Supplement: Methods to Address Confounding and Other Biases in Meta-Analyses

Contents

1. Supplemental Tables 1A-1B: Summary of two-stage and one-stage methods

Method	Summary	Other forms of bias accommodated	Required sensitivity parameters or estimates of confounding severity	Required data granularity*	Software	Distinctive assumptions†	Advantages	Disadvantages
Greenland & O'Rourke (12)	Adjust each study's estimate and variance using information on confounding severity from a comparator external study	–	For each study, the fully- and partially-adjusted estimates and variances from a comparator study	Study-level	–	Fully adjusted estimate in external study must be unbiased; confounding strength must be identical in external study's partially adjusted estimate and in study to be adjusted	Confounding severity is estimated rather than speculated; extends to any form of external adjustment method	Finding a comparator study (with both fully- and partially-adjusted estimates) for each meta-analyzed may be difficult; may under-adjust if comparator studies' "fully adjusted" estimates are still biased
Turner & Spiegelhalter (50)	Subjectively specify range of possible bias in each study on a visual analog scale; use ranges to adjust study's estimate and variance	Any‡	For each study, the subjective range of assumed bias	Study-level	–	Meta-analysts must be able to specify biases accurately	Extends to multiple independent biases	May be difficult for meta-analysts to accurately specify numerical ranges of bias for each study; moments-based meta-analysis estimation can perform poorly
Resche-Rigon et al. (40) (Similar: Audigier et al. (2); Jolani et al. (19))	Using IPD in each study, multiply impute missing confounders using information from studies that do measure the confounders	–	–	IPD	R package ("mice")	All confounders are measured in at least some studies; confounders are missing at random across studies	Confounding severity is estimated rather than speculated; multiple imputation naturally summarizes increases in uncertainty due to unmeasured confounding	Requires extensive IPD that may often be unavailable; may under-adjust if there are confounders that are unmeasured in every study; studies with few measured confounders may have little basis on which to impute unmeasured confounders; multiple imputation can be prone to non-convergence
Goto et al. (11) (Similar: Spiegelhalter & Best (47))	Adjust each study's estimate using analytical bias formula given sensitivity parameters	–	Sensitivity parameters regarding prevalence and confounding associations of a hypothesized confounder	Study-level	–	Unmeasured confounder must be a single categorical variable with known prevalences and confounding associations	Incorporating prevalence information in bias formula can be precise when this information is available; extends to adjustment with any analytical bias formula	Fairly restrictive assumptions on nature of confounding within studies
Manski (27)	Use partial-identification bounds on Y to calculate an interval for the possible causal effect in each study, then pool by taking the intersection of all studies' ranges	Missing data, noncompliance	The lower and upper limits of Y in each study	IPD (usually)	–	Each NRS must be large enough that statistical error is negligible; there must not be heterogeneity across studies	No assumptions on structure or nature of unmeasured confounding; extends to any partial-identification bound	Does not account for finite sample sizes in the NRS or for heterogeneity; intersection of all studies' ranges may often be empty for meta-analyses of more than a few studies; requires bounded Y; often requires IPD

Supplemental Table 1a: *Two-stage sensitivity analysis methods for unmeasured confounding in meta-analyses, listed chronologically. X: exposure variable. Y: outcome variable. –: None. *"Study-level" methods can be applied to a standard meta-analysis dataset consisting of studies' point estimates and variances; "IPD" methods require individual participant data or require statistical estimates at the study level beyond estimates and variances; "meta-level" methods require only summary estimates from the meta-analysis and do not require any study-level data. †Assumptions that are relatively distinctive to the method, excluding assumptions that are common to many methods (e.g., population effects are normal, number of studies is large, bias is independent of studies' standard errors, etc.). ‡Any within-study bias that conforms to the method's assumed scale on which bias affects point estimates (e.g., multiplicatively on the risk ratio scale).*

Method	Summary	Other forms of bias accommodated	Required sensitivity parameters or estimates of confounding severity	Required data granularity*	Software	Distinctive assumptions†	Advantages	Disadvantages
McCandless (36) (Similar: Wolpert & Mengersen (61))	Specify hyperpriors on sensitivity parameters and conduct Bayesian meta-analysis	–	Hyperprior means and variances of confounder's association with Y; hyperprior means and variances of its prevalences within strata of X	Study-level	R code for applied example in paper (http://www.sfu.ca/~lmccandl/Meta.r)	Unmeasured confounder must be a single binary variable that does not interact with X and is independent of measured covariates conditional on X; confounder strengths and prevalences must be independent of one another across studies, independent of studies' causal effects, and normal on the log or logit scale across studies	Bayesian approach naturally summarizes uncertainty about severity of confounding	Conditional independence assumption will always be violated when the unmeasured and measured confounders truly do affect X; other fairly restrictive assumptions on nature of confounding within studies; requires assumptions on distribution of bias across studies
VanderWeele & Ding (55) and Mathur & VanderWeele (33): E-value for pooled estimate and confidence interval	Calculate the average confounding associations that would required to explain away the results	–	–	Meta-level	R package ("EValue") and online tool (http://www.evalue-calculator.com/meta/)	–	No assumptions on nature of confounders within studies; relatively few assumptions about distribution of bias across studies; can be conducted with only meta-level data	Summarizes evidence in terms of only the pooled estimate (unlike E-value analogs below)
Mathur & VanderWeele (33): parametric $\widehat{P}_{>q}$	Specify distribution of bias across studies to estimate bias-corrected effect distribution	Any‡	Mean and variance of bias across studies (or mean and proportion of estimated heterogeneity due to bias)	Meta-level	R package ("EValue") and online tool (http://www.evalue-calculator.com/meta/)	Bias must be log-normal across studies and independent of studies' causal effects	No assumptions on nature of confounders within studies; can be conducted with only meta-level data; summarizes evidence in terms of heterogeneous distribution of effects rather than only point estimate	Requires assumptions on distribution of bias across studies
Mathur & VanderWeele (33)§: parametric $\widehat{T}(r,q), \widehat{G}(r,q)$	Calculate the severity of bias or confounding associations that would be required to explain away the results	Any (for $\widehat{T}(r,q)$ only)‡	Variance of bias across studies (or proportion of estimated heterogeneity due to bias)§	Meta-level	R package ("EValue") and online tool (http://www.evalue-calculator.com/meta/)	Bias must be log-normal across studies and independent of studies' causal effects§	No assumptions on nature of confounders within studies; can be conducted with only meta-level data; summarizes evidence in terms of heterogeneous distribution of effects rather than only point estimate	Requires assumptions on distribution of bias across studies
Mathur & VanderWeele (32): nonparametric $\widehat{P}_{>q}$	Specify homogeneous bias severity to estimate bias-corrected effect distribution	Any‡	Single bias factor	Study-level	R package ("EValue") and online tool (http://www.evalue-calculator.com/meta/)	Bias severity must be the same in all studies (except when the metric can interpreted as a conservative estimate)¶	No assumptions on nature of confounders within studies; fewer statistical assumptions about distribution of population effects (can be non-normal or clustered); allows clustering; summarizes evidence in terms of heterogeneous distribution of effects rather than only point estimate	Treats bias as equally severe in all studies (except when the metric can interpreted as a conservative estimate)¶
Mathur & VanderWeele (32): nonparametric $\widehat{T}(r,q), \widehat{G}(r,q)$	Calculate homogeneous bias severity or confounding associations that would be required to explain away the results	Any (for $\widehat{T}(r,q)$ only)‡	Variance of bias across studies (or proportion of estimated heterogeneity due to bias)§	Study-level	R package ("EValue") and online tool (http://www.evalue-calculator.com/meta/)	Bias severity must be the same in all studies (except when the metric can interpreted as a conservative estimate)¶	No assumptions on nature of confounders within studies; fewer statistical assumptions about distribution of population effects (can be non-normal or clustered); allows clustering; summarizes evidence in terms of heterogeneous distribution of effects rather than only point estimate	Treats bias as equally severe in all studies (except when the metric can interpreted as a conservative estimate)¶

