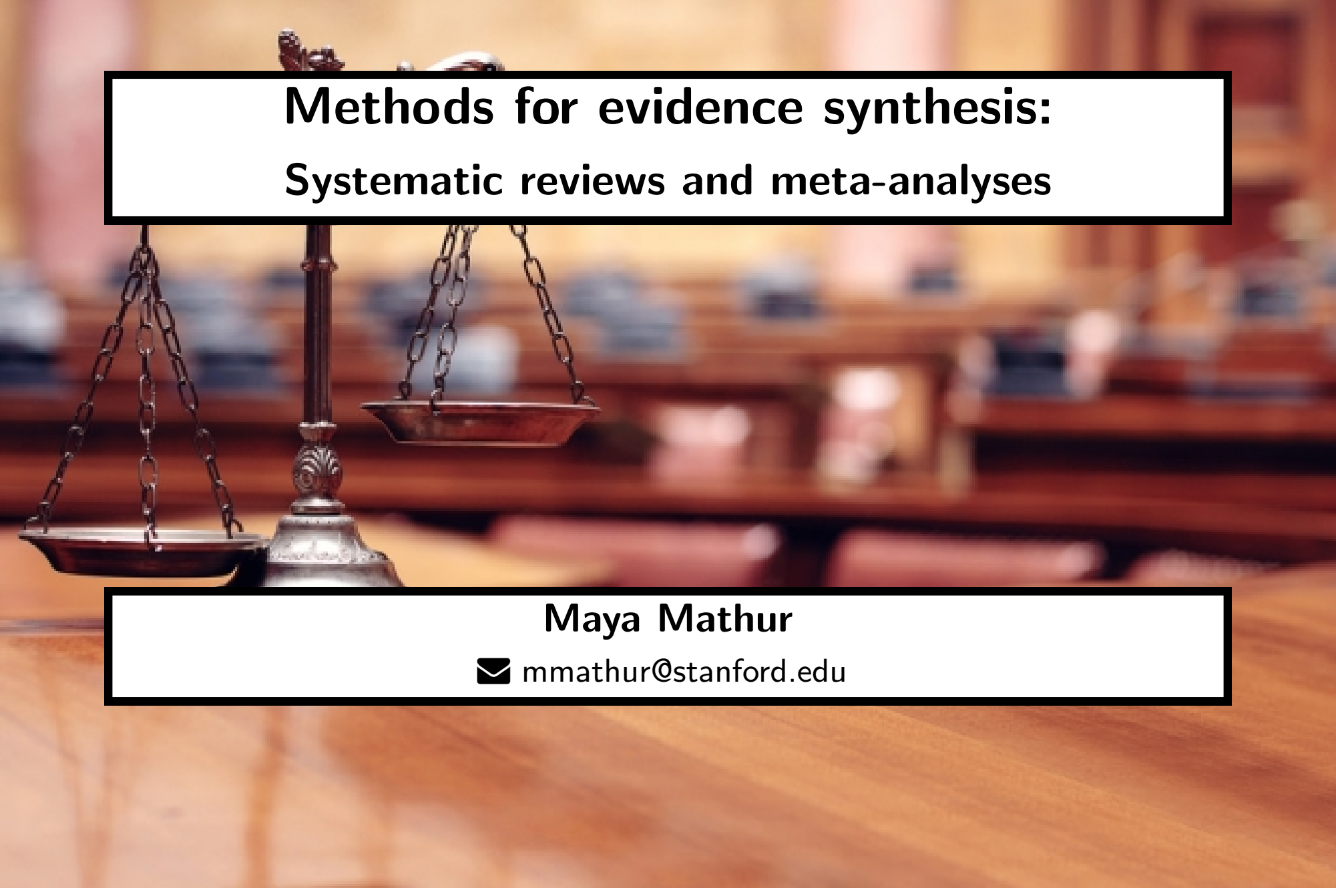


Methods for evidence synthesis:

Systematic reviews and meta-analyses

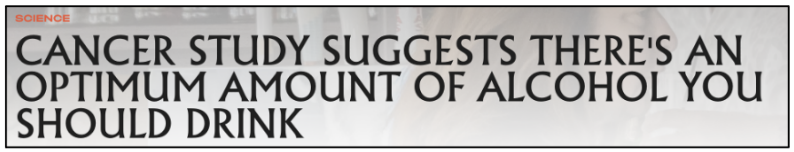
Maya Mathur

mmathur@stanford.edu

Evidence synthesis: The practice of qualitatively or quantitatively summarizing information across multiple studies.

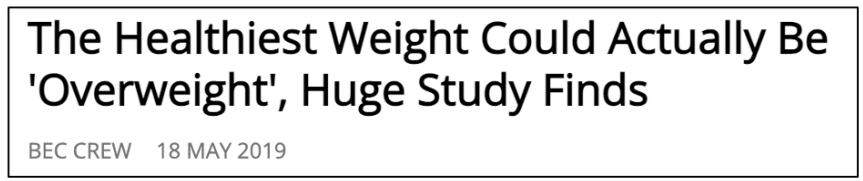
Why synthesize evidence?

