BMJ Open

Investigation of bias in meta-analyses due to selective inclusion of trial effect estimates: empirical study

Matthew J Page,[1,2] Andrew Forbes,[3] Marisa Chau,[4] Sally E Green,[1] Joanne E McKenzie[1]

To cite: Page MJ, Forbes A, Chau M, *et al*. Investigation of bias in meta-analyses due to selective inclusion of trial effect estimates: empirical study. *BMJ Open* 2016;**6**: e011863. doi:10.1136/bmjopen-2016-011863

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-011863).

Received 10 March 2016
Accepted 6 April 2016

[1]Australasian Cochrane Centre, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[2]School of Social and Community Medicine, University of Bristol, Bristol, UK
[3]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[4]National Trauma Research Institute, Central Clinical School, Monash University, Melbourne, Victoria, Australia

Correspondence to
Dr Joanne E McKenzie;
joanne.mckenzie@monash.edu

ABSTRACT

Objective: To explore whether systematic reviewers selectively include trial effect estimates in meta-analyses when multiple are available, and what impact this may have on meta-analytic effects.

Design: Cross-sectional study.

Data sources: We randomly selected systematic reviews of interventions from 2 clinical specialties published between January 2010 and 2012. The first presented meta-analysis of a continuous outcome in each review was selected (index meta-analysis), and all trial effect estimates that were eligible for inclusion in the meta-analysis (eg, from multiple scales or time points) were extracted from trial reports.

Analysis: We calculated a statistic (the Potential Bias Index (PBI)) to quantify and test for evidence of selective inclusion. The PBI ranges from 0 to 1; values above or below 0.5 are suggestive of selective inclusion of effect estimates more or less favourable to the intervention, respectively. The impact of any potential selective inclusion was investigated by comparing the index meta-analytic standardised mean difference (SMD) to the median of a randomly constructed distribution of meta-analytic SMDs (representing the meta-analytic SMD expected when there is no selective inclusion).

Results: 31 reviews (250 trials) were included. The estimated PBI was 0.57 (95% CI 0.50 to 0.63), suggesting that trial effect estimates that were more favourable to the intervention were included in meta-analyses slightly more often than expected under a process consistent with random selection; however, the 95% CI included the null hypothesis of no selective inclusion. Any potential selective inclusion did not have an important impact on the meta-analytic effects.

Conclusion: There was no clear evidence that selective inclusion of trial effect estimates occurred in this sample of meta-analyses. Further research on selective inclusion in other clinical specialties is needed. To enable readers to assess the risk of selective inclusion bias, we recommend that systematic reviewers report the methods used to select effect estimates to include in meta-analyses.

Strengths and limitations of this study

- We predefined our methods in a published protocol, used validated search strategies to identify eligible reviews and included a randomly selected sample of reviews.
- Our estimates of the Potential Bias Index and the impact of potential selective inclusion on the meta-analytic effects were robust to several assumptions and statistical approaches.
- We only examined two clinical specialties (with two-thirds of the index meta-analyses focused on pain or depression), so our results may have limited generalisability to systematic reviews of other conditions.
- Our findings may have been influenced by incomplete reporting by systematic reviewers, as we used the methods reported in the review protocols to select trial effect estimates; however, other selection methods may have been assumed but not documented at the protocol stage.
- We only investigated selective inclusion of completely reported effect estimates; the set of effect estimates that were actually considered by the systematic reviewers may have been larger, encompassing both reported and unreported estimates (in the case of the latter, obtained directly from the triallists).

INTRODUCTION

Systematic reviews and meta-analyses of healthcare intervention trials can contribute to improved patient care by providing valid and reliable information for healthcare decision-making.[1] However, the validity of systematic review findings can be compromised by challenges with undertaking meta-analysis. One challenge arises when there are multiple effect estimates in a trial report which could be included in a particular meta-analysis. For example, systematic reviewers performing a meta-analysis of depression scores may encounter mean differences for two depression scales at three time points.[2 3]

In such instances, it is preferable that the selection of effect estimates to include in the meta-analysis is based on predefined clinical or methodological rationale (eg, selecting the mean difference of the depression scale with the best measurement properties). However, in some cases, the selection may be based on the nature or direction of the estimates themselves;[4] this is analogous to when triallists perform multiple analyses of the same outcome, yet only report that which is most favourable. This data-driven selection by systematic reviewers is known as selective inclusion, and has the potential to bias meta-analytic effects.

Concerns about the potential for selective inclusion of trial effect estimates have been previously raised. In a study of 19 Cochrane reviews including 83 trials, a third of the trial reports had data for multiple measurement scales, intervention/control groups or time points that could have been included in a particular meta-analysis. There was potential for large and clinically important variability in the resulting meta-analytic effects depending on which trial effect estimates were included.[3] However, there has been no empirical investigation of whether systematic reviewers selectively include trial effect estimates in meta-analyses when multiple estimates are available, or the potential impact of selective inclusion on meta-analytic effects.[5] Therefore, it is unclear to what extent users of systematic reviews should be concerned about selective inclusion.

We investigated whether selective inclusion of trial effect estimates occurred in a sample of meta-analyses, and what impact this might have on meta-analytic effects.

METHODS

The study protocol is published elsewhere.[6] An overview of the methods is provided here, and deviations from the planned methods are presented in online supplementary appendix S1.

Eligibility criteria, searching and selection methods

We included systematic reviews meeting the following criteria: Cochrane or non-Cochrane systematic review published between January 2010 and January 2012; focusing on any intervention for rheumatoid arthritis (RA), osteoarthritis (OA), depressive disorders or anxiety disorders; written in English; reporting references of all included trials; and reporting, for at least one meta-analysis of a continuous outcome, the summary statistics (eg, means, SDs, sample sizes) or effect estimate and its precision for each included trial, and the meta-analytic effect estimate and its precision. We excluded systematic reviews which reported no meta-analyses of continuous outcomes, included non-randomised studies in all meta-analyses of continuous outcomes or used non-standard meta-analytical methods (eg, Bayesian, multiple treatments or individual patient data meta-analyses).

