# Online Supplement for: "New Metrics for Meta-Analyses of Heterogeneous Effects"

*Maya B. Mathur & Tyler J. VanderWeele*

## Contents

# Derivation of confidence interval

In Lemma 1, we first establish conditions under which $\widehat{\mu}$ and $\widehat{\tau}^2$ are asymptotically independent.

**Lemma 1.** *Let $\widehat{\mu}$ and $\widehat{\tau}^2$ denote maximum likelihood estimates under a normal specification in which, as usual, the within-study variances $\sigma_i^2$ are consideblack fixed and known (e.g., [1]). Suppose there are $k$ studies, with $\widehat{\theta}_i$ denoting the point estimate of the $i^{th}$ study, and assume that $E[\widehat{\theta}_i \mid \sigma_i^2] = E[\widehat{\theta}_i]$. Then $\widehat{\mu}$ and $\widehat{\tau}^2$ are asymptotically independent.*

*Proof.* The joint log-likelihood and partial derivatives are:

$$\log \mathcal{L}\left(\mu, \tau^2\right) = -\frac{1}{2}\sum_{i=1}^{k} \log\left(2\pi\left(\sigma_i^2 + \tau^2\right)\right) - \frac{1}{2}\sum_{i=1}^{k} \frac{(\widehat{\theta}_i - \mu)^2}{\sigma_i^2 + \tau^2}$$

$$\frac{\partial \log \mathcal{L}}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{k} \left(\sigma_i^2 + \tau^2\right)^{-1}\left(-2\widehat{\theta}_i + 2\mu\right)$$

$$\frac{\partial^2 \log \mathcal{L}}{\partial \mu \partial \tau^2} = \frac{1}{2}\sum_{i=1}^{k} \left(\sigma_i^2 + \tau^2\right)^{-2}\left(-2\widehat{\theta}_i + 2\mu\right)$$

$$= -\frac{1}{2}\sum_{i=1}^{k} \frac{2\widehat{\theta}_i - 2\mu}{\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4}$$

The off-diagonal element of the expected Fisher information matrix is therefore:

$$\mathcal{I}_{12} = -E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \mu \partial \tau^2}\right]$$

$$= \frac{1}{2}kE\left[\frac{2\widehat{\theta}_i - 2\mu}{\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4}\right]$$

By a second-order Taylor series expansion, we have, for general random variables $X$ and $Y$:

$$E[X/Y] \approx \frac{E[X]}{E[Y]} - \frac{\text{Cov}(X,Y)}{E[Y]^2} + \frac{\text{Var}(Y)E[X]}{E[Y]^3} \tag{1}$$

We have E[$2\widehat{\theta}_i - 2\mu$] = 0, so applying Equation (1) with the first and third terms equal to $0$ yields:

$$\begin{aligned}
\mathcal{I}_{12} &\approx \frac{1}{2}k \, \frac{E\Big[\Big(2\mu - 2\widehat{\theta}_i\Big)\left(\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4\right)\Big]}{E\Big[\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4\Big]^2} \\
&= \frac{1}{2}k \, \frac{2\mu E\big[\sigma_i^4\big] + 4\mu\tau^2 E\big[\sigma_i^2\big] + 2\mu\tau^4 - 2\tau^4 E\big[\widehat{\theta}_i\big] - 4\tau^2 E\big[\widehat{\theta}_i\sigma_i^2\big] - 2E\big[\widehat{\theta}_i\sigma_i^4\big]}{E\Big[\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4\Big]^2} \\
&= \frac{1}{2}k \, \frac{2\mu E\big[\sigma_i^4\big] + 4\mu\tau^2 E\big[\sigma_i^2\big] + 2\mu\tau^4 - 2\tau^4\mu - 4\tau^2\mu E\big[\sigma_i^2\big] - 2\mu E\big[\sigma_i^4\big]}{E\Big[\sigma_i^4 + 2\sigma_i^2\tau^2 + \tau^4\Big]^2} \\
&= 0
\end{aligned}$$

The penultimate line follows from the assumption that $E[\widehat{\theta}_i \mid \sigma_i^2] = E[\widehat{\theta}_i]$. Since the maximum likelihood estimates are asymptotically bivariate normal, asymptotic independence is established. □

We now derive an asymptotic confidence interval for $\widehat{P}\left(\theta < q^*\right)$, which is identical to that of $\widehat{P}\left(\theta > q\right)$ for a symmetric choice of $q$. We assume use of the standard Dersimonian-Laird estimator, $\widehat{\mu}$, and an arbitrary estimator $\widehat{\tau}^2$ such that, asymptotically:

$$\begin{bmatrix} \widehat{\mu} - \mu \\ \widehat{\tau}^2 - \tau^2 \end{bmatrix} \approx N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} \text{Var}\left(\widehat{\mu}\right) & \text{Cov}\left(\widehat{\mu}, \widehat{\tau}^2\right) \\ \text{Cov}\left(\widehat{\mu}, \widehat{\tau}^2\right) & \text{Var}\left(\widehat{\tau}^2\right) \end{bmatrix}}_{\Sigma/k}\right)$$

