

Application and interpretation of statistical analyses in systematic reviews of therapeutic interventions can improve: a cross-sectional analysis

Matthew J Page^{1,*}, Douglas G Altman², Joanne E McKenzie¹, Larissa Shamseer^{3,4}, Nadera Ahmadzai⁵, Dianna Wolfe⁵, Fatemeh Yazdi⁵, Ferrán Catalá-López^{5,6}, Andrea C Tricco^{7,8}, David Moher^{3,4}

1. School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria 3004, Australia
2. UK EQUATOR Centre, Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford OX3 7LD, United Kingdom
3. Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, K1H 8L6, Canada
4. School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada
5. Knowledge Synthesis Group, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, K1H 8L6, Canada
6. Department of Medicine, University of Valencia/INCLIVA Health Research Institute and CIBERSAM, Valencia, 46010, Spain
7. Knowledge Translation Program, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, M5B 1W8, Canada
8. Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, M5T 3M7, Canada

Correspondence to: Dr. Matthew Page, School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, Victoria, 3004, Australia. Telephone: +61 9903 0248. Email address: matthew.page@monash.edu

60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118

1

2

Keywords: Reporting, Systematic Reviews, Methodology, Quality

3

4

Word count: 4,196 (main text), 6 tables, 1 supporting file

5

119
120
121 1 **ABSTRACT**
122

123 2 **OBJECTIVES:** To investigate the application and interpretation of statistical analyses in a cross-
124
125 3 section of systematic reviews (SRs) of therapeutic interventions, without restriction by journal,
126
127 4 clinical condition, or specialty.
128

129 5 **STUDY DESIGN AND SETTING:** We evaluated a random sample of SRs assembled previously, which
130
131 6 were indexed in MEDLINE® during February 2014, focused on a treatment or prevention question,
132
133 7 and reported at least one meta-analysis. The reported statistical methods used in each SR were
134
135 8 extracted from articles and online appendices by one author, with a 20% random sample extracted
136
137 9 in duplicate.
138
139

140 10 **RESULTS:** We evaluated 110 SRs; 78/110 (71%) were non-Cochrane SRs, and 55/110 (50%)
141
142 11 investigated a pharmacological intervention. The SRs presented a median of 13 (interquartile range
143
144 12 5-27) meta-analytic effects. When considering the index (primary or first reported) meta-analysis of
145
146 13 each SR, just over half (62/110 [56%]) used the random-effects model, but few (5/62 [8%])
147
148 14 interpreted the meta-analytic effect correctly (as the average of the intervention effects across all
149
150 15 studies). A statistical test for funnel plot asymmetry was reported in 17/110 (15%) SRs, however, in
151
152 16 only 4/17 (24%) did the test include the recommended number of at least 10 studies of varying size.
153
154 17 Subgroup analyses accompanied 42/110 (38%) index meta-analyses, but findings were not
155
156 18 interpreted with respect to a test for interaction in 29/42 (69%) cases, and the issue of potential
157
158 19 confounding in the subgroup analyses was not raised in any SR.
159
160

161 20 **CONCLUSIONS:** There is scope for improvement in the application and interpretation of statistical
162
163 21 analyses in SRs of therapeutic interventions. Involvement of statisticians on the SR team, and
164
165 22 establishment of partnerships between researchers with specialist expertise in SR methods and
166
167 23 journal editors may help overcome these shortcomings.
168
169
170
171
172
173
174
175
176
177

1. Introduction

A key component of many systematic reviews (SRs) is meta-analysis, the statistical combination of data from independent studies (1). Meta-analysis yields an average effect estimate across studies, which is particularly helpful to decision makers (e.g. clinicians, policy makers) confronted with a large number of results to interpret (2, 3). Meta-analysis increases precision of the effect estimate and may allow a question to be answered that could not otherwise be answered from individual studies (4). Extensions to meta-analysis methods, such as subgroup analysis and meta-regression, can be used to identify factors that explain variation in the magnitude of intervention effects (treatment-by-covariate interaction). These analyses can provide greater insight into the way an intervention works in different populations and settings than a single study can provide (5). However, the benefits of meta-analytic methods are only realised if they are applied and interpreted correctly.

Researchers have previously investigated the application and interpretation of statistical methods reported in published SRs. Riley et al. examined 75 Cochrane Pregnancy and Childbirth Group SRs published up to March 2008 (6). They found that the process of quantifying, investigating, and accounting for statistical heterogeneity was often limited or inadequate, the potential that some unpublished results are missing from a meta-analysis because of the nature of the findings ("reporting bias") was rarely addressed, and random-effects estimates were not interpreted correctly (i.e. as the average of the intervention effects across studies). Other audits have examined the use of subgroup analyses and meta-regression (7), use of statistical tests to infer bias due to missing results (8-11), and handling of data from cluster-randomized trials (12) and crossover trials (13) in more recent samples of SRs. All of these studies noted deficiencies in application or interpretation, or both. However, these evaluations have been narrow in scope, focusing on one particular aspect, or restricting inclusion to SRs in a single clinical specialty (e.g. dentistry (8)), or to Cochrane reviews, which represent only 15% of all SRs of biomedical research (14).

There are several benefits of exploring deficiencies in the application and interpretation of various statistical methods in SRs. Doing so can highlight areas that are in most need of improvement in future SRs, and thus can identify what guidance needs to be emphasised in training materials for systematic reviewers. Also, if the frequency of deficiencies is high, readers will need to be alerted so that they can be critical of meta-analyses that they use to inform their clinical practice or research. The primary objective of this study was to investigate the application and interpretation of various statistical methods in a cross-section of SRs of therapeutic interventions, without restriction by journal, clinical condition or specialty. The secondary objective was to explore if application and interpretation of methods differed between Cochrane and non-Cochrane SRs.

2. Methods

We conducted this project in accordance with a study protocol, which is available on the Open Science Framework: <https://osf.io/523bq/>. This study was conducted concurrently with another project evaluating reproducible research practices in SRs, which is described elsewhere.

2.1. Selection of articles

We used a database of SRs previously assembled, which consists of SRs of studies of various designs that were indexed in MEDLINE® in February 2014. A full description of the eligibility criteria and search strategy to identify these SRs is available in Page et al. (14). Briefly, the database includes articles published in English that met the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols (PRISMA-P) definition of a SR (15, 16). That is, articles with an explicit description of methods used to identify, select, and synthesise (or summarise) studies. The SRs were identified from a search of Ovid MEDLINE®, restricted to February 2014, using the search strategy employed by Moher et al. to retrieve SRs (17).

1 All titles and abstracts were screened using the method of liberal acceleration (also known as “safety
2 first” (18)) via the web-based review software, DistillerSR. In this method, only one author needed to
3 classify a record as “include or unsure” for it to be marked for full text screening, while two authors
4 needed to independently classify a record as “exclude” for it to be excluded. Two authors then
5 independently screened each full text article retrieved. Any discrepancies in screening of full text
6 articles were resolved via discussion, with adjudication by a third author when necessary. A total of
7 684 SRs met the inclusion criteria. A sample of 300 SRs was randomly selected, and data on their
8 epidemiological and reporting characteristics (e.g. clinical condition examined, methods used to
9 appraise studies) were collected. For the current study, we restricted inclusion to the 110 SRs of
10 randomized trials or non-randomized studies of therapeutic (i.e. treatment or prevention)
11 interventions, which reported at least one meta-analysis.

13 **2.2.Data collection and verification**

14 In the current study, we collected data using a standardized data collection form created in
15 DistillerSR (S1 Form). The form included 61 items that characterised how various statistical analyses
16 were applied and interpreted. Selection and wording of items was influenced by the
17 recommendations in the Methodological Expectations of Cochrane Intervention Reviews (19), and
18 data collection forms used in previous studies (6, 7, 20). Items included whether fixed-effect or
19 random-effects meta-analysis models were applied, how tests of statistical heterogeneity were
20 interpreted, and which covariates were explored in subgroup and sensitivity analyses. Some items in
21 the form referred to the whole SR (e.g. “Were any sensitivity analyses reported in the SR?”), while
22 others were directed at one meta-analysis per SR, known as the index meta-analysis (e.g. “How
23 many studies were included in the index meta-analysis?”).