We selected the conditions RA/OA and depressive/anxiety disorders because we wanted to explore whether the existence of a core outcome measurement set for RA and OA trials[7 8] impacted on selective inclusion. Core outcome measurement sets are lists of measurement scales recommended for use in trials and systematic reviews of a particular health condition, and are designed to increase consistency in scale selection.[9] We therefore expected selective inclusion to be less common in reviews with a core outcome measurement set (RA and OA reviews) compared with reviews without a core outcome measurement set (depressive/anxiety disorders reviews). Depressive/anxiety disorders reviews were selected because of our familiarity with the measurement scales typically used in this specialty. Further, we only focused on continuous outcomes because there is greater scope for multiplicity of effect estimates for continuous rather than dichotomous outcomes in these clinical specialties (eg, arising from multiple measurement scales, adjusted vs unadjusted means, subscale scores).

We searched the Cochrane Database of Systematic Reviews and MEDLINE (PubMed interface), limiting the publication date from 1 January 2010 to 31 January 2012. Search strategies are provided in online supplementary appendix S2. One author (MJP) screened all titles and abstracts and retrieved the full-text reports of potentially eligible records. The citations of full-text reports were randomly sorted (using the random number generator in Microsoft Excel), and one author (MJP) then screened the full-text reports until at least 10 of each type of review (Cochrane RA/OA, non-Cochrane RA/OA, Cochrane depressive/anxiety disorder and non-Cochrane depressive/anxiety disorder) meeting the eligibility criteria were identified. Screening ceased once 44 eligible reviews were included. Any difficulties in determining eligibility were resolved by discussion with a second author (JEM). The first presented meta-analysis of an effect measure for a continuous outcome (either the mean difference or standardised mean difference (SMD)) in each review was selected for inclusion (which we denote the ‘index meta-analysis’). The index meta-analysis may have been identified from the abstract, summary of findings table or results section of the review, depending on where the meta-analytic effect estimate was first reported in the publication.

Data extraction

Published protocols of included reviews and reports of trials included in the index meta-analyses were retrieved. One author (MJP) extracted data from all review protocols, reviews and trial reports using a standardised pilot tested form. A second author (MC) independently extracted data from a random sample of 14 reviews, including the corresponding review protocols and included trials, to assess accuracy of the extraction. Discrepancies between the two sets of extracted data were resolved via discussion.

From each review protocol and review, we extracted descriptions of:

- The clinical condition, types of interventions and comparisons, outcomes of interest and funding source of the review.
- Any eligibility criteria to select effect estimates to include in the index meta-analysis. Eligibility criteria comprised lists of measurement scales, intervention/control groups, time points and analyses that were eligible for inclusion (eg, a statement that eligible depression scales included only the Hamilton Rating Scale for Depression (HRSD), the Montgomery-Asberg Depression Rating Scale (MADRS) and the Beck Depression Inventory (BDI)).
- Any decision rules to select effect estimates to include in the index meta-analysis. Decision rules comprised strategies to either select one effect estimate or combine effect estimates when multiple were available (eg, a statement that the HRSD would be selected over the BDI if data for both were available, or that data from two active intervention arms in multiarm trials would be combined for inclusion in a pairwise meta-analysis).

We extracted from the trial reports outcome data that could potentially be included in the index meta-analyses, according to the eligibility criteria and decision rules that were prespecified in the review protocol, and how the outcome was specified in the review (eg, 'depression score' or 'HRSD depression score at the end of treatment'). For reviews without a publicly available protocol, we assumed that no eligibility criteria and decision rules were prespecified, even if some were reported in the published review ('worst-case scenario' assumption), and extracted all outcome data based on how the outcome was specified in the review. For example, if the index meta-analysis was specified as 'mean difference in BDI scores' and there was no review protocol, we extracted from trials all results for the BDI (eg, at all time points, final and change from baseline values, unadjusted and covariate-adjusted analyses, regardless of whether decision rules for these measures/analyses were stated in the published review), but no results for any other depression scales. 'Outcome data' included summary statistics (eg, means, SDs, sample sizes) or effect estimate (eg, mean difference) and some measure of precision (eg, SE, 95% CI); or both if available. The predefined methods which we used to extract trial outcome data are summarised in online supplementary appendix S3. We only extracted trial outcome data that were completely reported, defined as reporting sufficient data for inclusion in a meta-analysis (ie, reporting of an effect estimate and a measure of precision, or summary statistics that enable calculation of these).[10] Unpublished data (eg, missing SDs) were not sought from triallists.

Statistical analyses

All analyses were conducted using Stata V.12 (Stata Corp. Stata Statistical Software: Release 12 [program]. College Station, Texas, USA: StataCorp LP, 2011). Index meta-analyses that included no trials with multiplicity of effect estimates were excluded from all analyses because they could not contribute to the assessment of selective inclusion. Characteristics of the index meta-analyses were summarised using frequencies and percentages for binary outcomes, and medians (with IQRs and ranges) for continuous outcomes. The frequencies and percentages of review protocols and reviews reporting eligibility criteria and decision rules to select trial effect estimates were calculated.

We calculated a statistic (the Potential Bias Index (PBI)) to quantify and test for evidence of selective inclusion. Mathematical details of the construction of the PBI are available in the study protocol,[6] and a worked example is provided in online supplementary appendix S4. In brief, this index is based on ordering effect estimates in each trial based on their magnitude and direction of effect, and determining the position within that order of the effect estimate selected for inclusion in the index meta-analyses. The PBI is the weighted average rank position of the selected effect estimates, where the weights are the inverse of the number of effect estimates available per trial. This weighting system therefore attributes greater priority to the rank positions of effect estimates where there are a larger number of effect estimates to choose from. To enable ranking of effect estimates that were measured using different measurement scales, all effect estimates were expressed in terms of SD units by dividing the mean difference on the raw measurement scale by the pooled SD (SMDs).[11] Further, the direction of intervention effects was standardised, so that larger negative values represented effects that are more favourable to the intervention.