Supplemental Table 1b: *One-stage sensitivity analysis methods for unmeasured confounding in meta-analyses, listed chronologically. X: exposure variable. Y: outcome variable. –: None. *"Study-level" methods can be applied to a standard meta-analysis dataset consisting of studies' point estimates and variances; "IPD" methods require individual participant data or require statistical estimates at the study level beyond estimates and variances; "meta-level" methods require only summary estimates from the meta-analysis and do not require any study-level data. †Assumptions that are relatively distinctive to the method, excluding assumptions that are common to many methods (e.g., population effects are normal, the number of studies is large, bias is independent of studies' standard errors, etc.). ‡Any within-study bias that conforms to the method's assumed scale on which bias affects point estimates (e.g., multiplicatively on the risk ratio scale). §: Statements in this row refer to the generalized expressions given in the present paper, which accommodate heterogeneous bias. ¶: Expressions given in this Supplement describe conditions under which these expressions can be interpreted as conservative estimates without the strict assumption of homogeneous bias.*

2. Using the website and R package to analyze the applied example

Here we provide a tutorial on how to use the R package `EValue` (version 4.1.2) and website (`www.evalue-calculator.com/meta`) to re-analyze Kodama et al. (2009)'s meta-analysis on aerobic capacity and mortality, as described in the main text.

2.1. Calculating an E-value for the point estimate and its confidence interval

To calculate E-values, we first fit a standard random-effects meta-analysis[a] to obtain the meta-analysis point estimate and confidence interval. Doing so in R yields the following output, with all estimates given on the log-relative risk scale (Viechtbauer et al., 2010):

```
Random-Effects Model (k = 16; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0.0395 (SE = 0.0258)
tau (square root of estimated tau^2 value):      0.1988
I^2 (total heterogeneity / total variability):   70.96%
H^2 (total variability / sampling variability):  3.44

Test for Heterogeneity:
Q(df = 15) = 38.6857, p-val = 0.0007

Model Results:

estimate      se    tval    pval   ci.lb   ci.ub
  0.5459  0.0668  8.1682  <.0001  0.4034  0.6883  ***
```

Then, in the first tab of the website `www.evalue-calculator.com/meta`, called "Sensitivity analysis for the point estimate", we input the pooled point estimate and confidence interval limits on the relative risk scale. We can obtain these by exponentiating the log-relative risks in the final line of the R output above, yielding a pooled point estimate of $\exp\{0.5459\} \approx 1.73$ with confidence interval bounds of $\exp\{0.4034\} \approx 1.50$ and $\exp\{0.6883\} \approx 1.99$. By default, the website considers shifting the point estimate and lower confidence interval limit to the null, which is usually taken to be 1 for ratio measures (but this default can be modified).

[a] In this meta-analysis, some papers or cohorts contributed multiple point estimates. We use "studies" to refer to the meta-analyzed point estimates. For illustrative purposes and to more closely reproduce the meta-analysts' reported results, we analyzed the dataset using a simple random-effects model estimated by restricted maximum likelihood and with standard errors estimated with the Knapp-Hartung adjustment. However, note that a best-practice meta-analysis would account for the clustering via, for example, robust estimation (Hedges et al., 2010) or multilevel modeling, or a combination (Pustejovsky & Tipton, 2021).

Alternatively, we could use the R package `EValue` by passing estimates from the meta-analysis object, `meta`. See also the standard R documentation and package vignettes for details (Mathur et al., 2018)).

```
install.packages("EValue")
library(EValue)
evalue( est = RR( exp(meta$b) ),
         lo = RR( exp(meta$ci.lb) ),
         hi = RR( exp(meta$ci.ub) ) )

            point    lower    upper
RR       1.726092 1.496936 1.990329
E-values 2.845602 2.359421       NA
```

Throughout this example, the numerical values produced by the R package (and as reported in the main text) differ slightly from those produced by the website. This occurs because we used rounded values as inputs to the website but used exact values as inputs to the R function.

2.2. Estimating the percentage of effects above $RR = 1.1$ and the strength of homogeneous confounding required to reduce the percentage to less than 15%

The sensitivity analyses we conducted above describe evidence strength in the meta-analysis in terms of only its pooled point estimate. To additionally characterize heterogeneity across the studies' true causal effects, we switch to the website tab called "Sensitivity analysis for the proportion of meaningfully strong effects". This tab has two options: "Robust estimation (homogeneous bias across studies)" and "Parametric estimation (allows heterogeneous bias)". The analyses shown in the main text all consider homogeneous bias across studies, so we use the "Robust" tab.

