, we used the random-effects meta-analysis model to pool effect estimates when calculating the distribution of possible meta-analytic SMDs. We performed a pre-specified sensitivity analysis to explore whether our primary analysis was modified when the distribution of meta-analytic SMDs were calculated using a fixed-effect model.

Supplementary Table S1: Number (%) of trials with different types of multiplicity of effect estimates (n=250)

Type of multiplicity	n (%)*
Any	118 (47)
Measurement scales	71 (28)
Intervention/control groups	27 (11)‡
Time points	30 (12)
Final and change from baseline values	0
Analyses undertaken on multiple samples (for example, intention-to-treat and per-protocol)	23 (9)
Unadjusted and covariate adjusted analyses	4 (2)
Period and paired analyses in cross-over trials	4 (2)

* Percentages do not sum to 100 because some trials had multiple sources of multiplicity of effect estimates.
‡ The unit of analysis was trial comparisons. Therefore, multi-arm trials that had been included in the meta-analysis through multiple comparisons were not counted in this estimate as having multiple intervention/control groups.

Supplementary Table S2: Number (%) of review protocols and reviews reporting eligibility criteria and decision rules to select effect estimates to include in meta-analyses

Type of eligibility criterion and decision rule	Review protocols n (%) *n = 12*	Reviews n (%) *n = 31*
Total		
At least one eligibility criterion	12 (100)	31 (100)
At least one decision rule	8 (67)	29 (94)
Measurement scales		
Eligibility criteria	11 (92)	21 (68)
Decision rule	2 (17)	12 (39)
Intervention/control groups		
Eligibility criteria	12 (100)	29 (94)
Decision rule	2 (17)	19 (61)
Time points		
Eligibility criteria	10 (83)	27 (87)
Decision rule	7 (58)	20 (65)
Analyses		
Eligibility criteria for any type of analysis	8 (67)	23 (74)
Decision rule for final versus change from baseline values	4 (33)	11 (35)
Decision rule for analyses undertaken on multiple samples (for example, intention-to-treat versus per-protocol)	7 (58)	13 (42)
Decision rule for unadjusted versus covariate adjusted analyses	0 (0)	1 (3)
Decision rule for period versus paired analyses in cross-over trials	7 (58)	9 (29)
Other decision rule	0 (0)	0 (0)

Supplementary Table S3: Sensitivity analyses for the PBI

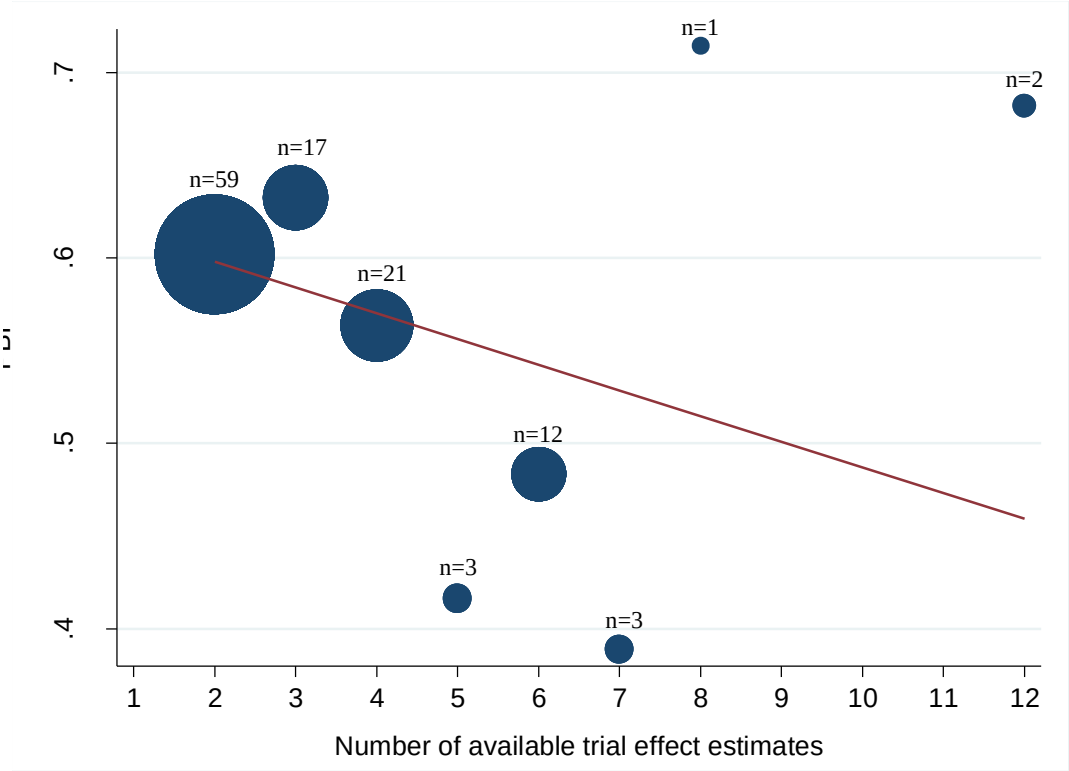
PBI analyses	Number of trials	Number of meta-analyses	PBI (95% CI*)
Primary analysis	118	31	0.57 (0.50, 0.63)
Sensitivity analyses			
1. Inclusion of the set of trial effect estimates that were compatible with the eligibility criteria and decision rules in the methods section of the review	88	27	0.58 (0.50, 0.66)
2. Calculation based on the rank positions of the *mean difference* for the subset of trial effect estimates that were measured on the same scale as the effect estimate included in the index meta-analysis	74	26	0.55 (0.44, 0.65)
3. Inclusion of trial effect estimates that required algebraic manipulation	121	31	0.55 (0.47, 0.62)
4. Exclusion of trial effect estimates in meta-analyses of head-to-head comparisons	114	29	0.57 (0.50, 0.65)

*Percentile-based confidence intervals for the PBI were constructed using bootstrap methods by resampling individual trials 2,000 times[1].
**The confidence limits and p-value for the difference in PBI between subgroups was constructed using bootstrap methods from 2,000 replicates.