(Asymptotic normality is theoretically justified for the maximum likelihood and restricted maximum likelihood estimators $\widehat{\tau}^2$ and, in simulations, also appears to hold under the same conditions for the estimators proposed by references 2, 3, 4, and 5.) Apply the delta method:

$$
\begin{aligned}
h\left(x_1, x_2\right) &= \widehat{P}\left(\theta < q^*\right) \\
&= \Phi\left(\frac{q^* - x_1}{\sqrt{x_2}}\right)
\end{aligned}
$$

$$
\nabla = \begin{bmatrix} \frac{\partial h}{\partial x_1} \\ \frac{\partial h}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -x_2^{-1/2}\phi\left(\frac{q^* - x_1}{\sqrt{x_2}}\right) \\ \frac{x_1 - q^*}{2} x_2^{-3/2}\phi\left(\frac{q^* - x_1}{\sqrt{x_2}}\right) \end{bmatrix}
$$

$$
\sqrt{k}\left[h\left(\widehat{\mu}, \widehat{\tau}^2\right) - h\left(\mu, \tau^2\right)\right] \rightarrow N\left(0, \nabla'\Sigma\nabla|_{\mu,\tau^2}\right)
$$

$$
\begin{aligned}
\nabla'\Sigma\nabla &= \nabla_1\left(\nabla_1\Sigma_{11} + \nabla_2\Sigma_{21}\right) + \nabla_2\left(\nabla_1\Sigma_{12} + \nabla_2\Sigma_{22}\right) \\
&= \nabla_1^2\Sigma_{11} + \nabla_2^2\Sigma_{22}
\end{aligned}
$$

for choices of estimators $\widehat{\tau}^2$ that are asymptotically independent of $\widehat{\mu}$. Thus:

$$
\widehat{\text{Var}}\left(\widehat{P}\left(\theta < q^*\right)\right) \approx \left(\phi\left(\frac{q^* - \widehat{\mu}}{\sqrt{\widehat{\tau}^2}}\right)\right)^2 \cdot \left(\frac{\widehat{\text{Var}}\left(\widehat{\mu}\right)}{\widehat{\tau}^2} + \frac{\widehat{\text{Var}}\left(\widehat{\tau}^2\right)\left(\widehat{\mu} - q^*\right)^2}{4\left(\widehat{\tau}^2\right)^3}\right)
$$

Thus, approximate 95% confidence limits for $\widehat{P}\left(\theta < q^*\right)$ are:

$$
\widehat{P}\left(\theta < q^*\right) \pm \Phi^{-1}\left(0.975\right) \cdot \phi\left(\frac{q^* - \widehat{\mu}}{\sqrt{\widehat{\tau}^2}}\right) \cdot \sqrt{\frac{\widehat{\text{Var}}\left(\widehat{\mu}\right)}{\widehat{\tau}^2} + \frac{\widehat{\text{Var}}\left(\widehat{\tau}^2\right)\left(\widehat{\mu} - q^*\right)^2}{4\left(\widehat{\tau}^2\right)^3}}
$$

An analogous derivation or argument from symmetry yields the same standard error for $\widehat{P}\left(\theta > q\right)$.

## Methods for choosing an effect size threshold

Much existing work, spanning a variety of disciplinary perspectives, has discussed how to choose thresholds for scientifically meaningful effect sizes. Reference 6 provides an excellent review and examples of numerous methods in the context of health outcomes. In particular, they discuss a variety of "anchoring-based" methods in which an effect size threshold is chosen by relating the outcome measure to external criteria bearing immediate scientific or clinical relevance. For example, a minimum effect size threshold for a composite scale outcome could be defined in relation to naturally occuring discrepancies in the outcome between patient groups with and without a diagnosis[7] or to effect sizes produced by existing interventions similar to the intervention under investigation[8]. In the telomere length example discussed in the main text, we

might define a threshold of scientific importance through comparison to the correlation strength of age with telomere length (approximately $r = -0.34$ to $-0.38$)[9] since age is perhaps the best-established determinant of telomere degradation. We might expect a psychological state, such as stress, to have a somewhat weaker effect size than the biological process of aging. Thus, to select an effect size threshold for psychological stress, we might blackuce the magnitude of the age-telomere length correlation to, for example, $r = -0.10$ or $r = -0.20$. Numerous other types of external "anchoring" criteria have also been used in the medical literature[6].

