


RESEARCH ARTICLE

New metrics for meta-analyses of heterogeneous effects

Maya B. Mathur^{1,2}  | Tyler J. VanderWeele¹

¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

²Quantitative Sciences Unit, Stanford University, Palo Alto, California

Correspondence

Maya B. Mathur, Quantitative Sciences Unit (c/o Inna Sayfer), Stanford University, 1070 Arastradero Road, Palo Alto, CA 94305.
Email: mmathur@stanford.edu

Funding information

National Defense Science and Engineering Graduate Fellowship, Grant/Award Number: 32 CFR 168a; National Institutes of Health, Grant/Award Number: CA222147 and ES017876

We provide two simple metrics that could be reported routinely in random-effects meta-analyses to convey evidence strength for scientifically meaningful effects under effect heterogeneity (ie, a nonzero estimated variance of the true effect distribution). First, given a chosen threshold of meaningful effect size, meta-analyses could report the estimated proportion of true effect sizes above this threshold. Second, meta-analyses could estimate the proportion of effect sizes below a second, possibly symmetric, threshold in the opposite direction from the estimated mean. These metrics could help identify if (1) there are few effects of scientifically meaningful size despite a “statistically significant” pooled point estimate, (2) there are some large effects despite an apparently null point estimate, or (3) strong effects in the direction opposite the pooled estimate also regularly occur (and thus, potential effect modifiers should be examined). These metrics should be presented with confidence intervals, which can be obtained analytically or, under weaker assumptions, using bias-corrected and accelerated bootstrapping. Additionally, these metrics inform relative comparison of evidence strength across related meta-analyses. We illustrate with applied examples and provide an R function to compute the metrics and confidence intervals.

KEYWORDS

effect sizes, heterogeneity, reporting

1 | INTRODUCTION

Random-effects meta-analyses aggregate evidence across studies measuring heterogeneous effects. Reporting usually focuses on the estimated mean of the distribution of true effects. However, under heterogeneity, others caution against exclusive focus on the estimated mean and recommend also reporting the estimated variance of true effects, not only the proportion of variance attributable to heterogeneity.^{1,2} Summarizing evidence strength by comparing only the estimated mean to a threshold of scientific importance is needlessly dichotomous and has, in the past, led authors to conflicting conclusions in meta-analyses reporting nearly identical point estimates (see the works of Kirsch et al³ and Turner et al⁴ with additional commentary on dichotomization by Turner and Rosenthal⁵). Others have proposed informative summary metrics that consider heterogeneity by characterizing the range of effects in the distribution or providing a prediction interval for a new effect in the population or in a subgroup.^{1,2,6-8}

Extending these previous recommendations regarding heterogeneity in meta-analyses, we recommend also considering questions such as “How common are effects of a size that is scientifically meaningful?” along with the traditional question

.....
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2018 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

“What is the average effect size?” We therefore propose simple metrics that, unlike traditional metrics, directly address the former question. These metrics make inference to the population of true effect sizes while accounting for differences in precision across studies. As discussed below, this is distinct from simply “vote counting” the significant p values or the observed effect sizes (ie, those measured with statistical error in the meta-analyzed studies) stronger than a threshold because the vote-counting approach, unlike our metrics, does not account for differences in precision and sample size across studies.

Reporting these new metrics along with traditional metrics may better identify situations in which (1) a “statistically significant” point estimate obscures the fact that the population of effect sizes contains few of scientifically meaningful size or, conversely, (2) a null point estimate disguises the fact that, due to heterogeneity, there are large effect sizes in some settings. Additionally, these metrics help identify whether the treatment or exposure of interest may have scientifically meaningful effects in the opposite direction in some populations, pointing to the need to further examine potential effect modifiers. Lastly, they can inform comparison of relative evidence strength across related meta-analyses.