First, we are prompted to upload our meta-analysis dataset as a .csv file. This dataset should contain

at least two columns: one containing studies' point estimates on the log-relative risk scale,[b] and one containing their variance estimates (i.e., squared standard errors). The dataset must have a single header row containing the variable names, like this:

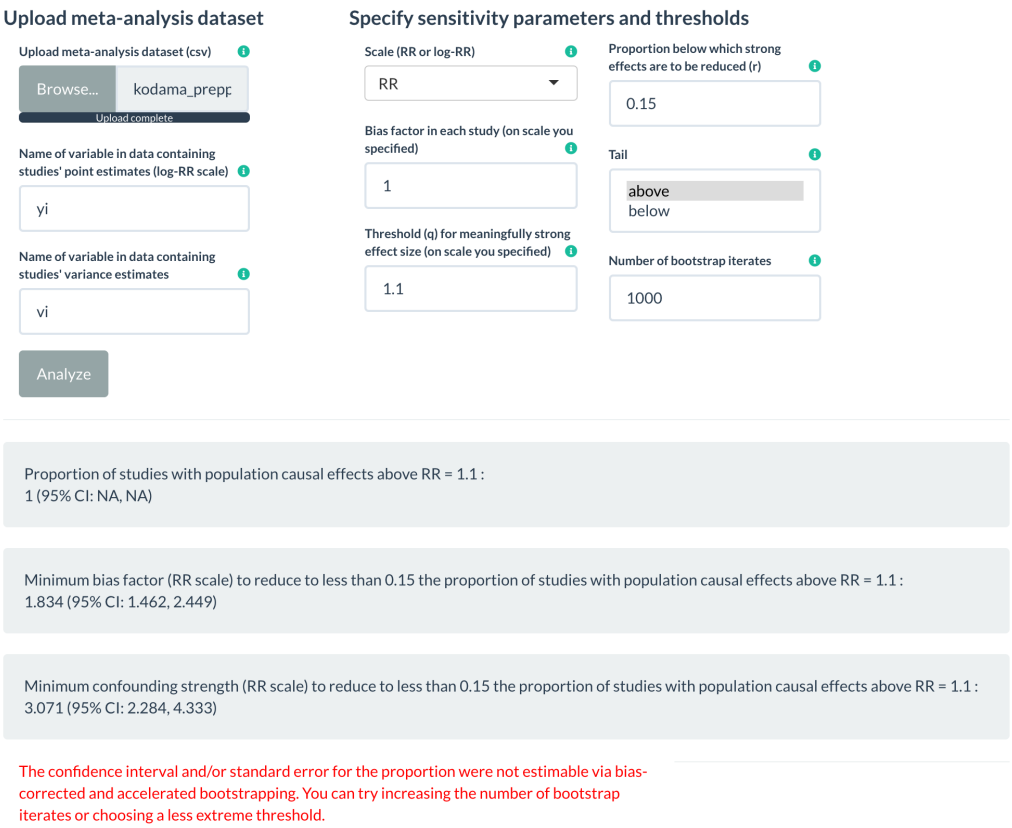
	yi	vi
1	0.385262400790645	0.000731026514084949
2	0.425267735404344	0.00938935228099761
3	0.22314355131421	0.00684180739201936
4	0.198850858745165	0.144504220605036
5	0.65232518603969	0.0233244877241552
6	0.22314355131421	0.0346130484520095
7	0.802001585472027	0.0399853247689023

We upload the dataset to the website and also input the column names for the point estimates and the variance estimates (here, "yi" and "vi").

Next we fill out the section "Specify sensitivity parameters and thresholds". We select a scale (log-relative risk or relative risk) on which to input the bias factor in each study and the threshold we have chosen to represent a meaningfully strong effect size; for this example we use the relative risk scale. To estimate the percentage of studies with meaningfully strong population effects prior to correction for unmeasured confounding, we type "1" for the bias factor in each study (because multiplying a relative risk by a bias factor of 1 would leave it unchanged, representing no bias due to unmeasured confounding).

At the same time, we can also estimate the strength of homogeneous confounding that would be required to reduce this percentage of meaningfully strong effects to less than 15%. To do so, we also type in our chosen threshold of $RR = 1.1$, the proportion below which the proportion of meaningfully strong effects is to be reduced (0.15), and we select the tail of effects we want to consider ("above" because we want to consider effects above, rather than below, $RR = 1.1$). The number of bootstrap iterates is used when estimating confidence intervals and defaults to 1,000, which is what is used in this example.

[b] As indicated in the relevant input box in the website, the studies' point estimates and variances should be provided on the log-relative risk scale regardless of whether the sensitivity parameters are provided on the relative risk or log-relative scale. This is because the meta-analysis should be conducted using estimates on the log scale and because variances of relative risks are usually estimated on the log scale. For meta-analyses that were conducted with other types of effect sizes (e.g., odds ratios with a common outcome or standardized mean differences), the estimates and variances should first be converted to the log-relative risk scale, for example using the function `convert_measures` in the R package `EValue`.



This analysis may take a minute or so to run, after which the metrics we want to estimate appear in the grey boxes. The output in the first box indicates that, prior to correction for unmeasured confounding, we estimate that the percentage of studies with meaningfully strong population effects ($RR > 1.1$) is nearly 100%. There is no confidence interval ("NA") because, per the red message at the bottom of the screen, we have chosen an extremely high threshold compared to the distribution of effects in the meta-analysis, and this can make it impossible to estimate confidence intervals. The output in the second box is not discussed in this tutorial paper, but is described in the instructions of the website. The output in the third box indicates that to bring this percentage of meaningfully strong effects below 15%, homogeneous unmeasured confounding associations with both higher aerobic capacity and lower all-cause mortality of $RR = 3.07$ (95% CI: [2.28, 4.33]) each could suffice, but weaker homogeneous confounding could not.

Alternatively, we could use the R package as follows:

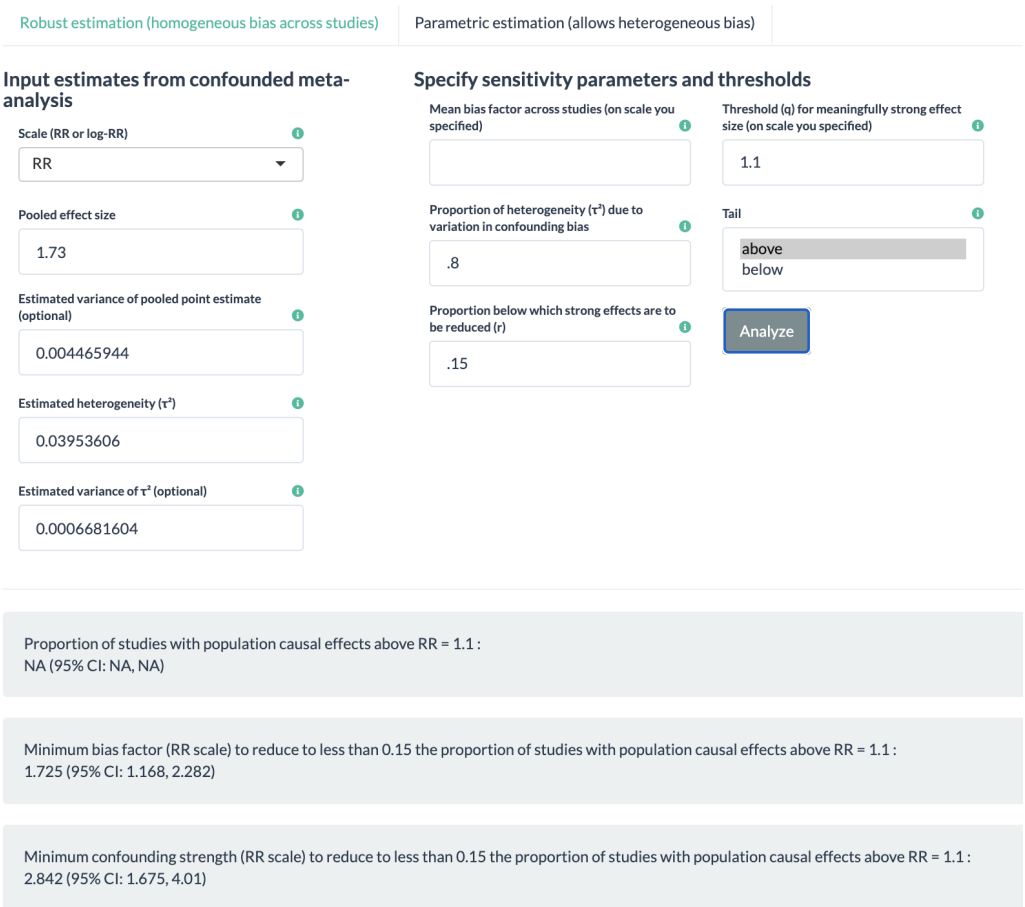
```
confounded_meta( method = "calibrated",
                 q = log(1.1),
                 r = 0.15,
                 tail = "above",
                 muB = 0,
                 dat = dat,
                 yi.name = "yi",
                 vi.name = "vi" )
```

