The magnitude of small-study effects in the *Cochrane Database of Systematic Reviews*: an empirical study of nearly 30 000 meta-analyses

Lifeng Lin [ORCID],[1] **Linyu Shi,**[1] **Haitao Chu** [ORCID],[2]
Mohammad Hassan Murad [ORCID][3]

Abstract

Publication bias, more generally termed as small-study effect, is a major threat to the validity of meta-analyses. Most meta-analysts rely on the p values from statistical tests to make a binary decision about the presence or absence of small-study effects. Measures are available to quantify small-study effects' magnitude, but the current literature lacks clear rules to help evidence users in judging whether such effects are minimal or substantial. This article aims to provide rules of thumb for interpreting the measures. We use six measures to evaluate small-study effects in 29932 meta-analyses from the Cochrane Database of Systematic Reviews. They include Egger's regression intercept and the skewness under both the fixed-effect and random-effects settings, the proportion of suppressed studies, and the relative change of the estimated overall result due to small-study effects. The cut-offs for different extents of small-study effects are determined based on the quantiles in these distributions. We present the empirical distributions of the six measures and propose a rough guide to interpret the measures' magnitude. The proposed rules of thumb may help evidence users grade the certainty in evidence as impacted by small-study effects.

Introduction

Meta-analyses are powerful tools to combine and compare information from multiple sources and provide the most comprehensive evidence for decision-making. They have been frequently applied to facilitate evidence-based medicine, and innovative approaches have been increasingly developed to meet contemporary needs of decision-makers and overcome various challenges.[1][2] An essential difficulty in meta-analyses is to reduce potential bias from individual studies, as well as from the process of combining them. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement guides the conduct and reporting of the systematic review process,[3][4] the Cochrane Collaboration provides instructions on assessing the risk of bias of individual studies,[5][6] and the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group has established approaches to rating the certainty of evidence.[7][8] Most efforts aim to avoid unwarranted strong inference based on misleading or low-quality evidence.[9] Nevertheless, compared with handling bias within

studies, it is more challenging to effectively detect or even correct publication bias. Studies with statistically significant results or results in certain directions may be more likely published,[10–13] which can seriously bias meta-analytic conclusions.

An ideal method to remedy publication bias is to retrieve unpublished studies from various sources besides published journal articles. These sources include clinical trial registries, drug or device approving agencies, scientific conference abstracts and proceedings.[14] However, this method is not always feasible. For example, the unpublished sources may not be available, and the retrieval process may be cumbersome and time-consuming. Also, unpublished databases have not been peer reviewed, and their quality is not guaranteed. Even if some unpublished data were incorporated into a meta-analysis, the likelihood of publication bias remains. Therefore, when using such non-statistical methods to deal with publication bias, meta-analysts need to pursue a tradeoff between time, effort, costs and the importance of unpublished data.[15]

Consequently, in addition to searching for unpublished databases, statistical methods have been popular to assess publication bias.[16–18] These include Egger's regression,[19] the trim-and-fill method[20] and a recently proposed skewness of the collected studies.[21] Meta-analysts usually determine the presence or absence of publication bias based on these tests' significance while the bias direction and magnitude can also deliver critical information.[22] Although Egger *et al*[19] suggested to measure the bias direction and magnitude using their proposed regression's intercept, most meta-analysts reported only this regression test's p value, not the intercept itself. Such a measure has been insufficiently reported primarily because researchers lack an intuitive guide to interpret the measure's magnitude.

Of note, most statistical methods, including Egger's regression, are designed based on examining the funnel plot's asymmetry. Such asymmetry is sometimes termed as small-study effect, rather than publication bias, to remind meta-analysts that publication bias may not be the only cause of the asymmetry.[18] Many other factors may also lead to an asymmetrical funnel plot (eg, due to heterogeneity or simply by chance). Consequently, we use the term small-study effects throughout this article to more accurately describe the results of statistical methods. We cautiously note that the

statistical methods may only assist the assessment of small-study effects, rather than ascertaining publication bias. The funnel plot's asymmetry should be carefully interpreted on a case-by-case basis from both statistical and clinical perspectives, for example, using the comprehensive guidelines provided by Sterne *et al*.[18]

This article summarises six measures for small-study effects and presents their empirical distributions based on many meta-analyses in the *Cochrane Database of Systematic Reviews* (CDSR). We provide rules of thumb to quantitatively help decision-makers evaluate the magnitude of small-study effects.