2 | METHODS

We make two recommendations to better characterize heterogeneous effect sizes represented by a given literature. First, in meta-analyses with heterogeneity (ie, a nonzero estimate of τ , the standard deviation of the true effects), the investigator could select a threshold above which an effect size might be considered of scientifically meaningful size and could report the estimated proportion of true effect sizes exceeding this threshold, along with a confidence interval. There is a large, interdisciplinary literature considering how to choose such thresholds, which we briefly summarize in the Online Supplement. For example, if the effect sizes are relative risks, then depending on the scientific context and informed by considerations such as those described in the Online Supplement, we might consider only those above 1.1 to be scientifically meaningful. Again, depending on context, if fewer than (for example) 10% of true effect sizes surpass this threshold, we might consider evidence strength for scientifically meaningful effect sizes to be fairly weak.

Second, given a chosen threshold of scientific importance, we recommend that meta-analyses report the proportion of effect sizes stronger than a second (possibly symmetric) threshold on the opposite side of the null hypothesis (eg, $1/1.1 \approx 0.9$). That is, we might estimate a pooled relative risk (RR) of 1.2, indicating that the mean effect is positive. Yet with enough heterogeneity, the population of effects may also contain a nonnegligible proportion of strong *inverse* associations (for example, 18% below a symmetric RR of $1/1.1 \approx 0.9$). In practice, for both metrics, it may be informative to report results at more than one threshold of scientific importance as well as to report the meta-analytic estimates (particularly inference on the heterogeneity estimate) required to allow a reader to compute our proposed metrics for an arbitrary threshold.

These two proportions with their confidence intervals can be computed using only the estimated mean, heterogeneity, and their standard errors from a random-effects meta-analysis fit with any estimation approach that yields unbiased estimates of the mean and variance of the effect distribution. Specifically, under standard assumptions of parametric random-effects meta-analysis,⁹ studies have true effect sizes* θ_i that are independently normal with a grand mean μ and variance τ^2 . Suppose we estimate μ with a point estimate denoted $\hat{\mu}$ and estimate τ^2 using one of several heterogeneity estimators,¹⁰ denoted $\hat{\tau}^2$. If $\hat{\mu}$ is above the null value (eg, an estimated log-OR of $\log(1.3)$), then the first proposed metric, namely, the estimated proportion of studies in the population with effect size greater than q (eg, $q = \log(1.1)$), is simply

$$\hat{P}(\theta > q) = 1 - \Phi\left(\frac{q - \hat{\mu}}{\sqrt{\hat{\tau}^2}}\right), \hat{\tau}^2 > 0, \quad (1)$$

where Φ denotes the standard normal cumulative distribution function. For the second metric, we can estimate the proportion below a second threshold q^* as

$$\hat{P}(\theta < q^*) = \Phi\left(\frac{q^* - \hat{\mu}}{\sqrt{\hat{\tau}^2}}\right), \hat{\tau}^2 > 0. \quad (2)$$

*Here, we reiterate the crucial distinction between the “true” effect sizes, which are unobservable statistical parameters, and the observed effect sizes measured with statistical error in each study. Variability in the true effect sizes comprises only heterogeneity, whereas variability in the observed effect sizes comprises both statistical error and heterogeneity. Thus, to estimate the proportion of true effects above or below a threshold, we must first use the observed effects to estimate the distribution of the true effects through the standard random-effects model.

In either case, asymptotic 95% confidence interval limits are[†]

$$\hat{P} \pm 1.96 \sqrt{\frac{\widehat{\text{Var}}(\hat{\mu})}{\hat{\tau}^2} + \frac{\widehat{\text{Var}}(\hat{\tau}^2)(\hat{\mu} - q)^2}{4(\hat{\tau}^2)^3}} \cdot \phi\left(\frac{q - \hat{\mu}}{\sqrt{\hat{\tau}^2}}\right), \quad (3)$$

where ϕ denotes the standard normal density function. \hat{P} in this expression can be computed from either Equation (1) or Equation (2), and q is replaced by q^* when considering the second metric. Simulation results (Online Supplement) indicate that when the number of studies is less than 10, $\hat{P} < 0.15$, or $\hat{P} > 0.85$, it is preferable to estimate the confidence interval via bias-corrected, accelerated bootstrapping.^{11,12} Applied examples in this paper for which $\hat{P} < 0.15$ or $\hat{P} > 0.85$ use bootstrapped confidence intervals, for which code is available online (<https://osf.io/pr2s9/>).