The science-journalism rollercoaster

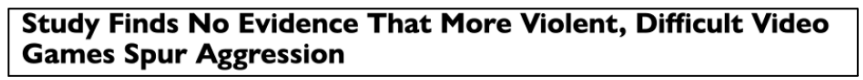












A few studies accrue most of the academic attention

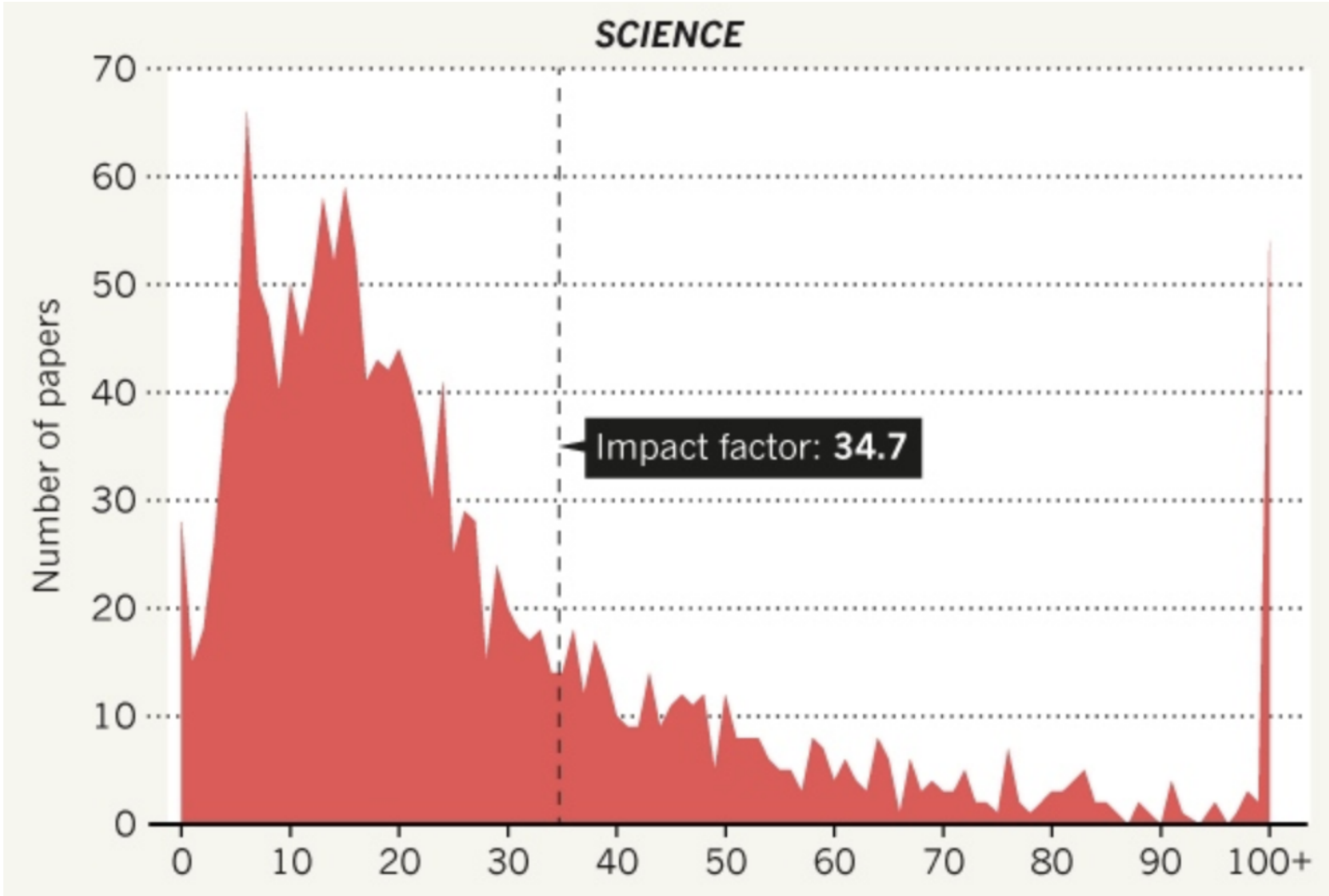


Figure: Citation distribution in 2016 of *Science* papers published in 2013-2014 (as reported by *Nature News*)

Is scientific knowledge really this fickle?

In the lecture on replication, you heard about many problems that contribute to instability (both real and illusory) in scientific results.

One of those problems: Our tendency to fetishize single studies and their p-values.

Instead: We can rigorously **synthesize evidence across studies**.

Goals of evidence synthesis

A rigorous evidence synthesis strives to:

- Capture all relevant literature on a topic, not just the most highly-cited studies
- Help qualitatively and quantitatively summarize all that is known on a topic and characterize remaining uncertainties
- Identify methodological strengths and weaknesses of the evidence base
- Provide informed guidance for future research, both substantively and methodologically
- Ultimately help shape recommendations to the lay public and policymakers

Steps in conducting evidence syntheses

1. Develop inclusion criteria.
2. Systematically search for studies.
3. Extract qualitative and/or quantitative information from studies.
4. (*Optional:* Quantitatively pool the effect sizes.)
5. Assess risks of bias in the literature.
6. Holistically assess the state of the evidence.

Steps in conducting evidence syntheses

1. Develop inclusion criteria.
2. Systematically search for studies.
3. Extract qualitative and/or quantitative information from studies.
4. (*Optional:* Quantitatively pool the effect sizes.)
5. Assess risks of bias in the literature.
6. Holistically assess the state of the evidence.

Systematic review:[1] A synthesis that performs all the qualitative steps, but potentially omits quantitative synthesis.

Meta-analysis: A type of systematic review that also conducts quantitative synthesis (i.e., all steps above).

[1] Some people use this term differently.

Systematic review methods

Running example

Updated systematic review and meta-analysis of the effects of n−3 long-chain polyunsaturated fatty acids on depressed mood[1–3]

Katherine M Appleton, Peter J Rogers, and Andrew R Ness

ABSTRACT

Background: The debate over a role for n−3 long-chain polyunsaturated fatty acids (n−3 PUFAs) in depressed mood continues.

Objective: The objective was to update a previous systematic review and meta-analysis of published randomized controlled trials investigating the effects of n−3 PUFAs on depressed mood and to explore potential sources of heterogeneity.

depressive symptomatology (6, 7), and randomized controlled trials (RCTs) are available suggesting beneficial effects of n−3 PUFAs compared with placebo (8–10).

Ecologic and observational studies, however, are also available that suggest no association between n−3 PUFAs and depressive illness (11, 12) or no association once confounders have been accounted for (13), cross-sectional evidence suggesting limited

Appleton et al. (2010) meta-analyzed 38 randomized trials on the effect of omega-3 fatty acid supplementation vs. placebo on depression symptoms.