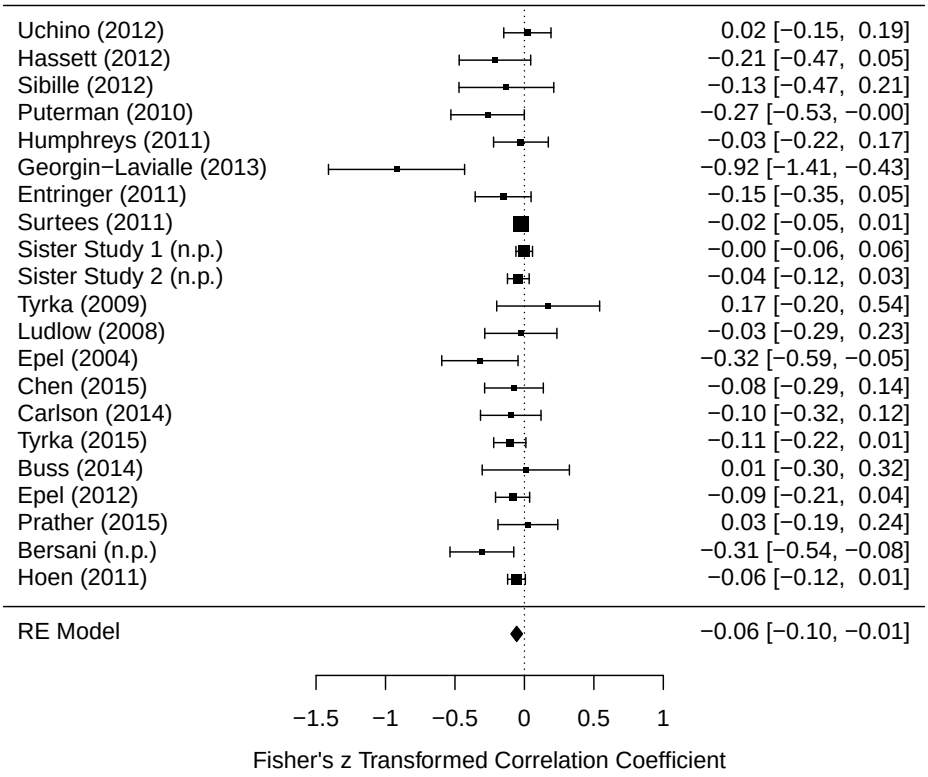
When the population public health impact of a health condition is the primary concern, investigators could draw upon the extensive literature on cost-effectiveness decision rules in selecting an effect size threshold. For example, much existing work has discussed or empirically quantified the cost threshold at which societies (or individuals) are willing to pay for a specific improvement in health, such as an addition of one quality-adjusted life-year (e.g.,[10,11]). Such findings could be used to "convert" hypothetical statistical effect sizes for a given health outcome to a concrete financial scale, such as dollars. A minimum effect size threshold could then be defined in relation to the utility, expressed in dollars, of the intervention or exposure of interest.

In contrast, in other scientific contexts, individuals' subjective experience of pain, distress, or disability may be the primary concern rather than (or in addition to) aggregate public health impact. In these cases, it may be useful to set the threshold as the minimum effect size that is subjectively perceptible[???,12,13]. A systematic review consideblack 62 studies that attempted to estimate such thresholds for a wide variety of health outcomes, for example by relating patients' subjective self-assessments to objective measurements of health condition severity[13]. This review found that $SMD = 0.50$ was a surprisingly consistent minimally detectable effect size for health outcomes, perhaps reflecting fundamental mechanisms of human sensory discrimination or constraints on categorical discrimination due to working memory capacity. For ease of comparison to other statistical measures of effect size, the threshold $SMD = 0.50$ is approximately equivalent (under some distributional assumptions)[14,15] to an odds ratio of 2.5 or to a risk ratio of 1.6. However, it is important to note that an intervention that has only small effects on the individual level, even ones that are not subjectively perceptible, may still have very substantial impacts on a population public health level; thus, as described above, much lower thresholds might often be consideblack.