24
25 The index meta-analysis was the first reported meta-analysis of the primary outcome. If no primary
26 outcome was defined, we selected the first outcome listed in the Objectives. If no outcome was

listed in the Objectives, we selected the first meta-analysis reported in the article. We categorised index meta-analysis outcomes as all-cause mortality, other objective outcome not requiring judgement (e.g. pregnancy, live births, laboratory outcomes), clinician-assessed outcome requiring judgement (e.g. events determined by clinical examination, cause-specific mortality), or patient-reported outcome (e.g. pain, mental health outcomes).

To count the number of meta-analyses in a SR, we summed the number of meta-analytic effect estimates presented in every forest plot, table, text, or web-based appendix. If a particular meta-analytic effect was presented in multiple locations (e.g. text and forest plot), we only counted it once. For subgroup analyses with an overall effect reported (which synthesised data across all subgroups), we counted each subgroup effect as well as the overall effect, as long as each subgroup included at least two studies.

One author (MJP) collected data from all SRs, from both the article and any available web-based appendices. One of three authors (NA, DW, FY) collected data from a 20% random sample of SRs (n=22) independently. Comparison of the data collected revealed 22 items where a discrepancy existed between two authors on at least one occasion (items marked in S1 Form). All discrepancies were resolved via discussion. All of the 22 items where a discrepancy had been identified were then checked for accuracy by one author (MJP) across the remaining 88 SRs.

2.3. Data Analysis

We summarised data as frequency and percentage for categorical items and median and interquartile range (IQR) for continuous items. We generated binomial exact confidence intervals for the following percentages – 1%, 5%, 10%, 25%, 50% and 75% – to illustrate to readers the precision of percentages calculated in our sample. We analyzed characteristics of all SRs and of SRs categorised as Cochrane or non-Cochrane. We explored whether the following eight practices, which

we considered fundamental to appropriate application and interpretation of statistical methods, differed between Cochrane and non-Cochrane SRs: the choice of meta-analysis model was not influenced by a heterogeneity statistic (e.g. Chi-squared test, I^2), use of the random-effects model was clinically justified, the random-effects estimate was interpreted correctly, a test for funnel plot asymmetry included at least 10 studies that were of varying size, rationale were reported for subgroup analyses, subgroup analyses were interpreted with respect to a statistical test for interaction, subgroup analyses were not interpreted by comparing the statistical significance of each subgroup effect estimate, and rationale were reported for sensitivity analyses. Comparisons were quantified using the risk ratio, with 95% confidence intervals (CIs), when the frequency was non-null in both groups. All analyses were performed using the statistical package Stata version 14 (21).

3. Results

3.1. Characteristics of SRs

Of the 110 included SRs, 32 (29%) were Cochrane SRs. Half of the SRs (55/110 [50%]) were led by authors based in China, UK or Canada, and nearly all (97/110 [88%]) were published in late 2013 (Table 1). Common clinical conditions addressed by the SRs included diseases of the digestive system, diseases of the circulatory system, infectious and parasitic diseases, and neoplasms. The interventions evaluated were pharmacological in 55/110 (50%) SRs, non-pharmacological in 43/110 (39%) SRs, or both in 12/110 (11%) SRs. The SRs included a median of 13 studies (IQR 7-23). The funding source was not reported in 38/110 (35%) SRs. Of the 72 SRs with funding source reported, 58 (81%) were funded by a non-profit source (e.g. University, government grant).

Table 1. General characteristics of 110 systematic reviews of therapeutic interventions

Characteristic	Data
Total number of journals	63
Journal impact factor (Thomson ISI 2012)	
0.0 – 5.0	59 (54%)
5.1 – 10	40 (36%)
10.1 – 15	0 (0%)
>15	2 (2%)
No impact factor	9 (8%)
Year of publication	
2014	11 (10%)
2013	97 (88%)
2012	2 (2%)
Country of corresponding author	
China	23 (21%)
UK	17 (15%)
Canada	15 (14%)
USA	13 (12%)
Other (21 countries with <10 SRs per country)	42 (38%)
Type of condition addressed by the SR (ICD-10 category)	
Diseases of the digestive system	14 (13%)
Diseases of the circulatory system	13 (12%)
Certain infectious and parasitic diseases	13 (12%)
Neoplasms (including cancers, carcinomas, tumors)	11 (10%)
Other (15 other ICD-10 classifications)	59 (54%)
Types of interventions addressed	
Pharmacological	55 (50%)
Non-pharmacological	43 (39%)
Both	12 (11%)
Number of included studies in the SR	13 (7-23)
Number of included participants in the SR	1,851 (630-5,540)
Use of PRISMA Statement mentioned	32 (29%)
SR protocol/registration mentioned	
SR registered (e.g. in PROSPERO)	6 (5%)
Protocol publicly available	36 (33%)
Source of funding	
Non-profit	58 (53%)
For-profit	1 (1%)
Mixed	1 (1%)
Specified there was no funding	12 (11%)
Not reported	38 (35%)

Data given as number (percent) or median (IQR). ICD-10 = International Classification of Diseases, Tenth Revision; SR = systematic review.

Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1% (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to 82%).

3.2. General characteristics of meta-analyses reported in the SRs

A median of 13 (IQR 5-27) meta-analytic effects, including those from subgroup meta-analyses and sensitivity analyses, were reported in the SRs (Table 2). These were calculated using a variety of effect measures, most commonly the risk ratio (48/110 [44%]) or mean difference (44/110 [40%]). Systematic reviewers most often reported using Review Manager (RevMan) (22) software to perform meta-analyses (80/110 [73%]), followed by Stata (21) (16/110 [15%]) and R (23) (7/110 [6%]). However, the particular package used in the latter two programs (e.g. *metan*, *meta*, *metafor*) was only stated in 7/23 (30%) SRs. There were no meta-analyses of dichotomous outcomes in 34/110 (31%) SRs and no meta-analyses of continuous outcomes in 55/110 (50%) SRs.

The index meta-analysis selected from each SR was described as a “primary” outcome in 68/110 (62%) cases (Table 2); in the remaining SRs there was no primary outcome specified so we examined the first reported meta-analysis. We classified most index meta-analyses as “other objective outcome not requiring judgement” or “clinician-assessed outcome requiring judgement” (each in 39/110 [35%] SRs). A median of six (IQR 3-11) studies and 593 (IQR 309-2,444) participants were included in the index meta-analyses. Nearly all index meta-analyses (102/110 [93%]) were presented using a forest plot. Few of these forest plots (20/102 [20%]) ordered studies by a potentially meaningful characteristic (e.g. year of publication, dosage of intervention), with most (61/102 [60%]) presenting studies in alphabetical order based on the surname of the lead author.

Table 2. General characteristics of meta-analyses reported in systematic reviews of therapeutic interventions

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
SR characteristics			
Number of meta-analytic effect estimates in the SR ^a	13 (5-27)	15 (5-30)	12 (5-24)
Statistical software used			
Review Manager (RevMan)	80 (73%)	32 (100%)	48 (62%)
Stata	16 (15%)	0 (0%)	16 (21%)
R	7 (6%)	0 (0%)	7 (9%)
Other (e.g. CMA, SPSS, WinBUGS)	25 (23%)	1 (3%)	24 (22%)
Effect measure used in at least one meta-analysis in the SR			
Risk ratio	48 (44%)	17 (53%)	31 (40%)
Odds ratio	28 (25%)	5 (16%)	23 (29%)
Risk difference	6 (5%)	2 (6%)	4 (5%)
Mean difference	44 (40%)	17 (53%)	27 (35%)
Standardized mean difference	21 (19%)	9 (28%)	12 (15%)
Ratio of means	0 (0%)	0 (0%)	0 (0%)
Hazard ratio	7 (6%)	2 (6%)	5 (6%)
Rate ratio	2 (2%)	2 (6%)	0 (0%)
Other (e.g. proportion, rate)	9 (8%)	0 (0%)	9 (12%)
No meta-analyses of binary outcomes	34 (31%)	10 (31%)	24 (31%)
No meta-analyses of continuous outcomes	55 (50%)	12 (38%)	43 (55%)
Handling of different study designs			
Data from RCTs and NRSI synthesized in at least one meta-analysis	12 (11%)	1 (3%)	11 (14%)
Data from RCTs and NRSI always analysed separately	3 (3%)	2 (6%)	1 (1%)
Not applicable – all had same design	95 (86%)	29 (91%)	66 (85%)
Index meta-analysis characteristics			
Index meta-analysis outcome type ^a			
All-cause mortality	3 (3%)	0 (0%)	3 (4%)
Other objective outcome not requiring judgement	39 (35%)	11 (34%)	28 (36%)
Clinician-assessed outcome requiring judgement	39 (35%)	10 (31%)	29 (37%)
Patient-reported outcome	29 (26%)	11 (34%)	18 (23%)
Index meta-analysis described as a primary outcome	68 (62%)	23 (72%)	45 (58%)
Number of studies included in index meta-analysis	6 (3-11)	4 (2-6)	7 (4-12)
Number of participants in index meta-analysis	593 (309-2,444)	419 (218-1,604)	817 (341-2,952)
Effect measure for index meta-analysis			
Risk ratio	36 (33%)	14 (44%)	22 (28%)
Odds ratio	20 (18%)	4 (13%)	16 (21%)
Risk difference	2 (2%)	0 (0%)	2 (3%)