In later slides, we'll re-analyze this meta-analysis. All code required to do so is at: `https://osf.io/ce36g/`.

Developing inclusion criteria

"PICO" criteria

- **Population:** What kinds of participants must the studies include? (Running example: All human participants)
- **Intervention:** What kinds of intervention (or exposure) must the studies assess? (Any dose and duration of an omega-3 fatty acid supplement or fish)
- **Comparison:** What kinds of comparator or placebo must the studies use? (Any placebo supplement, e.g., mineral oil)
- **Outcome:** How must studies define and measure the outcome? (Any measure of depressed mood)

Should choose PICO criteria such that synthesis makes sense substantively.

Higgins et al. (2019), Chapter 3

Developing inclusion criteria

Studies' methodological rigor

What study designs, wrt methodological rigor, will be included? This inclusion criterion is **critical to mitigate bias** in the review.

For example, in reviews assessing causal questions and in which confounding is a key concern:

- Cross-sectional studies should generally be excluded unless temporal ordering of exposure and outcome is clear.
- Default starting point: Include only randomized studies plus longitudinal nonrandomized studies that (i) measure exposure before outcome and (ii) control for baseline confounders and baseline outcome.
- Can also stipulate that studies with weaker designs will be included only in secondary analyses.

Searching for studies

Must decide which databases (plural!) to search. Important to also try to search the **gray literature** (e.g., conference proceedings, preprints, unpublished theses).

Also need to craft specific search terms, which should be highly sensitive (capture all relevant literature) but also reasonably specific (limit the number of irrelevant articles).

Developing sensitive search terms requires **a lot** of expertise. Every systematic review should have an expert on database searching (e.g., academic reference librarian) heavily involved.[2]

(On the horizon: AI saves the day?)

[2] I suggest offering co-authorship to this person: that is how much expertise and involvement they should have.

Screening studies against inclusion criteria

At least two reviewers should independently screen studies against the inclusion criteria.

Often makes sense to review in two stages:

1. Screen titles and abstracts to remove countless articles that are obviously irrelevant.

2. Skim full texts of the remaining articles.

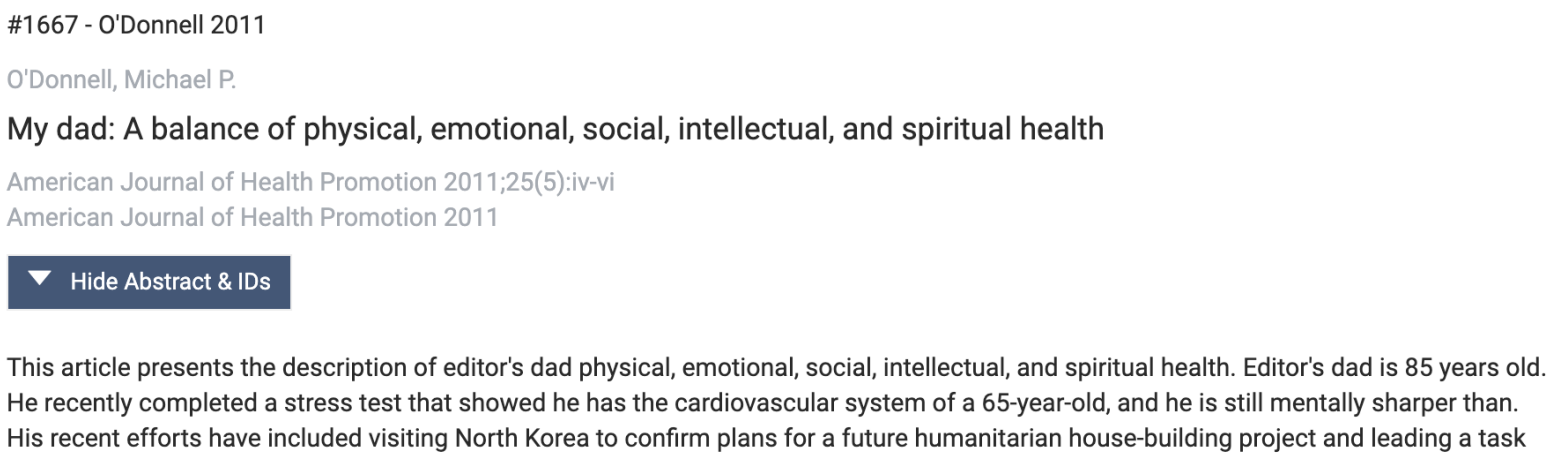


Figure: This one did not need full-text screening.

Extracting effect sizes

Effect-size measures

For a meta-analysis, need to extract quantitative results (i.e., effect sizes) that are reasonably comparable across studies.

Examples of effect-size measures:

- Risk ratios
- Odds ratios
- Pearson's correlations
- Standardized mean differences (SMD):

$$\text{Cohen's } d = \frac{\widehat{E}[Y \mid X = 1] - \widehat{E}[Y \mid X = 0]}{\widehat{SD}\,(Y \mid X)}$$

i.e., the estimated treatment effect in terms of standard deviations of the outcome.

Higgins et al. (2019), Chapter 6

Effect-size measures

As we'll see, standard meta-analysis methods assume the population effects are normal and independent of their standard errors.