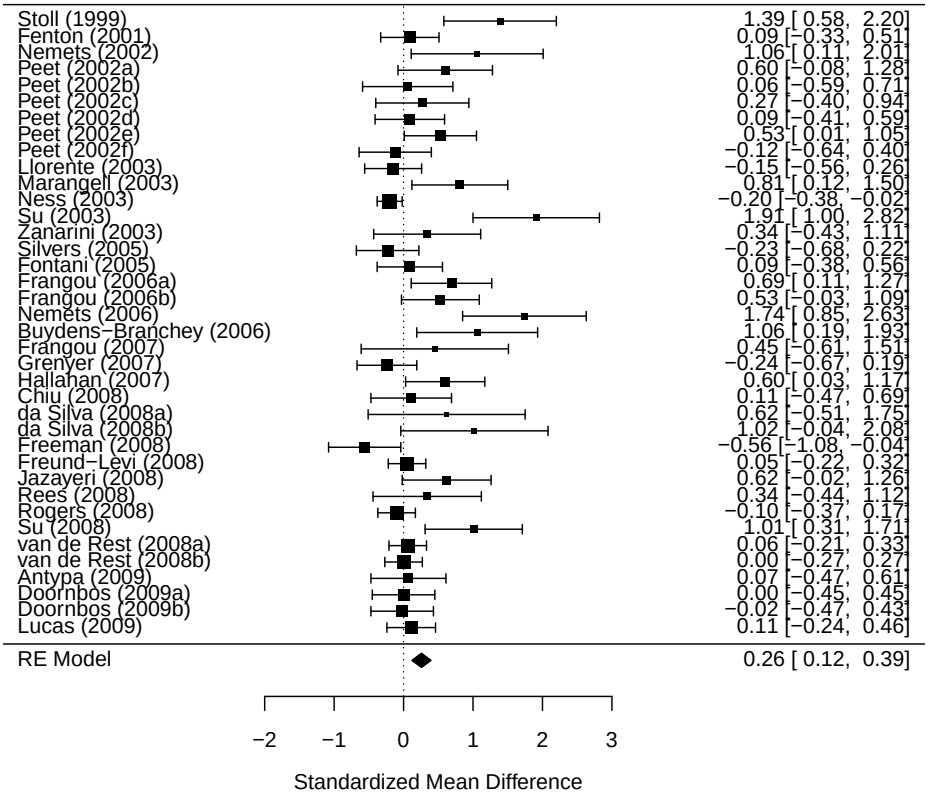
## Forest plots for applied examples

Below are forest plots corresponding to each applied example presented in the main text and Supplement.

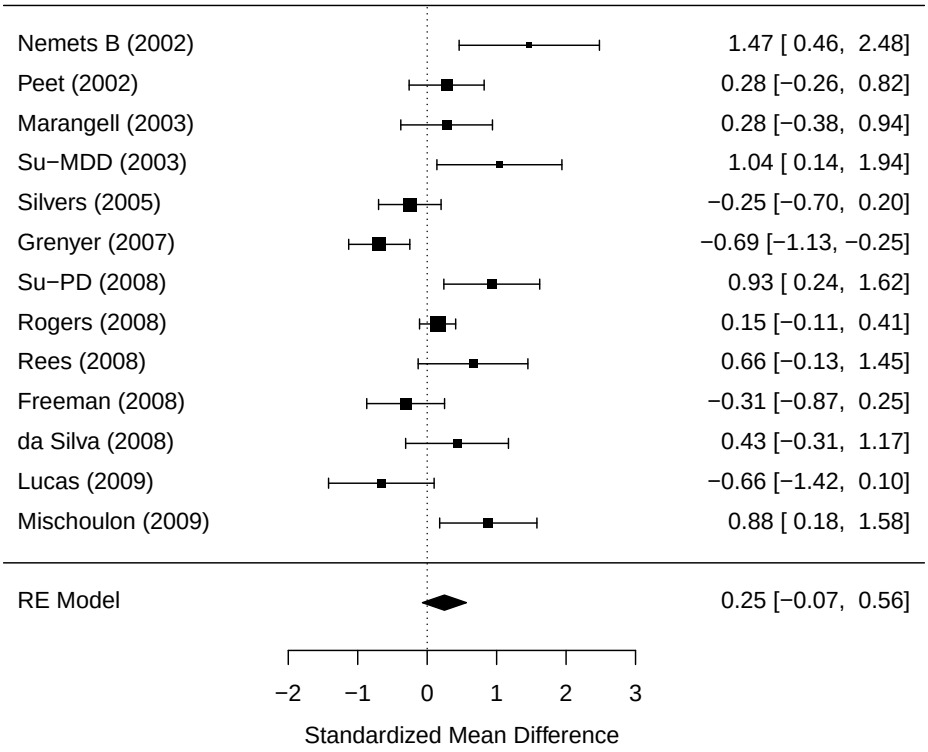
Supplementary Figure 1: Perceived stress and telomere length (Example 1)



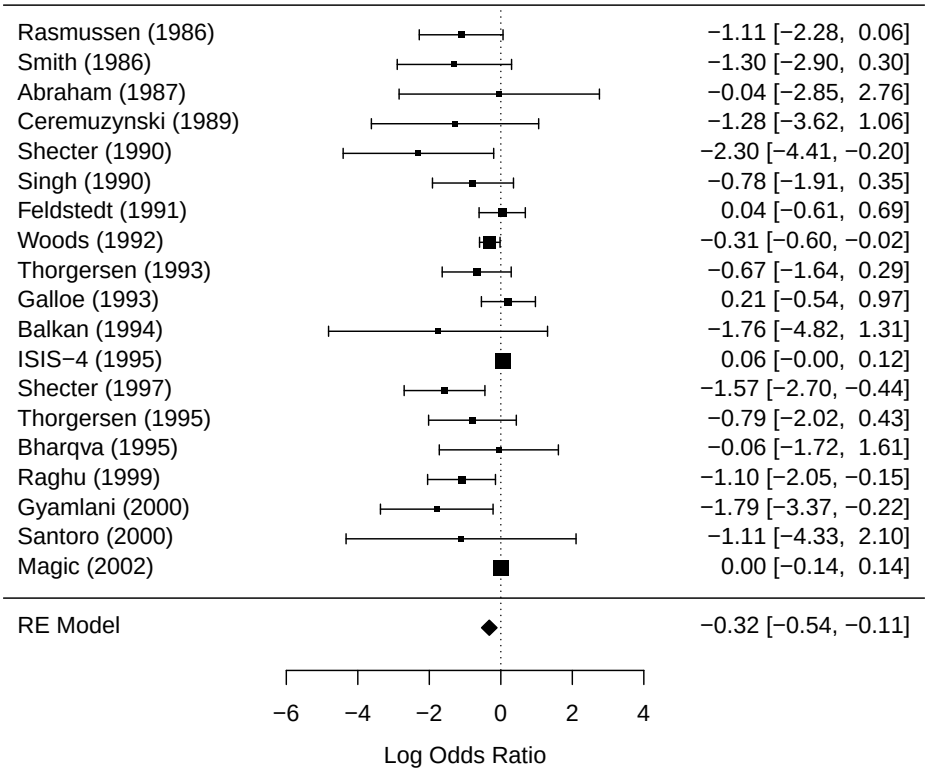
Supplementary Figure 2: Omega-3 supplementation and depression (Example 2, first meta-analysis)