Methods

Data collection

The Cochrane Collaboration provides a large collection of systematic reviews in healthcare. We searched for all reviews in the CDSR from 2003 Issue 1 to 2018 Issue 5 and downloaded their data on 27 May 2018 via the R package 'RCurl'. Each Cochrane review corresponded to a distinct healthcare-related topic and usually contained multiple meta-analyses on different outcomes and/or treatment comparisons. We pooled all meta-analyses from all reviews and classified them into two groups based on their outcomes (binary and non-binary). For meta-analyses with binary outcomes, regardless the effect sizes used in the original Cochrane reviews, we assessed their small-study effects based on (log) ORs so that the associated measures could be more consistent across meta-analyses. For meta-analyses with non-binary outcomes, we used the originally reported effect sizes (eg, mean differences, standardised mean differences, rate ratios) because it was impossible to transform them into a common type of effect size using the available data from the CDSR. Finally, it was difficult to justify small-study effects in meta-analyses with few studies,[16 23] so we only considered meta-analyses with at least five studies.

Meta-analyses in the same Cochrane review often shared some common studies, so their results were possibly correlated. In addition to the full database as collected above, we considered a reduced database by selecting the largest meta-analysis (with respect to the number of studies) from each Cochrane review. The

meta-analyses in the reduced database could be considered independent, and they were used as a sensitivity analysis.

Measures for small-study effects

We measured small-study effects using six methods briefly described in table 1 for each Cochrane meta-analysis. The first four measures were based on Egger's regression. We considered the regression intercept (originally introduced by Egger *et al*[19]) and the recently proposed skewness of the regression errors.[21] Egger's regression was originally described under the fixed-effect (FE) setting.[24] However, heterogeneity often existed between studies, and it was commonly modelled using random effects (RE) with additive between-study variances.[25 26] We considered the regression intercept and the skewness under both the FE and RE settings, denoted as T_I^{FE} and T_I^{RE}, and T_S^{FE} and T_S^{RE}, respectively. Under the RE setting, we first used the Q-test to examine the significance of heterogeneity. If its p value was <0.05, the between-study variance was estimated using the DerSimonian–Laird method[27]; otherwise, it was set to zero.

The other two measures were based on the trim-and-fill method,[20] which could use the observed n studies in a meta-analysis to impute the suppressed studies and thus adjust for small-study effects. We estimated the number of suppressed studies k_0, and the overall effect sizes before and after correcting small-study effects, denoted as $\theta_{published\ data}$ and $\theta_{total\ data}$, respectively. Consequently, small-study effects were measured using the proportion of suppressed studies, $P_{TF} = k_0 / \left(n + k_0 \right) \times 100\%$, and the relative change of the estimated overall effect size, $R_{TF} = \left(\theta_{published\ data} / \theta_{total\ data} - 1 \right) \times 100\%$.[14]

All six measures were theoretically zero when no small-study effects presented. The five measures except P_{TF} can be positive or negative; their signs indicated small-study effects' direction, and their absolute values implied the effects' magnitude. Positive regression intercepts or skewness indicated that studies with more negative results (ie, on the funnel plot's left side) tended