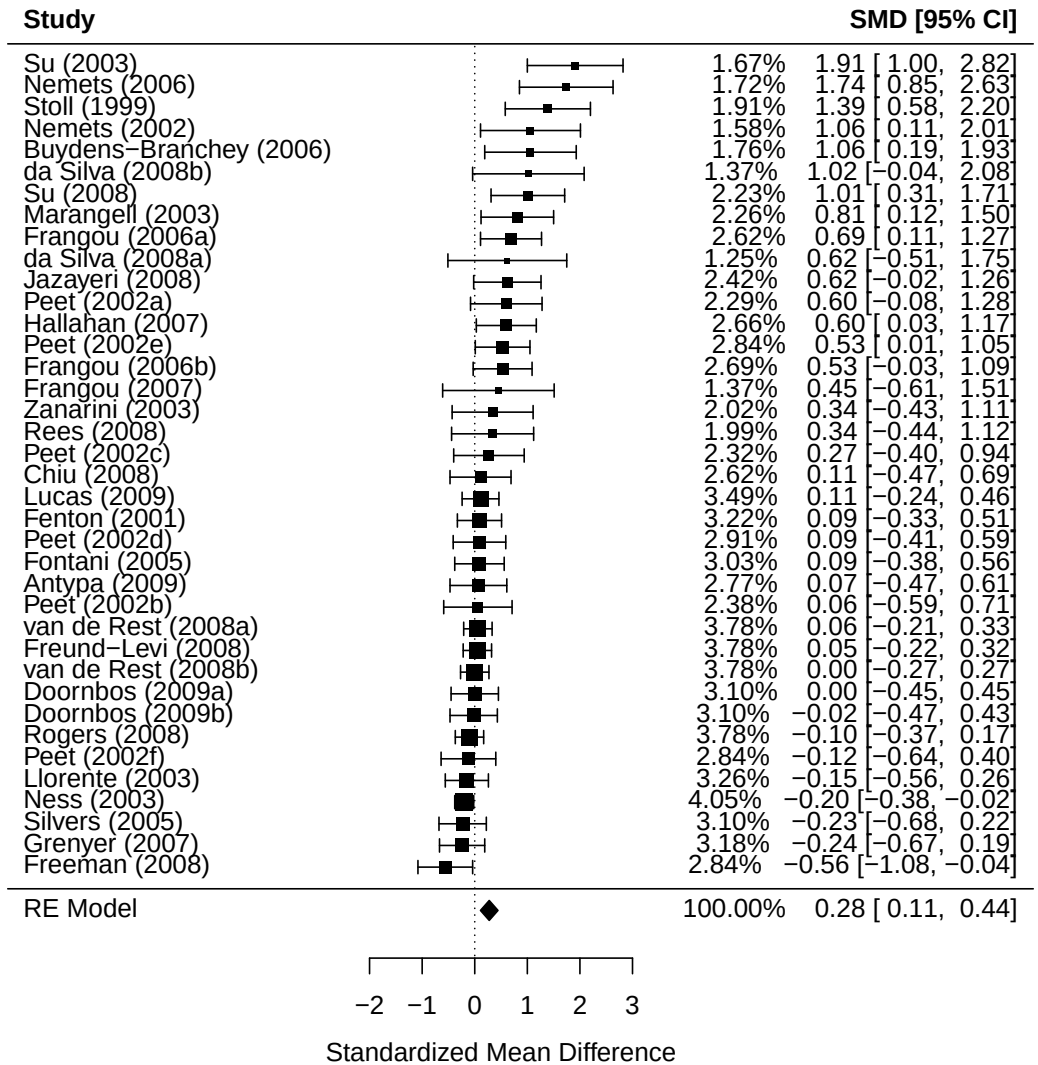
Some effect-size measures should be transformed before analysis to make these assumptions more reasonable.

For example, ratio measures should be logged, and correlations should be transformed via Fisher's z (which basically makes the correlation unbounded and less dependent on its variance).

Analyzing a meta-analysis

Visualizing evidence across studies

Forest plot: Figure depicting each study's estimate and confidence interval.



Quantitatively pooling the studies

What not to do

What not to do, #1: Count how many studies are significant.

Running example: 10 of 38 (26%) studies are significant and positive.[3] 2 studies (5%) are significant and negative.

Why shouldn't we do this?

[3] Technically, there were 35 studies (articles) that contributed 38 total point estimates.

Quantitatively pooling the studies

What not to do

What not to do, #2: Just average the studies' estimates.

Running example: Mean $SMD = 0.38$

Why shouldn't we do this?

Quantitatively pooling the studies

Notation

Instead of simply averaging studies' estimates, it would be better to **have larger studies count more** in the meta-analysis.

To make this idea precise, first...you guessed it! We need some notation.

k: Number of studies in meta-analysis

$\widehat{\theta}_i$, $\widehat{\sigma}_i$: Point estimate and estimated standard error of i^{th} study

μ: The true effect size in all the studies

$\widehat{\mu}$: Meta-analytic estimate (a.k.a "pooled estimate") of the effect size

Quantitatively pooling the studies

Fixed-effects meta-analysis

To appropriately upweight the more precise studies, weight each study's contribution to the meta-analytic average by its inverse-variance, $w_i = 1/\widehat{\sigma}_i^2$.

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} w_i \widehat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

This is called **fixed-effects meta-analysis**.

Quantitatively pooling the studies

Fixed-effects meta-analysis

Running example:

study	yi	vi	pval
Stoll (1999)	1.39	0.17079449	0.0007699012
Fenton (2001)	0.09	0.04592005	0.6744910916
Nemets (2002)	1.06	0.23493679	0.0287493005
Peet (2002a)	0.60	0.12037094	0.0837411095
Peet (2002b)	0.06	0.10998426	0.8564305645
Peet (2002c)	0.27	0.11685665	0.4296234159

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} w_i \widehat{\theta}_i}{\sum_{i=1}^{k} w_i} = \frac{(1.39/0.17) + (0.09/0.05) + \cdots}{1/0.17 + 1/0.05 + \cdots} = \mathbf{0.10}$$

Omega-3 supplementation is associated with decreased depression symptoms by $SMD = 0.10$ (i.e., 10% of a standard deviation).

Quantitatively pooling the studies

Fixed-effects meta-analysis

Fixed-effects meta-analysis does have an important limitation.

Imagine we actually had the full dataset of individual observations from each study.