which yields the following output:

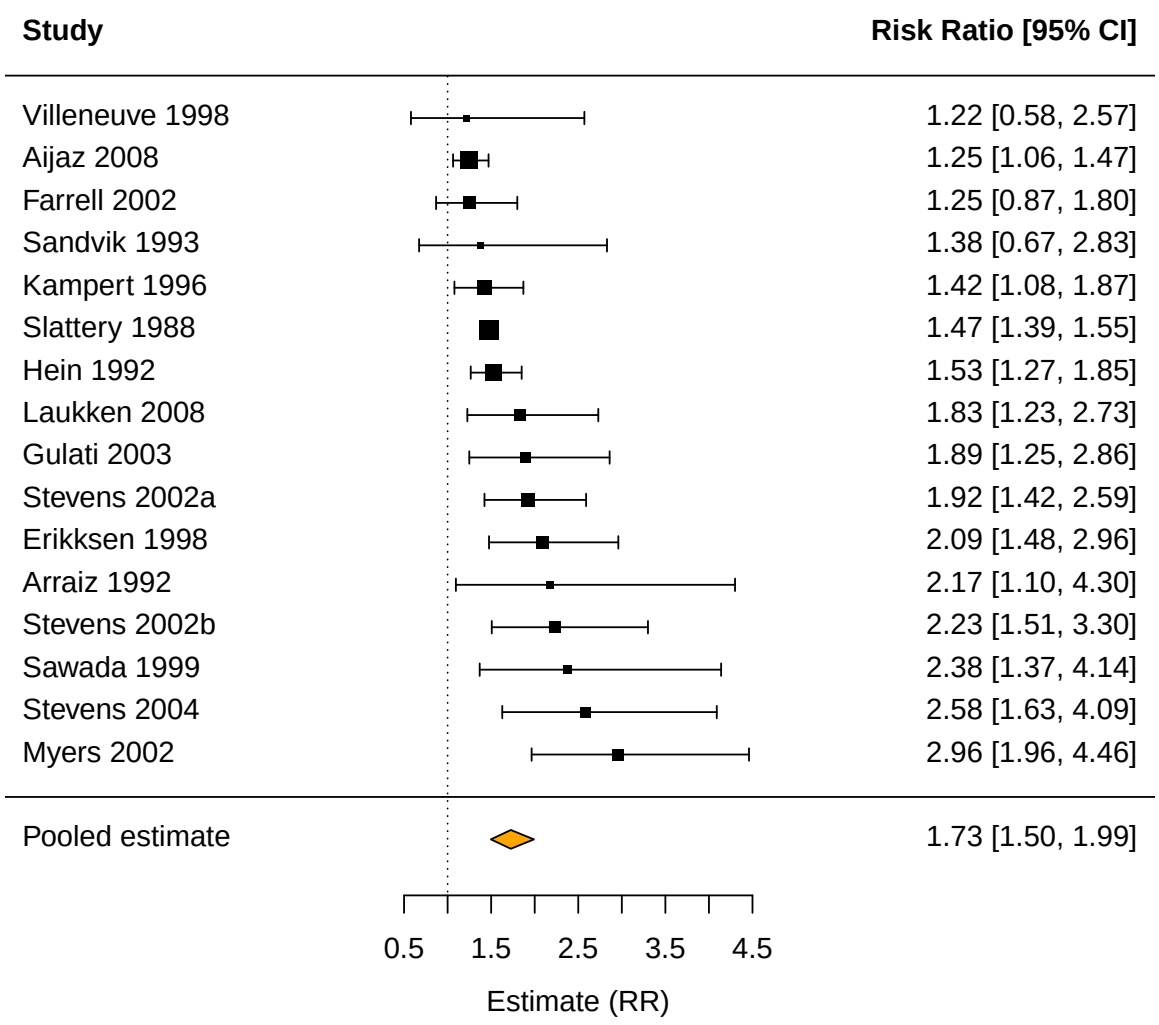
```
The confidence interval and/or standard error for the proportion were not estimable via
    bias-corrected and accelerated bootstrapping. You can try increasing R.
  Value      Est        SE    CI.lo    CI.hi
1  Prop 1.000000        NA       NA       NA
2  Tmin 1.833881 0.2452685 1.490533 2.420839
3  Gmin 3.070504 0.5084525 2.345609 4.275461
```

Because these confidence intervals are constructed by bootstrapping, they differ slightly from those provided on the website. In the main text, we report the confidence intervals given by the R package.