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
Mean difference	25 (23%)	6 (19%)	19 (24%)
Standardized mean difference	10 (9%)	4 (13%)	6 (8%)
Hazard ratio	6 (5%)	2 (6%)	4 (5%)
Rate ratio	2 (2%)	2 (6%)	0 (0%)
Other (e.g. proportion, rate)	9 (8%)	0 (0%)	9 (12%)
Index meta-analysis presented on forest plot	102 (93%)	32 (100%)	70 (90%)
How studies were ordered on forest plot			
By study author	61/102 (60%)	26 (81%)	35/70 (50%)
By year of publication	11/102 (11%)	1 (3%)	10/70 (14%)
By effect size	6/102 (6%)	2 (6%)	4/70 (6%)
By study weight	1/102 (1%)	1 (3%)	0/70 (0%)
By outcome measurement instrument	1/102 (1%)	1 (3%)	0/70 (0%)
By year of publication and condition	1/102 (1%)	0 (0%)	1/70 (1%)
Ordering factor not apparent	21/102 (21%)	1 (3%)	20/70 (29%)
Weight given to each study presented for index meta-analysis			
Numerically and graphically	94/102 (92%)	32 (100%)	62/70 (89%)
Numerically only	1/102 (1%)	0 (0%)	1/70 (1%)
Graphically only	4/102 (4%)	0 (0%)	4/70 (6%)
Weights not presented	3/102 (3%)	0 (0%)	3/70 (4%)
Labels indicating which group is favoured presented for index meta-analysis	89/102 (87%)	32 (100%)	57/70 (81%)

1 Data given as number (percent) or median (IQR). The denominator of fractions indicates the number
2 of reports where the variable concerned was considered relevant to the systematic review. CMA =
3 Comprehensive Meta-Analysis; NRSI = non-randomized study of intervention; RCT = randomized
4 controlled trial.

5 Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1%
6 (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to
7 82%).

8 NRSI = non-randomized studies of interventions; RCT = randomized controlled trial.

9 ^aExamples of “other objective outcome not requiring judgement” include pregnancy, live births, and
10 laboratory outcomes such as biochemical measurements or serologic tests. Examples of “clinician-
11 assessed outcome requiring judgement” include events determined by clinical examination, and
12 cause-specific mortality.

15 3.3. Application and interpretation of random-effects models and heterogeneity statistics

16 A random-effects model was used for at least one meta-analysis in 79/110 (72%) SRs (Table 3). In
17 61/110 (55%) SRs, systematic reviewers stated in the Methods section that a particular threshold of
18 the I^2 inconsistency statistic (24) denoted “substantial heterogeneity”; thresholds varied, ranging
19 from 25% to 75%. In 33/110 (30%) cases, systematic reviewers used a Chi-squared test for
20 heterogeneity (P-value for Cochran’s Q) or a particular value of I^2 to guide the choice of meta-

- 1 analysis model (e.g. "The fixed-effects model was used if I^2 was less than 50%"), despite this
- 2 approach being discouraged in guidance for meta-analyses (1, 19, 25, 26).

Table 3. Application and interpretation of random-effects models and heterogeneity statistics

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
SR characteristics			
Meta-analysis models used			
Fixed-effect model for all meta-analyses	26 (24%)	14 (44%)	12 (15%)
Random-effects model for all meta-analyses	47 (43%)	7 (22%)	40 (51%)
Varied across meta-analyses	32 (29%)	11 (34%)	21 (27%)
Not reported	5 (5%)	0 (0%)	5 (6%)
Minimum I ² value that was considered indicative of substantial heterogeneity			
25%	4 (4%)	1 (3%)	3 (4%)
50%	47 (43%)	16 (50%)	31 (40%)
75%	5 (5%)	2 (6%)	3 (4%)
Other (e.g. 30%, 35%, 70%)	5 (5%)	2 (6%)	3 (4%)
None reported	49 (45%)	11 (34%)	38 (49%)
Heterogeneity statistic guided choice of meta-analysis model (e.g. stated that fixed-effects model selected if I ² < 50%)	33 (30%)	6 (19%)	27 (35%)
Index meta-analysis characteristics			
Random-effects model used for the index meta-analysis	62 (56%)	13 (41%)	49 (63%)
Use of random-effects model for the index meta-analysis was clinically justified	18/62 (29%)	6/13 (46%)	12/49 (24%)
Between-study variance estimator used for the index meta-analysis			
DerSimonian-Laird	51/62 (82%)	13/13 (100%)	38/49 (78%)
Other	0/62 (0%)	0/13 (0%)	0/49 (0%)
Not reported	11/62 (18%)	0/13 (0%)	11/49 (22%)
Method to calculate confidence interval used for the index meta-analysis			
Wald-type method	51/62 (82%)	13/13 (100%)	38/49 (78%)
Other	0/62 (0%)	0/13 (0%)	0/49 (0%)
Not reported	11/62 (18%)	0/13 (0%)	11/49 (22%)
Index meta-analytic effect interpreted correctly ^a	5/62 (8%)	2/13 (15%)	3/49 (6%)
Prediction interval reported for the index meta-analysis	0/62 (0%)	0/13 (0%)	0/49 (0%)
Result for corresponding fixed-effect meta-analysis reported	7/62 (11%)	0/13 (0%)	7/49 (14%)

Data given as number (percent). The denominator of fractions indicates the number of reports where the variable concerned was considered relevant to the systematic review.

Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1% (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to 82%).

^aCorrect interpretation of the random-effects meta-analysis is as the average of the intervention effects across studies, not the best estimate of a common intervention effect across studies (27).

1 A random-effects model was used for 62/110 (56%) of the index meta-analyses (Table 3), but
2 without any clinical rationale in 44/62 (71%) SRs. Nearly all (51/62 [82%]) random-effects meta-
3 analyses were performed using the Dersimonian-Laird between-study variance estimator (28). Other
4 estimators (e.g. Sidik-Jonkman (29)) were not reported in any of the remaining SRs, but in 11/62
5 (18%) the choice of estimator was not stated and could not be inferred from the statistical package
6 used. None of the 62 random-effects meta-analyses were accompanied by a prediction interval. Few
7 systematic reviewers (5/62 [8%]) interpreted the meta-analytic effect correctly (i.e. as the *average* of
8 the intervention effects across studies, not the best estimate of a *common* intervention effect across
9 studies (27)).

11 3.4. Application and interpretation of funnel plots and tests for funnel plot asymmetry

12 In 45/110 (41%) SRs it was stated that a funnel plot had been generated, but the plot was presented
13 in only 28/45 (62%) of these SRs (Table 4). None of the funnel plots included contour lines
14 corresponding to perceived milestones of statistical significance, which can aid visual interpretation
15 (30). A statistical test for funnel plot asymmetry (e.g. Egger regression (31)) was reported in few SRs
16 (17/110 [15%]). However, in only 4/17 (24%) of these SRs did the test include the recommended
17 number of at least 10 studies that were of varying size (32). When interpreting and reporting these
18 plots and tests, only in 1/45 (2%) SR was there an acknowledgement that observed funnel plot
19 asymmetry may be due to a factor other than reporting bias (e.g. clinical heterogeneity, chance). In a
20 few SRs (8/45 [18%]), systematic reviewers reported that visual inspection of the funnel plot led
21 them to suspect reporting bias (i.e. that results were missing from a meta-analysis).