Fixed-effects meta-analysis is like gluing together all the studies' datasets and then calculating the effect size, ignoring that observations came from different studies.

But what if the **studies differ from one another?** Does this presence of clustering remind you of anything from longitudinal data analysis?

Quantitatively pooling the studies

Heterogeneity

Heterogeneity: When the population effects differ across studies. (This is not about the point estimates!)

Effects will often be heterogeneous if studies differ with respect to populations, interventions, outcomes, biases, etc.

In the running example, studies differed on:

- Whether subjects needed to be depressed at baseline
- Supplement dosage, formulation, and duration
- How depression was measured

Quantitatively pooling the studies

Random-effects meta-analysis

Think back to the idea of having full datasets for every study. How could you account in analysis for clustered (i.e., heterogeneous) observations?

(Hint: See slide subtitle.)

Quantitatively pooling the studies

Random-effects meta-analysis

More notation:

τ: The standard deviation of the population effects (not estimates) across studies. This describes the amount of heterogeneity.

μ: The **mean** effect **across studies**

Quantitatively pooling the studies

Random-effects meta-analysis

We can estimate μ and τ in practice by maximum likelihood[4] under these **assumptions**:

1. $\widehat{\theta}_i = \mu + \gamma_i + \epsilon_i$
   where $\mu + \gamma_i$ is study i's **own** population effect.

2. $\gamma_i \sim N(0, \tau^2)$ (study random effects)

3. $\epsilon_i \sim N(0, \widehat{\sigma}_i^2)$ (statistical error in estimates)

4. $\gamma_i \perp\!\!\!\perp \epsilon_i$ (e.g., $\text{corr}\left(\gamma_i, \widehat{\sigma}_i^2\right) = 0$)

[4] In fact, REML is a little better and is what we will use in the running example and accompanying code. I show the simpler ML estimators here for clarity.

Quantitatively pooling the studies

Connection to linear-model notation

- $\widehat{\theta}_i = \mu + \epsilon_i$ (fixed-effects meta-analysis)
- $\widehat{\theta}_i = \mu + \gamma_i + \epsilon_i$ (random-effects meta-analysis)
- $\widehat{\theta}_i = \mu + \gamma_i + \sum_j \beta_j X_{ij} + \epsilon_i$ (random-effects meta-regression)

where X_{ij} is some study-specific covariate j (e.g., omega-3 dose) for the i^{th} study.

Quantitatively pooling the studies

Random-effects meta-analysis

Under the random-effects model assumptions, marginally:
$\widehat{\theta}_i \sim N\left(\mu, \tau^2 + \widehat{\sigma}_i^2\right)$.
The log-likelihood is thus:

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

After a little calculus, the MLEs turn out to be:

$$\widehat{\mu} = \frac{1}{\sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2}} \times \sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2} \widehat{\theta}_i$$

$$\widehat{\tau}^2 = \max\left\{0, \frac{\sum_{i=1}^{k} \left(\frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2}\right)^2 \left(\left(\widehat{\theta}_i - \widehat{\mu}\right)^2 - \widehat{\sigma}_i^2\right)}{\sum_{i=1}^{k} \left(\frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2}\right)^2}\right\}$$

Quantitatively pooling the studies

Random-effects meta-analysis

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2} \widehat{\theta}_i}{\sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \widehat{\sigma}_i^2}} = \frac{\sum_{i=1}^{k} w_i \widehat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

Notice that this is just another inverse-variance-weighted average, except **now each study's (marginal) variance includes the heterogeneity**, $\widehat{\tau}^2$:

$$w_i = 1/(\widehat{\tau}^2 + \widehat{\sigma}_i^2)$$

Because, again, we assumed:

$$\widehat{\theta}_i \sim N\left(\mu, \tau^2 + \widehat{\sigma}_i^2\right)$$

Quantitatively pooling the studies

Random-effects meta-analysis

The MLEs $\widehat{\mu}$ and $\widehat{\tau}^2$ depend on each other, so in practice, software will estimate these MLEs by iterating between them.

Important: The standard error (and hence, p-value and CI) for $\widehat{\mu}$ should be adjusted using the Knapp-Hartung adjustment.[5]

[5] This helps account for uncertainty in estimating $\widehat{\tau}^2$.

Quantitatively pooling the studies

Random-effects meta-analysis

Running example:

```
> library(metafor); library(MetaUtility)
> ( modRE = rma.uni( yi = d$yi, vi = d$vi, method = "REML", knha = TRUE ) )

Random-Effects Model (k = 38; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0.1369 (SE = 0.0502)
tau (square root of estimated tau^2 value):      0.3699
I^2 (total heterogeneity / total variability):   71.36%
H^2 (total variability / sampling variability):  3.49

Test for Heterogeneity:
Q(df = 37) = 105.4651, p-val < .0001

Model Results:

estimate      se    tval    pval   ci.lb   ci.ub
  0.2755  0.0822  3.3531  0.0019  0.1090  0.4420  **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> MetaUtility::tau_CI(modRE)
[1] 0.1960740 0.4850527
```

Omega-3 supplementation is associated **on average across studies** with decreased depression symptoms by $SMD = 0.28$ (95% CI: [0.11, 0.44]). The estimated SD of effects across studies is $\widehat{\tau} = 0.37$ (95% CI: [0.20, 0.49]).