If $\hat{\mu}$ is below rather than above the null value (for example, we estimate a mean log-OR of $\log(0.85)$), then we would simply switch the two equations, using Equation (2) to estimate the proportion of effects below a threshold (eg, a log-OR of $\log(0.90)$) and Equation (1) to estimate the proportion of effects above a second threshold. (As usual, we use the log scale when considering odds ratios for approximate normality.) In practice, these proportions are easy to compute manually or using the R function `prop_stronger`, available in an open-source public repository (<https://osf.io/pr2s9/>). We now illustrate how they can facilitate interpretation through three examples.

3 | EXAMPLE 1: A “SIGNIFICANT” MEAN DESPITE LITTLE EVIDENCE FOR STRONG EFFECTS

Meta-analyses often achieve large pooled sample sizes and high power, and thus, it can happen that a very small estimated mean attains “statistical significance” at a given significance threshold. The proposed metrics may then illustrate that, despite a “significant” p value, few effects are in fact strong enough to warrant scientific interest. For example, a recent meta-analysis¹³ estimated a mean correlation of $r = -0.06$ ($p = 0.01$) between increased psychological stress and shorter telomeres, and subsequent literature largely interpreted this finding as supportive of an association. However, using the proposed methods, we can estimate that only 6% (95% CI: 0%, 71%) of true correlations between stress and telomere length are stronger than the modest threshold of $r = -0.10$ and that almost none (0%, 95% CI: 0%, 46%) is stronger than $r = -0.20$. It is important to note that the upper confidence interval limits for both choices of threshold represent nonnegligible proportions, so these results should not be interpreted as evidence *against* the occurrence of correlations stronger than these thresholds. The 95% prediction interval is $(-0.13, 0.02)$; this additionally suggests that, with high probability, a new effect drawn from the population will be near the null.^{6,7} These metrics perhaps better qualify the scientific relevance of the meta-analyzed studies. Note that comparing the point estimate alone to the threshold would not adequately convey evidence strength. This meta-analysis estimated little heterogeneity ($\tau = 0.03$), and hence, the small point estimate suggests a small proportion of strong true effects; however, if the heterogeneity estimate had been larger (for example, three-fold larger on the τ scale), then the estimated percent of true correlations stronger than $r = -0.10$ would increase from a negligible 6% to 30% (95% CI: 13%, 46%).[‡]

4 | EXAMPLE 2: META-ANALYSES WITH DIFFERENT CONCLUSIONS DESPITE SIMILAR EVIDENCE STRENGTH

These metrics can also inform comparison across meta-analyses, which is of interest when a meta-analysis is updated to reflect new literature, when applying different inclusion criteria or analysis methods to the same literature, or when

[†]This standard error applies for estimators $\hat{\tau}^2$ that are asymptotically normal and independent of $\hat{\mu}$. This holds, for example, for the maximum likelihood estimators under the assumption that $E[\hat{\theta}_i | \sigma_i^2] = E[\hat{\theta}_i]$, where $\hat{\theta}_i$ and σ_i^2 respectively denote the point estimate and squared standard error of the i th study (Online Supplement). In practice, this assumption can be verified by inspecting a funnel plot for symmetry. (Note that this is a weaker assumption than independence of $\hat{\theta}_i$ from σ_i^2 , which certainly does not hold by the definition of σ_i^2 .) When these assumptions are violated, the bias-corrected and accelerated confidence interval should be used instead.

[‡]It may appear surprising that the latter confidence interval under increased heterogeneity is narrower than the confidence interval with the observed estimates. This occurs for two reasons. First, $\widehat{\text{Var}}(\hat{\tau}^2)$ is not estimable for the hypothetical example with increased heterogeneity, so we held it constant to its observed value, resulting in a relatively small $\widehat{\text{Var}}(\hat{\tau}^2)$ compared to τ^2 . Second, an estimated true effect distribution with increased heterogeneity has flatter tails, resulting in more stable estimates of $\hat{P}(\theta < q^*)$ or $\hat{P}(\theta > q)$ in some neighborhoods of $\hat{\mu}$.