Table 4. Application and interpretation of funnel plots and tests for funnel plot asymmetry

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
SR characteristics			
Reported that a funnel plot was generated	45 (41%)	7 (22%)	38 (49%)
Funnel plot actually presented, not just referred to in text	28/45 (62%)	5/7 (71%)	23/38 (61%)
Number of funnel plots presented	1 (1-2)	2 (1-3)	1 (1-2)
Contour-enhanced funnel plot presented	0 (0%)	0 (0%)	0 (0%)
Possible reasons for funnel plot asymmetry suggested by systematic reviewers			
Reporting bias (missing results)	42/45 (93%)	7/7 (100%)	35/38 (92%)
Clinical heterogeneity	1/45 (2%)	0/7 (0%)	1/38 (3%)
Other (e.g. chance)	0/45 (0%)	0/7 (0%)	0/38 (0%)
None	2/45 (4%)	0/7 (0%)	2/38 (5%)
Funnel plot asymmetry test used (e.g. Egger test, Begg test)	17 (15%)	0 (0%)	17 (22%)
Possible reasons for statistically significant asymmetry test suggested by systematic reviewers			
Reporting bias (missing results)	17/17 (100%)	NA	17/17 (100%)
Other reason (e.g. clinical heterogeneity)	0/17 (0%)	NA	0/17 (0%)
Funnel plot asymmetry test included at least 10 studies that were of varying size	4/17 (24%)	NA	4/17 (24%)
Conclusion based on funnel plot about the risk of reporting bias (i.e. that results are missing from a meta-analysis)			
Suspected for at least one meta-analysis	8/45 (18%)	2/7 (29%)	6/38 (16%)
Not suspected for all meta-analyses	37/45 (82%)	5/7 (71%)	32/38 (84%)

Data given as number (percent) or median (IQR). The denominator of fractions indicates the number of reports where the variable concerned was considered relevant to the systematic review. NA = Not applicable.

Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1% (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to 82%).

3.5. Application and interpretation of subgroup and meta-regression analyses

At least one subgroup or meta-regression analysis was reported in 53/110 (48%) SRs (Table 5). The median number of subgroup or meta-regression analyses in these 53 SRs was four (IQR 2-7). A subgroup analysis accompanied 42/110 (38%) of the index meta-analyses, of which common covariates included intervention characteristics (e.g. dosage) (in 26/42 [62%]) and patient characteristics (e.g. age, sex) (in 23/42 [55%]).

1
2 There were several weaknesses in how subgroup analyses were applied and interpreted. Few SRs
3 (3/42 [7%]) included rationale for selection of the covariates. Further, the issue of potential
4 confounding in the subgroup analyses was not raised in any SR. That is, none of the systematic
5 reviewers acknowledged that a difference in subgroup effect estimates might be explained by some
6 other factor (e.g. a larger effect in trials conducted in high-income countries compared with trials
7 conducted in low- and middle-income countries may be confounded by administration of a co-
8 intervention that is only available in high-income countries) (5). For most subgroup analyses
9 accompanying the index meta-analysis (29/42 [69%]), results were not interpreted with reference to
10 a statistical test for interaction (e.g. Q-test for heterogeneity (33)). Also, in 11/42 (26%) SRs, some
11 subgroup analyses were inappropriately interpreted by comparing the statistical significance of each
12 subgroup effect (e.g. systematic reviewers incorrectly concluded that the intervention effect varied
13 per level of the covariate because one subgroup meta-analytic estimate was statistically significant
14 while the other was not) (34).

Table 5. Application and interpretation of subgroup and meta-regression analyses

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
SR characteristics			
Any subgroup or meta-regression analyses reported in the SR	53 (48%)	11 (34%)	42 (54%)
Subgroup-analyses reported in the SR	52 (47%)	11 (34%)	42 (54%)
Meta-regression reported in the SR	7 (6%)	0 (0%)	7 (9%)
Number of subgroup or meta-regression analyses reported per SR	4 (2-7)	4 (2-12)	4 (2-6)
Index meta-analysis characteristics			
Subgroup analysis accompanied the index meta-analysis	42 (38%)	9 (28%)	33 (42%)
Covariates included in subgroup analyses accompanying the index meta-analysis			
Study design (e.g. RCTs versus NRSI)	5/42 (12%)	0/9 (0%)	5/33 (15%)
Patient characteristic (e.g. age, sex)	23/42 (55%)	4/9 (44%)	19/33 (58%)
Intervention characteristic (e.g. dosage)	26/42 (62%)	5/9 (56%)	21/33 (64%)
Outcome characteristic (e.g. timing)	6/42 (14%)	0/9 (0%)	6/33 (18%)
Study risk of bias or "quality"	5/42 (12%)	1/9 (11%)	4/33 (12%)
Other	2/42 (5%)	2/9 (22%)	0/33 (0%)
Rationale for subgroup analyses accompanying the index meta-analysis reported			
Yes, for all analyses	3/42 (7%)	2/9 (22%)	1/33 (3%)
Yes, but only for some analyses	2/42 (5%)	1/9 (11%)	1/33 (3%)
No, not for any analysis	37/42 (88%)	6/9 (67%)	31/33 (94%)
Potential for confounding in subgroup analyses accompanying the index meta-analysis discussed	0/42 (0%)	0/9 (0%)	0/33 (0%)
Subgroup analyses accompanying the index meta-analysis interpreted with respect to a statistical test for interaction			
Yes, for all analyses	10/42 (24%)	6/9 (67%)	4/33 (12%)
Yes, but only for some analyses	3/42 (7%)	0/9 (0%)	3/33 (9%)
No, not for any analysis	29/42 (69%)	3/9 (33%)	26/33 (79%)
Subgroup analyses accompanying the index meta-analysis inappropriately interpreted by comparing the statistical significance of each subgroup effect			
Yes, for all analyses	8/42 (19%)	0/9 (0%)	8/33 (24%)
Yes, but only for some analyses	3/42 (7%)	0/9 (0%)	3/33 (9%)
No, not for any analysis	31/42 (74%)	9/9 (100%)	22/33 (67%)
Concluded that subgroup estimates differed in at least one subgroup analysis accompanying the index meta-analysis	19/42 (45%)	0/9 (0%)	19/33 (58%)

Data given as number (percent) or median (IQR). The denominator of fractions indicates the number of reports where the variable concerned was considered relevant to the systematic review. NRSI = non-randomized study of intervention; RCT = randomized controlled trial.

1 Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1%
2 (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to
3 82%).
4

6 **3.6. Application and interpretation of sensitivity analyses**

7 At least one sensitivity analysis was reported in 55/110 (50%) SRs, and the median number of
8 sensitivity analyses per SR was three (IQR 1-9). Almost half (51/110 [46%]) of the index meta-
9 analyses were accompanied by a sensitivity analysis. Several types of sensitivity analyses were
10 performed across the SRs. The most common investigations included using a different meta-analysis
11 model (i.e. fixed-effect versus random-effects) (17/51 [33%] SRs), removing studies with a particular
12 patient characteristic (e.g. studies with predominantly elderly participants) (14/51 [27%] SRs), and
13 removing each study one at a time (13/51 [25%] SRs). No rationale was provided for any of the
14 sensitivity analyses accompanying 37/51 [73%] index meta-analyses.
15
16

Table 6. Application and interpretation of sensitivity analyses

Characteristic	All (n = 110)	Cochrane (n = 32)	Non-Cochrane (n = 78)
SR characteristics			
Any sensitivity analyses reported in the SR	55 (50%)	15 (47%)	40 (51%)
Number of sensitivity analyses reported in the SR	3 (1-9)	7 (3-15)	2 (1-8)
Index meta-analysis characteristics			
Sensitivity analysis accompanied the index meta-analysis	51 (46%)	14 (44%)	37 (47%)
Variable explored in sensitivity analyses accompanying the index meta-analysis			
Study design (e.g. RCTs or NRSI)	4/51 (8%)	1/14 (7%)	3/37 (8%)
Patient characteristic	14/51 (27%)	6/14 (43%)	8/37 (22%)
Intervention characteristic	10/51 (20%)	2/14 (14%)	8/37 (22%)
Outcome characteristic	7/51 (14%)	2/14 (14%)	5/37 (14%)
Study risk of bias or "quality"	8/51 (16%)	3/14 (21%)	5/37 (14%)
Unpublished data	2/51 (4%)	2/14 (14%)	0/37 (0%)
Imputed data	1/51 (2%)	0/14 (0%)	1/37 (3%)
Different model or assumptions (e.g. fixed-effect versus random-effects)	17/51 (33%)	6/14 (43%)	11/37 (30%)
Removal of each study one at a time	13/51 (25%)	1/14 (7%)	12/37 (32%)
Other	8/51 (16%)	2/14 (14%)	6/37 (16%)
Rationale for sensitivity analyses accompanying the index meta-analysis reported			
Yes, for all analyses	12/51 (24%)	6/14 (43%)	6/37 (16%)
Yes, but only for some analyses	2/51 (4%)	0/14 (0%)	2/37 (5%)
No, not for any analysis	37/51 (73%)	8/14 (57%)	29/37 (78%)
Concluded that the index meta-analytic effect was robust according to all sensitivity analyses	43/51 (84%)	12/14 (86%)	31/37 (84%)