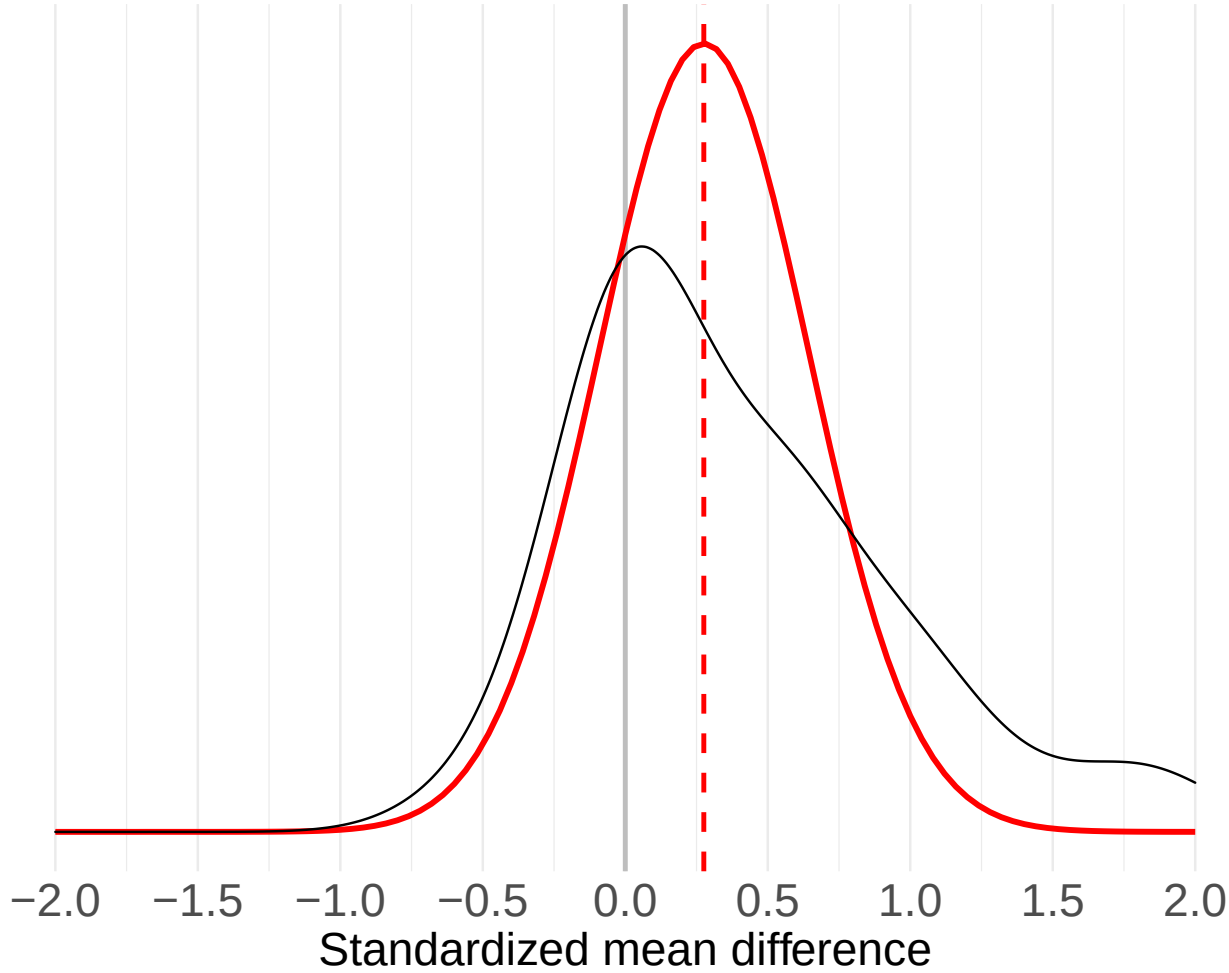


Figure: Omega-3 example. Heavy red curve: Estimated distribution of **population effects** across studies (i.e, normal distribution with mean $\widehat{\mu} = 0.28$ and standard deviation $\widehat{\tau} = 0.37$). Thin black curve: Estimated distribution of **estimates**.

Reporting on heterogeneity

Many meta-analyses report $\widehat{\mu}$ and its p-value, and pretty much stop there.

But it does not really make sense to summarize a heterogeneous distribution in terms of only its mean.

Some quantitative attention to, and reporting on, heterogeneity itself is needed anytime $\widehat{\tau} > 0$.

Reporting on heterogeneity

At minimum, report $\widehat{\tau}$ itself.

Can also report I^2, the proportion of overall variability across studies (heterogeneity + statistical error) that is due to heterogeneity. **But this should not be used as a standalone heterogeneity metric.**[6]

[6] Borenstein et al. (2017)

Reporting on heterogeneity

It's useful to supplement $\widehat{\tau}$ (and, optionally, I^2) with other metrics that summarize the heterogeneous distribution more intuitively:

95% prediction interval: An interval that contains the population effect of a randomly drawn new study (i.e., a draw from the red distribution above) with 95% probability.[7]

- Running example: [-0.49, 1.39]
- If we conducted a new study that was "drawn" from the same heterogeneous distribution as in the meta-analysis, then with 95% confidence, its **population** effect will be between $SMD = -0.49$ and 1.39.
- That's a pretty big range, even including detrimental effects of omega-3 on depression!

[7] Riley et al. (2011)

Reporting on heterogeneity

It's useful to supplement $\widehat{\tau}$ (and maybe I^2) with other metrics that summarize the heterogeneous distribution more intuitively:

Percentage of meaningfully strong effects: Can choose a threshold representing a meaningfully strong effect size (e.g., $SMD > 0.2$) and estimate $\widehat{P}_{>q}$, the percentage of effects above that threshold.[8]