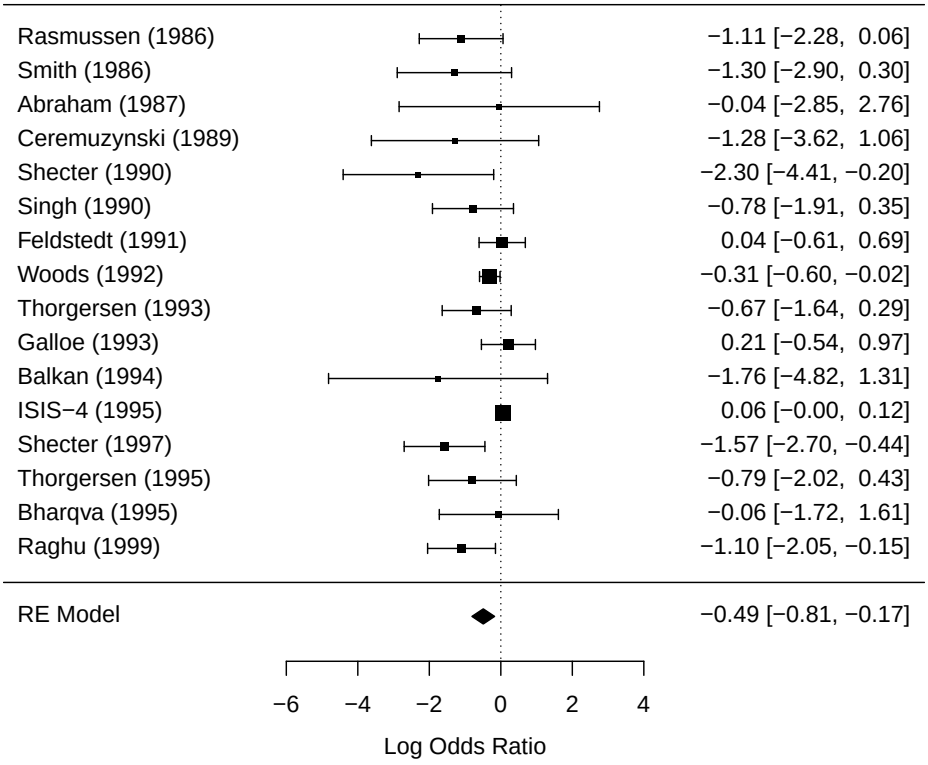
Supplementary Figure 3: Omega-3 supplementation and depression (Example 2, second meta-analysis)



Supplementary Figure 4: Magnesium and myocardial infarction (Example 3, 19-study meta-analysis)



Supplementary Figure 5: Magnesium and myocardial infarction (Supplementary Example, 16-study meta-analysis)