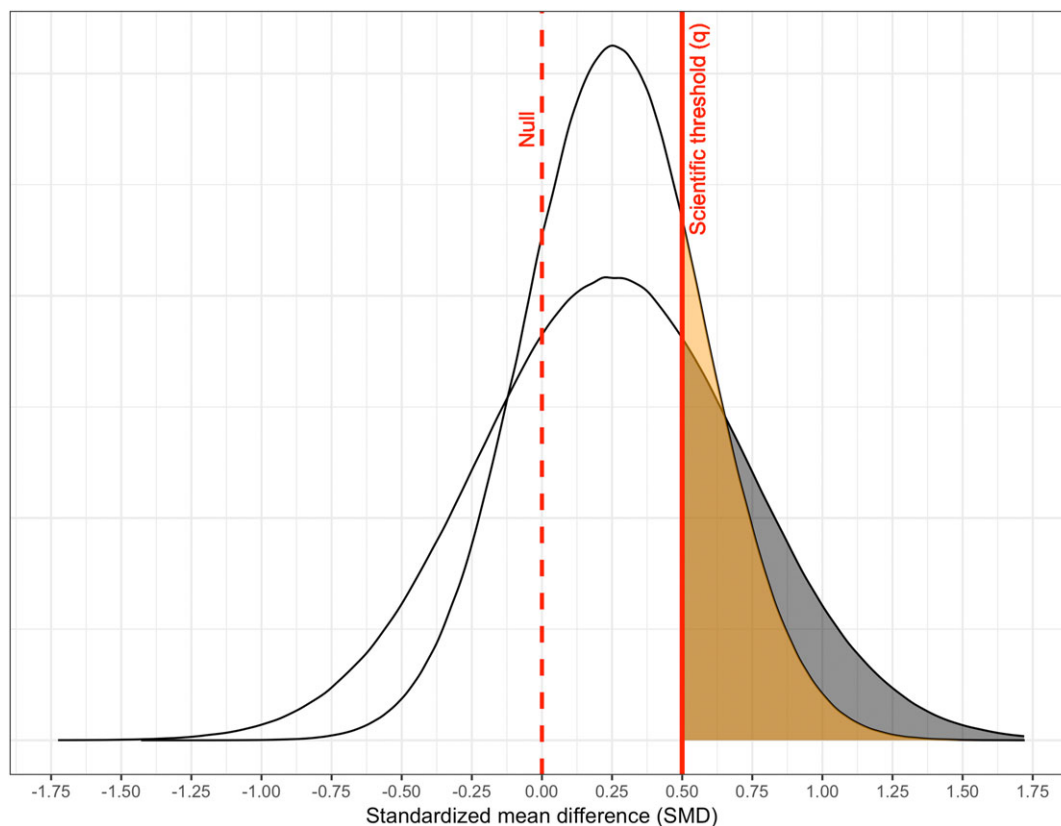


FIGURE 1 Estimated proportion of standardized mean differences (shaded) stronger than the threshold of scientific importance at $SMD = 0.50$ (solid red line) in two meta-analyses with differing “statistical significance.” (Dashed red line) reference null value ($SMD = 0$). [Colour figure can be viewed at wileyonlinelibrary.com]

meta-analyses assess different, but related, outcomes. A common, but unsatisfactory, approach is to compare only estimated means and “statistical significance.” However, around the p value cut-off (eg, 0.05), “statistical significance” can be highly sensitive to the inclusion of even one additional small study. Moreover, these metrics are not robust to the influence of large or outlying studies; they may therefore suggest a spuriously large discrepancy between meta-analyses. Yet even when estimated means differ, the more stable proposed metrics may be similar and have substantially overlapping confidence intervals, suggesting less dramatic discrepancies. Alternatively, despite comparable point estimates, differences in heterogeneity across meta-analyses can lead to different proportions of strong effect sizes.

