HHS Public Access

Author manuscript

Annu Rev Public Health. Author manuscript; available in PMC 2022 April 06.

Published in final edited form as:

Annu Rev Public Health. 2022 April 05; 43: 19–35. doi:10.1146/annurev-publhealth-051920-114020.

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

Maya B. Mathur[1], **Tyler J. VanderWeele**[2]

[1]Quantitative Sciences Unit and Department of Pediatrics, Stanford University, Stanford, CA, USA

[2]Department of Epidemiology, Harvard University, Boston, MA, USA

Abstract

Meta-analyses contribute critically to cumulative science, but they can produce misleading conclusions if their constituent primary studies are biased, for example by unmeasured confounding in nonrandomized studies. We provide practical guidance on how meta-analysts can address confounding and other biases that affect studies' internal validity, focusing primarily on sensitivity analyses that help quantify how biased the meta-analysis estimates might be. We review a number of sensitivity analysis methods to do so, especially recent developments that are straightforward to implement and interpret and that use somewhat less stringent statistical assumptions than earlier methods. We give recommendations for how these methods could be applied in practice and illustrate using a previously published meta-analysis. Sensitivity analyses can provide informative quantitative summaries of evidence strength, and we suggest reporting them routinely in meta-analyses of potentially biased studies. This recommendation in no way diminishes the importance of defining study eligibility criteria that reduce bias and of characterizing studies' risks of bias qualitatively.

Keywords

meta-analysis; bias; confounding; observational studies; sensitivity analysis

1. INTRODUCTION

Meta-analyses contribute critically to cumulative science (5, 14), but they can produce biased estimates and misleading conclusions if their constituent primary studies are themselves biased. Nonrandomized studies may be particularly prone to unmeasured confounding, misclassification, selection bias, and other biases. We provide practical guidance on how meta-analysts can address these biases that affect studies' internal validity, first briefly covering approaches for defining study eligibility criteria that reduce bias (Section 2) and for qualitatively assessing studies' risks of bias (Section 3). We then address our primary focus, namely methods for quantitatively assessing how sensitive

mmathur@stanford.edu .

REPRODUCIBILITY

All data, materials, and code required to reproduce the applied example (Section 5) and the re-analysis of empirical benchmarking data (Supplement) are publicly available and documented (https://osf.io/yd3ws/).

meta-analysis results may be to residual bias that cannot be eliminated by limiting study eligibility (Section 4). This last topic has received relatively less attention both in the methodological literature on meta-analysis and in empirical meta-analyses (11), partly because early methods were reasonably criticized for invoking strong statistical assumptions and requiring extensive statistical expertise to implement and interpret (19). However, as we discuss, several recently developed methods have made progress on these fronts – although they do still have limitations – and we therefore advocate for routinely using both qualitative and quantitative methods to assess risks of bias in individual studies and in the meta-analysis as a whole. We illustrate the use of selected quantitative sensitivity analyses in an applied example (Section 5). We focus primarily on meta-analyses whose research questions concern causation and in which the most critical bias is unmeasured confounding, although we comment on other biases throughout, especially in Section 4.3. We do not address publication bias and similar selection processes and so refer to biases affecting studies' internal validity simply as "bias".

2. DEFINING ELIGIBILITY CRITERIA THAT REDUCE BIAS

We recommend that attention to reducing bias in a meta-analysis begin as early as the protocol design stage, during which the eligibility criteria for studies' inclusion in the meta-analysis and in primary versus secondary analyses can be crafted to reduce bias. Preferably, these eligibility criteria, along with the rest of the meta-analysis protocol, should be preregistered formally, with any post hoc deviations disclosed in the final manuscript (19, 37).

2.1. Eligibility criteria for inclusion in the meta-analysis

First, the meta-analyst must decide whether to include non-randomized studies (NRS) at all, and if so, whether to include only certain types of NRS. If an initial scoping review identifies a number of relevant, well-conducted randomized studies (RS) on the topic of interest, limiting eligibility to RS may provide the least biased results and still permit reasonable statistical precision. However, there may be very few (or no) relevant RS, for example because it is not feasible or ethical to randomize the exposure. Alternatively, in some contexts, RS may be available but may be subject to limitations that NRS help mitigate. For example, if the available RS use less externally generalizable samples or shorter follow-up periods than NRS, then including NRS in the meta-analysis may better address the research question (19). When including NRS in a meta-analysis, we recommend that eligibility nevertheless be restricted to study designs that provide reasonably credible evidence given the specific biases that are relevant to a given scientific topic (19); however, when few well-designed NRS are available, deciding how stringent to be can create challenging tradeoffs between bias and precision.

Regarding confounding specifically, NRS are generally least susceptible to bias when they use longitudinal designs with the exposure measured before the outcome and when they control for confounders measured at baseline, ideally including baseline measures of the exposure and outcome themselves.[a] On the other hand, cross-sectional studies that measure the exposure, outcome, and any adjusted covariates contemporaneously are usually quite

prone to confounding because the temporal ordering of the variables is unclear. That is, the direction(s) of causation between the exposure and outcome often cannot be established and the adjusted covariates may not permit adequate control of confounding (53, 58). For this reason, we typically recommend that cross-sectional studies be excluded altogether in meta-analyses whose research questions concern causation, except if the exposure clearly precedes the outcome despite their contemporaneous measurement (e.g., if the exposure is fixed at birth or the outcome is mortality).[b] The Summary below provides a more detailed, but approximate, ranking of NRS study design features by the level of robustness to confounding that they typically provide (adapted with permission from (53)).

SUMMARY: A HIERARCHY OF NONRANDOMIZED DESIGNS FOR CONTROLLING CONFOUNDING

In ascending order of robustness to confounding:

1. Cross-sectional data with exposure and outcome measured contemporaneously
2. Longitudinal data with exposure preceding outcome and control for baseline confounders
3. Longitudinal data with control for baseline confounders and baseline outcome
4. Longitudinal data with control for baseline confounders, outcome, and exposure
5. Longitudinal data using time-varying exposures and confounding control

If the meta-analyst is concerned about biases besides confounding, risk-of-bias tools for NRS can provide guidance on what design features could be used as inclusion criteria (48). When reviewing articles for inclusion, these design features should preferably be assessed not based on study authors' own labels for study designs (e.g., "longitudinal study"), which are defined inconsistently, but rather based on studies' actual design features, such as those in the Summary above (19). Methods to conduct literature searches for NRS are discussed elsewhere (19).