Data given as number (percent) or median (IQR). The denominator of fractions indicates the number of reports where the variable concerned was considered relevant to the systematic review. NRSI = non-randomized study of intervention; RCT = randomized controlled trial. Illustrative binomial exact 95% confidence intervals for percentages when sample size is 110: 1% (0.02% to 5%); 5% (1% to 10%); 10% (5% to 17%); 25% (17% to 34%); 50% (39% to 59%); 75% (65% to 82%).

3.7. Influence of SR type on application and interpretation of statistical methods

Cochrane SRs were more likely than non-Cochrane SRs to not base the choice of meta-analysis model on the result of a heterogeneity statistic (26/32 Cochrane SRs versus 51/78 non-Cochrane SRs; RR 1.24, 95% CI 0.99-1.57), to provide clinical justification for use of the random-effects model (6/13 Cochrane SRs versus 12/49 non-Cochrane SRs; RR 1.88, 95% CI 0.88, 4.05), to interpret the

random-effects estimate correctly (2/13 Cochrane SRs versus 3/49 non-Cochrane SRs; RR 2.51, 95% CI 0.47-13.50), to provide rationale for subgroup analyses (2/9 Cochrane SRs versus 1/33 non-Cochrane SRs; RR 7.33, 95% CI 0.75-72.02), to interpret subgroup analyses with respect to a statistical test for interaction (6/9 Cochrane SRs versus 4/33 non-Cochrane SRs; RR 5.50, 95% CI 1.97-15.38), and to provide rationale for sensitivity analyses (6/14 Cochrane SRs versus 6/37 non-Cochrane SRs; RR 2.64, 95% CI 1.02-6.83). However, there was large uncertainty in the risk ratio estimates, with nearly all 95% CIs encompassing associations in both directions (i.e. favouring Cochrane or favouring non-Cochrane). An exception was the interpretation of subgroup analyses with respect to a statistical test for interaction, where the lower bound of the 95% CI indicated an important difference favouring Cochrane SRs.

4. Discussion

We identified several areas for improvement in the application and interpretation of statistical methods in SRs of therapeutic interventions. When considering the index (primary or first reported) meta-analysis of each SR, just over half (56%) used the random-effects model, but few (8%) interpreted the meta-analytic effect correctly, as the average of the intervention effects across all studies (27). A statistical test for funnel plot asymmetry was reported in 15% of SRs, however, in only 24% of these did the test include the recommended number of at least 10 studies that were of varying size. Subgroup analyses accompanied 38% of index meta-analyses, but most findings (69%) were not interpreted with respect to a statistical test for interaction, 26% were incorrectly interpreted by comparing the statistical significance of each subgroup effect, and the issue of potential confounding in the subgroup analyses was not raised in any SR. Sensitivity analyses accompanied 46% of the index meta-analyses, although no rationale was provided in 73% of cases. There were some differences between Cochrane and non-Cochrane SRs in the application and interpretation of statistical methods; however, the 95% CIs of nearly all risk ratio associations were wide and included the null.

4.1. Strengths and limitations of the study

There are several strengths of our methods. We included SRs that had been identified and selected previously using rigorous strategies (i.e. a systematic search and screening by two authors independently) (14). Our sample of SRs was not restricted by journal, clinical condition or specialty. This may enhance the generalisability of our results, since we were able to collect data on a broader cross-section of SRs. In addition, unlike previous studies with a more narrow scope (e.g. (7, 8)), we examined how several statistical methods were applied and interpreted.

Our findings should be considered in light of some limitations. The requirement that SRs meet the PRISMA-P definition of a SR is likely to have led to selection of SRs that are of higher methodological quality (35). If lower quality SRs have more shortcomings in application and interpretation of statistical methods, we will have underestimated the true scale of the problem. We are unsure of the extent to which our findings generalise to SRs indexed outside of MEDLINE® or SRs published in a language other than English. We only evaluated statistical methods used for standard pairwise meta-analyses; data on statistical methodology of network meta-analyses have been reported elsewhere (36). All SRs were indexed in February 2014. We could not identify any factors that would lead to a difference in how statistical methods are applied and interpreted in more recent SRs (e.g. no new or updated handbooks or guidelines for SR conduct and reporting have been disseminated). Most comparisons between Cochrane and non-Cochrane SRs were based on small samples, which yielded wide confidence intervals surrounding the estimated risk ratios.

Only one author collected data from 80% of SRs, so it is possible that there are some errors in our dataset. However, we suspect that the error rate is low because the discrepancy rate in the random sample of 22 SRs reviewed by two authors independently was low. There was opportunity for discrepancies in 2,442 fields (22 SRs x 111 responses options = 2,442), yet there were discrepancies in only 124/2442 fields; a discrepancy rate of 5%. Also, we minimized data collection errors in the

1 remaining 88 SRs by verifying all data collected for 22 of 61 items where at least one discrepancy
2 between two authors had been identified from the 22 randomly selected SRs. Having an
3 independent author verify data for the 22 items in the 88 SRs would have been ideal, but we decided
4 it was appropriate for the same author who coded the data the first time to verify the data, because
5 the original set of discrepancies were minor in frequency and nature. That is, there were no major
6 differences in interpretation for any items; rather, all discrepancies were due to simple mistakes
7 made by one of the data collectors. Such mistakes were identified and corrected by the data
8 collector on the second examination of the articles.

10 **4.2. Comparison with other studies**

11 Several of our findings are consistent with those observed in other cross-sectional analyses. For
12 example, Riley et al. found that in 75 Cochrane SRs published prior to March 2008, a heterogeneity
13 statistic guided the choice of meta-analysis model in 33% of SRs, interpretation of the random-
14 effects meta-analytic estimate was correct in 0% of SRs, and the issue of possible confounding in
15 subgroup analyses was discussed in 4% of SRs (6). The corresponding percentages in our sample
16 were 30%, 8%, and 0%, respectively, suggesting little or no improvement over time. Also, Donegan
17 et al. reported that subgroup analyses were interpreted with respect to a statistical test for
18 interaction in only 12% of 52 Cochrane SRs published up to August 2013 (7); there was a similarly
19 low percentage of SRs doing so in our sample (24%).

21 **4.3. Explanations of study results and implications**

22 There are several possible reasons for the poor statistical practices observed in our study. Many SRs
23 may be conducted by systematic reviewers without relevant statistical expertise or access to it,
24 which increases the chance for misunderstanding how methods should be applied and results
25 interpreted. Further, while user-friendly meta-analysis software (e.g. RevMan, Comprehensive Meta-
26 Analysis (CMA)) has the benefit of enabling a greater number of researchers to use meta-analytical
27 methods, this also makes it easy for those with limited knowledge to use these methods (6).

1 Reporting guidelines such as the PRISMA Statement cover some statistical considerations for
2 reporting (37). For example, item 14 recommends that systematic reviewers “Describe the methods
3 of handling data and combining results of studies, if done, including measures of consistency (e.g. I²)
4 for each meta-analysis” (38). However, the checklist items are designed primarily to encourage
5 transparent reporting of the methods used and results found, rather than to provide detailed
6 guidance on appropriate choice and interpretation of statistical methods.