The PBI ranges from 0 to 1. When the effect estimate that is most favourable to the intervention in each of the trials is always selected for inclusion, the PBI will have the value 1. Conversely, the PBI will have the value of 0 when the effect estimate that is least favourable to the intervention is always selected. Several methods for selecting effect estimates are acceptable in terms of not introducing bias, including (1) randomly selecting effect estimates, (2) selecting effect estimates based on some clinical or methodological rationale or (3) selecting the median effect estimate. If selection methods 2 and 3 are employed across the trials, we expect that the distribution of selected effect estimates would be consistent with what we would observe under purely random selection, so on average, the selected effect estimates would be at the middle rank position and the PBI would take the value of 0.5. A PBI of 0.5 therefore suggests that there is no selective inclusion of the most or least favourable effect estimates. We constructed a statistical test based on the PBI to test whether the observed selection of effect estimates is consistent with randomness of selection.[6] Confidence limits (95%) for the PBI were obtained by bootstrap resampling.[12]

We also investigated the impact of any potential selective inclusion of trial effect estimates on the magnitude of the resulting meta-analytic SMDs. For each meta-analysis, we calculated all possible meta-analytic SMDs from all combinations of available trial effect estimates. When the number of possible combinations was prohibitively large to calculate all combinations (ie, >30 000), we generated a random sampling distribution of 5000 meta-analytic SMDs. Each meta-analytic SMD was created by randomly selecting (with equal probability) an effect estimate for inclusion from each trial comparison, and meta-analysing the chosen effects. For each distribution of generated meta-analyses, we calculated (1) the percentile rank of the index meta-analytic SMD; (2) the median of all possible meta-analytic SMDs, which represented the median of a distribution where trial effect estimates were not selectively included and (3) the difference between the index meta-analytic SMD and the median meta-analytic SMD. When the difference between the index and median meta-analytic SMD is minimal, any potential selective inclusion had limited impact on the meta-analytic effect. Non-parametric statistics were used to describe these differences. We also meta-analysed these differences using a random-effects meta-analysis model, with the meta-analytic weights based on the variance of the index meta-analytic SMD estimate and the between-trial variability estimated using DerSimonian and Laird's[13] method of moments estimator.

We performed prespecified subgroup analyses to explore whether the availability of a systematic review protocol, and a core outcome measurement set for the clinical condition of the review, modified the PBI. We fitted a post hoc linear regression model to explore whether the PBI was modified by the number of available effect estimates in a trial. We undertook a series of prespecified and post hoc sensitivity analyses, the details of which are provided in online supplementary appendix S5.

RESULTS

Search results and data extraction discrepancies

Searching yielded a total of 2590 title and abstract records which were screened. A full-text article was retrieved for 264 records. Of these, 44 systematic reviews met the eligibility criteria, but only 31 included trials with multiplicity of effect estimates and thus could contribute to the assessment of selective inclusion (figure 1).

Of the subset of 14 of 44 reviews where data were extracted by two authors independently, 8 (26%) were in the final included sample of 31 reviews. Comparison of the double data extraction identified no errors in the classification of review text as an 'eligibility criterion' or 'decision rule', and the summary statistics and effect estimates extracted were consistent except for in one review where the second reviewer missed a decision rule in the review protocol, which resulted in the extraction of three ineligible trial effect estimates. Hence, no modifications to the first reviewers' extraction were required.

Characteristics of included meta-analyses

Of the 31 index meta-analyses, 4 were from Cochrane RA or OA reviews, 11 were from non-Cochrane RA or OA reviews, 9 were from Cochrane depressive or anxiety disorder reviews, and 7 were from non-Cochrane depressive or anxiety disorder reviews. Twelve (39%) meta-analyses were in Cochrane reviews with a protocol published between 2006 and 2011 (table 1). The most common outcome domains analysed in the index meta-analyses were depression (12 meta-analyses) or pain (8 meta-analyses). Nine other outcome domains were analysed in at least one meta-analysis. There was an approximately equal distribution of index meta-analyses that were labelled primary versus non-primary outcomes. The majority of index meta-analyses pooled SMDs, were fitted using a random-effects model, examined a placebo/no intervention controlled comparison and investigated the efficacy/effectiveness of a non-pharmacological intervention. Two (6%) reviews were funded by the pharmaceutical industry (table 1). There were a total of 250 trials included in the 31 index meta-analyses, with a median of 6 trials (IQR 3–10; range 2–28) per meta-analysis.

Multiplicity of effect estimates in trial reports and methods to select effect estimates

In 118 (47%) trials, there were multiple effect estimates that were eligible for inclusion in a particular meta-analysis. In these trials with multiplicity, there was a median of 3 (IQR 2–4; range 2–12) eligible effect estimates per trial. Details of the types of multiplicity (eg, multiple scales, time points) are available in online supplementary table S1. Descriptive statistics at the meta-analysis level were reflective of those at the trial level; per meta-analysis, a median of 50% of trials had multiplicity (IQR 22–67%; range 14–100%).

The types of methods to select effect estimates to include in meta-analyses that were documented in the review protocols and reviews varied considerably (see online supplementary table S2). For example, of 12 review protocols, only 2 (17%) included a prespecified decision rule to select measurement scales while 7 (58%) included a prespecified decision rule to select time points.