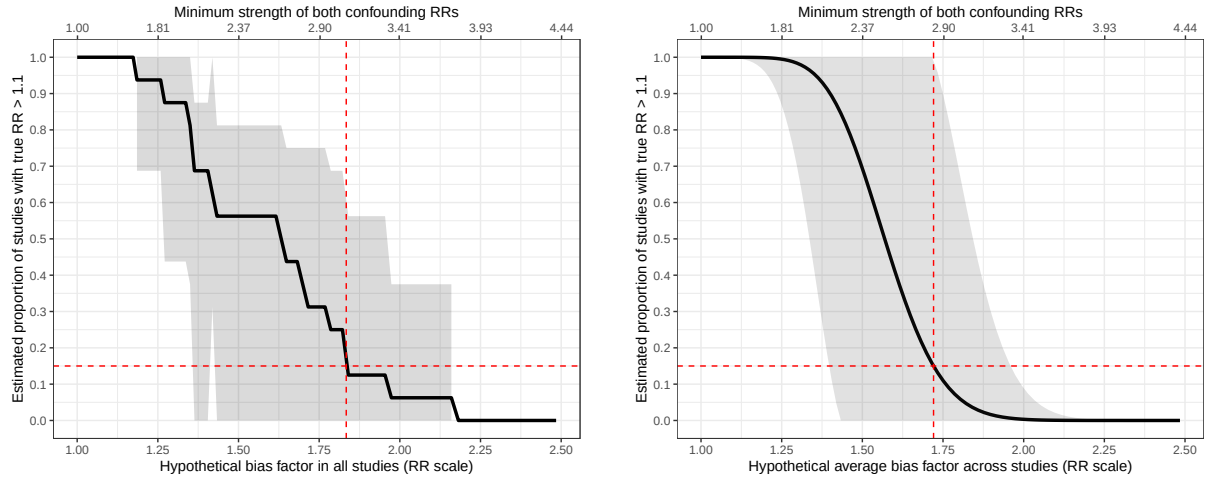
To consider heterogeneous bias as described in the main text, we can use the website tab "Parametric estimation (allows heterogeneous bias)". Here, rather than uploading the meta-analysis dataset, we input 4 estimates from our meta-analysis: the point estimate, its estimated variance (i.e., squared standard error, which can be found in the R output shown in Section 2.1), the estimated heterogeneity (called "`tau^2`" in the R output), and the estimated variance of the heterogeneity estimate. On the right side of the website, we again input the sensitivity parameters as we did in Section 2.2, but now, to characterize our assumed degree of heterogeneity in the confounding bias across studies, we specify what proportion of the observed heterogeneity (of $\widehat{\tau}_c^2 = 0.04$) is assumed to be due to variation in confounding bias rather than to genuine heterogeneity in studies' causal effects. If this proportion is 0, then confounding is assumed to be of homogeneous severity across studies. The closer the proportion is to 1, the more the severity of confounding is assumed to differ across studies.



3. Supplemental figures for the applied example



Supplemental Figure 1: *Forest plot of point estimates (RR) and uncorrected pooled estimate in our reanalysis of Kodama et al. (2009)'s meta-analysis on lower aerobic capacity and mortality. Studies are ordered from largest to smallest point estimate.*



(a) **(b)**

Supplemental Figure 2: *(a) In sensitivity analyses considering homogeneous bias across studies with nonparametric estimation, the estimated proportion of studies with meaningfully strong causal effects ($RR > 1.1$) as a function of the multiplicative bias in all studies (lower x-axis) or, equivalently, the confounding strength in all studies (upper x-axis). Horizontal dashed line: the threshold (r) at which only 15% of effects are meaningfully strong. Vertical dashed line: the confounding strength required to reduce to less than 0.15 the proportion of meaningfully strong effects ($\widehat{G}(r = 0.15, q = \log(1.1))$. (b) Counterpart for sensitivity analyses considering heterogeneous bias across studies. The confidence intervals in this panel are estimated parametrically and may not perform well for values of the proportion that are less than 0.15 or greater than 0.85 (Mathur & VanderWeele, 2020b).*

4. Empirical benchmarks on agreement between nonrandomized and randomized studies in meta-analyses

Although directly estimating the extent of confounding bias in meta-analyzed studies would be very difficult, several studies have addressed this issues indirectly by estimating the extent of agreement or disagreement between NRS and RS on the same topic. Most relevant to our discussion are meta-meta-analyses in which investigators sample existing meta-analyses that contain both NRS and RS, calculate metrics of agreement between the study designs for each meta-analysis, and then summarize these agreement metrics across meta-analyses. If, within a meta-analysis, the estimates from NRS were on average biased upward or downward due to systematic confounding, this would tend to decrease agreement between the two study designs. Critically, though, any such disagreements could reflect not only unmeasured confounding, but also other biases that might preferentially affect one study design (e.g., publication bias) as well as differing distributions of effect modifiers (e.g., populations, interventions, or outcome measures) between the study designs. Therefore, estimates from these meta-meta-analyses should not be interpreted as direct estimates of confounding bias, but rather of the aggregation of confounding bias plus any other systematic differences between study designs.

Supplemental Table 2 summarizes the results of 4 such meta-meta-analyses (Bun et al., 2020; Golder et al., 2011; Shikata et al., 2006; Ioannidis et al., 2001) that used ratio outcomes that were comparable to RRs. When possible, we re-analyzed data (Mathur & VanderWeele, 2020a) to estimate the percentage of discrepancy ratios (i.e., the pooled RR in NRS divided by that in RS) that were

more extreme than various thresholds (e.g., 1.25).[c] In the 2 meta-meta-analyses for which we could obtain study-level estimates to conduct these analyses, about 50% of meta-analyses had discrepancy ratios greater than 1.25 or less than 0.80, but few ($\leq 10\%$) had discrepancy ratios greater than 2 or less than 0.50. Again, these estimates potentially reflect multiple systematic differences between the study designs, not only unmeasured confounding bias.

Outside the context of meta-analyses, other studies have compared NRS to RS while more stringently minimizing systematic differences in populations, interventions, and outcomes. In the ongoing RCT-DUPLICATE project, existing medical RS are being "replicated" in NRS using claims data (Franklin et al., 2020). In the first 10 replications, one of 10 study pairs had a discrepancy with $p < 0.05$ (Franklin et al., 2020), and we estimated that none (0%) of the discrepancy ratios were greater than 1.25 or less than 0.80. These discrepancies appear to be less than seen in the meta-meta-analyses, perhaps partly reflecting extensive confounding control in the NRS conducted by RCT-DUPLICATE. A few studies have even randomly assigned subjects to participate in either an observational study or an RCT, which would eliminate average discrepancies in the characteristics of subjects participating in each study design (Shadish et al., 2008, 2011). We encourage further empirical work comparing NRS to RS, including in meta-analyses and accounting for differing distributions of suspected effect modifiers (Dahabreh et al., 2020; Mathur & VanderWeele, 2021).

[c] Our numerical results and qualitative interpretations sometimes disagree considerably with those reported in the meta-meta-analyses themselves for two reasons. First, some of the meta-meta-analyses reported much larger percentages of extreme discrepancy ratios than those we report. Those analyses had estimated these percentages by simply counting the number of *estimated* ratios that were above or below various thresholds. This method does not account for the substantial statistical error associated with ratio estimates, leading to substantially overestimated percentages of extreme ratios. We instead used methods that correct for this statistical error (Mathur & VanderWeele, 2020a), leading us to estimate smaller percentages of extreme ratios. Second, some meta-meta-analyses concluded that there was little disagreement based on the average discrepancy ratio across meta-analyses, but such results could occur even if discrepancies were extreme but occurred in different directions in different meta-analyses.