7
8 Various approaches might help overcome the shortcomings identified. Formal training in meta-
9 analytic methods could become a mandatory component of biomedical and public health course
10 curricula (39, 40). Partnerships could be formed between methodologists and statisticians with
11 specialist expertise in SR methods and journal editors to prepare guidance for authors submitting
12 SRs, and to implement quality standards (for example, using the model adopted by the Cochrane
13 Eyes and Vision Group partnering with ophthalmology journals, see
14 <http://eyes.cochrane.org/associate-editors-eyes-and-vision-journals>). Bodies that fund SRs could
15 ensure that the SR team includes someone with the necessary statistical expertise (41). In future, the
16 PRISMA checklist items could perhaps be revised or extended to provide more specific statistical
17 guidance.

18
19 Until initiatives such as these are in place, we recommend that systematic reviewers focus on the
20 following practices. Systematic reviewers should provide an explanation (ideally with reference to
21 clinical/methodological rationale) for their decision to use a fixed-effect or random-effects model,
22 their choice of covariates to explore in subgroup analyses, and their choice of sensitivity analyses (1).
23 When using a random-effects model, results from meta-analyses should be described as the *average*
24 of the intervention effects across studies; including a prediction interval to present the expected
25 range of true effects in similar studies, when appropriate to do so, may facilitate interpretation (27,
26 42). Systematic reviewers should understand that funnel plots can display whether there is a
27 tendency for the intervention effects estimated in smaller studies to differ from those estimated in

larger studies, but that there are different possible reasons for such funnel plot asymmetry (e.g. smaller studies with statistically non-significant results remain unpublished, a higher drug dosage was administered in smaller studies, or smaller studies were at higher risk of bias due to problems with the randomization process). Therefore, if using funnel plots, systematic reviewers should consider that any asymmetry detected can have occurred for reasons other than reporting bias (32). As a rule of thumb, statistical tests for funnel plot asymmetry should be used only when there are at least 10 studies of varying size in the meta-analysis; when the meta-analysis includes fewer than 10 studies, the statistical power of the test is usually too low to distinguish chance from real asymmetry (32). When conducting subgroup analyses, results should always be interpreted with respect to a test for interaction, because such a test indicates whether the subgroup effect estimates are statistically significantly different from each other (34). Also, the observational nature of subgroup analyses and potential for confounding should be acknowledged (5, 7).

5. Conclusion

We observed many deficiencies in the application and interpretation of statistical analyses in SRs of therapeutic interventions. There is an urgent need for strategies to overcome these shortcomings. Strategies worth exploring include formal training of the biomedical and public health research workforce in statistical methods for SRs, involvement of statisticians on the SR team, and establishment of partnerships between methodologists and statisticians with specialist expertise in SR methods and journal editors.

Acknowledgments

We thank Sean Harrison (University of Bristol) for assistance with data analysis.

Competing Interests

1 I have read the journal's policy and the authors of this manuscript have the following competing
2 interests: ACT is an Associate Editor for *Journal of Clinical Epidemiology* but had no involvement in
3 the peer review process or decision for publication. MJP and JEM are affiliates of Cochrane Australia.
4 MJP is a Co-Convenor of the Cochrane Bias Methods Group. JEM is a Co-Convenor of the Cochrane
5 Statistical Methods Group. ACT is an author of two of the systematic reviews included in this study,
6 but was not involved in eligibility assessment or data collection.

7 8 **Funding**

9 There was no direct funding for this study. MJP is supported by an Australian National Health and
10 Medical Research Council (NHMRC) Early Career Fellowship (1088535). DGA is a National Institute
11 for Health Research Senior Investigator. JEM is supported by a NHMRC Australian Public Health
12 Fellowship (1072366). FCL is supported by the Generalitat Valenciana (PROMETEOII/2015/021). ACT
13 is funded by a Tier 2 Canada Research Chair in Knowledge Synthesis. DM is supported in part by a
14 University Research Chair, University of Ottawa. The funders had no role in study design, data
15 collection and analysis, decision to publish, or preparation of the manuscript.

16 17 **Author Contributions**

18 All authors declare to meet the ICMJE conditions for authorship. MJP, DGA and DM conceived the
19 study design. JEM and LS provided input into the study design. MJP, DGA, JEM, LS and DM selected
20 items for inclusion in the data collection form. MJP, NA, DW, and FY collected data. MJP undertook
21 the statistical analyses. MJP wrote the first draft of the article. All authors contributed to revisions of
22 the article. All authors approved the final version of the submitted article.

23 24 **Data availability**

25 The study protocol, data collection form, and the raw data and statistical analysis code for this study
26 are available on the Open Science Framework: <https://osf.io/523bq/>

References

1. Deeks JJ, Higgins JPT, Altman DG. 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, 2011. Available from www.cochrane-handbook.org.
2. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*. 2008;336(7658):1413-5.
3. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, 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(2):171-9.
4. McKenzie JE, Beller EM, Forbes AB. Introduction to systematic reviews and meta-analysis. *Respirology*. 2016;21(4):626-37.
5. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ*. 2017;356:j573.
6. Riley RD, Gates S, Neilson J, Alfirevic Z. Statistical methods can be improved within Cochrane pregnancy and childbirth reviews. *J Clin Epidemiol*. 2011;64(6):608-18.
7. Donegan S, Williams L, Dias S, Tudur-Smith C, Welton N. Exploring treatment by covariate interactions using subgroup analysis and meta-regression in cochrane reviews: a review of recent practice. *PLoS One*. 2015;10(6):e0128804.
8. Koletsi D, Valla K, Fleming PS, Chaimani A, Pandis N. Assessment of publication bias required improvement in oral health systematic reviews. *J Clin Epidemiol*. 2016;76:118-24.
9. Atakpo P, Vassar M. Publication bias in dermatology systematic reviews and meta-analyses. *J Dermatol Sci*. 2016;82(2):69-74.
10. Hedin RJ, Umberham BA, Detweiler BN, Kollmorgen L, Vassar M. Publication Bias and Nonreporting Found in Majority of Systematic Reviews and Meta-analyses in Anesthesiology Journals. *Anesth Analg*. 2016;123(4):1018-25.

1594
1595
1596 1 11. Herrmann D, Sinnett P, Holmes J, Khan S, Koller C, Vassar M. Statistical controversies in
1597
1598 2 clinical research: Publication bias evaluations are not routinely conducted in clinical
1599
1600 3 oncology systematic reviews. *Ann Oncol*. 2016.
1601
1602 4 12. Richardson M, Garner P, Donegan S. Cluster Randomised Trials in Cochrane Reviews:
1603
1604 5 Evaluation of Methodological and Reporting Practice. *PLoS One*. 2016;11(3):e0151818.
1605
1606 6 13. Nolan SJ, Hambleton I, Dwan K. The Use and Reporting of the Cross-Over Study Design in
1607
1608 7 Clinical Trials and Systematic Reviews: A Systematic Assessment. *PLoS One*.
1609
1610 8 2016;11(7):e0159014.
1611
1612 9 14. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and
1613
1614 10 Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional
1615
1616 11 Study. *PLoS Med*. 2016;13(5):e1002028.
1617
1618 12 15. Moher D, Shamseer L, Clarke M, Ghera D, Liberati A, Petticrew M, et al. Preferred reporting
1619
1620 13 items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst*
1621
1622 14 *Rev*. 2015;4(1):1.
1623
1624 15 16. Shamseer L, Moher D, Clarke M, Ghera D, Liberati A, Petticrew M, et al. Preferred reporting
1625
1626 16 items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and
1627
1628 17 explanation. *BMJ*. 2015;349:g7647.
1629
1630 18 17. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting
1631
1632 19 characteristics of systematic reviews. *PLoS Med*. 2007;4:e78.
1633
1634 20 18. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the
1635
1636 21 efficiency of study identification methods in systematic reviews. *Syst Rev*. 2016;5(1):140.
1637
1638 22 19. Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of
1639
1640 23 Cochrane Intervention Reviews. London: Cochrane; 2016.
1641
1642 24 20. Schriger DL, Altman DG, Vetter JA, Heafner T, Moher D. Forest plots in reports of systematic
1643
1644 25 reviews: a cross-sectional study reviewing current practice. *Int J Epidemiol*. 2010;39(2):421-
1645
1646 26 9.
1647
1648 27 21. StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP; 2015.
1649
1650
1651
1652