Table 1 Statistical methods for assessing small-study effects

Method	Description
Funnel plot	Presenting the study-specific effect size against its standard error (or the inverse of SE). It is roughly symmetrical around the overall effect size if no small-study effects appear.
Regression test	SND = $\alpha + \beta \times$ precision + error. Under the fixed-effect setting, SND (standard normal deviate) = y/s and precision = $1/s$; under the random-effects setting, SND = $y/\sqrt{s^2 + \tau^2}$ and precision = $1/\sqrt{s^2 + \tau^2}$. Here, y and S are the study-specific effect size and its standard error within studies, respectively, and τ^2 is between-study variance due to heterogeneity. It tests for whether $\alpha = 0$.
Regression intercept (T_I^{FE} or T_I^{RE})	An estimate of the intercept α of the regression test under the fixed-effect (T_I^{FE}) or random-effects (T_I^{RE}) setting.
Skewness (T_S^{FE} or T_S^{RE})	An estimate of the skewness of the study-specific errors of the regression test under the fixed-effect (T_S^{FE}) or random-effects (T_S^{RE}) setting.
Trim-and-fill method	Estimating the suppressed studies and thus correcting small-study effects based on funnel plot's asymmetry.
Proportion of suppressed studies (P_{TF})	$P_{TF} = k_0 / \left(n + k_0 \right) \times 100\%$, where n is the number of studies in the original meta-analysis, and k_0 is the estimated number of suppressed studies using the trim-and-fill method.
Relative change of overall result by incorporating imputed suppressed studies (R_{TF})	$R_{TF} = \left(\theta_{published\ data} / \theta_{total\ data} - 1 \right) \times 100\%$, where $\theta_{published\ data}$ is the estimated overall result in the original meta-analysis of published studies, and $\theta_{total\ data}$ is that after incorporating imputed suppressed studies using the trim-and-fill method.
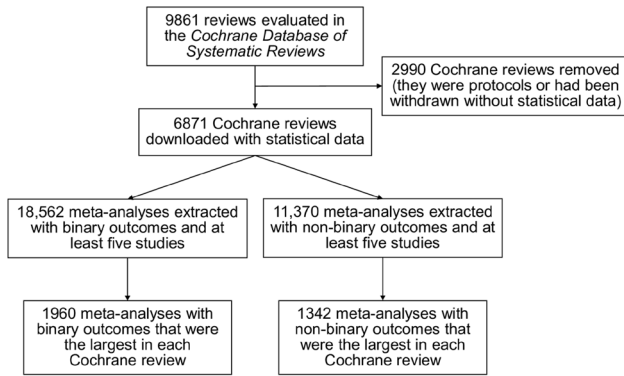
Figure 1 Flow chart of selecting the Cochrane meta-analyses.

to be suppressed; negative ones indicated that suppressed studies tended to be in the positive direction. When using R_{TF}, determining the small-study effects' direction depended on the overall effect size's direction. Moreover, unlike the above five measures, P_{TF} lied within 0%–100%; it informed only small-study effects' magnitude, not their direction.

Deriving rules of thumb for small-study effects' magnitudes
The six measures' empirical distributions were obtained using the Cochrane meta-analyses. We determined the cut-offs for different magnitudes of small-study effects based primarily on the quantiles in these distributions. We roughly classified the magnitudes into four levels: unimportant, moderate, substantial and considerable; each level contained roughly 30% of the distribution. Also, we permitted the levels' ranges to overlap because the importance of small-study effects depended on many factors (eg, disease outcome) and strict cut-offs without overlapped ranges might be misleading. Also, we rounded the selected quantiles to be in simple forms (with few digits after the decimal point) so that they could be easily summarised and applied. These labels of magnitudes and the approach of overlapping categories have been similarly used with the I^2 statistic for heterogeneity.[6 28]

Results
We obtained 18 562 eligible meta-analyses with binary outcomes and 11 370 ones with non-binary outcomes. The reduced database contained 1960 and 1342 meta-analyses with binary and non-binary outcomes, respectively. Figure 1 presents the flow chart of the selection process.

Meta-analyses with binary outcomes
Figures S1 and S2 in the Supplementary Material present the empirical distributions of measures for small-study effects and their absolute values (except P_{TF}) among meta-analyses with binary outcomes. The vertical dashed lines depict the null value (zero) of no small-study effects. The distributions of Egger's regression intercept and the skewness were approximately symmetrical around zero. The averages of the regression intercepts T_I^{FE} and T_I^{RE} were −0.14 and −0.16 with SDs 2.04 and 3.30, respectively; both had a median around −0.14. Also, 55.2% and 55.0% meta-analyses had negative T_I^{FE} and T_I^{RE}, respectively, and 44.8% and 45.0% meta-analyses had positive ones. The regression intercept was extreme in a few meta-analyses. Specifically, 279 (1.5%) and 254 (1.4%) meta-analyses had T_I^{FE} less than −4 and greater than 4, respectively. Using the regression intercepts under the RE setting, more meta-analyses had extreme measures: T_I^{RE} was less

than −4 in 560 (3.0%) meta-analyses and was greater than 4 in 495 (2.7%) meta-analyses.