Evidence of selective inclusion of trial effect estimates

The estimated PBI was 0.57 (95% CI 0.50 to 0.63; two-tailed p value of 0.10). This suggests that trial effect estimates that were more favourable to the intervention were included in meta-analyses slightly more often than expected under a process consistent with random selection; however, the 95% CI included the null hypothesis of no selective inclusion. The PBI was not modified by the availability of a systematic review protocol or a core outcome measurement set for the clinical condition of

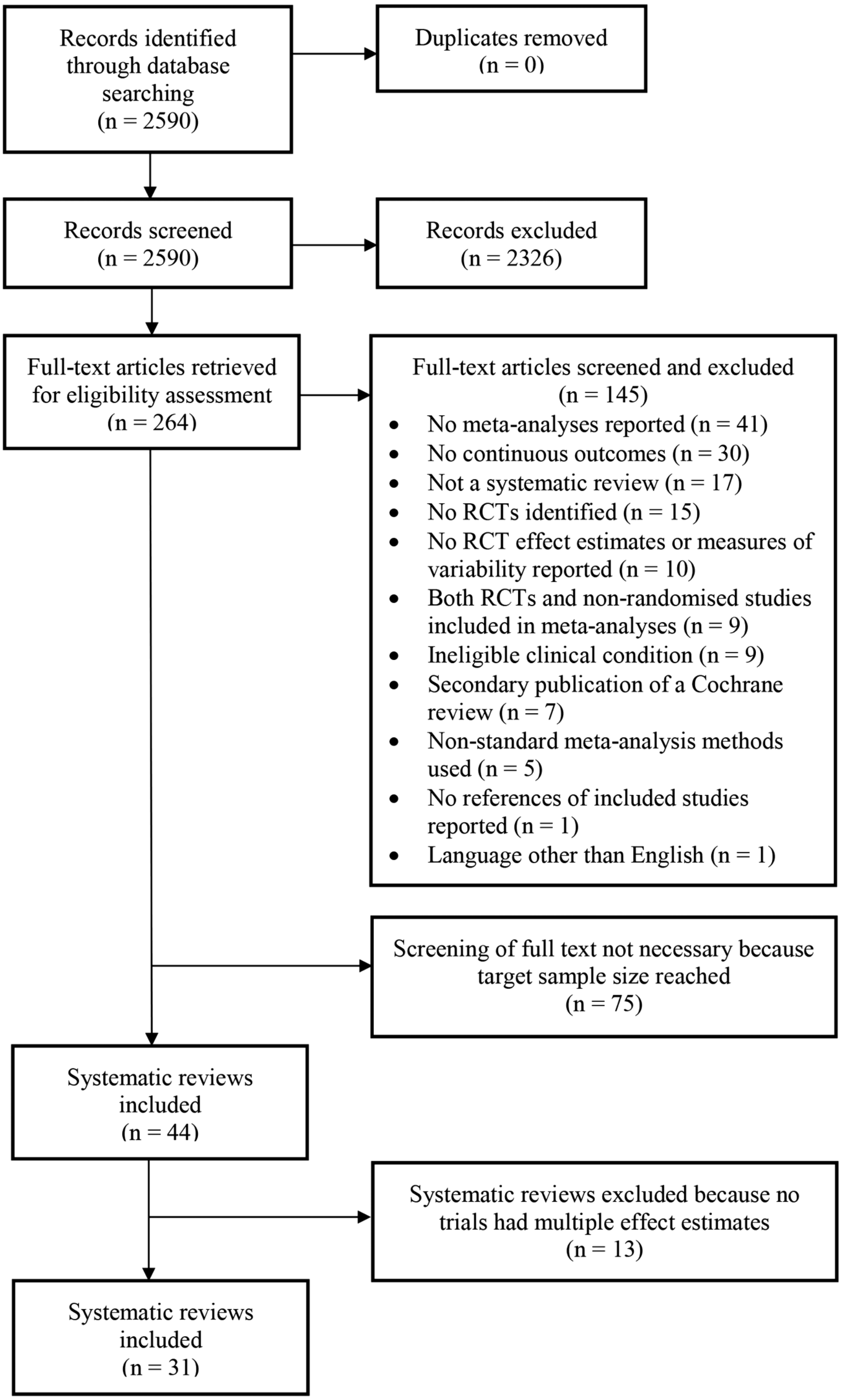

Records identified through database searching (n = 2590)
Duplicates removed (n = 0)
Records screened (n = 2590)
Records excluded (n = 2326)
Full-text articles retrieved for eligibility assessment (n = 264)
Full-text articles screened and excluded (n = 145)
• No meta-analyses reported (n = 41)
• No continuous outcomes (n = 30)
• Not a systematic review (n = 17)
• No RCTs identified (n = 15)
• No RCT effect estimates or measures of variability reported (n = 10)
• Both RCTs and non-randomised studies included in meta-analyses (n = 9)
• Ineligible clinical condition (n = 9)
• Secondary publication of a Cochrane review (n = 7)
• Non-standard meta-analysis methods used (n = 5)
• No references of included studies reported (n = 1)
• Language other than English (n = 1)
Screening of full text not necessary because target sample size reached (n = 75)
Systematic reviews included (n = 44)
Systematic reviews excluded because no trials had multiple effect estimates (n = 13)
Systematic reviews included (n = 31)

Figure 1 Flow diagram of identification, screening and inclusion of systematic reviews. RCT, randomised controlled trial.

the review (table 2), and was robust to the series of sensitivity analyses (see online supplementary table S3). The post hoc linear regression exploring the relationship between the number of available effect estimates and the PBI suggested that for every unit increase in the number of effect estimates, the PBI was predicted to reduce by −0.014 (95% CI −0.019 to −0.009; see online supplementary figure S1). For example, the predicted PBI was 0.60 (95% CI 0.59 to 0.61) when there were two available effect estimates, and 0.54 (95% CI 0.51 to 0.57) when there were six effect estimates.