1653
1654
1655 1 22. The Nordic Cochrane Centre (The Cochrane Collaboration). Review Manager (RevMan). 5.1.
1656
1657 2 Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration; 2011.
1658
1659 3 23. R Development Core Team. R: a language and environment for statistical computing. Vienna,
1660
1661 4 Austria: R Foundation for Statistical Computing; 2012.
1662
1663 5 24. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses.
1664
1665 6 BMJ. 2003;327(7414):557-60.
1666
1667 7 25. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication
1668
1669 8 No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January
1670
1671 9 2014. Chapters available at: www.effectivehealthcare.ahrq.gov.
1672
1673
1674 10 26. IOM (Institute of Medicine). Finding What Works in Health Care: Standards for Systematic
1675
1676 11 Reviews. Washington, DC: The National Academies Press; 2011.
1677
1678 12 27. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ.
1679
1680 13 2011;342:d549.
1681
1682 14 28. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88.
1683
1684 15 29. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. Journal of
1685
1686 16 the Royal Statistical Society Series C-Applied Statistics. 2005;54:367-84.
1687
1688 17 30. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis
1689
1690 18 funnel plots help distinguish publication bias from other causes of asymmetry. J Clin
1691
1692 19 Epidemiol. 2008;61(10):991-6.
1693
1694 20 31. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple,
1695
1696 21 graphical test. BMJ. 1997;315(7109):629-34.
1697
1698 22 32. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for
1699
1700 23 examining and interpreting funnel plot asymmetry in meta-analyses of randomised
1701
1702 24 controlled trials. BMJ. 2011;343:d4002.
1703
1704 25 33. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. West
1705
1706 26 Sussex, UK: John Wiley & Sons, Ltd; 2009.
1707
1708
1709
1710
1711

1712
1713
1714 1 34. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ.
1715
1716 2 2003;326(7382):219.
1717
1718 3 35. Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic
1719
1720 4 Reviews and Meta-analyses. Milbank Q. 2016;94(3):485-514.
1721
1722 5 36. Petropoulou M, Nikolakopoulou A, Veroniki AA, Rios P, Vafaei A, Zarin W, et al. Bibliographic
1723
1724 6 study showed improving statistical methodology of network meta-analyses published
1725
1726 7 between 1999 and 2015. J Clin Epidemiol. 2017;82:20-8.
1727
1728 8 37. Hutton B, Wolfe D, Moher D, Shamseer L. Reporting guidance considerations from a
1729
1730 9 statistical perspective: overview of tools to enhance the rigour of reporting of randomised
1731
1732 10 trials and systematic reviews. Evid Based Ment Health. 2017.
1733
1734 11 38. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for
1735
1736 12 Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med.
1737
1738 13 2009;6:e1000097.
1739
1740 14 39. Page MJ, Moher D. Mass Production of Systematic Reviews and Meta-analyses: An Exercise
1741
1742 15 in Mega-silliness? Milbank Q. 2016;94(3):515-9.
1743
1744 16 40. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing
1745
1746 17 value and reducing waste in research design, conduct, and analysis. Lancet.
1747
1748 18 2014;383(9912):166-75.
1749
1750 19 41. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, Korevaar DA, et al. Increasing value
1751
1752 20 and reducing waste in biomedical research: who's listening? Lancet. 2015.
1753
1754 21 42. Int'Hout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction
1755
1756 22 intervals in meta-analysis. BMJ Open. 2016;6(7):e010247.
1757
1758 23
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770

Competing Interests

I have read the journal's policy and the authors of this manuscript have the following competing interests: ACT is an Associate Editor for *Journal of Clinical Epidemiology* but had no involvement in the peer review process or decision for publication. MJP and JEM are affiliates of Cochrane Australia. MJP is a Co-Convenor of the Cochrane Bias Methods Group. JEM is a Co-Convenor of the Cochrane Statistical Methods Group. ACT is an author of two of the systematic reviews included in this study, but was not involved in eligibility assessment or data collection.

Supplementary File

Data extraction form

Data in each systematic review were collected using the following form by one author. A 20% random sample of SRs was extracted in duplicate, which revealed several items that were discrepant between two reviewers on at least one occasion. Data for these items (indicated by an asterisk * below) were verified for all systematic reviews by one author.

Question	Response options
RefID	[Free text]
Your initials	[Free text]
1. Were any meta-analyses performed?	Yes No
2. What statistical software was used to perform meta-analyses? [Check all that apply]	RevMan Stata [If yes, please specify which package(s) were used, e.g. metan, metaan, metareg] R [If yes, please specify which package(s) were used, e.g. meta, metafor] Comprehensive Meta-Analysis (CMA) OpenMeta[Analyst] SAS SPSS WinBUGS Other (please specify) Not reported

Question	Response options
<p>3. *How many meta-analyses were reported (either in the main paper or online supplement)?</p> <p>Sum the number of meta-analyses reported either in forest plots, tables or text. If the same meta-analytic estimate is presented in a forest plot and in text, only count this estimate once.</p> <p>For subgroup analyses with an overall effect reported, count each subgroup meta-analytic effect as well as the overall effect.</p> <p>Count each meta-analysis regardless of whether data for it were fully reported. For example, if the authors present a meta-analysis on a forest plot and then state that they performed a sensitivity analysis for this particular meta-analysis, count each as a separate meta-analysis (i.e. n=2 meta-analyses in this instance).</p>	[Free text]
<p>4. *Which effect measures were used for binary outcomes?</p> <p>[Check all that apply]</p>	<p>Risk ratio for at least one meta-analysis</p> <p>Odds ratio for at least one meta-analysis</p> <p>Risk difference for at least one meta-analysis</p> <p>No meta-analyses of binary outcomes</p>
<p>5. Which effect measures were used for continuous outcomes?</p> <p>[Check all that apply]</p>	<p>Mean difference for at least one meta-analysis</p> <p>Standardized mean difference for at least one meta-analysis</p> <p>Ratio of means for at least one meta-analysis</p> <p>No meta-analyses of continuous outcomes</p>
<p>6. Which of the following other effect measures were synthesized in meta-analyses? [Check all that apply]</p>	<p>Hazard ratio for at least one meta-analysis</p> <p>Rate ratio for at least one meta-analysis</p> <p>Other (please specify)</p> <p>None</p>

Question	Response options
7. *Were data from randomized controlled trials (RCTs) and non-randomized studies of interventions (NRSI) synthesized in any meta-analyses, or were data analysed separately by study design?	<p>Data from RCTs and NRSI were synthesized in at least one meta-analysis</p> <p>Data from RCTs and NRSI were always analysed separately</p> <p>Only RCTs were included in the SR</p> <p>Only NRSI were included in the SR</p> <p>Design of included studies was unclear/not reported</p> <p>Unclear/not stated</p>
8. *Did the authors refer to a categorization system to interpret the I-squared statistic? [Check all that apply]	<p>Yes - Higgins 2003 BMJ (http://www.bmj.com/content/327/7414/557)</p> <p>Yes - Cochrane Handbook (http://handbook.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm)</p> <p>Yes - Other (please specify)</p> <p>No categorization system reported</p>
9. Did the authors report that a measure of statistical heterogeneity (e.g. Chi-squared, I-squared, tau-squared or any other) was used to justify use of a fixed-effect or random-effects meta-analysis model (e.g. stated something like "If I-squared was above 50% we used the random-effects model; if less than 50% we used the fixed-effect model")?	<p>Yes</p> <p>No</p>
10. What criteria for choosing between a fixed-effect or random-effects model were reported? For example, "If I-squared>50%, we performed random-effects meta-analyses".	[Free text]
11. Were any treatment-by-covariate interaction analyses (e.g. subgroup analyses or	Yes

Question	Response options
meta-regression) reported in the SR?	No
12. What type of interaction analyses were reported?	Only subgroup analyses Only meta-regression Both subgroup analyses and meta-regression None
13. *How many interaction analyses were presented in the Results section or online appendices?	[Free text]
14. Were any sensitivity analyses reported in the SR?	Yes No
15. *How many sensitivity analyses were presented in the Results section or online appendices?	[Free text]
16. Did the authors report that they generated a funnel plot?	Yes No
17. Was a funnel plot presented (in either the main paper or online appendix)?	Yes No
18. How many funnel plots were presented (in either the main paper or online appendix)?	[Free text]
19. Which of the following possible reasons for asymmetry in a funnel plot were provided? [Check all that apply]	Publication bias Small study effects Heterogeneity