References

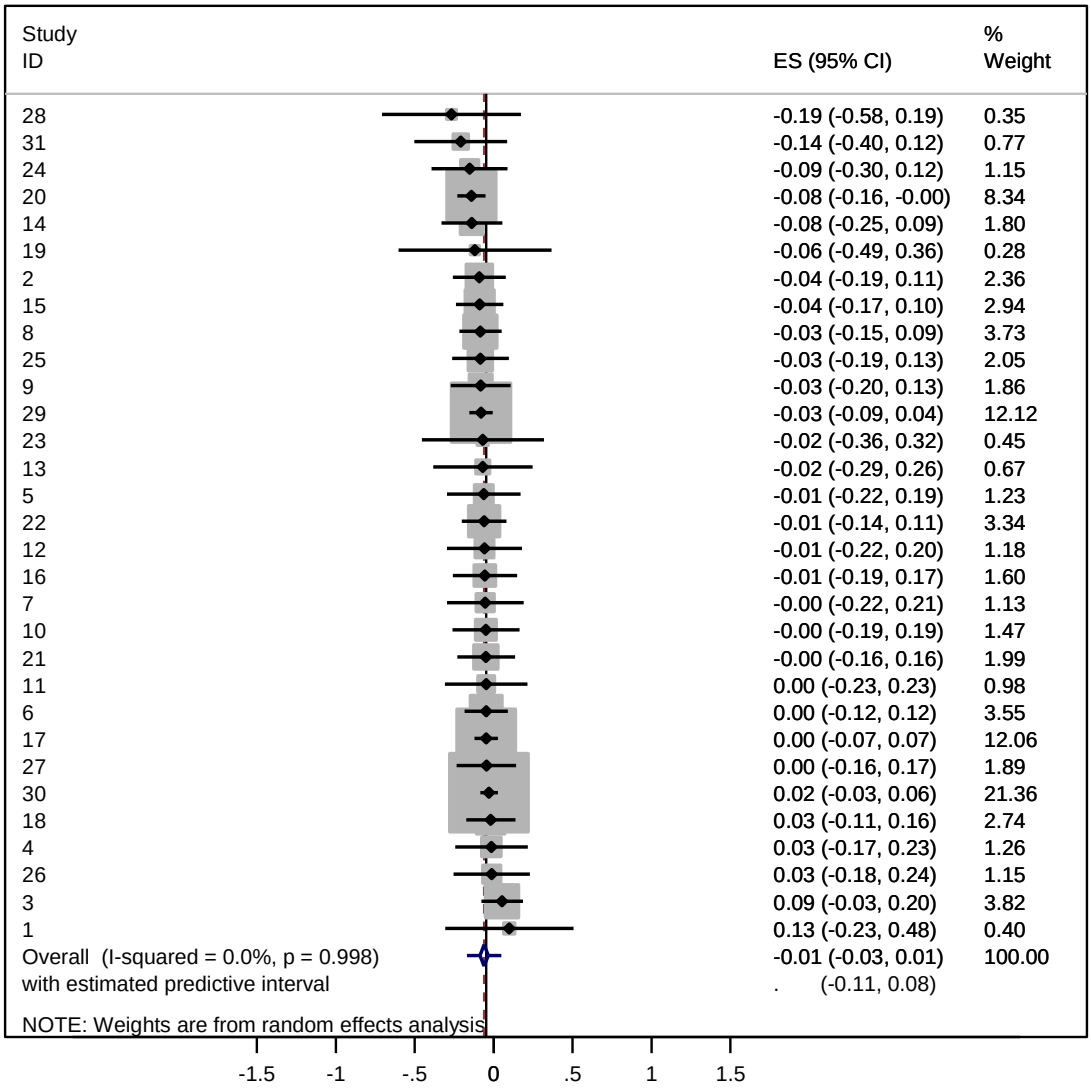
1. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.

Supplementary figures



Supplementary Figure S1. Scatterplot of the relationship between the PBI and the number of trial effect estimates available for inclusion in a meta-analysis. The observed PBI values are depicted by blue dots, the sizes of which are proportional to the number (n) of trials available. The red line represents the fitted regression line, weighted by the number of observations available per data point.

When all possible meta-analytic SMDs were calculated using a fixed-effect model, the median of the differences between the index meta-analysis and the median of all its possible meta-analytic SMDs was -0.01 standard deviation units (-0.04 to 0; -0.19 to 0.13). Meta-analysing these differences using a random-effects model yielded a pooled difference of -0.01 standard deviation units (95% CI -0.03 to 0.01; 95% prediction interval -0.11 to 0.08; I^2 = 0%) (Supplementary Figure 1).



Supplementary Figure S2. Meta-analysis of differences between the index meta-analytic SMD and median of all its possible meta-analytic SMDs (each calculated using the fixed-effect model). Differences less than zero indicate that the index meta-analysis SMD is more favourable to the intervention compared with the median meta-analytic SMD.