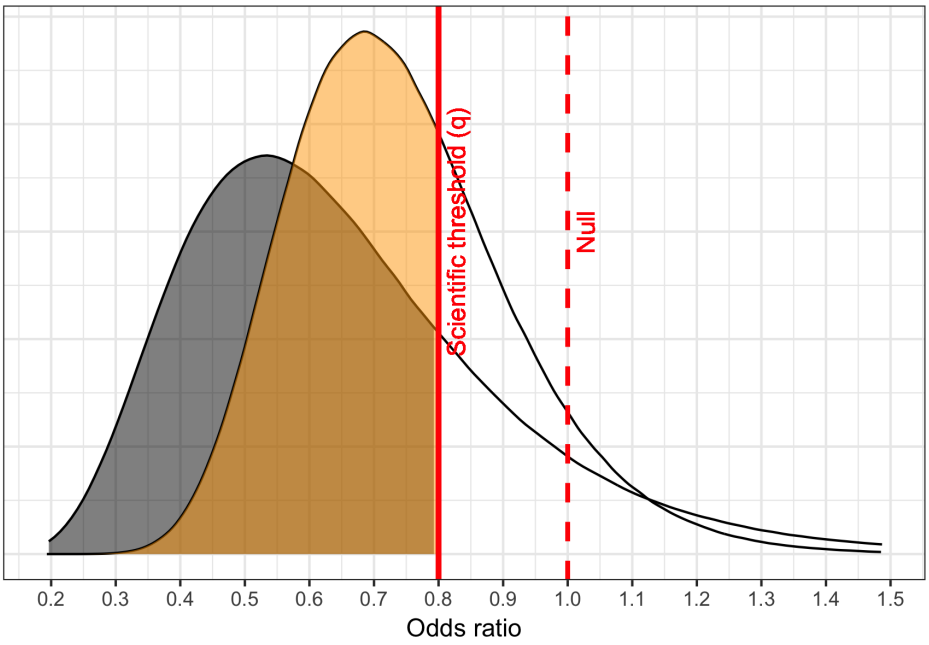



## Supplementary example: Meta-analyses with different estimated means, but similar evidence strength

Here, we extend Example 3 to illustrate comparison between meta-analyses in which estimated means differ somewhat, yet due to differences in heterogeneity, the proposed metrics suggest comparable proportions of scientifically meaningful effect sizes. Others have conducted a series of cumulative meta-analyses of the literature on intravenous magnesium on mortality following acute myocardial infarction[16]. In a meta-analysis of the earliest 16 trials versus that of all 19 trials published at the time of reference 16's analysis, the estimated mean shifted from an odds ratio ($OR$) of 0.61 (95% CI: 0.45, 0.84) to 0.72 (95% CI: 0.58, 0.9). Considering inverse associations below $OR = 0.8$ to be scientifically important, we reanalyzed data[16] to estimate that, in the population of effects represented by the first 16 studies, approximately $76\%$ are below $OR = 0.8$ (95% CI: 41%, 100%) versus approximately $66\%$ (95% CI: 28%, 100%) in the effects represented by all 19 studies (Supplementary Figure 1). Considering effects in the opposite direction, the two meta-analyses

would estimate, respectively, that approximately $4\%$ (95% CI: 0%, 6%) and $2\%$ (95% CI: 0%, 7%) are above an odds ratio of $1.2$.

Supplementary Figure 6: Estimated proportion of odds ratios (shaded) more protective than threshold of scientific importance at $OR = 0.8$ (solid black line) in cumulative meta-analyses of 16 (gray) or 19 (orange) studies. Dashed black line: reference null value ($OR = 1.0$). (As usual, normality was assumed on the log-$OR$ scale; the plotted distributions are skewed due to exponentiation.)




Under these thresholds of scientific importance, we might therefore consider both meta-analyses to have provided fairly strong evidence for scientifically important effects of magnesium again subject to methodological caveats mentioned in the main text[17]. This suggests a more stable view of evidence strength in an evolving literature than the difference in pooled point estimates alone might suggest.

## Simulation study

We performed a simulation study assessing the relative coverage of the proposed asymptotic confidence interval (CI) versus a bootstrapped confidence interval, including in meta-analyses of few studies or with relatively low-poweblack studies. We fixed the mean of the true effects to $\mu = 0.50$ on the mean difference scale while varying the number of studies ($k$) between 10 and 50, the heterogeneity ($\tau^2 \in \{0.01, 0.04, 0.25\}$), and the mean sample size in each study ($E[N] \in \{150, 850\}$). We used a bias-corrected and accelerated bootstrap (BCa) with 10,000 iterates to estimate the bootstrap confidence intervals[18]. (Additional simulations, not shown, suggested that basic or percentile bootstrap methods yielded substantially worse performance than BCa.) On some simulation repetitions, the BCa method failed to converge with fewer iterates; we selected $10,000$ iterates in order to ensure that every scenario had $< 10\%$ missing data due to convergence failures. Additional simulation results suggested that using fewer iterates (e.g., $5,000$) yielded nearly identical confidence intervals, albeit with more frequent convergence failures in some scenarios. Therefore, in practice, we believe that $5,000$ iterates is a reasonable choice unless the procedure fails to converge.

Supplementary Figures 7 and 8, respectively, show coverage and width of theoretical vs. bootstrapped confidence intervals. With a true proportion $P \geq 0.15$, the theoretical confidence interval had approximately nominal coverage when $k = 50$, regardless of the sample size of the meta-analyzed studies, and it usually had coverage $\geq 90\%$ when $k \geq 10$. When $P \geq 0.15$, its overall mean coverage was 91%, and its minimum coverage was 84%. For a smaller true proportion $P = 0.10$, the theoretical confidence interval sometimes showed fairly poor coverage for small $k$ (Supplementary Figure 7, Panel A). Across all scenarios, including $P = 0.10$, the mean coverage of the theoretical interval remained 91%, but its minimum coverage dropped slightly to 81%. In contrast, across all scenarios, the BCa interval had nominal mean coverage 95% and minimum coverage 86%. Accordingly, the bootstrap confidence interval were wider than the theoretical confidence interval, sometimes considerably so for small $k$ (Supplementary Figure 8). The bootstrap interval was sometimes overly conservative when the meta-analyzed studies were small (Supplementary Figure 8, Panel A, row 1).

Supplementary Figure 7: Coverage of theoretical vs. bootstrapped 95% confidence intervals. Dashed line indicates nominal coverage. True P = theoretical proportion of effects stronger than $q$ based on data generation parameters.