	Bun et al. (2020)	Golder et al. (2011)	Shikata et al. (2006)	Ioannidis et al. (2001)
Number of meta-analyses included	102	58	52	45
Type of meta-analyses included	Meta-analyses containing both RS and NRS on the efficacy or safety of a medical intervention	Meta-analyses containing both RS and NRS on adverse effects of health interventions	Meta-analyses of NRS on digestive surgery NRS paired with comparable meta-analyses of RS	Medical meta-analyses containing both RS and NRS, chosen non-systematically
Discrepancy $p < 0.05$	13%	N.R.	19%	16%
Percentages of extreme discrepancy ratios ($RR_{NRS} > RR_{RS}$)				
> 1.1**-fold (%)**		31% [10%, 48%]	46% [0%, 97%]	
> 1.25**-fold (%)**		19% [4%, 34%]	31% [0%, 100%]	
> 2**-fold (%)**		2% [0%, 10%]	2% [0%, 12%]	
Percentages of extreme discrepancy ratios ($RR_{NRS} < RR_{RS}$)				
< 0.91**-fold (%)**		48% [29%, 67%]	33% [0%, 60%]	
< 0.8**-fold (%)**		24% [5%, 43%]	23% [0%, 50%]	
< 0.50**-fold (%)**		2% [0%, 5%]	8% [0%, 19%]	

Supplemental Table 2: *Selected meta-meta-analyses that estimated discrepancies in results between NRS and of randomized controlled trials (RS) with outcome measures comparable to risk ratios.* RR_{NRS}*: pooled risk ratio in the NRS.* RR_{RS}*: pooled risk ratio in the RS. Discrepancy* $p < 0.05$*: Percentage of meta-analyses for which* RR_{NRS} *differed from* RR_{RS} *with* $p < 0.05$*. Percentages of extreme discrepancy ratios (*$RR_{NRS} > RR_{RS}$*): Among meta-analyses with* $RR_{NRS} > RR_{RS}$*, the estimated percentage of ratios* RR_{NRS}/RR_{RS} *that were greater than various thresholds. Percentages of extreme discrepancy ratios (*$RR_{NRS} < RR_{RS}$*): Among meta-analyses with* $RR_{NRS} < RR_{RS}$*, the estimated percentage of ratios* RR_{NRS}/RR_{RS} *that were less than various thresholds. N.R.: Not reported.*

5. Additional sensitivity analysis formulas

5.1. $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$ with log-normal bias

The expressions given the main text for $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$ apply if the confounded pooled estimate is apparently causative ($\widehat{\mu}^c > 0$). If instead the confounded pooled estimate is apparently preventive ($\widehat{\mu}^c < 0$), the expression for $\widehat{G}(r,q)$ in terms of $\widehat{T}(r,q)$ remains the same, but $\widehat{T}(r,q)$ itself becomes:

$$\widehat{T}(r,q) = \exp\left\{q - \widehat{\mu}^c - \Phi^{-1}(r)\sqrt{\widehat{\tau}_c^2 - \sigma_{B^*}^2}\right\} \tag{S.1}$$

$$\widehat{\text{SE}}\left(\widehat{T}(r,q)\right) = \exp\left\{q - \widehat{\mu}^c - \Phi^{-1}(r)\sqrt{\widehat{\tau}_c^2 - \sigma_{B^*}^2}\right\}\sqrt{\widehat{\text{Var}}\left(\widehat{\mu}^c\right) + \frac{\widehat{\text{Var}}\left(\widehat{\tau}_c^2\right)\left(\Phi^{-1}(r)\right)^2}{4\left(\widehat{\tau}_c^2 - \sigma_{B^*}^2\right)}}$$

These expressions, like those given in the main text, are straightforward generalizations of those given in Section 4.2 of Mathur & VanderWeele (2020b) for homogeneous bias.

5.2. Sensitivity analysis with weakened assumptions on the bias distribution

Here we provide a somewhat more technical explanation of how the E-value for the pooled estimate (VanderWeele & Ding, 2017; Mathur & VanderWeele, 2020b) and the nonparametric estimates $\widehat{P}_{>q}$, $\widehat{T}(r,q)$, and $\widehat{G}(r,q)$ (Mathur & VanderWeele, 2020a) can be calculated and interpreted under weakened assumptions about the distribution of bias. For the i^{th} meta-analyzed study, let $\widehat{\theta}_i^c$ denote its confounded estimate, θ_i^t its population causal effect, B_i^* its bias, and ϵ_i its statistical error, such that $\widehat{\theta}_i^c = \theta_i^t + B_i^* + \epsilon_i$. As in the main text, we use the superscript ^{c} to denote confounded estimates and parameters and the superscript ^{t} to denote their true (i.e., causal) counterparts, with $\widehat{\mu}^c$ denoting the confounded pooled log-RR. Let k denote the number of studies in the meta-analysis.

5.2.1 E-value for the pooled estimate and confidence interval

First, as discussed in the main text, the E-value calculated using the confounded pooled estimate is $\exp\left(\widehat{\mu}^c\right) + \sqrt{\exp\left(\widehat{\mu}^c\right)\left(\exp\left(\widehat{\mu}^c\right) - 1\right)}$. All the considerations discussed below also apply when calculating the E-value for the confidence interval limit. The E-value for the pooled estimate and confidence interval can be interpreted without assumptions on the distribution of the population causal effects or the distribution of bias across the population causal effects provided that:

- Condition (i). Any distributional assumptions of the meta-analysis model are fulfilled. This implies that if the population *confounded* effects, θ_i^c, are not normal, the meta-analysis must be conducted using an approach that does not make this distributional assumption (Hedges et al., 2010; Pustejovsky & Tipton, 2021).
- Condition (ii). The bias in each study is independent of its population causal effect, θ_i^t, and its standard error. That is, studies with larger population causal effects cannot have systematically

more or less bias than studies with smaller population causal effects, and more precise studies cannot have systematically more or less bias than less precise studies.