2.2. Eligibility criteria for inclusion in primary analyses

In meta-analyses that include both RS and NRS or that include NRS whose designs provide substantially different levels of robustness to confounding, we recommend pre-specifying

[a]Methods to control for confounders and baseline measures of the exposure and outcome include, for example, adjusting for covariates, using inverse-probability weighting, and using stratification or subset analyses. Specifically regarding the baseline outcome, another method is using within-subject change scores as the outcome in analyses. A common form of subset analysis to control for baseline outcome values is to recruit only individuals who have not yet experienced the outcome (e.g., myocardial infarcation or mortality).

[b]More stringently, meta-analysts could include only cross-sectional studies in which not only (i) the exposure temporally precedes the outcome; but also (ii) the confounders temporally precede the exposure (e.g., the confounders might be age, sex, and childhood socioeconomic status, and the exposure might be adulthood socioeconomic status). In terms of robustness to confounding, these 2 criteria are preferable to criterion (i) alone because covariates measured after the exposure are not structurally confounders, and adjusting for them may not adequately control for confounding (51). However, in practice, longitudinal studies often do not report when adjusted covariates were measured, so it may be difficult to apply criterion (ii) in meta-analyses.

in the meta-analysis protocol which designs will be included in primary and in secondary analyses (19). In general, we recommend analyzing RS and NRS separately, at least in secondary analyses (19). Regarding confounding, if the meta-analyst anticipates that the literature contains very few (if any) RS, primary analyses could be conducted using any available RS plus longitudinal NRS that measure the exposure before the outcome and that control for baseline confounders and the baseline outcome; secondary analyses could then stratify by randomization status using subset analyses or meta-regression methods (17, 19, 35, 39, 49). Similar secondary analyses could be conducted using risk-of-bias ratings, described in Section 3. Additionally, NRS often report estimates that adjust for different sets of confounders, and ideally, the meta-analysis protocol would also pre-specify which of these estimates will be extracted. In general, we recommend that the estimate that adjusts for the largest number of pre-exposure confounders be extracted for primary analyses, but when unadjusted estimates are also available, these could potentially also be extracted for secondary analyses.

3. QUALITATIVE METHODS FOR ASSESSING RISKS OF BIAS

We recommend that meta-analyses of NRS conduct detailed risk-of-bias (ROB) assessments on each study (19). The ROBINS-I tool provides particularly well-informed guidance on the design features that most contribute to risks of confounding and other biases (48); guidance on its use and reporting are provided elsewhere (19). We would suggest that meta-analyses of NRS report on risks of bias in at least 3 ways: (i) for the meta-analysis as a whole, the number and percent of studies occupying each level of the hierarchy given in the Summary box above; (ii) for each study, its summary and domain-specific ROB ratings assessed using ROBINS-I; and (iii) for each study, the list of pre-exposure confounders that were adjusted in the estimate that was extracted for primary meta-analyses.

These methods for detailing each study's risks of bias and design features are integral for meta-analyses of NRS, but it can be challenging to intuit how these individual characteristics contribute to the aggregate bias in the meta-analysis results. A common method to do so, and that required in Cochrane Collaboration reviews, is the GRADE approach (19, 43). In this approach, the meta-analyst first heuristically gauges the "proportion of information" in the meta-analysis that is contributed by studies at low versus high risks of various types of bias (19, 43). Using this heuristic assessment, the meta-analyst can choose to downgrade the overall certainty rating of the meta-analysis results from the default "high certainty" to "moderate", "low", or "very low". At the meta-analyst's discretion, the certainty rating could be upgraded again if the pooled estimate is large (GRADE suggests the criterion of risk ratio > 2 or < 0.5), if there is evidence of dose-response, or if the biases are thought to have attenuated rather than inflated estimates (43).

GRADE and other qualitative approaches to assessing aggregate risks of bias are useful, but have limitations. Intuiting how much "information" each study contributes to the meta-analysis is difficult when studies' standard errors and estimates differ, and considerable subjectivity is involved in deciding how to downgrade or upgrade the overall certainty rating, as the GRADE Working Group discusses (43). Additionally, the GRADE approach ultimately provides a 4-tiered qualitative rating of the overall certainty of the results, rather

than a quantitative summary of how numerical estimates might have been affected by bias. For these reasons, we encourage supplementing these qualitative methods with quantitative methods for assessing the sensitivity of meta-analysis results to bias (Section 4).

4. QUANTITATIVE METHODS FOR ASSESSING SENSITIVITY TO UNMEASURED CONFOUNDING AND OTHER BIASES

Sensitivity analyses are quantitative methods that characterize how numerical estimates might be affected by bias. We classify sensitivity analyses into two-stage and one-stage methods. Two-stage methods first adjust each study's point estimate and potentially also its variance, and then meta-analyze these bias-corrected estimates. In contrast, one-stage methods correct the meta-analysis holistically by specifying the distribution of bias across studies, rather than in each study individually. Below, we primarily discuss conceptually distinct methods that could be used in the context of unmeasured confounding (although many of these methods also accommodate other biases; Section 4.3) and are reasonably straightforward to implement in practice without extensive customization. Supplemental Tables 1A–1B provide additional details on the methods.

4.1. Two-stage methods: Adjusting each study individually before pooling

Two-stage methods begin by adjusting each individual study using any of 4 broad approaches. First, some methods use **subjective elicitation**, in which expert reviewers subjectively specify a numerical value for the severity of bias in each study as well as their own uncertainty in making each judgment (50). Each study's estimate is then corrected using the specified bias, and its variance estimate is inflated to accommodate subjective uncertainty (50).

Second, **external adjustment** methods adjust each study using information from an "external" study, which itself may or may not be included in the meta-analysis. For example, Greenland & O'Rourke (13) proposed adjusting each meta-analyzed NRS using information from a comparably designed external study that reports both an estimate that is thought to be fully adjusted for confounding (and so is unbiased) and a partially adjusted estimate that is subject to the same amount of confounding bias as the meta-analyzed estimate to be adjusted.

Third, methods based on **multiple imputation** are related to external adjustment, but apply in the special context in which the meta-analyst has access to individual participant data for all studies. Under the assumption that at least some studies are fully adjusted and that, in the remaining partially adjusted studies, confounder data are missing at random, individual participants' confounder values are imputed based on relationships among the observed variables, using multilevel models to account for heterogeneity across studies in the joint distribution of the observed variables (2, 20, 40). Each partially adjusted study can then be adjusted based on these imputed confounder values using any standard method for measured confounding, such as regression adjustment or propensity score methods.