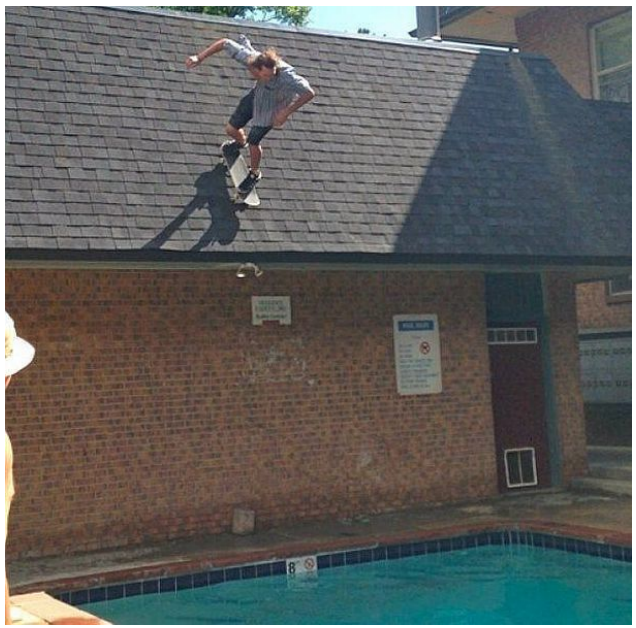
- Running example: 50% (95% CI: [21%, 63%]) of effects are above $SMD = 0.2$.
- We can also estimate the proportion of meaningfully strong detrimental effects ($SMD < -0.2$): 3% (95% CI: [0%, 8%])
- About half of effects are meaningfully strong and beneficial, but very few are meaningfully strong and detrimental.

[8] Mathur & VanderWeele (2020a, 2019)

Caveat about heterogeneity

Random-effects meta-analysis and the heterogeneity metrics above provide principled **statistical** means to account for heterogeneity.

But we still need to think on **substantive** grounds about what kinds of studies are similar enough conceptually to pool them. Don't pool recklessly.



Bias in meta-analyses

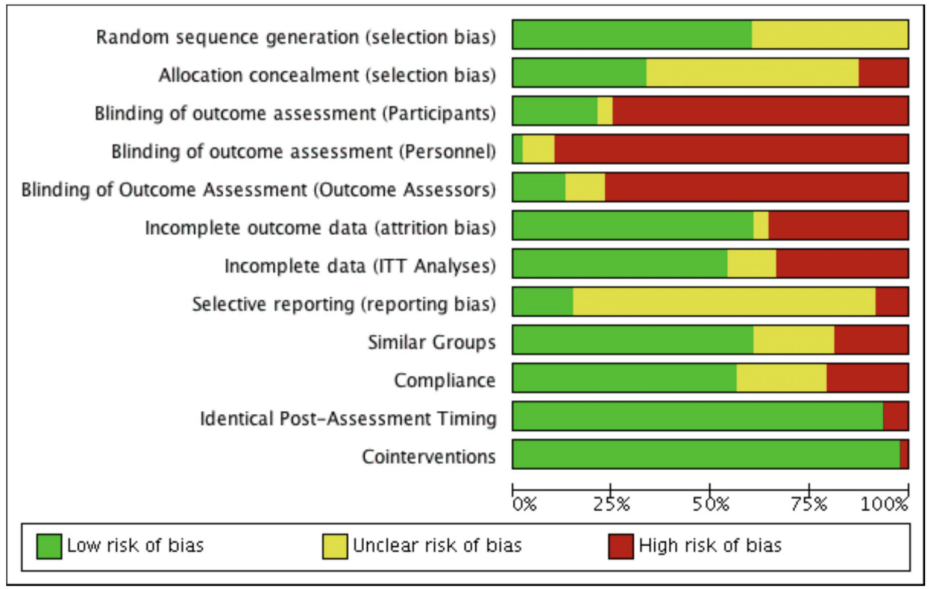
Meta-analyses can be affected by two kinds of bias:

- Biases within studies (e.g., confounding)
- Biases across studies (e.g., publication bias)

Within-study bias in meta-analyses

As noted earlier, should strive to prevent as much within-study bias as possible when designing inclusion criteria. But some within-study bias is usually inevitable, especially with non-randomized studies.

There are excellent tools for qualitatively **rating each study's risks of bias**, such as ROBINS-I. All meta-analyses should include some form of risk-of-bias ratings.



Sterne et al. (2016); "Doing meta-analysis in R", Mathias Harrer

Within-study bias in meta-analyses

It is also often useful to do quantitative sensitivity analyses that characterize how robust the meta-analysis is, holistically, to specific within-study biases.

This area is still nascent, but some methods exist (e.g., E-value analogs[9]).

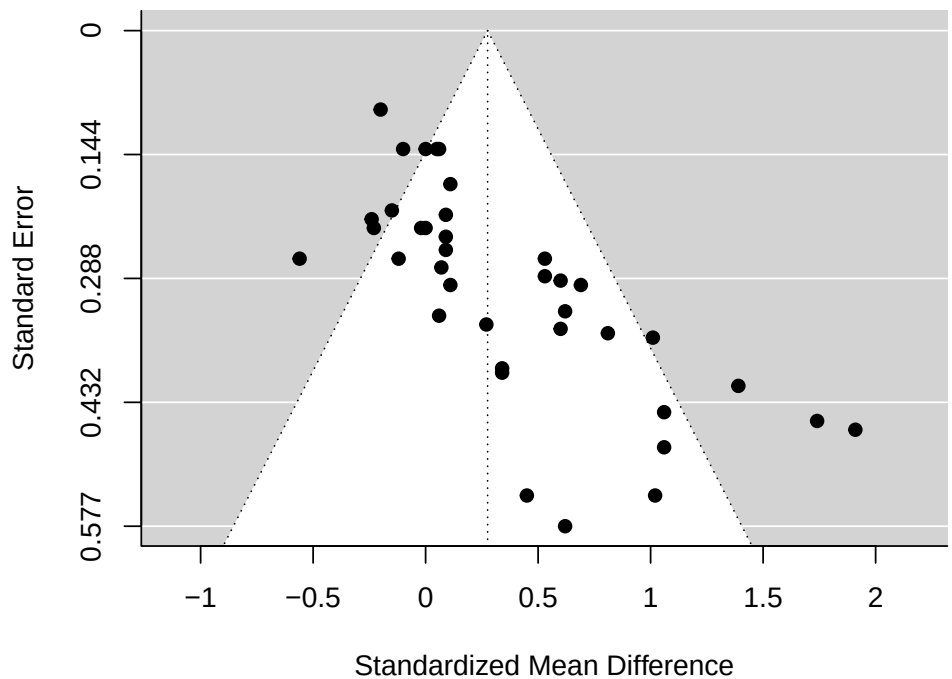
[9] Mathur & VanderWeele (2020a,b)

Across-study biases in meta-analyses

Funnel plots

Publication bias (i.e., selective publication of positive or "significant" results) can be a major threat to meta-analyses.