We illustrate using two meta-analyses investigating the effect of omega-3 fatty acid supplementation on depression in randomized trials; the first meta-analysis¹⁴ suggested a “significant” beneficial effect (standardized mean difference (SMD) = 0.26, 95% CI: 0.12, 0.39), whereas the second¹⁵ did not (SMD = 0.25, 95% CI: -0.07, 0.56). Using $SMD = 0.50$ as a threshold of scientific importance (based, for instance, on a minimum subjectively perceptible difference; see Online Supplement), our first metric estimates that 22% (95% CI: 6%, 38%) of SMD s in the former meta-analysis¹⁴ surpass this threshold, comparable to 30% (95% CI: 4%, 56%) in the latter¹⁵ (Figure 1). Using our second metric, neither meta-analysis suggests a high proportion of strong detrimental effects (1% below $SMD = -0.50$ with 95% CI: 0%, 3% in the former and 6% with 95% CI: 0%, 22% in the latter). Thus, evidence strength in these two meta-analyses is perhaps more concordant than their disagreement on “statistical significance” alone suggests. (A supplementary example regarding comparison between meta-analyses with differing point estimates appears in the Online Supplement.)

5 | EXAMPLE 3: A WIDE PREDICTION INTERVAL DESPITE EVIDENCE FOR MANY STRONG EFFECTS

Others recommend reporting a prediction interval representing a plausible range for a new effect drawn from the distribution underlying the meta-analyzed studies.^{2,6,7} We agree, while also noting that our metrics convey additional information.

For example, a meta-analysis¹⁶ of 19 trials on the effect of intravenous magnesium on mortality following acute myocardial infarction estimated an odds ratio of 0.72 (95% CI: 0.58, 0.90), corresponding to a 95% prediction interval of (0.42, 1.25). This interval is fairly wide and substantially overlaps the null, indicating considerable uncertainty about the size and direction of an effect in a new study. Nevertheless, if we consider inverse associations below $OR = 0.8$ to be scientifically meaningful, our proposed metrics suggest that a high proportion of true effects (66% with 95% CI: 28%, 100%) are more protective than this threshold and that few are comparably strong in the opposite direction (2% above $OR = 1.2$, 95% CI: 0%, 7%). These metrics suggest that there is strong evidence that a substantial proportion of studies have reasonably large protective effects, even though the prediction interval for any single true effect includes the null value of $OR = 1$. We provide this example for illustrative purposes only, noting that others have raised methodological concerns about this literature, including, for example, publication bias and possible data quality issues in some included studies.¹⁷

The findings of the prediction interval and of our metrics are not contradictory, but complementary: The prediction interval infers plausible values for a single effect by considering the middle 95% of the area of the effect distribution, whereas our metrics estimate the area of the lower and upper tails. Intuitively, as the effect distribution becomes more heterogeneous, the middle 95% widens, yielding a wider prediction interval. Simultaneously, the tails thicken, increasing the proportion of strong effects in both the same and the opposite direction as the estimated mean. Thus, with enough heterogeneity, an estimated mean near the null can belie the existence of meaningful effect sizes in some settings. In such cases, the first and second proposed metrics would likely indicate a substantial proportion of strong effect sizes both in the same and in the opposite direction from the estimated mean; this pattern of results would invite exploration of reasons for heterogeneity, for example, through individual patient data meta-analysis with covariates or meta-regression.

6 | TECHNICAL POINTS

The proposed metrics are distinct from antiquated “vote-counting” procedures based on the proportion of studies with significant p values. Such methods fail to differentiate small from large effect sizes (counting only those that are “statistically significant”) and fail to account for differences in precision and sample size across studies.¹ A modified procedure could consider the proportion of observed effect sizes (rather than p values) above a threshold of scientific interest, but this would still limit attention to the observed effect sizes despite that these are measured with sampling error. In contrast, the proposed metrics make appropriately weighted inferences regarding the distribution of true effect sizes in the population (rather than of the observed estimates). In the telomere length example, a basic vote count finds that 14% of studies (3 of 21) had $p < 0.05$. A count of *observed* effect sizes stronger than a threshold, but ignoring sample sizes and sampling variability, finds that 24% (5 of 21) are below $r = -0.20$. In contrast, our approach concerns the true effect sizes and estimates that $< 0.1\%$ of these are stronger than[§] $r = -0.20$.