Fourth, some methods use **analytical bias formulas** to adjust each study given hypothetical sensitivity parameters regarding the severity and distribution of unmeasured confounder(s).

For example, Goto et al.'s method (12) assumes that each study has a single, categorical unmeasured confounder; under this assumption, each study's estimate can be corrected using 4 sensitivity parameters characterizing the confounder's prevalences among the unexposed and among the exposed subjects as well as the confounder's strengths of association with the exposure and with the outcome (1).[c]

After obtaining bias-corrected estimates for each study using one of the above methods, two-stage methods then proceed to meta-analyze the bias-corrected estimates. Some methods simply conduct a standard meta-analysis in this second stage, without directly modeling additional uncertainty introduced by unmeasured confounding, because they essentially treat any sensitivity parameters used to obtain the corrected estimates as hypothetical fixed values (12). Other methods inflate the variances of the corrected estimates to reflect statistical error associated with the external data (13) or subjective uncertainty associated with subjective elicitation (50). In multiple imputation methods, increases in uncertainty due to unmeasured confounding are naturally captured by the between-imputation variance, which contributes to the final, pooled variance estimate (41). A conceptually unique partial identification approach first bounds each study's causal effect using bounds on the possible values of the outcome variable, then bounds the pooled estimate by taking the intersection of all the studies' bounds (27). This approach is unusual in that it provides an interval rather than a point estimate, assumes there is no effect heterogeneity across studies, and may often yield no interval at all in meta-analyses of more than a few studies.

4.1.1. Advantages and disadvantages.—The key advantage of two-stage methods is that they allow case-by-case adjustment of each study based on extensive information regarding its magnitude of bias. When such information is available, is accurate, and fulfills any necessary statistical assumptions (Supplemental Table 1A), these methods can allow for accurate and precise bias correction. With sufficiently detailed data from fully adjusted studies, some methods have the important advantage of directly and objectively correcting inference for uncertainty introduced by unmeasured confounding (2, 13, 20, 40). Additionally, there is a rich literature on methods to handle confounding and other biases in individual studies (reviewed in (23, 61)), and in principle two-stage methods could use any of these existing methods in the first stage.

However, two-stage methods' reliance on extensive information about each study is also a disadvantage, as this information may often be unattainable for any given study, let alone when meta-analyzing many existing studies. For example, external adjustment and multiple imputation methods require extensive data from fully adjusted studies, and if these "fully adjusted" studies in fact still have residual confounding, the methods may not adjust adequately. Two-stage methods using analytical formulas require fairly detailed information and assumptions about the unmeasured confounder(s), for example that there is a single, categorical unmeasured confounder with known prevalences (12). Of the two-stage methods, those using subjective elicitation require perhaps the fewest "inputs", but at the

[c]In practice, Goto et al. (12) in fact specified the same 4 sensitivity parameters for all studies, but the method would naturally allow specification of different sensitivity parameters for each study.

cost of relying critically on expert reviewers' ability to numerically estimate the severity of confounding bias in each study (50).

4.1.2. Software.—Multiple imputation methods can be implemented using well-established R packages (4, 20). To our knowledge, no software is available for the other methods, but all would be straightforward to implement in any command language for statistical analysis (e.g., R or SAS) by coding a few lines of analytical bias formulas.

4.2. One-stage methods: Adjusting the meta-analysis holistically after pooling

One-stage methods occupy 2 broad categories: bias correction methods and E-value analog methods (Supplemental Table 1B).

4.2.1. Bias correction methods.—These methods specify a hypothetical distribution of bias across studies and then obtain a bias-corrected pooled estimate or distribution of effects. McCandless et al. (36) proposed a Bayesian approach in which log- and logit-normal hyperpriors are specified on the distribution across studies of 3 sensitivity parameters: (i) the unmeasured confounder's strength of association with the outcome, conditional on the exposure and on any measured confounders; (ii) the confounder's prevalence among the unexposed group; and (iii) the confounder's prevalence among the exposed group. Assuming that, across studies, the sensitivity parameters are independent of one another and of studies' causal population effects, McCandless et al. (36) then obtained a bias-corrected likelihood and posterior for the meta-analysis by arithmetically correcting studies' estimates using these 3 sensitivity parameters. Critically, the bias formula they used to do so assumes that each study has a single, binary unmeasured confounder that does not interact with the exposure and that is independent of any measured confounders, conditional on the exposure (25). The latter assumption is highly problematic because it is in fact always violated when the measured and unmeasured confounders affect the exposure (18, 52); thus, while we do not recommend applying this method as-is, the general Bayesian approach could be adapted to use other bias formulas that obviate this assumption. Unlike two-stage methods that require the meta-analyst to specify sensitivity parameters for each study individually, the Bayesian framework requires the meta-analyst to specify only the means and variances of the sensitivity parameters' hyperpriors across studies.

Another method considers confounding bias that is additive on the scale on which studies' estimates are meta-analyzed (e.g., the log-risk ratio scale) and that is assumed to be distributed normally across studies, again independently of studies' causal population effects (34). This method characterizes evidence strength in the meta-analysis in terms of the proportion $\left(\hat{P}_{>q}\right)$ of causal population effects that are meaningfully strong, defined as effects above a threshold (q) that the meta-analyst has chosen to represent a meaningfully strong causal effect in the scientific context (e.g., risk ratio [RR] = 1.1 or some other threshold).[d] (For meta-analyses with pooled estimates in the apparently preventive direction, meaningfully strong causal effects could be defined as those *below* a threshold, such as

[d]Mathur & VanderWeele (31) discussed a number of methods to choose these thresholds, which included considering the size of discrepancies between naturally occurring groups of interest, effect sizes produced by well-evidenced interventions, cost-effectiveness analyses, or minimum subjectively perceptible thresholds.