Compared with the regression intercept, the skewness was more concentrated and symmetrical around zero. The averages of both T_S^{FE} and T_S^{RE} were close to zero with SDs around 0.50; their medians were also near zero. Moreover, 49.9% and 49.8% of the meta-analyses had negative T_S^{FE} and T_S^{RE}, respectively, and 50.1% and 50.2% had positive ones. Only 14 and 11 (<0.1%) meta-analyses had T_S^{FE} and T_S^{RE} greater than 2 in absolute magnitude.

When calculating P_{TF} and R_{TF}, the trim-and-fill algorithm did not converge in one meta-analysis with binary outcomes. Also, it did not identify any suppressed studies in 6099 (32.9%) meta-analyses; thus, their P_{TF} and R_{TF} were exactly zero and their empirical distributions had high frequencies massed at zero. The average of P_{TF} was 13.6% with SD 11.6%; its median was 14.3%. The P_{TF} was less than 40% in all meta-analyses. The average of R_{TF} was 57.0% with a huge SD 8087.9%, and its median was zero. The R_{TF} was negative in 28.9% meta-analyses and was positive in 38.2% meta-analyses. The empirical distribution of the absolute R_{TF} had a decreasing trend, which was similar to those of the absolute regression intercept and skewness while its frequency at zero was much higher. The empirical distribution of P_{TF} did not have an obvious decreasing trend as the measure increased.

Meta-analyses with non-binary outcomes
Figures S3 and S4 in the Supplementary Material present the measures' empirical distributions among meta-analyses with non-binary outcomes. Compared with those with binary outcomes, online supplementary figure S3 indicates that more meta-analyses had extreme measures for small-study effects, especially when using the regression intercept. The averages of T_I^{FE} and T_I^{RE} were −0.26 and −0.40 with SDs 2.74 and 17.89, respectively; their medians were around −0.24. The large SD of T_I^{RE} was due to its distribution's long tails. Many meta-analyses had extreme regression intercepts. Moreover, 56.2% and 55.8% meta-analyses had negative T_I^{FE} and T_I^{RE}, respectively, and 43.8% and 44.1% had positive ones. In one meta-analysis, all studies had equal effect sizes, so all measures for small-study effects were exactly zero.

The skewness was mostly lied between −2 and 2. The T_S^{FE} and T_S^{RE} had averages around −0.04 with SDs 0.64 and 0.61, respectively, and their medians were around −0.03. The T_S^{FE} and T_S^{RE} were negative in 52.4% and 52.5% meta-analyses and were positive in 47.6% and 47.4% ones, respectively. Only 1.1% and 0.8% meta-analyses had T_S^{FE} and T_S^{RE} greater than 2 in absolute magnitude. These proportions were slightly greater than those with binary outcomes.

The trim-and-fill method did not identify any suppressed studies in 5168 (45.5%) meta-analyses, whose P_{TF} and R_{TF} were zero. The P_{TF} had an average 11.0% with SD 11.6%, and its median was 10.0%. It was also less than 40% in all meta-analyses as in those with binary outcomes. The average of R_{TF} was 4.4% with SD 838.6%, and its median was 0%. The R_{TF} was negative in 26.1% and positive in 28.4% meta-analyses, respectively.

Rules of thumb for magnitudes of small-study effects
Table 2 provides rules of thumb for interpreting small-study effects' magnitudes based on the measures' empirical distributions. In absolute magnitude, small-study effects might be unimportant if the regression intercepts T_I^{FE} and T_I^{RE}, the skewness T_S^{FE} and T_S^{RE}, the proportion of suppressed studies P_{TF}, and the