These metrics are model based; they assume that the true effect sizes are normally distributed. In practice, this implies the sometimes testable assumption that the observed point estimates are normal. Because this distributional assumption may be inexact, our proposed metrics are perhaps most usefully treated as summary metrics of evidence strength for effects of scientifically meaningful size rather than as precise proportions. Additionally, these metrics and inference require the meta-analytic model to be statistically valid, but estimating the heterogeneity without bias, in particular, can be challenging.¹⁰ It is thus important to choose a heterogeneity estimator with good statistical properties for the chosen outcome type¹⁰ to enable good performance of our metrics and of traditional metrics. Lastly, publication bias can compromise our metrics in the same way that it compromises standard analyses. Therefore, when publication bias is suspected, we recommend conducting sensitivity analyses that estimate $\hat{\mu}$ and $\hat{\tau}^2$ in a manner that corrects for publication bias (eg, the work of Vevea and Hedges¹⁸). These bias-corrected estimates can then be used to unbiasedly estimate our proposed metrics.

[§]Due to the theoretical connection between random-effects meta-analysis and mixed models, this discrepancy can also be viewed as shrinkage of the random effect estimates toward $\hat{\mu}$. Indeed, in the telomere length meta-analysis, the best linear unbiased predictions of the study effect sizes were on average 71% as large absolutely as the corresponding point estimates.

7 | CONCLUSION

To better characterize heterogeneous effects in meta-analysis, we recommend supplementing standard reporting with two simple metrics regarding the proportion of effects above a threshold of scientific importance and below a second threshold on the opposite side of the null. These metrics account for effect size, heterogeneity, and statistical error, and they are easy to compute manually or using the R function `prop_stronger`.

ACKNOWLEDGEMENTS

We are grateful to Dr Ian Shrier for the insightful comments that significantly improved the narrative. MM was supported by National Defense Science and Engineering Graduate Fellowship 32 CFR 168a. TVW was supported by NIH grants ES017876 and CA222147. The funders had no role in the design, conduct, or reporting of this research.

REPRODUCIBILITY

All data and code required to reproduce the applied examples are publicly available (<https://osf.io/ksyq5/>), as is the full documentation for the telomere data set (<https://osf.io/6937j/>).

ORCID

Maya B. Mathur  <https://orcid.org/0000-0001-6698-2607>

REFERENCES

1. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons; 2009.
2. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J Roy Stat Soc Ser A (Stat Soc)*. 2009;172(1):137-159.
3. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine*. 2008;5(2):e45.
4. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England J Med*. 2008;358(3):252-260.
5. Turner EH, Rosenthal R. Efficacy of antidepressants. *BMJ*. 2008;336(7643):516.
6. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342:d549.
7. Int'Hout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*. 2016;6(7):e010247.
8. Ades AE, Lu G, Higgins JPT. The interpretation of random-effects meta-analysis in decision models. *Med Decis Mak*. 2005;25(6):646-654.
9. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. New York, NY: John Wiley & Sons Ltd; 2000.
10. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79.
11. Efron B. Better bootstrap confidence intervals. *J Amer Stat Assoc*. 1987;82(397):171-185.
12. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statist Med*. 2000;19(9):1141-1164.
13. Mathur MB, Epel E, Kind S, et al. Perceived stress and telomere length: a systematic review, meta-analysis, and methodologic considerations for advancing the field. *Brain Behav Immun*. 2016;54:158-169.
14. Appleton KM, Rogers PJ, Ness AR. Updated systematic review and meta-analysis of the effects of n-3 long-chain polyunsaturated fatty acids on depressed mood. *Amer J Clin Nutr*. 2010;91(3):757-770. <https://doi.org/10.3945/ajcn.2009.28313>
15. Bloch MH, Hannestad J. Omega-3 fatty acids for the treatment of depression: systematic review and meta-analysis. *Molecular Psychiatry*. 2012;17(12):1272.
16. Shrier I, Boivin J-F, Platt RW, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? *BMC Med Inform Decis Mak*. 2008;8(1):19.

17. Higgins JPT, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol.* 2002;31(1):96-104.
18. Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika.* 1995;60(3):419-435.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Mathur MB, VanderWeele TJ. New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine.* 2019;38:1336–1342. <https://doi.org/10.1002/sim.8057>