Impact of potential selective inclusion of trial effect estimates on meta-analytic SMDs

The median number of possible meta-analytic SMDs arising from meta-analysing all combinations of trial effect estimates was 8 (across the 31 meta-analyses); however, there was wide variation in this number (IQR 3–576, range 2–1.8 trillion; table 3). The median percentile rank of the index meta-analytic SMD was 0.74 (IQR 0.20–1; range 0–1; note that when the percentile rank is 1, the index meta-analytic SMD was the most favourable of all possible meta-analytic SMDs). For most

Table 1 Characteristics of index meta-analyses (n=31)

Characteristics	n (%)*
Review protocol status	
Published in the Cochrane Database of Systematic Reviews	12 (39)
Unavailable or no mention that a review protocol was used	19 (61)
Clinical condition	
Rheumatoid arthritis or osteoarthritis	15 (48)
Depressive or anxiety disorder	16 (52)
Trials	
Number of trials in the meta-analysis, median (IQR)	6 (3, 10)
Outcome domain	
Depression	12 (39)
Pain	8 (26)
Function	2 (6)
Swollen joint count	2 (6)
Other (anxiety, obsessive compulsive symptoms, aerobic capacity, fatigue, physical activity, quality of life, range of motion)	7 (23)
Outcome label	
Primary	16 (52)
Non-primary (secondary or not labelled)	15 (48)
Effect measure	
Mean difference	4 (13)
Standardised mean difference	27 (87)
Meta-analysis model	
Fixed-effects	5 (16)
Random-effects	24 (77)
Not reported†	2 (6)
Type of comparison	
Placebo/no intervention controlled comparison	29 (94)
Head-to-head comparison	2 (6)
Type of active intervention	
Pharmacological	12 (39)
Non-pharmacological	19 (61)
Source of funding for systematic review	
Pharmaceutical industry	2 (6)
Non-industry (governmental agency or other not-for-profit organisation)	11 (35)
No funding	9 (29)
Not reported	9 (29)
Meta-analysis specification	
RA/OA meta-analyses	*n=15*
Defined by scale	
Yes (eg, Health Assessment Questionnaire score)	3 (20)
No (eg, disability)	12 (80)
Defined by time point	
Yes (eg, pain at 6 weeks)	4 (27)
No (eg, pain)	11 (73)
Depressive/anxiety disorder meta-analyses	*n=16*
Defined by scale	
Yes (eg, Beck Depression Inventory score)	4 (25)
No (eg, depression score)	12 (75)
Defined by time point	
Yes (eg, anxiety at 3 months)	9 (56)
No (eg, anxiety)	7 (44)

*All values given as n (%) except where indicated.
†Meta-analysis model not stated and unclear because both a fixed-effects and random-effects model produced the same SMD and 95% CI.
OA, osteoarthritis; RA, rheumatoid arthritis; SMD, standardised mean difference.

meta-analyses, the range of possible meta-analytic SMDs which could be calculated (each fitted using a random-effects model) was narrow. The median difference between the largest and smallest possible meta-analytic SMD was 0.11 SD units (IQR 0.03–0.19; range 0–0.43; table 3 and figure 2).

Table 2 Primary and subgroup analyses for the PBI

PBI analyses	Number of trials	Number of meta-analyses	PBI (95% CI*)
Primary analysis	118	31	0.57 (0.50 to 0.63)
Subgroup analyses			
Availability of a systematic review protocol			
With a protocol	27	12	0.55 (0.39 to 0.72)
Without a protocol	91	19	0.57 (0.49 to 0.65)
Subgroup difference†=−0.02 (95% CI −0.19 to 0.18; two-tailed p value of test of interaction 0.87)			
Availability of a core outcome measurement set for the clinical condition of the review			
Available core set (RA/OA systematic reviews)	55	15	0.62 (0.53 to 0.71)
No core set (depressive/anxiety disorder systematic reviews)	63	16	0.51 (0.41 to 0.62)
Subgroup difference†=0.11 (95% CI −0.04 to 0.24; two-tailed p value of test of interaction 0.15)			

*Percentile-based CIs for the PBI were constructed using bootstrap methods by resampling individual trials 2000 times.[12]
†The confidence limits and p value for the difference in PBI between subgroups was constructed using bootstrap methods from 2000 replicates.
OA, osteoarthritis; PBI, Potential Bias Index; RA, rheumatoid arthritis.

The impact of any potential selective inclusion on the meta-analytic SMDs was negligible. The median of the differences between the index meta-analytic SMD and the median of all its possible meta-analytic SMDs (ie, the SMD expected when there is no selective inclusion) was −0.01 SD units (IQR −0.05 to 0.01; range −0.20 to 0.16). Meta-analysing these differences using a random-effects model yielded a pooled difference of −0.004 SD units (95% CI −0.03 to 0.03; I^2=0%; figure 3). Recalculating all possible meta-analytic SMDs using a fixed-effects model yielded nearly identical results (see online supplementary figure S2).

DISCUSSION

There was no clear evidence that systematic reviewers selectively included trial effect estimates in this sample of meta-analyses of interventions for RA, OA, and depressive or anxiety disorders. The PBI was not modified by the availability of a systematic review protocol or a core outcome measurement set for the clinical condition of the review, and was robust to several assumptions. Any potential selective inclusion did not have an important impact on the meta-analytic effects. To our knowledge, no other study has investigated selective inclusion of trial effect estimates in systematic reviews and its potential impact on meta-analytic effects.

Strengths and weaknesses of the study

Our study has several strengths. We predefined our objectives and methods in a published protocol.[6] We used validated search strategies to identify eligible reviews and included a randomly selected sample of reviews. We included both Cochrane reviews and reviews published in other sources to increase the generalisability of results. Our estimates of the PBI and the impact of potential selective inclusion on the meta-analytic SMDs were robust to several assumptions and statistical approaches.

There are some limitations of our study. We only examined two clinical specialties (with two-thirds of the index meta-analyses focused on self-reported pain or depression), and we only included Cochrane and MEDLINE indexed non-Cochrane reviews published in English, so our results may have limited generalisability to other types of systematic reviews. Our subgroup comparison of reviews with versus without a core outcome measurement set is confounded by clinical specialty. Our sample size calculation was based on estimating the extent of multiplicity in trials rather than the magnitude of selective inclusion bias, as power calculations for the latter type of analysis do not yet exist. However, the resulting CI for the magnitude of bias in the primary analysis was narrow and conclusive, in that it excluded an effect in the alternative direction (ie, the possibility of selective inclusion of less favourable effects). Screening of the systematic reviews and extraction of 74% of the data was completed by only one author. However, we believe single screening is unlikely to have biased our results because at the screening stage, included trials had not been retrieved, so review inclusion/exclusion decisions were unable to be influenced by knowledge of the results in the trial reports. Further, since the systematic reviews that were extracted by two authors were randomly selected, we expect that the data extraction error rate in the random sample (which was very low) is representative of that of the entire sample.