Along with standard assumptions of meta-analysis, these conditions imply that $\widehat{\mu}^c$ is consistent in k for μ^c. This consistency implies that the E-value calculated using $\widehat{\mu}^c$ is itself consistent for the population E-value that would be calculated using the confounded population mean μ^c. Specifically, standard meta-analysis methods based on inverse-variance weighting require studies' population effects to be independent of their standard errors. If we thus make the standard assumption that $\theta_i^t \amalg \epsilon_i$ and additionally assume that Condition (ii) holds (i.e., $B_i^* \amalg \epsilon_i$ and $B_i^* \amalg \theta_i^t$), then these assumptions imply that $\theta_i^c = \theta_i^t + B_i^* \amalg \epsilon_i$. Then, under standard regularity conditions, $\widehat{\mu}^c$ is consistent for μ^c without further assumptions on the distribution of $\theta_i^c = \theta_i^t + B_i^*$, subject to the choice of a meta-analysis model that fulfills Condition (i).

Note that even if θ_i^t are assumed to be normal, the first condition is still necessary to accommodate the possibility of non-normal distributions of bias. In practice, diagnostic plots and tests could be used to check the normality assumption (Hardy & Thompson, 1998; Wang & Lee, 2020). If normality appears to hold, then a standard parametric meta-analysis could still be used. Using parametric estimation facilitates estimating the standard error of $\widehat{\tau}_c^2$, which is used when constructing parametric confidence intervals for $\widehat{P}_{>q}$.

5.2.2 $\widehat{P}_{>q}$, $\widehat{G}(r,q)$, and $\widehat{T}(r,q)$

To estimate $\widehat{P}_{>q}$ under homogeneous bias B^*, the method of Mathur & VanderWeele (2020a) involves calculating a bias-corrected "calibrated" estimate (Wang & Lee, 2019) for each meta-analyzed study, $\widetilde{\theta}_i$, and then calculating the sample proportion of $\widetilde{\theta}_i$ that are above q as follows. Let $\tau_t^2 = \text{Var}\left(\theta^t\right)$ denote the heterogeneity of the population causal effects and $\widehat{\tau}_t^2$ its sample estimate based on the confounded meta-analysis, whose form is given below. Define a generic calibrated estimate as the function:

$$\widetilde{\theta}_i\left(b, v\right) = \widehat{\mu}^c - b + \sqrt{\frac{v}{\widehat{\tau}_c^2 + \widehat{\sigma}_i^2}}\left(\widehat{\theta}_i^c - \widehat{\mu}^c\right) \tag{S.2}$$

Assuming homogeneous bias of magnitude B^*, the calibrated estimates $\widetilde{\theta}_i\left(B^*, \widehat{\tau}_c^2\right)$ approximately match the first two moments of the marginal distribution of population effects (Wang & Lee, 2019). The proportion of meaningfully strong effects can then be estimated as (Mathur & VanderWeele, 2020a):[d]

$$\widehat{P}_{>q}\left(b, v\right) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\left\{\widetilde{\theta}_i\left(B^*, \widehat{\tau}_c^2\right) > q\right\} \tag{S.3}$$

The following lemma and corollary show that, under Conditions (i) and (ii) given in Section 5.2.1, the calibrated estimates could instead be calculated using an estimated lower or upper bound on τ_t^2. The lower bound on τ_t^2 is attained when the variance of the bias ($\sigma_{B^*}^2$) is the most that the meta-analyst

[d] Here, in contrast to Mathur & VanderWeele (2020a) we adopt the more explicit notation "$\widehat{P}_{>q}\left(b, v\right)$" to emphasize the dependence of this estimate on the choice of b and v used to calculate the calibrated estimates.

believes is plausible (called $UB\left(\sigma^2_{B^*}\right)$); the upper bound on τ^2_t is attained when the bias is homogeneous. The following results show that the calibrated estimates calculated using the lower bound on τ^2_t and assuming a mean bias of μ_{B^*} are $\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)\right)$ and are underdispersed compared to the distribution of the causal population effects. Conversely, the calibrated estimates calculated using the upper bound on τ^2_t are $\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}^2_c\right)$ and are overdispersed compared to the distribution of the causal population effects. This leads to estimates $\widehat{P}_{>q}$ that are typically, though not always, overestimates or underestimates depending on whether the bias-corrected pooled estimate is above or below q (**Supplemental Table 3**). These results could be used to calculate estimates, $\widehat{P}_{>q}\left(b, v\right)$, that are considered conservative in the context of whether one is claiming that a meta-analysis is robust or sensitive to unmeasured confounding.

Because $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$ are simply roots in b of the function $\widehat{P}_{>q}\left(b, v\right)$, they can also be interpreted similarly. For example, suppose $\widehat{\mu}^c = \log(2)$ and we estimate the severity of homogeneous bias (i.e., $\sigma^2_{B^*} = 0$, so $\widehat{\tau}^2_t = \widehat{\tau}^2_c$) required to reduce the proportion of causal effects above $q = \log(1.1)$ to $r = 0.15$. We might thus obtain $\widehat{T}(r, q) = 1.5$, which is equivalent to estimating that $\widehat{P}_{>q}\left(\log(1.5), \widehat{\tau}^2_c\right) = 0.15$. With this amount of bias, the bias-corrected mean is $\widehat{\mu}^t = \widehat{\mu}^c - \log(\widehat{T}(r, q)) \approx \log(1.33)$. Thus, $\widehat{\mu}^t > q$, so by consulting the third row and first column of **Supplemental Table 3**, we can conclude that the estimate $\widehat{P}_{>q}\left(\log(1.5), \widehat{\tau}^2_c\right) = 0.15$ would typically *under*estimate the true proportion of meaningfully strong effects if in fact the bias were heterogeneous across studies. Equivalently, our estimate $\widehat{T}(r, q) = 1.5$ is conservative in the sense that if the bias were in fact heterogeneous, then we would typically expect that *at least* 15% of effects would remain meaningfully strong.

	$\widehat{\mu}^t > q$	$\widehat{\mu}^t < q$
$\widehat{P}_{>q}\left(\mu_{B^*}, \widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)\right)$	Overestimate	Underestimate
$\widehat{P}_{<q}\left(\mu_{B^*}, \widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)\right)$	Underestimate	Overestimate
$\widehat{P}_{>q}\left(\mu_{B^*}, \widehat{\tau}^2_c\right)$	Underestimate	Overestimate
$\widehat{P}_{<q}\left(\mu_{B^*}, \widehat{\tau}^2_c\right)$	Overestimate	Underestimate

Supplemental Table 3: *Two methods of calculating $\widehat{P}_{>q}$ and $\widehat{P}_{<q}$ (i.e., the proportion of population causal effects below q) and the conditions under which they are typically overestimates or underestimates compared to the estimate that would be obtained using the true heterogeneity τ^2_t. These results apply regardless of the sign of $\widehat{\mu}^c$ (i.e., whether it is apparently causative or apparently preventive); as in Mathur & VanderWeele (2020b), we consider bias that operates on average away from the null with the convention $\mu_{B^*} > 0$ regardless of the sign of $\widehat{\mu}^c$, such that $\widehat{\mu}^t$ estimates μ^t and is equal to $\widehat{\mu}^c - \mu_{B^*}$ for $\widehat{\mu}^c > 0$ and $\widehat{\mu}^c + \mu_{B^*}$ for $\widehat{\mu}^c < 0$.*

The following lemma and corollary establish the claims made in **Supplemental Table 3**.