$RR = 0.90$.) Additionally, the meta-analyst can estimate the proportion of effects below a second, possibly symmetric, threshold in the opposite direction from the pooled estimate. These proportion metrics were recently introduced in the general context of random-effects meta-analysis in order to better convey evidence strength across heterogeneous effects than the pooled estimate alone (31).[e]

To bias-correct the proportion of meaningfully strong causal effects, Mathur & VanderWeele assumed that the additive bias is log-normal across studies (34). The meta-analyst would specify as sensitivity parameters the mean and variance across studies of these biases. Mathur & VanderWeele discuss methods to choose these parameters (34); for example, the variance of the biases could be calculated by first specifying the proportion of the confounded heterogeneity estimate $\left(\hat{\tau}_c^2\right)$ that is in fact due to heterogeneous bias. The metric $\hat{P}_{>q}$ can then be estimated using simple arithmetic expressions involving these sensitivity parameters along with estimates from the confounded meta-analysis (34). Comparable nonparametric methods can estimate $\hat{P}_{>q}$ without making the usual assumption in meta-analysis that the causal population effects are normal (33) or independent (35), and they provide inference that performs better in small meta-analyses or those with extreme true proportions. These methods specify a single fixed value for the bias in all studies ("homogeneous bias"). In some cases, assuming homogeneous bias yields a conservative estimate: for example, if the bias-corrected mean estimate is greater than the threshold q, then estimating $\hat{P}_{>q}$ under the assumption of homogeneous bias will typically be an underestimate (representing *greater* sensitivity to unmeasured confounding) if in fact the bias is heterogeneous. in the Supplement, we detail the conditions under which the nonparametric estimate $\hat{P}_{>q}$ is conservative, and we provide simple alternative expressions that are conservative under other conditions (e.g., when the bias-corrected mean estimate is less than q).

4.2.2. E-value analog methods.—As described above, bias correction methods specify the severity of bias across studies to obtain a corrected pooled estimate. Conversely, E-value analog methods characterize the severity of bias that would be required, hypothetically, to shift the pooled estimate to the null or to otherwise "explain away" the results of the meta-analysis. These methods are thus similar to the E-value, a recently introduced sensitivity analysis for unmeasured confounding in individual studies that does not require assumptions on the nature of unmeasured confounder(s) (8, 54). This standard E-value represents the minimum strength of association, on the RR scale, that unmeasured confounder(s) would need to have with both the exposure and the outcome, conditional on any measured covariates, to fully explain away the observed exposure-outcome association in an individual study (8, 54). When the confounded estimate in an individual study, $\widehat{RR^c}$, is apparently causative $\left(\widehat{RR^c} > 1\right)$, the E-value is:

[e]For example, these metrics can help identify if: (i) few effects of scientifically meaningful size exist despite a "statistically significant" pooled estimate; (ii) some large effects also exist despite an apparently null point estimate; or (iii) strong effects in the direction opposite of the pooled estimate also regularly occur (31). These metrics can also sometimes help adjudicate apparent conflicts between multiple meta-analyses (24, 30).

$$\text{E-value} = \widehat{RR^c} + \sqrt{\widehat{RR^c}\left(\widehat{RR^c} - 1\right)} \quad 4.1.$$

When instead $\widehat{RR^c} < 1$, one would first take its inverse before applying Eq. (4.1) (8, 54). Additional details on the E-value, including how to apply and interpret it for effect measures other than *RR*s, are discussed elsewhere (8, 54, 55, 57), and reporting guidelines are provided in (57). The same considerations apply for the meta-analysis analogs discussed below. It is critical to note that the E-value and its meta-analysis analogs do not estimate the actual severity of bias, but rather describe a hypothetical severity of bias that could suffice to explain away results. Additionally, the E-value is conservative in that it considers the *maximum* bias that could be generated by a given strength of confounder associations, but actual unmeasured confounders might not generate that much bias (8, 54).

As a simple E-value analog for a meta-analysis, Eq. (4.1) could be directly applied to the pooled estimate transformed to the *RR* scale (34). This E-value analog represents the *average* strengths of association across studies, on the *RR* scale, that unmeasured confounder(s) would need to have with studies' exposures and outcomes in order to shift the pooled estimate to the null. Additionally, one can consider the severity of confounding that would be required to shift the confidence interval for the pooled estimate to include the null; to do so, $\widehat{RR^c}$ in the above expression would simply be replaced with the confidence interval limit closer to the null (34, 54). These metrics do not make assumptions on the distribution of the bias in the confounded population effects, although again, the bias must be independent of studies' standard errors (Supplement).

Like most sensitivity analyses for meta-analyses, this simple E-value analog is limited to characterizing evidence strength only in terms of the pooled estimate and its confidence interval. Other E-value analogs instead characterize evidence strength in terms of the aforementioned proportion of meaningfully strong causal effects, $\widehat{P}_{>q}$ (34). For example, Mathur & VanderWeele (34) proposed a metric, $\widehat{G}(r, q)$, that represents the minimum average strengths of association on the *RR* scale that unmeasured confounder(s) would need to have with both the exposure and the outcome in order to reduce to less than some value r (e.g., 0.15) the proportion of studies with causal population effects stronger than q. The rationale for this approach is that, when effects are heterogeneous, one might define "explaining away" the results of the meta-analysis in terms of substantially reducing the proportion of meaningfully strong effects in this way. Letting $\hat{\mu}^c$ and $\hat{\tau}_c^2$ denote the pooled estimate and heterogeneity estimate from the confounded meta-analysis, σ^2_{B*} denote the across-study variance of the log-normal bias, and Φ denote the normal cumulative distribution function, the metric $\widehat{G}(r, q)$ can be estimated as follows for a confounded pooled estimate that is apparently causative ($\hat{\mu}^c > 0$ on the log-*RR* scale):[f]

[f]These expressions are straightforward generalizations of those given in (34). That paper had defined $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$ for the case of homogeneous bias $\left(\sigma^2_{B*} = 0\right)$; here we give expressions that accommodate heterogeneous bias.

$$\widehat{G}(r,q) = \widehat{T}(r,q) + \sqrt{\left(\widehat{T}(r,q)\right)^2 - \widehat{T}(r,q)}$$
$$\widehat{\mathrm{SE}}(\widehat{G}(r,q)) = \widehat{\mathrm{SE}}(\widehat{T}(r,q)) \cdot \left(1 + \frac{2\widehat{T}(r,q) - 1}{2\sqrt{\widehat{T}(r,q)^2 - \widehat{T}(r,q)}}\right) \quad 4.2.$$

where

$$\widehat{T}(r,q) = \exp\left\{\boldsymbol{\Phi}^{-1}(1-r)\sqrt{\widehat{\tau}_c^2 - \sigma_{B*}^2} - q + \widehat{\mu}^c\right\}$$
$$\widehat{\mathrm{SE}}(\widehat{T}(r,q)) = \exp\left\{\left(\boldsymbol{\Phi}^{-1}(1-r)\right)\sqrt{\widehat{\tau}_c^2 - \sigma_{B*}^2} - q + \widehat{\mu}^c\right\} \quad 4.3.$$
$$\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\mu}^c\right) + \frac{\widehat{\mathrm{Var}}\left(\widehat{\tau}_c^2\right)\left(\boldsymbol{\Phi}^{-1}(1-r)\right)^2}{4\left(\widehat{\tau}_c^2 - \sigma_{B*}^2\right)}}$$

(The intermediate estimate $\widehat{T}(r,q)$ represents multiplicative bias on the *RR* scale regardless of its origin, discussed further in Section 4.3.) Similar expressions for the apparently preventive case, $\widehat{\mu}^c < 0$, appear in the Supplement.