The most common method to assess for publication bias is the **funnel plot**, depicting studies' estimates vs. their standard errors.

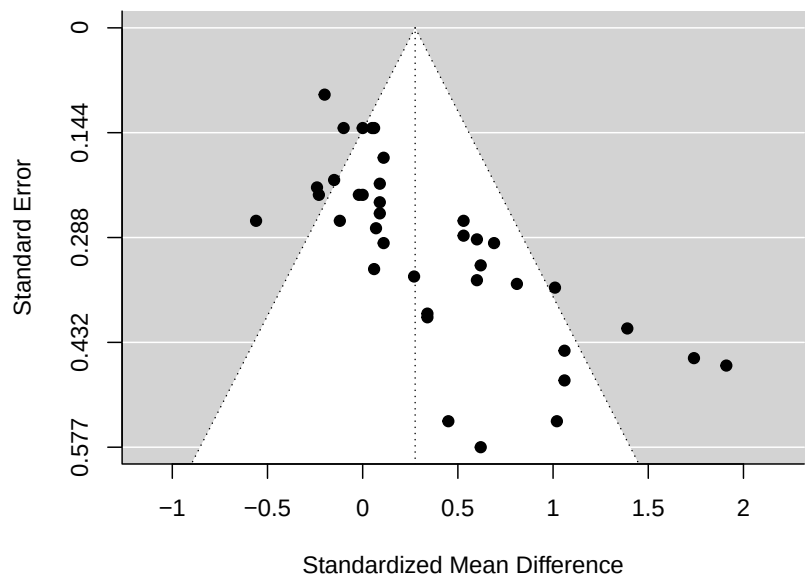


Across-study biases in meta-analyses

Funnel plots

The idea: Without publication bias, estimates should scatter randomly around $\widehat{\mu}$ for any given standard error, resulting in a symmetric funnel shape.

But if there is publication bias that **favors large positive estimates** and **affects small studies** more strongly than large ones, then we might expect the plot to be asymmetric.



Across-study biases in meta-analyses

Funnel plots

There are statistical methods to detect and correct for funnel plot asymmetry (e.g., Egger's test, Trim-and-Fill).

But funnel plots and these related methods have **major limitations** when used to detect publication bias:

- They may not detect publication bias that operates on statistical "significance" and/or that affects all studies regardless of size!
- They assume no heterogeneity.

Funnel plots and related methods **should not be the sole means** of assessing publication bias (and should be described as assessments of small-study effects, not publication bias).

Across-study biases in meta-analyses

Selection models

Selection models[10] assume, arguably more realistically, that publication bias favors "significant" positive results and affects all studies regardless of size.

The simplest selection models assume that **significant positive studies are more likely to be published** than nonsignificant or negative studies by some unknown ratio. The models then estimate that ratio as well as bias-correct $\widehat{\mu}$.

Related sensitivity analyses[11] characterize how severe publication bias would need to be to explain away the effect and provide an estimate corrected for hypothetical worst-case publication bias.

[10] e.g., Vevea & Hedges (1995); McShane et al. (2016)

[11] Mathur & VanderWeele (2020b)

Important but not covered here

Dependent effects: When papers contribute multiple estimates, meta-analysis should accommodate correlation of estimates within papers.[12]

Effect-size conversions: Studies often report different effect-size measures, which may need to be converted to a comparable scale (e.g., correlations to $SMDs$). There are many bad methods to do these conversions. Best overall reference so far is the Cochrane Handbook.[13]

[12] e.g., Hedges et al. (2010); Pustejovsky & Tipton (2020); Berkey et al. (1998)

[13] Higgins et al. (2019)

Summary

Summary

A well-conducted evidence synthesis can inform a balanced and appropriately evolving view of the scientific evidence on a topic.

Systematic reviews use rigorous search criteria to try to identify all relevant literature on a topic. Meta-analyses go a step farther by quantitatively pooling the studies' results.

Meta-analyses pool results by giving more precise studies more weight in the analysis.

By focusing on pooling effect sizes, meta-analyses can help reduce p-value fetishism, especially when there is also careful attention to summarizing heterogeneity.

Take-home resources and references

Some helpful resources

Code to reproduce the examples in this slide deck: `https://osf.io/n9frv/`

The Cochrane Handbook is an excellent, highly detailed (and free!) online guide to conducting evidence syntheses.[14]

DistillerSR is my favorite software for managing the search process.

R packages `metafor` and `robumeta` are my favorite meta-analysis tools.

One example of well-conducted and well-reported meta-analysis is Paluck et al. (2019).

[14] Higgins et al. (2019)

References

Appleton, K. M., Rogers, P. J., & Ness, A. R. (2010). Updated systematic review and meta-analysis of the effects of n- 3 long-chain polyunsaturated fatty acids on depressed mood. *The American Journal of Clinical Nutrition*, *91*(3), 757–770.

Berkey, C., Hoaglin, D., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, *17*(22), 2537–2550.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: i^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*(8), 1336–1342.

Mathur, M. B., & VanderWeele, T. J. (2020a). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, *31*(3), 356–358.

Mathur, M. B., & VanderWeele, T. J. (2020b). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749.

Paluck, E. L., Green, S. A., & Green, D. P. (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy*, *3*(2), 129–158.

Pustejovsky, J. E., & Tipton, E. (2020). Meta-analysis with robust variance estimation: Expanding the range of working models.

Riley, R. D., Higgins, J. P., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., ... others (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435.