Question	Response options
	Chance Risk of bias in the studies Other (please specify) None
20. Was a contour-enhanced funnel plot presented (in either the main paper or online appendix)?	Yes No
21. Was a statistical test for funnel plot asymmetry or small-study effects (e.g. Egger regression test) presented (in either the main paper or online appendix)?	Yes No
22. Which of the following possible reasons for a statistically significant test result were provided? [Check all that apply]	Publication bias Small study effects Heterogeneity Chance Risk of bias in the studies Other (please specify) None
23. What did the review authors conclude about the risk of publication bias?	Suspected for at least one meta-analysis (please specify the reason provided by the authors) Not suspected for any meta-analysis (please specify the reason provided by the authors) No quantitative assessment of the risk of publication bias

Question	Response options
24. *Did all funnel plots or statistical tests for funnel plot asymmetry include at least 10 studies of varying size?	Yes No Unsure
Answer the remaining questions for only one meta-analysis per review (the index meta-analysis) Select the first reported meta-analysis of the primary outcome. If no primary outcome was defined, select the first outcome listed in the Objectives. If no outcome is listed in the Objectives, select the first reported meta-analysis. Note that the first result may be identified from the Abstract or Results section of the review, depending on where it is first reported in the publication.	
25. What is the outcome (e.g. all-cause mortality, anxiety)?	[Free text]
26. Was the outcome labelled as 'primary' by the review authors?	Yes No
27. How many studies were included in the meta-analysis? State "Not reported" if not reported	[Free text]
28. What was the total number of participants included in the meta-analysis? State "Not reported" if not reported	[Free text]
29. What effect measure was used in the meta-analysis?	Risk ratio Odds ratio Risk difference Hazard ratio Rate ratio Mean difference

Question	Response options
	Standardized mean difference Ratio of means Other (please specify)
30. *What is the statistical significance of the meta-analytic intervention effect? If it is unclear which is the “intervention” and which is the “comparator” (e.g. because the authors compare two active interventions e.g. drug dose A versus drug dose B, or manual therapy versus exercise, WITHOUT specifying which one is the “active” intervention, select “Unclear”	Favourable, statistically significant (i.e. effect in favour of the intervention with $p \leq 0.05$) Favourable, non-statistically significant (i.e. effect in favour of the intervention with $p > 0.05$) Unfavourable, statistically significant (i.e. effect in favour of the comparator with $p \leq 0.05$) Unfavourable, non-statistically significant (i.e. effect in favour of the comparator with $p > 0.05$) Unclear Not reported Not applicable (e.g. no analysis of between-group difference)
31. Is the meta-analysis presented on a forest plot?	Yes No
32. *How are studies ordered/listed on the forest plot?	By study author By year of publication By effect size By sample size By risk of bias (or quality) criteria

Question	Response options
	Not reported
38. Which method was used to calculate the confidence interval for the summary effect?	<p>Not reported, but Wald-type method assumed (select when DerSimonian and Laird approach stated but there is no specific mention of an alternative method to calculate the confidence interval)</p> <p>Wald-type method</p> <p>t-distribution</p> <p>Hartung-Knapp</p> <p>Other (please specify)</p> <p>Not reported</p>
39. *Was the pooled result from the random-effects meta-analysis interpreted correctly (that is, as the average of the intervention effect across studies, not the best estimate of a common intervention effect)?	<p>Yes</p> <p>No</p>
40. Was a prediction interval reported for the meta-analysis?	<p>Yes</p> <p>No</p>
41. Was the pooled result from the meta-analysis and the corresponding fixed-effect meta-analysis reported?	<p>Yes</p> <p>No</p>
42. What was the I-squared value for the meta-analysis? State "Not reported" if not reported	[Free text]
43. What was the Chi-squared P-value for the meta-analysis? State "Not reported" if not reported	[Free text]
44. What was the tau-squared value for the meta-analysis? State "Not reported" if not reported	[Free text]

Question	Response options
45. Was a treatment-by-covariate interaction analysis (e.g. subgroup analysis or meta-regression) reported for this particular meta-analysis?	<p>Yes</p> <p>No</p> <p>Unsure</p>
46. *What type of interaction analyses for this particular meta-analysis did the authors report? [Check all that apply]	<p>Stratification based on study design (e.g. RCTs versus NRSI)</p> <p>Stratification based on year of publication</p> <p>Stratification based on patient characteristic (e.g. age, sex)</p> <p>Stratification based on intervention characteristic (e.g. dosage)</p> <p>Stratification based on outcome characteristic (e.g. measurement instrument, timing of follow-up)</p> <p>Stratification based on risk of bias (or “quality”)</p> <p>Stratification based on some other characteristic (please specify)</p>
47. *Were rationale provided for performing each interaction analysis? Only consider the interaction analyses relating to the index meta-analysis.	<p>Yes, for all interaction analyses</p> <p>Yes, but only for some interaction analyses</p> <p>No, not reported for any interaction analysis</p>
48. *In the interaction analyses reported, were there at least 10 studies per covariate? Only consider the interaction analyses relating to the index meta-analysis.	<p>Yes, for all interaction analyses</p> <p>Yes, but only for some interaction analyses</p> <p>No, not for any interaction analysis</p> <p>Unclear (number of studies per covariate not reported)</p>

Question	Response options
49. *Was due caution advised to readers due to the small number of studies per covariate? Only consider the interaction analyses relating to the index meta-analysis.	Yes No
50. *Was the issue of potential confounding across covariates raised? That is, do the authors acknowledge that the difference between subgroup effects may be wholly or partially accounted for by some other factor (for example, a difference in the effect of trials conducted in high income countries compared with trials conducted in low and middle-income countries may be confounded by administration of a co-intervention that is only available in high income countries).	Yes No
51. *Did authors interpret the result of the interaction analyses with reference to the test of interaction (e.g. P-value for the test of subgroup differences)? Only consider the interaction analyses relating to the index meta-analysis.	Yes, for all interaction analyses Yes, but only for some interaction analyses No, not for any interaction analysis
52. *[For reviews reporting subgroup analyses] Did authors draw inferences by comparing the statistical significance of each subgroup effect (i.e. infer a difference between subgroups because one had a statistically significant P-value while the other(s) had a non-significant P-value) Only consider the subgroup analyses relating to the index meta-analysis.	Yes, for all subgroup analyses Yes, but only for some subgroup analyses No, not for any subgroup analysis Not applicable – no subgroup analyses
53. *Did authors conclude that a subgroup effect existed or that a significant	Yes

Question	Response options
association was found using meta-regression? By “subgroup effect”, we mean that the authors concluded that the intervention effect differs across subgroups.	No
54. Was a sensitivity analysis reported for this particular meta-analysis?	Yes No Unsure
55. *What type of sensitivity analyses did the authors report for this particular meta-analysis? [Check all that apply]	Removal/inclusion of studies based on study design (e.g. removal of NRSI) Removal/inclusion of studies based on year of publication Removal/inclusion of studies based on patient characteristic (e.g. age, sex) Removal/inclusion of studies based on intervention characteristic (e.g. dosage) Removal/inclusion of studies based on outcome characteristic (e.g. measurement instrument, timing of follow-up) Removal/inclusion of studies based on risk of bias (“quality”) Removal/inclusion of studies with unpublished data Removal/inclusion of studies with imputed data (e.g. imputed standard deviations) Removal of each study one at a time Reanalysis using different model or assumptions (e.g. fixed-effect versus random-effects, or impute different set of missing standard deviations, or assume different within-subject correlation)

Question	Response options
	Other (please specify)
56. *Were rationale provided for performing each sensitivity analysis? Only consider the sensitivity analyses relating to the index meta-analysis.	Yes, for all sensitivity analyses Yes, but only for some sensitivity analyses No, not for any sensitivity analysis
57. *Did the authors conclude that the meta-analysis result was robust according to all sensitivity analyses? In other words, did the authors claim that there was no important change to the meta-analysis result in all sensitivity analyses?	Yes No
58. Was the meta-analytic effect estimate (e.g. risk ratio, mean difference) reported in the abstract?	Yes No
59. Was a confidence interval or P-value for the meta-analytic effect estimate reported in the abstract?	Yes – both a confidence interval and P-value reported Yes – only a confidence interval reported Yes – only a P-value reported No – neither reported
60. Was the number of studies contributing to the meta-analysis reported in the abstract?	Yes No
61. Was the number of participants contributing to the meta-analysis reported in the abstract?	Yes No

Question	Response options
General comments	[Free text]