Table 2 Proposed rules of thumb for assessing small-study effects' magnitudes

Measure	Database	Outcome	Measure range and proportion of Cochrane meta-analyses contained in the range among the full/reduced database with binary/non-binary outcomes			
			Unimportant small-study effects	Moderate small-study effects	Substantial small-study effects	Considerable small-study effects
T_I^{FE}			$\left|T_I^{FE}\right| \leq 0.6$	$0.4 \leq \left|T_I^{FE}\right| \leq 1$	$0.8 \leq \left|T_I^{FE}\right| \leq 2$	$\left|T_I^{FE}\right| \geq 1.8$
	Full	Binary	39.6%	31.2%	34.3%	18.7%
	Reduced	Binary	38.5%	31.1%	35.4%	18.7%
	Full	Non-binary	28.5%	24.7%	34.0%	32.8%
	Reduced	Non-binary	26.7%	24.1%	35.5%	33.9%
T_S^{FE}			$\left|T_S^{FE}\right| \leq 0.2$	$0.15 \leq \left|T_S^{FE}\right| \leq 0.4$	$0.35 \leq \left|T_S^{FE}\right| \leq 0.7$	$\left|T_S^{FE}\right| \geq 0.65$
	Full	Binary	32.0%	35.3%	30.7%	18.8%
	Reduced	Binary	32.2%	32.9%	29.9%	21.0%
	Full	Non-binary	28.8%	31.1%	30.2%	25.7%
	Reduced	Non-binary	26.8%	29.8%	30.8%	27.9%
T_I^{RE}			$\left|T_I^{RE}\right| \leq 0.6$	$0.4 \leq \left|T_I^{RE}\right| \leq 1$	$0.8 \leq \left|T_I^{RE}\right| \leq 2$	$\left|T_I^{RE}\right| \geq 1.8$
	Full	Binary	38.8%	30.2%	32.2%	21.8%
	Reduced	Binary	37.0%	29.9%	33.4%	22.8%
	Full	Non-binary	27.8%	23.3%	29.9%	38.2%
	Reduced	Non-binary	25.4%	21.9%	31.0%	40.5%
T_S^{RE}			$\left|T_S^{RE}\right| \leq 0.2$	$0.15 \leq \left|T_S^{RE}\right| \leq 0.4$	$0.35 \leq \left|T_S^{RE}\right| \leq 0.7$	$\left|T_S^{RE}\right| \geq 0.65$
	Full	Binary	32.3%	35.2%	30.7%	18.7%
	Reduced	Binary	33.0%	32.8%	30.1%	19.9%
	Full	Non-binary	29.9%	31.7%	30.5%	24.2%
	Reduced	Non-binary	27.8%	31.5%	31.4%	26.2%
P_{TF}			$P_{TF} \leq 4\%$	$2\% \leq P_{TF} \leq 18\%$	$16\% \leq P_{TF} \leq 26\%$	$P_{TF} \geq 24\%$
	Full	Binary	33.2%	31.1%	26.0%	24.6%
	Reduced	Binary	31.8%	31.9%	25.8%	23.2%
	Full	Non-binary	45.7%	25.3%	22.3%	19.6%
	Reduced	Non-binary	42.9%	28.0%	23.9%	17.7%
R_{TF}			$\left|R_{TF}\right| \leq 2\%$	$1\% \leq \left|R_{TF}\right| \leq 25\%$	$20\% \leq \left|R_{TF}\right| \leq 80\%$	$\left|R_{TF}\right| \geq 75\%$
	Full	Binary	34.0%	28.8%	26.3%	17.3%
	Reduced	Binary	32.0%	30.4%	28.5%	16.1%
	Full	Non-binary	47.2%	26.2%	21.7%	10.5%
	Reduced	Non-binary	44.2%	29.1%	21.9%	10.6%

relative change R_{TF} were less than 0.6, 0.2, 4% and 2%, respectively. Around 25%–45% meta-analyses with either binary or non-binary outcomes had measures within these ranges. The P_{TF} and R_{TF} generally indicated unimportant small-study effects in more meta-analyses than the other measures. If the six measures' absolute values were within 0.4–1, 0.15–0.4, 2%–18% and 1%–25% accordingly, small-study effects might be moderate; around 20%– 35% Cochrane meta-analyses had measures within the ranges. Small-study effects might be substantial if the absolute measures were within 0.8–2, 0.35–0.7, 16%–26% and 20%–80% accordingly, and might be considerable if they were at least 1.8, 0.65, 24% and 75% accordingly. The regression intercepts and skewness indicated substantial small-study effects in around 30%–35% meta-analyses, and considerable small-study effects in around 20%–40% ones. The P_{TF} and R_{TF} implied substantial and considerable small-study effects in less meta-analyses than the other measures.