Lemma 1 (Underdispersed calibrated estimates)**.** *Suppose $\mu_{B^*} = E[B^*]$ is considered known and that $\sigma^2_{B^*}$ is unknown, but has a known upper bound $UB\left(\sigma^2_{B^*}\right)$. Assume that $Corr\left(\theta^t_i, B^*_i\right) = 0$. If the calibrated estimates are calculated as:*

$$\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)\right) = \widehat{\mu}^c - \mu_{B^*} + \sqrt{\frac{\widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)}{\widehat{\tau}^2_c + \widehat{\sigma}^2_i}}\left(\widehat{\theta}^c_i - \widehat{\mu}^c\right) \tag{S.4}$$

then $E[\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}^2_c - UB\left(\sigma^2_{B^*}\right)\right)]$ approximately matches the expectation of the distribution of population

causal effects, and the calibrated estimates are underdispersed in the sense that $Var\left(\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2 - UB\left(\sigma_{B^*}^2\right)\right)\right)$ *is consistent for a value that is at most* τ_t^2.

Proof. Regarding the first moment, we have $E\left[\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2 - UB\left(\sigma_{B^*}^2\right)\right)\right] = \mu^c - \mu_{B^*}$. Regarding the second moment, first note that $\tau_t^2 = \tau_c^2 - \sigma_{B^*}^2$, so τ_t^2 is bounded by $\tau_c^2 - UB\left(\sigma_{B^*}^2\right) \leq \tau_t^2 \leq \tau_c^2$. (The lower bound, $\tau_c^2 - \sigma_{B^*}^2$, is attained when the bias explains the maximum plausible amount of the observed heterogeneity; the upper bound, τ_c^2, is attained when the bias explains none of the observed heterogeneity.)[e] Let "$\xrightarrow[k\to\infty]{p}$" denote convergence in probability in k. By invoking Conditions (i) and (ii) so that $\widehat{\mu}^c$ and $\widehat{\tau}_c^2$ are each consistent, we have:

$$\begin{aligned}
\text{Var}\left(\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2 - UB\left(\sigma_{B^*}^2\right)\right)\right) &\xrightarrow[k\to\infty]{p} \frac{\tau_c^2 - UB\left(\sigma_{B^*}^2\right)}{\tau_c^2 + \sigma_i^2} \times \text{Var}\left(\widehat{\theta}_i^c\right) \\
&= \tau_c^2 - UB\left(\sigma_{B^*}^2\right) \\
&\leq \tau_t^2
\end{aligned}$$

□

Corollary 1 (Overdispersed calibrated estimates)**.** *Under the same assumptions given above, if the calibrated estimates are calculated as:*

$$\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2\right) = \widehat{\mu}^c - \mu_{B^*} + \sqrt{\frac{\widehat{\tau}_c^2}{\widehat{\tau}_c^2 + \widehat{\sigma}_i^2}}\left(\widehat{\theta}_i^c - \widehat{\mu}^c\right) \tag{S.5}$$

then $E[\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2\right)]$ *approximately matches the expectation of the distribution of population causal effects and the calibrated estimates are overdispersed in the sense that* $Var\left(\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2\right)\right)$ *is consistent for a value that is at least* τ_t^2.

Proof. This follows immediately from the bound $\tau_t^2 \leq \tau_c^2$, attained when $\sigma_{B^*}^2 = 0$. □

The overdispersed calibrated estimates in Eq. (S.5) are equivalent to the calibrated estimates calculated by assuming homogeneous bias, $\widetilde{\theta}_i\left(B^*, \widehat{\tau}_c^2\right)$, when the homogeneous bias B^* is set equal to μ_{B^*}.

References

Bun, R.-S., Scheer, J., Guillo, S., Tubach, F., & Dechartres, A. (2020). Meta-analyses frequently pooled different study types together: a meta-epidemiological study. *Journal of Clinical Epidemiology*, *118*, 18–28.

[e] In principle, the assumption that the population causal effects are uncorrelated with their biases could be relaxed by specifying that correlation or an upper bound on the correlation. Then we have $\tau_t^2 \geq \tau_c^2 - UB\left(\sigma_{B^*}^2\right) - 2\text{Corr}\left(\theta_i^t, B_i^*\right)\sqrt{\widehat{\tau}_c^2 \times UB\left(\sigma_{B^*}^2\right)}$, where the correlation could be replaced by its own upper bound. The calibrated estimates would then be $\widetilde{\theta}_i\left(\mu_{B^*}, \widehat{\tau}_c^2 - UB\left(\sigma_{B^*}^2\right) - 2\text{Corr}\left(\theta_i^t, B_i^*\right)\sqrt{\widehat{\tau}_c^2 \times UB\left(\sigma_{B^*}^2\right)}\right)$.

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, *31*(3), 334–344.

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., ... others (2020). Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. *Circulation*.

Golder, S., Loke, Y. K., & Bland, M. (2011). Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: Methodological overview. *PLoS Medicine*, *8*(5), e1001026.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*(8), 841–856.

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.

Ioannidis, J. P., Haidich, A.-B., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. G., ... Lau, J. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association*, *286*(7), 821–830.

Kodama, S., Saito, K., Tanaka, S., Maki, M., Yachi, Y., Asumi, M., ... others (2009). Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: A meta-analysis. *Journal of the American Medical Association*, *301*(19), 2024–2035.

Mathur, M. B., Ding, P., Riddell, C. A., & VanderWeele, T. J. (2018). Website and R package for computing E-values. *Epidemiology*, *29*(5), e45.

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 unmeasured confounding in meta-analyses. *Journal of the American Statistical Association*, *115*(529), 163–172.

Mathur, M. B., & VanderWeele, T. J. (2021). Meta-regression methods to characterize evidence strength using meaningful-effect percentages conditional on study characteristics. *Research Synthesis Methods*. (Preprint retrieved from https://osf.io/bmtdq/)

Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 1–14.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1344.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*(2), 179.

Shikata, S., Nakayama, T., Noguchi, Y., Taji, Y., & Yamagishi, H. (2006). Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Annals of Surgery*, *244*(5), 668.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, *167*(4), 268–274.

Viechtbauer, W., et al. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wang, C.-C., & Lee, W.-C. (2019). A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*, *10*(2), 255–266.

Wang, C.-C., & Lee, W.-C. (2020). Evaluation of the normality assumption in meta-analyses. *American Journal of Epidemiology*, *189*(3), 235–242.