Our findings may be influenced by the set of trial effect estimates that we considered eligible for inclusion in the meta-analyses. We used the methods reported in the review protocols to select effect estimates; however, other selection methods may have been assumed but not documented at the protocol stage. For example, systematic reviewers may have assumed there was a specialty consensus about which particular scale, intervention

Table 3 Number (%) of trials with multiplicity, number of possible meta-analytic standardised mean differences (SMDs), and differences between meta-analytic SMDs for each index meta-analysis (n=31)

SR ID	Number of trials	Number (%) of trials with multiplicity	Number of meta-analytic SMDs	Largest minus smallest meta-analytic SMD	Index minus median meta-analytic SMD*	Percentile rank of index meta-analytic SMD†
1	2	2 (100)	6	0.34	0.16	0
2	6	3 (50)	24	0.09	−0.03	1
3	7	3 (43)	60	0.14	0.09	0.04
4	2	2 (100)	8	0.19	0.05	0.43
5	8	3 (38)	16	0.10	−0.01	0.53
6	3	1 (33)	2	0	0	0
7	5	1 (20)	2	0.03	0.01	0
8	20	10 (50)	3072	0.15	−0.05	0.95
9	7	1 (14)	3	0.04	−0.002	1
10	6	1 (17)	2	0.03	0.01	0
11	5	1 (20)	3	0.05	0	0.5
12	2	1 (50)	4	0.03	−0.02	1
13	2	1 (50)	2	0.04	−0.02	1
14	5	4 (80)	864	0.27	−0.09	0.93
15	13	11 (85)	2 239 488‡	0.43	−0.06	0.76
16	3	1 (33)	2	0.01	−0.01	1
17	10	4 (40)	36	0.03	0.001	0.37
18	16	9 (56)	34 992‡	0.11	0.03	0.05
19	3	1 (33)	3	0.18	−0.06	1
20	9	6 (67)	384	0.15	−0.06	0.98
21	5	1 (20)	3	0.01	−0.001	1
22	28	6 (21)	2304	0.10	−0.01	0.74
23	2	2 (100)	4	0.26	−0.10	0.67
24	5	3 (60)	24	0.19	−0.07	0.87
25	13	10 (77)	1 843 200‡	0.39	−0.04	0.73
26	6	2 (33)	6	0.11	0.03	0.20
27	9	2 (22)	6	0.03	0	0.40
28	3	2 (67)	4	0.42	−0.20	1
29	20	18 (90)	1.8 trillion‡	0.16	−0.03	0.85
30	21	4 (19)	576	0.03	0.01	0.01
31	4	2 (50)	576	0.28	−0.12	1
Median (IQR)	6 (3, 10)	50% (22%, 67%)	8 (3, 576)	0.11 (0.03, 0.19)	−0.01 (−0.05, 0.01)	0.74 (0.20, 1)

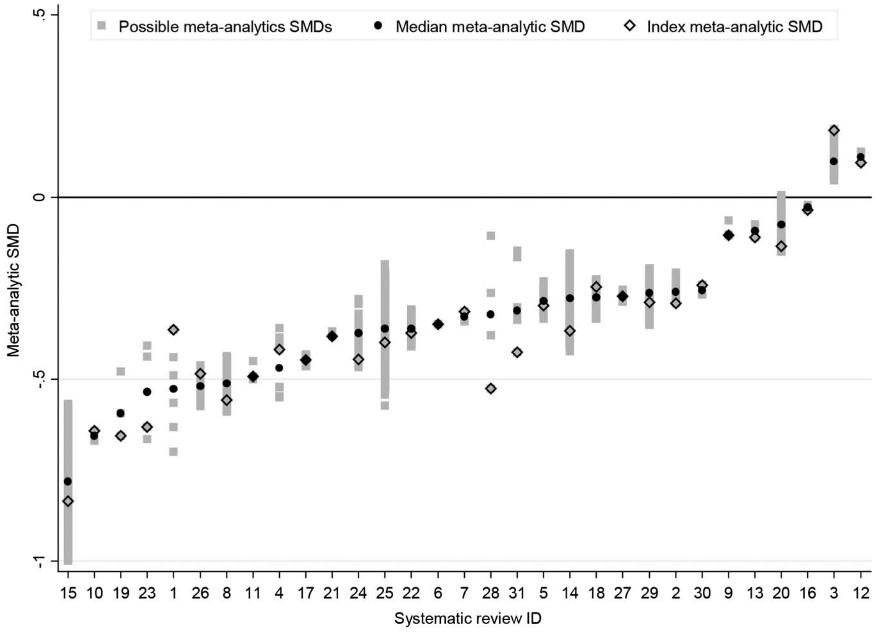
*Differences less than zero indicate that the index meta-analysis SMD is more favourable to the intervention compared with the median meta-analytic SMD.
†When the percentile rank is 1, the index meta-analytic SMD was the most favourable of all possible meta-analytic SMDs.
‡For these meta-analyses, we generated a sampling distribution of 5000 meta-analytic SMDs. Each meta-analytic SMD was created by randomly selecting (with equal probability) an effect estimate for inclusion from each trial, and meta-analysing the chosen effects.

dosage, time point or analysis to include in a review. In this instance, we may have extracted trial effect estimates that the reviewers considered ineligible. Further, we only investigated selective inclusion of completely reported effect estimates. The set of effect estimates that were actually considered by the systematic reviewers may have been larger, encompassing both reported and unreported estimates (in the case of the latter, obtained directly from the triallists). Empirical studies have found that trial outcomes with unfavourable results are less likely to be reported.[14 15] Therefore, we may have underestimated the PBI if effect estimates with unfavourable results were omitted from some of the trial reports, but retrieved by the systematic reviewers and subsequently excluded from the meta-analyses because of their unfavourable results.