As for $\widehat{P}_{>q}$, comparable nonparametric methods (33, 35) can estimate $\widehat{G}(r,q)$ in a wider range of settings than is possible using the above parametric methods. The nonparametric methods assume homogeneous bias across studies, but can sometimes be interpreted as a conservative estimate (Supplement).

4.2.3. Advantages and disadvantages.—In contrast to two-stage methods that require extensive information and assumptions about the severity of bias in each study, one-stage bias correction methods require specification of only a small number of sensitivity parameters that characterize the severity of bias across all studies. E-value analog methods require yet fewer, if any, sensitivity parameters to be specified because, conversely to bias correction methods, E-value methods solve for the severity of bias that would have to exist in order to explain away the results of the meta-analysis. This has the advantage of reducing "researcher degrees of freedom" associated with choosing sensitivity parameters post hoc, which might be especially problematic with two-stage methods (44) and with qualitative evidence-grading systems (43). Whereas all two-stage methods require access to at least study-level estimates and variances (Supplemental Table 1A), often precluding analysis by third parties, certain one-stage methods can be conducted using only statistical estimates from the meta-analysis itself, allowing for sensitivity analysis of some published meta-analyses for which study-level data are unavailable (34, 54).

However, by eliminating case-by-case specification of bias parameters, one-stage methods typically introduce assumptions about the distribution of bias across studies that are unnecessary for most two-stage methods. Most one-stage methods either assume that sensitivity parameters are homogeneous across studies or that they are are log- or logit-normal. Diagnostic plots and tests can sometimes be used to rule out severe violations of these assumptions; for example, the assumptions of Mathur & VanderWeele (34) imply that the population confounded effects are normal, so standard normality tests for meta-analyses could be used (16, 59). Nonparametric methods (33) can sometimes be calculated and

interpreted under weakened assumptions on the bias distribution (Supplement). Also, most one-stage methods make the important assumption that the bias in each study is independent of its population causal effect,[g] which could be violated if, for example, study authors who investigate small causal effects tend to adjust for only a few confounders in order to obtain "statistically significant" results. We give some practical guidance for navigating statistical assumptions in Section 4.4.

4.2.4. Software.—McCandless et al.'s (36) method could be implemented by modifying their example R code; as noted in Section 4.2, we consider it critical to use a different bias formula when applying the general Bayesian framework. All other one-stage methods discussed above (33, 34) can be implemented using the website http://www.evalue-calculator. com/meta/, or the R package EValue (28), for which vignettes are available (29). We provide a step-by-step tutorial for using this website and R package in the Supplement.

4.3. Biases other than confounding

Although we have focused on methods that can provide sensitivity analyses at least for unmeasured confounding, some of these methods also naturally accommodate other biases in NRS or RS, such as participant selection, measurement error, missing data, or noncompliance (Supplemental Tables 1A–1B). For example, in principle, meta-analysts could subjectively elicit any type of bias (50). Bayesian methods have been proposed for other biases (47, 60). Methods to handle psychometric artifacts arising from, for example, range restriction of the outcome or imperfect construct validity are detailed elsewhere (42).

One-stage sensitivity analyses for unmeasured confounding (33, 34) could be readily adapted for certain other biases for which expressions equivalent to the E-value are now available for individual studies (selection bias (46), differential measurement error (56), and combinations of these biases with unmeasured confounding (45)). These E-value equivalents represent the severity of bias, in terms of sensitivity parameters that are specific to the bias under consideration, that would be required shift the effect of an individual study to the null. To apply these results for a meta-analysis, the meta-analyst could first estimate $\hat{T}(r, q)$, which represents multiplicative bias on the RR scale regardless of origin, exactly as described above (33, 34), and then could calculate a bias-specific analog to $\hat{G}(r, q)$ by transforming $\hat{T}(r, q)$ using the relevant bias-specific "E-value" expression (45, 46, 56).

4.4. Overall sensitivity analysis recommendations

We recommend that meta-analyses of NRS routinely report one or more sensitivity analyses for unmeasured confounding and potentially for other biases as relevant to the scientific context and study designs. This recommendation in no way detracts from the importance of also implementing the recommendations in Sections 2–3. As noted above, all sensitivity analysis methods make statistical assumptions of varying stringency, and many sensitivity analyses require extensive information characterizing the amount of bias in each study. Additionally, many methods are not yet implemented in software.

[g]However, these methods do *not* assume that the bias is independent of the confounded estimates: naturally, studies with more severe bias may tend to estimate systematically larger or smaller effect sizes.

Given these considerations, one possible practical approach for choosing among the methods is as follows. First, the meta-analyst could calculate and report the following simple E-value analogs: (i) the E-value for the pooled estimate and its confidence interval limit closer to the null, which respectively represent the average severity of confounding across studies that would be required to shift the pooled estimate, and to shift its confidence interval, to the null (34, 54); and (ii) a nonparametric estimate of $\widehat{G}(r, q)$, which represents the severity of homogeneous confounding that would need to be present in each study in order to reduce to less than r the proportion of causal population effects stronger than a chosen threshold, q (33, 34). As discussed in Section 4.2 and the Supplement, the metric $\widehat{G}(r, q)$ is perhaps most informative when it is calculated and interpreted under conservative assumptions, rather than under the strict assumption that the bias truly is homogeneous.

We believe that, as a generic starting point, reporting these simple metrics is reasonable because these metrics apply to a fairly broad range of meta-analyses of NRS: (i) they do not make assumptions about the nature of unmeasured confounder(s) themselves within studies (e.g., the metrics accommodate multiple confounders, non-binary confounders, and confounders that interact with the exposure); (ii) they require no specification of sensitivity parameters; and (iii) they are straightforward to implement using standard study-level data and available software (28, 29). Additionally, these metrics characterize the sensitivity to unmeasured confounding of both the pooled estimate (and its confidence interval) and the proportion of meaningfully strong causal effects; they thus provide a straightforward way to summarize the distribution of causal effects in the meta-analysis in terms of both its mean and its variability. (If the heterogeneity estimate $\hat{\tau}_c^2$ is 0, then the metric $\widehat{G}(r, q)$ would be omitted.) As such, the metrics provide complementary information: depending on the distribution of population effects, the point estimate may be more or less sensitive to confounding than the percentage of meaningfully strong effects.