Sensitivity analysis

Figures S5-S8 in the Supplementary Material show the measures' histograms among the reduced database. The shapes of the histograms were similar to their counterparts in Figures S1-S4 using the full database.

Discussion

Main findings

We have presented the rules of thumb for interpreting measures for small-study effects based on their empirical distributions among many Cochrane meta-analyses. The results can aid decision-makers who are tasked with appraising evidence and determining the certainty warranted by the evidence. Using the GRADE approach, small-study effects lead to rating down certainty. Using the proposed rules of thumb, the certainty in evidence to support a decision may not need to be rated down if small-study effects' magnitude was small or if the effects' direction was inconsistent with exaggerated study results.[22]

As Egger's regression test is often considered more statistically powerful than the trim-and-fill method in many situations,[16] the regression intercept and skewness may be preferred measures than the trim-and-fill-based P_{TF} and R_{TF}, although the latter two may be more straightforward than the former ones. Also, despite its intuitive interpretation as the proportion of suppressed studies,

P_{TF} has several drawbacks. First, it accounts only for the number of suppressed studies, not their weights in the estimated overall result. If the suppressed studies are small, small-study effects may have little influence on the meta-analysis, and P_{TF} might exaggerate small-study effects. Second, P_{TF} does not take continuous values like other measures; instead, it must be fractions. If a meta-analysis has five studies, P_{TF} is at least 16.7% (1/6 x 100%) even when only one suppressed study is identified.

Moreover, the regression intercept might be extreme, possibly because of outliers while the skewness was mostly within a reasonable range. Under the RE setting, the regression intercept was more extreme in some meta-analyses. This may be due to poor between-study variance estimates affected by small-study effects.[29]

Limitations

This article has several limitations. First, many limitations are inherent to the available statistical methods for small-study effects. These methods are often underpowered and require strong assumptions. For example, the trim-and-fill method may perform poorly in the presence of high heterogeneity or when its assumption about suppressed studies is seriously violated.[30 31] In such cases, meta-analysts may avoid using P_{TF} and R_{TF} to quantify small-study effects. Also, Egger's regression may have inflated false-positive rates in certain cases.[17] The inflation is due to the intrinsic association between effect sizes and their within-study variances; such effect sizes include ORs, risk ratios, risk differences, as well as standardised mean differences.[17 32–37] These issues need to be carefully taken into account when exploring publication bias. Although alternative methods may control false-positive rates better than Egger's regression in some cases,[17] they may be also seriously underpowered and not applicable to generic effect sizes. As an illustrative article for interpreting small-study effects' magnitude based on empirical evidence, we did not comprehensively consider all available alternative methods; assessing the measures for small-study effects based on the alternatives will be one of our future studies.

Second, our analysis was restricted to the Cochrane meta-analyses, which focused on healthcare-related topics. However, many meta-analyses were performed to investigate diverse (eg, ecological and educational) topics.[2] Therefore, our findings may not be directly generalised to those meta-analyses.

Third, we considered Cochrane meta-analyses containing at least five studies because it was difficult to justify small-study effects in very small meta-analyses. However, it is infeasible to establish a widely accepted eligibility criterion on the number of studies for appropriately assessing small-study effects. The cut-off may be some other values, say 10.[18]

Finally, although we classified small-study effects' magnitudes into four levels, their true importance relates to the disease outcome's type and its context. For example, unimportant increase in mortality may be important in certain meta-analyses. Furthermore, our classifications were based roughly on the quantiles of the measures' empirical distributions; however, no gold standard of small-study effects can be feasibly applied to examine their true magnitudes in all meta-analyses in the large database. Therefore, the proposed rules of thumb may serve as auxiliary assessment of small-study effects; to ascertain such effects or publication bias, meta-analysts should carefully investigate individual studies on a case-by-case basis.[18]

ORCID iDs
Lifeng Lin http://orcid.org/0000-0002-3562-9816
Haitao Chu http://orcid.org/0000-0003-0932-598X
Mohammad Hassan Murad http://orcid.org/0000-0001-5502-5975

References

1 **Sutton AJ**, Higgins JP. Recent developments in meta-analysis. *Stat Med* 2008;**27**:625–50.