Explanations of study results

There are several possible reasons why the estimated PBI was close to 0.5 (the value indicative of no selective inclusion). It is possible that selective inclusion occurred infrequently in this sample, and that the majority of systematic reviewers who did not report their selection methods selected trial effect estimates based on factors other than the magnitude and direction of the effects (eg, based on some undisclosed clinical rationale). Alternatively, our estimate of the PBI may be influenced by the construction of the test statistic, which provides

Figure 2 Range of possible meta-analytic standardised mean differences (SMDs) per index meta-analysis, with median and index meta-analytic SMD.



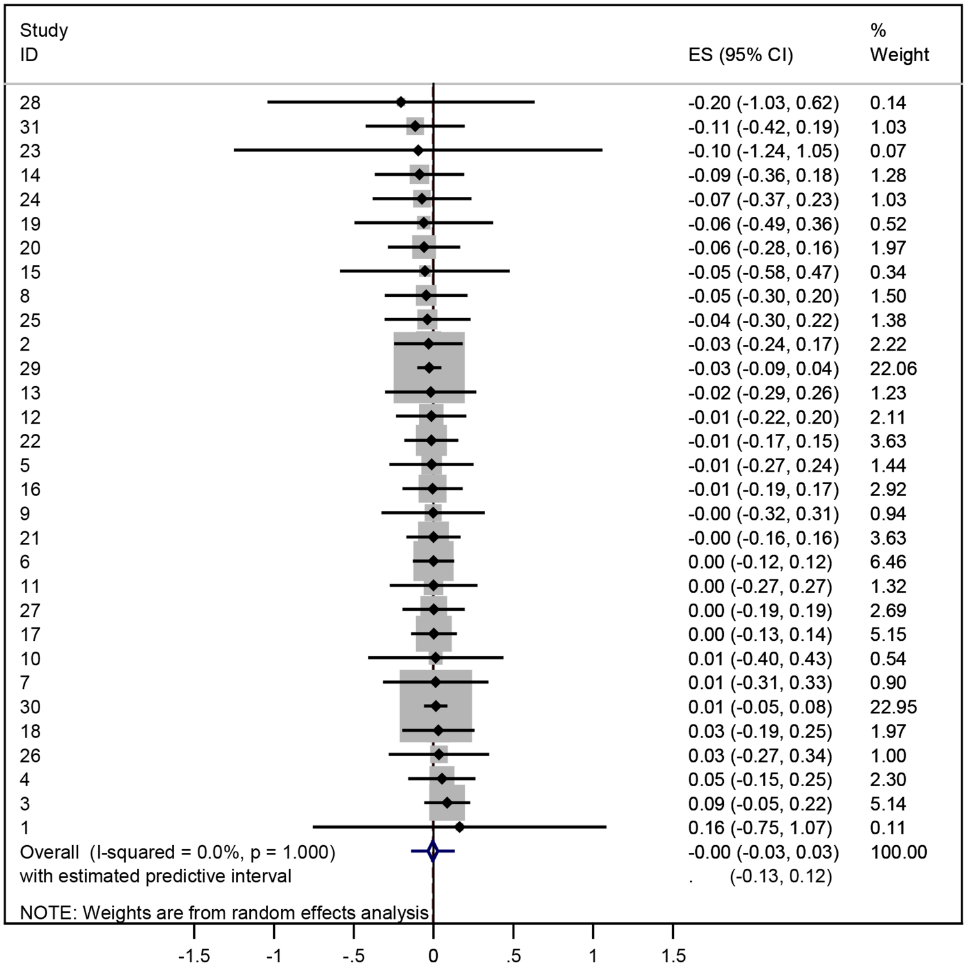


Figure 3 Meta-analysis of differences between the index meta-analytic SMD and median of all its possible meta-analytic SMDs (each calculated using the random-effects model). Differences less than zero indicate that the index meta-analysis SMD is more favourable to the intervention compared with the median meta-analytic SMD. ES, effect size; SMD, standardised mean difference.

more weighting to the rank position of effect estimates when there are more available. The post hoc analysis provides some evidence that when there are fewer effect estimates, the PBI is larger. One hypothesis for this result is that when there are only a few effect estimates, the reviewer can more easily compare the effects and select the most favourable, but when there are a large number of effect estimates, this comparison is more difficult, and so the reviewer selects effect estimates according to some other factor which is unrelated to the magnitude and direction. Contacting the authors to confirm the set of effect estimates they considered eligible for inclusion, and the methods they used to select effect estimates, could help rule out any suspected selective inclusion.

The negligible impact of potential selective inclusion on the meta-analytic effects may have been influenced by several factors. These include the percentage of trials with multiple effect estimates per meta-analysis, the extent to which the multiple effect estimates in a trial varied in magnitude and direction, the weights that trial effect estimates received in the meta-analysis or any combination of the above. For example, if one depression scale yields an SMD of −0.2 and another yields an SMD of −0.3, but the trial contributed little weight to the meta-analysis, selective inclusion of the most favourable SMD is unlikely to affect the meta-analysis in an important way. It would be useful to explore the association between the above factors and the magnitude of bias due to selective inclusion in future empirical studies.

Implications for systematic reviewers

We encourage systematic reviewers to predefine methods to select effect estimates to include in meta-analyses despite the negligible impact of selective inclusion that we found. Prespecification of eligibility criteria and decision rules to select effect estimates provides transparency in the review process and should make the process of data extraction and analysis easier, because not all available effect estimates may need to be extracted.[3 16] Current guidance documents recommend prespecification of eligibility criteria and decision rules for measurement scales and time points.[16–19] However, we believe decision rules for other sources of multiplicity (eg, unadjusted and covariate-adjusted analyses, analyses undertaken on multiple samples such as intention-to-treat and per-protocol) also deserve consideration.[20] When deciding which methods to prespecify, authors should consider issues such as the measurement properties of scales and clinical relevance of interventions and time points. If available, established guidance (eg, core outcome measurement sets or systematic reviews of the psychometric properties of scales) can inform the choice of methods to select effect estimates. Any rationale underlying the selection methods should be described in the protocol.