Given the results of these simple sensitivity analyses, we recommend that the meta-analyst then attempt to assess and report whether it is actually plausible that the meta-analyzed studies are subject to confounding as severe as that indicated by the E-values and by $\widehat{G}(r, q)$. Examples can be found in existing meta-analyses (3, 10, 26). This assessment would be based on substantive knowledge of the exposures and outcomes under consideration and the ROB assessments described in Section 2. For example, more unmeasured confounding would be plausible in a meta-analysis of cross-sectional studies than in an otherwise comparable meta-analysis of longitudinal studies that control for an ample set of baseline confounders, including baseline values of the exposure and outcome (53, 58). Additionally, examining the confounding associations of *measured* confounders with the exposure and outcome, for example from studies that report both adjusted and unadjusted estimates, can also help inform assessments of the plausible severity of *unmeasured* confounding. However, even if measured confounders have strong confounding associations, residual unmeasured confounding above and beyond these strong measured confounders may be considerably less severe. Also, because the E-value considers maximum bias, even if unmeasured confounders do in fact have confounding associations similar in magnitude to the E-value, this does not necessarily mean that these confounders could actually explain away the effect, only that the evidence is less clear. Last, some empirical studies have more

broadly assessed the extent of agreement or disagreement between NRS and RS on the same topic (e.g., those included in the same meta-analysis); we summarize several such results in the Supplement. However, it is critical to note that disagreements between NRS and RS cannot be interpreted as direct estimates of confounding bias, but rather of the aggregation of confounding bias plus any other systematic differences between study designs. Furthermore, the severity of confounding differs across meta-analyses and scientific topics.

In general, we would advise also conducting sensitivity analyses that consider more precise forms of heterogeneous bias across studies. One possible approach is to apply one-stage methods that assume log-normal bias across studies and do not make assumptions about the nature of confounder(s) *within* each study; as discussed above, these methods also characterize the heterogeneous distribution of population effects (34). If the meta-analyst is concerned about a specific, single unmeasured confounder with known prevalences, one-stage Bayesian methods that similarly assume log-normal and logit-normal sensitivity parameters across studies could be adapted (36), again replacing the existing bias formula with one that obviates the problematic conditional independence assumption. If the meta-analyst has access to the specific forms of external data or individual participant data required by two-stage methods (Supplemental Table 1A), then these methods could be applied to obviate distributional assumptions on the bias and to provide potentially more accurate bias-corrected estimates, albeit by introducing different statistical assumptions.

SUMMARY: SENSITIVITY ANALYSIS RECOMMENDATIONS

1. When meta-analyzing NRS, report sensitivity analyses for unmeasured confounding and potentially other biases, even when following the principles in Sections 2–3.
2. As a starting point, consider reporting (i) the E-value for the pooled estimate and its confidence interval; and (ii) the amount of homogeneous confounding, $\hat{G}(r, q)$, that would be required to substantially reduce the proportion of meaningfully strong causal effects.
3. Consider also conducting further one-stage or two-stage sensitivity analyses that accommodate more precisely specified forms of heterogeneous bias.
4. Interpret and report the results of these sensitivity analyses in the context of studies' risks of bias.

5. APPLIED EXAMPLE

We now illustrate the use and interpretation of selected one-stage sensitivity analyses by applying them to a published meta-analysis. Kodama et al. (22) meta-analyzed longitudinal studies that assessed the association of lower versus higher maximal aerobic capacity with all-cause mortality (Supplemental Figure 1). Prior to correction for unmeasured confounding, our replication of their meta-analysis using 16 studies yielded a pooled *RR* of 1.73 (95% confidence interval [CI]: [1.50, 1.99]; $p<0.001$; heterogeneity $\hat{\tau}^c = 0.20$). All studies were longitudinal and adjusted for some, but not all, possible confounders. The

aerobic capacity measure was arithmetically adjusted for average age and sex differences, but many studies did not adjust for other possible confounders, such as smoking, body mass index, physical activity, and underlying diseases. Kodama et al. (22) did not report on whether each study measured confounders at baseline and did not rate studies' risks of bias.

We first conducted the simple sensitivity analyses described in Section 4.4. (The Supplement provides a step-by-step tutorial on how to conduct all analyses described below.) Computing the E-value for the pooled estimate indicated that unmeasured confounder(s) associated with both lower aerobic capacity and higher all-cause mortality by average risk ratios of 2.85-fold each could potentially shift the pooled estimate to the null; average confounding associations of 2.36-fold each could potentially shift the confidence interval to the null. To characterize the heterogeneous distribution of causal effects, we considered effects larger than $RR = 1.1$ to represent meaningfully strong detrimental effects of lower aerobic capacity. We therefore chose $q = \log(1.1)$ because we conducted analyses on the log-RR scale. Prior to correction for unmeasured confounding, we estimated that the percentage of studies with meaningfully strong population effects ($RR > 1.1$) was nearly 100%.

We then estimated that to reduce this percentage to 15%, homogeneous unmeasured confounding RRs with both lower aerobic capacity and higher all-cause mortality of $\widehat{G}(r = 0.15,\ q = \log(1.1)) = 3.07$ (95% CI: [2.35, 4.28]) each could suffice, but weaker homogeneous confounding could not (Supplemental Figure 2A) (33). Although control of confounding in the meta-analyzed studies was quite limited, these sensitivity analyses seem to suggest reasonably robust evidence for effects of aerobic capacity on mortality: it seems somewhat implausible that, above and beyond measured confounding, each study had sufficiently severe unmeasured confounding (e.g., associations of $RR = 2.36$ to 3.07 with both aerobic capacity and mortality) to shift the pooled estimate or its confidence interval to null, or to reduce the percentage of meaningfully strong effects to only 15%.