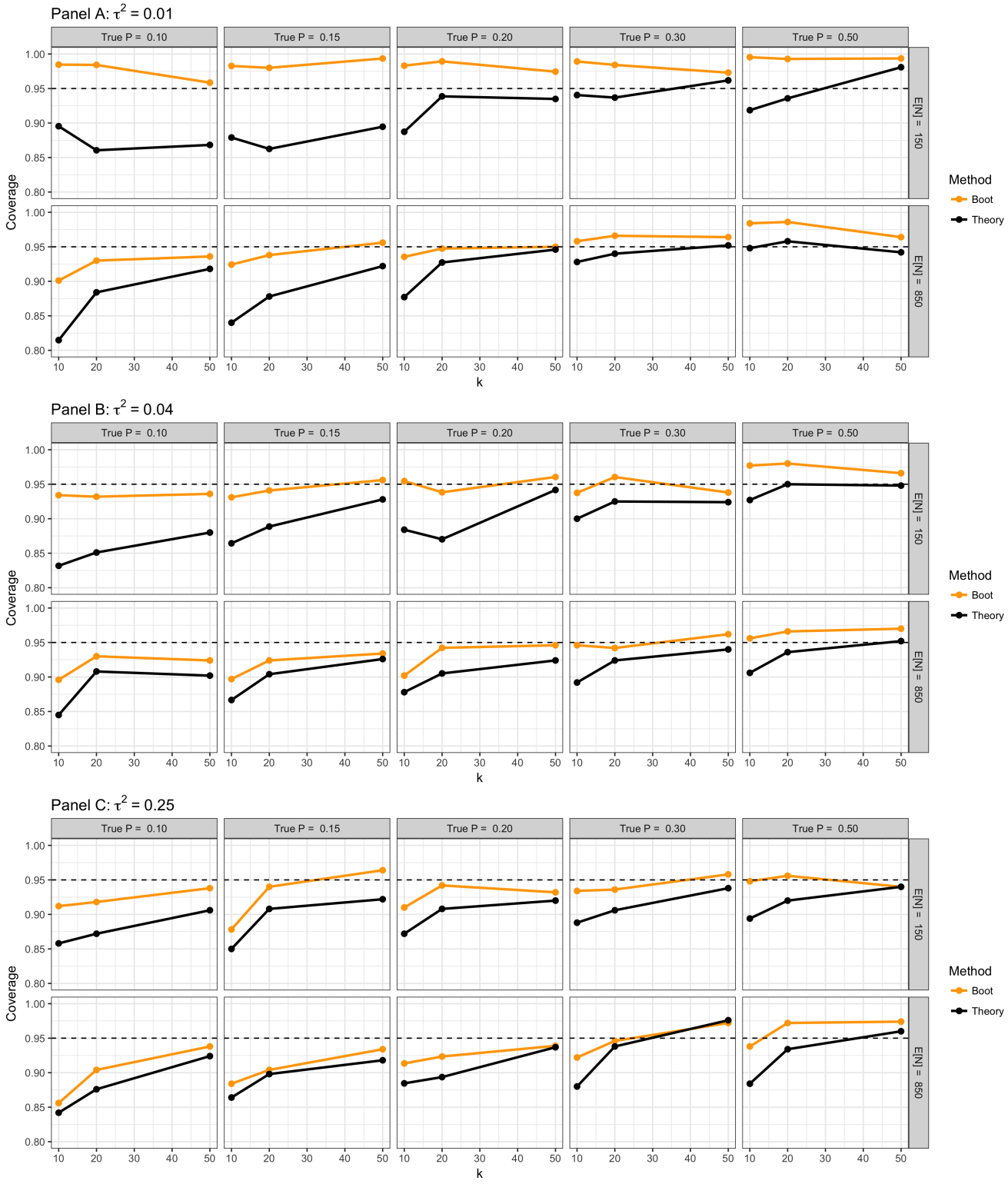
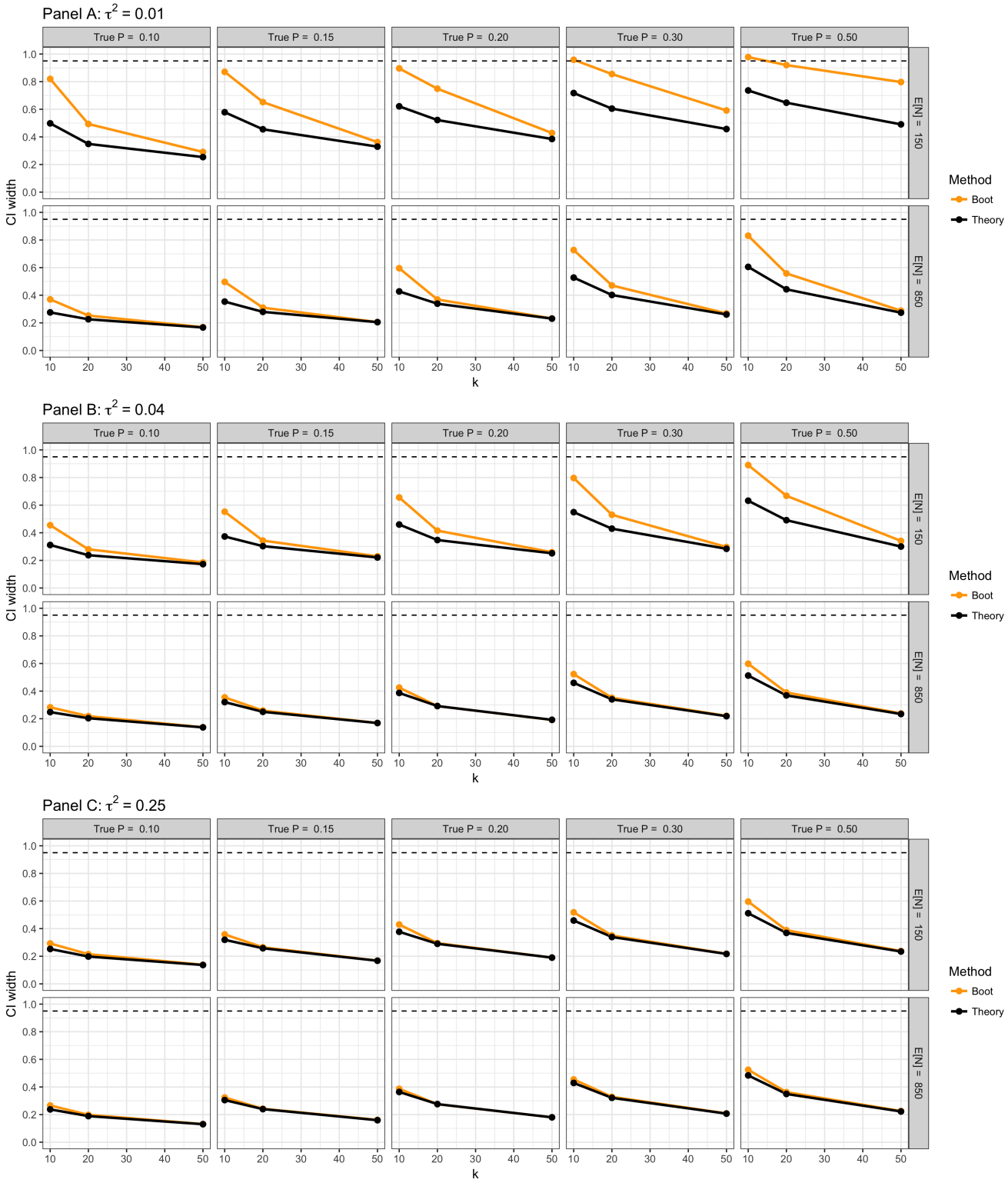