At the review stage, we recommend that systematic reviewers report the following: whether multiple trial effect estimates were available for inclusion in particular meta-analyses; if so, the methods used to select effect estimates; and any post hoc additions, omissions or modifications to methods to select effect estimates, along with justification for any discrepancies between the review protocol and review. Reporting such information may involve extra work for systematic reviewers, but will certainly help readers assess the risk of selective inclusion bias in meta-analyses. We recommend that a standardised table which facilitates reporting of information on multiplicity and selection methods be developed.

Future research

It is important to investigate whether selective inclusion of trial effect estimates occurs in meta-analyses which address other clinical specialties than those examined in our study. Most of the outcomes we examined were subjective. It would be valuable to examine in future studies mortality or other objective outcomes, which are not free from the risk of selective inclusion bias as they can be measured and analysed by triallists in multiple ways. In addition, it would be useful to investigate whether there is evidence of selective inclusion in meta-analyses of dichotomous outcomes, which have some unique types of multiplicity (eg, when events are defined using multiple diagnostic criteria, or defined by dichotomising measurement scales using different cut-points). Also, rather than just investigating selective inclusion of results fully reported in journal articles, it would be valuable to explore the frequency of selective inclusion in reviews with access to data in multiple sources (eg, conference abstracts, regulatory documents, clinical study reports). Further, given the evidence of an association between industry funding and research outcomes,[21] there is benefit in exploring whether selective inclusion is more common in reviews funded by the sponsor of the product under investigation. It would be ideal if future studies adopt the methods we have used, so that the findings of each investigation can be synthesised in meta-analyses.

CONCLUSION

There was no clear evidence that systematic reviewers selectively included trial effect estimates in this sample of meta-analyses of interventions for RA, OA, and depressive or anxiety disorders. Any potential selective inclusion did not have an important impact on the meta-analytic effects. To enable readers to assess the risk of selective inclusion bias, we recommend that systematic reviewers report whether multiplicity of effect estimates was encountered in trial reports, the methods used to select effect estimates to include in meta-analyses, and whether these methods were predefined or developed post hoc. Further research on selective inclusion in other clinical specialties is needed.

Contributors MJP conceptualised the study, screened literature for inclusion, extracted data, contributed to the statistical analyses, drafted the manuscript and approved the final manuscript. AF contributed to the design of the study,

developed the Potential Bias Index, test statistics and simulations, edited the manuscript and approved the final manuscript. MC extracted data, resolved discrepancies, edited the manuscript and approved the final manuscript. SEG edited the manuscript and approved the final manuscript. JEM conceptualised the study, conducted statistical analyses, drafted sections of the manuscript and approved the final manuscript.

Funding This work was conducted as part of a PhD undertaken by MJP, which was funded by an Australian Postgraduate Award administered through Monash University, Australia. MJP is supported by an Australian NHMRC Early Career Fellowship (1088535). JEM is supported by an Australian NHMRC Public Health Fellowship (1072366).

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Murad MH, Montori VM, Ioannidis JP, *et al*. How to read a systematic review and meta-analysis and apply the results to patient care: Users' guides to the medical literature. *JAMA* 2014;312:171–9.
2. Bender R, Bunce C, Clarke M, *et al*. Attention should be given to multiplicity issues in systematic reviews. *J Clin Epidemiol* 2008;61:857–65.
3. Tendal B, Nüesch E, Higgins JP, *et al*. Multiplicity of data in trial reports and the reliability of meta-analyses: empirical study. *BMJ* 2011;343:d4829.
4. Page MJ, McKenzie JE, Forbes A. Many scenarios exist for selective inclusion and reporting of results in randomized trials and systematic reviews. *J Clin Epidemiol* 2013;66:524–37.
5. Page MJ, McKenzie JE, Kirkham J, *et al*. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions (Review). *Cochrane Database Syst Rev* 2014;10:MR000035.
6. Page MJ, McKenzie JE, Green SE, *et al*. An empirical investigation of the potential impact of selective inclusion of results in systematic reviews of interventions: study protocol. *Syst Rev* 2013;2:21.
7. Kirkham JJ, Boers M, Tugwell P, *et al*. Outcome measures in rheumatoid arthritis randomised trials over the last 50 years. *Trials* 2013;14:324.
8. Jüni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. *Best Pract Res Clin Rheumatol* 2006;20:721–40.
9. Boers M, Kirwan JR, Wells G, *et al*. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol* 2014;67:745–53.
10. Chan AW, Hróbjartsson A, Haahr MT, *et al*. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457–65.
11. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]*. The Cochrane Collaboration. http://www.cochrane-handbook.org
12. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York, NY: Chapman & Hall, 1993.
13. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
14. Dwan K, Gamble C, Williamson PR, *et al*., Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLoS ONE* 2013;8:e66844.
15. Dwan K, Altman DG, Clarke M, *et al*. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS Med* 2014;11:e1001666.
16. Fu R, Vandermeer BW, Shamliyan T, *et al*. Chapter 14: handling continuous outcomes in quantitative synthesis. *In: Methods guide for effectiveness and comparative effectiveness reviews, AHRQ Publication No. 10(14)-EHC063-EF*. Rockville, MD: Agency for Healthcare Research and Quality, 2014. http://www.effectivehealthcare.ahrq.gov
17. Chandler J, Churchill R, Higgins J, *et al*. *Methodological standards for the conduct of new Cochrane Intervention Reviews*. Version 2.3, *2 December* 2013. http://www.editorial-unit.cochrane.org/mecir
18. Institute of Medicine (IOM). *Finding what works in health care: standards for systematic reviews*. Washington DC: The National Academies Press, 2011.
19. Booth A, Clarke M, Ghersi D, *et al*. Establishing a minimum dataset for prospective registration of systematic reviews: an international consultation. *PLoS ONE* 2011;6:e27319.
20. Page MJ, McKenzie JE, Chau M, *et al*. Methods to select results to include in meta-analyses deserve more consideration in systematic reviews. *J Clin Epidemiol* 2015;68:1282–91.
21. Lundh A, Sismondo S, Lexchin J, *et al*. Industry sponsorship and research outcome. *Cochrane Database Syst Rev* 2012;12:MR000033.