To supplement these simple sensitivity analyses, we also assessed the sensitivity of these results to unmeasured confounding under the assumption that bias was highly heterogeneous across studies such that it accounted for 80% of the estimated total between-study variance (Supplemental Figure 2B) (34). We estimated that, to reduce the percentage of meaningfully strong causal effects to 15%, unmeasured confounding RRs with both higher aerobic capacity and lower all-cause mortality of on average $\widehat{G}(r = 0.15,\ q = \log(1.1)) = 2.83$ (95% CI: [1.67, 4]) across studies would suffice to explain away the meta-analysis results in this sense, but weaker confounding would not (34). Again, this severity of unmeasured confounding seems somewhat implausible in these longitudinal studies, and thus the conclusion that there is a considerable percentage of studies with meaningfully large effects seems fairly robust to even substantial degrees of heterogeneous unmeasured confounding. This final analysis assumes that the bias was log-normal across studies; diagnostic plots did not suggest any severe violation of this assumption.

6. CONCLUSION

Our overall recommendations have been as follows:

SUMMARY: OVERALL RECOMMENDATIONS

1. Pre-specify study eligibility criteria that reduce risks of bias. Meta-analyses addressing causal questions should usually exclude cross-sectional studies unless the exposure clearly precedes the outcome temporally.
2. Pre-specify which study designs will be included in primary and in secondary analyses. Stratify on designs that provide substantially differing levels of evidence, at least in secondary analyses.
3. Qualitatively characterize risks of bias in terms of, at minimum: (i) the number and percent of studies occupying each level of robustness to confounding; (ii) for each study, its summary and domain-specific ROB ratings using ROBINS-I (48); and (iii) for each study, the list of pre-exposure confounders that were adjusted.
4. Quantitatively assess sensitivity to residual biases and interpret the results in light of the qualitative risk-of-bias assessments.

We have recommended routinely applying quantitative sensitivity analyses on the grounds that they provide informative, relatively objective quantitative summaries of evidence strength that complement more widespread qualitative approaches (Section 3). This recommendation may prove controversial: others have reasonably argued that sensitivity analysis methods require unrealistic statistical assumptions and are difficult for non-statisticians to implement and interpret (19). However, as we have discussed, more recently developed sensitivity analyses relax some – though certainly not all – important assumptions and are straightforward to implement and interpret. We therefore believe that these methods make progress toward resolving these concerns and that the methods, when reported responsibly, can contribute substantially to characterizing the credibility of a meta-analysis.

To further advance this field, several future directions seem particularly impactful. First, it would be valuable to continue **extending quantitative methods**, for example to more flexibly model the propagation of uncertainty in sensitivity parameters to meta-analysis results, to characterize evidence strength using metrics that summarize heterogeneous effect distributions rather than only the pooled estimate, and to further accommodate heterogeneous bias, especially bias that is correlated with the causal population effects. Second, because interpreting sensitivity analyses requires assessing the severity of bias that is plausible in the meta-analyzed studies, it would be valuable to continue **establishing empirical benchmarks** for the actual severity of bias in studies on different topics and of different designs. We have reviewed some such work in the Supplement, but we particularly encourage further developments that more rigorously parse genuine bias from other systematic differences between study designs (6, 7).

Third, **making datasets publicly available** for both original research (with appropriate deidentification) and meta-analyses would resolve a critical and largely unnecessary limiting factor on meta-analysts' ability to handle bias. If individual participant data were routinely available, much more sophisticated quantitative methods with fewer assumptions could be

developed. Meta-analysts themselves often do not make even study-level data available publicly or on request (32, 38), largely preventing third parties from conducting sensitivity analyses except sometimes by a single method (34). Simple policies and incentives by journals can sometimes rapidly improve data availability when ethical (15, 21), with many collateral benefits for the credibility and efficiency of both original research and meta-analyses.

We hope that the methods and recommendations discussed in this review, along with the suggested future directions, will help inform a balanced and nuanced view of the credibility of meta-analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Sebastian Schneeweiss and the other RCT-DUPLICATE authors (9) for providing study-level data for re-analysis (Supplement).

FUNDING

This research was supported by (1) NIH grants R01 LM013866R01 and R01 CA222147; (2) the NIH-funded Biostatistics, Epidemiology and Research Design (BERD) Shared Resource of Stanford University's Clinical and Translational Education and Research (UL1TR003142); (3) the Biostatistics Shared Resource (BSR) of the NIH-funded Stanford Cancer Institute (P30CA124435); and (4) the Quantitative Sciences Unit through the Stanford Diabetes Research Center (P30DK116074).

LITERATURE CITED

1. Arah OA, Chiba Y, Greenland S. 2008. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Annals of Epidemiology 18:637–646 [PubMed: 18652982]
2. Audigier V, White IR, Jolani S, Debray TP, Quartagno M, et al. 2018. Multiple imputation for multilevel data with continuous and binary variables. Statistical Science 33:160–183
3. Baumeister SE, Leitzmann MF, Linseisen J, Schlesinger S. 2019. Physical activity and the risk of liver cancer: a systematic review and meta-analysis of prospective studies and a bias analysis. JNCI: Journal of the National Cancer Institute 111:1142–1151 [PubMed: 31168582]
4. Sv Buuren, Groothuis-Oudshoorn K. 2010. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software :1–68
5. Cumming G. 2014. The new statistics: Why and how. Psychological Science 25:7–29 [PubMed: 24220629]
6. Dahabreh IJ, Petito LC, Robertson SE, Hernán MA, Steingrimsson JA. 2020. Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. Epidemiology 31:334–344 [PubMed: 32141921]
7. Dahabreh IJ, Robins JM, Hernán MA. 2020. Benchmarking observational methods by comparing randomized trials and their emulations. Epidemiology 31:614–619 [PubMed: 32740470]
8. Ding P, VanderWeele TJ. 2016. Sensitivity analysis without assumptions. Epidemiology (Cambridge, Mass.) 27:368
9. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, et al. 2020. Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. Circulation