Figure 8: Width of theoretical vs. bootstrapped 95% confidence intervals. True P = theoretical proportion of effects stronger than $q$ based on data generation parameters.



# References

1. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*. 2001;20(6):825-840.

2. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986;7(3):177-188.

3. Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*. 1982;87(5):377-385.

4. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2005;54(2):367-384.

5. Hedges L, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press; 1985.

6. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology*. 2003;56(5):395-407.

7. Johnson PA, Goldman L, Orav EJ, Garcia T, Pearson SD, Lee TH. Comparison of the medical outcomes study short-form 36-item health survey in black patients and white patients with acute chest pain. *Medical Care*. 1995:145-160.

8. Hill CJ, Bloom HS, Black AR, Lipsey MW. Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*. 2008;2(3):172-177.

9. Müezzinler A, Zaineddin AK, Brenner H. A systematic review of leukocyte telomere length and age in adults. *Ageing Research Reviews*. 2013;12(2):509-519.

10. Braithwaite RS, Meltzer DO, King Jr JT, Leslie D, Roberts MS. What does the value of modern medicine say about the $50,000 per quality-adjusted life-year decision rule? *Medical Care*. 2008;46(4):349-356.

11. Eichler H-G, Kong SX, Gerth WC, Mavros P, Jönsson B. Use of cost-effectiveness analysis in health-care resource allocation decision-making: How are cost-effectiveness thresholds expected to emerge? *Value in Health*. 2004;7(5):518-528.

12. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*. 1989;10(4):407-415.

13. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*. 2003;41(5):582-592.

14. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*. 2000;19(22):3127-3131.

15. VanderWeele TJ. On a square-root transformation of the odds ratio for a common outcome. *Epidemiology*. 2017;28(6):e58-e60.

16. Shrier I, Boivin J, Platt R, et al. The interpretation of systematic reviews with meta-analyses: An objective or subjective process? *BMC Medical Informatics and Decision-Making*. 2008;8(1):19.

17. Higgins JP, Spiegelhalter DJ. Being sceptical about meta-analyses: A Bayesian perspective on magnesium

trials in myocardial infarction. *International Journal of Epidemiology*. 2002;31(1):96-104.

18. Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*. 1987;82(397):171-185.