2 **Gurevitch J**, Koricheva J, Nakagawa S, *et al*. Meta-analysis and the science of research synthesis. *Nature* 2018;**555**:175–82.

3 **Moher D**, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;**6**:e1000097.

4 **Stewart LA**, Clarke M, Rovers M, *et al*. Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;**313**:1657–65.

5 **Higgins JP**, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928.

6 **Higgins JPT**, Green S. *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: John Wiley & Sons, 2008.

7 **Guyatt GH**, Oxman AD, Vist G, *et al*. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;**64**:407–15.

8 **Hultcrantz M**, Rind D, Akl EA, *et al*. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;**87**:4–13.

9 **Bonner A**, Alexander PE, Brignardello-Petersen R, *et al*. Applying GRADE to a network meta-analysis of antidepressants led to more conservative conclusions. *J Clin Epidemiol* 2018;**102**:87–98.

10 **Chavalarias D**, Wallach JD, Li AH, *et al*. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA* 2016;**315**:1141–8.

11 **Cristea IA**, Ioannidis JPA. P values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PLoS One* 2018;**13**:e0197440.

12 **Dwan K**, Altman DG, Arnaiz JA, *et al*. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008;**3**:e3081.

13 **Dickersin K**. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990;**263**:1385–9.

14 **Eyding D**, Lelgemann M, Grouven U, *et al*. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ* 2010;**341**:c4737.

15 **Schmucker CM**, Blümle A, Schell LK, *et al*. Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLoS One* 2017;**12**:e0176210.

16 **Jin ZC**, Zhou XH, He J. Statistical methods for dealing with publication bias in meta-analysis. *Stat Med* 2015;**34**:343–60.

17 **Peters JL**, Sutton AJ, Jones DR, *et al*. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006;**295**:676–80.

18 **Sterne JA**, Sutton AJ, Ioannidis JP, *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002.

19 **Egger M**, Davey Smith G, Schneider M, *et al*. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34.

20 **Duval S**, Tweedie R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* 2000;**95**:89–98.

21 **Lin L**, Chu H. Quantifying publication bias in meta-analysis. *Biometrics* 2018;**74**:785–94.

22 **Murad MH**, Chu H, Lin L, *et al.* The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evid Based Med* 2018;**23**:84–6.

23 **Higgins JPT**, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;**172**:137–59.

24 **Moreno SG**, Sutton AJ, Ades AE, *et al.* Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009;**9**:2.

25 **Higgins JP**. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008;**37**:1158–60.

26 **Riley RD**, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;**342**:d549.

27 **DerSimonian R**, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88.

28 **Higgins JP**, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.

29 **Jackson D**. The implications of publication bias for meta-analysis' other parameter. *Stat Med* 2006;**25**:2911–21.

30 **Terrin N**, Schmid CH, Lau J, *et al.* Adjusting for publication bias in the presence of heterogeneity. *Stat Med* 2003;**22**:2113–26.

31 **Peters JL**, Sutton AJ, Jones DR, *et al.* Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Stat Med* 2007;**26**:4544–62.

32 **Macaskill P**, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 2001;**20**:641–54.

33 **Lin L**. Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One* 2018;**13**:e0204056.

34 **Zwetsloot PP**, Van Der Naald M, Sena ES, *et al.* Standardized mean differences cause funnel plot distortion in publication bias assessments. *Elife* 2017;**6**:e24260.

35 **Hamman EA**, Pappalardo P, Bence JR, *et al.* Bias in meta-analyses using Hedges' d. *Ecosphere* 2018;**9**:e02419.

36 **Doncaster CP**, Spake R. Correction for bias in meta-analysis of little-replicated studies. *Methods Ecol Evol* 2018;**9**:634–44.

37 **Pustejovsky JE**, Rodgers MA. Testing for funnel plot asymmetry of standardized mean differences. *Res Synth Methods* 2019;**10**:57–71.