10. Fu R, Sekercioglu N, Mathur MB, Couban R, Coyte PC. 2020. Dialysis initiation and all-cause mortality among incident adult patients with advanced ckd: A meta-analysis with bias analysis. Kidney Medicine
11. Golder S, Loke YK, Bland M. 2011. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: Methodological overview. PLoS Medicine 8:e1001026 [PubMed: 21559325]
12. Goto A, Arah OA, Goto M, Terauchi Y, Noda M. 2013. Severe hypoglycaemia and cardiovascular disease: systematic review and meta-analysis with bias analysis. BMJ 347
13. Greenland S, O'Rourke K. 2008. Meta-analysis. In Modern epidemiology. Lippincott Williams & Wilkins, 3rd ed.
14. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. 2018. Meta-analysis and the science of research synthesis. Nature 555:175–182 [PubMed: 29517004]
15. Hardwicke TE, Mathur MB, MacDonald K, Nilsonne G, Banks GC, et al. 2018. Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Royal Society Open Science 5:180448 [PubMed: 30225032]
16. Hardy RJ, Thompson SG. 1998. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 17:841–856 [PubMed: 9595615]
17. Hedges LV, Tipton E, Johnson MC. 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 1:39–65 [PubMed: 26056092]
18. Hernan M, Robins J. 1999. Letter to the editor of Biometrics. Biometrics 55:1316–1316 [PubMed: 11315091]
19. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, et al. 2019. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons
20. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. 2015. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice. Statistics in Medicine 34:1841–1863 [PubMed: 25663182]
21. Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, et al. 2016. Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. PLoS Biology 14:e1002456 [PubMed: 27171007]
22. Kodama S, Saito K, Tanaka S, Maki M, Yachi Y, et al. 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 301:2024–2035 [PubMed: 19454641]
23. Lash TL. 2021. Bias analysis. In Modern epidemiology. Wolters Kluwer, 4th ed.
24. Lewis M, Mathur MB, VanderWeele TJ, Frank MC. 2020. The puzzling relationship between multi-lab replications and meta-analyses of the published literature Preprint: https://psyarxiv.com/pbrdk/
25. Lin DY, Psaty BM, Kronmal RA. 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics :948–963 [PubMed: 9750244]
26. Ling S, Brown K, Miksza JK, Howells L, Morrison A, et al. 2020. Association of type 2 diabetes with cancer: A meta-analysis with bias analysis for unmeasured confounding in 151 cohorts comprising 32 million people. Diabetes Care 43:2313–2322 [PubMed: 32910779]
27. Manski CF. 2020. Toward credible patient-centered meta-analysis. Epidemiology 31:345–352 [PubMed: 32079834]
28. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. 2018. Web site and R package for computing E-values. Epidemiology 29:e45–e47 [PubMed: 29912013]
29. Mathur MB, Ding P, VanderWeele TJ. 2020. EValue 4.1.1: Sensitivity analyses for unmeasured confounding in observational studies and meta-analyses. Https://cran.r-project.org/web/packages/EValue/
30. Mathur MB, VanderWeele TJ. 2019. Finding common ground in meta-analysis "wars" on violent video games. Perspectives on Psychological Science 14:705–708 [PubMed: 31188714]
31. Mathur MB, VanderWeele TJ. 2019. New metrics for meta-analyses of heterogeneous effects. Statistics in Medicine 38:1336–1342 [PubMed: 30513552]

32. Mathur MB, VanderWeele TJ. 2020. Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. Research Synthesis Methods
33. Mathur MB, VanderWeele TJ. 2020. Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. Epidemiology 31:356–358 [PubMed: 32141922]
34. Mathur MB, VanderWeele TJ. 2020. Sensitivity analysis for unmeasured confounding in meta-analyses. Journal of the American Statistical Association 115:163–172 [PubMed: 32981992]
35. Mathur MB, VanderWeele TJ. 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/
36. McCandless LC. 2012. Meta-analysis of observational studies with unmeasured confounders. The International Journal of Biostatistics 8:1
37. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, et al. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews 4:1–9 [PubMed: 25554246]
38. Polanin JR, Hennessy EA, Tsuji S. 2020. Transparency and reproducibility of meta-analyses in psychology: A meta-review. Perspectives on Psychological Science 15:1026–1041 [PubMed: 32516081]
39. Pustejovsky JE, Tipton E. 2021. Meta-analysis with robust variance estimation: Expanding the range of working models. Prevention Science :1–14
40. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group PIS. 2013. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Statistics in Medicine 32:4890–4905 [PubMed: 23857554]
41. Rubin DB. 2004. Multiple imputation for nonresponse in surveys, vol. 81. John Wiley & Sons
42. Schmidt FL, Hunter JE. 2015. Methods of meta-analysis: Correcting error and bias in research findings. Sage, 3rd ed.
43. Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, et al. 2019. Grade guidelines: 18. how robins-i and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. Journal of Clinical Epidemiology 111:105–114 [PubMed: 29432858]
44. Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22:1359–1366 [PubMed: 22006061]
45. Smith LH, Mathur MB, VanderWeele TJ. In press. Multiple-bias sensitivity analysis using bounds. Epidemiology Preprint: https://arxiv.org/abs/2005.02908/
46. Smith LH, VanderWeele TJ. 2019. Bounding bias due to selection. Epidemiology 30:509 [PubMed: 31033690]
47. Spiegelhalter DJ, Best NG. 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Statistics in Medicine 22:3687–3709 [PubMed: 14652869]
48. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, et al. 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355
49. Thompson SG, Higgins JP. 2002. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 21:1559–1573 [PubMed: 12111920]
50. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. 2009. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172:21–47
51. VanderWeele T. 2015. Explanation in causal inference: methods for mediation and interaction. Oxford University Press
52. VanderWeele TJ. 2008. Sensitivity analysis: distributional assumptions and confounding assumptions. Biometrics 64:645–649 [PubMed: 18482060]
53. VanderWeele TJ. 2021. Causal inference with time-varying exposures. In Modern epidemiology. Wolters Kluwer, 4th ed.
54. VanderWeele TJ, Ding P. 2017. Sensitivity analysis in observational research: introducing the E-value. Annals of Internal Medicine 167:268–274 [PubMed: 28693043]

55. VanderWeele TJ, Ding P, Mathur M. 2019. Technical considerations in the use of the e-value. Journal of Causal Inference 7

56. VanderWeele TJ, Li Y. 2019. Simple sensitivity analysis for differential measurement error. American Journal of Epidemiology 188:1823–1829 [PubMed: 31145435]

57. VanderWeele TJ, Mathur MB. 2020. Commentary: developing best-practice guidelines for the reporting of e-values. International Journal of Epidemiology 49:1495–1497 [PubMed: 32743656]

58. VanderWeele TJ, Mathur MB, Chen Y, et al. 2020. Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. Statistical Science 35:437–466

59. Wang CC, Lee WC. 2020. Evaluation of the normality assumption in meta-analyses. American Journal of Epidemiology 189:235–242 [PubMed: 31781756]

60. Wolpert RL, Mengersen KL, et al. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. Statistical Science 19:450–471

61. Zhang X, Stamey JD, Mathur MB. 2020. Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. Pharmacoepidemiology and Drug Safety 29:1219–1227 [PubMed: 32929830]